A Systematic Evaluated Recommendation on Performance Enhancement Factors and Procedures of Relational Data Anonymization

Kishore Verma S*  Rajesh A  Adeline Johnsana J S

1Research Scholar, Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Kanchipuram, Tamilnadu, India.
*Corresponding author’s Email: kishore.saj3@gmail.com
2Principal, C. Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamilnadu, India.
Email: amrajesh73@gmail.com
3Research Scholar, Department of Computer Science and Engineering, St. Peter’s Institute of Higher Education and Research, Avadi, Tamilnadu, India.
Email: adeline.j.s@gmail.com

Abstract:

The huge quantity of information being gathered among people has brought new demanding situations in ensuring their privacy while this information is mined. Thus privacy preserving data mining has come to be an energetic research arena, in which numerous anonymization approaches have been proposed. Though an extensive number of approaches are available, confined information about their quality of performance has made hard to recognize and select the most suitable approach in given specific mining situations, particularly for experts. In this perspective, we denote quality of privacy preserving data mining in two aspects, privacy, and utility. In this work, we derived two novel metrics null value count and transformation pattern loss that measures privacy and utility and also implemented an efficient examination procedures to evaluate Cell oriented Anonymization (CoA), Attribute oriented Anonymization (AoA) and Record oriented Anonymization (RoA). We explore the novelty of assessment by utilizing a more far-reaching set of situations, distinctive privacy parameters and utility measurements applying on an openly available implementation of those approaches. We framed a series of experiments and complete evaluation to become aware of utility and privacy factors that influence the data mining performances. So as to direct the experts in a choice of approaches we exhibit thorough experimental evaluation, the situation in which one approaches outperforms the other with respect to null value count and transformation pattern loss. Our results facilitate and prompt the need of developing methods that delivers endorsement on choosing scenario based optimum approaches.

Key Terms: K-Anonymity, Transformation Pattern Loss, Null Values Count.

1. Introduction

As of now, the amount of data being produced grows day by day rapidly [1]. From these data, there resides an increasing quantity of privacy records. This reality has pulled in the consideration of the researchers keen on making more custom-made and customized administration of statistical information accessibility. Aimed at this purpose, trade industries of many sectors gather personal data that might be examined under various conditions (for either money related, societal or lawful reasons). Yet this situation has introduced new difficulties to guard the privacy of the individuals involved in the published datasets.
Therefore, Privacy Preserving Data mining has emerged into a region of concern for scholars and experts. The crucial notion of the PPDM framework is that trespassers will be realized as part of data miner’s community, who assumed to reveal peoples information that seems to be sensitive. In this way, PPDM approaches strive to regulate the information of the individuals in coarse granular manner, such that people's privacy is ensured as well as holding the utility of those anonymized data sets. The main principle of PPDM is to generate anonymized data sets that can be used by the variety of data mining tasks. For an open ground of information activities [2], it is difficult to recognize every one of the information beneficiaries. In this way, any data coordinator/ controller concerned within the sharing of non-public knowledge has to adopt privacy preserving techniques [3]. Nevertheless, this is not a trivial process, as data miner's perspective, they are not pure professionals in data privacy [4, 5].

In addition, usually the case that no systems remain to guarantee Anonymization is performed efficaciously in an organization. This pushes the professionals to utilize basic techniques of anonymization(e.g eliminating all direct identifiers that include names and social security numbers), before publishing the data. Regardless, it has been demonstrated that this approach alone isn't sufficient to preserve privacy [6, 7 and 8]. This dispute arises that it is yet conceivable to associate with other data sets or pertaining background knowledge about the person, with the specific goal to make implications about their personality. The re-identity of the person is carried out by way of linking attributes, referred to as quasi-identifiers (QIDs) consisting of gender, date of birth, occupation, ZIP code etc.

Accordingly, different anonymization approaches were proposed in the area of PPDM under the sub-categories of Cell oriented Anonymization (CoA), Attribute oriented Anonymization (AoA) and Record oriented Anonymization (RoA). Still, picking the suitable techniques for these categories are cumbersome for the executors. Not just there is a plenty of anonymization procedures from which one can pick, yet every recently presented procedure guarantees a specific prevalence over the others. This drags up for the situation where the scope of proving the assessment is narrowed in their experimental procedures (using imperfect single metric, skipping other aspectedual environments etc.). Furthermore, there are situations where the implementers propose their own novel metric which intends in favoring the proposed methodology under one side perception. These circumstances may often complicate the executors to misinterpret one specific methodology sole purposes and the way it can be used. Conversely, this attitude could not assure the performances of different algorithms under various circumstances.

Considering these demanding situations, we trust there is a solid necessity to broaden the prevailing assessments on the anonymization procedures to hide a lot of complete set of experimental formations. Necessarily, the goal of this analysis is to offer the experts with a detailed report on the motives that stood as causes for the performance deviations of the anonymization procedures with respect to two novel metrics null Value count and transformation pattern loss. The main theme behind this analysis is to facilitate practical application and implementation of anonymization under CoA, AoA, and RoA.

The experimental results of this work demonstrate the choice of desirable approaches that appropriates the given circumstances (i.e. Data mining Scenarios) relies upon various elements, which includes the property of input data set, preferred privacy necessity (Transformation Pattern loss) and utility necessity( Null values count). In
addition, our outcomes inspire the need of making philosophies that could facilitate data
miners in choosing the most appropriate methodologies in versatile situations. This might
be done with the aid of guidelines approximating the best performing algorithms with
respect to privacy and utility metrics. In this paper the main contributions are:

- A wide comparison of k-Anonymization algorithms in terms of Privacy
  (Transformation Pattern Loss) and Utility (Null value count).
- A widespread examination of the effects of the privacy/utility parameters in Cell
  Oriented Anonymization, Attribute Oriented Anonymization and Record Oriented
  Anonymization.
- Identification and investigation of significant features to consider when choosing an
  anonymization algorithm and data utility metrics.
- A realistic demonstration that the "optimum" procedure in a certain state is subjective
  through numerous factors. This paper is organized as follows: Section 2 delivers some
  basic concepts and survey on related approaches. Section 3 gives the description of k-
  anonymization procedures of CoA, AoA, and RoA. Section 4 explains our comparative
  analysis. Section 5 deliberates the experimental assessment and outcomes. Finally,
  Section 6 appeals conclusions and delivers guidelines for future work.

2. Basic Concepts and Survey on Related Approaches

This section discusses the concepts and methods that form as the fundamental
for our analysis work.

2.1 ARX's K-anonymization

ARX anonymization tool is an open source tool available [9]. This tool
provides feasible and powerful anonymization strategies that can be used by the data
provider to protect their data from privacy disclosure.
ARX supports K-anonymity, K-map, Average Risk, Population Uniqueness, Sample

2.2 ARX K-anonymization data quality models

The output of the anonymization requires an objective function to optimize the
process which term to be data quality model. Thus ARX anonymization tool implements
quality on three perspectives (i) on individual cells, (ii) attributes and (iii) Records often
referred as Cell oriented Anonymization (CoA), Attribute oriented Anonymization (AoA)
and Record oriented Anonymization (RoA). CoA and AoA can be parameterized with the
following aggregate functions, each of them has its own significance

**Rank:** Lexicographically compared measurements information loss of all attribute

**Geometric mean:** Applies geometric mean of information loss to all attribute

**Arithmetic Mean:** Applies arithmetic mean of information loss to all attribute

**Sum:** Sum up all the information loss of the attributes

**Maximum:** Denotes the maximum information loss obtained among all the attribute

The following general purpose transformation model was supported by CoA, AoA,
and RoA.

Cell oriented Anonymization – Loss and Precision
2.3 Literature survey

The choice of suitable methods of anonymization to protect the privacy of published record set is the significant challenge for privacy preserving data publisher. Accordingly, the examination of different anonymization procedures to assess the trade-off between privacy and utility they offer embodies to be the important one in the research. [10] compares clustering based k anonymization algorithms and deliberate the effectiveness of each algorithm with respect to the target application. Likewise, performance comparison for statistical disclosure SDC methods was done by [11,12] by having information loss and disclosure risk as a performance measure.[13] presented adaptive utility based anonymization AVA for big data to assess the quality of data on data mining applications. They compared three different classifications on five different data sets with respect to classification accuracy, precision, f-measure, percent −correct and entropy.[14] the authors have done the evaluation of four anonymization algorithms with respect to runtime and information loss( by varying the suppression limit to 10% and 20%). The information loss is calculated with two metrics Loss and discernibility. However, the information loss and runtime comparative assessment does not reveal about the utility of the approaches.[15] surveyed the classification performance with respect to classification accuracy, precision, and recall. But this method does not account for the null values that are generated during anonymization in assessing the accuracy of the classifier.[16] proposed the methodology that is used to analyze the impact of anonymization on data mining results. The assessment is purely based on the scenario of input data sets that required to be anonymized.[17] assessed the performance of five different anonymization strategy with two efficiency and one privacy metric. The efficiency is computed based on the size of the equivalence class. [18] discussed various measures that are used to assess the quality of information being derived in privacy preserving data mining. The analysis was done on part of anonymization, classification, and clustering. But this approach fails to deliver a common noteworthy measure across anonymization approaches.[19] derived a framework to assess and evaluate the PPDM algorithms with respect to right criteria, privacy loss, information loss, data mining task, modifying data mining algorithm, preserved property data type, in distinguishability level, data dimension etc but presentation of that evaluation criteria does not reveals any experimental proof for their assessment and it is bounded to theoretical evaluation.[20] proposed an evaluation method that is applicable to anonymization procedures of big data. The evaluation is subjected to clustering on part of accuracy, efficiency and bit rate. [21] detailed an assessment report on various k- anonymization strategies with respect to the equivalence class size measured using average class partitioning metric.[22] discussed a theoretical approach to PPDM specifying the merits and demerits of k anonymization with generalization and suppression, randomization and condensation approaches.[23] analyzed the anonymized data set on the classification process. This method relies on ethical releasing the properties of quasi-identity data to have proper gain in privacy preserving data mining, but ethical releasing of statistics may sometime fails to satisfy the major role.[24] proposed a framework to assess the privacy risk that exists with retail data using
Probability of re-identification. [25] presented a classification task on anonymized streaming data. In their experimentation part, they did not consider the suppressed values in the classification process accuracy computation. [26] proposed a framework to analyze the effectiveness of the anonymized data, how far it is useful in data mining application and also evaluated disclosure risk (privacy) of the anonymized data using information theory. [27] proposed the framework that analyses the classification results on anonymized data. Here they analyzed decision tree, logistic regression and SVM classifier and attempt to deliver the best classifier, but this method does not account null values. [28] presented a systematic approach to evaluating the publicly available anonymization approaches. Their results reported suggestions on the impact of the factors that mainly influences the performance of k anonymization. [29] proposed a lightening algorithm for handling high dimensional data. And they performed experimental analysis for 5 different datasets with four different approaches with distinct parameters, this work initiated us to use ARX implementation in our analysis process.

3. Comparison Methodology

3.1 Null Value Impact

For a recordset $S$ and the generalized form $S'$, if $S'$ is achieved by $k$-anonymized. Then if there exist $N(S') = *$, means the original values are replaced with * null value. Null value impact means the presence of null values (null Values count $nV_C$) in the anonymized recordset $S'$ will terribly affect the accuracy of the data mining task. So it is advisable to have a method that possesses less number of null values after anonymization. Thus null Values count $nV_C$ is indirectly proportional to accuracy and directly proportional to Information Loss (IL).

\[ nV_C \propto 1/\text{Classification Accuracy} \]  \hspace{1cm} (1)

\[ nV_C \propto IL \]  \hspace{1cm} (2)

On these two perceptions, the algorithms are analyzed.

3.2 Transformation pattern Loss ($T.p.L$)

Transformation pattern is the reference string represented as a level of generalization hierarchy adopted by each attribute involved in anonymization. Transformation pattern Loss is the cosine distance between the expected transformation pattern and actual transformation pattern.

\[ P = \prod_{i=1}^{n}(Ae) \]  \hspace{1cm} (3)

Transformation pattern is formed by appending each attributes transformation level string after anonymization. Let $P$ be the vector of the expected Transformation pattern representation. $Ae$ is each attribute’s level of transformation in the expected form.

\[ Q = \prod_{i=1}^{n}(Aa) \]  \hspace{1cm} (4)

$Q$ be the vector of actual transformation pattern representation. $Aa$ is each attribute's levels of transformation in actual form.

Transformation pattern Loss ($T.p.L$) = Cosine distance($P,Q$)  \hspace{1cm} (5)

3.3 Method of Evaluation

Our proposed approach of evaluation to assess the performance of anonymization algorithms purely based on two factors i) Number of null Values ($nV_C$) and ii) Transformation pattern Loss ($T.p.L$). In this assessment, the anonymization
methodology that possesses less number of null values and low (nearly equivalent to 0) transformation pattern loss is characterized as the optimized one in terms of performance. This assessment process is inspired and stimulated by a perception that anonymizing the recordset without specifying the expected level of generalization in the hierarchy lead to produce more number of null value i.e. The ultimate aim of this review research is to generalize the relational record set with minimum null values and transformation pattern loss. In this analysis we come to know that generalization with minimum null values and transformation pattern Loss can be achieved by two predefined factors a) Attribute weight modulation and b) Attribute generalization Level modulation.

a) **Attribute Weight Modulation**
Each attribute involved in the generalization process can be assigned a weight scaling from (0 to 1). Based on which actual generalization lattice may be generated nearer to the expected generalization lattice.

**Steps to Modulate Attribute Weight**
1. Analyse the significance of each attribute in the recordset, select the appropriate one.
2. Anonymize and compute the Transformation pattern without modulating the attribute weights.
3. Modulate the attribute weights accordingly, analyze the effect of Transformation pattern after anonymization.
4. If the generated Transformation pattern is approximately equivalent to the expected Transformation pattern. Declare it is the optimum one.

b) **Attribute Generalization Level Modulation**
Each attribute consists of generalization levels according to the created generalization hierarchy. Based on the level of generalization can be predefined before anonymization without affecting or violating the k-anonymization property. Usually modulating the attribute generalization level before anonymization, may some extent attempt to violate the k-anonymization property. Where these need to be keenly monitored and the records that violate are suppressed.

**Modulating the generalization levels can be done as follows**
(i) For each attribute
   Identify the minimum and maximum generalization level from the generalization hierarchy
   (ii) Configure each attribute’s generalization level of anonymization Min and Max values as per the identified Minimum and Maximum generalization level.
   (iii) Anonymize the recordset
   (iv) Check for the violation of k-anonymization property, if true suppress the target records.
   (v) Compute the actual Transformation pattern

**Comparative Analysis**
Anonymize the recordset for different k values (k=2 to 20, k=k+2 on each iteration) with and without modulating attribute weights and generalization level.

(a) Compute the number of null values present in the anonymized recordset $S^*$. 
$$nV_c = |S^* = *|.$$
(b) Compute the expected Transformation pattern \( P \) from the generalization hierarchy.

(c) Compute the actual Transformation pattern \( Q \) and for each \( k \) anonymized recordset ranging from 2 to \( n \), for CoA, AoA and RoA.

Where \( n \) denotes the user parameter of \( k \) upper bound of experimentation.

(d) Calculate the Transformation pattern Loss by equation 5

(e) Compare and declare the optimum \( k \)-anonymization strategy in CoA, AoA, and RoA.

4. Experimental Evaluation

This section discusses the experimental settings adopted and the results obtained in this research work,

4.1 Execution Platform

To enrich our analysis we developed and utilized two procedures i) Null value computation and ii) \( k \)-anonymization. Null value computation is developed in Dot net framework having SQL server as the backend. \( k \)-anonymization is executed by the widely used open source anonymization tool ARX available in[30]. The experiments were executed on a machine running 64-bit windows 8.1, Intel Core i5 processor with 8GB.

4.2 Experimental Initiatives

In this experimental assessment, UCI Machine learning repository -Adult data set is used. This dataset consists of 30162 records with 9 attributes. According to our \( k \)-anonymization strategy, the dataset is categorized as

(i) Qid – Quasi Identifier, are the attributes which are considered as the linking attributes that are exposed to linking attacks. (ii) Sa- Sensitive attributes are the attributes which should not be correlated with the specific individual as an account of linking attacks. (iii) Ia-Identifying attributes are the direct signifiers of the records i.e. explicitly reveals the identity of the individual. Each attribute needs special consideration in \( k \)-anonymization process Qid’s need to generalized or suppressed to support \( k \)-anonymization, Sa’s need to protected from correlating with Qid’s and Ia’s need to be eliminated from publishing. Here in our experimentation from 9 attributes, first eight attributes are taken as Qid’s, the last attribute is considered as Sa. The detailed description of attributes with their generalization level as in the created hierarchy is given in table 1. In Cell oriented Anonymization and Attribute oriented Anonymization two kinds of data quality models are executed under five aggregate functions namely Geometric Mean, Arithmetic Mean, Sum, Rank, and Maximum. Whereas in Record oriented Anonymization four kinds of data quality models are executed and this approach will not support any aggregate functions. The experiment is carried with and without modulating attribute’s weight and with and without modulating attribute’s generalization levels. For these experiments, Number of Null values \((nV_C)\) and Transformation pattern Loss \((T.p.L)\) that arise on account of anonymizing the recordset with \( k=2 \) to 20, \( k=k+2 \) on each iteration is computed.

<table>
<thead>
<tr>
<th>Table 1 Data Set Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of the Attribute</td>
</tr>
<tr>
<td>1. Sex</td>
</tr>
<tr>
<td>2. Age</td>
</tr>
<tr>
<td>3. Race</td>
</tr>
<tr>
<td>4. Marital Status</td>
</tr>
</tbody>
</table>
4.3 Results

4.3.1 Attribute Weight Modulation

The experimentation is performed in analyzing the best method that can deliver minimum $nV_C$ and $T.p.L$ on varying the attribute’s weight with different values. In this analysis, we inferred that Precision/ Maximum is the only data quality model that highly responds ably to attribute weight modulation scenario and able generate optimal transformation pattern with less number $nV_C$ and $T.p.L$. Other models also vary in performance, when subjected to attribute weight modulation, but not upon significances achieved in modulating the attribute generalization levels.

4.3.2 Attribute Generalization Level Modulation

4.3.2.1 Cell oriented Anonymization (CoA)

In Cell oriented Anonymization two kinds of data quality models i) Loss and ii) Precision are experimented with and shown in table 1 and 2. From table1 it is clearly inferable that $nV_C$ and $T.p.L$ for the anonymized recordset without modulating attribute generalization levels is much more than the $nV_C$ and $T.p.L$ of anonymized recordsets by modulated attribute generalization levels. Usually, the increase in the null values count will terribly degrade the accuracy of the data mining tasks. In this experimentation, null value count $nV_C$ of loss (maximum, rank) and precision (maximum, rank) anonymization with modulated attribute generalization level seems to be lesser, from this fact alone is not possible to derive the best one. From Table 2 the Transformation pattern Loss ($T.p.L$) of Precision (Geometric Mean, Arithmetic Mean, and Sum) seems to be minimum. On part of this analysis of two factors, $nV_C$ and $T.p.L$ different data quality models seem to be the best, however, to recommend the experts with the optimum model that are well-being with respect to $nV_C$ and $T.p.L$ is required. According to this perception we recommend loss (Geometric Mean, Arithmetic Mean, and Sum) is the optimum data quality model that generates second optimistic values on these two measures $nV_C$ and $T.p.L$. 

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.</td>
<td>Education</td>
<td>Qid</td>
</tr>
<tr>
<td>6.</td>
<td>Native Country</td>
<td>Qid</td>
</tr>
<tr>
<td>7.</td>
<td>Work Class</td>
<td>Qid</td>
</tr>
<tr>
<td>8.</td>
<td>Occupation</td>
<td>Qid</td>
</tr>
<tr>
<td>9.</td>
<td>Salary</td>
<td>Sa</td>
</tr>
</tbody>
</table>
The document contains tables and figures discussing the effects of data anonymization techniques on various metrics such as precision, number of null values, arithmetic mean, geometric mean, and maximum values. The tables include comparisons between Cell Oriented Anonymization and other methods. The figures illustrate the transformation pattern loss and comparative results.

### Figures

**Figure 1** Null value Comparative results of Cell oriented Anonymization Approaches with and without modulating attribute levels.

**Figure 2** Transformation pattern Loss Comparative results of Cell oriented Anonymization Approaches with and without modulating attribute levels.

### 4.3.2.2 Attribute-oriented Anonymization (AoA)

In Attribute oriented Anonymization, suppose two data quality models namely (i) Non-Uniform Entropy and (ii) Normalised Non-Uniform Entropy are experimented with and shown in table 2 and 3. From table 3 the geometric mean aggregate function of Non-uniform Entropy with modulating attribute levels of anonymization seems to produce lesser nFre whereas from Table 4 it can be seen that Normalized Non-Uniform Entropy (Rank and Maximum) seems to generate less Tp.L for the anonymized recordset with modulating attribute generalization levels. As with results shown in table 3 and table 4, the best approach with respect to nFre and Tp.L varies. Therefore as default, the second best approach that is common to both measures will be the optimistic one. In this context, we state the Normalized Non-Uniform Entropy (Rank and Maximum) would be the optimal anonymization technique for AoA.
4.3.2.3 Record-oriented Anonymization (RoA).

Record-oriented anonymization experiments four data quality models namely (i) Discrimibility, (ii) Average Equivalence Class, (iii) Ambiguity and (iv) Entropy-Based Model. The results obtained are represented in Table 5 and Table 6. In Table 5 Discrimibility and Ambiguity data quality models generate a minimum number of $n_V^C$ whereas the other two models whereas on the computation of $T, p, L$. Entropy-Based Model causes less $T.p.L$. Thus these measures have two contradictory data quality models with minimum values, the obvious process is to analyze the optimal one on these measures. Accordingly, Discrimibility and Ambiguity are the best models on $n_V^C$ and the second
minimum on $T, p, L$. Hence we state that Discernibility and Ambiguity data quality models with modulating attribute levels are the optimal methodology of anonymization with higher data utility rate of RoA.

<table>
<thead>
<tr>
<th>K Value</th>
<th>Discernibility</th>
<th>Average Equivalence Class</th>
<th>Ambiguity</th>
<th>Entropy based Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>151062</td>
<td>153506</td>
<td>181029</td>
<td>71762</td>
</tr>
<tr>
<td>4</td>
<td>181232</td>
<td>153792</td>
<td>181549</td>
<td>72328</td>
</tr>
<tr>
<td>6</td>
<td>181309</td>
<td>160095</td>
<td>120976</td>
<td>94764</td>
</tr>
<tr>
<td>8</td>
<td>181776</td>
<td>172954</td>
<td>151122</td>
<td>10777</td>
</tr>
<tr>
<td>10</td>
<td>181879</td>
<td>157518</td>
<td>151524</td>
<td>105526</td>
</tr>
<tr>
<td>12</td>
<td>181879</td>
<td>162533</td>
<td>213118</td>
<td>108612</td>
</tr>
<tr>
<td>14</td>
<td>181879</td>
<td>170773</td>
<td>213466</td>
<td>98412</td>
</tr>
<tr>
<td>16</td>
<td>211978</td>
<td>199519</td>
<td>151396</td>
<td>83602</td>
</tr>
<tr>
<td>18</td>
<td>211978</td>
<td>153506</td>
<td>151396</td>
<td>107395</td>
</tr>
<tr>
<td>20</td>
<td>211334</td>
<td>167456</td>
<td>151338</td>
<td>103190</td>
</tr>
</tbody>
</table>

On modulating attribute levels of quasi-identifiers in generalization

<table>
<thead>
<tr>
<th>K Value</th>
<th>Discernibility</th>
<th>Average Equivalence Class</th>
<th>Ambiguity</th>
<th>Entropy based Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.151332</td>
<td>0.282863</td>
<td>0.168478</td>
<td>0.641431</td>
</tr>
<tr>
<td>4</td>
<td>0.168478</td>
<td>0.236237</td>
<td>0.151332</td>
<td>0.475858</td>
</tr>
<tr>
<td>6</td>
<td>0.168478</td>
<td>0.266201</td>
<td>0.092885</td>
<td>0.494924</td>
</tr>
<tr>
<td>8</td>
<td>0.304299</td>
<td>0.386059</td>
<td>0.208743</td>
<td>0.323877</td>
</tr>
<tr>
<td>10</td>
<td>0.289953</td>
<td>0.363604</td>
<td>0.177808</td>
<td>0.323877</td>
</tr>
<tr>
<td>12</td>
<td>0.199359</td>
<td>0.255792</td>
<td>0.213786</td>
<td>0.283885</td>
</tr>
<tr>
<td>14</td>
<td>0.319586</td>
<td>0.346725</td>
<td>0.213643</td>
<td>0.208845</td>
</tr>
<tr>
<td>16</td>
<td>0.331847</td>
<td>0.46967</td>
<td>0.183503</td>
<td>0.208845</td>
</tr>
<tr>
<td>18</td>
<td>0.331847</td>
<td>0.46967</td>
<td>0.331847</td>
<td>0.208845</td>
</tr>
<tr>
<td>20</td>
<td>0.331847</td>
<td>0.361085</td>
<td>0.198612</td>
<td>0.208845</td>
</tr>
</tbody>
</table>

Mean: 0.2597026 0.3437906 0.1940637 0.3379232

Figure 5 Null value Comparative results of Record oriented Anonymization Approaches with and without modulating attribute levels

<table>
<thead>
<tr>
<th>K Value</th>
<th>Discernibility</th>
<th>Average Equivalence Class</th>
<th>Ambiguity</th>
<th>Entropy based Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.151332</td>
<td>0</td>
<td>0.105773</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.151332</td>
<td>0</td>
<td>0.105773</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.151332</td>
<td>0</td>
<td>0.105773</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.031037</td>
<td>0.043817</td>
<td>0.031037</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.031037</td>
<td>0.043817</td>
<td>0.031037</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0.019419</td>
<td>0.056544</td>
<td>0.019419</td>
<td>0.035236</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0.109129</td>
<td>0.026271</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0.109129</td>
<td>0.026271</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0.109129</td>
<td>0.026271</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.109129</td>
<td>0.026271</td>
<td>0</td>
</tr>
</tbody>
</table>

Mean: 0.0398812 0.0580694 0.0398812 0.014032

Mean: 0.0398812 0.0580694 0.0398812 0.014032

Figure 6 Transformation pattern Loss Comparative results of Record oriented Anonymization Approaches with and without modulating attribute levels

5. Conclusion and future work

In this work, we employed a well-defined data quality assessment procedures with respect to two newly derived terms null value count $nV_c$ and Transformation pattern Loss ($T, p, L$). ARX open source tool is utilized in implementing various k anonymization data quality models and the evaluation of these measures is done for Cell oriented Anonymization (CoA), Attribute orientation Anonymization (AoA) and Record oriented Anonymization (RoA). Our evaluation is the systematic procedures and excels in recommending the experts with the optimistic methodology of CoA, AoA, and RoA. This work will provide an opportunity for the experts to gain more insight in selecting the appropriate methods with higher utility rate for CoA, AoA, and RoA on different scenarios.

In the future, we plan to assess the quality of the published anonymized dataset and also assess the quality of the anonymizing the recordset with null values.
References


