ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Achieving Privacy Preservation Constraints based on K-Anonymity in conjunction with Adjacency Matrix and Weighted Graphs

Surapon Riyana[1], Kittikorn Sasujit[2] and Nigran Homdoung[3]

ABSTRACT

A well-known privacy preservation model is k-anonymity. It is simple and widely applied in several real-life systems. To achieve k-anonymity constraints in datasets, all explicit identifiers of users are removed. Furthermore, the unique quasi-identifiers of users are distorted by their less specific values to be at least k indistinguishable tuples. For this reason, after datasets are satisfied by k-anonymity constraints, they can guarantee that all possible query conditions to them always have at least k tuples that are satisfied. Aside from achieving privacy preservation constraints, the data utility and the complexity of data transformation are serious issues that must also be considered when datasets are released. Therefore, both privacy preservation models are proposed in this work. They are based on k-anonymity constraints in conjunction with the weighted graph of correlated distortion tuples and the adjacency matrix of tuple distances. The proposed models aim to preserve data privacy in datasets. Moreover, the data utility and data transform complexities are also considered in the privacy preservation constraint of the proposed models. Furthermore, we show that the proposed data transformation technique is more efficient and effective by using extensive experiments.

Article information:

Keywords: k-anonymity, Adjacency Matrix, Data Distortion, Privacy Preservation, Weighted Correlated Tuple Graph, Symmetric Matrix

Article history:
Received: July 13, 2023
Revised: November 23, 2023
Accepted: December 28, 2023
Published: January 20, 2024 (Online)

DOI: 10.37936/ecti-cit.2024181.253483

1. INTRODUCTION

Privacy violation is a serious issue that the data holder must consider when datasets are released to utilize in the outside scope of data-collecting organizations [1], i.e., the data holder must ensure that when the datasets are released, they must not have any concern of privacy violation issues. To achieve these aims in released datasets, k-anonymity is proposed. An example of privacy preservation is based on k-anonymity constraints. Let Table 1 be the original dataset such that Age and Gender are the quasi-identifier attributes. Another attribute, Disease, is the sensitive attribute. Suppose the value of k is 2. In this situation, a released data version of Table 1 is shown in Table 2. Although Table 2 is more secure in terms of privacy preservation than its original data version (Table 1), we can see that it loses some data meaning in terms of data utilization. In addition, k-anonymity is more complex in terms of data transformation. To rid these vulnerabilities of k-anonymity, both privacy preservation models are proposed in this work. They are based on k-anonymity constraints in conjunction with the weighted graph of correlated distortion tuples and the adjacency matrix of tuple distances.

Table 1: *An example of original dataset.*

#ID	Age	Gender	Disease
d_1	45	Male	Flu
d_2	46	Female	Fever
d_3	47	Male	Cancer
d_4	48	Male	HIV
d_5	48	Female	Flu
d_6	42	Female	HIV
d_7	42	Female	Fever

Table 2: *A 2-Anonymity data version of Table 1.*

#ID	Age	Gender	Disease
d_1	45 - 48	*	Flu
d_2	45 - 48	*	Fever
d_5	45 - 48	*	Flu
d_3	47 - 48	Male	Cancer
d_4	47 - 48	Male	HIV
d_6	42	Female	HIV
d_7	42	Female	Fever

[1,2,3] The authors are the School of Renewable Energy, Maejo University, Sansai, Chiangmai, Thailand, 50290, E-mail: surapon_r@mju.ac.th, k.sasujit@yahoo.com and nigranghd@gmail.com

The organization of this paper is as follows. In this section, privacy preservation issues and the traditional k-anonymity are presented. In the next section, we illustrate other existing privacy preservation models, data distortion techniques, and data transformation algorithms for transforming datasets to satisfy privacy preservation constraints. In Section 3, we propose the data distortion techniques for the numerical and non-numeric quasi-identifier, the distance of tuples, and both privacy preservation models that are based on k-anonymity constraints in conjunction with the weighted graph of correlated distortion tuples and the adjacency matrix of tuple distances. Then, the data utility metrics for evaluating the data utility of released datasets are presented (Section 4). The experimental results can indicate the effectiveness and efficiency of both proposed privacy preservation models. They are discussed in Section 5. Finally, the conclusion of this work will be discussed in Section 6.

2. BACKGROUND AND RELATED WORK

Data privacy, data utility, and the complexity of data transformation are serious issues that the data holder must consider when datasets are released to utilize in the outside scope of data-collecting organizations. For these reasons, several privacy preservation models have been proposed. Aside from privacy preservation models, we can also see data distortion techniques and data transformation algorithms are rapidly presented.

2.1 Privacy Preservation Model

Generally, privacy preservation models are the data framework such that they are used to address privacy violation issues when datasets are released to utilize in the outside scope of data-collecting organizations.

One of the most well-known privacy preservation models is k-anonymity [2]. It is a simple privacy preservation model. Moreover, it is widely applied in several real-life systems. For privacy preservation, all explicit identifier values of users are removed. Furthermore, the unique quasi-identifier values are distorted by their less specific values to be at least k indistinguishable tuples. Therefore, after datasets satisfy k-anonymity constraints, they can guarantee that all possible query conditions through the quasi-identifier attributes always have at least k tuples that are satisfied. However, in [3], the authors demonstrate that only removing the explicit identifier values and distorting the unique quasi-identifier values are not enough to address privacy violation issues in datasets because the sensitive data of users in datasets can still be violated by using identity and attribute linkage attacks.

To rid these vulnerabilities of k-anonymity, an extended k-anonymity model, l-diversity [3], is proposed. With l-diversity, aside from removing the explicit identifier values and distorting the unique quasi-identifier values, the number of distinct sensitive values is further considered in privacy preservation constraints such that every group of indistinguishable quasi-identifier values must relate to at least l different sensitive values. Therefore, after datasets are satisfied by l-diversity constraints, they can guarantee that all possible query conditions through the quasi-identifier attributes always have at least l distinct sensitive values that are satisfied.

Unfortunately, the datasets satisfy l-diversity constraints. They still have concerns about privacy violation issues from considering the distance of sensitive values. To rid this vulnerability of l-diversity, t-closeness [4] is proposed. With t-closeness, the parameter t enables one to trade-off between data utility and data privacy in datasets. We limit the gain from dis_1 to dis_2 by the limitations of the distance between both values as v_1 and v_2. Intuitively, if dis_1 equals dis_2, v_1 and v_2 are the same. If dis_1 and dis_2 are close, v_1 and v_2 are close. Also, when dis_1 and dis_2 are more different, it means that v_1 and v_2 are very different. For privacy preservation, the unique quasi-identifier values are distorted to be indistinguishable such that every group of distorted quasi-identifier values must relate to the set of the protected sensitive values that have the distance to be at least t. Therefore, after datasets satisfy t-closeness constraints, they can guarantee that the distance of re-identifying the protected sensitive values is at least t. Although datasets satisfy t-closeness constraints to be more secure in terms of privacy preservation than k-anonymity and l-diversity, they have data utility issues that must be addressed.

To address the data utility issues in high-dimension datasets, k^m-anonymity [5] is proposed. This privacy preservation model assumes that the adversary has the limitation of the background knowledge about the target user, i.e., the adversary has the background knowledge about the target user to be at most m values. Thus, only the m-size of unique quasi-identifier values are distorted to be at least k indistinguishable tuples. For this reason, after the datasets are satisfied by k^m-anonymity constraints, they can guarantee that every possible query condition through the m-size quasi-identifier attributes always has at least k tuples that are satisfied.

LKC-privacy [6] is a privacy preservation model. It is also proposed to address privacy violation issues in high-dimension datasets. For privacy preservation, all at most L sizes of the unique quasi-identifier values are distorted to be at least K indistinguishable tuples. Moreover, every protected sensitive value in every group of the distorted quasi-identifier values must have the confidence of data re-identification to be at most C.

Aside from the mentioned-above privacy preserva-

tion models, other well-known models are also proposed to address privacy violation issues in released datasets such as (k, e)-anonymous [7] and (α, k)-anonymity [8].

In [7], (k, e)-anonymous is proposed. For privacy preservation, before datasets are released, the tuples are firstly re-sorted by the sensitive values ascending. Then, the tuples are partitioned. That is, every partition must include at least k tuples. Moreover, every partition collects the sensitive values that have a different range between the lower and upper bounds to be at least e. Finally, the quasi-identifier tuples or the sensitive values of each partition are shuffled. For this reason, after datasets are satisfied by (k, e)-anonymous constraints, they can guarantee that the sensitive values have the confidence of the data re-identification to be at most e. Moreover, they can further guarantee that all possible queried results always have at least k tuples that are satisfied.

With (α, k)-anonymity [8], it is proposed to address privacy violation issues by using probability inference linkage attacks. With this privacy preservation model, α and k are both privacy preservation constraints. That is, before datasets are released, the unique quasi-identifier values are distorted to be at least k indistinguishable tuples. Moreover, every sensitive value must have the confidence of data re-identification from using probability inference linkage attacks to be at most α. Therefore, after datasets satisfy (α, k)-anonymity constraints, they can guarantee that the sensitive values cannot have any concern of privacy violation issues from using probability inference linkage attacks. Moreover, they can also guarantee that all possible queried results always have at least k tuples that are satisfied.

In brief, the privacy preservation model is the data model that is used to address privacy violation issues in datasets when they are released to the outside scope of data-collecting organizations.

2.2 Data Distortion

The data distortion technique is a data structure for distorting the unique values in released datasets to satisfy privacy preservation constraints. A few examples of well-known data distortion techniques are as follows.

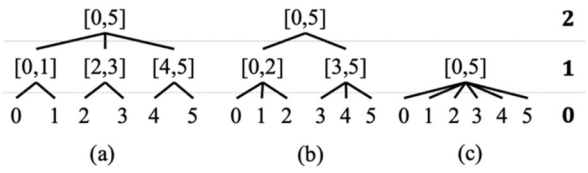

Fig.1: *A DGH of the ratting scores as [0, 5].*

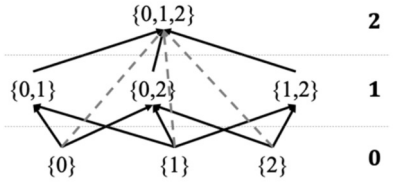

Fig.2: *A NDGH of the ratting scores as [0, 2].*

Domain Generalization Hierarchies (DGH) [2] is a well-known data structure that is proposed to describe the level of specificity quasi-identifier values that are available in datasets. They are based on tree data structures such that the values of the low level are more specific than the values that are available at the high level. Three examples are shown in Figure 3. They are proposed to present the level of data specifications for ratting scores as [0, 5].

Natural Domain Generalization Hierarchies (NDGH) [10] is also a well-known data structure that is proposed to describe the level of specificity quasi-identifier values that are available in datasets. With NDGH, it is based on directed graphs such that the label of each node presents a set of quasi-identifier values. Moreover, the label values of the low level are more specific than the label values of the high level. An example of NDGH is shown in Figure 2. It is proposed to present the level of data specifications for ratting scores as [0, 2].

Data swapping [7] is a data distortion technique that is widely acknowledged to apply in privacy preservation models. With this data distortion, the unique values in datasets are distorted by swapping their positions.

The range of data [10] is a data distortion technique that is often used to distort the unique numerical in datasets. That is, the uniquely numerical values do not have any concern of privacy violation issues when they are distorted by their range between the lower and upper bounds to be indistinguishable.

Another data distortion technique is often available in privacy preservation models. It is additive noises [11]. That is, the unique data is distorted by using some appropriate noises.

In brief, the data distortion technique is the data structure that is proposed to describe the level of the specificity values that are available in datasets.

2.3 Data Transformation

The data transformation is an algorithm. It is used for transforming datasets to satisfy privacy preservation constraints such as brute force [12] and clustering [13]. With brute force algorithms, they are a simple and classical idea for transforming datasets to satisfy privacy preservation constraints. Generally, they can guarantee that the released data version of datasets is optimal. However, they are highly complex. Thus,

they are unsuitable for addressing privacy violation issues in larger-size or high-dimension datasets. To rid this vulnerability of brute force algorithms, the privacy preservation models are based on clustering algorithms to be proposed. That is, before datasets are released to utilize in the outside scope of data-collecting organizations, the similar tuples are firstly clustered to be the groups. Finally, the unique values of each group are distorted by their less specific values to be indistinguishable. The most well-known privacy preservation model, k-member [13], is based on clustering algorithms in conjunction with k-anonymity constraints. For privacy preservation, the at least k size of tuple groups is constructed by considering their distance scores. Another well-known privacy preservation model is based on clustering algorithms. It is Mondrian Multidimensional k-anonymity [14]. With this privacy preservation model, the most similarity of tuples in each group is constructed by a strict multi-dimensional partitioning algorithm. However, we can see that these privacy preservation models still are complex. To rid this vulnerability, a privacy preservation technique is based on the weighted graph of correlated generalization tuples and the adjacency matrix of tuple distances to be proposed in this work.

In brief, data transformations are algorithms that can transform datasets to satisfy privacy preservation constraints. A desired data transformation algorithm is low complexity and can maintain the data utility of datasets as much as possible.

3. THE PROPOSED MODEL

In this section, we propose both privacy preservation models. They are based on k-anonymity in conjunction with a weighted graph of correlated generalization tuples and an adjacency matrix of tuple distances. Before the proposed technique will be presented. We first define the necessary basic definitions of this work.

Definition 1 (Dataset) Let $U = \{u_1, u_2, \ldots, u_n\}$ be the set of users. Let $D = \{d_1, d_2, \ldots, d_n\}$ be the original dataset that is constructed from n user tuples such that every $d_i \in D$ represents the profile tuple of the user $u_i \in U$, where $1 \leq i \leq n$. Moreover, every d_i constructs from m attributes as $A = \{a_1, a_2, \ldots, a_m\}$. That is, $QI = \{qi_1, qi_2, \ldots, qi_{m-1}\} \subset A$ is the set of quasi-identifier attributes. Another attribute $S \in A$ (i.e., $A - QI$) is the sensitive attribute. In addition, let $D[qi_y]$, where $1 \leq y \leq |QI|$, be the data projection of qi_y of D. Let $D[S]$ be the data projection of S of D. Let $D[QI]$ be the data projection of all quasi-identifier tuples that are available in D. Let $d_i[QI]$ be the data projection of the quasi-identifier tuple of d_i. Moreover, let $d_i[S]$ be the data projection of the sensitive value of d_i.

For example, let Table 1 be the original dataset D that has seven user profile tuples (i.e., $d_1, d_2, d_3, d_4, d_5, d_6$, and d_7) and three attributes, i.e., Age, Gender, and Disease. Let Age and Gender be the quasi-identifier attributes. The Disease is a sensitive attribute. Therefore, the $D[d_1]$ of Table 1 is shown in Table 3. The $D[S]$ of Table 1 is shown in Table 4. The $D[QI]$ of Table 1 is shown in Table 5. The $d_1[QI]$ of Table 1 is shown in Table 6. The $d_1[S]$ of Table 1 is shown in Table 7.

Table 3: *The data projection of $D[d_1]$ of Table 1.*

#ID	Age	Gender	Disease
d_1	45	Male	Flu

Table 4: *The data projection of D[S] of Table 1.*

#ID	Disease
d_1	Flu
d_2	Fever
d_3	Cancer
d_4	HIV
d_5	Flu
d_6	HIV
d_7	Fever

Table 5: *The data projection of D[QI] of Table 1.*

#ID	Age	Gender
d_1	45	Male
d_2	46	Female
d_3	47	Male
d_4	48	Male
d_5	48	Female
d_6	42	Female
d_7	42	Female

Table 6: *The data projection of $d_1[QI]$ of Table 1.*

#ID	Age	Gender
d_1	45	Male

Table 7: *The data projection of $d_1[S]$ of Table 1.*

#ID	Disease
d_1	Flu

Definition 2 (Privacy Violation Issues) Let D be the specified dataset. Let k be the minimum number of tuples that cannot have any concerns about privacy violation issues. The meaning of privacy violation issues for a tuple $d_i \in D$ is that $d_i[QI]$ is duplicated by other quasi-identifier tuples in $D[QI]$ to be at most $k-1$ tuples.

In addition, the data domain of $qi_y \in QI$ is generally numerical and non-numerical. The numerical data is the data that can be quantifiable, e.g., age and salary. With the non-numerical data, it is categorical data such as gender, position, and education.

For this reason, they can be distorted to satisfy privacy preservation constraints through a different data transformation technique.

3.1 Data Distortion Based on Numerical Data

Let $qi_y \in QI$ be a specified quasi-identifier attribute of D such that the data domain of qi_y is numerical. Given $\vartheta \subseteq qi_y$is the set of the particular quasi-identifier values from qi_y. With ϑ, it can be distorted to be indistinguishable from using their range between $f_{MIN}(\vartheta)$ and $f_{MAX}(\vartheta)$. More range between $f_{MIN}(\vartheta)$ and $f_{MAX}(\vartheta)$ is more secure in terms of privacy preservation. Therefore, the distorted value distance of ϑ can be defined by $f_{dis}(\vartheta) : f_{MAX}(\vartheta) - f_{MIN}(\vartheta)$.

For example, let the Age attribute of Table 1 be the specified quasi-identifier attribute. Let 45, 46, and 48 be the particular quasi-identifier values. $f_{MIN}(\{45, 46, 48\})$ is 45. $f_{MAX}(\{45, 46, 48\})$ is 48. Therefore, the distorted value of them is "45-48". The distorted value distance of them is 3, i.e., 48-45=3.

3.2 Data Distortion Based on Non-Numerical Data

This section is proposed to explain the technique that can be used for distorting non-numerical quasi-identifier values to be indistinguishable. Let $qi_y \in QI$ be a quasi-identifier attribute. Let $DO^{qi_y} = \{do_1^{qi_y}, do_2^{qi_y}, \ldots, do_z^{qi_y}\}$ be the data domain of qi_y. Let $f_DGH(DO_L^{qi_y} : DO_L^{qi_y} \rightarrow DO_{L+1}^{qi_y}$ be the distortion function of DO^{qi_y} from the level L to the level $L+1$ such that all values of the level L are more specific than all values that are available in the level $L+1$. From the function, the domain distortion hierarchy of qi_y, DGH_{qi_y}, can be defined from a distortion sequence as $DO_0^{qi_y} \xrightarrow{f_{DGH}(DO_0^{qi_y})} DO_1^{qi_y} \xrightarrow{f_{DGH}(DO_1^{qi_y})} DO_2^{qi_y} \ldots DO_{L-2}^{qi_y} \xrightarrow{f_{DGH}(DO_{L-2}^{qi_y})} DO_{L-1}^{qi_y} \xrightarrow{f_{DGH}(DO_{L-1}^{qi_y})} DO_L^{qi_y}$, where $DO_0^{qi_y} = DO^{qi_y}$. That is, the most specific values are available in the level 0 and the lowest specific values are available in the level L. Let $\vartheta \subseteq qi_y$ be the set of the particular quasi-identifier values in qi_y. Let ω be the distorted value of ϑ such that ω is available in the level Φ of DGH_{qi_y}. Thus, the distorted value distance of ϑ is Φ, i.e., $f_{dis}(\vartheta) : \Phi$.

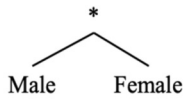

Fig.3: *DGH_{Gender}.*

For example, let the Gender attribute of Table 1 be the specified quasi-identifier attribute. Let Figure 3 be DGH_{Gender}. Moreover, suppose that the gender value is available in d_1 and d_2 to be the particular quasi-identifier values. Thus, the appropriately distorted value of the particular quasi-identifier values is *. The distorted value distance of the particular quasi-identifier values is 1 because * is available in the level 1 of DGH_{Gender}.

3.3 The Distance of Tuples

Let $D_{SEL} \subseteq D$ be the set of the specified tuples. Let $f_{dis}(D_{SEL}[qi_1]), f_{dis}(D_{SEL}[qi_2]), \ldots, f_{dis}(D_{SEL}[qi_{m-1}])$ be the functions that represent the distorted value distance of $D_{SEL}[qi_1], D_{SEL}[qi_2], \ldots, D_{SEL}[qi_{m-1}]$, respectively. Therefore, the distorted value distance of D_{SEL} can be defined by $f_{DIS}(D_{SEL}) : f_{dis}(D_{SEL}[qi_1]) + f_{dis}(D_SEL[qi_2]) + \ldots + f_{dis}(D_{SEL}[qi_{m-1}])$. More value of $f_{DIS}(D_{SEL})$ is more secure in terms of privacy preservation.

For example, let Table 1 be the original dataset D. Let d_1 and d_2 be both of the particular tuples in Table 1, i.e., $D_{SEL} = \{d_1, d_2\}$. Let Figure 3 be DGH_{Gender}. In this situation, a distorted data version of Table 1 is shown in Table 8. That is, the Age attribute, $D_{SEL}[Age]$, is distorted by the range of users' ages that are available in the Age attribute of d_1 and d_2, i.e., 45 - 46. Moreover, the unique gender values in the Gender attribute, $D_{SEL}[Gender]$, are distorted by an appropriate value, *, that is available in DGH_{Gender}. Therefore, Table 8 has the distorted value distance to be 2, i.e., $1 + 1 = 2$.

Table 8: *A distorted data version of d_1 and d_2 of Table.*

#ID	Age	Gender	Disease
d_1	45 - 46	*	Flu
d_2	45 - 46	*	Fever

3.4 Privacy preservation is Based on k-anonymity in conjunction with Weighted Graphs and Adjacency Matrix

This section is devoted to presenting both privacy preservation models. They are based on the weighted graph of correlated distortion tuples and the adjacency matrix of tuple distances to be proposed. The proposed technique aims to transform datasets to satisfy k-anonymity constraints.

Definition 3 (The Weighted Graph of Correlated Distortion Tuples) Let $G = (V, E)$ be a weighted graph that proposes to present the correlated distortion tuples of D. That is, V is the set of vertices such that every $v_i \in V$ represents the tuple d_i of D, while $E = \{\{v_\alpha, v_\beta\} | v_\alpha, v_\beta \in V$ and $v_\alpha \neq v_\beta\}$ is the set of edges that represent the relationship between the vertices v_α and v_β. Moreover, each edge of v_α and v_β has an associated numerical value (a weight) that represents $f_{dis}(v_\alpha, v_\beta)$.

Definition 4 (The Adjacency Matrix of Distorted Tuple Distances) Let $G = (V, E)$ be a weighted graph of correlated distortion tuples. An adjacency matrix of $G = (V, E)$ is $GM = [gm_\alpha, gm_\beta]_{n*n}$, where $0 \leq \alpha \leq n$ and $0 \leq \beta \leq n$, such that $GM = [gm_\alpha, gm_\beta]_{n*n}$ is in the form of $\begin{bmatrix} 0 & \Delta GM^T \\ \Delta GM & 0 \end{bmatrix}$, i.e., the (gm_α, gm_β)-position of GM collects $f_{dis}(gm_\alpha, gm_\beta)$, otherwise, it collects 0. Therefore,

$$GM[gm_\alpha, gm_\beta = \begin{bmatrix} gm_0, gm_\beta \\ gm_1, gm_\beta \\ \vdots \\ gm_n, gm_\beta \end{bmatrix}$$

Is equal to

$$\mathrm{GM}[\mathrm{gm}_\beta, gm_\alpha]^T = [gm_\beta, gm_0 \quad gm_\beta, gm_1 \quad \cdots \quad gm_\beta, gm_n].$$

Definition 5 (Equivalence Class) Let a positive integer k, where $k \in I^+$ and $k \geq 2$, be the privacy preservation constraint. Let $D_{SEL} \subseteq D$ be the set of the specified tuples such that the size of D_{SEL} is at least k, i.e., $|D_{SEL}| \geq k$. Let ec be an equivalence class that can be constructed from D_{SEL} by using $f_A(D_{SEL}) : D_{SEL} \rightarrow D'_{SEL}$, i.e., $ec = D'_{SEL}$. That is, the unique quasi-identifier values of D'_{SEL} are distorted by their less specific values that are presented in the form of the DGH or the value range of them.

Definition 6 (The Error of Equivalence Classes) Let ec be an equivalence class that is constructed from $D_{SEL} \subseteq D$. Thus, the error of ec can be defined by $f_err(ec) : f_{DIS}(D_{SEL}) * |D_{SEL}|$.

3.4.1 Privacy Preservation Based on The Adjacency Matrix of Distorted Tuple Distances

Let $\boldsymbol{D = \{d_1, d_2, \ldots, d_n\}}$ be the original dataset. Let $\boldsymbol{G = (V, E)}$ be a weighted graph of correlated distortion tuples of $\boldsymbol{D}$. Let $\boldsymbol{GM}$ be an adjacency matrix of $\boldsymbol{G = (V, E)}$. With $\boldsymbol{GM}$, we know that it is a symmetric matrix, i.e., $\boldsymbol{GM[gm_\alpha, gm_\beta] = GM[gm_\beta, gm_\alpha]^T}$. Thus, only the lower or upper triangular matrix of the adjacency matrix of $\boldsymbol{GM}$ is considered for constructing the desired released dataset $\boldsymbol{D'}$ of $\boldsymbol{D}$, i.e., $\boldsymbol{\Delta GM}$. Let $\boldsymbol{f_A(D, \Delta GM, k) : D \rightarrow_{\Delta GM,k} D'}$ be the function for transforming $\boldsymbol{D}$ to become $\boldsymbol{D'}$. That is, the users' unique quasi-identifier values in $\boldsymbol{D'}$ are distorted to be at least $\boldsymbol{k}$ indistinguishable quasi-identifier tuples by using their ranges between the lower and upper bounds or the less specific values that are available in the $\boldsymbol{DGH}$ of them. Moreover, the summation of every group of indistinguishable quasi-identifier tuples in $\boldsymbol{\Delta GM}$ must be minimized. In addition, aside from the list of tuples, $\boldsymbol{D'}$ further has another view that is in the form of the set of its equivalence classes, i.e., $\boldsymbol{EC = \{ec_1, ec_2, \ldots, ec_e\}}$. Without loss of generality, every equivalence class $\boldsymbol{ec_p}$, where $\boldsymbol{1 \leq p \leq e}$, of $\boldsymbol{D'}$ must satisfy the data limitations that are $\boldsymbol{\bigcup_{p=1}^{e} ec_p = D}$ and $\boldsymbol{\bigcap_{p=1}^{e} ec_p = \emptyset}$.

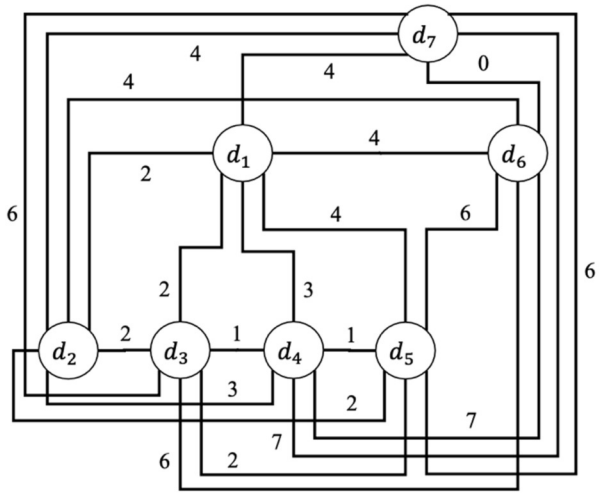

Fig.4: *A weighted graph of the correlated distortion tuples.*

For example, let Table 1 be the original dataset D. Moreover, the DGH of the Gender attribute is shown in Figure 3. A correlated distortion tuple graph of Table 1 is shown in Figure 4. Thus, the adjacency matrix of Table 1 is shown in Figure 5. While the lower triangular matrix of the adjacency matrix is shown in Figure 6 and the upper triangular matrix of the adjacency matrix is shown in Figure 7. In addition, in this example, only the lower triangular matrix of the adjacency matrix is used to consider constructing the appropriate released data version of Table 1. Suppose the value of k is 2. A released data version of Table 1 is shown in Table 9. Table 9 has three equivalence classes. The first equivalence class constructs from d_6 and d_7. The cause of d_6 and d_7 is available in the first equivalence class, it is that the distance between d_6 and d_7 is closer than other tuples, i.e., the distance between d_6 and d_7 is 0 (zero). In addition, after the first equivalence class is constructed successfully, d_6 and d_7 will not be considered. Currently, Table 1 only remains five tuples, i.e., d_1, d_2, d_3, d_4, and d_5. With the remaining tuples, we can see that (d_3, d_4)-position and (d_3, d_5)-position collect the same tuple distance and they are closest than other remaining tuples, i.e., they have the tuple distance to be 1. Moreover, (d_3, d_4)-position and (d_4, d_5)-position are associated through d_4, so, d_3, d_4, and d_5 are constructed to be the second equivalence class of the released dataset. In this situation, only d_1 and d_2 are not investigated. Moreover, they are fit for only constructing an equivalence class. Thus, the third equivalence class is constructed from two remaining tuples as d_1 and d_2. For this reason, the error

of Table 9 is 10, i.e., ((1+1)*2)+((1+1)*3)=4+6=10.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	0	2	2	3	4	4	4
d_2	2	0	2	3	2	4	4
d_3	2	2	0	1	2	6	6
d_4	3	3	1	0	1	7	7
d_5	4	2	2	1	0	6	6
d_6	4	4	6	7	6	0	0
d_7	4	4	6	7	6	0	0

Fig.5: *An adjacency matrix of Table1.*

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	0						
d_2	2	0					
d_3	2	2	0				
d_4	3	3	1	0			
d_5	4	2	2	1	0		
d_6	4	4	6	7	6	0	
d_7	4	4	6	7	6	0	0

Fig.6: *A lower triangular matrix of the adjacency matrix of Table 1.*

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	0	2	2	3	4	4	4
d_2		0	2	3	2	4	4
d_3			0	1	2	6	6
d_4				0	1	7	7
d_5					0	6	6
d_6						0	0
d_7							0

Fig.7: *A upper triangular matrix of the adjacency matrix of Table 1.*

Table 9: *A distorted data version of Table 1 satisfies $k = 2$ by using the proposed data transformation technique.*

#ID	Age	Gender	Disease
d_6	42	Female	HIV
d_7	42	Female	Fever
d_3	47 - 48	*	Cancer
d_4	47 - 48	*	HIV
d_5	47 - 48	*	Flu
d_1	45 - 46	*	Flu
d_2	45 - 46	*	Fever

Another example of privacy preservation is based on k-anonymity constraints in conjunction with the adjacency matrix of distorted tuple distances. Also, let Table 1 be the original dataset D and let Figure 3 be DGH_{Gender}. Let the value of k be 3. In this situation, the first equivalence class of the released dataset of Table 1 is constructed from the tuples d_3, d_4, and d_5 because the tuple distance of them is only 3, i.e., $2 + 1 = 3$. Another equivalence class of the dataset is constructed from the tuples d_1, d_2, d_6, and d_7. Therefore, Table 10 is a released data version of Table 1 such that it satisfies 3-anonymity constraints. With this released data version of Table 1, it has the error to be 26, i.e., ((1+1)*3)+((4+1)*4)=6+20=26.

Table 10: *A distorted data version of Table 1 satisfies $k = 2$ by using the proposed data transformation technique.*

#ID	Age	Gender	Disease
d_3	47 - 48	*	Cancer
d_4	47 - 48	*	HIV
d_5	47 - 48	*	Flu
d_1	42 - 46	*	Flu
d_2	42 - 46	*	Fever
d_6	42 - 46	*	HIV
d_7	42 - 46	*	Fever

Aside from the adjacency matrix of distorted tuple distances, the distorted data versions of D can also be constructed from the weighted graph of correlated distortion tuples of D directly.

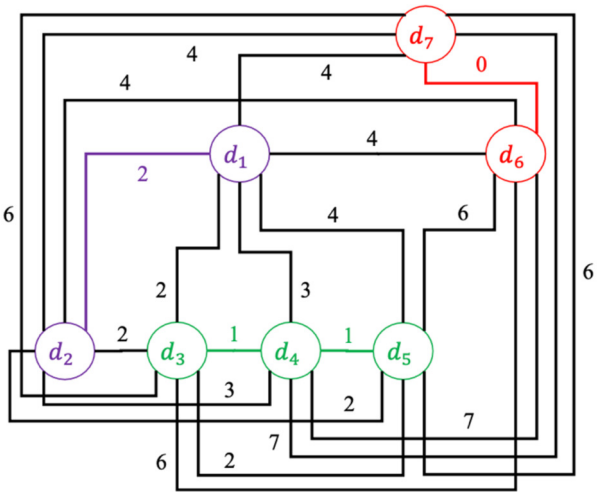

Fig.8: *A graphic info for demonstrating the relationship of the similar tuples for constructing the suitable equivalence classes of Table 1, where $k = 2$.*

Table 11: *A partitioned data version of Table 1.*

#ID	Age	Gender	Disease
d_6	42	Female	HIV
d_7	42	Female	Fever
d_3	47	Male	Cancer
d_4	48	Male	HIV
d_5	48	Female	Flu
d_1	45	Male	Flu
d_2	46	Female	Fever

3.4.2 Privacy Preservation Based on The Weighted Graph of Correlated Distortion Tuples

Let $G = (V, E)$ be a weighted graph of correlated distortion tuples of D. Let $G_{SUB} = \{G_1, G_2, \ldots, G_q\}$ be the set of disjoint sub-graphs of G. Let $E_r = \{\{v_\alpha, v_\beta\} | v_\alpha, v_\beta \in G_r \text{ and } v_\alpha \neq v_\beta\}$ be the set of edges in G_r, where $1 \leq r \leq q$. For privacy preservation, G_{SUB} is first satisfied the limitations as follows,

- $G_1 \cup G_2 \cup \ldots \cup G_q = V$
- $G_1 \cap G_2 \cap \ldots \cap G_q = \emptyset$
- $|G_r| \geq k$, where $1 \leq r \leq q$, and
- the summation of the weighted edges of $G_1, G_2, \ldots, G_q$ is minimized.

Then, the tuples of D are partitioned to correspond G_{SUB}. Finally, all users' unique quasi-identifier values in each partition are distorted by using their ranges between the lower and upper bounds or a less specific value that is available in the *DGH* of them.

Table 12: *The data version of Table 1 is after inserted d_8 .*

#ID	Age	Gender	Disease
d_1	45	Male	Flu
d_2	46	Female	Fever
d_3	47	Male	Cancer
d_4	48	Male	HIV
d_5	48	Female	Flu
d_6	42	Female	HIV
d_7	42	Female	Fever
d_8	45	Female	Cancer

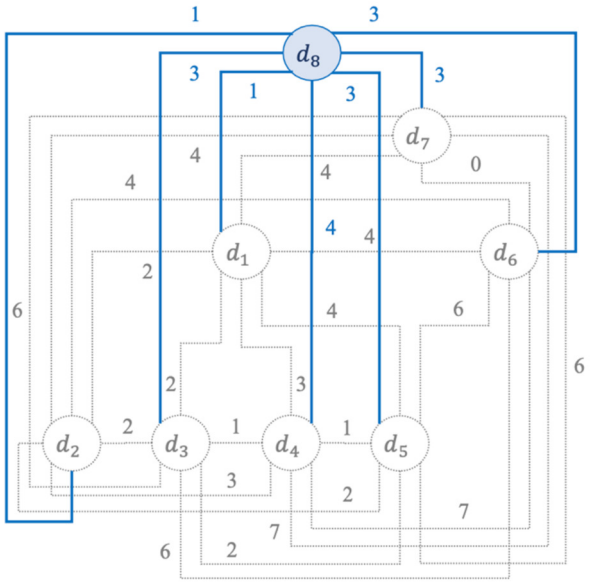

Fig.9: *The weighted graph of correlated generalization tuples after inserts d_8.*

For example, let Table 1 be the original dataset D. Let the value of k be 2. Let Figure 4 be the weighted graph of correlated distortion tuples of Table 1. Therefore, an appropriate G_{SUB} version of Figure 4 such that it is satisfied by $k = 2$ to be shown in Figure 8. That is, the weighted graph of correlated distortion tuples of Table 1 can be separated to be three appropriate sub-graph versions. The first sub-graph version includes both vertices as d_6 and d_7 (the red vertices). While three vertices construct the second sub-graph version as d_3, d_4, and d_5 (the green vertices). Another appropriate sub-graph version consists of d_1 and d_2 (the purple vertices). Thus, the partitioned data version of Table 1 is accorded to the sub-graph of its weighted graph to be shown in Table 11. After that, the unique values are available in each quasi-identifier of every partition to be distorted by their less specific values, i.e., the unique users' ages of each partition are distorted by their range, and the unique users' genders are distorted by an appropriately less specific value of them such that it is available in its DGH as shown in Figure 3. Therefore, a released data version of Table 1 is Table 9.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
d_1	0	2	2	3	4	4	4	1
d_2	2	0	2	3	2	4	4	1
d_3	2	2	0	1	2	6	6	3
d_4	3	3	1	0	1	7	7	4
d_5	4	2	2	1	0	6	6	3
d_6	4	4	6	7	6	0	0	3
d_7	4	4	6	7	6	0	0	3
d_8	1	1	3	4	3	3	3	0

Fig.10: *The adjacency matrix of Table 12 is after d_8 to be inserted.*

In addition, the examples are discussed in Sections 3. 4.1 and 3.4.2. We can see that the weighted graph of correlated distortion tuples and the adjacency matrix of Table 1 do not have any data changes. For this reason, we can conclude that if the data of D is not changed, only the first time has the cost for building the weighted graph of correlated distortion tuples and the adjacency matrix of D. Although the data of D is changed, the cost of building the weighted graph of correlated distortion tuples and the adjacency matrix of D only has the effect of the number of tuple changes.

For example, suppose that the new tuple d_8 is inserted into Table 1 such that it is shown in Table 12. Thus, the weighted graph of correlated distortion tuples and the adjacency matrix are shown in Figures 9 and 10, respectively. In this example, it is clear that when the data of D is changed when the new data becomes available, only the related new data are changed in the weighted graph of correlated distortion tuples and the adjacency matrix of D.

4. DATA UTILITY METRIC

Although D' is generally higher security in terms of privacy preservation than its original D. We can see that D' loses some data utility. For this reason, only D' is highly data utilities to be desired. Thus, the data utility metric is a necessity. Since privacy preservation based on k-anonymity constraints has been presented, several data utility metrics are proposed, e.g., Precision metric (PREC) [2], Discernibility metric (DM) [15], and Relative error [7].

4.1 Precision Metric (PREC) [2]

With the proposed models, D' is based on data distortions. For this reason, the penalty cost of D' depends on the distance and the number of distorted values. For the non-numerical quasi-identifier values, the penalty cost of data distortion for D' can be defined by Equation 1.

$$f_{D_1}(d_i, QI, DGH_{qi_y}) = \sum_{y=1}^{|QI|} \frac{f_{dis}(d_i[qi_y])}{|DGH_{qi_y}|}$$

$$f_{GEN1}(D', DGH_{qi_y}) = \frac{\sum_{i=1}^{|D'|} f_{D_1}(d_i, QI, DGH_{qi_y})}{|D'| * |QI|} \quad (1)$$

Where,

■ $|QI|$ is the number of quasi-identifier attributes that are available in D.

■ $f_{dis}(d_i[qi_y])$ is the distance of the distorted value that is available in qi_y of d_i.

■ $|DGH_{qi_y}|$ is the height of the DGH for DO^{qi_y}

■ $|D'|$ is the number of tuples that are available in D.

For the numerical quasi-identifier values, the penalty cost of data distortion for D' can be defined by Equation 2.

$$\mathrm{f}_{D_2}(d_i, QI, DGH_{qi_y}) = \sum_{y=1}^{|QI|} \frac{f_{dis}(d_i[qi_y])}{f_{MAX}(qi_y) - f_{MIN}(qi_y)}$$

$$f_{GEN2}(D', DGH_{qi_y}) = \frac{\sum_{i=1}^{|D'|} f_{D_2}(d_i, QI, DGH_{qi_y})}{|D'| * |QI|} \quad (2)$$

Where,

■ $f_{MAX}(qi_y)$ is the maximum value that is available in qi_y of D.

■ $f_{MIN}(qi_y)$ is the minimum value that is available in qi_y of D.

4.2 Discernibility Metric (DM) [15]

The DM metric is a data utility metric that can also be used to define the penalty cost or the data utility of D'. With the DM metric, the penalty cost of D' depends on the size of equivalence classes. The larger size of equivalence classes leads to more penalty cost DM. Therefore, the DM penalty cost of D' can be defined by Equation 3.

$$f_{DM}(D') = \sum_{p=1}^{|EC|} |ec_p|^2 \quad (3)$$

Where,

■ $|EC|$ is the number of equivalence classes that are available in D'.

4.3 Relative Error [7]

The relative error is a metric that can also define the penalty cost of D'. With this metric, the data utility of D' depends on the difference in the query results between D' and D. The more relative error means that D' has less data utility. For query results that are numerical data, their relative errors can be defined by Equation 4.

$$f_{RE1}(v, v_0) = \frac{|v - v_0|}{v} \quad (4)$$

Where,

■ v is the result that is queried from D.

■ v_0 is the relative result of v such that it is queried from D'.

With query results that are not numerical data, their relative errors can be defined by Equation 5.

$$f_{RE1}(n(v), n(v_0)) = \frac{|n(v) - n(v_0)|}{n(v)} \quad (5)$$

Where,

■ $n(v)$ is the number of values that are queried from D.

■ $n(v_0)$ is the number of the relative values of $n(v)$ such that they are queried from D'.

5. EXPERIMENT

In this section, the effectiveness and efficiency of the proposed privacy preservation models are discussed by comparison with k-member [13] and Mondrian Multidimensional k-anonymity [14].

5.1 Experimental Setup

All experiments are proposed to evaluate the effectiveness and efficiency of the proposed privacy preservation model, they are conducted on both Intel(R) Xeon(R) Gold 5218 @2.30 GHz CPUs with 64 GB memory and six 900 GB HDDs with RAID-5. Furthermore, all implementations are built and executed on Microsoft Windows Server 2019 in conjunction with Microsoft Visual Studio 2019 Community Edition and Microsoft SQL Server 2019. Moreover, they are discussed and conducted on the Adult dataset which is available at the UCI Machine Learning Repository [16]. This dataset is constructed from 32561 user profile tuples. Each user profile tuple consists of 14 attributes, i.e., Age, Workclass, Fnlwgt,

Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-loss, Hours-per-week, and Native-country. To conduct effective experiments, only the attributes Age, Education, Marital-status, Occupation, Sex, Capital-loss, and Native-country are available in the experimental dataset. The attributes Age, Education, Marital-status, Occupation, Sex, and Native-country are set to be the quasi-identifier attributes, and another attribute, Capital-loss, is set to be the sensitive attribute. Moreover, all user profile tuples include the values "?" and "0", they are removed. Therefore, the experimental dataset only includes 1428 user profile tuples.

5.2 Experimental Results and Discussion

5.2.1 Effectiveness

In this section, the effectiveness of the proposed privacy preservation models is evaluated by the PREC, DM, and relative error metrics.

In the first experiment, we propose to evaluate the effect of the given value of k that influences the data utility of datasets. All experimental results are discussed. They are based on the PREC penalty cost. Moreover, all tuples and all attributes are available in the experiments, and the value of k is varied from 2 to 20.

From the experimental results shown in Figure 11, we can see that the value of k influences the data utility of datasets (i.e., the number and the level of distorted values in the datasets). That is, the value of k is when increased. The number and the level of distorted values in the datasets also increase. Because the value of k directly influences the size of equivalence classes and the level of distorting quasi-identifier values. Moreover, we can see that both of the proposed privacy preservation models are more effective than both of the compared privacy preservation models, and we can further see that k-member is more effective than Mondrian Multidimensional k-anonymity. That is because the equivalence classes of both proposed privacy preservations are built from more similarity tuples than the equivalence classes that are constructed by both of the compared privacy preservation models.

In the second experiment, we propose to evaluate the effect of the number of quasi-identifier attributes that influence the data utility of datasets. All experimental results are based on the PREC penalty cost. Moreover, all tuples are available in the experiments. The value of k is fixed to be 10. The number of quasi-identifier attributes is varied from 1 to 6.

From the experimental results shown in Figure 12, we can see that the number of quasi-identifier attributes also influences the number and the level of distorted values in the datasets. It is when the number of quasi-identifier attributes that are increased. The number and the level of distorted values in the datasets also increase. However, the number of quasi-identifier attributes influences the level of distorted values in the equivalence classes to be less than the value of k. Also, both proposed privacy preservation models are more effective than the compared privacy preservation models in all experiments.

In the fourth experiment, we propose to evaluate the effect of the value of k that influences the size of equivalence classes in datasets. All experimental results are based on the DM penalty cost. Moreover, all tuples and all attributes are available in the experiments. The value of k is varied from 2 to 20.

From the experimental results shown in Figure 14, we can see that the value of k directly influences the DM penalty cost of datasets or the size of equivalence classes of datasets. The large size of equivalence classes leads to more DM penalty cost of datasets. Also, both proposed privacy preservation models are more effective than the compared privacy preservation models in all experimental results.

In the third experiment, we propose to evaluate the effect of the size of datasets that influences the data utility of datasets. All experimental results are also based on the PREC penalty cost. Moreover, all quasi-identifier attributes are available in the experiments. The value of k is fixed to be 10. The size of datasets is varied from 200 to 1400.

From the experimental results that are shown in Figure 13, we can see that when the size of datasets is increased, the number and the level of distorted values in the datasets is decreased, or we can say that when the size of datasets is increased, the data utility of datasets is also increased. That is because when the size of datasets is increased, the variety of values in the datasets is also increased. Also, both proposed privacy preservation models are more effective than the compared privacy preservation models in all experimental results of the third experiment.

In the fifth experiment, we propose to evaluate the effect of the number of quasi-identifier attributes that influence the size of equivalence classes in datasets. All experimental results are based on the DM penalty cost. Moreover, all tuples are available in the experiments. The value of k is fixed to be 10. The number of quasi-identifier attributes is varied from 1 to 6.

From the experimental results shown in Figures 14 and 15, the number of quasi-identifier attributes also influences the size of equivalence classes. However, they are less effective than the value of k. Also, both proposed privacy preservation models are more effective than the compared privacy preservation in all experimental results.

In the sixth experiment, we propose to evaluate the effect of the number of quasi-identifier attributes that influence the query results that are queried by the AND and OR query operations and the range of query conditions. All experimental results are based on relative errors. Moreover, the number of quasi-

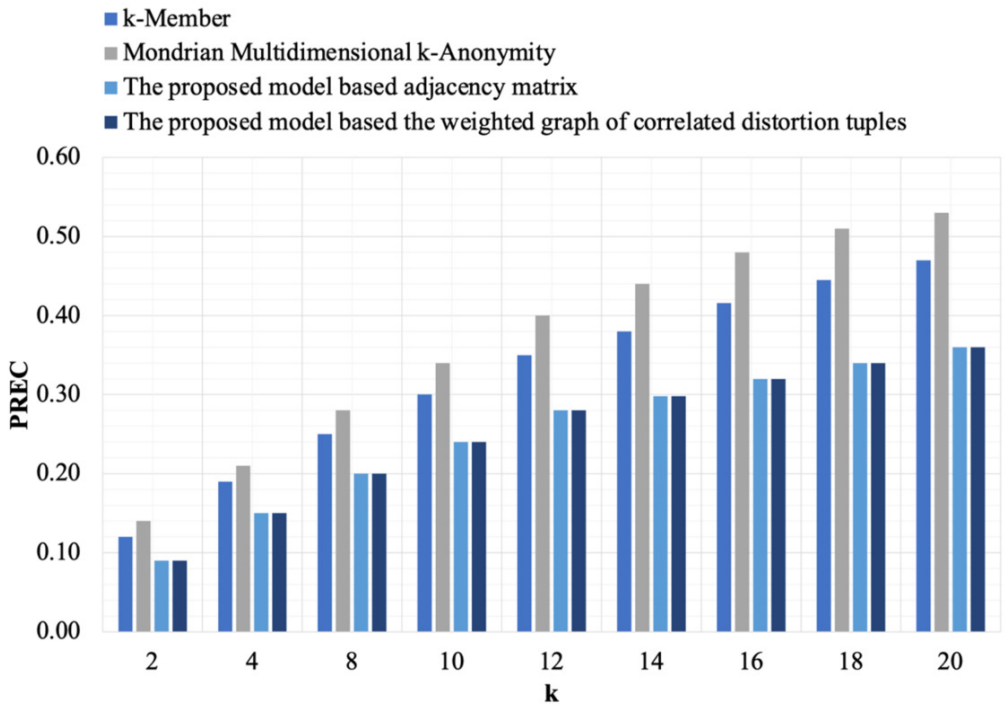

Fig.11: *The effectiveness based on the parameter k in conjunction with PREC.*

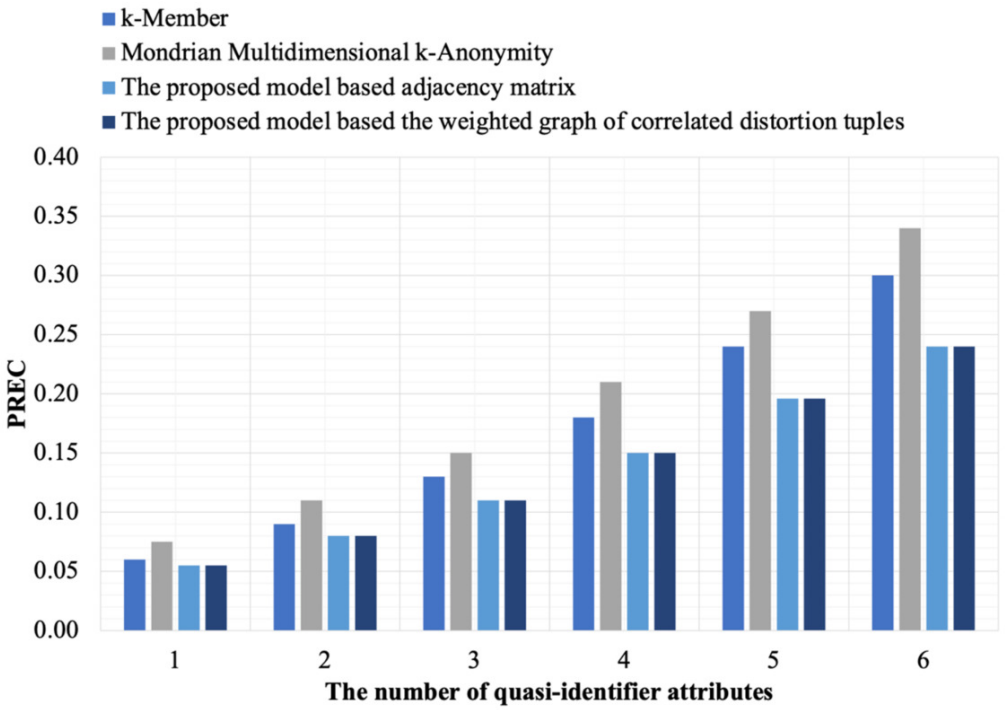

Fig.12: *The effectiveness based on the number of quasi-identifier sttributes in conjunction with PREC.*

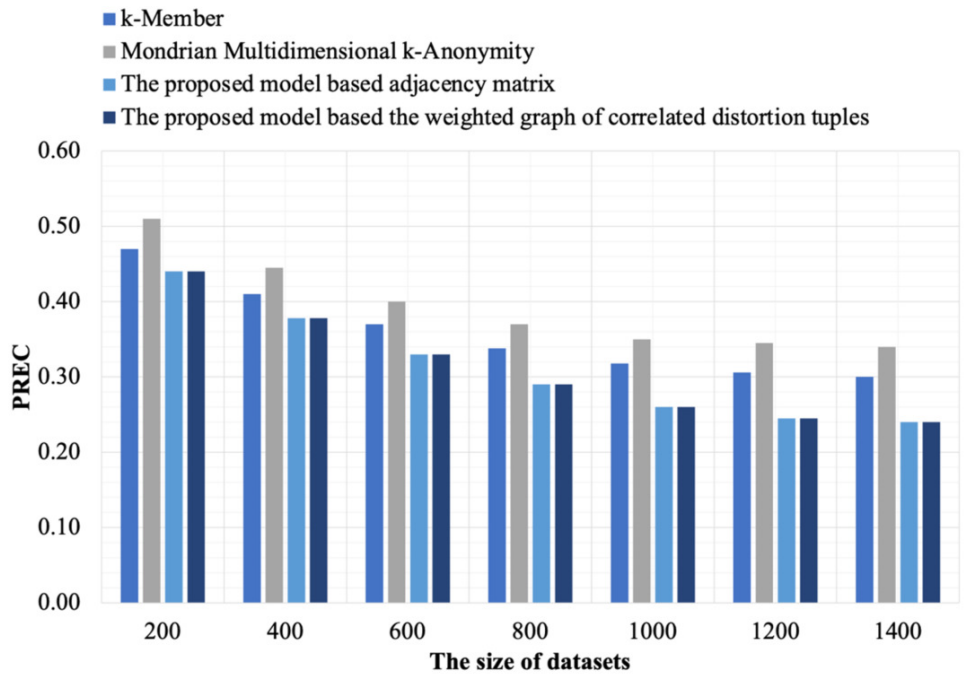

Fig.13: *The effectiveness based on the size of datasets in conjunction with PREC.*

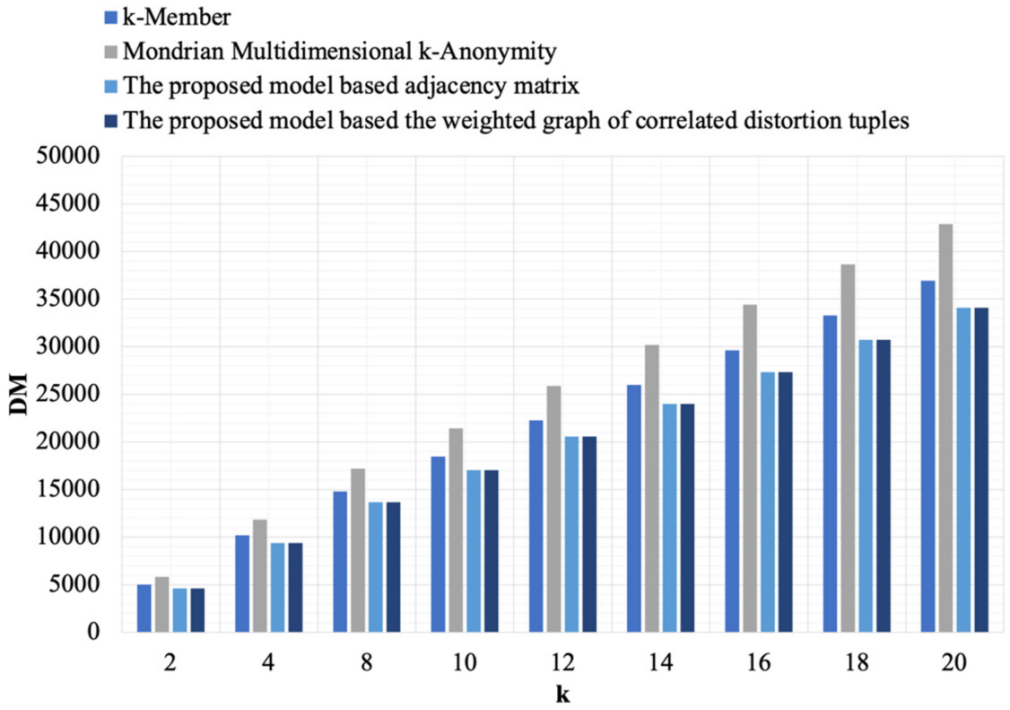

Fig.14: *The effectiveness based on the parameter k in conjunction with DM.*

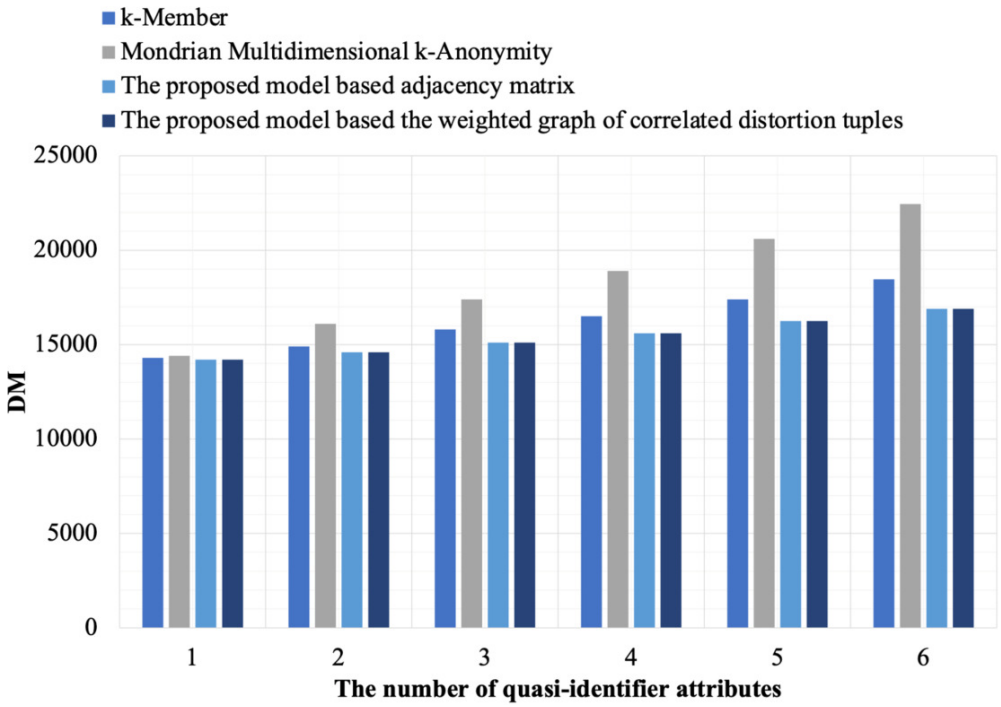

Fig.15: *The effectiveness based on the number of quasi-identifier attributes in conjunction with DM.*

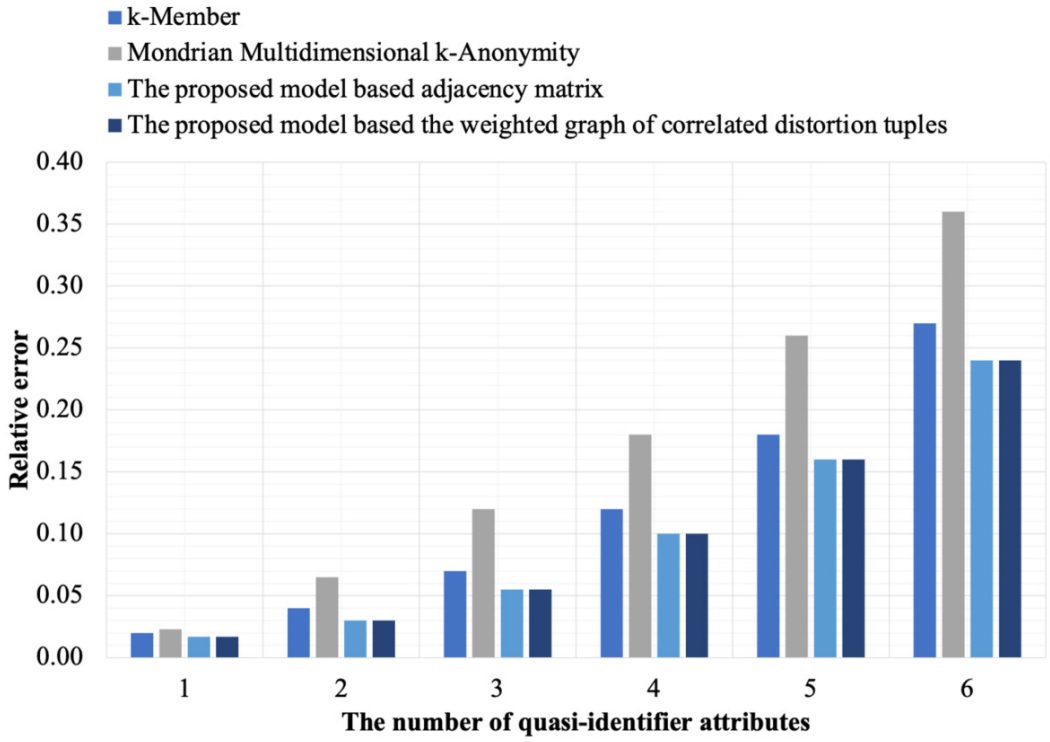

Fig.16: *The effectiveness based on the AND query operation in conjunction with relative errors.*

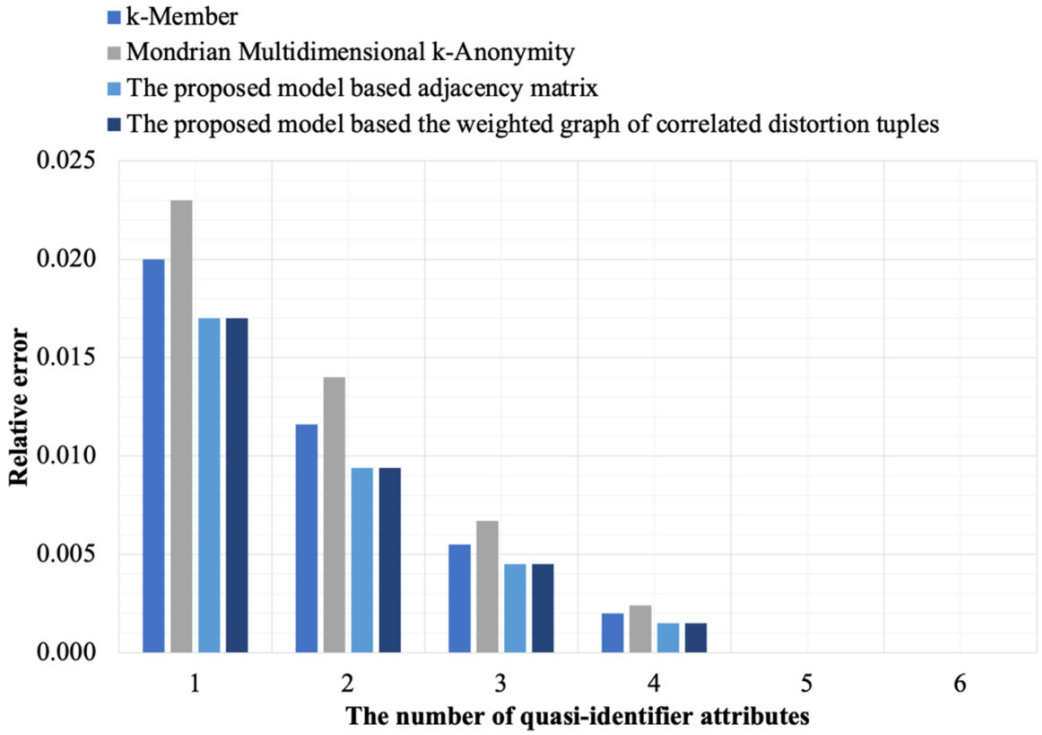

Fig.17: *The effectiveness based on the OR query in conjunction with relative errors.*

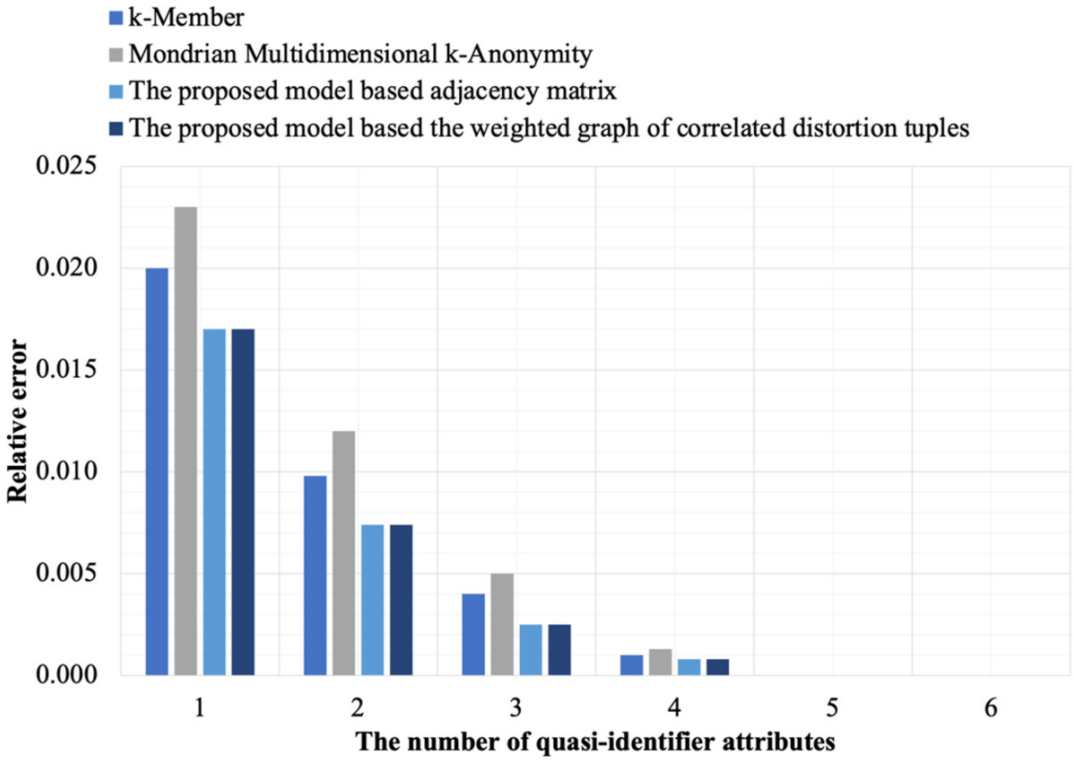

Fig.18: *The effectiveness based on the range of queries in conjunction with relative errors.*

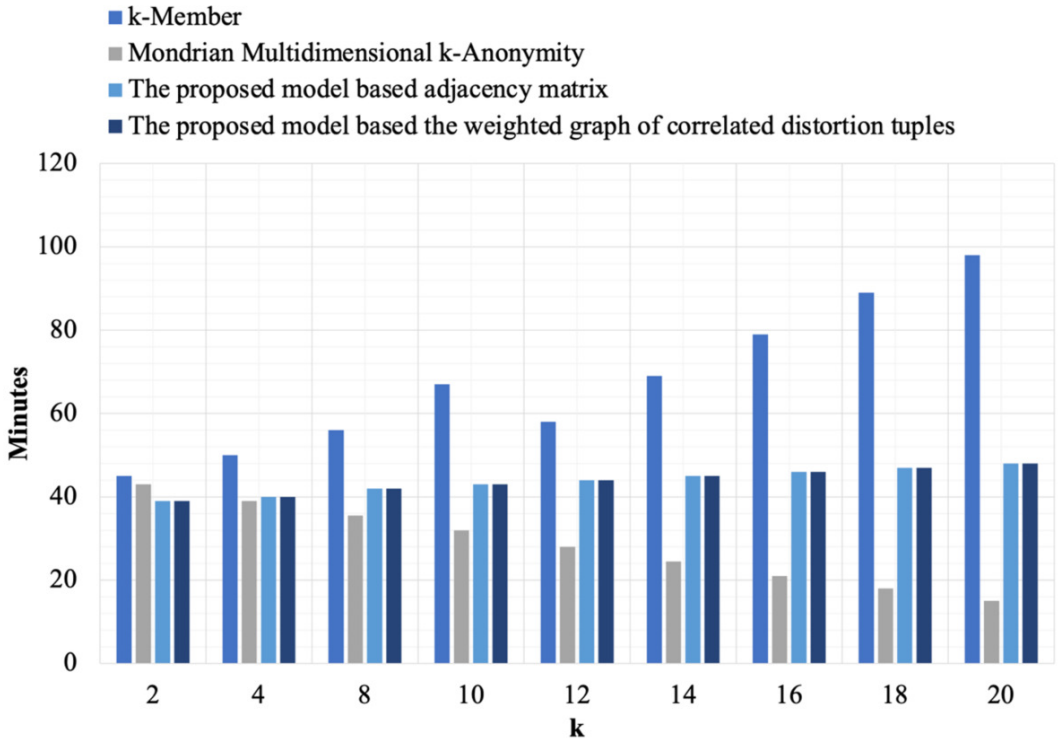

Fig.19: *Execution time.*

identifier attributes in queried conditions is varied from 1 to 6.

From the experimental results shown in Figures 16, 17, and 18, the number of quasi-identifier attributes is more influential to the query results. With the AND operation, Figure 16, we can see that the number of condition attributes has more effect on the penalty of query results, i.e., when the number of condition attributes is increased, the quality of query results is decreased. This effect of query results is the limitation of the distortion options that all experimental privacy preservation models have. When the number of condition attributes increases, the satisfied values for constructing the query results are more different from the original. Contrastively, when the number of condition attributes increases, the quality of query results of the OR query operations and the range of query conditions are increased (they are shown in Figures 17 and 18). This is because when the number of condition attributes is increased, the satisfied values of the query results from the distorted datasets and their original lead are more similar.

5.2.2 Efficiency

In this section, the efficiency of the proposed privacy preservation models is evaluated. Moreover, all tuples and all attributes are available in the experiments. The value of k is varied from 2 to 20. From the experimental results shown in Figure 19, we can see that when the value of k is increased, the execution time for choosing the suitable tuples to construct the equivalence classes of the proposed technique and k-member is also increased. That is because the higher value of k leads to the larger size of the equivalence classes in datasets. Thus, the number of iterations for constructing the equivalence classes of datasets is increased. However, both proposed privacy preservation models are more efficient than k-member. In addition, the value of k is 10. We can see that k-member uses the execution time for constructing its released datasets to be more than when k is set to be 11 because k-member has both sub-algorithms for considering the suitable tuples to equivalence classes, i.e., the best tuple selection algorithm and the best equivalence class selection algorithm. The best tuple selection algorithm is the main algorithm of k-member, while the best equivalence class selection algorithm is the optional algorithm of k-member. Thus, when the best equivalence class selection algorithm is enabled, the execution time of the k-member is increased. Contrastively, the trend of Mondrian Multidimensional k-anonymity uses the execution time to be different from the proposed privacy preservation models and k-member, i.e., when the number of k is increased, the execution time for constructing the released datasets of Mondrian Multidimensional k-anonymity is decreased. The cause of decreasing the execution time of Mondrian Multidimensional k-anonymity is that the search space for considering the suitable tuples for constructing the equivalence classes of datasets is reduced by half of the previous process.

6. CONCLUSION

In this work, both privacy preservation models can transform datasets for satisfying k-anonymity constraints to be proposed. They are based on the weighted graph of correlated distortion tuples and the adjacency matrix of tuple distances. That is, every equivalence class of datasets is constructed from the group of the more similar tuples, i.e., the group of tuples has the summation of distance tuples to be minimum. Moreover, all experimental results indicate that both proposed privacy preservation models are more effective and efficient than the compared privacy preservation models. In addition, the proposed privacy preservation models can be further applied in l-diversity, (k, e)-anonymous, (α, k)-anonymity, t-closeness, k^m-anonymity, and LKC-privacy.

References

[1] S. Riyana, K. Sasujit and N. Homdoung, "Privacy-enhancing data aggregation for big data analytics," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 17, no. 3, pp. 452–468, 2023.

[2] L. Sweeney, "K-anonymity: A model for protecting privacy," *International Journalon Uncertainty, Fuzziness and Knowledge-basedSystems*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

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

[4] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, pp. 106-115, 2007.

[5] M. Terrovitis, N. Mamoulis and P. Kalnis, "Privacy-preserving anonymization of set-valued data," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 115–125, Aug. 2008.

[6] B. C. M. Fung, M. Cao, B. C. Desai and H. Xu, "Privacy protection for RFID data," in *Proceedings of the 2009 ACM Symposium on Applied Computing, SAC '09*, New York, NY, USA, pp. 1528–1535, 2009.

[7] Q. Zhang, N. Koudas, D. Srivastava and T. Yu, "Aggregate Query Answering on Anonymized Tables," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, pp. 116-125, 2007.

[8] R. C.-W. Wong, J. Li, A. W.-C. Fu and K. Wang., "$(\alpha,$ k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, Association for Computing Machinery, New York, NY, USA, pp. 754–759, 2006.

[9] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. 96–96, 2006.

[10] S. Riyana, "$(l^{p_1}, \ldots, l^{p_n})$-privacy: privacy preservation models for numerical quasi-identifiers and multiple sensitive attributes, *Journal of Ambient Intelligence and Humanized Computing*, vol.12, pp. 9713–9729, 2021.

[11] K. Nissim, S. Raskhodnikova and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

[12] D. J. Bernstein, "Understanding brute force," in *Workshop record of ECRYPT STVL workshop on symmetric key encryption, eSTREAM report*, Citeseer, vol. 36, pp. 2005., 2005.

[13] J.-W. Byun, A. Kamra, E. Bertino and N. Li, "Efficient k-anonymization using clustering techniques," in *Proceedings of the 12th International Conference on Database Systems for Advanced Applications, DASFAA'07*, Berlin, Heidelberg, pp. 188–200, 2007.

[14] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, pp. 25–25, Apr. 2006.

[15] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *21st International Conference on Data Engineering (ICDE'05)*, pp. 217–228, Apr. 2005.

[16] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, AAAI Press, pp. 202–207, 1996.

Surapon Riyana received a B.S. degree in computer science from Payap University (PYU), Chiangmai, Thailand, in 2005. Moreover, He further received a M.S. degree and a Ph.D. degree in computer engineering from Chiangmai University (CMU), Thailand, in 2012 and 2019 respectively. Currently, he is a lecturer in Smart Farming and Agricultural innovation Engineering (Continuing Program), School of Renewable Energy, Maejo University (MJU), Thailand. His research interests include data mining, databases, data models, privacy preservation, data security, databases, and the internet of things.

Kittikorn Sasujit (Assistant Professor) received a B.Eng (Environmental Engineering) in 2004 from Rajamangala University of Technology Lanna, Thailand, and an M. Eng and Ph.D. (Energy Engineering) in 2008 and 2020, respectively, from Chiang Mai University, Thailand. His studies will include biomass technology, wind energy technology, NTP applications for biomass tar removal, and renewable energy.

Nigran Homdoung received a B.S. degree in mechanical engineering from King Mongkut's University of Technology Thonburi (KMUTT), Thailand, in 2001. He received a M.Eng. in Energy Engineering from Chiang Mai University (CMU), Thailand, in 2007. Moreover, he received a D.Eng. In Mechanical Engineering from Chiang Mai University (CMU), Thailand, in 2015. His research interests include biomass technology (gasification and pyrolysis process) and application Internal combustion Engine to biofuels. machine learning, data science, and artificial intelligence.