Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210088

An Extended Mondrian Algorithm – XMondrian to Protect Identity Disclosure

R.Padmaja[a,1] and V. Santhi[b]
[a,b]*Professor, School of CSE, VIT University, Vellore - 632014, Tamil Nadu, India*

Abstract. In recent days**,** Privacy Preserving Data Publishing (PPDP) is considered as vital research area due to rapid increasing rate of data being published in the Internet day by day. Many Organizations often need to publish their data in internet for research and analysis purpose, but there is no guarantee that those data would be used only for ethical purposes. Hence data anonymization comes into picture and play a vital role in preventing identity disclosure, also it restricts the amount of data that can be seen or used by the external users. It is an extensively used PPDP technique among data encryption, data anonymization and data perturbation methods. Mondrian is considered as one such data anonymization technique that has outperformed compare to many anonymization algorithms, because of its fast and scalable nature. However, the algorithm insists to encode the categorical values into numerical values and decode it, to generalize the data. To overcome this problem, a new extended version of Mondrian algorithm is proposed, and it is called XMondrian algorithm. The proposed algorithm can handle both numerical and categorical attributes without encoding or decoding the categorical values.The effectiveness of the proposed algorithm has been analysed through experimental study and observed that the proposed XMondrian algorithm outshine the existing Mondrian algorithm in terms of anonymization time and Cavg. Cavg is one of the metric used to quantify the utility of data.

Keywords. Data Anonymization, Mondrian, Privacy Preserving Data Publishing, X-Mondrian, pseudonymous data, K-anonymity, Identity Disclosure

1. Introduction

Many organizations often need to release their data for research and other public analysis purpose.For instance, medical organizations need to publish their patient data in the internet for medical researchers to carry out their research. But the patient data contains personal information of the patients which needs to be preserved before the data gets released in the internet. Initially, organizations were publishing their data by removing the recognizing attributes like Social Security Number, Phone Number, Email Id etc.., to protect the identity of individuals[1]. Dataset in which direct identifiers are eliminated and indirect identifiers remain intact are called pseudonymous data.Even then the research has proved that identity of individuals is getting disclosed because of quasi identifiers in the publisheddataset.Quasi identifiers are set of attributes to identify tuples [2].A classical situation of conventional privacy preserving data publishing (PPDP) is depicted in the belowFig.1, which demonstrates diverse

[1]V.Santhi ,professor ,school of CBSE,VIT university, vellore
Email:vsanthi@vit.ac.in

phases of data publishing. But this conventional data publishing method will not preserve the individuals privacy as the data is linked to public dataset like voter database, census data etc.., through which the privacy of individuals will be revealed out [3].

Figure 1. Conventional Model of PPDP for Pseudonymous Data

As a result, privacy preserving data publishing techniques has turned into a critical area of research for analyser's and specialists. Hence, the objective of privacy preserving techniques is to mask or encrypt or add noise to the original dataset to protect the identity disclosure without compromising the data utility. Thus, the spirit of privacy preserving techniques is to publish datasets without compromising the data usefulness. Manipulating quasi identifiers mathematically and technically guarantees to prevent re-identification is called **Anonymous Data**. This reality made many researchers consider new confronts to preserve the personal or sensitive information of the individuals in the published datasets. A contemporary model of PPDP is depicted in Figure. 2

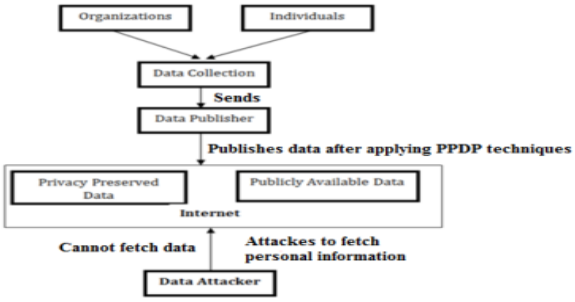

Figure 2. Contemporary Model of PPDP for generating Anonymous Data.

In general, Privacy preserving techniques are categorized into three types 1) Data Encryption 2) Data Anonymization and 3) Data Perturbation. Data anonymization technique, which uses generalization and suppression for anonymization, are the broadly studied and extensively accepted approach for protecting the identity disclosure. In data anonymization, the data is generalized or suppressed so that individual's

identity is not revealed. As a result, multiple data anonymization algorithms like datafly, incognito,Mondrian, info gainMondrian, LSD Mondrian have being projected in the region of data anonymizationtechnique.The Mondrian algorithm is a widely adopted anonymization algorithm because it reaches low data distortion. As per the literature review carried out, it is clear that Mondrian outperforms the other algorithms in terms of two metrics named DM and C_{AVG}, which is used to calculate the size and number of equivalence classes created. The Mondrian's equivalence groups have a greater granularity, which helps to improve data usefulness, but it takes time for converting categorical values into numerical values.

In this paper, an improved version of the current Mondrian algorithm is proposed and it is named as, XMondrian algorithm. In Mondrian algorithm, if the dataset contains any categorical values, then the algorithm needs to convert categorical values into numerical values. In order to overcome this impediment, an extension to the Mondrian algorithm called XMondrian algorithm is proposed.

2. Methods and Materials

2.1 Privacy Risks and Models

In this segment, the most significant risks associated with the disclosure of personal information is has been presented elaborately. The models that can be used to defend against these threats are then addressed.

2.1.1 Privacy Threats

Privacy threats are identified with 3 distinct kinds of identifiers and they are direct attribute, quasi identifies and sensitive attribute. In Figure 3, a model dataset is presented, in which SSN is an identifying identifier, marital status, age and gender are quasi-identifiers, and income is a sensitive attribute [2].

	ID	QIDs			SA
Tuple #	SSN	Marital Status	Age	Gender	Income
1	1234	Separated	38	Female	<=50000
2	2341	Separated	43	Female	>50000
3	2567	Married-civ-spouse	54	Male	>50000
4	1452	Widowed	43	Female	>50000
5	1765	Married-AF-spouse	50	Male	<=50000
6	1356	Never-Married	39	Female	<=50000

Figure 3. Micro Data from Census Dataset

The three classes of privacy threats based on above three types of attributes are **identity disclosure, membership disclosure and attribute disclosure**[5].

2.1.2 Privacy Models

The popular privacy models have been discussed with respect to the privacy threats as mentioned in section 2.1. The classification based on privacy models is presented in Figure 4.

Attack Type	Privacy Models
Identity Disclosure	k-Anonymity, K-Map etc..
Membership disclosure	δ-presence, c-confident δ-presence etc ..
Attribute disclosure	l-diversity, t-closeness

Figure 4. Privacy models against Privacy Threats[5]

The most established model for protecting identity disclosure is k-anonymity.

2.2 Privacy Preserving Operations

Privacy operations to achieve various privacy models are generalization, suppression, anatomization, permutation, perturbation [6]. Among these operations generalization and suppression are data anonymization technique operators. Generalization substitutes the original attribute value into a less specificvalue,but more general value.Suppression works by replacing some values with a special character like ``*". Suppression operation is generally used where it is not possible to generalize.

2.3 Privacy Preserving Techniques

The techniques to preserve the privacy of individuals are classified into3types : The data encryption,data perturbation and data anonymization .Anonymization is a optimal technique to protect Identity Disclosure.The proposed algorithm, XMondrian is one such data anonymization Algorithm [7]. In data anonymization, the user identity is modified to less specific values to protect the sensitive information.

2.4 Mondrian, A Data Anonymization Algorithm to prevent Identity Disclosure

Data anonymization is the process of taking the personal data and modifying it in such a way that it can no longer be used to identity an individual. Data anonymization is most extensively used technique to publish privacy preserved Data because of two reasons. 1) it can be applied on any type of data including big data and real time applications data and 2) it is very easy to implement.

Most anonymization algorithms rely on generalization and suppression to turn original datasets into anonymous data sets.The popular algorithms where generalization operations are mostly applied are urgus, Datafly, Median Mondrian, Infogain Mondrian, Incognito etc...Among these Algorithms, Median Mondrian Algorithm outperforms other Algorithms. The privacy model to protect Identity disclosure is K-Anonymity. Every Data Anonymization Algorithm that tries to protect identity disclosure should satisfy the k-anonymity principle.

2.4.1 k-anonymity

Sweeney and Samarati has introduced the k-anonymity principle [5].This privacy principle demands that each tuple in a discharged table cannot be connected to a probability greater than 1/k, meaning that each tuple is indistinguishable from at least k - 1 other tuples A dataset iscalledto be k anonymous dataset if each equivalence class have atleast k-1 records with respect to quasi identifiers[8]. Therefore for every dataset T, k-anonymization is performed to produce a new dataset T* that verifies the k-anonymity property on the set of quasi-identifiers. In K-anonymous table, the probability of discovering an individual from set of k tuples will be 1/k which is the degree of uncertainty[3].

2.4.2 Mondrian Algorithm

Mondrian is a k-anonymity data anonymization algorithm that recursively partitioning the dataset by finding the median of the quasi identifier that has largest number of unique values[9]. This algorithm partitions recursively until the equivalence class size is less than 2k-1.

Algorithm: *Mondrian[14]*

```
Input: Quasi identifiers qid and a Dataset D.
Output: An Anonymized Dataset D*
Anonymize (partition)
if (no allowable multidimensional cut for partition)
return φ : partition → summary
else
dim ← choose-dimension ()
fs ← frequency-set (partition, dim)
splitAttribute ← largestWideRange(fs)
splitVal ← find-median(splitAttribute)
lhs ← {t € partition: t.dim ≤ splitVal}
rhs ← {t € partition: t.dim>splitVal}
return Anonymizer(rhs) U anonymizer(lhs)
```

Figure 5. Mondrian Algorithm

3. Extended Mondrian Algorithm (XMondrian)

Even though the existing Mondrian algorithm is the widely used algorithm to protect identity disclosure but the algorithm can be applied only on dataset that contains numerical attributes. If a dataset contains categorical attributes, then the algorithm needs to convert the categorical attribute values into numerical values in order to apply the Mondrian algorithm. Again the corresponding numerical values of the categorical attribute need to convert into categorical values. This process of encoding and decoding takes anonymization time. To overcome this limitation, we extend the Mondrian algorithm as XMondrian algorithm which can handle both numerical and categorical attribute without encoding and decoding.

XMondrian algorithm recursively partitions the dataset into equivalence classes based on quasi identifier that has largest normalized range called splitattribute. Once split attribute is identified, split Value need to be identified. If the splitAttribute is categorical,for each distinct value of the attribute, a partition is created and if the split attribute is numerical, the partition depends on the median of the split attribute. The

process continues until the size of the partition is less than 2k-1.Thus, XMondrianproduces smaller size equivalence classes when compare to Mondrian, which indicates less information loss and less anonymization time.

Algorithm: *Xmondrian*

```
Anonymize(partition)
if (no allowable multidimensional cut for partition)
return ϕ : partition → summary
else
        dim ← choose-dimension()
fs ← frequency-set (partition, dim)
SplitAttribute ← largestWideRange(fs)
if splitAttribute is Numerical
splitVal ← find-median(splitAttribute)
lhs ← {t € partition: t.dim ≤ splitVal}
rhs ← {t € partition: t.dim>splitVal}
return Anonymizer(rhs) U anonymizer(lhs)
else
splitValue ← distinct-values(SplitAttribute)
partition ← {t € partition: t.dim =splitValue}
return Anonymizer (partition)
```

Figure 6.XMondrian Algorithm

4. Dataset

The ADULT dataset acquired from the UCI Repository is utilized to compare diverse sorts of k-anonymity algorithms. The dataset consist the statisticsof 30,162 people from the US enumeration in 1990.The narrative for the ADULT Dataset is shown in Figure 7[9].

Number	Attribute	Card	Type
1	Age	74	Numerical
2	Gender	2	Categorical
3	Race	5	Categorical
4	Marital Status	7	Categorical
5	Native Country	41	Categorical
6	Work Class	8	Categorical
7	Occupation	14	Categorical
8	Education	16	Categorical
9	Income	2	Numerical

Figure 7. Description of ADULT Dataset

The below figure presents a descriptive example of10 records of the ADULT Dataset.

ID	QID's				SA
Tuple No.	Age	Workclass	Marital_Status	Native_country	Income
1	39	State-gov	Never-married	US	<=50K
2	50	Self-emp-not-inc	Married-civ-spouse	US	<=50K
3	38	Private	Divorced	US	<=50K
4	53	Private	Married-civ-spouse	US	<=50K
5	28	Private	Married-civ-spouse	Cuba	<=50K
6	37	Private	Married-civ-spouse	US	<=50K
7	49	Private	Married-spouse-absent	Jamaica	<=50K
8	52	Self-emp-not-inc	Married-civ-spouse	US	>50K
9	31	Private	Never-married	US	>50K
10	42	Private	Married-civ-spouse	US	>50K

Figure 8. Sample Adultdata

Figure 9 demonstrates a 2-anonymous edition of Figure 8, means each equivalence class has at least 2*2-1 records same with respect to quasi identifiers.To attain anonymity, the direct identifier is expelled and the QIDs have been generalized .

EQs	QIDS				SA
1	[31-52]	Employed	*	US	>50K
	[31-52]	Employed	*	US	>50K
	[31-52]	Employed	*	US	>50K
2	[31-52]	Employed	*	US	<=50K
	[31-52]	Employed	*	US	<=50K
3	[31-52]	Private	Spouse-absent	North America	<=50K
	[31-52]	Private	Spouse-absent	North America	<=50K
4	[25-55]	Private	Married-civ-spouse	North America	<=50K
	[25-55]	Private	Married-civ-spouse	North America	<=50K
	[25-55]	Private	Married-civ-spouse	North America	<=50K

Figure 9. A2-Anonymous Version of Figure 8 after Applying Mondrian Algorithm

Figure 10.A 2-Anonymous Version of Figure 8 after Applying XMondrian Algorithm

EQs	QIDS				SA
1	[31-52]	Employed	*	US	>50K
	[31-52]	Employed	*	US	>50K
	[31-52]	Employed	*	US	>50K
2	39	State-gov	Never-married	US	<=50K
3	50	Self-emp-not-inc	Married-civ-spouse	US	<=50K
4	38	Private	Divorced	US	<=50K
5	49	Private	Married-spouse-absent	Jamaica	<=50K
6	[25-55]	Private	Married-civ-spouse	North America	<=50K
	[25-55]	Private	Married-civ-spouse	North America	<=50K

5.Comparison metrics

In terms of information loss and anonymization time, we equate our proposed algorithm XMondrian to the current algorithm. In certain cases, privacy-preserving algorithms alter datasets by adding false data or generalising and suppressing the original values. It's obvious that the more data is disguised, the less accessible it is for data analysts and researchers. As a result, the most important aspect of the metrics is quantifying the data quality after anonymization. There are a number of data quality indicators that can be used to measure the utility of data after it has been anonymized. 1) Average Equivalence Class Size Metric (CAVG) and 2) Anonymization Time are two common metrics for assessing data quality after anonymization.

5.1. Average Equivalence Class Size Metric (C_{AVG})

This metric is to measure the average size of the equivalence classes (EQs) in the anonymized dataset. The aim of this metric is to reduce the penalty, so if C_{AVG} is 1, it

implies that the anonymization is fine, with the size of the EQs equal to the provided K value. [2] .The total C_{AVG}scorefor an anonymized table T* is given by:

$$C_{AVG}(T^*)=|T|/(|EQs|.k)$$

Where T denotes the original table, |T| denotes the number of documents, |EQs| denotes the total number of equivalence classes generated, and k denotes the privacy requirement.[11].Figure 9 which is obtained from Mondrian shows 4 EQs, and thus the C_{AVG} score for the whole table is calculated as: 10/(4*2) = 1.25 and consider Figure 10 which is obtained by applying XMondrian Algorithm shows 6 EQs, and thus the C_{AVG} score for the whole table is calculated as :10/(6*2) = 0.8333

5.2 Anonymization Time in seconds

Performance of the algorithms can be evaluated using anonymizationtime.Theanonymization time of XMondrian algorithm is less than Mondrian because XMondrian algorithm doesn't require encoding and decoding of categorical values. We achieved 0.009 sec for XMondrain while Mondrain algorithm was executed in 0.014 sec. This shows that XMondrain performs better for any number of records. This is further explained in detail in Section 5.2.

6. Experimental Evaluation

6.1 Experimental Setup

To Experiment both Mondrian and XMondrian algorithms, we used the adult datasets as described in Section 3. Theconfigurations used in these experiments are shown in Figure 11: Various parameters inthese experiments are:

- #QIDs : Defines Number of Quasi Identifiers
- k-value : Defines the privacy level
- Dataset size : Defines the number of records in the dataset.

Figure 11. Parameters chosen for Experimentation

Number	Experiment	Parameter Setting	Dataset(Size)
1	Varied #QIDs	k-value = 2 #QIDs € [2..4]	ADULT (10)
2	Varied k-value	k-value €[3...7]&#QIDs= 4	ADULT (50)
3	Varied Size	k-value=2, #QIDs= 4	ADULT Datasize€[50...250]

6.2 Results

6.2.1. Varied Dataset Size with constant K Value =2

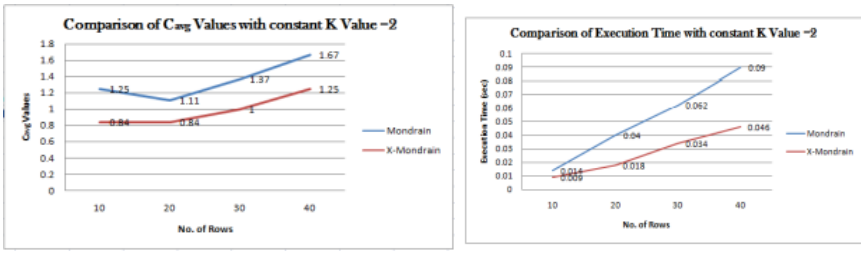

Figure 12. Varied Dataset Size with constant K Value =2

We can see that the Cavg values are getting closer to 1 and are slowly increasing as the number of rows increases, with a constant k value of 2.We can also compare the execution time with varied number of records and constant k value. The execution time always lesser for XMondrian compared to Mondrian. This proves that XMondrian is efficient when the data size is varied keeping the number of QIDs and K value to be constant.

6.2.2 Varied K Values with constant Number of Rows

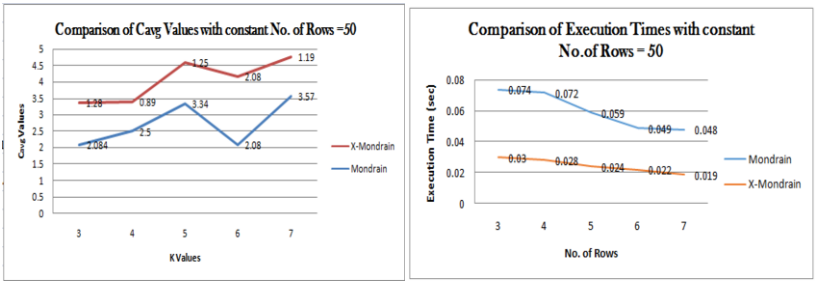

Figure 13.Varied K Values with constant Number of Rows

We can see that the Cavg values are closer to 1 and with varied trends with increase in the number of k values keeping number of rows and number of QIDs value as constant, while the Cavg values of Mondrian are extremely away from the equilibrium in case of increase in K value. We can also compare the execution time with varied K value and constant number of records = 50. The execution time is always lesser for XMondrian compared to Mondrian. The difference in the execution time between Mondrian and XMondrian is also significant. This proves that XMondrian is efficient when the K value is varied keeping the number of QIDs and data size to be constant.

6.2.3 Varied No. of QID

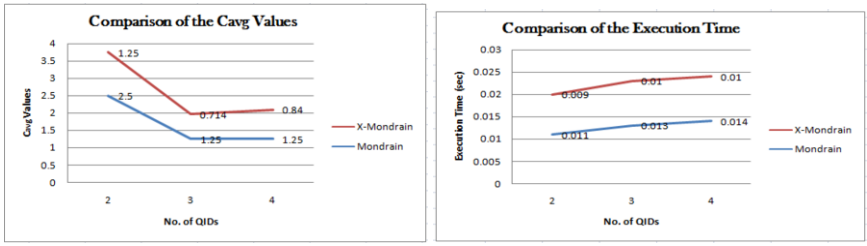

Figure 14. Varied No. of QID

We can see that the Cavg values are closer to 1 and with varied trends with increase in the number of QID values keeping number of rows and K value as constant, but it is not the case in Mondrian. We can also compare the execution time with varied number of QIDs value and constant number of records = 10 and constant K value being 2. The execution time is always lesser for XMondrian compared to Mondrian. This proves that XMondrian is efficient when the number of QIDs value is varied keeping the number of K value and data size to be constant.

7. Conclusion

We have conducted few experiments on both the algorithms to occur at conclusions on 3 important parameters to measure the data privacy and data loss. Cavg is (number of rows)/k*(number of equivalent classes)When ‘k’ is made constant the number of equivalent classes determine the difference in Cavg between Mondrian and XMondrian. The number of partitions in XMondrian will be greater than or equal to the number of partitions in Mondrian generally. This can be based on the working ofXMondrian. This algorithm groups like records together of categorical nature, whereas Mondrian splits the dataset into two parts per iteration depending on the median value. While the number of equivalent classes generated per iteration in XMondrian is not fixed. This tends to give ‘x’ number of partitions for a partition which might be greater than or equal to 2. This way XMondrian generates more partitions compared to Mondrian, which in turn gives lower Cavg value compared to Mondrian. This Cavg tends to have better proximity to 1, which is a more desired quality. The execution time of XMondrian is always lesser than Mondrian’s. This reflects the fact that there is no need for encoding and decoding for categorical attributes. This time is reduced in the XMondrian algorithm. The excessive data that is stored online is invariably in the danger of being exploited. The privacy of data needs to be preserved, which is a topic of ethical considerations. With the implementation of our proposed algorithm, we can end the many issues related to unpreserved data. Throughout our paper we have tested the extended model of Mondrian called "The XMondrian", that has excelled in every aspect. It outshone the existing algorithm called Mondrian in all ways possible. Our test results have refined most of the features of Mondrian to give better results with our new algorithm . We intend to extend the scope of this experiment to the BigData and implement MapReduce version of XMondrian. This enables us to preserve the BigData with higher potency, in an efficient way[10].

References

[1] Machanavajjhala, A., Gehrke, J., Kifer, D., &Venkitasubramaniam, M. (2006). ℓ-Diversity, Privacy beyond k-anonymity, in Proceedings of the 22nd International Conference on Data Engineering, ICDE '06 (pp. 24).[1617392] (Proceedings - International Conference on Data Engineering; Vol. 2006). https://doi.org/10.1109/ICDE.2006.1

[2] N. Li, T. Li and S. Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 2007, pp. 106-115, doi: 10.1109/ICDE.2007.367856.

[3] Fung, Benjamin & Wang, ke& Chen, Rui& Yu, Philip. (2010), Privacy-Preserving Data Publishing: A Survey of Recent Developments, ACM Comput. Surv..42. 10.1145/1749603.1749605.

[4] Ghinita, G., Karras, P., Kalnis, P., &Mamoulis, N. (2007), Fast Data Anonymization with Low Information Loss. VLDB, 2007.

[5] ArisGkoulalas-Divanis, Grigorios Loukides, Jimeng Sun, Publishing data from electronic health records while preserving privacy: A survey of algorithms, Journal of Biomedical Informatics, Volume 50, 2014, Pages 4-19, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2014.06.002.

[6] Natgunanathan, Iynkaran& Xiang, Yong &Hua, Guang&Guo, Song. (2016). Protection of Big Data Privacy. IEEE Access. 4. 1-1. 10.1109/ACCESS.2016.2558446.

[7] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, Protection of Big Data Privacy, in IEEE Access, vol. 4, pp. 1821-1834, 2016, doi: 10.1109/ACCESS.2016.2558446.

[8] Latanyasweeney,. (2012), k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10.10.1142/S0218488502001648.

[9] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, Mondrian Multidimensional K-Anonymity, *22nd* International Conference on Data Engineering *(ICDE'06)*, Atlanta, GA, USA, 2006, pp. 25-25, doi: 10.1109/ICDE.2006.101

[10] Khan R, Tao X, Anjum A, Kanwal T, Malik SuR, Khan A, Rehman Wu, Maple C, θ-Sensitive *k*-Anonymity: An Anonymization Model for IoT based Electronic Health Records. Electronics*, 2020; 9(5):716. https://doi.org/10.3390/electronics9050716.*

[11] Ayala-Rivera, Vanessa &Mcdonagh, Patrick &Cerqueus, Thomas & Murphy, Liam. (2014), A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners.Transactions on Data Privacy, 7.337-370.

[12] Samarati.P . Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 6, pp. 1010-1027, Nov.-Dec. 2001, doi: 10.1109/69.971193.

[13] Wei Fang, XueZhi Wen, Yu Zheng& Ming Zhou (2017), A Survey of Big Data Security and Privacy Preserving, IETE Technical Review, 34:5, 544-560, DOI: 10.1080/02564602.2016.1215269.

[14] A. Gao and L. Diao, Privacy preservation for attribute order sensitive workload in medical data publishing, 2009 IEEE International Symposium on IT in Medicine & Education, Jinan, China, 2009, pp. 1140-1145, doi: 10.1109/ITIME.2009.5236250.

[15] Q. Gong, M. Yang, Z. Chen and J. Luo, Utility enhanced anonymization for incomplete microdata, 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Nanchang, China, 2016, pp. 74-79, doi: 10.1109/CSCWD.2016.7565966.