I. J. Computer Network and Information Security, **2015, 12, 42-51**
Published Online November 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2015.12.05

Extended K-Anonymity Model for Privacy Preserving on Micro Data

Masoud Rahimi
Sheikhbahaee University, Isfahan, Iran
Email: m.rahimi@shbu.ac.ir

Mehdi Bateni
Sheikhbahaee University, Isfahan, Iran
Email: bateni@shbu.ac.ir

Hosein Mohammadinejad
Sheikhbahaee University, Isfahan, Iran
Email: mohammadi.n@shbu.ac.ir

Abstract—Today, information collectors, particularly statistical organizations, are faced with two conflicting issues. On one hand, according to their natural responsibilities and the increasing demand for the collected data, they are committed to propagate the information more extensively and with higher quality and on the other hand, due to the public concern about the privacy of personal information and the legal responsibility of these organizations in protecting the private information of their users, they should guarantee that while providing all the information to the population, the privacy is reasonably preserved. This issue becomes more crucial when the datasets published by data mining methods are at risk of attribute and identity disclosure attacks. In order to overcome this problem, several approaches, called p-sensitive k-anonymity, p+-sensitive k-anonymity, and (p, α)-sensitive k-anonymity, were proposed. The drawbacks of these methods include the inability to protect micro datasets against attribute disclosure and the high value of the distortion ratio. In order to eliminate these drawbacks, this paper proposes an algorithm that fully protects the propagated micro data against identity and attribute disclosure and significantly reduces the distortion ratio during the anonymity process.

Index Terms—Privacy preservation, data mining, k-anonymity, micro data.

I. Introduction

The researchers and policy makers` increasing demand for more comprehensive statistical information has led to a paradox. Statistical organizations collect a huge amount of data for statistical purposes. Responders can only satisfy the informational requests of a statistical organization when they are assured that their information is meticulously protected, used for producing general statistics, and their privacy is not at risk. Therefore, statistical data usage by external users should not disclose the individuals` identity and endanger their data privacy.

Subsets of statistics that deal with such issues include statistical disclosure control, statistical data disclosure limitation, statistical data protection, statistical privacy or data anonymity. Data providers usually propagate data in two forms of tables and micro data. Tabular data is the common product of data providers. These tables include categorized data that is propagated in two forms of magnitude and frequency tables [1], whose security is out of the scope of this paper. Another form of data propagation is micro data. The datasets called micro data contain records comprised of a large amount of information about responders. In other words, each record includes several variables for a responder. Depending on who that responder is, included variables in the micro data can be the location, occupation, production, activity type, etc. [2].

Disclosure refers to the inappropriate extraction of the information regarding an individual or an organization. Generally, there are two types of data disclosure: identity disclosure and attribute disclosure [2].

In identity disclosure, which is the most important type of disclosure, first the individual is identified and then based on his identity, his private information is extracted from the data. However, determining the identity is not always necessary condition to disclose the sensitive information about a responder. In some cases, only knowing that a responder is a member of a group, without knowing which one he is, is sufficient to disclose his data. In attribute disclosure, having information regarding the propagated table records is also essential. An instance of this type of disclosure, called attribute disclosure, is as follows. Assume it is known that individual I is in group G and from the propagated data, it is recognized that the revenue of each individual in group G is more than T. we can conclude that the revenue of individual I is more than T.

Data disclosure occurs when an intruder can find the value of a sensitive variable related to the target

Copyright © 2015 MECS

individual after identifying him. A variable is called sensitive if the data propagator does not wish to reveal its value for one or several specific responders to an intruder (for instance the revenue). The responder can be identified using the identifying variable. A variable that is individually or in combination to other variables, used to identify a responder is called an identifying or identifier variable [1].

If the responder can be identified only by this variable, it is called the direct identifying variable (e.g. name, address, etc.) and if the identity is recognized through a combination of variables, they are called, indirect identifying variable (e.g. age, education, location, etc.). Whenever a data provider wants to propagate a micro dataset, he removes the direct identifying variables from the file; therefore, there is no direct approach to determine whether a particular record belongs to a specific responder [1, 2].

For instance, assume the data provider wants to propagate a dataset consisting of micro data including information about location, occupation, and criminal history of responders. In addition, a record appears in the micro dataset in combination with the following values: "location: Isfahan, occupation: Mayor, criminal history: one case".

If it is assumed that the name and the address of the responder is not propagated, many individuals in Iran can recognize who the responder is. They can particularly conclude that the responder, i.e. the mayor of Isfahan, has a criminal history. Any responder having a combination of values that occur infrequently in the population is in danger of being identified.

This research aims to propose an extended model of k-anonymity that preserves the privacy of micro data based on l-diversity and t-closeness algorithms, makes the propagated datasets secure against attribute disclosure, and reduces the distortion ratio. The rest of this paper is organized as follows: in the second section, the previous works and major approaches to preserving the privacy of micro data is presented. Section 3 introduces the main concepts used in the proposed method and section 4 presents the proposed approach. Section 5 investigates the evaluation measures and section 6 tests the proposed method according to these evaluation measures. Finally, section 7 presents the conclusions and future works.

II. T. Defining Basic Concepts

In this section, table T is presented as the initial table and T' is the table consisting of the propagated micro data. T' is comprised of a set of records from the set of attributes. The attributes of micro data is categorized in three groups as follows:

Identifier attributes (direct identifier variables) like first name, last name, and national security number used to identify a record. Since, the goal is to eliminate the connection of sensitive information to specific responders, it is assumed that the identifier attributes of the micro data are removed or encrypted in a preprocessing stage.

Indirect identifier variables or quasi-identifiers (QI) include postal code and age, which are used in combination to other external information to identify the sensitive information of the responders in the micro data table. In contrast to identifier attributes, QI attributes cannot be removed from the micro data table, since each attribute can potentially be a QI attribute. Sensitive attributes, including diseases or revenue, are the ones that require protection and an intruder is not aware to whom they belong.

Next, it is assumed that identifier attributes are removed from the micro data table and quasi-identifier and sensitive attributes usually remain the initial micro data table and the published table. Another assumption is that sensitive variables are not accessible through an external source. These assumptions ensure that an intruder cannot manipulate sensitive attributes to increase his chance in disclosing the responders` identity. Unfortunately, an intruder may exploit the linkage techniques (linking attack) between quasi-identifier attributes and the information in external sources, which were collected from different locations, to identify records and individuals in the micro data table. In what follows, due to their application in the proposed method, some basic concepts are briefly explained:

Definition 1 (k-anonymity): the modified micro data table T' satisfies k-anonymity if and only if each combinations of quasi-identifier attributes occur at least k times in T'.

Definition 2 (p-sensitive k-anonymity): the modified micro data table T' satisfies the p-sensitive k-anonymity constraint if it satisfies the k-anonymity constraint and for each group-identity in T', the number of different values for each sensitive attribute is at least p in a similar group of quasi-identifiers.

Definition 3 (p+-sensitive k-anonymity): the modified micro data table T' satisfies the p+-sensitive k-anonymity constraint if it satisfies the k-anonymity constraint and each group-quasi-identifier in T' has at least p different levels of each sensitive attribute.

Definition 4 ((p, α)-sensitive k-anonymity): the modified micro data table T' satisfies the (p, α)-sensitive k-anonymity constraint if it satisfies the k-anonymity constraint and for each group-quasi-identifier in ((p, α)-sensitive k-anonymity, the number of different values for each sensitive attribute is at least p in each similar group of quasi-identifiers with at least a total weight of α.

Definition 5 (l-diversity): a class is called l-diversity if there is at least 1 value for a sensitive variable in that lass and a table is called l-diversity if each class is an equivalent of that table [11, 12].

Definition 6 (t-closeness): an equivalence class is called t-closeness if the distribution distance between sensitive variables in this class and the sensitive variables in the table is not higher than threshold t; a table is called t-closeness, if all equivalence classes are t-closeness [13].

K-means clustering algorithm is one of the simplest and most well-known unsupervised learning algorithms.

Copyright © 2015 MECS

In K-Means, the dataset is practically divided into predefined clusters. The main idea of this algorithm is defining k centroids for k clusters. The best choice for cluster centroids in k-means is placing them as far from each other as possible. Subsequently, each record of the dataset is assigned to the closest cluster centroid. The drawbacks of this algorithm is as follows:

- The final answer strongly depends on the initial cluster centroids.
- There is no specific procedure to compute the initial cluster centroids.
- If in an iteration, the number of members in a cluster (except for the cluster centroid) becomes zero, there is no way to change and improve the current state.
- In this method, it is assumed that the number of clusters is predetermined; however, in most application, the number of clusters is not known a priori.

One of the solutions to eliminate the problem of selecting initial cluster centroids and finding the optimal k in k-means algorithms is running the algorithm several times. This leads us to x-means clustering algorithm.

X-means clustering algorithm [22] was proposed in 2000 by Andrew Moore and Dan Pelleg, which effectively searches the cluster space and the number of clusters to optimize the values and employs AIS and BIC metrics to measure this optimization. X-means first considers two values to determine the range of K. subsequently, lets K the smallest value in that range and runs k-means. The algorithm proceeds by adding to k until k reaches the maximum number. During this process the set of centroids with the highest score are selected as the output. The stages of x-means are as follows:

1. Improving parameters.
2. Improving the structure.
3. Stopping the algorithm if K> Kmax and reporting the best score, as well as the obtained centroids. Otherwise go to stage 1.

In what follows, the attribute weighting concepts are discussed. The set of attribute weighting operators are used to employ a certain mechanism to specify the value of each attribute in identifying the corresponding group. For instance, if a dataset is applied to this model to determine the diabetic patients, factors like the eye color is insignificant against factors like blood sugar. With appropriate weighting, we can specify such effects in the data. The weighting operator used in this research is called Chi Squared Statistic. This operator computes the relationship of each attribute in the input dataset with the label attribute using Chi-squared method and weights attributes accordingly.

III. Related Works

Currently, the privacy preserving technologies in database applications are mostly focused on data mining and statistical domains. There are several statistical methods in this context that include global recording, local suppression, perturbation, and micro aggregation.

Global recording: when a string variable is recorded in a micro data file, the strings of the variable are transformed into new strings that are less comprehensive and wider [14].

Local suppression: the value of a variable in one or more records is exchanged with a missing value. This is called local since it does not apply to all records, but to the number of records that are recognized as unsafe [14].

Perturbation: this is a general term for different modifications that exchange (not missing) values of a variable with (not missing) values of another. Some instances of this method include: adding random noise to continuous variables or using post randomizing method (PRAM) for string variables [15].

The post randomization method (PRAM) was first proposed by Sarandal et al. in 1992 that used an approach similar to the common randomized response technique in sampling to protect micro data files against disclosure through randomizing personal records. In PRAM, we are faced with micro data with many fields and strings that makes PRAM a NP-Hard problem [1, 15].

Micro aggregation: this is another statistical disclosure control approach in the group of perturbation methods that is employed for quantitative variables. One of the problems of using micro aggregation is performing optimal multi-variable micro aggregation with the least amount of lost data, which is a NP-Hard problem. Chin et al. proposed a method, which can obtain an approximate solution in $O(n/k^2)$ [16].

There are many privacy preservation methods in data mining. These methods can be categorized from different aspects including data distortion, data mining algorithms, rules and data hiding, protecting privacy, etc. we can say that the main goal of privacy protection using data mining approaches is modifying the main data through some methods, so that their privacy and validity is preserved [3].

Anonymization is one of the data mining methods for preserving the privacy of micro data that aims to propagate micro data for research purposes. This research is focused on data anonymization techniques to make a trade-off between ensuring the privacy of individuals and propagating the desired data.

One recent study in USA has estimated that about 87% of the population of the United States can be uniquely identified through a "linking attack" using unique safe characteristics like (age, birthday, and the 5 digit postal code). In order to prevent this attack, a technique called k-anonymity was proposed by Samarati and Sweeney [5].

A table is called k-anonymous if its records can be recognized by at least k-1 other records through indirect identifier variables. In other words, each record in the main data table is mapped to k-1 record in the transformed table as we can see in table 1 [5].

Copyright © 2015 MECS

It should be noted that increasing k also increases the privacy protection; however, it also increases the lost information. In other words, the information validity is reduced [3]. K-anonymity technique guarantees that individuals are not uniquely identifiable by linking attacks and they are protected against identity disclosure attacks with probability of 1/k [4].

Table 1. Example of k-anonymity where k=2 and QI = {Race, Birth, Gender, ZIP}

Race	Birth	Gender	Zip	Problem
Black	1965	m	0214*	Short breath
Black	1965	m	0214*	chest pain
Black	1965	f	0213*	hypertension
Black	1965	f	0213*	hypertension
Black	1964	f	0213*	obesity
Black	1964	f	0213*	chest pain
White	1964	m	0213*	chest pain
White	1964	m	0213*	obesity
White	1964	m	0213*	Short breath
White	1967	m	0213*	chest pain
White	1967	m	0213*	chest pain

Recent algorithms trying to solve the privacy preservation problem of micro data using k-anonymity aim to propose a method to reduce the lost information [6]. Meyerson and Williams [7] and Aggarawal et al. [8] have proved that optimal k-anonymity (based on the number of cells and attributes that are global recorded and local suppressed) is a NP-hard problem and used approximation algorithms to find the optimal k-anonymity [10]. In order to achieve the optimal k-anonymity, global recording and local suppression are employed [2, 6].

The drawback of k-anonymity is that although it makes data records secure against identity disclosure attacks with probability of 1/k, it cannot secure them against attribute disclosure. In other words, the low diversity in the values of sensitive variables allows strong attribute disclosure attacks. In order to eliminate this issue, the values of sensitive variables are diversified using a method called l-diversity. In k-anonymity, the generated data are not still secure enough to ensure privacy against attribute disclosure attacks using sensitive variables. In order to overcome this drawback, l-diversity introduces the concept of equivalent classes each of which has at least one different value for a sensitive variable [6, 9, 11]. It should be noted that in this method, records are not fully protected against attribute disclosure attacks, for instance similarity attacks [6].

After proposing l-diversity, researcher recently discovered that the distribution of personal information having similar diversity levels may provide different levels of privacy preservation. This is due to the semantic connection between the values of sensitive attributes and their different sensitivity level values. They also believe that preserving privacy is connected to the overall distribution. This made some researchers propose t-closeness [6, 11].

Accordingly, algorithms were proposed to apply some constraints to k-anonymity, l-diversity, and t-closeness that are investigated in this paper. One of these algorithms is k-anonymity by Samarati and Sweeney [4].

Guillermo et al exploited micro aggregation to develop a k-anonymity algorithm and consequently, created privacy preservation on query logs. In fact, they applied clustering, one of the data mining techniques, to k-anonymity. The advantages of this method includes efficiency and simplicity in comparison to other methods [2, 18].

In 2010, Chih-Hua et al. proposed a method called k-support anonymity that enhanced privacy preservation in comparison to previous research and reduced the overhead of generating safe data. In 2011, Katerian Doka et al. developed k-anonymity on distributed data and succeeded to increase privacy, as well as reduce the lost information. It should be noted that in these algorithms, there is a high rate of lost data. Moreover, the constraints of l-diversity and t-closeness were not considered (which shows a weakness of preserving privacy in micro data). Moreover, p-sensitive k-anonymity, p+-sensitive k-anonymity, (p, α)-sensitive k-anonymity were proposed [9] that considered the constraints of l-diversity; however, the problem of attribute disclosure was not fully addressed and these models are vulnerable to similarity attacks. Another disadvantage of these models is not applying the constraints of t-closeness.

IV. The Proposed Method

In this section a three stage algorithm is proposed that simultaneously applies l-diversity, p-sensitive k-anonymity, p+-sensitive k-anonymity, and t-closeness constraints to the dataset and also significantly reduces the lost information. Figure 1.Shows the proposed method and its three main stages.

Stage.1- Attribute weighting: The first stage to begin the process of data anonymization is attribute weighting.

Copyright © 2015 MECS

This stage affects the results of the data preprocessing stage. We can say that the quality of the generated clusters in the second stage fully depends on this stage. The output of the attribute weighting stage is a set of numbers that specify the effect of each attribute in the clustering operation. The attributes with larger Weights have the most effect on the clustering operation. The output of this stage is used in the preprocessing stage of the clustering operation.

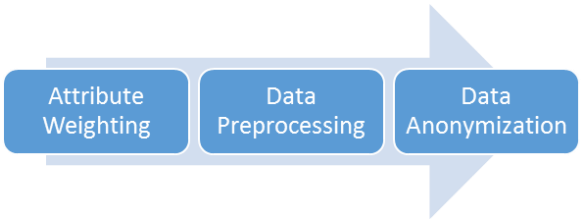

Fig 1. The stages of the proposed method.

Stage.2- Data preprocessing: After executing the attribute weighting stage, the anonymization process enters the preprocessing phase. The goal of this stage is to run the clustering operations on the dataset such that the number of members in each clusters are not more than k. if the number of members in any cluster is higher than k, the lost information in the anonymization stage is significantly reduced. This stage is comprised of three parts:

A- Selecting quasi-identifiers
B- Level 1 clustering operations
C- Level 2 clustering operations

In what follows, each of these stages are explained in details.

A- Selecting Quasi-Identifiers:

In this section, according to the weights of quasi-identifiers, a subset of them having the highest weight and the lowest diversity is selected for the clustering operations. It should be noted that the equal attribute values in a cluster is significantly effective in reducing the lost information in the anonymization stage.

B- Level 1 Clustering Operations:

In this section, quasi-identifiers are used for the clustering operations. In this research, all clustering operations are implemented by x-means. In order to run x-means algorithm, must select the minimum and maximum values for the number of clusters and the algorithm will select the best number of clusters in this range for the dataset. Subsequently, each record is labeled with a cluster and according to the selected range of k, clusters with members larger than k are considered as dataset D1 and clusters with members less than k are considered as dataset D2.

C- Level 2 Clustering Operations:

The level 2 clustering operations is explained in two parts:

C.1- Clustering Operations on Dataset D1:

In Dataset D1, according to the weights of quasi-identifiers in the weighting stage, one or more attributes, which were not selected for the clustering operation are investigated and the attribute(s) with the largest weight, the lowest diversity in the remaining quasi-identifiers is added to the previous attributes, and the clustering operations are applied to each of their clusters again. In this stage, the label of the second clustering is assigned to the records in dataset D1.

C.2- Clustering Operations on Dataset D2:

Since the number of members in the clusters of dataset d2 is smaller than k, it is not appropriate to perform anonymization operations in the next stage. In the preprocessing stage, the number of members in each cluster should be larger than k; otherwise, the lost information in the anonymization stage is increased. Therefore, the number of members in each cluster in dataset D2 must increase more than k. In order to do so, the records` labels, by the level 1 clustering operations, is removed and clustering proceeds by eliminating some quasi-identifiers. The executive steps of this section are presented in figure 2.

Stage.3- Data anonymization: Finally, after finishing the clustering operations on dataset D2 and assigning a new cluster label to it, clusters with members larger than k are stored in dataset D21 and clusters with members lower than k are stored in dataset D22; datasets D1, D21, and D22 are considered the inputs of the anonymization stage.

The anonymization stage uses the three datasets D1, D21, and D22, which are the outputs of the preprocessing stage. In what follows, the specifications of these datasets are explained. Datasets D1 is comprised of the records of quasi-identifiers and sensitive variables, as well as cluster labels and a sub-cluster for each record. Dataset D21 is similar to dataset D1; however, each record has only the cluster label. Dataset D22 is also similar to dataset D21 with the difference that records are not labeled.

The anonymization operations are divided into three parts:

A- Operations on dataset D1
B- Operations on dataset D21
C- Operations on dataset D22

A- Operations on Dataset D1:

In this section, based on users` requirements (users may only require one or more constraints), the algorithm first applies k-anonymity, p-sensitive k-anonymity, p+-sensitive k-anonymity, and t-closeness constraints to the corresponding sub-cluster of each cluster.

If it is not possible to apply these constraints to a sub-cluster, the distance of the sub-cluster`s centroid is computed with the centroids of surrounding clusters, it is merged with the cluster with the minimum distance, and the constraints are applied to the sub-cluster again. If it is still not possible to apply the constraints the merging continues until they are successfully applied. These operations are performed on sub-clusters of each cluster

Copyright © 2015 MECS

and finally, the output of this stage is the confidential dataset P (D1).

B- Operations on Dataset D21:

The anonymization operations on dataset D21 is slightly different from the operations in stage 1. The only different between dataset D21 and dataset D1 is the sub-cluster label attribute. Records in dataset D21 are only labeled with the cluster attribute. The algorithm of this section is similar to the algorithm for D1, except for performing on clusters instead of sub-clusters.

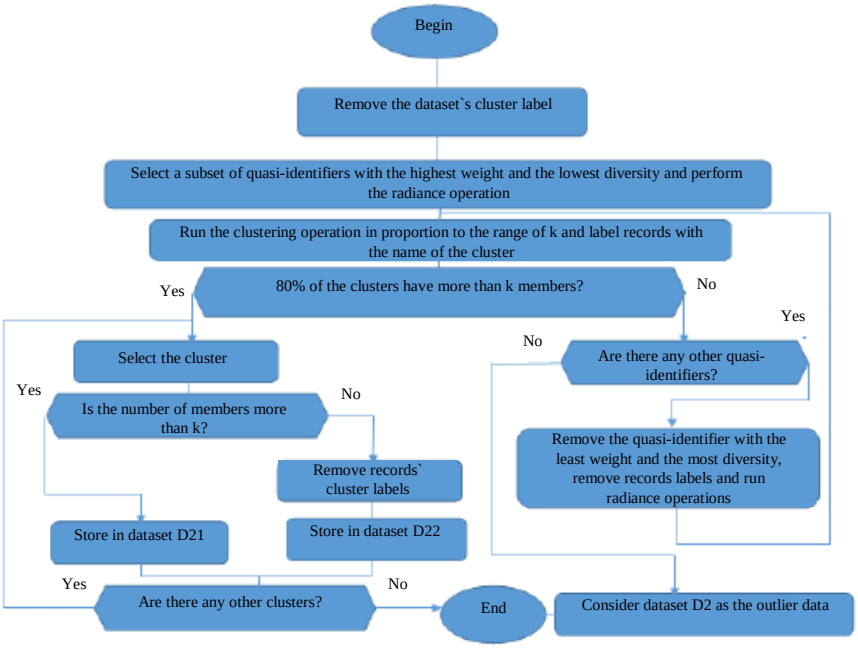

Fig. 2. The flowchart of execution steps of the clustering operations using dataset D2

C- Operations on Dataset D22:

The anonymization operations of this section is different from stages 1 and 2. Dataset D22 does not have cluster or sub-cluster attributes, since during clustering, the number of members in its clusters was less than k.

In this stage, the k-anonymity, p-sensitive k-anonymity, p^+-sensitive k-anonymity, and t-closeness constraints are applied to the entire dataset. If these constraints are not applicable, the elimination process is performed on the whole dataset and the amount of lost information reaches 100%.

Dataset D22 is usually small and its removal does not significantly affect the amount of lost information. This set is full of outliers and removing this data can considerably affect the security of the dataset.

After finishing the anonymization stage, a confidential dataset is created with all k-anonymity, p-sensitive k-anonymity, p^+-sensitive k-anonymity, and t-closeness constraints, which is completely secure against identity disclosure and attribute disclosure attacks. The pseudo code of figure 3 presents the proposed algorithm.

V. Evaluation Measures

This section discusses the three general evaluation measures to measure the quality of the anonymized data.

Distortion ratio: consider the value of an attribute in a record that is not generalized (global recorded). In this case, there has been no alterations. While, if the value of an attribute in a record is extended to a more generalized value in the taxonomy tree (hierarchical generalization), the alterations made on this attribute is in proportion to the degree of generalization performed on that attribute. For instance, if the updated value in the taxonomy tree is a node close to the root, the distortion ratio is higher; therefore, the distortion ratio of the values of the attributes depend on what height of the generalization tree they are located. For instance, the value that is not generalized is placed at height 0 of the tree. If a value is generalized on only one level, it is placed at height 1 of the generalization tree.

VI. Evaluation

In this research, Adult dataset is used [21] for evaluation, it was developed in UC Irvine Machine Learning Repository. This dataset is used as a metric for the evaluation of k-anonymity models regarding efficiency (similarity attacks), execution time, and the distortion ratio. The settings were adjusted similar to X. Sun L. Sun, and H. Wang, 2011.

Records with unknown values in dataset were removed and as a result, a dataset with 45222 records were produced. Seven attributes were selected as quasi-identifiers, column Health Condition with the sensitive values including {HIV, Cancer, Indigestion, Flu, Asthma, Obesity, Hepatitis, and Phthisis} was added to Adult dataset, and a sensitive value was randomly assigned to

Copyright © 2015 MECS

each data record. Table 2 presents a brief description of the modified dataset, which includes the used attributes and their type, number of distinct values, and hierarchical generalization height.

In this experiment, p-sensitive k-anonymity, p+-sensitive k-anonymity, (p, α)-sensitive k-anonymity, (k, p, t) Anonymizing3Layer (the proposed method), and (k, p+, t) 3LayerAnonymizing (the proposed method) respectively considered model 1, 2, 3, 4, and 5. All experiments were conducted on one system running Windows XP with a 2.00 GHz processor and 1GB RAM.

Algorithm: (K, L, T) 3LayerAnonymizing
Input: Dataset D
Parameter1: A quasi identifiers: QI
Parameter2: Limitations values: K, L and T
Outputs: Data sets: P (D)
Proccess:
1. Weighting (QI)
2. Select Subset of the QI
3. Execute Clustering level 1
4. Product Dataset D_1
5. Clustering level 2-1
6. Product Dataset D_{21}
5. Clustering level 2-2
6. Product Dataset D_{22}
7. Anonymizing D_1(K, L, T)
8. Anonymizing D_{21}(K, L, T)
9. Anonymizing D_{22}(K, L, T)
10. Join (D_1, D_{21}, D_{22})
11. Save change D
12. Return D

Fig. 3. The pseudo code of the proposed algorithm

A. Experimental Results

The proposed algorithm was evaluated using metrics like similarity attacks, execution time, and distortion ratio. The experimental results are presented as follows:

A.1- Similarity Attacks (Attribute Disclosure): points out to cases in which sensitive attributes have similar values in QIs and the intruder can obtain important information about the responders. The first seven attributes in table 2 were considered quasi-identifier attributes and Health condition was selected as the sensitive attribute. The eight values of attribute Health Condition were divided into 4 predefined groups based on their confidentiality level. Table 3 presents this division. Each quasi-identifier, in which the values of all sensitive attributes are in one group, are vulnerable against similarity attacks.

Table 2. A brief description about the dataset and the height of the generalization tree in dataset Adult.

Attribute	Type	Height
Age	Numeric	5
Workclass	Categorical	3
Education	Categorical	4
Country	Categorical	3
Marital status	Categorical	3
Race	Categorical	3
Gender	Categorical	2
Health condition	Sensitive	-

According to the experiment implemented in [9, 11, 20], the following results were obtained. First a 2-sensitive 2-anonymous confidential table was created, which transformed into a total of 21 minimized datasets. Thirteen of these datasets were vulnerable to similarity attacks. In one of these anonymized tables, in sum, 916 records were derived from the class of the sensitive values and only four of these datasets were vulnerable to similarity attacks. Similar results were obtained about the p+-sensitive k-anonymity model; thus, we can conclude that the two p+-sensitive k-anonymity and (p, α)-sensitive k-anonymity models are significantly effective in reducing the possibility of similarity attacks.

Subsequently, the confidential table of the proposed method, (2, 2+, 0.2) Anonymizing3Layer, was created, which was transformed into a total of 65 minimized clusters. In this 65 clusters, there was no vulnerability against similarity attacks (0/65=0.0%). In the confidential table of (2, 2+, 0.2) Anonymizing3Layer, 19 records were identified and removed as outliers and we can say that eliminating these 19 records was considerably effective in reducing the similarity attacks to zero.

A.2- Execution Time: the execution time was compared for the four main privacy preservation metrics, p-sensitive k-anonymity, p+-sensitive k-anonymity, (p, α)-sensitive k-anonymity, (k, p, t) Anonymizing3Layer (the proposed method), and (k, p+, t) Anonymizing3Layer (the proposed method). Results of the experiments are presented in tables 3 and 4. In figure

Copyright © 2015 MECS

3, α=4, P=4, k=4, and t=0.5. In this figure, the horizontal axis shows the number of quasi-identifiers and the vertical axis indicates the execution time. The number of quasi-identifiers changed between 2 and 7.

Table 3. The confidentiality degree of Health Condition

Category ID	Sensitive values	Sensitivity
One	HIV, cancer	Top secret
Two	Phthisis, hepatitis	Secret
Three	Obesity, asthma	Less secret
Four	Flu, indigestion	Non secret

Figure 4 shows that increasing the number of quasi-identifiers also slightly increases the execution time of the proposed algorithm, which can indicate that the execution time of the algorithms are almost independent of the number of quasi-identifiers.

In figure 4, results show that the proposed algorithm is considerably slower than the other three models. However, an issue is ignored regarding the time optimization of the proposed algorithms, which is not showable in the figures. As it was mentioned, in these algorithms, there are stages called weighting and preprocessing. These stages prepare the dataset for the anonymization stage. The important point about these stages is that they are performed only once for a range of k.

For instance, if the range of k is between 2 to 100 and we aim to produce 98 anonymized datasets respectively with 2-anonymity to 100-anonymity, weighting and preprocessing stages are performed only once and the anonymization stage is performed 98 times. This shows that in large anonymization scales, the weighting and preprocessing stages are significantly effective in reducing the execution time and we can observe that the proposed algorithm is suitable for commercial purposes and large scale anonymization; in such cases, a reduction in the execution time of the proposed algorithm can be observed in comparison to other algorithms. However, due to the low number of anonymization for different values of k and p, the overhead of the preprocessing stage is considerable and the proposed algorithm is shown to be slower than the other algorithms.

Figure 5 presents the effect of increasing k on the execution time. In the experiment of figure 5, the value of t and l are respectively 0.2 and 2. As we can see, by increasing k, the execution time first has a linear increase to a specific value. However, after that value, the execution time remains constant and even starts to decrease.

In conclusion, this experiment shows that increasing k after passing the initial values results in k decreasing execution times with a very small slope. The experiment above indicates that regarding the execution time, the proposed algorithm is considerably more efficient than other methods.

In this section, the productivity of the proposed algorithm was compared with the most well-known models. In the next section the distortion ratio measure is compared and discussed.

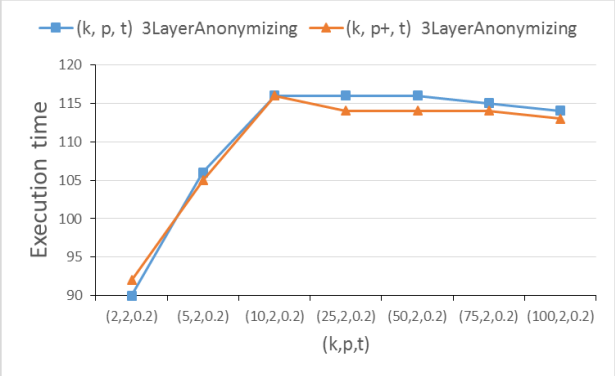

Fig. 4. The results of the effect of increasing k on the execution time of the proposed algorithm in the first experiment.

A.3 - The Distortion Ratio: the distortion ratio of four privacy preservation measures were compared: p-sensitive k-anonymity, p+-sensitive k-anonymity, (p, α)-sensitive k-anonymity, (k, p, t) Anonymizing3Layer, and (k, p+, t) Anonymizing3Layer. Results of the experiment is presented in figure 6 and 7. In figure 6, α=4, P=4, k=4, t=0.5.

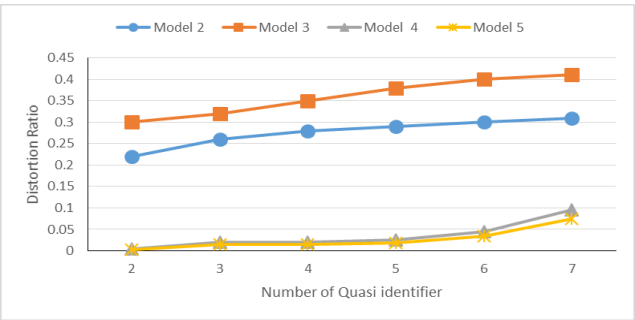

Fig 5. The effect of increasing the number of quasi-identifiers on the distortion ratio

In figure 6, the horizontal axis shows the number of quasi-identifiers and the vertical axis shows the distortion ratio. The number of quasi-identifiers changes between 2 to 7. As we can see, the distortion ratio values in the proposed algorithm is considerably less than 2 and 3.

In table 6, we can see that increasing the number of quasi-identifiers also increases the distortion ratio. This is natural, since increasing quasi-identifiers also increases the lost information, which in turn increases the distortion ratio. Figure 7 presents the effect of increasing k on the distortion ratio. In the experiment of figure 7, values t and l are respectively considered 0.2 and 2.

As we can see in figure 7, increasing k also increases the distortion ratio, since increasing k makes data more secure against identity disclosure attacks and thus more information is lost. Increasing k is directly related to increasing the security level of the output tables.

Copyright © 2015 MECS

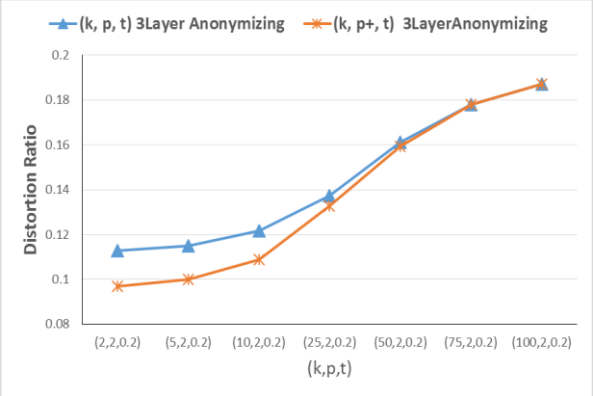

Fig. 6. Results of increasing k on the distortion ratio of the proposed algorithm

VII. Conclusions and Future Works

This research proposes methods to preserve privacy and increase the safety of the propagated micro datasets. In this research, using clustering and attribute weighting concepts, a three stage approach was proposed. The proposed method fully protects the micro data against attribute disclosure attacks and significantly reduces distortion ratio. The proposed method can be improved as follows:

- The proposed method can be improved regarding the execution time and the lost values. Investigating different weighting algorithms and applying them in the preprocessing stage can improve these issues.
- One of the well-known clustering algorithms of WEKA, called X-Means, was used in this research. The provided results of the proposed algorithm can be improved by customizing this clustering algorithm.
- Another issue is related to the anonymity stage of the proposed method. It seems that it is still possible to make optimizations in this stage to reduce the lost information and the execution time of the algorithm.

References

[1] L. Willenborg and T.Waal "Elements of Statistical Disclosure Control", Springer, 2001.

[2] S. Morton, "An Improved Utility Driven Approach Towards k-anonymity Using Data Constraint Rules", Submitted to the faculty of the University Graduate School in partial fulfillment of the requirements for the degree Doctor of Philosophy in the School of Informatics, Indiana University , 2012.

[3] X. Qi, and M. Zong, "An Overview of Privacy Preserving Data Mining", International Conference on Environmental Science and Engineering, ICESE, p. 1341-1347, 2012.

[4] D. Agrawal, and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Volume NY, pp. 247-255, 2001.

[5] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty Fuzziness and Knowledge Based Systems", vol. 10, no. 5, pp. 557-570, 2002.

[6] J. Yongcheng and S.Jiajin Le, "A Survey on Anonymity-based Privacy Preserving", Proceedings of the E-Business and Information System Security, p. 1 - 4, 2009.

[7] A. Meyerson, and R. Williams, "On the complexity of optimal k-anonymity", Proceedings of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, Paris, France, pp. 223-228, 2004.

[8] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Anonymizing tables", Proceedings of the 10th International Conference on Database Theory (ICDT'05), Edinburgh, Scotland, pp. 246-258, 2006.

[9] X. Sun, L. Sun, and H. Wang, "Extended k-anonymity models against sensitive attribute disclosure", Computer Communications, Volume 34, Issue 4, Pages 526–535, 2011.

[10] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation algorithms for k-anonymity", Proceedings of the 2007 ACM SIGMOD international conference on Management of data, p. 67-78, 2007.

[11] C. Moque, A. Pomares, and R. Gonzalez, "AnonymousData.co: A proposal for Interactive Anonymization of Electronic Medical Records", Proceedings of the 4th Conference of ENTERprise Information Systems-aligning technology, 2012.

[12] A. Machanavajjhala, J.Gehrke, and D. Kifer, "L-diversity: Privacy beyond k-anonymity", Proceedings of the ICDE", p.24, 2006.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-anonymity and L-Diversity", Proceedings of the ICDE, pp.106-115, 2007.

[14] T. Waal, and L. Willenborg, "Information loss through global recoding and local suppression", Proceedings of the Netherlands Official Statistics, Volume 14, p. 17-20, spring 1999.

[15] P. Wolf, J. Gouweleeuw, P. Kooiman, and L. Willenborg, "Reflections on PRAM", Proceedings of the Statistics Netherlands Department of Statistical Methods 1998.

[16] C. Chang, Y. Li, and W. Huang, "TFRP: An efficient micro aggregation algorithm for statistical Disclosure control", The Journal of Systems and Software, Volume 80, Issue 11, p.1866-1878, 2007.

[17] Z. FeiFei, D. LiFeng, W. Kun, and L. Yang, "Study on Privacy Protection Algorithm Based on K-Anonymity", International Conference on Medical Physics and Biomedical Engineering, Volume 33, p. 483 - 490, 2012.

[18] G. Torra, A. Erola, and J. Roca,"User k-anonymity for privacy preserving data mining of query logs", Information Processing and Management, Volume 48, Issue 3, p.476-487, 2012.

[19] C. Tai, P. Yu, and M. Chen," k-Support Anonymity Based on Pseudo Taxonomy for Outsourcing of Frequent Item set Mining", KDD'10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 473-482, 2010.

[20] K. Doka, D. Tsoumakos, and N. Koziris," KANIS: Preserving k-anonymity Over Distributed Data", Research Center Athena, Athens, Greece, 2011.

Copyright © 2015 MECS

[21] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases. Available at <www.ics.uci.edu/-mlearn/ MLRepository.html, University of Califonia, Irvine, 2015/05/25.
[22] D. Pelleg and A. Moore: "X-Means: Extending K-Means with Efficient Estimation of the number clusters", in: ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, pp 727-734, 2000.

Authors' Profiles

Masoud Rahimi, received the B.Sc. degree in computer engineering in 2009 from University of PNU, Farsan, Iran, and the M.Sc. degree in computer engineering from University of Sheikhbahaee (SHBU), Isfahan, Iran, in 2015.

Mehdi Bateni, is an assistant professor of computer engineering at the Faculty of Engineering of the Sheikhbahaee University (SHBU). He received his B.Sc. in Computer Engineering in 1997 from University of Isfahan, Isfahan, Iran and his M. Sc. in Computer Engineering from Ferdowsi University of Mashhad, Mashhad, Iran in 2000. He received his Ph.D. in Computer Engineering in 2012 from University of Isfahan, Isfahan, Iran.

Hosein Mohammadinejad, received the B.Sc. degree in computer engineering in 2000 from University of Tehran, Tehran, Iran, and the M.Sc. degree in computer engineering from University of Isfahan, Isfahan, Iran, in 2003. He is currently working toward a Ph.D. degree with the Department of Computer Engineering, University of Isfahan, Isfahan, Iran. He has worked as a Lecturer at Sheikhbahaee University, Isfahan, Iran, since 2003.

How to cite this paper: Masoud Rahimi, Mehdi Bateni, Hosein Mohammadinejad,"Extended K-Anonymity Model for Privacy Preserving on Micro Data", IJCNIS, vol.7, no.12, pp.42-51, 2015.DOI: 10.5815/ijcnis.2015.12.05

Copyright © 2015 MECS