Computers, Materials & Continua
DOI: 10.32604/cmc.2022.024663
Article

Tech Science Press

Slicing-Based Enhanced Method for Privacy-Preserving in Publishing Big Data

Mohammed BinJubier[1], Mohd Arfian Ismail[1], Abdulghani Ali Ahmed[2,*] and Ali Safaa Sadiq[3]

[1]Faculty of Computing, Universiti Malaysia Pahang, Kuantan, Pahang, Malaysia
[2]School of Computer Science and Informatics, De Montfort University, Leicester, LE1 9BH, United Kingdom
[3]School of Engineering, Computing and Mathematical Sciences, University of Wolverhampton, Wulfruna Street Wolverhampton, WV1 1LY, United Kingdom
*Corresponding Author: Abdulghani Ali Ahmed. Email: aa.ahmed@dmu.ac.uk
Received: 26 October 2021; Accepted: 12 January 2022

Abstract: Publishing big data and making it accessible to researchers is important for knowledge building as it helps in applying highly efficient methods to plan, conduct, and assess scientific research. However, publishing and processing big data poses a privacy concern related to protecting individuals' sensitive information while maintaining the usability of the published data. Several anonymization methods, such as slicing and merging, have been designed as solutions to the privacy concerns for publishing big data. However, the major drawback of merging and slicing is the random permutation procedure, which does not always guarantee complete protection against attribute or membership disclosure. Moreover, merging procedures may generate many fake tuples, leading to a loss of data utility and subsequent erroneous knowledge extraction. This study therefore proposes a slicing-based enhanced method for privacy-preserving big data publishing while maintaining the data utility. In particular, the proposed method distributes the data into horizontal and vertical partitions. The lower and upper protection levels are then used to identify the unique and identical attributes' values. The unique and identical attributes are swapped to ensure the published big data is protected from disclosure risks. The outcome of the experiments demonstrates that the proposed method could maintain data utility and provide stronger privacy preservation.

Keywords: Big data; big data privacy preservation; anonymization; data publishing

1 Introduction

The vast influence of emerging computing techniques has encouraged the generation of large data volumes in the past few years, leading to the trending concept known as "big data" [1,2]. Data publishing assists many research institutions in running big data analytic operations to reveal the information embedded and provide several opportunities with great unprecedented benefits in many

fields [3]. This process helps organizations improve their efficiency and future plans [1,4–6]. Analyzing big data and extracting new knowledge while protecting sensitive information is now considered as one of the imperative needs [7]. Moreover, much attention has been paid to potential data privacy violations and data misuse; hence, the proper protection of released data must be ensured because failure may lead to harmful situations impact to individuals and organizations [4]. Many establishments such as educational institutes and healthcare centers need to publish data in different formats to extract new knowledge [7].

Data publication is the easiest method for data sharing that helps research entities run data mining operations on published databases to extract knowledge from the published data. Such knowledge can represent, interpret, or discover interesting patterns [7,8]. However, the potentials of published partial data derived from big or a series of datasets are yet to be realized. Scholars face several problems during knowledge extraction process from the published data. One of such challenges is the issue related to data privacy that leads to the disclosure of individuals' identities. This issue is threatening the secure propagation of private data over the web. It has been the reason to limit the availability of large datasets to researchers [9]. One of the common practices and most widely used for providing privacy for individuals is the anonymization approach of data prior to its publication. Data anonymization aims to reduce the associated risk of disclosing information of individuals and preserves the possible utilization of published data [10]. Though, this approach remains holds two main open questions: 1) Can anonymized data be effectively used for data mining operations? 2) What protection is needed to prevent private information disclosure while preserving data utility? [11].

There are two popular models have been proposed for data publication [12]: (1) Multiple publication models from the same data publisher. Multiple data publications refer to a series of datasets in distinct timestamps that are all extensions in certain aspects (e.g., quarterly released data) [8,13]. When the datasets come from the same publisher, this implies that the publisher knows all the original data. (2) Single publication model from several data publishers. Several privacy approaches exist [14] for preserving data privacy. However, majority of these approaches focus mainly on a single publication [12,15,16], where the publisher anonymizes the dataset without considering other datasets that have been published.

In both models, there are two fundamental methods for releasing the published data. The first method is an interactive setting in which the data collector computes some function on the big data to answer the queries posed by the data analyzer. The second method is the non-interactive setting in which the big data is sanitized and then published [17]. It is worth noting that in our study, we consider the scenario of a single publication model in the non-interactive setting where the big data are sanitized and independently published by many organizations (data collectors) that share several common individual records. The issue with this assumption is that in several cases, the information of an individual may be published by more than one organization [18], and an attacker may launch a composition attack [12,19] on the published data to alter their privacy.

The attributes that cover more than one organization may publish to create links, such as sex, age, and zip code, are called quasi-identifiers (QIs). A composition attack is a situation where an intruder tries to identify an individual by linking several available attributes (QIs) in the published data to an external database to exploit sensitive information [12,20–22]. Therefore, anonymization can only be achieved by altering these attributes to conceal the linkage between the individual and specific values to avoid such attacks and preserve the possible utilization of the published data [12]. The common method to sanitize the database while maintaining data utility is data anonymization, which is defined in [11] as a set of methods to reduce the risk of disclosing information on individuals, businesses, or

other organizations. Most of the existing anonymization-based methods work by setting protection methods, such as perturbing [22,23], suppressing or generalizing variable values [13], or preserving privacy based on measures of correlation [12,24]. The main aim of these methods is to create some sort of uncertainty in assessing identity inference or sensitive value [11]. Besides, this protection method aims to weaken the linkage between the QI values and sensitive attribute (SA) values such that an individual cannot be identified with his/her sensitive values.

The single publication model has several correlated attributes rather than a single column distribution to achieve exceptional new knowledge results [3]. Suppressing or generalizing methods rearrange the data distributions to execute mining for privacy preservation, which involves analyzing each dimension separately, overlooking the correlations among various attributes (dimensions) [25]. Preserving privacy based on the perturbation method alters the original values of a dataset to its anonymized version, which leads to data utility problems depending on the amount and type of noise or the specific properties of data that are not preserved [7].

The clever approach to resolving these problems is to measure correlation to improve the protection and enrich data utility. The association is measured by a correlation coefficient, denoted by r, which plays a major role in data science techniques that measure the strength of association between variables; hence, the choice of a particular similarity measure can be a major cause of the success or failure in some of classification and clustering algorithms [26].

The Pearson Correlation Coefficient (PCC) and Mean Square Contingency Coefficient (MSCC) are the two commonly used measures in identifying association [24,27,28]. PCC is used to determine the strength of a linear relationship between two continuous variables. The value of the coefficient r ranges from $[-1, +1]$ [27]. When the value of r is -1 or $+1$, a perfect linear relationship exists between the considered variables. However, if the value is 0, this infers no linear relationship exists between the pairs of variables. An MSCC is a chi-square measure of the correlation between two categorical attributes. Unlike PCC, chi-square measures the extent of the significance of the relationship instead of measuring the strength of the relationship.

The idea behind the measure of correlation is to keep data utility via grouping highly correlated attributes together in columns and preserving the correlations between these attributes. The correlation measure protects privacy as it breaks the associations between uncorrelated attributes in other columns via anonymization approaches such as randomly permutated and generalization [12,24].

In this study, ideas are pooled from [12,24] to propose an effective method of determining the level of data protection needed and knowing the optimal manner to achieve this protection level whilst preserving data utility. Both are achieved by using slicing in the anonymization approach for data publishing using vertical partitioning (attribute grouping) and horizontal partitioning (tuple partition). The lower protection level (LPL) and upper protection level (UPL) are used to overcome the unique attributes and presence of identical data for data privacy protection whilst preserving data utility. LPL overcomes the unique attribute values, whereas UPL overcomes the high identical attribute values. LPL and UPL define the level of protection around the attribute values and ensure that an attacker cannot obtain the sensitive information needed to identify the record owner within such interval. This work also relies on value swapping to ensure a lower risk of attribute disclosure and l-diverse slicing. The proposed approach ensures that the published big data is protected from disclosure risks. The outcome of the experiments show that the UL method could keep more data utility and provide a stronger privacy preservation.

This paper's major contribution is the proposed upper and lower-level-based protection method (UL method) for data anonymization. The UL method better balances privacy, information loss,

and utility. That is why the level of protection required and the optimal manner of achieving it are determined while preserving data utility using the lower and upper protection levels. This work also relies on rank swapping to guarantee a lower risk of attribute disclosure, achieve aggregate query and l-diverse inside the table and solve the problem of creating invalid tuples.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 presents in detail the UL method. Section 4 discusses the experimental analysis. Finally, Section 5 concludes the paper and highlights the key findings.

2 Related Work

The most favourable approaches for preserving privacy based on the suppressing or generalizing method and anonymization of the data include the k-anonymity approach [29], l-diversity approach [30], and the T-closeness approach [15]. These approaches were proposed for privacy preservation in one-time data publishing. These methods take personal data and anonymise it and make it unattributable to any specific source or person by breaking the relations amongst the attribute values. High dimensionality renders these approaches ineffective because the identities of the primary record holders can be unmasked by merging the data with either public (composition attack) or background information [12,31]. Readers can refer to [5,7,32–34] for more comprehensive understanding of these approaches.

In the last decade, the probabilistic approach [35], e-differential privacy approach (e-DP) [36], hybrid approach [31], and composition [37]-preserving privacy based on the perturbation method were proposed for multiple independent data publishing. Composition is the first privacy model to prevent composition attacks in multiple independent data publishing [12]. The proposed approach in [37] has integrated two novel concepts: (ρ, α)-anonymization by sampling and composition-based generalization for independent datasets to protect against composition attacks. The proposed approach in [31] combined sampling, generalisation and perturbation by adding Laplacian noise to the count of every sensitive value in each equivalence class. The probabilistic approach suggests a new method called (d, α)-linkable. It tries to limit the likelihood of an adversary completing a composition attack by ensuring that the d personal values are associated with a quasi-identifying group with a probability of α by exploring the correlation between the QI attributes and SAs.

Mohammed [36] proposed the first noninteractive-based approach called e-DP based on the generalization method. The proposed solution produces a generalized contingency table and adds noise to the counts. The e-DP provides a strong privacy guarantee for statistical query answering and protection against composition attacks by differential privacy-based data anonymization [12,19,31,38,39] showed that using e-DP to protect against composition attacks generates substantial data utility losses during anonymization.

The most recent measure correlation-based methods are slicing [24] and merging [12]. Slicing has received substantial attention for privacy-preserving data publishing, which is considered a novel data anonymization approach. The authors presented a risk disclosure prevention concept that is devoid of generalization. Random slicing permutates the values of attributes in the bucket to annul the column-wise relationships. This method protects the privacy of the published records from attribute and membership disclosure risks. In addition, slicing is recommended for high-dimensional data anonymization because it keeps more data utility than the generalization of attribute values. Therefore, slicing ensures data privacy and preserves data utilities because the attribute values are not generalized. It uses vertical partitioning (attribute grouping) and horizontal partitioning (tuple partition), and its sliced table should be randomly permutated [24] (see Tab. 1).

Table 1: Published data by slicing

(Age, gender)	(Zip code, disease)
(30, F)	(130350, ovarian cancer)
(23, M)	(130350, heart disease)
(28, F)	(130352, Flu)
(53, F)	(130350, heart disease)
(39, F)	(130352, Flu)
(60, M)	(130351, heart disease)

However, slicing can cause data utility and privacy-preserving problems, as slicing randomly permutates attribute values in each bucket, creating invalid tuples that negatively affect the utility of the published microdata. The invalid tuples may easily result in several errors and incorrect results in process challenges. An attacker can rely on the analysis of the fake tuples in the published table to capture the concept of the deployed anonymization mechanism, having the chance to violate the privacy of published data [5,7,40].

For instance, in Tab. 1, tuple t_1 has just one matching equivalence class that is linked with two sensitive values for zip code 130350. Here, any person may be linked with sensitive values with a probability of not more than 1/l via l-diverse slicing because slicing has been shown to satisfy l-diverse slicing by being linked with the sensitive values by 1/2. If the QI attribute, namely, the zip code is revealed because it has high identical attribute values (sufficient variety) and an adversary relies on background knowledge and has a knowledge of (23, M), then the adversary can determine the SA for the individual. Moreover, if the slicing algorithm switches the sensitive value (randomly) between t_1 and t_2, then incompatibility is created between the SA and QI attribute values, as mentioned in [40].

Hasan et al. [12] designed the merging approach to protecting personal identity from disclosure. It is considered an extension of slicing approach. The primary aim of the merging approach is privacy preservation in multiple independent data publications via cell generalization and random attribute value permutation to break the linkage between different columns. To compute data utility and privacy risks, the merging approach that preserves data utility has minor risks because it increases the false matches in the published datasets. However, the major drawback of merging is the random permutation procedure for attribute values to break the association between columns. Besides increasing the false matches for unique attributes in the published datasets, these procedures may generate a small fraction of fake tuples but result in many matching buckets (more than the original tuples). This will eventually lead to loss data utility and can produce erroneous or infeasible extraction of knowledge through data mining operations [41,42]. Therefore, the primary reason for revealing people's identity is the existence of unique attributes in the table or allowing several attributes in the row to match the attributes in other rows, leading to the possibility of accurately extracting the attributes of a person [7,12,24].

Other studies [8,24] proved the importance of allowing a tuple to match multiple buckets to ensure protection against attribute and membership disclosure. This finding implies that mapping the records of an individual to over one equivalence class results in the formation of a super equivalence class from the set of equivalence classes.

In this study, the proposed UL method preserves the privacy of the published data while maintaining its utility. The UL method uses the upper level to overcome the identical high values

in every equivalence class. However, it uses the lower level to overcome the unique attributes found in every equivalence class. It also uses swapping to break the linkage between the unique attributes and the attributes with identical high values to improve the diversity in our work and increase personal privacy. Worth mentioning that the attributes have been generalised to the switching ability of them and the associated issues. The primary goal of swapping or generalizing the attribute values is to get anonymized data.

3 The Proposed Method

This section presents the UL method to be used in the enhanced protection method of data publishing while maintaining data utility. The proposed method reduces the risk of a composition attack when multiple organizations independently release anonymized data. The primary goal of this work is to get a specified level of privacy with minimum information loss for the intended data mining operations. The UL method proposed comprises four main stages, as illustrated in Fig. 1. The following four subsections describe these four stages.

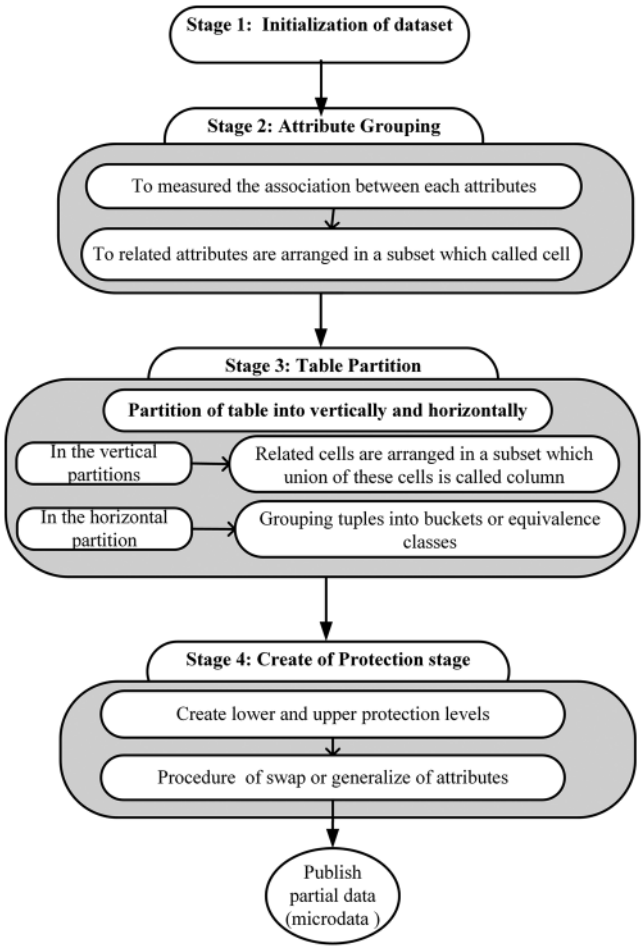

Figure 1: General block diagram of the UL method

3.1 Dataset Initialisation Stage

A standard machine learning dataset known as the "Adult" dataset was used for the experiments. This dataset was assembled by Ronny Kohavi and Barry Becker and drawn from the 1994 United States Census Bureau data [43]. The dataset comprised 48,842 tuples with fifteen QI attribute values.

3.2 Attribute Grouping Stage

The utilized table T has a_i attributes, where $i = 1, 2, \ldots$ n. The highly correlated attributes are clustered into columns and uncorrelated attributes are in the other columns, such that each attribute a_i belongs to one subset. Col_i columns $\{col_{1,}\ col_{2,}\ \ldots col_{n,}\}$ contain all the attributes a_i. The grouping of the related attributes is based on the inter attribute relationship measurement, which is ideal for privacy and utility. Regarding data utility, the grouping of highly correlated attributes ensures the preservation of their interattribute relationships. However, in terms of privacy, the identification risk is relatively higher due to the association of uncorrelated attributes compared with the association of highly correlated attributes because of the less frequent association of uncorrelated attribute values; hence, they are more identifiable. For privacy protection, breaking the linkages between the uncorrelated attributes is better [24]. The appropriate measure of association for this situation is MSCC because most of the attributes are categorical. Assume attribute a_1 with value domain $\{v_{11,}\ v_{12,}\ \ldots v_{1d1,}\}$, attribute a_2 with value domain $\{v_{21,}\ v_{22,}\ \ldots v_{2d2,}\}$, and their domain sizes are d_1 and d_2, respectively. The MSCC between a_1 and a_2 is defined as follows:

$$r^2(a_1, a_2) = \frac{1}{\min\{d_1, d_2\}} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} \tag{1}$$

where $r^2(a_1, a_2)$ is the MSCC between a_1 and a_2 attributes; $f_{i.}$ and $f_{.j}$ are the fractions of occurrence of v_{1i} and v_{2j} in the data, respectively; and f_{ij} is the fraction of cooccurrence of v_{1i} and v_{2j} in the data. Therefore, $f_{i.}$ and $f_{.j}$ are the marginal totals of f_{ij} : $f_{i.} = \sum_{j=1}^{d_2} f_{ij}$ and $f_{.j} = \sum_{i=1}^{d_1} f_{ij}$. $0 \leq r^2(a_1, a_2) \geq 1$.

3.3 Table Partition (Vertical and Horizontal) Stage

Given the computation of correlation (r) for each pair of the attributes, the dataset is vertically and horizontally partitioned in the table. In vertical partitions, the k-medoid clustering algorithm, also known as the partitioning around medoids algorithm [44], is used to arrange the related attributes into columns such that each attribute belongs to one column. This algorithm ensures each attribute is determined just as a point in the cluster space, and the inter-attribute distance in the cluster space is given as, which ranges from 0 to 1. When the two attributes are strongly correlated, the distance between the related data points will be smaller in the clustering space.

Tab. 2 shows three partitions for the columns: (1) T^* contains all columns with highly correlated attributes col^*, where $col^* = \{col^*_{1,}, \ col^*_2, \ldots col^*_i\}$, and $col^* \in T^*$. (2) T^{**} contains all columns with uncorrelated attributes col^{**}, where $col^{**} = \{col^{**}_{1,}, \ col^{**}_2, \ldots col^{**}_i\}$. (3) T^c contains columns with SA col^c when a single SA exists, and its SA is placed in the last position for easy representation, where $col^c \in T^c$ and $(T^* \cap T^{**}) \cap T^c = T$, $(i = 1, 2 \ldots n)$. If the data contain multiple SAs, one can either consider them separately or consider their joint distribution [45].

Table 2: Example of partitions in table T

T^* contains all columns with highly correlated attributes		T^{**} contains all columns with uncorrelated attributes		T^c contains column with sensitive attributes
$col^*_{1,}$	col^*_2	col^{**}_1	col^{**}_2	col^c
(a_1,a_2)	(a_3,a_4)	(a_5,a_6)	(a_7,a_8)	(a_s)

In the horizontal partition, all tuples that contain identical values are grouped into buckets or equivalence classes. Each individual is linked with one distinct sensitive value, such that an attacker could not have access to the person's sensitive values with a probability of over 1/l. The Mondrian [46] algorithm is used to group the tuples.

3.4 Protection Stage

This stage explains the data protection method proposed in this study using the UL method. It is an opportunity to improve the protection level and resolve issues on privacy with the preservation of data utility via two steps:

3.4.1 Creation of Protection Levels

The key parameters used to improve the protection level in slicing include *LPL* and *UPL*. This study used *LPL* and *UPL* to overcome the unique attributes, and the attributes have identical high values in every equivalence class. Both protection levels define the protection interval around the unique attribute values and identical high values, that fall within this period in T^{**}, such that the attacker would find deducing sensitive information difficult and identifying the record owner within such an interval impossible. The lower levels overcome the unique attribute values, whereas the upper levels overcome the high values identical to the individual's privacy protection. Suppose the cells have high values of the correlation coefficient (r). In that case, the probability of cells is in the same equivalence class, and by linking these cells with other cells in T^*, the adversary has high confidence around the SA, leading to a privacy breach. The rest of the cells are protected from attribute disclosure and membership disclosure because of their presence in over one equivalence class. The proposed privacy goal further requires the range of the rest of the cell groups to be larger than a certain threshold (containing diversity that is at least ≥ 2 in each equivalence class (see algorithm 1). The upper and lower protection levels (*UPL* and *LPL*) aim to find a set of unique cell values and high identical values for cells from T^{**}, which are presumed known to any attacker:

$$\overline{C_{col,E}} = \Phi \leq UPL < 1.0, \text{and} \underline{C_{col,E}} = 0.0 < LPL \leq \Phi$$

The attributes that fall within this period, which will be swapped, are called the swapping attributes, and $|\overline{C_{col,E}}|$ and $|\underline{C_{col,E}}|$ are the numbers of cells that fall within this period. Values that have been initially marked to be swapped are called swap rate, denoted by Φ. Typically, Φ is of the order of 1%–10%; thus, the fraction of attributes swapped will be less than one.

Definition 1 (Cell): A cell is a pair of attributes, such as (age, gender), where any cell $C_{col,E}$ is identified by the number of columns Col_i and the number of an equivalence class E_j. For example, in Tab. 1, any cell in column {(Age, Gender)} is identified by Col_i and E_j, where $1 \leq i \leq col$ and $1 \leq j \leq E$ and the first equivalence class consists of tuples $t = \{t_1, t_2, t_3, t_4\}$.

Definition 2 (Matching Buckets): Let col^{**} be the columns, where $col^{**} = \{col_1^{**}, col_2^{**}, \ldots col_n^{**}\}$, and $col^{**} \in T^{**}$. Let t^{**} be a tuple, and $t^{**}|col_i^{**}|$ be the col_i^{**} value of t^{**}. Let E^* be an equivalence class in the table T^{**}, and $E^{**}|col_i^{**}|$ be the multiset of col_i^{**} values in the equivalence class E^{**}. E^{**} is a matching bucket of t^{**} iff for all $1 \leq i \leq col^{**}$, $t^{**}|col_i^{**}| \in E^{**}|col_i^{**}|$.

Definition 3 (Lower and Upper Protection Level): LPL and UPL are correlation coefficient (r) values for each cell $C_{col,E}^{**}$ in $col_i^{**} LPL\ and\ UPL \in r$.

Algorithm 1: Creation of Protection Levels attributes

Input: Table T

Output: Defining a set of attributes a_i^{**} that contain value of r that fall in $\overline{C_{col,E}}$ and $\underline{C_{col,E}}$

1. attribute grouping-stage 2
2. table partition-stage 3
3. **for** each equivalence class E in $\boldsymbol{T^{**}}$ **do**
4. $\overline{C_{col,E}}$: correlation coefficient (r) for attributes in $\Phi \leq a_i^{**} < 1.0$
5. $\underline{C_{col,E}}$: correlation coefficient (r) for attributes is in $0.0 < a_i^{**} \leq \Phi$
6. $C_{col,E}$ **:** correlation coefficient (r) for attributes in $\underline{C_{col,E}} < a_i^{**} < \overline{C_{col,E}}$
7. Swapping or generalisation of attributes a_i^{**} in $\overline{C_{col,E}}$ (algorihm 2)
8. Swapping or generalisation of attributes a_i^{**} in $\underline{C_{col,E}}$ (algorithm 2)
9. *Ensure the l-diversity of all equivalence classes to satisfy privacy requirement as in* [24]**.**

Given the computation of correlation (r) for each pair of the attributes, attribute a_i^{**} values are grouped into three groups: **1)** $\overline{C_{col,E}}$ contains all attribute values, that have a correlation coefficient (r) falling within this period $\Phi \leq a_i^{**} < 1.0$ (see **Line 4**). **2)** $\underline{C_{col,E}}$ contains all unique attribute values, that have a correlation coefficient (r) falling within this period $0.0 < a_i^{**} \leq \Phi$ (see **line 5**). **3)** $C_{col,E}$ contains the rest of the cells that contain a distant association value from $\overline{C_{col,E}}$ and $\underline{C_{col,E}}$ and fall within this period $\underline{C_{col,E}} \leq a_i^{**} < \overline{C_{col,E}}$ (see **line 6**). $C_{col,E}$ are characterized by the presence probability it's in multiple equivalence classes, which leads to the prevention of attribute disclosure. **Line 9** is a check for the l-diversity privacy requirement as in slicing [24]. Moreover, these cells must contain diversity that is at least greater than or equal to two (diversity ≥ 2) and distributed in each equivalence class.

3.4.2 Swapping or Generalisation of Attributes

Swapping or generalization of attributes is the anonymization stage, where randomly permutated values in an equivalence class may not be protected from attribute or membership disclosure because the permutation of these values increases the risk of attribute disclosure, rather than ensuring privacy [40]. Therefore, the proposed algorithm in this study ensures the privacy requirement in each equivalence class. Rank swapping is used to break the linkage between the unique attributes and the cells with identical high values to improve the diversity in slicing and increase personal privacy. Attribute swapping alters the tuple data with unique attribute values or identical high values by switching the values of attributes across pairs of records in a fraction of the original data. With not being able to swap attributes, the attributes have to be generalized. The primary goal of switching or generalizing the attribute values is to get the anonymized table T, which would not have any nonsensical combinations in the record (invalid tuples) and would satisfy the l-diverse slicing (see Algorithm 2).

Algorithm 2: Swapping or Generalisation of Attributes

Input: Table T

Output: Obtain the Anonymized Table T^*

1. Check if swapped attributes are in the same rank group.
2. Check if the tuple does not have any nonsensical combination.
3. Swap the attributes values to satisfy k-anonymity.
4. **else**
5. Generalize the attributes value to satisfy k-anonymity.

To ensure the integrity of attribute swapping, the values of an attribute a_i^{**} are ranked in groups, for example, $\boldsymbol{Level}_0$ in Fig. 2 has two groups: {Federal − gov, Local − gov, State − gov} and {Self − emp − inc, Self − emp − not − inc}. **In line 3**, the value is swapped between two attributes if the two attributes are in the same rank group and have no nonsensical combinations. If the two attributes are in different groups or if the records have any nonsensical combination, the attribute values are generalized to satisfy k-anonymity (see **line 5**). Hence, the whole equivalence class is not generalized during attribute generalization; hence, it provides an opportunity to improve data utility compared with full table or column generalization. It also improves the utility of the published dataset. In addition, attribute swapping or generalization provides greater information veracity when deciding. Veracity is the reliability of data and represents the meaningfulness of depending on such data for data mining operations [12,40].

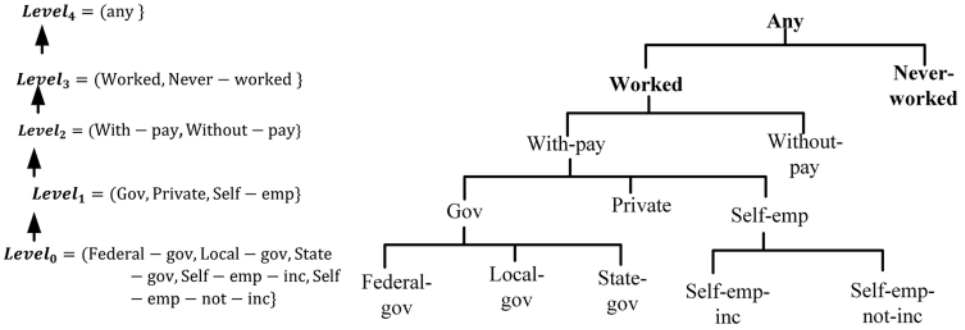

Figure 2: Example of domain (left) and value (right) generalization hierarchies for the work–class attributes

Definition 4 (Attribute Generalization): Let T^{**} be part of table T, and a_i^{**} be a QI attribute set in T^{**}. Generalisation replaces the QI attribute values with their generalised version. Let d_i^{**} and d_j^{**} be two domains with dimensional regions $\{d_{i1}^{**}, d_{i2}^{**} \ldots, d_{in}^{**}\}$ and $\{d_{J1}^{**}, d_{j2}^{**}, \ldots d_{Jn}^{**}\}$, respectively, where $\cup_{d_{in}^{**}} = d_i^{**}$ and $d_i^{**} \cap d_j^{**} = \Phi$. If the values of d_j^{**} are the generalisation of the values in domain d_i^{**}, denote $d_i^{**} < d_j^{**}$ (a many-to-one value generalisation approach). Generalisation is based on a domain generalisation hierarchy that is defined as a set of domains whose ordering is totally based on the relationship $d_i^{**} < d_j^{**}$ (see Fig. 2).

Fig. 2 (right) shows a domain generalization hierarchy for the work–class (WC) attributes. No generalization is applied at the bottom of the domain generalization hierarchy for the WC attributes. However, the WC is increasingly more general in the higher hierarchy levels. The maximal domain level element is a singleton, which signifies the possibility of generalizing the values in each domain to a single value.

4 Experiment and Implementation

The Adult dataset, which included a real dataset, was used [43]. This experiment was implemented using the Python language. To perform the experiments, independent datasets were needed to simulate the actual independent data publishing scenario. Five disjoint datasets of different sizes were pooled from the Adult dataset and extracted into two independent datasets called the Education and Occupation dataset with eight QI attribute values: age (continuous, 74), marital status (categorical, 7), sex (categorical, 2), work class (categorical, 8), salary (categorical, 2), relationship (categorical, 6), education (categorical, 16) and occupation (categorical, 14). The values in the parenthesis show the type of attribute and the number of classifiers for each attribute.

Each dataset has 4 K tuples that were randomly selected. The remaining 8 K tuples were used to generate the overlapping tuple pool and check for composition attacks. Five copies were made for each group of the remaining tuple pool by inserting 100, 200, 300, 400, and 500 tuples into the Education and Occupation datasets to generate datasets with sizes of 4.1, 4.2K, 4.3K, 4.4K, and 4.5 (where K = 1000) for the Education and Occupation dataset, respectively.

The experiments on real datasets were presented in two parts. The first part has measured the desired level of protection. In the second part, the proposed method was tested for effectiveness against composition attacks, and the effectiveness of the proposed method in data utility preservation and privacy were evaluated compared with other existing works. The experimental results showed that the proposed method provided privacy protections against the considered attacks by maintaining good level of data utility.

4.1 Measuring Protection Level

The desired level of protection was determined by determining the unique attributes and grouping the identical data (matching of attributes) into tables. As mentioned earlier, the correlation coefficient (r) plays a significant role in determining the strength of the relationship between attributes. The LPL determines all cells with unique attribute values, and the values of the attributes fall between the range of $0.0 < LPL \leq \Phi$. The value of r for unique attributes is always close to 0 but does not equal 0. UPL determines all cells with many matching attributes of which their values fall within $\Phi \leq UPL < 1.0$. The value of r for these matching attributes is always close to 1 but not equal to 1. The rest of the cells containing a distant association value from 1 and 0 are characterized by the probability of multiple equivalence classes, leading to the prevention of attribute disclosure. Moreover, these cells must contain at least greater than or equal two (diversity ≥ 2) and distribute in each equivalence class.

The attributes that fall within this period (LPL and UPL) that will be swapped are called swapping attributes, whereas the values that are marked for swapping and considered a measure of privacy are called the swap rate and denoted by Φ. The decision-maker must specify this based on the disclosure risk and data utility by looking at the measures of the strength of the relationship between attributes.

Using the experiment datasets partitioned according to Tab. 2 and based on the Education dataset, five swap rates were performed on partitions T^{**} to find the number of cells and tuples in each LPL and UPL. Tab. 3 tabulates the number of cells that fall in the tuples that contain the swapping attributes. Cells with unique attributes or near-unique attributes are potentially riskier than other elements. Tab. 4 lists the number of cells that fall in the tuples with the matching attributes, not variety. Cells with matching attributes or near matching attributes are riskier than other elements because almost all tuples are in the same equivalence class. The adversary has more confidence around the SAs by linking these attributes with the highly correlated attributes or other datasets.

Table 3: Five changes of swap rates for ***LPL*** to calculate the number of cells and tuples in each change

Data set	$0.0 < LPL \leq \Phi$																			
	$\Phi = 0.01$				$\Phi = 0.02$				$\Phi = 0.05$				$\Phi = 0.10$				$\Phi = 0.15$			
	# of cells		# of tuples for each cell		# of cells		# of tuples for each cell		# of cells		# of tuples for each cell		# of cells		# of tuples for each cell		# of cell		# of tuples for each cell	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
4.1K	5	16	11	156	7	28	24	654	13	44	111	1082	17	62	324	1132	21	68	542	1230
4.2K	6	35	9	488	8	59	18	1982	13	103	128	2412	17	115	342	2824	22	153	611	2910
4.3K	6	16	13	16	7	34	20	238	10	53	107	1386	17	67	398	1988	20	76	631	2088
4.4K	7	39	12	1619	7	58	12	1674	14	102	166	2590	17	146	362	2703	22	146	639	2703
4.5K	6	22	17	22	8	37	27	273	12	50	117	467	15	67	229	721	19	72	526	743

Table 4: Five changes of swap rates for ***UPL*** to calculate the number of cells and tuples in each change

Data set	$\Phi \leq UPL < 1.0$																			
	$\Phi = 0.99$				$\Phi = 0.98$				$\Phi = 0.95$				$\Phi = 0.90$				$\Phi = 0.85$			
	# of cells		#of tuples for each cell		# of cells		#of tuples for each cell		# of cells		#of tuples for each cell		# of cells		#of tuples for each cell		# of cell		#of tuples for each cell	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
4.1K	1	0	1362	0	2	0	1626	0	2	0	1626	0	2	1	1626	264	2	3	1626	302
4.2K	1	0	1412	0	2	0	1674	0	2	0	1674	0	2	0	1674	0	2	0	1674	0
4.3K	2	0	1412	0	2	0	1412	0	2	0	1412	0	2	0	1412	0	2	3	1412	34
4.4K	1	0	1501	0	2	0	1787	0	2	0	1787	0	2	0	1787	0	2	0	1787	0
4.5K	1	0	1555	0	2	0	1780	0	2	0	1780	0	2	0	1780	0	2	3	1780	36

The strength of the association between attributes was used because the strength and variety of data was known. Then, *LPL* and *UPL* were used to find the specific attributes to swap between them instead of a random approach to breaking the correlations between the attribute values. This method provided more variety of data in the equivalence class. A higher swap rate (Φ) in Tab. 3 or a lower swap rate in Tab. 4 means higher privacy but decreased data utility.

4.2 Comparison Evaluation

From many data publishers, the Single publication model is considered a non-interactive data publishing used in experimental analysis. The experiment was carried out in non-interactive privacy settings. However, most of the work in differential privacy [47] is in line with the interactive settings; a user can gain access to the data set using a numerical query as the anonymization technique will add noise to query answers. The environment may not always favor this phenomenon because datasets are usually known to be published in public. As a result, the non-interactive setting was chosen for the experiment on differential privacy, which is highlighted in [36].

This section contains the assessment of the proposed work, which is achieved through the measurement of its efficiency using the hybrid [31], merging [12], e-DP [36], probabilistic [35], Mondrian [46], and composition [37] approaches, in the non-interactive privacy settings. The quasi-identifier equivalence class was given as k-anonymity [16] by the merging, probabilistic, e-DP, hybrid, composition, and Mondrian approaches. To create an equivalence class, $k = 6$ was chosen, where L-diversity is also given as 6. The main purpose of L-diversity is the preservation of privacy by expanding sensitive values' diversity. The Laplacian noise in an equivalence class for differential privacy is appended to the count of sensitive values [35] with $e = 0.3$ for the e-differential privacy budget. There are basically two factors upon which comparison can be made. One is the data utility, while the other is the risk disclosure. These factors are discussed in the subsections below.

4.2.1 Data Utility Comparison

Privacy preservation is an essential issue in table T publication; hence, data utility must also be considered. Data quality is measured based on distortion ratio (*DR)*. The *DR* in published data can be measured using several methods [13] to quantify the effect of anonymization on the overall data distortion for data mining. The generalized distortion ratio *GDR* is one appropriate measure for calculating the [42]. The swap and generalize method are used to break the association of the attributes because most of the attributes are categorical. For any two categorical attributes ($a_1^{**}, a_2^{**} \in T^{**}$), where t is its taxonomy tree and a node p in t is used to swap or generalise the attributes, the *DR* with p is defined as follows:

$$DR(a_1^{**}, a_2^{**}) = \begin{cases} 0, & a_1^{**} = a_2^{**} \\ \dfrac{|Common(a_1^{**}, a_2^{**})|}{|N|}, & a_1^{**} \neq a_2^{**} \end{cases} \tag{2}$$

where $|N|$ denotes the set of all the leaf nodes in t, and $|Common(a_1^{**}, a_2^{**})|$ is the set of leaf nodes in the lowest common tree of a_1^{**} and a_2^{**} in t.

Fig. 2 denotes the taxonomy of the WC attributes. If the values of a_i^{**} and a_j^{**} are in the same rank group and have no nonsensical combinations, then their swap values are equal, and the *DR* is 0. Moreover, if the values of a_i^{**} and a_j^{**} are not in the same rank group or have any nonsensical

combinations, then, their generalised values are equal to $\frac{|Common(a_1^{**},a_2^{**})|}{|N|}$, and the DR is equal to $\sum_{j=1,k=1}^{n,m} d_{j,k}$, where $d_{j,k}$ is the distortion of the attribute a_j^{**} of tuple t_k.

DR is proportional to the distortion of the generalised dataset over the distortion of the fully generalized dataset. Data utility can be estimated by subtracting DR in Eq. (3) [13] as follows:

$$Data\ utility = (100 - DR)\% \tag{3}$$

Figs. 3 and 4 show the results of the experiments on data utility, that is made based on data loss on the Education dataset. The proposed work in Fig. 3 had a swap rate (Φ) of 2% using LPL and 98% using UPL. The proposed work had a swap rate (Φ) of 5% using LPL and 95% using UPL in Fig. 4. Decision-makers must select the swap rate to determine the protection level required by looking into the changes in swap rates, which helps know the number of cells in each swap rate (Φ) (see Tabs. 3 and 4). An increase in swap rate (Φ) in LPL or decrease in the swap rate (Φ) in UPL enhances the privacy while the data utility becomes lower. The assessment of the proposed work, done through its comparison with hybrid [31], merging [12], e-DP [36], probabilistic [35], Mondrian [46], and composition [37] approaches revealed that the data utility obtained by the UL is higher than that of all the known works. Whereas, merging approach had N fake tuples with the same QI values as in the original table, and the sensitive values were assigned to them based on the sensitive value distribution in the initial dataset. Therefore, the proposed approach resulted in lesser data loss than the merging method. The UL method employs selective generalization within the cell when satisfying the privacy requirements is essential; hence, more data utility is preserved.

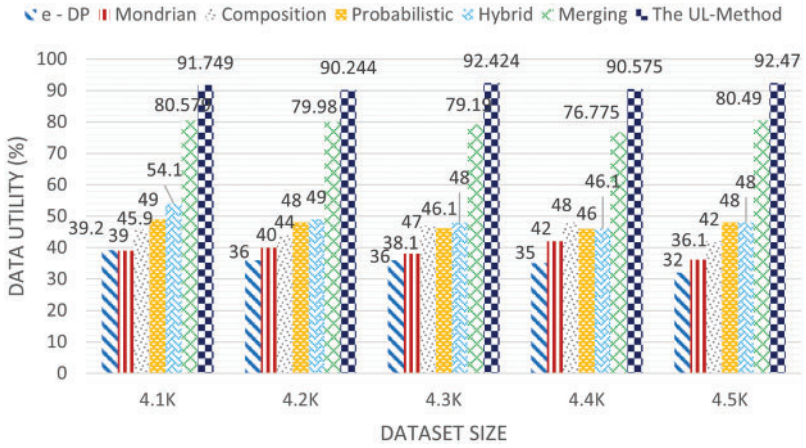

Figure 3: Data utility on the Education dataset (swap rate ($\boldsymbol{\Phi}$) of 2% using ***LPL*** and 98% using ***UPL***)

4.2.2 Measuring Risks

A composition attack is a situation where an intruder tries to identify an individual in the table T by linking several available records in the microdata to an external database to exploit sensitive information, especially when the intruder has much background information about the relationship between the QI and SAs [48]. Therefore, measuring disclosure risk is essentially measuring the rareness of a cell in data publishing. The methods employed for assessing risk disclosure in table T during a composition attack are discussed in this section.

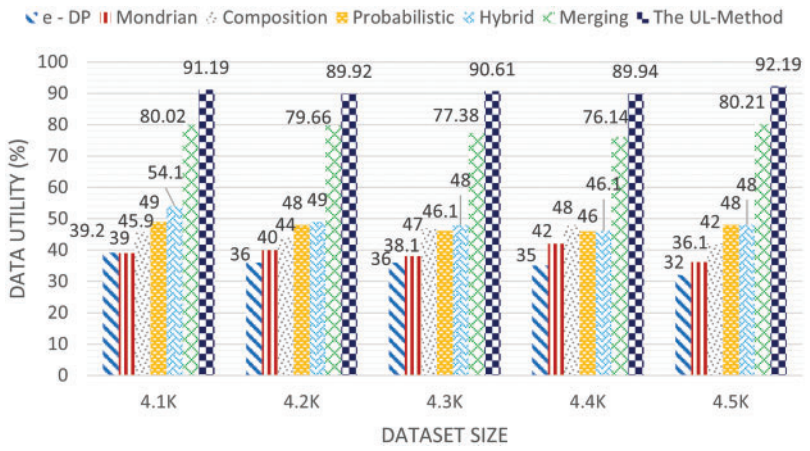

Figure 4: Data utility on the Education dataset (swap rate (**Φ**) of 5% using ***LPL*** and 95% using ***UPL***)

Data publishers should strive to measure the risk disclosure of anonymization approach outputs to ensure privacy preservation. This step is key in defining the level of protection needed. Therefore, differentiating the risk disclosure measures is important because the quantity must not depend on how the data representation method is selected. Risk disclosure can be measured by determining the proportion of the genuine matches to the total matches, as expressed in Eq. (4).

$$\textit{Disclosure risk ratio } (DRR) = \frac{\textit{Matched records}}{\textit{Total records}} \times 100\% \tag{4}$$

The experimental results for the Education datasets are shown in Fig. 5, while that of Occupation datasets are shown in Fig. 6. The experimental results represented are for disclosure risk ratio (*DRR*), which is known to define the confidence level of an adversary and can be used to understand the sensitive values on the Education and Occupation dataset. Amongst the approaches, the e-DP approach [36] provided the lowest privacy risks for composition. The proposed solution in [36], probabilistically generated a generalized contingency table and then added noise to the counts. However, it reduced data utility, as discussed in Section A (Data Utility Comparison) and Figs. 3 and 4.

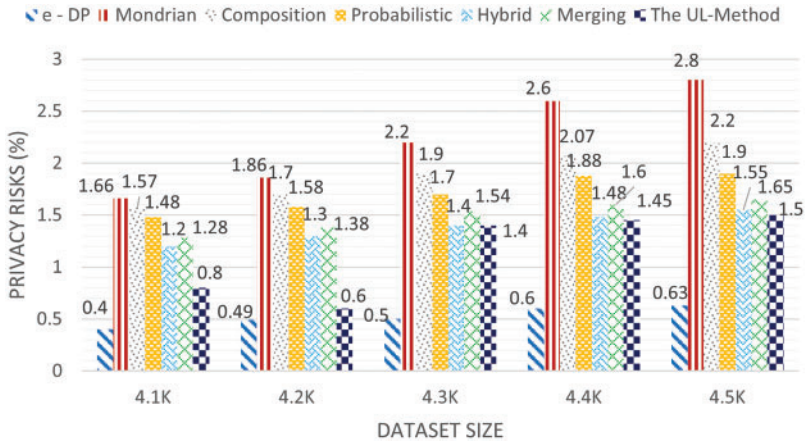

Figure 5: Privacy risk for Education dataset (k = 6, l = 6)

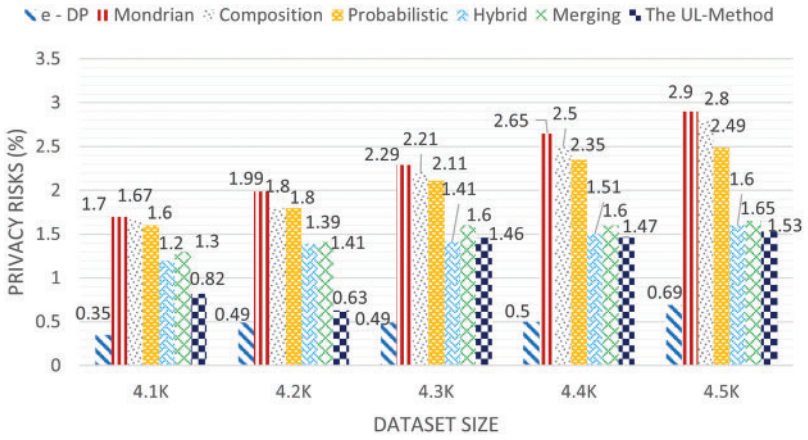

Figure 6: Privacy risk for Occupation dataset (k = 6, l = 6)

In addition, the hybrid [31] approach yielded a lower probability of inferring the user's private information than the probabilistic [35], composition [37], Mondrian [46], and merging [12] approaches. The merging approach reduced the probability of composition attack on the published datasets compared with the probabilistic [35], composition [37], and Mondrian [46] approaches. The proposed work could successfully reduce the probability of composition attack on the published datasets by overcoming the unique attributes and high identical attribute values using *UPL* and *LPL*, and providing multiple matching cells in each equivalence class, which led the protection against identity disclosure.

In Fig. 7, the experimental results are summarized for disclosure risk ratio (*DRR*) for *LPL* and *UPL* when $\Phi = \{(1\%, 99\%), (2\%, 98\%), (5\%, 95\%), (10\%, 90\%), (15\%, 85\%)\}$. As Fig. 7 and Tabs. 3 and 4 illustrate, when increasing the swap rate (Φ) in *LPL* or decreasing the swap rate in *UPL* means a higher the privacy but decreased data utility. In this study, the special risk ratio for the composition attack was decreased by overcoming the unique attributes and high identical attribute values by using *UPL* and *LPL*, and providing the multiple matching cells, which confer protection from identity disclosure. Intuitively, a cell is at risk for disclosure if it can be singled out from the rest [49].

4.2.3 Aggregate Query Error

An aggregate query is a mathematical computation that involves a set of values and results in a single value expressing the significance of the data. An aggregate query aims to estimate data utility in the published datasets. Aggregate query operators are often used as 'COUNT', 'MAX', and 'AVERAGE' to provide key numbers representing the estimated data utility to verify the effectiveness of the proposed work [50]. In the experiment, only the 'COUNT' operator was tested in this experiment, and the query was considered in the following form:

SELECT COUNT(∗)

FROM Unknown − Table T

$Where\ v_{i1} \in V_{i1}\ and \ldots . v_{idim} \in V_{idim}\ and\ s \in V_s$

where v_{ij} $(1 \leq j \leq d)$ is the QI value for attribute a_{ij}, $v_{ij} \subseteq d_{ij}$ and d_{ij} is the domain for attribute a_{ij}, s is the SA value, $s \subseteq d_s$ and d_s is the domain for the SA. Predicate dimension d and query selectivity *sel* are two characteristics of a query predicate; d indicates the amount of QI in the predicate, and

sel indicates the number of values in each v_{ij}, $1 \leq j \leq d$. The size of v_{ij}, $1 \leq j \leq d$ was chosen at random from 0, 1, $\ldots set * |d_{ij}|$. Each query was run on the original table as well as those generated by the proposed work and other existing works. The original and anonymized table each had *a count*, with the original count denoted by org_{count} and the anonymised count denoted by anz_{count}, where anz_{count} denotes the proposed work and other existing works, respectively. All queries were computed using Equation [50] to determine the average relative error in the anonymized dataset:

$$Relative\ error = \frac{org_{count} - anz_{count}}{org_{count}} * 100\% \tag{5}$$

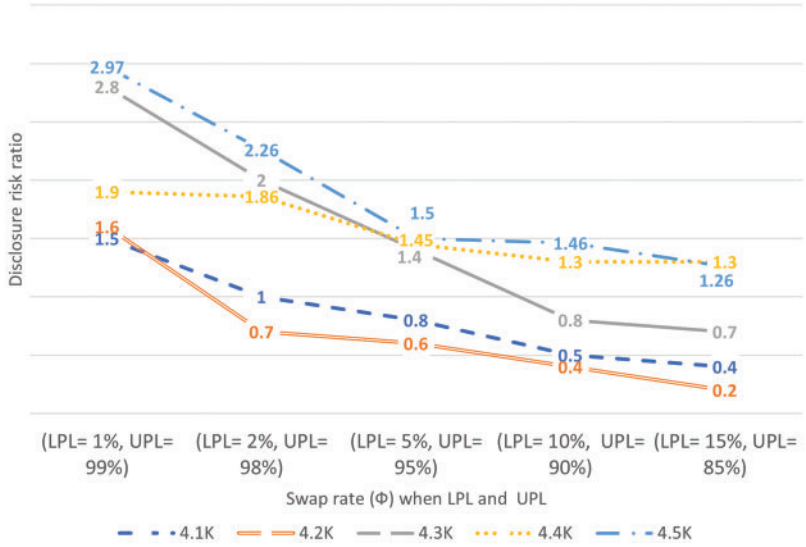

Figure 7: Experimental result for ***DRR*** for ***LPL*** and ***UPL*** when $\mathbf{\Phi} = \{(1\%, 99\%), (2\%, 98\%), (5\%, 95\%), (10\%, 90\%), (15\%, 85\%)\}$

Based on the QI selection, the relative query error was plotted on the y-axis in Fig. 8. For the Mondrian, hybrid, e-DP, probabilistic, and composition approaches, the value of k was set to 6, and I-diversity was set to 6 for merging and the proposed work, with the value of ($LPL = 5\%$ and $UPL = 95\%$). The relative query error was calculated on anonymized tables created by the proposed work and other existing works, and one, two, three, four, or five attributes were chosen as QI. Furthermore, for the 4.5 K Occupation dataset, all possible variations of the query were created and executed across the anonymization tables. Fig. 8 depicts the relative query error, with the value on the y-axis denoting the relative percentage error and the values on the x-axis denoting different QI choices. The experimental results show that the swapping approach (proposed work) consistently outperforms generalization in answering aggregate queries. For anonymized datasets, the competing approaches show a higher relative query error. Furthermore, the experimental results show that the proposed work has a slight relative error as compared to all other approaches. Because in the case of not being able to switch attributes, then they must be generalized.

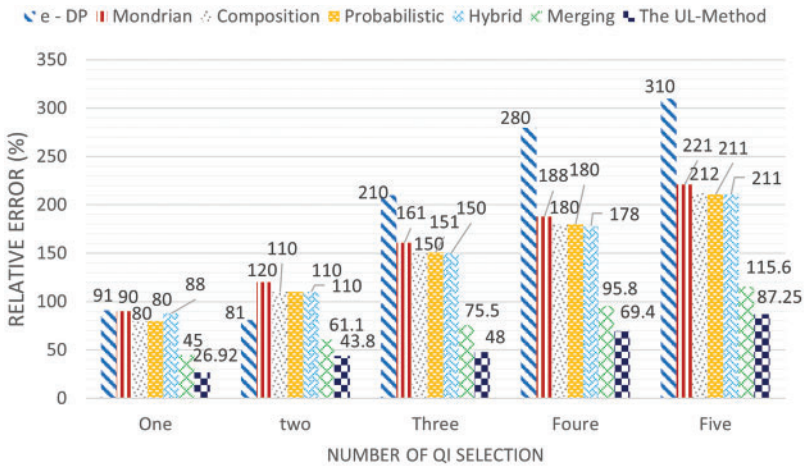

Figure 8: Aggregate query error

5 Conclusions

This study started with investigating problems associated with slicing and merging approaches that are related to the random permutation of attribute values, which is used as a way to break the association between different columns in the table. Therefore, the UL method, which confers protection by finding the unique attribute values and high identical attribute values and swapping them to decrease the attribute disclosure risk and ensure attainment of l-diverse in the published table, is proposed against composition attacks. The keyword behind that is selecting group of cells to enhance published data privacy and maintain good data utility. The results of the experiments show that the UL method could improve data utility and provide a stronger privacy preservation. In terms of data utility, the UL method achieves approximately 92.47% data utility higher than works when the percentage of swap rate is 2% using *LPL* and 98% using *UPL* with Education dataset size of 4.5 K. It achieves (92.19%) when the percentage of swap rate is 5% using *LPL* and 95% using *UPL* with Education dataset size of 4.5 K. Moreover, the UL method potentially reduces risk disclosure compared with other existing works. The achieved performance using our proposed method helps researchers, decision-makers, and technological experts to benefit from the published big data for extracting knowledge in many fields, such as education, healthcare. In future, the proposed work could be extended to several promising directions that may focus on speeding up the performance of UL method using parallel techniques. Moreover, the effectiveness of UL method has been tested against composition attacks, and by using Adult dataset, thus, it is important to test its performance against different attacks and by using different type of datasets.

Funding Statement: This work was supported by Postgraduate Research Grants Scheme (PGRS) with Grant No. PGRS190360.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. N. Maniam and D. Singh, "Towards data privacy and security framework in Big data governanc," *International Journal of Software Engineering and Computer Systems*, vol. 6, no. 1, pp. 41–51, 2020, https://doi.org/10.15282/ijsecs.6.1.2020.5.0068.

[2] P. K. Premkamal, S. K. Pasupuleti and P. J. A. Alphonse, "Efficient escrow-free CP-ABE with constant size ciphertext and secret key for big data storage in cloud," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 10, no. 1, pp. 28–45, 2020.

[3] M. BinJubeir, M. A. Ismail, S. Kasim, H. Amnur and S. S. Defni, "Big healthcare data: Survey of challenges and privacy," *International Journal on Informatics Visualization*, vol. 4, no. 4, pp. 184–190, 2020.

[4] L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information security in Big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014, https://doi.org/10.1109/access.2014.2362522.

[5] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017, https://doi.org/10.1109/ACCESS.2017.2706947.

[6] A. A. Ahmed and C. Li, "Analyzing data remnant remains on user devices to determine probative artifacts in cloud environment," *Journal of Forensic Sciences*, vol. 63, no. 1, pp. 112–121, 2018.

[7] M. BinJubeir, A. A. Ahmed, M. A. Ismail, A. S. Sadiq and M. K. Khan, "Comprehensive survey on Big data privacy protection," *IEEE Access*, vol. 8, pp. 20067–20079, 2020, https://doi.org/10.1109/ACCESS.2019.2962368.

[8] A. G. Divanis and G. Loukides, *Medical Data Privacy Handbook*, Cham: Springer International Publishing, 2015.

[9] M. Siddique, M. A. Mirza, M. Ahmad, J. Chaudhry and R. Islam, "A survey of big data security solutions in healthcare," in *Int. Conf. on Security and Privacy in Communication Systems*, Singapore, pp. 391–406, 2018.

[10] A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021, https://doi.org/10.1109/ACCESS.2020.3045700.

[11] T. A. Lasko and S. A. Vinterbo, "Spectral anonymization of data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 437–446, 2010, https://doi.org/10.1109/TKDE.2009.88.

[12] A. Hasan, Q. Jiang, H. Chen and S. Wang, "A new approach to privacy-preserving multiple independent data publishing," *Applied Sciences*, vol. 8, no. 5, pp. 783, 2018, https://doi.org/10.3390/app8050783.

[13] R. C. W. Wong and A. W. C. Fu, "Privacy-preserving data publishing: An overview," *Synthesis Lectures on Data Management*, vol. 2, no. 1, pp. 1–138, 2010, https://doi.org/10.2200/S00237ED1V01Y201003DTM002.

[14] B. C. M. Fung, K. Wang, R. Chen and P. S. Yu, "Privacy-preserving data publishing," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, 2010, https://doi.org/10.1145/1749603.1749605.

[15] N. Li, T. Li and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in *IEEE 23rd Int. Conf. on Data Engineering, 2007. ICDE 2007*. Istanbul, Turkey, pp. 106–115, 2007.

[16] L. Sweeney, "K-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[17] A. Narayanan, "Data privacy: The non-interactive setting", Ph.D. dissertation, The University of Texas Austin, 2009.

[18] B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems," *Journal of Biomedical Informatics*, vol. 37, no. 3, pp. 179–192, 2004, https://doi.org/10.1016/j.jbi.2004.04.005.

[19] S. R. Ganta, S. P. Kasiviswanathan and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Las Vegas, NV, United States, pp. 265–273, 2008.

[20] A. A. Ahmed and N. A. K. Zaman, "Attack intention recognition: A review," *International Journal of Network Security*, vol. 19, no. 2, pp. 244–250, 2017.

[21] A. A. Ahmed, A. Jantan and T. C. Wan, "Filtration model for the detection of malicious traffic in large-scale networks," *Computer Communications*, vol. 82, pp. 59–70, 2016.

[22] Z. Yu, C. Gao, Z. Jing, B. B. Gupta and Q. Cai, "A practical public key encryption scheme based on learning parity with noise," *IEEE Access*, vol. 6, pp. 31918–31923, 2018.
[23] C. K. Liew, U. J. Choi and C. J. Liew, "A data distortion by probability distribution," *ACM Transactions on Database Systems (TODS)*, vol. 10, no. 3, pp. 395–411, 1985.
[24] T. Li, N. Li, J. Zhang and I. Molloy, "Slicing: A New approach for privacy preserving data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 561–574, 2012, https://doi.org/10.1109/TKDE.2010.236.
[25] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in *Proc.-IEEE Int. Conf. on Data Mining, ICDM*, Houston, TX, USA, pp. 589–592, 2005. https://doi.org/10.1109/ICDM.2005.121.
[26] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, San Francisco, United States, Elsevier, 2012. https://doi.org/10.1016/C2009-0-61819-5.
[27] H. Cramér, *Mathematical Methods of Statistics (PMS-9)*, Princeton University Press, vol. 9, 2016.
[28] A. A. Ahmed and M. F. Mohammed, "SAIRF: A similarity approach for attack intention recognition using fuzzy min-max neural network," *Journal of Computational Science*, vol. 25, pp. 467–473, 2018.
[29] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc.-Int. Conf. on Data Engineering*, Tokyo, Japan, pp. 217–228, 2005. https://doi.org/10.1109/ICDE.2005.42.
[30] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *Data Engineering, 2006. ICDE'06. Proc. of the 22nd Int. Conf. on*, Atlanta, GA, USA, pp. 24, 2006.
[31] J. Li, M. M. Baig, A. H. Sattar, X. Ding, J. Liu *et al.,* "A hybrid approach to prevent composition attacks for independent data releases," *Information Sciences*, vol. 367–368, pp. 324–336, 2016, https://doi.org/10.1016/j.ins.2016.05.009.
[32] Y. A. Aldeen, M. Salleh and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, no. 1, pp. 1–36, 2015, https://doi.org/10.1186/s40064-015-1481-x.
[33] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin *et al.,* "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Record*, vol. 33, no. 1, pp. 50–57, 2004, https://doi.org/10.1145/974121.974131.
[34] N. Zhang and W. Zhao, "Privacy-preserving data mining systems," *Computer*, vol. 40, no. 4, pp. 52–58, 2007, https://doi.org/10.1109/MC.2007.142.
[35] A. H. Sattar, J. Li, J. Liu, R. Heatherly and B. Malin, "A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments," *Knowledge-Based Systems*, vol. 67, pp. 361–372, 2014, https://doi.org/10.1016/j.knosys.2014.04.019.
[36] N. Mohammed, R. Chen, B. C. Fung and P. S. Yu, "Differentially private data release for data mining," in *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, pp. 493–501, 2011.
[37] M. M. Baig, J. Li, J. Liu, X. Ding and H. Wang, "Data privacy against composition attack," in *Int. Conf. on Database Systems for Advanced Applications*, Busan, South Korea, pp. 320–334, 2012.
[38] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava and T. Yu, "Empirical privacy and empirical utility of anonymized data," in *2013 IEEE 29th Int. Conf. on Data Engineering Workshops (ICDEW)*, Brisbane, QLD, Australia, pp. 77–82, 2013.
[39] R. Sarathy and K. Muralidhar, "Evaluating Laplace noise addition to satisfy differential privacy for numeric data," *Transaction on Data Privacy*, vol. 4, no. 1, pp. 1–17, 2011.
[40] A. S. Hasan, Q. Jiang, J. Luo, C. Li and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Security and Communication Networks*, vol. 9, no. 16, pp. 3219–3228, 2016, https://doi.org/10.1002/sec.1527.
[41] A. Sharma, G. Singh and S. Rehman, "A review of Big data challenges and preserving privacy in Big data," *Advances in Data and Information Sciences*, vol. 94, pp. 57–65, 2020.
[42] S. Rohilla, "Privacy preserving data publishing through slicing," *American Journal of Networks and Communications*, vol. 4, no. 3, pp. 45, 2015, https://doi.org/10.11648/j.ajnc.s.2015040301.18.

[43] R. Kohavi and B. Becker, "UMI machine learning repository: Adult data Set," *Irvine*, CA: University of California, School of Information and Computer Science, 2019. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Adult. [Accessed: 04-May-2020].
[44] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Hoboken, NJ, USA: John Wiley & Sons, vol. 344, 2009, https://doi.org/10.1002/9780470316801.
[45] 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.
[46] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *22nd Int. Conf. on Data Engineering (ICDE'06)*, Atlanta, GA, USA, pp. 25, 2006.
[47] C. Dwork, "Differential privacy," *Information Security and Cryptography*, Springer, Berlin, Heidelberg, vol. 4052, pp. 1–12, 2006.
[48] B. C. Chen, D. Kifer, K. LeFevre and A. Machanavajjhala, "Privacy-preserving data publishing," *Foundations and Trends in Databases*, vol. 2, no. 1–2, pp. 1–167, 2009, https://doi.org/10.1561/1900000008.
[49] L. Taylor, X. H. Zhou and P. Rise, "A tutorial in assessing disclosure risk in microdata," *Statistics in Medicine*, vol. 37, no. 25, pp. 3693–3706, 2018, https://doi.org/10.1002/sim.7667.
[50] Q. Zhang, N. Koudas, D. Srivastava and T. Yu, "Aggregate query answering on anonymized tables," in *IEEE 23rd Int. Conf. on Data Engineering*, Istanbul, Turkey, pp. 116–125, 2007.