applied sciences

Article

A New Approach to Privacy-Preserving Multiple Independent Data Publishing

A S M Touhidul Hasan [1,2], Qingshan Jiang [1,*], Hui Chen [1] and Shengrui Wang [3]

1 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; touhidul.hasan@siat.ac.cn (A.S.M.T.H.); hui.chen1@siat.ac.cn (H.C.)
2 Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China
3 Department of Computer Science, University of Sherbrooke, Sherbrooke, QC J1K2R1, Canada; Shengrui.Wang@Usherbrooke.ca
* Correspondence: qs.jiang@siat.ac.cn; Tel.: +86-0755-8639-2340

Received: 17 April 2018; Accepted: 10 May 2018; Published: 14 May 2018

Featured Application: The Merging method might apply to publish the datasets sequentially from the different organizations where it will ensure more data utility and privacy.

Abstract: We study the problem of privacy preservation in multiple independent data publishing. An attack on personal privacy which uses independent datasets is called a composition attack. For example, a patient might have visited two hospitals for the same disease, and his information is independently anonymized and distributed by the two hospitals. Much of the published work makes use of techniques that reduce data utility as the price of preventing composition attacks on published datasets. In this paper, we propose an innovative approach to protecting published datasets from composition attack. Our cell generalization approach increases both protection of individual privacy from composition attack and data utility. Experimental results show that our approach can preserve more data utility than the existing methods.

Keywords: anonymization; composition attack; privacy preservation; data publishing

---

1. Introduction

Data sharing helps the individual researcher and research organizations to run data analytics operations on published databases. However, the publishing of data may jeopardize personal privacy and disclose the sensitive values [1]. In recent decades, the sharing of personal data has resulted in numerous incidents involving data privacy breaches [2,3], with disastrous results for the reputations and finances of organizations. Privacy-preserving data publishing methods are anonymizing the published data to preserve user privacy while allowing organizations to release their datasets [4].

Personal privacy is ensured by privacy-preserving data publishing methods and anonymization of the data at the time of widespread publication. Although identifying attributes like social security numbers and names are never published for data mining purposes, sensitive data may still flow due to linking attacks, whereby an attacker may reveal hidden identities or sensitive data by linking the published data attributes with other publicly available data sources [5]. The attributes that can be efficiently used to create such links, such as sex, zip code, and age, are called quasi-identifiers (*QIs*). Anonymization requires the alteration of these attributes to prevent such attacks while preserving the maximum possible utility of the released data.

k-anonymity [6] is the first privacy model for privacy-preserving data publishing which generalizes the attribute values of the quasi-identifiers so that each of the released records becomes indistinguishable from at least k-1 other records when predicted on those attributes. As a result, each person can be associated only with sets of records of size at least k in the anonymized table. While the goal of *k-anonymity* is to prevent identity disclosure, the later privacy models *l-diversity* [7] and *t-closeness* [8] aim at preventing disclosure of sensitive attributes by requiring restrictions on the distribution of sensitive values in each subset of records that are indistinguishable by their *QIs*.

Existing anonymization methods mainly concentrate on one-time data publication [6,7,9], in which a data publisher anonymizes a dataset without considering other published datasets. In many cases, multiple views of a dataset [10,11] or a series of datasets in distinct time stamps [12–16] are published. An example of the former case is the publication of data with different generalization schemes for different purposes, and an example of the latter is a quarterly publication of hospital data. Both examples are multiple-time data publications. Our previously published research [17] concentrates on multiple-time data publishing for bike sharing datasets. When the information of an individual remains in multiple datasets, an adversary may examine the intersection of some anonymized datasets to reveal the individual's private information even though it is preserved in each separate publication [12,18].

Let us look at an example of how multiple publications can lay information open to a composition attack, which uses the intersection of some published datasets to deduce the sensitive values of individuals whose records are in multiple datasets. Tables 1 and 2 contain data segments from two hospitals, both including the same person's health records. Assume that Boby's personal information (Age = 22, Sex = Male, Zipcode = 47905), is known to the adversary. The adversary also knows that Boby visited two hospitals for medication. We can assume that the two hospitals published their data without consulting each other. Tables 3 and 4 are the anonymized tables published by the two hospitals. We will see that this would result in an increased probability of breaching Boby's privacy from their published data. It is true that the adversary cannot find a person's sensitive information in either dataset since both satisfy *k-anonymity* or *l-diversity*. However, the intersection of Tables 3 and 4 shown in Table 5 comprises only those individuals who have visited both hospitals or have the same *QI* and sensitive values. Now, from Table 5, the adversary can link Boby's *QI* values with the sensitive value, breaching Boby's personal privacy.

Multiple independent data publishing poses new challenges for data privacy and the utility of the published data. In multiple independent data publications, a data owner does not know which published dataset may be used for a composition attack. Multiple independent data publications are different from traditional multiple-time data publications, such as multiple-view data publication [10,11] and series data publication [12–14], in which a data publisher is familiar with all the datasets (different views or previous versions of the current dataset) that could be used for composition attacks and can use information in the known datasets to anonymize the current dataset. Since there is no communication or information sharing between data owners in multiple independent data publications, collaborative privacy-preserving data publishing techniques [19–21] cannot be used to protect privacy in this case. In addition, we published earlier research on privacy-preserving data publishing [15,17,22]. To reduce the likelihood of composition attacks on published datasets, existing anonymization techniques [2,23,24] utilize generalization and perturbation, which decrease the data utility.

Table 1. Microdata of hospital A.

Name	Age	Sex	Zipcode	Disease
Boby	22	M	47905	Gastritis
Alisa	22	F	47905	Ovarian Cancer
Tina	33	F	47901	Breast Cancer
Doug	52	M	47901	Flu
Kevin	54	M	47902	Dyspepsia
Sandy	60	F	47902	Fever
Molly	60	F	47308	Cancer
Dolly	64	F	47308	Fever

Table 2. Microdata of hospital B.

Name	Age	Sex	Zipcode	Disease
Boby	22	M	47905	Gastritis
Aron	22	M	47905	Blood Cancer
Angela	33	F	47907	Breast Cancer
Arnold	58	M	47903	Flu
Blake	30	M	47902	Dyspepsia
Sandy	60	F	47902	Fever
Camila	65	F	47308	Flu
Easter	65	F	47309	Cancer

Table 3. Anonymized table of hospital A (Mondrian).

Age	Sex	Zipcode	Disease
22	*	47905	Gastritis
22	*	47905	Ovarian Cancer
33–52	*	47901	Breast Cancer
33–52	*	47901	Flu
54–60	*	47902	Dyspepsia
54–60	*	47902	Fever
60-64	F	47308	Cancer
60-64	F	47308	Fever

* denotes the generalization of the attribute value.

Table 4. Anonymized table of hospital B (Mondrian).

Age	Sex	Zipcode	Disease
22	M	47905	Gastritis
22	M	47905	Blood Cancer
33–58	*	4790 *	Breast Cancer
33–58	*	47905	Flu
30–60	*	47902	Dyspepsia
30–60	*	47902	Fever
65	F	4730 *	Flu
65	F	4730 *	Cancer

* denotes the generalization of the attribute value.

Table 5. Intersection of Tables 3 and 4.

Age	Sex	Zipcode	Disease
22	M	47905	Gastritis
33	*	4790 *	Breast Cancer
			Flu
60	*	47902	Dyspepsia
		47902	Fever

* denotes the generalization of the attribute value.

In this paper, we use some ideas from [17] and propose a new approach called Merging for protection against composition attack in various independent data publications while preserving better data utility. It partitions the data both vertically and horizontally. In the vertical partitions, highly correlated attributes are grouped into columns and each resulting column will then contain a subset of attributes. In the horizontal partition, the tuples are grouped in buckets or equivalence classes. In an equivalence class, the attribute values are randomly permuted to break the association between different columns. We introduce the cell generalization approach to increase the privacy of the published dataset. Hence each *QI* value will be linked with *l* distinct sensitive values, reducing the confidence that the adversary will have when breaching personal privacy. Tables 6 and 7 are the published tables from the two hospitals, generated using the anonymization technique proposed in this paper.

The idea behind our approach is to increase the probability of false matches by linking the *QI* values with the *l* distinct sensitive values [25]. When a person's record is similar in two datasets, there will be common values in the intersection of the anonymized datasets, including *QI* values and sensitive values. When a person's record is not in the two datasets, there may still be a common record in both anonymized datasets, induced by two different patients having the same *QI* and sensitive values. Such a match is called a false match. We consider the example as mentioned earlier where user privacy is breached by the intersection of two published datasets. Tables 6 and 7 are published by the Merging method, and Table 8 is the intersection of Tables 6 and 7. From the intersection, the adversary cannot discover the actual *QI* values of the user. In fact, the first bucket or equivalence class of Table 8 contains 44 *QI* values. Now the adversary will need another publicly available data source to match his desired *QI* values, which we call a true match. The adversary will want to link this true match with the sensitive value. However, in the sensitive values column, i.e., Disease, there are three distinct values. It will thus be difficult for him to deduce an exact sensitive value for the particular *QI* values.

Table 6. Anonymized table of hospital A (Merging).

(Age, Sex)	Zipcode	Disease
(22, *)	47905	Ovarian Cancer
(22, *)	47901	Gastritis
(33–52, *)	47905	Flu
(33–52, *)	47901	Breast Cancer
(54, M)	47308	Fever
(60, F)	47902	Dyspepsia
(60, F)	47308	Cancer
(64, F)	47902	Fever

* denotes the generalization of the attribute value.

Table 7. Anonymized table of hospital B (Merging).

(Age, Sex)	Zipcode	Disease
(22, *)	4790 *	Blood Cancer
(22, *)	47905	Gastritis
(33–58, *)	4790 *	Breast Cancer
(33–58, *)	47905	Flu
(30, M)	4730 *	Flu
(60, F)	47902	Fever
(65, F)	4730 *	Cancer
(65, F)	47902	Dyspepsia

* denotes the generalization of the attribute value.

Table 8. Intersection of Tables 6 and 7.

(Age, Sex)	Zipcode	Disease
(22, *)	47905	Gastritis
(33, *)	4790 *	Breast Cancer
		Flu
(60, F)	47902	Fever
	4730 *	Cancer
		Dyspepsia

* denotes the generalization of the attribute value.

The essential aim of our proposed Merging anonymization technique is to increase the probability of false matches during a composition attack. In a real-world scenario, since there will be more records, there will be a stronger probability of producing a false match for a *QI* value. When we consider the ambiguity behind true and false matches, it is conceivable that the probability of a false match is higher than that of a true one. When a significant difference in such probabilities is achieved, the privacy of an individual is protected in multiple independent data publications. Putting this principle in the differential privacy context, the appearance of a common record in the published datasets is independent of whether or not the common record belongs to the same individual, and hence an adversary cannot be sure whether the sensitive value in the common record belongs to the person.

The main contributions of the paper are summarized as follows. Equivalence classes are created, and attribute values are randomly permuted in the equivalence class to break the cross-column relation to increasing the published data privacy. A cell generalization approach is introduced to protect the published datasets from composition attacks. In addition, we present the anonymization algorithm which can successfully anonymize the dataset to ensure the protection from composition attack and increase the data utility as well. The proposed method can protect the anonymized data from privacy breach by satisfying the *l*-diversity privacy requirements. We conduct the extensive experiments on real-world data to compare with the other state-of-art techniques to support the effectiveness of the Merging method.

The remainder of this paper is structured as follows. Background and related work are reviewed in Section 2. Sections 3 and 4 give the details of the proposed system and anonymization algorithm. We present experimental analysis in Section 5 and conclude in Section 6.

2. Background and Related Work

In this section, we review the existing anonymization techniques, focusing on multiple independent data publishing, and discuss the background knowledge for the published microdata tables.

2.1. Privacy-Preserving Data Publishing Context

In order to preserve user privacy, an essential privacy context needs to be defined for the privacy-preserving data publishing. For determining a particular privacy setting, recently published studies [26–29] classified the essential privacy terminology for the cyberspace, and these include sender, recipient, attacker, anonymity, pseudonymity, identifiability, identity confidentiality, unlinkability, undetectability, unobservability, and identity management. Pfitzmann and Hansen [27–29] represent a privacy context that illustrates the relationship between fundamental privacy terms.

In the privacy setting, a sender transfers his dataset to a recipient where an attacker will not be able to gain any knowledge about that dataset. This privacy setting is followed for the privacy-preserving data publishing event. In the privacy-preserving data publishing context, a data publisher publishes the data to the public, and it is open to everybody. An attacker, i.e., adversary also receives the published data, and he might use previously gained background knowledge to identify a person by linking with some publicly available data sources [6]. Hence, the requirement for anonymity is significantly present in the privacy-preserving data publishing context [26]. The anonymity of an individual is the anonymous properties of the particular individual in the dataset in which an attacker cannot recognize the record owner within a set of other records, which is called the anonymity set [29]. By applying some anonymization techniques on the published dataset, the anonymity set can be created to protect the dataset from making such a link to identify a person. Therefore, the anonymous dataset will be protected from privacy attacks, and it will ensure the identity confidentiality in the published dataset.

2.2. Related Work

Numerous privacy-preserving data publishing methods have been reported in the last few decades, based on partitioning and randomization. In partitioning methods, the data values of quasi-identifiers (e.g., age, sex and zipcode) are generalized to create an equivalence class, so that individuals cannot be identified with their sensitive values in the equivalence class. By contrast, in randomization methods, the original values are changed by adding noise to make it difficult to pinpoint an individual in a published dataset. Some popular anonymization techniques such as k-anonymity [6], l-diversity [7], t-closeness [8], JS-reduce [30] have been developed for privacy preservation in one-time data publishing. Among these methods, JS-reduce [30] is the only method which models sequential background knowledge attack. The proposed method provides a better model which considers sequential background knowledge attack and anonymizes data, which gives better privacy protection to the individual.

For the multiple sensitive values in the dataset, a multidimensional framework was proposed for the privacy-preserving data publishing [31]. The multidimensional framework partitions the QI attributes and sensitive attributes in the bucket and ensures the privacy by satisfying l-diversity and k-anonymity privacy requirements. For the multiple numerical sensitive values, the multi sensitive bucketiazation method was proposed based on cluster technology for privacy-preserving data publishing [32]. In addition, several recent approaches [33–35] have been proposed to anonymize and publish a dataset while preserving more data utility. However, these methods are vulnerable to composition attack.

Several privacy-preserving data publishing methods have been designed to take into account known releases of related datasets, such as earlier publications by the same data owner (called sequential, serial or incremental releases) [13,14] and multiple views of the same dataset [11]. These methods explicitly deal with composition attacks which, by applying the intersection of two or more published datasets, reveal the sensitive information of any individual. Generalization [36] is a popular method that has been widely used for data anonymization. While generalization may protect personal data privacy, it results in serious data utility issues during data publication.

In the last decade, *Hybrid* [2], *Probabilistic* [23] and *Composition* [24] privacy methods have been proposed for multiple independent data publishing. Composition is the first privacy model to prevent against composition attacks in the multiple independent data publishing context. The proposed method in [23] uses sampling and generalization for independent datasets to protect composition attacks. The probabilistic approach tries to establish the linkability of sensitive values shared between multiple independent datasets by exploring correlation the *QI* attributes and sensitive attributes to simulate the anonymized data from another organization. The method in [24] works on the top of *k-anonymization* [6] privacy model to protect datasets from composition attacks. Hybrid method combines random sampling, perturbation, and generalization to protect the independent datasets from composition attacks. Hybrid provides better privacy protection and data utility than the Composition and Probabilistic methods.

Recently, $\epsilon - differential\, privacy\, (\epsilon - DP)$ [37] has received substantial attention for privacy-preserving data publishing. In differential privacy, $\epsilon - DP$ provides a strong privacy guarantee for statistical query answering. A composition attack can be protected by differential privacy based data anonymization [18]. A survey on differential privacy can be found in [38]. An anonymization mechanism on a dataset satisfies $\epsilon - DP$ privacy if the deletion or insertion of a single record from the dataset has only a small effect on the output of the randomization technique. In $\epsilon - DP$ privacy, if the independent datasets privacy are preserved by the privacy budget ϵ, then the smaller ϵ value provides the higher privacy protection. It is observed in [2,39,40] that using $\epsilon - DP$ to protect from composition attacks generates a significant amount of loss of data utility during anonymization. Most of the differential privacy methods support interactive settings to satisfy the $\epsilon - DP$ requirements. *Mohammed* [41] proposed the first non-interactive based approach for differentially private data release that protects information for published datasets.

2.3. Background Knowledge

Background knowledge can be described as an experience that has already been learned formally from previous rules regulated in the microdata of different data publishers or gleaned informally from life experience. For example, some of the sensitive attribute values (breast cancer, ovarian cancer) are linked with females only. Background knowledge helps the adversary to learn relevant sensitive information from the published microdata tables. Background knowledge helps in finding records and breaching individual privacy in published microdata tables.

An adversary may have formal and informal knowledge about the published dataset. An adversary may know that a person visits hospital A and hospital B for medication for the same disease. Thus, even before the datasets from A and B are published, the adversary already has the person's QI values. The adversary can then use the QI values of the person, that is person P in the table T, and the following facts to breach the sensitive value S:

1. P is also in another table F
2. F^* and T^* are the published tables of F and T

From the facts mentioned earlier and also using some absolute facts, the adversary can determine the individual's sensitive value.

In this paper, we assume that data publisher A has no knowledge about data publisher B's datasets with respect to concurrent tuples that A's dataset may contain. However, publisher A assumes that the QI values and sensitive values S of B's dataset follow the identical distributions as A's dataset. In addition, B's dataset is anonymized by the same data anonymization techniques as those used by data publisher A.

3. Preliminaries and Problem Definition

Given a microdata table T of records with d+1 attributes, $AT = \{AT_1, AT_2, ..., AT_d, S\}$ and the attribute domains are $\{D[AT_1], D[AT_2], ..., D[AT_d], D[S]\}$. A tuple $t \in T$ can be expressed as $t = (t[AT_1], t[AT_2], ..., t[AT_d], t[S])$, where $t[AT_i] (1 \leq i \leq d)$ is the quasi-identifier of t and $t[S]$ is the sensitive value of t.

We assume that a number of other microdata tables $F = \{F_1, F_2, ..., F_n\}$ are published by different independent publishers whose microdata tables are also defined by the same schema as T and published using the same anonymization method. This is a reasonable assumption since public and private groups may have a standard method for data distribution. For example, all health care institutions in the USA follow one law, the HIPAA, for their distribution of *d-identified* data [2].

3.1. Equivalence Class and Match

In a published *l-diversity* [25] dataset, an equivalence class or bucket consists of l distinct sensitive values from a particular sensitive value domain S and highly similar QI attribute values. In an equivalence class, any individual is linked with l distinct sensitive values so that the adversary cannot learn the sensitive values of the individual with a probability greater than $1/l$.

Let E_{T^*} and E_{F^*} be two equivalence classes in the published datasets T^* and F^*, respectively. These two equivalence classes match if their attribute value pairs $QI(E_{T^*})$ and $QI(E_{F^*})$ are equal or have a non-empty intersection, and sensitive values $S(E_{T^*}) \cap S(E_{F^*}) \neq \emptyset$. For example, E_{T^*} and E_{F^*} are two equivalence classes if $QI(E_{T^*})$ = (Age = 22, Sex = *, Zipcode = 47905) and $QI(E_{F^*})$ = (Age = 22, Sex = *, Zipcode = 47905). Suppose that the sensitive value domains are $S(E_{T^*})$ = (Ovarian Cancer, Gastritis, Flu, Breast Cancer) and $S(E_{F^*})$ = (Blood Cancer, Gastritis, Breast Cancer, Flu). The QI values of the two equivalence classes are 22, Male, 47905 or 22, Female, 47905. Therefore, $QI(E_{T^*}) \cap QI(E_{F^*}) \neq \emptyset$. In addition, $S(E_{T^*}) \cap S(E_{F^*})$ = (Gastritis, Breast Cancer, Flu)$\neq \emptyset$, and therefore, the equivalence classes E_{T^*} and E_{F^*} are matched.

Let t be the tuples of a user in T and F, s the possible sensitive value for t. The probability of a true match is defined as $P_T = P(t \in E_{T^*} \wedge t \in E_{F^*})$. In other words, the probability of a true match is the probability that the records of t are in both equivalence classes.

A match can be generated by two independent persons: this is called a false match. Even If the user is not in two published datasets, there is still a probability that two records match. This is the probability of a match in F^* and T^* that is generated by the uncertainty of two independent individuals, denoted by P_F. For example, in the published dataset, two persons may have the same age, live in the same zip code area and even have suffered from the same disease, without having visited the same hospital for medication.

3.2. Composition Attack and Privacy Breach

Given two independently published tables, T^* and F^*, a composition attack consists in examining the intersection of the two tables to find the common QI values and the corresponding sensitive value s of a person. In the composition attack, T^* and F^* are from two different data publishers, and there is no information shared prior to data publishing. If there is only one common sensitive value shared by two equivalence classes of s in T^* and F^*, the privacy of the person is breached with 100% likelihood.

From the background knowledge, it is certain that the adversary already knows the QI values because the adversary has the published tables T^* and F^*. In addition, the adversary may gather knowledge about a particular individual from publicly available data sources such as voter registration lists. By using the QI values from the voter registration lists, the adversary may try to find the QI values from the intersection of tables T^* and F^* and finally uncover the sensitive values.

In order to protect the privacy of the individual, we break the association between QI values by segmenting the microdata table respectively into columns and rows. Breaking the association between the QI values will confuse an adversary looking for exact QI values from the intersection of published tables T^* and F^*.

4. Merging Method

In this section, we formulate solutions from the problem definitions and design anonymization algorithms. The objective of anonymization is to obtain *l-diversity* in possible intersections of independently published datasets and to enable higher probabilities of false matches than of true matches, i.e., $P_F > P_T$.

4.1. Formulation

In this subsection, we sketch the ideas behind our algorithm development to meet the anonymization objectives. The probability of false matches will be increased by grouping the attributes in the equivalence class and permutation among the attribute values to break the correlations between attribute values. Specifically, the anonymization algorithm consists of the following steps: creation of fake tuples, attribute separation, tuple separation, and cell generalization.

4.1.1. Creation of Fake Tuples

A published table with generalization has less data utility than the published table with fake tuples. The presence of fake tuples does not have any effect on the utility of the published dataset, but will increase the probability of false matches during a composition attack on the published table [25]. In our anonymization method, we create *n (n=1,2,3,...)* fake tuples with the same *QI* values as in the original table and assigned sensitive values to them based on the sensitive value distribution in the initial dataset.

4.1.2. Attribute Separation

For attribute separation, related attributes are arranged in a subset, such that each attribute belongs to one subset. Each grouped subset is called a column. Specifically, in a microdata table T there will be c columns $C_1, C_2, ..., C_c$ satisfying $\bigcup_{i=1} C_i = AT$ and for any $1 \leq i_1 \neq i_2 \leq c$, $C_{i_1} \cap C_{i_2} = \emptyset$. The sensitive attribute S forms the last column C_c, called the sensitive column. The remaining columns $\{C_1, C_2, ..., C_{c-1}\}$ contain *QI* attributes.

Related attributes are grouped by measuring the association between attributes. The widely used method to measure the association between categorical attributes is the mean square contingency coefficient [25,42]. Given two categorical attributes AT_1 and AT_2 with value domains $\{v_{11}, v_{12}, ..., v_{1d_1}\}$ and $\{v_{21}, v_{22}, ..., v_{2d_2}\}$ respectively and domain sizes d_1 and d_2, the mean square contingency coefficient between attributes AT_1 and AT_2 is

$$\phi^2(AT_1, AT_2) = \frac{1}{min\{d_1, d_2\} - 1} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}}$$

where, $f_{i\cdot}$ and $f_{\cdot j}$ are the fractions of occurrence of v_{1i} and v_{2j} in the data, respectively. f_{ij} is the fraction of cooccurrence of v_{1i} and v_{2j} in the data. Therefore, $f_{i\cdot}$ and $f_{\cdot j}$ are the marginal totals of $f_{ij} : f_{i\cdot} = \sum_{j=1}^{d_2} f_{ij}$ and $f_{\cdot j} = \sum_{i=1}^{d_1} f_{ij}$. It can be shown that $0 \leq \phi^2(AT_1, AT_2) \leq 1$.

For continuous attributes, we have applied discretization to partition the domain of continuous attributes. Equal-width discretization is used for partitioning the domain into some k equal-sized intervals.

After computation of association between attributes, the widely employed *k-medoid* clustering algorithm Partition Around Medoids (PAM) [43] is used for partitioning attributes into the columns. In the anonymization algorithm, each attribute is determined as a point in the cluster space. The dissimilarity of two attributes is defined as $d(AT_1, AT_2) = 1 - \phi^2(AT_1, AT_2)$, which lies in 0 and 1. In the cluster space, two correlated attributes have a lower dissimilarity between the corresponding data points.

We arrange attributes to bring the highly correlated attributes in the same column. For the data utility and privacy, this approach performs well. In terms of data utility, grouping highly correlated attributes protect the relationships among those attributes. With regard to privacy, the association between correlated attributes reduce the identification risk. Since the association between uncorrelated attribute values are not common among the dataset and hence more distinctive, it is suitable to break the relation between uncorrelated attributes to preserve the user privacy. That way, when an adversary examines the intersection of two or more published datasets, the resulting intersection data will lead to more false matches, reducing the adversary's confidence in breaching data privacy.

4.1.3. Tuple Separation

This entails generating different subsets of T, such that each tuple is assigned to one subset. Every subset of tuples is called an equivalence class. Specifically, let there be e equivalence classes $E_1, E_2, ..., E_e$ such that $\bigcup_{i=1}^{e} E_i = T$ and for any $1 \le i_1 \neq i_2 \le e$; $E_{i_1} \bigcap E_{i_2} = \emptyset$.

The tuple separation operation divides the records horizontally into a number of partitions, called buckets or equivalence classes. The Mondrian [44] algorithm is applied, and it follows the top-down approach without generalization feature to separate tuples in the equivalence classes. Within each equivalence class, the values in each column are randomly permuted to break the cross-column associations. Therefore, the tuple separation will minimize the linkage between the sensitive values with the QI values which will reduce the adversary's confidence in linking with the sensitive values.

4.1.4. Cell Generalization

A cell is the cross-section of a column and a row. In our problem definition, a cell consists of a column and an equivalence class. Given a microdata table T, a column C_c and an equivalence class E_e make a cell $CE_{(i,j)}$, where $(1 \le i \le c)$ and $(1 \le j \le e)$. For example, in Table 3 above, the column {(Age, Sex)} and the first equivalence class which consists of tuples $\{t_1, t_2, t_3, t_4\}$ form the first cell of the table. A cell generalization for $CE_{(i,j)}$ generalizes each attribute value of $CE_{(i,j)}$ to satisfy privacy requirements.

In the anonymization, the attribute values in the equivalence class will be shuffled randomly to break the association of each tuple, in order to increase personal privacy. Random shuffling will break the association of the tuple but it will create some invalid records and, in some cases, it may increase the likelihood of privacy breach for a particular set of sensitive values [22]. For these special circumstances, we have introduced cell generalization to enhance the privacy of the equivalence class. Because cell generalization does not generalize the whole equivalence class, it allows better data utility than column generalization or full generalization of the table.

Cell generalization satisfies the *k-anonymity* and *l-diversity* privacy requirements for the particular cell and increases the probability of false matches of the attributes. In addition, cell generalization helps to reduce the curse of dimensionality associated with full generalization of the microdata table, and it certainly increases the data utility of the published dataset.

4.2. Algorithms

In this subsection, we present our merging algorithm for protecting published tables from composition attack. Two algorithms are introduced to perform the anonymization process. The primary objective is to increase the number of false matches by satisfying the *l-diversity* [25] requirement, thereby protecting the published table from composition attack and increasing data utility.

4.2.1. Anonymization Algorithm

Our anonymization algorithm (Algorithm 1) performs the anonymization process as follows: It creates n fake tuples at line 2 and adds them to the original microdata table T. It maintains two data structures: a queue of equivalence classes Q and a set of anonymized equivalence classes ET. Initially, Q contains only one equivalence class, and ET is empty. In each iteration (lines 4 to 10), the anonymization algorithm removes an equivalence class from Q and breaks the equivalence class into two equivalence classes according to the Mondrian [44] criterion. In line 7, privacy is checked by the Privacy-Check algorithm and the two equivalence classes are appended at the end of the queue Q (for more breaks of the equivalence class this is in line 8). If the equivalence class can no longer be broken, then the anonymization algorithm puts the equivalence class into ET in line 9. Finally, in line 12, the anonymized table T^* is published.

Algorithm 1 Anonymization

Input: Microdata Table T
Output: Anonymized Table T^*

For a given table T, generate an anonymized table T^*, satisfying privacy requirement R of l-diversity;
Generate datasets X with n fake tuples and combine with the dataset $T = T \cup X$;
$Q = T$; $ET = \emptyset$;
while Q is not empty **do**
    remove first equivalence class E from Q; $Q = Q - E$;
    separate E into two equivalence classes E_1 and E_2 as in Mondrian [44].
    if Privacy Checking $(T, Q \cup \{E_1, E_2\} \cup ET, R)$ **then**
        $Q = Q \cup \{E_1, E_2\}$;
    else
        $ET = ET \cup E$;
$T^* = ET$;
return T^*.

4.2.2. Algorithm for Privacy Checking

Our privacy checking algorithm assures the privacy requirement R in each equivalence class. In the anonymization, column values are permuted randomly to break cross-column associations. There is a possibility of creating some invalid records [22] or incompatible tuples in the process. In line 2, tuple incompatibility is checked as in [22]. If there are incompatible tuples, we generalize the particular cell values to satisfy *k-anonymity*. In line 5 we check the *l-diversity* privacy requirement as in slicing [25].

Algorithm 2 Privacy checking(T, T^*, R)

Input: Microdata Table T
Output: TRUE, if the equivalence class satisfies privacy requirement R

1: **for** each equivalence class E in T^* **do**
2:     check-out the tuple compatibility as in [22] incompatible tuple chek;
3:     **if** Incompatible tuple exists **then**
4:         generalize the cell values to satisfy *k-anonymity*;
5:     ensure the *l-diversity* of all equivalence classes to satisfy privacy requirement R as executed in [25];
6: **return** TRUE.

4.2.3. Anonymization Algorithm Time Complexity

To compute the time complexity of the anonymization algorithm, we need to consider the time complexity of Mondrian method [44], incompatible tuple check [22] technique, and the *l-diversity* verification technique. In the Mondrian algorithm, it requires $O(nlogn)$ times because the whole dataset must be scanned $O(n)$ times and the Mondrian algorithm needs n heights of the taxonomy tree, which is $O(logn)$. To check incompatible tuples it requires $O(n)$, and verifying the *l-diversity* requirement takes $O(n^2)$. Therefore, the whole time complexity of the anonymization algorithm is presented by $O(n^2logn)$.

4.3. Discussion on the Merging Method

In this subsection, we illustrate how the Merging method is able to protect the anonymization table from composition attack while increasing the data utility and maintaining l-diversity in the intersection of two published tables. Tables 6 and 7 are the published tables from Hospitals A and B.

Consider a tuple t with QI values (22, M, 47905), where tuple t visits both hospitals for medication. We can create a search for the QI values of tuple t in the intersection shown in Table 8. To determine the *QI* and sensitive values of *t*, its matching equivalence class is examined. In the first column, the attributes (Age, Sex) have the values (22,*), and in the second column, Zipcode has the value 47905 in the first equivalence class. Therefore, we can say the person may exist in the first equivalence class. Because the Sex attribute value is generalized, there is a possibility that the individual could be female even though the attribute Sex is considered to be male; the *QI* values (22, M, 47905) are linked to three sensitive values. Based on negative association rules [22], a male cannot suffer from breast cancer, so we deduct one more value and there are two values, satisfying *l-diversity* [25].

5. Experimental Analysis

In this section, we present experiments on real-world datasets. The experiments are divided into two parts: the first part was designed to test the effectiveness of the proposed anonymization algorithm against composition attack, in comparison with the $\epsilon - DP$ [41], Hybrid [2], Probabilistic [23], Composition [24] and Mondrian [44] methods. Our experimental results show that the Merging method also provides smaller privacy risks for the composition attacks. The results of this experiment are presented in *Composition Attack* subsection.

In the second part, we evaluated the effectiveness of our Merging method in preserving data utility, as compared to the same set of competing methods. The experiment demonstrates that the Merging method preserves more data utility than the other methods. In addition, it has smaller relative query error and better classification accuracy than the competing methods. The results of this experiment are presented in *Data Utility* section.

5.1. Data Set

The US-Census Adult dataset is derived from the UC Irvine Machine Learning Repository [45], which is composed of data accumulated from the US census. Data sets are described in Table 9. In our experiments, we extracted two independent datasets from the Adult dataset (i) Occupation and (ii) Education. The Adult dataset has 48842 tuples with six *QI* attribute values: Age, Sex, Marital status, Work class, Relationship, and Salary. Occupation is used as the sensitive attribute value for the Occupation dataset, and Education for the Education dataset.

Table 9. Description of US-Census Adult dataset.

	Attribute	Type	Number of Values
1	Age	Continuous	74
2	Sex	Categorical	2
3	Marital status	Categorical	7
4	Work class	Categorical	8
5	Relationship	Categorical	6
6	Salary	Categorical	2
7	Occupation	Categorical	14
8	Education	Categorical	16

For the experiment, we needed independent datasets to simulate the independent data publishing environment. Therefore, 10 disjoint datasets were composed from each of the Education and Occupation datasets, each with 4 K randomly selected tuples. The remaining 8 K tuples were used to generate the overlapping tuples pool to check for composition attack. From the remaining tuples pool, we made five copies of each group, respectively inserting 100, 200, 300, 400, and 500 tuples into the Education and Occupation datasets. Finally, we obtained datasets with sizes of (4.1 K, 4.1 K), (4.2 K, 4.2 K), (4.3 K, 4.3 K), (4.4 K, 4.4 K) and (4.5 K, 4.5 K) for Education and Occupation datasets respectively.

In the experiment, each group of datasets was used as input to the $\epsilon - DP$ [41], Hybrid [2], Probabilistic [23], Composition [24], Mondrian [44] and Merging algorithms to calculate the privacy risks and corresponding data utility. Privacy-preserving multiple independent data publishing is a non-interactive data publishing context, and for experimental analysis, we conduct the experiment on the non-interactive privacy settings. Conversely, the majority of the work in differential privacy [37] follows the interactive settings and a user accesses a dataset through a numerical query, while the anonymization technique appends noise to the query answers. The interactive environments might not always support the entire situation because in most cases datasets need to be published in public. Therefore, we select a non-interactive setting for differential privacy experiment as discussed in [41].

We compared the proposed Merging method with the $\epsilon - DP$ [41], Hybrid [2], Probabilistic [23], Composition [24] and Mondrian [44] methods. To compute privacy risks, we execute all algorithms on the non-interactive privacy-preserving data publishing environment. In the non-interactive privacy settings, $\epsilon - DP$, Hybrid, Probabilistic, Composition and Mondrian create *quasi-identifier* equivalence class as *k-anonymity* method [6]. For creating equivalence class, we select k = 4 and k = 6. In an equivalence class for the differential privacy, Laplacian noise is appended to the count of sensitive values [23]. To create the equivalence class for Merging method *l-diversity* [7] is chosen as discussed earlier, and we select l = 4 and l = 6. We select $\epsilon = 0.3$ for the $\epsilon - differential\ privacy$ budget.

The Merging method creates n = 1 fake tuples. For this reason, the output size of the Merging method will be larger than the all other methods. In the experimental comparison, we have therefore calculated the percentage of the respective output for each group of datasets.

In our experiments, we used the anonymization algorithm to anonymize the datasets. Composition attack was performed on all pairs of datasets, and data utility was measured after anonymization of the datasets.

5.2. Composition Attack

We checked the effectiveness of an anonymization algorithm in reducing the privacy risk due to composition attack. Privacy risk is measured by the ratio of true matches to total matches and expressed as [2]:

$$Privacy\ risk = \frac{True\ Matches\ (P_T)}{Total\ Matches\ (P_T + P_F)} \times 100\%$$

Composition attacks were measured by calculating privacy risk for all pairs of the extracted dataset with identical overlapping records. In the Merging method, the false matches will be increased because of l distinct sensitive values linked with the QI values, and it will decrease the privacy risk.

Figures 1–4 present the experimental results on the Occupation and Education datasets, respectively. They illustrate the privacy risk resulting from different anonymization techniques. Privacy risk indicates how confidently an adversary can learn sensitive values of a user from the multiple independent datasets. $\epsilon - DP$ [41] provides the lowest privacy risk for composition attacks among all the compared methods. Privacy risk gets smaller by increasing the false matches in the published datasets. The breaking of cross-column relation increases the probability of false matches in the anonymized datasets by the Merging method. As reported by the privacy risk shown in the result, Merging yields a lower probability of inferring the user's private information than the Probabilistic [23], Composition [24] and Mondrian [44] methods. It has almost identical privacy risk to that of the Hybrid [2] method. Therefore, we can say Merging also reduces the probability of composition attack on published datasets.

According to the privacy risk shown in Figures 1–4, we see that the $\epsilon - DP$ method achieves the best result for composition attacks because in an equivalence class it appends Laplacian noise to the count of sensitive values. It thus has the highest probability of yielding false matches. However, it reduces data utility, as discussed in Section 5.3. Conversely, the Merging method preserves more data utility compared with all the other methods.

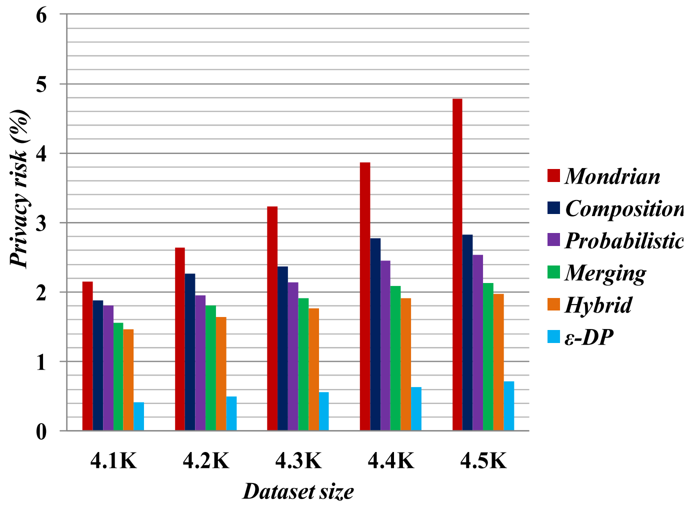

Figure 1. Privacy risk on Occupation dataset ($k = 4$, $l = 4$).

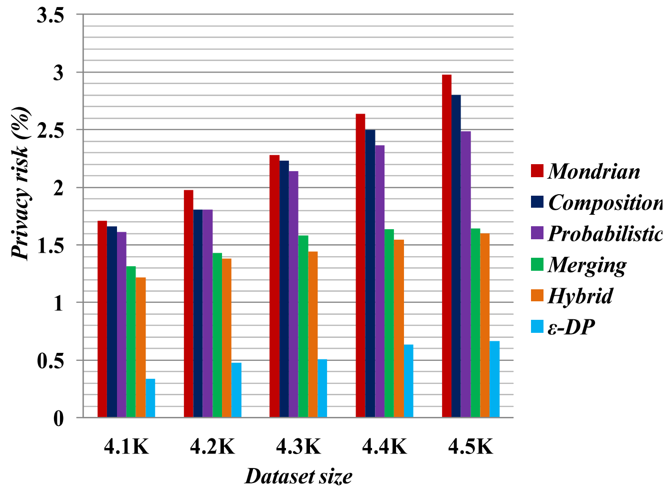

Figure 2. Privacy risk on Occupation dataset ($k = 6$, $l = 6$).

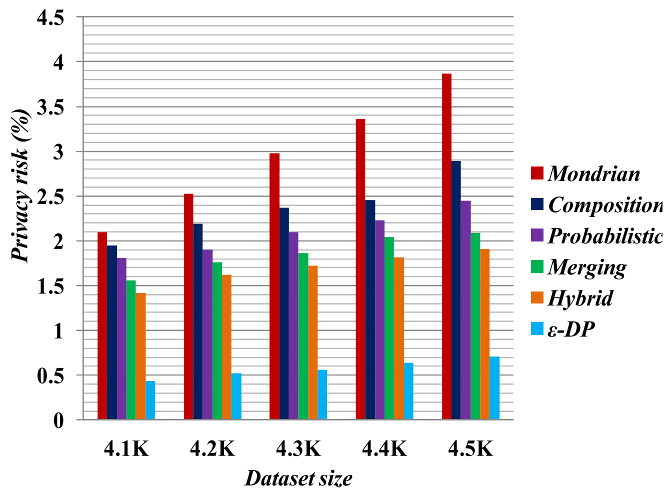

Figure 3. Privacy risk on Education dataset ($k = 4$, $l = 4$).

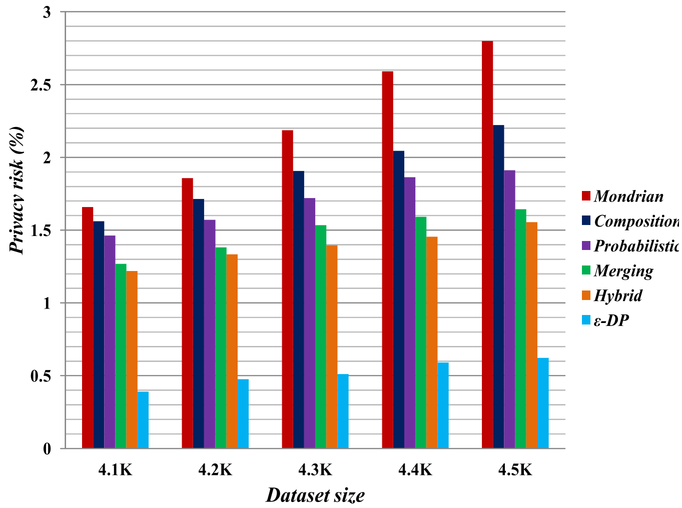

Figure 4. Privacy risk on Education dataset (k = 6, l = 6).

5.3. Data Utility

In the experiments in this subsection, we measured the data quality by the distortion ratio and the aggregate query answering error. In addition, we conducted an experiment on classifier learning of the published datasets. We observed that the Merging method has lower data loss, less relative error and better classification accuracy than the all other methods.

5.3.1. Data Utility Comparison

There are many methods [46] for calculating the information loss in the published data. Here we describe a simple method to illustrate the basic information loss metric.

We estimate that each attribute value of the microdata table is correlated with a generalized taxonomy tree. The cost is calculated from the published dataset and is called the distortion ratio. If the attribute value of a tuple is at the leaf node of the taxonomy tree, then the value is not generalized, and the distortion of that value is 0. Consequently, if the attribute value is generalized and does not represent the leaf node, then the distortion is defined by the position of the generalized attribute value and the height of the taxonomy tree. For instance, the age 22 is not generalized, and it stands at the leaf node; therefore, the height is 0, and the distortion is likewise equal to 0. While the attribute value is generalized one level up in the taxonomy tree, the distortion is equal to *1/H*. Here H represents the height of the taxonomy tree. Let $d_{j,k}$ is the distortion of the attribute A_j of tuple t_k. The distortion of the entire published microdata table is equal to the sum of the distortions of all values in the generalized dataset. In addition, the distortion is defined as discussed in [46]:

$$\sum_{j=1,k=1}^{n,m} d_{j,k}$$

The distortion ratio is $D_R = \frac{D_P}{D_G}$, where D_R is the distortion ratio, D_P is the distortion of the published table, and D_G is the distortion of the fully generalized (i.e., all attribute values are generalized by the root of the taxonomy trees) dataset.

Figure 5 illustrates the experimental result for data utility, based on data loss in the published datasets. For the data loss experiment, we selected a 4.5K dataset with *6-anonymity* for the Hybrid [2], $\epsilon - DP$ [41], Probabilistic [23], Composition [24] and Mondrian [44] methods, and *6-diversity* for the Merging method. The results show that the Merging method yields less data loss than all the other methods. We know the full generalization of the attribute values reduces the published data utility [25]. The Merging method employs selective generalization in the cell if it is necessary to meet the privacy requirements. Therefore, it preserves more data utility than the other methods.

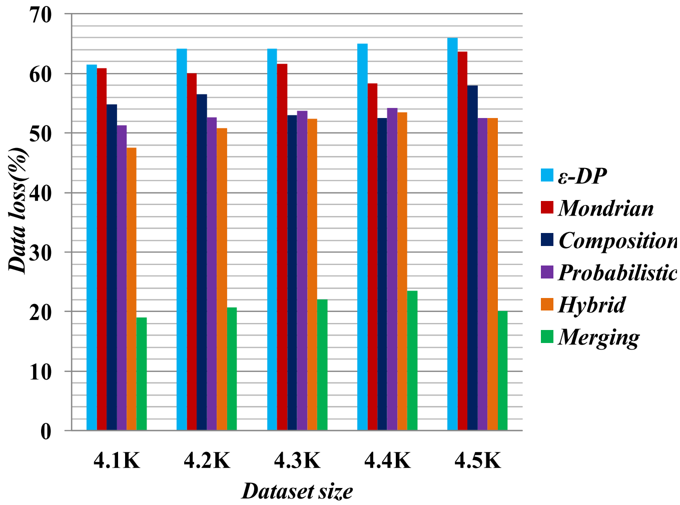

Figure 5. Data loss in the published dataset.

5.3.2. Aggregate Query Answering Error

In this experimental analysis, the accuracy of aggregate query answering [47] was also evaluated as a measure of data utility. It is possible to compute aggregate query operators such as "COUNT", "MAX", "AVERAGE" and so on. In the experiment, only the "COUNT" operator was evaluated, for queries whose predicates involved the sensitive values. The query is considered in the following form:

SELECT COUNT(*) FROM Table
WHERE $v_{i_1} \in V_{i_1}$ AND$\ldots v_{i_{dim}} \in V_{i_{dim}}$AND $s \in V_s$

where $v_{i_j}(1 \leq j \leq dim)$ is the quasi-identifier value for attribute AT_{i_j}, $V_{i_j} \subseteq D_{i_j}$ and D_{i_j} is the domain for attribute AT_{i_j}, s is the sensitive attribute value, $V_s \subseteq D_s$ and D_s is the domain for the sensitive attribute S. A query predicate is characterized by predicate dimension dim and query selectivity sel, dim indicating the number of quasi-identifiers in the predicate and sel indicating the number of values in each V_{i_j}, $(1 \leq j \leq dim)$. The size of V_{i_j}, $(1 \leq j \leq dim)$ was randomly chosen from $0, 1, \ldots, sel * |D_{i_j}|$. Each query was executed on seven tables: the original and those generated by the Merging, $\epsilon - DP$, Hybrid, Probabilistic, Composition and Mondrian methods. Count is indicated for the original and anonymized tables, the original count denoted by org_{count} and the anonymized count by anz_{count}, where anz_{count} is Merging, $\epsilon - DP$, Hybrid, Probabilistic, Composition and Mondrian respectively. To measure the average relative error in the anonymized dataset, we compute all queries as described in [47]:

$$Relative\ error = \frac{|anz_{count} - org_{count}|}{org_{count}} \times 100\%$$

In Figure 6, relative query error is plotted on the Y-axis based on the quasi-identifier selection. In the experiment, we selected one, two, three, four or five attributes as quasi-identifiers and calculated the relative query error on the anonymized tables generated by Merging, $\epsilon - DP$ [37], Hybrid [2], Probabilistic [23], Composition [24] and Mondrian [44] methods. For example, suppose we want to calculate the relative query error by Merging for Table 6, and the corresponding query is

SELECT COUNT(*) FROM Table 3
WHERE (sex='F') AND (Disease='Fever')

From the query answer, there is only one female person suffering from Fever. However, from the original table, i.e., Table 1, the query answer will be two persons. Using a relative error formula, it could be shown that Merging has a 50% relative query error for one attribute selection. For the experiment, all possible combinations of the query were generated and executed across anonymization tables for the 4.5 K Occupation dataset and the average relative query error was calculated, with k set to 6 for the Mondrian, Hybrid, $\epsilon - DP$, Probabilistic, Composition methods and l set to 6 for Merging. The relative query error was calculated and is shown in Figure 6, where the value on the *Y-axis* denotes relative error percentage and those on *X-axis* stand for different quasi-identifier selections. While the Merging method creates fake tuples in the anonymization, it can be seen from the experimental result that Merging still has small relative error compared with all other methods. For the generalization of attribute values, one needs to consider all possible combinations for a particular query answer. Therefore, the competing methods demonstrate the higher relative query error for the anonymized datasets.

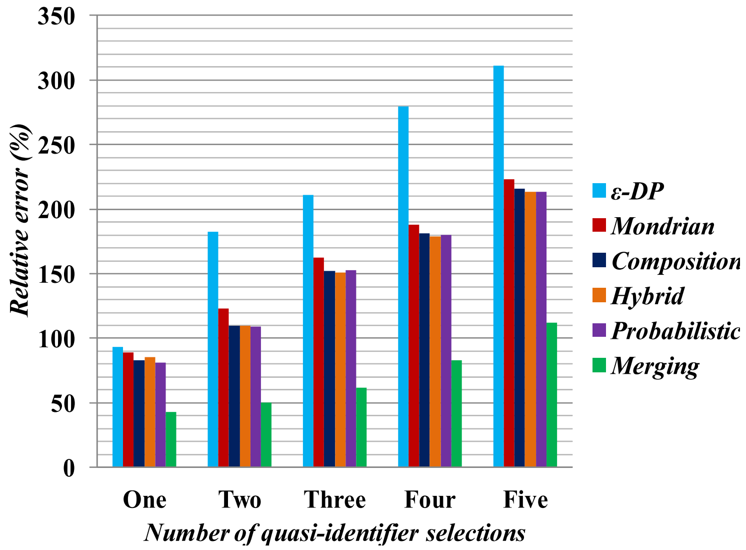

Figure 6. Aggregate query answering error.

5.3.3. Classifier Learning

Some preprocessing steps were applied on the anonymized data for classifier learning. The data anonymized by Merging contains multiple columns, and the linking between columns is broken. In Merging, attributes are partitioned into two or more columns. For an equivalence class that contains k tuples and c columns, the k tuples were generated as follows: first, randomly permute the values in each column; second, generate the $i^{th}(1 \leq i \leq k)$ tuple by linking the i^{th} value in each column. This procedure was applied to all equivalence classes and generated all of the tuples. The procedure generates the linking between the two columns in a random fashion.

We measure the quality of anonymized data for classifier learning, which has been used in [25]. The Weka [48] software package was used to evaluate the classification accuracy for C4.5 (J48). Default settings were used to do the classification task. In the experiment, we used 10-fold cross-validation. In each experiment, one attribute was chosen as the target attribute and the others served as predictor attributes. For performance evaluation, we selected 4,6 diversity of the Merging table and the 4,6-anonymized versions of the $\epsilon - DP$, Hybrid, Probabilistic, Composition and Mondrian tables. In the classifier learning, Education was chosen as the sensitive attribute and Relationship was chosen as the QI attribute.

In this experiment, we built two classifiers based on sensitive attribute Education and QI attribute Relationship. All other attributes are predictor attributes. Tables 10 and 11 present the classifier learning for sensitive attribute Education and QI attribute Relationship, respectively. Figure 7 compares the quality of the anonymized datasets with the comparison of original data when the target attribute is Education. Figure 8 presents the quality of the anonymized datasets with the comparison of original data when the target attribute is Relationship. Classifier learning indicates the quality of the dataset in terms of the attribute associations. We know that in the generalization, the attribute values have to consider all possible combinations of association. Therefore, it shows the lower performance of the classifier learning. In all experiments for classifier learning, the Merging method shows the better classification accuracy than the other methods because it only generalizes the required cell to satisfy the privacy requirement.

Table 10. Learning the sensitive attribute (Target: Education).

	$k = 4, l = 4$	$k = 6, l = 6$
Original Data	33	33
Merging	20.23	17.23
Mondrian	19.33	16.43
Composition	19.11	16.45
Probabilistic	18.66	15.59
Hybrid	18.43	15.23
$\epsilon - DP$	15.78	13.89

Table 11. Learning the *QI* attribute (Target: Relationship).

	$k = 4, l = 4$	$k = 6, l = 6$
Original Data	40	40
Merging	37.48	36.13
Mondrian	37.13	35.89
Composition	36.13	35.45
Probabilistic	37.22	35.87
Hybrid	36.76	35.75
$\epsilon - DP$	32.92	29.04

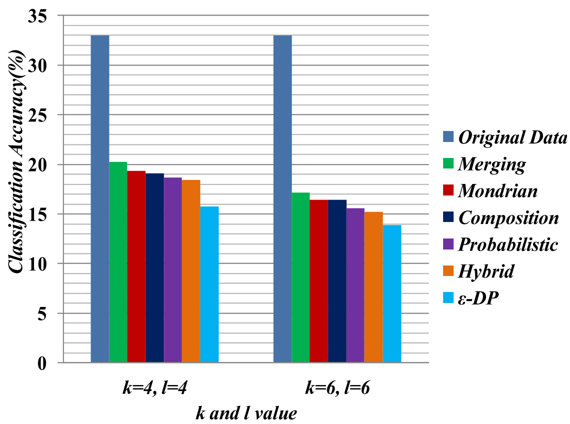

Figure 7. Learning the sensitive attribute (Target: Education).

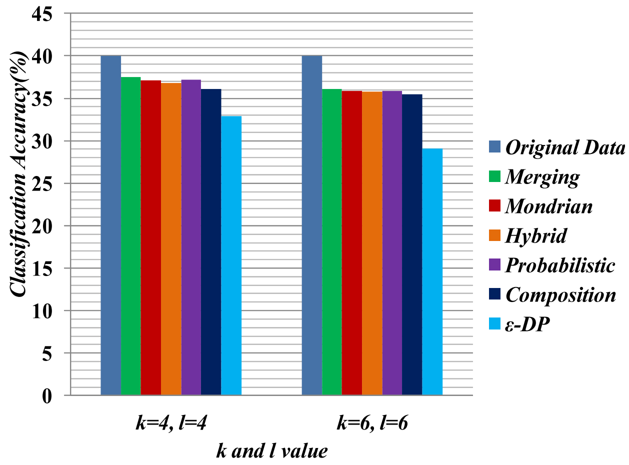

Figure 8. Learning the *QI* attribute (Target: Relationship).

5.4. Execution Time

We measure the scalability of the Merging method by evaluating the computation time to run the anonymization algorithm. To measure the computation time, we fix $l = 6$ and increase the dataset sizes for the execution time. Figure 9 presents the computation time as a function of the number of records. The results show that the Merging algorithm scales well with the data sizes.

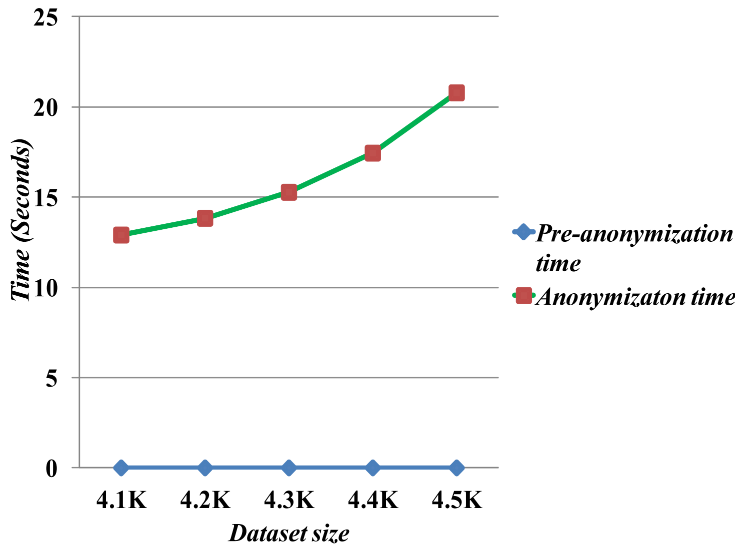

Figure 9. Execution time.

6. Conclusions

This paper presents an anonymization technique using cell generalization to limit the probability of a successful composition attack when independent data publishers are unable to coordinate before data publication. We experimentally demonstrated that the proposed Merging method satisfies the l-diversity privacy requirements after the intersection of independently published datasets. In contrast to most of the existing techniques, which reduce the data utility of the published datasets due to generalization and perturbation, our approach generalizes only the required cells, and thus results in less information loss and provides better data utility of the anonymized dataset. The experimental results illustrate that the Merging method offers higher data utility and smaller relative query error as compared to the state-of-the-art techniques.

Author Contributions: A.S.M.T.H. and S.W. conceptualized and composed the experiments and conducted the experiments; A.S.M.T.H., Q.J. and H.C. examined the data and contributed analysis tools; A.S.M.T.H., Q.J., H.C. and S.W. wrote the manuscript.

Acknowledgments: This research work was supported by Shenzhen Technology Development Grant No. CXZZ20150813155917544, Guangdong Province Research Grants No. 2015A080804019 and 2015A030310364; and sponsored by the CAS-TWAS President's Fellowship for International Ph.D. students.

Conflicts of Interest: The authors of the manuscript declare no conflict of interest.

References

1. Elliot, M.; Mackey, E.; O'Hara, K.; Tudor, C. *The Anonymisation Decision-Making Framework*; UK Anonymisation Network: Manchester, UK, 2016.
2. Li, J.; Baig, M.M.; Sattar, A.S.; Ding, X.; Liu, J.; Vincent, M. A hybrid approach to prevent composition attacks for independent data releases. *Inf. Sci.* **2016**, *367–368*, 324–336. [CrossRef]
3. Narayanan, A.; Shmatikov, V. Shmatikov how to break anonymity of the netflix prize dataset. *arXiv* **2006**, arXiv:cs/0610105.
4. Bee-Chung, C.; Daniel, K.; Kristen, L.; Ashwin, M. *Privacy-Preserving Data Publishing*; Now Publishers Inc.: Hanover, MA, USA, 2009; Volume 2, pp. 1–167.
5. Yamaoka, Y.; Itoh, K. k-presence-secrecy: Practical privacy model as extension of k-anonymity. *IEICE Trans. Inf. Syst.* **2017**, *100*, 730–740. [CrossRef]
6. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzz. Knowl. Based Syst.* **2002**, *10*, 557–570. [CrossRef]
7. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3. [CrossRef]
8. Li, N.; Li, T.; Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 23rd International Conference on Data Engineering ICDE, The Marmara Hotel, Istanbul, Turkey, 15–20 April 2007; pp. 106–115.
9. Sattar, A.S.; Li, J.; Ding, X.; Liu, J.; Vincent, M. A general framework for privacy preserving data publishing. *Knowl. Based Syst.* **2013**, *54*, 276–287. [CrossRef]
10. Yao, C.; Wang, X.S.; Jajodia, S. Checking for k-anonymity violation by views. In Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, 30 August–2 September 2005; pp. 910–921.
11. Yang, B.; Nakagawa, H.; Sato, I.; Sakuma, J. Collusion-resistant privacy-preserving data mining. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 483–492.
12. Wang, K.; Fung, B. Anonymizing sequential releases. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 414–423.
13. Wong, R.C.-W.; Fu, A.W.-C.; Liu, J.; Wang, K.; Xu, Y. Global privacy guarantee in serial data publishing. In Proceedings of the IEEE 26th International Conference on Data Engineering (ICDE), Long Beach, CA, USA, 1–6 March 2010; pp. 956–959.
14. Xiao, X.; Tao, Y. M-invariance: towards privacy preserving re-publication of dynamic datasets. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, 12–14 June 2007; pp. 689–700.
15. Hasan, A.S.M.T.; Jiang, Q. A general framework for privacy preserving sequential data publishing. In Proceedings of the 1st International Conference on Advanced Information Networking and Applications Workshops, Taipei, Taiwan, 27–29 March 2017; pp. 519–524.
16. Srisungsittisunti, B.; Natwichai, J. An incremental privacy-preservation algorithm for the (k, e)-anonymous model. *Comput. Electr. Eng.* **2015**, *41*, 126–141. [CrossRef]
17. Hasan, A.S.M.T.; Jiang, Q.; Li, C. An effective grouping method for privacy-preserving bike sharing data publishing. *Future Internet* **2017**, *9*, 65. [CrossRef]

18. Ganta, S.R.; Kasiviswanathan, S.P.; Smith, A. Composition attacks and auxiliary information in data privacy. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 265–273.
19. Jiang, W.; Clifton, C. A secure distributed framework for achieving k-anonymity. *Int. J. Very Large Data Bases* **2006**, *15*, 316–333. [CrossRef]
20. Jurczyk, P.; Xiong, L. Privacy-preserving data publishing for horizontally partitioned databases. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, CA, USA, 26–30 October 2008; pp. 1321–1322.
21. Mohammed, N.; Fung, B.; Wang, K.; Hung, P.C. Privacy-preserving data mashup. In Proceedings of the 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, 24–26 March 2009; pp. 228–239.
22. Hasan, T.; Jiang, Q.; Luo, J.; Li, C.; Chen, L. An effective value swapping method for privacy preserving data publishing. *Secur. Commun. Netw.* **2016**, *9*, 3219–3228. [CrossRef]
23. Sattar, A.S.; Li, J.; Liu, J.; Heatherly, R.; Malin, B. A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. *Knowl. Based Syst.* **2014**, *67*, 361–372. [CrossRef] [PubMed]
24. Baig, M.M.; Li, J.; Liu, J.; Ding, X.; Wang, H. Data privacy against composition attack. In Proceedings of the 17th International Conference Database Systems for Advanced Applications, Busan, Korea, 15–19 April 2012; pp. 320–334.
25. Li, T.; Li, N.; Zhang, J.; Molloy, I. Slicing: A new approach for privacy preserving data publishing. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 561–574 . [CrossRef]
26. Kambourakis, G. Anonymity and closely related terms in the cyberspace: An analysis by example. *J. Inf. Secur. Appl.* **2014**, *19*, 2–17. [CrossRef]
27. Pfitzmann, A.; Köhntopp, M. Anonymity, unobservability, and pseudonymity—A proposal for terminology. In *Designing Privacy Enhancing Technologies*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 1–9.
28. Pfitzmann, A.; Hansen, M. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. 2010. Available online: http://www.citeulike.org/user/isp/article/12731327 (accessed on 4 April 2018).
29. Hansen, M.; Smith, R.; Tschofenig, H. Ca privacy terminology and concepts. In *Internet Draft, March 2012*; Technical Report; Network Working Group, IETF: Fremont, CA, USA, 2011.
30. Thomas, C.; Thomas, D. An enhanced method for privacy preservation in data publishing. In Proceedings of the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 4–6 July 2013.
31. Luo, F.; Han, J.; Lu, J.; Peng, H. ANGELMS: A privacy preserving data publishing framework for microdata with multiple sensitive attributes. In Proceedings of the 2013 International Conference on Information Science and Technology (ICIST), Yangzhou, China, 23–25 March 2013; pp. 393-398.
32. Liu, Q.; Shen, H.; Sang, Y. Privacy-preserving data publishing for multiple numerical sensitive attributes. *Tsinghua Sci. Technol.* **2015**, *20*, 246–254.
33. Sánchez, D.; Domingo-Ferrer, J.; Martínez, S.; Soria-Comas, J. Utility-preserving differentially private data releases via individual ranking microaggregation. *Inf. Fusion* **2016**, *30*, 1–14. [CrossRef]
34. Hua, J.; Tang, A.; Fang, Y.; Shen, Z.; Zhong, S. Privacy-preserving utility verification of the data published by non-interactive differentially private mechanisms. *IEEE Trans. Inf. Forens. Secur.* **2016**, *11*, 2298–2311. [CrossRef]
35. Lee, H.; Kim, S.; Kim, J.W.; Chung, Y.D. Utility-preserving anonymization for health data publishing. *BMC Med. Inf. Decis. Mak.* **2017**, *17*, 104. [CrossRef] [PubMed]
36. Samarati, P.; Sweeney, L. Generalizing data to provide anonymity when disclosing information. *PODS* **1998**, *98*, 188. [CrossRef]
37. Dwork, C. Differential privacy. In *IN ICALP*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
38. Dwork, C. Differential privacy: A survey of results. In Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, Xi'an, China, 25–29 April 2008; Springer: Berlin/Heidelberg, Germany; pp. 1–19.
39. Cormode, G.; Procopiuc, C.M.; Shen, E.; Srivastava, D.; Yu, T. Empirical privacy and empirical utility of anonymized data. In Proceedings of the 29th IEEE International Conference on Data Engineering, ICDE, Brisbane, Australia, 8–12 April 2013; pp. 77–82.

40. Sarathy, R.; Muralidhar, K. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Priv.* **2011**, *4*, 1–17.
41. Mohammed, N.; Chen, R.; Fung, B.; Yu, P.S. Differentially private data release for data mining. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 493–501.
42. Cramér, H. *Mathematical Methods of Statistics (PMS-9)*; Princeton University Press: Princeton, NJ, USA, 2016; Volume 9.
43. Kaufman, L.; Rousseeuw, J. *Finding Groups in Data: An Introduction to Custer Analysis*; John Wiley & Sons: Princeton, NJ, USA, 2009; Volume 344.
44. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering, Atlanta, GA, USA, 3–8 April 2006; p. 25.
45. Lichman, M. UCI Machine Learning Repository. 2013. Available online: http://archive.ics.uci.edu/ml/datasets/Adult (accessed on 4 April 2018).
46. Wong, R.C.-W.; Fu, A.W.-C. Privacy-preserving data publishing: An overview. *Synth. Lect. Data Manag.* **2010**, *2*, 1–138. [CrossRef]
47. Zhang, Q.; Koudas, N.; Srivastava, D.; Yu, T. Aggregate query answering on anonymized tables. In Proceedings of the IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 116–125.
48. Hall, M.A.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *SIGKDD Explor.* **2009**, *11*, 10–18. [CrossRef]