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

Privacy-Enhancing Data Aggregation for Big Data Analytics

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

ABSTRACT

Data utility and data privacy are serious issues that must be considered when datasets are utilized in big data analytics such that they are traded off. That is, the datasets have high data utility and often have high risks in terms of privacy violation issues. To balance the data utility and the data privacy in datasets when they are provided to utilize in big data analytics, several privacy preservation models have been proposed, e.g., k-Anonymity, l-Diversity, t-Closeness, Anatomy, k-Likeness, and $(l^{p1}, \ldots, l^{pn})$-Privacy. Unfortunately, these privacy preservation models are highly complex data models and still have data utility issues that must be addressed. To rid these vulnerabilities of these models, a new privacy preservation model is proposed in this work. It is based on aggregate query answers that can guarantee the confidence of the range and the number of values that can be re-identified. Furthermore, we show that the proposed model is more efficient and effective in big data analytics by using extensive experiments.

Article information:

Keywords: Dynamic Datasets, Data Streaming, Aggregate Query Frameworks, Privacy Violation Issues, Privacy Preservation Models

Article history:
Received: June 7, 2023
Revised: August 17, 2023
Accepted: September 21, 2023
Published: September 30, 2023 (Online)

DOI: 10.37936/ecti-cit.2023173.252952

1. INTRODUCTION

Data is one of the key elements for developing and improving the policy of organizations, e.g., organization management, financial management, employee management, customer service, and marketing strategy. Thus, an organization has up-to-date, volume, accurate, varied, and covered datasets. It can be advantageous in terms of business competition. To achieve these aims in datasets, one of the well-known data definitions, big data [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11], is proposed such that they are based on the data properties that are volume, velocity, variety, veracity, and value. Aside from big data, we can see that data analysis tools are also proposed constantly. They are tools for extracting data patterns and unknown data correlations in datasets. A few examples of the well-known data analysis tools are R-Programming [12] [13], Apache Hadoop [14] [15] [16], RapidMiner [17] [18] [19], and Microsoft Azure [20] [21] [22] [23]. Aside from data utilization, data privacy [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] must also be considered when datasets are allowed to be utilized in big data analytics. For this reason, balancing the data utility and the data privacy is a challenge in datasets when datasets are released and provided to the data analyst. To achieve data utility and data privacy, there are several well-known privacy preservation models to be proposed, i.e., data anonymization models [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44], data anatomization models [45] [46] [47] [48], and aggregate query frameworks [49]. In addition, some privacy preservation models are based on data anonymization in conjunction with aggregate query frameworks such as k-Likeness [50] and $(l^{p1}, \ldots, l^{pn})$-Privacy [51] [52]. Although these models can address privacy violation issues in datasets, they still have various serious vulnerabilities, e.g., data utility issues and complexity. To rid these vulnerabilities, a new privacy preservation model is based on aggregate query answers and can guarantee the confidence of the range and the number of values that can be re-identified to be proposed in this work.

The organization of this work is as follows. The first section, Section 2, is proposed to explain the motivation of this work. Then, we present the proposed model (Section 3). The experimental results for describing the efficiency and effectiveness of the proposed model will be discussed in Section 4. Finally, the conclusion and the future work are discussed in Sections 5 and 6, respectively.

[1,2,3] The authors are with Maejo university, Sansai, Chiangmai, Thailand, 50290, E-mail: surapon_r@mju.ac.th, kittikorn@mju.ac.th and nigran@mju.ac.th

2. MOTIVATION

2.1 Data anonymization

2.1.1 k-Anonymity [34]

In 2002, L.Sweeney proposed a well-known privacy preservation model, i.e., k-Anonymity. For privacy preservation, given the parameter k is privacy preservation constraints, where $k \in I^+$ and $k \geq 2$. That is, before datasets are released for public use, all explicit identifier values of users are removed. Moreover, the unique quasi-identifier values of users are suppressed or generalized by their less specific values to be at least k indistinguishable tuples.

For example, let Table 1 be the raw dataset and the value of k set to be 2. For privacy preservation, all explicit identifier values of users are removed, i.e., SSN. Moreover, the unique quasi-identifier values, i.e., Gender, Age, and Position, of users are generalized by their less specific values to be at least two indistinguishable tuples. Thus, Table 2 is a released data version of Table 1. We can see that Table 2 can guarantee that all possible query conditions always have at least two satisfied tuples. Thus, Table 2 is more secure than its original data version (Table 1). However, we can further see that Table 2 loses some data meaning in terms of data utilization, and it is highly complex in terms of data transformation processes [39] [50] [53]. Furthermore, k-Anonymity still has serious privacy violation issues that must be addressed, e.g., the privacy violation issues from using the adversary background knowledge about the target user and the privacy violation issues when the sensitive values in the sensitive attribute are homogeneity.

Table 1: *An example of raw datasets.*

SSN	Gender	Age	Position	Salary
000-00-0001	Male	37	Programmer	\$50,000
000-00-0002	Female	40	Programmer	\$50,000
000-00-0003	Male	40	Doctor	\$55,000
000-00-0004	Male	55	Doctor	\$68,000
000-00-0005	Female	55	Accounting	\$48,000
000-00-0006	Female	55	Marketing	\$49,000

Table 2: *A released data version of Table 1 is satisfied by 2-Anonymity constraints.*

Gender	Age	Position	Salary
*	37-40	Programmer	\$50,000
*	37-40	Programmer	\$50,000
Male	40-55	Doctor	\$55,000
Male	40-55	Doctor	\$68,000
Female	55	*	\$48,000
Female	55	*	\$49,000

An example of privacy violation issues in datasets that are satisfied by k-Anonymity constraints. We suppose that the adversary receives Table 2. Let Bob be the target user of the adversary. The adversary needs to reveal Bob's salary collected in Table 2. We assume that the adversary highly believes a user profile tuple that is available in Table 2 to be Bob's profile tuple. The adversary further knows that Bob is a male person who is 37 years old. In this situation, the adversary can infer that Bob's salary is \$50,000 because only \$50,000 relates to a male person who is 37 years old. Therefore, we can conclude that although datasets satisfy k-Anonymity constraints, they still have privacy violation issues that must be addressed. To rid this vulnerability of k-Anonymity, l-Diversity was proposed. It will be explained in Section 2.1.2.

2.1.2 l-Diversity [35]

An extendedly well-known privacy preservation model of k-Anonymity is l-Diversity [35]. It was proposed by Machanavajjhala et al in 2006. With this privacy preservation model, a positive integer l is privacy preservation constraints, where $l \in I^+$ and $l \geq 2$. That is before datasets are released for public use. All explicit identifier values of users are first removed. Finally, the unique quasi-identifier values of users are suppressed or generalized such that they are related to at least l distinct sensitive values.

For example, let Table 1 be the raw dataset and the value of l set to be 2. For privacy preservation, all explicit identifier values of users are removed in the first step. Finally, the unique quasi-identifier values of users are suppressed or generalized by their less specific values to be indistinguishable such that every group of indistinguishable quasi-identifier values must relate to at least two distinct sensitive values. Therefore, a released data version of Table 1 is shown in Table 3. With this released dataset, we can see that it can guarantee that all possible conditions always have at least two distinct sensitive values that are satisfied. Thus, Table 3 is more secure than Table 2. Therefore, we can conclude that the datasets of l-Diversity constraints are more secure than the datasets of k-Anonymity. However, the datasets of l-Diversity often lose the data utility more than the datasets of k-Anonymity. Moreover, we can see that the datasets of l-Diversity constraints still have privacy violation issues that must be addressed when the adversary has enough background knowledge about the target user and the sensitive values related to the indistinguishable quasi-identifier values to be the close values.

Table 3: *A released data version of Table 1 is satisfied by 2-Diversity constraints.*

Gender	Age	Position	Salary
*	37 - 55	*	\$50,000
*	37 - 55	*	\$49,000
Male	40 - 55	Doctor	\$55,000
Male	40 - 55	Doctor	\$68,000
Female	40 - 55	*	\$50,000
Female	40 - 55	*	\$48,000

An example of privacy violation issues is when the sensitive values in datasets are close. We suppose that the adversary received Table 3, and we assume that Bob is the target user of the adversary such that the adversary needs to reveal Bob's salary. Moreover, the adversary highly believes that a user profile tuple of Table 3 is Bob's profile tuple. The adversary further knows that Bob is a male person who is 37 years old. Therefore, the adversary can infer that Bob's salary is in the narrow range between $49,000 and $50,000. To rid this vulnerability of l-Diversity, t-Closeness is proposed.

2.1.3 t-Closeness [36]

To address privacy violation issues by considering the close sensitive values. In [36], the authors propose a well-known privacy preservation model, t-Closeness. With this privacy preservation model, the parameter t is privacy preservation constraints. It is the minimum distance of sensitive values that have the concern of privacy violation issues. More value of t is higher security in terms of privacy preservation. It enables one to trade-off between data utility and data privacy in datasets. We limit the gain from v_1 to v_2 by limiting the distance between both values as s_1 and s_2. Intuitively, if $s_1 = s_2$, v_1 and v_2 are same. If s_1 and s_2 are close, v_1 and v_2 are close. Also, if s_1 and s_2 are more different, v_1 and v_2 are more distant. For privacy preservation, before datasets are released for public use. All explicit identifier values of users are removed. The unique quasi-identifier values of users are suppressed or generalized to be indistinguishable. The sensitive values of each indistinguishable quasi-identifier group must have the distance to be at least t.

Table 4: *A released data version of Table 1 is satisfied by 2,000-Closeness constraints.*

Gender	Age	Position	Salary
*	37 - 55	*	$48,000
*	37 - 55	*	$50,000
*	37 - 55	*	$55,000
*	40 - 55	*	$49,000
*	40 - 55	*	$50,000
*	40 - 55	*	$68,000

For example, let Table 1 be the raw dataset. Let the value of t be 2,000. To achieve 2,000-closeness constraints in Table 1, all explicit identifier values of users are removed in the first step. The unique quasi-identifier values of users are generalized to be indistinguishable. Moreover, the sensitive values of every indistinguishable quasi-identifier value group must have at least 2,000 distances. Therefore, a released data version of Table 1 is shown in Table 4. This table can guarantee that the query result of all possible query conditions always obtains the sensitive values that have at least 2,000 distances. In this situation, we can conclude that Table 4 is more secure than Tables 2 and 3.

In addition, aside from k-Anonymity, l-Diversity, and t-Closeness, other well-known data anonymization models have been proposed, e.g., (α, k)-Anonymous [54], k$^{\mathrm{m}}$-Anonymity [55], and LKC-Privacy [56].

2.2 Data anatomization [45]

Data anatomization constraints are also based on data distortions. That is, datasets are not concerned about privacy violation issues when all explicit identifiers of users are removed. At least l distinct sensitive values partition the tuples of datasets. Moreover, the partitions anatomize to be both tables, i.e., the quasi-identifier table and the sensitive table. In addition, the relationship of tuple partition in anatomized tables is defined by a partition identifier.

Table 5: *A partitioned data version of Table 1 is satisfied by l=2.*

Gender	Age	Position	Salary	PID
Male	37	Programmer	$50,000	1
Female	40	Programmer	$50,000	
Male	40	Doctor	$55,000	
Male	55	Doctor	$68,000	2
Female	55	Accounting	$48,000	
Female	55	Marketing	$49,000	

An example of privacy preservation is based on data anatomization constraints [45]. Let Table 1 be the raw dataset, and the value of l is set to be 2. At first, all explicit identifiers of users are removed. Then, the tuples are partitioned such that every partition must include at least l distinct sensitive values. Moreover, the identifier of tuple partitions is also defined by this step. Therefore, a partitioned data version of Table 1 is shown in Table 5. Finally, the tuples of every tuple partition are anatomized to be the quasi-identifier table and the sensitive table. A version of the quasi-identifier and the sensitive tables of Table 1 are shown in Tables 6 and 7, respectively.

Table 6: *The quasi-identifier table of Table 1 is satisfied by l=2.*

Gender	Age	Position	PID
Male	37	Programmer	1
Female	40	Programmer	
Male	40	Doctor	
Male	55	Doctor	2
Female	55	Accounting	
Female	55	Marketing	

Tables 2, 6, and 7 are clear that data anatomization constraints can maintain the data utility of datasets to be better than data anonymization. For example, suppose the data analyst needs to know how

Table 7: *The sensitive table of Table 1 is satisfied by l=2.*

Salary	PID
$50,000	1
$55,000	
$55,000	
$68,000	2
$48,000	
$49,000	

many programmers who are 37 years old in Tables 2, 6, and 7. With this query condition, the data analyst can see that the query result of Table 2 has two satisfied tuples. The query result of Table 1 (the raw dataset) and the result of joining between Tables 6 and 7 do not have any satisfied tuples. However, data anatomization constraints have vulnerabilities that are the same as the datasets that satisfy l-Diversity constraints. They still have privacy violation issues from considering the adversary's background knowledge about the target user when the sensitive values in the specified tuple partition are the close values.

2.3 Aggregate query frameworks [49] [50] [51]

Aside from data anonymization and data anatomization, aggregate query frameworks can also address privacy violation issues in datasets. That is, datasets are not concerned about privacy violation issues when they accord to the data limitation as follows.

- Every query result must be queried from a particular sensitive attribute by using an appropriate aggregate function.
- (Optional) the query condition can only determine through the specified quasi-identifier attribute(s).

An example of data utilization is satisfied by the given data limitations of aggregate query frameworks. It is shown in Query 1.

- **Query 1:** SELECT AVERAGE(Salary) FROM Table 1 WHERE Gender = 'Male'

With Query 1, the query result is $57,666.67 as the average salary. This query is an aggregate query answer, i.e., ($50000 + $55000 + $68000) / 3 = $57, 666.67. In this situation, the privacy violation issues in Table 1 seem impossible. Unfortunately, in [50] and [51], the authors illustrate that Table 1 still has privacy violation issues that must be addressed.

For example, suppose that Bob is the target user of the adversary. The adversary needs to reveal Bob's salary from Table 1. We assume that the adversary strongly believes that a tuple of Table 1 is Bob's profile tuple, and the adversary knows that Bob is a male person who is 37 years old. To reveal Bob's salary, the adversary first uses the COUNT function to verify and identify the desired risk query condition. An example of verifying and identifying the risk query condition is shown in Query 2.

- **Query 2:** SELECT COUNT(*) AS NumberOfRows FROM Table 1 WHERE Gender = 'Male' AND Age = 37

With Query 2, only a tuple in Table 1 can be satisfied. Thus, the adversary can be confident that the satisfied tuple must be Bob's profile tuple. Finally, the adversary uses an appropriate aggregate query function, e.g., MAX, MIN, AVERAGE, and SUM, to reveal Bob's salary. In addition, only the AVERAGE function is used in this example. It is shown in Query 3.

- **Query 3:** SELECT AVERAGE(Salary) FROM Table 1 WHERE Gender = 'Male' AND Age = 37

The query result of Query 3 is $50,000 as the average salary. Therefore, the adversary can infer that Bob's salary is $50,000 because this query result is not aggregate query answers, i.e., $50,000 / 1 = $50,000. From this example, it is clear that aggregate query frameworks still have privacy violation issues that must be addressed. For this reason, both enhanced aggregate query framework version is based on data suppression and data generalization to be proposed. They will be presented in Sections 2.3.1 and 2.3.2, respectively.

2.3.1 Aggregate query frameworks based on data suppression

To address the vulnerabilities of aggregate query frameworks, in [49], the authors suggest that before datasets are provided through aggregate query frameworks, all explicit identifier values of users are removed.

An example of aggregate query frameworks is based on data suppression. Let Table 1 be the raw dataset. For privacy preservation, the SSN of users is removed. The output of this step is shown in Table 8. Then, the unique quasi-identifier values are suppressed. The output of this step is shown in Tables 9, 10, and 11. Table 9 is focused on data utilization via Gender, but Tables 10 and 11 are focused on data utilization via Age and Position, respectively. With these tables, they can guarantee that all possible query always have at least two satisfied tuples. But we can see that they are different. Thus, they can be different in terms of data utilization. The data holder must choose an appropriate data version to provide in aggregate query frameworks. Finally, the suppressed data version provides aggregate query frameworks.

2.3.2 Aggregate query frameworks based on data generalization

A well-known privacy preservation model, k-Likeness, is enhanced from aggregate query frameworks. It is proposed in [50]. With this model,

Table 8: *The data version of Table 1 without SSN.*

Gender	Age	Position	Salary
Male	37	Programmer	$50,000
Female	40	Programmer	$50,000
Male	40	Doctor	$55,000
Male	55	Doctor	$68,000
Female	55	Accounting	$48,000
Female	55	Marketing	$49,000

Table 9: *A data version of Table 8 suppresses the unique quasi-identifier values of users such that it is focused on utilizing data via Gender.*

Gender	Age	Position	Salary
Male			$50,000
Female			$50,000
Male		Doctor	$55,000
Male		Doctor	$68,000
Female	55		$48,000
Female	55		$49,000

datasets are not concerned about privacy violation issues when the unique query conditions are generalized to be at least k indistinguishable tuples.

For example, let Table 8 be the raw dataset. Let k set to be 2. To achieve 2-Likeness constraints, the unique quasi-identifier values of users are generalized to be at least two indistinguishable tuples. Therefore, Table 2 is a released data version of Table 8. We can see that Table 2 can guarantee that all possible query conditions always have at least two satisfied tuples, and Table 2 has more data utility than Tables 9, 10, and 11. Therefore, we can conclude that the aggregate query framework based on data generalization is more effective than the aggregate query framework based on data suppression. Moreover, the proposed aggregate query framework of [50] can be more secure than the aggregate query framework of [49] because it can guarantee that all possible query conditions to datasets always have at least k satisfied tuples.

Table 10: *A data version of Table 8 suppresses the unique quasi-identifier values of users such that it is focused on utilizing data Age.*

Gender	Age	Position	Salary
			$50,000
	40		$50,000
	40		$55,000
			$68,000
Female	55		$48,000
Female	55		$49,000

Table 11: *A data version of Table 8 suppresses the unique quasi-identifier values of users such that it is focused on utilizing data utilizing data via Position.*

Gender	Age	Position	Salary
		Programmer	$50,000
		Programmer	$50,000
Male		Doctor	$55,000
Male		Doctor	$68,000
Female	55		$48,000
Female	55		$49,000

From the examples illustrated in Sections 2.3.1 and 2.3.2, the privacy violation issues in aggregate query frameworks seem impossible because provided datasets can guarantee that all possible query conditions always have at least two and k satisfied tuples, respectively. Unfortunately, these aggregate query frameworks still have privacy violation issues from using identical attacks. It will be presented in Section 2.3.3.

2.3.3 $(l^{p1}, \ldots, l^{pn})$-Privacy [51]

In [51], the author demonstrates that although datasets can guarantee that all possible query conditions always have at least two or k satisfied tuples, they still have privacy violation issues from using the MAX and MIN functions in conjunction with the adversary's background knowledge about the target user. To rid this vulnerability of aggregate query frameworks, $(l^{p1}, \ldots, l^{pn})$-Privacy [51] is proposed. That is, before datasets will be provided through aggregate query frameworks, all query conditions have the number of distinctly satisfied sensitive values in every sensitive attribute px to be at least l^{px} values, they are removed. Therefore, after datasets are satisfied by $(l^{p1}, \ldots, l^{pn})$-Privacy constraints, they can guarantee that all possible query results always have at least l^{px} distinctly satisfied sensitive values. Therefore, we can conclude that the aggregate query framework based on $(l^{p1}, \ldots, l^{pn})$-Privacy constraints are more secure than the aggregate query frameworks that are proposed in [50]. However, in [57], the author illustrates that $(l^{p1}, \ldots, l^{pn})$-Privacy is a privacy preservation model that is highly complexity. It further has data utility issues that must be addressed.

In addition, to the best of our knowledge about privacy preservation models based on equivalent classes (or data partitions), we observe that they are highly complex when the data holder needs to form the optimized equivalence classes of datasets. The complexity of transforming datasets to satisfy privacy preservation constraints can be separated as follows.

At first, the complexity of constructing all equivalent classes, EC, is presented by Equation 1

$$O(EC) = 2^{|T|} \tag{1}$$

Where,

- $|T|$ is the size of the provided datasets.

The complexity of generating all data versions is presented in Equation 2.

$$O(T'_{ALL}) = 2^{|EC|} \tag{2}$$

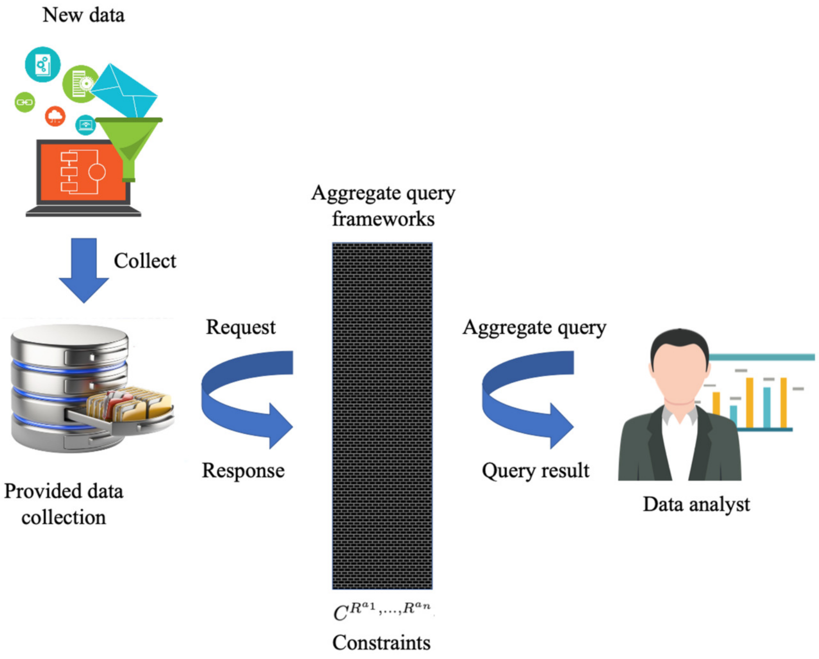

Fig.1: *The workflow of the proposed aggregate query framework.*

Another complexity of finding the desired data version, T'_{DES}, is presented by Equation 3.

$$O(find(T/_{DES})) = |T'_{ALL}| \tag{3}$$

Therefore, the complexity of transforming datasets to satisfy privacy preservation constraints can be presented by Equations 4 and 5.

$$O(T'_{DES}) = O(EC) + O(T'_{ALL}) + O(find(T'_{DES})) \tag{4}$$

Or

$$O(T'_{DES}) = 2^{|T|} + 2^{|EC|} + |T'_{ALL}| \tag{5}$$

With the examples that are illustrated in this section, we can conclude that the existing data anonymization, data anatomization, and aggregate query frameworks could be insufficient to address privacy violation issues in big data analytics because they have various data utility issues that must be addressed, and they are further high complexity. Moreover, they are the static privacy preservation model. That is, before datasets are released or provided to the data analyst, they must be transformed by the data holder to satisfy the given privacy preservation constraint. To rid these vulnerabilities of these models, a dynamic privacy preservation model based on aggregate query frameworks is proposed in this work. With the proposed model, aside from privacy preservation constraints, the complexity of data transformation processes and the data utility are also considered. It will be presented in Section 3.

3. THE PROPOSED MODEL

In this section, a privacy preservation model, $C^{R_{a1},\ldots,R_{an}}$-Privacy, for addressing privacy violation issues in dynamic datasets and big data analytics is proposed. It is based on aggregate query frameworks that can guarantee the confidence of the range and the number of values that can be re-identified by the adversary. Moreover, the proposed model allows the data analyst to query and define the query condition through all attributes of datasets. Aside from privacy preservation constraints, the complexity of data transformation processes and the data utility are also considered.

3.1 Problem definitions

Definition 1 (Dataset): Let A=$\{a_1, a_2, \ldots, a_n\}$ be the set of data attributes such that every a_x, where $1 \le x \le n$, of A must not be the explicit identifier of users. Let $Do^{a_x} = \{do_1^{a_x}, do_2^{a_x}, \ldots, do_m^{a_x}\}$ be the data domain of a_x. Let $U = \{u_1, u_2, \ldots, u_j\}$ be the set of users. Let $T = \{t_1, t_2, \ldots, t_j\}$ be the dataset that allows to change the data when the new data become available. Each $t_i \in T$, where $1 \le i \le j$, is represented by the profile tuple of $u_i \in U$ such that it is in the form of $t_i = \{do_1^{a_{x1}}, do_2^{a_{x2}}, \ldots, do_m^{a_{xn}}\}$, where $a_{x_1}, a_{x_2}, \ldots, a_{x_n} \in A$.

Definition 2 (Data query): Every query result must be queried from an attribute a_x by an appropriate aggregate query function such as COUNT, MAX, MIN, SUM, and AVERAGE. While the query condition(s) can be defined to all attributes.

For example, let Table 1 without SSN be the provided dataset in aggregate query frameworks. Both data queries satisfy Definition 2. They are shown in

Queries 4 and 5.

- **Query 4:** SELECT SUM(Salary) FROM Table 1 without SSN WHERE Age > 37
- **Query 5:** SELECT MIN(Salary) FROM Table 1 without SSN

Definition 3 (Privacy violation issues with non-diverse values): Let C be a positive integer. It is the maximum number of the adversary's knowledge about the target user in a_x. The scenario of privacy violation issues with non-diverse values is that the query result of a_x is constructed from at most C distinct values of a_x.

For example, let Table 1 without SSN be the provided dataset. Let the value of C be 5. An example of query results does not have any concern of privacy violation issues with non-diverse values. It is shown in Query 6. With Query 6, its query result is \$54,000 as the average salary, i.e., (\$50,000 + \$55,000 + \$68,000 + \$48,000 + \$49,000) / 5 = \$54,000.

- **Query 6:** SELECT AVERAGE(Salary) FROM Table 1 without SSN WHERE Age > 37

An example of query results about the concern of privacy violation issues with non-diverse values is shown in Query 7. With this query, the result is \$52,500 as the average salary, i.e., (\$50,000 + \$55,000) / 2 = \$52,500.

- **Query 7:** SELECT AVERAGE(Salary) FROM Table 1 without SSN WHERE Age = 40

In addition, although the adversary cannot be directly allowed to access the data that is available in provided datasets, the adversary can still use the DISTINCT COUNT function to verify and identify a risk query condition and further use the identified risk query condition in conjunction with an appropriate aggregate query function, i.e., COUNT, MAX, MIN, SUM, and AVERAGE, to violate the sensitive data of the target user. For this reason, we recommend that a suitable value of C is equal to or greater than 2.

Definition 4 (Privacy violation issues with the narrow range of query results): Let $a_x \in A$ be the specified attribute such that $D0^{a_x}$ is numerical. Let R^a_x be the maximum confidence range of data re-identification for a_x. The meaning of privacy violation issues with the narrow range of query results is that the difference between the maximum value and the minimum value of the query result to be at most $R^{a_x} - \beta^{a_x}$, where β^{a_x} is the closest value of R^{a_x}.

For example, let Table 1 without SSN be the provided dataset. Let the value of R^{Salary} be \$15,000, i.e., R^{Salary}=15,000. An example of query results that do not have any privacy violation issues with the narrow range of query results, it is shown in Query 8. With Query 8, its query result is \$68,000 as the maximum salary such that it is constructed from five satisfied query values that are \$50,000, \$55,000, and \$68,000, so, the range of the satisfied query values is \$50,000 - \$68,000 = \$18,000.

- **Query 8:** SELECT MAX(Salary) FROM Table 1 without SSN WHERE Gender = 'Male'

An example of query results that have privacy violation issues with the narrow range, it is shown in Query 9. With this query, its query result is \$55,000 as the average salaries such that it is constructed from three satisfied query values that are \$50,000, \$50,000, and \$55,000, so, the range of the satisfied query values is \$55,000 - \$50,000 = \$5,000.

- **Query 9:** SELECT MAX(Salary) FROM Table 1 without SSN WHERE Age >= 37 AND Age <= 40

3.2 $C^{R^{a_1},\dots,R^{a_n}}$-Privacy principle

Let T be the specified dataset to be provided in aggregate query frameworks. T does not include any explicit identifier value of users, and it allows to change the data when new data becomes available. Moreover, the data analyst can only utilize the data of T in the form of aggregate query answers that are described in Definition 2. Moreover, every possible query result must be constructed from at least C distinct query values. In addition, an arbitrary attribute $a_x \in A$ has the data domain, $D0^{a_x}$, to be numeric. It further considers the range between the lower and upper bounds of the satisfied query values to be at least R^{a_x}. If the query values do not accord to C and R^{a_x}, they do not return from aggregate query frameworks. The characteristic of the proposed aggregate query framework is shown in Figure 1.

Theorem 1: If every result is querying from $a_x \in A$ of T through aggregate query frameworks such that it is constructed from at least C distinct values, it has the re-identifiable confidence to be at most $\frac{1}{C}$.

Proof: Suppose that $DF(v_1^{a_x}, \dots, v_s^{a_x})$ be the function for getting the number of distinct values from the set of $v_1^{a_x}, \dots, v_s^{a_x} \subseteq D0^{a_x}$. We know that the result is querying from every $a_x \in A$ of T such that it does not have any concern of privacy violation issues, it must make from $v_1^{a_x}, \dots, v_s^{a_x}$ such that $v_1^{a_x}, \dots, v_s^{a_x}$ have $DF(v_1^{a_x}, \dots, v_s^{a_x})$ to be at least C. In this situation, we can say that the re-identifiable confidence for $v_1^{a_x}, \dots, v_s^{a_x}$ is at most $\frac{1}{C}$. Suppose not. When we can say that the solution of privacy preservation to the whole problem with $\delta_1^{a_x}, \dots, \delta_s^{a_x}$ such that $\delta_1^{a_x}, \dots, \delta_s^{a_x}$ have $DF(\delta_1^{a_x}, \dots, \delta_s^{a_x})$ at most $C - 1$. Let the solution of privacy preservation can define from $DF(v_1^{a_x}, \dots, v_s^{a_x}) - C \geq 0$ for the desired solution to the whole problem associated with a_x. Since we can construct the query result of a_x from $\delta_1^{a_x}, \dots, \delta_s^{a_x}$, then we can determine $DF(\delta_1^{a_x}, \dots, \delta_s^{a_x}) < DF(v_1^{a_x}, \dots, v_s^{a_x})$. But then we can use this scheme to get a solution for the whole problem of values, as shown in Equation 6, which contradicts the desirability of our original solution.

$$DF(\delta_1^{a_x},\ldots,\delta_s^{a_x}) < DF(v_1^{a_x},\ldots,v_s^{a_x}) - C \geq 0 \quad (6)$$

Theorem 2: If every result is querying from $a_x \in A$ of T through aggregate query frameworks such that it is constructed from the set of values that have the distance between the lower bound and the upper bound to be at least R^{a_x}, it has the re-identifiable confidence to be at most R^{a_x}.

Proof: Suppose that $v_1^{a_x},\ldots,v_s^{a_x} \subseteq D0^{a_x}$ is the set of values that satisfy the specified query condition such that $D0^{a_x}$ is numerical. Let $UB(v_1^{a_x},\ldots,v_s^{a_x})$ be the function for getting the upper bound of $v_1^{a_x},\ldots,v_s^{a_x}$. Let $LB(v_1^{a_x},\ldots,v_s^{a_x})$ be the function for getting the lower bound of $v_1^{a_x},\ldots,v_s^{a_x}$. We know that the query result does not have any concern of privacy violation issues to be constructed from $v_1^{a_x},\ldots,v_s^{a_x}$ such that the distance between $LB(v_1^{a_x},\ldots,v_s^{a_x})$ and $UB(v_1^{a_x},\ldots,v_s^{a_x})$ is equal to and greater than R^{a_x}. Suppose not. When we can say that the solution of privacy preservation to the whole problem with $\delta_1^{a_x},\ldots,\delta_s^{a_x}$ such that $\delta_1^{a_x},\ldots,\delta_s^{a_x}$ have $UB(v_1^{a_x},\ldots,v_s^{a_x}) - LB(v_1^{a_x},\ldots,v_s^{a_x})$ to be at most $R^{a_x} - \beta^{a_x}$, where β^{a_x} is the closest value of R^{a_x}. Let the solution of privacy preservation can define from $UB(v_1^{a_x},\ldots,v_s^{a_x}) - LB(v_1^{a_x},\ldots,v_s^{a_x}) - R^{a_x} \geq 0$ for the desired solution to the whole problem associated with a_x. Since we can construct the query result of a_x from $\delta_1^{a_x},\ldots,\delta_s^{a_x}$, then we can determine $UB(\delta_1^{a_x},\ldots,\delta_s^{a_x})$-$LB(\delta_1^{a_x},\ldots,\delta_s^{a_x}) < UB(v_1^{a_x},\ldots,v_s^{a_x})$−LB($\mathrm{v}_1^{a_x},\ldots,v_s^{a_x}$). But then we can use this scheme to get a solution for the whole problem of values, as shown in Equation 7, which contradicts the desirability of our original solution.

$$\begin{aligned} &UB(\delta_1^{a_x},\ldots,\delta_s^{a_x}) - LB(\delta_1^{a_x},\ldots,\delta_s^{a_x}) < \\ &UB(v_1^{a_x},\ldots,v_s^{a_x}) - LB(v_1^{a_x},\ldots,v_s^{a_x}) - R^{a_x} \geq 0 \end{aligned} \quad (7)$$

For example, let Table 1 without SSN be the dataset that is provided through proposed aggregate query frameworks. Let C and R^{Salary} set to be 2 and 10000, respectively. Suppose that Queries 10, 11, and 12 are the particular data queries.

- **Query 10:** SELECT MAX(Salary) FROM Table 1 without SSN WHERE Age >= 50
- **Query 11:** SELECT MAX(Salary) FROM Table 1 without SSN WHERE Age = 40
- **Query 12:** SELECT MAX(Age) FROM Table 1 with- out SSN WHERE Position ≠ 'Programmer'

Query 10 has the users' three salaries, i.e., \$68,000, \$48,000, and \$49,000. Thus, the different range of the satisfied values of Query 10 is \$20,000, i.e., \$68,000 - \$48,000 = \$20,000. In this situation, the aggregate query framework returns \$68,000 as the query result to the data analyst because it accords to C and R^{Salary}, respectively.

As Query 11, the aggregate query framework does not return any query result because the different range between the lower and upper bounds of the satisfied values does not accord to R^{Salary}.

Another query, Query 12, obtains 55 as the query result because the number of distinctly satisfied query values accords the given value of C, and the different range between the lower and upper bounds of the satisfied values does not require in Age.

Although datasets are provided through the proposed aggregate query framework, they can be higher secure than their original. However, we can see that they lose some data utility. For this reason, the data utility metric is necessary for the proposed model.

3.3 Relative error [38] [58]

The relative error is a data utility metric that can be used to define the penalty cost of provided datasets of aggregate query frameworks. The penalty cost of each query result is based on the difference between the original query result and its related experiment query result. The more relative error means that the query result has less data utility. For query results that are numerical data, their relative errors can be defined by Equation 8.

$$f_{REI}(v, v_0) = \frac{|v - v_0|}{v} \quad (8)$$

Where,

- v is the query result from the raw dataset.
- v_0 is the related query result of v such that it is queried from the aggregate query framework.

With query results that are not numerical data, their relative errors can be defined by Equation 9.

$$f_{REC}(n(v), n(v_0)) = \frac{|n(v) - n(v_0|}{n(v)} \quad (9)$$

Where,

- $n(v)$ is the query result from the raw dataset.
- $n(v_0)$ is the related query result of $n(v)$ such that it is queried from the aggregate query framework.

4. EXPERIMENTS

In this section, the effectiveness and efficiency of the proposed model, $C^{R_{a1},\ldots,R_{an}}$-Privacy, are discussed by comparing with the comparable privacy preservation models as k-Anonymity [34], l-Diversity [35], t-Closeness [36], Anatomy [45], Anatomy for multiple sensitive attributes (MSA-Anatomy) [46], k-Likeness [50], and $(l^{p1},\ldots,l^{pn})$-Privacy [51].

4.1 Experimental setup

All experiments are proposed to evaluate the effectiveness and efficiency of the proposed model They

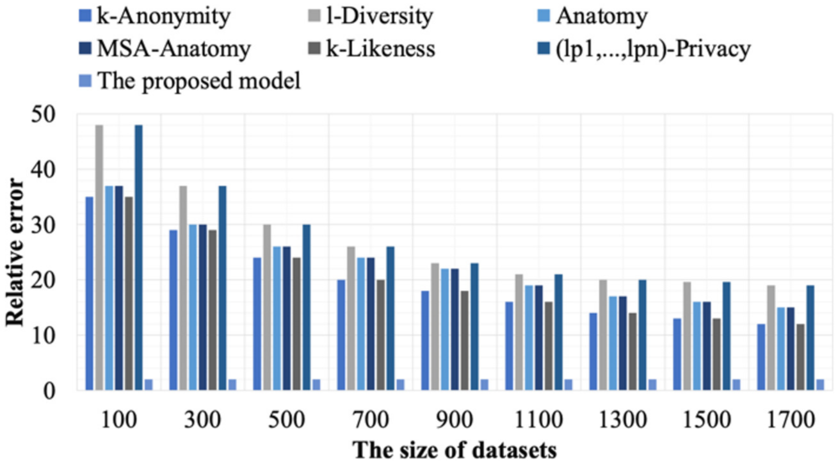

Fig.2: *The data utility based on the size of datasets.*

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 [59]. 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 tributes Age, Workclass, Education, Marital-status, Occupation, Relationship, Sex, Capital-loss, Hours- per-week, and Native-country are available in the experimental dataset. For k-Anonymity, l-Diversity, Anatomy, MSA-Anatomy, t-Closeness, k-Likeness, and $(l^{p1}, \ldots, l^{pn})$-Privacy, the attributes Age, Education, Marital-status, Occupation, Sex, and Native-country are set to be the quasi-identifier attributes, and the other attributes (i.e., Workclass, Capital-loss, Hours-per-week, and Relationship) are set to be the sensitive attributes.

4.2 Effectiveness

This section is proposed to discuss the effectiveness of the proposed model.

4.2.1 The data utility based on the size of datasets

The first experiment is proposed to evaluate the effect of dataset sizes that influence the data utility of datasets that are satisfied by k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, and the proposed model. For experiments, all quasi-identifier attributes are available in the experimental datasets. However, only Capital-loss is the sensitive attribute of the experimental datasets. Moreover, 100 tuples of the experimental datasets are randomly selected to be the initial tuples, thereafter, 200 tuples are randomly increased for each experiment until the experimental dataset collects 1700 tuples. The parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-Privacy, and the parameter C of the proposed model is set to be 4. In addition, the parameter R^{a_x} of the proposed model is not defined.

From the experimental results shown in Figure 2, we observe when the size of the experimental datasets increases, the data utility of datasets increases. Moreover, the experimental results indicate that the proposed model is more effective than the compared models. The cause of increasing the data utility is the variety of values in datasets.

4.2.2 The data utility based on the number of quasi-identifier attributes

The second experiment is proposed to evaluate the effect of the number of quasi-identifier attributes that influence the data utility of datasets that are satisfied by k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, and the proposed model. For experiments, all tuples are available in the experimental datasets, and only Capital-loss is the sensitive attribute of the experimental datasets. The parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-Privacy, and the parameter C of the proposed model is set to be 4. In addition, the parameter R^{a_x} of the proposed model is not defined. Moreover, the number of quasi-identifier attributes varies from 1 to 6.

From the experimental results shown in Figure 3, we observe when the number of quasi-identifier attributes of the datasets increases, the data utility

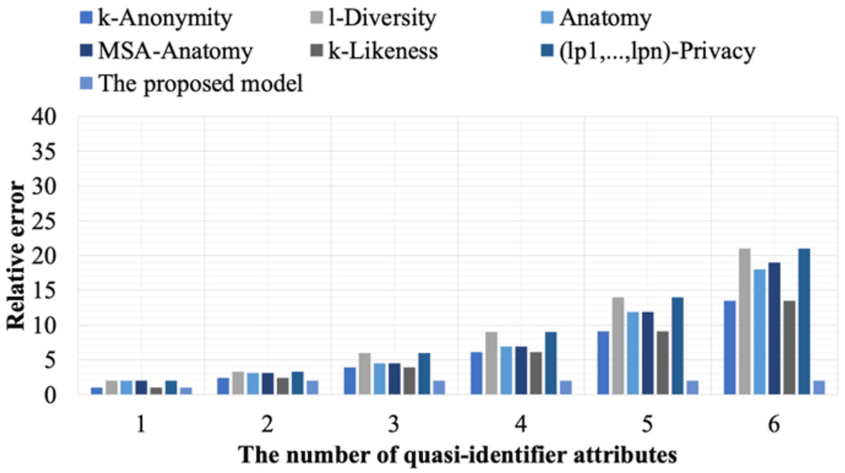

Fig.3: *The data utility based on the number of quasi-identifier attributes.*

of datasets decreases. The cause of decreasing the data utility is the larger size of the indistinguishable quasi-identifier groups, i.e., when the number of quasi-identifier attributes is increased, the size of the indistinguishable quasi-identifier groups is also increased. However, the number of quasi-identifier attributes has less effect on the proposed models because the proposed model is not based on the group of indistinguishable quasi-identifier values.

4.2.3 The data utility based on the number of sensitive attributes

The third experiment is proposed to evaluate the effect of the number of sensitive attributes that influence the data utility of datasets that are constructed by k-Likeness, l-Diversity, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, the proposed model. For experiments, the parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-Privacy, the parameter C of the proposed model is set to be 4. In addition, the parameter R^{a_x} of the proposed model is not defined. Moreover, the number of sensitive attributes is varied from 1 to 4.

From the experimental results shown in Figure 4, we observe that the number of sensitive attributes also has an effect on the data utility of datasets that are satisfied by k-Likeness, l-Diversity, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, and the proposed model. When the number of sensitive attributes increases, the data utility of datasets decreases. However, the number of sensitive attributes has less effect on the proposed models because it is not based on data partitions and data distortions.

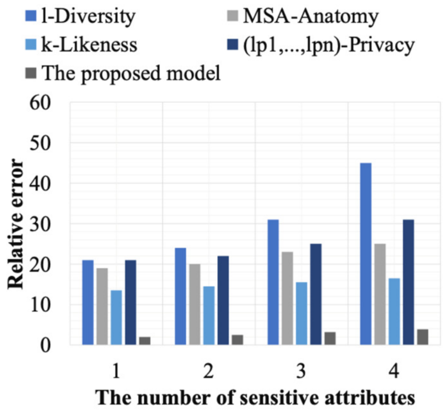

Fig.4: *The data utility based on the number of sensitive attributes.*

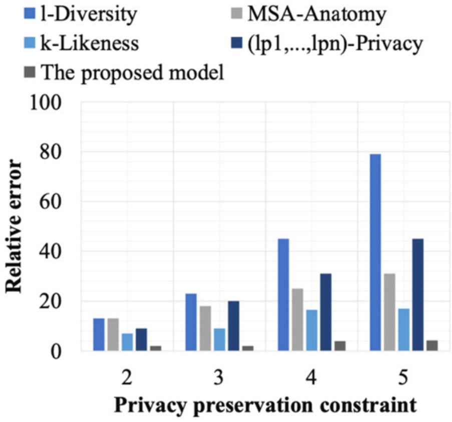

Fig.5: *The data utility based on privacy preservation constraints.*

4.2.4 The data utility based on privacy preservation constraints

The fourth experiment is proposed to evaluate the effect of the privacy preservation constraints of k-Likeness, l-Diversity, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, and the proposed model that influence the data utility of datasets. For experiments, the parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-

Privacy, the parameter C of the proposed model is varied from 2 to 5. In addition, the parameter R^{a_x} of the proposed model is not defined.

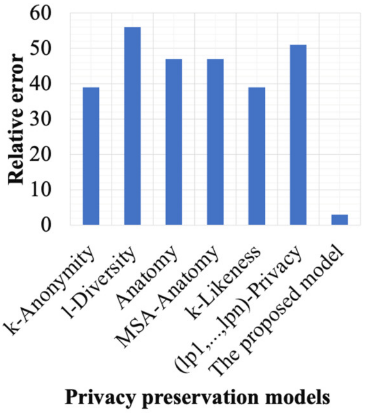

Fig.6: *The query results based on quasi-identifier attributes.*

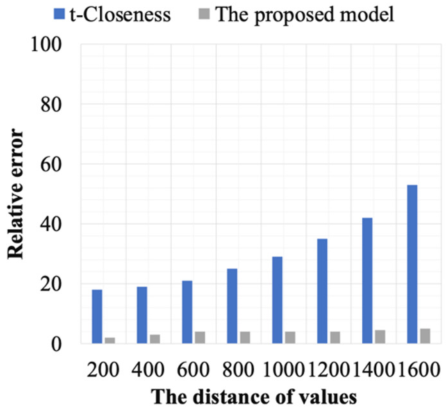

Fig.7: *The data utility based on the distance of values.*

From the experimental results shown in Figure 5, we observe that when the level of privacy preservation is increased, the data utility of datasets is decreased. Straightforwardly, the data utility and privacy are traded off. Moreover, the proposed model is more effective than the compared models. A cause of the effectiveness of the proposed model is that each query result of data queries is independently considered. Contrastively, every group of indistinguishable values is available in the datasets of the compared models, it must be constructed before it will be released and provided for public use.

4.2.5 The query results based on quasi-identifier attributes

The fifth experiment is proposed to evaluate the data utility in the part of the quasi-identifier attributes of datasets that are satisfied by k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \dots, l^{pn})$-Privacy, and the proposed model. For experiments, only Capital-loss is the sensitive attribute of the experimental datasets. The parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \dots, l^{pn})$-Privacy, the parameter C of the proposed model is varied from 2 to 5. In addition, the parameter R^{a_x} of the proposed model is not.

From the experimental results shown in Figure 6, we observe that the quasi-identifier values have more effect on the compared models than the proposed model because the compared models are based on data distortions and the attributes of datasets are grouped to be the quasi-identifier attributes and the sensitive attribute(s).

4.2.6 The data utility based on the distance of values

The sixth experiment is proposed to evaluate the data utility of datasets that are constructed by t-Closeness and the proposed model. For experiments, only Capital-loss is set to be the sensitive attribute. In addition, the parameter C of the proposed model is not defined. Moreover, the parameter t and $R^{Capital-loss}$ are varied from 200 to 1600.

From the experimental results shown in Figure 7, we can observe that the value of t and $R^{Capital-loss}$ influence the data utility of datasets. Moreover, the experimental results indicate that the proposed model is more effective than t-Closeness. A cause of the effectiveness of the proposed model is that the attributes of datasets are not grouped to be quasi-identifier attributes and sensitive attributes. Moreover, the data is available in dataset, it is not distorted but it is deprived of privacy violation issues from using data ignoration techniques.

4.2.7 The data utility based on the AND query operator

The seventh experiment is proposed to evaluate the effect of the AND query operator that influences the data utility of datasets of k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \dots, l^{pn})$-Privacy, and the proposed model. For experiments, the parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \dots, l^{pn})$-Privacy, and the parameter C of the proposed model are set to be 4. In addition, the parameter R^{a_x} of the proposed model is not Moreover, the number of quasi-identifier attributes is varied from 1 to 6. defined.

From the experimental results shown in Figure 8, we can observe that the number of the query condition attributes (the quasi-identifier attributes) has more effect on the compared models. When the number of query condition attributes increases, the data

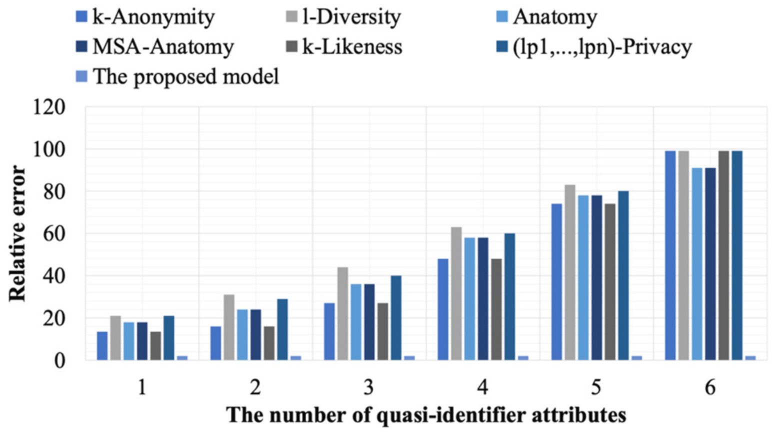

Fig.8: *The data utility based on the AND query operator.*

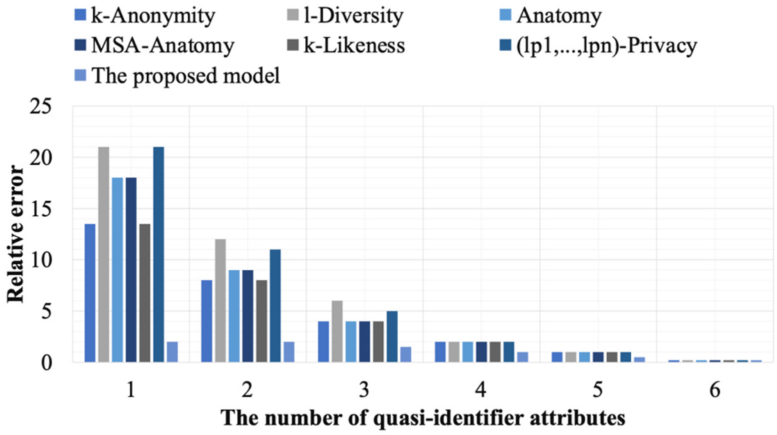

Fig.9: *The data utility based on the OR query operator.*

utility of datasets based on the AND query operator decreases. A cause of the ineffectiveness of the compared models is that when the number of query condition attributes increases, the number of fake values in datasets also increases.

4.2.8 The data utility based on the OR query operator

The eighth experiment is proposed to evaluate the effect of the OR query operator that influences the data utility of datasets that are satisfied by k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \dots, l^{pn})$-Privacy, and the proposed model. All experiments in this section are set up to be the same as the experimental setups of Section 4.2.7.

From the experimental results shown in Figure 9, we can observe that the number of query condition attributes can affect the query result of the OR query operator. When the number of query condition attributes is increased, the data utility of datasets is also increased. Moreover, all experiments indicate that the proposed model is more effective than the compared models. A cause of increasing the data utility of datasets is when increasing the number of query condition attributes, the values in datasets can have more opportunity to be the query results. Therefore, fewer query result errors can be obtained.

4.2.9 The data utility based on the range of query conditions

The ninth experiment is proposed to evaluate the effect of the range of query conditions that influence the data utility of datasets that are satisfied by k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \dots, l^{pn})$-Privacy, and the proposed model. All experiments in this section are set up to be the same as the experimental setups of Section 4.2.7.

From the experimental results shown in Figure 10, we observe that the range of query conditions has an effect on the datasets of k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, and $(l^{p1}, \dots, l^{pn})$-Privacy, and the proposed model. When the range of query conditions is increased, the data utility of datasets is also increased. Moreover, the experiments indicate that the proposed model is more effective than the compared models. Also, a cause of more effectiveness is when increasing the range of query conditions. It is the opportunity of values that

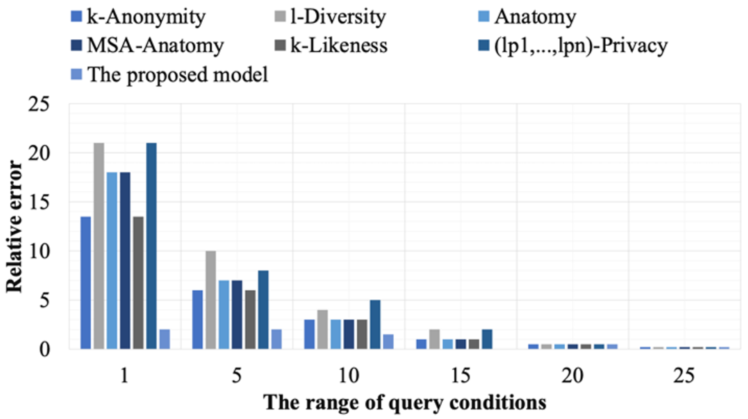

Fig.10: *The data utility based on the range of query conditions.*

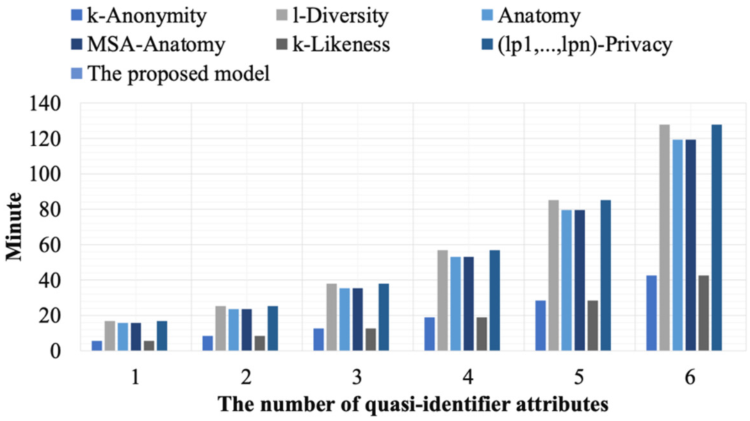

Fig.11: *The execution time based on the size of datasets.*

can be the query results.

4.3 Efficiency

In the section, the experiments for evaluating the efficiency of the proposed model are proposed.

4.3.1 The execution time based on the size of datasets

The tenth experiment is proposed to evaluate the effect of dataset sizes that influence the execution time for transforming datasets to satisfy k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, and the proposed model. For experiments, Capital-loss is set to be the sensitive attribute. Moreover, 100 tuples of the experimental datasets are randomly selected to be the initial tuples for the experiments, thereafter, 200 tuples are randomly increased for each experiment until the experimental dataset collects 1700 tuples. The parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter lpx, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-Privacy, and the parameter C of the proposed model are set to be 4. In addition, the parameter R^{a_x} of the proposed model is not defined.

From the experimental results shown in Figure 11, we observe that the number of tuples directly influences the execution time for transforming datasets to the satisfaction of the privacy preservation constraint of the compared models, i.e., when the tuples of datasets are increased, the execution time for transforming datasets is also increased. Straightforwardly, every tuple generally uses a transformed time to satisfy privacy preservation constraints. Contrastively, the proposed model does not use any execution time for transforming datasets because datasets can be provided through the proposed model after all explicit identifier values of users are removed. That is, the datasets of the proposed model are not distorted, they can be deprived of privacy violation issues by using data ignoration techniques.

4.3.2 The execution time based on the number of quasi-identifier attributes

The eleventh experiment is proposed to evaluate the effect of the number of quasi-identifier attributes that influence the execution time for transforming datasets to satisfy k-Anonymity, k-Likeness, l-Diversity, Anatomy, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, and the proposed model. For experiments, only Capital-loss is the sensitive attribute of the ex-

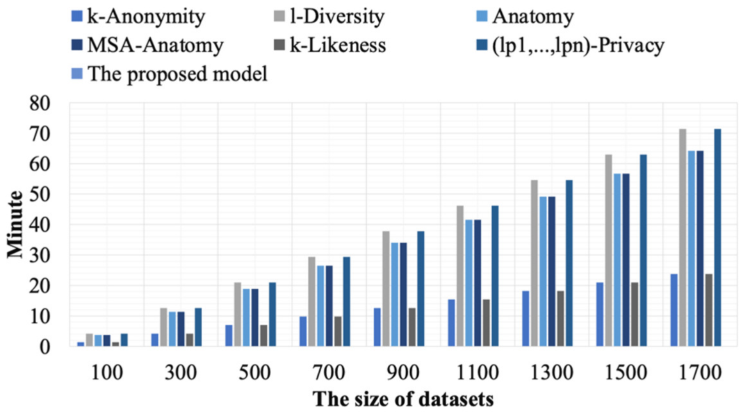

Fig.12: *The execution time based on the number of quasi-identifier attributes.*

perimental datasets. The parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter lpx, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-Privacy, and the parameter C of the proposed model is set to be 4. In addition, the parameter R^{a_x} of the proposed model is not defined. Moreover, the number of quasi-identifier attributes varies from 1 to 6.

From the experimental results shown in Figure 12, we observe that the number of quasi-identifier attributes affects the execution time for transforming datasets to satisfy the privacy preservation constraint of the compared models. Straightforwardly, the number of quasi-identifier attributes directly influences the quasi-identifier data dimension and the number of quasi-identifier values that must be considered to satisfy privacy preservation constraints. Also, the proposed model does not use any execution time for transforming datasets because datasets can be provided through the proposed model after all explicit identifier values of users are removed. That is, the datasets of the proposed model are not distorted, but they are deprived of privacy violation issues by using data ignoration techniques.

4.3.3 The execution time based on the number of sensitive attributes

The twelfth experiment is proposed to evaluate the effect of the number of sensitive attributes that influence the execution time for transforming datasets to satisfy k-Likeness, l-Diversity, MSA-Anatomy, $(l^{p1}, \ldots, l^{pn})$-Privacy, the proposed model. For experiments, the parameter k of k-Anonymity and k-Likeness, the parameter l of l-Diversity, Anatomy, and MSA-Anatomy, the parameter l^{px}, where $1 \leq x \leq n$, for $(l^{p1}, \ldots, l^{pn})$-Privacy, and the parameter C of the proposed model is set to be 4. In addition, the parameter R^{a_x} of the proposed model is not defined. Moreover, the number of sensitive attributes varies from 1 to 4.

From the experimental results shown in Figure 13, we can observe that the number of sensitive attributes affects the execution time for transforming datasets to satisfy the privacy preservation constraint of the compared models. Straightforwardly, the number of sensitive attributes directly influences the search space of the sensitive values that must be considered to satisfy privacy preservation constraints. Also, the proposed model does not use any execution time for transforming datasets because datasets can be provided through the proposed model after the explicit identifier values of users are removed. That is, the datasets of the proposed model are not distorted, but they are deprived of privacy violation issues by using data ignoration techniques.

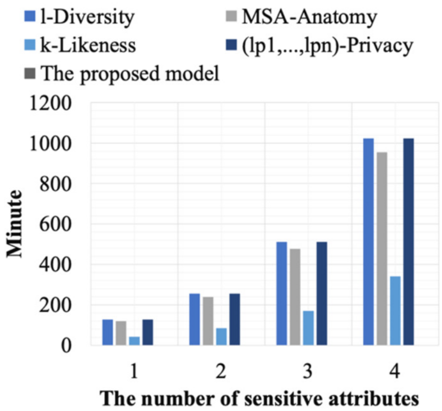

Fig.13: *The execution time based on the number of sensitive attributes.*

5. CONCLUSION

In this work, a privacy preservation model, $C^{R_{a1},\ldots,R_{an}}$-Privacy, is proposed. It can guarantee the confidence of the range and the number of values that can be re-identified by the adversary. Moreover, it can address privacy violation issues in dynamic datasets (datasets are allowed to change the data when new data become available) and big data analytics. Aside from privacy preservation constraints, the complexity of transformation processes and data utility are also considered in the proposed model. To achieve the aims of the proposed model, the data ignoration technique is applied, i.e., every query result cannot be satisfied by C and R^{a_x}, it is ignored. Moreover, the experimental results indicate that the proposed model is a more effective and efficient privacy preservation model.

6. FUTURE WORK

Although the proposed model can address privacy violation issues in big data analytics, an adversary will discover a new privacy violation approach that can be used to attack the privacy data that is available in big data analytics in the future. Thus, an appropriate privacy preservation model that can address the newly discovered privacy violation issue should also be proposed.

References

[1] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, *et al. Challenges and opportunities with big data 2011-1.* 2011.

[2] T. H. Davenport, P. Barth and R. Bean, *How'big data'is different*, 2012.

[3] Z. Zheng, J. Zhu and M. R. Lyu, "Service-generated big data and big data-as-a-service: an overview," in *2013 IEEE international congress on Big Data*, pp. 403–410, 2013.

[4] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 international conference on collaboration technologies and systems (CTS)*, pp. 42–47, 2013.

[5] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97– 107, 2013.

[6] M. Chen, S. Mao and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.

[7] A. Mohamed, M. K. Najafabadi, Y. B. Wah, E. A. K. Zaman and R. Maskat, "The state of the art and taxonomy of big data analytics: view from new big data framework," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 989–1037, 2020.

[8] Y. Cui, S. Kara and K. C. Chan, "Manufacturing big data ecosystem: A systematic literature review," *Robotics and computer-integrated Manufacturing*, vol. 62, no. 0101861, 2020.

[9] W. Haoxiang, *et al.* "Big data analysis and perturbation using data mining algorithm, *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 1, pp. 19–28, 2021.

[10] J. Wang, C. Xu, J. Zhang and R. Zhong, "Big data analytics for intelligent manufacturing systems: A review," *Journal of Manufacturing Systems*, vol. 62, pp. 738–752, 2022.

[11] M. Naeem, *et al.* "Trends and future perspective challenges in big data," in *Advances in intelligent data analysis and applications*, pp. 309–325. Springer, 2022.

[12] B. K. Chan, "Data analysis using r programming," in *Biostatistics for Human Genetic Epidemiology*, pp. 47–122. Springer, 2018.

[13] E. Kaya, M. Agca, F. Adiguzel and M. Cetin, "Spatial data analysis with r programming for environment," *Human and ecological risk assessment: An International Journal*, vol. 25, no. 6, pp. 1521– 1530, 2019.

[14] J. Nandimath, E. Banerjee, A. Patil, P. Kakade, S. Vaidya and D. Chaturvedi, "Big data analysis using apache hadoop," in *2013 IEEE 14th International Conference on In- formation Reuse & Integration (IRI)*, pp. 700–703, 2013.

[15] O. Azeroual and R. Fabre, "Processing big data with apache hadoop in the current challenging era of covid-19," *Big Data and Cognitive Computing*, vol. 5, no.1:12, 2021.

[16] E. Nazari, M. H. Shahriari and H. Tabesh, "Big-data analysis in healthcare: apache hadoop, apache spark and apache flink," *Frontiers in Health Informatics*, vol. 8, no. 1:14, 2019.

[17] M. Hofmann and R. Klinkenberg, "RapidMiner: Data mining use cases and business analytics applications," *CRC Press*, 2016.

[18] V. Kotu and B. Deshpande, "Predictive anaytics and data mining: concepts and practice with rapidminer," *Morgan Kaufmann*, 2014.

[19] A. Boranbayev, G. Shuitenov and S. Boranbayev, "The method of analysis of data from social networks using rapidminer, in *Science and Information Conference*, pp. 667–673. Springer, 2020.

[20] M. Copeland, J. Soh, A. Puca, M. Manning and D. Gollob, "Microsoft azure," New York, NY, USA:: Apress, pp. 3–26, 2015.

[21] Ro. Barga, V. Fontama, W. H. Tok and L. Cabrera-Cordon, "Predictive analytics with Microsoft Azure machine learning," Springer, 2015.

[22] B. Gupta, P. Mittal and T. Mufti, "A review on amazon web service (aws), microsoft azure & google cloud platform (gcp) services," in *Proceedings of the 2nd International Conference on ICT for Dig- ital, Smart, and Sustainable Development, ICIDSSD 2020*, 27-28 February 2020, Jamia Hamdard, New Delhi, India, 2021.

[23] A. K. Sandhu, "Big data with cloud computing: Discussions and challenges," *Big Data Mining and Analytics*, vol. 5, no. 1, pp. 32–40, 2021.

[24] Y. Gahi, M. Guennoun and H. T. Mouftah, "Big data analytics: Security and privacy challenges," in *2016 IEEE Symposium on Computers and Communication (ISCC)*, pp. 952–957, 2016.

[25] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao and Z. Huo, "Privacy Preservation in Big Data From

the Communication Perspective—A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 753-778, Firstquarter 2019.

[26] Q. Zhang, L. T. Yang and Z. Chen, "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning," in *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1351-1362, 1 May 2016.

[27] D. S. Terzi, R. Terzi and S. Sagiroglu, "A survey on security and privacy issues in big data," in *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pp. 202–207, 2015.

[28] O. Tene and J. Polonetsky, "Big data for all: Privacy and user control in the age of analytics," *Northwestern Journal of Technology and Intellectual Property*, vol. 11, no. 5, pp. 240-273, 2012.

[29] Z. Lv and L. Qiao, "Analysis of health-care big data," *Future Generation Computer Systems*, vol. 109, pp. 103–110, 2020.

[30] H.-N. Dai, H. Wang, G. Xu, J. Wan and M. Imran, "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies," *Enterprise Information Systems*, vol. 14, no. 9-10, pp. 1279–1303, 2020.

[31] A. D. Dwivedi, G. Srivastava, S. Dhar and R. Singh, "A decentralized privacy- preserving healthcare blockchain for iot," *Sensors*, vol. 19, no. 2:326, 2019.

[32] E. Bertino and E. Ferrari, "Big data security and privacy," in *A comprehensive guide through the Italian database research over the last 25 years*, pp. 425–439. Springer, 2018.

[33] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature medicine*, vol. 25, no. 1, pp. 37– 43, 2019.

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

[35] A. Machanava jjhala, 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.

[36] 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.

[37] S. Riyana, N. Harnsamut, T. Soontornphand and J. Natwichai, "(k, e)-Anonymous for Ordinal Data," *2015 18th International Conference on Network-Based Information Systems*, Taipei, Taiwan, pp. 489-493, 2015.

[38] S. Riyana, N. Riyana and S. Nanthachumphu, " Enhanced (k, e)-anonymous for categorical data," in *Proceedings of the 6th International Conference on Software and Computer Applications*, pp. 62–67, 2017.

[39] S. Riyana, S. Nanthachumphu and N. Riyana, "Achieving privacy preservation constraints in missing-value datasets," *SN Computer Science*, vol. 1, no. 4, pp. 1–10, 2020.

[40] S. Riyana, N. Riyana and S. Nanthachumphu, "An effective and efficient heuris- tic privacy preservation algorithm for decremental anonymization datasets," in *International Conference on Image Processing and Capsule Networks*, pp. 244–257. Springer, Cham, 2020.

[41] N. Riyana, Surapon Riyana, S. Nanthachumphu, S. Sittisung and D. Duangban, "Privacy violation issues in republication of modification datasets," in *International Conference on Intelligent Computing & Optimization*, pp. 938–953. Springer, Cham, 2021.

[42] S. Riyana, N. Riyana, and S. Nanthachumphu, "Privacy preservation techniques for sequential data releasing," in *The 12th International Conference on Advances in Information Technology*, pp. 1–9, 2021.

[43] S. Riyana and N. Riyana, "A privacy preservation model for rfid data-collections is highly secure and more efficient than lkc-privacy," in *The 12th International Conference on Advances in Information Technology*, pp. 1–11, 2021.

[44] S. Riyana, *Privacy preservation models for the independent data release of high-dimensional datasets*, 2023.

[45] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," *VLDB '06*, pp. 139–150. VLDB Endowment, 2006.

[46] S. Riyana, N. Riyana and W. Sujinda, "An anatomization model for farmer data collections," *SN Computer Science*, vol. 2, no. 5, pp. 1–11, 2021.

[47] S. Riyana and N. Riyana, "Achieving anonymization constraints in high-dimensional data publishing based on local and global data suppressions," *SN Computer Science*, vol. 3, no. 1, pp. 1–12, 2022.

[48] S. Riyana, N. Ito, T. Chaiya, U. Sriwichai, N. Dussadee, T. Chaichana, R. Assawarachan, T. Maneechukate, S. Tantikul and N. Riyana, "Privacy threats and privacy preservation techniques for farmer data collections based on data shuffling," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 16, no. 3, pp. 289–301, 2022.

[49] N. Ramakrishnan, B. J. Keller, B. J. Mirza, A. Y. Grama and G. Karypis, "Privacy risks in recommender systems," in *IEEE Internet Computing*, vol. 5, no. 6, pp. 54-63, Nov.-Dec. 2001.

[50] S. Riyana and J. Natwichai, "Privacy preservation for recommendation databases," *Service Oriented Computing and Applications*, vol. 12, no.3, pp. 259– 273, 2018.

[51] S. Riyana, "$(l^{p1}, \ldots, l^{pn})$-privacy: privacy preservation models for numerical quasi-identifiers and multiple sensitive attributes," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2021.

[52] S. Riyana, K. Sasujit, N. Homdoung, T. Chaichana and T. Punsaensri, "Effective privacy preservation models for rating datasets," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 17, no. 1, pp. 1– 13, 2023.

[53] H. Liang and H. Yuan, "On the complexity of t-closeness anonymization and related problems," in *International Conference on Database Systems for Advanced Applications*, pp. 331–345. Springer, 2013.

[54] R. C.-W. Wong, J. Li, A. W.-C. Fu and K. Wang, "(α, 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*, New York, NY, USA, , pp. 754–759, 2006.

[55] 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.

[56] 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.

[57] S. Riyana and N. Riyana, *Simple, effective, and efficient privacy preservation models for rating datasets*, 2023.

[58] Q. Zhang, N. Koudas, D. Srivastava and T. Yu, "Aggregate query answering on anonymized tables," in *2007 IEEE 23rd International Conference on Data Engineering*, pp. 116–125, April 2007.

[59] 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*, pp. 202–207. AAAI Press, 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.