Privacy Threats and Privacy Preservation in Multiple Data Releases of High-Dimensional Datasets

Surapon Riyana *

Posted Date: 15 August 2025

doi: 10.20944/preprints202508.1159.v1

Keywords: privacy preservation models; privacy violation issues; independent data releases; high-dimensional datasets; multiple sensitive attributes; data comparison attacks

 doi:10.20944/preprints202508.1159.v1

Article

Privacy Threats and Privacy Preservation in Multiple Data Releases of High-Dimensional Datasets

Surapon Riyana

School of Renewable Energy, Maejo University, Thailand; surapon_r@mju.ac.th

Abstract

Determining how to balance data utilities and data privacy when datasets are released to be utilized outside the scope of data-collecting organizations constitutes a major challenge. To achieve this aim in data collection (datasets), several privacy preservation models have been proposed, such as k-Anonymity and l-Diversity. Unfortunately, these privacy preservation models may be insufficient to address privacy violation issues in datasets that do not have high-dimensional attributes. For this reason, we propose a privacy preservation model, LKC-Privacy, to address privacy violation issues in high-dimensional datasets. With this model, these issues are not a concern, as all L-size distinct quasi-identifier values are distorted (suppressed or generalized) to be at least K indistinguishable tuples. Moreover, every protected sensitive value relates to each group of indistinguishable quasi-identifier values; it must have sufficient re-identification confidence to be at most C. Although LKC-Privacy is more efficient and effective than k-Anonymity and l-Diversity, it is generally not efficient or effective in addressing privacy violation issues in location-based datasets. Moreover, while datasets satisfy LKC-Privacy constraints, they still exhibit privacy violation issues from using data comparison attacks, and they also have data utility issues that must be addressed. Therefore, our proposed privacy preservation model can address privacy violation issues in high-dimensional datasets, such that there are no concerns about privacy violations in its released datasets from using data comparison attacks, and it is highly efficient and effective in data maintenance. Furthermore, we show that the proposed model is more efficient and effective than k-Anonymity and l-Diversity through extensive experiments.

Keywords: privacy preservation models; privacy violation issues; independent data releases; high-dimensional datasets; multiple sensitive attributes; data comparison attacks

1. Introduction

User privacy violation is a serious issue that data holders must consider when collected data is released to be utilized outside the scope of data-collecting organizations [1–9]. Therefore, in this paper, *k*-Anonymity [10] is proposed to address this issue before datasets are released. All explicit user identifier values are available in the datasets to be removed. Moreover, users' unique quasi-identifier values are suppressed or generalized by their less-specific values to be at least *k* indistinguishable tuples.

Here, we provide an example of privacy preservation based on *k*-Anonymity constraints in conjunction with data generalization [10,11]. With *k*-Anonymity, the parameter *k* represents the privacy preservation consultant such that it is represented by a positive integer that is equal to or greater than 2, i.e., $k \in I^+$ and $k \geq 2$. Suppose that *k* is set to 2, i.e., $k = 2$. Let Table 1 be the raw dataset. In Table 1, there are two explicit identifier attributes (*SSN* and *Name*), three quasi-identifier attributes (*Age*, *Gender*, and *Zipcode*), and a sensitive attribute (*Disease*). For privacy preservation, the *SSN* and *Name* data are removed in the first step. Finally, all unique quasi-identifier values are generalized by their less-specific values to be at least two indistinguishable tuples. The resulting released version of the data from Table 1 is shown in Table 2.

 doi:10.20944/preprints202508.1159.v1

Table 1. An example of a raw dataset.

SSN	Name	Age	Gender	Zip code	Disease
000-00-0001	Jacob	45	Male	60636	Flu
000-00-0002	Jessica	46	Female	60632	Fever
000-00-0003	David	47	Male	60635	Cancer
000-00-0004	Bob	48	Male	60639	Cancer
000-00-0005	Amelia	48	Female	60632	Flu
000-00-0006	Sophia	42	Female	60632	HIV
000-00-0007	Isabella	42	Female	60632	Fever

Table 2. The released version of the data in Table 1, which satisfies 2-Anonymity constraints.

Age	Gender	Zip code	Disease	EC
45 - 46	*	6063*	Flu	ec_1
45 - 46	*	6063*	Fever	
47 - 48	Male	6063*	Cancer	ec_2
47 - 48	Male	6063*	Cancer	
42 - 48	Female	60632	Flu	ec_3
42 - 48	Female	60632	HIV	
42 - 48	Female	60632	Fever	

Table 2 shows that all possible data utilization conditions, owing to their quasi-identifier attributes, always have at least two tuples that are satisfied. In this situation, privacy violation in Table 2 seems impossible. Unfortunately, in [12], the authors demonstrate that Table 2 still has privacy violation issues that must be addressed. For example, let *Bob* be the target user such that *Bob*'s disease is the target data of the adversary. We assume that the adversary believes the user profile tuple in Table 2 to be *Bob*'s. Moreover, the adversary knows that *Bob* is a 48-year-old male. Thus, the adversary can be confident based on Table 2 that *Bob* has *Cancer*. Although two user profile tuples match the adversary's knowledge about *Bob*, the adversary can see that only *Cancer* is shown in the *Disease* attribute of these tuples. From this example, we can conclude that although the released datasets guarantee that all possible data utilization conditions, owing to their quasi-identifier attributes, always have at least k satisfied tuples, they still have privacy violation issues that must be addressed. To address this vulnerability of k-Anonymity, l-Diversity [12] is proposed. For privacy preservation with l-Diversity, in addition to removing the explicit identifier values and distorting (suppressing or generalizing) all unique quasi-identifier values, the number of distinct sensitive values available in each sensitive attribute is also considered, i.e., every group of indistinguishable quasi-identifier values must relate to at least l different sensitive values, where $l \in I^+$ and $l \geq 2$, in every sensitive attribute.

Here, an example of privacy preservation is given based on l-Diversity, where Table 1 is the raw dataset. For privacy preservation, let the value of l be set to 2. The released version of the data from Table 1 satisfying the 2-Diversity constraints is shown in Table 3. In Table 3, it is guaranteed that all possible data utilization conditions, owing to the quasi-identifier attributes, always have at least l different sensitive values to be satisfied. Therefore, we can conclude that l-Diversity is more secure in terms of privacy preservation than k-Anonymity.

 doi:10.20944/preprints202508.1159.v1

Table 3. The released version of the data from Table 1, which satisfies 2-Diversity constraints.

Age	Gender	Zip code	Disease	EC
45 - 48	*	6063*	Flu	ec_1
45 - 48	*	6063*	Fever	
45 - 48	*	6063*	Cancer	
45 - 48	*	6063*	Cancer	
42 - 48	Female	60632	Flu	ec_2
42 - 48	Female	60632	HIV	
42 - 48	Female	60632	Fever	

Unfortunately, to the best of our knowledge about l-Diversity, it is generally insufficient to address privacy violation issues in datasets that have high-dimensional attributes and are independently released when new data becomes available [13–17]. To eliminate these vulnerabilities in l-Diversity in high-dimensional datasets, k^m-Anonymity [18] and LKC-Privacy [19–21] are proposed. These privacy preservation models assume that the adversary has limited background knowledge about the target user. That is, in terms of the adversary's background knowledge about the target user the m and L values are limited by k^m-Anonymity and LKC-Privacy, respectively, so m or L sizes of the unique quasi-identifier values are suppressed or generalized to be k or K indistinguishable tuples. However, the effectiveness of these privacy preservation models is questioned, because they are based on an estimation of the adversary's level of background knowledge about the target user, and they are inadequate for addressing privacy violation issues in datasets that do not allow the quasi-identifier attributes to determine $NULL$ or the empty value. Moreover, these privacy preservation models can also be inadequate for addressing privacy violation issues in datasets that are independently released and have multiple sensitive attributes [22–28].

To address privacy violation issues in datasets that have multiple sensitive attributes, two well-known privacy preservation models have been proposed, i.e., aggregate query frameworks [29–31] and data anonymization models for multiple sensitive attributes [32–34]. For privacy preservation with aggregate query frameworks, the data analyst is not allowed to utilize the data available in datasets directly, i.e., they can only utilize the data via aggregate query frameworks. Another well-known privacy preservation solution is distorting users' unique quasi-identifier values in datasets to be indistinguishable. Moreover, the number of distinct sensitive values and re-identifiable sensitive values in every sensitive attribute of each group of indistinguishable quasi-identifier values are also considered when establishing privacy preservation constraints.

For preserving data privacy in independently released datasets, in [26,35], the authors recommend that, in addition to releasing datasets that satisfy privacy preservation constraints, all results that could be compared between the released and original datasets must also satisfy the privacy preservation constraints.

In addition, to the best of our knowledge about the above-mentioned privacy preservation models, they have serious vulnerabilities that must be improved. That is, they are inadequate for addressing privacy violation issues in datasets that have high-dimensional quasi-identifier attributes, multiple sensitive attributes, and independent data releases. These vulnerabilities in these privacy preservation models will be explained in Section 2.

This paper is organized as follows. The motivation for this work is explained in Section 2. Then, our privacy preservation model for high-dimensional datasets is proposed in Section 3. Subsequently, the experimental results are discussed in Section 4. Finally, our conclusion and directions for future work in this field are discussed in Sections 5 and 6, respectively.

2. Motivation

Before we explain the motivation for this work, we present the necessary definitions.

 doi:10.20944/preprints202508.1159.v1

Definition 1 (High-dimensional datasets). *Let $QI = \{qi_1, qi_2, \ldots, qi_p\}$ be the set of quasi-identifier attributes. Let $DO^{qi_r} = \{do_1^{qi_r}, do_2^{qi_r}, \ldots, do_v^{qi_r}\}$ be the data domain of $qi_r \in QI$, where $1 \leq r \leq p$. Let $S = \{s_1, s_2, \ldots, s_q\}$ be the set of sensitive attributes. Let $DO^{s_o} = \{do_1^{s_0}, do_2^{s_0}, \ldots, do_w^{s_0}\}$ be the data domain of $s_o \in S$, where $1 \leq o \leq q$. Let $D = \{d_1, d_2, \ldots, d_n\}$ be the high-dimensional dataset. Every d_i of D, where $1 \leq i \leq n$, represents the profile tuple of the user u_i such that it is in the form $QI \cup S$, i.e., $d_i = (do_\omega^{qi_1}, do_\psi^{qi_2}, \ldots, do_\varphi^{qi_p}, do_\varrho^{s_1}, do_\xi^{s_2}, \ldots, do_\vartheta^{s_q})$, where $do_\omega^{qi_1} \in DO^{qi_1}$, $do_\psi^{qi_2}, \in DO^{qi_2}$, $do_\varphi^{qi_p} \in DO^{qi_p}$, $do_\varrho^{s_1} \in DO^{s_1}$, $do_\xi^{s_2} \in DO^{s_2}$, and $do_\vartheta^{s_q} \in DO^{s_q}$. Let $D^{\Gamma \cup \Delta}$ be a sub-data version of D such that it is constructed from $\Gamma \cup \Delta$, where $\Gamma \subseteq QI$ and $\Delta \subseteq S$, i.e., $\Gamma = \{qi_{r_1}, qi_{r_2}, \ldots, qi_{r_p}\} \subseteq QI$ and $\Delta = \{s_{o_1}, s_{o_2}, \ldots, s_{o_q}\} \subseteq S$. Thus, every $d_i^{\Gamma \cup \Delta}$ of $D^{\Gamma \cup \Delta}$, where $1 \leq i \leq n$, is in the form $(do_\omega^{qi_{r_1}}, do_\psi^{qi_{r_2}}, \ldots, do_\varphi^{qi_{r_p}}, do_\varrho^{s_{o_1}}, do_\xi^{s_{o_2}}, \ldots, do_\vartheta^{s_{oq}})$. The data projection on Γ and Δ of $D^{\Gamma \cup \Delta}$ is D^Γ and D^Δ, respectively. The data projection on qi_{r_β} and s_{o_α} of $D^{\Gamma \cup \Delta}$ is $D^\Gamma[qi_{r_\beta}]$ and $D^\Delta[s_{o_\alpha}]$, respectively.*

Definition 2 (Data generalization hierarchy). *Let $f_{DGH}(DO^{qi_r}) : DO_\zeta^{qi_r} \rightarrow DO_{\zeta+1}^{qi_r}$ be the generalized function of DO^{qi_r} from the level ζ to $\zeta + 1$ such that all values of the level ζ are more specific than their related values at the level $\zeta + 1$. Therefore, we can write the data generalization sequence of DO^{qi_r}, $DGH_{DO^{qi_r}}$, from the level 0 to L, as $DO_0^{qi_r} \xrightarrow{f_{DGH}(DO_0^{qi_r})} DO_1^{qi_r} \xrightarrow{f_{DGH}(DO_1^{qi_r})} DO_2^{qi_r} \ldots DO_{L-2}^{qi_r} \xrightarrow{f_{DGH}(DO_{L-2}^{qi_r})} DO_{L-1}^{qi_r} \xrightarrow{f_{DGH}(DO_{L-1}^{qi_r})} DO_L^{qi_r}$. That is, all values at the level 0 are more specific than at other levels, and the values at level L are less specific than at other levels.*

Definition 3 (Data generalization). *Let Λ be the set of specified quasi-identifier values in qi_r of D. The meaning of data generalization is that Λ is distorted by an appropriately less-specific value that is presented by $DGH_{DO^{qi_r}}$ as indistinguishable.*

Definition 4 (Data suppression). *Let d_i be an arbitrary tuple of D. The meaning of data suppression is that d_i is not available in the released version of the data in D.*

Definition 5 (Adversary's background knowledge about the target user). *Let $G_{u_i} = \{g_1, g_2, \ldots, g_b\}$ be the information about the user u_i. If B_{u_i} is the adversary's background knowledge about the user u_i in D, B_{u_i} must satisfy both limitations as follows: $B_{u_i} \subseteq G_{u_i}$ and $B_{u_i} \subseteq d_i[QI]$.*

Definition 6 (Privacy violation concerns). *Let l be a positive integer such that it is equal to or greater than two, i.e., I^+ and $I \geq 2$. Let $s_o \in S$ be a specific sensitive attribute. If the number of distinct sensitive values available in s_o is at most $l - 1$, s_o has privacy violation concerns that must be addressed.*

Definition 7 (l-Diversity). *Let l be the privacy preservation constraint. Let $f_A(l, D, DGH_{DO^{qi_1}}, DGH_{DO^{qi_2}}, \ldots, DGH_{DO^{qi_p}}) : D \rightarrow_{l, DGH_{DO^{qi_1}}, DGH_{DO^{qi_2}}, \ldots, DGH_{DO^{qi_p}}} \mathbb{D}$ be the function for transforming D into $\mathbb{D}$. That is, all unique quasi-identifier values are available in QI of $\mathbb{D}$, and they are suppressed or generalized by their less-specific values in $DGH_{DO^{qi_1}}, DGH_{DO^{qi_2}}, \ldots, DGH_{DO^{qi_p}}$ to be indistinguishable. Moreover, every group of indistinguishable quasi-identifier tuples must be related to at least l distinct sensitive values of every sensitive attribute $s_o \in S$. In addition, every group of indistinguishable quasi-identifier tuples satisfies l-Diversity constraints, and they are referred to as equivalence classes ec of $\mathbb{D}$. Hence, $\mathbb{D}$ can be presented by the set of its equivalence classes, i.e., $EC = \{ec_1, ec_2, \ldots, ec_a\}$.*

Here, we give an example of privacy preservation based on l-Diversity constraints in datasets that have multiple sensitive attributes. Let Table 4 be the raw dataset D. Let the value of l be set to 2. With these instances, the released version of the data $\mathbb{D}$ in Table 4 is satisfies the 2-Diversity constraints, as shown in Table 5. In Table 5, there are three equivalence classes, i.e., ec_1, ec_2, and ec_3. Moreover, Table 5 guarantees that all possible data utilization conditions, owing to the quasi-identifier attributes, always have at least l distinctly satisfied sensitive values in every sensitive attribute. However, Table 5 has

 doi:10.20944/preprints202508.1159.v1

data utility issues that must be addressed, i.e., the meaning of the data in Table 5 is much less clear than in Table 4.

Table 4. An example of raw datasets that have high-dimensional quasi-identifiers and sensitive attributes.

Position	Education	...	Age	Gender	Zip code	Disease	Salary	...
Accounting	Bachelor	...	45	Male	60636	Flu	USD 10000	...
Accounting	Master	...	46	Female	60632	Flu	USD 13000	...
Programmer	Doctor	...	47	Male	60635	Cancer	USD 14000	...
Programmer	Master	...	48	Male	60639	Cancer	USD 15000	...
Lecturer	Doctor	...	48	Female	60632	Flu	USD 16000	...
Lecturer	Doctor	...	42	Female	60632	HIV	USD 17000	...
Lecturer	Master	...	42	Female	60632	Fever	USD 18000	...

Table 5. The released version of the data in Table 4, which satisfies 2-Diversity constraints.

Position	Education	...	Age	Gender	Zip code	Disease	Salary	...	EC
*	*	...	45 - 47	*	6063*	Flu	USD 10000	...	ec_1
*	*	...	45 - 47	*	6063*	Flu	USD 13000	...	
*	*	...	47 - 47	*	6063*	Cancer	USD 14000	...	
*	*	...	48	*	6063*	Cancer	USD 15000	...	ec_2
*	*	...	48	*	6063*	Flu	USD 16000	...	
Lecturer	*	...	42	Female	60632	HIV	USD 17000	...	ec_3
Lecturer	*	...	42	Female	60632	Fever	USD 18000	...	

2.1. Data Utility Issues

2.1.1. Data Utility Issues Based on the Number of *QI* Attributes

Let the value of l be set to 2, and let Table 4 without *Disease* be the raw dataset D. For these instances, the released version of the data $\mathbb{D}$ in Table 4 satisfies the 2-Diversity constraints, as shown in Table 6. In another situation, let Table 4 without *Education*, *Age*, *Zipcode*, and *Disease* be the raw dataset D. For these instances, the released version of the data $\mathbb{D}$ in Table 4 satisfies the 2-Diversity constraints, as shown in Table 7. In Tables 6 and 7, we can see that Table 7 utilizes more data than Table 6. Therefore, we can conclude that the number of quasi-identifier attributes directly influences the data utility of $\mathbb{D}$.

 doi:10.20944/preprints202508.1159.v1

Table 6. The released version of the data in Table 4 without *Disease*, which satisfies 2-Diversity constraints.

Position	Education	Age	Gender	Zip code	Salary	EC
Accounting	*	45 - 46	*	6063*	USD 10000	ec_1
Accounting	*	45 - 46	*	6063*	USD 13000	
*	*	47 - 48	*	6063*	USD 14000	ec_2
*	*	47 - 48	*	6063*	USD 15000	
*	*	47 - 48	*	6063*	USD 16000	
Lecturer	*	42	Female	60632	USD 17000	ec_3
Lecturer	*	42	Female	60632	USD 18000	

Table 7. The released version of the data in Table 4 without *Education*, *Age*, *Zipcode*, and *Disease*, which satisfies 2-Diversity constraints.

Position	Gender	Salary	EC
Accounting	*	USD 10000	ec_1^7
Accounting	*	USD 13000	
Programmer	Male	USD 14000	ec_2^7
Programmer	Male	USD 15000	
Lecturer	Female	USD 16000	ec_3^7
Lecturer	Female	USD 17000	
Lecturer	Female	USD 18000	

2.1.2. Data Utility Issues Based on the Number of S Attributes

Let the value of l be set to 2, and let Table 4 without *Education*, *Age*, and *Zipcode* be the raw dataset D. For these instances, the released version of the data $\mathbb{D}$ in Table 4 satisfies the 2-Diversity constraints, as shown in Table 8. In Tables 7 and 8, we can see that Table 7 utilizes more data than Table 8. Therefore, we can conclude that the number of sensitive attributes also influences the data utility of $\mathbb{D}$.

Table 8. The released version of the data in Table 4 without *Education*, *Age*, and *Zipcode*, which satisfies 2-Diversity constraints.

Position	Gender	Disease	Salary	EC
*	*	Flu	USD 10000	ec_1
*	*	Flu	USD 13000	
*	*	Cancer	USD 14000	
*	*	Cancer	USD 15000	
Lecturer	Female	Flu	USD 16000	ec_2
Lecturer	Female	HIV	USD 17000	
Lecturer	Female	Fever	USD 18000	

Based on Sections 2.1.1 and 2.1.2, it is clear that the number of QI and S attributes influences the data utility of $\mathbb{D}$. For this reason, only the utilized QI and S attributes of D should be available in $\mathbb{D}$ in D. For example, let Table 4 be the raw dataset D. Suppose that if the data holder only needs to show a statistical report of employee salaries based on gender and position, only *Position*, *Gender*, and *Salary* are available in the released version of the data $\mathbb{D}$ in Table 4. Although this privacy preservation solution addresses the data utility issues of $\mathbb{D}$, it often leads to privacy violation from data comparison attacks when the adversary has background knowledge about the target user and has received enough released versions of the data in D.

 doi:10.20944/preprints202508.1159.v1

2.2. Privacy Violation from Data Comparison Attacks

In this section, a data privacy attack (violation) on independently released datasets D using data comparison is demonstrated. It is based on the assumption that datasets can be changed when new data becomes available. Moreover, we assume that the adversary has received the corresponding released version of the data in D. The adversary believes that the tuple of the received datasets is the profile tuple of the target user. In addition, we assume that the adversary has enough background knowledge about the target user, and that the sensitive attribute targeted by the adversary is available in all received datasets. Privacy violation is considered to have occurred when the comparison result of the targeted sensitive attribute in the received datasets does not satisfy the given value of l.

Let $D^{\Gamma_x \cup \Delta_x}$ be a sub-data version of D such that it is constructed from $\Gamma_x \subseteq QI$ and $\Delta_x \subseteq S$. Let $D^{\Gamma_y \cup \Delta_y}$ be a sub-data version of D such that it is constructed from $\Gamma_y \subseteq QI$ and $\Delta_y \subseteq S$, where $\Gamma_x \cap \Gamma_y \neq \emptyset$, $|\Gamma_x \cup \Gamma_y| < (|\Gamma_x| + |\Gamma_y|)$, and $\Delta_x \cap \Delta_y \neq \emptyset$. For privacy preservation, let $f_A(l, D^{\Gamma_x \cup \Delta_x}, DGH_{DO^{qi_{x_1}}}, DGH_{DO^{qi_{x_2}}}, \ldots, DGH_{DO^{qi_{x_p}}}): D^{\Gamma_x \cup \Delta_x} \rightarrow_{l, DGH_{DO^{qi_{x_1}}}, DGH_{DO^{qi_{x_2}}}, \ldots, DGH_{DO^{qi_{x_p}}}} \mathbb{D}^{\Gamma_x \cup \Delta_x}$, where $qi_{x_1}, qi_{x_2}, \ldots, qi_{x_p} \in \Gamma_x$, be the function for transforming $D^{\Gamma_x \cup \Delta_x}$ into $\mathbb{D}^{\Gamma_x \cup \Delta_x}$. Let $f_A(l, D^{\Gamma_y \cup \Delta_y}, DGH_{DO^{qi_{y_1}}}, DGH_{DO^{qi_{y_2}}}, \ldots, DGH_{DO^{qi_{y_p}}}): D^{\Gamma_y \cup \Delta_y} \rightarrow_{l, DGH_{DO^{qi_{y_1}}}, DGH_{DO^{qi_{y_2}}}, \ldots, DGH_{DO^{qi_{y_p}}}} \mathbb{D}^{\Gamma_y \cup \Delta_y}$, where $qi_{y_1}, qi_{y_2}, \ldots, qi_{y_p} \in \Gamma_y$, be the function for transforming $D^{\Gamma_y \cup \Delta_y}$ into $\mathbb{D}^{\Gamma_y \cup \Delta_y}$. That is, $\mathbb{D}^{\Gamma_x \cup \Delta_x}$ and $\mathbb{D}^{\Gamma_y \cup \Delta_y}$ are satisfied by Definition 7. Let $s_o \in (\Delta_x \cap \Delta_y)$ be the sensitive attribute targeted by the adversary such that it is available in $D^{\Gamma_x \cup \Delta_x}$ and $D^{\Gamma_y \cup \Delta_y}$. Let u_i be the target user of the adversary. Let B_{u_i} be the adversary's background knowledge about the target user u_i. Let the values in $ec_{z_1}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[\Gamma_x] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[\Gamma_x]$ and $ec_{z_1}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[\Gamma_y] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[\Gamma_y]$ match those in B_{u_i}. Moreover, $ec_{z_1}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[\Gamma_x], \ldots, ec_{z_c}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[\Gamma_x]$ and $ec_{z_1}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[\Gamma_y], \ldots, ec_{z_c}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[\Gamma_y]$ without data generalities satisfy the limitations $(ec_{z_1}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[\Gamma_x] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[\Gamma_x]) \subset (ec_{z_1}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[\Gamma_y] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[\Gamma_y])$ and $|(ec_{z_1}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[s_o] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[s_o]) - (ec_{z_1}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[s_o] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[s_o])| < l - 1$. Therefore, the targeted value of the target user u_i is available in s_o in $\mathbb{D}^{\Gamma_x \cup \Delta_x}$ and $\mathbb{D}^{\Gamma_y \cup \Delta_y}$, and it can be revealed by $(ec_{z_1}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[s_o] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_y \cup \Delta_y}}[s_o]) - (ec_{z_1}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[s_o] \cup \ldots \cup ec_{z_c}^{\mathbb{D}^{\Gamma_x \cup \Delta_x}}[s_o])$.

Here, we provide an example of privacy violation issues in an independently released dataset D from data comparison attacks. Let Table 4 be the raw dataset. Let the value of l be set to 2. Let Table 7 be the released version of the data in a sub-table of Table 4 such that it is constructed from *Position*, *Gender*, and *Salary*, and it further satisfies the 2-Diversity constraints. Moreover, let Table 9 be the released version of the data in a sub-table of Table 4 such that it is constructed from *Gender*, *Zipcode*, and *Salary*. In addition, Table 9 satisfies the 2-Diversity constraints. Suppose that the adversary has received Tables 7 and 9. Let ***John*** be the target user of the adversary such that they strongly believe that the tuple in Tables 7 and 9 is *John*'s profile tuple. Moreover, the adversary knows that ***John* is a male accountant**. Furthermore, we assume that the adversary needs to reveal ***John*'s salary**, which is given in Tables 7 and 9. In this situation, the adversary can be confident that the tuple of ec_1^7 and ec_1^9 must be *John*'s profile tuple because only the quasi-identifier values of these equivalence classes match the adversary's background knowledge about *John*. Moreover, the adversary can see that ec_1^9 relates to ec_1^7 and ec_2^7, and they also see that the ec_2^9 relates to ec_1^7 and ec_3^7. Therefore, the adversary can infer from Tables 7 and 9 that **USD 10,000** is ***John*'s salary**. The data relationships between Tables 7 and 9 are shown in Figure 1.

 doi:10.20944/preprints202508.1159.v1

Table 9. The released version of the data in Table 4 without *Position*, *Education*, *Age*, and *Disease*, which satisfies 2-Diversity constraints.

Gender	Zip code	Salary	EC
Male	6063*	USD 10000	ec_1^9
Male	6063*	USD 14000	
Male	6063*	USD 15000	
Female	60632	USD 13000	ec_2^9
Female	60632	USD 16000	
Female	60632	USD 17000	
Female	60632	USD 18000	

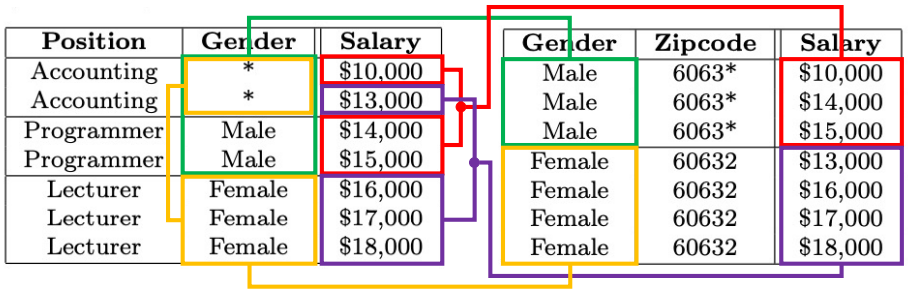

Figure 1. An example of the data relationships between Tables 7 and 9

Based on Sections 2.1 and 2.2, it is clear that although the datasets satisfy l-Diversity constraints, they still have two serious issues that must be addressed, i.e., data utility and privacy violation. To eliminate these vulnerabilities in l-Diversity, we propose a new extended privacy preservation model of l-Diversity. It will be presented in Section 3.

3. The Proposed Model

In this section, we describe a new privacy preservation model that can address privacy violation issues in high-dimensional datasets that are allowed to change and independently release data when new data becomes available, such that released datasets are not susceptible to privacy violation from data comparison attacks.

3.1. Privacy preservation in high-dimensional datasets

Let l be a privacy preservation constraint represented by a positive integer that is equal to or greater than two. Let $D^{\Gamma_j \cup \Delta_j}$ be the specific raw dataset that is released at the timestamp j. Let $\mathbb{D}^{\Gamma_1 \cup \Delta_1}$, $\mathbb{D}^{\Gamma_2 \cup \Delta_2}$, ..., $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ be the released versions of the data D such that they relate to $D^{\Gamma_j \cup \Delta_j}$ and are released from the timestamp 1 to $j-1$. For privacy preservation, let $f_A^j(l, D^{\Gamma_j \cup \Delta_j}, \mathbb{D}^{\Gamma_1 \cup \Delta_1}, \mathbb{D}^{\Gamma_2 \cup \Delta_2}, \ldots, \mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}, DGH_{do^{qir_1}}, DGH_{do^{qir_2}}, \ldots, DGH_{do^{qir_p}}) : D^{\Gamma_j \cup \Delta_j} \rightarrow_{l, \mathbb{D}^{\Gamma_1 \cup \Delta_1}, \mathbb{D}^{\Gamma_2 \cup \Delta_2}, \ldots, \mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}, DGH_{do^{qir_1}}, DGH_{do^{qir_1}}, \ldots, DGH_{do^{qir_p}}} \mathbb{D}^{\Gamma_j \cup \Delta_j}$ be the function for transforming $D^{\Gamma_j \cup \Delta_j}$ into $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. That is, all unique quasi-identifier values are suppressed or generalized by their less-specific values in $DGH_{do^{qir_1}}$, $DGH_{do^{qir_2}}$, ..., $DGH_{do^{qir_p}}$ to be indistinguishable. Moreover, every group of indistinguishable quasi-identifier tuples must relate at least l distinct sensitive values in every $s_{o_z} \in \Delta_j$. In addition, every group of tuples in $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is then satisfied by the given value of l to call an equivalence class ec^j of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. Thus, we can say that $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is the set of its equivalence classes, i.e., $EC^j = \{ec_1^j, ec_2^j, \ldots, ec_a^j\}$. In addition to suppressing or generalizing all unique quasi-identifiers and considering the number of distinct sensitive values, the comparison result between the sensitive values in $ec_z^j[s_{o_\varkappa}]$ and every related $ec_z^t[s_{o_\varkappa}]$ in EC^t, where $1 \leq t \leq (j-1)$, must also be satisfied by the given value of l, i.e., $|ec_z^j[s_{o_\varkappa}] - ec_z^t[s_{o_\varkappa}]| \geq l$.

 doi:10.20944/preprints202508.1159.v1

In addition, after datasets satisfy the proposed privacy preservation constraint, they are often more secure in terms of privacy preservation than their corresponding raw datasets, but they lose some data utility. Moreover, a given privacy preservation constraint for each dataset generally has various released versions of the data that can be satisfied. For example, let Table 4 without *Education*, *Age*, *Zip code*, and *Disease* be the raw dataset for public use. Suppose that only Table 7 is the previously released version of the data that relates to the specific raw dataset. Let the value of l be set to 2. For these instances, Tables 10 and 11 are both released versions of the data that are not susceptible to privacy violation from data comparison attacks. However, Tables 10 and 11 are different, so they could be different in terms of data utilization. Only the released version of the data has the desired high data utility. Therefore, the data utility metric is a necessity in the proposed model; this will be discussed in Section 3.2.

Table 10. The released version of the data in Table 4 without *Position*, *Education*, *Age*, and *Disease*, which satisfies the proposed privacy preservation constraint, where $l = 2$.

Gender	Zip code	Salary	EC
*	6063*	USD 10000	ec_1
*	6063*	USD 13000	
Male	6063*	USD 14000	ec_2
Male	6063*	USD 15000	
Female	60632	USD 16000	ec_3
Female	60632	USD 17000	
Female	60632	USD 18000	

Table 11. The released version of the data in Table 4 without *Position*, *Education*, *Age*, and *Disease*, which also satisfies the proposed privacy preservation constraint, where $l = 2$.

Gender	Zip code	Salary	EC
*	6063*	USD 10000	ec_1
*	6063*	USD 14000	
*	6063*	USD 15000	
*	6063*	USD 13000	
Female	60632	USD 16000	ec_2
Female	60632	USD 17000	
Female	60632	USD 18000	

3.2. Data Utility Metric

Although $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ satisfies the proposed privacy preservation constraint, it is generally higher in terms of privacy preservation than $D^{\Gamma_j \cup \Delta_j}$. However, $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ often loses some data utility. For this reason, only $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ has sufficiently high data utility. Thus, the data utility metric is a necessity in the proposed privacy preservation model. Since privacy preservation based on data suppression and generalization was established, several data utility metrics have been proposed, e.g., the Precision metric (*PREC*) for data suppression in conjunction with data generalization [36], the Discernibility metric (*DM*) [37], and Relative error [38,39]. These metrics will be explained in Sections 3.2.1, 3.2.2, and 3.2.3, respectively.

3.2.1. Precision Metric (PREC) for Data Suppression in Conjunction with Data Generalization [36]

With the proposed privacy preservation model, $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ can satisfy the privacy preservation constraints using data suppression in conjunction with data generalization. For this reason, $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ has two data penalty costs that must be considered, i.e., the penalty costs of data suppression and data generalization. With data generalization, the penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ depends on the level and the number of generalized values, such that a high level and a larger number of generalized values lead to

 doi:10.20944/preprints202508.1159.v1

a higher penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. Therefore, the penalty cost of data generalization for $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ can be defined by Equation 1, i.e., the penalty cost of data generalization for $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is between 0 and $|\mathbb{D}^{\Gamma_j \cup \Delta_j}| \cdot |\mathbb{D}^{\Gamma_j}|$. A higher penalty cost of $f_{\mathbb{GEN}}$ means that $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is more generalized.

$$f_{\mathbb{GEN}}(\mathbb{D}^{\Gamma_j \cup \Delta_j}) = \sum_{i=1}^{|\mathbb{D}^{\Gamma_j \cup \Delta_j}|} \sum_{r=1}^{|\mathbb{D}^{\Gamma_j}|} \frac{\zeta}{| DGH_{DO^{qi_r}} |} \tag{1}$$

where

- $|\mathbb{D}^{\Gamma_j}|$ is the number of quasi-identifier attributes that are available in $\mathbb{D}^{\Gamma_j \cup \Delta_j}$;
- ζ is the generalized level of the quasi-identifier value that is available in qi_r of d_i;
- $|DGH_{DO^{qi_r}}|$ is the height of the data generalization hierarchy for DO^{qi_r};
- $|\mathbb{D}^{\Gamma_j \cup \Delta_j}|$ is the number of tuples that are available in $\mathbb{D}^{\Gamma_j \cup \Delta_j}$.

Another penalty cost for $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is data suppression, which depends on the number of suppressed tuples and the size of $D^{\Gamma_j \cup \Delta_j}$. Thus, the penalty cost of data suppression for $D^{\Gamma_j \cup \Delta_j}$ can be defined by Equation 2. That is, the penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is between 0 and $|D^{\Gamma_j \cup \Delta_j}|^2$. A higher penalty cost of $f_{\mathbb{SUP}}$ means that $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is more suppressed.

$$\begin{aligned} f_{\mathbb{SUP}}(D^{\Gamma_j \cup \Delta_j}, \mathbb{D}^{\Gamma_j \cup \Delta_j}) = \\ | D^{\Gamma_j \cup \Delta_j} - \mathbb{D}^{\Gamma_j \cup \Delta_j} | \cdot | D^{\Gamma_j \cup \Delta_j} | \end{aligned} \tag{2}$$

Therefore, the total penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ can be defined based on the penalty cost of $f_{\mathbb{GEN}}$ and $f_{\mathbb{SUP}}$, as shown in Equation 3. In addition, a higher penalty cost of $f_{\mathbb{PREC}}$ means that $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ has lower data utility.

$$\begin{aligned} f_{\mathbb{PREC}}(D^{\Gamma_j \cup \Delta_j}, \mathbb{D}^{\Gamma_j \cup \Delta_j}) = \\ f_{\mathbb{GEN}}(\mathbb{D}^{\Gamma_j \cup \Delta_j}) + f_{\mathbb{SUP}}(D^{\Gamma_j \cup \Delta_j}, \mathbb{D}^{\Gamma_j \cup \Delta_j}) \end{aligned} \tag{3}$$

3.2.2. Discernibility Metric (DM) [37]

The DM metric is a data utility metric that can also be used to define the penalty cost or the data utility of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. With the DM metric, the penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ depends on the size of its equivalence classes. That is, it can be defined by Equation 4. The DM penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is between 0 and $|\mathbb{D}^{\Gamma_j \cup \Delta_j}|^2$. A higher DM penalty cost means that $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ has lower data utility.

$$f_{\mathbb{DM}}(EC^j) = \sum_{z=1}^{|EC^j|} | ec_z^j |^2 \tag{4}$$

3.2.3. Relative Error [38,39]

The relative error is the metric that is used to define the penalty cost of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. With this metric, the data utility of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ depends on the difference in the queried results between $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ and its original dataset $D^{\Gamma_j \cup \Delta_j}$. A higher relative error means that $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ has lower data utility. For query results that are represented by numerical data, their relative errors can be defined by Equation 5.

$$f_{\mathbb{REI}}(\nu, \nu_0) = \frac{| \nu - \nu_0 |}{\nu} \tag{5}$$

where

- ν is the result that is queried from $D^{\Gamma_j \cup \Delta_j}$;
- ν_0 is the relative result of ν such that it is queried from $\mathbb{D}^{\Gamma_j \cup \Delta_j}$.

With query results that are not represented by numerical data, their relative errors can be defined by Equation 6.

 doi:10.20944/preprints202508.1159.v1

$$f_{\mathbb{REC}}(n(\nu), n(\nu_0)) = \frac{\mid n(\nu) - n(\nu_0) \mid}{n(\nu)} \tag{6}$$

where

- $n(\nu)$ is the number of values that are queried from $D^{\Gamma_j \cup \Delta_j}$;
- $n(\nu_0)$ is the number of relative values of $n(\nu)$ such that they are queried from $\mathbb{D}^{\Gamma_j \cup \Delta_j}$.

3.3. The Proposed Algorithm

In this section, a new privacy preservation algorithm, l^{HD}-*Diversity*(l, $D^{\Gamma_j \cup \Delta_j}$, $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$, $DGH_{do^{qi}r_1}$, $DGH_{do^{qi}r_2}$, ..., $DGH_{do^{qi}r_v}$), is presented that can address privacy violation issues in high-dimensional datasets that are allowed to change and independently release data when new data become available. With the proposed algorithm, in addition to privacy preservation, the data utility and exclusion time are also maintained where possible. To achieve the aims of the proposed algorithm, greedy [40–43] and data clustering [44–47] are applied. Moreover, the proposed algorithm is based on the assumption that all corresponding released versions of the data $D^{\Gamma_j \cup \Delta_j}$ are released from the timestamp 1 to $j-1$, i.e., $\mathbb{D}^{\Gamma_1 \cup \Delta_1}$, $\mathbb{D}^{\Gamma_2 \cup \Delta_2}$, ..., $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$, and they always satisfy the proposed privacy preservation model. Thus, only $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ is considered to construct $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ of $D^{\Gamma_j \cup \Delta_j}$. The proposed algorithm (Algorithm 1) is shown below.

 doi:10.20944/preprints202508.1159.v1

Algorithm 1 l^{HD}-*Diversity*$(l, D^{\Gamma_j \cup \Delta_j}, \mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}, DGH_{do^{qi}r_1}, DGH_{do^{qi}r_2}, \ldots, DGH_{do^{qi}r_v})$

Require: A positive integer l, The sub-data version $D^{\Gamma_j \cup \Delta_j}$ of D, the relatedly released data version $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ of $D^{\Gamma_j \cup \Delta_j}$, $DGH_{do^{qi}r_1}$, $DGH_{do^{qi}r_2}$, $\ldots$, and $DGH_{do^{qi}r_v}$.
Ensure: A released data version $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ of $D^{\Gamma_j \cup \Delta_j}$.
Let $TMPT_1$ and $TMPT_2$ be the set of temporal tuples.
Let $TMPS_1$ and $TMPS_2$ be the set of temporal penalty costs.
if $|D^{\Gamma_j \cup \Delta_j}| < l$ **then**
  return *Failure*
else if $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ is $NULL$ **then**
  $TMPT_1 \leftarrow d_i$, $D^{\Gamma_j \cup \Delta_j} \leftarrow D^{\Gamma_j \cup \Delta_j} - d_i$, $TMPS_2 \leftarrow \infty$, $g \leftarrow 1$
  while $D^{\Gamma_j \cup \Delta_j}[s_{o_1}], D^{\Gamma_j \cup \Delta_j}[s_{o_2}], \ldots, D^{\Gamma_j \cup \Delta_j}[s_{o_q}]$ satisfy l **do**
    $TMPS_1 \leftarrow f_{\mathbb{PREC}}(f_A(TMPT_1 \cup d_i))$
    if $TMPS_1 < TMPS_2$ **then**
      $TMPT_2 \leftarrow d_i$
      $TMPS_2 \leftarrow TMPS_1$
    end if
    if $g = |D^{\Gamma_j \cup \Delta_j}|$ **then**
      $TMPT_1 \leftarrow TMPT_1 \cup TMPT_2$
      $D^{\Gamma_j \cup \Delta_j} \leftarrow D^{\Gamma_j \cup \Delta_j} - TMPT_2$, $TMPS_2 \leftarrow \infty$, $g \leftarrow 1$
      if $TMPT_1[s_{o_1}], TMPT_1[s_{o_2}], \ldots, TMPT_1[s_{o_q}]$ satisfy l **then**
        $\mathbb{D}^{\Gamma_j \cup \Delta_j} \leftarrow \mathbb{D}^{\Gamma_j \cup \Delta_j} \cup f_A(TMPT_1, DGH_{do^{qi}r_1}, DGH_{do^{qi}r_2}, \ldots, DGH_{do^{qi}r_v})$
        $TMPT_1 \leftarrow d_i$, $D^{\Gamma_j \cup \Delta_j} \leftarrow D^{\Gamma_j \cup \Delta_j} - d_i$
      end if
    end if
    $g \leftarrow g + 1$
  end while
else
  $TMPT_1 \leftarrow d_i$, $D^{\Gamma_j \cup \Delta_j} \leftarrow D^{\Gamma_j \cup \Delta_j} - d_i$, $TMPS_2 \leftarrow \infty$, $g \leftarrow 1$
  while $D^{\Gamma_j \cup \Delta_j}[s_{o_1}], D^{\Gamma_j \cup \Delta_j}[s_{o_2}], \ldots, D^{\Gamma_j \cup \Delta_j}[s_{o_q}]$ satisfy l **do**
    $TMPS_1 \leftarrow f_{\mathbb{PREC}}(f_A(TMPT_1 \cup d_i))$
    if $TMPS_1 < TMPS_2$ **then**
      $TMPT_2 \leftarrow d_i$
      $TMPS_2 \leftarrow TMPS_1$
    end if
    if $g = |D^{\Gamma_j \cup \Delta_j}|$ **then**
      $TMPT_1 \leftarrow TMPT_1 \cup TMPT_2$
      $D^{\Gamma_j \cup \Delta_j} \leftarrow D^{\Gamma_j \cup \Delta_j} - TMPT_2$, $TMPS_2 \leftarrow \infty$, $g \leftarrow 1$
      if $TMPT_1[s_{o_1}], TMPT_1[s_{o_2}], \ldots, TMPT_1[s_{o_q}]$ satisfy l **then**
        for $z \leftarrow 1$ to $|EC^{j-1}|$ **do**
          if Every compared result between each $TMPT_1[s_{o_\varkappa}]$ and its comparable $ec_z[s_{o_\varkappa}]$ in EC^{j-1} of $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$, where $1 \leq \varkappa \leq q$ and $1 \leq z \leq |EC^{j-1}|$, satisfies l **then**
            $\mathbb{D}^{\Gamma_j \cup \Delta_j} \leftarrow \mathbb{D}^{\Gamma_j \cup \Delta_j} \cup f_A(TMPT_1, DGH_{do^{qi}r_1}, DGH_{do^{qi}r_2}, \ldots, DGH_{do^{qi}r_v})$.
            $TMPT_1 \leftarrow d_i$, $D^{\Gamma_j \cup \Delta_j} \leftarrow D^{\Gamma_j \cup \Delta_j} - d_i$
          end if
        end for
      end if
    end if
    $g \leftarrow g + 1$
  end while
end if
return $\mathbb{D}^{\Gamma_j \cup \Delta_j}$

The inputs of the proposed privacy preservation algorithm are a positive integer l, the sub-data version $D^{\Gamma_j \cup \Delta_j}$ of D, the corresponding released version of the data $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ in $D^{\Gamma_j \cup \Delta_j}$, and the data generalization hierarchies $DGH_{do^{qi}r_1}$, $DGH_{do^{qi}r_2}$, $\ldots$, and $DGH_{do^{qi}r_v}$. The output of the proposed

 doi:10.20944/preprints202508.1159.v1

privacy preservation algorithm is the released version of the data $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ in $D^{\Gamma_j \cup \Delta_j}$ such that it satisfies the proposed privacy preservation constraints that are presented in Section 3.1.

For privacy preservation, $D^{\Gamma_j \cup \Delta_j}$ is first investigated to answer the question "can $D^{\Gamma_j \cup \Delta_j}$ be transformed to satisfy the given value of l?". If $D^{\Gamma_j \cup \Delta_j}$ cannot be transformed to satisfy the given value of l, the algorithm returns $Failure$. If it can be transformed, the second or third part of the algorithm is enabled. In the second part, the algorithm investigates the question "is there a corresponding released version of the data $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ in $D^{\Gamma_j \cup \Delta_j}$?". That is, if $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ is $NULL$, it means that $D^{\Gamma_j \cup \Delta_j}$ does not have a corresponding released version of the data . Thus, $D^{\Gamma_j \cup \Delta_j}$ can be transformed to satisfy the given value of l, without considering privacy violation from data comparison attacks, using the following steps:

- In the first step, an arbitrary tuple $d_i \in D^{\Gamma_j \cup \Delta_j}$ is chosen to be the initialized tuple for constructing the first equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. Moreover, d_i is removed from $D^{\Gamma_j \cup \Delta_j}$ and kept by $TMPT_1$.
- In the second step, the maximum penalty cost for constructing the first equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is determined such that it is kept by $TMPS_2$, i.e., $TMPS_2 = \infty$.
- In the third step, all tuples of $D^{\Gamma_j \cup \Delta_j}$ are iterated until they cannot satisfy the given value of l. In each iteration, a tuple d_i of $D^{\Gamma_j \cup \Delta_j}$ is assigned to its appropriate equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ and removed from $D^{\Gamma_j \cup \Delta_j}$. In addition, to construct each new equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$, an arbitrary tuple $d_i \in D^{\Gamma_j \cup \Delta_j}$ is chosen to be the initialized tuple, and the maximum penalty cost, $TMPS_2 = \infty$, for constructing the equivalence class is set to be maximized.
- In the fourth step, the unique quasi-identifier values are made available in qi_1, qi_2, ..., and qi_p are generalized by their less-specific values, which are represented by $DGH_{do^{qir_1}}$, $DGH_{do^{qir_2}}$, ..., and $DGH_{do^{qir_v}}$, respectively.
- Finally, $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is returned.

In addition, the tuples of $D^{\Gamma_j \cup \Delta_j}$ cannot be transformed to satisfy the given value of l; they are suppressed.

Another part of the algorithm is enabled when $D^{\Gamma_j \cup \Delta_j}$ can be satisfied by the given value of l, and when the corresponding released version of the data $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ in $D^{\Gamma_j \cup \Delta_j}$ is available. In this part of the algorithm, in addition to generalizing the unique quasi-identifier values and considering the number of unique sensitive values, all compared results between each equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ and its comparable equivalence class in $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ are also considered to satisfy the given value of l. Moreover, the tuples of $D^{\Gamma_j \cup \Delta_j}$ cannot be transformed to satisfy the given value of l; they are also suppressed. Finally, $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is returned.

3.3.1. The Complexity of the Proposed Algorithm

In this section, we discuss the complexity of the proposed algorithm. With the proposed algorithm, we can see that before every equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ is constructed, its most similar tuples are first determined such that they are $f_{\mathbb{PREC}}(TMPT_1)$ as minimized as possible, and each of their sensitive attributes must collect at least l distinct sensitive values. In addition, the most similar tuples are determined and removed from $D^{\Gamma_j \cup \Delta_j}$ in each iteration of the proposed algorithm. For this reason, the tuples of $D^{\Gamma_j \cup \Delta_j}$ are reduced by one in every iteration. Therefore, the cost of determining the most similar tuples in the proposed algorithm can be defined by Equation 7.

$$\begin{aligned} &| D^{\Gamma_j \cup \Delta_j} | + (| D^{\Gamma_j \cup \Delta_j} | - 1) + (| D^{\Gamma_j \cup \Delta_j} | \\ &- 2) + \cdots + (| D^{\Gamma_j \cup \Delta_j} | - (| D^{\Gamma_j \cup \Delta_j} | - 1)) = \\ &\frac{| D^{\Gamma_j \cup \Delta_j} |^2 + | D^{\Gamma_j \cup \Delta_j} |}{2} \end{aligned} \tag{7}$$

For example, suppose that six tuples are available in $D^{\Gamma_j \cup \Delta_j}$. An infographic illustrating the cost of determining the most similar tuples in the proposed algorithm is shown in Figure 2, where the blue square is the number of tuples that are considered by each iteration of the proposed algorithm.

 doi:10.20944/preprints202508.1159.v1

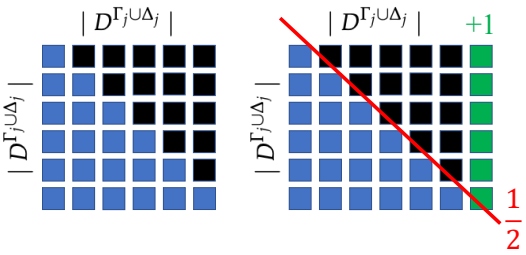

Figure 2. An infographic illustrating the cost of determining the most similar tuples to construct equivalence classes of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$.

In addition to the cost of determining the most similar tuples, the proposed algorithm has two further costs that must be considered, i.e., the cost of data generalization and the cost of comparing the results between each constructed equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ and its comparable equivalence classes in $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$.

The cost of data generalization depends on the number of quasi-identifier attributes, the height of the data generalization hierarchy of each quasi-identifier attribute, and the number of equivalence classes of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$. Therefore, the cost of generalizing the unique quasi-identifier values of the proposed algorithm can be defined by Equation 8.

$$\begin{aligned} &MAX(\mid DGH_{do^{qir_1}} \mid, \mid DGH_{do^{qir_2}} \mid, \\ &\ldots, \mid DGH_{do^{qir_v}} \mid) \cdot \mid D^{\Gamma_j} \mid \cdot \mid EC^j \mid \end{aligned} \tag{8}$$

Another cost of the proposed algorithm is that of comparing the results between each constructed equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ and its comparable equivalence classes in $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$. This cost depends on the number of sensitive attributes, the number of equivalence classes of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$, and the number of equivalence classes of $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$. Therefore, the cost of comparing the results between each constructed equivalence class of $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ and its comparable equivalence class in $\mathbb{D}^{\Gamma_{j-1} \cup \Delta_{j-1}}$ can be defined by Equation 9.

$$\mid D^{\Delta_j} \mid \cdot \mid EC^j \mid \cdot \mid EC^{j-1} \mid \tag{9}$$

Thus, the total cost (the complexity) for the proposed algorithm of constructing the released version of the dataset $\mathbb{D}^{\Gamma_j \cup \Delta_j}$ in $D^{\Gamma_j \cup \Delta_j}$ can be defined by Equation 10.

$$\begin{aligned} &\frac{\mid D^{\Gamma_j \cup \Delta_j} \mid^2 + \mid D^{\Gamma_j \cup \Delta_j} \mid}{2} \cdot MAX(\\ &\mid DGH_{do^{qir_1}} \mid, \mid DGH_{do^{qir_2}} \mid, \ldots, \mid DGH_{do^{qir_v}} \mid) \cdot \\ &\mid D^{\Gamma_j} \mid \cdot \mid D^{\Delta_j} \mid \cdot \mid EC^{j-1} \mid \cdot \mid EC^j \mid^2 \end{aligned} \tag{10}$$

4. Experiment

This section is focused on evaluating the effectiveness and efficiency of the proposed privacy preservation model by comparing it with *l*-Diversity [12] and *LKC*-Privacy [19].

4.1. Experimental Setup

All experiments were performed to evaluate the effectiveness and efficiency of the proposed privacy preservation model; they were 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 were built and executed using Microsoft Windows Server 2019 in connection with Microsoft Visual Studio 2019 Community Edition and Microsoft SQL Server 2019.

Moreover, all of the experimental results discussed were obtained from the *Adult* dataset, which is available at the *UCI* Machine Learning Repository [48]. This dataset is constructed from 32561 user profile tuples. Each user profile tuple consists of 14 attributes, i.e., *Age*, *Workclass*, *Fnlwgt*,

 doi:10.20944/preprints202508.1159.v1

Education, *Education-num*, *Marital-status*, *Occupation*, *Relationship*, *Race*, *Sex*, *Capital-gain*, *Capital-loss*, *Hours-per-week*, and *Native-country*. To effectively conduct the experiments, only the attributes *Age*, *Workclass*, *Education*, *Marital-status*, *Occupation*, *Relationship*, *Sex*, *Capital-loss*, *Hours-per-week*, and *Native-country* were made available in the experimental datasets. The attributes *Age*, *Education*, *Marital-status*, *Occupation*, *Sex*, and *Native-country* were set as the quasi-identifier attributes. Other attributes (i.e., *Workclass*, *Capital-loss*, *Hours-per-week*, and *Relationship*) were set as the sensitive attributes. Moreover, in this dataset, all user profile tuples include the values "?" and "0"; for the purposes of this study, they were removed. Therefore, the experimental dataset only included 1428 user profile tuples. We assumed that all experimental datasets had been released twice. Histograms and the cumulative percentages of each quasi-identifier attribute and each sensitive attribute are shown in Figures 3 and 4, respectively.

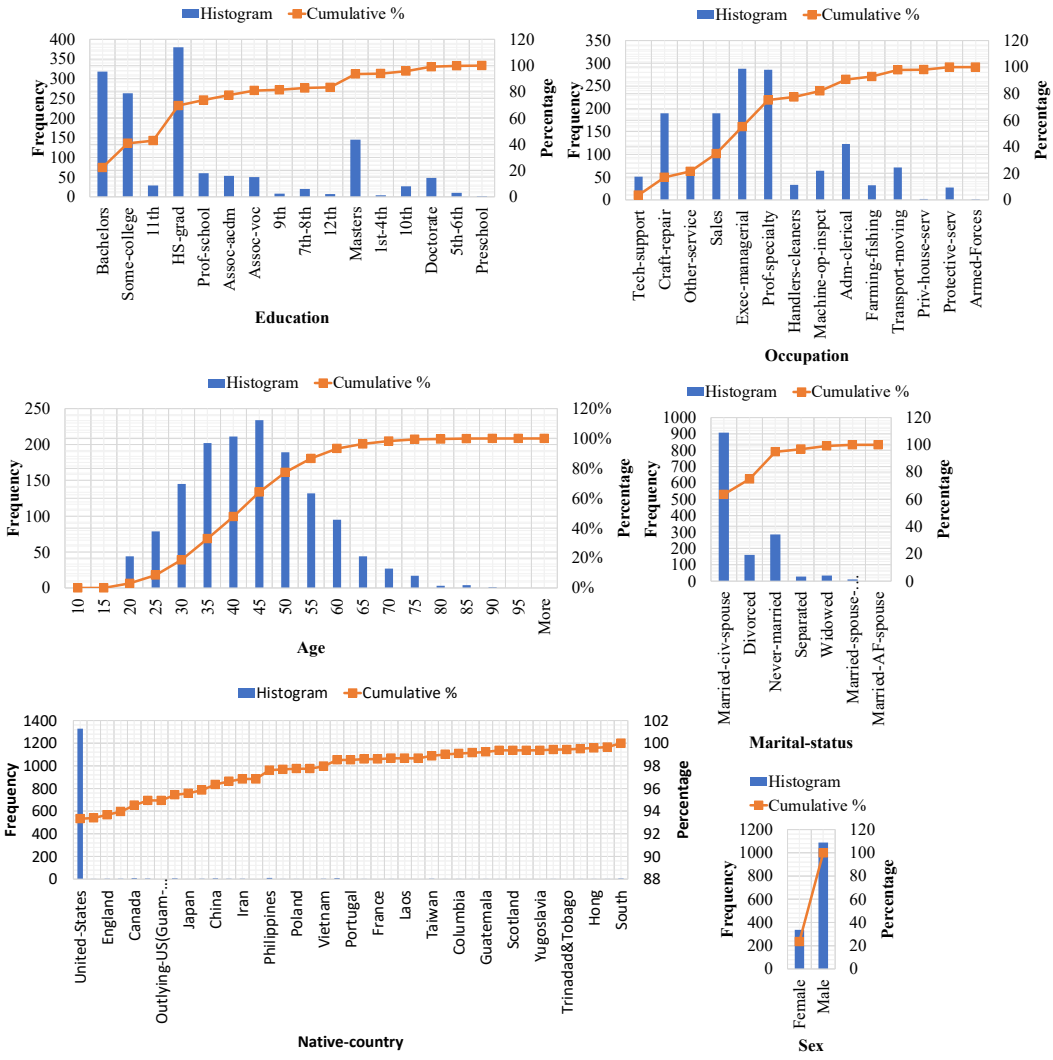

Figure 3. Histograms and the cumulative percentages of each quasi-identifier attribute of the experimental dataset.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 August 2025 doi:10.20944/preprints202508.1159.v1

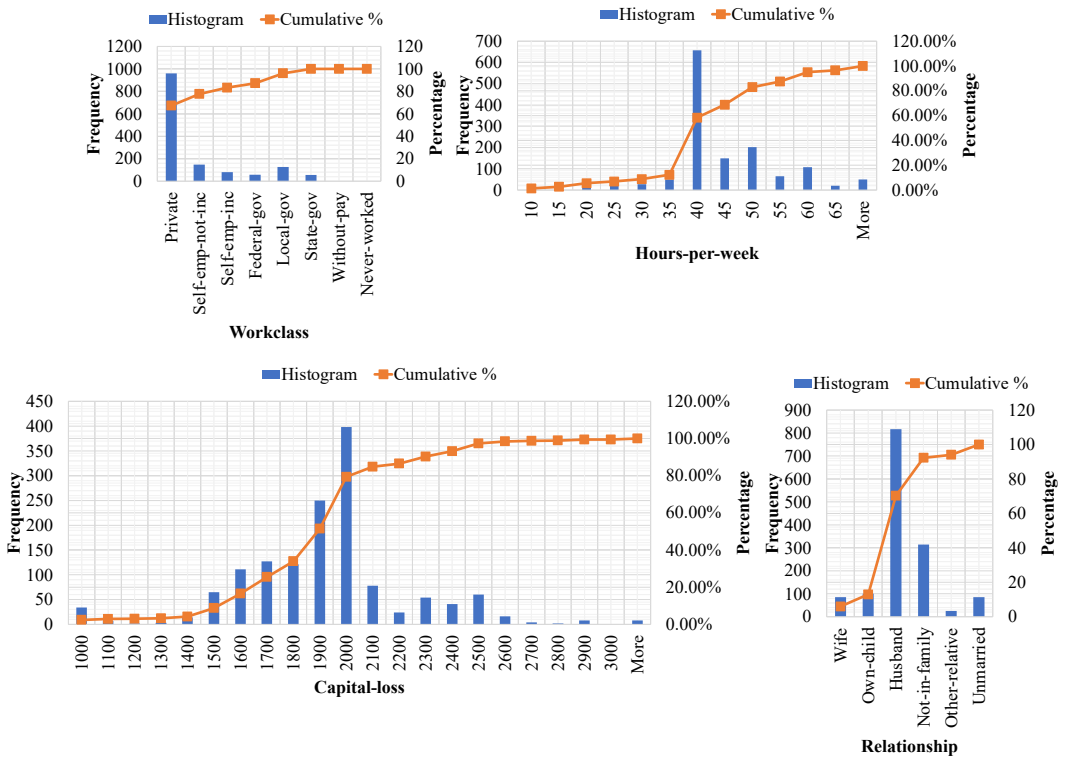

Figure 4. Histograms and the cumulative percentages of each sensitive attribute of the experimental dataset.

4.2. Experimental Results and Discussion

4.2.1. Effectiveness of the Model

In the first experiment, we evaluate the effect of the number of quasi-identifier attributes on the data utility of the datasets constructed by the proposed model and the compared models such that they are based on *PREC* and *DM* penalty costs. For the experiment, the value of l is fixed at 2 for the proposed model and l-Diversity. For *LKC*-Privacy, the values of L, K, and C are set to the number of quasi-identifier attributes, l, and $1/l$, respectively. Furthermore, all sensitive values available in the experimental datasets are protected sensitive values. Only *Capital-Loss* is set as a sensitive attribute. The number of quasi-identifier attributes varies from 1 to 6. The process of varying the number of quasi-identifier attributes is as follows.

- Initially, only *Native-country* is a quasi-identifier attribute.
- In the second experiment, the experimental dataset only contains *Native-country* and *Sex* as quasi-identifier attributes.
- *Native-country*, *Sex*, and *Marital-status* are the quasi-identifier attributes in the third experiment.
- The fourth experiment has the quasi-identifier attributes *Native-country*, *Sex*, *Marital-status*, and *Occupation*.
- The quasi-identifier attributes in the fifth experimental dataset are *Native-country*, *Sex*, *Marital-status*, *Occupation*, and *Education*.
- In the final experimental dataset, *Age*, *Education*, *Marital-status*, *Occupation*, *Sex*, and *Native-country* are all set as quasi-identifier attributes.

As shown in Figures 5 and 6, when the number of quasi-identifier attributes is increased, the *PREC* and *DM* penalty costs of all experimental datasets also increase. Moreover, l-Diversity and *LKC*-Privacy are equally effective and exhibit higher performance than the proposed model. However, they are slightly different. The higher *PREC* and *DM* penalty costs in the experimental datasets with an increasing number of quasi-identifier attributes is due to the increase in the number of quasi-identifier attributes, and the size of the equivalence classes also increases. In addition, larger equivalence classes generally lead to more suppressed or generalized values. Moreover, larger equivalence classes often

 doi:10.20944/preprints202508.1159.v1

lead to a large number of generalized values in datasets. The reason *l*-Diversity and *LKC*-Privacy are equally effective in terms of maintaining the data utility of the experimental datasets with every experiment is that when the experimental datasets do not allow for the retrieval of missing values and all sensitive values are protected sensitive values, the released version of the data from the datasets based on *l*-Diversity is not different from that based on *LKC*-Privacy. The reason for the reduced effectiveness of the proposed model compared to the other models is that, in addition to datasets needing to satisfy the privacy preservation constraints, the compared results between datasets and their corresponding released versions must also satisfy the privacy preservation constraints. For this reason, the datasets satisfy the privacy preservation constraint of the proposed model; they are not susceptible to privacy violation from data comparison attacks. However, datasets constructed from *l*-Diversity and *LKC*-Privacy are still susceptible to such attacks.

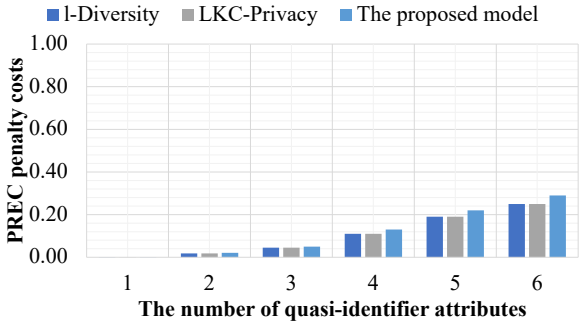

Figure 5. The effectiveness of the models based on the number of quasi-identifier attributes and *PREC*.

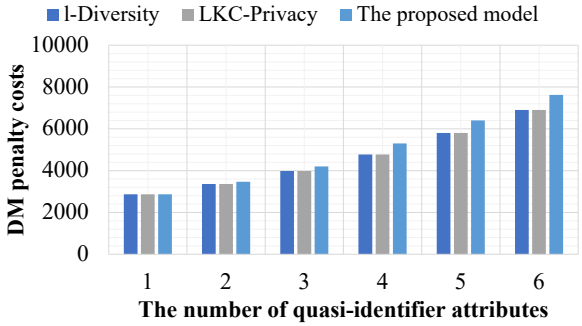

Figure 6. The effectiveness of the models based on the number of quasi-identifier attributes and *DM*.

In the second experiment, we evaluate the effect of the number of sensitive attributes on the data utility of datasets constructed by the proposed model and the compared models, such that they are based on *PREC* and *DM*. In this experiment, the proposed model is only evaluated by comparison with *l*-Diversity, because *LKC*-Privacy cannot address privacy violation issues in datasets with multiple sensitive attributes. In this experiment, the value of l is fixed at 2; all quasi-identifier attributes are available in the experimental datasets; and the number of sensitive attributes varies from 1 to 4. The process of varying the number of sensitive attributes is as follows.

- The first experimental dataset only has *Capital-loss* as a sensitive attribute.
- The second experimental dataset only contains *Capital-loss* and *Relationship* as sensitive attributes.
- *Capital-loss*, *Relationship*, and *Workclass* are the sensitive attributes in the third experimental dataset.
- In the final experimental dataset, *Capital-loss*, *Relationship*, *Workclass*, and *Hours-per-week* are all set as sensitive attributes.

Figures 7 and 8 show that when the number of sensitive attributes is increased, the *PREC* and *DM* penalty costs of all experimental datasets are also increased. The reason for the higher *PREC* and

 doi:10.20944/preprints202508.1159.v1

DM penalty costs in the experimental datasets with an increasing number of sensitive attributes is that when the number of sensitive attributes is increased, the size of the equivalence classes also increases. Moreover, *l*-Diversity is more effective than the proposed model. However, they are slightly different. The reason for the reduced effectiveness of the proposed model compared to *l*-Diversity is that, in addition to datasets needing to satisfy the privacy preservation constraints, the compared results between datasets and their corresponding released versions must also satisfy the privacy preservation constraints, but this privacy preservation constraint is not considered by *l*-Diversity. For this reason, although the proposed model is less effective than *l*-Diversity, it is more secure in terms of privacy preservation.

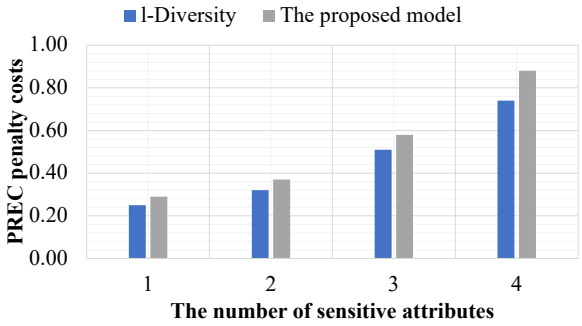

Figure 7. The effectiveness of the models based on the number of sensitive attributes and *PREC*.

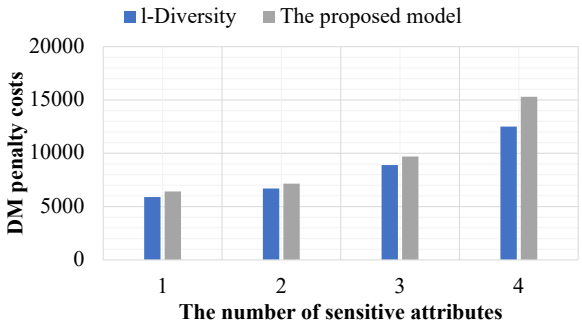

Figure 8. The effectiveness of the models based on the number of sensitive attributes and *PREC*.

In the third experiment, we evaluate the effect of the value of l on the data utility of datasets constructed by the proposed model and the other models, such that they are based on *PREC* and *DM*. In this experiment, only *Capital-Loss* is set as a sensitive attribute, and all quasi-identifier attributes are available in the experimental datasets. The value of l is varied from 2 to 10 for the proposed model and *l*-Diversity. In *LKC*-Privacy, the values of L, K, and C are set to the number of quasi-identifier attributes, l, and $1/l$, respectively. Furthermore, all sensitive values available in the experimental datasets are protected sensitive values.

Figures 9 and 10 show that when the value of l is increased, the *PREC* and *DM* penalty costs of all experimental datasets are also increased. This is because when the value of l is increased, the size of the equivalence classes is also increased. Moreover, the compared models are more effective than the proposed model. However, they are slightly different. The reason the proposed model is less effective than the compared models is that in addition to datasets needing to satisfy the privacy preservation constraints, the compared results between the datasets and their corresponding released versions must also satisfy the privacy preservation constraints, but this privacy preservation constraint is not considered by the compared models.

 doi:10.20944/preprints202508.1159.v1

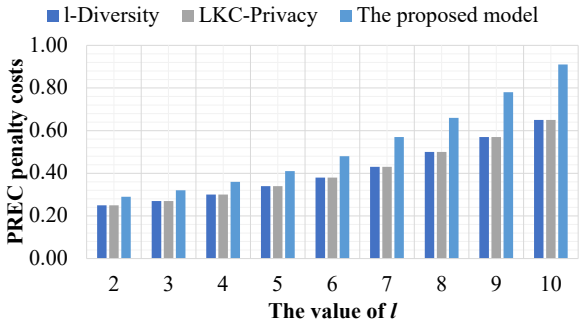

Figure 9. The effectiveness of the models based on the value of l and *PREC*.

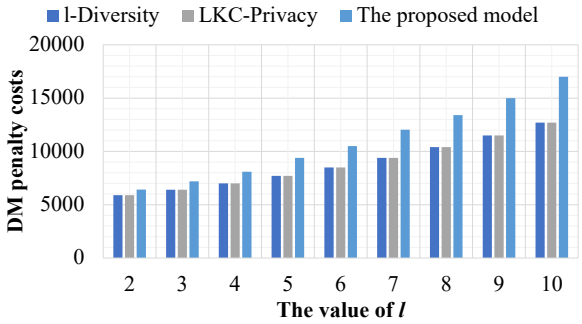

Figure 10. The effectiveness of the models based on the value of l and *PREC*.

Figures 5–10 show that the number of sensitive attributes and the value of l have a greater effect on the data utility in the datasets than the number of quasi-identifier attributes; this is because the privacy preservation constraint of the proposed model and of the compared models is based on the number of distinct sensitive values.

In the fourth experiment, we evaluate the effect of a limited number of quasi-identifier attributes on the data utility of datasets constructed by the proposed model and the compared models, such that they are based on *PREC* and *DM*. In this experiment, we assume that the data holder needs to limit the number of quasi-identifier attributes for data release from 1 to 6 attributes. In this experiment, the value of l is fixed at 2 for the proposed model and l-Diversity. In *LKC*-Privacy, the values of L, K, and C are set to the number of quasi-identifier attributes, l, and $1/l$, respectively. Furthermore, all sensitive values available in the experimental datasets are protected sensitive values. Only *Capital-Loss* is set as a sensitive attribute. All quasi-identifier attributes are available in the experimental datasets.

Figures 11 and 12 show that the proposed model is more effective than the compared models in all experiments constructed from experimental datasets with five quasi-identifier attributes at most. The reason for this is that it supports separation of the quasi-identifier attributes to preserve the privacy of data in datasets, while the compared models do not consider this property in their privacy preservation constraints. For this reason, the compared models must consider all quasi-identifier attributes in every experiment. However, when the experimental dataset has six quasi-identifier attributes, the proposed model is less effective than the compared models. This is because the experimental datasets are the same size, and in addition to datasets needing to satisfy the privacy preservation constraints, the compared results between datasets and their corresponding released versions must also satisfy the privacy preservation constraint of the proposed model.

 doi:10.20944/preprints202508.1159.v1

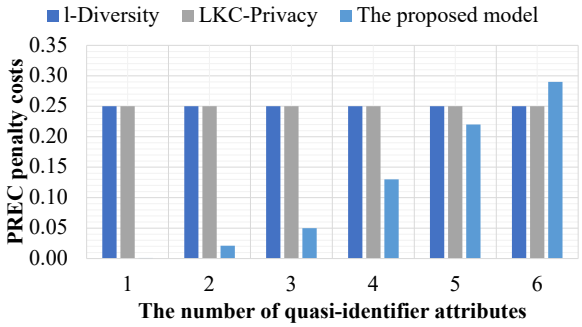

Figure 11. The effectiveness of the model based on the limited number of quasi-identifier attributes and *PREC*.

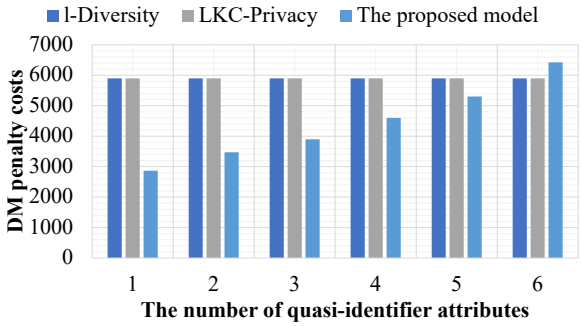

Figure 12. The effectiveness of the models based on a limited number of quasi-identifier attributes and *DM*.

In the fifth experiment, we evaluate the effect of a limited number of sensitive attributes on the data utility of datasets constructed by the proposed model and the compared models, such that they are based on *PREC* and *DM*. In this experiment, the proposed model is only evaluated by comparison with *l*-Diversity, because *LKC*-Privacy cannot address privacy violation issues in datasets that have multiple sensitive attributes, and we assume that the data holder needs to limit the number of sensitive attributes for data release from 1 to 4 attributes. For this experiment, the value of l is fixed at 2. All quasi-identifiers and sensitive attributes are available in the experimental datasets.

Figures 13 and 14 show that the proposed model is more effective than the compared models in all experiments constructed from the experimental datasets with three sensitive attributes at most. The reason for this is that it supports separation of the sensitive attributes to preserve the privacy of data in datasets, while the compared models do not consider this property in their privacy preservation constraints. For this reason, the compared models must consider all sensitive attributes in every experiment. However, when the experimental dataset has four sensitive attributes, the proposed model is less effective than the compared models. The reason for this is that the experimental datasets are the same size, and in addition to datasets needing to satisfy the privacy preservation constraints, the compared results between datasets and their corresponding released versions must also satisfy the privacy preservation constraint of the proposed model.

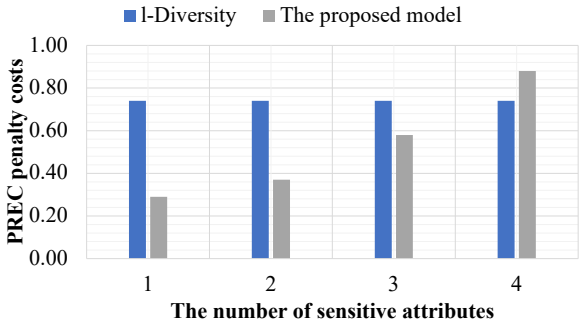

Figure 13. The effectiveness of the models based on a limited number of sensitive attributes and *PREC*.

 doi:10.20944/preprints202508.1159.v1

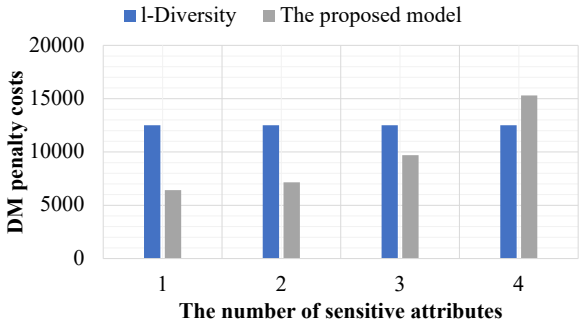

Figure 14. The effectiveness of the models based on a limited number of sensitive attributes and *DM*.

Figures 11–14 clearly indicate that the proposed model is more secure in terms of privacy preservation and better in terms of maintaining the data utility of datasets compared to the other models.

In the sixth experiment, we evaluate the data utility of datasets that satisfy the privacy preservation constraint of the proposed model, *l*-Diversity, and *LKC*-Privacy, such that they are based on the *AVERAGE* query function in conjunction with the *AND* or *OR* query operator and the range of queries, and they are evaluated by the relative error metric presented in Section 3.2.3. In this experiment, the value of l is fixed at 2 for the proposed model and *l*-Diversity. With *LKC*-Privacy, the values of L, K, and C are set to the number of quasi-identifier attributes, l, and $1/l$, respectively. Furthermore, all sensitive values available in the experiment datasets are protected sensitive values. Only *Capital-Loss* is a sensitive attribute, and all quasi-identifier attributes are available in the experimental datasets. Moreover, all of the experimental results are shown in Figures 15 and 16, and are presented as the mean of the average of the results of 15 randomized queries in the form of *Query 1*. The experimental results are shown in Figure 17, and are presented as the mean of the average results of 15 randomized queries in the form of *Query 2*.

- *Query 1: SELECT AVERAGE(Capital-loss) WHERE* $qi_1 = qiv_1$ *[AND/OR] ... [AND/OR]* $qi_p = qiv_p$*;*
- *Query 2: SELECT AVERAGE(Capital-loss) WHERE Age BETWEEN LB AND UB;*

The elements of these queries are defined as follows:

- $qi_1 \ldots qi_p$ are the specified quasi-identifier attributes *Age*, *Education*, *Marital-status*, *Occupation*, *Sex*, and *Native-country*.
- $qiv_1 \ldots qiv_p$ are the specified values for querying the data from the datasets.
- LB is the lower bound for querying the data from the datasets.
- UB is the upper bound for querying the data from the datasets.

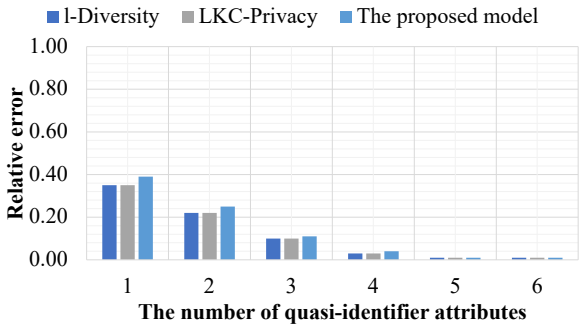

Figure 15. The effectiveness of the models based on the *OR* query operation.

In Figure 15, we show the data utility of query results affected by the *OR* query operation. The experimental results show that the number of query-condition attributes inversely influences the data utility of query results; i.e., a larger number of query-condition attributes leads to higher data utility of the query results. This is because a larger number of query-condition attributes gives all experimental models more options for generalizing the data in datasets, thus resulting in fewer errors.

 doi:10.20944/preprints202508.1159.v1

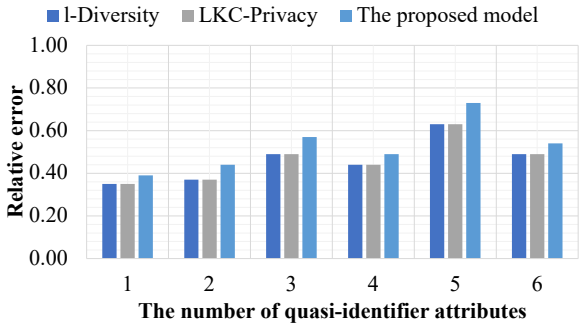

Figure 16. The effectiveness of the models based on the *AND* query operation.

Figure 16 shows the effect of using the *AND* query operation on the query results. Obviously, when the number of query-condition attributes is increased, the relative errors of the query results are also increased. This is because all experimental models have limitations regarding the values satisfied in data queries.

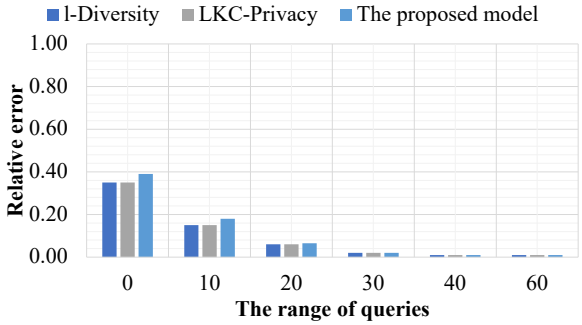

Figure 17. The effectiveness of the models based on the range of queries.

Figure 17 shows the effect of using the range of queries on the query results. Note that when the query range condition is set to 0, it means that the exact same value is applied. The trend of the experimental results in Figure 17 is similar to that of the experimental results shown in Figure 15 for the same reason, i.e., a wide range of query conditions often leads to more options for generalizing the data in datasets, thus resulting in fewer errors.

4.2.2. Efficiency

In the seventh experiment, we evaluate the efficiency of the proposed model, which is based on the number of quasi-identifier attributes. In this experiment, the value of l is fixed at 2 for the proposed model and l-Diversity. With *LKC*-Privacy, the values of L, K, and C are set to the number of quasi-identifier attributes, l, and $1/l$, respectively. Furthermore, all sensitive values available in the experimental datasets are set as protected sensitive values. Only *Capital-Loss* is a sensitive attribute, and the number of quasi-identifier attributes varies from 1 to 6.

Figures 18–20 show that the proposed model is less efficient than l-Diversity, but it is more efficient than *LKC*-Privacy. The reason l-Diversity is more efficient than all other experimental models is that its privacy preservation constraints are simpler than those of the other models. That is, with the proposed model, in addition to the datasets needing to satisfy the privacy preservation constraints, the compared results must also satisfy the model's privacy preservation constraints. Thus, in addition to the cost of considering the data from datasets to satisfy privacy preservation constraints, the model incurs the cost of data comparison between it and its corresponding datasets. The reason *LKC*-Privacy is less efficient than the other experimental models is that it must consider sub-datasets with a size of L at most.

 doi:10.20944/preprints202508.1159.v1

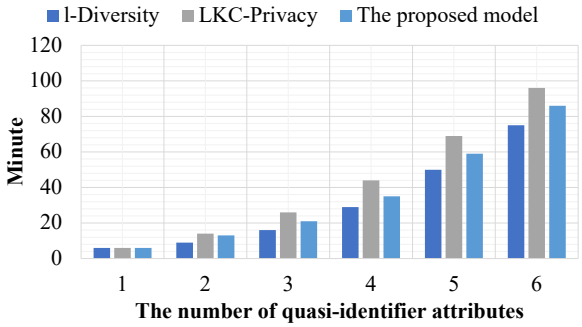

Figure 18. The efficiency of the models based on the number of quasi-identifier attributes.

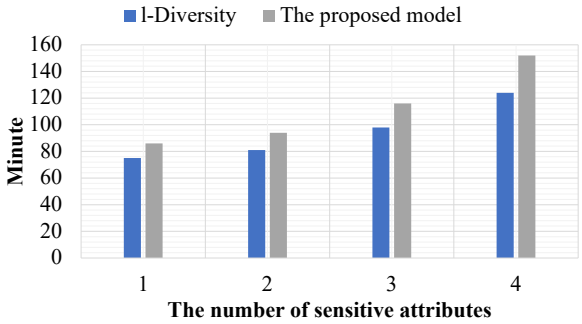

Figure 19. The efficiency of the models based on the number of sensitive attributes.

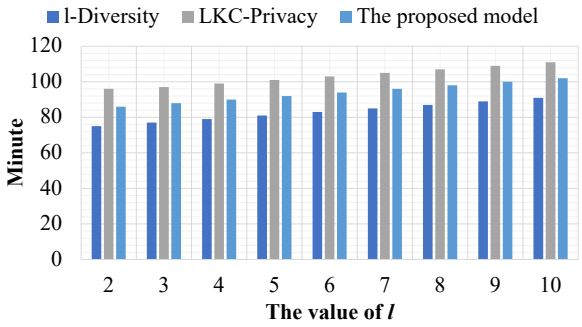

Figure 20. The efficiency of the models based on the value of l.

5. Conclusion

This work enumerates and explains the vulnerabilities of privacy preservation models to data comparison attacks when datasets are independently released. To address the vulnerabilities of privacy preservation models, we propose a new model that can address privacy violations caused by data comparison attacks on datasets. Moreover, our experimental results indicate that released datasets are satisfied by the proposed model, which is found to be more secure in terms of privacy preservation and better in terms of maintaining the data utility of datasets compared to the other models.

6. Future Work

Although the proposed model can address privacy violation issues resulting from data comparison attacks on independently released datasets, adversaries will discover new approaches to compromising the privacy of data. Thus, an appropriate privacy preservation model that can address newly discovered privacy violation issues should be proposed.

Author Contributions: Surapon Riyana wrote this manuscript.

Funding: Not applicable.

 doi:10.20944/preprints202508.1159.v1

Ethical Approval: Not applicable.

Competing Interests: Not applicable.

Availability of Data and Materials: All data can be found online in the *Adult* dataset, which is available at the *UCI* Machine Learning Repository.

References

1. Alferidah, D.K.; Jhanjhi, N. A review on security and privacy issues and challenges in internet of things. *International Journal of Computer Science and Network Security IJCSNS* **2020**, *20*, 263–286.
2. Alwarafy, A.; Al-Thelaya, K.A.; Abdallah, M.; Schneider, J.; Hamdi, M. A survey on security and privacy issues in edge-computing-assisted internet of things. *IEEE Internet of Things Journal* **2020**, *8*, 4004–4022.
3. Deep, S.; Zheng, X.; Jolfaei, A.; Yu, D.; Ostovari, P.; Kashif Bashir, A. A survey of security and privacy issues in the Internet of Things from the layered context. *Transactions on Emerging Telecommunications Technologies* **2022**, *33*, e3935.
4. Hathaliya, J.J.; Tanwar, S. An exhaustive survey on security and privacy issues in Healthcare 4.0. *Computer Communications* **2020**, *153*, 311–335.
5. Joshi, A.P.; Han, M.; Wang, Y. A survey on security and privacy issues of blockchain technology. *Mathematical foundations of computing* **2018**, *1*, 121.
6. Newaz, A.I.; Sikder, A.K.; Rahman, M.A.; Uluagac, A.S. A survey on security and privacy issues in modern healthcare systems: Attacks and defenses. *ACM Transactions on Computing for Healthcare* **2021**, *2*, 1–44.
7. Zhi, Y.; Fu, Z.; Sun, X.; Yu, J. Security and privacy issues of UAV: a survey. *Mobile Networks and Applications* **2020**, *25*, 95–101.
8. Riyana, S.; Sasujit, K.; Homdoung, N.; Chaichana, T.; Punsaensri, T. Effective Privacy Preservation Models for Rating Datasets. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)* **2023**, *17*, 1–13. https://doi.org/10.37936/ecti-cit.2023171.250111.
9. Riyana, S. Achieving Anatomization Constraints in Dynamic Datasets. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)* **2023**, *17*, 27–45. https://doi.org/10.37936/ecti-cit.2023171.248877.
10. Sweeney, L. K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. https://doi.org/10.1142/S0218488502001648.
11. Riyana, S.; Nanthachumphu, S.; Riyana, N. Achieving privacy preservation constraints in missing-value datasets. *SN Computer Science* **2020**, *1*, 1–10.
12. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: privacy beyond k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), 2006, pp. 24–24. https://doi.org/10.1109/ICDE.2006.1.
13. Aggarwal, C.C.; Yu, P.S. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*; Springer, 2008; pp. 11–52.
14. Aggarwal, C.C.; Yu, P.S. Privacy-preserving data mining: a survey. In *Handbook of database security*; Springer, 2008; pp. 431–460.
15. Gascón, A.; Schoppmann, P.; Balle, B.; Raykova, M.; Doerner, J.; Zahur, S.; Evans, D. Privacy-Preserving Distributed Linear Regression on High-Dimensional Data. *Proc. Priv. Enhancing Technol.* **2017**, *2017*, 345–364.
16. Wang, R.; Zhu, Y.; Chang, C.C.; Peng, Q. Privacy-preserving high-dimensional data publishing for classification. *Computers & Security* **2020**, *93*, 101785.
17. Wang, W.; Chen, L.; Zhang, Q. Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation. *Computer Networks* **2015**, *88*, 136–148.
18. Terrovitis, M.; Mamoulis, N.; Kalnis, P. Privacy-Preserving Anonymization of Set-Valued Data. *Proc. VLDB Endow.* **2008**, *1*, 115–125. https://doi.org/10.14778/1453856.1453874.
19. Fung, B.C.M.; Cao, M.; Desai, B.C.; Xu, H. Privacy Protection for RFID Data. In Proceedings of the Proceedings of the 2009 ACM Symposium on Applied Computing, New York, NY, USA, 2009; SAC '09, p. 1528–1535. https://doi.org/10.1145/1529282.1529626.
20. Riyana, S.; Riyana, N. A Privacy Preservation Model for RFID Data-Collections is Highly Secure and More Efficient than LKC-Privacy. In Proceedings of the The 12th International Conference on Advances in Information Technology, New York, NY, USA, 2021; IAIT2021. https://doi.org/10.1145/3468784.3469853.
21. Riyana, S.; Riyana, N. Achieving Anonymization Constraints in High-Dimensional Data Publishing Based on Local and Global Data Suppressions. *SN Computer Science* **2022**, *3*, 1–12.

 doi:10.20944/preprints202508.1159.v1

22. Anjum, A.; Ahmad, N.; Malik, S.U.; Zubair, S.; Shahzad, B. An efficient approach for publishing microdata for multiple sensitive attributes. *The Journal of Supercomputing* **2018**, *74*, 5127–5155.
23. Cassa, C.A.; Miller, R.A.; Mandl, K.D. A novel, privacy-preserving cryptographic approach for sharing sequencing data. *Journal of the American Medical Informatics Association* **2013**, *20*, 69–76.
24. Gal, T.S.; Chen, Z.; Gangopadhyay, A. A privacy protection model for patient data with multiple sensitive attributes. *International Journal of Information Security and Privacy (IJISP)* **2008**, *2*, 28–44.
25. Lu, D.; Zhang, Y.; Zhang, L.; Wang, H.; Weng, W.; Li, L.; Cai, H. Methods of privacy-preserving genomic sequencing data alignments. *Briefings in Bioinformatics* **2021**, *22*, bbab151.
26. Riyana, S.; Riyana, N.; Nanthachumphu, S. Privacy Preservation Techniques for Sequential Data Releasing. In Proceedings of the The 12th International Conference on Advances in Information Technology, 2021, pp. 1–9.
27. Wang, R.; Zhu, Y.; Chen, T.S.; Chang, C.C. Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness. *Journal of Computer Science and Technology* **2018**, *33*, 1231–1242.
28. Ye, Y.; Liu, Y.; Wang, C.; Lv, D.; Feng, J. Decomposition: privacy preservation for multiple sensitive attributes. In Proceedings of the International Conference on Database Systems for Advanced Applications. Springer, 2009, pp. 486–490.
29. Riyana, S. (lp1,...,lpn)-Privacy: privacy preservation models for numerical quasi-identifiers and multiple sensitive attributes. *Journal of Ambient Intelligence and Humanized Computing* **2021**, pp. 1–17.
30. Riyana, S.; Natwichai, J. Privacy preservation for recommendation databases. *Service Oriented Computing and Applications* **2018**, *12*, 259–273.
31. Yilmaz, E.; Ferhatosmanoglu, H.; Ayday, E.; Aksoy, R. Privacy-Preserving Aggregate Queries for Optimal Location Selection. *IEEE Transactions on Dependable and Secure Computing* **2017**, *PP*, 1–1. https://doi.org/10.1109/TDSC.2017.2693986.
32. Khan, R.; Tao, X.; Anjum, A.; Sajjad, H.; Khan, A.; Amiri, F.; et al. Privacy preserving for multiple sensitive attributes against fingerprint correlation attack satisfying c-diversity. *Wireless Communications and Mobile Computing* **2020**, *2020*.
33. Riyana, S.; Ito, N.; Chaiya, T.; Sriwichai, U.; Dussadee, N.; Chaichana, T.; Assawarachan, R.; Maneechukate, T.; Tantikul, S.; Riyana, N. Privacy Threats and Privacy Preservation Techniques for Farmer Data Collections Based on Data Shuffling. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)* **2022**, *16*, 289–301.
34. Riyana, S.; Riyana, N.; Sujinda, W. An Anatomization Model for Farmer Data Collections. *SN Computer Science* **2021**, *2*, 1–11.
35. Bourahla, S.; Laurent, M.; Challal, Y. Privacy preservation for social networks sequential publishing. *Computer Networks* **2020**, *170*, 107106. https://doi.org/https://doi.org/10.1016/j.comnet.2020.107106.
36. Riyana, S.; Riyana, N.; Nanthachumphu, S. An effective and efficient heuristic privacy preservation algorithm for decremental anonymization datasets. *Advances in Intelligent Systems and Computing* **2021**, *1200 AISC*, 244–257.
37. Bayardo, R.J.; Agrawal, R. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), April 2005, pp. 217–228. https://doi.org/10.1109/ICDE.2005.42.
38. Riyana, S.; Riyana, N.; Nanthachumphu, S. Enhanced (k,e)-Anonymous for categorical data. *ACM International Conference Proceeding Series* **2017**, pp. 62–67.
39. Zhang, Q.; Koudas, N.; Srivastava, D.; Yu, T. Aggregate Query Answering on Anonymized Tables. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, April 2007, pp. 116–125. https://doi.org/10.1109/ICDE.2007.367857.
40. Chekuri, C.; Pal, M. A recursive greedy algorithm for walks in directed graphs. In Proceedings of the 46th annual IEEE symposium on foundations of computer science (FOCS'05). IEEE, 2005, pp. 245–253.
41. Feldman, M.; Naor, J.; Schwartz, R. A unified continuous greedy algorithm for submodular maximization. In Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. IEEE, 2011, pp. 570–579.
42. Korte, B.; Lovász, L. Mathematical structures underlying greedy algorithms. In Proceedings of the Fundamentals of Computation Theory; Gécseg, F., Ed., Berlin, Heidelberg, 1981; pp. 205–209.
43. Koutsoupias, E.; Papadimitriou, C.H. On the greedy algorithm for satisfiability. *Information processing letters* **1992**, *43*, 53–55.
44. Fung, G. A Comprehensive Overview of Basic Clustering Algorithms. 2001.

 doi:10.20944/preprints202508.1159.v1

45. Hammouda, K.; Karray, F. A comparative study of data clustering techniques. *University of Waterloo, Ontario, Canada* **2000**, *1*.
46. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: a review. *ACM computing surveys (CSUR)* **1999**, *31*, 264–323.
47. Yong Wang.; Hodges, J. Document Clustering with Semantic Analysis. In Proceedings of the Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 2006, Vol. 3, pp. 54c–54c. https://doi.org/10.1109/HICSS.2006.129.
48. Kohavi, R. Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In Proceedings of the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996, KDD'96, p. 202–207.