Variable Distinct l-diversity Algorithm Applied on Highly Sensitive Correlated Attributes

Zakariae El Ouazzani and Hanan El Bakkali
Information Security Research Team - ISeRT
ENSIAS-Mohammed V University
Rabat, Morocco
email: zakariae.elouazzani@gmail.com and h.elbakkali@um5s.net.ma

***Abstract*—In this information age, large amount of data is available online. These data are used by both internal and external sources for analysis and research purposes. The collected data is stored into huge data sets containing sensitive and Non−Sensitive Attributes. For the reason that attributes are generally separated, the correlation between these various attributes is lost. Thus, it will be necessary to prevent attributes from losing the correlation between them or at least reduce the correlation loss. As a solution, correlated attributes are grouped together. Although, the data utility is preserved by reducing the correlation loss between Sensitive Attributes, privacy protection remains a serious concern. The main problem here is publishing data sets without revealing the sensitive information of individuals and in the same time preserving data utility. Most of the current researches on ensuring privacy in big data are centered on data anonymization. L-diversity is an anonymization technique that can be applied on a data set with one or multiple Sensitive Attributes. This paper proposes an algorithm that deals with sensitive numerical and non−numerical attributes. The algorithm applies the principle of l-diversity technique after grouping highly correlated attributes together through a vertical partitioning. Our proposed algorithm makes a balance between privacy and data utility.**

***Index Terms*—big data; anonymization; l-diversity technique; non−numerical attributes; correlation; Pearson.**

I. Introduction

In recent years, the data collected by public and private organizations are increasing every day and stored in electronic repositories. The collected data includes various types of attributes, especially sensitive ones [1]. Besides, Big data sets can be used in different sectors, for example, biology, online banking, medical research and so on [2]. However, more challenges are rising since the collected data includes sensitive information [1]. The first challenge is preserving data utility. Because each attribute is universally separated, the correlations between different Sensitive Attributes are lost. This will be a major problem when performing analysis about data utility [3]. Thus, we have to reduce the correlation loss between attributes by grouping highly correlated attributes together. However, even if the data utility is preserved by dividing the huge data set into various data sets containing only highly correlated attributes, the challenge of ensuring privacy remains a crucial issue when sharing a data set that contains personal information [4]. Current information technologies create vast amount of data characterized by velocity, volume and veracity. So, disseminating this data increases the possibility of violating the privacy of individuals. That's why privacy protection is considered as one of the most important issues in big data processing [5]. In order to ensure privacy, data has to be sanitized and the best way of sanitization is data anonymization. There are several anonymization techniques treating Sensitive Attributes in the literature, one of them is called "l-diversity" using horizontal partitionning. The main idea behind "l-diversity" is that the values of the Sensitive Attributes are well−represented in each bucket [6]. In this paper, a new algorithm of data anonymization is proposed. It is a variable distinct l-diversity algorithm applied on highly sensitive correlated attributes whatever its type: numerical or non-numerical. The algorithm makes a balance between data privacy and data utility. Besides, it is divided into two main parts. The first one is intended for preserving data utility by grouping highly correlated attributes together in several data sets. We used "Pearson" correlation tool to determine the highly correlated attributes. Although, "Pearson" tool processes numerical values only, we used an algorithm that converts non−numerical values into numerical ones to process non−numerical attributes too. The second part used the l-diversity principle by splitting the data set horizontally into buckets including distinct values in order to ensure privacy. In this paper, we try to prove that the l-diversity principle must only be applied on data sets including highly correlated attributes; otherwise, l-diversity will not be an effective anonymization technique.

The reminder of the paper is organized as follows: in Section 2, we will make an overview of some works found in literature using l-diversity technique in order to ensure privacy in big data. Next, in Section 3, we will present the proposed technique including the algorithm. Later, in Section 4, we give our experimental results applied on a part of a real data set. Finally, we conclude our paper and give some perspectives in Section 5.

II. Related Work

Generally, l-diversity technique aims to ensure privacy in huge data sets. In most of cases, l-diversity is applied on a data set while the threshold l is fixed to a specific value. Besides, the degree of correlation between attributes is not considered. For instance, Priyadarsini et al. in [7] proposed an Enhanced l-diversity algorithm able to diversify several Sensitive Attributes without dividing the data set. The proposed

ISBN: 978-1-61208-719-1

algorithm attempts to support multiple Sensitive Attributes for l-diversity by applying certain conditions to determine the size of the bucket. Moreover, Priyadarsini et al. in [7] accommodate the values corresponding to the sensitive categorical attributes within each bucket by setting the value of the threshold l based on the occurrence of distinct values in the whole column. Besides, Sei et al. in [8] suggested a privacy model called $(l_1, ..., l_q)$-diversity, which can deal with databases including various sensitive Quasi-Identifier (QI) attributes. The proposed method in [8] does not make any modifications on the original data set but adds various random values to each attribute in the data set to realize $(l_1, ..., l_q)$-diversity. Therefore, the threshold l is set to a fixed value. Moreover, Oishi et al. in [4] presented (l, d)-semantic diversity algorithm considering the resemblance of sensitive attribute values within each bucket by adding distances to settle the problem of impossibility to satisfy the threshold l of l-diversity. The algorithm in [4] satisfies l-diversity through a method based on adding a Boolean indicator to every sensitive attribute without generalizing the Quasi-Identifier attributes. Also, Gaoming et al. in [9] proposed a (k, l, θ)-diversity model based on clustering to reduce information loss and increase the usefulness of data. The algorithm in [9] takes as input three parameters, the thresholds k and l correspond to k-anonymity and l-diversity techniques respectively and the parameter θ corresponds to the degree of privacy preserving. Additionally, A new technique using the principle of l-diversity is presented by Y. Sei and Ohsuga in [10], which randomizes the Sensitive Attributes belonging to each individual. The method in [10] is divided into two parts; the first one concerns the data holder where $l-1$ random values are generated and added to a sensitive attribute in the whole original data set. The second one concerns the data user where the user has the possibility to identify the QI attributes that should be analyzed based on the relation between QI attributes and sensitive ones. Furthermore, Chakraborty et al. in [11] proposed (α, l) and recursive (α, c, l) diversity techniques. Both eigenvector centrality and noise node addition concepts are used in the process in order to create an anonymized network. In other sector, Tu et al. in [12] proposed a heuristic algorithm in order to get an approximate solution. The algorithm meets l-diversity principle for protecting trajectory privacy through specific generalization, while guaranteeing the smallest loss of spatiotemporal granularity. Besides, R. Yogesh Kulkarni and Murugan in [13] proposed an algorithm called, CPGEN (C-mixture based Privacy GENetic algorithm) in order to ensure privacy. The method in [13] combines the genetic algorithm with C-mixture theory for privacy measurements. The C-mixture is a new privacy measure, which integrates various privacy constrains belonging to both k-anonymity and l-diversity principles. Moreover, Susan and Christopher in [14] suggested an anonymization technique by combining the advantages of anatomization, and an improved slicing technique using both k-anonymity and l-diversity principles to treat high dimensional data sets, which include various Sensitive Attributes. The anatomization approach reduces the information loss and slicing algorithm preserves the correlation and utility.

The main idea of this paper is inspired from the previous works and we assume that l-diversity principle has to be applied only on highly correlated attributes in order to ensure privacy and preserve utility. In the next section, we will present our proposed technique including the algorithm that applies the principle of l-diversity on a data set containing attributes having strong correlation between them.

III. The proposed Technique and Some Related Concepts

Our proposed technique ensures privacy by applying l-diversity principle on highly correlated attributes. Besides, the technique preserves data utility by grouping every two highly correlated attributes in a data set.

A. L-diversity and Correlation

1) Correlation analysis: With data analysis techniques, precious information could be extracted from big data. In data analysis, big data technologies includes data mining, machine learning and correlation analysis [15]. Correlation is a well-known mathematical and statistical method for analyzing the compatibility of huge data sets [15]. Since each attribute is generally separated and thus distinguishable, the correlation between various attributes is lost. This is considered as an inherent issue to make efficient analysis of attribute correlations [3]. In order to reduce the correlation loss, a partitioning approach is proposed in [16] based on the lexicographic and Non-Sensitive Attributes (NSAs) sorted by correlation between NSAs and Sensitive Attributes (SA). Besides, this approach preserves the published data utility [16]. Authors in [3], [16]–[18] used vertical partitioning by grouping attributes into columns according to the correlations existing between these attributes where only highly correlated ones are grouped into columns. The main idea is to break the association between columns while preserving the relationship within each column [3] and [17]. The fact of grouping highly correlated attributes together minimizes the high dimensionality of the data set [17] and [18], moreover, it preserves better utility than generalization and bucketization approaches [17]. Besides, as mentioned in [3], [14], [17], [18], slicing technique preserves data utility because highly correlated attributes are grouped together while conserving the correlations between such attributes. The evaluation of the correlations between the pairs of attributes could be realized through several correlation tools depending on the type of the treated attributes. For instance, Pearson correlation coefficient is utilized to evaluate the correlation between two continuous attributes [18], whereas mean-square contingency coefficient is a chi-square measure of correlation between two categorical attributes [16] and [18].

In this paper, we used Pearson tool to identify the highly correlated attributes whether they are numerical or not. In the case we have non-numerical attributes in the data set; we convert non numerical values into numerical ones through a proposed converting algorithm. The algorithm gives the same number to similar values in the data set. This conversion will give us the opportunity to process different types of data. The Pearson correlation tool is used to calculate the degree of linear correlation between two numerical attributes through 1 [19].

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x}) \sum_i (y_i - \bar{y})}} \tag{1}$$

Where $\bar{x}$= the mean of x variable.
and $\bar{y}$= the mean of y variable.

The correlation here is the sum of the multiplication between corresponding numbers related to the treated attributes. Besides, the resulting correlation values are in the range $[-1.0, +1.0]$. After calculation Pearson correlation coefficient for all the pairs of attributes existing in the data set, we identify those corresponding to the highest value in order to apply the l-diversity principle on a data set containing only highly correlated attributes.

2) L-diversity principle: Most of anonymization techniques existing in the literature are applied before publishing the data set [20]. Some of these techniques deal with quasi identifier attributes and others deal with sensitive ones. In this paper, a technique using the principle of distinct l-diversity is suggested dealing with Sensitive Attributes. Besides, the proposed variable l-diversity technique doesn't take into consideration any prior value of the threshold l. Furthermore, the principle of l-diversity has been introduced to improve traditional data mining that preserves privacy. l-diversity is considered as an important technique in privacy protection [21]. L-diversity is a group based form of anonymization used to ensure privacy in huge data sets by minimizing the huge scale of big data in term of representation [21]. The l-diversity model (Distinct, Entropy, Recursive) is an extension of the k-anonymity technique, which deal with QI attributes [22] and [23]. L-diversity ensures that an adversary needs $l-1$ values using background knowledge to deduce $l-1$ possible values of a sensitive attribute in order to violate privacy [9] and [24]. In other words, an equivalence class (EC), also called bucket is deemed to satisfy l-diversity if there are at least l "well-represented" values related to the treated Sensitive Attributes (SAs) [6], [24], [25]. Then, the whole data set is deemed to satisfy l-diversity when every bucket existing in that data set satisfies l-diversity [24] and [25]. Moreover, l-diversity helps to mitigate both homogeneity and background knowledge attacks [22] and [25]. Existing methods for l-diversity only take into consideration l "well represent" sensitive values. However, they omit the size of every bucket in the data set. Thus, the loss of information in the published data sets is much larger, which lead to a decrease concerning the data utility [9]. Our proposed algorithm applies the principle of distinct l-diversity without a prior value of the threshold "l". That means that the value of l is not fixed, so there is an opportunity to maximize this value in order to ensure privacy as much as possible. In the next part of this section, we will present our proposed algorithm, which applies the principle of variable distinct l-diversity on highly correlated attributes.

B. The proposed Algorithm

Algorithm 1 L-diversity on highly correlated attributes algorithm

```
 1: procedure ANONYMIZATION
 2:     $OriginalTable[1 \rightarrow N]$ $struct$ $attr1(String)$ $attr2(String)...attrL(String)$ $end$ $struct$
 3:     $D1[1 \rightarrow N]$ $struct$ $attr1(String)$ $attr2(String)...attrL(String)$ $end$ $struct$
 4:     $D2[1 \rightarrow N]$ $struct$ $attr1(String)$ $attr2(String)...attrL(String)$ $end$ $struct$
 5:     $RT[1 \rightarrow N]$ $struct$ $attr1(String)$ $attr2(String)...attrL(String)$ $end$ $struct$
 6:     Conversion($OriginalTable$)
 7:     $hc \leftarrow 0$
 8:     $indi \leftarrow 0$
 9:     $indj \leftarrow 0$
10:     $p \leftarrow 0$
11:     $find \leftarrow 0$
12:     $i \leftarrow 1$
13:     while $i<L-1$ do
14:         $j \leftarrow i+1$
15:         while $j<L$ do
16:             $p = pearson(OriginalTable[.].attr[i], OriginalTable[.].attr[j])$
17:             if $hc<p$ then
18:                 $hc \leftarrow p$
19:                 $indi \leftarrow i$
20:                 $indj \leftarrow j$
21:             $j++$
22:         $i++$
23:     repeat
24:         $D1.put(OriginalTable[0])$
25:         $i \leftarrow 1$
26:         while $i<N$ do
27:             $find \leftarrow 0$
28:             if $D1.Contains(OriginalTable[i].attr[indi])$ then
29:                 $find \leftarrow 1$
30:                 if $find = 1$ then
31:                     $RT.put(OriginalTable[i])$
32:                 else
33:                     $D1.put(OriginalTable[i])$
34:             $i++$
35:         $D2.put(D1[0])$
36:         $i \leftarrow 1$
37:         while $i<D1.length()$ do
38:             $find \leftarrow 0$
39:             if $D2.Contains(D1[i].attr[indj])$ then
40:                 $find \leftarrow 1$
41:                 if $find = 1$ then
42:                     $RT.put(D1[i])$
43:                 else
44:                     $D2.put(D1[i])$
45:             $i++$
46:         $Clear(OriginalTable)$
47:         $Copy(OriginalTable, RT)$
48:     until $RT.isEmpty()$
```

Our algorithm is divided into two parts. The first one identifies the two highly correlated attributes among all the attributes in the Original Table. The second one presents the process of applying l-diversity principle. The anonymization process is applied on a Table containing N tuples and L attributes. The attributes are the fields of a structure.

In the first part, from line 6 to line 22 in the algorithm, the identification of the two highly correlated attributes is realized through a correlation tool called "Pearson". Since the data set could contain both numerical and non-numerical attributes and also Pearson tool processes only numerical attributes, we convert non-numerical attributes into numerical ones. Then, we calculate the correlation coefficient between every two attributes p in the Original Table. After that, we save the indexes $indi$ and $indj$ of the attributes corresponding to the highest correlation coefficient hc.

In the second part, from line 23 to line 48 in the algorithm, we apply l-diversity on the two attributes, which have the highest correlation. We start by identifying the distinct values corresponding to the first attribute, then, we put these tuples in $D1$ Table, the remaining tuples are put in RT Table. However, Table $D1$ may still contain non distinct values with respect to the second attribute. Then, we copy distinct tuples in Table $D1$ with respect to the second attribute in Table $D2$, besides, we add the remaining tuples in $D1$ to RT Table. Thus, $D2$ is the l-diversity Table with distinct values with respect to both highly correlated attributes. Once the process ends, we clear the Original Table and we copy the content of RT Table in the original table and we repeat the process of l-diversity until RT Table is empty. The complexity of the anonymization part of the algorithm is of the order of $N^{2*NbBuckets}$ where $NbBuckets$ is the number of buckets existing in the data set and N is the number of tuples in the same data set. Therefore, when we analyse the bloc ”repeat-until”, we find that there are two loops inside. The first while loop processes $NbBuckets$ operations because we identify distinct values corresponding to the first attribute in the $OriginalTable$. Later, in the second while loop we identify distinct values corresponding to the second attribute based on the result table of the previous loop. Then, we analyse the ”repeat-until” bloc and according to the algorithm, we notice that the process is repeated N times, which is the number of lines in the $OriginalTable$.

In the next section, we will highlight the different steps of the proposed algorithm applied on a part of real data set.

IV. Experimental Results

Now, we will implement our algorithm on a test table related to health sector. We have developed our algorithm with Java tool. Table I is a part of a real data set called ”careplans” [26], which contains several attributes like ”Disease”, ”Treatment”, ”Date of diagnosis” and ”Cure date”. I mention that I have randomly selected 9 tupes from the ”careplans” real data set.

TABLE I
The Original Table.

Id	Disease	Treatment	Date of diagnosis	Cure date
1	Whiplash injury to neck	Recommendation to rest	04/09/2015	27/09/2015
2	Whiplash injury to neck	Musculoskeletal care	15/02/2008	17/03/2008
3	Fracture of forearm	Recommendation to rest	18/12/2007	04/02/2008
4	Gout	Healthy diet	18/01/1968	24/09/1975
5	Gout	Musculoskeletal care	18/01/1968	24/09/1975
6	Rheumatoid arthritis	Ice therapy	16/12/2005	13/08/2010
7	Whiplash injury to neck	Recommendation to rest	28/12/1942	05/02/1943
8	Gout	Healthy diet	18/01/1968	24/09/1975
9	Rheumatoid arthritis	Healthy diet	16/12/2005	13/08/2010

Table I represents our original test table. Besides, Table II represents Table I after the application of the conversion process. We substitute non-numerical values (String and Date types) by numerical ones.

TABLE II
The original table after anonymization.

Id	Disease	Treatment	Date of diagnosis	Cure date
1	1	5	9	15
2	1	6	10	16
3	2	5	11	17
4	3	7	12	18
5	3	6	12	18
6	4	8	13	19
7	1	5	14	20
8	3	7	12	18
9	4	7	13	19

After converting non numerical values, we calculate the correlation between every two attributes in the data set. First, we calculate the correlation between ”Disease” attribute and the other attributes in the data set. In the following, we give the corresponding Pearson correlation coefficients:

$$r(Disease,\ Treatment) = 0.8431$$

$$r(Disease,\ Date\,of\,diagnosis) = 0.5103$$

$$r(Disease,\ Cure\,date) = 0.5103$$

The correlation between ”Disease” and ”Treatment” attributes is strong and positive. However, there is a moderate positive correlation between ”Disease” and ”Date of diagnosis”, the same moderate correlation is between ”Disease” and ”Cure date” attributes.

The Pearson correlation coefficient between ”Disease” and ”Treatment” equals 0.8431, which is the highest value among the three correlation values calculated between ”Disease” attribute and the other attributes in the data set. The l-diversity principle will be applied on a part of Table I, which contains only ”Disease” and ”Treatment” attributes.

Second, we calculate the correlation between ”Treatment”, ”Date of diagnosis” and ”Cure date” attributes. Here are the values of the calculated Pearson correlation coefficients:

$$r(Treatment,\ Date\,of\,diagnosis) = 0.3983$$

$$r(Treatment,\ Cure\,date) = 0.3983$$

We remark that the correlation value between ”Treatment” and ”Date of diagnosis” attributes equals the correlation value between ”Treatment” and ”Cure date”. The relationship between the attributes is weak because the correlation value is near zero value.

Finally, we calculate the correlation between the last two attributes ”Date of diagnosis” and ”Cure date”.

$$r(Date\,of\,diagnosis,\ Cure\,date) = 1$$

The calculation of Pearson correlation coefficient between ”Date of diagnosis” and ”Cure date” gives a value of 1, which means that there is a strong positive correlation between these two attributes.

Now, we will process by applying 1-diversity on Table I with respect to ”Disease” and ”Treatment” attributes corresponding to the highest value of Pearson correlation coefficient (0.8431).

We are going to highlight through different tables the whole steps until we obtain an anonymized table satisfying 1-diversity. Table III represents Bucket 1 where all the tuples contain distinct values when treating both ”Disease” and ”Treatment” attributes.

TABLE III
BUCKET 1.

Id	Disease	Treatment	Bucket
1	Whiplash injury to neck	Recommendation to rest	1
4	Gout	Healthy diet	1
6	Rheumatoid arthritis	Ice therapy	1

In the first step, we collect the distinct values from ”Treatment” attribute column, which are ”Recommendation to rest”, ”Musculoskeletal care”, ”Healthy diet” and ”Ice therapy”. Then, we put in Bucket 1 the tuples corresponding to the already mentioned distinct values with ascendant order. We can see that ”Recommendation to rest” and ”Musculoskeletal care” values correspond to ”Whiplash injury to neck” value, then we will retain only ”Recommendation to rest” attribute because it is the first value in the order. However, ”Healthy diet” and ”Ice therapy” correspond to distinct values, which are ”Gout” and ”Rheumatoid arthritis” values. Then, we obtain the first bucket satisfying 1-diversity as mentioned in Table III. In the next step, we put the remaining tuples from Table I in another table called Rest of table $RT1$.

TABLE IV
REST OF TABLE RT 1.

Id	Disease	Treatment
2	Whiplash injury to neck	Musculoskeletal care
3	Fracture of forearm	Recommendation to rest
5	Gout	Musculoskeletal care
7	Whiplash injury to neck	Recommendation to rest
8	Gout	Healthy diet
9	Rheumatoid arthritis	Healthy diet

Table IV is called $RT1$; it contains tuples other than those existing in Bucket l. This table takes the place of the Original Table in the remaining of the proposed algorithm. Table V corresponds to Bucket 2.

TABLE V
BUCKET 2.

Id	Disease	Treatment	Bucket
2	Whiplash injury to neck	Musculoskeletal Mcare	2
3	Fracture of forearm	Recommendation to rest	2
8	Gout	Healthy diet	2

Table V includes three tuples containing distinct values with respect to ”Disease” and ”Treatment” attributes. Consequently, Table V satisfies 3-diversity.

TABLE VI
BUCKET 3 AND REST OF TABLE $RT2$.

Id	Disease	Treatment	Bucket
5	Gout	Musculoskeletal care	3
7	Whiplash injury to neck	Recommendation to rest	3
9	Rheumatoid arthritis	Healthy diet	3

Table VI represents the Rest of table $RT2$ and in the same time Bucket 3 since all the tuples existing in this table are all of them containing distinct values. And here we obtain 3 buckets satisfying l-diversity.

TABLE VII
THE ORIGINAL TABLE AFTER ANONYMIZATION.

Disease	Treatment	Date of diagnosis	Cure date	Bucket
Whiplash injury to neck	Recommendation to rest	04/09/2015	27/09/2015	1
Gout	Healthy diet	18/01/1968	24/09/1975	1
Rheumatoid arthritis	Ice therapy	16/12/2005	13/08/2010	1
Whiplash injury to neck	Musculoskeletal Mcare	15/02/2008	17/03/2008	2
Fracture of forearm	Recommendation to rest	18/12/2007	04/02/2008	2
Gout	Healthy diet	18/01/1968	24/09/1975	2
Gout	Musculoskeletal care	18/01/1968	24/09/1975	3
Whiplash injury to neck	Recommendation to rest	28/12/1942	05/02/1943	3
Rheumatoid arthritis	Healthy diet	16/12/2005	13/08/2010	3

We notice that we will reapply all the steps of l-diversity algorithm on a Table containing ”Date of diagnosis” and ”Cure date” attributes since there is a strong correlation between them.

Since Tables III, V and VI satisfy the principle of distinct l-diversity, we could say that Table VII satisfies distinct l-diversity too. Besides, we remark that at least there exist three tuples within each bucket in Table VII. Consequently, the resulting table after the anonymization process is called 3-diversity table.

V. CONCLUSION AND PERSPECTIVES

This paper presents a new approach for data anonymization. The approach focuses on anonymizing data sets while preserving the data utility. First, we applied a conversion process on values in the data set by transforming non-numerical values into numerical ones. After that, we grouped the pairs of attributes with the highest correlation together into several data sets through the calculation of Pearson correlation coefficient. Consequently, the data utility is preserved by reducing the correlation loss between the grouped highly correlated attributes. Later and in order to ensure privacy, we apply a variable distinct l-diversity on highly correlated attributes algorithm throughout a horizontal partitioning until treating all the buckets in data set. Besides, our proposed algorithm makes a balance between privacy and data utility. As a perspective, we plan to compare our proposed technique with other anonymization techniques existing in the literature. In addition, we will test our algorithm on the large real data set ”Careplans”. Moreover, we plan to deal also with QI attributes by applying k-anonymity technique instead of l-diversity one.

References

[1] M. Prakash and G. Singaravel, ”An approach for prevention of privacy breach and information leakage in sensitive data mining,” Computers and Electrical Engineering, vol. 45, July 2015, pp. 134-140. DOI: http://dx.doi.org/10.1016/j.compeleceng.2015.01.016.

[2] Y. Qu, S. Yu, L. Gao and J. Niu, ”Big data set privacy preserving through sensitive attribute-based grouping” The 2017 IEEE International Conference on Communications (ICC 2017) Piscataway, N.J., 2017, pp. 4887-4892, doi: 10.1109/ICC.2017.7996330.

[3] K. Kiruthika, M.S Kavitha and S. Gayathiri, ”Publishing High-Dimensional Micro Data Using Anonymization Technique,” Imperial Journal of Interdisciplinary Research (IJIR), vol. 2(8), pp. 86-96, 2016, ISSN: 2454-1362.

[4] K. Oishi, Y. Tahara, Y. Sei and A. Ohsuga, ”Proposal of l-diversity algorithm considering distance between sensitive attribute values” The 2017 IEEE Symposium Series on Computational Intelligence (SSCI 2017), 2017, pp. 1-8. 10.1109/SSCI.2017.8280973

[5] J. Jeyanthi and J. Antony, “Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data,” Advances in Computational Sciences and Technology, vol. 10(2), 2017, pp. 247-253, ISSN 0973-6107.

[6] A. Shah, H. Abbas, W. Iqbal and R. Latif, ”Enhancing E-Healthcare Privacy Preservation Framework through L-Diversity” The 14th International Wireless Communications and Mobile Computing Conference (IWCMC 2018), June 2018 pp. 394-399. DOI:10.1109/IWCMC.2018.8450306

[7] R. Praveena Priyadarsini, S. Sivakumari and P. Amudha, ”Enhanced – Diversity Algorithm for Privacy Preserving Data Mining” Communications in Computer and Information (CSI), vol. 679, pp. 14-23, Nov. 2016, DOI:https://doi.org/10.1007/978-981-10-3274-5-2

[8] Y. Sei, H. Okumura, T. Takenouchi and A. Ohsuga, ”Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness” The IEEE Transactions on Dependable and Secure Computing (TDSC 2017), Apr. 2017, pp(99). 1-1. DOI: 10.1109/TDSC.2017.2698472

[9] Y. Gaoming, L. Jingzhao, Z. Shunxiang and Y. Li, ”An enhanced l-diversity privacy preservation,” The 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2013), Shenyang, 2013, pp. 1115-1120. DOI: 10.1109/FSKD.2013.6816364

[10] Y. Sei and A. Ohsuga, ”Randomized addition of sensitive attributes for l-diversity,” The 11th International Conference on Security and Cryptography (SECRYPT 2014), Vienna, Aug. 2014, pp. 1-11. ISBN: 978-9-8985-6595-2

[11] S. Chakraborty and B.K. Tripathy, ”Alpha-anonymization techniques for privacy preservation in social networks” Social Network Analysis and Mining Journal, vol. 6(29), Dec. 2016, pp. 1-11, DOI: 10.1007/s13278-016-0337-x.

[12] Z. Tu et al., ”Protecting Trajectory from Semantic Attack Considering k-Anonymity, l-diversity and t-closeness” The IEEE Transactions on Network and Service Management (TNSM 2018), Oct. 2018, pp(99). 1-1. 10.1109/TNSM.2018.2877790

[13] R. Yogesh Kulkarni and T. Senthil Murugan, ”C-mixture and multi-constraints based genetic algorithm for collaborative data publishing,” Journal of King Saud University-Computer and Information Sciences, vol. 30(2), pp. 175–184, Apr. 2018, https://doi.org/10.1016/j.jksuci.2016.06.001.

[14] V. Shyamala Susan and T. D. Dickman Christopher, ”Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes,” SpringerPlus, vol. 5: 964, July 2016, https://doi.org/10.1186/s40064-016-2490-0.

[15] H. Jiang, K. Wang, Y. Wang, M. Gao and Y. Zhang, ”Energy big data: A survey,” IEEE Access, vol. 4, 2016, pp. 3844-3861. doi: 10.1109/ACCESS.2016.2580581

[16] H. Zhu, S. Tian and M. Xie, ”Anonymization on refining partition: Same privacy, more utility” The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), Shanghai, 2014, pp. 998-1005. doi: 10.1109/ICSAI.2014.7009431

[17] P.Sreevani, P.Niranjan, P.Shireesha, ”A Novel Data Anonymization Technique for Privacy Preservation of Data Publishing,” International Journal of Engineering sciences and Research Technology (IJESRT), vol. 3(11), pp. 201-205, Nov. 2014.

[18] P.Nithya1, V.Karpagam, ”Improving Privacy And Data Utility For High-Dimensional Data By Using Anonymization Technique,” International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), vol. 2(1), pp. 2874-2881, Mar. 2014, ISSN: 2320-9801

[19] W. Feng, Q. Zhu, J. Zhuang and Y. Shimin, ”An expert recommendation algorithm based on Pearson correlation coefficient and FP-growth” Cluster Computing, Jan. 2018, DOI: 10.1007/s10586-017-1576-y.

[20] F. Li, X. Zou, P. Liu, JY. Chen, ”New threats to health data privacy,” BMC Bioinformatics, vol. 12 (Suppl 12):S7, Nov. 2011, DOI:10.1186/1471-2105-12-S12-S7.

[21] P. Jain, M. Gyanchandani and N. J. Khare, ”Big data privacy: a technological perspective and review,” Journal of Big Data, vol. 3(25), July 2016, DOI: 10.1186/s40537-016-0059-y.

[22] P. Dabas and S. Sharma, ”Privacy and Security Issues in Social Networks with Prevailing Privacy Preserving Techniques,” Journal of Network Communications and Emerging Technologies (JNCET), vol. 8(2), pp. 54-56, Feb. 2018, ISSN: 2395-5317.

[23] Z. EL Ouazzani, H. El Bakkali, ”A new technique ensuring privacy in big data: k-anonymity without prior value of the threshold k,” ELSEVIER First International Conference on Intelligent Computing in Data Sciences. (ICDS), Vol. 127, 2018, pp. 52-59.

[24] E. Elabd, H. Abd elkader and A. Mubarak Alhamodi, ”L–Diversity-Based Semantic Anonymaztion for Data Publishing,” International Journal of Information Technology and Computer Science (IJITCS), vol. 7, pp. 1-7, Sep. 2015, DOI:10.5815/ijitcs.2015.10.01.

[25] L. Philippe Sondeck, M. Laurent and V. Frey, ”The Semantic Discrimination Rate Metric for Privacy Measurements which Questions the Benefit of t-closeness over l-diversity” The 14th International Conference on Security and Cryptography (ICSC 2017), Jan. 2017, pp. 285-294. DOI:10.5220/0006418002850294

[26] https://syntheticmass.mitre.org/downloads/2017_1106/synthea_sample _data_csv_nov2017.zip accessed on 15 May 2019.