[1]Rinkalben J. Prajapati
[2]Dr. Jaykumar Shantilal Patel

A Systematic Review: Cluster based k-Anonymization Approaches for Big Data Privacy

Journal of Electrical Systems

Abstract: - The growing adoption of Internet-based applications in diverse areas, such as marketing, analysis and research continuously expanded the volume of data, resulting in the accumulation of big data. Available Online information can endanger individuals by exposing their sensitive data to privacy risk. Numerous privacy preservation techniques exist in literature including cryptography-based, differential privacy and k-anonymization based methods, are accessible to safeguard individual privacy. K-anonymity is an approach used to achieve data privacy by applying generalization and suppression to group of data by making them anonymous. This can be achieved using traditional way or via clustering algorithms. This paper presents an in-depth review of state-of-the-art cluster-based approaches for achieving k-anonymity of structural and stream based data. Moreover, detailed analysis of the effect of clustering approaches on measuring parameters such as data utility and information loss. It also discusses the challenges faced while using clustering techniques to grouping data for data k-anonymization

Keywords: Bigdata, privacy, Quasi-identifier, Sensitive-Attribute, k-anonymization, clustering, Information loss

I. INTRODUCTION

We cannot imagine today's world without big data. The widespread use of internet and different internet based applications adds data to existing data in form of big data and makes it more complex in form of volume, Variety and velocity. From the data mining perspective, mining big data has opened many new opportunities for researchers but existing data mining techniques are not scalable for big data due to 3 V's. Aim of Big data is not only to store data but to use these data for different purpose like analytics, marketing and to make important decisions by extraction knowledge in the field of public health, commercial trade, business and social networks. Big data processing deals with large and complex data sets which cannot be processed by traditional operations efficiently. In the last few years, several big data frameworks like Distribute file system architecture Hadoop using MapReduce Programming are developed as a solution to cope with the problem [1]. For analytics purpose mainly data are collected from people in direct or indirect way and published it to extract valuable knowledge from data Published data does not contains direct identifying information but there is a chance of reviling the identity or sensitive information of individual if they are linked with other publicly available information which puts anonymity at risk [2][3]. In this situation either people are aware of data collected about them but not aware about where these data is going to be used or sometimes they are not aware of the data collected about themselves [3]. In any condition their anonymity is always at risk and it should be preserved. There many techniques available in past to preserve the data privacy.

Big data analytics involve the analysis of data from various sources or surveys for decision-making purposes. Privacy can be ensured by data providers, data collectors, data miners, and decision makers. Data collectors collect data from data providers to support the mining process, but the information collected may be sensitive. If the data collector doesn't take sufficient step for data privacy before releasing it for mining purpose then sensitive information may be disclosed. To prevent sensitive information from being disclosed, data collectors must modify the original data before releasing it for mining purposes. Sensitive information of data providers can neither directly be found in the modified data nor be inferred by anyone with malicious intent. The process of modifying data to provide privacy and utility simultaneously is known as privacy-preserving data publishing (PPDP). The concept of privacy disclosure is demonstrated through the example of two data tables. Table 1 contains medical data with identifiers removed, while Table 2 contains publicly available voter information. Linking these tables can reveal an individual's identity. For instance, linking the table 1 and table 2 reveals that "Sue J. Carlson, 1458 Main Street, Cambridge," suffers from "shortness of breath." With the amount of publicly available information increasing, it is

[1] * Research Scholar, Computer/IT Engineering, Gujarat Technological University, Ahmedabad-382424, Gujarat, India.
Email: rinkal10feb@gmail.com
[2] *Professor, Chaudhari Technical Institute, Gandhinagar-382007, Gujarat, India.
Email: jay_sp_mca@yahoo.co.in
*Corresponding author

important to reform data tables so that linking attacks do not reveal sensitive information and the data remains useful for analytics. Several approaches have been proposed in the literature to protect the sensitive information of individual persons [6]. These methods include the following:

Table 1 Medical data [11]

ID	Name	Ethnicity	Birth Date	Sex	Zip	Marital Status	Disease
		Asian	09/27/64	female	02139	divorced	hypertension
		Asian	09/30/64	female	02139	divorced	Obesity
		Asian	04/18/64	male	02139	married	Chest pain
		Asian	04/15/64	male	02139	married	Obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	Shortness of breath
		black	09/13/64	female	02141	married	Shortness of breath
		Black	09/07/64	female	02141	married	Obesity
		White	05/14/61	male	02138	Single	Chest pain
		White	05/08/61	male	02138	single	Obesity
		White	09/15/61	female	02142	Widow	Shortness of breath

Table 2 publically available Voter List [11]

Name	Address	City	Zip	DoB	Sex	party
…	…	…	…	…	…	
…	…	…	…	…	…	
Sue J. Carlson	1458 main St.	Cambridge	02142	09/15/61	F	democrat
…	…	…	…	…	…	

A. Cryptography based Techniques

This technique provides privacy by encrypting the data with defined access structure or policy. The data users can access the data only if their attributes satisfy the access structure in the cipher-text.[7].There are two types of cryptography based techniques ,The symmetry key and asymmetry key Cryptography. In Symmetry key cryptography the same key is used for encryption and decryption, while in asymmetry key cryptography different keys are used for encryption and decryption process.

B. Anonymization Based techniques

To preserve the anonymity of individual data must be k-anonymized. To achieve the anonymization, need to apply generalization and suppression on data such that newly generated equivalence class has at least k similar records [3].

C. Differential Privacy

Differential Privacy is a technique that offers investigators and server developers, an opportunity to receive valuable information through repositories comprising personally identifiable information without exposing the individual's identity [6]. It can be achieved by adding randomized “noise” to an aggregate query result to protect individual entries without significantly changing the result.

The objective of the study is to understand the significance working of k-anonymization and the importance of clustering techniques in attaining data privacy through k-anonymization. To protect individual’s privacy, it is necessary to create equivalence classes with comparable Quasi-identifier values in order to anonymize the records. Clustering techniques efficiently establish equivalence classes, but in k-anonymization, their effectiveness lies in generating equivalence classes with minimal modification of Quasi-identifier values. Clustering algorithm come up with different characteristics based on its type. They possess the ability to form equivalence groups among various data entries, determined by chosen criteria of similarity or dissimilarity. Consequently, clustering algorithms prove invaluable in anonymization efforts, yet they can compromise data utility, leading to information loss. In safeguarding individual privacy through k-anonymization aided by clustering, it becomes imperative to construct equivalence classes that minimize information loss while ensuring substantial data privacy. This paper will explore various privacy preservation approaches based on clustering techniques for static data and stream data.

Section II of the paper will provide an overview of the k-anonymization approach along with other extensions derived from k-anonymity. Section III will delve into various clustering techniques for grouping data into equivalence classes, including different categories of basic clustering methods. In Section IV, we will review research on privacy preservation approaches using cluster-based anonymization conducted in previous years and a comparative study summarizing their findings in tabular form. Section V will discuss the diverse performance parameters employed to evaluate various privacy-preserving techniques. Section VI will present summary of the cluster-based k-anonymity approach, leading to the Conclusion.

II. K- ANONYMITY FOR PRIVACY PRESERVATION

Data k-anonymization is achieved if it satisfies the k-anonymity property. K-anonymity [3] concept inspired from real world systems Data-fly [8], μ-Argus [9] and k- Similar [10] motivate this approach .some terms related to k-anonymization are:

A. *Quasi-Identifier(QI)[11]:*For a Table T(A_1,A_2,…A_n) with n attributes having Quasi-identifier is set of attributes (A_i,...,A_j) from (A_1,…,A_n) whose release must be controlled to protect privacy of individual. In table 1 attributes Ethnicity, Birthdate, Sex, Zip and Marital Status can be categorized as Quasi-identifier.

B. *Sensitive Attribute(SA) [11]:* For Table T(A_1,A_2,…A_n), Sensitive attribute is one or more attribute that describes sensitive information of individual and need to be protected at the highest level. Here Attribute Disease is categorized as sensitive attribute.

C. *Non-Sensitive Attribute [11]:* For a Table T(A_1,A_2,…A_n) with n attributes, Non-Sensitive Attribute that does not affect the privacy of Individual.

D. *Generalization [3]:* It substitutes the values of a given attribute with less specific or more general value that is faithful to the original. Generalization of value '02138' for QI attribute Zip and 'single' for QI attribute Marital status of table 1 can be modified with value '0213*' and 'Never_married' , respectively as shown in Fig.1 . Generalization is applied on QI attribute value based on defined generalization hierarchy provided for each quasi-identifier. Generalization of Quasi-identifier causes the Information loss [12].

Generalization [13] can be categorized as global generalization, also known as full-domain generalization, and local generalization. Global generalization applies the same generalization step to every attribute with its corresponding value. In contrast, local generalization allows specific attributes to be generalized while others with the same value remain unchanged. This approach facilitates more precise data processing to achieve k-anonymity with minimal data utility loss. Data-fly [8] and μ-Argus [9] employ global generalization, whereas the Mondrian Multidimensional algorithm [14] and KACA algorithm [15] have achieved k-anonymization using local generalization.

D. *Suppression [3]:* It does not release the value of tuple referring a QI or replace with '*'. In some of the researches, it removes data from the table. To moderate the generalization suppression is performed mostly on tuple [3][11].

E. *K-anonymity [3][11]:* Let T(A_1,…, A_n) be a table, and QI be a set of quasi-identifier associated with it. T is said to satisfy k-aonymity with respect to QI iff each sequence of values in T [QI] appears at least with k occurrences in T[QI][3]. It can be achieved by Generalization [3][16][17] and suppression [11].

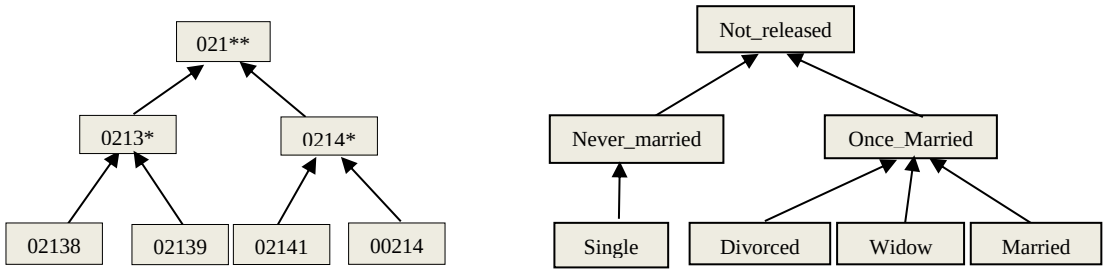

Fig.1 Generalization hierarchy: (a) numeric Quasi-identifier attribute zip (b) categorical Quasi-identifier attribute marital status

F. *Equivalence class [11]:* In k-anonymity, each equivalent class contains at least k records. Quasi-identifier groups of all records should have the same value in one equivalent class. Table 3 shows the data after applying generalization on Ethnicity, birth date, Sex, zip and Marital status to achieve 3-anonymous table. Using Table 3 one cannot directly identified that Sue J. Carlson, is suffering from which problem.

Table 3 3-anonumous Table for Medical data [11]

ID	Name	Ethnicity	Birth Date	Sex	Zip	Marital Status	Disease
		asian	**/**/64	Person	02139	Once_married	hypertension
		asian	**/**/64	Person	02139	Once_married	Obesity
		asian	**/**/64	Person	02139	Once_married	Chest pain
		asian	**/**/64	Person	02139	Once_married	Obesity
		black	**/**/**	Person	021**	married	hypertension
		black	**/**/**	Person	021**	married	Shortness of breath
		black	**/**/**	Person	021**	married	Shortness of breath
		Black	**/**/**	Person	021**	married	Obesity
		White	**/**/61	Person	021**	Not_released	Chest pain
		White	**/**/61	Person	021**	Not_released	Obesity
		White	**/**/61	Person	021**	Not_released	Shortness of breath

K-anonymity is suffering from Homogeneity Attack [16] and Background Knowledge Attack [16]. To address this problem, ℓ-Diversity [16] approach is introduced which uses the concept of Bayes-optimal privacy and also adapts basic k-anonymity algorithm. It provides privacy even when data publisher does not know the kind of knowledge is possessed by the adversary. The main idea behind ℓ-Diversity is the requirement that the values of the sensitive attributes are well-represented in each group [16].

G. *Distinct ℓ-Diversity [16][18]:* The values of sensitive attributes in each equivalence class in the anonymized data set have at least ℓ different values. It does not possess any limit to the probability proportion of a single sensitive attribute.

H. Entropy L-Diversity[16][18]: A table is said to have entropy L-diversity if for every equivalence class e,

$$\text{Entropy}(e) \geq \log L \qquad (1)$$

Where, entropy L-diversity of a equivalence class (e) for domain SA of sensitive attribute can be defined as:

$$\text{Entropy}(e) = -\sum_{s \in SA} p(e,s) \log p(e,s) \qquad (2)$$

and p(e, s) is the fraction of records in e that have sensitive value s.

I. Recursive(C,L)-Diversity[16][18]: In a given equivalence class e, if r_i denote the number of times the i^{th} most frequent sensitive value appears in that it .Given a constant C, the equivalence class e satisfies recursive (c, ℓ) diversity if

$$r1 < c\,(r_\ell + r_{\ell+1} + \ldots + r_m) \qquad (3)$$

A table said to be satisfy recursive(C, L)-diversity if every equivalence classes in a table satisfies recursive(C, L)-diversity. If there are multiple sensitive attribute then it is hard for equivalence class to satisfy ℓ-Diversity, however it is possible by satisfying Multi-Attribute ℓ -Diversity via suppression and generalization [16]. ℓ-Diversity is insufficient to prevent attribute disclosure. It is unable to protect against Skewness attack [19] and similarity attack [19] on l-diverse table. t-closeness [19] formalized to overcome the limitations of ℓ-Diversity. It aims to distribute sensitive attribute in an equivalence class such that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. In simple way, that the distance between the two distributions should be no more than a threshold t. Idea behind approach in [19] is that they separate the information gain an observer can get from a released data table into two parts: that about all population in the released data and that about specific individuals. They use the Earth Mover Distance metric [19] to measure the distance between the two distributions. k-anonymity, ℓ-Diversity[16] and .t-closeness[19] uses the tradition approach to creates equivalence class of size k requires more time and suffers from high information loss due to dependency on pre-defined generalization hierarchies[3][14] of QI or total order [14][17] defined for each attribute domain.

III. CLUSTERING ALGORITHMS AND IT'S ROLE IN K-ANONYMIZATION

Clustering is an unsupervised learning technique which groups the unlabeled data based on their similarity or dissimilarity to each other. Clustering technique can be categorized as hard/Exclusive clustering and soft/overlapping clustering based on characteristics to assign single data point to only one cluster or more than two clusters, respectively. There have been 100 clustering algorithms suggested in previous research [20] by S.

Kaushik, and these can be classified into five distinct categories [21], such as Partitional Clustering, Hierarchical Clustering, Density-Based Clustering, Grid-Based Clustering, and Model-Based Clustering. Among all categories, Partitional, Hierarchical, and Density-Based Clustering exhibit the capability to manage large datasets with numerical or categorical attributes, depending on their operational principles [22]. The selection of a clustering algorithm should be based on chosen criteria linked to the 3V's of big data, dataset types, dimensionality, stability, and time complexity [22]. In this paper it was examined that for privacy preservation, k-means, distributed, combined clustering, and hierarchical clustering could prove advantageous. To overcome the problem of high information loss through generalization of QI in traditional approach, clustering algorithm assign similar records to the same cluster as much as possible. Similar records in a cluster assure less distortion while applying generalization to the QI. To satisfy k-anonymity property each equivalence class must contains at least k records having similar values of Quasi-identifier. Byun, J.-W., Kamra, A., Bertino, E., and Li, N. [23] considered limitations of k-anonymization and observed it as a clustering problem and K-means stood as the first clustering algorithm employed for generating equivalence classes for anonymization. Clustering algorithms have capability to put similar records in same Equivalence class based on similarity/distance between the records or selected centroid record. To adopt the clustering technique for k-anonymization, Distance function need to be specify to measure the similarity/dissimilarity between the QI of records and cost function to measure the cost of clustering the records The distance function is defined based on the type of quasi-identifier of records. As k-anonymity provides privacy to person specific data records containing combination of numeric and categorical attributes, therefore distance between two records can be calculated as total of distance between each QI attribute value. For each QI in record, there is need to define a separate distance function depending on type of QI. In k-member clustering algorithm [23] to create cluster with similar records, the distance between tuples need to be calculated for numerical attribute and categorical attribute.

Distance between numeric type values [23]: for finite numeric value domain D, the normalized distance between two values $v_i, v_j \in D$ is calculated as:

$$\delta\text{numeric}(vi, vj) = {v_i - v_j}/{|MAX_N - MIN_N|} \quad (4)$$

Where v_{max} and v_{min} defines maximum and minimum value for given domain D, respectively.

Straight formula cannot be applicable to measure the distance between two categorical attributes values because there exist some semantic relationships among the values. This type of relationship can be represented in form of taxonomy tree. The taxonomy tree of any domain is a balanced tree having leaf nodes representing all the distinct values present in the domain. For example consider Fig.2 representing taxonomy tree for attribute country. Taxonomy tree for categorical attribute is same as generalization Hierarchy defined for categorical attribute to achieve k-anonymity.

Distance between two categorical values [23]: Let consider domain D for categorical attribute and T_D be a taxonomy tree defined for D. The normalized distance between two values $v_i, v_j \in D$ is defined as:

$$\delta\text{categorical}\ (vi, vj) = {H\left(\Lambda(v_i, v_j)\right)}/{H(T_D)} \quad (5)$$

Where $\Lambda(v_i, v_j)$ represents sub-tree rooted at the lowest common ancestor of v_i and v_j, and $H(T_D)$ represents the height of tree T.

*Distance between two records[23]:*Consider Quasi-identifier set $QT = \{N_1, \ldots, N_m, C_1, \ldots, C_n\}$ of table T-having numeric attributes with numeric domain $N_i (i = 1,. .,m)$ and categorical attribute with categorical domain C_j (j = 1, . ., n).The distance of two records r_1, r_2 of T is defined as:

$$\Delta(r_1, r_2) = \sum_{i=1,\ldots,m} \delta_N(r_1[N_i], r_2[N_i]) + \sum_{j=1,..,m} \delta_C\left(r_1[C_j], r_2[C_j]\right) \quad (6)$$

where $r_i[A]$ represents the value of attribute A in r_i, and δ_N and δ_C are the distance functions for numeric and categorical attribute, respectively.

The cost function for clustering is defined as a cost to achieve k-anonymization of data with help of clustering. In this process, all the quasi-identifier (QI) values of records in a cluster are generalized to the same value. This involves generalizing numeric values to a range and replacing categorical values with a common value according to a generalization hierarchy. While achieving privacy through generalization It distorts the actual values of the quasi-identifiers in each record within the cluster. The utility of the data may also be affected which leads to fewer values can be extracted from the data. To quantify this distortion, a cost is defined as an information loss metric.

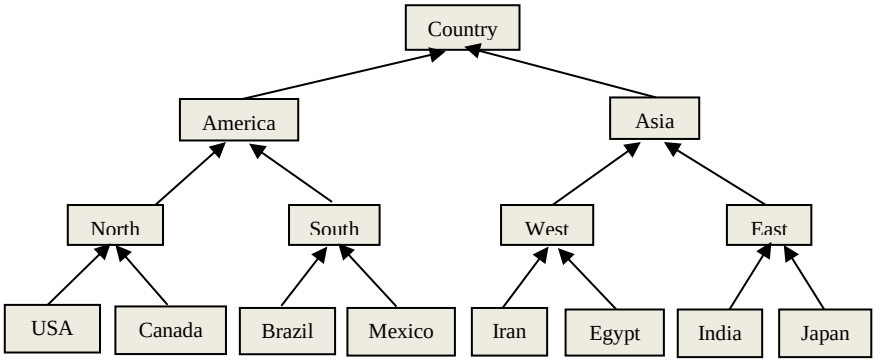

Fig.2 Taxonomy Tree for Attribute Country [23]

Information loss measures the amount of distortion introduced by the generalization process to a cluster or as a reduction in data utility. High loss of information considers as less data utility of anonymized information. In context of k-anonymization cost function should minimize the information loss. Information loss need to be calculated for each cluster to identify overall Information loss. Clustering algorithms exhibit promising performance in big data platforms like Map-Reduce, Spark, and Hadoop also they can be applied to imbalanced data [24].

IV. CLUSTER BASED K-ANONYMITY METHODS FOR PRIVACY PRESERVATION

In this section, our focus is on the methods used for ensuring privacy through the generation of equivalence classes for k-anonymity via clustering techniques. There are numerous cluster-based approaches present in the literature, and we will examine some of these approaches herein. The fundamental requirement for k-anonymity [3] is to partition the data table into equivalent groups such that the quasi-identifiers' value of tuples in each group is equal. In the field of data mining, clustering is a technique that partitions the data table based on the similarity of the attributes, so the tuples in the same cluster have similar values. Meanwhile, records in different clusters are largely dissimilar from one another. Unlike k-anonymity, where all but one of the tuple values in a group must be suppressed or generalized, cluster-based anonymization techniques enable us to select a cluster center whose value for the attribute is the same as the common value. This approach allows for the disclosure of more information without compromising privacy.

Byun, J.-W., Kamra, A., Bertino, E., and Li, N. have proposed the greedy k-member clustering algorithm to satisfy k-anonymization to maximize data quality in paper [23] .It initially select the random record as the cluster centroid and add the records to the cluster such records in a cluster to be as similar to each other as possible so information loss is minimum. When no. of records in cluster reached to k then algorithm selects new record far from existing cluster centroid. The algorithm repeats until all records are assigned to clusters. If records less than k still not assigned to any cluster the each record is individually assign to closest cluster. The experimental results analyses less information loss compared to Mondrian [14].This algorithm suffers with high execution time O(n2) for dataset with n records and it increases the information loss if outlier is selected as cluster centroid.

Loukides, G., and Shao, J. proposed greedy k-anonymization [25] , creates one cluster at a time following the concept of k-member clustering however it selects a random record as a seed to create first cluster. It repeatedly select and add record to cluster until cluster , when building a cluster, this algorithm continues selecting and adding records to the clusters until the diversity of the cluster reaches beyond the user-defined threshold. Algorithm removes entire cluster if the number of records in this cluster is fewer than k. Unlike [23] Outlier does not affect the performance of algorithm still it suffers from information loss due to deletion of clusters having number of records less than k and proper value of threshold is hard to identify. The time complexity of algorithm is O (n2 log (n)/c), where c is the average number of records in each cluster. This algorithm uses any local recoding generalization method to satisfy k-anonymity [23]. Suppression is used for rejected cluster tuples [26]. This algorithm is compared with Mondrian [14] and K-Members [27]. Experimental results on both synthetic and real-world data show that proposed algorithm is able to produce high quality anonymization and also balancing usefulness and protection requirements.

In paper[28] authors Chiu, C.-C., and Tsai, C.-Y have used the weighted feature C-means clustering algorithm for k-anonymization of dataset containing only numeric Quasi identifier. It selects first ⌊n/k⌋ records randomly as seed of clusters and start generating all clusters simultaneously. Algorithm repeatedly adds record based on the closeness with the cluster centroid and updates weights to minimize information loss. To enhance the clustering quality, Algorithm adjusts the weight of each quasi-identifier feature based on the importance of the feature to

clustering quality. Its working is similar to C-means clustering algorithm. Process of weight updating repeats until the record assignment to clusters stops changing. As the end of process algorithm merges cluster with less than k records to large satisfying k-anonymity constraint. It has runtime complexity of $((cn)^2/k)$, having c no. Of iterations for creating clusters of n records. The experiment result is compared with the results of three hierarchical clustering methods. They are single-link, complete-link, and average-link clustering methods [29].The results shows that proposed method has less information distortion and Computation complexity is superior compared to the hierarchical clustering method for the k-anonymity model.

Lin, J.L., Wei and M.C have proposed One pass k-means for k-anonymization[30] having time complexity of $O(n^2/k)$.It works in two-stages, in first stage It uses the K-means algorithm to implement One pass K-means Algorithm (OKA).In first stage it clusters data only once. During the second stage, It moves records from cluster(shrinking cluster) having more than k records to cluster(growing cluster) having less than k records satisfying constraints that no cluster has records less than k. OKA has better results compared to k-means algorithm for both total information loss and execution time. Experimental results showed that shrinking clusters have smaller information loss than growing clusters. First stage of algorithm works perfectly in terms of information loss but there is a scope of improvement in second stage. To solve this problem authors proposed hybrid methods in [31] by combining OKA and the k-member algorithm. Initially OKA is used to generate some clusters with low information loss then it uses k-member clustering to generate the remaining clusters reduce the information loss. Experimental results are compared with the k-member algorithm and OKA in terms of total information loss; had a smaller variance of .It has less information loss than OKA and less time than the k-member algorithm.

In 2009, Lin, J. L., Wei and M. C have proposed a genetic algorithm-based clustering approach [32] for k-anonymization. It initially partitioned the dataset using Hybrid method [31] .Based on the concept that a genetic algorithm can be used as a searching mechanism to fine-tune the setting of another clustering method and can be applied directly to the clustering problem. Propose GA solves the problem of GA that partitions the domain of each quasi-identifier for k-anonymization at the attribute level [33] but not at cell level. It also adopts the Michigan approach to encode a population of partitions. Each partition contains no fewer than k records, where each record indicates the index of a record in the dataset. After partitioning crossover operations are applied on them based on a rank-based selection strategy until final partitions are found. Once partition done, each quasi identifier of the record in partition generalized to the same value to satisfy k-anonymity. Experiments were performed on adult dataset and results are compared with Hybrid method [31]. Results showed that proposed method has 2-5% less information loss compared to Hybrid Method.

In paper [34], authors Aggarwal et al. Implemented R-Gather Clustering to anonymize the dataset. They used three features for clusters: the selection of quasi-identifying attribute values for the cluster center, the number of points within the cluster, and a set of values taken by the sensitive attributes .to publish and bound the error due to clustering use maximum cluster radius. It was similar to k-cantered clustering but they put restriction on the size of cluster to r records instead of no. Of clusters. They also implement r-cellular clustering o publish the radius of each cluster in addition to its center and the number of points within it. Also they generalized algorithms to allow a ϵ-fraction of points to remain un-clustered.

Authors Kabir, M. E., Wang, H. and Bertino, E. of Paper [12] tried to cope with the challenge of data utility and information loss. K-member algorithm [23] takes more time in the selection of record while generating clusters. To reduce time they proposed new clustering technique called systematic clustering. It focuses on selection of records for clustering. First it sorts all records in the whole data set with respect to quasi identifiers. Then set of clusters are constructed in such a way that the clusters are mutually exclusive, the sum of records of all clusters is equal to the total number of records and the size of each cluster is at least k which satisfies the criteria of k-anonymization. No.of clusters created depends on the value of anonymity parameter k .i.e. if number of records are 9 and k=3 then resulted clusters are 9/3=3. If the total number of records is not exactly divisible by the k-anonymity parameter, then the rest of the records will be included in the similar clusters where information loss is minimum and this process continues until the number of records in a particular cluster is k to satisfy the k-anonymity requirement. The problem tries to minimize the sum of all intra-cluster distances, where an intra-cluster distance of a cluster is defined as the maximum distance between any two records in the cluster. It has less information loss and time requirement O (n^2/k) which is better compared to k-member clustering algorithm.

P. R. Bhaladhare and D. C. Jinwala have proposed an algorithm in [35] first categorizes the attributes such as identifier, quasi-identifier and sensitive attributes. It then removes identifier attribute and sorts the dataset based on selected QI. Then it groups and clusters the original dataset. Based on first approach, they generated a sub-dataset based equal combination of quasi-identifier and sensitive attribute and second approach they created an unequal

combination of quasi-identifier (QI) and sensitive attribute (SA) and partitioned the sub-dataset records into k groups. They used a Systematic clustering algorithm [12] to generate the first cluster by selecting random record from the group and calculated Information loss for the cluster. Similarly, remaining clusters are created form other groups. The process repeated to add record to cluster based on low information loss. Experimental results gives less execution time and less information loss compared to Systematic clustering [12] and greedy k-member algorithm.

In paper [36], Ni, S., Xie, M. and Qian, Q. implemented Clustering Based K-anonymity Algorithm for Privacy Preservation and optimize it with parallelization. Algorithm works in for stages: Grading stage grades the tuple based on score calculated for categorical and numeric attribute and sort the records based on score and then Centering stage selects the centroid. Third clustering stage generates equivalence classes by placing record to cluster based on distance from centroid. Last generalization stage generalize all the tuples of cluster based on the numeric and categorical attributes. They use multithreading concept to achieve parallelization. The experimental results shows that they their serial version of algorithm GCCG has less information loss compared Incognito[17] and KACA[39] and execution time is nearer to Incognito[17] but it's a faster compared to KACA. They also compared serial and parallel version of their algorithm which shows parallel version executes faster than serial version but suffers from more information loss.

In paper [37] authors Zheng, W., Wang, Z., Lv, T., Ma, Y. and Jia, C. have implemented improved clustering algorithm by opting the idea of well-known Machine Learning classification algorithm, k-nearest neighbour algorithm to minimize the effect of random selection for the initial centroids which reduces information loss. They also considered the impact of positions of existed clusters on the selection of new centroid. They used the position of each generated cluster as a reference point to select new centroid to merge records with the closest distance. Improved Clustering algorithm iterates one time to reduce the over generalization of categorical attributes. Experiments were performed on Adult data set and results has less information loss compared with the k-member clustering algorithm [23], Mondrian multidimensional k-anonymity algorithm [14] and one-time k-means anonymity algorithm [30] for different value of n and k.

Ashoka K and Poornima B. presented organized clustering approach [38] along with equal combination of quasi-identifier and sensitive attributes using Systematic clustering [12] .Initially algorithm sort dataset based on QI selected then it identifies the number of cluster and creates sub dataset by partitioning the dataset into equal combinations of QI and SA groups. It starts creating cluster by selecting random record as centroid and repeats the same process for other groups. Then It places the records from group to cluster based on low information loss. Experimental results are compared with greedy k-member algorithm [23] and systematic clustering algorithm [12], it shows algorithm achieves less information loss and faster execution time.

Authors W. Zheng, Y. Ma, Z. Wang, C. Jia, and P. Li have proposed an algorithm to achieve k-anonymization via clustering and distributing sensitive attribute using L-diversity approach [39]. Algorithm starts by selecting random record as cluster centroid and start generating cluster by putting record to a cluster based on closeness to the centroid while keep L-diversity for sensitive attributes achieved or not. If no then continue placing record to the cluster and go for the second iteration of the cluster. Algorithm achieves k-anonymity and L-diversity at same time. They conducted experiments on Adult data set and compared the results with traditional (K,L)-member algorithm. Theoretical analysis and the experimental results shows that the improved L-diversity algorithm reduces the information loss and improve the privacy protection degree of sensitive data.

Yan, Y., Herman, E. A., Mahmood, A., Feng, T., and Xie, P. have proposed a weighted K-member clustering algorithm (WKMCA) [40] to reduce the effect of outliers by removing them based on AR-score calculation in initial phase, then adding record to the cluster to satisfy k-anonymity based on distance between records. The final adjustment phase places the remaining records and outliers into existing clusters which are closer to them to reduce information loss. Algorithm is compared with greedy K-member clustering algorithm[23], one pass K-means algorithm (OKA) [42], and the improved K-anonymity Algorithm based on clustering (IKA) [41].It has better clustering result compared to existing algorithm [23][41][42]. It has also lower information loss compared to [23][41][42] which provides more data availability. It performs faster compare to [23][41] but slower compared to [42].

Ferrao, M. E., Prata, P., and Fazendeiro P., have used partition based clustering algorithm in paper[43] to clusters data and analysed clustering validity parameters to investigate whether data structure is preserved after application of anonymization and clustering methods ,if yes up to what extent it preserves the utility of data. The process was consists of three steps: applying a privacy model; quantifying the risk of disclosure or re-identification; assessing data utility. They used several clustering validity indices to understand the utility of the data preserved

or not after data k-anonymization. The algorithms attempt to directly decompose the data set into a collection of disjoint clusters. This partition is built during an iterative optimization process repeated until its associated cost function reaches a minimum. The cost function, also designed performance index or objective function is a mathematical criterion expressing some desired features of local or global level of data structure of the resulting partition. Combining some heuristics with an adequate formulation of the objective function, it is possible to design an optimization process which can determine at least suboptimal partitions. The c-Means is used [46].The results suggest that for low dimensionality/cardinality datasets will put anonymization procedure to the risk of re-identification and maintaining the maximum usefulness of the data [44][45].Results also shows that increase in no. Of Quasi Identifier increase the performance issue and also increase the complexity.

In paper [47], M. Shin, Sunyong Y., Kwang H. Lee, and Doheon L. implemented the k-member cluster seed selection algorithm (KMCSSA) by introducing the closeness centrality concept to find the seed for selection of centroid of k-means clustering algorithm. The k-member cluster. The algorithm initially calculates the closeness centrality for all records and creates a seed list containing records with centrality values and from the seed list record with large centrality value is selected as an initial seed. Then k-1 nearest records is added to create a cluster with k records and simultaneously the same record is removed from the seed list and dataset. The process repeated for the record which has the record with next largest closeness centrality and cluster build as same way previously created. The Remaining records that are left are added to closest cluster. This algorithm avoids the recalculation of distance to build next cluster which leads to the less execution time and less information less. Experimental results analyses the information loss for different seed selection strategy and results shows for low k-values the information loss does not affected but on increasing the k-value information loss increase. Also The KMCSSA has less information loss compared to k-member [24] and one-pass k-mean clustering problem [28].

Parameshwarappa, P., Chen, Z., and Koru, G. have proposed approach called multi-level clustering [48] based on MDAV [49] and applied two anonymization techniques: k-anonymity and differential privacy. Multi-level clustering with aggregation initially computes the centroid of the data set. It finds a record r which has highest distance from the centroid and called record then after it finds k−1 nearest records to both r and s. If still 2*k records are remaining again process repeated for remaining records. At the end of all cluster build up and still records k to 2k-1 remaining then new cluster is formed. Then generalization is performed to achieve k-anonymity and used Fourier Perturbation Algorithm to achieve differential privacy. Experimental results show that proposed approach is better than MDAV[49].

Fan, Y., Shi, X., Zhang, S. and Tong, Y. proposed Multi-Attribute Clustering and Generalization Constraint (k,l)-Anonymity (MCKL) Algorithm[50] aiming to solve the problems of overgeneralization and insufficient diversity constraints. The generalization hierarchy quality directly affects the information loss of data in the anonymization process. A greedy partition strategy used to create generalization hierarchy for multidimensional data to reduce the degree of information loss. For this they used entire multidimensional data attribute space and sorted the attribute values by width first. Then they selected largest width value attribute as the division dimension and recursively partition the subspaces until all subspaces were indivisible to generate the generalization hierarchy of the attributes. Authors also used the concept of improved KNN clustering based on inter-tuple distance metric. Initially tuple with the width first is selected as cluster center and QIs are partitioned into KNN clusters and adding k-1 records with closest distance. While generating equivalence classes to solve the insufficient diversity problem of sensitive attribute authors applied the frequency–diversity constraints based on frequency of sensitive attribute boundary condition was set for it. According to that the frequency of S attributes was not greater than f calculated as division of no. of sensitive attribute in equivalence class to l diversity. Using this approach equivalence classes are create with no less than k records and satisfying the diversity value should not less than l. Then QI of equivalence classes are generalized based on generalization hierarchy. The experiments for MCKL were performed and analyze against cluster-based k-anonymity algorithm (CKA) and the cluster-l-diversity-based k-anonymity (CKL) algorithm .Results showed that MCKL has less information loss with varied value of QI and dataset size compared to CKA and CKL and less execution time than CKA but not CKL.

The 3V’s of big data expands it beyond static storage, encompassing real-time streams generated by various applications such as e-commerce, social media, healthcare, and telecommunications with benefits in time-sensitive and IoT-based applications. While there exist numerous effective privacy preservation methods for structured data, their direct application to dynamic real-time or streaming data poses challenges due to its constantly changing nature [51][52]. A data stream can be depicted as an infinite ordered sequence of tuples, each with a unique arrival sequence, which can be used to determine the priority for processing incoming tuples. Data streams clustering requires unique considerations compared to static data clustering. It presents challenges, including the fact that

stream records (instances) can only be read once and in sequence, without the option to store them. Additionally, there's a limited time interval for processing and publishing, necessitating outlier detection and processing [53]. Sliding Window Anonymization framework SWAF [54] was introduce to anoymize the stream data by protecting privacy and utility with consideration of limited processing time for each tuple of data steam, Small memory requirement. Experimental results showed larger value of k leads to more information loss while increase in sliding window size increases the anonymization quality.

In paper [53], Zubaroglu, A. and Atalay V. covered basic concept of data stream clustering and illustrated the concept drift, data structure of data stream, different time window models, outlier's detections and distinct algorithms for data stream clustering. They also compared clustering algorithm to cluster the data stream and discussed the open challenges for data stream clustering.

Authors Cao, J., Carminati, B., Ferrari, E., and Tan, K. L. presented Continuously Anonymizing streaming data via adaptive clustering (CASTLE)[55] providing guarantee to publish the anonymized data streams on-the-fly while satisfying the specified delay constraints and l- diversity. Initially algorithm start by adding incoming tuple to a cluster based on interval closest to cluster. If it is not possible to add tuple to existing cluster it creates new one by using minimum cluster enlargement to decrease the information loss. It also satisfies delay constraints by merging clusters and adding tuple to it based on value of k. They also reused the anonymized clusters to increase information quality. Experimental results showed CASTLE is efficient and effective compared to [56] but due to no restriction on maximum number of tuples to be inserted in cluster and merging of cluster increased the information loss. B-CASTLE [57] was introduced by Wang, P et al. to resolve these limitations of CASTLE. Algorithm used α parameter which adds the restriction on the maximum tuples number in each cluster. B-CASTLE requires that a tuple can be inserted into a cluster only if its size is less than threshold α. The clusters produced by B-CASTLE are more evenly balanced in size compared to CASTLE. Tuple with the closest to cluster is merged based on Correlation Distance. Results were compared to CASTLE showing less average information losses at each step and lower running time.

Zakerzadeh, H. and Osborn, S.L. proposed a cluster-based k-anonymity algorithm for numeric data stream called FAANST [58] by adapting k-member clustering algorithm. It accepts the input tuples up to given buffer size and placed them to cluster using k-member clustering. When cluster has k tuple based on information loss less than the given threshold algorithm publishes it and save it for future use. It number of tuples is less than k then tuples are suppressed. Experimental results show that FAANST outperforms compared to CASTLE [55] in terms of data loss, running time, and the number of suppressed tuples.

Guo, K. and Zhang, Q. have implemented a Fast clustering-based anonymization FADS [59] approaches with time constraints for data streams. It starts by reading a tuple and adding them to cluster and anonymized it and publish when no more tuples arrived. If the algorithm not found k-1 tuples to be published then it reuse the older anonymized tuple to satisfy constraint and publish it. Algorithm has a low average information loss and low running time compared to CASTLE [55] and FAANST [58]. FADS suffers by publishing a newly arrived tuple early before its time expiration just because it is one k-1 nearest neighbors of a tuple waiting for publication.

Mohammadian, E., Noferesti, M. and Jalili, R., have presented a parallel anonymization algorithm for big data stream called FAST [60]. A new proactive heuristic estimated - round - time is proposed in order to publish data before a specific expiration-time passed to overcome the problem occurred in FADS [59]. The results of experiments are compared with FADS [59] and results are efficiency and effectiveness for anonymizing big data stream. This algorithm has less the information loss and less cost metric for varying parameter compared to FADS.

Tekli J., Al Bouna B., Issa Y. B., Kamradt M. and Haraty R. have proposed (k,l)-Clustering for Transactional Data Streams Anonymization[61] using a bucketization technique. It releases l-diverse group based on QI which is created from a subset of clusters having disjoint centroids. It performs anonymization using two functions safe clustering and tuple assignment. They also performed experiment for supervised and unsupervised learning based clustering. Result shows that implemented approaches satisfy privacy constraints for stream data.

Dagadu Puri, G. and Haritha D. have implemented an anonymization method for health data stream [62]. Due to limitation of l-diverse group's repetition in stream data can cause re-identification of individual. In this method group are made based on similarity of incoming sensitive value and sensitive value of earlier published tuples. Result shows the data loss is decreased by using synonyms of the sensitive value in the group and it helps to avoid similarity attack for stream data.

Sopaoglu, U. and Abul, O. focused on minimizing average delay and keeping data quality high for stream data. They presented data stream k-anonymization framework and Utility Based approach for Data Stream Anonymization (UBDSA) [63] by introducing Cardinality Aware Information Loss metric. UBDSA assigns the

cluster to incoming tuple without delaying in buffer more than given delay constraint due to limited buffer capacity. Experimental results were compared with CASTLE [55] and FADS [59] for average delay and information loss. Algorithm gives better performance than CASTLE for cluster assignment distance. UBDSA has a better balanced performance metrics with the compared work and it is able to tune and balance the information loss and average delay metrics.

,Sopaoglu, U. and Abul, O have again presented and developed Classification Utility Aware Data Stream Anonymization(CUDSA) a k-anonymization method for data streams which protects the sensitive data and enables effective classification models[64]. Algorithm creates a cluster out of at least k number of tuples waiting in the buffer in such a way that it reduces the loss while combining two clusters. It works for both numeric and categorical attributes. Algorithm takes k and δ (delay) as hard constraint parameters. Other parameters like limit of how many anonymized clusters will be stored in the system to reuse, the window size, etc. It also use two function publish and Form cluster function to publish data after anonymization and to build the cluster, respectively. Experimental results were compared with CASTLE [55] and FADS [59] with respect to Information loss on the QI-attributes (IL), Classification accuracy (CA), results shows accuracy less proposed algorithm has more information loss compare to FADS.

Yang, L., Chen, X., Luo, Y., Lan, X., and Wang, W., have implemented a slide-window-based processing framework called utility-enhanced approach for Incomplete Data stream Anonymization (IDEA) in [65].This framework aims to balance privacy and utility while also considering missing data. It continuously reads tuples from the data stream and assigns them to clusters based on similarity measures of Quasi-identifier. When the window size is reached, old tuples are published. If no tuples are found, the remaining tuples are published. Clusters having less information loss than the threshold value are stored for reuse, which reduces the time for searching reusable clusters. The algorithm's results have less average information loss and execution time compared to IoT anonymization [66] and K-VARP [67]. Table 4 presents a summary of cluster-based anonymization approaches used for privacy preservation of structured data and stream data based on the literature review above.

Table 4 Summary of different cluster based k-Anonymization approach for static and stream data

SR. NO.	PAPER TITLE	SIMULATION TOOL/ TECHNOLOGY AND DATASET	PERFORMANCE CRITERIA	CLUSTERING/ CENTROID SELECTION TECHNIQUE	RESULT	LIMITATION
1.	Efficient k-anonymization using clustering techniques. [23]	Java 2 Platform, Adult	Information loss Execution time	greedy k-member clustering algorithm Random record as the cluster centroid	Less information loss compared to Mondrian [14]. Slower than the Partition based algorithm.	Information loss increases if outlier is selected as cluster centroid.
2.	Capturing data usefulness and privacy protection in K-anonymisation[25]	Java Adult	Discernibility Usefulness and protection	greedy k-member clustering algorithm based on similarity of QI and Random record as cluster centroid	Better anonymization compared to Mondrian [14] and K-Members [27]	Performance can be improved by considering heuristics.
3.	A k-Anonymity Clustering Method for Effective Data Privacy Preservation[28]	Excel VBA programming language Iris, Wine, and Zoo dataset	Information Distortion	Weighted Feature C-Means Clustering Algorithm Random selection of centroid	Compared to hierarchical clustering less computational efficiency and less information distortion	Numerical attributes only.

SR. NO.	PAPER TITLE	SIMULATION TOOL/ TECHNOLOGY AND DATASET	PERFORMANCE CRITERIA	CLUSTERING/ CENTROID SELECTION TECHNIQUE	RESULT	LIMITATION
4.	An Efficient Clustering Method for k-Anonymization [30]	Java Adult	Information loss Execution Time	Single iteration of K-Means algorithm. Random record selected as centroid	less information loss and much less execution time than the k-member [23]	Extended to l-diversity,t-closeness ,(α, k)-anonymity .
5.	Genetic algorithm-based clustering approach for k-anonymization [32]	Java Adult	Information loss	assignment-oriented method	Lower total information loss than the Hybrid method by 2–5%.	Extended to l-diversity,t-closeness ,(α, k)-anonymity .
6.	Efficient systematic clustering method for k-anonymization [12]	VBA programming Adult	Average information loss Execution time	Systematic clustering. Random selection of record for centroid	Less execution time and information loss compared to k-member [23] and Mondrian [14]	Extended to l-diversity,t-closeness ,(α, k)-anonymity .
7.	Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model[35]	Java Adult	information loss Execution time	Systematic clustering based on Unequal and Equal combination of QI and SA approach Random selection of record for centroid	Less IL and execution time Compared to Greedy k-member [6] and Systematic clustering [16]. Equal approach is better than unequal approach.	To investigate the performance on a combination of multiple SA and QI attributes.
8.	Clustering Based K-anonymity Algorithm for Privacy Preservation[36]	Java Multithreading Framework for Parallel Programming Adult	information loss Execution time	GCCG based serial/Parallel clustering tuple grading based Centroid selection.	Less IL and better execution time compared to KACA[15] and incognito[17]	Should be big data oriented
9.	K-Anonymity Algorithm based on Improved Clustering [37]	Python Adult	information loss Execution Time	Centroid updated Iteratively for clustering	20% less IL compared to k-member clustering [23], Mondrian multi-dimensional k-anonymity [14] and OKA [30]. Execution time decrease with k as no.of cluster decreases.	Cluster based approach only performed for categorical data.

SR. NO.	PAPER TITLE	SIMULATION TOOL/ TECHNOLOGY AND DATASET	PERFORMANCE CRITERIA	CLUSTERING/ CENTROID SELECTION TECHNIQUE	RESULT	LIMITATION
10.	Organized Clustering Method for Privacy Preserving Data Publishing[38]	Java Adult	information loss execution time	QI and sensitive information partition based systematic clustering. Random record as cluster centroid.	Less execution time, IL compared to greedy k-member [23] and systematic clustering [12].	Should be applicable to multiple SA and QI combination.
11.	Effective L-Diversity Anonymization Algorithm Based on Improved Clustering [39]	adult	Information loss	Random record as initial centroid and then greedy selection based approach to select centroid.	Less information loss compared to (K,L)-member diversity	-
12.	A weighted K-member clustering algorithm for K-anonymization [40]	R studio adult	information loss execution time	Clustering method based on weight indicators for numerical and categorical attributes. random record as cluster centroid	Less IL compared to Greedy K-member [23] and OKA [30], IKA [37]. Less execution time compare to IKA and Greedy but not better compare to OKA	-
13.	Anonymous Methods Based on Multi-Attribute Clustering and Generalization Constraints[50]	Adult	Information loss	Improved KNN clustering with Highest width Attribute as centroid. QI selection based on Greedy approach.	Less information loss and higher execution time compared to CKA and CKL	
14.	Electronic Medical Records Privacy Preservation through k-Anonymity Clustering Method[47]	Java Adult	Information loss	Based on closeness centrality seed list of all records and large centrality value is selected as an initial centroid from seed list.	Less information loss compared to k-member [24] and one-pass k-mean clustering problem[30]	Data distribution for closeness centrality can lead to IL. Can use to reduce Reidentification Risk l-diversity [8], t-closeness[9], Optimal k-anonymization [5]
15.	A clustering-based anonymi-zation approach for privacy-preserving in the healthcare cloud[68]	ARX /EMRbots R programmin g Tool	information loss, scalability, execution time	K-means++ clustering Distribution function used to remove less frequent data.	Scalable, IL reduced by 1.5 times and run time 3.5 times compared to AKA and GCCG.	Categorical data is converted to numerical. K-means++ only work for numeric data

SR. NO.	PAPER TITLE	SIMULATION TOOL/ TECHNOLOGY AND DATASET	PERFORMANCE CRITERIA	CLUSTERING/ CENTROID SELECTION TECHNIQUE	RESULT	LIMITATION
16.	Privacy Preserving Parallel Clustering Based Anonymization for Big Data Using MapReduce Framework [69]	Hadoop with HDFS, YARN and MapReduce frameworks Heart Disease Kasandra Statlog HEPMASS Adult	Kullback–Leibler divergence, F-measure, classification accuracy discernibility cost.	K-means used to divide dataset and Parallel cluster based algorithm. Frequency counted for sensitive attribute using map reduce and equivale-nce class created,	Less execution time, better utility and accuracy compared with KNN(G,S) and (G,S)	can be more scalable and speedy.
17.	Adaptive k-Anonymity Approach for Privacy Preserving in Cloud[70]	Python Adult	Execution Time information loss	Used systematic approach for seed selection to cluster the records	Less Execution time and information loss than enhanced clustering[84] method and Hybrid method [31]	Optimization technique can be used to decrease information loss.
18.	Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop [71]	Hadoop-1.2.1 Java Weka Tool adult synthesis dataset	degree of privacy data utility Scalability execution time.	KNN—(G,S) clustering algorithm	Scalable with increasing the data set size.	Need to identify cluster size to reduce the time.
19.	S-CPM: Semantic-Similarity Cluster based Privacy Preservation Model with Cell Generalization Principle[72]	apache spark Income-Census (KDD) Bank Credit Card	Data distortion compatibility to privacy breach attack scalability	t-centroid for first phase clustering and neighborhood-aware clustering for merging the clusters.	More scalable and Time efficient for large-scale data set Compared to TDS-PP	Data anonymization through a bottom-up approach also to protect against privacy breach attacks.
20.	DHkmeans- ℓ-diversity:distribut ed hierarchical K-means for satisfaction of the ℓ-diversity privacy model using Apache Spark[73]	Apache Spark Eclipse IDE. pocker hand dataset	Information loss Accuracy F-measure Execution Time	Hierarchical kmeans-based data clustering. Seed selection is based on less similarity between QI and more similarity between SA.	Less information loss Compared to MapReduce-based Mondrian methods. Privacy increased.	other clustering method and other privacy model can be applied can do it for high dimension big data
21.	Evaluating the Effectiveness of Clustering-Based K-Anonymity and KNN Cluster for Privacy Preservation[74]	Adult	NCP Execution time	comparison of k-member ,KNN and k-anonymity	It has less execution time compared to Greedy QS k-anonymity w.r.t value of k, no.of attributes and no.of records.	Integration with machine learning will give better results.

SR. NO.	PAPER TITLE	SIMULATION TOOL/ TECHNOLOGY AND DATASET	PERFORMANCE CRITERIA	CLUSTERING/ CENTROID SELECTION TECHNIQUE	RESULT	LIMITATION
22.	CASTLE: A continuously - Anonymizing data streams[55]	Java Adult SFU-Adult	Utility Average Information loss Run time	cluster based anonymization algorithm for data streams to satisfy l-diversity	Less efficient. Frequent merging and splitting operations increase the time complexity	No restriction on buffering of stream or Size of cluster.
23.	B-CASTLE: An efficient publishing algorithm for k-anonymizing data streams[57]	JAVA	Run time Information loss	Avoid splitting and merging of cluster like CASTLE.	Less execution time and Less Information loss than CASTLE [55].	Can be implemented to satisfy l-diversity.
24.	FAANST: Fast anonymizing algorithm for numerical streaming data[58]	C++ using Visual Studio 2005 Pen-Based Recognition of Handwritten Digits Dataset	Information loss Run time Anonymity degree, window size, and quality of clusters.	k-means clustering	Less runtime than CASTLE [55].	Tuple delay constraints can be soften. It considers numerical data so, Medoid for categorical data.
25.	Fast clustering-based anonymization approaches with time constraints for data streams[59]	Java Adult	Average Information loss Run time	k-means clustering Without splitting and merging based on the new tuple and its neighbors. Delay factor introduce which Increase tuples in buffer	Limits the no. of anonymized clusters. The time and space complexity linear with the size of the data. information loss is less compared to CASTLE[55] and FAANST[58]	Information loss can be reduces and other technique like t-closeness and differential privacy can be applied.
26.	FAST: Fast Anonymization of Big Data Streams[60]	Java adult	Average information loss Average execution time	Cluster based stream anonymization	Increase in no.of clusters decreases IL, decreases run time by increase in no.of threads.	distributed cloud-based framework can be used to gain cloud computation power and achieve high scalability
27.	A Novel Method for Privacy Preservation of Health Data Stream [62]	Human disease ontology used to find synonyms of disease terms.	information loss	Stream data anonymization using l-diversity	It avoids similarity attack. Low information loss.	-
28.	A utility based approach for data stream Anonymization [63]	Java Adult TELCO	Information Loss Average Delay	CAIL distance based clustering.	Better compared to CASTLE [55] ,FADS. IL and Average Delay is Balanced.	Need adopted with l-diversity.

SR. NO.	PAPER TITLE	SIMULATION TOOL/ TECHNOLOGY AND DATASET	PERFORMANCE CRITERIA	CLUSTERING/ CENTROID SELECTION TECHNIQUE	RESULT	LIMITATION
29.	Classification utility aware data stream anonymization [64]	Java and Python adult nursery Telco	information loss classification accuracy	CUDSA based on minimum loss while generating clusters	better Compare to CASTLE and FADS but IL is high compare to FAANST	
30.	IDEA: A Utility-Enhanced Approach to Incomplete Data Stream Anonymization [65]	JAVA Adult INFORMS	Information loss Run time	Clustering of tuple with missing value.	compared to K-VARP[67] and IoT anonymity [66] less IL and more runtime	Need of data stream processing framework in a distributed environment.
31.	Advancing Data Privacy: A Novel K-Anonymity Algorithm with Dissimilarity Tree-Based Clustering and Minimal Information Loss[75]	Adult	NCP Execution time	Use of Dissimilarity Tree-based strategy to select the centroids	More accurate clusters reduces NCP and computing time compared to KNN .Less NCP and 20% reduction in IL compared to KNN.	May apply for larger dataset and can be go with differential privacy techniques.
32.	K-Anonymity Privacy Protection Algorithm for Multi-Dimensional Data against Skewness and Similarity Attacks[76]	Python Dataset from cooperative medical information technology company in Hangzhou, China	clustering accuracy diversity anonymity	Multi-dimensional sensitive data clustering based on sensitive attribute using fuzzy c-means clustering. Improved African vultures optimization used to distribute SA.	Improves clustering accuracy, diversity, anonymity compared to CPPA [78] , SKAM[79] and PAMS[80] , under skewness and similarity attacks, no improvement in IL.	Need to try efficient anonymization algorithms for dynamic data with data analytics and machine learning methods

IV. PERFORMANCE EVALUATION PARAMETER

Several cluster-based k-anonymity approaches have been proposed for k-anonymization, and their performance has -been evaluated using various performance metrics. This section outlines key metrics such as information loss, accuracy, and execution time used to measure their effectiveness.

A. Information Loss[23][73]

As we know anonymization would be beneficial if we can achieve data privacy. While achieving privacy, the utility of the data may also be affected which leads to fewer values can be extracted from the data. Information loss is defined as reduction in data utility. High loss of information considers as less utility of anonymized information. To calculate information loss, assume n and a are the number of data records and attributes, respectively. $Upper_{ij}$ and $lower_{ij}$ are the upper and lower bounds of the j^{th} attribute in the i^{th} record after the anonymization process. max_j and min_j are the maximum and minimum values of j^{th} attribute in the original dataset. Then formula to calculate Information loss [73]:

$$\text{Information loss}=\frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{|\text{Upperij}-lowerij|}{maxj-minj} \quad (7)$$

Information loss calculation for cluster [23]: Let equivalence class or cluster generated using clustering is e = {r_1, … , r_k} contains record r_1 to r_k . Record r_i has quasi-identifier consists of numeric attributes N_1,…, N_m where MIN_{Ni} and MAX_{Ni} be the min and max values in e with respect to attribute Ni and categorical attributes C_1,⋯, C_n. Let TC_i is the taxonomy tree defined for the domain of categorical attribute C_i. Now consider $\cup C_i$ be the union set of values in e with respect to attribute C_i. Information loss to generalize cluster e is calculated as:

$$(\text{IL}(e)) = (|e|) * \left(\sum_{i=1,\ldots,m} \frac{(\text{MAX}_{Ni} - \text{MIN}_{Ni})}{|Ni|} + \sum_{j=1,..,n} \frac{H\left(\Lambda\left(\cup C_j\right)\right)}{H\left(T\left(C_j\right)\right)} \right) \quad (8)$$

|e| defines the number of records in e, |Ni| represents the size of numeric domain Ni, Λ ($\cup C_j$) represents the sub-tree rooted at the lowest common ancestor of every value in $\cup C_j$, height of taxonomy tree T is defined as H (T) .The cumulative information loss of set ε containing cluster e_i i=1,…,k, resulting from anonymization can be calculated by aggregating the information loss from each cluster as:

$$\text{Total Information loss IL} = \sum_{e_i \in \varepsilon} \text{IL}(e_i) \quad (9)$$

Several clustering techniques exist for grouping similar records. The equations mentioned above for computing the distance between records and information loss may differ depending on the specific clustering technique employed for record clustering. Information loss also be considers as Utility Loss. Different matrices are used to measure the Data Utility loss. Some of matrices are mentioned below:

B. Discernibility Metric (DM)[69][77]

The Discernibility Metric (DM) cost measures data utility loss by measuring size of the equivalence classes. Lower value of DM refers as small size of equivalence class which results to less utility loss. The higher the DM value the utility loss would be high. .It can be evaluate as [77]:

$$\text{DM} = \sum_{|E| \epsilon D'} |E|^2 \quad (10)$$

Where D' is represent anonymized and privacy matrices dataset and E is the equivalence class generated by aonymized approach. We need to minimize the amount of tuples that are indistinguishable in an equivalence class to satisfy the k-anonymity criterion.

C. Kullback-Leibler-Divergence (KLD)[69][77]

KLD concept was introduced in probability theory and information theory. It measures the difference between two probability distributions over the same variable x. In anonymization process it computes the distance between the distribution before and after the anonymization of data. Lower value of KLD denotes the lower distortions and it is easy to identify the original value from the matching anonymized value i.e., low privacy. If both the distribution are same then KL will be considered as Zero. KL can be calculated for two distribution p(x) and q(x) as[77]:

$$KL = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \quad (11)$$

KLD can also be evaluate as entropy [77]. It is utilized to measure the uncertainty or randomness for a given probability distribution [81].Entropy can be used in two way in the anonymization process. One is to assign priority to QI used for anonymization process [81-83]. it identifies the nature of association that an attribute share with the sensitive attribute, thereby quantifying the uncertainty associated with a prediction. Attribute with low entropy suggest high information content to identify the sensitive attribute. Second is used to identify the amount of information that is lost by anonymization. KLD can be computed in terms of entropy:

$$KL=(B)- H(B|A) \quad (12)$$

Where B represents the sensitive attribute and A represents set of QI attributes. Entropy of the sensitive attribute B is H(B) and H(B|A) is the conditional entropy of B conditioned on the QI attributes A.

D. Average Equivalence Class Size (CAvg)[69][77]

It is also used as measure for data utility loss [69]. Number of more Equivalence classes represent less records available in Equivalence class, requires less generalization to make them k anonymous, describing low value of C_{Avg}, representing less information loss.[72, 77],It can be calculated as[77] :

$$C_{Avg}=\left(\frac{|D|/N_E}{k}\right) \quad (13)$$

for dataset D with N_E number of equivalence classes having k as anonymization parameter.

C. Accuracy [83]

Accuracy is metric for evaluating classification models. Classification Accuracy (CA) defines the correctness of classification achieved by the algorithm. CA is calculated to evaluate data accuracy of anonymized dataset and the higher values of classification accuracy are preferred.CA values which are closer to the original values mean that the information loss is low which refers to higher data utility. The classification accuracy for clustering based algorithm can be calculated as:

$$CA=\frac{\text{Number of clusters correctly classified}}{\text{Total number of clusters}} \quad (14)$$

D. F-Measure[69]

It is used to evaluate any clustering algorithms effectively and classification algorithms for comparisons of records before and after anonymization. Higher value of F-Measure are denotes high accuracy. It is calculated based on precision and recall which are measures of exactness and completeness, respectively. The F-measure is calculated based on value of precision=$\frac{TP}{TP+FP}$ and Recall= $\frac{TP}{TP+FN}$as:

$$\text{F-measure}= \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recal}} \quad (15)$$

Where, P is the number of positive tuples and N is the number of negative tuples. True positives (TP) and True negatives (TN) represents correctly labelled as positive tuples and correctly labelled as negative tuples, respectively.Same way false positives (FP) are the negative tuples which are mislabelled as positive and False negatives (FN) are the positive tuples which are incorrectly labelled as negative.

There is a possibility that Outliers are present in the dataset. This outlier also affects the performance of anonymization process and affects the data utility and privacy. So, while designing one should consider the outliers exists in dataset and required to perform Outlier Analysis and detection without ignoring it.

E. Execution Time (Time Complexity)

Time required by the algorithm to complete the process of anonymization. In anonymization process the it depends on the various parameters like anonymization value (k), number of records, number of clusters, etc[73].

V. SUMMARY AND DISSCUSION

From the study of section II, III, IV and V, we can summarized that clustering algorithm plays important role to generate equivalence classes for the anonymization of dataset. Equivalence classes generated from clustering algorithm lead to less information loss compared to traditional way because of the equivalence classes generated based on similarity or distance calculation, nevertheless, there is still need for better algorithm due to their limitations. Some of them are:

Real world datasets are the mixed of numerical and categorical attribute and clustering algorithm can operate on specific type data. Majority of approaches perform clustering based on numeric attribute, to cluster records based on categorical attribute ,it need to be transform in numeric type. Building clusters based on only numeric attribute is not preferable for real world dataset. Exploring clustering algorithms suitable for categorical data is an ongoing area of research for anonymization purposes.

Equivalence class size plays the critical role in anonymizaion process. Also, cluster size depend on the anonymity parameter k. Large cluster size needs to place more records in equivalence class and for anoymization they are generalized to same centroid or generalization value increases the information loss, whereas less records in cluster sacrifices the privacy of data holder. So, there is a requirement to select appropriate value of k for cluster size to balance the privacy and utility trade-off. When considering privacy and utility, the selection of the appropriate set of quasi-identifiers is crucial in the trade-off between privacy and utility, as these identifiers are

publicly disclosed and are at a high risk of re-identification. Moreover, having numerous quasi-identifiers within an equivalence class necessitates more extensive generalization, leading to increased information loss.

Secondly, Most of the approaches select random record as the cluster centroid, if randomly selected record is outlier then it directly affects the increase in information loss within the cluster. Hence, there is need of proper technique to select the cluster centroid among the available records as it affect the information loss and data utility. Additionally, it is essential to ensure proper management of outliers as they can significantly impact the utility and information loss of anonymized data.

From the literature on stream data anonymization, we can deduce that stream data refers to real-time data subject to time delay constraints. The challenges encountered in clustering based k-anonymity approach for static data also pertain to stream data with additional time delay constraint. Since, Stream data must be assigned to suitable cluster before their expiration and anonymized prior to publication.. Occasionally, newly arrived tuples may not fit into any existing cluster, necessitating the creation of new cluster or merging with the existing anonymized cluster. However, if a new cluster lacks a predefined size k, it cannot be processed, potentially resulting in tuple expiration before anonymization and publication. In such instances, publishing tuples without anonymization poses risks of privacy breaches or re-identification.

Combining an incoming tuple with any existing cluster may increase information loss to fulfil k-anonymity criteria. To integrate incoming tuples with existing clusters, an appropriate cluster must be available in memory. However, due to memory constraints in streaming data, it is not feasible to keep a large number of clusters in memory, which can lead to the unavailability of a suitable cluster when needed. Given these memory limitations in streaming data, achieving high-performance processing is crucial.

Once more, k-anonymity alone does not adequately protect against background knowledge attacks and other re-identification risks. With the widespread adoption of the internet and internet-based applications, data collection and sharing occur in distributed environments. Hence, there is a pressing need to develop a scalable approach for anonymizing streaming data or its extended methods capable of handling large datasets in distributed environments.

VII. CONCLUSION

Numerous k-anonymity solutions are documented in the literature for safeguarding individual privacy. The k-anonymity approach necessitates the creation of equivalence classes or groups, followed by anonymization through generalization or suppression. Clustering presents an efficient method for forming equivalence groups with reduced information loss compared to traditional approaches. This paper presents a comprehensive exploration of clustering-based anonymization techniques for both structured and streaming data. It discusses the challenges associated with using clustering algorithms to generate equivalence classes that meet k-anonymity requirements. Before choosing the clustering algorithm, it is essential to accurately distinguish between quasi-identifiers and sensitive attributes in the dataset according to privacy requirements. The selection of clustering algorithms should be guided by the types of attributes in the dataset, and appropriate centroid selection strategies are crucial for achieving a balance between privacy preservation and information loss.

For scalable k-anonymization of streaming data, clustering algorithms must generate and anonymize equivalence classes while considering time delays and limited memory resources. Considering current research trends, the implementation of cluster-based k-anonymity approaches and their extensions such as l-diversity and t-closeness in distributed big data environments remains an open research challenge.

REFERENCES

[1] P. Wayner, “7 top tools for taming big data,” 2012. [Online]. Available: https://goo.gl/XsmfWM

[2] Nie, J.-Y., Institute of Electrical and Electronics Engineers, & IEEE Computer Society. (n.d.). 2017 IEEE International Conference on Big Data : proceedings : Dec 11- 14, 2017, Boston, MA, USA.

[3] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,vol. 10, no. 05, pp. 557–570, 2002.

[4] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, & Yong Ren. “Information Security in Big Data: Privacy and Data Mining.” IEEE Access, 2, (2014), 1149–1176.

[5] V.Ciriani, S. De Capitani di Vimercati, S.Foresti, and P.Samarati,”k-Anonymous Data Mining:A Survey “.Advances in Database Systems,volume 34,pp.105-136,2008

[6] Surana, J., Khandelwal, A., Kothari, A., Solanki, H., & Sankhla, M,”Big Data Privacy Methods”, International Journal of Advance Research and Innovative Ideas in Education,3(2), . (2017) 5500-5507

[7] Premkamal, P. K., Pasupuleti, S. K., & Alphonse, P. J. A.”A new verifiable outsourced ciphertext-policy attribute based encryption for big data privacy and access control in cloud.” Journal of Ambient Intelligence and Humanized Computing, 10(7), (2019) 2693–2707.

[8] L. Sweeney,”Guaranteeing anonymity when sharing medical data, the Datafly system.” Proceedings, Journal of the American Medical Informatics Association. Washington, DC: Hanley & Belfus, Inc., 1997.

[9] A. Hundepool and L. Willenborg.” μ- and μ-argus: software for statistical disclosure control.” Third International Seminar on Statistical Confidentiality. Bled: 1996.

[10] L. Sweeney. “Towards the optimal suppression of details when disclosing medical data, the use of sub-combination analysis.” Proceedings, MEDINFO 98. International Medical Informatics Association. Seoul, Korea. North-Holland, 1998.
[11] P. Samarati and L. Sweeney, “Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression,” Technical Report SRI-CSL-98-04, 1998.
[12] Kabir, M. E., Wang, H., & Bertino, E..” Efficient systematic clustering method for k-anonymization.” Acta Informatica, 48(1), (2011) 51–66. https://doi.org/ 10.1007/s00236-010-0131-6
[13] Ayala-Rivera V ,McDonagh P ,Cerqueus T ,Murphy L . ,”A systematic comparison and evaluation of k -anonymization algorithms for practitioners.” Trans. Data Privacy 2014;7(3):337–70 .
[14] LeFevre, K., DeWitt, D. J., & Ramakrishnan, R.. “Mondrian multidimensional k-anonymity.” In Proceedings of the 22nd IEEE international conference on data engineering (ICDE’06), 2006
[15] J. Y. Li, R. C. Wong, A. W. Fu, and J. Pei, “Achieving k-anonymity by clustering in attribute hierarchical structures," in 8th International Conference on Data Warehousing and Knowledge Discovery, pp. 405-416, Krakow, Poland, Sept. 2006.
[16] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke,M. Venkita subramaniam"l-Diversity: Privacy Beyond k-Anonymity", ACM Transactions on Knowledge Discovery from Data,Volume 1,Issue 1, ,2007
[17] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito:Efficient full-domain k-anonymity," in Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (Sigmod'05), (2005) pp. 49-60, Baltimore, Maryland, USA
[18] W Zheng, Y Ma, Z Wang, C Jia, P Li, ,”Effective L-Diversity Anonymization Algorithm Based on Improved Clustering”, Cyberspace Safety and Security: 11th International Symposium, CSS 2019 ,318-329
[19] Li, N., Li, T., & Venkatasubramanian, S.,” t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”. IEEE 23rd International Conference on Data Engineering. 2007.
[20] S. Kaushik, “An introduction to clustering & different methods of clustering (2016, December10). Retrieved July5,2018, https://www.analyticsvidhya.com/blog/ 2016/11/anintroduction to-clustering-and-different-methods-of-clustering/
[21] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou, A. Bouras, “A survey of clustering algorithms for big data: taxonomy and empirical analysis. “,IEEE Trans. Emerg. Top. Comput. 2, (2014) 267–279
[22] Ghosal, A., Nandy, A., Das, A. K., Goswami, S., & Panday, M., “A Short Review on Different Clustering Techniques and Their Applications”,Advances in Intelligent Systems and Computing, 937, (2020) 69–83. https://doi.org/10.1007/978-981-13-7403-6_9
[23] Byun, J.-W., Kamra, A., Bertino, E., & Li, N.. Efficient k-anonymization using clustering techniques. In 12th International conference on database systems for advanced applications (DASFAA), Bangkok, Thailand, (2007) pp. 188-200.
[24] Somasundaram and U. S. Reddy, ``Data imbalance: Effects and solutions for classification of large and highly imbalanced data," in Proc.1st Int. Conf. Res. Eng., Comput. Technol., 2016, pp. 1-16.
[25] Loukides, G., & Shao, J.,” Capturing data usefulness and privacy protection in k-anonymisation.”, In Proceedings of the 2007 ACM symposium on applied computing, (2007)
[26] R. Bayardo and R. Agrawal,”Data privacy through optimal k-anonymization”, in Proceedings of the 21st International Conference on Data Engineering, (2005) 217-228.
[27] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, “Efficient k-anonymity using clustering technique” Purdue University-CERIAS Tech Report 2006. 10.1007/978-3-540-71703-4_18
[28] Chiu, C.-C., & Tsai, C.-Y.. “A k-anonymity clustering method for effective data privacy preservation.” In Third international conference on advanced data mining and applications (ADMA), (2007) pg. 89-99.
[29] Jain, A.K., Murty, M.N., Flynn, P.J., ,”Data Clustering: A Review.”, ACM Computer Survey 31, (1999) 264–323.
[30] Lin, J.L., Wei, M.C.,”An efficient clustering method for k-anonymization”, In International Workshop on Privacy and Anonymity in Information Society.(ACM) (2008) pp. 46-50.
[31] Lin, J.-L., Wei, M.-C., Li, C.-W., & Hsieh, K.-C., “A hybrid method for k-anonymization”, In Proceedings of IEEE Asia-Pacific Services Computing Conference (APSCC’08) (2008) pp. 385–390.
[32] Lin, J. L., & Wei, M. C.,”Genetic algorithm-based clustering approach for k-anonymization”, Expert Systems with Applications, 36(6), (2009) 9784–9792. https://doi.org/10.1016/j.eswa.2009.02.009
[33] Lunacek, M., Whitley, D., & Ray, I.,”A crossover operator for the k-anonymity problem”, In GECCO’06: Proceedings of the 8th annual conference on Genetic and evolutionary computation, New York, NY, USA, (2006) pp. 1713–1720.
[34] Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., & Zhu, A. “Achieving anonymity via clustering.”,ACM Transactions on Algorithms, 6(3) (2010). https://doi.org/10.1145/1798596. 1798602
[35] P. R. Bhaladhare and D. C. Jinwala ,“Novel approaches for privacy preserving data mining in k- anonymity model," Journal of Information Science and Engineering, vol. 32, no. 1, (2016) pp. 63-78
[36] Ni, S., Xie, M., & Qian, Q.,”Clustering based k-anonymity algorithm for privacy preservation.” International Journal of Network Security, 19(6), (2017) 1062–1071. https://doi.org/10.6633/IJNS.201711.19(6).23
[37] Zheng, W., Wang, Z., Lv, T., Ma, Y., & Jia, C.,”K-anonymity algorithm based on improved clustering.”, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11335 LNCS,(2018) 462–476. https://doi.org/10.1007/978-3-030-05054-2_36
[38] Ashoka K, Poornima B.”Organized Clustering Method for Privacy Preserving Data Publishing.”, In International Journal of Engineering and Manufacturing Science (Vol. 8, Issue 1). (2018) http://www.ripublication.com 2019_10.1007@978-3-030-37352-8
[39] W Zheng, Y Ma, Z Wang, C Jia, P Li,, ”Effective L-Diversity Anonymization Algorithm Based on Improved Clustering”, Cyberspace Safety and Security: 11th International Symposium, CSS (2019) 318-329
[40] Yan, Y., Herman, E. A., Mahmood, A., Feng, T., & Xie, P.,”A weighted K-member clustering algorithm for K-anonymization. Computing,” 103(10), (2021) 2251–2273. https://doi.org/10.1007/s00607-021-00922-0
[41] Zheng W.T., Zhongyue W., Tongtong L.v., Ma Y., Jia C,”K-anonymity Algorithm Based on Improved Clustering” In: Proceedings of the 18th International Conference on Algorithms and Architectures for Parallel Processing ,Guangzhou, China, November, (2018) 462-476.
[42] Lin J., MengCheng W.,”An Efficient Clustering Method for k-Anonymization.” In: Proceedings of the 11th International Conference on Extending Database Technology, Nantes, France, (2008) 46-50.
[43] Ferrao, M. E., Prata, P., & Fazendeiro, P.,”Utility-driven assessment of anonymized data via clustering”, In Scientific data (Vol. 9, Issue 1, (2022) p. 456. NLM (Medline). https://doi.org/10.1038/s41597-022-01561-6
[44] Prasser, F., Eicher, J., Spengler, H., Bild, R. & Kuhn, K. A., ”Flexible data anonymization using ARX—Current status and challenges ahead.” Softw. Pract. Exp. 50, (2020) 1277–1304.
[45] Churi, P., Pawar, A. & Moreno-Guerrero, A. J.,” A comprehensive survey on data utility and privacy: Taking indian healthcare system as a potential case study.”Inventions 6, (2021) 1–30
[46] Bharara, S., Sabitha, S. & Bansal ,” A. Application of learning analytics using clustering data mining for students’ disposition analysis.”Educ. Inf. Technol. 23, (2018) 957–984
[47] M. Shin,Sunyong Yoo,Kwang H. Lee,Doheon Lee, "Electronic Medical Records Privacy Preservation through k-Anonymity Clustering Method", in The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems, (2012) 1119-1124.
[48] Parameshwarappa, P., Chen, Z., & Koru, G. (n.d.). A Multi-level Clustering Approach for Anonymizing Large-Scale Physical Activity Data.

[49] Solanas, A., Martinez-Balleste, A., and Domingo-Ferrer, J. "V-mdav: a multivariate microaggregation with variable group size.", In 17th COMPSTAT Symposium of the IASC, Rome, (2006) pages 917–925

[50] Fan, Y., Shi, X., Zhang, S., & Tong, Y. "Anonymous Methods Based on Multi-Attribute Clustering and Generalization Constraints.", Electronics (Switzerland),12(8) (2023) .https://doi.org/10.3390/electronics12081897

[51] R. Patil, P. D. Patil, S. Kanase, N. Bhegade, V. Chavan, and S. Kashetwar, "System for analyzing crime news by mining live data streams with preserving data privacy," in Sentimental Analysis and Deep Learning. Singapore: Springer, (2022) pp. 79-811

[52] .J. Kumar,,"Slide window method adapted for privacy-preserving: Transactional data streams," Eur. J. Mol. Clin. Med., vol. 8, no. 2, (2021)pp. 2528-2538,

[53] Zubaroğlu, A., & Atalay, V., "Data stream clustering: a review." Artificial Intelligence Review, 54(2), (2021) 1201–1236. https://doi.org/10.1007/s10462-020-09874-x

[54] Wang, W., Li, J., Ai, C., & Li, Y.."Privacy protection on sliding window of data streams." International Conference on Collaborative Computing: Networking, Applications and Worksharing (2007). doi:10.1109/colcom.2007.4553832

[55] Cao, J., Carminati, B., Ferrari, E., & Tan, K. L.. "CASTLE: Continuously anonymizing data streams." IEEE Transactions on Dependable and Secure Computing, 8(3), (2011) 337–352. https://doi.org/10.1109/TDSC.2009.47

[56] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Int'l Conf. Extending Database Technology (EDBT), (2004) pp. 183-199.

[57] Wang, P., Lu, J., Zhao, L., & Yang, J."B-CASTLE: An efficient publishing algorithm for k-anonymizing data streams", Proceedings - 2010 2nd WRI Global Congress on Intelligent Systems, GCIS 2010, 2, (2010) 132–136. https://doi.org/10.1109/GCIS.2010.196

[58] Zakerzadeh, H., Osborn, S.L.,"FAANST: fast anonymizing algorithm for numerical streaming DaTa.", In: J. Garcia-Alfaro et al. (Eds.): DPM 2010 and SETOP 2010, LNCS 6514, (2011) pp. 36–50, https://doi.org/10.1007/ 978-3-642 -19348-4 4

[59] Guo, K., & Zhang, Q.,"Fast clustering-based anonymization approaches with time constraints for data streams", Knowledge-Based Systems, 46, (2013) 95–108. https://doi.org/10.1016/j.knosys.2013.03.007

[60] Mohammadian, E., Noferesti, M., & Jalili, R.,"FAST: Fast anonymization of big data streams.", ACM International Conference Proceeding Series, (2014) 04-07- https://doi.org/10.1145/2640087.2644187

[61] Tekli, J., al Bouna, B., Bou Issa, Y., Kamradt, M., & Haraty, R.," (K, l)-clustering for transactional data streams anonymization", Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11125 LNCS, (2018) 544–556. https://doi.org/10.1007/978-3-319-99807-7_35

[62] Dagadu Puri, G.,"A Novel Method for Privacy Preservation of Health Data Stream", International Journal of Advanced Trends in Computer Science and Engineering, 9(4), (2020) 4959–4963

[63] Sopaoglu, U., & Abul, O.,"A utility based approach for data stream anonymization", Journal of Intelligent Information Systems, 54(3), (2020) 605–631. https://doi.org/10.1007/s10844-019-00577-6

[64] Sopaoglu, U., & Abul, O.,"Classification utility aware data stream anonymization",Applied Soft Computing, 110, (2021) 107743. doi:10.1016/j.asoc.2021. 107743

[65] Yang, L., Chen, X., Luo, Y., Lan, X., & Wang, W.,"IDEA: A Utility-Enhanced Approach to Incomplete Data Stream Anonymization",TSINGHUA SCIENCE AND TECHNOLOGY, Volume 27, Number 1,(2022) pp.127–140.

[66] A. Otgonbayar, Z. Pervez, and K. Dahal,, " Toward anonymizing IoT data streams via partitioning", in Proc. of 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems, Brasilia, Brazil, (2016) pp. 331–336.

[67] A. Otgonbayar, Z. Pervez, K. P. Dahal, and S. Eager., " KVARP: k-anonymity for varied data streams via partitioning",Inf. Sci., vol. 467, (2018) pp. 238–255.

[68] Abbasi, A., & Mohammadi, B., "A clustering-based anonymization approach for privacy-preserving in the healthcare cloud",Concurrency and Computation: Practice and Experience, 34(1) (2022). https://doi.org/10.1002/ cpe. 6487

[69] Usha Lawrance, J., & Nayahi Jesudhasan, J. V., "Privacy Preserving Parallel Clustering Based Anonymization for Big Data Using MapReduce Framework", Applied Artificial Intelligence, 35(15), (2021) 1587–1620.

[70] Arava, K., & Lingamgunta, S."Adaptive k-Anonymity Approach for Privacy Preserving in Cloud". Arabian Journal for Science and Engineering, 45(4), (2020) 2425–2432.

[71] Nayahi, J. J. V., & Kavitha, V., "Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop", Future Generation Computer Systems, 74, (2017) 393–408.

[72] Basapur, S. B., Shylaja, B. S., & Venkatesh., "S-CPM: Semantic-Similarity Cluster based Privacy Preservation Model with Cell Generalization Principle", Journal of Computer Science, 18(3), (2022) 138–150

[73] Ashkouti, F., Khamforoosh, K., Sheikhahmadi, A., Khamfroush,H.,"DHkmeans-ℓdiversity: distributed hierarchical K-means for satisfaction of the ℓ-diversity privacy model using Apache Spark.", Journal of Supercomputing, 78(2), (2022) 2616–2650. https://doi.org/ 10.1007/ s11227 -021-03958-3

[74] Kanade, D. M. ., & Sane, S. S.."Evaluating the Effectiveness of Clustering-Based K-Anonymity and KNN Cluster for Privacy Preservation.", International Journal of Intelligent Systems and Applications in Engineering, 11(11s), (2023) 85–93.

[75] Patil, A., & Wang, B.," Advancing Data Privacy: A Novel K-Anonymity Algorithm with Dissimilarity Tree-Based Clustering and Minimal Information Loss", International Journal on Recent and Innovation Trends in Computing and Communication, 11(8), (2023) 323–330. https://doi.org/10.17762/ijritcc.v11i8.8005

[76] Su, B., Huang, J., Miao, K., Wang, Z., Zhang, X., & Chen, Y.," K-Anonymity Privacy Protection Algorithm for Multi-Dimensional Data against Skewness and Similarity Attacks", Sensors, 23(3) (2023) . https://doi.org/ 10.3390/s23031554

[77] Nayahi, J. J. V., & Kavitha, V., " An Efficient Clustering for Anonymizing Data and Protecting Sensitive Labels.", International Journal of Uncertainty, Fuzziness and Knowldege-Based Systems, 23(5), (2015) 685–714. https://doi.org/10.1142/S0218488515500300

[78] Piao, C., Liu, L., Shi, Y., Jiang, X., & Song, N."Clustering-based privacy preserving anonymity approach for table data sharing.", International Journal of System Assurance Engineering and Management, 11(4),(2019) 768–773. https://doi.org/10.1007/s13198-019-00834-5

[79] Thaeter, F.; Reischuk, R., " Scalable k-anonymous microaggregation: Exploiting the tradeoff between computational complexity and information loss", In Proceedings of the 18th International Conference on Security and Cryptography (SECRYPT), Setubal, Portugal, 6–8 July 2021; Springer: Berlin, Germany, (2021) pp. 87–98

[80] Wang, R.; Zhu, Y.; Chen, T.; Chang, C., "Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness.", J. omput. Sci. Technol. 2018, 33, (2018) 1231–1242.

[81] Ashish Ranjan, Prabhat Ranjan , "Two-phase Entropy based approach to Big Data Anonymization", Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016) (2016) 29-30

[82] Majeed, A., & Lee, S., "Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data", Applied Intelligence, 50(8), (2020) 2555–2574. https://doi.org/10.1007/s10489-020-01656

[83] Eyupoglu, C., Aydin, M. A., Zaim, A. H., & Sertbas, A., "An efficient big data anonymization algorithm based on chaos and perturbation techniques.", Entropy, 20(5) (2018).

[84] Pranamik, M.I.; Lau, R.Y.K.; Zhang, W., " k-anonymity through the enhanced clustering method ", In: IEEE International Conference for e-Business Engineering, (2016)