Journal of Information Security, 2015, 6, 179-196
Published Online July 2015 in SciRes. http://www.scirp.org/journal/jis
http://dx.doi.org/10.4236/jis.2015.63019

Utility-Based Anonymization Using Generalization Boundaries to Protect Sensitive Attributes

Abou-el-ela Abdou Hussien[1]*, Nagy Ramadan Darwish[2], Hesham A. Hefny[2]

[1]Department of Computer Science, Faculty of Science and Arts, Shaqra University, Shaqra, KSA
[2]Department of Computer and Information Sciences, Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt
Email: *abo_el_ela_2004@yahoo.com, drnagyd@yahoo.com, hehefny@hotmail.com

Received 7 May 2015; accepted 12 June 2015; published 15 June 2015

Abstract

Privacy preserving data mining (PPDM) has become more and more important because it allows sharing of privacy sensitive data for analytical purposes. A big number of privacy techniques were developed most of which used the *k*-anonymity property which have many shortcomings, so other privacy techniques were introduced (*l*-diversity, *p*-sensitive *k*-anonymity, (*α*, *k*)-anonymity, *t*-closeness, etc.). While they are different in their methods and quality of their results, they all focus first on masking the data, and then protecting the quality of the data. This paper is concerned with providing an enhanced privacy technique that combines some anonymity techniques to maintain both privacy and data utility by considering the sensitivity values of attributes in queries using sensitivity weights which determine taking in account utility-based anonymization and then only queries having sensitive attributes whose values exceed threshold are to be changed using generalization boundaries. The threshold value is calculated depending on the different weights assigned to individual attributes which take into account the utility of each attribute and those particular attributes whose total weights exceed the threshold values is changed using generalization boundaries and the other queries can be directly published. Experiment results using UT dallas anonymization toolbox on real data set adult database from the UC machine learning repository show that although the proposed technique preserves privacy, it also can maintain the utility of the publishing data.

Keywords

Privacy, Privacy Preserving Data Mining, *K*-Anonymity, Generalization Boundaries, Suppression

[*]Corresponding author.

How to cite this paper: Hussien, A.A., Darwish, N.R. and Hefny, H.A. (2015) Utility-Based Anonymization Using Generalization Boundaries to Protect Sensitive Attributes. *Journal of Information Security*, **6**, 179-196.
http://dx.doi.org/10.4236/jis.2015.63019

1. Introduction

Many organizations collect and hold very large volumes of data like hospitals, credit card companies, real estate companies and search engines. They would like to publish the data for the purposes of data mining. Data mining is a technique for automatically and intelligently extracting information or knowledge from very large amount of data [1] [2]. When these data are released, it contains a lot of sensitive information. So we would like to preserve the privacy of the individuals represented in these data. In order to make any public system secure, we must not only be sure that private sensitive data have been maintained, but also make sure that inferences caused by certain channels have been blocked as well. Therefore, privacy preserving data mining (PPDM) [3]-[5] is becoming an increasingly important field of research and it has been drawn increasing attention. The main goals that a PPDM algorithm should enforce are [6]:

1) A PPDM algorithm should have to prevent the discovery of sensitive information.
2) It should be resistant to the various data mining techniques.
3) It should not compromise the access and the use of non sensitive data.
4) It should not have an exponential computational complexity.

A number of effective techniques for PPDM have been proposed. Most techniques use some form of transformation on the original data in order to maintain the privacy preservation. The transformed dataset could be available for mining and must achieve privacy requirements without affecting the mining benefits.

2. Related Research Areas

In recent years, a lot of techniques have been proposed for implementing *k*-anonymity [7] via generalization and suppression [8]. We introduce some of them in next sections.

2.1. *K*-Anonymity Technique

K-anonymity classifies the attributes of tables into four different classes [6]. *First class is Explicit_Identifier*, a set of attributes, such as name and social security number SSN, containing information that explicitly identifies record owners; *Second Class is key or Quasi_Identifier*, set of attributes that could potentially identify record owners, as Gender, Zip code and Age, that may be known by an intruder (which are generally linked with publicly available database to re-identify the individual); *Third Class is Sensitive _Attributes*, consists of sensitive person-specific information such as disease, salary, and disability status and *Fourth Class is Non-Sensitive_ Attributes*, contains all attributes that do not fall into the previous three categories. The Four sets of attributes should be disjoint.

Suppose that we have two tables, **Table 1** contains classification of attributes for *k*-anonymity and **Table 2** anonymization of **Table 1** [6].

2.2. *K*-Anonymity, Generalization and Suppression

Let *IM* be the initial microdata and *MM* be the released (masked) microdata [9]-[11]. We divide attributes characterizing *IM* into the following Classes as mentioned before: *Identifier attributes*, *Key or quasi-identifier attri-*

Table 1. Classification of attributes for *k*-anonymity.

Identifier attribute	Quasi-identifier				Sensitive attributes
Name	Race	Birth	Sex	Zip code	Disease
Alice	Blank	1965	M	02141	Flu
Bob	Blank	1965	M	02142	Cancer
David	Blank	1966	M	02135	Obesity
Helen	Blank	1966	M	02137	Gastritis
Jane	White	1968	F	02139	HIV
Paul	White	1968	F	02138	Cancer

Table 2. Anonymization of **Table 1**.

Quasi-identifier				Sensitive attributes
Race	Birth	Sex	Zip code	Disease
Blank	1965	M	0214*	Flu
Blank	1965	M	0214*	Cancer
Blank	1966	M	0213*	Obesity
Blank	1966	M	0213*	Gastritis
White	1968	F	0213*	HIV
White	1968	F	0213*	Cancer

butes, *Sensitive or confidential attributes* as shown in **Table 1**. While the *Identifier attributes* are removed from the released microdata, *the Quasi-Identifier and confidential attributes* are usually released to the researchers/ analysts as shown in **Table 2**. A general assumption is that the values for the confidential attributes are not available from any external source. Unfortunately, an intruder may use record linkage techniques [12] [13] between quasi-identifier attributes and external available information to disclose the identity of individuals from the masked microdata. To avoid this possibility of disclosure, one solution is to modify the initial microdata, specifically the quasi-identifier attributes values, in order to maintain the *k*-anonymity property. To express the *k*-anonymity property, we can use the following concept:

Definition 1 (*QI-Cluster*): consists of all the records with identical combination of quasi-identifier attribute values in M [11]. This term was not defined when *k*-anonymity was introduced [14] [15]. More recent papers use different terminologies such as *QI-group* [16], and equivalence class [17].

Definition 2 (*K*-Anonymity Property): The *k*-anonymity property for an *MM* is satisfied if every *QI-cluster* from *MM* contains k or more tuples (records) [11] [18]. *K-anonymity* technique is generally use generalization or suppression of QI-*group* attributes masking initial microdata. First Generalization consists in replacing the actual value of the attribute with a less specific, but more general value. Generalization was extended for numerical attributes either by using predefined hierarchies [19] or a hierarchy-free model [20]. The values from different domains of this hierarchy are represented in a tree called value generalization hierarchy. They illustrate domain and value generalization hierarchy in **Figure 1** for attributes Zip code and Sex.

There are two ways to achieve generalization, namely global recoding and local recoding. Another name for global recoding is domain generalization. The generalization happens at the domain level. When an attribute value is generalized, every occurrence of the value is replaced by the new generalized value. A global recoding method may over-generalize a table. In global recoding, all occurrences of an attribute value are recoded to the same value. In contrast local-recoding method generalizes attribute values at cell level. A local recoding method does not over-generalize a table and hence may minimize the distortion of an anonymous view. If local recoding is adopted, occurrences of the same value of an attribute may be recoded to different values.

Second Suppression involves not releasing a value at all [13]. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.

2.3. Constrained *K*-Anonymity

This technique [11] is used to explain how the resulted masked microdata would still be useful although generalization process for *QI-groups*. The data owners are often care of how researchers are mining their data and, consequently, they could identify maximum allowed generalization values. So there is need to specify a generalization boundary, which determines maximum allowed generalization value that is associated with each possible *QI-group* attribute value from *IM*.

Definition 3: (Maximum Allowed Generalization Value): Let *Q* be a quasi-identifier attribute (numerical or categorical), and *HQ* its predefined value generalization hierarchy [11]. For every leaf value $v \in HQ$, the maximum allowed generalization value of *v*, denoted by *MAGVal*(*v*), is the value (leaf or not-leaf) in *HQ* situated on the path from *v* to the root, such that:

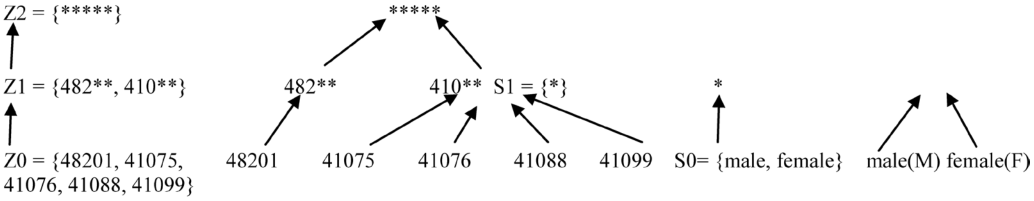


Figure 1. Examples of domain and value generalization hierarchies.

- For any released microdata, the value *v* is permitted to be generalized only up to *MAGVal*(*v*) and
- When different *MAGVals* exist on the path between *v* and the hierarchy root, then the *MAGVal*(*v*) is the first *MAGVal* that is reached when following the path from *v* to the root node.

From **Figure 2** we can introduce an example of defining maximal allowed generalization values for a subset of values for the Location attribute. The *MAGVals* for the leaf values "Geza", "Helwan" and "Alex" are "South Egypt", "Mid Egypt" and "North Egypt", respectively, (the *MAGVals* are marked by * characters that delimit them). This means that the *QI*-Location's value "Geza" may be generalized to itself or "South Egypt", but not to "A.R.E", Also, "Helwan" may be generalized to itself, or "Mid Egypt", but not to "A.R.E", and "Alex" may be generalized to itself, or "North Egypt", but not to "A.R.E".

The second restriction in the *MAGVal's* definition specifies that the hierarchy path between a leaf value *v* and *MAGVal*(*v*) can contain no node other than *MAGVal*(*v*) that is a maximum allowed generalization value. This restriction is forced in order to avoid any ambiguity about the *MAGVals* of the leaf values in a sensitive attribute hierarchy. Note that several *MAGVals* may exist on a path between a leaf and the root as a result of defining *MAGVals* for other leaves within that hierarchy.

Definition 4: (Maximum Allowed Generalization Set): The set of all *MAGVals* for attribute *Q* is called *Q's* maximum allowed generalization set, and it is denoted by $MAGSet(Q) = \{MAGVal(v)| \forall v \in \text{leaves}(HQ)\}$ (The notation leaves(*HQ*) represents all the leaves from the *HQ* value generalization hierarchy).

From **Figure 2** the hierarchy for the attribute Location, *MAGSet* (Location) = {South Egypt, Mid Egypt, North Egypt}. Usually, the data owner/user only has generalization restrictions for some of the *QIs* in a microdata that is to be masked. If for a particular *QI*-attribute *Q* there are not any restrictions in respect to its generalization, then no maximal allowed generalization values are specified for *Q's* value hierarchy; in this case, each leaf value in *HQ* is considered to have the *HQ's* root value as its maximal allowed generalization value.

Definition 5: (Constraint Violation): the masked microdata *MM* has a constraint violation if one *QI*-value, *v*, in *IM*, is generalized in one record in *MM* beyond its specific maximal generalization value, *MAGVal*(*v*).

Definition 6: (Constrained *K*-Anonymity): The masked microdata *MM* satisfies the constrained *k*-anonymity property if it satisfies *k*-anonymity and it does not have any constraint violation.

2.4. Sensitivity-Based Anonymity

Automatic detection of sensitive attribute in PPDM [21] could be introduced in next steps:

1) Client's query is analyzed then taking in account the sensitive values in sensitive queries.

2) Queries having sensitive values are to be scrambled only depending on threshold value.

3) Different weights assigned to individual attributes are used to calculate threshold value.

4) The attributes whose total weights exceed the threshold values is scrambled depending on swapping techniques.

5) Queries whose total weights under the threshold values can be directly released.

6) The data owner is responsible for predetermine threshold limit.

Now we introduce the mechanism steps:

1) The number of weights assigned to attributes in a given database is equal to the total number of attributes.

2) The weights of attributes are allocated according to how much a particular attribute is involved in revealing the identity of an individual.

3) The attribute which can directly defines individual and leads to privacy violation is assigned highest weight.

4) Based on their level of priority similar weights can be assigned for some attributes.

How we apply the mechanism steps:

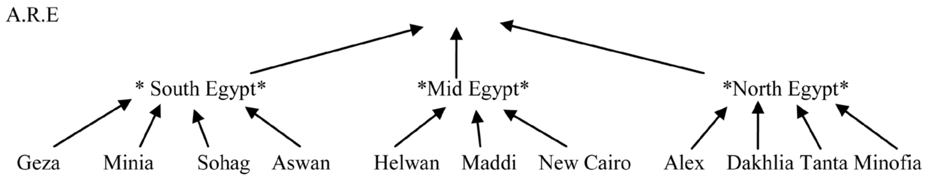


Figure 2. Examples of *MAGVals*.

1) Highest value of weights are assigned for The attributes which directly identify an individual like name, personal identification number, social security number, etc.

2) The sensitive attribute or attributes are identified by summing up weights of the attributes submitted in the client's query.

3) If the total weight exceeds the threshold limit then the attribute values are modified using modification technique in PPDM. **Table 3** gives an example showing different attributes and their weights. In this example they considered maximum threshold limit as 7.

The Architecture for sensitive attribute identification in PPDM [21] shown in **Figure 3** is very useful where group of parties or companies want to share their data bases for their mutual benefit. The parties also want to protect the privacy of sensitive information available in their databases. Each party is actually a data owner and maintains its own database.

In this technique they focus on determining the sensitive attributes before releasing the data required to the client. To determine the sensitive attributes the weights are assigned to each attribute depending on some facts as [21]:

1) If single attribute could disclosure identity of an individual.

2) If there is a possibility that group of attributes can indirectly disclosure the identity of an individual.

According to above facts weights are given to each attribute, only the sensitive data available in the database is modified for those attribute or group of attributes whose values exceeds the threshold limit of sensitiveness.

2.5. Utility-Based Anonymization

Anonymized data is often for the purposes of analysis and data mining [22]. As in many data analysis applications we could recognize that different attributes may have different utility. For example, consider disease analysis in anonymized data set about patients in certain steps as follows:

1) To achieve *k*-anonymity suppose that we can generalize from a five-digit full zip-code to a four-digit prefix (e.g., from 53712 to 5371*).

2) We can also generalize alternatively, attribute age to age groups (e.g., from to [19]-[27]).

3) In many cases we find that the age information is critical to disease analysis, while the information loss on the accurate location is often acceptable (a four digit prefix still identifies a relatively local region).

4) Therefore the age attribute has more utility than the zipcode attribute, and should be retained as accurate as possible in anonymization process.

Could we make the anonymization utility aware? Previous anonymization techniques has not consider attributes utility , so the researchers have intent to focus their attention on this important issue in the current paper to benefit from it in determining attributes weights.

2.6. UT DALLAS Anonymization Toolbox

This Toolbox is designed for anonymize datasets [24]. Next Section 2.6.1, explains input file format and Section 2.6.2, provides the details on various output file format choices made available to the user.

2.6.1. Input Format

This toolbox only supports unstructured text files which could be represented by ASCII files where each line represents a record that consists of attributes values separated by some character sequence (e.g., comma, tab, and semi-column), specified in the configuration file or passed as an argument to the toolbox. All descriptive information will be collected through the configuration file regarding attribute types and indices. So a header with such description will not be necessary.

Table 3. Database of employee with different attributes and weights.

SSNo weight = 7	Name weight = 6	Age weight = 4	Location weight = 5	Salary weight = 3	Gender weight = 2
ASD1	Ali	26	KSA	26 K	M
QDF4	Soha	33	Bahrin	33 k	F
GTR3	Alaa	45	Qatar	66 k	F
AEF4	Moustafa	33	Egypt	34 k	M

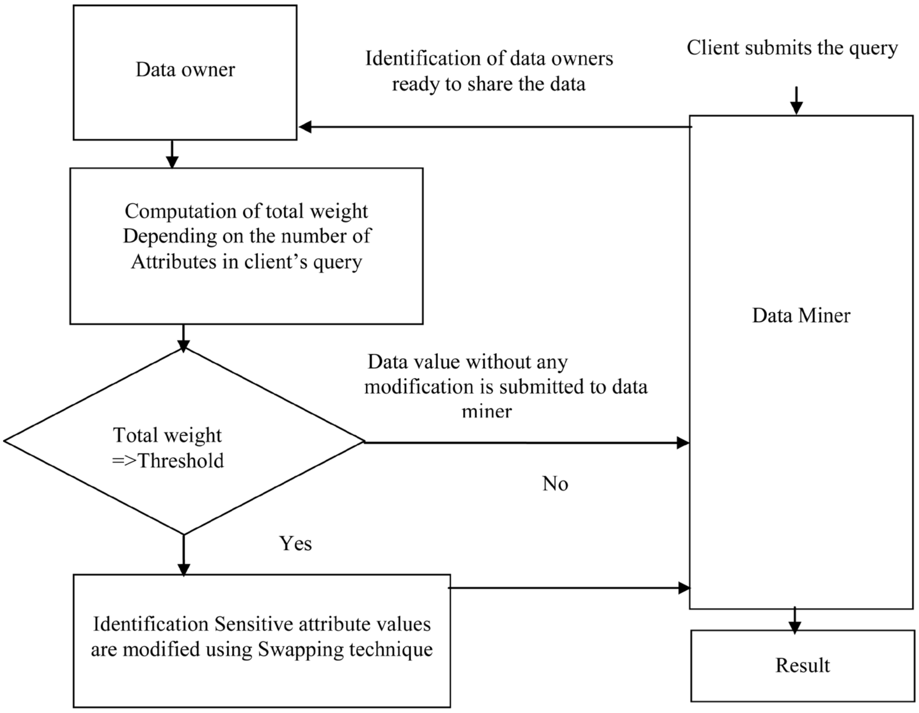

Data owner
Identification of data owners ready to share the data
Client submits the query
Computation of total weight Depending on the number of Attributes in client's query
Data Miner
Data value without any modification is submitted to data miner
Total weight =>Threshold
No
Yes
Identification Sensitive attribute values are modified using Swapping technique
Result

Figure 3. Architecture for sensitive attribute identification in PPDM.

2.6.2. Output Format

The output format likes the input format. There is only one difference between them caused by anonymization for quasi-identifier attributes through replacement of specific values with their generalized values. This output format is referred to as *genVals* in the toolbox. The toolbox permits releasing anonymized records with additional information as described in some researches likes, the first approach, using anonymized data for classification, proposed in [25] that intent to increase the accuracy of classification models built on anonymized data by releasing equivalence-wide statistics on each quasi-identifier attribute, referred to as *QI-statistics*. For categorical attributes the complete distribution is provided. For numerical attributes, such statistics contain the mean and variance across the equivalence class. Anatomization is the last output format supported by the toolbox approach described in, Anatomy: simple and effective privacy preservation, proposed in [26].

2.7. Data Utility Measures

There are two global utility measures presented here which capture differences in the distributions of the original and masked data. The first measure is *Cluster Analysis Measure* [27] [28] uses cluster analysis to determine whether records in the original and masked data have similar values. The second measure is *Empirical CDF Measures* use Kolmogorov-Smirnov-type statistics to evaluate differences between the empirical distribution functions of the original and masked data.

2.7.1. Cluster Analysis Measure

Cluster analysis, a shape of unsupervised machine language, divides records into groups according to having similar values for selected variables. A cluster can be considered as a group of objects which are similar between them and are dissimilar to the objects belonging to other clusters. When proportion of observations from original or masked data for each cluster is constant we could say that the two data sets have the same distributions. The distribution of original and masked data are compared by assigning observations in pooled data to clusters then computing differences between the number of observations from original and masked data for each cluster. Let g be the number of clusters. Then, cluster utility is defined by [28]:

$$U_{\text{cluster}} = \sum_{i=1}^{g} W_i \left(\frac{n_{i1}}{n_i} - c \right)^2$$

where n_i is the total number of observations grouped in i-th cluster, n_{i1} is the number of observations from original data for i-th cluster, W_i is the weight assigned to i-th cluster and c is a constant. One value is obtained for each cluster in cluster utility, so W_i in Equation explains how much each cluster contributes to the utility. Clustering methods can be clustered in many different ways. This measure has weaknesses, when two masking strategies have the same value of U_C; it is not necessarily that the masked data sets they produce are equally.

2.7.2. Empirical CDF Measures

These measures completely describe probability function and tests the differences between the empirical distribution functions obtained from the original and masked data which could be appropriate for measuring data utility. Let S_X and S_Y be the empirical distributions obtained from the original data, X, and the masked data, Y, respectively. When X has dimension $N_x \times d$, we have, [28]

$$S_x \left(X_1 \cdots, X_d \right) = \frac{1}{N_x} \sum_{i=1}^{N_x} I \left(X_{i1} \le X_1, \cdots, X_{id} \le X_d \right)$$

where x_{ij} equals the value of the *j-th* variable for the *i-th* observation, and $I(\cdot)$ equals one when the condition inside the parentheses is true and equals zero otherwise The $S_Y \left(y_1, \cdots, y_d \right)$ is defined similarly.

3. Proposed Technique

In this paper the researchers proposes Utility-Based Anonymization using Generalization Boundaries to protect Sensitive Attributes Depending on Attributes Sensitivity Weights. In this technique researchers start with considering the sensitivity of values in queries and then only quires having sensitive values (taking in account Utility-based Anonymization) are generalized using Generalization Boundaries and the other quires that doesn't have sensitive values can be directly published.

The objective of proposed model is trying to achieve Privacy and in the same time maintain Data utility and minimum information loss to enable efficient statistical analysis using data mining as follows [29]:

- Data utility—The goal is to eliminate the privacy violation (how much an adversary Learn from the released data) and increase utility (accuracy of data mining task) of published database. This is achieved by generalizing quasi-identifiers of only those Attributes having high sensitive attribute values taking in account Utility-based Anonymization as mentioned before.
- Privacy—To provide the individual data privacy by generalization in such a way that data re-identification cannot be possible
- Minimum information loss—The loss of information is minimized by giving Sensitivity level for sensitive attribute values, and attributes which has high sensitive level is only generalized and the rest of attributes are released as it is.

Proposed approach will be illustrated by an algorithm in Section 3.2.

3.1. Main Procedures

The operation of this technique is as below:

First: we use *automatic detection of sensitive attribute in PPDM* [21] to calculate Threshold.

Second: we use *utility-based anonymization* [22] to avoid more generalization for critical attributes using previous step.

Third: we use *constrained k-anonymity* [11] for these sensitive attributes using generalization boundaries to maintain data utility and minimum information loss.

3.2. Proposed Algorithm

The Algorithm of proposed technique is explained in **Figure 4** and its steps are as follows:

1) Client submit query to the data miner.

2) The data miner identifies various data owner who are ready to share their databases.

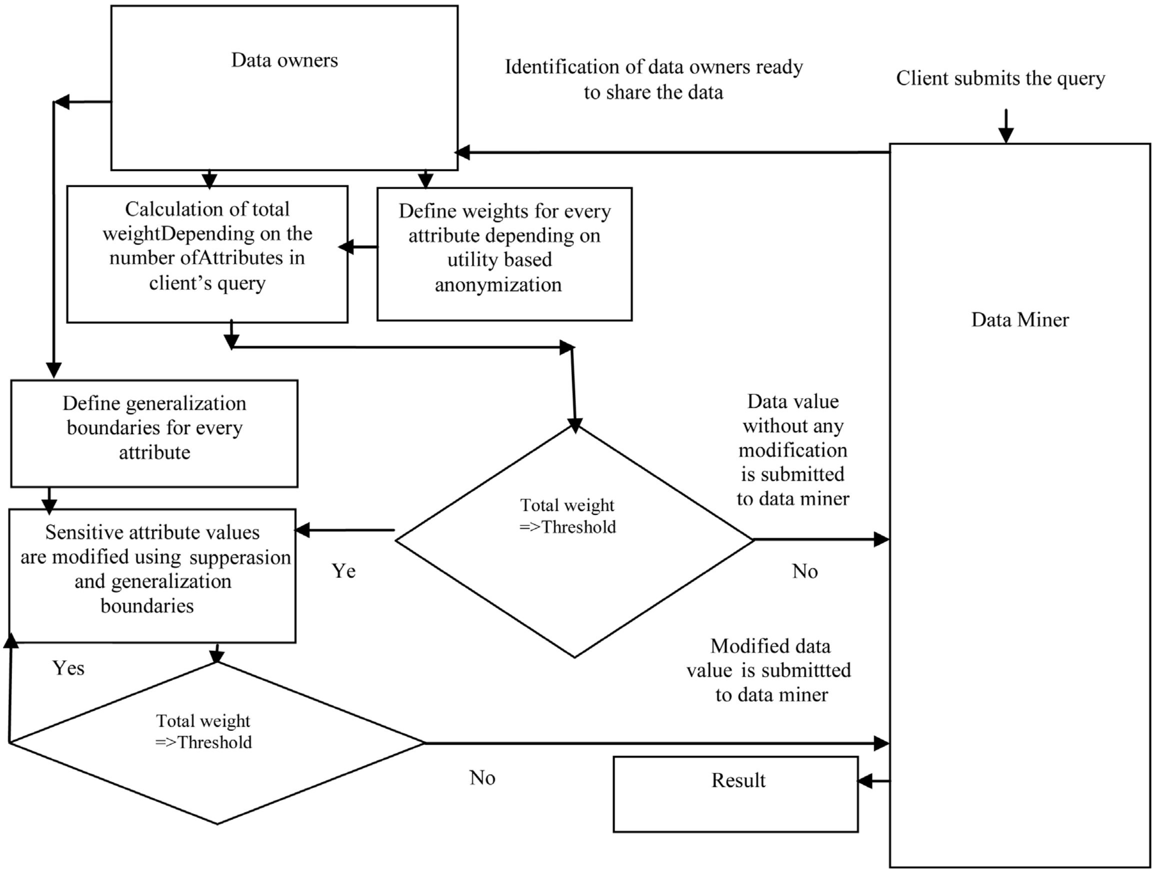
3) The data miner sends the client's query to data owner.

4) The data owner accepts the client's query and calculates the total weight assigned to these attributes, *i.e*. as informed in **Table 4**, taking in account next conditions:

- *Identifier attributes* such as Name and SSN that can be used to identify a record we give it weights equal to threshold to be suppressed first.
- *Quasi-identifier attributes* that has more utility needs like Age takes big weights near from threshold to be latest generalized one if we need to generalize them and others takes different weights according to their utility-based needs.
- *Sensitive or confidential attributes* doesn't take any weights to keep them without generalization for statistical analysis.

5) If the total weight is under the predefined threshold limit of sensitiveness then the data owner released the data requested in the client's query.

6) If the total weight is above the predefined threshold limit of sensitiveness then the related attributes are considered as sensitive attributes and the data values under such Attributes are modified using generalization



Data owners
Identification of data owners ready to share the data
Client submits the query
Calculation of total weightDepending on the number ofAttributes in client's query
Define weights for every attribute depending on utility based anonymization
Data Miner
Define generalization boundaries for every attribute
Data value without any modification is submitted to data miner
Total weight =>Threshold
Sensitive attribute values are modified using supperasion and generalization boundaries
Ye
No
Yes
Modified data value is submittted to data miner
Total weight =>Threshold
No
Result

Figure 4. Proposed algorithm architecture for sensitive attribute identification & generalization using generalization boundaries & utility based anonymization.

Table 4. Shows the data base table, initial microdata set *IM* and their attributes weights.

Attribute	Distinct values	Generalizations	Height	Weights
Name	---------	------------	-----	6
Age	100	10-, 20-, 30-, ····	4	5
Marital status	8	taxonomy tree	2	3
Race	5	Taxonomy tree	2	4
Sex	2	Person	1	2
Salary	2	Sensitive attribute	1	Without

boundaries and suppression as follows:

- Starting with suppress for the identifier attributes which has weight equal to threshold.
- Then starting generalization for quasi-identifier attributes using generalization boundaries as in **Figures 5-8** with attributes that has lowest weight then making *k*-anonymity test and if anonymization doesn't exist we start with next higher weight attributes till reaching suggested *k*-anonymity maintaining *l*-diversity.

7) The modified database portion is transferred to the data miner to perform various data mining operation as per client's request.

Now we consider sample query as sample examples and how researchers treat them according to proposed model.

***Example*: *Query* 1: Select * From Sample database Table.**

From query 1 we find that **Table 5** shows Sample database Table where the client requests for values under these attributes are Name, Marital_status, Age, Race, Sex and Salary. As the total weight of these Attributes is 20 which exceeds threshold limit (6). Hence, the values under these particular attributes are generalized using generalization boundaries and suppressed according to their weights in ascending arrange as follows:

- According to the weights for the five attributes there is one attribute equal to threshold, Name = 6, so we suppress it first.
- Other five attributes are less than threshold limit because all of them under "6", so we use only generalization boundaries according to **Figures 5-8**.
- We start generalization with Sex attribute which has lowest weight "2" using **Figure 8** then making *K*-anonymity test then we complete with next attribute.
- Next attribute is Marital-status which has next weight "3" using **Figure 6** then making *K*-anonymity test then we complete with next attribute.
- Next attribute is Race which has next weight "4" using **Figure 7** then making *K*-anonymity test then complete with next attribute.
- Next attribute is Age which has next weight "5" using **Figure 5** then making *K*-anonymity test which give us *K*-anonymity **Table 6**, which maintain also 2-diversity so we stop here and The modified database portion as it is in **Table 6** is transferred to the data miner to perform various data mining operation as per client's request.

3.3. Experimental Results and Analysis

The researchers compare three models as follows:

- First: Original Data Model as on Real Data Set Adult Database from the UC machine learning repository. The Adult Database contains 32561 records from US Census data. After preprocessing data and removing records containing missing values 30162 records are selected. This database contains 42 attributes from that only 5 attributes are used as **Table 5**. From that four attributes are consider as quazi-identifier and one attribute "Salary" as a sensitive attribute.
- Second: Modified Data Model According to our Utility-Based Anonymization Using Generalization Boundaries to protect Sensitive Attributes Depending on Attributes Sensitivity Weights developed in this paper.
- Third: *K*-anonymity Data Model According to Automatic Detection of Sensitive Attribute Without using Generalization Boundaries and without determining attributes weight; we refer to it as old model in next sections.

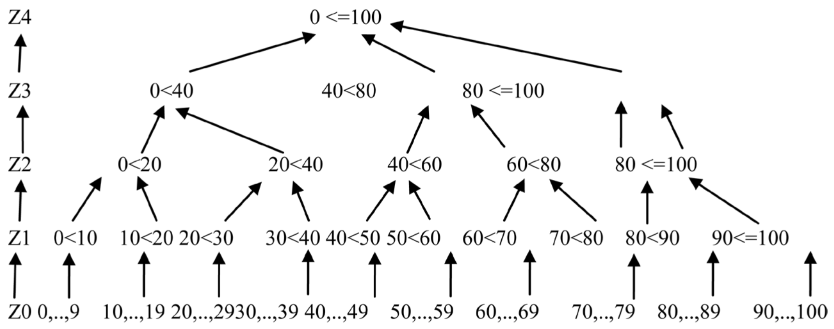


Figure 5. Examples of domain and value generalization hierarchies for age.

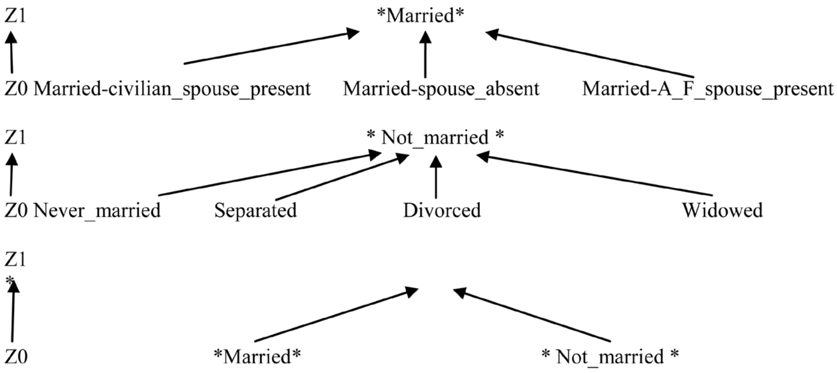


Figure 6. Examples of domain and value generalization hierarchies for marital status.

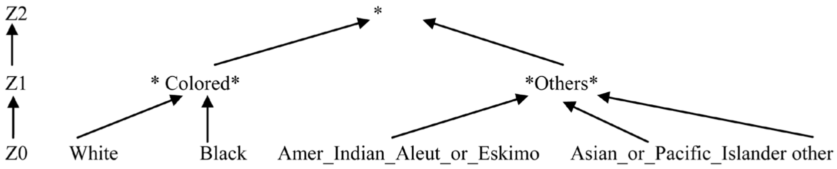


Figure 7. Examples of domain and value generalization hierarchies for race.

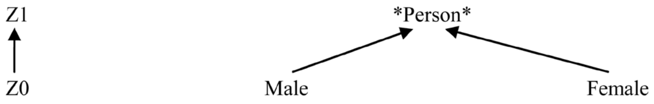


Figure 8. Examples of domain and value generalization hierarchies for sex.

Table 5. Sample database table.

Name weight = 6	Marital_status weight = 3	Age weight = 5	Race weight = 4	Sex weight = 2	Salary
Alice	Never_married	32	White	M	50000+
Lelyan	Divorced	30	Black	F	−50000
Charley	Married-spouse_absent	42	Amer_Indian_Aleut_or_Eskimo	M	50000+
Dave	Married-civilian_spouse_present	40	Asian_or_Pacific_Islander	M	−50000
John	Never_married	20	other	M	−50000
Casey	Widowed	25	Asian_or_Pacific_Islander Wichita	F	50000+

Table 6. Modified database table: MM.

Name weight = 6	Marital_status weight = 3	Age weight = 5	Race weight = 4	Sex weight = 2	Salary
*	Not_married	30 - 40	Colored	Person	50000+
*	Not_married	30 - 40	Colored	Person	−50000
*	Married	40 - 50	Others	Person	50000+
*	Married	40 - 50	Others	Person	−50000
*	Not_married	20 - 30	Others	Person	−50000
*	Not_married	20 - 30	Others	Person	50000+

3.3.1. Cluster Analysis Measure

A cluster can be regarded as a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

The researchers concentrate here to make our experiments on numerical data such as Age then we can generalize the results for all numerical and categorical data. **Figure 9** represents Histogram of original Age attribute data, **Figure 10** represents histogram of proposed model modified data for Age attribute and **Figure 11** represents Histogram of old model data for Age.

From **Figure 9** and **Figure 10** we find that our proposal model has similar histogram as original data which means that we maintain data from being changed. From **Figure 9** and **Figure 11** we find that old model has dissimilar histogram from original data which means that it doesn't maintain data from being changed. So we can say that our proposal model is better with a big degree than old one. **Figure 12** represents Rplot of original and proposed models side by side to explain similarity between them which means that data doesn't changed. **Figure 13** represents Rplot of original and old models side by side to explain dissimilarity between them which means that data extensively changed.

3.3.2. Empirical CDF Measures

These measures assess the differences between the empirical distribution functions obtained from the original and masked data. **Figure 15** represents empirical CDF Measure of original Age attribute data, **Figure 14** represents empirical CDF Measure of proposed modified data for Age attribute and **Figure 16** represents empirical CDF Measure of old Age attribute data. From **Figure 14** & **Figure 15** we see that they has similar CDF distribution which means that our proposed model maintain data distribution from being changed. From **Figure 15** and **Figure 16** we find that old model has dissimilar distribution from original data because there is a big difference between them which means that it doesn't maintain data distribution from being changed. So we can say that our proposal model is better than old one.

3.3.3. Test Analysis for Both Original and Proposed Technique Data

The researchers operate three kinds of test analysis comparing original data (c1) and proposed technique data (c1mm) as follows:

ks-test (Kolmogorov-Smirnov test), *t-test* (Welch Two Sample *t*-test), and *var-test* (F test to compare two variances)

- ***ks*-test (Kolmogorov-Smirnov test):**

The Kolmogorov Goodness-of-Fit Test (Kolmogorov-Smirnov one-sample test) [23]:

a) A test for goodness of fit usually involves examining a random sample from some unknown distribution in order to test the null hypothesis that the unknown distribution function is in fact a known, specified function.

b) We usually use Kolmogorov-Smirnov test to check the normality assumption in Analysis of Variance.

c) A random sample X_1, X_2, $\cdots$, X_n is drawn from some population and is compared with $F^*(x)$ in some way to see if it is reasonable to say that $F^*(x)$ is the true distribution function of the random sample.

d) One logical way of comparing the random sample with $F^*(x)$ is by means of the empirical distribution function $S(x)$.

- ***t*-test (Welch two sample *t*-test) [30]-[33]:**

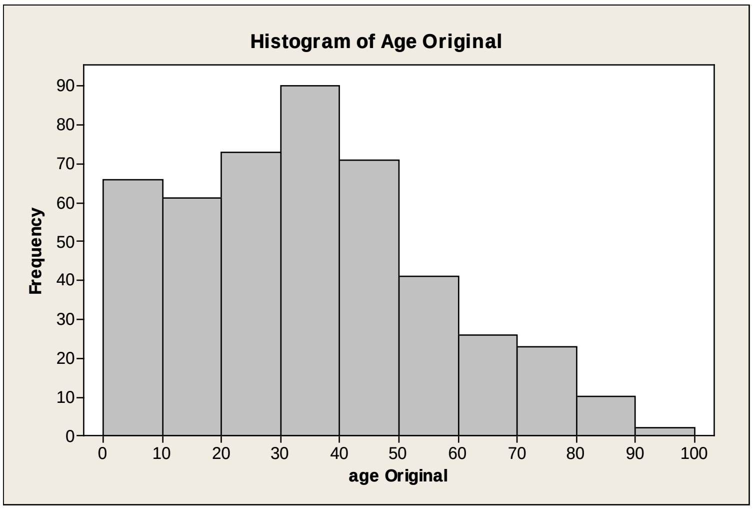


Figure 9. Histogram of original.

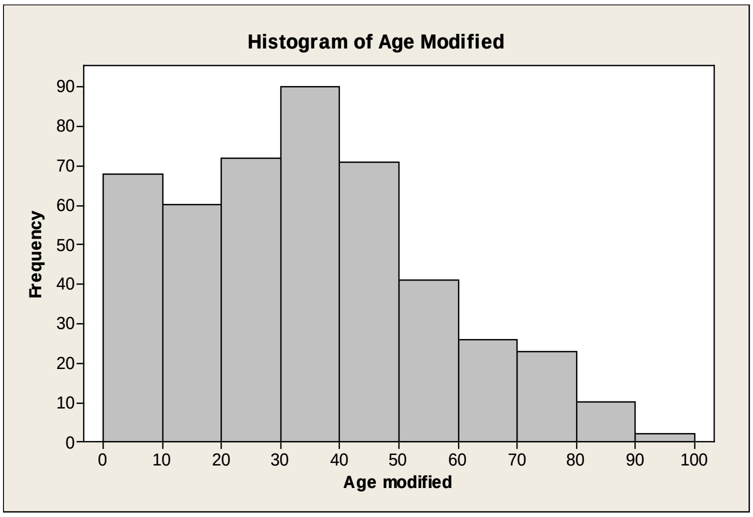


Figure 10. Histogram of proposed model.

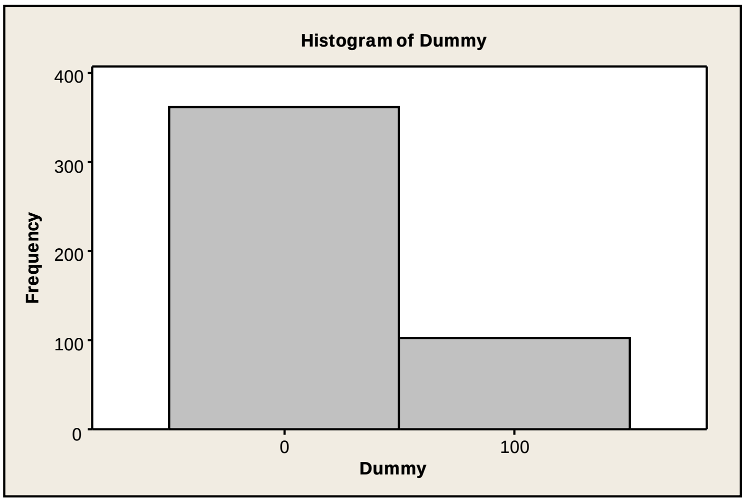


Figure 11. Histogram of old model.

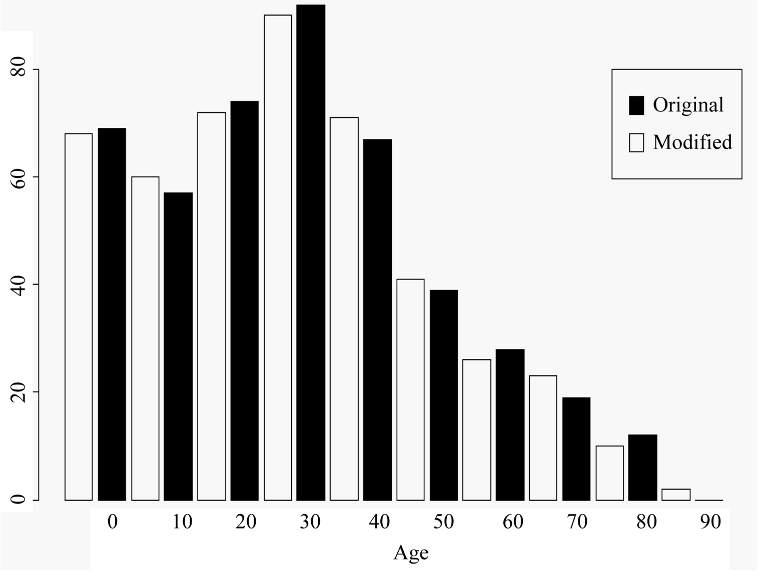


Figure 12. Rplot of orig & proposed.

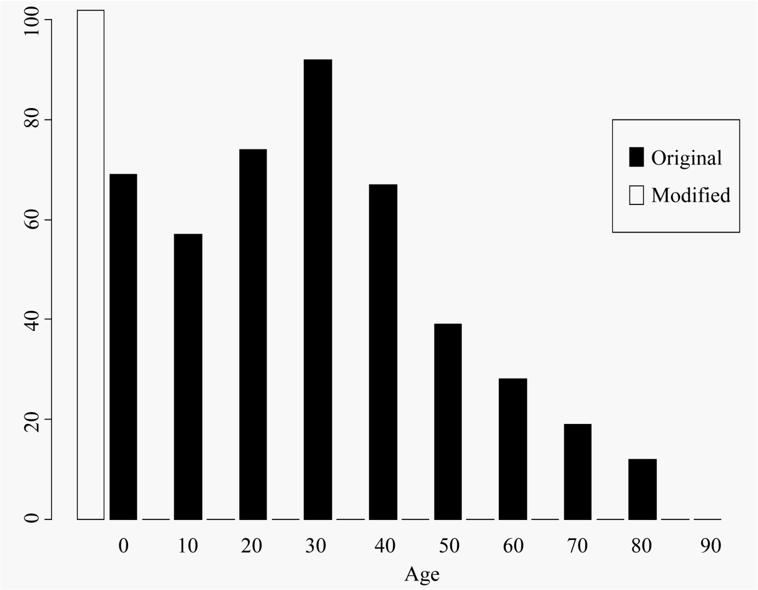


Figure 13. Rplot of orig & old model.

a) *t-test* performs *t-tests* on the equality of means. In the first form, *t-test* tests that *varname* has a mean of #.

b) In the second form, *t*-test tests that *varname* has the same mean within the two groups defined by *groupvar*.

c) In the third form, *t-test* tests that *varname* 1 and *varname* 2 have the same mean, assuming unpaired data.

d) In the fourth form, *t-test* tests that *varname* 1 and *varname* 2 have the same mean, assuming paired data.

e) *t-test* is the immediate form of *t-test*. For the equivalent of a two-sample *t-test* with sampling weights (*pweights*), use the svy: mean command with the over ( ) option, and then use lincom.

f) The *t-test* with Welch (1936, 1938) correction was compared to parametric and permutational *t-tests* for two types of data distributions and for equal and unequal population variances. The result of a *t*-test is identical to that of an anova computed for two groups; the *t-statistic* is the square root of the F-statistic used in anova.

g) The Welch correction was designed to provide a valid *t-test* in the presence of unequal population variances. It consists of using a corrected number of degrees of freedom to assess the significance of the *t-statistic* computed as usual. *n* is the next smaller integer of the value obtained.

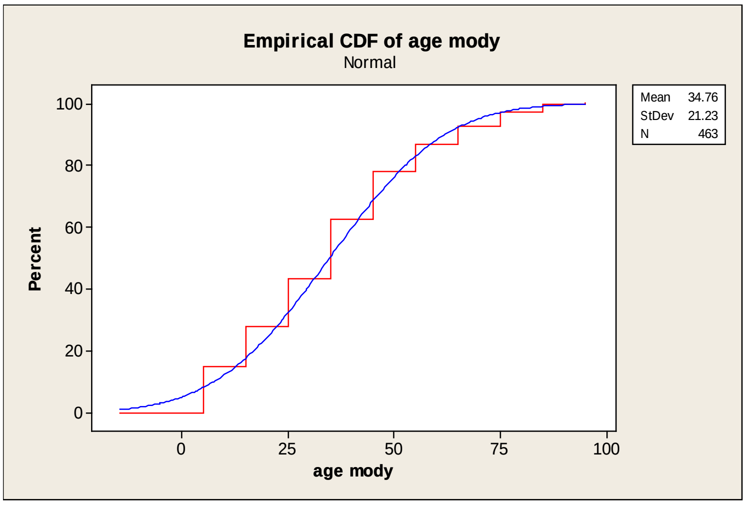


Figure 14. Proposed CDF.

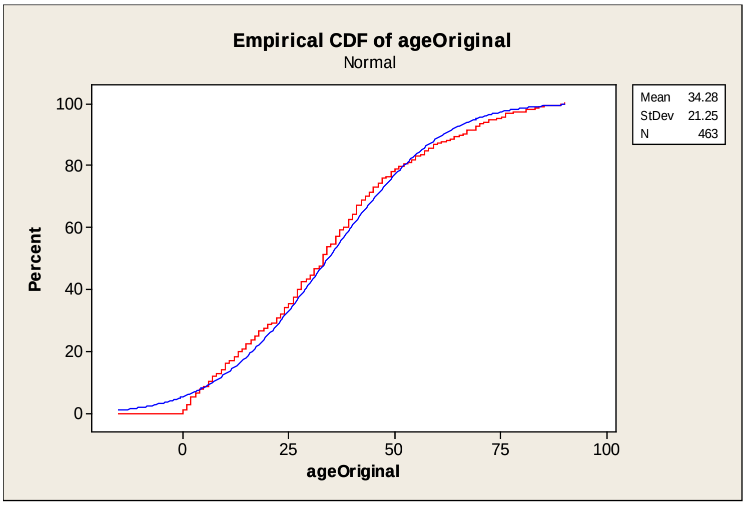


Figure 15. Original CDF.

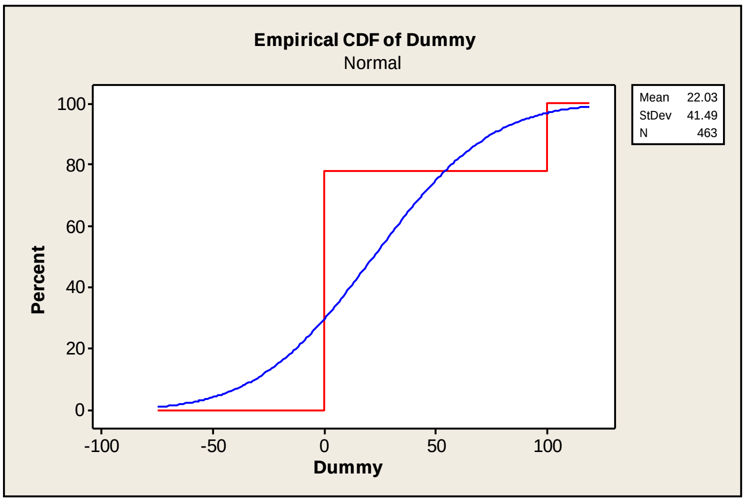


Figure 16. Old model CDF.

- **Var-test (f test to compare two variances):**

a) Performs an *F-test* to compare variances of two samples from normal populations.

b) In *R* the function *var-test* allows for the comparison of two variances using an *F-test.* Although it is possible to compare values of s^2 for two samples, there is no capability within *R* for comparing the variance of a sample, s^2, to the variance of a population, σ^2. There are two ways to interpret the results provided by *R*.

c) First, the *p-value* provides the smallest value of α for which the *F-ratio* is significantly different from the hypothesized value. If this value is larger than the desired α, then there is insufficient evidence to reject the null hypothesis; otherwise, the null hypothesis is rejected.

d) Second, *R* provides the desired confidence interval for the *F-ratio*; if the calculated value falls within the confidence interval, then the null hypothesis is retained.

```
Actual test output:
>ks-test (c1, c1mm)
Two-sample Kolmogorov-Smirnov test
data: c1 and c1mm
D = 0.1037, p-value = 0.0138
Alternative hypothesis: two-sided
Warning message: In ks-test (c1, c1mm): p-values will be approximate in the presence of ties
```

ks-test can be used to conclude that *p-value* = 0.0138 < 0.05 (Significant level) therefore, the distributions are different from one another but not in mean and variances as shown in *t*-test & *f*-test so the different between the two distributions is negligible.

```
Actual test output:
>t-test (c1,c1mm) Welch Two Sample t-test
data: c1 and c1mm
t = −0.3481, df = 923.999, p-value = 0.7278
Alternative hypothesis: true difference in means is not equal to 0
   −3.225762    2.253839  : 95 percent confidence interval
Sample estimates: mean of x 34.27646 mean of y 34.76242
```

t-test can be used to conclude that the two distributions have the same mean because *p*-value = 0.7278 > 0.05

```
Actual test output:
>var-test (c1, c1mm)
F-test to compare two variances
data: c1 and c1mm
F = 1.0017, num df = 462, denom df = 462, p-value = 0.9857
Alternative hypothesis: true ratio of variances is not equal to 1
95percent confidence interval:        0.834491 1.202329
Sample estimates:    ratio of variances         1.001665
```

f-test can be used to conclude that the two distributions have the same variance because *p-value* = 0.9857 > 0.05.

ks-test can be used to conclude that the two distributions are, indeed, significantly different from one another (*p-value* < 0.05) but not in their means and variances. Therefore, the researchers conclude that using generalization boundaries could maintain data utility and doesn't affect data.

3.3.4. Test Analysis for Both Original and Old Technique Data

The researchers operate here three kinds of test analysis comparing original (c1) and old models data (D) as follows:

```
Actual Test Output:
>ks-test (c1, D)
Two-sample Kolmogorov-Smirnov test
data: c1 and D
D = 0.7667, p-value < 2.2e−16
Alternative hypothesis: two-sided
Warning message: In ks-test (c1, D): p-values will be approximate in the presence of ties
Actual Test Output:
>t-test (c1, D) Welch Two Sample t-test
data: c1 and D
t = 5.6528, df = 688.778, p-value 2.312e−08
Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 7.992702 16.499738
Sample estimates: mean of x mean of y 34.2764579 0.2203024
```

t-test can be used to conclude that the two distributions have different mean.

```
Actual Test Output:
>var-test (c1,D)
F test to compare two variances
data: c1 and D
f = 0.2623, num df = 462, denom df = 462, p-value < 2.2e−16
Alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval: 0.2185396 0.3148702
Sample estimates: ratio of variances        2623.197
```

ks-test can be used to conclude that the two distributions are more significantly difference from one another where *p*-value < $2.2e^{-16}$ = 0 and there is different between means and variances as stated in *t-test* & *f-test* with the same above *p*-value.

3.3.5. Test Analysis for Both Original and Proposed Techniques Data

The researchers operate analysis test to show the difference between the three tests according to age intervals and record the output in **Table 7** as follows:

- *ks.test* can be used to conclude that the test is more significant difference as well as the Age interval increases.
- *t.test* and *f.test* can be used to conclude that the two distributions are in significantly difference in means and variances with high *p*-value in the age interval 10 years.
- The above table can be used to conclude that as the Age interval decreases as the test results are more insignificant, *i.e.*, better results. So as we decrease generalization according to our boundaries as we have more utility.

4. Conclusions and Future Work

There are many threats in traditional *k*-anonymity privacy preserving algorithms which consider all of sensitive

Table 7. Age intervals three tests.

Interval	*ks.test* (*p*-value)	*t.test* (*p*-value)	*f.test* (*p*-value)
10 years	0.0138	0.7278	0.9857
20 years	$4.814e^{-07}$	0.5555	0.8743
30 years	$3.331e^{-16}$	0.193	0.6996
40 years	$<2.2e^{-16}$	0.2965	0.7264
50 years	$<2.2e^{-16}$	0.2081	0.6054
0 < +100 (dummy)	$<2.2e^{-16}$	$2.312e^{-08}$	$<2.2e^{-16}$

attribute values at the same level and apply generalization on all, this leads to some issues like, information loss, data utility, and privacy measure.

So there is a need to develop a method which provides the privacy with minimum information loss and maximum data utility. As shown in the related work, many techniques have been developed to automatically determine which part of database needs scrambling; others have been developed to scrambling database using generalization and suppression at all which leads to very height information loss and others making scrambling using generalization boundaries. In this paper, the researchers aimed to propose a new utility-based anonymity model based on sensitivity of attributes in queries which determine which part of database needs changes according to sensitivity weights for each attribute; then only queries having sensitive values are to be changed depending on threshold value. Therefore, according to these weights we can determine which attributes must be generalized first using generalization boundaries, so information loss is reduced and data utility is increased because only sensitive attributes in queries are not only generalized at all but also generalized using generalization boundaries. So both privacy of individual and data utility are preserved. Proposed model combines all previous techniques starting from automatic detection of sensitive attributes by calculating their sensitivity weights using generalization boundaries taking in account utility-based anonymization. Utility based anonymization reduces information loss by starting generalization from low weights in ascending weights manner. After finishing each attribute generalization we can make test for anonymity which decide when we stop generalization for next weight. So our technique maintains the previous three issues, reducing information loss, increasing data utility and maintains privacy, together in one model which proves with practical experiments. The researchers conclude that all privacy techniques should not only give attention to privacy but also take into account information loss and data utility to avoid the trade-offs between them. There are many issues and ideas that can be tackled in future, most of which are related to enhance privacy techniques by combining some anonymity techniques taking into account their advantages and avoiding their disadvantages to maintain both privacy and data utility for useful extracting knowledge.

References

[1] Huang, Z., Du, W. and Chen, B. (2005) Deriving Private Information from Randomized Data. *Proceedings of the ACM SIGMOD Conference on Management of Data*, Baltimore, 37-48. http://dx.doi.org/10.1145/1066157.1066163

[2] Fung, B., Wang, K. and Yu, P. (2005) Top-Down Specialization for Information. *Conference on Data Engineering* (*ICDE*05), 205-216.

[3] Aggarwal, C. and Yu, P. (2008) Models and Algorithms: Privacy-Preserving Data Mining. Springer, Berlin. http://dx.doi.org/10.1007/978-0-387-70992-5

[4] Burnett, L., Barlow-Stewart, K., Pros, A. and Aizenberg, H. (2003) The Gene Trustee: A Universal Identification System That Ensures Privacy and Confidentiality for Human Genetic Databases. *Journal of Law and Medicine*, **10**, 506-513.

[5] Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K. (2003) On the Privacy Preserving Properties of Random Data Perturbation Techniques. *Proceedings of the* 3rd *International Conference on Data Mining*, 19-22 November 2003, 99-106. http://dx.doi.org/10.1109/icdm.2003.1250908

[6] Hussien, A.A., Hamza, N., Shahen, A.A. and Hefny, H.A. (2012) A Survey of Privacy Preserving Data Mining Algorithms. *Yanbu Journal of Engineering and Science*, **5**, 1433H.

[7] LeFevre, K., DeWitt, D. and Ramakrishnan, R. (2005) Incognito: Efficient Full Domain *k*-Anonymity. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Baltimore, 49-60. http://dx.doi.org/10.1145/1066157.1066164

[8] Dayal, U. and Hwang, H.Y. (1984) View Definition and Generalization for Database Integration in Multi-Database Systems. *IEEE Transactions on Software Engineering*, **10**, 628-645. http://dx.doi.org/10.1109/TSE.1984.5010292

[9] Samarati, P. (2001) Protecting Respondents Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, **13**, 1010-1027. http://dx.doi.org/10.1109/69.971193

[10] Pei, J., Xu, J., Wang, Z.B., Wang, W. and Wang, K. (2007) Maintaining *k*-Anonymity against Incremental Updates. *Proceedings of the* 19*th International Conference on Scientific and Statistical Database*.

[11] Miller, J., Campan, A. and Marius, T. (2008) Constrained *K*-Anonymity: Privacy with Generalization Boundaries. *P*3*DM*'08, 26 April 2008, Atlanta.

[12] Winkler, W. (1995) Matching and Record Linkage. Business Survey Methods. Wiley, New York, 374-403. http://dx.doi.org/10.1002/9781118150504.ch20

[13] Hussien, A.A., Hamza, N. and Hefny, A. (2013) Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. *Journal of Information Security*, **4**, 101-112. http://www.scirp.org/journal/jis http://dx.doi.org/10.4236/jis.2013.42012

[14] Sweeney, L. (2002) *K*-Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty*, *Fuzziness*, *and Knowledge-Based Systems*, **10**, 557-570. http://dx.doi.org/10.1142/S0218488502001648

[15] Sweeney, L. (2002) Achieving *K*-Anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty*, *Fuzziness*, *and Knowledge-Based Systems*, **10**, 571-588. http://dx.doi.org/10.1142/S021848850200165X

[16] Xiao, X.K. and Tao, Y.F. (2006) Personalized Privacy Preservation. *Proceedings of the ACM SIGMOD International Conference*, 229-240. http://dx.doi.org/10.1145/1142473.1142500

[17] Wong, R.C.-W., Li, J.Y., Fu, A.W.-C. and Wang, K. (2006) (α, K)-Anonymity: An Enhanced *K*-Anonymity Model for Privacy Preserving Data Publishing. KDD, 754-759.

[18] Bayardo, R. and Agrawal, R. (2005) Data Privacy through Optimal *K*-Anonymity. *Proceedings of the* 21*st International conference on Data Engineering* (*ICDE*), Tokyo, 5-8 April 2005, 217-228. http://dx.doi.org/10.1109/ICDE.2005.42

[19] Iyengar, V. (2002) Transforming Data to Satisfy Privacy Constraints. *Proceedings of the ACM SIGMOD International Conference*, 279-288. http://dx.doi.org/10.1145/775047.775089

[20] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R. (2006) Mondrian Multidimensional *K*-Anonymity. *Proceedings of the* 22*nd International Conference on Data Engineering*, 3-7 April 2006, 25. http://dx.doi.org/10.1109/icde.2006.101

[21] Kamakshi, P. and Vinaya Babu, A. (2012) Automatic Detection of Sensitive Attribute in PPDM. 2012 *IEEE International Conference on Computational Intellgence and Computing Research*.

[22] Xu, J., Wang, W., Pei, J., Wang, X.Y., Shi, B.L. and Fu, A.W.-C. (2006) Utility-Based Anonymization Using Local Recoding. *KDD*'06, Philadelphia, 20-23 August.

[23] Conover, W.J. (1999) Practical Nonparametric Statistical. 3rd Edition, John Wiley & Sons Inc., New York, 428-433.

[24] cs.utdallas.edu/dspl/toolbox/

[25] Kantarcioglu, I.M. and Bertino, E. (2009) Using Anonymized Data for Classification. ICDE, 429-440.

[26] Xiao, X.K. and Tao, Y.F. (2006) Anatomy: Simple and Effective Privacy Preservation. *Proceedings of the* 32*nd International Conference on Very Large Data Bases.* VLDB Endowment, 139-150.

[27] Woo, M.-J., Reitery, J.P., Oganianz, A. and Karr, A.F. (2009) Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *Journal of Privacy and Confidentiality*, **1**, 111-124.

[28] Kaar, A.F., Oganian, A., Reiter, J.P. and Woo, M.-J. (2006) New Measures of Data Utility.

[29] Abad, B. and Kinariwala S.A. (2012) A Novel Approach for Privacy Preserving in Medical Data Mining Using Sensitivity Based Anonymity. *International Journal of Computer Applications*, **42**, 13-16.

[30] Acock, A.C. (2014) A Gentle Introduction to Stata. 4th Edition, Stata Press, College Station.

[31] Boland, P.J. (2000) William Sealy Gosset—Alias "Student" 1876-1937. In: Houston, K., Ed., *Creators of Mathematics*: *The Irish Connection*, University College Dublin Press, Dublin, 105-112.

[32] Dixon, W.J. and Massey Jr., F.J. (1983) Introduction to Statistical Analysis. 4th Edition, McGraw-Hill, New York.

[33] Gleason, J.R. (1999) Sg101: Pair Wise Comparisons of Means, Including the Tukey Wsd Method. *Stata Technical Bulletin*, **47**, 31-37.