Journal of Computer Science

Original Research Paper

S-CPM: Semantic-Similarity Cluster based Privacy Preservation Model with Cell Generalization Principle

[1]Satish B Basapur, [1]B S Shylaja and [2]Venkatesh

[1]*Department of Information Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru-560056, India*
[2]*Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, India*

Article history
Received: 05-09-2021
Revised: 16-02-2022
Accepted: 22-02-2022

Corresponding Author:
Satish B Basapur
Department of Information Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru-560056, India
Email: satish.basapur@gmail.com

Abstract: Timely data analysis on a wide variety and a large volume of data unveil valuable information or new insights. The analysis results could be used to innovate new avenues in health care service, business and e-service, etc. However, releasing, storing and reusing sensitive data to third parties results in breaching the data privacy of the individual. To combat privacy breach invasion, privacy-preserving techniques such as suppression, generalization and encryption-based privacy models have been proposed in the literature. The widely used privacy preservation model k-anonymity model prevents record-linkage invasions but fails to satisfy monotonicity property. It has more data distortion and fails to defend semantic-similarity, closeness, nearest-neighborhood data privacy breaches. Moreover, existing approaches are not scalable for the large-scale data set. The paper proposes a semantic similarity two-phase cluster based privacy preservation model. The proposed model considers both numerical and categorical attribute values for data anonymization. Two-phase clustering contains two phases. In the first phase, the t-centroid clustering algorithm is designed and used to partition a set of transaction records of data set D into a set of t-centroids based on the Euclidean distance between transaction records. In the second phase, the neighborhood-aware hierarchical clustering algorithm is designed. It is used to split a set of transaction records within clusters based on neighborhood aware attribute values. Two-phase clustering operations are carried out in parallel and scalable for Big Data sets. The proposed privacy model relies on cell generalization to combat records linkage and semantic-similarity, closeness, nearest-neighborhood privacy breach invasion. All experiments are carried out on two different datasets: Income-Census (KDD) and Bank Credit Card dataset. The experimental results demonstrate that the proposed privacy model can combat privacy breach invasion with cell generalization principles. The proposed privacy model is scalable and time efficient for large-scale data sets.

Keywords: Privacy Preservation Model, Cell Generalization, Transaction Records, Clusters, Quasi-Identifiers and Sensitive Attributes

Introduction

The rapid technological development in information, communication and proliferation of mobile devices enabled millions of users to use social networks, sensors surveillance systems, IoT-enabled healthcare applications, e-Learning and e-Commerce applications for various purposes (Lv *et al*., 2017; Ang *et al*., 2020 and Zheng *et al*., 2020). All these applications are a source of data deluge in different formats (i.e., text, audio, video, image, etc.) (Tsui *et al*., 2019 and D'Alconzo *et al*., 2019). The data with different formats are generated at a higher speed and it is referred to as Big Data. Big Data is characterized by volume, velocity, variety and traditional methods are not appropriate to handle data that explode at an exponential rate (Manyika *et al*., 2011). Moreover, the data generated from different sources could be unstructured, semi-structured, or structured and make it difficult to process, store and maintain privacy (L'heureux *et al*., 2017). The systematic and time-bound analysis of Big Data gives actionable and profitable insights; these insights could be more useful in enhancing

Science Publications

business, defining new strategies, take profitable management decisions (Liang *et al*., 2018). The research challenges and issues in the systemic analysis of Big Data have attracted research and the scientific community. In recent years, Big Data analytics in the cloud environment privacy preservation has been a hot research topic.

Despite the difficulty in storing and processing Big Data, Big Data could be effectively utilized to understand the trends of users on social networks, trends in business, proliferate new research solutions to these complex problems. With the great potential of Big Data, it is easy to gather store user personal information. However, commercial social network platforms have started sharing user personal information with the purpose of profit. The reuse/misuse of User personal information by social network platforms is a violation of personal data privacy and a breach of data integrity. For example, most common social network platforms such as Amazon, Flipkart, e-bay perform analytics on user data to extract user shopping frequency, pattern, priorities, likes and dislikes. Social media like Facebook do extensive analytics on user habits, social status, social relationships, list out family members, friends, colleagues and store user personal data. YouTube suggests the videos to the user based on the user search track on the browser. Under various circumstances, social network platforms breach the user's privacy (Liu Zhang, 2020; Mehmood *et al.*, 2016; Zhou *et al*., 2019; and Bhaskar and Shylaja, 2021):

- To find the user preferences over product or services, business companies retrieve user personal information from social networks platforms
- Secretive personal information is stored in a public database and new inference from the public database may reveal confidential information of the user to others
- Storing and processing of user personal information in an unprofessionally and unsecured manner may result in the disclosure of the user's data privacy

To preserve the privacy of user data, extensive research work has been going in recent years. A well-known and widely accepted approach has been presented to protect Big Data privacy or hide private personal information, while some agglomerated data are open for data analysis purposes. The existing privacy models are either not scalable or inefficient due to velocity, volume and variety of data (Li *et al*., 2009; Aggarwal *et al*., 2010; Fung *et al*., 2010). Moreover, privacy models introduce noise and falsify data to protect the privacy of data. The existing privacy models cannot withstand record-linkage, sensitive attribute attacks, data distortion and are unable to maintain monotonicity properties. Therefore, designing a privacy model that preserves privacy with low distortion and combat sensitivity attribute and linkage attribute attack is a challenging and open research problem for largescale datasets. This study proposes a privacy protection model that combat privacy breach attacks, data distortion and satisfying monotonicity property. The paper proposes a two phase cluster-based privacy model that minimizes both data distortion and privacy breach attacks; the paper considers both numerical and categorical attribute values for data anonymization. Two-phase clustering contains two phases; in the first phase, *t-centroid clustering algorithm* is designed and used to partition a set of transaction records of data set D into a set of t-centroids based on Euclidean distance (i.e., the similarity between quasi-identifiers of transaction records) Between two transactions records. In the second phase, neighborhood aware hierarchical clustering algorithm is designed and used to split a set of transaction records within clusters based on neighborhood-aware attributes values (i.e., the similarity between categorical and sensitive attribute values). Two-phase clustering operations are executed in parallel and are scalable for Big data set.

Motivation and Background

Existing privacy models can thwart attacks such as record linkage (i.e., linking nameless transaction records to external non-traceable records) but fail to prevent attacks such as uniformity, asymmetry and closeness attributes linkage. The existing approach violates data privacy by allowing an intruder who has an awareness of the domain to link records that contain a set of similar sensitive values to external non traceable records. However, there are plenty of models that have the ability to tackle the attacks mentioned above but not enough privacy models that adopted cell generalization for data anonymization and thwart record and attribute linkage attacks.

Existing Incognito model computationally slow to handle large scale data set. Existing privacy model based on *k*-anonymity fails to combat attacks when the data set is split in clustering pattern and fails to satisfy monotonicity property. Designing a privacy model through cell generalization thwart record and attribute attacks is a challenging task. This study leverage Apache Spark to address scalability issue while data anonymization using a clustering approach.

Important Contributions of Proposed Research Work

- Unlike *k*-anonymity-based privacy model, the proposed semantic similarity cluster-based privacy preservation model considers multiple sensitive and semantic similarities between categorical values
- Cell generalization principal-based privacy preservation model combat semantic similarity breaches
- Two-phase cluster-based privacy preservation model is presented to parallelized cell generalization
- Parallel jobs execution at different job nodes in Apache spark cloud environment to perform parallel computations to cope with large scale data set

Literature Survey

In this section, rigorous reviews on existing and well known approaches have been carried out. The global, full domain, sub-tree, multidimensional and local-recording for data anonymization have been studied extensively. The privacy models that thwart privacy breaches and linkage attributes are surveyed.

The most common and widely accepted approaches accomplish data privacy preservation through data anonymization. In data anonymization, personal and sensitive data are hidden. But, aggregated data could be disclosed for mining and analysis purposes. The extensive literature survey is carried out on data privacy models and studies their merits demerits (Fung *et al*., 2010; Li *et al*., 2009; Aggarwal *et al*., 2010). The privacy models found in the literature are not efficient for Big Data anonymization due to volume, velocity, the veracity of Big Data. Moreover, privacy models are not scalable. Authors in (Zhang *et al*., 2013) proposed Map Reduce based approach for data anonymization. The proposed approach is scalable and useful to multidimensional or subtree data anonymization with the scalability issues that have been addressed successfully.

From the point of the individual user, storing individual information in a cloud has a lot of benefits: No worry of storage management, anytime, anywhere data accessibility and no capital investment on the storage device, etc. But, data on the cloud relinquishes user control over their personal information. Third-party data auditors perform operation and maintenance of user actual data instead of data owner; therefore, user personal information can be disclosed (Bhagyashri and Gurav, 2014). The approaches proposed in (Zhang *et al*., 2013a; Bhagyashri and Gurav, 2014) have not been tried on a commercial, public cloud such as Amazon cloud service. The privacy preservation mechanism adopts top-down generalization. Moreover, the algorithms are not robust and do not cope with a very large-scale data set. Authors in (Sun *et al*., 2020) proposed authentication and verification algorithms that minimize the discloser of user personal information. The authentication, verification and encryption techniques cannot ensure privacy preservation. The approach suggested in (Sun *et al*., 2020) was utilized to authenticate a stream of Big Data and check data privacy leakage while the data was being audited by a third party. The presence of non-zero entries in the privacy matrix indicates the existence of nodes and edges between nodes. The attacker retrieves information on a number of nodes and edges from a social graph and introduces a privacy attack (Zhou *et al*., 2008).

The techniques in literature (Sharma *et al*., 2018) preserve the privacy of data by disguising data by the data owner. To disguise data, the owner of data uses either Additive or somewhat Homomorphic Encryption (AHE or SHE) and announces the key. Data owners add fake nodes with indistinguishable values for encryption. The matrix with a fake node does not leak any information to adversaries. It is imperative to preserve privacy in email content and detect spam from genuine email. The author in (Kanwal *et al*., 2021) has discussed the relation between privacy models and privacy techniques. It emphasizes the trade-off between the data privacy-data utility. The relevant privacy techniques can be adapted to achieve data privacy in HER (Kanwal *et al*., 2021). In recent years, researchers have extensively studied Discrete Wavelet Transform (DWT) and rule-based association mining to safeguard the privacy of data. The experimental results and theoretical analysis proved that association rules are exceptional in maintaining data privacy. The data privacy preservation techniques have been proposed to provide mechanisms for maintaining data privacy while publishing or mining valuable data. The methodologies discussed in the survey paper are aimed towards either a multidimensional or a sub-tree scheme. The survey paper gives an insight on privacy preservation models prevent assaults on attribute linkage. These models handle categorical attributes only and fail to thwart privacy breaches in numerical sensitive attributes (Wang *et al*., 2018).

Quite a large number of research papers on privacy preservation through association rule mining are found in the literature. The rule-based association mining is applied for categorical data, Boolean data and association rules for centralized or distributed environments are studied in (Verykios *et al*., 2004; Kantarcioglu and Clifton, 2004; Zhang *et al*., 2013b).

The author has designed proximity-aware local-recording anonymization using the Map Reduce framework. However, the Map Reduce framework is not suitable for privacy preservation for knowledge discovery, data sharing and data analysis (Zhang *et al*., 2014). The proposed work uses the Apache Spark framework to perform an intensive examination on a larger dataset to generalize data and increase data utility. The data distortion after anonymization is more in the existing approach compared to the proposed work. Authors in (Zhang *et al*., 2013a) have proposed a scalable multidimensional anonymization approach for privacy preservation using Map Reduce on the cloud. However, designing suitable Map Reduce jobs for complicated applications is a challenging task. Map Reduce is a constrained software framework. Developing proper Map Reduce jobs for complicated applications remains difficult. Therefore, this research work leverage the Apache-Spark framework to address the scalability issues and issues in the Map Reduce framework. (Mehmood *et al*., 2016) extensively studied various techniques for protecting personalized data at different phases of Big Data to reduce risk of disclosing data privacy by falsifying data (Xu *et al*., 2014). The research issues related to storing and processing of Big Data and techniques to alter the data while storing Big Data on the cloud are discussed in (Fung *et al*., 2010; Li *et al*., 2008).

Existing data privacy protection models are built on the *k*-anonymity principle. These models take categorical values into account; however, they don't account for privacy breaches in numerical values of sensitive variables. Moreover, data skewness and scalability issues of data anonymization for big data set are not addressed. By assessing extra privacy satisfiability during each phase of the top-down anonymization process, *k*-anonymity-based approaches and their extensions models can prevent attribute linkage assaults. However, these approaches incur data distortion. The differential privacy model fails to satisfy the monotonicity property. This study considers the cell generalization approach for data anonymization problems with limited available memory (i.e., executing on Apache Spark). The proposed method in this study splits large-scale data set into different clusters based on the similarity of quasi identifiers and similar numerical and categorical values are mapped to different anonymized values in each cluster. The existing approaches are based on global recording and single dimensional (i.e., same generalized rule and only categorical values).

Preliminaries and Problem Definition

Definition: Privacy

In the framework of personal data, data privacy means ensuring confidentiality and integrity of data. For example, user A should not know user B's age, salary, account number, etc. If user A is adequate enough to disclose B's personal information, then data integrity and confidentiality is breached and user B's data privacy is at risk. In this study, user data are private and sensitive.

Definition: Attributes

The attributes of each transaction record in the dataset are Identifier, Quasi-Identifier, Sensitive and Non-Sensitive:

- Identifier attributes are unique and shall be used to distinguish a record from other records in the dataset. For example: Driving license number, mobile number
- Quasi-Identifier attributes shall not identify a record in the dataset but it can be used to identify if it is linked with other external records. The identifier attribute value is removed and quasi-identifier attributes are used during data anonymization
- *Sensitive* attributes are private, contain sensitive information. The sensitive attribute values to be concealed. For example, disease, ATM pin number, passwords etc. The sensitive attributes are used extensively for data analytics or data mining but not for anonymization
- Non-sensitive attribute value can be disclosed and no need to protect data privacy

In this study, both categorical and numerical values of the quasi-identifier attribute are considered. The set of quasi identifier attributes are called quasi-group and denoted as *QuasiIG* and identified by *QuasiI*. Transaction records of data set D are represented as *points* in multidimensional space.

Word monotonicity refers to quantity that never decreases or increases (Li *et al*., 2008). The monotonicity property in privacy model refers to data set. Let us consider two disjoint data subset D_1, D_2 of data set D. If data subset D_1, D_2 satisfies constrains of privacy model then agglomeration of subset D_1, D_2 also satisfies constrains of privacy model.

Problem Definition

Given large scale dataset D, partition data set D into group of clusters $C = \{C_1, C_2..,C_m\}$, where cluster size at least k and cluster contain k transaction records r such that $C=\{r_1,r_2..r_k\}$. Privacy model must prevent transaction record linkage, semantic similarity attributes privacy breach attacks. The privacy model must converge at faster rate and must have minimum data distortion. Formally, problem is formulated as maximization problem as follows.

Maximize $r_x \neq r_{y\, \neq \epsilon QuasiIG}$ min (Sim(r_x, r_y)), Subject to:

1. Minimize(Los(QuasiIG))
2. $\bigcup_{i=1..m} C_i = D and\, C_i \bigcap C_j = 0, i \neq j$
3. $\forall$Ci ϵ $\mathbb{C}$ and | $\mathbb{C}$ |$\geq$k

Semantic Similarity-Aware Privacy Model

In this study, multiple categorical or numerical sensitive attributes are considered in the privacy model. The numerical sensitive attributes have more sense of closeness, similarity and neighborhood semantics than categorical sensitive attributes. The existing privacy model examines only categorical attributes to check exact similarity or dissimilarity between records. Identification of closeness similarity semantics between numerical sensitive values is more important to prevent data privacy breach attacks. The similarity semantic between categorical values must be considered to prevent privacy breach attacks. It is possible to find similarities between categorical values based on domain knowledge. The privacy model considers closeness similarity semantics for numerical and categorical attribute values. The dissimilarity and similarity between two transaction records is determined using dissimilarity and similarity between two numerical sensitive attributes and two categorical sensitive attributes. The distance between two numerical sensitive attributes is given in Eq. 1:

$$Dst_N = \left(sen_{ual},\ sen'_{ual}\right) = \left|sen_{ual} - sen'_{ual}\right| / Dom \qquad (1)$$

The variable *Dom* represent domain of attribute with maximum and minimum values. Similarly, distance between categorical sensitive values is computed using Eq. 2:

$$Dst_{Cat} = \left(sen_{val}, sen'_{val}\right) = len\left(sen_{val}, sen'_{val}\right) / 2.H\left(Tr\right) \qquad (2)$$

The distance between two categorical values is computed using the length of the path between two categorical values and the height of the hierarchy tree. All leaf nodes in the hierarchy tree have the same depth (i.e., the similar transaction records should be at the same level of the hierarchy tree and be part of the same cluster class). The maximum length of the path between any two nodes in the hierarchy tree is 2× H (Tr). For instance, a simple hierarchy tree for the Quasi-Attribute *Disease*. The hierarchy tree is accessible to the public and classifies attributes of diseases into tree leaves. The diseases in its sub-tree are described by the name of an intermediate node.

For multiple sensitive values of transaction record *r*, the distance between multiple sensitive values of two transaction records is given by Eq 3. Let us consider the transaction record r_x= {sen_{val1},...., sen_{valm}} and transaction record r_y={sen'_{val1},...,sen'_{valn} }. The weight w_N and w_C represents priority to be assigned to the numerical, categorical attributes and satisfy condition $0 \leq w_N \leq 1$, $0 \leq w_C \leq 1$.

$$d\left(r_x, r_y\right) = \sum_{i=1}^{m} w_N.Dst_N + \sum_{i=m+1}^{m+n} w_C.Dst_{Cat} \qquad (3)$$

It is essential to find the similarity or closeness between two transaction records containing multiple sensitive values. It is ascertained that similar transaction records belong to the same cluster or group. To prevent privacy breach attacks, it is also necessary to hold the condition that two transaction records belonging to the same cluster must be dissimilar in an anonymous data set. The sufficient and necessary condition to prevent privacy attacks is holding at least one transaction record containing multiple sensitive values in a cluster that must be dissimilar with other transaction records of the cluster. The minimum size of the cluster to be *k*. The sufficient and necessary condition to prevent privacy attacks is given in Eq. 4. Equation 4 gives the similarity index between two transaction records containing multiple sensitive values:

$$Sim(r_x, r_y) = \sum_{i=1}^{m} w_N.Dst_N + \sum_{i=m+1}^{m+1} w_C.Dst^*_{cat} \qquad (4)$$

The variable Dst^*_N and Dst^*_{Cat} represent similarity index for numerical and categorical values respectively.

Any quasi-group (*QuasiIG*) which is identified by *QuasiIG* must contain at least one transaction record that must be dissimilar with other transaction of cluster or group. Semantic similarity between two transactions records within a cluster is given by Eq. 5:

$$Sim\left(QuasiIG\right) = \max_{r_x \neq r_y \in QuasiIG} \min(Sim\left(r_x, r_y\right) \qquad (5)$$

A privacy model must ensure minimum data distortion, because it should be possible to extract exact information even after data anonymization. The data distortion is also called information loss. For each transaction record, the information loss is given Eq. 6:

$$Los(r) = \sum_{i=1}^{m} \frac{ual^{\max} - ual^{\min}}{Dom} + \sum_{i=m+1}^{m+n} \frac{len\left(ual, ual^{CA}\right)}{H(Tr)} \qquad (6)$$

The variable val^{max}, val^{min} represent maximum and minimum value of attributes. *len* (*ual*, ual^{CA}) is the path length between sen_{val} and sen'_{val}. Information loss for quasi-group *QuasiIG* is given by Eq. 7:

$$Los\left(QuasiIG\right) = \sum_{r \in QuasiIG} Los(r) \qquad (7)$$

Similarity or closeness among transaction records of the cluster is accomplished through clustering. The cluster-based cell generalization converges for large-scale data set and satisfy the monotonicity property.

Methodology

Cell Generalization Anonymization

Cell generalization is commonly known as local-recording, is one of the schemes mentioned in (Terrovitis *et al*., 2011). Authors have outlined other schemes: Sub-tree, optimal anonymization, full-domain and multidimensional anonymization. All other schemes mentioned are global recording schemes. Cell generalization relies on generalizing values at the local or cell level, but global recoding generalizes all or none of the values of the transaction at the domain level. Cell generalization minimizes data distortion that occurs during data anonymization. Cell generalization anonymizes data by replacement only in a neighborhood of data. Finding the most appropriate neighborhood data with similar values is a challenging task in high-dimensional data. Cell generalization defines a group of functions to process a set of attributes in an overlapping transaction. Overlapping transaction means that set of transactions may contain similar quasi-identifiers. Typically, function *φ*: $T_i \longrightarrow QI_D$ where T_i is transaction indexed by *i*.

B. Semantic Similarity-aware Cluster-based Privacy model.

To do data anonymization for large-scale data set, two-phase clusters are constructed for given datasets. Some key observations while selecting clustering method to construct cluster. Firstly, the value of *k* in *k*-anonymity privacy preservation model is meager compared to the size of the data set for Big Data applications. For cell generalization anonymization, the ceiling limit on cluster size is 2*k*-1; correspondingly, there would be quite a large number of clusters with small size clusters. Secondly, a cluster with a size not less than *k* is ideal and preferred because it leads to less data distortion. Thirdly, data set with clustering architecture is more suitable for cell generalization anonymization. With key points mentioned above and with known benefits of H: Hierarchical and *P: point assignment*. The proposed two-phase H-P clustering method blends the merits of hierarchical and point assignment for cell generalization anonymization. The proposed two-phase H-P clustering method is most appropriate for cell generalization anonymization because the stopping condition is applied when cluster size reaches 2*k*-1 and accomplishes less distortion while merging two clusters. Moreover, clustering in the two phase H-P method is performed in parallel and scalable for large-scale data set. The two-phase H-P clustering algorithm has two phases. In the first phase, the t-centroid algorithm is executed to get a set of clusters. In the second phase, the neighborhood-aware hierarchical clustering algorithm is executed by considering two linkage distances as criteria to merge two clusters. Linkage distance means that two clusters' distance is the same as the weighted distance between two transaction records belonging to two different clusters and these two transaction records are at most distant from each other. When a cluster with a size less than *k* is unmerged, then the transaction records are assigned to cluster to the nearest cluster whose size is less than 2*k*-1.

The two-phase H-P clustering algorithm begins by representing each transaction record of data set as t-centroids. Generally, a t-centroid is at the center of a cluster and the attribute value of the categorical quasi-identifier is lowest among other original values in the cluster. Additionally, the numeral quasi identifier of the centroid is the median of original values. Transaction records of the data set consist of quasi-identifier attributes and sensitive attributes. The transaction records are considered to form a cluster and ensure the privacy of sensitive values at all stages of clustering. The new transaction records are assigned to the cluster based on the Euclidean distance between a new instance of the transaction record and the t-centroid of the cluster. A new transaction record is then assigned to a cluster having a minimum distance to the centroid. The distance between quasi-identifier records is given by Eq. 3.

Typically, t-centroid is a randomly selected transaction record that is far away from other records. Concretely, the selection of t-centroid is achieved via the t-centroid function defined in algorithm 2. In this algorithm, the first t-centroid record is chosen at random and then repeatedly chosen next transaction record whose minimum Euclidian distance is maximum to the existing t-centroid of the cluster. Algorithm 2 terminates when the size of the t-centroid cluster reaches *2k-1*.

Each iteration of the t-centroid clustering algorithm consists of two steps: Formation (F) and Shifting (S). In the formation step, a transaction record is attached to their nearby t-centroid and constitutes β-cluster. In the shifting step, the t-centroid is shifted or recomputed accordingly transaction record attributes and the new t-centroid is used in the next round of formation step. Iteration continues till it converges and stops when the t-centroid no longer changes. In this study, a widely accepted stopping condition is used to stop cluster formation. In the first condition, the difference of two t-centroid positions in two successive rounds of iteration must be smaller than the threshold value (i.e., the median value of categorical and numerical attributes is computed). The second condition, the number of iteration rounds, reaches the predefined number. Let S^i and S^{i+1} denotes the two t-centroid in i^{th} and $(i+1)^{th}$ round respectively. The difference of two t-centroid position in two successive rounds of iteration is given in Eq 8. Generally, this difference is represented as the average distance between transaction records.

$$d\left(S^i, S^{i+1}\right) = \left(\sum\nolimits_{j=1}^{t} d\left(r_j^{i}, r_j^{i+1}\right) / t\right) \tag{8}$$

The difference of two t-centroid positions in two successive rounds of iteration must be less than τ, where parameter τ is the threshold value. The parameter θ represents the highest iteration rounds. The t-centroid clustering algorithm stops execution when either of the stopping condition occurs.

Similarity-Semantics-Aware Hierarchical Clustering Algorithm Phase-II

In a neighborhood-aware clustering algorithm, initially, a transaction-record is considered as a cluster. In each round of iteration of the algorithm, two clusters are chosen and merged till the terminating condition is satisfied. In this study, the linkage distance is used as the criterion to merge two clusters. Linkage distance means that two clusters' distance is the same as the weighted distance between two transaction records belonging to two different clusters and these two transaction records are at most distant from each other. The merged cluster diameter is equal to the distance between two clusters considered for merging. The distance between cluster C_i and C_{i+1} is given by Eq 9. The cluster is not chosen for merging if the cluster size is greater than or equal to *k*. The process of merging clusters ends when there are no two clusters with a size less than *k*. The size of the merged cluster must be less than or equal to 4*k*-2. In rare cases, if a cluster with a size less than *k* remains unmerged, then the transaction records of the unmerged cluster are assigned to the nearest cluster whose size is less than *2k1*. In the worst case, if all clusters are of size *2k-1* and a cluster remains unmerged, then choose cluster randomly and take out some transaction records and assign to cluster that being unmerged to make its cluster size much better. At the end of Neighborhood-Aware Hierarchical Clustering Algorithm execution, all clusters shall have at least *k* transaction records but less than *4*k-*2* transaction records.

$$d\left(C_x,C_y\right)=\left(\left|C_x\right|+\left|C_y\right|\right)\max_{rx\in}C,ry\in Cd(r_{x,}r_y) \quad (9)$$

Algorithm 1: Two-Phase Clustering

```
Input: Data set D, k-Anonymization, parameter k
Output: Anonymized Data Set D'
Obtain β-clusters $\square^{\beta}=\{C_1^{\beta},...C_t^{\beta}\}$ by Executing t-centroid clustering algorithm on Dataset D
For every β-clusters $Ci\in\square^{\beta}\forall_i 1\le i\le\beta$ Execute Neighborhood-Aware Hierarchical Clustering algorithm on $C^{\beta}_i$ to get group of cluster $Ci=\{C_{i1},..., C_{im}\}$ 0
Each cluster $C_j\in\square,\square=\bigcup_{i=1}^{l}$ generalize $C_j$ to $C_j$ by replacing attributes value with general values
Generate $D'=\bigcup_{j=1}^{m_j}C_j, m_j=\sum_{i=1}^{l}m_i$
```

Algorithm 2: t-centroid algorithm: Clustering phase-I

```
Input: Data set D, Transaction records r, r∈ D, k, threshold value τ, θ set of centroids at round $S^{(i+1)}=\{r_1^{(i+1)},...r_t^{(i+1)}\}$
Output: set of β- clusters $C_i = \{C_{i1},..., C_{im}\}$, set t centrodes $S = \{r_1,...r_t\}$, set of centroids at round $S^{(i+1)}=\{r_1^{(i+1)},...r_t^{(i+1)}\}$.
Generate a rand random value, assign rand← 0 to 1 and rand ≤ |D|
Arbitrarily pick up transaction records from list (transaction records r) and assign S ←r
while (| S | ≤ k) do
    Determine r list(transaction records r) and has Max($\min_{r'\in S}(d(r, r'))$)
    S ←r
    return(S)
end while
while $(d(S^i,S^{i+1})\ge\tau$ and $i\le\theta$ ) do
    initialize $d^{min}\leftarrow\infty$
    for j← 1 to k do
        if $(d(r,r_j)\le d^{min})$ then
            $d^{min}\leftarrow d(r,r_j)$ and assign $j^{min}\leftarrow j$
        end if
        return $(j^{min}, j)$
    end for
    for l←1 to mQuasiI do
        if $\left(att_j^{QuasiI} is\, numerical\right)$ then
            $v_l$=Find Median (set of transaction records
        else
            $v_l$=t-centroid (Set of transaction records
        end if
        $return\left(j,r_j^{(i+1)}=\left(v_1,....,v_{mQuasiI}\right)\right)$
    end for
end while
return β-cluster with t-centroid $S^{(i+1)}s$
```

Results and Performance Analysis

Experimental Setup

It is a tedious task to share data among multiple Map Reduce functions. Data sharing among clusters through disk storage in Map-Reduce is achieved by disk operation and disk scheduling. In this study, all experiments have been conducted on Apache Spark, an open-source framework for Big Data application and it provides driver program to begin execution with the main module, processing nodes for parallel execution of clusters, cluster reformation and similarity index of transaction records and memory abstraction for sharing data set. Apache-spark allows us to create a Resilient Distributed dataset after the spark session.

All experiments were conducted on two data sets: Census Income (KDD) (Terran Lane and Ronny Kohavi) and Bank Credit Card dataset (Dal Pozzolo *et al*., 2015). Census-Income (KDD) dataset contains 1,99,523 records with 40 attributes relevant to employment and has both categorical and numerical values. The data processing is performed on given data set and extracted 88,560 records with 26 attributes from the dataset. The sampled data set contain 20 quasi identifier attribute and 06 sensitive attributes. Both categories and numerical values of Quasi-identifier and sensitive attributes are considered.

Algorithm 3: Neighborhood - Aware Hierarchical Algorithm: Clustering phase-II

```
Input: Data set $\mathbb{C}^{\beta}$, k-Anonymization, parameter k
Output: set of C ={C_1,..., C_n}
Consider record in $C^{\beta}$ as a cluster $\square^0=\{C_1^0,...,C_n^0\}$ and do $\mathbb{C}^0\leftarrow 0$, $i\leftarrow 0$
```

```
For all $C_x^0, C_y^0 \in □^0$, PriQueue ← $\left(C_x^0, C_y^0, d\left(C_x^0, C_y^0\right)\right)$
  and $x \neq y$
while (PriQueue 6 = null) do
    $(C^{'}_x, C^{'}_y, d(C^{'}_x, C^{'}_y))$← PriQueue, Find $C^{'}_z \leftarrow (C'_x$
    $U\ C'_y)$
    $\mathbb{C}(i+1) = \mathbb{C}(i) \setminus (C'_x, C'_y)$
    Discard $\mathbb{C}'_x$ or $\mathbb{C}'_y$ from PriQueue
    if $(|C'_z| \geq k)$ then
        $\mathbb{C} = \mathbb{C} \cup C^{'}_z$
    else
        $\mathbb{C}^{(i+1)} = \mathbb{C}^{(i+1)} \cup C'_z$
        $\forall C' \epsilon \mathbb{C}^{(i+1)}$, Do PriQueue $\leftarrow (C^{'}, C^{'}_z, d(C^{'}, C^{'}_z))$
    end if
end while
if $(|C^{(i+1)}| == 1)$ and $C'' \epsilon \mathbb{C}^{(i+1)}$ then
   $\forall$ r $\epsilon C'''$, Determine cluster $C\ \epsilon$ and $\mathbb{C}|C| \leq 4k-2$,
   min d({r}, C) and C ← C $\cup \{r\}$
end if
```

The credit card dataset consists of transactions performed by the owner of credit cards in the month of September 2013. Transactions made by the credit card holder in two days of September 2013 are considered. The preprocessed and obtained records are 14,200 with 10 attributes from the dataset and there are 8 quasi-identifier attributes and 2 sensitive attributes that include both categorical and numerical values. The proposed Semantic Similarity-aware Cluster-based Privacy model is implemented in Python and executed on the Apache-Spark framework.

Performance Parameters

To evaluate the performance of the proposed algorithm Semantic Similarity-aware Cluster based Privacy model, two performance parameters are used, namely, compatibility to privacy breach attack and data distortion, second parameter is *scalability* of proposed model for large scale data set *D*. The compatibility to privacy breach attack is measured by minimum similarity between two transaction records (*r*) within a cluster *C* and Average distance between two clusters that have similarity in Quasi-identifier group *Quasi IG*. The minimum distance between two transaction records of *Quasi IG* is determined by Eq. 10:

$$QuasiIG^{Min} = \min_{r_x \neq r_y \in QuasiI} Gd\left(r_x, r_y\right) \tag{10}$$

The average distance between two quasi-identifier groups or clusters in dataset *D* is given by Eq. 11:

$$QuasiIG^{Aug} = \frac{2 \times \sum_{r_x \neq r_y \in QuasiI} G \times d\left(r_x, r_y\right)}{|QuasiIG|} \tag{11}$$

A Quasi-identifier group(*QuasiIG*) that satisfy necessary and sufficient condition (i.e., minimum dissimilarity condition) is apprehended by parameter $QuasiIG^{Min}$. For whole data set *D*, necessary and sufficient condition (i.e., the minimum dissimilarity between clusters of group *QuasiIG* is apprehended by the cumulative distribution of QuasiIG^{Min}. An average distance between groups *QuasiIG* is calculated using Eq. 12:

$$QuaisIG^{Avg} = \frac{1}{|QuaisIG|} \sum_{QuasiIG \subseteq D} QuaisIG^{Avg} \tag{12}$$

The metric *I loss* means Information loss that indicates data distortion after data anonymization. The efficiency of the proposed system is verified through execution time for different computing nodes and different data set sizes *D*.

Results and Discussion

All experiments are carried out on two different data set: Census-Income (KDD) (Terran Lane and Ronny Kohavi) and bank credit card dataset (Dal Pozzolo *et al.*, 2015). To test the performance and efficiency of proposed technique *Semantic Similarity-aware cluster based Privacy (S-CPM)* sampled data set D_1 and D_2 are considered for each execution. Sampled data set D_1 with 1000 transaction records and Dataset D_2 with 1000 transaction records are taken into consideration for each iteration of experiments. The weight w_N and w_C represents priority that is being assigned to numerical sensitive or quasi-identifier attributes and categorical numerical sensitive or quasi-identifier attributes respectively. Weight w_N varies from 0 to 1.0 with successive interval of 2. The performance and efficiency of proposed privacy model (i.e., *Semantic Similarity-aware Cluster based Privacy Model* (S-CPM) is compared with Top-Down Specialization Privacy Preservation (TDS-PP) (Zhang *et al.*, 2013a), a data anonymization technique with top-down approach. The cluster-size *k* parameter is set i.e., $k = 50$ for data set D_1 and $k = 10$ for dataset D_2 and parameter t is set to 10 and $\tau = 0.001$. No technical reason for selection of these values. On each iteration proposed S-CPM try to assign dissimilar transaction records to a cluster.

The dissimilarity distance between a set of transaction records belongs *Quasi IG*-group and the average distance between quasi-identifier group or clusters of data set are shown in Fig. 1. The average distance between clusters increases linearly with datasets. A cluster that has minimum *k* records and not more than *2k-1* transaction records are chosen for merging. Two clusters that have maximum similarity between transaction records are merged. While merging clusters, if a cluster with transaction records not less than *k* is remained alone, then select a cluster that has transaction records not more than *2k*-1 and that most similarity is selected for merging clusters. As dissimilarity between clusters increases, then privacy breach attacks can be prevented and produce anonymous data. The proposed privacy model outperforms TDS (Zhang *et al*., 2016) because TDS split the dataset based on domain and the global recording approach cannot combat sensitive attribute privacy breach attacks.

The values of information loss are shown in Fig. 2. Information loss quantifies the occurrence of data distortion during the data anonymization process. Figure 2 illustrates rise in information loss as value of weights (i.e., w_N and w_C) changes from 0 to 1. Ideal value for weights (i.e., w_N and w_C) is 0.6 for optimized information loss. Unlike the proposed approach, TDS a top-down approach, checks attribute-linkage invasion at each iteration. The top-down anonymization process has relatively small information loss. The dividend of dissimilarity between clusters is achieved at the cost of data distortion in the proposed approach because the proposed approach performs data anonymization only to the nearest neighborhood records. Moreover, finding semantic similarity between sensitive attributes of transaction records in the cluster is hard in the large-scale data set. To reduce information loss, select a proper weight value to balance between thwarting privacy breaches and information loss.

Figure 3 and 4 illustrate compatibility against privacy breach attacks for data set D_1 *and* D_2 respectively. The privacy breach attack can be thwarted by finding a group of clusters that have minimum similarity between two transaction records (r). In others words, a cluster must have a maximum number of dissimilar transaction records will have minimum privacy breach invasion. The proposed approach determines a set of similar transaction records based on semantics similarity in sensitive attributes (i.e., it includes both numerical and categorical sensitive attributes). It is observed from Fig. 3 and 4 that the curve moves towards as rising in values of weights (i.e., w_N and w_C), demonstrating that distance of semantic dissimilarity between transaction records in the cluster. When the value of weight (i.e., w_N and w_C) is 0.8 or 1.0, then more numbers of dissimilar transaction records in clusters and it also indicates that semantic-similarity based transaction records are merged in clusters. The process of merging two transaction records continues till the size of cluster size reaches 2k-1. The leftmost curve in Fig. 3 and 4 is TDS and it indicates that two transaction records can be merged when transaction records contain only quasi-identifier attributes. The proposed approach considers both quasi-attributes similarity and nearest neighbor transaction that has semantic similarity. A cluster will have a set of transaction records that contains both quasi-attributes and sensitive attributes. Due to these reasons, the proposed approach can effectively combat the privacy breach invasion.

The scalability and efficiency of the proposed approach are examined by the total execution time taken to execute the proposed algorithm for sampled records of the dataset. The proposed approach performs data anonymization based on cell generalization. Cell generalization splits the dataset and each partitioned dataset is processed locally on processing nodes in the Apache spark cloud environment. Cell generalization permits similar transaction records (i.e., based on semantic similarity, nearest-neighborhood, closeness similarity) mapped to distinguishable generalized values. Figure 5 illustrates execution time for partitioning data set, cluster formation and merging transaction records based on semantic similarity and nearest neighborhood principle. All these operations are scalable and carried out in parallel in Apache Spark environment by controlling cluster size k (cluster size is not less than k and not more than $2k$-1).

Transaction records varies from 10k to 100k. The change in execution time for the proposed approach is not exponential, but it is stable; it is due to conduction of operation in parallel fashion and distribution of data set. Figure 5 shows that the proposed approach is able to handle large-scale data set.

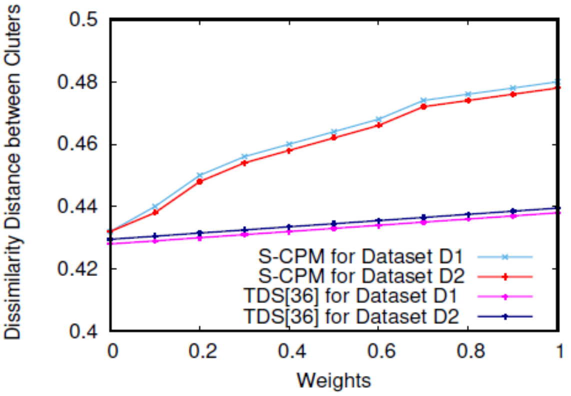

Fig. 1: Average Dissimilar Distance between clusters vs Weights w_N, w_C

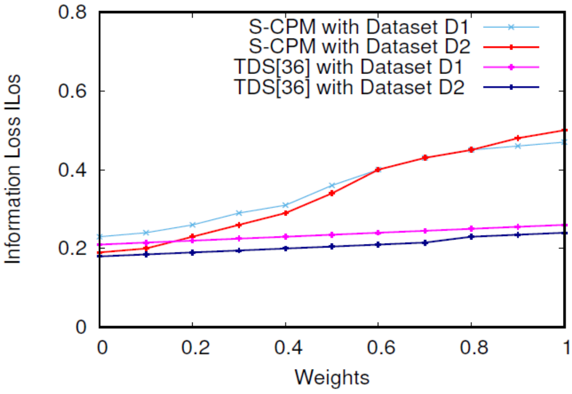

Fig. 2: Information Loss quantify data distortion

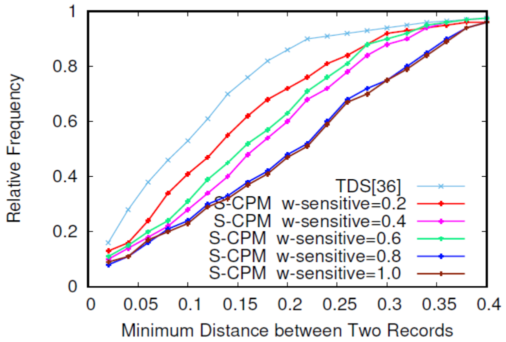

Fig. 3: Cumulative distribution Quasi IG Avg for Dataset D1

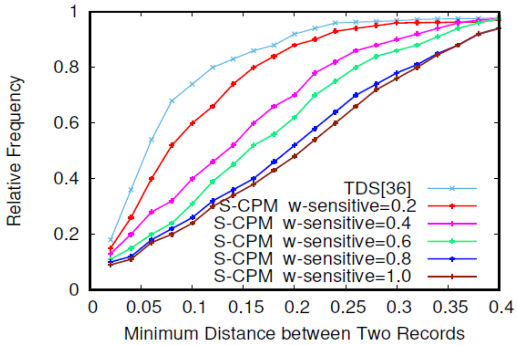

Fig. 4: Cumulative distribution Quasi IG Avg for Dataset D2

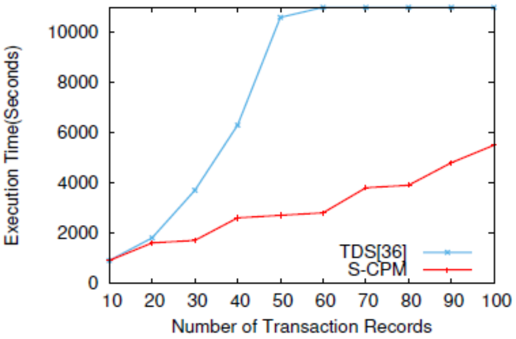

Fig. 5: Execution time(Seconds) vs number of records in dataset D

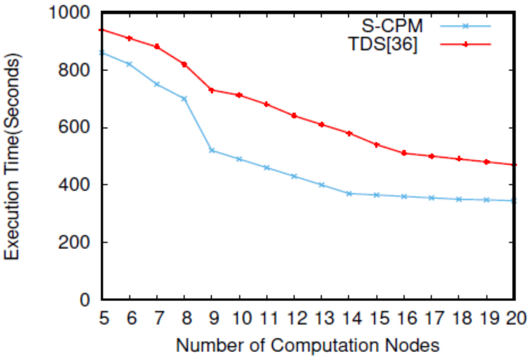

Fig. 6: Required number of computing nodes Vs execution time (seconds)

The number of processing nodes in Apache Spark environment play a vital role in total execution time. It seen from Fig. 6 that as number of computation nodes increases then the execution time required declines linearly. Whenever the number of computing nodes increases, the execution time decreases in a fairly linear manner, as shown in Fig. 6. In terms of scalability, the proposed model is linearly scalable in terms does not affect the execution time in the existing system TDS (Zhang *et al.*, 2013a). Notably, execution time in the proposed model decreases linearly as the computing nodes increase, while execution time in TDS (Zhang *et al.*, 2013b) goes stable. The dramatic decrease in execution time in the proposed model illustrates the ability to handle large-scale datasets. The difference in execution time of the proposed model and TDS (Zhang *et al.*, 2013a) becomes more significant as the computing nodes increase. The execution time difference illustrates that the proposed model is more scalable and efficient than TDS (Zhang *et al.*, 2013b). The time and space complexity of the proposed model is shown in Table 2. In the first phase of point assignment *t* transaction records are selected. The transaction records that are far away from others are chosen. The record selection is accomplished by the worker node of Spark. *N* records are emitted to the worker node to make the model scalable to Big Data. The worker node randomly picks up the first transaction record and then repeatedly picks transaction records whose distance to the already chosen record is the maximum. Each round of cluster phase contains Expectation (E) and Maximization (M) steps. In the expectation step, the cluster manager of Spark assigns transaction records to the (1) the difference among two ancestor transaction records in two successive iterations is less than a threshold; (2) the number of iterations reached the predefined maximum number of iterations. Typically, Worker node is responsible for assigning transaction records to the nearest ancestor transaction record in the E step, while Cluster Manager is responsible for updating the ancestor transaction record in the M step.

Table:1 Symbols and notations

Symbols	Description
D	A dataset consists of Transaction Records
Dst_N	Distance b/w two numerical sensitive attr.
Dst*Cat*	Distance b/w two categorical sensitive attr.
d(rx,ry)	Closeness b/w two trans. records having multiple sen.
Sim(r_x,r_y)	Similarity index b/w trans. records having multiple sen.
Los(r)	Information loss each transaction record.
Los(Quasi IG)	Information loss for quasi-group *Quasi IG*
θ	Highest iteration rounds
τ	Threshold value
C	Cluster
Quasi IG^{Min}	Minimum distance b/w two trans. records of Quasi IG
Quasi IG^{Avg}	average distance b/w two quasi-iden. groups/clusters

Table: 2 Time and space complexity analysis

Tasks/Nodes	Time	Space	Traffic	Rounds
1Worker node	O(1)	O(1)	O(N)	O (1)
Cluster Manager	$O(t^2(N-t))$	O(N)		
2Worker node	O(t)	O(t)	O(n)	O(l)
Cluster Manager	$O(m^{QI}(n/t)^2)$	O(n/t)		
3Worker node	O(1)	O(1)	O(n)	O (1)
Cluster Manager	$O(n/t)^2 \log n/t)$	$O((n/t)^2)$		

Table 2 illustrates the time and space complexity of the proposed model. Specifically, the numbers 1, 2 and 3 denoted transaction record selection, updating ancestor transaction records and forming clustering, respectively. Variable l and P denotes the number of iterations and data split size fed to a cluster manager. Thus, the time and space complexity of worker node and cluster manager has a constant and it depends on N The number of worker nodes linearly increases as the data set size. Thus, the proposed model is scalable with an appropriate value of τ. Table 2 gives insights into the nature of first, second and third jobs. The second job has a maximum number of iterative rounds *O(l)*. The value of l is determined by the stopping condition.

Experiments are not conducted on the Apache Mahout platform, which uses machine learning algorithms for classification, clustering and item-set data mining. It is planned to execute the proposed model on Apache Mahout to achieve higher efficiency and scalable privacy.

The proposed work is based on cell-level generalization. The proposed work generalizes the values of the quasi identifier attributes at the local or cell level and minimizes data distortion. Moreover, the proposed model is single dimensional anonymization techniques. A multidimensional anonymization technique recodes the array of values associated with the vector of quasi-identifier attribute values using the local generalization rule defined. It is a limitation of the proposed model. It is NP-hard to achieve optimal multidimensional anonymization. Based on the model proposed herein, future work is to design a multidimensional anonymization model.

Conclusion

This study proposes a Semantic-aware Cluster-based Privacy Model (S-CPM) that adopts cell generalization for anonymization and to thwart data privacy breach invasion with cell generalization, the proposed privacy model can combat privacy breach invasion, scalable and time-efficient. The proposed model includes multiple numerical, categorical sensitive attributes. This study proposes a scalable two-phase cluster based privacy model to protect privacy breach invasion with cell generalization. The two-phase clustering combines the benefits of the point-assignment and hierarchical clustering approach. In the first phase, this study leverages the point assignments technique to split the dataset and nearest neighbor, or closest records are grouped to form a cluster. In the second phase, quasi-identifiers attribute similarity and semantic-similarity of sensitive values of transaction records are considered to merge clusters. A Series of experiments are conducted to investigate the efficiency and scalability of the proposed approach.

The data set size is large enough to assess the effectiveness of the proposed model. Approximately the size of the cluster is 1000 for different sizes of the dataset. The values of k-anonymity parameter (i.e., $k = 10$), weight of semantic similarity (i.e., $w_C = 0.5$, $w_N = 0.5$), stopping condition (i.e., $\theta = 5$, $\tau = 0.001$) and ten computation nodes make a model to combat privacy breach invasion with cell generalization principles. The proposed privacy model is scalable and time efficient for large-scale data set. Future research explores the adoption of proposed research work for data anonymization through a bottom-up approach. Future plans to investigate scalable and robust data anonymity privacy solutions against adversaries' privacy breach attacks.

Author's Contributions

Satish B Basapur: Problem definition, designing solution, implementation on Apache spark. Interpretation of data and validating the results.

B S Shylaja: Research problem discussion, reviewing manuscript critically for technical content and experiments results. Final version of manuscript approval

Venkatesh: Dataset collection, data preprocessing. Interpretation of data, testing and validating the results.

Ethics

Authors should address any ethical issues that may arise after the publication of this manuscript

References

Aggarwal, G., Panigrahy, R., Feder, T., Thomas, D., Kenthapadi, K., Khuller, S., & Zhu, A. (2010). Achieving anonymity via clustering. ACM Transactions on Algorithms (TALG), 6(3), 1-19. doi.org/10.1145/1798596.1798602

Ang, K. L. M., Ge, F. L., & Seng, K. P. (2020). Big educational data & analytics: Survey, architecture and challenges. IEEE access, 8, 116392-116414. doi.org/10.1109/ACCESS.2020.2994561

Bhagyashri, S., & Gurav, Y. B. (2014). Privacy-preserving public auditing for secure cloud storage. IOSR Journal of Computer Engineering (IOSR-JCE), 16(4), 33-38. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.933.623&rep=rep1&type=pdf

Bhaskar, R., & Shylaja, B. S. (2021). Dynamic Virtual Machine Provisioning in Cloud Computing Using Knowledge-Based Reduction Method. In Next Generation Information Processing System (pp. 193-202). Springer, Singapore. doi.org/10.1016/j.suscom.2018.01.002

D'Alconzo, A., Drago, I., Morichetta, A., Mellia, M., & Casas, P. (2019). A survey on Big Data for network traffic monitoring and analysis. IEEE Transactions on Network and Service Management, 16(3), 800-813. doi.org/10.1109/TNSM.2019.2933358

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence (pp. 159-166). IEEE. doi.org/10.1109/SSCI.2015.33

Fung, B. C., Wang Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (Csur), 42(4), 1-53. doi.org/10.1145/1749603.1749605

Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE transactions on knowledge and data engineering, 16(9), 1026-1037 doi.org/10.1109/TKDE.2004.45

Kanwal, T., Anjum, A., & Khan, A. (2021). Privacy preservation in e-health cloud: Taxonomy, privacy requirements, feasibility analysis and opportunities. Cluster Computing, 24(1), 293-317. doi.org/10.1007/s10586-020-03106-1

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with Big Data: Challenges and approaches. Ieee Access, 5, 7776-7797. doi.org/ 10.1109/ACCESS.2017.2696365

Li, J., Tao, Y., & Xiao, X. (2008, June). Preservation of proximity privacy in publishing numerical sensitive data. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 473-486). doi.org/10.1145/1376616.1376666

Li, N., Li, T., & Venkatasubramanian, S. (2009). Closeness: A new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering, 22(7), 943-956. doi.org/0.1109/TKDE.2009.139

Liang, F., Yu, W., An, D., Yang, Q., Fu, X., & Zhao, W. (2018). A survey on Big Data market: Pricing, trading and protection. Ieee Access, 6, 15132-15154 doi.org/10.1109/ACCESS.2018.2806881

Liu, Z., & Zhang, A. (2020). Sampling for Big Data profiling: A survey. IEEE Access, 8, 72713-72726. doi.org/10.1109/ACCESS.2020.2988120

Lv, Z., Song, H., Basanta-Val, P., Steed, A., & Jo, M. (2017). Next-generation Big Data analytics: State of the art, challenges and future research topics. IEEE Transactions on Industrial Informatics, 13(4), 1891-1899. doi.org/10.1109/TII.2017.2650204

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition and productivity. McKinsey Global Institute. https://catalog.lib.kyushu-u.ac.jp/opac_detail_md/?lang=0&amode=MD824&bibid=3144682

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. IEEE access, 4, 1821-1834. doi.org/10.1109/ACCESS.2016.2558446

Wong, R. C. W., & Fu, A. W.-C. (2010). 'Privacy-Preserving Data Publishing: An Overview', *Synthesis Lectures on Data Management,* ,vol. 2, no. 1, pp. 1-38,2010.

Sharma, S., Powers, J., & Chen, K. (2018). Private Graph: Privacy-preserving spectral analysis of encrypted graphs in the cloud. IEEE Transactions on Knowledge and Data Engineering, 31(5), 981-995. doi.org/10.1109/TKDE.2018.2847662

Sun, Y., Liu, Q., Chen, X., & Du, X. (2020). An adaptive authenticated data structure with privacy-preserving for big data stream in cloud. IEEE Transactions on Information Forensics and Security, 15, 3295-3310. doi.org/10.1109/TIFS.2020.2986879

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2011). Local and global recoding methods for anonymizing set-valued data. The VLDB Journal, 20(1), 83-106. doi.org/10.1007/s00778-010-0192-8

Tsui, K. L., Zhao, Y., & Wang, D. (2019). Big data opportunities: System health monitoring and management. IEEE Access, 7, 68853-68867. doi.org/10.1109/ACCESS.2019.2917891

Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y., & Dasseni, E. (2004). Association rule hiding. IEEE Transactions on knowledge and data engineering, 16(4), 434-447. doi.org/10.1109/TKDE.2004.1269668

Wang, T., Zheng, Z., Rehmani, M. H., Yao, S., & Huo, Z. (2018). Privacy preservation in Big Data from the communication perspective-A survey. IEEE Communications Surveys & Tutorials, 21(1),753-778. doi.org/10.1109/COMST.2018.2865107

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. Ieee Access, 2, 1149-1176. doi.org/10.1109/ACCESS.2014.2362522

Zhang, F., Rong, C., Zhao, G., Wu, J., & Wu, X. (2013a, December). Privacy-preserving two-party distributed association rules mining on horizontally partitioned data. In 2013 International Conference on Cloud Computing and Big Data (pp. 633-640). IEEE. doi.org/10.1109/CLOUDCOMASIA.2013.87

Zhang, X., Yang, C., Nepal, S., Liu, C., Dou, W., & Chen, J. (2013b, September). A MapReduce based approach of scalable multidimensional anonymization for Big Data privacy preservation on cloud. In 2013 International conference on cloud and green computing (pp. 105-112). IEEE. doi.org/10.1109/CGC.2013.24

Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2013c). A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. IEEE Transactions on Parallel and Distributed Systems, 25(2), 363-373. doi.org/10.1109/TPDS.2013.48

Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., & Chen, J. (2014). Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. IEEE transactions on computers, 64(8), 2293-2307. doi.org/10.1109/TC.2014.2360516

Zheng, X., Tian, L., Luo, G., & Cai, Z. (2020). A collaborative mechanism for private data publication in smart cities. IEEE Internet of Things Journal, 7(9), 7883-7891. doi.org/10.1109/JIOT.2020.2991798

Zhou, B., Pei, J., & Luk, W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM Sigkdd Explorations Newsletter, 10(2), 12-22. doi.org/10.1145/1540276.1540279

Zhou, P., Wang, K., Guo, L., Gong, S., & Zheng, B. (2019). A privacy-preserving distributed contextual federated online learning framework with Big Data support in social recommender systems. IEEE Transactions on Knowledge and Data Engineering, 33(3), 824-838. doi.org/10.1109/TKDE.2019.2936565