Towards Optimization of Privacy-Utility Trade-off Using Similarity and Diversity based Clustering

Abdul Majeed, Safiullah Khan, and Seong Oun Hwang, *Senior Member, IEEE*

Abstract—Most data owners publish personal data for information consumers, which is used for hidden knowledge discovery. But data publishing in its original form may be subjected to unwanted disclosure of subjects' identities and their associated sensitive information, and therefore, data is usually anonymized before publication. Many anonymization techniques have been proposed, but most of them often sacrifice utility for privacy, or vice versa, and explicitly disclose sensitive information when original data have skewed distributions. To address these technical problems, we propose a novel anonymization method using similarity and diversity-based clustering that effectively preserves both the subjects' privacy and anonymous-data utility. We identify influential attributes from the original data using a machine learning algorithm that assists in preserving a subject's privacy in imbalanced clusters, and that remained unexplored in previous research. The objective function of the clustering process considers both similarity and diversity in the attributes while assigning records to clusters, whereas most of the existing clustering-based anonymity techniques consider either similarity or diversity, thereby sacrificing either privacy or utility. Attribute values in each cluster set are minimally generalized to effectively achieve both competing goals. Extensive experiments were conducted on four real-world benchmark datasets to prove the feasibility of proposed method. The experimental results showed that the common and AI-based privacy risks were reduced by 13.01% and 24.3% respectively in contrast to existing methods. Data utility was augmented by 11.25% and 20.21% on two distinct metrics compared to its counterparts. The complications (e.g., # of iterations) of the clustering process were 2.25$\times$ lower than the state-of-the-art methods.

Index Terms—privacy preserving data publishing, similarity, diversity, utility, privacy, clustering, generalization, personal data.

1 Introduction

Most organizations and service providers, such as banks, hospitals, and government agencies, collect a huge amount of data about their customers/subscribers on a daily basis to improve the quality of service. This data often includes basic (i.e., demographic) as well as sensitive information (e.g., income, diseases suffered, and political/religious views) about individuals. The collected data can be utilized for content recommendation, decision-making, healthcare services, etc. Although this data can offer a wide range of benefits to the society and economy when analyzed with data mining tools, privacy issues can restrict its use on a large scale. Therefore, privacy preservation has become an important problem to be addressed while handling personal data encompassing basic, as well as sensitive information about individuals. According to one survey in the United States [1], identification of unique individuals is possible at significantly higher %ages based on the following combinations of three attribute values:

- Gender, zip code (5-digits), and date of birth $\rightarrow$ **87%**
- Gender, place of residence, and date of birth $\rightarrow$ **50%**
- Gender, country of origin, and date of birth $\rightarrow$ **18%**

User attributes, i.e., gender, date of birth, race, and zip code are called quasi-identifiers (QIDs). As cited above, each QID has a distinct effect on privacy, and some QIDs (i.e., zip codes and places) are more likely to identify someone than others. The presence of such QIDs in personal data increases the chances of identity and associated sensitive attribute (SA) disclosures [2]. To address this privacy issue, data is usually anonymized before publication. The practical approaches for preserving privacy are encryption, anonymization, pseudonymization, and obfuscation. Anonymization has been widely used in commercial environments for privacy-preserving data publishing (PPDP) because of low computing overheads and was recently legislated by laws in some advanced countries [3].

The anonymization techniques used to preserve individual privacy in PPDP fall mainly into two major categories: syntactic and semantic. In the former category, the original data is divided into different classes/clusters, and anonymity operations (generalization, suppression, randomization, etc.) are applied. The well-known anonymity mechanisms in this category are k-anonymity [4], ℓ-diversity [5], t-closeness [6], and their upgraded versions. In the latter category, individual privacy is mostly preserved by injecting an appropriate amount of noise into the original data [7]. The famous algorithmic solution for this category is differential privacy (DP) [8] and its enhancements.

Due to the lower defense of syntactic anonymity methods against recent privacy attacks and utility issues of DP, the emerging clustering-based anonymization (CBA) methods have been experimentally tested to fulfill the privacy and utility requirements in PPDP. The major developments of the past five years in the CBA are summarized as follows. In 2016, the SBC-based anonymity [9] model was experimentally tested to solely improve the utility aspect in the PPDP. In 2017, SWAP anonymity [10] was proposed to accomplish privacy-preserving clustering on data stored at distinct locations. A PPC algorithm [11] was proposed to accomplish anonymity and clustering tasks on Hadoop systems. In 2018, a new k-means clustering-based model that is

Abdul Majeed and Seong Oun Hwang are with the Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea.
Safiullah Khan is with Department of IT Convergence Engineering, Gachon University, Seongnam 13120, Republic of Korea.
Manuscript received April 19, 2005; revised August 26, 2015.

superior to traditional k-anonymity in terms of utility was proposed [12]. Zouinina et al. [13] evaluated k-constrained clustering concept to control heavier changes in data sanitization. However, in these studies, SA was not given special attention due to which SA disclosure can inevitably occur. In 2020, a distributed k means-based algorithm in conjunction with DP was proposed to attain higher utility in data publishing [14]. In 2021, the k-member clustering [15], the BHA-based k-anonymity [16], and the k-means++-based anonymity [17] have shown improved results in either utility or privacy. Despite the success of these methods, in most cases, either the user's SA is inferred with a much higher probability through linking with auxiliary data, or the resulting anonymized data has poor quality, leading to poor utility in the knowledge discovery process. In addition, they give equal importance to each QID; thereby, data/value variability issues (distribution, sparsity, density, etc.) cannot be handled effectively. Although CBA algorithms have been applied to anonymize personal data and experimental results have been gathered, our key motivation is to answer these three open research questions. *(1), can we offer a generic and practical solution for anonymizing original data regardless of its structure whether it is in balanced or imbalanced form? (2), can we adapt and leverage multidisciplinary techniques (e.g., random forest, similarities, and diversity) in order to improve various critical aspects of the anonymization process? (3), can we design an objective function for clustering to make the anonymized data protected from various contemporary privacy threats and still offer a higher utility for secondary purposes (e.g., data analytics)?*

To answer the research questions cited above, we propose a novel and generic anonymization method using the clustering concept. This new method well suits the healthcare and banking sectors in which data with a greater balance of privacy and utility need to be shared frequently.

1.1 Main Contributions

The main contributions of this work are listed below.

- In order to transform original data into anonymized data while effectively solving the privacy and utility trade-off, a novel anonymization method using similarity and diversity-based clustering is proposed. It resolves the five problems (see Sec. 2.2) of traditional models by exploiting and using valuable information about attribute values from the original data.
- The objective function of the clustering mechanism is designed based on both similarity and diversity, which ensures homophily based on QIDs as well as heterophily (diversity) regarding SA values in each cluster to control excessive generalization issues.
- We perform the minimal required generalization in each cluster that assists in enhancing utility, as well as safeguarding individual privacy against contemporary threats by having more diversity in SA values.
- The proposed method preserves an individual's privacy—even within clusters where diversity requirements cannot be met due to less heterogeneity in the SA values in T—by identifying influential QIDs that remained unexplored in previous anonymity studies.
- This is the first work that dynamically extracts valuable information in the original data from three different aspects (similarity, diversity, and influence) in a clustering process to achieve the stated goals.

Experiments were performed on four real-world and benchmark datasets under different conditions in order to verify the efficacy of our method. The anonymous data produced by our method can be highly applicable to knowledge-based systems because it can avoid technical problems (such as garbage-in garbage-out, infeasible query results, and data-specific biases) when given as input to such systems. Additionally, this work has higher significance in responsible data science [1] which is a recent trend in the big data era.

The remaining sections of this paper are structured as follows. Section 2 presents background and related work concerning PPDP. Section 3 discusses the proposed clustering-based anonymization method in detail. Section 4 presents the datasets used in experiments, the corresponding results, and the comparisons of our method with prior state-of-the-art work. Section 5 concludes this paper.

2 Background and Related Work

2.1 Preliminaries

In facile words, privacy is all about keeping private information away from the public [18]. The scope of privacy can fall into four major categories [19]. This work falls under information privacy, and is about acquiring, storing and managing, analyzing, and outsourcing personal data. We consider the data enclosed in a table, T , and we illustrate in Figure 1 an example of the tabular data to be anonymized.

(a) Overview of user's attributes in a relational data

	DI	NSA	Quasi Identifiers (QIDs)			SA
ID	Name	Height	Age	Zip Code	Martial Status	Salary
1	Joe	6	29	32042	Separated	>50K
2	Jill	5	20	32021	Single	>50K
3	Sue	4	24	32024	Widowed	≤50K
4	Abe	5	28	32046	Separated	≤50K
5	Bob	7	25	32045	Widowed	>50K

(b) Treatment in the anonymization process

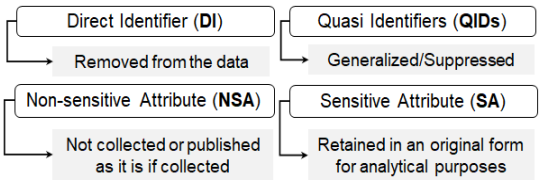


Fig. 1: Overview of user attributes and their treatment in the anonymization process.

There exist three main privacy threats (membership, identity, and SA disclosures) that can emerge via QIDs and SA values after the data release. Anonymization employs different operations to provide sufficient resilience against all these threats. The most widely used operations are generalization, noise addition, and suppression [19]. All these operations modify/shuffle the QID values in T to protect a user's privacy and augment the utility of anonymous data.

Many techniques have resulted from three pioneer studies of PPDP: k-anonymity [4], ℓ-diversity [5], and t-closeness

1. https://redasci.org/

[6] for a non-interactive setting (e.g., a setting in which the whole dataset is published in anonymous form). Zouinina et al. [20] proposed a two-step anonymity method for PPDP based on ML combined with k-anonymity. An anonymity model to avoid privacy attacks resulting from the inherent properties of equivalence classes was proposed by Huang et al. [21]. Recently, many enhanced versions of the ℓ-diversity model [5] have emerged, such as DHkmeans ℓ-diversity [22], and distributed ℓ-diversity [23], to name a few. Although these methods are handy in terms of realizing the ℓ-diversity model's property in practical scenarios, various data utility-related issues can still occur frequently.

2.2 Previous Research Status

Recently, due to a significant rise in the emergence of data from different sources (the IoT, the IoMT, epidemic systems, etc.), many clustering-based anonymity solutions have been proposed for PPDP [24], [25]. These solutions eliminate the hard constraint/parameter enforcement in anonymization to yield better privacy and utility. Lin [26] proposed a clustering-based algorithm named *kRPP*. The proposed algorithm makes use of k-means clustering integrated with noise addition to preserve utility. Canbay et al. [27] proposed an anonymity method based on distributed k-means clustering for PPDP. Few studies have explored related solutions for PPDP, mixed-feature weighted clustering for k-anonymity (MWCK) [28], random k-anonymous (RKA) [29], and effective ℓ-diversity (ELD) [30] algorithms. The MWCK algorithm [28] provides support for data publishing with reduced information loss, but diversity is not considered, and thereby, SA disclosures can occur with a higher probability. The RKA algorithm [29] preserves privacy and utility by using noise addition and shuffling operations. Meanwhile, that proposed method can be subjected to higher information loss when the data comprises only numerical attributes. An ELD algorithm [30] improves privacy and utility in data sharing, but it can lead to SA and identity disclosures when clusters cannot fulfill the ℓ-constraints due to an imbalance in SA values. There are five major problems with the aforementioned techniques.

- The existing literature gives preference to either QID or SA values in the anonymization process, and thereby, effective resolution of privacy and utility cannot be guaranteed simultaneously in most cases [5], [9], [15]. Solving the privacy-utility trade-off (PUT) is quite challenging to achieve, but imperative scientific/business goal.
- There is a lack of methods that can extract some valuable information (i.e., attribute influences) regarding attribute values/distributions in order to lessen explicit disclosure of SA/identity when the class/data is highly imbalanced (i.e., SA value distributions are uneven) [27], [30].
- Privacy models that enforce constraints regarding SA values often reduce the degree of reliability in the information. Therefore, adoption of the anonymized data produced by them is not as effective in knowledge-based applications/systems [5], [21].
- Most clustering-based anonymity methods delete data that is less frequent, which can hinder knowledge discovery from all perspectives [17].
- Some clustering-based methods are not well suited to privacy protection, because objective functions of clustering are not tailored to the characteristics/values of attributes, which can limit their use in real-world scenarios.

3 The Proposed Anonymization Method

In this section, we describe the workflow of our clustering-based anonymization method. We demonstrate the conceptual overview of the proposed anonymization method in Figure 2. The proposed method comprises five main phases: identification of attribute types present in T, pre-processing of T, influential QID identification from T by using ML, similarity and diversity-aware clustering of users, and generalization of the cluster set to yield $T^{'}$. Comprehensive descriptions of each phase are provided in subsections 3.1-3.5. This method is devised to effectively preserve user privacy and utility in the relational data, and to reduce privacy risks caused by highly influential QIDs and low-diversity clusters. Table 1 presents the key notations used in the proposed anonymization method.

TABLE 1: Main notations used in the proposed method.

Symbols	Concise detail
T	Original data
$T^{'}$	Anonymized data
N	Number of individuals in T, where $N = \|T\|$
u_i	ith user/record/tuple in T
A	Set of attributes in T, where $A = \{a_1, a_2, \ldots, a_n\}$
Q	Set of QIDs, where $Q = \{QI_1, QI_2, \ldots, QI_p\}$
S	SA values set, where $S = \{v_1, v_2, \ldots, v_{p'}\}$
$v_{QI_1}^{u_i}$	Value of the first QID (e.g., QI_1) for user u_i
I_{QI_i}	Influence value of an ith QID present in T
$CS(u_i, u_j)$	Similarity value between users i and j
k	Privacy parameter (i.e., at least k tuples in a cluster)
ξ	Similarity-wise ranked user matrix, where $\xi \subseteq T$
Y	Initial cluster-centers matrix computed from ξ
T_{CS}	Threshold for similarity while allocating records
$Div(u_i, u_j)$	Diversity between users i and j based on SA values
Z	Set of clusters, where $Z = \{Z_1, Z_2, Z_3, \ldots, Z_{\|Y\|}\}$

3.1 Identification of attributes types present in T

Generally, T can contain four types of attributes about individuals—QIDs, SAs, non-sensitive attributes (NSAs), and direct identifiers (DIs)—as shown in Figure 1. These types are usually determined based on the association with someone's identity, availability at external sources/databases, and any privacy requirements [31]. For example, most people want to hide their finance/disease information from others, and therefore, it is regarded as SA information. Similarly, DIs and QIDs, respectively, can reveal someone's identity, fully or partially. All other attributes that do not fall into these three classes are called NSAs. Usually, NSAs do not encompass critical information about individuals that can assist in re-identification. Depending upon the situation, the availability of attribute types varies in T, and some types

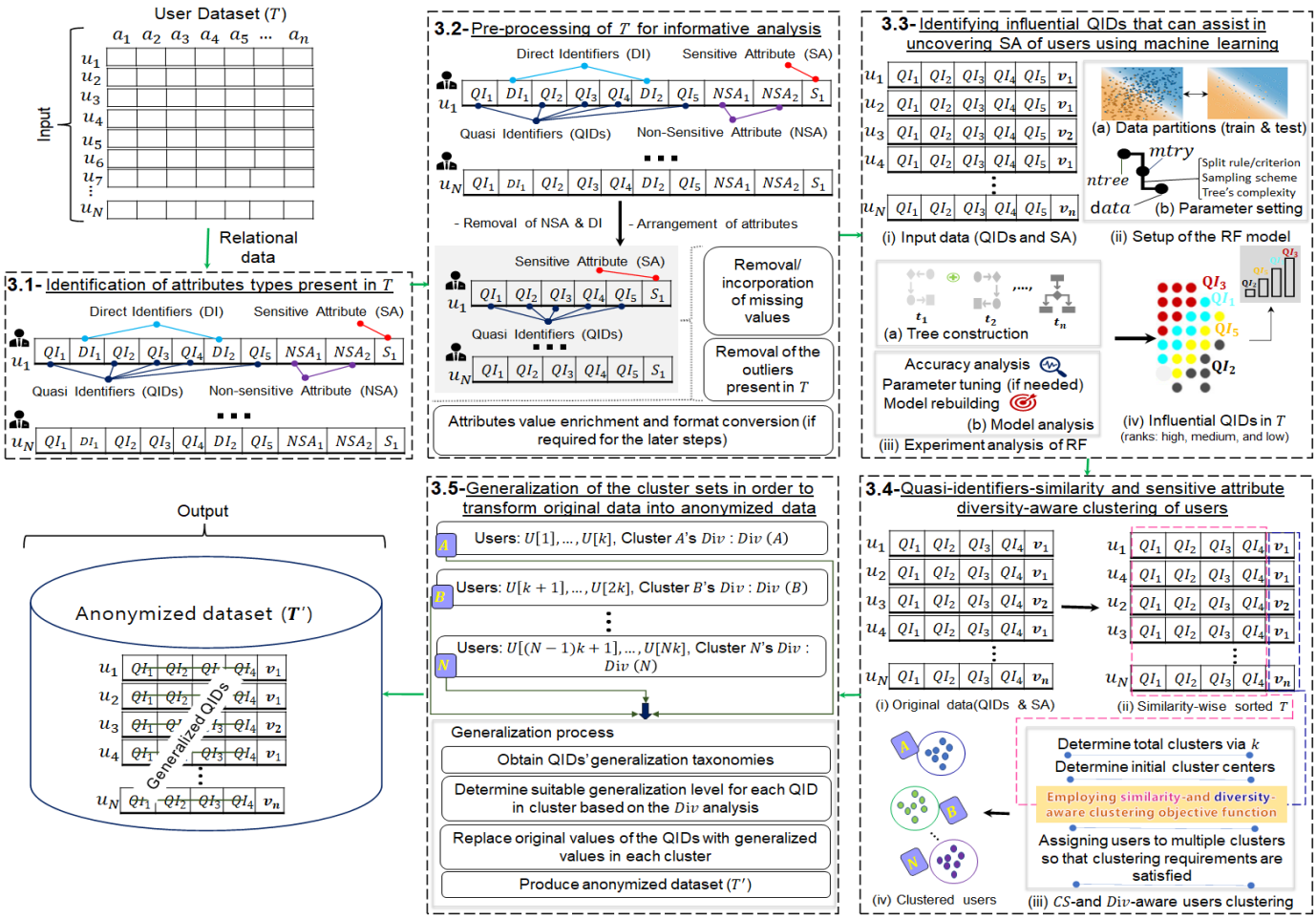


Fig. 2: Conceptual overview of the similarity-and diversity-aware clustering-based anonymization method for tabular data.

of attributes (i.e., DIs, NSAs) may or may not be present in T. Therefore, we classify as follows the four possible groupings regarding the availability of attribute types in T.

- Possibility I: {(QIs, ✓), (SA, ✓), (NSA, ✓), (DI, ✓)}
- Possibility II: {(QIs, ✓), (SA, ✓), (NSA, ×), (DI, ✓)}
- Possibility III: {(QIs, ✓), (SA, ✓), (NSA, ✓), (DI, ×)}
- Possibility IV: {(QIs, ✓), (SA, ✓), (NSA, ×), (DI, ×)}

After determining the availability of attribute types in T, we assign labels and determine the indexes of each attribute type, as shown in Figure 2 with an example: $\{\forall A \in T : a_1, a_3, a_4, a_5, a_7 \in QIDs, a_2, a_6, a_9 \in DIs, a_8, a_{10} \in NSA, a_{11} \in SA\}$. After the identification of attribute types, and their indexes, pre-processing is performed on T to clean it for further processing.

3.2 Applying a five-step method to pre-process T

Data pre-processing has become a de-facto standard in most applications, and it makes data interpretation easier. Therefore, we pre-process T in five steps, as described below.

3.2.1 Removal of NSAs and DIs from T

In the beginning, we eliminate two types of attributes, NSAs and DIs, per the standard criteria for PPDP [19]. DIs are mainly removed to prevent explicit identity disclosure of individuals, whereas NSAs have a negligible effect on data utility. In some cases, NSAs are published as they are—if they provide useful analysis in combination with QIDs. Both of these types of attributes can be removed at any stage, but earlier removal can significantly assist in saving computing overhead. After the removal of these two types of attributes (i.e., $A - \{NASs, DIs\}$), set A contains only two other attribute types, namely, QIDs (denoted as Q) and SAs (denoted as S). For the sake of similarity, we use Q and S hereafter to refer to both types of attributes in T.

3.2.2 Arrangement of the attributes

Since NSAs and DIs are removed from the different indexes, QID and SA may not be well separated from each other in some cases. To address these issues (i.e., grouping QIDs and separating them from SA), we arrange the indexes of QIDs and SAs. In most cases, SA is usually moved to the last column of a T. The structure of T after the steps in Section 3.2.1 becomes $T\{Q, S\}$, where $Q = \{QI_1, QI_2, \ldots, QI_p\}$, and $S = \{v_1, v_2, \ldots, v_{p'}\}$. In set Q, each element has a label and a value. For example, QI_1 can refer to age, and each individual can have the same (or a different) value for age in their respective cells. Similarly, v_1 denotes the SA value in a tuple concerning an individual. An overview of a simplified structure for all tuples after attribute arrangements is provided in Eq. 1.

$$T_{users,attributes} = \begin{pmatrix} u_i & QI_1 & QI_2 \cdots & QI_p & S \\ u_1 & v^{u_1}_{QI_1} & v^{u_1}_{QI_2} \cdots & v^{u_1}_{QI_p} & v_1 \\ u_2 & v^{u_2}_{QI_1} & v^{u_2}_{QI_2} \cdots & v^{u_2}_{QI_p} & v_2 \\ u_3 & v^{u_3}_{QI_1} & v^{u_3}_{QI_2} \cdots & v^{u_3}_{QI_p} & v_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_N & v^{u_N}_{QI_1} & v^{u_N}_{QI_2} \cdots & v^{u_N}_{QI_p} & v_{p'} \end{pmatrix} =$$

$$\begin{pmatrix} u_i & QI_1 = age & QI_2 = gender & QI_p = race & S = disease \\ 1 & 29 & M \cdots & Black & Flu \\ 2 & 38 & F \cdots & White & Leukemia \\ 3 & 59 & F \cdots & Black & Flu \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 10,000 & 38 & M \cdots & White & Rhinitis \end{pmatrix} \quad (1)$$

where each row shows complete information of a user, encompassing basic attributes (i.e., QIDs) as well as SAs. Each column shows one item (e.g., salary or age) related to that particular user. After giving an understandable structure to T, some pre-processing operations on actual values of the QIDs and SAs are performed.

3.2.3 Removal of outliers

Generally, data is obtained from different sources and individuals; therefore, it often includes outliers (i.e., unrealistic values for some attributes; for example, age cannot be 300, but is more likely 30). Outliers can emerge in data due to flaws in collection methods, human errors, measuring device limitations, etc. Outliers corrupt analysis results, therefore, they should be removed from T. To remove outliers, we visually inspect each attribute value and analyze its domain. In addition, we perform a $min - max$ analysis to check whether each attribute's values are within the desirable range or not. After outlier removal, T is further analyzed for any missing values.

3.2.4 Removal/estimation of missing values

In most real-world T, it is highly likely that some information about individuals will be incomplete or missing. Missing values can appear in T due to data validation or privacy preferences because some people prefer to not disclose some personal attributes. There exist two common mechanisms to deal with missing values: elimination and estimation. In this work, we eliminated records with missing values. Alternatively, missing values can be estimated by performing proximity analysis with other available values [25]. Apart from removing records with missing values, we ensure that values for each attribute are consistent with the relevant data type (i.e., categorical, numerical, etc.). Duplicate records are also removed at this stage to improve clustering quality and to lower computational overhead.

3.2.5 Attribute value enrichment and format changes (if needed)

In the last pre-processing step, attribute value enrichment and format conversion are carried out to lessen complications with some mathematical operations. For example, similarity and distance computations can become easier when all attributes are numerical. Therefore, we convert some attribute types by using the key–value pair concept. For example, gender information can be converted/encoded from categorical to numerical using two key–value pairs like: $(M, 0)$ and $(F, 1)$. In addition, in some cases, data is not in a unified format due to direct scanning from documents. Therefore, the format of the values can be changed accordingly in order to execute the desired operations on them. By applying all five steps, an error-free T is obtained.

3.3 Identification of influential QIDs by using ML

As shown in Sec. 1, each QID has a distinct impact on identification, and some items have a greater identity-revealing ability than others. QIDs with that greater ability can assist adversaries in compromising someone's identity, leading to SA disclosure as well [32]. We refer to these QIDs as influential QIDs, because they can have a strong correlation with SA values, and can thereby accelerate identity and SA disclosure. Studies have highlighted the significance of identifying and giving alternate weights to each QID during the anonymization process in order to control privacy issues [2]. Moreover, the identification of influential QIDs using ML and their utilization in CBA processes has remained unexplored. To that end, we propose a mechanism for identifying influential QIDs from T to better address the privacy issues by using an ML method named random forest (RF) [33]. The rationale behind the RF selection is its ability to yield better accuracy, and most parameter values can be specified at one time. Hence, parameter tuning is relatively easy. Further, RF can handle high-dimensional data, and it gives output based on multiple trees rather than relying on a single tree's results. The procedure employed to identify and rank influential QIDs using RF is given in Figure 3.

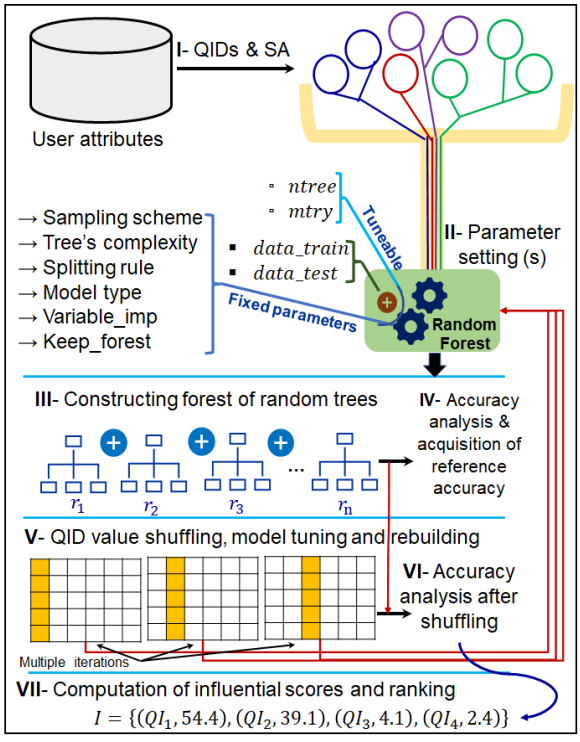


Fig. 3: Workflow of the method utilized to identify influential QIDs in T.

The complete procedure comprises seven main steps. In the first step, T encompassing QIDs and SAs of users is provided as input. In the next step (i.e., parameter setting), T is partitioned into d_{train} and d_{test}. The former is used

to train the RF model, and the latter is used for validation. We divide the RF parameters into two categories: tuneable and fixed. The tuneable parameters are $ntree$ (the number of classification/regression trees to be made from the data) and $mtry$ (the variable required to split each tree's nodes in the tree construction process). These two parameters have an impact on model performance and the resulting accuracy; therefore, their values are chosen carefully based on the problem size. The value of fixed parameters can be specified all at one time. For example, the RF model type can be chosen based on the data type of an SA (e.g., regression, if the SA is numerical, otherwise, classification). Furthermore, the splitting rule/criterion value can be selected as Gini index or impurity. Some parameter values (i.e., $variable_{imp}$) can be set to true/false depending upon the requirements. Once all parameter values are specified, trees are constructed from data samples, and accuracy values are obtained by aggregating each tree error or vote (steps 3 and 4). Later, validation is performed on the test data, and the reference accuracy value, denoted as δ_{ref}, is recorded. The value of δ_{ref} can be determined using Eq. 2.

$$\delta_{ref} = 1 - \sum_{r=1}^{ntree} \mu_r \tag{2}$$

where μ_r denotes the error (a.k.a. an out-of-bag (OOB) error for tree r). After obtaining δ_{ref}, each QID value is permutated column-wise (one QID at a time), and the RF model is built from the modified data in Step 5. Due to permutation, the association between S and $QI_i^{'}$ is broken. Therefore, distinct impacts on accuracy can be observed when the model is built from permutated and non-permutated QIDs together. The new accuracy value can be obtained, which we refer to as modified/shuffled data accuracy, and is denoted with δ_m. Later, we take the difference, d, where $d = \delta_{ref} - \delta_m$, between the two accuracy values to analyze the shuffling effect. If d is large, it means the QID in which values were shuffled is more influential, and the opposite is also true. The influence, I, of a QID (QI_i) induced in tree r by value shuffling can be determined using Eq. 3.

$$I^r_{QI_i} = \frac{\sum_{x \in \mu_r} I\left(v_x = \hat{v}_x^{(r)}\right)}{|\mu_r|} - \frac{\sum_{x \in \mu_r} I\left(v_x = \hat{v}_{x,\pi_j}^{(r)}\right)}{|\mu_r|} \tag{3}$$

where μ_r denotes the OOB error for tree r, x denotes an observation, v represents an SA value (a.k.a. class in ML), v_x denotes # of votes for correct class in xth observation, $\hat{v}_x^{(r)}$ is the predicted value of the SA in the xth observation before permutation, and $\hat{v}_{x,\pi_j}^{(r)}$ is the predicted value of the SA in the xth observation after QI_i value permutation. Numerical value V of I_{QI_i} in r can be of the following two types:

$$V(I^r_{QI_i}) = \begin{cases} I^r_{QI_i}, & \text{if } QI_i \in r \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

In Eq. 4, $V(I^r_{QI_i}) = 0$ implies that either QI_i is not a part of tree r or it has all identical values. To find the influence score of the ith QID in T, we use Eq. 5.

$$I_{QI_i} = \frac{\sigma}{\zeta} \tag{5}$$

where σ and ζ, respectively, denote the standard deviation and the mean of the OOB error determined collectively from all trees. The values of ζ and σ can be determined by using Eqs. 6 and 7, respectively:

$$\zeta = \frac{\sum_{r=1}^{ntree} I^r_{QI_i}}{ntree} \tag{6}$$

$$\sigma = \sqrt{\frac{1}{ntree - 1} \sum_{r=1}^{ntree} (I^r_{QI_i} - \zeta)^2} \tag{7}$$

With the help of Eq. 5, we can determine the I value of QI_i. Similarly, we can determine the I values for other QIDs, and can store them for subsequent steps. These values represent the value of I from a global perspective (i.e., the whole T). The set I of influential QIDs obtained from RF can be mathematically expressed as

$$I = \{(QI_1, I_{QI_1}), (QI_2, I_{QI_2}), \ldots, (QI_p, I_{QI_p})\} \tag{8}$$

In Eq. 8, based on the I values, we rank the QIDs in order to pay ample attention to those QIDs that are most influential in imbalanced clusters to address privacy issues.

The higher I value for a particular QID indicates that most values in that QID are distinct (e.g., cardinality is very high), and therefore, the chances of identity revelation can be high during published data analytics [34]. In contrast, a lower I means that a particular QID has only a few distinct values, and one/two values constitute $\geq$ 80% of T, and therefore, the probability of re-identification is less due to general patterns [35]. We explain experimental results of I obtained from each benchmark dataset in Section 4.2.

3.4 Similarity-and diversity-aware clustering

This subsection presents the method proposed for clustering based on similarity and diversity. Our proposed method considers the similarity of QIDs and the diversity of the SAs as criteria in the clustering process to flexibly maintain the balance between privacy and $T^{'}$ utility. The clustering method comprises the five main phases discussed below.

3.4.1 Ranking of records for ensuring higher information availability in $T^{'}$

In order to reduce information loss from $T^{'}$, records in each cluster must have the same QID values to the extent possible. By ignoring higher similarities among subjects in each cluster, excessive distortion in $T^{'}$ can occur during anonymization, resulting in poor data quality. Higher cohesion in clusters can be achieved by exploiting the similarities among records. To find and rank similar records, we employed cosine similarity (CS). It is a highly reliable and one of the most widely used measures in recommender systems for similarity measurement among users/items [36]. Its value ranges between 0 and 1, $CS \in [0, 1]$, where 1 means two records have identical QIDs, and 0 indicates that nothing is common. The CS value between two distinct users, u_1 and u_2, can be computed using Eq. 9.

$$CS(u_1, u_2) = \frac{\sum_{i=1}^{p} u_{1_i} \times u_{2_i}}{\sqrt{\left(\sum_{i=1}^{p} u_{1_i}\right)^2} \times \sqrt{\left(\sum_{i=1}^{p} u_{2_i}\right)^2}} \tag{9}$$

where i represents the QIDs, and p denotes the total # of QIDs. From Eq. 9, CS values among all the subjects are computed, and results are provided as input to the clustering process. The pseudo-code used to compute similarities and to rank the most similar records from T is shown in Algorithm 1.

Algorithm 1: Ranking of similar records for ensuring higher information availability in $T^{'}$.

Input : (1) Dataset T encompassing N records. Each u_i has multiple QIDs and one SA.
Output : Matrix ξ of highly similar records.
Procedure:

1 $N = |T|$
2 **if** $(N == 0)$ **then**
3 Return "Error: no records in T".
4 **else**
5 **Initialize**, matrix ξ = zeros($|T|, p+1$)
6 **Initialize**, matrix χ = zeros($|T|, |T|$) ▷ Adj. matrix
7 **Initialize**, matrix h = zeros($|T|, 1$) ▷ Index matrix
8 **for** $i = 1$ to N do
9 **for** $j = 1$ to N do
10 Compute $CS(u_i, u_j)$ between users i and j based on QID values ($QI_1 \rightarrow QI_p$) via Eq. 9.
11 Add the computed $CS(u_i, u_j)$ values in χ.
12 Repeat steps 8-9 for subsequent values of i and j.
13 **End for**
14 **End for**
15 Sort χ in descending order (e.g., records with higher CS values appear on the top, and vice versa) based on computed CS values.
16 Find the new indexes of records and store in h.
17 Get the complete records values from T based on new indexes stored in h.
18 Add results in ξ, where $\xi = T(h, :)$.
19 **end**
20 **return** ξ

In Algorithm 1, dataset T is provided as input. Matrix ξ containing the most similar records is obtained as output. The CS values in the records are determined, and results are saved in the user matrix χ (Lines $8-14$). Later, the records are sorted (e.g., the CS value decreases down the order in ξ) based on the mutual analysis of CS values among records, new indexes for all records are determined, and QID values from T are updated based on newly determined indexes (Lines $15-18$). Finally, matrix ξ containing records ranked by CS values is returned (Line 20).

These ξ's results will be employed for the identification of initial cluster centers, as well as for assigning records to each cluster. During this similarity computing process, the placement of records usually gets changed based on CS values, as shown in Eq. 10.

$$T = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}, \xi = \begin{pmatrix} u_5 \\ u_{N-1} \\ \vdots \\ u_2 \end{pmatrix} \tag{10}$$

The CS values can fall into one of three cases (i.e., 0, 1 and $0 < CS < 1$) based on the QIDs.

3.4.2 Selection of the initial cluster centers

The initial cluster centers have a significant impact on clustering quality (i.e., homogeneity within a cluster, and heterogeneity between clusters) and computing/convergence speed. To compute the initial cluster–centers matrix Y that can augment convergence speed and improve the compactness of the clusters, we use two types of information (i.e., CS and I) already extracted about the QID values. By employing this information, the initial cluster centers are determined in a greedy fashion that is more reliable, and that needs less updating in the clustering process, thereby contributing to preserving utility in the data. The procedure to determine Y comprises three key steps: determining the total number of clusters considering size ξ and privacy parameter k, a well-defined slicing of ξ based on the k value, and selection of the cluster centers exploiting I and CS information. Since the most influential QIDs mostly have distinct values, they can assist in ensuring low intra-cluster similarity. Furthermore, by means of slicing, high intra-cluster similarity can be maintained. By jointly using slicing and influence information, initial cluster centers are extracted that provide a solid foundation for final clusters that can ensure uniformity in a global scope. The pseudo-code used to determine initial cluster centers is given in Algorithm 2.

Algorithm 2: Selecting initial cluster centers.

Input : (1) Matrix ξ encompassing N similarity-wise ranked records.
(2) Privacy parameter k, where $k > 1$.
(3) Influence score information I.
Output : Matrix Y of initial cluster centers
Procedure:

1 **if** $(N \leq k)$ **then**
2 Return "Error: fewer records than k".
3 **else**
4 **Initialize**, matrix $Y = \emptyset$
5 Determine total number of clusters C using formula $C = |\xi|/k$
6 Create multiple slices set τ from ξ, where $\tau = \{\tau_1, \tau_2, \tau_3, \ldots, \tau_g\}$
7 **Assign**, matrix $Y = y_1$, a random tuple
8 **for** $i = 1$ to $|\tau|$ do
9 Identify tuple y_i from τ_i that has CS value $\geq T_{CS}$ from al teast $k-1$ other tuples in the neighbourhood.
10 Check the dissimilarity between inflential QID values already stored in Y.
11 If infleuntial QID values are different, then $y_1 = y_i$.
12 **Include**, y_i in Y, $Y = Y \cup y_i$
13 **Repeat**, $\forall\ \tau$, and get a center y in each round.
14 **End for**
15 Ensure the required number of centers is determined using Eq. $|Y| == C$ from ξ.
16 **end**
17 **return** Y

In Algorithm 2, matrix ξ encompassing N similarity-wise ranked records, k (a.k.a. the privacy parameter), and

the QIDs' influential information are provided as input. Matrix Y of the initial cluster centers is gathered as output. At the beginning, the total number of clusters is determined, ξ is sliced, and an initial cluster center is determined randomly (Lines $5-7$). Later, from each slice, cluster center y is determined by exploiting neighborhood similarity and influential QID value dissimilarity information (lines $8-14$). Upon identification of the required cluster centers, matrix Y encompassing $|C|$ centers is returned as output. Since Algorithm 2 uses pre-computed CS and I values, the computing overhead is very small. The Y resulting from Algorithm 2 is used in the cluster formation process.

3.4.3 Designing the similarity- and diversity-aware objective function

The proposed anonymization method aims to ensure greater data availability without compromising privacy. Thus, a similarity- and diversity-aware objective function is designed for clustering that ensures greater similarity when records are mapped to initial cluster centers while also increasing the diversity. The core idea is to divide data into clusters in such a way that both similarity (CS) and diversity (Div) increase simultaneously. The former measure (i.e., CS) contributes to significantly lowering information loss, and thereby, $T^{'}$ can preserve more utility. The latter measure (Div) assists in consistently protecting against SA disclosure from malevolent adversaries. More specifically, CS and Div are the two decisive measures that can affect privacy and utility. Given an initial cluster center, $y_i \subseteq Y$, the decision to couple a new tuple, t, with y_i to form two-record cluster Z_i (i.e., $k = 2$) depends on the following objective function:

$$Z_i(y_i, t) = \alpha_{pr} \cdot Div(v_i, v_t) + \beta_{ut} \cdot CS(y_i, t) \qquad (11)$$

where y_i represents an initial cluster center (in some cases, it is chosen randomly), and t is a new tuple to be joined with y_i (if the clustering requirements are satisfied) to form cluster Z_i. $Div(v_i, v_j)$ denotes the Div increases if $v_i \neq v_t$, where v_i and v_t denote the SA value of the initial cluster center and the new tuple, respectively. Moreover, while assigning tuples to the clusters, ample attention is paid to prevent $Div(v_i, v_t) = 0$ in order to control explicit SA disclosures. $CS(y_i, t)$ represents the CS value between both tuples based on the QID values, and α_{pr} and β_{ut} are the privacy and utility parameters used to adjust the balance between the two competing goals. The main reason to introduce two coefficients is to augment the flexibility of the proposed method and to empower data owners in deciding whether user privacy or data utility is more crucial. By choosing distinct values for these two coefficients, the structure of the clusters and corresponding privacy and utility results will be altered. Hence, these coefficients are imperative in adjusting privacy and utility levels in real-world applications. The CS computing process was explained in Section 3.4.1, and the Div values between tuples can be computed using the Shannon index [37]. It is a popular index mostly employed to analyze the dynamics of communities/biodiversity in forests. We employed this concept to measure Div between tuples using Eq. 12.

$$Div(y_i, t) = -\sum_{i=1}^{|S|}[(p_i) \times ln(p_i)] \qquad (12)$$

where p_i denotes the proportion of each unique SA value. The Div value ranges between 0 and 1 (i.e., $Div \in [0, 1]$). An example of computing Div values between two tuples having similar/dissimilar SA values is shown in Figure 4.

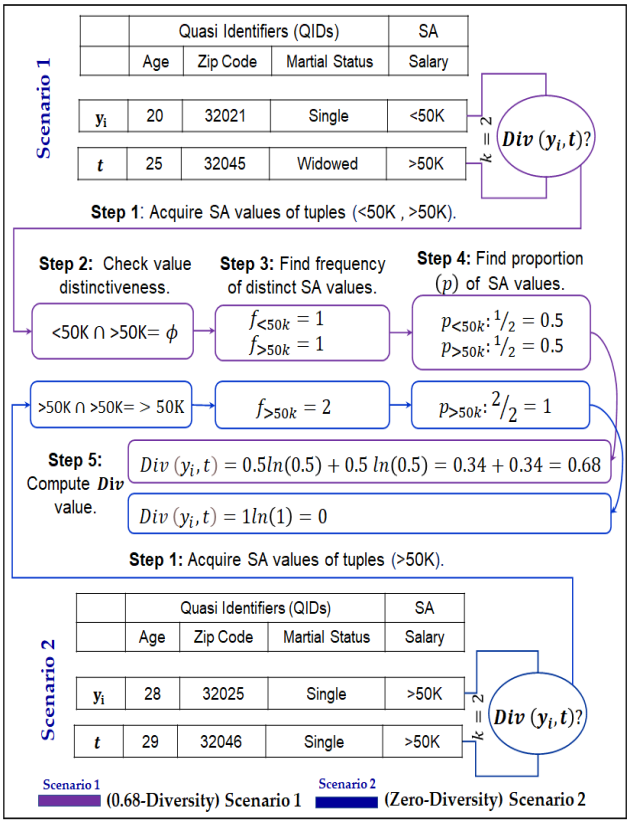


Fig. 4: Example of computing Div values in the clustering process.

By designing and using the CS-and Div-aware objective function for clustering, most requirements of the traditional anonymity models (e.g., ℓ-diversity, β-likeness, p-sensitive k-anonymity, etc.) can be met, while $T^{'}$ retains more utility for data mining tasks. Furthermore, it changes the placement of tuples, and thereby, privacy is guaranteed against linking and background knowledge attacks.

3.4.4 Assigning records to clusters based on clustering requirements

In this step, we explain the mechanism based on which records are assigned to their respective clusters while satisfying the clustering requirements (i.e., higher CS and Div values, and at least k records in each cluster Z_i). We ensure dynamism while dividing data into clusters so that most clusters can satisfy the clustering requirements. The pseudo-code used to assign records to their respective clusters based on the clustering requirements is shown in Algorithm 3.

In Algorithm 3, matrix ξ encompassing N records, k (the privacy parameter), and matrix Y of the initial cluster centers are provided as input. The set of clusters, Z , where each cluster in Z contains at least k tuples, is yielded as

Algorithm 3: Assigning tuples to their respective clusters.

Input : (1) Matrix ξ encompassing N similarity-wise ranked records.
(2) Matrix Y of initial cluster centers.
(3) Privacy parameter k, where $k > 1$.

Output : Set Z of clusters, where each cluster has at least k tuples.

Procedure:

1 **if** $(|\xi| \leq k)$ **then**
2 Return "Error: insufficient records in ξ".
3 **else**
4 **Let**, set $Z = \emptyset$
5 **Generate** raw clusters set Z_{raw} with one tuple as $Z_{raw} = \{y_1, y_2, y_3, \ldots, y_C\} \leftarrow$ initial centers
6 **for** $i = 1$ to $len(Z_{raw})$ do
7 Select raw cluster Z^i_{raw} to be converted into Z_i.
8 While (no. of tuples in $Z^i_{raw} < k$) do
9 Extract a close candidate tuple t from ξ
10 $t =$ Arg-max $Z_i(y_i, t) t \in \xi \leftarrow$ using Eq. 11
11 **if** *(t satisfies clustering criteria)* **then**
12 $\xi = \xi$- t, $Z^i_{raw} = Z^i_{raw} \cup t$
13 **else**
14 $\xi = \xi \cup t$, and repeat steps 9 onward.
15 **end**
16 Once at least k tuples are mapped to Z^i_{raw}.
17 End While
18 $Z_i = Z_i \cup Z^i_{raw} \leftarrow$ formation of final clusters.
19 **Include**, Z_i in Z, $Z = Z \cup Z_i$
20 **Repeat** steps $7 - 19$, $\forall\ len(Z_{raw}) - 1$, and get cluster of at least size k (e.g., $\forall\ len(Z_i) \geq k$)
21 **Update** Z in each iteration with a new cluster Z_i
22 **End for**
23 **if** *(there remain some tuples in ξ (i.e., $|\xi| \neq 0$))* **then**
24 Add residual tuples to relevant clusters in Z based on CS until k does not exceed by much.
25 Produce final clusters set Z to be anonymized.
26 **else**
27 No residuals left in ξ (i.e, $|\xi| = 0$)).
28 Z is a set of final clusters to be anonymized.
29 **end**
30 **end**
31 **return** Z

output. Steps $4 - 30$ are the key steps in the clustering algorithm. Step 5 generates a set of raw clusters (e.g., clusters that have fewer records than k, and all records that are pre-extracted initial cluster centers). Step 6 ensures that the loop will continue until raw clusters are converted into final clusters. One raw cluster (i.e., ith from Z_{raw}) to be processed further is extracted in Step 7. Line 8 determines the suitability of tuple t to be included in a cluster by using Eq. 11, where, as long as the k-anonymity property is not met, the loop will progress. Steps 9-10 decide the relevant tuple to be included in a cluster. Lines 11-15 check whether t is a good choice for a particular cluster. Line 16 ensures that k tuples are included in a particular cluster. Lines 18-21 convert a raw cluster into a final cluster, including it in a cluster set. The residual records are assigned in lines 23-29 to relevant clusters based on the clustering requirements (i.e., CS). While assigning residual tuples to the clusters, the strength of the tuples is maintained with factors $((k + k/2), (k + k/3), (k + k/4)))$ depending upon the k-value. Finally, the set Z of clusters is obtained in Line 31 and is anonymized afterward. In some clusters, due to residual tuple additions, the Div value can decrease, leading to imbalanced cluster formation. Hence, we pay ample attention to these clusters during QIDs' generalization.

Recently, a new branch of the clustering methods named, fair clustering (FC) has been proposed, which aims to hide SAs in the clustering results [38]. FC ensures four characteristics such as fairness, compactness, informative features, and balancedness while partitioning data into clusters. FC is well suited to image data and enables the partition of data that are debiased. However, the use of FC in data anonymization scenarios is yet to be explored. In contrast, the proposed clustering method creates compact clusters of size k while ensuing higher diversity in each cluster concerning SA values. The proposed method is well suited to tabular data, and its efficacy in anonymization scenarios has been rigorously tested. The proposed method is imperative to foster the secondary use of data while solving PUT.

3.5 Needs-based generalization of cluster sets to produce anonymous data

Once the clusters of size k or $\geq k$ are formed, needs-based generalization of QID values in each cluster set is performed to produce anonymous data, $T^{'}$. Generalization is the process of converting real values into generalized values. The generalized values have two main properties to the original values: less specificity and semantic consistency. For example, a QID(age) value of 29 can be generalized to $25 - 30$, < 30, > 20, etc. However, it is desirable to employ precise operations to preserve more utility in $T^{'}$. We perform the generalization of QID values in clusters using generalization hierarchies. In Figure 5, we present an example of hierarchy H for QID (age).

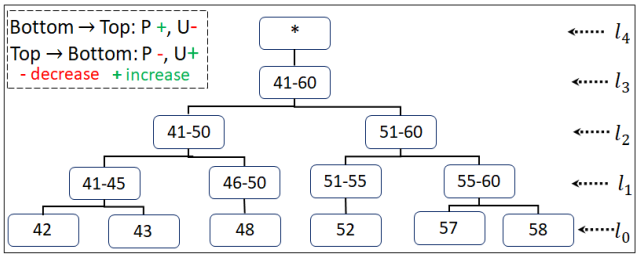


Fig. 5: Overview of the generalization hierarchy for age.

As shown in Figure 5, there can be different levels in each QID's H. The first level (e.g., l_0) represents the original QID value to be anonymized; the intermediate levels (i.e., l_1, l_2, l_3) are generalized forms of the original value, and the last level (i.e., l_4) is called suppression because it hides the QID value fully. Furthermore, privacy (P) and utility (U) have distinctive effects on each level. For instance, the levels close to l_0 in H have higher utility but provide less privacy,

and vice versa. The selection of appropriate generalization is of paramount importance to satisfy the two important (yet conflicting) requirements. In order to maintain the balance between privacy and utility, we select generalization levels that are close to the original values by paying ample attention to Div values in each cluster. Furthermore, the proposed method only performs the necessary generalization, meaning that higher-level generalization is prevented due to CS-based clustering, and if needed, only influential QID values are anonymized to the higher levels. Hence, an effective resolution of two competing goals is achieved consistently. The pseudo-code used to perform needs-based generalization to produce $T^{'}$ is shown in Algorithm 4.

Algorithm 4: Needs-based generalization of clusters.

Input : (1) Cluster set Z.
(2) Hierarchy set H of QIDs.
(3) Set I of QID influence scores.
(4) Privacy parameter k value.

Output : Anonymized data $T^{'}$.

Procedure:

1 **Let**, $T^{'} = \{\}$
2 **for** each cluster $Z_i \in Z$ do
3 **if** $(|Z_i| > k)$ **then**
4 Z_i can be imbalanced cluster with high probability.
5 Compute Div value of Z_i using the formula below. $Div(Z_i) = -\sum_{i=1}^{|S^{'}|}[(p_i) \times ln(p_i)]$
6 Compare Div value with Div threshold T_{Div}.
7 **if** $(Div > T_{Div})$ **then**
8 Perform anonymization using the outer *else* box.
9 **else**
10 **for** each QID $QI_e \in Z_i$, where $e = 1, 2, \ldots, p$
11 Obtain hierarchy $H_{QI_e} \in H$ of QI_e
12 Acquire influence score I_{QI_e} of the QID
13 If I is higher, generalize QI_e values at higher level of H_{QI_e}
14 **Repeat:** $\forall$ QIDs, $QI_1, QI_2, \ldots, QI_p \in Z_i$
return $Z_i^{'}$
15 **End for**
16 **end**
17 **else**
18 **for** each QID $QI_e \in Z_i$, where $e = 1, 2, \ldots, p$
19 Obtain hierarchy $H_{QI_e} \in H$ of QI_e
20 Acquire influence score I_{QI_e} of the QID
21 Generalize QI_e values at lower level of H_{QI_e}
22 **Repeat:** $\forall$ QIDs, $QI_1, QI_2, \ldots, QI_p \in Z_i$
return $Z_i^{''}$
23 **End for**
24 **end**
25 **End for**
26 Combine results of both cases, (i.e., $Res =$ combine $(Z_1^{''} + Z_2^{''} +, \ldots, Z_n^{''}$ and $Z_1^{'} + Z_2^{'} +, \ldots, Z_n^{'})$.
27 $T^{'} = T^{'} \cup Res$
28 **return** $T^{'}$

In Algorithm 4, cluster set Z, hierarchy set H of the QIDs, set I of QID influential scores, and privacy parameter k are given as input. Anonymized data $T^{'}$ is obtained as output. Line 1 initializes $T^{'}$ as an empty set. In lines 2-3, every cluster extracted from set Z undergoes a check for balance/imbalance. If a cluster turns out to be imbalanced (i.e., it has more records than k), then the Div value is computed and compared with the threshold (lines 4-6). If Div is higher, the cluster is treated as balanced concerning SA values (lines 7-8) and is sent for lower anonymization (lines 18-23). If a cluster is imbalanced (i.e., required SA values are not present, and distribution is skewed), then higher-level anonymization for influential QIDs is performed to preserve privacy (lines 10-15). The balanced clusters always undergo lower-level anonymization to yield higher utility (lines 18-23). Later, the results of both categories are combined and stored in $T^{'}$. Finally, $T^{'}$ is obtained as output.

To resolve the PUT, three distinct algorithms were used. The main reason to use RF for I computation is its capability to produce higher accuracy, parameters setting is easy, and acceptable success in identical tasks. This task cannot be accomplished via decision trees due to less reliability in outcomes (i.e., dependence on a single tree's result) and the inability to handle multivariate interactions among QIDs. Further, SVM may not give a desirable performance while computing I, in particular, when the domain of SA is higher than $|T|$. The reason to utilize CS for computing homophily among users is its ability to yield reliable results and co-working with the optimization algorithms. Furthermore, it is one of the most widely used measures in recommender systems for an identical task (i.e., similarity estimation). The other measures such as hamming distance and domain-based grouping cannot ensure the consistent ranges of QID's values. For computing Div, communities/biodiversity concepts are adapted to ensure desired privacy protection levels in each cluster. The alternate solutions for this task are the enforcement of hard constraints (i.e., ℓ, α, t, etc.) with the pre-determined value that can destroy the structure of T. In some cases, enforcement of these constraints becomes impossible due to imbalance/sparsity issues in T. All three chosen algorithms contributed significantly to adjust the degree of anonymity, leading to the effective resolution of two conflicting goals.

4 Experimental Evaluation

This section demonstrates the results obtained from the experiments. We conducted detailed experiments on four real-world benchmark datasets, and compared the results of our method with three SOTA anonymization algorithms and one SOTA anonymity model. In the following subsections, we present the dataset descriptions, experiments setup, the evaluation metrics for comparing the method's performance, and the results' comparisons with prior solutions.

4.1 Descriptions of Datasets

During experimental evaluations, we used a relational T encompassing individuals' information (QIDs $\rightarrow$ identity information, SA $\rightarrow$ sensitive information). We evaluated and compared our proposed method's performance on the Adults [39], Bkseq [40], Careplans [41], and Diabetes 130-US hospitals [42] datasets. The Adults dataset encompasses

four QIDs and a single SA. The latter dataset contains three QIDs and one SA. In Bkseq dataset, weight was classified as QID because some data items of this dataset were omitted at the time of public release. The form of the tuple was: user-id , age, gender, weight, omit, omit, omit, omit, and disease. Hence, to perform experimental analysis from a broader scope, all available information was utilized. The careplans dataset belongs to "SyntheticMAss" (an online repository hosting about 1 million medical records). We used its $10k\_synthea\_covid19\_csv$ folder that contains a variety of public and private information about patients. We used five QIDs and one SA from it in our experiments. We ignored some DIs and other non-QID attributes from it. In this dataset, some fields were empty, and therefore, pre-processing was carried out to make it error-free. The Diabetes 130-US hospitals dataset is massive healthcare data concerning diabetes diagnosis in the US. We employed the substantial # of records having four QIDs and one SA from this dataset. We present a brief overview of all four datasets in Table 2. These datasets are publicly available and have been widely used for evaluating the feasibility of anonymity solutions. All datasets were pre-processed as described earlier before actual utilization in the experiments. After applying sophisticated pre-processing, the dimensions of the datasets were $32,561 \times 5$, $16,160 \times 4$, $12,352 \times 6$, and $20,501 \times 5$, respectively.

TABLE 2: Overview of the datasets used in experiments.

Dataset	Total records	Name of QID (Cardinality, Type, H levels)	SA name (Unique values)
Adults [39]	32,561	Age (74, Numerical, 7) Gender (2, Categorical, 2) Race (5, Categorical, 3) Country (41, Categorical, 4)	Salary/Income (2)
Bkseq [40]	16,160	Age (30, Numerical, 5) Gender (2, Categorical, 2) Weight (30, Numerical, 3)	Medical test results (19)
Careplans [41]	12,352	Relationship (3, Categorical, 3) Gender (2, Categorical, 2) State (14, Categorical, 3) Race (5, Categorical, 3) Zip_code (291, Numerical, 5)	Healthcare_coverage (9,535)
Diabetes [42]	20,501	Race (5, Categorical, 3) Gender (2, Categorical, 2) I_status (4, Categorical, 3) Age (32, Numerical, 7)	DiabetesMed (2)

4.2 Details of the Experimental Settings

All experiments were carried out with an Intel Core i5-3320M CPU @ 2.60GHZ running Windows 10 Professional and 8GB RAM. The results were produced with the assistance of two reliable software packages: Matlab version 9.11.0.1687835 (R2021b) 64-bit (win64) and RTools with R version 4.0.0 ($x64$) with built-in package support. The two main libraries used in computing influence were randomForest [2] and ranger [3] (a fast implementation of RF). Descriptions of other useful parameters and other required variables employed for the QID influence computations are given in Table 3. We used concise form (e.g., A,B,C, and D) to refer to each dataset in Table 3. Besides the values mentioned in Table 3, some parameters' default values were also used. For example, bootstrapping was used as a default sampling scheme, and tree complexity was controlled using the default value for node size (e.g., 1). We determined the

2. https://cran.r-project.org/web/packages/randomForest/
3. https://cran.r-project.org/web/packages/ranger/

TABLE 3: Parameters/variables used for computing the influence of QIDs in T.

Parameters/variable name	Parameter's values (Adults [39], Bkseq [40], Careplans [41], Diabetes [42])	
	Numerical (A,B,C,D)	Non-numerical
No. of records in d_{train}	21,815, 10,827, 8,285, 13,720	-
No. of records in d_{test}	10,746, 5,333, 4, 067, 6,781	-
$ntree$ value	491, 264, 221, 291	-
Desired model type	-	Classification (A,B,D) & Regression (C)
Splitting rule	-	Impurity
QID importance	-	true
$mtry$ value	4, 3 , 5, 4	-
Minimum node size	1	-
Keep forest	-	true
Sample. fraction	0.8	-
Predictors' label	-	age, gender, race, country (A)
	-	age, gender, weight (B)
	-	relationship, gender, state, race, zip_code (C)
	-	gender, age, race, ins-status (D)
Target class label	-	SA value of the respective dataset

values of parameters/variables from extensive experiments under different settings. Figure 6 presents the I values of different QIDs present in all four datasets listed in Table 2.

The reason behind the highest I value for age in adults and bkseq datasets is the higher distinctiveness in the values. However, for the race QID in the adults dataset, the majority of the records had white as the race value, and the remaining records listed four other races that likely appear only when k is small. Therefore, the I value of race was lower, and had a lower effect on individual privacy. Similarly, in the Bkseq dataset, gender had lower I values because most records shared the same gender information. In this dataset, the difference in I values was not sufficiently large for one QID due to the more unique values of the SA (e.g., 19) than in the Adults dataset (e.g., 2). In addition, this dataset had balanced distributions for most QID values. In the last two datasets, relationship and insulin status are the most influential QIDs based on higher distinctiveness in their values. We verified these I values by analyzing each QID's original value distribution and domain in T. We performed validation to ensure the correctness of the I values. The validation results showed that the I values found by RF were highly reliable and accurate. The α_{pr} and

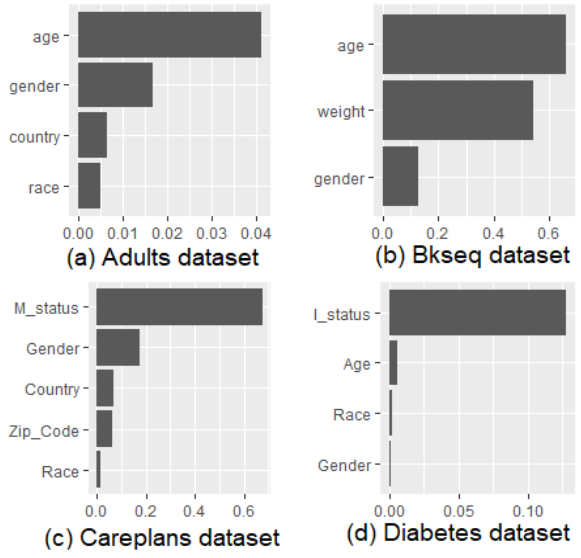


Fig. 6: Computed I scores of QIDs present in four datasets.

β_{ut} values were both set to 0.5. T_{CS} and T_{Div} values were set to 0.9 and 0.7, respectively, but k values were chosen differently in each test (i.e., from 2∼400). However, these

hyper-parameters (i.e., α_{pr}, β_{ut}, T_{CS}, and T_{Div}) can influence privacy strength and data utility. Therefore, we present the numerical evaluation results in Figure 7 based on which the optimal values were chosen. The optimal values satisfy the privacy-utility curve properties and ensure narrow gaps between privacy and utility results.

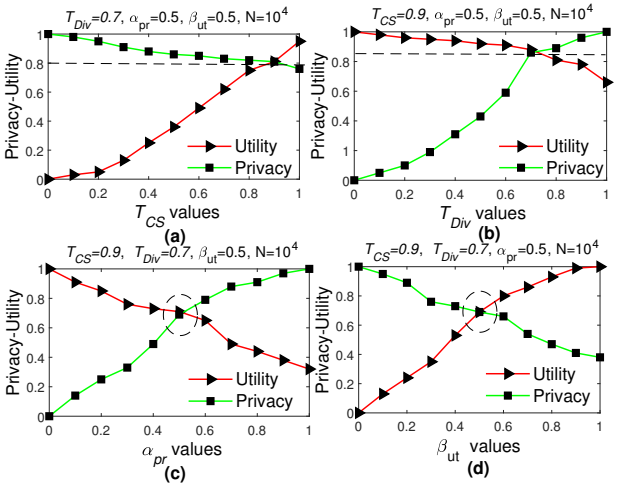


Fig. 7: Numerical analysis of privacy strength and $T^{'}$ utility.

This study involves various hyper-parameters that were determined through four different ways: (i) domain knowledge, (ii) multidisciplinary techniques applied, (ii) target objectives (privacy vs. utility), and (iv) datasets used in the evaluation. The optimal values were determined through extensive experimentation and sensitivity analysis of hyper-parameters w.r.t. target objectives. The optimal hyperparameter set (i.e., α_{pr}, β_{ut}, T_{CS}, and T_{Div}) is the same across different datasets. Moreover, some parameter values used in computing I can vary depending on the type of the dataset.

4.3 Evaluation Metrics and Feasibility Testing Criteria

To verify the feasibility of our method, we employed four metrics. Two metrics were used to measure the privacy of the individuals, and two metrics were used for utility evaluation. The chosen metrics are most appropriate for quantifying privacy and utility technically considering the nature of data (i.e., relational) [43]. To evaluate privacy, we used disclosure risk (i.e., re-identification probability) and exposure of dominant SA values in clusters. The disclosure risk (which measures the probability that an attacker can infer any sanitized record's SA values successfully), D, can be computed using Eq. 13.

$$D(T^{'}) = \frac{\sum_{u \in T'} D(u^{'})}{|T'|} \tag{13}$$

where $D(u^{'})$ denotes the probability of successfully extracting any SA of $u^{'}$,which is formulated below.

$$D(u^{'}) = \begin{cases} 0, & \text{if } u^{'} \notin Z^{'} \\ \frac{\sum_{v \in S_{Z'}} max\{1/|z^{'}|, Pr_{Z'}(v)\}}{|S_{Z'}|}, & \text{if } u^{'} \in Z^{'} \end{cases} \tag{14}$$

where $Pr_{Z'}(v)$ denotes the frequency of SA value v in anonymized cluster $Z^{'}$, and $S_{Z'}$ is the set of SA values in $Z^{'}$. There are two important concepts in this function, record identity and SA disclosure. The SA value is also inferred if an attacker can determine the record of the target. This can be expressed as the probability $1/|z^{'}|$. Besides, an attacker can also infer the SA value v with probability $Pr_{Z'}(v)$. So the resulting probability is $max\{1/|z^{'}|, Pr_{Z'}(v)\}$. More knowledge about the above equations can be obtained from [44]. $D = 1$ only when all users in a cluster share the same SA. We considered worst-case analysis (journalist scenarios) while comparing and computing D values. The value of the second metric (i.e., exposure E of dominant SA value dv_i) can be determined using Eq. 15.

$$E_{dv_i} = \frac{\sum_{i|dominant\ value\ of\ the\ SA} dv_i}{\sum_{i=1}^{|Z|} dv_i} \tag{15}$$

The numerator in Eq. 15 denotes the dominant value frequency in a cluster/sample, while the denominator represents the frequency of the dominant value in T. In other words, the numerator is an SA's dominant frequency in the sample, and the denominator is the SA's dominant frequency in the population. Both these metrics were used to analyze the level of privacy offered by the proposed method.

To evaluate the utility of $T^{'}$, we used information loss (IL) and a count of the clusters that can lead to an infeasible query result (IQR). To measure IL, we used distortion measure (DM), which is a highly reliable measure [43]. DM values can be calculated by analyzing the level of H on which QID values are generalized. The value of DM is calculated using Eq. 16.

$$DM = \sum_{QI=1}^{p} \frac{l_a}{l_t} \times I_{QI} \tag{16}$$

where l_a denotes the level at which the QID value is generalized, and l_t denotes the total # of levels in the hierarchy, while I_{QI} represents the influence value of the QID. The DM values from each QID and record are summed for analysis and comparisons. Since the proposed method is k-anonymity-based, and therefore, the DM for each record is the same in a particular cluster. We used the average DM value to compare our method with SOTA methods. If the value of a particular QID is retained as it was originally in T, the DM value will be 0. The value of the other metric is determined via Eq.17.

$$IQR = \sum_{i|cluster\ leading\ to\ infeasible\ query\ results} Z_i^{'} \tag{17}$$

The criteria on which the results were computed and compared were 64 anonymized versions of four datasets resulting from different scales (e.g., small, and large) of k.

4.4 Performance against Existing Anonymity Solutions

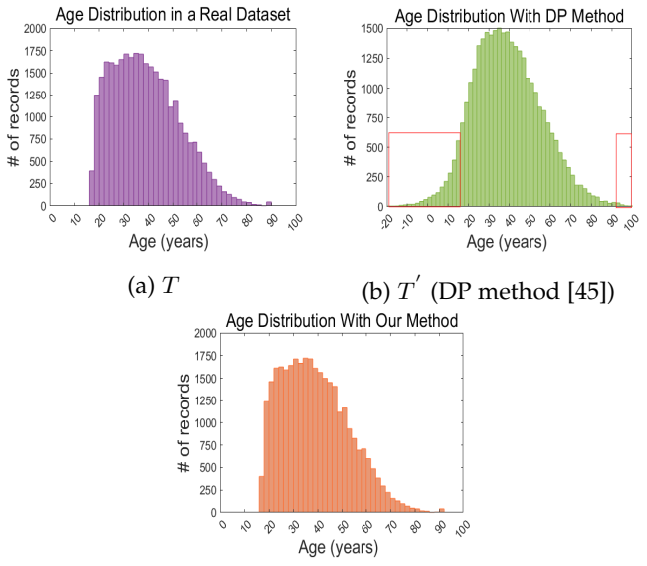
To benchmark our method, we compared the results with three prior state-of-the-art algorithms: the MWCK algorithm [28], the RKA algorithm [29], and the ELD algorithm [30]. All these algorithms are recent and yield competitive results w.r.t. privacy and utility. The RKA algorithm [29] employs noise addition and randomization operations, and is an extended version of the DP model. According to the DP model,

a randomized function F yields ϵ-DP if for all original data sets D_1 and D_2 that differ by at most one tuple, and for all $S \subseteq Range(F)$,

$$Pr[F(D_1) \in S] \leq exp(\epsilon) \times Pr[F(D_2) \in S] \quad (18)$$

DP can be realized via Laplace and exponential mechanisms. The former is suitable for DP application to numerical data and the latter is appropriate for categorical data. Since RKA algorithm [29] is the first DP implementation for achieving k-anonymity, therefore, we compared our results with this version of the DP model. To enhance the persuasiveness of our work, we used the ℓ-diversity [5] privacy model as a baseline in the experimental evaluation and compared results against it. Although ℓ-diversity has higher recognition in data privacy, it performs anonymization by enforcing constraints on SA values that degrade utility.

Many ramifications of the DP technique have been proposed, but most of them optimize either privacy or utility. Furthermore, the ϵ parameter greatly affects the equilibrium between privacy and utility, and the quality of resulting $T^{'}$ can be degraded. To prove the superiority of our method against DP techniques, we compared the $T^{'}$ quality with the SOTA method named, MRMondrian [45]. The results of the experiment concerning T and $T^{'}$ in terms of stability in QID values are shown in Figure 8.



(a) T (b) $T^{'}$ (DP method [45])

(c) $T^{'}$ (Proposed method)

Fig. 8: Comparison of stability in QID values in T and $T^{'}$.

From the results, it can be noticed that our method (shown in Fig. 8(c)) has preserved higher stability in QID values (i.e., the distributions are highly alike to that of T). In contrast, the MRMondrian method (shown in Fig. 8(b)) has lower stability in QID values (i.e., the distributions are minimally aligned to that of T), and some values were produced as negative (marked with a red square) for age QID, which can change the semantics of knowledge enclosed in T. These results fortify the efficacy of the proposed method in real-life scenarios. To provide reasonable utility in PPDP, a large value of ϵ is desirable which can lower the privacy guarantees [46]. Considering the difficulty in specifying the optimal ϵ value, and the inability to favorably resolve the PUT in data releases, our method can be handy to make data broadly available while effectively resolving PUT. In the next subsections, we present quantitative results obtained through extensive tests on four real-life datasets.

4.4.1 Comparisons of Privacy Preservation

In this subsection, we evaluate and present the performance of our method based on two privacy metrics: disclosure risk (i.e., D), and exposure of SA dominant values (i.e., E). The formalization of these metrics is given in Eqs. 13 and 15.

(i) Reduction in D: The first experimental analysis evaluated and compared D values yielded from experiments on the four chosen datasets. Before presenting the numerical value of D obtained from each anonymized dataset with varying k, we illustrate in Figure 9 an overview of the results generated from a small subset of T.

(**a**) Original data.

	Quasi Identifiers		SA
ID	Age	Country	Disease
1	75	Greenland	Leukemia
2	75	Canada	Rhinitis
3	78	Belize	Leukemia
4	80	Belize	Leukemia
5	77	Canada	Rhinitis
6	77	Canada	Rhinitis

(**b**) 2-anonymous data (MWCK Algorithm).

	Quasi Identifiers		SA
Cluster	Age	Country	Disease
Z'_1	75-77	North America	Leukemia
	75-77	North America	Rhinitis
Z'_2	78-80	Central America	Leukemia
	78-80	Central America	Leukemia
Z'_3	75-77	North America	Rhinitis
	75-77	North America	Rhinitis

(**c**) 2-anonymous data (Proposed Method).

	Quasi Identifiers		SA
Cluster	Age	Country	Disease
Z'_1	75-80	North America	Leukemia
	75-80	North America	Rhinitis
Z'_2	75-80	America	Leukemia
	75-80	America	Leukemia
Z'_3	75-80	America	Rhinitis
	75-80	America	Rhinitis

Privacy results analysis	(**a**) Original data.	(**b**) 2-anonymous data (MWCK Algorithm).	(**c**) 2-anonymous data (Proposed Method).
	Disclosure risk in record linkage attack : Scenario I	**Name** = Lin, **Age** = 79, **Country** = Belize **D** (Disease is **Leukemia**) = 2/2= **1**	**Name** = Lin, **Age** = 79, **Country** = Belize **D** (Disease is **Leukemia**) = 2/4= **0.5**
	Disclosure risk in record linkage attack: Scenario II	**Name**= Tuan, **Age** = 75, **Country**= Canada **D** (Disease is **Rhinitis**) = ¾= **0.75**	**Name**= Tuan, **Age** = 75, **Country**= Canada **D** (Disease is **Rhinitis**) = ½ or 3/6 = **0.5**

Fig. 9: Comparative analysis of individual privacy preservation in $T^{'}$.

In this analysis, Figure 9 (a) represents the T to be anonymized, and Figures 9 (b) and (c) show the corresponding $T^{'}$ versions created by the MWCK algorithm and our method, respectively. Although SA value distribution is uniform (e.g., 50, 50) in this small subset, Div varies in each cluster according to the k value, leading to two clusters ($Z^{'}_2$ and $Z^{'}_3$) with no diversity in SAs. Hence, privacy protection in these clusters is challenging; most existing methods often overlook Div, and they do not pay ample attention to protecting privacy in imbalanced clusters based on QID information. Meanwhile, our method uses the I information to address privacy issues in these circumstances by using higher-level anonymization. Hence, in both scenarios, our method had a lower D compared to the prior methods. The numerical values of D obtained from different versions of $T^{'}$ created with distinct k values are shown in Figure 10. In these experiments, we assume the attacker already has some individuals' data, and he/she executes linking to infer SAs. Generally, for the higher value of k, the avg. disclosure risk should be low, however, we used the most dominant SA values in computing D which is why it does not show a decreasing trend. Also, the higher value of k can lower identity disclosure risk, the SA disclosure risk is subject to the distribution of SA value in the respective cluster.

During evaluation, we conducted experiments on two biomedical datasets, named careplans [41] and diabetes [42]. The obtained D results, and their comparisons on fifteen different values of k are plotted in Figure 11. From the results, we can note that our method has yielded better results

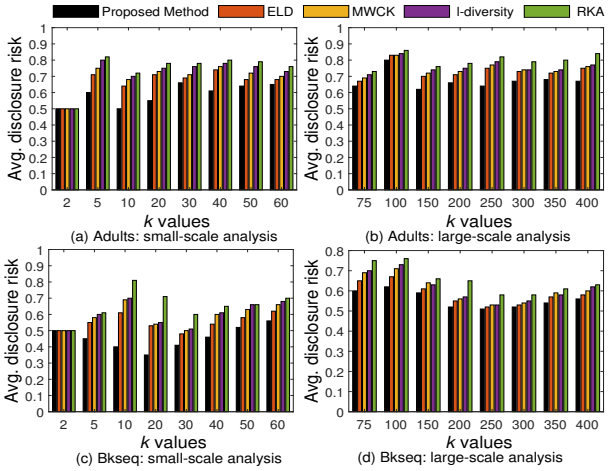


Fig. 10: Average D: Proposed method versus existing algorithms.

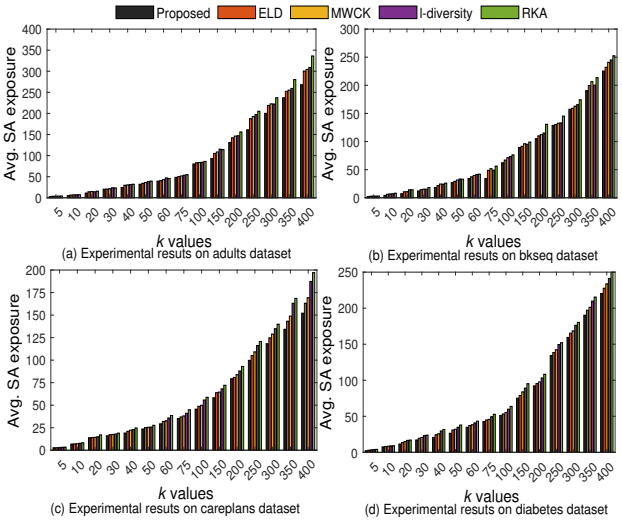


Fig. 12: Average E: Proposed method versus existing algorithms.

than existing methods. The lower D values in the care plans dataset are due to the higher variation in SA values and the numerical nature of SA. In contrast, the diabetes dataset is relatively skewed, and the categorical nature of SA. In our method, the maximum values of D on both these datasets were 0.58 and 0.68, respectively. However, in the existing methods, the maximum values of D on these datasets were reached up to 0.68 and 0.84, respectively.

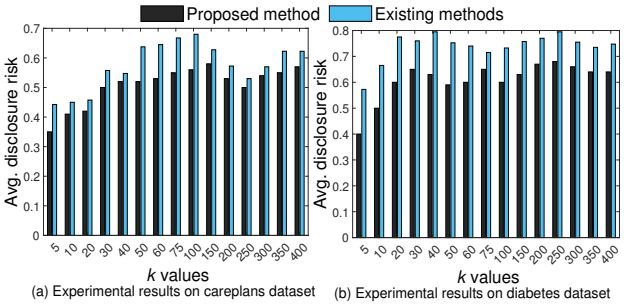


Fig. 11: Average D: Proposed method versus existing algorithms.

(ii) Reduction in E: In the next experiments, we computed and compared the E values from 32 versions of $T^{'}$. To assess the effectiveness of methods regarding privacy protection in generic cases, we selected the most dominant SA values that likely appear with abundance in clusters. Although such values do not affect someone's privacy directly, they provide sufficient information about the characteristics of T. Hence, explicit disclosure of someone's SA, or a prediction of it, can be made regarding an unknown community (e.g., a marketer scenario). As shown in Figure 12, E increases with k in all four datasets. However, our method does not expose the SA more frequently, and therefore, it effectively preserves individual privacy in most cases. From the results presented in Figures 10, 11, and 12, we can see that the proposed method offers a higher-level safeguard against linking and SA-exposure attacks, and it addresses those issues in imbalanced clusters, which is not properly addressed by previous clustering-based solutions. These results prove the significance of our method in safeguarding privacy.

Performance against AI-powered attacks: Recently, AI-based attacks are creating unexpected privacy risks by extracting sensitive data from large and high dimensional T. These attacks have abilities to compromise people's privacy by statical matching $T^{'}$ with auxiliary data. To enhance the persuasiveness of our work concerning major privacy risks, we evaluate our method against two AI-powered attacks namely, SA prediction and T reconstruction. In the former attack, the attacker trains an AI model with $T^{'}$ and external sources data and predicts the SA of unknown people based on QIDs. In the later attack, the attacker re-constructs the parts (or whole) of T using $T^{'}$ and other auxiliary data. The results of both attacks in terms of accuracy and probability are given in Table 4. Due to the imbalance in SA values, attacks on A and D can be launched more accurately.

TABLE 4: Results against AI-powered attacks: our method versus prior methods.

Methods	Attack I (SA Prediction)				Attack II (T Reconstruction)			
	A	B	C	D	A	B	C	D
ℓ-diversity Model [5]	31.21	23.21	22.09	26.01	0.24	0.19	0.20	0.22
MWCK Algorithm [28]	25.43	19.72	18.06	20.31	0.15	0.11	0.14	0.18
RKA Algorithm [29]	33.01	26.45	28.43	31.62	0.13	0.08	0.11	0.14
ELD Model [30]	21.75	15.95	14.61	17.22	0.19	0.15	0.18	0.19
Proposed Method	17.05	9.09	13.09	15.05	0.10	0.06	0.08	0.09

4.4.2 Comparisons of the Utility of $T^{'}$

In this subsection, we evaluate the feasibility of our method concerning utility enhancement in $T^{'}$ based on two metrics: DM (Eq. 16) and IQR (Eq. 17). Although utility loss is inevitable when QID values change from T to $T^{'}$, the change can be controlled by generalizing values to lower levels of H by making use of Div and I information. Our method performs minimal and required generalization in a cluster set to enhance the utility of $T^{'}$. Higher-level generalization is performed only when the SA values are less diverse in the cluster. However, the existing methods suppress some records and use imprecise operations (i.e., $\geq$, $\leq$, **, etc.), and

in most cases, utility loss is very higher. Most prior methods perform over-generalization of data leading to poor utility. In contrast, our method resolves these issues by performing only the required generalizations of QIDs.

(i) Reduction in IL: To verify the efficacy of our method concerning utility preservation, we computed IL using the DM metric from 16 versions of each $T^{'}$. In each test, we measured and compared the results with prior algorithms. Numerical analysis is presented in Figure 13. From the results, we observe that IL increases with k. However, our method yields a lower IL than the existing algorithms for most k values. The $T^{'}$ produced by our method preserves more information for researchers that, in return, can help them perform analyses of $T^{'}$ with minimal post-processing. These improvements have been achieved by controlling the excessive generalization issues. The IL results and their comparisons on the two biomedical datasets are shown in Table 5. These results were obtained and compared with the four prior SOTA methods for sixteen different values of k.

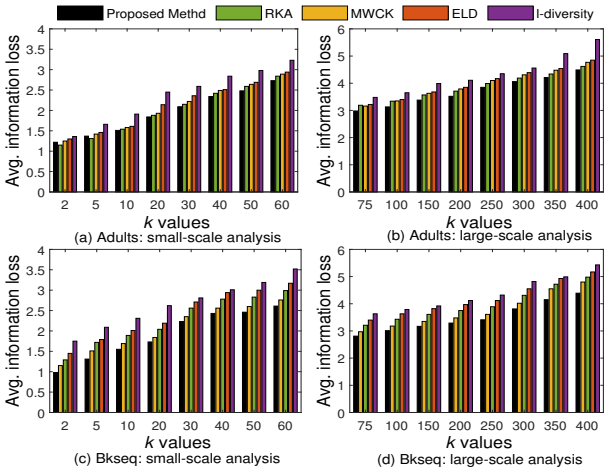


Fig. 13: Average IL: Proposed method versus existing algorithms.

TABLE 5: Avg. IL: Proposed method versus prior methods.

Dataset	k	Anonymity solutions and their IL results				
		Proposed Method	ℓ-diversity [5]	MWCK [28]	RKA [29]	ELD [30]
Careplans [41]	2	1.42	1.56	1.51	1.47	1.53
	5	1.57	1.77	1.67	1.63	1.71
	10	1.76	2.01	1.89	1.82	1.95
	20	1.97	2.22	2.11	2.05	2.15
	30	2.23	2.51	2.34	2.29	2.43
	40	2.43	2.76	2.59	2.52	2.63
	50	2.75	3.26	2.94	2.85	3.02
	60	2.94	3.45	3.11	2.99	3.17
	75	3.11	3.61	3.25	3.19	3.35
	100	3.42	3.92	3.56	3.49	3.68
	150	3.57	4.17	3.73	3.65	3.87
	200	3.71	4.42	3.92	3.82	4.15
	250	3.96	4.58	4.28	4.19	4.39
	300	4.13	4.72	4.37	4.29	4.55
	350	4.61	5.21	4.73	4.69	4.84
	400	5.09	5.76	5.38	5.21	5.63
Diabetes [42]	2	1.31	1.45	1.41	1.38	1.43
	5	1.45	1.60	1.51	1.49	1.55
	10	1.59	1.73	1.69	1.64	1.70
	20	1.78	1.91	1.85	1.83	1.87
	30	1.97	2.26	2.17	2.11	2.21
	40	2.02	2.43	2.27	2.21	2.34
	50	2.26	2.83	2.47	2.41	2.54
	60	2.61	3.04	2.79	2.74	2.84
	75	2.76	3.19	2.91	2.88	2.96
	100	2.91	3.33	3.21	3.14	3.27
	150	3.04	3.99	3.43	3.37	3.58
	200	3.19	4.24	3.49	3.41	3.75
	250	3.33	4.45	3.67	3.59	3.86
	300	3.67	4.64	3.75	3.71	3.98
	350	4.01	4.91	4.18	4.14	3.34
	400	4.23	5.25	4.47	4.32	4.65

(ii) Reduction in IQR: In the last set of analyses, we analyzed and compared the frequency of clusters that can lead to infeasible results when queries are executed on them. Due to imprecise operations used in an anonymization process ($*, \leq, \geq$, etc.), in most cases, the clusters yield query results that have a poor link (or no link) with the underlying QID/SA values in T (garbage-in garbage-out). We present in Figure 14 a comparative analysis of the average number of clusters leading to an IQR.

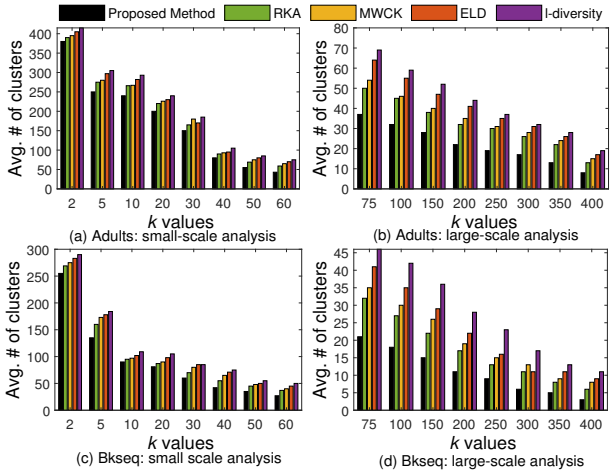


Fig. 14: Average IQR (i.e., the number of clusters leading to infeasible query results): Proposed method versus existing algorithms.

From the results in Figure 14, we note that the proposed method has lower IQR values in most cases. Although the proposed method does not use imprecise operations, we consider infeasible those clusters in which the generalization interval is relatively wide (e.g., a relatively higher offset from the original QID values), and the probability of infeasible query results can increase. These results fortify the efficacy of our method regarding utility preservation, and the $T^{'}$ resulting from our method can be highly applicable in knowledge-based applications such as analyzing income relationships with different combinations of demographics, disease analysis w.r.t certain age groups, and general data mining tasks. In addition, we performed experiments on two biomedical datasets. The obtained IQR results and their comparative analysis is shown in Table 6. As shown in Table 6, IQR values decrease with an increase in k due to a decrease in the # of clusters. Better results on both T prove the efficacy of our method in the healthcare sector.

TABLE 6: Average IQR on biomedical T: our method versus prior methods.

Methods	Sixteen different values of the privacy parameter (i.e., k) and corresponding IQR results.															
	2	5	10	20	30	40	50	60	75	100	150	200	250	300	350	400
ℓ-diversity [5]	180.1	168.2	161.1	144.9	135.8	120.5	112.0	104.5	62.6	59.5	49.9	39.4	31.1	21.2	19.1	15.9
MWCK [28]	173.2	160.8	154.9	132.8	124.6	115.8	104.2	94.3	58.4	48.2	41.1	29.3	24.5	18.3	13.8	11.54
RKA [29]	170.8	155.9	149.7	128.3	121.3	110.8	99.3	89.1	54.4	46.1	38.8	26.0	23.9	16.4	10.3	9.1
ELD [30]	176.2	165.1	157.1	138.9	128.2	119.3	109.7	98.1	60.1	56.8	45.1	37.3	27.1	20.5	16.0	12.2
Our Method	159.2	134.4	122.7	98.0	92.3	87.2	79.1	65.9	35.4	29.1	19.2	16.8	12.0	5.1	4.3	3.1
ℓ-diversity [5]	321.1	213.45	118.4	111.5	98.2	86.7	67.2	62.9	43.7	39.5	33.3	24.5	21.1	15.8	14.2	12.9
MWCK [28]	316.2	198.3	111.5	104.3	83.2	70.3	60.1	56.3	39.2	33.7	29.4	17.3	15.9	12.6	10.3	8.1
RKA [29]	308.1	190.3	108.8	99.3	82.4	69.7	49.2	44.3	36.5	30.8	25.2	16.9	12.2	10.4	9.5	7.3
ELD [30]	318.9	201.4	113.8	108.4	96.3	83.2	63.4	58.7	40.9	35.3	31.7	20.5	17.6	12.3	11.0	9.6
Our Method	295.1	161.3	101.9	92.6	71.1	55.4	47.8	42.6	34.2	27.9	17.3	13.2	10.8	7.5	6.1	5.4

It is important to note that most studies remove NSAs and DIs from T. However, DIs are something that we need

to remove to prevent explicit identity disclosure, but NSAs can enhance data utility in some cases. To verify this claim, we performed experiments on three real-world datasets in which NSAs exist, and analyzed their impact on data utility with respect to original T. The experimental results obtained from three datasets using accuracy as an evaluation metric are shown in Table 7. The difference in accuracy for D dataset before and after NSA removal is relatively large due to the higher # of NSAs in it.

TABLE 7: Impact of NSAs on data utility with respect to T.

Utility analysis	Three real-world benchmark datasets		
	Adults [39]	Careplans [41]	Diabetes [42]
Accuracy (with NSAs)	88.14	82.05	84.69
Accuracy (without NSAs)	86.31	81.58	80.89
%age difference in accuracy	1.83	0.47	3.80

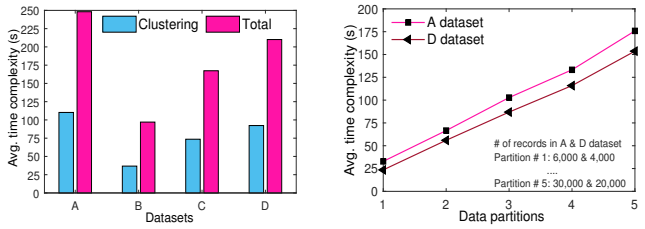
Through extensive testing and comparisons with the SOTA methods, our method yielded better privacy and utility results on all four datasets. The average privacy improvements on two metrics (i.e., D and E) than existing SOTA algorithms on each T are listed as: (A,16.3%, 13.1%), (B,13.9%, 10.1%), (C,16.3%, 12.5%), and (D,16.6%, 9.9%). The average utility enhancements on two metrics (i.e., DM and IQR) than existing SOTA algorithms on each T are listed as: (A,15.6%, 17.5%), (B, 8.5%,23.1%), (C,9.4%,17.6%), and (D,10.6%,22.7%). A comparative analysis of overall results in terms of numbers and percentages are shown in Table 8.

TABLE 8: Summary of improved results in %age compared to SOTA methods.

Objective	Metric/Method	Average improvements (%)
Privacy preservation (Traditional)	D & E	14.42 % & 11.39 %
Privacy preservation (AI-powered)	Accuracy & Probability	20.39 % & 24.57 %
Utility enhancements (data analytics)	DM & IQR	11.25 % & 20.21 %
Privacy results comparison	Against Syntactic Method [5]	18.22 %
	Against Clustering Methods [28], [30]	14.62 %
	Against DP Method [29]	11.35%
Utility results comparison	Against Syntactic Method [5]	16.86 %
	Against Clustering Methods [28], [30]	13.06 %
	Against DP Method [29]	15.11%
Reduction in clustering's complications	# of iterations	2.25× lower

Limitations: Although our method yielded better results compared to SOTA methods, the highly skewed distribution of most attributes can degrade its performance. The accuracy of hypothesis generation using $T^{'}$ can be low when the T is poisoned (e.g., a malicious/influential subset of the users aims to corrupt the decision-making towards some minority groups). Lastly, computing complexity can increase with the horizontal and vertical expansion of T.

Complexity analysis: The most critical parts of the proposed method are I computation, CS computation, CS and Div based clustering, and the need-based generalization. Since the fast implementation of RF was used in experiments therefore the complexity of the whole I computing process in $O(n)$. The complexity of algorithm 1 highly depends on the two 'for' loops (e.g., lines 6 and 7), and the set of instructions between them. Therefore, the overall complexity is $O(\lambda n^2)$. However, the upper bound of λ is constant, hence the overall complexity is $O(n^2)$. The complexity of algorithm 2 lies in steps 8 to 14. The 'for' loop iterates n times depending on the k value, and a set of instructions is executed in each iteration. The total complexity is $O(\Lambda n/k) = O(n/k)$ because Λ has a constant upper bound. The complexity of algorithm 3 lies in steps 6 to 28. However, most of the statistics are pre-computed, the complexity of the algorithm 3 is $O(n)$. The complexity of algorithm 4 mainly lies in steps 2 to 25. For each cluster balance/imbalance check is performed initially, and anonymization is performed accordingly. The 'for' loop performs n iterations, and in each iteration, a set Φ is executed. Therefore, the complexity of the whole process is $O(\Phi n) = O(n)$, because Φ has a constant upper bound. With the pre-computed values of three statistics (CS, Div, and I), computing overheads are not significantly high. The quantitative results of time complexity (in seconds) that were obtained from real-world datasets are demonstrated in Figure 15. For better analysis, we arranged the results into two settings: utilizing whole records of each dataset and by varying # of records from two datasets (e.g., A and D) that are relatively large. We partitioned these datasets based on records and analyzed the time complexity. Figure 15 (a) presents the total time as well as the clustering time which is one of the expensive operations in the proposed method.



(a) Time complexity using whole T.

(b) Time complexity w.r.t # of records.

Fig. 15: Analysis of time complexity of proposed method.

From the results, it can be observed that the proposed method has acceptable time complexity in both settings, and time complexity does not blow up significantly while varying # of records. Based on these results, it can be concluded that the proposed method can scale well to larger datasets.

5 Conclusion

In this paper, we proposed a novel anonymization method using similarity and diversity based clustering. Specifically, we determined answers to three important research questions concerning personal data anonymization via detailed theoretical and experimental analysis. To the best of our knowledge, these questions and their answers are not already known in the literature. *First*, most existing approaches ($\sim$ 2022) often assume that underlying data to be anonymized is always in a balanced (e.g., values distributions are uniform) form. However, this assumption rarely holds, and most real-world datasets can be highly imbalanced (interestingly, we found that the adults and diabetes datasets are also imbalanced for some attributes), and therefore, the existing solutions can have limited applicability in anonymizing them. In contrast, our method relaxes this assumption and applies to any real-world dataset balanced/imbalanced. *Second*, the current studies have various

limitations in critical aspects e.g., some do not extract useful information about QIDs to concentrate more on identity vulnerable QIDs, mainly focused on one part of T either QIDs or SA, used wider generalization levels while transforming T to $T^{'}$ that can degrade $T^{'}$ quality, and overlooked powerful information about QIDs that is needed to ensure needs-based generalization. To address these limitations, we adapted multidisciplinary concepts that improved various critical parts of the anonymization and made our solution flexible, robust, and applicable to most real-world scenarios. *Third*, existing methods often degrade $T^{'}$ utility while providing safeguards against privacy threats (e.g., preference is given to one metric at the expense of another). However, the careful balance of both privacy and utility is extremely important in accomplishing multiple scientific and business goals such as conducting medical research, validating a hypothesis, and/or generating a new hypothesis. To this end, our method offers the first generic solution for consistently optimizing the privacy-utility trade-off using the objective function of clustering. Our experimental findings from four datasets are expected to improve/rectify the anonymization process from both technical and theoretical perspectives.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2020R1A2B5B01002145).

REFERENCES

[1] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.

[2] S. K. Kroes, M. P. Janssen, R. H. Groenwold, and M. van Leeuwen, "Evaluating privacy of individuals in medical data," *Health Informatics Journal*, vol. 27, no. 2, p. 1460458220983398, 2021.

[3] F. YAĞAR, "Growing concern during the covid-19 pandemic: Data privacy," *Turkiye Klinikleri J Health Sci*, vol. 6, no. 2, pp. 387–92, 2021.

[4] 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.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

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

[7] A. V. Bogdanov, A. Degtyarev, N. Shchegoleva, V. Korkhov, V. Khvatov, N. Zaynalov, J. Kiyamov, A. Dik, and A. Faradzhov, "Protection of personal data using anonymization," in *International Conference on Computational Science and Its Applications*. Springer, 2021, pp. 447–459.

[8] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[9] M. R. S. Aghdam and N. Sonehara, "Achieving high data utility k-anonymization using similarity-based clustering model," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 8, pp. 2069–2078, 2016.

[10] M. Sheikhalishahi and F. Martinelli, "Privacy preserving clustering over horizontal and vertical partitioned data," in *2017 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2017, pp. 1237–1244.

[11] J. J. V. Nayahi and V. Kavitha, "Privacy and utility preserving data clustering for data anonymization and distribution on hadoop," *Future Generation Computer Systems*, vol. 74, pp. 393–408, 2017.

[12] W. Zheng, Z. Wang, T. Lv, Y. Ma, and C. Jia, "K-anonymity algorithm based on improved clustering," in *International conference on algorithms and architectures for parallel processing*. Springer, 2018, pp. 462–476.

[13] S. Zouinina, N. Grozavu, Y. Bennani, A. Lyhyaoui, and N. Rogovschi, "Efficient k-anonymization through constrained collaborative clustering," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2018, pp. 405–411.

[14] C. Xia, J. Hua, W. Tong, and S. Zhong, "Distributed k-means clustering guaranteeing local differential privacy," *Computers & Security*, vol. 90, p. 101699, 2020.

[15] Y. Yan, E. A. Herman, A. Mahmood, T. Feng, and P. Xie, "A weighted k-member clustering algorithm for k-anonymization," *Computing*, pp. 1–23, 2021.

[16] L. Kacha, A. Zitouni, and M. Djoudi, "Kab: A new k-anonymity approach based on black hole algorithm," *Journal of King Saud University-Computer and Information Sciences*, 2021.

[17] A. Abbasi and B. Mohammadi, "A clustering-based anonymization approach for privacy-preserving in the healthcare cloud," *Concurrency and Computation: Practice and Experience*, p. e6487, 2021.

[18] S. Mahmood, "The anti-data-mining (adm) framework-better privacy on online social networks and beyond," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 5780–5788.

[19] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10 562–10 582, 2017.

[20] S. Zouinina, Y. Bennani, N. Rogovschi, and A. Lyhyaoui, "A two-levels data anonymization approach," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2020, pp. 85–95.

[21] X. Huang, J. Liu, Z. Han, and J. Yang, "A new anonymity model for privacy-preserving data publishing," *China Communications*, vol. 11, no. 9, pp. 47–59, 2014.

[22] F. Ashkouti, K. Khamforoosh, A. Sheikhahmadi, and H. Khamfroush, "Dhkmeans-ℓ-diversity: distributed hierarchical k-means for satisfaction of the ℓ-diversity privacy model using apache spark," *The Journal of Supercomputing*, pp. 1–35, 2021.

[23] M. Jeon, O. Temuujin, J. Ahn, and D.-H. Im, "Distributed l-diversity using spark-based algorithm for large resource description frameworks data," *The Journal of Supercomputing*, pp. 1–17, 2021.

[24] J. A. Onesimu, J. Karthikeyan, and Y. Sei, "An efficient clustering-based anonymization scheme for privacy-preserving data collection in iot based healthcare services," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1629–1649, 2021.

[25] L. Yang, X. Chen, Y. Luo, X. Lan, and W. Wang, "Idea: A utility-enhanced approach to incomplete data stream anonymization," *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 127–140, 2021.

[26] C.-Y. Lin, "A reversible privacy-preserving clustering technique based on k-means algorithm," *Applied Soft Computing*, vol. 87, p. 105995, 2020.

[27] Y. Canbay, A. Kalyoncu, M. Ercimen, A. Dogan, and S. Sagiroglu, "A clustering based anonymization model for big data," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2019, pp. 720–725.

[28] X. Qian, X. Li, and Z. Zhou, "An efficient privacy-preserving approach for data publishing," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2021.

[29] F. Song, T. Ma, Y. Tian, and M. Al-Rodhaan, "A new method of privacy protection: random k-anonymous," *IEEE Access*, vol. 7, pp. 75 434–75 445, 2019.

[30] W. Zheng, Y. Ma, Z. Wang, C. Jia, and P. Li, "Effective l-diversity anonymization algorithm based on improved clustering," in *International Symposium on Cyberspace Safety and Security*. Springer, 2019, pp. 318–329.

[31] D. Sadhya and B. Chakraborty, "Quantifying the effects of anonymization techniques over micro-databases," *IEEE Transactions on Emerging Topics in Computing*, 2022.

[32] C. Zhang, H. Jiang, Y. Wang, Q. Hu, J. Yu, and X. Cheng, "User identity de-anonymization based on attributes," in *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 2019, pp. 458–469.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] C. Zhang, S. Wu, H. Jiang, Y. Wang, J. Yu, and X. Cheng, "Attribute-enhanced de-anonymization of online social net-

works," in *International Conference on Computational Data and Social Networks*. Springer, 2019, pp. 256–267.
[35] M. Strobel and R. Shokri, "Data privacy and trustworthy machine learning," *IEEE Security & Privacy*, vol. 20, no. 5, pp. 44–49, 2022.
[36] F. Fkih, "Similarity measures for collaborative filtering-based recommender systems: Review and experimental comparison," *Journal of King Saud University-Computer and Information Sciences*, 2021.
[37] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
[38] P. Zeng, Y. Li, P. Hu, D. Peng, J. Lv, and X. Peng, "Deep fair clustering via maximizing and minimizing mutual information," *arXiv preprint arXiv:2209.12396*, 2022.
[39] D. Newman, "Uci repository of machine learning databases, university of california, irvine," *http://www. ics. uci. edu/mlearn/MLRepository. html*, 1998.
[40] F. Amiri, N. Yazdani, A. Shakery, and A. H. Chinaei, "Hierarchical anonymization algorithms against background knowledge attack in data releasing," *Knowledge-Based Systems*, vol. 101, pp. 71–89, 2016.
[41] Z. El Ouazzani, A. Braeken, and H. El Bakkali, "Proximity measurement for hierarchical categorical attributes in big data," *Security and Communication Networks*, vol. 2021, 2021.
[42] K. Sujatha and V. Udayarani, "Chaotic geometric data perturbed and ensemble gradient homomorphic privacy preservation over big healthcare data," *International Journal of System Assurance Engineering and Management*, pp. 1–13, 2021.
[43] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010.
[44] Y.-Y. Wu, Z.-X. Shen, and W.-Y. Lin, "Anonymizing periodical releases of srs data by fusing differential privacy," *arXiv preprint arXiv:2211.10648*, 2022.
[45] X. Zhang, L. Qi, W. Dou, Q. He, C. Leckie, R. Kotagiri, and Z. Salcic, "Mrmondrian: scalable multidimensional anonymisation for big data privacy preservation," *IEEE Transactions on Big Data*, vol. 8, no. 1, pp. 125–139, 2017.
[46] J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "The limits of differential privacy (and its misuse in data release and machine learning)," *Communications of the ACM*, vol. 64, no. 7, pp. 33–35, 2021.

Abdul Majeed received the Ph.D. degree in Computer Information Systems & Networks from the Korea Aerospace University, Korea. He is currently working as an Assistant Professor with the Department of Computer Engineering, Gachon University, Korea.

Safiullah Khan received the M.Sc. degree in Electrical Engineering from COMSATS University Islamabad, Pakistan. He is currently pursuing the Ph.D. degree in computer engineering with Gachon University, Seongnam, South Korea.

Seong Oun Hwang received the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST),Korea. He is currently working as a Professor with the Department of Computer Engineering, Gachon University, Korea.