International Journal of Pure and Applied Mathematics
Volume 120 No. 5 2018, 1175-1188
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue
http://www.acadpubl.eu/hub/

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

Kishore Verma S[1]* Rajesh A[2]Adeline Johnsana J S[3]

[1]Research Scholar,Department of Computer Science and Engineering, Sri ChandrasekharendraSaraswathiViswaMahavidyalaya University, Kanchipuram, Tamilnadu, India.
*Corresponding author's Email: kishore.saj3@gmail.com
[2] Principal, C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamilnadu, India.
Email: amrajesh73@gmail.com
[3]Research 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 tonull 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 adoptprivacy 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 aspectual 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 uniqueness,L-diversity,T-closeness. K anonymization transformation modelsgeneralization, suppression, and micro aggregation.

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

Attribute-oriented Anonymization – Non-Uniform Entropy and Normalized Non-Uniform Entropy
Record-oriented Anonymization – Average equivalence class size, Discernibility, Ambiguity, and Entropy-based model

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 statisticsmay 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]presentedan analysis of the 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 $S$of 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).

$$nV_C \propto {}^{1}/_{Classification_{Accuracy}} \quad (1)$$

$$nV_C \propto IL \quad (2)$$

On these two perceptions, the algorithms are analyzed.

3.2 Transformation pattern Loss($\boldsymbol{T.p.L}$)

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.

$$P = \coprod_{i=1}^{n}(Ae) \quad (3)$$

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.

$$Q = \coprod_{i=1}^{n}(Aa) \quad (4)$$

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

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

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 recordsetS^*.

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 1 Data Set Description

	Name of the Attribute	Attribute Category	Generalization Levels in Hierarchy
1.	Sex	Qid	0
2.	Age	Qid	1-3
3.	Race	Qid	1
4.	Marital Status	Qid	1

5.	Education	Qid	1-2
6.	Native Country	Qid	1
7.	Work Class	Qid	1
8.	Occupation	Qid	1
9.	Salary	Sa	0

4.3 Results

4.3.1 Attribute Weight Modulation

The experimentation is performed in analyzingthe best method that can able 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 $T.p.L$.

	Cell Oriented Anonymization									
	Loss					Precision				
	Number of Null Values					Number of Null Values				
K Value	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum
2	22329	26982	26982	9522	9522	54930	54930	54930	86157	86157
4	19584	19584	19584	20835	20835	83879	58802	58802	24390	24390
6	26091	26091	26091	29025	29025	81247	65842	65842	32868	32868
8	32616	32616	32616	35946	35946	84866	70666	70666	40572	40572
10	36072	36072	36072	41283	41283	87358	87358	87358	46107	46107
12	50642	40977	40977	47016	47016	90214	90214	90214	52128	52128
14	52346	31437	31437	51489	51489	106782	92300	92300	55548	55548
16	53866	34299	34299	55872	55872	107316	93819	93819	60606	60606
18	55442	36702	36702	59877	59877	108810	95310	95310	64467	64467
20	58378	38655	38655	64692	64692	110130	98206	98206	68472	68472
Mean	40736.6	32341.5	32341.5	41555.7	41556	91553.2	80744.7	80744.7	53132	53131.5
On Modulating Attribute Levels of Quasi Identifiers in generalization										
2	8352	8352	8352	3492	3492	12951	12951	12951	3492	3492
4	13194	13194	13194	9153	9153	28440	28440	28440	9153	9153
6	18099	18099	18099	12753	12753	26937	26937	26937	12753	12753
8	15669	15669	15669	15669	15669	22581	22581	22581	15669	15669
10	18261	18261	18261	18261	18261	26388	26388	26388	18261	18261
12	20898	20898	20898	20898	20898	30258	30258	30258	20898	20898
14	24687	24687	24687	24687	24687	33840	33840	33840	24687	24687
16	27297	27297	27297	27297	27297	37728	37728	37728	27297	27297
18	29547	29547	29547	29547	29547	40536	40536	40536	29547	29547
20	31554	31554	31554	31554	31554	44208	44208	44208	31554	31554
Mean	20755.8	20755.8	20755.8	19331.1	19331	30386.7	30386.7	30386.7	19331	19331.1

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

	Cell Oriented Anonymization									
	Loss					Precision				
	Transformation pattern Loss					Transformation pattern Loss				
K Value	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum
2	0.39391	0.4302	0.4302	0.52191	0.52191	0.54165	0.54165	0.54165	0.62204	0.62204
4	0.24407	0.24407	0.24407	0.2662	0.2662	0.47085	0.46548	0.46548	0.11808	0.11808
6	0.24407	0.24407	0.24407	0.2662	0.2662	0.42265	0.46548	0.46548	0.11808	0.11808
8	0.13037	0.13037	0.13037	0.12294	0.12294	0.24093	0.25464	0.25464	0.05132	0.05132
10	0.13037	0.13037	0.13037	0.12294	0.12294	0.24093	0.24093	0.24093	0.05132	0.05132
12	0.13037	0.09861	0.09861	0.15385	0.15385	0.27373	0.27373	0.27373	0.0755	0.0755
14	0.1092	0.05719	0.05719	0.08479	0.08479	0.17705	0.17705	0.17705	0.05719	0.05719
16	0.08075	0.08713	0.08713	0.08479	0.08479	0.12462	0.17705	0.17705	0.05719	0.05719
18	0.08075	0.05719	0.05719	0.08479	0.08479	0.12462	0.17705	0.17705	0.05719	0.05719
20	0.08075	0.08713	0.08713	0.08479	0.08479	0.12462	0.17705	0.17705	0.05719	0.05719
Mean	0.16246	0.15663	0.15663	0.17932	0.17932	0.27417	0.29501	0.29501	0.12651	0.12651
On Modulating Attribute Levels of Quasi Identifiers in genaralization										
2	0.04382	0.04382	0.04382	0.10577	0.10577	0	0	0	0.10577	0.10577
4	0.12169	0.12169	0.12169	0.10577	0.10577	0	0	0	0.10577	0.10577
6	0.12169	0.12169	0.12169	0.10577	0.10577	0.04382	0.04382	0.04382	0.10577	0.10577
8	0.03104	0.03104	0.03104	0.03104	0.03104	0.0658	0.0658	0.0658	0.03104	0.03104
10	0.03104	0.03104	0.03104	0.03104	0.03104	0.0658	0.0658	0.0658	0.03104	0.03104
12	0.01942	0.01942	0.01942	0.01942	0.01942	0.06905	0.06905	0.06905	0.01942	0.01942
14	0	0	0	0	0	0.02627	0.02627	0.02627	0	0
16	0	0	0	0	0	0.02627	0.02627	0.02627	0	0
18	0	0	0	0	0	0.02627	0.02627	0.02627	0	0
20	0	0	0	0	0	0.02627	0.02627	0.02627	0	0
Mean	0.03687	0.03687	0.03687	0.03988	0.03988	0.03496	0.03496	0.03496	0.03988	0.03988

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 supports 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 nV_C whereas from Table 4 it can be seen that Normalized Non-Uniform Entropy (Rank and Maximum) seems to generate less $T.p.L$ 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 nV_C and $T.p.L$ 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.

	Attribute oriented Anonymization (AoA)									
	Non Uniform Entropy					Normalised Non Uniform Entropy				
	Number of Null Values					Number of Null Values				
K Value	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum
2	150842	83102	83102	126522	126522	73792	54930	54930	31986	31986
4	180972	117516	117516	147698	147698	76242	76242	76242	29583	29583
6	180984	105894	105894	160593	160593	69991	50754	50754	36999	36999
8	180984	110628	110628	181494	181494	72588	54450	54450	44361	44361
10	180984	116586	116586	187026	187026	98136	76669	76669	48744	48744
12	181017	133493	133493	187810	187810	99348	76501	76501	42156	42156
14	181017	135728	135728	159494	159494	100308	80148	80148	45747	45747
16	181017	138563	138563	161514	161514	100830	80085	80085	48771	48771
18	181068	141438	141438	163174	163174	101526	101526	101526	51597	51597
20	181068	143298	143298	165006	165006	102414	102414	102414	55233	55233
Mean	177995	122625	122625	164033	164033	89517.5	75371.9	75371.9	43518	43517.7
	On Modulating Attribute Levels of Quasi Identifiers in generalization									
2	3492	12951	12951	8748	8748	12951	12951	12951	5688	5688
4	9153	28440	28440	19926	19926	13194	13194	13194	13185	13185
6	12753	28962	28962	28962	28962	18099	18099	18099	18441	18441
8	15669	36459	36459	36459	36459	15669	22581	22581	15669	15669
10	18261	43758	43758	43758	43758	18261	18261	18261	18261	18261
12	20898	47925	47925	47925	47925	20898	20898	20898	20898	20898
14	24687	52533	52533	52533	52533	24687	24687	24687	24687	24687
16	27297	39528	39528	58149	58149	27297	27297	27297	27297	27297
18	29547	42525	42525	62577	62577	29547	29547	29547	29547	29547
20	31554	46827	46827	66753	66753	31554	31554	31554	31554	31554
Mean	19331.1	37990.8	37990.8	42579	42579	21215.7	21906.9	21906.9	20523	20522.7

Figure 3 Null value Comparative results of Attribute oriented Anonymization Approaches with and without modulating attribute levels

	Attribute oriented Anonymization(AoA)									
	Non Uniform Entropy					Normalised Non Uniform Entropy				
	Transformation pattern Loss					Transformation pattern Loss				
K Value	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum	Geometric Mean	Arithmetic Mean	Sum	Rank	Maximum
2	0.192219	0.191878	0.191878	0.37006	0.370059	0.47085	0.541651	0.54165	0.15485	0.154846
4	0.114578	0.219601	0.219601	0.2662	0.266201	0.407001	0.407001	0.407	0.11808	0.118083
6	0.114578	0.14958	0.14958	0.2662	0.266201	0.319664	0.355342	0.35534	0.11808	0.118083
8	0.061657	0.288488	0.288488	0.42845	0.428452	0.177808	0.19096	0.19096	0.05132	0.051317
10	0.061657	0.288488	0.288488	0.42845	0.428452	0.19171	0.208845	0.20885	0.05132	0.051317
12	0.047074	0.213204	0.213204	0.42711	0.427108	0.139172	0.16795	0.16795	0.03923	0.039231
14	0.042927	0.331352	0.331352	0.40519	0.405188	0.092782	0.092782	0.09278	0.04742	0.047421
16	0.042927	0.331352	0.331352	0.40519	0.405188	0.096304	0.104331	0.10433	0.04742	0.047421
18	0.042927	0.331352	0.331352	0.40519	0.405188	0.096304	0.092782	0.0963	0.04742	0.047421
20	0.042927	0.331352	0.331352	0.33135	0.331352	0.096304	0.092782	0.0963	0.04742	0.047421
Mean	0.076347	0.267665	0.267665	0.37334	0.373339	0.20879	0.225443	0.22615	0.07226	0.072256
	On Modulating Attribute Levels of Quasi Identifiers in generalization									
2	0.105773	0	0	0.04382	0.043817	0	0	0	0.05654	0.056544
4	0.105773	0	0	0.04382	0.043817	0.12169	0.12169	0.12169	0.05654	0.056544
6	0.105773	0.043817	0.043817	0.04382	0.043817	0.12169	0.12169	0.12169	0.05654	0.056544
8	0.031037	0.1	0.1	0.1	0.1	0.020204	0.020204	0.0202	0.03104	0.031037
10	0.031037	0.1	0.1	0.1	0.1	0.03107	0.03107	0.03107	0.03104	0.031037
12	0.019419	0.03526	0.03526	0.03526	0.03526	0.01949	0.01949	0.01949	0.01942	0.019419
14	0	0.105573	0.105573	0.10557	0.105573	0	0	0	0	0
16	0	0.019419	0.019419	0.10557	0.105573	0	0	0	0	0
18	0	0.019419	0.019419	0.10557	0.105573	0	0	0	0	0
20	0	0.019419	0.019419	0.10557	0.105573	0	0	0	0	0
Mean	0.039881	0.044291	0.044291	0.0789	0.0789	0.031414	0.031414	0.03141	0.02511	0.025113

Figure 4 Transformation pattern Loss Comparative results of Attribute oriented Anonymization Approaches with and without modulating attribute levels

4.3.2.3 Record-oriented Anonymization (RoA).

Record-oriented anonymization experiments four data quality models namely (i) Discernibility, (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 Discernibility and Ambiguity data quality models generate a minimum number of nV_C than 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 most minimum values, the obvious process is to analyze the optimal one on these measures. Accordingly, Discernibility and Ambiguity are the best models on nV_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.

	Record oriented Anonymization(RoA) Number of Null Values			
K Value	Discernibility	Average Equivalence Class	Ambiguity	Entropy Based Model
2	151062	115386	181029	71762
4	181212	135792	181149	70298
6	181509	160593	120973	94764
8	181716	171963	151122	100777
10	181479	157618	151126	105635
12	181668	164333	211318	108610
14	181854	170273	151266	96164
16	211478	195614	151346	98369
18	211478	199498	211478	101001
20	211514	187458	151538	103080
Mean	187497	165852.8	166234.5	95046
	On Modulating Attribute Levels of Quasi Identifiers in generalization			
2	3492	12951	3492	12951
4	9153	28440	9153	28440
6	12753	41112	12753	41112
8	15669	51390	15669	34722
10	18261	59787	18261	39771
12	20898	66267	20898	45891
14	24687	72360	24687	33840
16	27297	79056	27297	37728
18	29547	85356	29547	40536
20	31554	89514	31554	44208
Mean	19331.1	58623.3	19331.1	35919.9

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

	Record oriented Anonymization(RoA) Transformation pattern Loss			
K Value	Discernibility	Average Equivalence Class	Ambiguity	Entropy Based Model
2	0.151332	0.282863	0.168478	0.641431
4	0.168478	0.236237	0.151332	0.475858
6	0.168478	0.266201	0.092885	0.494924
8	0.304299	0.386059	0.208743	0.323877
10	0.289953	0.363604	0.177808	0.323877
12	0.199359	0.255792	0.213786	0.283885
14	0.319586	0.346725	0.213643	0.208845
16	0.331847	0.46967	0.183503	0.208845
18	0.331847	0.46967	0.331847	0.208845
20	0.331847	0.361085	0.198612	0.208845
Mean	0.2597026	0.3437906	0.1940637	0.3379232
	On Modulating Attribute Levels of Quasi Identifiers in generalization			
2	0.105773	0	0.105773	0
4	0.105773	0	0.105773	0
6	0.105773	0	0.105773	0
8	0.031037	0.043817	0.031037	0
10	0.031037	0.043817	0.031037	0
12	0.019419	0.056544	0.019419	0.035236
14	0	0.109129	0	0.026271
16	0	0.109129	0	0.026271
18	0	0.109129	0	0.026271
20	0	0.109129	0	0.026271
Mean	0.0398812	0.0580694	0.0398812	0.014032

Figure 6Transformation 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

[1] J. Gantz and D. Reinsel. The digital universe in 2020: Big Data, Bigger DigitalShadows, and Biggest Growth in the Far East. Technical report, IDC, sponsored by EMC, 2012.

[2] OpenData websites. http://www.data.gov/, http://data.gov.uk/.

[3] Information Commissioner's Office. Data Sharing Code of Practice. Technical report, ICO, 2011

[4] K. El Emam. Data Anonymization Practices in Clinical Research: A Descriptive Study. Technical report, Access to Information and Privacy Division of Health Canada, Ottawa, 2006.

[5] D. Goodin. Poorly anonymized logs reveal NYC cab drivers' detailed whereabouts.http://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cabdrivers-detailed-whereabouts/.

[6] M. Barbaro and T. Zeller. A Face Is Exposed for AOL Searcher No . 4417749, 2006.

[7] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In Proceedingsof the 2008 IEEE Symposium on Security and Privacy, SP '08, pages 111–125, 2008.

[8] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. Int. J. Uncertain. Fuzziness Knowl.Based Syst., 10(5):557–570, 2002.

[9] https://arx.deidentifier.org/downloads/

[10] M. E. Nergiz and C. Clifton. Thoughts on k-Anonymization. Data and Knowledge Engineering,63(3):622–645, 2007.

[11] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure. In Proceedings of ETK-NTTS 2001, Luxemburg: Eurostat, pages 807–826, 2001.

[12] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60:224–232, 2006.

[13] Jisha Jose Panackal,Anitha S Pillai, "Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets", In: Proc.International Symposium on Big Data and Cloud Computing,Procedia Computer Science, pp: 347-352, 2015.

[14] DevyaniPatiland Dr. Ramesh K. Mohapatra, "Evaluation of Generalization Based K-Anonymization Algorithms" In: Proc.International Conference on Sensing, Signal Processing and Security, pp.: 171-175, 2017.

[15] Paranthaman and Dr. T. Aruldoss Albert Victoire, "Performance Evaluation of K-Anonymized Data",Global Journal of Computer Science and Technology Software & Data Engineering, Volume 13 Issue 8, pp: 1-6,2013.

[16] Ines Buratović, Mario Miličević and KrunoslavŽubrinić, "Effects of Data Anonymization on the Data Mining Results", In: Proc.International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, pp.: 1619-1623, 2012.

[17] MdNurul Huda, Shigeki Yamada, and Noboru Sonehara, "On Enhancing Utility in k-Anonymization", International Journal of Computer Theory and Engineering, Vol. 4, No. 4, pp.: 527-833, 2012.

[18] Sam Fletcher and MdZahidul Islam, “Measuring Information Quality for Privacy Preserving Data Mining”, International Journal of Computer Theory and Engineering, Vol. 7, No. 1, pp.: 22-28,2015.

[19] Mohammad Reza Keyvanpourand SomayyehSeifiMoradi, “Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework”, “International Journal on Computer Science and Engineering (IJCSE)", Vol. 3 No. 2, pp.: 862-870, 2011.

[20] Gladiss Merlin. N, “Performance Evaluation of Anonymization Approach Using Clustering Algorithm In Big Data Application”,International Journal of Advanced Engineering Research and Technology (IJAERT)”,Volume 5 Issue 11,pp.:878-882,2017.

[21] Deepak Narula, Pradeep Kumar, and ShuchitaUpadhaya, “ Performance Explanation of K-Anonymization Algorithms for Average Class Partitioning Metric”, International Journal of Advanced Research in Computer Science, Vol.9, No.1, pp.:700-705,2018.

[22] MadhanSubramaniam and Senthil R, “An Analysis on Preservation of Privacy in Data Mining”, International Journal on Computer Science and Engineering, Vol. 02, No. 05,pp.: 1697-1699,2010

[23] Ali Inan, Murat Kantarcioglu, and Elisa Bertino, “Using Anonymized Data for Classification”, In: Proc.International Conference on Data Engineering,pp.:1-12,2009.

[24] Roberto Pellungrini, Francesca Pratesi, and Luca Pappalardo, “Assessing Privacy Risk in Retail Data”, In: Proc.International Workshop on Personal Analytics and Privacy. An Individual and Collective Perspective, pp.:17-22, 2017.

[25] AradhanaNyati and DivyaBhatnagar, “Performance Evaluation of Anonymized Data Stream Classifiers”, International Journal of Computer Science and Network,Volume 5, Issue 2, pp:381-387,2016

[26] Josep Domingo-Ferrer and David Rebollo-Monedero, “Measuring Risk and Utility of Anonymized Data Using Information Theory”, In: Proc.International Conference on Extending Database Technology, pp.:126-130, 2009.

[27] Abdul ‘Azim Mohammad and Maheyzah Md. Siraj, “Privacy Preserving Data Mining Based on K-Anonymity and Decision Tree Classification”, In: Proc.Innovations in Computing Technology and Applications,pp.:1-5,2016.

[28] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, and Liam Murphy, “A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners”, Transactions On Data Privacy,Volume 7 Issue 3,pp.:337-330,2014.

[29] Fabian Prasser, RaffaelBild, Johanna Eicher, Helmut Spengler, Florian Kohlmayer and Klaus A. Kuhn, “Lightning: Utility-Driven Anonymization of High-Dimensional Data”, Transactions On Data Privacy, Volume 9 Issue 2,pp.: 161-185,2016.

[30] https://arx.deidentifier.org/downloads/