Wireless Communications and Mobile Computing

# Quasi-Identifiers Recognition Algorithm for Privacy Preservation of Cloud Data Based on Risk Re-Identification

Huda O. Mansour [1, 2], Maheyzah M. Siraj[2], Fuad A. Ghaleb[1*], Fasial Saeed[3], Eman H. Alkhammash[4], and Mohd A. Maarof [1]

[1] Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia (UTM), Johor, 81310, Malaysia.
[2] Department of Computer Science, Faculty of Computer Science and Information Technology, University of Kassala, Kassala, 31111, Sudan.
[3] College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia.
[4] Department of Computer Science, College of Computers and Information Technology, Taif University, P.O.Box 11099, Taif 21944, Saudi Arabia.

*Corresponding Author: Fuad A. Ghaleb (abdulgaleel@utm.my)

## Abstract

Cloud computing plays an essential role as a source for outsourcing data to perform mining operations or other data processing, especially for data owners who do not have sufficient resources or experience to execute data mining techniques. However, the privacy of outsourced data is a serious concern. Most data owners are using anonymization-based techniques to prevent identity and attribute disclosures to avoid privacy leakage before outsourced data for mining over the cloud. In addition, data collection and dissemination in a resource-limited network such as sensor cloud require efficient methods to reduce privacy leakage. The main issue that caused identity disclosure is Quasi-Identifiers (QIDs) linking. But most researchers of anonymization methods ignore the identification of proper QIDs. This reduces the validity of the used anonymization methods and may thus lead to a failure of the anonymity process. This paper introduces a new quasi-identifier recognition algorithm that reduces identity disclosure resulted from QIDs linking. The proposed algorithm is comprised of two main stages: (1) Attributes Classification (or QIDs Recognition), and (2) QID's-Dimension Identification. The algorithm works based on the re-identification of risk rate for all attributes and the dimension of QIDs where it determines the proper QIDs and their suitable dimensions. The proposed algorithm was tested on a real dataset. The results demonstrated that the proposed algorithm significantly reduces privacy leakage and maintaining the data utility compared to recent related algorithms.

## Introduction

In the modern information age, many companies are using external sources of data for processing, storing, or obtaining some services such as data mining. Unlimited computational resources, reduced costs, non-burden of maintenance, and non-diligence to learn the skills of proficiency in certain services, all of these were temptations to advance to the modern change. However, there are still security and privacy concerns that hinder the use of the features offered by the cloud [1]. Numerous studies clarified that attackers often reveal the information from third-party services or third-party clouds [2]. For example, one of the security breaches in October 2014 was a breakthrough for Dropbox. The attackers stole 700 user passwords to obtain cash values of its Bitcoins (BTC). In 2015, a lot of users' information, exceeds 4 million, such as the user's name, date of birth, address, e-mail, phone number, and other sensitive data were leaked through the TalkTalk service provider in the UK. In 2016, Time Warner one of the largest cable television companies in the United States has announced that about 32 million passwords and e-mail of the users have been stolen via an attacker. In 2017, more than 200 million data of the users containing users' names, phone numbers, e-mail addresses, home addresses, and other data have been disclosed through the API of McDelivery Company in India [2]–[4]. A fresh security violation in Google displayed that any administrator of the server who has access to the secret information can misuse it easily. The worst problem is that administrator of the honest-but-curious server can violate privacy without being discovered [5].

Three kinds of the disclosure can cause privacy leakage, identity disclosure, attribute disclosure, and membership disclosure [6]. In attribute disclosure and identity disclosure, the intruder identifies that the tuple of the target individual is found in the released dataset and he aims to acquire some private/sensitive data about that individual from the released dataset [7]. Serious issues that lead to identity disclosure are Quasi-Identifiers (QIDs) values linking and the attacker's knowledge background. The QIDs are the dataset attributes that if each of them is considered separately does not distinguish the individual, but when several attributes are combined they can give a distinctive identification of individuals [8]. For example, when looking at the attributes of date of birth, gender, and ZIP code together, one can re-identification the individuals as stated in [9]. Re-identification of the individuals through linking their QIDs leads to what are called linking-attacks. Therefore, the careless publication of QIDs will lead to leakage of privacy [3].

One of the popular practices to avoid privacy leakage is anonymization. The anonymization can perform via several types of transformations, by removing the values, changing the structure, replacing the values by taxonomy, and combine the values. The anonymization-based methods use one or a combination of operations to accomplish an optimum level of concealment [10]. A commonly utilized privacy criterion of anonymization is k-anonymity has been introduced by Sweeney 2002 [9]. The K-Anonymization model aims to make any record in the released dataset that cannot be distinguished from at least (K-1) other records [1], [11]. To avoid the linking-attacks K- Anonymization can be used. The effective method to determine the real QIDs is the primary issue for privacy-preserving methods based on K-anonymity or other anonymization models seek to prevent QIDs linking. While most of the current

methods neglected this issue or just determine QIDs manually, this reduces the validity of the anonymization method as well as negatively affects the usefulness of anonymous data [3]. This study aims at overcoming the identity disclosure resulting from QIDs linking and reduce the leakage of privacy by proposing a QIDs Recognition (QIR) algorithm based on risk rate re-identification. The proposed algorithm comprises two main stages: (1) Attributes Classification (or QIDs Recognition), and (2) QID's-Dimension Identification. The algorithm works based on the re-identification of risk rate for all attributes and the dimension of QIDs where it determines the proper QIDs and their suitable dimensions. Figure 1 shows the causes effect diagram of privacy leakage. The dark boxes in Figure 1 explain the privacy leakage causes addressed by the proposed QIDs Recognition (QIR) algorithm in this study. As shown in Figure 1, it is essential to properly identify the QIDs attributes to overcome the identity disclosure to reduce the leakage of privacy resulting from QIDs linking. This paper is made up of 5 sections. Section 2 describes the state-of-art of privacy-preserving data mining (PPDM) over the cloud, whereby present some of the current methods and algorithms that address the issue of identification QIDs accurately to avoid identity disclosure. A detailed description of the proposed algorithm has been provided in Section 3. Section 4 demonstrates the experimental evaluation, discussion, and comparison with related work. Section 5 concludes this work.

**Related Work**

The research of privacy-preserving outsourced data focus on anonymization-based methods [12]–[18], cryptographic-based methods [19]–[24], hybrid methods [2], [25]–[27] and methods seek to improve the data utility [26], [28], [29]. Some recent studies have demonstrated the privacy requirements of incremental datasets [30]–[32], multiple sensitive attributes [33]–[35]. However, most of these studies neglected the issue of identification of the right QIDs, despite its importance in the success of the anonymity process. Few of these studies have attempted to introduce methods so that identification of the QIDs is required in the anonymization process, as presented in the next section.

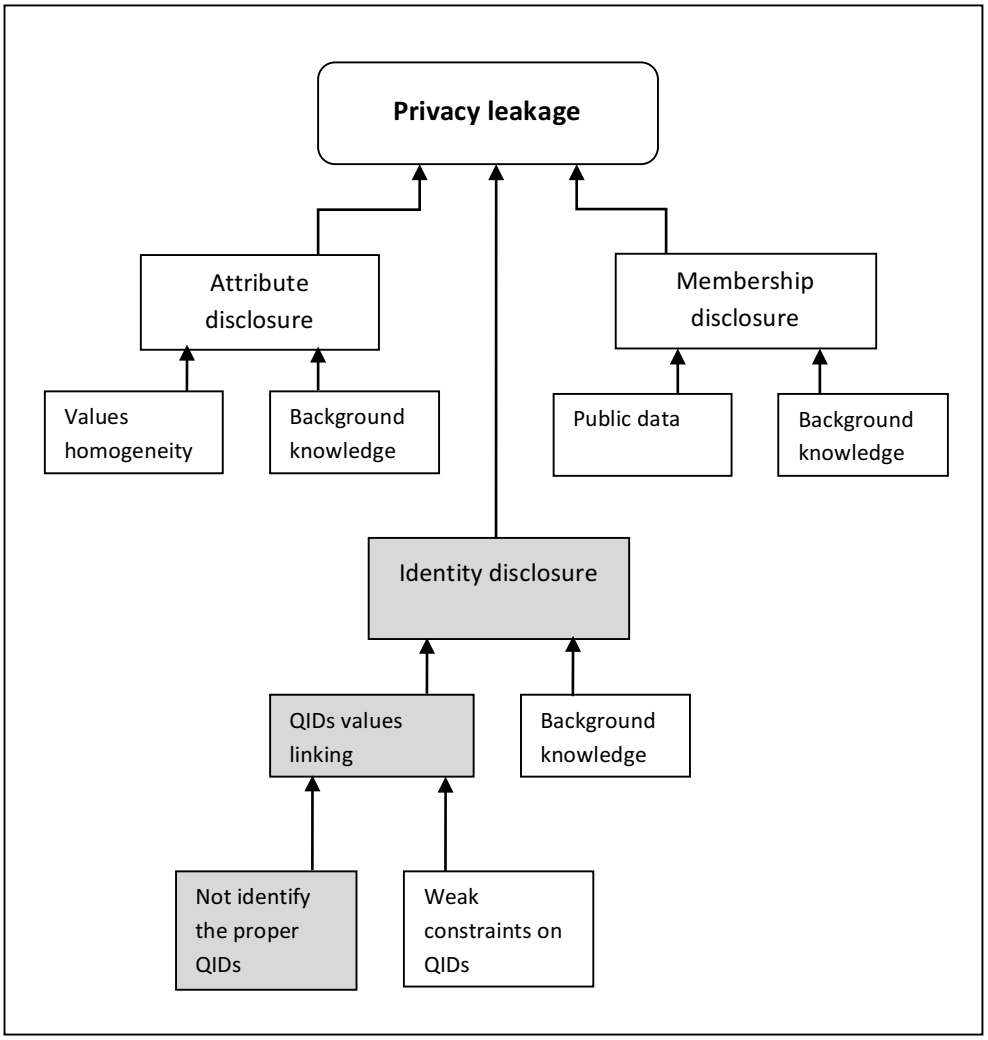


Figure 1: Privacy leakage causes addressed by the QIR algorithm.

Huang and others [36] introduce a new method that depends on the hyper-graph to finding a group of related views and QIDs set. This method maps the group of related views into a hyper-graph and includes all paths available between every two nodes instead of finding the group of related views. The weakness of this method is that the QIDs group produced may include so many attributes. Further, it has high computational complexity resulted from the process of degeneration of the common graph from the hyper-graph.

Omer and Mohamad [37] introduce a new method to select a quasi-identifier (QID) to achieve k-anonymity. Selective and Decompose algorithms depend on nominating multiple attributes as a set, and then generate power set *P(S)* for them. Following that, the distinct values of the power set *P(S)* elements were computed and listed in a table. Finally, the candidate element from the power set is the element with the maximum distinct value. The main problem in this method is selecting the primary nominate set of

attributes, where the accuracy of the selection depends on the user experience [3]. Furthermore, is impractical to generate P(S) if the number of attributes is big (e.g., more than 8).

Y. J. Lee and Lee [38] examine the factors and the likelihood of an individual re-identified for medical information through inferable QIDs. The QIDs were considered as database variables that enable the re-identification of individuals by linking their QIDs with available external information or a specific individual. They selected five factors to form QIDs attributes to prevent patient privacy violations. The factors were selected based on their influence on the likelihood of re-identification and the possibility of inferring it from background knowledge. One of the disadvantages of this study is that the QIDs that can be extracted to re-identification patients' records may exceed 5. Besides, the paper focused only on the problem of re-identification of patients' records and avoiding leakage of privacy in the medical records, lacking a public method that could be used for general data publishing. Bampoulidis and others [8] assume that some QIDs are more important than others (i.e. in data mining/analysis) and, therefore, should be distorted as little as possible in the anonymization process. They present a tool to address the issue of QIDs by utilizing a local recoding algorithm for k-anonymity. The tool outperforms the ARX (data anonymization tool) in terms of dataset quality. The major problem with this method is that it depends on the user in defining the QIDs attributes, giving priority to each attribute, as the user relies on his personal experience in determining the QIDs attributes, which are usually not accurate.

Kaur and Agrawal [10] study the impact of QIDs on the anonymization process. They gave new ways to consider before choosing the quasi-identifiers. The re-identification risks have been examined using different QIDs, diverse parameters, and different sizes of a data sample. The results of their work showed that when making the variance in selecting the QIDs for anonymization operation, note that the risk of re-identification increases when the number of QIDs increases, and it decreases when using QIDs that contain fewer categories. Although it is good to take into account these observations before starting the anonymity process, it should be noted that these observations extracted by the study are not fixed and may change from one dataset to another.

Wong and others [39] do not reveal the complete set of quasi-identifiers (QID) to the data collector before and after the data anonymization process. They believed that the QIDs can be both sensitive values and identifying values, they allow the respondents/data owners to hide sensitive-QIDs attributes from other parties. The first issue with this method is that the QIDs attributes that respondents consider them are sensitive may contain data that are very useful in mining or may adversely affect mining outcomes. The second issue is if respondents submit inaccurate data, there is no guarantee of the usefulness of the results obtained from data analysis.

Sei and others [40] consider that some QIDs are regarded as sensitive QIDs and they propose novel privacy models, namely, $(l1,\dots,lq)-diversity$ and $(t1,\dots,tq)-closeness$, and a method that can treat sensitive QIDs. Their proposed method comprises of two algorithms: anonymization and reconstruction algorithms that can treat sensitive QIDs. Although this method can perform anonymity while preserving the quality of the data, it suffers from the problem of the Wong [39] method, this is because there is no effective method to accurately determine which of the QIDs attributes is considered sensitive QIDs.

Victor and Lopez [41] offer a $(k, n, m)$ anonymity method for sensitive/private data based on the k-anonymity. The graph algorithms were used to perform QIDs and are moreover has been improved by selecting similar QIDs based on the composite and derived attributes. The set of QIDs gets from the methods in [36], [41] may include too many attributes, which is increases the information loss in models based on generalizations like the K-anonymity [3].

## The Proposed QIDs Recognition Algorithm

There are two main stages involved in the QIDs Recognition algorithm (QIR) to prevent privacy leakage of outsourced data. *First*, classification of the dataset attributes into Quasi-Identifiers (QIDs ), Sensitive Attributes (SAs), and Non-Sensitive attributes ( NSs). That is, each attribute in the dataset is classified into one of the aforementioned group (QIDs, SAs, or NSs). In the attributes' classification (QIDs Recognition) stage, the IDs (Identifiers attributes) are usually removed from the dataset by the data owner. The Quasi-Identifiers (QIDs) are the attributes that, when linked together, define the individual. For example, age, gender, and ZIP. The Sensitive Attributes (SAs) are the attributes that explain sensitive/private information about an individual such as medical information, financial records, and location. Meanwhile, the Non-Sensitive attributes (NSs) are the other attributes in the dataset that do not fall under the previously mentioned categories, as they do not help reidentify the identity of the individual for example state and religious attributes. In the basic privacy models (such as k-anonymity [3,8,9,11-13,18,28], l-diversity [40,52], and t-closeness [34, 53]), the attributes of a dataset were categorized into two groups: sensitive and non-sensitive. Meanwhile, most of the recent researchers such as in [3], [42]–[45] divide the datasets attributes into three types: QID, SA, and NS (not including identifiers) directly. Accordingly, the classification of dataset attributes in this study is divided into three types of QID, SA, and NS (not including identifiers) with utilizing the same definitional meaning of each category as in the previous work in [3], [42]–[45].

*Second*, determine the actual dimension of QIDs that should be used in an anonymization operation that will achieve optimum case. If the set of QIDs contains too many attributes, the loss of information caused by generalization will be exacerbated. Nonetheless, sometimes the minimal set of QID does not imply the most appropriate privacy protection setting because the method does not consider what attributes the adversary could potentially have [37]. Therefore, we need a mechanism that determines the appropriate dimension of the QIDs to avoid these problems. In the QIDs-Dimension Determining stage, the proposed algorithm performs this task. Figure 2 illustrates the general procedure of the two main phases of the QIR Algorithm. The following subsections explain these two stages in more detail.

### QIDs Recognition stage

In this stage, the algorithm classifies the attributes depend on the re-identification risk rate for each attribute in the dataset, then the risk rate of the attribute is compared to the threshold values of the

classification. As shown in Figure 2, the attributes classification stage comprises four main activities. These activities include (1) data set preprocessing, (2) computing risk rate for all attributes, (3) selecting the classification thresholds, and (4) Classify the attributes according to the selected thresholds.

In the first activity, the dataset is preprocessed which include filling the missing values, fixing the inconsistencies in the dataset, and data normalization. Then in the second activity, the risk rate is computed according to the g-distinct which is adopted in computing the re-identification risk rate [46]. A detailed description of the g-distinct method is presented in the next section. In third activity, the classification thresholds were selected based on the maximum and minimum risk of re-identification as follows. These thresholds are denoted by β and α in this study, α threshold represents the maximum risk of re-identification of the individual while β represents the minimum risk of re-identification. The threshold values can be determined by the user or the data owner after calculating the re-identification risk for all attributes. Based on percentages of the highest and lowest attributes risk one can choose the α value to be less than the highest risk value and choose the β value to be less than the lowest risk value. The nature of the data and the degree of importance of each attribute affect the selection of the threshold values. So, these thresholds are adjustable and differ from one dataset to another. For instance, let the dataset $(D)$ contains attributes $(A_1, A_2, \dots, A_n)$, i.e., $D = A_1, A_2, \dots, A_n$ let $\beta = 0.05\%$ and $\alpha = 30\%$. Let $Rrisk_{A_i}$ be the re-identification risk of attribute $A_i$, and $Rrisk_{A_i} = 35\%$. As $Rrisk_{A_i} > \alpha$, then the $A_i$ is classified as SA. Suppose $Rrisk_{A_3}$ and $Rrisk_{A_5}$ $are$ 23 $and$ 0.01 respecively, then $A_3$ is classified as $QID$ $w\square ile$ $A_5$ will be classified as NS, respectively. Re-identification risk rate of attribute $A_i$ computes the degree that makes the records distinguished based on this attribute. Finally, the fourth activity includes classify the attributes according to the selected thresholds using rules represented by *if-else* testaments (see Algorithm 1, Lines 27 – 39). In the following subsection, a details description of computing the re-identification risk rate (g-Distinct) is presentedMore explanation of the QIDs Recognition stage are also presented.

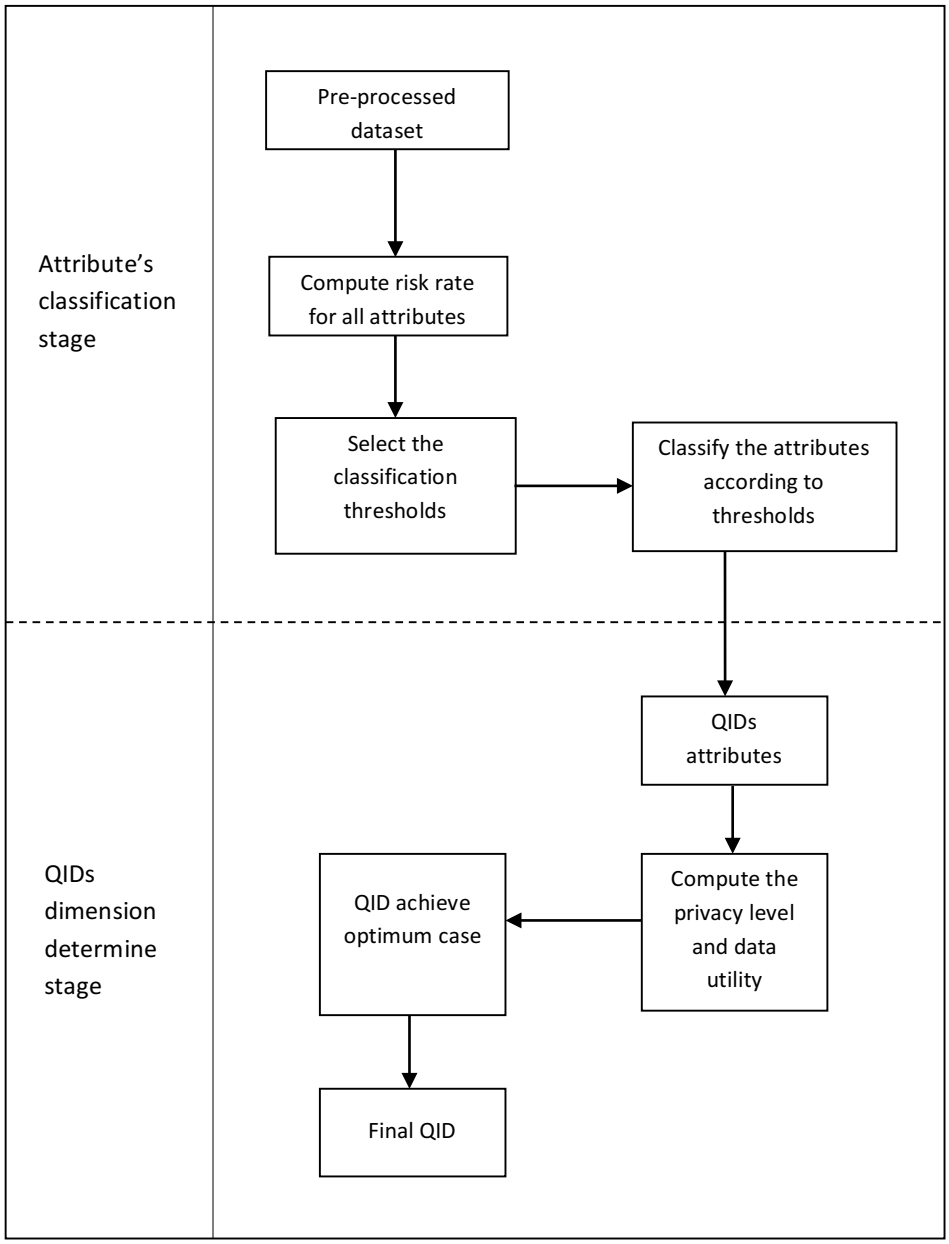


Figure 2: The general procedure of the proposed QIR algorithm.

## g-Distinct

The g-distinct is adopted in computing the re-identification risk rate [46]. A person or record in any dataset is said to be unique if he/she or it has a combination of attributes that is not for someone/record else. The person/record is g-distinct if their combination of attributes is matching to g-1 or less than other

people/records in the dataset [46]. Thus, uniqueness is the base situation of 1-distinct. In general, g-distinct is the total of the number of subgroups with $i$ individuals, which is computed as:

$$h_n(g) = \sum_{i=1}^{g} i \,.\, f_n(i) \qquad (1)$$

Where $f_n(i)$ refers to the expected number of subgroups with $i$ individuals that can be derived from a given aggregated group, and $g$ represents the whole number of individuals in a subgroup. That is, $g$ is associated with the g-distinct to represents the number of distinguished individuals in the subgroup. For example, when we say 3-distinct, it means that three individuals have common QIDs characteristics out of the total number of people $g$ in the subgroup. The Sum of all g-distinct of individuals in a specific attribute represents the Re-identification risk rate that the attribute potential to cause it. We can compute the general risk of the whole dataset through equation (2) where $b$ is the number of possible subgroups.

$$R_n^j(g) = \binom{j}{n} b^{1-n} \left(b^n - (b-1)^n\right) \qquad (2)$$

Finally, the attributes classification stage returns the re-identification risk rate for each attribute in the dataset. Based on the resulted re-identification risk rates, the dataset attributes are classified to sensitive and nom sensitive according to the rate of the re-identification risk for each attribute in addition to threshold values $\beta, \alpha$. The outcomes of this stage will be as input into the QIDs dimension identification stage to determine the dimension of QIDs that suitable to achieve optimal privacy requirements. The practical steps of the classification stage are explained by algorithm 1. The lines 2-16 in algorithm 1 are to compute the g-distinct for all dataset attributes while lines 18-26 are to calculate the re-identification risk rate based on the attributes' g-distinct. Finally, lines 28-40 addressed the process of attributes classification using the re-identification risk rate of each attribute to produce three categories of attributes: QIDs, SAs, and NSs.

The importance of this stage of the proposed algorithm represented by algorithm 1 is that it contributes to reducing the attribute disclosure resulting from linking the QIDs values due to a weakness/failure in defining the QID characteristics correctly. This contribution helps in minimizing the leakage of information and avoid privacy violations.

**Algorithm 1** Attributes Classification

**Input**: Dataset D, β, α.
**Output**: Classified dataset.

*//Compute g-distinct for all dataset tuples for each attribute.*
$Dg_{Attr}$ ⟵ g-distinct of the attribute ($Attr$)
$n$ ⟵ attributes domain
$m$ ⟵ tuples domain
$Attr \in n$
$g \in m$
$tv$ ⟵ attribute value of a specific tuple
$Attr\_Dg\ [i][j] = 0$
**For** $i := 1\ to\ n.length$ **do**
    **For** $j := 1\ to\ m.length$ **do**
        $Dg_{Attr\ [i]} = \frac{1}{\int (tv)_j}$
        $Attr\_Dg\ [i][j] = Attr\_Dg\ [i][j] + Dg_{Attr\ (i)}$ ;
        $j = j + 1;$
    **End**
    $i = i + 1;$
**end**
*//Compute re-identification risk rate for all dataset attributes.*
$Rrisk\_Attr[i] = 0$
$Rrisk_{Attr}$ ⟵ re-identification risk rate of Attr
**For** $i := 1\ to\ Attr\_Dg\ [i].length$ **do**
    **For** j:= 1 to m.length **do**
        $Rrisk_{Attr}$ [i]= $Rrisk_{Attr}$[i ]+$Dg_{Attr\ [i][j]}$
        $j = j + 1;$
    **End**
    $i = i + 1;$
**End**
*//Classified the attributes based on risk rate and threshold values.*
$QIDs\ [\ ] = 0$
$SAs\ [\ ] = 0$
$NSs\ [\ ] = 0$
**For** $i := 1\ to\ Rrisk_{Attr}\ [i]\ .length$ **do**
    **If** ( $Rrisk_{Attr}\ [i]\ in\ range(\beta)$ )
        $QIDs\ [i] = QIDs\ [\ ] + Rrisk_{Attr}\ [i];$
    **Else If** ( $Rrisk_{Attr}\ [i]\ in\ range(\alpha)$ )
        $SAs\ [i] = SAs\ [\ ] + Rrisk_{Attr}\ [i];$
    **Else**
        $NSs\ [i] = SAs\ [\ ] + Rrisk_{Attr}\ [i];$
    $i = i + 1;$
**end**
**Return** ( $QIDs\ [\ ],\ SAs\ [\ ], NSs\ [\ ]$ )

## QIDs-Dimension Identification Stage

This stage of the algorithm aims to determine the best dimension of QIDs that will achieve optimum cases. The optimum case gives high privacy with a high/reasonable percentage of preserving data quality.

In other words, it has high Privacy Gain (PG) with high/reasonable Non-uniform entropy (NUE). Algorithm 2 describes the implementation steps for this stage. The algorithm takes a sample of data with the QID that has the highest re-identification risk rate. Following that, the QIR calculates the PG and NUE base on k-anonymity through equations 3 and 4. In the next step, the QIDs number is increased, the PG and NUE are calculated again and so on until finish all QIDs.

**Algorithm 2** QIDs Dimension Identification

**Input**: Dataset sample d, QIDs [ ], privacy parameter k .
**Output**: Optimal Dimension of QIDs.

1: $QidD \longleftarrow dimension\ of\ QIDs$
2: $QidD \in QIDs\ [\,]$
3: $Optimal\_QidD \longleftarrow Optimal\ dimension\ of\ QIDs$
4: $QidD[\,] = 0$
5: **For** $i := 1\ to\ QIDs\ [\,].length$ **do**
6: $QidD[i] = QidD[\,] + QIDs\ [i]$;
7: $Anonymized\_data\ [i] = k - anonymity(d,\ QidD[i],\ k)$;
8: $PG\ [i] = Privacy\_gain(\ Anonymized\_data\ [i])$;
9: *NUE [i] = Non-Uniform_Entroy ( Anonymized_data [i]);*
10: $Difference\ [i] = PG\ [i] - EIL\ [i]$;
11: $i = i + 1$;
12: **end**
13: **If** $((\ PG\ [\,] == \boldsymbol{max})\ \&\&\ (NUE\ [\,] == \boldsymbol{max}))$
14: $Optimal\_QidD\ [\,] = QidD[i]$;
15: **Return** $(Optimal\_QidD\ [\,])$.

Finally, the algorithm determines the optimum case is that gives high privacy with a high/reasonable percentage of preserving data quality. The best QIDs dimension is the QIDs with the optimum case. Algorithm 2 provides the executive steps of this stage; lines 5-12 implement the anonymization by k-anonymity on a sample of the dataset. It begins with QID that has the highest re-identification risk rate. After that, the algorithm calculates the Privacy Gain (PG) and Non-uniform entropy (NUE) through equations 3 and 4. Then, the QIDs number is increased; the PG and NUE have been calculated repeatedly until all the QIDs are finished. Lastly, in lines 13-15 the algorithm determines the best QIDs dimension (QidD) that achieves the optimum case to be involved in the anonymization process.

It was observed in study [3] that in most cases, when the QIDs dimension is large, the data loss increases. However, when the QID dimension is small, the privacy protection is not applied optimally because one cannot know what the actual QIDs an attacker possesses [37]. Therefore, determining an appropriate QIDs dimension is important to reduce data loss.

## Performance Measures

Two performance evaluation measures were used in this study: the Privacy Gain (PG) and the Non-Uniform Entropy (NUE). More explanation and the derivation of these measures are presented in the following sub-sections.

### The Privacy Gain

To evaluate the privacy level for the proposed algorithm, Equation 3 and Definition 1 are used as follows.

$$PG = A_{t(gen)} - A_{b(gen)} \tag{3}$$

Where $A_{t(gen)}$ is anonymity after generalization (gen), and $A_{b(gen)}$ is anonymity before generalization [27], [47], [48].

**Definition (1)** Anonymity quasi-identifier: A quasi-identifier $qid$ is an anonymity quasi-identifier if $|QIG(qid)| = min_{qid' \in QID} |QIG(qid')|$, where $||$ represents the size of a QI-group [48].

### Non-Uniform Entropy

In the context of data de-identification, the Non-Uniform entropy is to compare the frequencies of attribute values in the transformed dataset according to frequencies in the input dataset; it was originally introduced as a model for measuring the loss of information [49]. When a dataset D is transformed into another dataset $D'$, Non-Uniform entropy is defined as:

$$\Delta(D, D') = \sum_{x \in D} -\log\left(\frac{f(D, x)}{f(D', x)}\right) \tag{4}$$

## Experimental Evaluation

In this section, the experimental evaluation of our implementation algorithm will be presented in terms of PG and NUE. In the Dataset Setup subsection, we describe the datasets we have used for running the experiments and the experimental environment setup. In the experimental results subsection, we present the first set of experiments and provide the results from our algorithm. In the performance benchmark and discussion subsection, we provide benchmark and discussion results of our algorithm against a close recent algorithm introduced by Omer & Mohamad 2016 [37].

## Dataset Setup

Two real-life datasets from the University of California – Irvine were used in this study to demonstrate the performance of the proposed algorithms. The first is the Bank Direct Marketing dataset [50]. The bank dataset consists of 17 attributes and 45,211 tuples and does not include any missing values. The dataset attributes are divided into three divisions are 1) Data of bank clients: age, job, marital, education, default, balance, housing, and loan. In this paper, we will consider these attributes because these attributes are significant for bank clients and re-identification purposes. 2) Data related to the last contact of the current campaign. 3) Other attributes like the campaign and days. The second dataset is the Adult dataset [51] uses as a standard for anonymization algorithms evaluation [8] consist of 48842 census records and 15 attributes.

ARX data anonymization software is open source introduced and developed by Fabian Prasser et al. [52] for data anonymization, we used it to implement the algorithms as explained in the following sections. The experiments were executed on a machine with an Intel Core i7 2.7 GHz processor with 8 GB RAM, under Windows 10.

## Experimental Results

The first experiment is to classify the datasets' attributes according to their risk rate. Figures 3 and 4 illustrate the risk rate for bank attributes and Adult attributes, respectively.

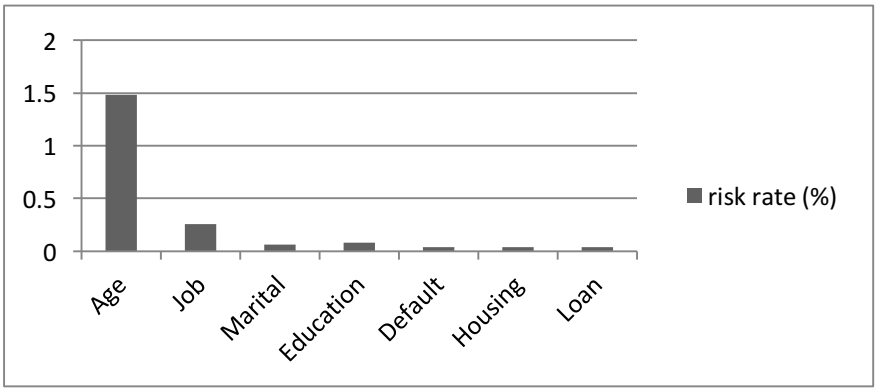


Figure 3: Risk rate of the Bank dataset attributes

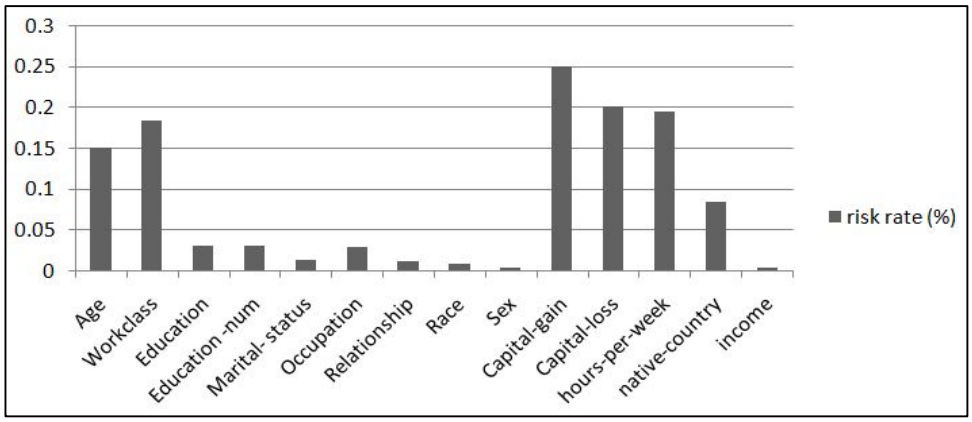


Figure 4: risk rate of the Adult dataset attributes

For the Bank dataset, we identify $\alpha$ and $\beta$ as $\alpha = 30, \beta = 0$. Table 1 demonstrates Bank attributes classification. In the Adult dataset, we add $\alpha = 0.2, \beta = 0.01$ to classify the attributes. Table 2 demonstrates the classification of the adult dataset. Because the "Balance" attribute has a risk is 52.04 %, which is large comparing to other attributes, it is excluded from Figure 3 to highlight the difference between the attributes that have relatively small risk values.

After calculating the risk rate of each attribute in the dataset, the attribute is classified according to the selected threshold $\alpha$ $and$ $\beta$ as was explained in the QIDs Recognition stage section. Table 1 and Table 2 show the classification results of the Bank dataset and Adult dataset, respectively, according to the selected classification thresholds $\alpha$ $and$ $\beta$ for each dataset. After the classification stage, the best dimension of QIDs that achieve optimum case should be determined. In the Bank dataset, the QIDs Dimension($QidD$) is four ($QidD$ = 4) while in the Adult dataset $QidD$ is 10 ($QidD$ = 10). For each dataset, the initial value of QID's Dimension is set to one ($QidD = 1$) to be used as input into the proposed QIDs-Dimension Identification algorithm (as explained in Algorithm 2). Identification of QIDs-Dimension begins with the initial value of $QidD$ and it is incremented until the maximum number of QID's Dimension. Identification of QIDs-Dimension begins also with a sample size equal to 10% of the dataset with k-anonymity of 5 and it is incriminated until K=25 for each $QidD$ value (sample size is changeable). Then, the privacy gain (PG) and the Non-Uniform entropy (NUE) are calculated for each sample and each new $QidD$ until $QidD$ values reach four($QidD$ = 4) for the Bank dataset and $QidD$ = 10 for the Adult dataset.

Table 1: Classification of the Bank dataset

| **Classification** | **Threshold value** $\alpha = 30, \beta = 0$ | **Attributes** |
|---|---|---|
| SAs | Rrisk $>$ α | Balance. |
| QIDs | β ≤ Rrisk $<$ α | Age, Job, Education, and Marital. |
| NSs | Rrisk $<$ β | Default, Housing, and Loan. |

Table 2: Classification of the Adult dataset

| **Classification** | **Threshold value** $\alpha = 0.2, \beta = 0.01$ | **Attributes** |
|---|---|---|
| SAs | $Rrisk > \alpha$ | Capital-gain, Capital-loss. |
| QIDs | $\beta \leq Rrisk < \alpha$ | Hours-per-week, Work-class, Age, Native-country, Education, Education-num, Occupation, Marital-status, Relationship, and Race. |
| NSs | $Rrisk < \beta$ | Sex, Income. |

Finally, the proposed algorithm returns the $QidD$ that achieves the optimum case to be as best dimension will be used in the anonymization process. Table 3 demonstrates the results of finding the best $QidD$ for the Adult dataset.

Table 3: Experiments results for select the best QidD in the Adult dataset.

| QID value | K = 5 | | K = 15 | | K = 25 | |
|---|---|---|---|---|---|---|
| | PG % | NUE % | PG % | NUE % | PG % | NUE % |
| 1 | 33.8 | 30.41 | 38.35 | 21.05 | 38.35 | 21.05 |
| 2 | 55.34 | 44.65 | 76.86 | 23.13 | 76.86 | 23.13 |
| 3 | 77.94 | 22.05 | 83.17 | 16.82 | 83.17 | 16.82 |
| 4 | 79.53 | 20.46 | 84.39 | 15.6 | 84.39 | 15.6 |
| 5 | 83.62 | 16.37 | 87.48 | 12.51 | 87.48 | 12.51 |
| 6 | 85.91 | 14.08 | 89.56 | 10.43 | 89.56 | 10.43 |
| 7 | 86.65 | 13.34 | 86.65 | 13.34 | 89.69 | 10.3 |
| 8 | 90.51 | 9.48 | 90.51 | 9.48 | 90.51 | 9.48 |
| 9 | 92.68 | 7.31 | 92.68 | 7.31 | 92.68 | 7.31 |
| 10 | 91.59 | 8.4 | 91.59 | 8.4 | 91.59 | 8.4 |

According to Table 3, we observed that QidD = 2 is the optimum case that increases the privacy gain as well as the NUE. Moreover, we can notice that the privacy level also increases when QidD value increases. The privacy gain reaches 91.59 % when the QidD is 10. On the other hand, the NUE decreases, and accordingly, the data utility decreases when QidD increases. Figures 5a, 5b, and 5c demonstrate the selection of the best QidD for the Bank dataset by the proposed QIR algorithm on different k-anonymity values, 5, 15, and 25, respectively. In the bank dataset, the proposed algorithm the selected QIDs attributes are Work-class and Hours-per-week (HPW). These two attributes achieve the highest re-identification risk, thus, they must be involved in the anonymization process (See Figure 4).

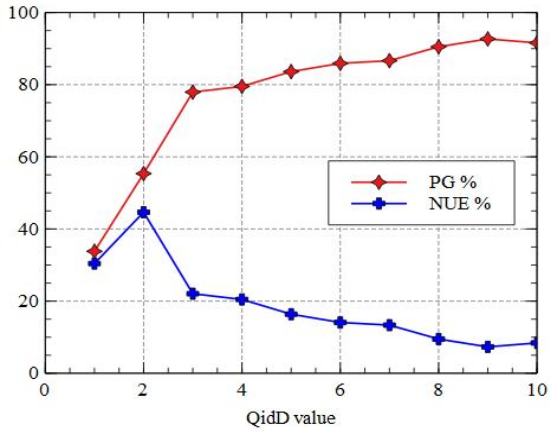


Figure 5 (a): k = 5

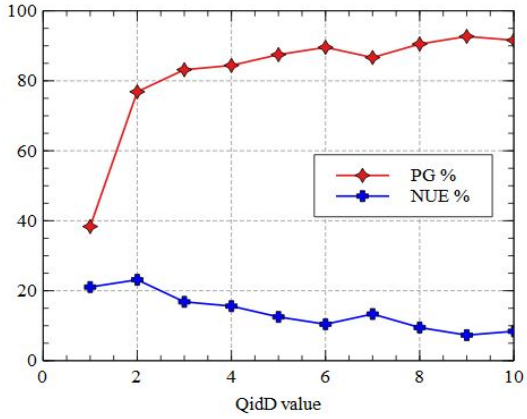


Figure 5 (b): k = 15

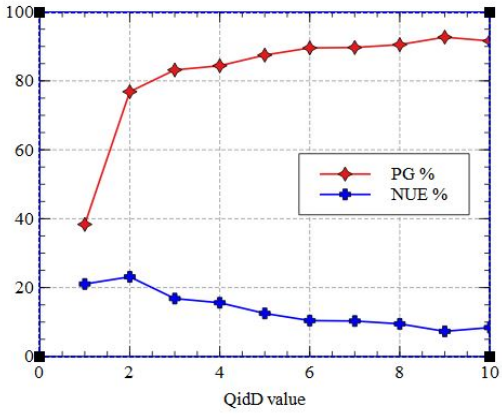


Figure 5 (c): k = 25

Figure 5: (a, b, and c). Best QidD selection for Bank dataset by QIR on different k values.

To determines the best QidD in the Bank dataset track Table 4 and Fig. 6 (a, b, and c), it is clear that when QidD = 1 the proposed algorithm achieves the optimum case as it gives high privacy in several cases of k values. It can be also observed in Table 4 that the NUE drops from 45.28% when K=5 to

17.27% when K increases above 15. It is also noticeable in the Bank database that privacy decreases at increase the value QidD which is normal with the level of privacy provided.

Table 4: Experiments results for select the best QidD in Bank dataset

| QidD | QID | K = 5 | | K = 15 | | K = 25 | |
|---|---|---|---|---|---|---|---|
| | | PG % | NUE % | PG % | NUE % | PG % | NUE % |
| 1 | age | **23.89** | **45.28** | **36.12** | **17.27** | **36.12** | **17.27** |
| 2 | age, job | 21.83 | 36.65 | 21.83 | 36.65 | 21.83 | 36.65 |
| 3 | age, job, marital | 15.83 | 40.35 | 16.67 | 37.18 | 17.94 | 32.37 |
| 4 | age, job, marital, education | 14.88 | 35.93 | 14.88 | 35.93 | 16.43 | 29.26 |

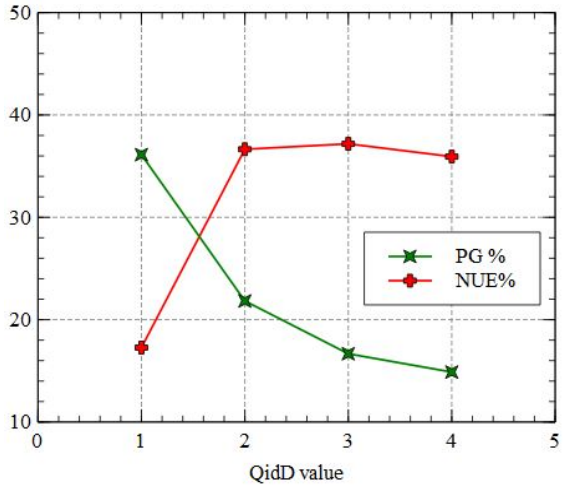


Figure 6 (a): k = 15

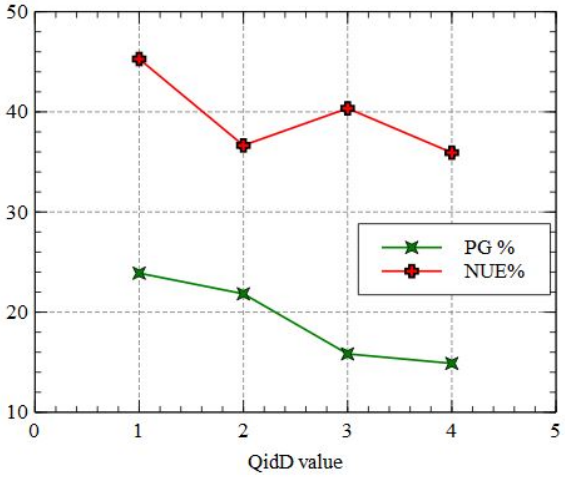


Figure 6 (b): k = 5

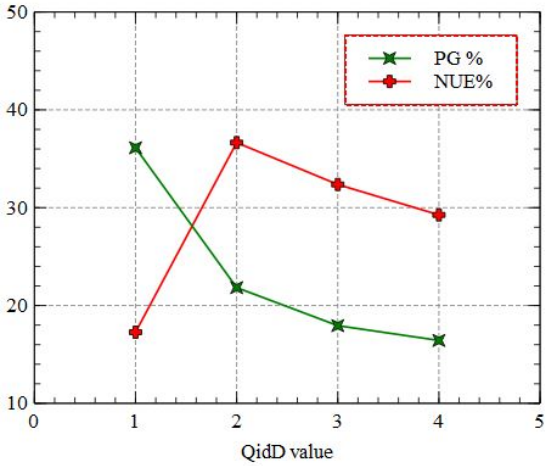


Figure 6 (c): k = 25

Figure 6: (a, b, and c). Best QidD selection for Bank dataset by QIR on different k values.

**Performance Benchmark and Discussion**

To evaluate the proposed QIR algorithm, we compare it with a recent based on k-anonymity against recent similar work SQI algorithm [37]. The comparison was conducted in terms of their Privacy Gain (PG) and Non-uniform Entropy (NUE). . Multiple *k* values and different dataset sizes of the Adult will be used. In Figure 7 and Figure 8 . the privacy provided by QIR is more than the privacy achieved by SQI, where the improvement average exceeds 23%. Although SQI outperformed the QIR in data utility represented by NUE at k = 26, 29, 35, with a privacy rate of 9.57%, this is considered a deficiency because QIR provided data utility higher than that with much higher privacy at k = 4, 6, 10, 17 and 20.

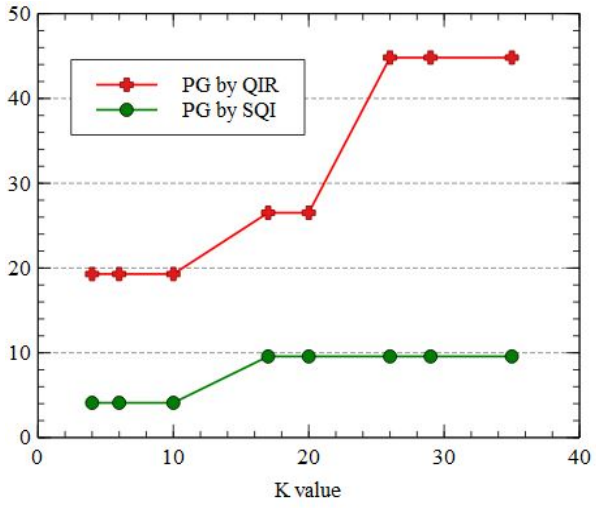


Figure 7: PG at several k values

Dataset: Adult dataset 48842 tuple
QidD of QIR = 2 (Work class, HPW)
QidD of SQI = 1 (Age)

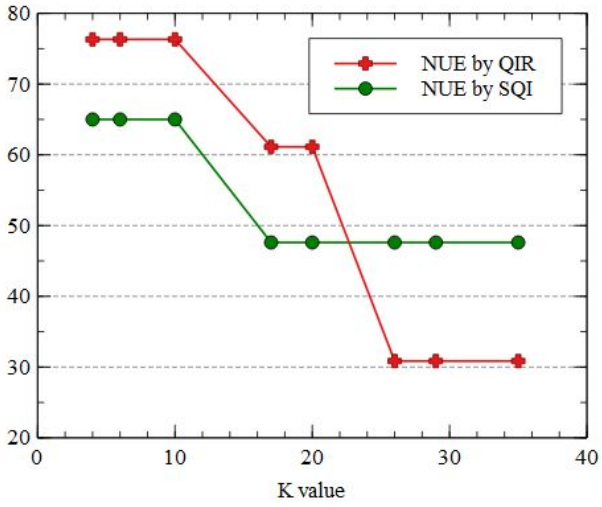


Figure 8: NUE at several k values

Dataset: Adult dataset 48842 tuple
QidD of QIR = 2 (work class, HPW)
QidD of SQI = 1 (Age)

In Figures 9 and 10, it can be observed that at 10% of the dataset and k =10 the privacy achieved by the proposed QIR algorithm is more than double the privacy achieved by the SQI algorithm with slight increases in data utility. That is the proposed QIR algorithm outperforms the SQI algorithm in terms of preserving privacy and data utility. With data size, 20% and k = 20, the NUE obtained by SQI, QIR is 30.27, 31.66 % respectively while the privacy given by SQI is 20.52% and by QIR is 51.82 which is twice time more than that is achieved by SQI. Similar results were obtained at k=20 and data size =30% and 90%, respectively. In most cases, when data size increases the privacy decreases, and therefore the data utility increases.

Generally, for the whole Adult data, results of the experiments at k = 10, and k=20 show that the average privacy percentage presented by SQI is 10.17% with 48.62% data utility, while the average privacy percentage offered by the proposed QIR is 46.49% with 41.04% data utility. As well for the whole Adult dataset and all K values experimented the average privacy provided by SQI is 7.51% against 54.13% data utility. While the average privacy percentage achieved by QIR is 30.67% against 55.46% data utility, hence, using QIR for identification of the real QIDs is considered more ideal.

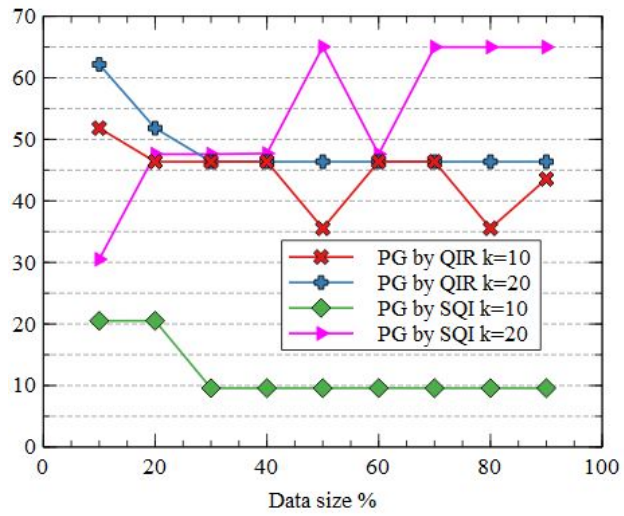


Figure 9: PG at several data sizes

Dataset: Adult dataset
QidD of QIR = 2 (workclass, HPW)
QidD of SQI = 1 (Age)
K = 10, 20

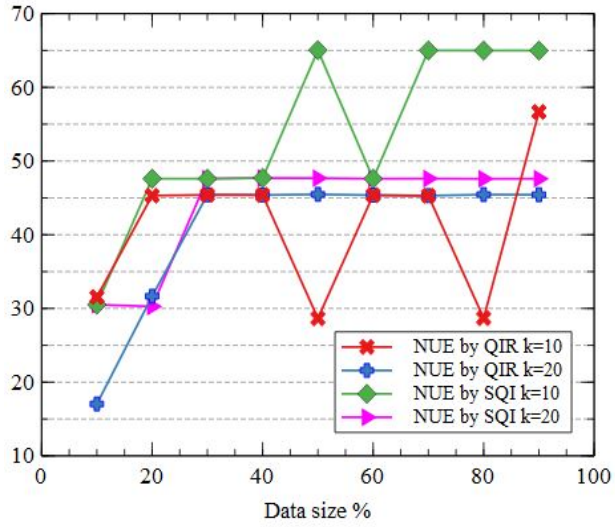


Figure 10: NUE at several data sizes

Dataset: Adult dataset
QidD of QIR = 2 (workclass, HPW)
QidD of SQI = 1 (Age)
K = 10, 20

## Conclusions

Accurate Identification of QIDs is an important issue for the success and validity methods of privacy-preserving outsourced data that seek to avoid privacy leakage caused by QIDs linking. This paper aims to classify dataset attributes before the anonymization process and determines the proper QIDs that should be involved in anonymity operation. A new algorithm is proposed based on the calculation of the re-identification risk for dataset attributes to classify attributes to SAs, QIDs, and NSs based on pre-specified thresholds. In addition to attributes classification, the algorithm determines the actual dimension of QIDs that is required in the anonymization process depending on the amount of privacy provided versus a loss of the quality of the data. The experiment results indicated that the proposed identification algorithm has better performance and is more perfect in terms of privacy provided against data utility when compared with other work. Although the proposed algorithm is suitable to be used with any method or privacy model concerned with QIDs attributes, in this paper we have relied on the K-anonymity model.

## Data Availability

All data that used in this article are available in the machine learning repository at the University of California, Irvine (UCI): https://archive.ics.uci.edu/ml/datasets/.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgment

The authors would like to acknowledge Taif University Researchers Supporting Project number (TURSP-2020/292) Taif University, Taif, Saudi Arabia.

## Funding Statement

This research was funded by Taif University Researchers Supporting Project number (TURSP-2020/292) Taif University, Taif, Saudi Arabia.

## References

[1] J. Domingo-Ferrer, O. Farràs, J. Ribes-González, and D. Sánchez, "Privacy-preserving cloud computing on sensitive data: A survey of methods, products, and challenges," *Computer Communications*, vol. 140–141. Elsevier B.V., pp. 38–60, 01-May-2019.

[2] S. Aldeen Yousra and S. Mazleena, "A New Heuristic Anonymization Technique for Privacy Preserved Datasets Publication on Cloud Computing," *J. Phys.*, vol. 1003, no. 1, pp. 0–15, 2018.

[3] Z. L. Yan Y, Wang W, Hao X, "Finding quasi-identifiers for k-anonymity model by the set of cut-vertex," 2018.

[4] Contel Bradford, "7 Most Infamous Cloud Security Breaches - StorageCraft," *storagecraft*, 2020. [Online]. Available: https://blog.storagecraft.com/7-infamous-cloud-security-breaches/. [Accessed: 18-Oct-2020].

[5] Chen B, Cheung P, Cheung P, and Kwok Y, "Cypherdb: A novel architecture for outsourcing secure database processing," *ieeexplore.ieee.org*, 2018.

[6] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, 2010.

[7] S. A. Abdelhameed, S. M. Moussa, and M. E. Khalifa, "Privacy-preserving tabular data publishing: A comprehensive evaluation from web to cloud," *Comput. Secur.*, 2018.

[8] A. Bampoulidis, I. Markopoulos, and M. Lupu, "PrioPrivacy: A local recoding K-anonymity tool for prioritised Qasi-identifiers," in *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence Workshops, WI 2019 Companion*, 2019, pp. 314–317.

[9] L. Sweeney, "Achieving k-anonymity Privacy Protection using Generalization and Suppression," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 571–588, 2002.

[10] G. Kaur and S. Agrawal, *Differential Privacy Framework : Impact of Quasi-identifiers on Anonymization*, vol. 46. Springer Singapore, 2019.
[11] D. Wei, K. Natesan Ramamurthy, and K. R. Varshney, "Distribution-preserving k-anonymity," *Stat. Anal. Data Min.*, vol. 11, no. 6, pp. 253–270, Dec. 2018.
[12] P. R. Bhaladhare and D. C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model," *J. Inf. Sci. Eng.*, vol. 32, no. 1, pp. 63–78, Jan. 2016.
[13] M. S. Simi, K. S. Nayaki, and M. S. Elayidom, "An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity," in *IOP Conference Series: Materials Science and Engineering*, 2017.
[14] H. Kaur, N. Kumar, and S. Batra, "ClaMPP: a cloud-based multi-party privacy preserving classification scheme for distributed applications," *J. Supercomput.*, vol. 75, no. 6, pp. 3046–3075, 2019.
[15] G. G. Dagher, B. C. M. Fung, N. Mohammed, and J. Clark, "SecDM: privacy-preserving data outsourcing framework with differential privacy," *Knowl. Inf. Syst.*, vol. 62, no. 5, pp. 1923–1960, May 2020.
[16] A. F. Westin, "Privacy and Freedom.," *Am. Sociol. Rev.*, vol. 33, no. 1, p. 173, 1968.
[17] M. Templ, *Statistical disclosure control for microdata: Methods and applications in R*. 2017.
[18] W. Mahanan, W. A. Chaovalitwongse, and J. Natwichai, "Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT," in *Service Oriented Computing and Applications*, 2020, vol. 14, no. 2, pp. 89–100.
[19] S. Mayil, M. Vanitha, C. Science, J. J. College, and T. St, "A Survey on Privacy Preserving Data Mining Techniques for Clinical Decision Support System," vol. 5, no. 5, pp. 6054–6056, 2016.
[20] N. Uttarwar and M. A. Pradhan, "K-NN DATA CLASSIFICATION TECHNIQUE USING SEMANTIC SEARCH ON ENCRYPTED RELATIONAL DATA BASE," in *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*, 2017.
[21] K. El Makkaoui, A. Beni-Hssane, A. Ezzati, and A. El-Ansari, "Fast Cloud-RSA Scheme for Promoting Data Confidentiality in the Cloud Computing," *Procedia Comput. Sci.*, vol. 113, pp. 33–40, Jan. 2017.
[22] W. Wang, L. Chen, and Q. Zhang, "Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation," *Comput. Networks*, vol. 88, pp. 136–148, Sep. 2015.
[23] K. El Makkaoui, A. Beni-Hssane, and A. Ezzati, "Speedy Cloud-RSA homomorphic scheme for preserving data confidentiality in cloud computing," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 12, pp. 4629–4640, 2019.
[24] D. Chandravathi and P. V. Lakshmi, "Privacy preserving using extended euclidean algorithm applied to RSA-homomorphic encryption technique," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 10, pp. 3175–3179, 2019.
[25] P. Shyja Rose, J. Visumathi, and H. Haripriya, "Research paper on privacy preservation by data anonymization in public cloud for hospital management on big data," *Int. J. Control Theory Appl.*, 2016.
[26] Y. A. A. S. Aldeen and M. Salleh, "Privacy Preserving Data Utility Mining Architecture," in *Smart Cities Cybersecurity and Privacy*, Elsevier Inc., 2019, pp. 253–268.
[27] Y. A. A. S. Aldeen and M. Salleh, "Techniques for Privacy Preserving Data Publication in the Cloud for Smart City Applications," in *Smart Cities Cybersecurity and Privacy*, Elsevier Inc., 2019, pp. 129–145.
[28] Y. A. A. S.Aldeen and M. Salleh, "A Hybrid K-anonymity Data Relocation Technique for

Privacy Preserved Data Mining in Cloud Computing," *J. Internet Comput. Serv.*, vol. 17, no. 5, pp. 51–58, Oct. 2016.
[29] H. Lee, S. Kim, J. W. Kim, and Y. D. Chung, "Utility-preserving anonymization for health data publishing," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, p. 104, Jul. 2017.
[30] Y. A. A. S. Aldeen, M. Salleh, and Y. Aljeroudi, "An innovative privacy preserving technique for incremental datasets on cloud computing," *J. Biomed. Inform.*, vol. 62, pp. 107–116, Aug. 2016.
[31] S. R. . Reddy, K. V. S. V. . Raju, and V. Valli Kumari, "Personalized privacy preserving incremental data dissemination through optimal generalization," *J. Eng. Appl. Sci.*, 2018.
[32] R. V. Sudhakar and T. C. M. Rao, "Security aware index based quasi–identifier approach for privacy preservation of data sets for cloud applications," *Cluster Comput.*, 2020.
[33] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale, and J. A. Oguntuase, "KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes," *Inf. Secur. J.*, vol. 26, no. 3, pp. 121–135, May 2017.
[34] R. Wang, Y. Zhu, T.-S. Chen, and C.-C. Chang, "Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness," *J. Comput. Sci. Technol.*, vol. 33, no. 6, pp. 1231–1242, Nov. 2018.
[35] S. Srijayanthi, T. Sethukarasi, and A. Thilagavathy, "Efficient anonymization algorithm for multiple sensitive attributes," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 4961–4963, Nov. 2019.
[36] L. Huang, J. Song, Q. Lu, X. Liu, and C. Zhang, "Hypergraph-based Solution for Selecting Quasi-identifier," vol. 6, no. November, pp. 597–606, 2012.
[37] A. M. Omer and M. M. Bin Mohamad, "Simple and effective method for selecting quasi-identifier," *J. Theor. Appl. Inf. Technol.*, vol. 89, no. 2, pp. 512–517, 2016.
[38] Y. J. Lee and K. H. Lee, "Re-identification of medical records by optimum quasi-identifiers," pp. 428–435, 2017.
[39] K. S. Wong, N. A. Tu, D. M. Bui, S. Y. Ooi, and M. H. Kim, "Privacy-Preserving Collaborative Data Anonymization with Sensitive Quasi-Identifiers," in *2019 12th CMI Conference on Cybersecurity and Privacy, CMI 2019*, 2019.
[40] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness," *IEEE Trans. DEPENDABLE Secur. Comput.*, vol. 16, no. 4, pp. 580–593, 2019.
[41] N. Victor and D. Lopez, "Privacy preserving sensitive data publishing using (k,n,m) anonymity approach," *J. Commun. Softw. Syst.*, vol. 16, no. 1, pp. 46–56, Mar. 2020.
[42] H. Y. Tran and J. Hu, "Privacy-preserving big data analytics a comprehensive survey," *J. Parallel Distrib. Comput.*, vol. 134, pp. 207–218, Dec. 2019.
[43] E. E. Brown, "Improving privacy preserving methods to enhance data mining for correlation research," in *Conference Proceedings - IEEE SOUTHEASTCON*, 2017, pp. 3–6.
[44] X. Jiang, A. D. Sarwate, and L. Ohno-Machado, "Privacy technology to support data sharing for comparative effectiveness research: a systematic review.," *Med. Care*, vol. 51, no. 8 Suppl 3, pp. S58-65, Aug. 2013.
[45] K. Patel and G. B. Jethava, "Privacy Preserving Techniques for Big Data: A Survey," in *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, 2018, pp. 194–199.
[46] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," pp. 169–177, 2010.
[47] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, "Combining Top-Down and Bottom-

Up: Scalable Sub-tree Anonymization over Big Data Using MapReduce on Cloud," in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2013, pp. 501–508.

[48] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud," *J. Comput. Syst. Sci.*, vol. 80, no. 5, pp. 1008–1020, 2014.

[49] F. Prasser, R. Bild, and K. A. Kuhn, "A Generic method for assessing the quality of De-Identified health data," *Stud. Health Technol. Inform.*, vol. 228, pp. 312–316, 2016.

[50] S. Moro, P. Cortez, and P. Rita, "UCI Machine Learning Repository: Bank Marketing Data Set," 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. [Accessed: 04-Oct-2020].

[51] Ronny Kohavi and Barry Becker, "Adult Census Income | Kaggle," 2016. [Online]. Available: https://www.kaggle.com/uciml/adult-census-income. [Accessed: 04-Oct-2020].

[52] F. Prasser, K. A. Kuhn, and J. Eicher, "Flexible data anonymization using ARX — Current status and challenges ahead," no. August 2019, pp. 1277–1304, 2020.

[53] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "*L-diversity: privacy beyond k-anonymity*," 22nd International Conference on Data Engineering (ICDE'06), 2006, pp. 24-24, doi: 10.1109/ICDE.2006.1.

[54] N. Li, T. Li and S. Venkatasubramanian, "t-*Closeness: Privacy Beyond k-Anonymity and l-Diversity,*" 2007 IEEE 23rd International Conference on Data Engineering, 2007, pp. 106-115, doi: 10.1109/ICDE.2007.367856.