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 Anatomization Constraints in Dynamic Datasets

Surapon Riyana[1]

ABSTRACT

Anatomy is one of the well-known privacy preservation models that are proposed to address privacy violation issues in released datasets. Unfortunately, we found that Anatomy is often sufficient to address privacy violation issues in datasets that are focused on performing a time of data releases. Thus, if datasets are dynamic (i.e., the data is updated when new data becomes available) and they are independently released, Anatomy can be insufficient. That is, released datasets are satisfied by Anatomy constraints, they still have privacy violation issues from data comparison attacks such as $iFRCA$, $iMRCA$, $iMRcMLA$, $dFLCA$, $dMLCA$, $dMLcMRA$, SVM, and partition changing such that they are presented in this work. To address these privacy violation issues in released datasets, a new privacy preservation model is proposed in this work. Furthermore, we show that the proposed model is higher secure in terms of privacy preservation than Anatomy with extensive experiments.

Article information:

Keywords: Privacy Preservation Models, Anatomization Models, Dynamic Datasets, and Sequential Datasets

Article history:
Received: June 28, 2022
Revised: October 8, 2022
Accepted: November 6, 2022
Published: Februry 18, 2023
(Online)

DOI: 10.37936/ecti-cit.2023171.248877

1. INTRODUCTION

Data utility and data privacy are both aspects that must be considered when datasets are released for public use [14] [16] [19] [20] [23] [24] [26]. For this reason, data anatomization models [4] [28] have been proposed. The privacy preservation idea of anatomization models is that before datasets will be released for public use, the tuples of them are partitioned such that every partition must include l distinct sensitive values, where $l \geq 2$. Moreover, the tuples of every partition are anatomized to be the quasi-identifier table and the sensitive table. The relationship of each partition in the anatomized tables is its defined partition identifier. Data anatomization models are generally proposed to address privacy violation issues in datasets that are focused on performing a time of data releases. For this reason, if datasets are dynamic (i.e., the data is updated when new data becomes available) and they are independently released, Anatomy can be insufficient. That is, released datasets are satisfied by Anatomy constraints, they still have privacy violation issues because they are vulnerable to privacy data attacks that are presented in Section 3. To rid this vulnerability of anatomization models, a new anatomization model for dynamic datasets is proposed in this work, it will be presented in Section 4.2.

The organization of this work is as follows. At first, related works are described in Section 2. Then, the vulnerabilities of anatomization models are identified in Section 3. Then, the basic definitions and the proposed algorithm are presented in Section 4. Then, the experimental results are discussed in Section 5. Then, the conclusion of this work is given in Section 6. Finally, possible future works for this work are presented in Section 7.

2. RELATED WORK

k-Anonymity [26] is a well-known privacy preservation model. For privacy preservation, the attributes of datasets are grouped to be the explicit identifier attributes, the quasi-identifier attributes, and a sensitive attribute. Then, all explicit identifier values of users are removed. Finally, all unique quasi-identifier values are generalized or suppressed to be at least k indistinguishable tuples. Every group of indistinguishable quasi-identifier tuples calls an equivalence class, EC, of released datasets.

For example, let Table 1 be the specified raw dataset such that SSN and $Name$ are the explicit identifier attributes. Age, Sex, and $Zipcode$ are the quasi-identifier attributes. $Disease$ is the sensitive attribute. Let the value of k be set to 2. To achieve 2-Anonymity constraints in Table 1, SSN and $Name$

[1] The author is with Maejo University, Sansai, Chiangmai, Thailand, 50290, E-mail: surapon_r@mju.ac.th

Table 1: *An example of raw datasets.*

SSN	Name	Age	Sex	Zipcode	Disease
000-00-0001	Jacob	45	Male	60636	Flu
000-00-0002	Thomas	46	Male	60632	Fever
000-00-0003	John	47	Male	60635	Cancer
000-00-0004	David	48	Male	60639	Cancer
000-00-0005	Amelia	48	Female	60632	Cancer
000-00-0006	Sophia	42	Female	60632	HIV
000-00-0007	Isabella	42	Female	60632	Fever
000-00-0008	Emily	41	Female	60636	Cancer
000-00-0009	Olivia	40	Female	60636	HIV
000-00-0010	Victoria	39	Female	60636	HIV
000-00-0011	Jessica	38	Female	60639	Fever
000-00-0012	Jennifer	37	Female	60639	Cancer

Table 2: *The data version of Table 1 without the explicit identifier attributes.*

Age	Sex	Zipcode	Disease
45	Male	60636	Flu
46	Male	60632	Fever
47	Male	60635	Cancer
48	Male	60639	Cancer
48	Female	60632	Cancer
42	Female	60632	HIV
42	Female	60632	Fever
41	Female	60636	Cancer
40	Female	60636	HIV
39	Female	60636	HIV
38	Female	60639	Fever
37	Female	60639	Cancer

Table 3: *A released data version of Table 1 satisfies 2-Anonymity constraints.*

Age	Sex	Zipcode	Disease	EC
45 - 46	Male	6063*	Flu	1
45 - 46	Male	6063*	Fever	
47 - 48	Male	6063*	Cancer	2
47 - 48	Male	6063*	Cancer	
42 - 48	Female	60632	Cancer	3
42 - 48	Female	60632	HIV	
42 - 48	Female	60632	Fever	
39 - 41	Female	60636	Cancer	4
39 - 41	Female	60636	HIV	
39 - 41	Female	60636	HIV	
37 - 38	Female	60639	Fever	5
37 - 38	Female	60639	Cancer	

are removed, i.e., Table 2 is the output of this step. Finally, *Age*, *Sex*, and *Zipcode* are generalized by their less specific values to create at least two indistinguishable tuples. Therefore, Table 3 is a released data version of Table 1 such that it is satisfied by 2-Anonymity constraints. With Table 3, we can see that all possibly re-identified conditions through the quasi-identifier attributes always have at least two tuples to be satisfied. For this reason, Table 3 does not seem to have any privacy violation issue. Unfortunately, in [11], the authors demonstrate that Table 3 still has privacy violation issues that must be addressed. If the adversary has enough background knowledge about the target user and can match his/her background knowledge to an *EC* of Table 3 such that the sensitive value is available in the matched *EC* to be homogeneous, the privacy data of the target user can be violated by the adversary.

To show an example of the vulnerability of k-Anonymity constraints, we suppose that *David* is the target user of the adversary. We suppose that the adversary knows that *David* is a male person who is 48 years old. Moreover, the adversary strongly believes that *David*'s profile tuple is available in Table 3. In this situation, the adversary can be highly confident that a tuple of *EC* 2 of Table 3 is *David*'s profile tuple because only this *EC* of Table 3 can match to the adversary's background knowledge about *David*. Moreover, the adversary can observe that the sensitive attribute of the matched *EC* only collects *Cancer*. Therefore, the adversary can infer that *David*'s disease is *Cancer*. From this example, it is clear that although released datasets can guarantee that all possibly re-identified conditions always have at least k tuples to be satisfied, they still have privacy violation issues that must be addressed. To rid this vulnerability of k-Anonymity, in [11], l-Diversity is proposed. For privacy preservation with l-Diversity, aside from removing all explicit identifier values and distorting (generalizing or suppressing) all unique quasi-identifier values, the number of distinct sensitive values in each *EC* is also considered, i.e., every *EC* of released datasets is further required to have at least l different sensitive values.

As an example of privacy preservation based on l-Diversity, let Table 2 be the specified raw dataset for public use. Let the value of l be set to 3. With these instances, Table 4 is a data version of Table 1 that is satisfied by 3-Diversity constraints. If the adversary tries to reveal the disease of users from Table 4, the adversary always sees that every possibly re-identified

Table 4: *A released data version of Table 1 satisfies by 3-Diversity constraints.*

Age	Sex	Zipcode	Disease	EC
45 - 48	Male	6063*	Flu	1
45 - 48	Male	6063*	Fever	
45 - 48	Male	6063*	Cancer	
45 - 48	Male	6063*	Cancer	
42 - 48	Female	60632	Cancer	2
42 - 48	Female	60632	HIV	
42 - 48	Female	60632	Fever	
37 - 41	Female	6063*	Cancer	3
37 - 41	Female	6063*	HIV	
37 - 41	Female	6063*	HIV	
37 - 41	Female	6063*	Fever	
37 - 41	Female	6063*	Cancer	

Table 5: *A partitioned data version of Table 2, where $l = 3$.*

Age	Sex	Zipcode	Disease	PID
45	Male	60636	Flu	1
46	Male	60632	Fever	
47	Male	60635	Cancer	
48	Male	60639	Cancer	
48	Female	60632	Cancer	2
42	Female	60632	HIV	
42	Female	60632	Fever	
41	Female	60636	Cancer	3
40	Female	60636	HIV	
39	Female	60636	HIV	
38	Female	60639	Fever	
37	Female	60639	Cancer	

condition has at least three different sensitive values that are satisfied. For this situation, we can conclude that if released datasets satisfy l-Diversity constraints, they can guarantee that all possibly re-identified conditions always have at least l different sensitive values that are satisfied.

Aside from k-Anonymity and l-Diversity, there are other well-known anonymization models that are available such as t-Closeness [10], (α, k)-Anonymity [15] [27], and k-Likeness [18]. However, to the best of our knowledge about anonymization models, all of them still could be insufficient to address privacy violation issues in dynamic datasets [1] [7] [13] [15] [21] [22] [30].

In [1] and [30], the authors show privacy violation scenarios in released datasets that allow adding new tuples. Moreover, in [15] and [21], the authors demonstrate that released datasets are in privacy violation scenarios when the data of released datasets is deleted. In [13], the authors illustrate scenarios in that released datasets are allowed to update the data and are independently released, they also have privacy violation issues that must be addressed. Furthermore, in [22], the authors emphasize that released datasets of dynamic datasets still have privacy violation issues from data comparison that must be addressed. To address these privacy violation issues in released datasets, in [1], [13], [15], [21], [22], and [30], the authors suggest that aside from requiring released datasets to satisfy privacy preservation constraints, all possible data comparative results between the released dataset and its previously related release datasets must also be satisfied by privacy preservation constraints.

However, anonymization models still have a serious vulnerability that must be improved. They often have data utility issues [28] in released datasets. To rid this vulnerability of anonymization models, data anatomization models [4] [28] were proposed. For privacy preservation, before datasets are released, all users' explicit identifier values are removed. Then, the tuples are partitioned such that every partition must include at least l distinct sensitive values. Finally, every partition is anatomized to be the quasi-identifier table and the sensitive table such that the relationship of each partition in the anatomized tables is presented by its defined identifier PID.

We now consider an example of privacy preservation based on data anatomization constraints. Let Table 2 be the specified raw dataset for public use. We suppose that the value of l is set to 3. For privacy preservation, the tuples of Table 2 are partitioned by the given value of l in the first step such that every partition must include at least three distinct sensitive values. Moreover, the identifier, PID, of each partition is defined by this step. Table 5 is the output of the first step. Finally, the tuples of each partition are anatomized to be the quasi-identifier table and the sensitive table such that they are shown in Tables 6 and 7 respectively. The relationship of partitions in Tables 6 and 7 is their defined PID. For this reason, Tables 6 and 7 can guarantee that all possibly re-identified conditions always have at least l distinct sensitive values that are satisfied. Moreover, we can see that Tables 6 and 7 have better data utility than Table 4. Suppose that "$Age = 45$ *and* $Sex = Female$" is the specified query condition. With Table 4, we can see that there are three tuples that are satisfied. But Table 1 (the raw table) and the query result of data joining between Table 6 and 7 does not have any tuple which matches the query condition.

From the example, it is clear that data anatomization models can be more effective than other data anonymization models. Data anatomization models are generally proposed to address privacy violation issues in datasets that are focused on performing a time of data releases. For this reason, if datasets are dynamic and independently released, data anatomization models can be insufficient because they still have privacy violation issues that must be addressed. That is, data anatomization models are vulnerable to attacks from using data comparison that will be pre-

Table 6: *The quasi-identifier table of Table 2, where $l = 3$.*

Age	Sex	Zipcode	PID
45	Male	60636	1
46	Male	60632	
47	Male	60635	
48	Male	60639	
48	Female	60632	2
42	Female	60632	
42	Female	60632	
41	Female	60636	3
40	Female	60636	
39	Female	60636	
38	Female	60639	
37	Female	60639	

Table 7: *The sensitive table of Table 2, where $l = 3$.*

Disease	PID
Flu	1
Fever	
Cancer	
Cancer	
Cancer	2
HIV	
Fever	
Cancer	3
HIV	
HIV	
Fever	
Cancer	

sented in Section 3. To rid the vulnerabilities of data anonymization models in dynamic datasets, a new data anatomization model is proposed in Section 4.

3. MOTIVATION

Before privacy violation issues in anatomization tables are presented, the significant basic definitions of several terms are given.

Definition 1 (Raw dataset) Let QI be the set of quasi-identifier attributes. Let S be a sensitive attribute. Let D be the raw dataset such that every tuple $d_r \in D$ is constructed from $QI \cup S$. Let D^1, D^2, $\ldots$, D^{z-1}, and D^z be the data versions of D at the timestamps 1, 2, $\ldots$, $z-1$, and z respectively. Let $D^j[QI]$ be the data projection over on QI of D^j, where $1 \leq j \leq z$. Let $D^j[S]$ be the data projection over on S of D^j. Moreover, let $d_r^j[QI]$ and $d_r^j[S]$ be the data projection over on QI and S of d_r^j in D^j respectively.

Definition 2 (Data anatomization [28]] Let a positive integer l, where $l \geq 2$, be the privacy preservation constraint. Let $f_{Ana}(D^j, l) : D^j \rightarrow_l D^j_{QI}, D^j_S$ be the anatomization function for transforming D^j to become D^j_{QI} and D^j_S. That is, the tuples of D^j are partitioned to be par^j_1, par^j_2, $\ldots$, par^j_x, where $\cup^x_{g=1} par^j_g = D^j$ and $\cap^x_{g=1} par^j_g = \emptyset$. Moreover, each partition par^j_g must include at least l distinct sensitive values. $1, 2, \ldots, x$ are the partition identifier, PID. Furthermore, par^j_1, par^j_2, $\ldots$, par^j_x are anatomized to be both tables, i.e., the quasi-identifier table D^j_{QI} and the sensitive table D^j_S such that each related partition of D^j_{QI} and D^j_S is connected by its PID.

Definition 3 (Adversary background knowledge) Let u^j_r be the target user of the adversary in D^j such that the tuple d^j_r of D^j is the profile tuple of the user u^j_r. Moreover, let $B_{u^j_r}$, where $B_{u^j_r} \subseteq d^j_r[QI]$ and $d^j_r \in D^j$, be the adversary's background knowledge about the target user u^j_r in D^j. If $B_{u^j_r}$ is unique in the anatomization tables of D^j, the adversary can use $B_{u^j_r}$ to identify the profile tuple of u^j_r and reveal the sensitive value of u^j_r from the anatomization tables of D^j.

Aside from the basic definitions, to aid the readability of the focused issues in anatomization tables, the reader is supplied with the notation symbol used in Table 8.

3.1 Privacy violation issues in anatomization tables based on data increasing

In this section, we demonstrate privacy violation issues that could occur in anatomization tables when new data is added to them.

3.1.1 Full Right Coverage data Attack ($iFRCA$)

The assumption of this privacy data attack in anatomization tables is that only par^j_g of D^j matches the adversary's background knowledge about the target user. Moreover, there is only par^{j-1}_g of D^{j-1} such that it matches par^j_g. For violating the privacy data of the target user in anatomization tables, let u^j_r be the target user of the adversary. Let $B_{u^j_r}$ be the adversary's background knowledge about the target user u^j_r. Let $par^j_g[D^j_{QI}]$ collect the quasi-identifier tuple that matches $B_{u^j_r}$. Let $par^j_g[D^j_S]$ be the sensitive partition that relates to $par^j_g[D^j_{QI}]$. Moreover, let $par^{j-1}_g[D^{j-1}_{QI}]$ be the quasi-identifier partition that is fully covered by $par^j_g[D^j_{QI}]$, i.e., $par^{j-1}_g[D^{j-1}_{QI}] \subset par^j_g[D^j_{QI}]$. Let $par^{j-1}_g[D^{j-1}_S]$ be the sensitive partition that relates to $par^{j-1}_g[D^{j-1}_{QI}]$. If the compared result between $par^{j-1}_g[D^{j-1}_S]$ and $par^j_g[D^j_S]$ does not satisfy the given value of l, the privacy data of the target user u^j_r in $par^j_g[D^j_S]$ can be inferred from $par^j_g[D^j_S]$ $-$ $par^{j-1}_g[D^{j-1}_S]$. The characteristic of privacy violation issues in incremental anatomization tables from using $iFRCA$ attacks is shown in Figure 1.

Table 8: Summary of notation used.

Notation	Definition
QI	The quasi-identifier attributes
S	The sensitive attribute
D^j	The raw dataset at the timestamp j, where $1 \leq j \leq z$
d^j_r	An arbitrary tuple d^j_r of D^j
l	The privacy preservation constraint, where $l \in I^+$ and $l \geq 2$
D^j_{QI}	The quasi-identifier table of D^j
D^j_S	The sensitive table of D^j
par^j_g	An arbitrary partition par^j_g of D^j
$par^j_g[D^j_{QI}]$	An arbitrary partition par^j_g of D^j_{QI}
$par^j_g[D^j_S]$	An arbitrary partition par^j_g of D^j_S
PID	The partition identifier
$B_{u^j_r}$	The adversary's background knowledge in D^j

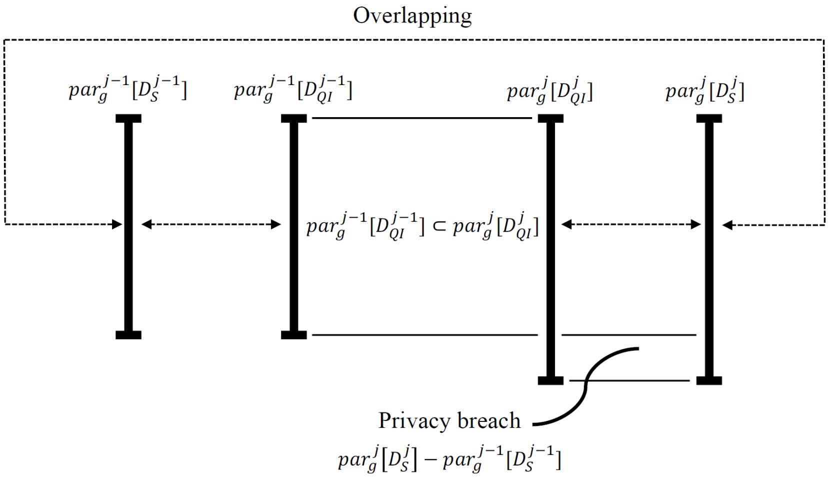
Overlapping
$par_g^{j-1}[D_S^{j-1}]$
$par_g^{j-1}[D_{QI}^{j-1}]$
$par_g^j[D_{QI}^j]$
$par_g^j[D_S^j]$
$par_g^{j-1}[D_{QI}^{j-1}] \subset par_g^j[D_{QI}^j]$
Privacy breach
$par_g^j[D_S^j] - par_g^{j-1}[D_S^{j-1}]$

Fig.1: *The characteristic of privacy violation issues in incremental anatomization tables from using iFRCA attacks.*

3.1.2 Merged Right Coverage data Attack (*iMRCA*)

The assumption of this privacy data attack in anatomization tables is that only par^j_g match the adversary's background knowledge about the target user and par^j_g only relates to par^{j-1}_g. Moreover, par^{j-1}_g relates to par^j_b, ..., par^j_c of D^j. For violating the privacy data of the target user in anatomization tables, let u^j_r be the target user of the adversary. Let $B_{u^j_r}$ be the adversary's background knowledge about the target user u^j_r. Let $par^j_a[D^j_{QI}]$ includes the quasi-identifier tuple which matches $B_{u^j_r}$. Moreover, let $par^j_b[D^j_{QI}]$, ..., $par^j_c[D^j_{QI}]$ be other specified quasi-identifier partitions of the adversary in D^j_{QI}. Let $par^{j-1}_g[D^{j-1}_{QI}]$ be the related quasi-identifier partition of $par^j_a[D^j_{QI}]$, $par^j_b[D^j_{QI}]$, ..., $par^j_c[D^j_{QI}]$ such that they satisfy the limitations as $(par^j_b[D^j_{QI}] \cup \ldots \cup par^j_c[D^j_{QI}]) \subset par^{j-1}_g[D^{j-1}_{QI}]$ and $par^{j-1}_g[D^{j-1}_{QI}] \subset (par^j_a[D^j_{QI}] \cup par^j_b[D^j_{QI}] \cup \ldots \cup par^j_c[D^j_{QI}])$. Let $par^j_a[D^j_S]$, $par^j_b[D^j_S]$, ..., $par^j_c[D^j_S]$ be the sensitive partitions that relate to $par^j_a[D^j_{QI}]$, $par^j_b[D^j_{QI}]$, ..., $par^j_c[D^j_{QI}]$ respectively. Also, let $par^{j-1}_g[D^{j-1}_S]$ be the sensitive partition that relates to $par^{j-1}_g[D^{j-1}_{QI}]$. If the compared result between $par^j_a[D^j_S]$, $par^j_b[D^j_S]$, ..., $par^j_c[D^j_S]$ and $par^{j-1}_g[D^{j-1}_S]$ does not satisfy the given value of l, the privacy data of the target user u^j_r in D^j_S can be inferred from $(par^j_a[D^j_S] \cup par^j_b[D^j_S] \cup \ldots \cup par^j_c[D^j_S]) - par^{j-1}_g[D^{j-1}_S]$. The characteristic of privacy violation issues in incremental anatomization tables from using *iMRCA* attacks is shown in Figure 2.

3.1.3 Merged Right fully covers Merged Left data Attack (*iMRcMLA*)

The assumption of this privacy data attack in anatomization tables is that par^j_g of D^j match the adversary's background knowledge about the target user and par^j_b, ..., par^j_c of D^j are identified by the adversary. Moreover, the adversary can see that par^j_g, par^j_b, ..., par^j_c relate to par^{j-1}_d, ..., par^{j-1}_e of D^{j-1}. For violating the privacy data of the target user in anatomization tables, let u^j_r be the target user of the adversary. Let $B_{u^j_r}$ be the adversary's background

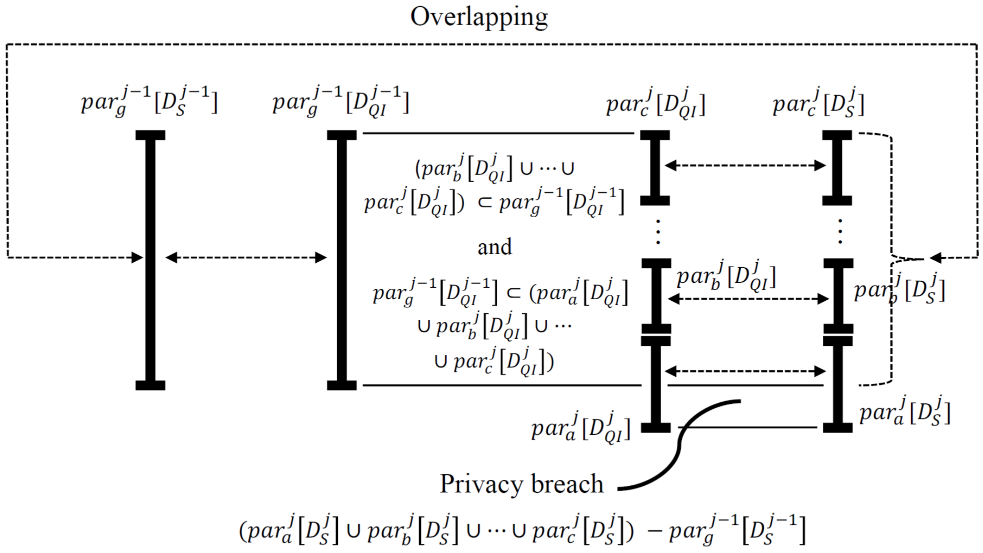
Overlapping
$par_g^{j-1}[D_S^{j-1}]$
$par_g^{j-1}[D_{QI}^{j-1}]$
$par_c^j[D_{QI}^j]$
$par_c^j[D_S^j]$
$(par_b^j[D_{QI}^j] \cup \cdots \cup par_c^j[D_{QI}^j]) \subset par_g^{j-1}[D_{QI}^{j-1}]$
and
$par_g^{j-1}[D_{QI}^{j-1}] \subset (par_a^j[D_{QI}^j] \cup par_b^j[D_{QI}^j] \cup \cdots \cup par_c^j[D_{QI}^j])$
$par_b^j[D_{QI}^j]$
$par_b^j[D_S^j]$
$par_a^j[D_{QI}^j]$
$par_a^j[D_S^j]$
Privacy breach
$(par_a^j[D_S^j] \cup par_b^j[D_S^j] \cup \cdots \cup par_c^j[D_S^j]) - par_g^{j-1}[D_S^{j-1}]$

Fig.2: *The characteristic of privacy violation issues in incremental anatomization tables from using iMRCA attacks.*

knowledge about the target user u_r^j. Let $par_a^j[D_{QI}^j]$ be the quasi-identifier partition that matches $B_{u_r^j}$. Moreover, let $par_b^j[D_{QI}^j]$, ..., $par_c^j[D_{QI}^j]$ be other identified quasi-identifier partitions of the adversary such that they are also available in D_{QI}^j. Let $par_d^{j-1}[D_{QI}^{j-1}], \ldots, par_e^{j-1}[D_{QI}^{j-1}]$ be the related quasi-identifier partitions of $par_a^j[D_{QI}^j]$, $par_b^j[D_{QI}^j]$, ..., $par_c^j[D_{QI}^j]$ such that they satisfy the limitations as $(par_b^j[D_{QI}^j] \cup \ldots \cup par_c^j[D_{QI}^j]) \subset (par_d^{j-1}[D_{QI}^{j-1}] \cup \ldots \cup par_e^{j-1}[D_{QI}^{j-1}])$ and $(par_d^{j-1}[D_{QI}^{j-1}] \cup \ldots \cup par_e^{j-1}[D_{QI}^{j-1}]) \subset (par_a^j[D_{QI}^j] \cup par_b^j[D_{QI}^j] \cup \ldots \cup par_c^j[D_{QI}^j])$. Let $par_a^j[D_S^j]$, $par_b^j[D_S^j]$, ..., $par_c^j[D_S^j]$ be the related sensitive partition of $par_a^j[D_{QI}^j]$, $par_b^j[D_{QI}^j]$, ..., $par_c^j[D_{QI}^j]$ respectively. Also, let $par_d^{j-1}[D_S^{j-1}]$, ..., $par_e^{j-1}[D_S^{j-1}]$ be the related sensitive partition of $par_d^{j-1}[D_{QI}^{j-1}]$, ..., $par_e^{j-1}[D_{QI}^{j-1}]$ respectively. Thus, if the compared result between $par_a^j[D_S^j] \cup par_b^j[D_S^j] \cup \ldots \cup par_c^j[D_S^j]$ and $par_d^{j-1}[D_S^{j-1}] \cup \ldots \cup par_e^{j-1}[D_S^{j-1}]$ does not satisfy the given value of l, the privacy data of the target user u_r^j in D_S^j can be inferred from $(par_a^j[D_S^j] \cup par_b^j[D_S^j] \cup \ldots \cup par_c^j[D_S^j]) - (par_d^{j-1}[D_S^{j-1}] \cup \ldots \cup par_e^{j-1}[D_S^{j-1}])$. The characteristic of privacy violation issues in incremental anatomization tables from using *iMRcMLA* attacks is shown in Figure 3.

3.2 Privacy violation issues in anatomization tables based on removing data

In this section, we demonstrate privacy violation issues that could occur in anatomization tables when some data of them is deleted.

3.2.1 Full Left Coverage data Attack (*dFLCA*)

The assumption of this privacy data attack in anatomization tables is that only par_g^{j-1} of D^{j-1} matches the adversary's background knowledge about the target user. Moreover, there is only par_g^j of D^j such that it matches par_g^{j-1}. For violating the privacy data of the target user in anatomization tables, let u_r^{j-1} be the target user of the adversary. Let $B_{u_r^{j-1}}$ be the adversary's background knowledge about the target user u_r^{j-1}. Let $par_g^{j-1}[D_{QI}^{j-1}]$ be the identified quasi-identifier partition of the adversary such that $par_g^{j-1}[D_{QI}^{j-1}]$ includes the quasi-identifier tuples which match $B_{u_r^{j-1}}$. Moreover, let $par_g^j[D_{QI}^j]$ be the related quasi-identifier partition of $par_g^{j-1}[D_{QI}^{j-1}]$, where $par_g^j[D_{QI}^j] \subset par_g^{j-1}[D_{QI}^{j-1}]$. Let $par_g^j[D_S^j]$ be the sensitive partition which relates to $par_g^j[D_{QI}^j]$, and the sensitive partition $par_g^{j-1}[D_S^{j-1}]$ be the related sensitive partition of $par_g^{j-1}[D_{QI}^{j-1}]$. If the compared result between $par_g^j[D_S^j]$ and $par_g^{j-1}[D_S^{j-1}]$ does not satisfy the given value of l, the privacy data of the target user u_r^{j-1} in D_S^{j-1} can be inferred from $par_g^{j-1}[D_S^{j-1}] - par_g^j[D_S^j]$. The characteristic of privacy violation issues in decremental anatomization tables from using *dFLCA* attacks is shown in Figure 4.

3.2.2 Merged Left Coverage data Attack (*dMLCA*)

The assumption of this privacy data attack in anatomization tables is that only par_g^{j-1} of D^{j-1} matches the adversary's background knowledge about the target user. There is only par_g^j of D^j such that it match to par_g^{j-1}. Moreover, the adversary can see that par_g^j of D^j further relates to par_b^{j-1}, ..., par_c^{j-1} of D^{j-1}. For violating the privacy

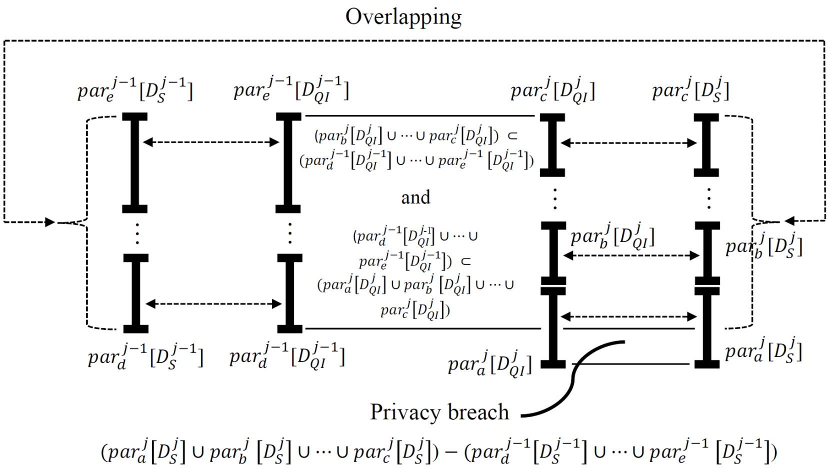
Overlapping
$par_e^{j-1}[D_S^{j-1}]$
$par_e^{j-1}[D_{QI}^{j-1}]$
$par_c^j[D_{QI}^j]$
$par_c^j[D_S^j]$
$(par_b^j[D_{QI}^j] \cup \cdots \cup par_c^j[D_{QI}^j]) \subset (par_d^{j-1}[D_{QI}^{j-1}] \cup \cdots \cup par_e^{j-1}[D_{QI}^{j-1}])$
and
$(par_d^{j-1}[D_{QI}^{j-1}] \cup \cdots \cup par_e^{j-1}[D_{QI}^{j-1}]) \subset (par_a^j[D_{QI}^j] \cup par_b^j[D_{QI}^j] \cup \cdots \cup par_c^j[D_{QI}^j])$
$par_b^j[D_{QI}^j]$
$par_b^j[D_S^j]$
$par_d^{j-1}[D_S^{j-1}]$
$par_d^{j-1}[D_{QI}^{j-1}]$
$par_a^j[D_{QI}^j]$
$par_a^j[D_S^j]$
Privacy breach
$(par_a^j[D_S^j] \cup par_b^j[D_S^j] \cup \cdots \cup par_c^j[D_S^j]) - (par_d^{j-1}[D_S^{j-1}] \cup \cdots \cup par_e^{j-1}[D_S^{j-1}])$

Fig.3: *The characteristic of privacy violation issues in incremental anatomization tables from using iMRcMLA attacks.*

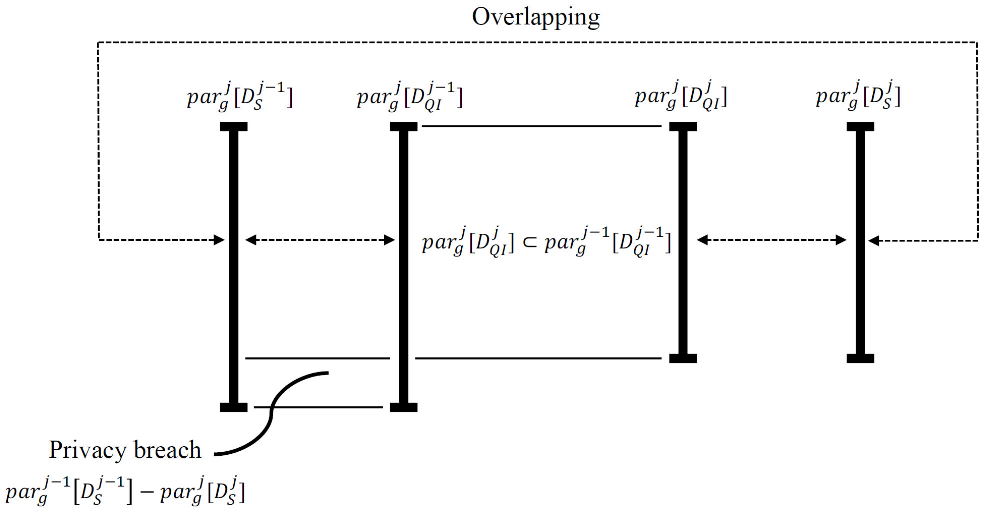

Fig.4: *The characteristic of privacy violation issues in decremental anatomization tables from using dFLCA attacks.*

data of the target user in anatomization tables, let u_r^{j-1} be the target user of the adversary. Let $B_{u_r^{j-1}}$ be the adversary's background knowledge about the target user u_r^{j-1}. Let $par_a^{j-1}[D_{QI}^{j-1}]$ be an identified quasi-identifier partition of the adversary such that $par_a^{j-1}[D_{QI}^{j-1}]$ contains the quasi-identifier tuple which matches $B_{u_r^{j-1}}$. Moreover, let $par_b^{j-1}[D_{QI}^{j-1}]$, …, $par_c^{j-1}[D_{QI}^{j-1}]$ be other identified quasi-identifier partitions of the adversary in D_{QI}^{j-1}. Let $par_g^j[D_{QI}^j]$ be the related quasi-identifier partition of $par_a^{j-1}[D_{QI}^{j-1}]$, $par_b^{j-1}[D_{QI}^{j-1}]$, …, $par_c^{j-1}[D_{QI}^{j-1}]$ such that $par_g^j[D_{QI}^j]$ satisfies the limitations as $(par_b^{j-1}[D_{QI}^{j-1}] \cup \ldots \cup par_c^{j-1}[D_{QI}^{j-1}]) \subset par_g^j[D_{QI}^j]$ and $par_g^j[D_{QI}^j] \subset (par_a^{j-1}[D_{QI}^{j-1}] \cup par_b^{j-1}[D_{QI}^{j-1}] \cup \ldots \cup par_c^{j-1}[D_{QI}^{j-1}])$. Let $par_a^{j-1}[D_S^{j-1}]$, $par_b^{j-1}[D_S^{j-1}]$, …, $par_c^{j-1}[D_S^{j-1}]$ be the related sensitive partition of $par_a^{j-1}[D_{QI}^{j-1}]$, $par_b^{j-1}[D_{QI}^{j-1}]$, …, $par_c^{j-1}[D_{QI}^{j-1}]$ respectively. Also, let $par_g^j[D_S^j]$ be the sensitive partition that relates to $par_g^j[D_{QI}^j]$. If the compared result between $par_a^{j-1}[D_S^{j-1}] \cup par_b^{j-1}[D_S^{j-1}] \cup \ldots \cup par_c^{j-1}[D_S^{j-1}]$ and $par_g^j[D_S^j]$ does not satisfy the given value of l, the privacy data of the target user u_r^{j-1} in D_S^{j-1} can be inferred from $(par_a^{j-1}[D_S^{j-1}] \cup par_b^{j-1}[D_S^{j-1}] \cup \ldots \cup par_c^{j-1}[D_S^{j-1}]) - par_g^j[D_S^j]$. The characteristic of privacy violation issues in decremental anatomization tables from using *dMRCA* attacks is shown in Figure 5.

3.2.3 Merged Left fully covers Merged Right data Attack (*dMLcMRA*)

The assumption of this privacy data attack in anatomization tables is that only par_g^{j-1} of D^{j-1} matches the adversary's background knowledge about the target user and par_b^{j-1}, …, par_c^{j-1} of D^{j-1} are further identified by the adversary. Moreover, par_g^{j-1}, par_b^{j-1}, …, par_c^{j-1} relate to par_d^j, …, par_e^j of D^j. For violating the privacy data of the target user in

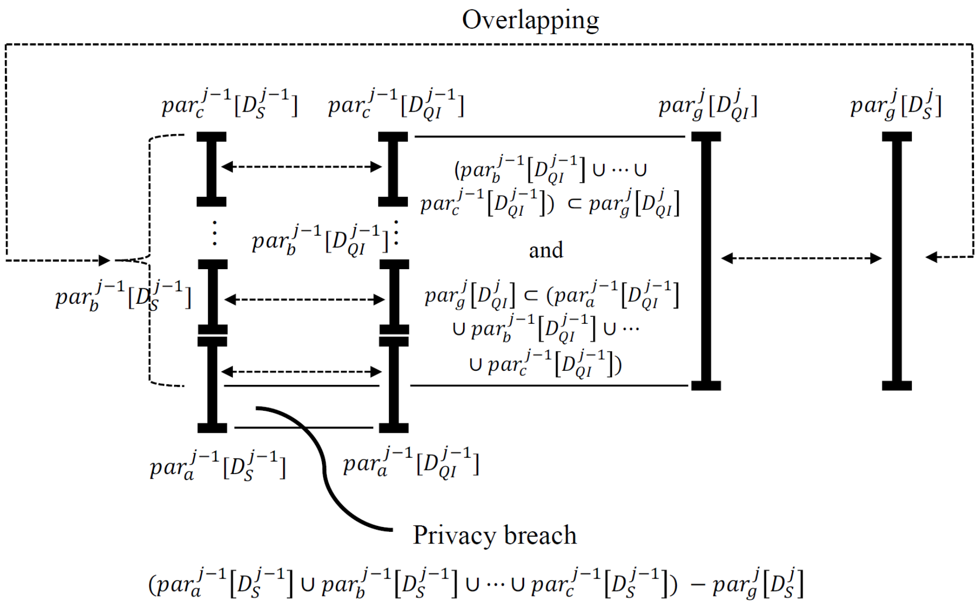
Overlapping
$par_c^{j-1}[D_S^{j-1}]$
$par_c^{j-1}[D_{QI}^{j-1}]$
$par_g^j[D_{QI}^j]$
$par_g^j[D_S^j]$
$(par_b^{j-1}[D_{QI}^{j-1}] \cup \cdots \cup$
$par_c^{j-1}[D_{QI}^{j-1}]) \subset par_g^j[D_{QI}^j]$
$par_b^{j-1}[D_{QI}^{j-1}]$
and
$par_b^{j-1}[D_S^{j-1}]$
$par_g^j[D_{QI}^j] \subset (par_a^{j-1}[D_{QI}^{j-1}]$
$\cup\, par_b^{j-1}[D_{QI}^{j-1}] \cup \cdots$
$\cup\, par_c^{j-1}[D_{QI}^{j-1}])$
$par_a^{j-1}[D_S^{j-1}]$
$par_a^{j-1}[D_{QI}^{j-1}]$
Privacy breach
$(par_a^{j-1}[D_S^{j-1}] \cup par_b^{j-1}[D_S^{j-1}] \cup \cdots \cup par_c^{j-1}[D_S^{j-1}]) - par_g^j[D_S^j]$

Fig.5: *The characteristic of privacy violation issues in decremental anatomization tables from using $dMLCA$ attacks.*

anatomization tables, let u_r^{j-1} be the target user of the adversary. Let $B_{u_r^{j-1}}$ be the adversary's background knowledge about the target user u_r^{j-1}. Let $par_a^{j-1}[D_{QI}^{j-1}]$ be an identified quasi-identifier partition of the adversary such that it contains the quasi-identifier tuple which matches $B_{u_r^{j-1}}$. Moreover, let $par_b^{j-1}[D_{QI}^{j-1}]$, $\ldots$, $par_c^{j-1}[D_{QI}^{j-1}]$ be other identified quasi-identifier partitions of the adversary such that they are also available in D_{QI}^{j-1}. Let $par_d^j[D_{QI}^j]$, $\ldots$, $par_e^j[D_{QI}^j]$ be the related quasi-identifier partitions of $par_a^{j-1}[D_{QI}^{j-1}]$, $par_b^{j-1}[D_{QI}^{j-1}]$, $\ldots$, and $par_c^{j-1}[D_{QI}^{j-1}]$ such that $par_d^j[D_{QI}^j]$, $\ldots$, $par_e^j[D_{QI}^j]$ satisfy the limitations as $(par_b^{j-1}[D_{QI}^{j-1}] \cup \ldots \cup par_c^{j-1}[D_{QI}^{j-1}]) \subset (par_d^j[D_{QI}^j] \cup \ldots \cup par_e^j[D_{QI}^j])$ and $(par_d^j[D_{QI}^j] \cup \ldots \cup par_e^j[D_{QI}^j]) \subset (par_a^{j-1}[D_{QI}^{j-1}] \cup par_b^{j-1}[D_{QI}^{j-1}] \cup \ldots \cup par_c^{j-1}[D_{QI}^{j-1}])$. Let $par_a^{j-1}[D_S^{j-1}]$, $par_b^{j-1}[D_S^{j-1}]$, $\ldots$, $par_c^{j-1}[D_S^{j-1}]$ be the related sensitive partition of $par_a^{j-1}[D_{QI}^{j-1}]$, $par_b^{j-1}[D_{QI}^{j-1}]$, $\ldots$, par_c^{j-1} $[D_{QI}^{j-1}]$ respectively. Also, let $par_d^j[D_S^j]$, $\ldots$, $par_e^j[D_S^j]$ be the related sensitive partition of $par_d^j[D_{QI}^j]$, $\ldots$, $par_e^j[D_{QI}^j]$ respectively. If the compared result between $par_a^{j-1}[D_S^{j-1}] \cup par_b^{j-1}[D_S^{j-1}] \cup \ldots \cup par_c^{j-1}[D_S^{j-1}]$ and $par_d^j[D_S^j] \cup \ldots \cup par_e^j[D_S^j]$ does not satisfy the given value of l, the privacy data of the target user u_r^{j-1} in D_S^{j-1}, it can be inferred from $(par_a^{j-1}[D_S^{j-1}] \cup par_b^{j-1}[D_S^{j-1}] \cup \ldots \cup par_c^{j-1}[D_S^{j-1}]) - (par_d^j[D_S^j] \cup \ldots \cup par_e^j[D_S^j])$. The characteristic of privacy violation issues in decremental anatomization tables from using $dMLcMRA$ attacks is shown in Figure 6.

3.3 Privacy violation issues in anatomization tables based on data modifications

This section presents privacy violation issues that can occur in anatomization tables when the sensitive values are updated. To aid the readability of privacy violation issues that are presented in this section, we call them to be SVM. For violating the privacy data of the target user in anatomization tables, let u_r be the target user of the adversary such that the profile tuple of u_r is available in both related anatomization data versions which are released at the timestamps $j-1$ and j respectively, i.e., D_{QI}^{j-1}, D_S^{j-1}, D_{QI}^j, and D_S^j. Assume that after D_{QI}^{j-1} and D_S^{j-1} are released, the sensitive data of the target user u_r is updated from s_{old} to become s_{new}. Let B_{u_r}, where $B_{u_r} \subset d_r^{j-1}[QI]$, $B_{u_r} \subset d_r^j[QI]$ and $d_r^{j-1}[QI] = d_r^j[QI]$, be the adversary's background knowledge about the target user u_r. Let $par_g^{j-1}[D_{QI}^{j-1}]$ and $par_g^j[D_{QI}^j]$ be both identified quasi-identifier partitions of the adversary such that they contain the quasi-identifier tuple which matches B_{u_r}. Let $par_g^{j-1}[D_S^{j-1}]$ and $par_g^j[D_S^j]$ be the related sensitive partition of $par_g^{j-1}[D_{QI}^{j-1}]$ and $par_g^j[D_{QI}^j]$ respectively. If the compared result between $par_g^{j-1}[D_S^{j-1}]$ and $par_g^j[D_S^j]$ does not satisfy the given value of l, the adversary can infer that the sensitive value of u_r is changed from $par_g^{j-1}[D_S^{j-1}]$ $-$ $par_g^j[D_S^j]$ to become $par_g^j[D_S^j]$ $-$ $par_g^{j-1}[D_S^{j-1}]$. The characteristic of privacy violation issues based on data modifications in anatomization tables from using SVM attacks is shown in Figure 7.

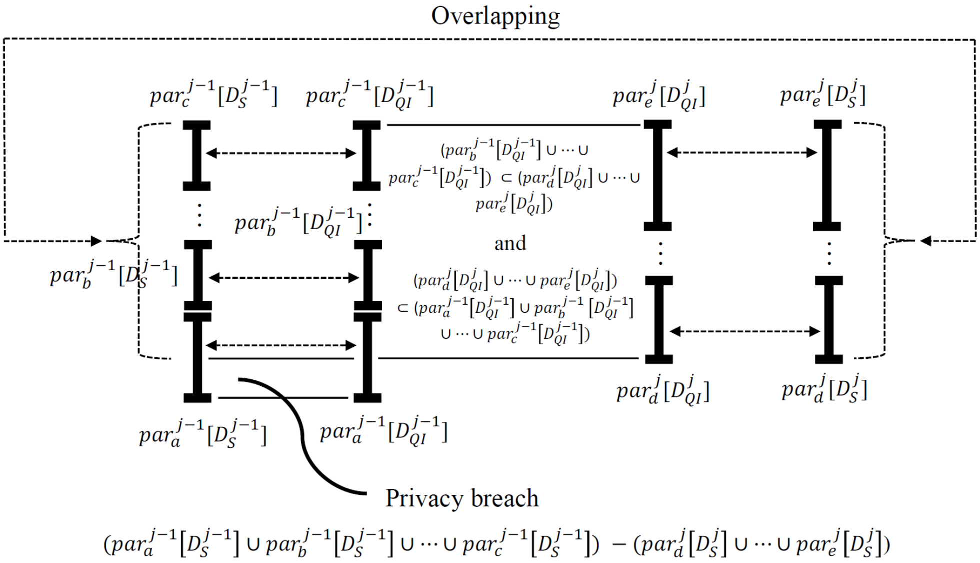
Overlapping
$par_c^{j-1}[D_S^{j-1}]$
$par_c^{j-1}[D_{QI}^{j-1}]$
$par_e^{j}[D_{QI}^{j}]$
$par_e^{j}[D_S^{j}]$
$(par_b^{j-1}[D_{QI}^{j-1}] \cup \cdots \cup par_c^{j-1}[D_{QI}^{j-1}]) \subset (par_d^{j}[D_{QI}^{j}] \cup \cdots \cup par_e^{j}[D_{QI}^{j}])$
$par_b^{j-1}[D_{QI}^{j-1}]$
and
$par_b^{j-1}[D_S^{j-1}]$
$(par_d^{j}[D_{QI}^{j}] \cup \cdots \cup par_e^{j}[D_{QI}^{j}]) \subset (par_a^{j-1}[D_{QI}^{j-1}] \cup par_b^{j-1}[D_{QI}^{j-1}] \cup \cdots \cup par_c^{j-1}[D_{QI}^{j-1}])$
$par_d^{j}[D_{QI}^{j}]$
$par_d^{j}[D_S^{j}]$
$par_a^{j-1}[D_S^{j-1}]$
$par_a^{j-1}[D_{QI}^{j-1}]$
Privacy breach
$(par_a^{j-1}[D_S^{j-1}] \cup par_b^{j-1}[D_S^{j-1}] \cup \cdots \cup par_c^{j-1}[D_S^{j-1}]) - (par_d^{j}[D_S^{j}] \cup \cdots \cup par_e^{j}[D_S^{j}])$

Fig. 6: *The characteristic of privacy violation issues in decremental anatomization tables from using dMLcMRA attacks.*

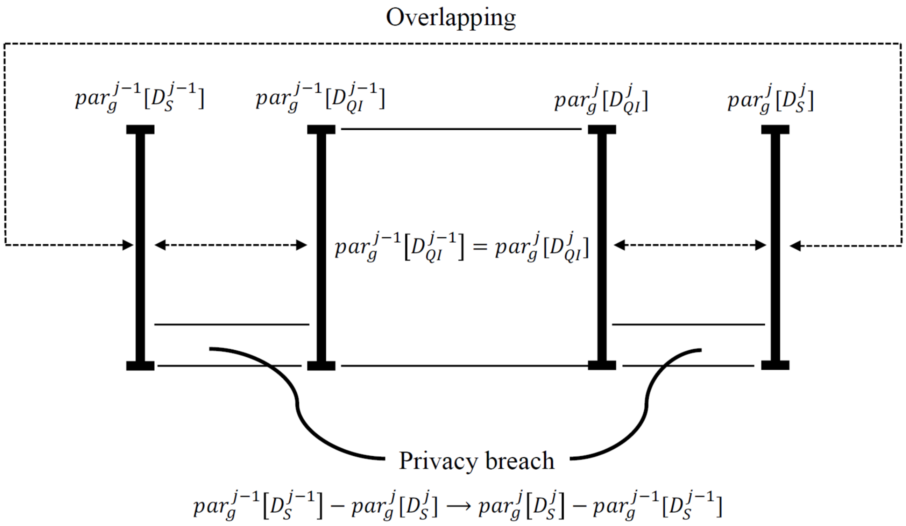
Overlapping
$par_g^{j-1}[D_S^{j-1}]$
$par_g^{j-1}[D_{QI}^{j-1}]$
$par_g^{j}[D_{QI}^{j}]$
$par_g^{j}[D_S^{j}]$
$par_g^{j-1}[D_{QI}^{j-1}] = par_g^{j}[D_{QI}^{j}]$
Privacy breach
$par_g^{j-1}[D_S^{j-1}] - par_g^{j}[D_S^{j}] \longrightarrow par_g^{j}[D_S^{j}] - par_g^{j-1}[D_S^{j-1}]$

Fig. 7: *The characteristic of privacy violation issues based on data modification in anatomization tables from using SVM attacks.*

3.4 Privacy violation issues in anatomization tables based on partition changing

Aside from $iFRCA$, $iMRCA$, $iMRcMLA$, $dFRCA$, $dMRCA$, $dMRcMLA$, and SVM, anatomization tables are susceptible to privacy violation issues when the tuples are moved to be in different partition.

For example, let Table 1 be the specified raw dataset at the timestamp $j-1$, denoted as D^{j-1}, and its anatomization tables are shown in Tables 6 and 7,

so denoted as D_{QI}^{j-1} and D_S^{j-1} respectively. We suppose that after Tables 6 and 7 are released, the tuples (49, $Male$, 60639, $Cancer$), (50, $Male$, 60639, HIV), and (51, $Male$, 60639, Flu) are inserted into D, so denoted as D^j. For privacy preservation, let the value of l be set to 3. Thus, Tables 9 and 10 are an anatomization data version of Table 1 at the timestamp j, denoted as D_{QI}^j and D_S^j respectively. Let $David$ be the target user of the adversary who needs to reveal $David$'s disease from Tables 7 and 10. We assume that the adversary knows that $David$ is a male person who is 48 years old, $B_{David} = (48, Male)$. Moreover, the adversary strongly believes that $David$'s profile tuple is available in Tables 6 and 7, 9, and 10. In this situation, the quasi-identifier partition $PID = 1$ of Table 6 and the quasi-identifier partition $PID = 2$ of Table 9 are both desired quasi-identifier partitions of the adversary because they contain the quasi-identifier tuple, (48, $Male$, 60639), which matches B_{David}. Moreover, the adversary can observe that the quasi-identifier partition $PID = 1$

Table 9: *The quasi-identifier table of Table 1, where $l = 3$, after the tuple* (49, $Male$, 60639, $Cancer$), (50, $Male$, 60639, HIV)*, and* (51, $Male$, 60639, Flu) *has been inserted.*

Age	Sex	Zipcode	PID
45	Male	60636	1
46	Male	60632	
47	Male	60635	
48	Male	60639	2
49	Male	60639	
50	Male	60639	
51	Male	60639	
48	Female	60632	3
42	Female	60632	
42	Female	60632	
41	Female	60636	4
40	Female	60636	
39	Female	60636	
38	Female	60639	
37	Female	60639	

Table 10: *The sensitive table of Table 1, where $l = 3$, after the tuple* (49, $Male$, 60639, $Cancer$), (50, $Male$, 60639, HIV)*, and* (51, $Male$, 60639, Flu) *has been inserted.*

Disease	PID
Flu	1
Fever	
Cancer	
Cancer	2
Cancer	
HIV	
Flu	
Cancer	3
HIV	
Fever	
Cancer	4
HIV	
HIV	
Fever	
Cancer	

of Table 6 fully covers the quasi-identifier partition $PID = 1$ of Table 9. With the quasi-identifier partition $PID = 1$ of Table 6, it relates to Flu, $Fever$, $Cancer$, and $Cancer$ in Table 7. The quasi-identifier partition $PID = 1$ of Table 9 relates to Flu, $Fever$, and $Cancer$ in Table 10. Thus, the adversary can be highly confident that the quasi-identifier tuple (48, $Male$, 60639) of Table 6 relates to $Cancer$ in Table 7. In addition, after the adversary compares all quasi-identifier tuples that are available in Table 6 and 9, the adversary can ensure that the quasi-identifier tuples (49, $Male$, 60639), (50, $Male$, 60639), and (51, $Male$, 60639) are inserted into Table 1 after Table 6 and 7 are released. That is because these quasi-identifier tuples (49, $Male$, 60639), (50, $Male$, 60639) to only be available in Table 9. Furthermore, after the adversary compares the sensitive values that are available in Table 7 and 10, the adversary can see that there are three diseases to be different, i.e., $Cancer$, HIV, and Flu. In this situation, the adversary can infer that the quasi-identifier tuple (48, $Male$, 60639) of Table 9 relates to $Cancer$ of Table 10. Therefore, the adversary strongly believes that $David$'s disease in 7 and 10 to be $Cancer$.

At this point, it is clear that anatomization tables allow to change (insert, delete, and update) of the data and are independently released. They have privacy violation issues that must be addressed. For this reason, a new anatomization model is proposed in Section 4.

4. THE PROPOSED PRIVACY PRESERVATION MODEL

In this section, an appropriate model for addressing privacy violation issues in dynamic anatomization tables based on anatomization constraints [28] in conjunction with additive noise (data holding) [2] [6] [12] and data suppression [25] [5] [17] is presented.

4.1 Basic definition

Definition 4 (Data suppression) Let D^j be the raw dataset D at the timestamps j, and its anatomization tables are denoted as D^j_{QI} and D^j_S. Let d^j_r be an arbitrary tuple that is available in D^j. The meaning of data suppressions for d^j_r is that d^j_r cannot be available in D^j_{QI} and D^j_S.

Definition 5 (Data holding) Let D^{j-1} and D^j be the raw dataset D at the timestamps $j - 1$ and j respectively. Let d^{j-1}_r be an arbitrary tuple which is available in D^{j-1}. The meaning of data holding for d^{j-1}_r is that d^j_r of D^j is replaced by d^{j-1}_r of D^{j-1} or d^{j-1}_r is inserted into D^j.

Definition 6 (Partition of tuples) Let the positive integer l, where $l \in I^+$ and $l \geq 2$, be the privacy preservation constraint. Let D^1_{PAR}, D^2_{PAR}, $\dots$, D^{j-1}_{PAR} be the set of partitions that are available in the anatomization tables of D such that they are released at the timestamps $1, 2, \dots, j - 1$ respectively. Let $D^\beta_{PAR} = \{par^\beta_1, par^\beta_2, \dots, par^\beta_x\}$ be the set of partitions of D^β, where $1 \leq \beta \leq j-1$. For privacy preservation, let $f_{PAR}(D^1_{PAR}, D^2_{PAR}, \dots, D^{j-1}_{PAR}, D^j, l) : D^j \rightarrow_{D^1_{PAR}, D^2_{PAR}, \dots, D^{j-1}_{PAR}, l} D^j_{PAR}$ be the function for partitioning the tuples of D^j to become D^j_{PAR} where $f_{PAR}(D^1_{PAR}, D^2_{PAR}, \dots, D^{j-1}_{PAR}, D^j, l)$ is based on data suppressions in conjunction with data holding. D^j_{PAR} is the set of partitions of D^j, i.e., $D^j_{PAR} = \{par^j_1, par^j_2, \dots, par^j_x\}$. In addition, $1, 2, \dots, x$ are the defined partition identifiers, PID, for the partitions par^{j-1}_1, par^{j-1}_2, $\dots$, par^{j-1}_x respectively. Moreover, par^j_1, par^j_2, $\dots$, par^j_x must be satisfied by limitations as follows:

Table 11: *The data version of Table 1 without all explicit identifier values of users at the timestamp j.*

Age	Sex	Zipcode	Disease
45	Male	60636	Flu
46	Male	60632	Fever
47	Male	60635	Cancer
48	Male	60639	Cancer
48	Female	60632	Cancer
41	Female	60636	Cancer
40	Female	60636	HIV
39	Female	60636	HIV
38	Female	60639	Fever
37	Female	60639	Cervical cancer
51	Female	60635	Cancer
52	Female	60635	HIV
53	Female	60635	Flu

Table 12: *A partitioned data version of Table 11, where $l = 3$, satisfies Definition 6.*

Age	Sex	Zipcode	Disease	PID
45	Male	60636	Flu	1
46	Male	60632	Fever	
47	Male	60635	Cancer	
48	Male	60639	Cancer	
41	Female	60636	Cancer	2
40	Female	60636	HIV	
39	Female	60636	HIV	
38	Female	60639	Fever	
37	Female	60639	Cancer	
51	Female	60635	Cancer	3
52	Female	60635	HIV	
53	Female	60635	Flu	

Table 13: *A version of quasi-identifier tables is constructed from Table 11, where $l = 3$, such that it satisfies Definition 7.*

Age	Sex	Zipcode	PID
45	Male	60636	1
46	Male	60632	
47	Male	60635	
48	Male	60639	
41	Female	60636	2
40	Female	60636	
39	Female	60636	
38	Female	60639	
37	Female	60639	
51	Female	60635	3
52	Female	60635	
53	Female	60635	

Table 14: *A version of sensitive tables is constructed from Table 11, where $l = 3$, such that it satisfies Definition 7.*

Disease	PID
Flu	1
Fever	
Cancer	
Cancer	
Cancer	2
HIV	
HIV	
Fever	
Cancer	
Cancer	3
HIV	
Flu	

- $par_g^j \subseteq D^j$, where $1 \leq g \leq x$,
- $par_1^j \cap par_2^j \cap \ldots \cap par_x^j = \emptyset$,
- The number of distinct sensitive values is available in every $par_g^j[S]$ to be equal to or greater than l values, i.e., $|par_g^j[S]| \geq l$, and
- All possibly compared results between $par_g^j[S]$ and its related sensitive partitions that are available in $par_g^\beta[S]$ must also contain at least l distinct sensitive values.

Definition 7 (Anatomization tables) Let D_{PAR}^j be the set of partitions of D^j such that it satisfies Definition 6. Let $f_{Ana}(D_{PAR}^j) : D_{PAR}^j \rightarrow D_{QI}^j, D_S^j$ be the function for anatomizing D_{PAR}^j to become D_{QI}^j and D_S^j. The relationship of each partition in D_{QI}^j and D_S^j is represented by its PID.

Definition 8 (The error of partitions) Let $par_g^j[D_{QI}^j]$ and $par_g^j[D_S^j]$ be the specified partition par_g^j of D_{QI}^j and D_S^j respectively, where $par_g^j[D_{QI}^j]$ and $par_g^j[D_S^j]$ are constructed from $par_g^j \in D_{PAR}^j$. For this reason, the penalty cost of D_{QI}^j and D_S^j can be defined from par_g^j, i.e., the penalty cost of D_{QI}^j and D_S^j can be defined by $PL(D^j, par_g^j)$, as shown in Equation 1. More penalty cost of $PL(D^j, par_g^j)$ implies that par_g^j is less data utility.

$$PL(D^j, par_g^j) = |par_g^j|^2 + (|D^j| \cdot (|HD|) + |SP|)) \quad (1)$$

Where, $|HD|$ is the number of the tuples which are held, and $|SP|$ is the number of the tuples which are suppressed.

Definition 9 (The error of anatomization tables) Let D_{QI}^j and D_S^j be constructed from $D_{PAR}^j = \{par_1^j$, par_2^j, $\ldots$, $par_x^j\}$. Thus, the penalty cost of D_{QI}^j and D_S^j can be defined by $DL(D^j, D_{PAR}^j)$, as shown in Equation 2. More penalty cost of $DL(D^j, D_{PAR}^j)$ implies implies that D_{PAR}^j is less data utility.

$$DL(D^j, D_{PAR}^j) = \sum_{g=1}^{x} PL(D^j, par_g^j) \quad (2)$$

For example, let Table 1 be the raw dataset at the timestamp $j-1$, so denoted as D^{j-1}, and its anatomization tables are shown in Tables 6 and 7, so denoted as D_{QI}^{j-1} and D_S^{j-1} respectively. After Tables 6 and 7 are released, we suppose that the tuples (42, $Female$, 60632, HIV) and (42, $Female$, 60632, $Fever$) are deleted from Table 1, and the disease of the tuple (37, $Female$, 60639) of Table 1 is further updated from $Cancer$ to become $Cervical\ cancer$. Moreover, the tuples (51, $Male$, 60635, $Cancer$), (52, $Male$, 60635, HIV), and (53, $Male$, 60635, Flu), are inserted into Table 1. The new resulting data version of Table 1 at the timestamp j is shown in Table 11, denoted as D^j. For privacy preservation, let the value of l be set to 3. At first, the tuples of Table 11 are partitioned by the given value of l to satisfy Definition 6, as shown in Table 12. In Table 12, we can see that every partition contains at least three different diseases. Moreover, the compared result between each specified partition of Table 12 and its related partition(s) that is available in Table 5, always has at least three different diseases. Furthermore, we can observe that the tuple (37, $Female$, 60639, $Cancer$) of Table 1 is in Table 12, and its updated tuple version, (37, $Female$, 60639, $Cervical\ cancer$), in Table 11 is suppressed. That is because if the tuple (37, $Female$, 60639, $Cancer$) of Table 2 is not present and the tuple (37, $Female$, 60639, $Cervical\ cancer$) of Table 11 is not suppressed, their diseases can be disclosed by using SVM after they are released for public use. Furthermore, we can see that the tuple (48, $Female$, 60636, $Cancer$) of Table 11 is also suppressed because it can lead to privacy violation issues from using a privacy data attack described in Section 3.4 Finally, the tuples of Table 12 are anatomized to become Table 13 and 14. Table 13 and 14 are an anatomization data version of Table 11 which is not vulnerable to any of the data attacks which are proposed in Section 3. Moreover, the error of these anatomization tables, $DL(Table\ 12)$, is only 89.

4.2 Dynamic anatomization algorithm (DAA)

In this section, a privacy preservation algorithm for addressing privacy violation issues in dynamic anatomization tables is proposed. The algorithm is based on the assumption that all relatedly released data versions of D^j are released from the timestamp 1 to $j-1$, they are always satisfied by the limitations of the algorithm. Thus, only the anatomization tables are released at the timestamp $j-1$ to be considered for constructing D^j_{QI} and D^j_S. With this algorithm, aside from privacy preservation, the data utility and the execution time are also maintained as much as possible. To achieve the aims of the algorithm, greedy [9] and data clustering [3] [29] are applied, as shown in Algorithm 1.

The inputs of the algorithm are a positive integer l, the partitioned data version D^{j-1}_{PAR}, and the raw dataset D^j. The output of the proposed algorithm is the quasi-identifier table D^j_{QI} and the sensitive table D^j_S such that they satisfy Definition 7.

For privacy preservation, the changed tuples of D^j are investigated in the first step. That is, if they do not satisfy the given value of l, they are suppressed or replaced by the old version of them because they can lead to privacy violation issues from using data attacks that are presented in Section 3.

In the second step, all tuples are available in TMP_0 to be iterated until they cannot satisfy the given value of l. In each iteration, an arbitrary tuple $d^j_r \in TMP_0$ is assigned to be the initial tuple for constructing each partition of D^j_{PAR}. Moreover, the best co-tuples, ct_g, of the initial tuple are also determined by the sub-algorithm FBT that is shown in Algorithm 2. That is, $ct_g \cup d^j_r$ includes at least l distinct sensitive values, and all possible compared results between $ct_g \cup d^j_r$ and its related partitions which are available in D^{j-1}_{PAR} must also satisfy the given value of l. Furthermore, $ct_g \cup d^j_r$ is removed from TMP_0. Finally, DP is returned to the main algorithm DAA. In this situation, we can see that the size of $ct_g \cup d^j_r$ directly influences the data search space (the size of CT) of the next data partition processes.

For example, we suppose that TMP_0 collects five user profile tuples, i.e., the tuples 1, 2, 3, 4, and 5. For privacy preservation, let the value of l be set to 2. We suppose that the algorithm assigns tuple 1 to be the initial tuple for constructing the first partition. With this partition, the number of possible partitions must be considered by FBT, i.e., 15 partitions. For this reason, we can claim that the cost of constructing the first partition of TMP_0 can be denoted by Equation 3, i.e., $1 + ((5-1)!/(1! * ((5-1)-1)!)) + ((5 - 1)! / (2! * ((5 - 1) - 2)!)) + ((5-1)!/(3! * ((5-1)-3)!)) = 15$.

$$SubFBTCost = 1 + \sum_{PSize=l-1}^{|TMP_0|-2} \frac{(|TMP_0|-1)!}{PSize! \cdot ((|TMP_0|-1)-PSize)!} \qquad (3)$$

In addition, we assume that tuple 2 is the best co-tuple of tuple 1. Tuples 1 and 2 are the first tuple partition that is constructed from TMP_0. After that, the tuples 1 and 2 are removed from TMP_0. Thus, TMP_0 only remains the tuples 3, 4, and 5. Then, the remained tuples of TMP_0 are partitioned by using FBT again because they still satisfy $l = 2$. To construct the second partition of TMP_0, we suppose that the algorithm assigns tuple 3 to be the initial tuple. The cost of constructing this tuple partitions of TMP_0 can also be defined by Equation 3, i.e, $1 + ((3-1)!/(1! * ((3-1)-1)!)) = 3$. We assume that tuple 4 is chosen by the algorithm to be the best co-tuple of tuple 3, so, tuples 3 and 4 are the second tuple partition. Also, after the second partition is constructed, tuples 3 and 4 are removed from TMP_0. Thus, TMP_0 contains only tuple 5. In this situa-

Algorithm 1 DAA(l, $PAR_{D^{j-1}}$, D^j)

```
Let $TMP_0$ and $TMP_1$ be temporaries.
Let $Tuple_{Deleted}$ be the set of deleted tuples.
Let $Tuple_{Updated}$ and $Tuple_{Old}$ be the set of updated tuples and the set of considering the previous tuple
  of them respectively.
Let $Tuple_{Inserted}$ be the set of inserted tuples.
Let $D^j_{PAR}$ be the set of desired partitions of $D^j$.
$TMP_0 = D^j$
if $Tuple_{Deleted}$ does not satisfy $l$ then
    $TMP_0 = TMP_0 \cup Tuple_{Deleted}$
end if
if $Tuple_{Updated}$ does not satisfy $l$ then
    $TMP_0 = (TMP_0 \cup Tuple_{Old}) - Tuple_{Updated}$
end if
if $Tuple_{Inserted}$ does not satisfy $l$ then
    $TMP_0 = TMP_0 - Tuple_{Inserted}$
end if
while $|TMP_0[S]| \geq l$ do
    $TMP_1 = FBT(D^j, D^{j-1}_{PAR}, l, TMP_0 - d^j_r, d^j_r)$, where $d^j_r \in TMP_0$
    $D^j_{PAR} = D^j_{PAR} \cup TMP_1$
    $TMP_0 = TMP_0 - TMP_1$
end while
while $|TMP_0| \geq 1$ and $D^j_{PAR} \neq \emptyset$ do
    $D^j_{PAR} = FBP(D^j, D^{j-1}_{PAR}, D^j_{PAR}, l, d^j_r)$, where $d^j_r \in TMP_0$
    $TMP_0 = TMP_0 - d^j_r$
end while
if $D^j_{PAR} \neq \emptyset$ then
    $D^j_{PAR}$ is anatomized to be $D^j_{QI}$ and $D^j_S$
    return $D^j_{QI}$ and $D^j_S$
else
    return $NULL$
end if
```

tion, we can see that the remained tuple of TMP_0 does not accord $l = 2$, so, the new partition cannot be formed, but the algorithm can assign a suitable partition by FBP (Find the Best of Partition), Algorithm 3, which will be explained in the next step. In addition, the complexity of FBT can be defined by Equation 4.

$$FBTCost = \sum_{loop=1}^{|TMP_0|\ mod\ l} 1 + \sum_{PSize=l-1}^{(|TMP_0|-(l*(loop-1)))-2} \frac{(|TMP_0|-1)!}{PSize! \cdot (((|TMP_0|-1) - PSize)!} \quad (4)$$

Finally, the algorithm investigates the remaining tuples of TMP_0, if they do not accord $l = 2$, FBP is enabled. To assign each remaining tuple d^j_r to the appropriate tuple partition by FBP, all remaining tuples are iterated. In each iteration, the error of d^j_r $\cup$ par^j_g is calculated. If the error of $d^j_r \cup par^j_g$ is minimized and the compared result between $d^j_r \cup par^j_g$ and every related partition $par^{j-1}_g \in D^{j-1}_{PAR}$ satisfies the given value of l, d^j_r is assigned to par^j_g and it is removed from TMP_0. For this reason, the complexity of assigning the appropriate tuple partition of each remaining tuple d^j_r can be computed with Equation 5.

$$SubFBPCost = |D^j_{PAR}| * |D^{j-1}_{PAR}| \quad (5)$$

Therefore, the complexity of FBP can be calculated by Equation 6.

$$FBPCost = SubFBPCost * |\Psi| \quad (6)$$

That is, $|\Psi|$ is the number of tuples that are available in TMP_0 such that they cannot be assigned to their appropriate partition by FBT.

For example, let tuples 1 and 2 be the first partition and let the second partition be constructed from tuples 3 and 4. Let tuple 5 be the remaining tuple that cannot be assigned to any tuple partition by FBT. To assign the appropriate tuple partition for tuple 5, we assume that the error of tuples 1, 2, and 5 is represented by Δ. Moreover, Θ represents the error of tuples 3, 4, and 5. If $\Delta > \Theta$ then the second partition is the appropriate tuple partition for tuple

Algorithm 2 FBT(D^j, D^{j-1}_{PAR}, l, TMP_0, d^j)

```
Let $CT$ be temporary.
let $DP$ and $TDP$ be the desired partition and the temporary desired partition.
let $Err_0$ and $Err_1$ be the error of the specific tuples.
Let $IsDesired$ be the partition state.
$Err_0 = \infty$
for $PSize = l - 1; PSize \leq |TMP_0|; PSize{+}{+}$ do
    $CT = CT \cup Combine(TMP_0, PSize)$
end for
for $g = 1; g \leq |CT|; g{+}{+}$ do
    $TDP = ct_g \cup d^j_r$, where $ct_g \in CT$
    if $TDP[S]$ satisfies $l$ then
        $Err_1 = \sum_{\forall qi \in QI} |TDP[qi]|$
        if $Err_0 > Err_1$ then
            $IsDesired = True$
            for $gg = 1; gg \leq |D^{j-1}_{PAR}|; gg{+}{+}$ do
                if The compared result between $TDP[S]$ and $par^{j-1}_{gg}[S]$ does not satisfy $l$ then
                    $IsDesired = False$
                    Break
                end if
            end for
            if $IsDesired = True$ then
                $DP = TDP$
                $Err_0 = Err_1$
            end if
        end if
    end if
end for
return $DP$
```

5. Therefore, we can conclude that the complexity of the proposed algorithm, DAA, can be determined with Equation 7. That is, the complexity of DAA depends on the complexity of its sub-algorithms FBT and FBP.

$$DDACost = FBTCost + FBPCost \tag{7}$$

5. EXPERIMENT

In this section, the experimental results show the effectiveness and efficiency of the proposed model by comparing it with the comparative anatomization model, Anatomy, that is proposed in [28].

5.1 Experiment setup

All experiments used to evaluate the effectiveness and efficiency of the proposed model were conducted on both Intel(R) Xeon(R) Gold 5218 @2.30 GHz CPUs together with 64 GB memory and six 900 GB HDDs with RAID-5. All implementations were built with Microsoft Visual Studio 2019 Community Edition in conjunction with MSSQL Server 2019.

The experimental results are discussed in this section conducted on the *Adult* dataset which is available in *UCI* Machine Learning Repository [8]. 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 *Age*, *Occupation*, *Sex*, *Race*, and *Capital-loss* are used such that *Age*, *Occupation*, *Sex*, *Race* are the quasi-identifier attributes, and *Capital-loss* is the sensitive attribute. Moreover, all user profile tuples include the values as "?" and "0", they are removed. Thus, only 2181 user profile tuples remained in the experimental dataset, so-called D^{ALL}_{EXP}.

5.2 Experimental results and discussions

5.2.1 Effectiveness

In the first experiment, privacy violation issues in anatomization tables are evaluated by using privacy data attacks which are presented in Section 3. For experiments, the value of l is fixed to be 8. The tuples of D^{ALL}_{EXP} are first randomized to select 1000 tuples, denoted as $D^{1000}_{EXP_{j-1}}$, and $D^{1000}_{EXP_{j-1}}$ is anatomized to satisfy the given value of l, denoted as $D'^{1000}_{EXP_{j-1}}$. For evaluating the effect of privacy violation issues in anatomization tables after tuples were deleted, the tuples of $D^{1000}_{EXP_{j-1}}$ are randomly deleted from 5 to

Algorithm 3 FBP(D^j, D^{j-1}_{PAR}, D^j_{PAR}, l, d^j)

```
Let $DP$, $TDP$, and $ODP$ be the temporary partition.
Let $Err_0$ and $Err_1$ be the error of the specific tuples.
Let $IsDesired$ be the partition state.
$Err_0 = \infty$
for $g = 1; g \le |D^j_{PAR}|; g{+}{+}$ do
    $TDP = par^j_g \cup d^j_r$
    $Err_1 = \sum_{\forall qi \in QI} |TDP[qi]|$
    if $Err_0 > Err_1$ then
        $IsDesired = True$
        for $gg = 1; gg \le |D^{j-1}_{PAR}|; gg{+}{+}$ do
            if The compared result between $par^j_g[S]$ and $par^{j-1}_{gg}[S]$ does not satisfy $l$ then
                $IsDesired = False$
                Break
            end if
        end for
        if $IsDesired = True$ then
            $DP = TDP$
            $ODP = par^j_g$
            $Err_0 = Err_1$
        end if
    end if
end for
if $DP \neq NULL$ and $ODP \neq NULL$ then
    $D^j_{PAR} = (D^j_{PAR} - ODP) \cup DP$
end if
return $D^j_{PAR}$
```

160 tuples. Each deleted data version of $D^{1000}_{EXP_{j-1}}$ denotes $D^{1000,D_x}_{EXP_{DEL_j}}$, where $5 \le x \le 160$. Also, every $D^{1000,D_x}_{EXP_{DEL_j}}$ is anatomized to satisfy the given value of l and is denoted as $D'^{1000,D_x}_{EXP_{DEL_j}}$. To evaluate the effect of privacy violation issues in anatomization tables after new tuples were inserted, the tuples of D^{ALL}_{EXP} unavailable in $D^{1000}_{EXP_{j-1}}$ are randomly selected by varying from 5 to 160 tuples and are then inserted into $D^{1000}_{EXP_{j-1}}$. Each data version of $D^{1000}_{EXP_{j-1}}$ after inserting the selected tuples is denoted as $D^{1000,D_x}_{EXP_{INS_j}}$, where $5 \le x \le 160$. Also, $D^{1000,D_x}_{EXP_{INS_j}}$ is anatomized to satisfy the given value of l, and is denoted as $D'^{1000,D_x}_{EXP_{INS_j}}$. To evaluate the effect of privacy violation issues in anatomization tables after tuples are updated, the sensitive value of tuples available in $D^{1000}_{EXP_{j-1}}$ randomly updated by varying from 5 to 160 tuples. Each data version of $D^{1000}_{EXP_{j-1}}$ after updates the sensitive value to be denoted as $D^{1000,D_x}_{EXP_{UPD_j}}$, and $D^{1000,D_x}_{EXP_{UPD_j}}$ is anatomized to satisfy the given value of l, denoted as $D'^{1000,D_x}_{EXP_{UPD_j}}$.

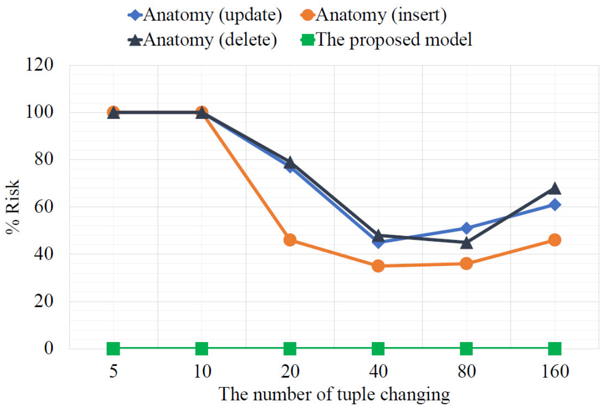

Fig.8: *The percentage of data risks based on the number of tuple changing*

The experimental results as shown in Figure 8 show that after anatomization tables satisfy the proposed privacy preservation constraints, they do not have any vulnerabilities to privacy violation issues from using any of the privacy data attacks that are presented in Section 3. The cause of this effectiveness of the proposed privacy preservation constraints is that aside from anatomization tables that satisfy privacy preservation constraints, all compared results between anatomization tables and their related anatomization tables which were released at the timestamp $j-1$ must also be satisfied by privacy preservation constraints. However, we see that even when the anatomization tables satisfy the comparative anatomization constraints, they still are vulnerable to privacy data attacks are presented in Section 3.

In the second experiment, the effect of the value of l is evaluated. For experiments, the value of l

is varied from 2 to 10. The 100-tuples of $D^{1000}_{EXP_{j-1}}$ are randomly deleted, and the 100-tuples of $D^{1000}_{EXP_{j-1}}$ are randomly to update the sensitive value of them. Moreover, the 100-tuples of D^{ALL}_{EXP} are not available in $D^{1000}_{EXP_{j-1}}$, they are randomly to select and insert into $D^{1000}_{EXP_{j-1}}$. Then, the experimental dataset is anatomized to satisfy the given value of l.

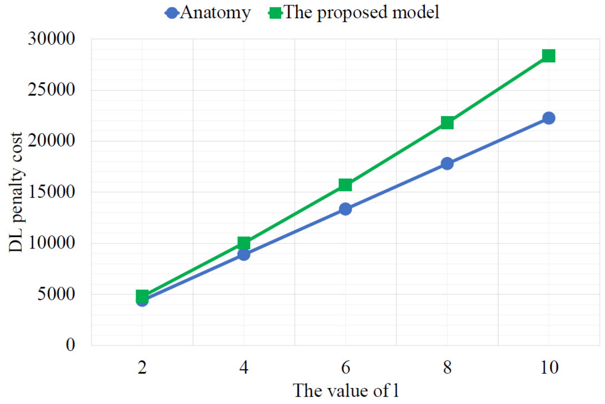

Fig.9: *Data utility based on the value of l*

The experimental results as shown in Figure 9 show that when the value of l is increased, the penalty cost of the experimental anatomization tables also increases. The cause of increasing the penalty cost of the experimental anatomization tables when the value of l increases is the size of partitions that are available in the experimental anatomization tables. The size of partitions often becomes larger when the value of l is increased. Although all experimental results indicate that the proposed privacy preservation constraints are less effective than the comparative anatomization constraints, they are only a little difference. In general, there is a trade-off between data privacy and data utility. An increase in privacy leads to a decrease in utility and vice versa.

The third experiment evaluates the effect of the number of quasi-identifier attributes. For experiments, all experimental datasets are the same as the experimental datasets used in the second experiment. The value of l is fixed to be 8. Moreover, the number of quasi-identifier attributes varies from 1 to 4 attributes.

From the experimental results as shown in Figure 10, we conclude that the number of quasi-identifier attributes also influences the data utility of the experimental anatomization tables. Although all experimental results indicate that the proposed privacy preservation constraints are less effective than the comparative anatomization constraints, they are only a little difference. The proposed privacy preservation constraints often lead to data being more secure in terms of privacy preservations than the comparative anatomization constraints.

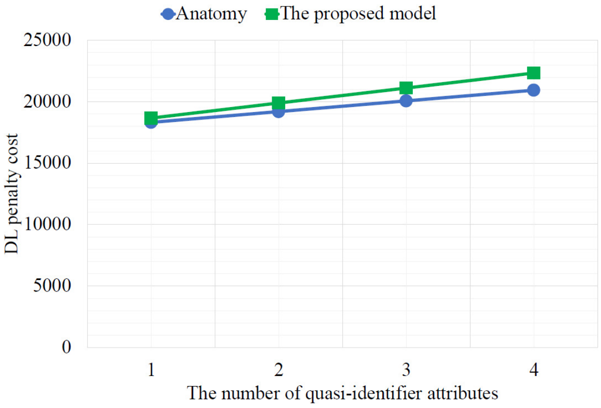

Fig.10: *Data utility based on the number of quasi-identifier attributes*

5.2.2 Efficiency

The fourth experiment evaluates the effect of the size of datasets on the execution time required for transforming the datasets to satisfy privacy preservation constraints. For experiments, the value of l is fixed to be 8. The tuples are available in D^{ALL}_{EXP} to be randomly selected by varying from 100 to 1000 tuples to be the raw datasets. The raw datasets are transformed to satisfy the given value of l.

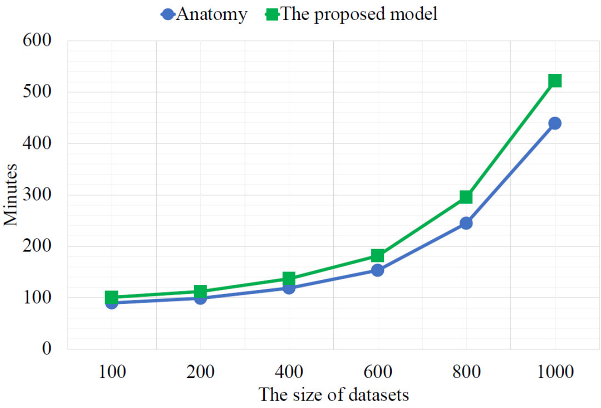

Fig.11: *Execution time based on the size of datasets*

From the experimental results as shown in Figure 11, we see that when the size of the raw datasets is increased, the execution time for the data transformations is also increased. The execution time increases because when the size of the raw datasets is increased, it is the data search space increases.

The fifth experiment evaluated the effect of the value of l on the execution time for transforming the raw datasets to satisfy privacy preservation constraints. For experiments, the value of l is varied from 2 to 8. All experimental datasets used in this experiment are the same as the experimental datasets used in the second experiment.

From the experimental results as shown in Figure 12, we see that when the value of l is increased, the execution time for the data transformations is also

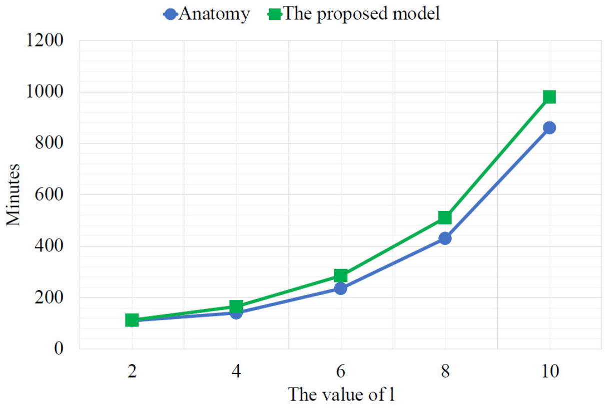

Fig.12: *Execution time based on the value of l*

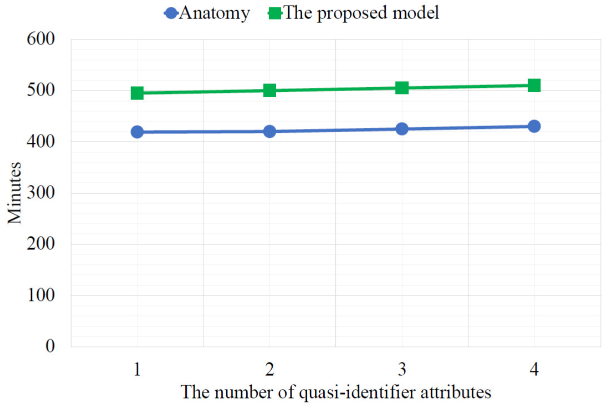

Fig.13: *Execution time based on the number of quasi-identifier attributes*

increased. Also, it is the influences of the data search spaces, i.e., it is when the value of l is increased, the data search space is also increased.

The final experiment evaluates the effect of the number of quasi-identifier attributes that influence the execution time for transforming the raw datasets to satisfy privacy preservation constraints. For experiments, the number of quasi-identifier attributes varies from 1 to 4. Also, all experimental datasets used in this experiment are set to be the same as the experimental datasets which are used in the second experiment.

From the experimental results as in Figure 13 show that when the number of quasi-identifier attributes is increased, the execution time for transforming the raw dataset to the satisfaction of the given experimental privacy preservation constraints is also increased. However, the number of quasi-identifier attributes has less of an effect on the execution time for transforming the raw dataset to the satisfaction of the given experimental privacy preservation constraints than the size of datasets and the value of l. That is because the experimental algorithm of the proposed model and Anatomy is based on tuple partitioning. Thus, their execution times depended on the size of datasets and the value of l (the number of partitions) more than the number of quasi-identifier attributes.

6. CONCLUSION

This work enumerates and explains the vulnerabilities of anatomization models, i.e., Full right coverage data attack ($iFRCA$), Merged right coverage data attack ($iMRCA$), Merged right fully covers merged left data attack ($iMRcMLA$), Full left coverage data attack ($dFLCA$), Merged left coverage data attack ($dMLCA$), Merged left fully covers merged reft data attack ($dMLcMRA$), privacy violation issues in anatomization tables based on data modifications (SVM), and privacy violation issues in anatomization tables based on partition changing. To rid privacy violation issues that are based on the explained vulnerabilities of anatomization models, a new privacy preservation model is proposed in this work. Moreover, we show experimental results that can indicate anatomized datasets that satisfy the proposed model are more secure in terms of privacy preservation than anatomized datasets that are satisfied by the comparative anatomization model.

7. FUTURE WORK

Although the proposed model can address privacy violation issues in anatomization tables from attackers using all attacks that are explained in this work, an adversary could discover a new approach that can be used to violate the privacy data that is available in anatomization tables in the future. Thus, an appropriate privacy preservation model that addresses newly discovered privacy violation issues should also be proposed in the future.

8. COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interest Author declares that they have no conflict of interest. **Ethical approval** This paper does not contain any studies with human participants or animals performed by any of the authors.

References

[1] J.W. Byun, Y. Sohn, E. Bertino and N. Li, “Secure anonymization for incremental datasets,” *Secure Data Management. SDM 2006. Lecture Notes in Computer Science*, vol. 4165, pp. 48–63, 2006.

[2] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov and M. Naor, “Our data, ourselves: Privacy via distributed noise generation,” *Annual International Conference on the Theory and Applications of Cryptographic Techniques - Advances in Cryptology (EUROCRYPT 2006)*, vol. 4004, pp. 486–503, 2006.

[3] G. Fung, “A comprehensive overview of basic clustering algorithms,” 2001.

[4] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang and B. Shi, "Permutation anonymization: Improving anatomy for privacy preservation in data publication," *New Frontiers in Applied Data Mining. PAKDD 2011. Lecture Notes in Computer Science*, vol.7104, pp. 111–123, 2012.

[5] V.S. Iyengar, "Transforming data to satisfy privacy constraints," *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 279–288, 2002.

[6] M.A. Kadampur and S. D.V.L.N, "A noise addition scheme in decision tree for privacy preserving data mining," *Journal of computing*, vol.2, no. 1, pp.137-144, 2010.

[7] S. Kim and Y.D. Chung, "An anonymization protocol for continuous and dynamic privacy-preserving data collection," *Future Generation Computer Systems*, vol.93, pp. 1065–1073, 2019.

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

[9] B. Korte and L. Lovász, "Mathematical structures underlying greedy algorithms," *Fundamentals of Computation Theory. FCT 1981. Lecture Notes in Computer Science*, vol. 117, pp. 205–209, 1981.

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

[11] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol.1, no.1, pp. 3–es, 2007.

[12] K. Mivule,: Utilizing noise addition for data privacy, an overview (2013)

[13] N. Riyana, S. Riyana, S. Nanthachumphu, S. Sittisung and D. Duangban, "Privacy violation issues in re-publication of modification datasets," *Intelligent Computing and Optimization. ICO 2020. Advances in Intelligent Systems and Computing*, pp. 938–953, 2021.

[14] S. Riyana, "(lp1,. . .,lpn)-privacy: privacy preservation models for numerical quasi-identifiers and multiple sensitive attributes," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2021.

[15] S. Riyana, N. Harnsamut, U. Sadjapong, S. Nanthachumphu and N. Riyana, "Privacy preservation for continuous decremental data publishing," *Image Processing and Capsule Networks. ICIPCN 2020. Advances in Intelligent Systems and Computing*, vol. 1200, pp. 233–243, 2021.

[16] 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.

[17] S. Riyana, S. Nanthachumphu and N. Riyana, "Achieving privacy preservation constraints in missing-value datasets," *SN Computer Science*, vol.1, no.4, pp.227, 2020.

[18] S. Riyana and J. Natwichai, "Privacy preservation for recommendation databases," *Service Oriented Computing and Applications*, vol.12, no.(3–4), pp.259–273, 2018.

[19] S. Riyana and N. Riyana, "A privacy preservation model for RFID data-collections is highly secure and more efficient than LKC-privacy," *IAIT2021: The 12th International Conference on Advances in Information Technology*, no.15, pp. 1–11, 2021.

[20] 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.

[21] S. Riyana, N.Riyana and S. Nanthachumphu, "An effective and efficient heuristic privacy preservation algorithm for decremental anonymization datasets," in *Chen, J.IZ., Tavares, J.M.R.S., Shakya, S., Iliyasu, A.M. (eds) Image Processing and Capsule Networks. ICIPCN 2020. Advances in Intelligent Systems and Computing*, vol.1200 , pp. 244–257, 2021.

[22] S. Riyana, N. Riyana and S. Nanthachumphu, "Privacy preservation techniques for sequential data releasing," *IAIT2021: The 12th International Conference on Advances in Information Technology*, no. 24, 2021.

[23] 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.

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

[25] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 19 pages, 1998.

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

[27] 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 publish-

ing," *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 754–759, 2006.

[28] X. Xiaokui and Y. Tao, "Anatomy: Simple and effective privacy preservation," *Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB '06*, pp. 139–150, 2006.

[29] Yong Wang and J. Hodges, "Document clustering with semantic analysis," *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, vol.3, pp. 54c–54c, 2006.

[30] X. Zhang, C. Liu, S. Nepal and J. Chen, "An efficient quasi-identifier index-based approach for privacy preservation over incremental data sets on cloud," *Journal of Computer and System Sciences*, vol.79, no.5, pp. 542–555, 2013.

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.