Articulation Point Based Quasi Identifier Detection for Privacy Preserving in Distributed Environment

Ila Chandrakar[1], Vishwanath R Hulipalled[2]

[1]Presidency University & Research Scholar, REVA University, Bangalore, India
[2]School of C&IT, REVA University, Bangalore, India

Abstract: These days, huge data size requires high-end resources to be stored in IT organizations premises. They depend on cloud for additional resource necessities. Since cloud is a third-party, we cannot guarantee high security for our information as it might be misused. This necessitates the need of privacy in data before sharing to the cloud. Numerous specialists proposed several methods, wherein they attempt to discover explicit identifiers and sensitive data before distributing it. But, quasi-identifiers are attributes which can spill data of explicit identifiers utilizing background knowledge. Analysts proposed strategies to find quasi-identifiers with the goal that these properties can likewise be considered for implementing privacy But, these techniques suffer from many drawbacks like higher time consumption and decreased data utility. The proposed work overcomes this drawback by extracting minimum required quasi attributes with minimum time complexity.

Keywords: Articulation point, Privacy Preserving, Quasi Identifier

1. Introduction

The speedy development in information technology gave birth to many social media websites which uses our personal data for their usage. These websites may not be trust worthy and leak user’s data. Similarly, due to digitization, there is huge increase in data size in all the organization which became a big problem for data owners. They are unable to afford huge requirement of resources and are using cloud for the need of extra resources. Since, cloud is a third party, it can be curious and try to leak data. Also, data are published by the organization for the purpose of the research, hence there is a need to implement privacy for the data before publishing it.

In past years, many incidents happened pertaining to leakage of data from organization. In 2019, personal contents of many email accounts were exposed from many accounts of Microsoft Office 365. In 2018, another huge data leak happened by Facebook in which 50 million user account personal details were exposed and another 40 million were exposed because of those accounts. In 2013, approximately all the users of Yahoo were affected as their username and password were leaked by Russian hackers. In 2014, the data like name, contact, and passport number etc. of 500 million customers of Marriot International was leaked and many more cases are there in which huge amount of personal data is leaked.

Personal data privacy is very important for any individual. This can be achieved by hiding the personal sensitive information before publishing it. The attributes in a data set can be categorized into three types. First is explicit identifiers which directly give information of the subject of interest i.e. from which one can easily identify a person and his details like name, social security number etc. Second is sensitive attributes which give identification of a person, but it is a private information of person which if gets leaked, may cause harm. For example, salary of an employee or type of disease of any person in health data set is sensitive information. Third type of attributes is quasi identifier attributes which are subset of attributes or combination of attributes which together can identify a person. For example, gender, zip code and date of birth can identify maximum population in USA. The earlier privacy preserving techniques were concerned about hiding explicit identifiers and sensitive attributes in the data set but it is still vulnerable as the quasi identifiers can be used to extract personal information of an individual using linking attacks. These quasi identifiers can easily leak information of explicit identifiers. Many researchers have used k-anonymity to overcome linking referred to as link between explicit identifiers and quasi identifiers [36] [37]. In k-anonymity method, generalization method has been applied on quasi identifiers to convert it into more generalized form. This can provide privacy to quasi attributes but only to some extent. But the problem is how to find the optimum number of quasi attributes in a data set. If too many quasi attributes are identified and if we apply privacy techniques like k-anonymity on those big number attributes (which is not exact set), it decreases data utility of the data set. Also, if number is too less, it causes privacy leak. In many research work, the quasi attributes are found from experts based their personal experience. But, finding quasi attributes in this way is not very accurate.

The objective of this work is to find the minimum or optimum number of quasi attributes in the data set in optimal time and complexity. This improves the performance in implementing privacy as it is just optimal number of quasi attributes.

2. Related Work

Owing to advancement in Web 2.0 technologies, the user’s data are openly accessible in social platforms[28]. The accessed data are misused by the third parties for commercial purposes [1]. The k-anonymity l-diversity schemes do optimize the published data, but only partially. Hence, this issue has to be addressed [2–4]. The first research on providing privacy in data mining was limited to implementing privacy on centralized data. Later on, it was implemented for distributed data. Many data distortion techniques like perturbation, adding noise to the data to change original data, generalization, k-anonymity, L-diversity etc. were used [28] [29] [38]. Other researchers used cryptographic techniques to provide privacy like homomorphic encryption, order preserving encryption, Elliptic curve cryptography, Pallilier system [31][32][33].

Researchers model the social networks as graphs in which nodes and edges denote entities and links respectively [5]. They have proposed various anonymous models based on k-anonymity to achieve privacy protection in existing research [6-9]. However, in some cases, k-anonymity cannot protect against “attribute disclosure”.

The relation between safety and information utility is still new in the field of social networks [4]. The current methodologies may anticipate leakage of some security data in distributing networks, yet may result in utility loss without understanding its sensitivity. Exhaustive surveys carried out in this field exposes the issue of excessive anonymity for preserving sensitive attributes [10].

The models of l-diversity and t-closeness have been widely studied for protecting privacy. Most existing studies assume that they can separate Quasi-identifiers (QIDs) from sensitive attributes, but we cannot always make such assumptions in real world situations.

In recent years, many organization opts for preserving privacy data to be published by third party using data anonymization [24][27][30]. The third party collects data from data owners, anonymize it and sends it to cloud for analysis. Given a data set, data can be divided into different sets which can cause leakage of sensitive data [25]. Explicit identifiers (EI) which explicitly identifies record owners are typically removed from the released data. QID can be linked with external information to re-identify individual record owners, and sensitive attributes that needs to be protected.

The existing work on privacy preserving data mining used l-diversity [26] and t-closeness considered only explicit sensitive attributes to apply privacy techniques. But practically, many other attribute sets which are related to sensitive attributes can also leak information of sensitive attributes. These attributes are called quasi attributes.

From above analysis, it appears that we may benefit by completely omitting quasi-identifier attributes from a dataset if the rarity of their values is highly correlated with many numeric attributes. In doing so, the loss of information caused by the exclusion of the attributes must be taken into account. Suppressing quasi-identifier columns whose value scarcity is highly correlated with numerical attributes has the same effect as increasing survey bias. An improvement in sampling methods in order to ensure the balance of background covariates (and thereby quasi-identifier attributes) is a way to reduce the amount of bias introduced to a dataset. A release of data is said to adhere to k-anonymity, if each released record has at least (k-1) other records whose values are indistinct over a special set of fields called the quasi-identifier [16]. The quasi-identifier contains those fields that are likely to appear in other known data sets. Therefore, k-anonymity provides privacy protection by guaranteeing that each record relates to at least k individuals even if the released records are directly linked (or matched) to external information. Generalizations based on attributes with higher generalization hierarchies tend to be more accurate than generalizations based on attributes with shorter hierarchies. In addition, higher hierarchies can provide different precision measurements for the same plate. The construction of hierarchies of generalization is therefore part of the criteria for choice. Algorithms may provide generalizations that are not k-minimal distortions because they both enforce generalization at the attribute level. This leads to inaccurate precision measures. There may exist values in the table that when generalized at the cell level, satisfy k without modifying all values in the attribute. In summary, more work is needed to correct these heuristic-based approaches.

Data Privacy Preservation has shown important data services. K-anonymization is a data privacy solution and has been the focus of research over the past few years. Currently anonymized privacy has received considerable attention to the preservation of data publication. The multi dimension bucketization is discussed in [17] to anonymize multiple sensitive attributes. It proposes bucketizeation method, finds l-diversity for sensitive attribute, and presents the techniques to generalize quasi-identifiers that prevents attacks. By modifying factors that affect the relationship between correlations of quasi-identifier value frequencies with other numeric attributes, the possibility that correlations between the frequencies of quasi-identifier attributes with other numeric attributes contribute to bias and loss of utility is confirmed. Specifically, we investigated the following three factors that contributes to bias.

I. Increasing the value of k in k-anonymity.

In a situation where a quasi-identifier field’s frequency of values is correlated to a numeric attribute, increasing the value of k is imperative to change the quasi-identifier rarity threshold at which a given record must be cut. If a quasi-identifier’s rarest values are more often tied to either high or low numeric attributes, this therefore would suggest that higher levels of k (which correspond to greater anonymity requirements) would create biases in the resulting de-identified dataset’s numeric attributes.

II. Eliminating quasi-identifier fields with high correlations of value frequency to numeric attributes. Given a quasi-identifier field with a high correlation between the frequencies of its values with numeric attributes, there may be value in simply omitting the entire field rather than allowing it to create bias within the dataset. In this case, a balance must be maintained between the values of the information encoded in the quasi-identifier field versus the bias created in the dataset.

III. Increasing the correlation between quasi-identifier value frequencies with given numeric attributes. The manual alteration of quasi-identifier values confirms the amount of bias introduced during the de-identification process may be related to the magnitude of the correlation between quasi-identifier value frequencies and its attribute. Due to the fact that numeric attributes are highly skewed toward values near 0, situations in which rare quasi-identifier values are associated with high values of attributes causes more bias in the data. The research works in [22] and [23] have also taken quasi attributes as part of sensitive data by applying two phase anonymization.

To summarize this related works, we can say that the existing techniques suffer from many drawbacks like higher time consumption and extract more quasi identifiers leading to decrease in data utility. The proposed method is discussed in next section.

3. Proposed Work

We consider a scenario where organizations depend on cloud for additional resource necessities. Since cloud is a third-party, the data shared to it might be leaked, hence this issue needs to be resolved. One of the solution that we propose is the identification of quasi attributes in an optimal way to deal with this problem.

We use the concept of articulation point in graph to find quasi attributes. A graph's articulation point is the attribute in the graph that disconnects a graph if removed from it. In other words, the relationship between two attributes in the graph is eliminated. Articulation point is usually used to define vulnerabilities in the network, i.e. the point in the graph failure from which the network can be disconnected. Figure 1 represents the graph and articulation point in the graph. Its easy to show that if we remove node '1' from the graph, it disconnects the whole graph.

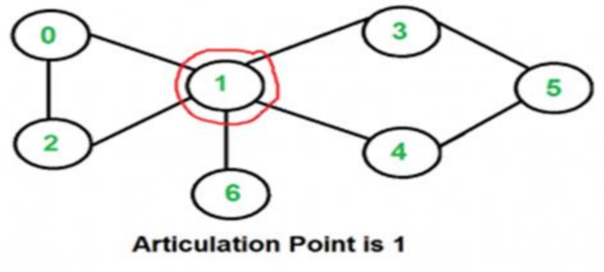

Figure 1: Articulation Point

The relationship between quasi attributes, explicit identifiers and sensitive attributes should be understood to find quasi attributes. Using this relationship or dependency, we can deduce how much sensitive information can be leaked using these quasi attributes. The steps to find quasi attributes in the proposed method are:

Step 1: Draw the attribute graph.
Step 2: Select the explicit identifier attribute and sensitive attribute.
Step 3: Find all the paths between explicit identifier attribute and sensitive attribute.
Step 4: Find the articulation points which are the quasi attributes.

3.1 Construction of Attribute Graph

Graph is a structure made up of objects and some objects have a relationship between them. The graph of the attribute is the structure in which the data set attributes are represented as vertices and the relationship between these attributes are represented as edges. Let us consider a database which has m number of attributes, so the vertices of the attribute graph can be represented as V={ A1,A2,……Am} and the set of edges as E={ AiAj , for i,j=1,2,…m}. This graph can be represented mathematically as G= {V, E}. There are two types of graph, directed and undirected graph. Directed graph can be constructed if there is no dependency on the values of different attributes i.e. they are independent to each other. But generally in database, dependency exists between attributes. So, for this situation, directed graph need to be constructed which represents the dependency between the values of attributes. Figure 2 represents directed graph.

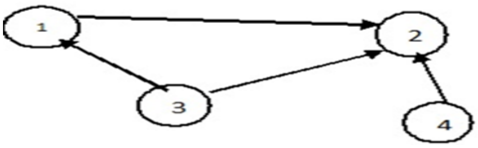

Figure 2: Directed Graph

Algorithm 1 shows the steps to construct attribute graph:

Algorithm 1:
Input: Set of n Attributes
Output: Graph with attributes as vertices and edges as dependency between attributes.
Steps:

```
For i= 0 to n
    For j= 0 to n
        Check dependency between jth and other n-1 attributes
        If there is a dependency, draw the edge between the attributes.
        Else go to next attribute.
```

We have used online retail dataset from UCI repository. It is a transaction dataset which contains transactions in UK retail shops for a period of time. It contains following attributes: Customer-id, Invoice-No, Stock-code, Stock-Description, Quantity, Invoice-date, Unit-Price and Country. The graph is constructed by taking the dependency among the attributes. Figure 3 shows the graph for the dataset considered.

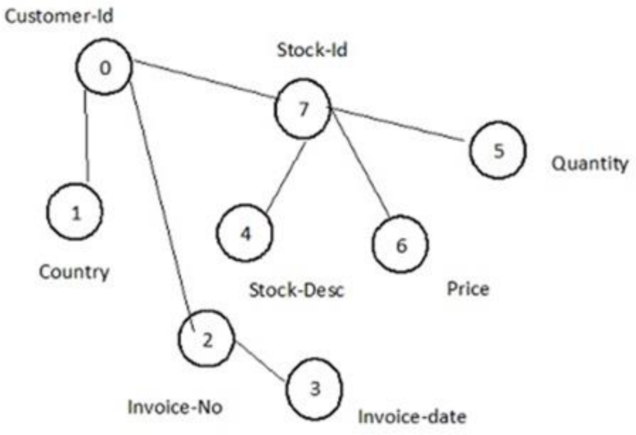

Figure 3: Attribute Graph for Online Retail Data set

This graph represents the dependency among attributes in data set and this information is used to find the articulation point.

3.2 Selection of Explicit Identifier and Sensitive Attribute:

In the online retail dataset, the explicit identifier attribute is Customer-id because it uniquely identifies a customer. The sensitive attributes are Stock-Description and Quantity as they give the information of the purchase history of a customer.

3.3 Find Paths between Explicit Identifier and Sensitive Attribute and articulation point in the paths:

In this step, we find all the paths between explicit attribute and the sensitive attributes through depth-first-search (DFS) traversal technique. While searching for the path, we consider the adjacent node same as adjacent node in any other previous found path. Then, the search stops there. Algorithm 2 gives the steps involved in finding all the paths and the attributes adjacent to explicit attributes.

```
Algorithm 2:
Input: Graph with attributes as vertices
Output: All the paths between EI to SA
Steps:
Source = EI, Destination = SA
i) Start from source vertex EI and visit next vertex and store the vertex in adjacent array and visited array.
ii) Now the second vertex is source vertex and find next vertex and store in visited array and so on.
iii) Now again take EI as source and check for next path adjacent node:
    For i=0 to n where n= size of Adjacent array:
        If(Adjacent attribute = Adjacent[i]):
            Stop search of that path and go to next path.
iv) Find the articulation point:
    For all the paths between EI to SA:
        For i=0 to m:  where m is the number of attributes in the path.
            a) Remove each vertex one by one and check whether it disconnects the path between EI and SA.
            b) If yes, save the vertex else add it back to graph.
```

Therefore taking the online retail dataset example, the paths calculated is [Customer-id, Stock-id, Stock-Desc] and the articulation point is Stock-id which disconnects the path between Customer-id and Stock-Desc. Similarly, for sensitive attribute Quantity, all the paths between EI and SA is [Customer-id, Stock-id, Quantity] so here also, articulation point attribute is Stock-id. Hence, the quasi attribute in this data set is Stock- id.

4. Experimental Analysis

The proposed method is analyzed using different datasets and compared with the existing techniques. We have considered two parameters to compare our method with existing works. The result of comparison is explained in this section. The parameters are: number of quasi attributes determined and running time.

Consider the example of data set Used Car which has seven attributes, namely: company, years, cost price, selling price, condition, distance completed and owner. It has 10000 rows and 7 columns. Figure 4 shows the attribute graph for the stated dataset.

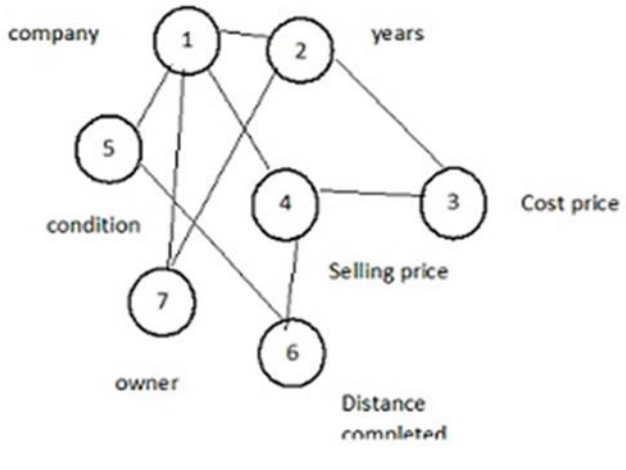

Figure 4: Attribute Graph

1. 7 1 4
2. 7 1 ---- stop here
3. 7 1 --- stop here
4. 7 1 ---- stop here
5. 7 1 ---- stop here
6. 7 1 ----- stop here
7. 7 1 ------ stop here
8. 7 1----- stop here
9. 7 1 ---- stop here
10. 7 2 4
11. 7 2 ---- stop here
12. 7 2----- stop here
13. 7 2 ---- stop here
14. 7 2 ---- stop here
15. 7 2 ---- stop here
16. 7 2 ---- stop here
17. 7 2 ---- stop here
18. 7 2 ---- stop here
19. 7 2 ---- stop here
20. 7 2 ---- stop here
21. 7 2 ---- stop here
22. 7 2 ---- stop here

The articulation point is calculated as [1, 2] i.e. company and years which are the quasi attributes. In proposed work, when graph is created, if there are n number of attributes then there can be maximum of n-2 paths between EI and SA. In other words, there can be maximum n-2 paths in worst case. In best case, there can be just one articulation point i.e one quasi attribute.

4.1 Comparative Analysis

In [18], research discusses the concept of distinct ratio and separation ratio. It chooses $\kappa = \log_{\frac{1}{1-c}} \frac{2^m}{\delta}$ pair of tuples and then applies the greedy algorithm. It finds the attributes as quasi attributes, if its distinct ratio is less than αβ where $\alpha = 1 - \frac{\sqrt{2\ln(2m/\delta)}}{\beta\kappa}$. The work shows that minimum number of quasi attributes generated is ≈ ln(log 2m) and the maximum number of quasi attributes is ≈ m.

In [19], the authors have used the data set of medical records. They analyzed the factors which are affecting the re-identification of medical records on the basis of outside environment factors. The minimum number of quasi attributes is 1 and maximum is 5 or greater. So it is not flexible to all types of datasets. In [20], hyper graph is constructed from the attributes. In this work, first different views are determined for attributes and the hyper-graph is created for each view. Then the path between two attributes are determined. Since the number of views are more, there are too many number of hyper-graph. It extracts many attributes as quasi attributes which is more than required. It creates m different hyper-graphs and hence in worst case, it can have m number of quasi attributes. Table 1 shows the comparison of minimum and maximum number of QIs possible with existing and proposed work.

Table 1: Comparison of Number of Quasi Attributes

Technique	Minimum Number of QIs	Maximum Number of QIs
Rajeev *et.al.* [18]	≈ ln(log 2m	≈ m
Yong Ju *et.al.* [19]	1	5
Huang *et.al.* [20]	1	*m*
Proposed Work	1	*m-2*

4.2 Running Time and Complexity

The proposed work is also compared for running time and complexity in implementation with existing works. The work in [18] is an improvement over traditional greedy algorithms. In traditional greedy algorithm, the running time is proportional to n where n is the number of tuples in data set, but in [18] the running time is O (m4) where m is the number of attributes. It calculates distinct ratio and separation ratio. The algorithm to calculate quasi attributes, which is a complex calculation, uses the values of the distinct ration and separation ratio. The implementation of the algorithm is therefore complicated. In [19], whole data set is scanned many times to find the set of quasi attributes. So, if n is the number of tuples in data set, then the running time is proportional to n and considering most datasets, n>>> m, so the running time is greater than the proposed work. The work discussed in [19] is less complex than the work mentioned in [18]. But, it is more complex as it scans the data set for particular value of particular attribute, unlike the proposed method where data need not be scanned.

In [20], first hyper-graph is constructed from attributes and then it is converted into common graph. So if there are m nodes, considering all the views, the time complexity is equivalent to O (m5). It uses hyper-graph in which it considers many views. Then, it constructs common graph so it is less complex than [18] but more complex than the proposed work.

The work in [35] uses method of finding different sets and subsets of attributes to find quasi attributes and check with the number of maximum tuples. So, it needs scan of tuples and needs time proportional to m3 if m is the number of attributes. The complexity of implementation is also medium.

The patent work in [34] used the method of creating the sets from the given attributes. It first makes set of two attributes including all attributes and scans the values of all rows of these attributes to find the quasi relation between them. Then it makes set of three attributes, four attributes and so on. So, if n is the number of tuples in data set and m is the number of attributes then its time complexity is O (m*n). Also, this method is difficult to implement.

A noteworthy aspect of our work is that we must find all the paths between EI to SA but stop the search if we get same adjacent node. This leads to the time complexity is O (m2 /2). Then, to find articulation point, it takes O (m2). So, total time is O (m2 + m2 /2). An accuracy of the proposed method is better as it considers all the possible sets to find quasi attributes. Table 2 shows comparison w.r.t running time and complexity in the existing and proposed work.

Table 2: Comparison of Running Time and Complexity

Technique	Running Time	Complexity
Rajeev *et.al.*	$O(m^4)$	High
Yong Ju *et.al.* [19]	*n*	Medium
Huang *et.al.* [20]	$O(m^5)$	Medium
Amani [35]	$O(m^3)$	Medium
Braghin [34]	$O(m*n)$	High
Proposed Work	$O(m^2 + m^2 /2)$	Low

5. Conclusions

We discussed the model for organizations that depend on cloud for additional resource necessities. Since cloud is a third-party, the data shared to it might be leaked. We discussed an optimal solution by identification of quasi attributes through the concept of articulation point to overcome this problem.

The proposed work overcomes this drawback by extracting minimum required quasi attributes with minimum time complexity. On comparison with existing works, the parameters such as an accuracy, complexity and range of QIs values are optimized by the proposed method as it considers all the possible sets to find quasi attributes. In future, we would like to apply similar model for Big data sets for privacy preserving.

References

[1] Li, Y.; Li, Y.; Yan, Q.; Deng, R.H., "Privacy leakage analysis in online social networks", Computer Security, Vol 49, pp. 239–254, 2015.

[2] O Gutcu; Testik, Ö.M.; Chouseinoglou, O., "Analysis of personal information security behavior and awareness", Computer Security, Vol 56, pp. 83–93, 2016.

[3] Dunning, L.A.; Kresman, "R. Privacy preserving data sharing with anonymous id assignment", IEEE Trans. Inf. Forensics Security, Vol 8, pp. 402–413, 2013.

[4] WANG, Yazhe; XIE, Long; ZHENG, Baihua; and LEE, Ken C. K., "Utility-Oriented K-Anonymization on Social Networks", Database Systems for Advanced Applications: 16th International Conference, DASFAA 2011, Hong Kong, China, April 22-25, 2011, Proceedings, Part I. 6587, 78-92.

[5] Liu, X.Y.; Wang, B.; Yang, X.C., "Survey on Privacy Preserving Techniques for Publishing Social Network Data", Journal of Software, Vol. 25, Issue 3, pp. 576–590, 2014.

[6] Sweeney, L., "K-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 05, pp. 557-570, 2002.

[7] Thompson, B.; Yao, D.F., "The union-split algorithm and cluster-based anonymization of social networks", In Proceedings of the 4th International Symposium on Information Computer and Communications Security, Sydney, Australia, 10–12 March 2009.

[8] Li, F.; Shin, R.; Paxson, V., "Exploring privacy preservation in outsourced k-nearest neighbors with multiple data owners", In Proceedings of the 2015 ACM Workshop on Cloud Computing Security Workshop, Denver, CO, USA, pp. 53-64, 12–16 October 2015.

[9] Yuan, M.X.; Chen, L.; Philip, S.Y.; Yu, T., "Protecting sensitive labels in social network data anonymization", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No 3, pp. 633–637, 2011.

[10] Fu, Y.Y.; Zhang, M.; Feng, D.G., "Attribute privacy preservation in social networks based on node anatomy", Journal of Software, Vol. 25, pp. 768–780, 2014.

[11] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," IEEE Transactions on Knowledge Data Engineering, Vol. 27, No. 11, pp. 3098–3110, 2015.

[12] P. Shi, L. Xiong, and B. Fung, "Anonymizing data with quasi-sensitive attribute values Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, pp. 1389–1392, October 26-30, 2010.

[13] M. Terrovitis, N. Mamoulis, J. Liagouris, and S. Skiadopoulos, "Privacy preservation by disassociation," Proceedings of VLDB Endowment, Vol. 5, No. 10, pp. 944–955, 2012.

[14] J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic K-anonymity through microaggregation and data swapping, "IEEE International Conference on Fuzzy Systems, pp. 1–8, 2012.

[15] K. Wang, Y. Xu, A. W. C. Fu, and R. C. W. Wong, "FF-anonymity: When quasi-identifiers are missing," IEEE 25th International Conference on Data Engineering, pp. 1136–1139, 2009.

[16]] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 571-588, 2002.

[17] Dharavathu Radha, and Prof. Valli Kumari Vatsavayi, "Bucketize: Protecting Privacy on Multiple Numerical Sensitive Attribute", Advances in Computational Sciences and Technology, Vol. 10, No. 5, pp. 991-1008, 2017.

[18] Rajeev Motwani, Ying Xu. "Efficient Algorithms for Masking and Finding Quasi-Identifiers", Proceedings of VLDB, Vienna, Austria, pp. 758-769, 2007.

[19] Yong Ju LEE, Kyung Ho LEE, "Re-identification of medical records by optimum quasi-identifiers", International Conference on Advanced Communication Technology (ICACT), Bongpyeong, South Korea, pp. 428-335, 2017.

[20] Huang L M, Song J L, Lu Q C, et al. "Hypergraph-based solution for selecting quasi-identifier", International Journal of Digital Content Technology and its Applications, Vol.6, No. 20, pp. 597-606, 2012.

[21] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, No. 3, pp. 1-12, 2007.

[22] Wong, R. C. W., Fu, A. W. C., Wang, K., and Pei, J., "Minimality attack in privacy preserving data publishing", In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), pp. 543-554, 2007.

[23] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu, "Anonymizing transaction databases for publication", KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 767-775, 2008.

[24] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys, Vol. 42, No. 4, pp. 14.1-14.53, 2010.

[25] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity", 22nd IEEE International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, USA, 2006.

[26] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity", 2007 IEEE 23rd International Conference on Data Engineering, pp. 106-115, 2007.

[27] Nergiz ME, Gök MZ, " Hybrid k-anonymity", Computers and Security, Vol. 44, pp. 51-63, 2014,

[28] Jhon Francined Herrera-Cubides, Paulo Alonso Gaona-García, Carlos Enrique Montenegro-Marín, Diego Mauricio Cataño, Rubén González-Crespo' "Security Aspects in Web of Data Based on Trust Principles. A brief of Literature Review", International Journal of Communication Networks and Information Security, Vol. 11, No. 3, pp. 365-379, 2019.

[29] G, Yang Y, Chen J, " A historical probability-based noise generation strategy for privacy protection in cloud Computing", Journal of Computer and System Sciences, Vol 78, No. 5 pp: 1374–138, 2012.

[30] Mahesh, R., & Meyyappan T, "Anonymization technique through record elimination to preserve privacy of published data", International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), Salem , India, pp. 328-332, 2013.

[31] Usha, P., Shriram, R., & Sathishkumar, S, " Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining", International Conference on Information Communication and Embedded Systems, Chennai, India, 2014.

[32] Nirali Nanavati, Devesh Jinwala,, "Privacy Preservation for Global Cyclic Associations in Distributed Databases", Procedia Technology, Vol. 6 ,pp. 962-969, 2012.

[33] Ibrahim A, Jin H, Yassin AA, Zou D, "Towards privacy preserving mining over distributed cloud databases", IEEE second international conference on cloud and green computing (CGC), pp. 130-136, Xiangtan, China, 2012.

[34] Sankita J. Patel, Dharmen Punjani and Devesh C. Jinwala," An Efficient Approach for Privacy Preserving Distributed Clustering in Semi-honest Model Using Elliptic Curve Cryptography", International Journal of Network Security, Vol 17, No. 3, pp 328-339, 2015.

[35] Stefano Braghin, Gkoulalas-Divanis, Michael Wurst," Detecting Quasi-Identifiers in Data Sets", United States Patent Application Publication, US 2016/0342636A1, Nov. 24, 2016.

[36] Amani Mahagoub Omer, 2 Mohd Murtadha Bin Mohamad, "Simple And Effective Method For Selecting Quasi-Identifier", Journal of Theoretical and Applied Information Technology, Vol. 89, No. 2, pp: 512-51, 2016.

[37] Sweeney L. "K-Anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol.10, No. .5, pp 557-570, 2002.

[38] Sweeney L. "Achieving 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.