Received: July 2, 2018

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Utility Enhancement of Deficient Relational Recordset Anonymization

Kishore Verma Samraj[1]* **Rajesh Appusamy[2]** **Ramya Ravi Shankar[3]**

[1] *Department of Computer Science and Engineering,*
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Kanchipuram, Tamilnadu, India
[2] *C.Abdul Hakeem College of Engineering and Technology,*
Melvisharam, Tamilnadu, India
[3] *Department of Computer Science and Engineering,*
C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamilnadu, India
* Corresponding author's Email: kishore.saj3@gmail.com

Abstract: Increase in the use of digital platforms generate a very huge amount of user data which on processing reveals more valuable information about users while mining or it may also reveal some future events. Privacy preserving data mining (PPDM) is the current developing area of research that precisely ensures a certain level of privacy by increasing the utility of information. Data Anonymization is a PPDM technique that protects sensitive information in the recordset with high confidence. Anonymizing recordsets that consists of null values habitually referred as deficient recordsets, suffer from serious information loss owing to null value contamination. This would be arisen because of the null values present in the original recordset. In this paper, we proposed an enhanced ARX framework to anonymize the deficient recordsets through a higher level of privacy and utility. We explore the characteristics of null value contamination on generalization, driven by these characteristics, we formulate an Enhanced ARX Anonymization- (K, R_n^N)Anonymity Model to preserve data utility on deficient recordsets. The results obtained by executing our framework over the real - world dataset proved to be optimized in providing better trade-off between utility and privacy for deficient recordsets in Cell oriented Anonymization (CoA), Attribute oriented Anonymization (AoA) and Record oriented Anonymization (RoA) than the existing procedures.

Keywords: K Anonymity, (K, R_n^N)Anonymity, Null value contamination, Deficient recordset, Privacy, Utility.

1. Introduction

Now a day's society has faced massive growth in data acquisition, data analytics and utilization. These practises are done on the individual's records for a variety of reasons like Digital Marketing, Social Media Analyses, E-commerce, M-commerce, and Health care systems, etc. During data mining, an individual's data may get revealed by joining the one/more records utilized for the above practises with multiple publicly available data sources. Anyhow, most of the recordsets may contain identifiable information about an individual that leads to privacy risk, when they are exposed to external records joining process. In a sense, [1] analysed and stated that removing personal identifiable attributes alone is not enough to protect the recordset from linking attacks and proposed k anonymization to progress the practise without linking attacks. Unfortunately, this method was exposed to Homogeneity and Background Knowledge attack. To overcome these attacks [2] presented a method called l-diversity, which maintains a threshold in scattering sensitive attributes in each equivalence class but fails in gripping semantics within the attribute values of a record. The semantic issues of l-diversity is overwhelmed by [3] t-closeness. Likewise, number of k-anonymization strategies is proposed like (∝, k) anonymization, work load aware anonymization, k anonymity with the access control policy, etc. Condensation [4] and Micro aggregation [5] are also proposed to ensure the recordsets be protected from linkage attacks under different environmental backgrounds.

Mostly the conventional k anonymization strategies concentrates much on producing k anonymized record sets that strictly abide k anonymity criteria and some of the approaches focus on reducing information loss that incurs due k-anonymization. The main issue we would like to determine here is the previous approaches do not turn up their deep focus on null values that are present in the original recordset, often represented as a deficient recordset. Moreover, all the conventional anonymization procedures [6-8] is implemented by assuming the recordsets that do not contain any null values. It means that they are perfect recordsets at any point of implementation. Due to complexity most of the existing approaches do not prefer to use deficient recordsets. For theoretical aspects, anonymization of deficient recordsets without considering null values may be interesting, whereas in real-world practise the recordsets used for privacy preserving data mining do contain null values in the recordsets. Only few of the existing approaches attempted to consider the null values present in the recordsets as outliers.

[9 - 11] approaches include null values in their anonymization procedures and publish them along with normal values, but mostly suffer from huge information loss. The following example illustrates the impact of null values in anonymization procedure of exposing more information loss and suppressed records, which is termed as null value contamination.

Now let's consider a scenario where the hospital wants to release patient's medical records. The dataset contains three attributes (i.e.,) age, zip code, disease.

Table 1. Null Record sets and anonymization with existing K anonymity (Mondrian method)

A)Tuple Id	**Age**	**Zip Code**	**Disease**
1)Alex	14	632002	Hypertension
2)Bob	18	632006	Cancer
3)Lucy	25	*	Birth Defects
4)Janet	26	*	Bird Flu
5)Alice	*	632007	Bird Flu
6)Simon	37	632009	Heart Disease
7)John	*	632012	Hypertension
8)Tom	48	632014	Bird Flu
B) Mondrian[6]			
	*	*	Hypertension
	*	*	Cancer
	*	*	Birth defects
	*	*	Bird flu
	*	*	Bird flu
	*	*	Heart disease
	*	*	Hypertension
	*	*	Bird flu

The first two attributes are considered as a quasi-identifier (Qid). Disease attribute is considered as a sensitive attribute (Sa).While collecting data of some patients, they do not want their personal information to be published or to be public. Thus the record set will be provided as shown in Table 1A**.**

Here the null value is denoted as *. In Table 1A 25%of values are null and up to 50% of records contain null values. Removing the records with the null values will lose 50% of records that is nearly 50% of information is lost. Evidently the solution will lose more information if there is higher deficient record rate.

If the null values are allowed to participate in anonymization, where it leads to another issue called null value contamination. Null value contamination is the problem that arises while performing anonymization based on generalization, the normal values will get contaminated as *(i.e.) Null value, due to the presence of null values in certain attributes. E.g., if the widely used Mondrian [6] algorithm is applied to Table 1A, then the entire Qid nearly 67% information will be lost or contaminated because of only 4 null values, as shown in Table 1A and 1B.

One of the recent literature work [12] developed an enhanced Mondrian and semi partition approach, where they divide the given dataset into null value and normal value set with respect to attribute. The partitions are merged and anonymized in top down strategy. The conceptual idea of segregating the null value records from the whole recordsets are inferred in this approach. The proposed Enhanced ARX Anonymization - (K, R_n^N) Anonymity Model works with distinct perspective and contributions as follows

(i) (K, R_n^N) Anonymity Model is framed to generate quality k anonymized data for the record set (R_n^N) which contains both null value and Normal values.

(ii) The proposed approach is experimented in all dimensional aspects of the recordset i.e. Cell oriented Anonymization, Attribute oriented Anonymization and Record oriented Anonymization.

(iii) This approach evaluates the obtained results with our novel utility parameters a) Suppressed Record count SR_C b) Null Value Information rate NVI_r and c) Null Value Record rate NVR_r. Since this approach is applicable to all dimensional anonymizations, information loss alone cannot be taken as the utility measurement parameter. Each of the dimensional anonymization follows different perception in computing the information loss.

(iv) In [12] work top down anonymization is used, whereas in this approach ,null value segregation is appropriately combined with ARX Anonymization often referred to be Enhanced ARX Anonymization- (K, R_n^N) Anonymity Model to anonymize the recordset, which is capable enough to generate k anonymized data set with increased data utility.

(v) The proposed Enhanced ARX Anonymization -(K, R_n^N) Anonymity Model is executed at least with a minimum of two data quality parameters for each of the dimensional oriented approaches to identify the optimistic one.

The rest of the paper is organised as Section 2 discusses the related approaches, Section 3 introduces the basic definitions, concepts and precepts that are required to have better foundations. Section 4 presents the proposed approach and discusses its importance. Section 5 shows the implementation and experimental analysis of the proposed approach. Section 6 concludes the benefits and extension procedures for our proposed approach.

2. Related works

Preserving privacy in the publication of data is an important problem in the privacy literature. In 2002, the authors of [1, 13] initially discovered that removing personal identities is lacking to guard privacy throughout data publication due to the existence of Qid. To handle this issue they proposed k-anonymity which needs every record as indistinguishable with a minimum of k-1 alternative records on Qid. In [2] authors argued and proved that k-anonymity is vulnerable to homogeneity attack and background knowledge attack. They presented l-diversity model with reserves diversity constraint to boost privacy protection. In [3] authors found that skewness attack and similarity attack on l-diversity are possible and further proposed t-closeness model with distribution constraint to preserve privacy. [14] Demonstrated that t-closeness cannot sufficiently shield the occasional Sensitive attribute values, and structured β-likeness with sturdy (strong) constraints on relative confidence gain to attain anonymity. The authors of [15] pointed out that competitor have partial knowledge regarding high dimensional transactions may lead to privacy threat and proposed k anonymous to stop the opponent in gaining knowledge of m items in particular transaction. The authors of [16] state that multiple records of particular individual might result privacy breaches and large information loss and proposed (k, l)-diversity to preserve privacy and utility with restricted assumptions. Intuitively, the initial micro data typically cannot satisfy the aforesaid privacy models, unless they are properly anonymized. According to [3, 17, 18], achieving optimal anonymization with minimum data loss beneath existing models, is NP-hard. Therefore all existing anonymization approaches attempt to come through nearly-optimal anonymity with the approximation algorithms. In 2006, [6] presented a top-down greedy approximation k-anonymization algorithm called Mondrian which was based upon local recoding. This uncomplicated scheme has been widely used in different literatures [19, 20]. [7] authors presented two clustering-based algorithms that outperformed Mondrian on data loss by sacrificing potency. [8] mapped multi-dimensional microdata to one-dimension, and proposed two systematic microdata algorithms named Hilb and iDist to achieve k-anonymity.[21] proposed a clustering-oriented methodology to stay nearest neighbourhood structures of data points throughout anonymization. The authors of [22] developed a clustering-based anonymization approach to preserve the characteristics of streaming data. [23] framed a privacy protective sub-feature choice approach mostly based on fuzzy possibilities. Recently, in [24] the authors found that k-anonymity will be achieved by non-homogeneous generalization, and proposed a method named ring generalization to realize higher utility whereas providing identical privacy guarantee. The authors of [25] adapted ring generalization for anonymizing high-dimensional data, and proposed a non-reciprocal recording anonymization theme for such information/data. [26] developed the optimal-utility k-anonymization and proposed freeform generalization for higher utility. The problem of deficient records having null values arises frequently [27]. It is stated that this sort of data encompasses a negative result on data processing [28]. To deal with this issue, researchers have frame worked a serial of approaches that appropriately handles null values. One of the foremost used approaches is, not to include records with null values in anonymization. Another popular methodology assigns any value in place of null values based on the assumption e.g. mean substitution, regression imputation. However unwanted imputations can create a major bias between real and assumed data. To the most effective inference, no other previous works uses null values throughout anonymization. That is as a result, anonymization is usually separated from usage of data. By using null values throughout anonymization which could change the essence of raw dataset and misdirect the data recipient. In the all existing works [6-8] and [16, 29], have removed records that has null values in the pre-processing step. As shown in the Section 1, this will result in more information loss. To ignore this problem. The authors of [9] frame worked cell-based suppression, but this approach causes serious null value contamination. To avoid the null value contamination controlled anatomized based approach for deficient datasets was framed in [10]. Conversely their solution is not that much competent enough to be adopted by existing anonymization. [11] Proposed two methods basic match and extend match based on generalization to anonymize the deficient records but often generates more information loss. In [12] the authors formulated an enhanced Mondrian method, where they split the known recordset into two groups i.e. records with null value and records with normal value with respect to each attributes. These partitions are combined and anonymized in top down scheme, which comparatively suffers from higher information loss and reduced utility than the proposed scheme. ARX anonymization handles null values of the recordset as equivalent to other values [30]. We employed ARX anonymization tool with various quality models to assess the effectiveness comparison of existing approaches [9 - 11] with our proposed scheme. Here ARX anonymization is meshed with Records partition procedure to form Enhanced ARX Anonymization - (K, R_n^N) Anonymity Model to create k anonymized records set with decreased information loss and increased utility.

3. Definitions, concepts and precepts

A. K anonymization

An anonymized record set is said to have k anonymity property, if the attribute values of each record contained in the published record set cannot be distinguished from at least k-1 individual records. K anonymization can be done in two methods as mentioned below

(a) Generalization

Modifying each record's attribute values with their domain specific coarse granular values until resulting in k anonymity property.

(b) Suppression

Modifying the record's attribute values with a null value (*), that are not able to resultant in k anonymity property.

B. Data quality models

Three kinds of data quality models as mentioned below are experimented

(a) Cell oriented anonymization (CoA)

Cell oriented Anonymization procedure works on the principle of generalization that is taking place with respect to cell values. This method generalizes any given attribute value with varying generalization structure on each execution until the K anonymization is maintained.

(b) Attribute oriented anonymization (AoA)

This procedure works on the principle of generalization done with respect to each column/attribute of a given dataset/recordset. AoA maps value of the attribute with the same set of coarse granular value on every appearance of those values.

(c) Record oriented anonymization (RoA)

Record Oriented Anonymization (RoA) works on the principle of applying anonymization procedure with respect to all quasi identifier's domain hierarchy. Here the anonymized dataset/record set is obtained by applying the generalisation procedure to the vector of quasi identifier value in each record of the recordset. The equivalence classes or partitions that are created should not contain any overlapping values. i.e. more than one kind of coarse granular representation of a particular quasi identifier of one equivalence class is not applicable to other similar quasi identifier belonging to another equivalence class in a recordset.

The quasi identifier values that are anonymized is purely based on the consideration of all the attribute values of the record.

C. Data utility parameters

(a) Suppressed record count $(\boldsymbol{SR_C})$

It computes the number of records that are masked with *, for not having k anonymity property.

$$SR_C = \mid R^* \mid where\ \forall\ R^{*\prime}s, A(i) = * \qquad (1)$$

R^* - The anonymized form of record set R; $A(i)$ - attribute of each records.

(b) Null value information rate $(\boldsymbol{NVI_r})$

Let N be the set that collects all null values .i.e. N(S) $=\{q_i[\mathrm{j}]\ \in\ \mathrm{S}, q_i[\mathrm{j}]\ = *, \#, -\}$.

The Null Value Information rate is defined as

$$NVI_r(S) = {}^{\mid N(S) \mid}\!\big/_{\mid Q \mid n} \qquad (2)$$

$\mid N(S) \mid$ - The number of null values present in the anonymized record set R^*; Q is the set of all quasi identifiers; $\mid Q \mid$ is number of quasi identifiers in the record set R and n denotes total number of tuples present in the record set R.

(c) Null value record rate$(\ \boldsymbol{NVR_r})$

Let R_n be the set that collects all records with null values i.e. $R_n = (q_i \in S, \exists q_i(j) = *, \#, -)$

The Null value Record rate

$$NVR_r(S) = {}^{\mid R_n(S) \mid}/_{n} \qquad (3)$$

$\mid R_n(S) \mid$ - Indicates the number of records that contains null values and n denotes the total number of tuples present in the record set R.

(d) Null value contamination

For a record set R and the generalised form of R is R^*, if $R^*$is achieved by k anonymization, Then if there exist $\mid N(R^*) \mid \geq \mid N(R) \mid$, that particular recordset is said to be contaminated due to the presence of null values.

D. Precepts

(a) Precept 1

For a recordset group S in R, $q_i \in S$, if there exist $V(A_i) = *, \#, - .etc. S^*$ will make all $V(A_i)$ as $*, \#, -$.

$V(A_i)$ - denotes value of each attributes A.

Each quasi identifier q_i of S forms a group of values called Q. Anonymization of the group Q.i.e. Generalize(Q), then according to precept 1 the null values *, #, - of any q_i in Q will make all q_i in Q to become Null as shown in the Table 1B. We address this issue to be the impact that raised due to null value threat. To overcome this drawback and prevent from null value contamination. The deficient record sets are necessary to be segregated from the entire recordset and generalize them distinctly.

(b) Precept 2

For a record set group S in R , anonymization of S , i.e. $G(S) = S^*$, if there exists more number of Suppressed Record count SR_C , that makes the data not usable, leads to increase in the information loss (IL).

$$SR_C \propto IL \qquad (4)$$

For any recordset group $S \in R$, the anonymization procedure on S of R caused suppressed records, the SR_C need to be small, else it will make the entire recordset as worthless. The anonymized recordset that possesses more number of suppressed records will vigorously impact the data mining results accuracy. Thus by precept 2 the anonymization procedure that can return minimum number of suppressed record primes to have greater accuracy in utility.

Algorithm 4.1 Enhanced ARX Anonymization- (K, R_n^N)Anonymity Model

```
Precondition: Raw Record set: $R_n^N \Leftrightarrow$ Record set that comprises both Normal values and null values
Post condition: Generalised Record set $R_n^{N*}$
Recordset-Partition ($R_n^N$)
/* Splits the record sets into two distinct groups Normal Value Group $NVG$ and null Value Group $nVG$.
If the Recordset cannot be partitioned then append the partition to the global return list
Else
For each quasi-Identifier of the Recordset check-for null values and form $nVG$
  $nVG \leftarrow \{R_n^N \leftarrow Recordset: v[quai - Identifier] = *, \#, -\}$
Return $nVG$
Then for each quasi-Identifier of the Recordset check-for not null values and form $NVG$.
  $NVG \leftarrow \{R_n^N \leftarrow Recordset: v[quai - Identifier] \neq *, \#, -\}$
Return $NVG$
/* anonymizes both Normal Value Group and null Value Group recordsets independently. */
Generalize-Partitions ($nVG, NVG$)
Return $nVG*$ and $NVG*$
Build Anonymized Recordset on the records that consist of both Normal and null Values.
$R_n^{N*}$ = Merge both group as Union ($nVG*$,$NVG*$)
Publish $R_n^{N*}$
Apply Utility parameters
    Null Value Information rate ($NVI_r$)
    Null Value Record rate( $NVR_r$)
Evaluate data quality models
End
```

4. Enhanced ARX anonymization $-$ (K, R_n^N) anonymity model for deficient record set

R_n^N denotes the record set that contains both null values and Normal values. Thus anonymization of R_n^N with usual k anonymization procedure will anonymize the record set with less utility and not supportable to data miner's community. The primary aim of (K, R_n^N)Anonymity Model as shown in Algorithm 4.1 is to anonymize the deficient recordset and analyse the impact null values contamination in our proposed method with respect to existing procedures. Our Enhanced ARX anonymization is capable of generating anonymized $R_n^{N\ *}$ that possess more balanced privacy –utility trade-off than the existing ones.

4.1 Null value segregation

In this work, null value contamination needs to be avoided. Thus the Recordset -Partition (R_n^N) will segregate records with null values from R_n^N and make them as null Value Group nVG (i.e. the set of records comprise null values) and Normal Value record Group NVG (i.e. the set of records comprise normal values/ without null values).

Recordset-Partition (R_n^N) selects each quasi identifier q_i from Q and checks if there exists any null values like $*, \#, -$, then those records are segregated as distinct group called nVG. NVG is created by checking each q_i from Q having attribute values not equal to null. According to precept 1 our proposed Recordset-Partition (R_n^N) successfully segregates the null value record set from whole record set. The null Value Group nVG and Normal Value Group NVG which have been created by Recordset-Partition procedure are then subjected to anonymization.

4.2 Partition anonymization

Both partitions nVG andNVG, created by Recordset-Partition are anonymized using Algorithm 4.2. This Generalize-Partition function anonymize nVG / NVG recordset group only if the number of records in each of the group is greater than K because it is not possible to create K anonymized records with the records count less than K. Suppose if the number of records is less than the K, those records can directly be appended in the anonymized table or can be suppressed. This is up to the feasibility of the data provider according to his/her required level of privacy/utility.

```
Algorithm 4.2 Generalize-Partitions ($nVG, NVG$)
Input: null Value Group $nVG$ / Normal Value Group $NVG$
Output: Anonymized null Value group $nVG*$/ Normal Value Group $NVG*$
If | $nVG/ NVG$ |$< K$ then throw exception “The
   resultant Group cannot be Anonymized” append to
   the global resultant list / suppress.
Else
For all $nVG/NVG$ $K = 2; K \leq n; K + 2\}$
                    Note: n is the user defined parameter
Choose Data quality Model for (Cell Oriented
   Anonymization / Attribute Oriented Anonymization
   / Record Oriented Anonymization) from given
   Quality measure.
Compute $SR_C \longrightarrow Suppressed\ Record\ Count.$
Compute Score ⟶ Information Loss
Return $nVG*$ / $NVG*$
End
```

4.3 Evaluation of utility measures

Here K-anonymized $nVG *$ and $NVG *$ are merged using union function and formed the resultant anonymized Recordset $R_n^{N\ *}$. Then this $R_n^{N\ *}$ is reported towards the utility analysis by applying the measures SR_C, NVI_r and NVR_r as mentioned in Eqs. (1), (2) and (3). For each k values ranging from k=2 to n, the recordsets are anonymized and the utility measures are calculated and tabulated individually. An average function is taken for SR_C, NVI_r and NVR_r on different null value record set size N for all k values. The computation is followed and analysed by all privacy metrics that are available under CoA, AoA and RoA. The privacy metric that owns lowest SR_C, NVI_r and NVR_r are adequate enough to support the data mining process with higher utility.

5. Experimental evaluation

In this section the experimental results are assessed in term of data utility. Recordset-Partition procedure that comprises NVI_r and NVR_r measures are implemented in .Net framework. Generalize-Partitions procedure are implemented using open source ARX anonymization tool by configuring the tool according to our analytical model. These algorithms run on Intel® core™ i5-5200U CPU @ 2.20 GHZ, 8 GB RAM and 64-bit Windows 8.1 operating system. Here the real world data set ADULT is used. This data set is widely used in all k-anonymization procedures and retrieved from publicly available website https://archive.ics.uci.edu/ml/datasets/adult. The dataset description is shown Table 2.

We compute the utility of the anonymized record set by $\mathrm{SR_C}$(Suppressed Record Count), $\mathrm{NVI_r}$ (Null Value Information rate) and $\mathrm{NVR_r}$(Null Value Record rate) . Decreased $\mathrm{SR_C}$, $\mathrm{NVI_r}$ and $\mathrm{NVR_r}$ implies to have desirable quality on anonymized record set. We measured $\mathrm{SR_C}$, $\mathrm{NVI_r}$ and $\mathrm{NVR_r}$ in various parametrical setup and the results are demonstrated in Figs. 1 to 9. To deliberate that the utility of the record set is sensitive to the null values record set population N in whole record set size. We created a series of subsets from the whole recordset with different null values population i.e. 100,200,300…, 1000.

For each series of subsets, 10 sample set of experimentation for k=2 to 20, by k=k+2 on each iterations are anonymized. The represented outcome is the average of the results obtained from those 10 sample sets. Likewise totally 16 experiments were executed in CoA, AoA and RoA comprising existing and proposed methodologies.

Table 2. Data set description

Data Set	**No.of Tuples**	**QID**	**Type**	**S.A**

Adult	30162	sex,
		age,
		race,
		marital status, education, native country
		Work class,
		occupation
			Categorical(2)
			Numerical
			Categorical(5)
			Categorical(7)
			Categorical(16)
			Categorical(41)
			Categorical(8)
			Categorical(14)
				Salary

5.1 Cell oriented anonymization (CoA)

This approach performs cell oriented anonymization with two different utility measures namely i) Loss and ii) Precision. In this experimentation, *sum* is used as default aggregator function based on the fruitful results obtained in our previous analysis with respect to generalization lattice.

Tables 3, 4 and 5 represent the results obtained on experimenting CoA. ARX Anonymization (Precision) anonymizes recordsets with null values by extend match method as performed in [11]. ARX Anonymization -Loss anonymizes recordsets with null values by suppressing the records that contains null values during anonymization as done in [9]. From the resultant figures it is clearly represented that Loss and Precision metric of Enhanced ARX Anonymization generates minimum number of SR_C , NVI_r and NVR_r .Enhanced ARX Anonymization-Precision(Highlighted with green colour in Tables 3, 4 and 5) seems to be better than [11]. Enhanced ARX Anonymization-Precision is much sensitive to null value population where SR_C , NVI_r and NVR_r values do not have normal increasing trend, when size of the null value population N increases. On complete assessment of these measures Enhanced ARX Anonymization-Loss (Highlighted with blue colour in Tables. 3, 4 and 5) is the best method in generating quality anonymized data that possess better utility in CoA.

This analytical results will help the data provider to anonymize the data with guaranteed privacy and greater utility. From this, a fact can be derived i.e. on anonymizing record set containing null values, size of the null values N in the record set is directly proportional to SR_C, NVI_r and NVR_r where these values linearly scales in increasing trend.

5.2 Attribute oriented anonymization (AoA)

Tables. 6, 7 and 8 represent the results that are obtained on experimenting Attribute Oriented Anonymization (AoA). AoA is executed on two utility measures called Non Uniform Entropy (NUE) and Normalised Non uniform Entropy (NNUE). ARX Anonymization-Non Uniform Entropy (NUE) is the extension method of [31] by implementing extend match procedure of [11]. And ARX Anonymization-Normalized Non Uniform Entropy (NNUE) is the normalised variant of NUE.

Table 3. Cell oriented anonymization SR_C utility comparison

N	Enhanced ARX Anonymization – Proposed System		ARX Anonymization	
	Suppressed Record count (SR_C)			
	Loss	Precision	Loss[9]	Precision[11]
100	2898.7	3567.5	2947.8	4024.5
200	2973.5	3668.5	3224.9	4118.2
300	3067.9	3763.3	3211.3	4208.6
400	3181	3861.3	3199	4180.1
500	3268.1	4251.5	3401.1	4303.6
600	3329.8	4284.2	3406.3	4309.6
700	3378.7	4361.5	3541.1	4487.8
800	3427.5	4219.1	3679.9	4413.4
900	3486.9	4474.3	3724.1	4681.3
1000	3514.4	4420.7	3834.3	4500.5

Table 4. Cell oriented anonymization NVI_r utility comparison

N	Enhanced ARX Anonymization – Proposed System		ARX Anonymization	
	Null Value Information Rate (NVI_r)			
	Loss	Precision	Loss[9]	Precision[11]
100	0.1921	0.2364	0.1955	0.2669
200	0.1965	0.2426	0.2139	0.2731
300	0.2029	0.248	0.2127	0.2792
400	0.2123	0.2534	0.2134	0.2774
500	0.2173	0.2813	0.2214	0.2824
600	0.22	0.2852	0.2216	0.2864
700	0.2208	0.2886	0.2253	0.2902
800	0.2269	0.2684	0.2289	0.2941
900	0.2326	0.2961	0.2331	0.2984
1000	0.233	0.2751	0.2355	0.3007

Table 5. Cell oriented anonymization NVR_r utility comparison

N	Enhanced ARX Anonymization – Proposed System		ARX Anonymization	
	Null Value Record Rate (NVRr)			
	Loss	Precision	Loss[9]	Precision[11]
100	0.0961	0.119	0.0978	0.1335
200	0.0986	0.1216	0.1072	0.1368
300	0.1017	0.1244	0.107	0.14
400	0.1053	0.128	0.1075	0.1396
500	0.1086	0.1352	0.1106	0.1427
600	0.1088	0.1353	0.1137	0.1455
700	0.1151	0.1377	0.1168	0.1486
800	0.117	0.1399	0.1202	0.152
900	0.1202	0.145	0.1232	0.1552
1000	0.1209	0.1466	0.1264	0.1583

From the Tables 6, 7 and 8, it can be seen that our proposed method Enhanced ARX Anonymization for both NUE and NNUE out performs than ARX Anonymization of NUE and NNUE. With respect to SR_C(Suppressed Record Count), NVI_r (Null Value Information rate) and NVR_r(Null Value Record rate) Enhanced ARX Anonymization- Normalized Non Uniform Entropy (Highlighted with blue colour in Figs. 4, 5 and 6) is best in generating minimum values of above said measures and capable of producing quality anonymized data sets. From the analysis Enhanced ARX Anonymization –NUE (Highlighted with green colour in Tables 6, 7 and 8) seem to be more sensitive to null value record population, since the derived values on those measures has more fluctuations in trend.

Table 6. Attribute oriented anonymization SR_C utility comparison

N	Enhanced ARX Anonymization – Proposed System		ARX Anonymization	
	Suppressed Record count (SR_C)			
	NUE	NNUE	NUE	NNUE
100	3772.5	2956.6	5156.9	3054.6
200	3869.7	3146.9	5257.1	3156.1
300	3959.8	3229.2	5148.5	3250.2
400	4061.8	3309.5	5424.9	3345.1
500	5348.1	3313.2	5509.3	3441.4
600	5358	3441.2	5385.8	4002.6
700	5451.5	3490.1	5462.5	3631.7
800	4635.7	3474	5534.4	3787.7
900	5516.3	3533.7	5602.5	3828.3
1000	4873.6	3561.5	5658.5	3925

Table 7. Attribute oriented anonymization NVI_r utility comparison

N	Enhanced ARX Anonymization – Proposed System		ARX Anonymization	
	Null Value Information Rate (NVI_r)			
	NUE	NNUE	NUE	NNUE
100	0.25	0.2025	0.3826	0.2026
200	0.2559	0.2086	0.3486	0.2088
300	0.261	0.214	0.3415	0.2143
400	0.2662	0.2193	0.3599	0.2198
500	0.3505	0.2202	0.3656	0.224
600	0.3575	0.229	0.3938	0.3084
700	0.3574	0.2319	0.3628	0.2326
800	0.2883	0.1215	0.3677	0.2382
900	0.364	0.1248	0.3724	0.2396
1000	0.3049	0.128	0.3732	0.242

Table 8. Attribute oriented anonymization NVR_r utility comparison

N	Enhanced ARX Anonymization – Proposed System		ARX Anonymization	
	Null Value Record Rate (NVR_r)			
	NUE	NNUE	NUE	NNUE
100	0.1251	0.1013	0.3334	0.1105
200	0.1283	0.1036	0.1745	0.1046
300	0.1313	0.1068	0.1711	0.1078
400	0.1347	0.1099	0.1805	0.111
500	0.1773	0.1131	0.1837	0.1119
600	0.1727	0.115	0.18	0.1172
700	0.1817	0.1184	0.1832	0.1203
800	0.1537	0.12	0.1863	0.1215
900	0.1839	0.1234	0.1892	0.1248
1000	0.1616	0.1235	0.192	0.128

5.3 Record oriented anonymization (RoA)

RoA experimentation is performed with four privacy parameters namely Discernibility (DM, Ambiguity (AM), Average Equivalence Class (AEC), Entropy based model (EBM). Tables 9, 10 and 11 represent the results that are obtained in RoA. In this experimentation, the recent literature work [9] is implemented as ARX Anonymization – DM (Highlighted in red colour in Tables 9, 10 and 11) and [12] is implemented as Enhanced ARX Anonymization- AEC (Highlighted in red colour in Tables 9, 10 and 11). ARX Anonymization- DM [9] is done by suppressing the records that constitutes null values by making the suppression limit as 100%. It is shown that our proposed Enhanced ARX Anonymization-DM(Highlighted in green colour in Tables 9, 10 and 11) and Enhanced ARX Anonymization – AM(Highlighted in blue colour in

Table 9. Record oriented anonymization SR_C utility comparison

N	Enhanced ARX Anonymization- Proposed System				ARX Anonymization			
	Suppressed Record count (SR_C)							
	DM	AEC [12]	AM	EBM	DM [9]	AEC	AM	EBM
100	2870.3	3772.5	2870.3	3678.7	2873	7186.5	2873	4424.9
200	2961.7	3869.7	2961.7	3779.8	2973.5	7270.6	2973.5	4515.5
300	3041.1	3959.8	3041.1	3874.2	3067.9	7351.3	3067.9	4606.8
400	3149.1	4052.6	3119.1	3973.5	3165.2	7427.3	3165.2	4693
500	3188.6	5631.8	3188.6	4774.2	3262.7	7505.6	3260.7	4956.9
600	3249.2	6637.3	3249.2	4680.9	4642.1	7569.8	4660.1	5565.7
700	3295.9	7905.7	3295.9	4739	3452.1	7641.6	3451.6	5145
800	3343.4	4637.5	3343.4	4791.4	3540.9	7710.9	3540.9	4858.3
900	3401.9	6141.7	3401.9	4853.4	3648.6	7782.7	3439.6	5346.2
1000	3429.9	4873.6	3429.9	4776.6	3745.7	7848.5	3745.7	4883.3

Table 10. Record oriented anonymization NVI_r utility comparison

N	Enhanced ARX Anonymization –Proposed System				ARX Anonymization			
	Null Value Information Rate (NVI_r)							
	DM	AEC [12]	AM	EBM	DM [9]	AEC	AM	EBM
100	0.1904	0.25	0.1879	0.2438	0.1904	0.4765	0.1909	0.2935
200	0.1965	0.2449	0.1965	0.25	0.1965	0.4822	0.1965	0.2995
300	0.2019	0.261	0.2018	0.2554	0.2019	0.4876	0.2019	0.3055
	0.2072	0.2635	0.2071	0.2608	0.2072	0.4927	0.2072	0.3114
500	0.212	0.3677	0.212	0.3169	0.212	0.4979	0.212	0.3245
600	0.2163	0.4341	0.2163	0.3109	0.3018	0.5023	0.3018	0.363
700	0.2199	0.5022	0.2199	0.3151	0.2199	0.5071	0.2199	0.3321
800	0.2233	0.2983	0.2234	0.2904	0.2234	0.512	0.2234	0.319
900	0.2275	0.3942	0.2255	0.3233	0.2276	0.5169	0.2276	0.3402
1000	0.2299	0.3062	0.2291	0.2964	0.2301	0.5215	0.2301	0.3259

Table 11. Record oriented anonymization NVR_r utility comparison

N	Enhanced ARX Anonymization –Proposed System				ARX Anonymization			
	Null Value Record Rate (NVR_r)							
	DM	AEC [12]	AM	EBM	DM [9]	AEC	AM	EBM
100	0.0858	0.1251	0.0841	0.122	0.0953	0.2383	0.0952	0.1467
200	0.0919	0.1283	0.0879	0.1253	0.0986	0.2413	0.0986	0.1499
300	0.0924	0.1313	0.0917	0.1284	0.1017	0.2442	0.1015	0.1536
400	0.1	0.1344	0.0939	0.1317	0.1049	0.2469	0.1046	0.1563
500	0.1003	0.1867	0.0991	0.1463	0.1078	0.2499	0.1081	0.1598
600	0.1062	0.2201	0.1027	0.1472	0.1112	0.2524	0.1112	0.1574
700	0.1106	0.2535	0.1033	0.1506	0.1145	0.255	0.1145	0.1605
800	0.1107	0.1538	0.1074	0.1611	0.1174	0.2584	0.1176	0.1778
900	0.1138	0.2036	0.112	0.1562	0.1202	0.2613	0.121	0.1669
1000	0.1178	0.1616	0.1242	0.1584	0.1245	0.2644	0.1376	0.1701

Tables 9, 10 and 11) can able to generate minimum values of SR_C, NVI_r and NVR_r than [9] and [12]. Enhanced ARX Anonymization – AEC [12] seems to be very sensitive to the null value population. By evaluating all the method with SR_C, NVI_r and NVR_r, two of our proposed methodologies seems to be optimum in RoA i) Enhanced ARX Anonymization-DM and ii) Enhanced ARX Anonymization- AM.

6. Conclusion

In this paper, the impact of null value present in the recordset during anonymization with respect to data utility are studied and examined. This issue (often referred as null value contamination) is addressed by our proposed Enhanced ARX anonymization-(K, R_n^N)Anonymity model, which is capable enough to anonymize the deficient recordset with good level of privacy and at the same time with greater utility. Through the experimental results, it is shown that the utility measures $\mathrm{SR_C}$ (Suppressed Record Count), $\mathrm{NVI_r}$ (Null Value Information rate) and $\mathrm{NVR_r}$(Null Value Record rate) of our proposed system is minimum than the existing procedures of [9, 11, and 12]. Hence we have proved that our proposed approach is much more adequate for privacy preservation of relational recordset R_n^N that consists of null values with higher utility.

Our proposed Enhanced ARX Anonymization - (K, R_n^N) Anonymity model have been implemented with better privacy and utility results under Cell oriented Anonymization(CoA), Attribute oriented Anonymization(AoA) and Record oriented Anonymization(RoA).

From our experimentation it has been shown that our proposed strategy is dimensional independent, since our results have shown better utility in CoA, AoA (single dimensional) and RoA (multi-dimensional).

Enhanced ARX Anonymization-(K, R_n^N)Anonymity model generates less information loss (Suppressed Record Count), good privacy and more data utility for deficient record set i.e. Record set with null values. This approach can be extended to other K-Anonymization strategies like l-diversity and t-closeness. Moreover this approach is implemented for single sensitive attribute setting which can be extended to multiple sensitive attribute settings.

References

[1] L. Sweeney, "k-Anonymity: Privacy Protection Using Generalization and Suppression", *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, Vol. 10, No. 5, pp. 571-588, 2002.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity", In: *Proc. of International Conf. on Data Engineering*, pp. 24-35, 2006.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", In: *Proc. of International Conf. on Data Engineering*, pp. 106-115, 2007.

[4] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining", In: *Proc. of International Conf. on Extending Database Technology*, pp. 183-199, 2004.

[5] J. Domingo-Ferrer, A. Oganian, A. Torres and J.M. Mateo Sanz, "On the security of micro-aggregation with individual ranking: Analytical attacks", *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 477–491, 2002.

[6] K. LeFevre, D.J. DeWitt., and R. Ramakrishnan, "Mondrian multi-dimensional k-anonymity", In: *Proc. of International Conf. on Data Engineering*, pp.25. 2006.

[7] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.C. Fu, "Utility-based anonymization using local recoding", In: *Proc. of International Conf. on Knowledge Discovery and Data Mining*, pp. 785–790. 2006.

[8] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss", In: *Proc. of International Conf. on Very Large Data Bases*, VLDB Endowment, pp. 758–769, 2007.

[9] M. Nergiz, C. Clifton, and A. Nergiz, "Multirelational k-anonymity", *IEEE Transaction on Knowledge Data Engineering*, Vol.21, No.8, pp.1104–1117, 2009.

[10] Q. Gong, J. Luo, and M. Yang, "Aim: a new privacy preservation algorithm for incomplete microdata based on anatomy", In: *Proc. of International Conf. on Pervasive Computing and the Networked World*, pp. 194–208, 2012.

[11] C. Margreta, E. Johann, and K. Christian, "Anonymization of Datasets with Null Values", In: *Proc. of Special Issue on Database- and Expert-Systems Applications on Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIV*, Vol. 9510, pp 193-220, 2016

[12] Q. Gong, M. Yan, Z. Chen, W. Wu, and J. Luo, "A framework for utility enhanced incomplete microdata Anonymization", *Cluster Computing*, Vol. 20, No. 2, pp 1749-1764, 2017.

[13] L. Sweeney, "k-anonymity: a model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp. 557–570, 2002.

[14] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee", In: *Proc. of VLDB Endowment*, Vol. 5, No. 11, pp. 1388–1399, 2012.

[15] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data", In: *Proc. of VLDB Endowment*, Vol. 1, No. 1, pp. 115– 125, 2008.

[16] Q. Gong, J. Luo, M. Yang, W. Ni, and X.B. Li, "Anonymizing 1: m micro data with high utility", *Knowledge Based Systems*, Vol. 115, pp.15-26, 2017.

[17] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity", In: *Proc. of SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 223–228, 2004.

[18] X. Xiao, K. Yi, and Y. Tao, "The hardness and approximation algorithms for l-diversity", In: *Proc. of International Conf. on Extending Database Technology, ACM*, pp. 135–146, 2010.

[19] Y. He and J.F. Naughton, "Anonymization of set-valued data via top-down, local generalization", In: *Proc.of VLDB Endowment*, Vol.2, No.1, pp. 934–945, 2009.

[20] H. Zakerzadeh, C.C. Aggarwal, and K. Barker, "Privacy-preserving big data publishing", In: *Proc. of International Conf. on Scientific and Statistical Database Management*, pp. 1–26, 2015.

[21] W. Ni and Z. Chong, "Clustering-oriented privacy-preserving data publishing", *Knowledge Based Systems*, Vol.35, pp. 264-270, 2012.

[22] K. Guo and Q. Zhang, "Fast clustering-based anonymization approaches with time constraints for data streams", *Knowledge Based System*, Vol. 46, pp. 95-108, 2013.

[23] H.K. Bhuyan and N.K Kamila, "Privacy preserving sub-feature selection based on fuzzy probabilities", *Cluster Computing*, Vol. 17, No. 4, pp. 1383– 1399, 2014.

[24] W.K. Wong, N. Mamoulis, and D.W.L. Cheung, "Non-homogeneous generalization in privacy preserving data publishing", In: *Proc. of International Conf. on Management of Data*, pp. 747–758, 2010.

[25] M. Xue, P. Karras, C. Raïssi, J. Vaidya, and K.L. Tan, "Anonymizing set-valued data by nonreciprocal recoding", In: *Proc. of International Conf. on Knowledge Discovery and Data Mining*, pp. 1050–1058, 2012.

[26] K. Doka, M. Xue, D. Tsoumakos, and P. Karras, "k-anonymization by freeform generalization", In: *Proc. of Symposium on Information, Computer and Communications Security*, pp. 519–530, 2015.

[27] D. Rubin, "Inference and missing data", *Biometrika*, Vol. 63, No. 3, pp. 581–592, 1976.

[28] M.L. Brown and J.F. Kros, "Data mining and the impact of missing data", *Industrial Management and Data System*, Vol. 103, No. 8, pp.611–621, 2003.

[29] X. Zhang, C. Leckie, W. Dou, J. Chen, R. Kotagiri, and Z. Salcic, "Scalable local-recoding anonymization using locality sensitive hashing for big data privacy preservation", In: *Proc. of International on Conf. on Information and Knowledge Management*, pp.1793–1802, 2016.

[30] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K.A. Kuhn "ARX - A Comprehensive Tool for Anonymizing Biomedical Data", In: *Proc. of AMIA Annual Symposium*, pp.984-993, 2014.