ENHANCED SLICING-BASED ANONYMIZATION APPROACH FOR PRIVACY-PRESERVING DATA PUBLISHING WITH IMPROVED DATA UTILITY

MOHAMMED MAHFOUDH KHAMIS
BINJUBEIR

DOCTOR OF PHILOSOPHY

UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : MOHAMMED MAHFOUDH KHAMIS BINJUBEIR

Date of Birth : 18/08/1977

Title : ENHANCED SLICING-BASED ANONYMIZATION APPROACH FOR PRIVACY-PRESERVING DATA PUBLISHING WITH IMPROVED DATA UTILITY

Academic Session : SEMESTER II 2023/2024

I declare that this thesis is classified as:

□	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
□	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☒	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang Al-Sultan Abdullah reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang Al-Sultan Abdullah
2. The Library of Universiti Malaysia Pahang Al-Sultan Abdullah has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

New IC/Passport Number
Date: 13 / 5 / 2024

(Supervisor's Signature)

MOHD ARFIAN ISMAIL

Name of Supervisor
Date: 13 / 5 / 2024

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

SUPERVISOR’S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy.

____

(Supervisor’s Signature)

Full Name : TS. DR. MOHD ARFIAN BIN ISMAIL

Position : SENIOR LECTURER

Date : 13/05/2024

اونيۏرسيتي مليسيا ڤهڠ السلطان عبدالله
UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

STUDENT’S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang Al-Sultan Abdullah or any other institutions.

____

(Student’s Signature)

Full Name : MOHAMMED MAHFOUDH KHAMIS BINJUBEIR
ID Number : PCS17014
Date : 13/05/2024

ENHANCED SLICING-BASED ANONYMIZATION APPROACH FOR PRIVACY-PRESERVING DATA PUBLISHING WITH IMPROVED DATA UTILITY

MOHAMMED MAHFOUDH KHAMIS BINJUBEIR

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy

Faculty of Computing

UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

MAY 2024

ACKNOWLEDGEMENTS

First and foremost, praises and thanks to Allah, the Almighty, for His showers of blessings throughout my research work and the grace to complete the research successfully. Secondly, there are a number of people without whose help this thesis would not have been possible. First, I would like to thank my supervisor Dr. Mohd Arfian Ismail for his insightful advice and encouragement in the writing and completion of this thesis. His understanding of the subject and supervision skills have helped me during my research. It was a great privilege and honour to work and study under his guidance.

My gratitude also goes to Universiti Malaysia Pahang for granting me the opportunity to pursue my study. I would like to express my gratitude to the Faculty of Computing and all staff in the faculty for their support during my study.

My profound gratitude goes to my family here and in Yemen for their love, understanding, prayers, and continuing support to complete this research work.

My special thanks go to Dr. Abdulghani Ali Ahmed, for the keen interest he showed towards the successful completion of this study.

I am also grateful to Sana'a Community College, Yemen, for providing me with the scholarship to pursue my Ph.D. degree in Universiti Malaysia Pahang. I would like to thank my friends in Universiti Malaysia Pahang for their insightful comments and support.

ABSTRAK

Penerbitan data merupakan kaedah yang banyak digunakan untuk berkongsi data, terutamanya dalam bidang penyelidikan, kerana ia membolehkan operasi perlombongan data untuk mengekstrak pengetahuan berharga daripada pangkalan data yang diterbitkan. Pengetahuan ini boleh digunakan untuk representasi, interpretasi, atau penemuan corak menarik. Walau bagaimanapun, potensi penuh data parsial yang diterbitkan, berasal dari set data besar atau siri set data, belum sepenuhnya diwujudkan, terutamanya disebabkan oleh pelbagai cabaran yang dihadapi oleh para sarjana semasa mengekstrak pengetahuan dari data yang diterbitkan. Salah satu cabaran yang penting berkaitan dengan privasi data, yang sering menyebabkan pendedahan identiti individu, akses tidak dibenarkan kepada maklumat peribadi, dan penyalahgunaan data peribadi untuk tujuan yang tidak diingini. Isu ini telah menjadi halangan besar kepada kemajuan data yang diterbitkan. Bagi mengatasi kebimbangan ini dan memastikan kegunaan data, beberapa pendekatan berdasarkan penyamaran telah dibangunkan dalam bidang Penerbitan Data Penjagaan Privasi (PPDP). Kekberkesanan pendekatan anonimisasi data bergantung kepada pelbagai metode perlindungan yang digunakan untuk mencapai privasi. Walau bagaimanapun, metode perlindungan ini sering mengubah data secara berlebihan atau menuntut tahap kepercayaan yang tidak praktikal dalam pelbagai skenario berkongsi data. Melindungi data peribadi daripada orang-orang yang tidak boleh mengakses maklumat ini dan kemampuan individu untuk menentukan atau menarik kesimpulan mengenai identiti individu yang dapat mengakses maklumat peribadi mereka adalah aspek penting dalam metode perlindungan data. Meningkatkan metode perlindungan untuk publikasi data adalah penting untuk mencapai keseimbangan antara utiliti data dan privasi individu, yang merupakan cabaran besar. Untuk mencapai anonimisasi data yang berkesan, kajian ini mencadangkan pendekatan ditingkatkan yang dikenali sebagai pendekatan perlindungan berdasarkan tahap Upper Lower (UL), berdasarkan pendekatan pemotongan. Pendekatan UL bertujuan untuk mencapai keseimbangan yang lebih baik antara utiliti dan privasi. Kajian ini mencadangkan satu metodologi yang melibatkan pembahagian data kepada bahagian mendatar dan menegak, serta memanfaatkan Tahap Perlindungan Rendah (LPL) dan Tahap Perlindungan Tinggi (UPL) untuk mengira atribut yang unik dan serupa. Dengan menukar atribut-atribut ini, data yang diterbitkan dapat dijaga daripada risiko pendedahan sambil memastikan kepelbagaian yang mencukupi. Idea utamanya adalah memilih set atribut untuk menentukan tahap perlindungan yang diperlukan dan menukar di antara mereka untuk meningkatkan privasi data yang dipublikasikan sambil mengekalkan utiliti data yang tinggi. Dataset Dewasa, yang mengandungi dataset sebenar, digunakan, dan menurut keputusan, pendekatan UL dapat mengekalkan kegunaan data sambil menawarkan perlindungan privasi yang ditingkatkan. Pendekatan yang dicadangkan memberikan utiliti data sekitar 92.47%, yang lebih tinggi daripada yang dicapai apabila peratusan tahap pertukaran adalah 2% menggunakan LPL dan 98% menggunakan UPL dengan dataset pendidikan berukuran 4.5K. Dengan kadar pertukaran 5%, pendekatan yang dicadangkan mencapai 92.19% menggunakan LPL dan 95% menggunakan UPL. Secara kesimpulannya, pendekatan UL mengurangkan risiko pendedahan data berbanding dengan kerja-kerja sedia ada seperti penggabungan, $e - DP$, Mondrian, komposisi, probabiliti, dan kaedah hibrid. Dengan menggunakan pendekatan ini, publikasi data dapat dilaksanakan dengan cara yang memastikan kegunaan data secara praktikal sambil melindungi privasi individu. Menjaga keseimbangan antara utiliti dan privasi adalah penting, dan pendekatan UL menawarkan penyelesaian yang berjanji untuk mencapai keseimbangan ini. pendekatan UL menawarkan penyelesaian yang menjanjikan untuk mencapai keseimbangan ini.

ABSTRACT

Data publication is a widely used method for sharing data, particularly in research fields, as it allows for data mining operations to extract valuable knowledge from published databases. This knowledge can be utilized for representation, interpretation, or the discovery of interesting patterns. However, the full potential of published partial data, derived from large datasets or a series of datasets, is yet to be realized, primarily due to various challenges faced by scholars during the extraction of knowledge from published data. One significant challenge is related to data privacy, which often results in the disclosure of individuals' identities, unauthorized access to private information, and the misuse of personal data for unintended purposes. This issue has become a major hindrance to the advancement of published data. To address these concerns and ensure data utility, several anonymization-based approaches have been developed in the field of Privacy-Preserving Data Publishing (PPDP). The effectiveness of data anonymization approaches relies on different protection methods employed to achieve privacy. However, these protection methods often either excessively falsify data or demand an impractically high level of trust in different data-sharing scenarios. Protecting private data from people who must not access this information and the individuals' capability to determine or infer the identity of individuals who can access their personal information are crucial aspects of data protection methods. Improving protection methods for data publication is crucial to strike a balance between data utility and individuals' privacy, presenting a significant challenge. To achieve effective data anonymization, this study proposes an enhanced approach called the Upper Lower (UL) level-based protection approach, based on the slicing approach. The UL approach aims to strike a better balance between utility and privacy. The study proposes a methodology involving the division of data into horizontal and vertical partitions and leveraging the Lower Protection Level (LPL) and Upper Protection Level (UPL) to compute unique and identical attributes. By swapping these attributes, the published data can be effectively safeguarded against disclosure risks while still preserving adequate diversity. The key idea is to choose a set of attributes to determine the required level of protection and swap between them to improve published data privacy while preserving high data utility. The Adult dataset, which included a real dataset, was used, and according to the results, the UL approach could maintain the data's usefulness while offering improved privacy preservation. The proposed approach delivers about 92.47% data utility, which is more than what is achieved when the percentage of exchange level is 2% using LPL and 98% using UPL with a 4.5K education dataset. With a 5% swap rate, the proposed approach obtains 92.19% using LPL and 95% using UPL. In conclusion, the UL approach minimizes the risk of data disclosure compared to existing works such as merging, $e-DP$, Mondrian, composition, probabilistic, and hybrid methods. By employing this approach, data publication can be carried out in a manner that ensures practical usability of data while protecting individuals' privacy. Striking a balance between utility and privacy is crucial, and the UL approach offers a promising solution to achieve this balance.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CIA	Confidentiality, integrity, and availability
Col	Column
CP	Certainty Penalty
EC	Equivalence class
E-DP	E-Differential Privacy
GWO	Grey Wolf Optimizer
IA	Identity Attribute
IT	Information Technology
LPL	Lower Protection Level
MSCC	Mean square contingency coefficient
NCP	Normalized Certainty Penalty
PAM	Partitioning around medoids algorithm
PCC	Pearson correlation coefficient
PPDM	Privacy-Preserving Data Mining
PPDP	Privacy-Preserving Data Publishing
QI	Quasi-Identifier
SA	Sensitive Attribute
T	Table
t	Tuples
UL method	Upper and Lower Level-Based Protection approach
UPL	Upper Protection Level

CHAPTER 1

INTRODUCTION

1.1 Background

Over the past few years, the increasing prevalence of advanced computing techniques has led to the creation of vast amounts of data, commonly referred to as "big data" (Maniam and Singh, 2020; Majeed and Lee, 2021). Numerous organizations across various sectors, including government, banking, medical, insurance, and public and private institutions, have actively sought to make their data published to the public (Binjubeir et al., 2020). This effort involves collecting data from clients or users for research, analysis, exploration, or other purposes (Maniam and Singh, 2020; Lei Xu et al. 2014; Mendes and Vilela, 2017). This data has been recognized as a transformative force in the digital age, with experts likening its significance to that of "new oil" in society (Jayapradha et al., 2022). The data collected is often unstructured or complex and sourced from multiple channels, such as sales records, internet-of-things sensors, social media, healthcare patient records, and image or video archives (Binjubeir et al., 2020).

Data publishing helps researchers in data analytics to extract new and valuable information from partial data derived from big data sets (BinJubier et al., 2022). This information can be used for diverse purposes, such as representing, interpreting, or discovering intriguing patterns that can improve organizational productivity and assist in their potential plans. Analyzing and extracting patterns or knowledge from published data sets is a crucial practice that organizations use to obtain new or useful information (Jubeir et al., 2020; Jeba et al., 2022; Maniam and Singh, 2020; Lei Xu et al. 2014; Mendes and Vilela, 2017). The ability to analyze data, discover novel insights, and protect sensitive data is a crucial competitive advantage for organizations today (BinJubier et al., 2022; Cavanillas, Curry, and Wahlster, 2016). However, the potentials of published partial data are yet to be realized as scholars are facing several problems during knowledge extraction from the published data (Nasiri and Keyvanpour, 2020). One of such challenges is the issue related to data privacy that leads to the disclosure of individuals' identity. In addition, recent advancements in the field of learning technology have significantly threatened the secure propagation of private data over the web (Jayapradha et al., 2022; BinJubier et al., 2022; Zigomitros et al., 2020). It has led to the limited availability of datasets to researchers (Yu, 2016; Siddique et al., 2018; Olatunji et al., 2022). With the internet being a prime source of privacy breaches and cyber-attacks, hackers have been presented with numerous opportunities, resulting in the insecurity of data and information (Kumar et al., 2019).

In general, two common models for data publication have been proposed: the multiple publication model from the same data publisher and the single publication model from several data publishers (Hasan et al., 2018). Multiple data publications involve a series of datasets released at different times, each providing extensions to certain aspects (e.g., quarterly released data) (Gkoulalas-Divanis and Loukides, 2015; Wong et al., 2010; Wong and Fu, 2010). When datasets are from the same publisher, this implies that the publisher has knowledge of all the original data. Several privacy approaches exist (Chen et al., 2009) to maintain data privacy. However, these approaches mainly focus on the single publication model (Li, Li, and Venkatasubramanian, 2007; Machanavajjhala et al., 2007; Sweeney, 2002b; Hasan et al., 2018), where the dataset is anonymized by the publisher without considering other published datasets. An attacker can then launch a composition attack (Ganta, Kasiviswanathan, and Smith, 2008; Hasan et al., 2018) on the published data to compromise their privacy. The single publication model can be exemplified in a scenario where patients visit hospitals A and B for specialized procedures or follow-up care. In this case, both hospitals anonymize their original data by generalizing or replacing non-sensitive attributes with new values. The anonymized data is subsequently made available to the intended recipients without taking into account the published datasets from the other hospital. The composition attack is a way used by attackers to link available records in the microdata to an external database to identify individuals and exploit sensitive information. This attack can be carried out by combining the quasi-identifier (QI) attributes of different datasets to identify individuals. The attack becomes more complex with the availability of more datasets from different data collectors (Chen et al. 2009b; BinJubier et al. 2022). Assume that the following information of the patient is known to the attacker (Age = 25 years old, Sex = Male, and lives in Zipcode = 132000). The attacker is aware of the hospitals, which were visited by the patient, and that the hospitals have individually published their data without consulting each other. This situation increases the risk of breaching the patient's privacy. However, exploiting the sensitive information of the patient becomes difficult for the attacker in either dataset, as both datasets satisfy k-anonymity and l-diversity (Jayapradha et al., 2022).

Two fundamental approaches exist for revealing published data in both models. The first approach is an interactive setup in which the data collector performs certain functions on the data in order to respond to the data analyzer's queries. The second approach is a non-interactive setting, where data sanitization is carried out before being published (Gao and Zhou, 2020; Narayanan, 2009). Furthermore, the type of privacy varies depending on the data and how it is used, where many methods are used to provide privacy (Shah and Gulati, 2016b; Binjubeir et al., 2020). Anonymizing data before publication is one of the most common and used practices for protecting individuals' privacy. Data anonymization intends to reduce the threat of disclosing personal information while preserving the possibility of using published data (Majeed and Lee, 2021).

1.2 Motivation

The widespread use of the Internet and information technology makes it easier for people to expand their virtual presence online (Aldeen, Salleh, and Razzaque, 2015). Meanwhile, the Internet and information technology have delivered a deluge of information, making it easy to collect, share, and exchange data (Yu and Member, 2016; Aldeen et al., 2015; Eom et al., 2020). This huge amount of data makes it crucial to develop tools that can discover the extraction of hidden knowledge to generate new or useful information that can represent, interpret, or discover interesting patterns. These tools are called data mining tools (Lei Xu et al., 2014; Han, Kamber, and Pei, 2012; Zorarpacı and Özel, 2021). Data mining promises to reveal what is hidden, but the owners will be upset if sensitive information is exposed to the public or adversaries. Additionally, publishing data is the most straightforward way to share data, enabling research organizations to extract information from public datasets using data mining activities.

With this information, intriguing patterns may be represented, interpreted, or discovered (Binjubeir et al., 2020; Gkoulalas-Divanis and Loukides, 2015).

Besides the importance of data mining operations in many applications, there has been a growing focus on the privacy risks associated with data publication. Consequently, potential breaches and abuses of data privacy have received a lot of attention. As a result, it is essential to ensure proper data protection because failure could lead to circumstances that could harm both people and organizations (Binjubeir et al., 2020; Lei Xu et al., 2014; Zorarpacı and Özel, 2021). For this reason, many establishments are caught between sharing information and protecting their privacy to get this valuable information (Binjubeir et al., 2020; Jubeir et al., 2020). Consequently, researchers have developed a novel research field called privacy-preserving data publishing (PPDP) (BinJubier et al. 2022; Chen et al. 2009), which has garnered considerable attention recently. PPDP focuses on eliminating privacy risks of individuals while preserving the utility of released data for data mining (Jayapradha et al., 2022; Jeba et al., 2022; Lei Xu et al., 2014; Yu and Member, 2016; Mendes and Vilela, 2017; Aldeen et al., 2015).

1.3 Problem Statement

The practice of data publishing allows research institutions to easily share their data with others, enabling research organizations to perform data mining operations and extract valuable insights from the published data. These insights can subsequently be utilized to represent, translate, or uncover new and exciting forms of information (Binjubeir et al., 2020; Gkoulalas-Divanis and Loukides, 2015). However, the potential of published data has yet to be explored because scholars face several challenges when extracting information from published data. One of these challenges is data privacy, which results in the exposing of individuals' identities, unauthorized access to information and private data, and the use of personal information for unintended purposes (Yu, 2016; Siddique et al., 2018; Hasan et al., 2018; BinJubier et al., 2022; Binjubeir et al., 2020; Zorarpacı and Özel, 2021; Jayapradha et al., 2022). Despite removing identity attributes (IAs) like names and social security numbers to protect data, in many cases, the remaining data can still be used to identify the person. Additionally, even if sensitive attributes (SA) are not explicitly disclosed, they may still be inferred through linking attacks, where the remaining attributes are linked with other data sources. This kind of attack is known as a composition or intersection attack (Zigomitros et al., 2020;Gkoulalas-Divanis and Loukides, 2015; Hasan et al., 2018; BinJubier et al., 2022; Binjubeir et al., 2020).

To address these concerns while still retaining data utility, various anonymization-based approaches have been developed in Privacy-Preserving Data Publishing (PPDP), such as slicing (Li et al., 2012), merging (Hasan et al., 2018), $e{-}DP$ (Mohammed et al., 2011), Mondrian (LeFevre et al., 2006), composition (Baig et al., 2012), probabilistic (Sattar et al., 2014) and hybrid (Li et al., 2016). The aim of data anonymization is to reduce the likelihood of revealing personal information while preserving the possibility of using published data and causing uncertainty in identity inference or sensitive value estimation (Lasko and Vinterbo, 2010; Majeed and Lee, 2021; BinJubier et al., 2022; Zorarpacı and Özel 2021; Olatunji et al., 2022). Anonymization-based approaches recurrently resort to using different protection methods, such as grouping methods, perturbation methods, and measurement correlation (similarity) methods (Jeba et al., 2022; BinJubier et al., 2022; Li et al., 2016; Hasan et al., 2018; Mohammed et al., 2011; Sattar et al., 2014; LeFevre et al., 2006; Baig et al., 2012; Wong and Fu, 2010), and causing uncertainty in identity inference or sensitive value estimation (Lasko and Vinterbo, 2010). Protection methods used with anonymization-based approaches aim to avoid attempts to identify the record owner's identity by converting a dataset's original values to the anonymized dataset. When performing the extraction of knowledge through data mining operations, the anonymous dataset is used instead of the original dataset (Jeba et al., 2022; BinJubier et al., 2022).

Anonymization approaches based on grouping methods are commonly used for privacy protection but face challenges in defending sensitive values against composition attacks (BinJubier et al., 2022). To mitigate this issue, the perturbation method is employed, modifying the real values of the dataset to create an anonymized version. The $e - DP$, composition, probabilistic, and hybrid approaches are highly respected for their effectiveness in privacy preservation through data perturbation and anonymization. However, this modification can impact data utility depending on the amount and type of noise or the specific properties of that data are not preserved (Jeba et al., 2022; Mivule, 2013). Introducing a correlation measure as a solution to enhance protection and preserve data utility is a promising protection method. By grouping highly correlated attributes together in columns and preserving their correlations, the correlation measure safeguards privacy. It achieves this by breaking associations between uncorrelated attributes in other columns through protection methods based on anonymization approaches, such as random permutation and generalization (Jayapradha et al., 2022; A. Hasan et al., 2018). However, recent correlation-based methods like slicing and merging face the drawback of relying on random permutation, which may not provide reliable protection against attribute or membership disclosure. Additionally, merging procedures can generate fake tuples, resulting in a loss of data utility and incorrect knowledge extraction (BinJubier et al., 2022; Majeed and Lee, 2021; Jeba et al., 2022; Anil Sharma et al., 2020; Rohilla, 2015). The following 2.4 section (Protection Methods Based on Anonymization) will describe in detail each of these protection methods.

Evidently, this gap is an opportunity to improve privacy protection through the design of a slicing-based approach. This approach would effectively prevent attackers from identifying individuals or disclosing sensitive information in a table, while simultaneously determining the optimal balance between privacy and data utility. Achieving this balance would require improved protection methods that identify specific attributes capable of detecting potential disclosure, enhancing the privacy of published data without compromising its utility. Protection methods can convert original dataset values into anonymized ones, which can then be used in data mining operations to prevent the identification of record owners (Jayapradha et al., 2022; Cunha et al., 2021). Nevertheless, most current approaches fail to adequately address the effectiveness of using anonymized data for data mining and determine the level of protection necessary to prevent the disclosure of private information while maintaining data utility (Lasko and Vinterbo, 2010; Binjubeir et al., 2020; Majeed and Lee, 2021; BinJubier et al., 2022; Eom et al., 2020; Cunha et al., 2021). Table 1.1 outlines this problem and underscores the need for the improvement of more effective protection methods.

Table 1.1 Summary of problem statement

No.	Problem	Description	Affect
1	Although numerous approaches have been proposed to address privacy concerns, many of them do not adequately consider the effectiveness of anonymized data when attempting to attain a high level of privacy. This critical issue has been highlighted in studies by (Mehmood et al., 2016; Hasan et al., 2016; Jeba et al., 2022; BinJubier et al., 2022). However, there is still a need for further research to identify the most effective approaches for achieving this balance between privacy and utility.	Data anonymization is a crucial practice that helps to mitigate the risks of revealing personal information when publishing data. This process involves using various protection methods to alter or mask sensitive values, thereby reducing the chances of identity inference.	By implementing anonymization approaches to mitigate the risk of disclosing individuals' information, a trade-off often arises between higher privacy and decreased data utility or increased data utility and decreased privacy.
2	The majority of current approaches lack adequate consideration of the effectiveness of using anonymized data for data mining and determining the level of protection necessary to prevent the disclosure of private information while preserving data utility (Lasko and Vinterbo, 2010; Binjubeir et al., 2020; Majeed and Lee, 2021; BinJubier et al., 2022; Eom et al., 2020). To address these issues, it is crucial to develop improved protection methods that accurately determine the level of protection required and identify specific attributes that can help detect potential disclosure.	The protection of specific attributes within data is crucial to ensure privacy. Anonymization-based approaches are commonly employed to transform original attribute values into anonymized ones, employing various protection methods. These anonymized values can subsequently be utilized in data mining operations, effectively safeguarding the identification of record owners.	Effect on the data utility and privacy

Table 1.1 shows that this study will address two significant issues: achieving a balance between privacy and utility in data applications, and determining the amount of protection required to achieve published data privacy while retaining more data utility.

1.4 Aim and Research Objectives

The primary goal of this research is to protect data privacy by lowering the risk of disclosing personal information and preserving the potential use of published data. The slicing approach has been designed to overcome the limitations of previous works in preventing unauthorized disclosure of individuals' identities, and it is widely used for data anonymization while maintaining data utility. Slicing can be performed through vertical and horizontal data partitioning. Vertical partitioning is applied when highly correlated attribute values are grouped into columns, while horizontal partitioning is applied when tuples are grouped into buckets; the attribute values are permutated randomly to break the linkages between different columns. However, a significant drawback of the slicing approach is the reliance on random permutation as a protection method, as it does not always guarantee adequate protection against attribute or membership disclosure. Moreover, striving for a high level of privacy using random permutation-based protection methods inevitably leads to some information loss.

To address these limitations, the proposed study introduces an enhanced approach called the Upper Lower (UL) level-based protection approach, based on the slicing approach. The proposed approach consists of two steps. In the first step, it uses horizontal partitioning (tuple partition) and vertical partitioning (attribute grouping) to anonymize published data, similar to the slicing approach. The second step involves the implementation of an improved protection method using Lower Protection Level (LPL) and Upper Protection Level (UPL). This method aids in identifying specific attributes that have the potential to reveal personal information and determines the appropriate level of protection required to prevent the disclosure of private information. By doing so, the UL approach effectively mitigates identity disclosure while achieving a balance between preserving privac and maintaining data utility. The aim is to safeguard personal information while still allowing the data to remain useful for various purposes. This overarching goal can be subdivided into the following specific objectives:

i. To improve the current slicing approach, which incorporates the randomized protection method, a novel approach is introduce termed Upper Lower (UL) level-based protection. This approach integrates horizontal partitioning (tuple partitioning) and vertical partitioning (attribute grouping) to anonymize the published data, resembling the slicing approach while eschewing the randomization protection method. The primary objective of this approach is to strike a balance between preserving privacy and optimizing data utility,

ii. To anonymize data using the Lower Protection Level (LPL) and Upper Protection Level (UPL) aims to enhance the effectiveness of the UL approach in determining the necessary protection level. This is achieved by selecting specific cell values that aid in preventing identity disclosure and broken the link between it through a rank swapping protection method. The utilization of the rank swapping protection method is intended to enhance data privacy, utility, and ensure l-diversity in the dataset.

iii. To evaluate the efficiency of the proposed approach by comparing its performance with other existing works.

Table 1.2 summarizes the mapping between research questions (RQ), research objectives (RO), and research contributions (RC) for this study. The first question led to the design of a slicing-based enhanced approach called the Upper Lower (UL) level-based protection approach that can be used for data anonymization of published data. The enhanced approach would effectively prevent attackers from identifying individuals or disclosing sensitive information in a table while simultaneously determining the optimal balance between privacy and data utility. The second objective answers the second research question, which introduces an improved protection method called the Lower Protection Level (LPL) and Upper Protection Level (UPL) for the anonymization approach to be more effective in determining the amount of protection required through selecting particular cell values that help prevent identity disclosure. By doing so, the UL approach effectively mitigates identity disclosure and breaks the linkages between the attribute values that compromise the privacy of published data, resulting in enhanced privacy for published data and the preservation of data utility. The third objective answers the third and fourth research questions and leads to the evaluation of the efficiency of the proposed approach by comparing its performance with other existing works.

Table 1.2 Mapping of research questions to objectives and contributions in this thesis

Research Question	**Research Objective**	**Contribution**
What approach(es) can be used for data anonymisation while simultaneously determining the optimal balance between privacy and data utility?	To design a slicing-based enhanced approach called the Upper Lower (UL) level-based protection approach that can be used for data anonymization of published data. The enhanced approach aims to achieve a better balance of utility and privacy before releasing any data product.	Propose an enhanced approach based on slicing that can be used for data anonymization of published data. This enhanced approach aims to effectively minimize the risk of identity disclosure and sever the connections between attribute values that could potentially compromise the privacy of the published data, resulting in enhanced privacy for published data and the preservation of data utility.
What amount of protection is needed to prevent private information disclosure whilst preserving data utility?	To propose an improved protection method called the Lower Protection Level (LPL) and Upper Protection Level (UPL) for the anonymization approach, which is more effective in determining the amount of protection required through selecting particular cell values that help prevent identity disclosure. By doing so, the UL approach effectively mitigates identity disclosure and breaks the linkages between the attribute values that compromise the privacy of published data, resulting in enhanced privacy for published data and the preservation of data utility.	rks by effectively determining the amount of protection required to prevent personal information disclosure and striking a balance between privacy and utility in the published data.
How do you evaluate the efficiency of the anonymization approach? and Can anonymized data be effectively used for data mining operations?	This study involved comparing the performance of the proposed approach to that of other existing works to assess its effectiveness.	The evaluation of the proposed approach revealed that this method has a high capacity to preserve more data utility and provide stronger privacy protection.

1.5 Contribution

This research introduces an enhanced approach called the Upper Lower (UL) level-based protection approach, which utilizes slicing to achieve data anonymization. Privacy-Preserving Data Publishing (PPDP) encompasses various protection methods and approaches, with anonymization being one of the most widely used. The primary objective of this research is data anonymization while maintaining an optimal balance between privacy and data utility. The contributions of this research can be summarized as follows:

i. Designing the UL approach, an enhanced slicing-based method that effectively prevents attackers from identifying individuals or disclosing sensitive information in the data table. This approach achieves a better balance between privacy and data utility,

ii. Introducing the UPL and LPL methods as improved protection methods within the anonymization approach. These methods address limitations present in existing works by accurately determining the level of protection required to prevent the disclosure of personal information. Unlike random way used in other approaches to break attribute value correlations, the UPL and LPL methods identify specific attributes for swapping. Selectively swapping specific cells is crucial for enhancing data privacy, preserving valuable information, and ensuring l-diversity in the published microdata table,

iii. Validating the proposed approach by utilizing existing data from related works to evaluate its effectiveness against composition attacks. A comprehensive evaluation is performed to compare the proposed approach with existing methods in terms of preserving data utility and privacy.

1.6 Scope and Limitation

The motivation behind preserving published data privacy stems from the proposal of an enhanced approach known as the UL approach, based on slicing. This approach aims to prevent attackers from identifying individuals or disclosing sensitive values in the table, while achieving a better balance between privacy, information loss, and utility. To ensure the efficient execution of the proposed UL approach, the implementation environment is carefully selected by simulating all relevant scenarios. The research scope encompasses data, attack, and privacy types. Table 1.3 provides a summary of the study's scope and limitations.

Table 1.3 Research Scope and Limitations

No	Items	Scope of Research	Reason
1	dataset	Adult dataset (Kohavi and Becker, 2019)	The "Adult" dataset was utilized in the experiments to evaluate and compare the results with other existing works.
2	Data form	microdata form	Microdata contains the actual information and offers the advantage of conducting analyses that may not be feasible with other data forms.
3	Privacy type	Information privacy	Information privacy is concerned with ensuring that sensitive values are not disclosed and that the identities of individuals or groups cannot be deduced from the information collected by data collectors.
4	Attack type	A composition attack	When addressing privacy preservation in anonymization approaches, it is crucial to assess the potential threats presented by adversaries with access to externally accessible data and diverse inference methods. Understanding these risks is key to effectively safeguarding privacy.

1.7 Thesis Outline and Organization

Chapter 1 discusses the introduction of the research, encompassing the study's background, motivation, problem statements, aims and objectives, contributions, scope, and limitations. It concludes by summarizing the thesis outline and organization.

Chapter 2 comprises the literature review, providing an overview of the approaches and methods used for privacy-preserving. It delves into the concept of privacy, anonymization approaches, protection methods based on anonymization, and data utility, along with measuring risks.

Chapter 3 outlines the research procedures, presenting the stages, methods, and their relationships to clarify the design of the proposed solution. The research methodology explains the proposed approach and the involved algorithms at each stage, followed by dataset initialization and performance evaluation.

Chapter 4 introduces the design of the UL approach and the improved protection methods, UPL and LPL. It also includes the measurement of the required protection level. The chapter evaluates the effectiveness of the approach in preserving data utility, comparing it with other existing works, and assessing its privacy protection against considered attacks.

Chapter 5 provides the thesis conclusion, summarizing the work, discussing key findings, and suggesting future research directions.

Except for the first and final chapters, each chapter begins with an introduction and concludes with a summary.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews privacy, anonymization approaches, and protection methods based on anonymization approaches. It begins with the definition of privacy and describes the approaches to privacy-preserving data publishing (PPDP). Following that, the anonymization approach is explained, and a discussion of the protection methods based on anonymization approaches ensues, comparing their advantages and disadvantages. Additionally, the approaches used under each protection method are compared in this chapter. Next, measures of risk disclosure and data utility are described. Finally, a summary is provided to recapitulate the key points of this chapter.

2.2 Privacy Definition

Although various definitions of privacy exist, providing an accepted standard definition of this concept is difficult (Cranor et al., 2016; Banisar and Davies, 1999; Mendes and Vilela, 2017; Gan et al., 2018). In 1948, the Universal Declaration of Human Rights (Banisar and Davies, 1999) established privacy as a fundamental right. However, due to the fact that privacy can vary in scope, such as in communication, at home, and with family, its definition may be perceived as having a very restricted application. According to (Banisar and Davies, 1999; Mendes and Vilela, 2017), privacy is commonly categorized into bodily privacy, communication privacy, information privacy, and territorial privacy, as illustrated in Figure 2.1.

The collection and management of personally identifiable information is part of information privacy. Individuals' bodily privacy protects them from invasive procedures, such as drug testing and others. Communication privacy entails protecting the privacy of all forms of communication. Territorial privacy is intended to limit intrusion into a regional environment.

The study of information privacy is primarily concerned with interaction and content privacy. Content privacy involves the prevention of identity disclosure from anonymized or encrypted databases, such as the extraction of credit card information from a state or national database. Interaction privacy, on the other hand, refers to preventing the disclosure of an individual's content, such as checking victims' encrypted web traffic or using a voice fingerprint to access services (Yu, 2016).

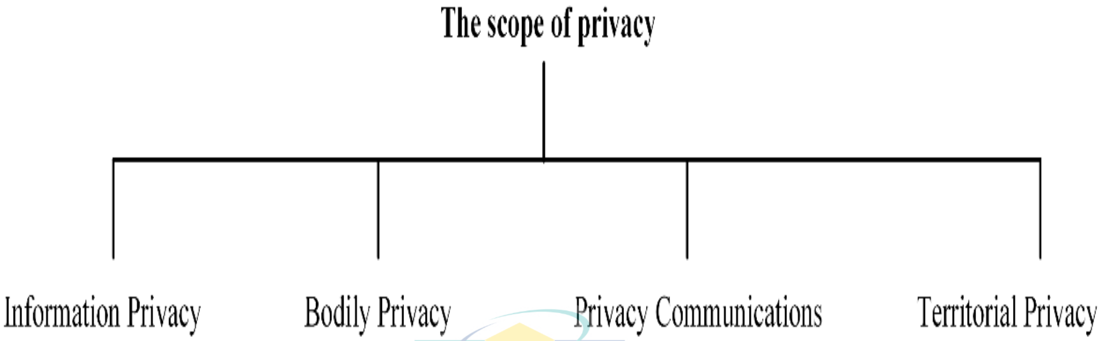


Figure 2.1 Scope of privacy

As a result, the current study adopts a definition of privacy in content and interaction (Mendes and Vilela, 2017; Banisar and Davies, 1999), which is relevant to the research path involving the collection and analysis of individual data, ultimately assisting organizations in improving their efficiency and supporting future plans.

Despite the overlap between confidentiality and privacy in some contexts, their respective concepts and protection methods are distinct. Confidentiality is viewed as data-related, focusing on protecting data from unauthorized access, modification, or loss when transferred across a network (Senosi and Sibiya, 2017; Binjubeir et al., 2020). On the other hand, privacy has an additional "data owner-oriented" concept as it deals with the data owners and aims to protect the private information of the data owners (Wang et al., 2019). The variation in privacy arises from the types of data collected and their uses, leading to the adoption of various approaches to ensure privacy (Shah and Gulati, 2016b).

When addressing privacy concerns, multiple methods are employed, but some information loss is inevitable, and there is a trade-off between data utility and information loss. Data utility refers to the extent to which modified data can be effectively used for in-depth analysis while satisfying privacy requirements. However, existing generic solutions fail to address all privacy concerns associated with protecting sensitive information while preserving data utility. Previous studies have focused on finding effective protocols for specific problems, considering the trade-offs between data utility and information loss during privacy-enhancing data mining (Ding and Klein, 2010; Shah and Gulati, 2016a; Bhaladhare and Jinwala, 2016; Yu et al., 2018; Yu et al., 2016).

In recent years, a subfield of data mining known as PPDP (Privacy-Preserving Data Publishing) has grown to address privacy concerns during data publication. A crucial aspect of PPDP is manipulating the data using a specific method so that a good data mining model can be developed on modified data that satisfies a specified privacy requirement with minimal information loss for the intended data analysis task. Here, data mining operations will be used without jeopardizing the security of sensitive individual data, especially at the record level (Wong and Fu, 2010).

2.2.1 Privacy Preservation Data Publishing (PPDP)

Data collectors collect vast quantities of data from data providers on data warehouse servers to support subsequent data mining operations. The data collected may contain sensitive personal information about individuals. Consequently, the purpose of PPDP (Privacy-Preserving Data Publishing) is to protect privacy during data collection and dissemination to various data mining servers (Zhang and Zhao, 2007; Yin et al., 2017). In addition to protecting privacy, PPDP considers the utility of data for effective data mining operations, ensuring that sensitive information remains unintelligible, and record owners remain unidentifiable within a set of other records, making it difficult for attackers to link them to specific records. Since ideas about privacy vary, so do the ways they protect it. No existing generic solutions can handle all privacy issues while effectively protecting sensitive information from unwanted disclosure (Binjubeir et al., 2020). Hence, several methods are used to provide privacy (Shah and Gulati, 2016b). Three approaches have been developed for privacy-preserving information (Bertino, Lin, and Jiang, 2008; Binjubeir et al., 2020). As illustrated in Figure 2.2, these approaches are data exchange (Conway and Strip, 1976), data cryptography (Yang, Zhong, and Wright, 2005), and data anonymization (Agrawal and Srikant, 2000; Agrawal and Haritsa, 2005; Binjubeir et al., 2020).

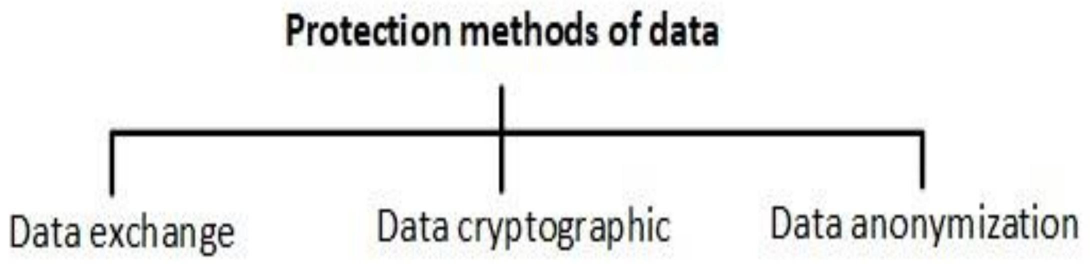


Figure 2.2 Data protection methods in PPDP

Using a data exchange approach, private information may be disseminated from (at least) one data source to another. Therefore, this approach only works for systems with reliable data sources, where none of the data providers have any intention to compromise the disseminated private information. In several real-world systems, data providers are rarely trusted because they could seek to undermine the distributed private information. Hence, private information cannot be fully protected from compromise with this data exchange approach (Clifton et al., 2004; Zhang, 2006).

When using data cryptography, multiple parties (referred to as data providers) collaborate to compute results or analyze non-sensitive information. Each data provider possesses a pair of public and private keys. Furthermore, all parties, including the data collectors (servers for the data warehouse), must have access to the public keys of all data providers. Initially, all data providers are provided with the combined sum of public keys, which they use to encrypt their data and transmit it securely to the data warehouse servers. As a result, no participant gains access to information beyond their own input. Once the encrypted data is received, the data warehouse servers can utilize mathematical manipulations to generate precise models. These models serve to address privacy concerns among competitors or other untrustworthy entities (Vaghashia and Ganatra, 2015; Lindell, 2011; Andrew et al., 2019). However, the complexity of this approach may lead to high computational costs for data providers and warehouse servers, making it impractical (ElGamal, 1985; Luo and Wen, 2014).

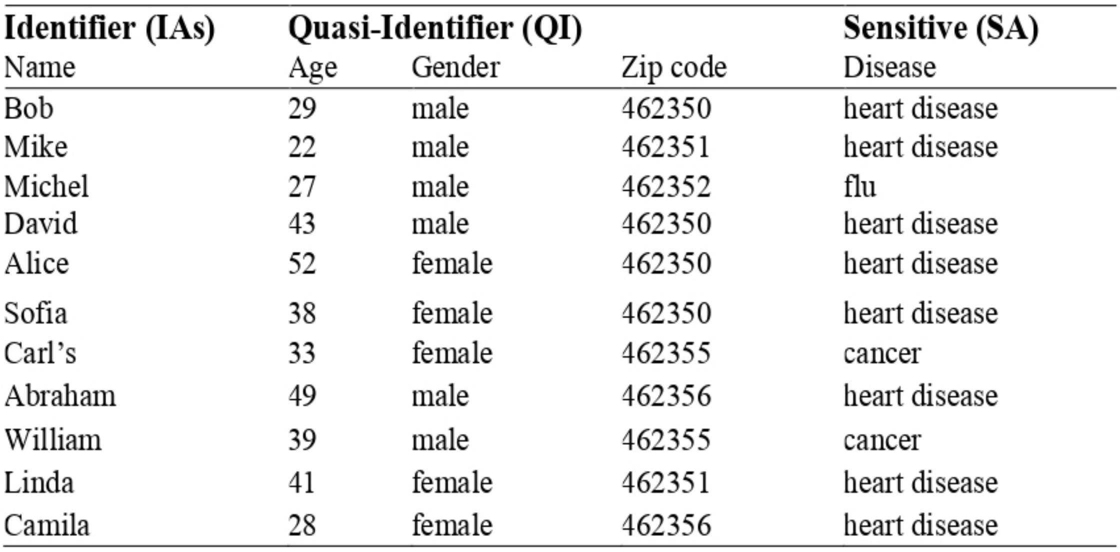
To counteract attempts to identify the owners of specific records, the data anonymization approach is employed (Binjubeir et al., 2020; Hasan et al., 2018). This approach aims to preserve the usefulness of the data while ensuring the protection of individual identities. The upcoming section delves further into the discussion of the data anonymization approach, as it holds the primary focus of this thesis.

2.3 Data Anonymization Approaches

Anonymization of data is an approach used to safeguard data in a manner that prevents attackers from making inferences about the identities of individuals or disclosing private information to those individuals. It is necessary to treat the collected data as a private table called the microdata table T. This table comprises a set of tuples, and each tuple t is regarded as a client, possessing various attributes linked to clients (Jubeir et al., 2020; Aldeen et al., 2015; Olatunji et al., 2022). Typically, these attributes are categorized as Identity Attribute (IA), which explicitly identifies the records of the client (e.g., name, cellular numbers, and driver's license numbers); Quasi-identifier (QI) attributes refer to a sequence of individuals' non-explicit attributes (e.g., gender, nationality, ZIP code, and age) where no single attribute can give specific identification of the person; instead, all the attributes must be combined to identify the person. The QI attributes include two types: numeric and categorical, as depicted in Figure 2.3. Additionally, there are Sensitive Attributes (SA), which denote confidential information (see Table 2.1) (Sharma, 2017; A Machanavajjhala et al., 2006; Jeba et al., 2022; Zigomitros et al., 2020).

Therefore, it is expected that organizations publish only partial data derived from their datasets in the form of microdata. This approach can enhance an organization's reputation or support future plans without divulging the proprietorship of the sensitive data. Even though the attributes (IAs) that identify users in the table, such as names, cellular numbers, and driver's license numbers, are removed under the assumption that anonymity is maintained, the remaining data can, in most cases, be used to re-identify the person. Additionally, sensitive attributes (SA) may still be exposed through linking attacks, where sensitive attributes are associated with other public data sources. This situation is known as a "composition attack" or "intersection attack." As a result, anonymization can only be accomplished by modifying these attributes to obscure the relation between the individual and specific values, thereby preventing such attacks and preserving the potential application of the published data (Hasan et al., 2018; Jeba et al., 2022).

Table 2.1 Medical database of patients



Identifier (IAs)	Quasi-Identifier (QI)			Sensitive (SA)
Name	Age	Gender	Zip code	Disease
Bob	29	male	462350	heart disease
Mike	22	male	462351	heart disease
Michel	27	male	462352	flu
David	43	male	462350	heart disease
Alice	52	female	462350	heart disease
Sofia	38	female	462350	heart disease
Carl’s	33	female	462355	cancer
Abraham	49	male	462356	heart disease
William	39	male	462355	cancer
Linda	41	female	462351	heart disease
Camila	28	female	462356	heart disease

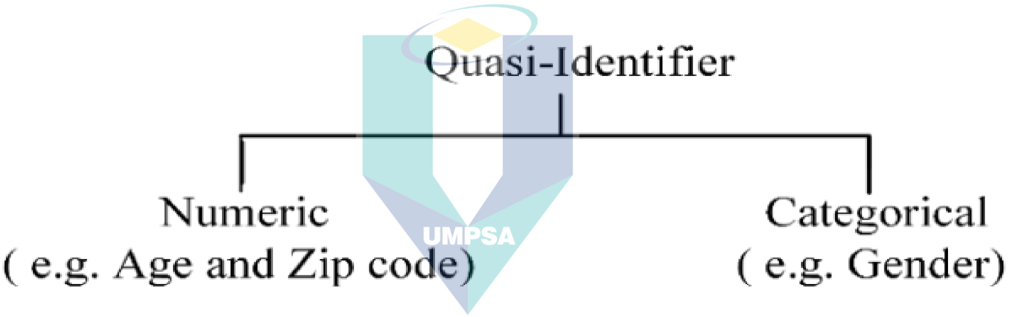


Figure 2.3 Types of QI attributes

The composition attack is the result of mixing different publicly available datasets. Since datasets are rarely isolated, attackers rely on the intersection of datasets to exploit sensitive information, given that they are a combination of other datasets. The complexity of this problem increases with the availability of more datasets from various data collectors (Gkoulalas-Divanis and Loukides, 2015). Records in published datasets are typically arranged in small groups known as an equivalence class. All individuals in the equivalence class are similar and associated with sensitive values, depending on the protection method used (Gambs, Killijian, and del Prado Cortez, 2010; Hasan, Jiang, and Li, 2017). A person's privacy is compromised if the adversary's confidence is significantly greater than a random guess.

Anonymization approaches have proven to be an effective means of preserving privacy. In recent years, protecting published data through anonymization approaches has been extensively studied (Binjubeir et al., 2020; Hasan et al., 2018). These approaches aim to modify the attributes (QI values) to weaken the linkage between QI and SA values, thereby preventing such attacks and ensuring the preservation of privacy for published data. To achieve effective privacy preservation, a variety of protection methods based on anonymization approaches are employed prior to data publication. These protection methods are carefully designed to prevent the identification of record owners while simultaneously maintaining the utility of the data (Jeba et al., 2022; Zigomitros et al., 2020). The following section will describe these protection methods based on anonymization approaches.

2.4 Protection Methods Based on Anonymization Approaches

Anonymization approaches employ various protection methods, which can be used and combined within the same approach to introduce uncertainty into identity inference or sensitive value estimation. Protection methods based on anonymization approaches aim to avoid attempts to identify the record owner's identity by converting a dataset's original values to an anonymized dataset. When performing data mining operations, the anonymous dataset is used instead of the original dataset. In this study, these protection methods based on anonymization approaches are classified into three, as illustrated in Figure 2.4 grouping methods, perturbation methods, and measurement correlation (similarity) methods (Binjubeir et al., 2020; Hasan et al. 2018). The following subsections describe these methods in further detail.

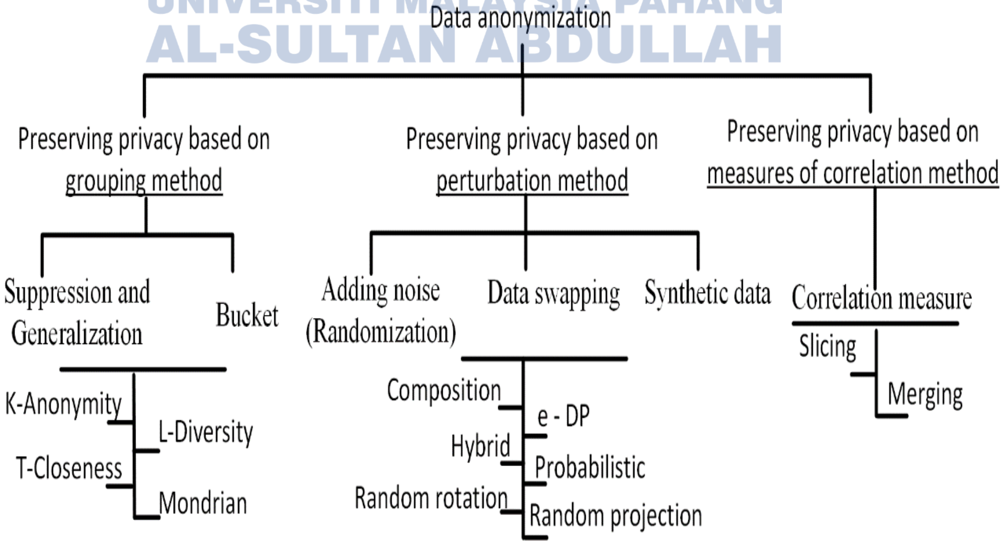

Data anonymization
Preserving privacy based on grouping method
Preserving privacy based on perturbation method
Preserving privacy based on measures of correlation method
Suppression and Generalization
Bucket
Adding noise (Randomization)
Data swapping
Synthetic data
Correlation measure
Slicing
Merging
K-Anonymity
L-Diversity
T-Closeness
Mondrian
Composition
e - DP
Hybrid
Probabilistic
Random rotation
Random projection

Figure 2.4 Classification of the protection methods of the data anonymisation approaches

2.4.1 Preserving Privacy Based on The Grouping Method

This method divides the entire records horizontally into several groups or partitions and only allows each tuple to belong to one group (Xiao and Tao, 2006; Sweeney, 2002b). The goal of this procedure is to make it harder for a person to identify with their SA values in the group by weakening the linkage between the QI and SA values. Grouping is often implemented using suppression and generalization and bucketization and/or combined.

Suppression and generalization are effective sensitive data protection methods against unauthorized access as they hide or replace some details of the attributes. Furthermore, both methods address different attributes individually, meaning that they only adjust the selected values that will minimize utility loss (Verykios et al., 2004; Olatunji et al., 2022). In the suppression way, the values of the attributes are replaced in the table with ANY, denoted by "*". This means some or all values of the attribute are replaced by "*". The second operation is the generalization method. In the generalization method, a specific value of the attributes is replaced by a more general value according to a given taxonomy, thereby making the QI less identifying (Wong and Fu, 2010; Cunha et al., 2021).

There are two major ways of anonymizing information using the generalization method: global recoding and local recoding. For global recoding, once an attribute value is generalized, each value occurrence should be replaced by a new generalized value. For example, all values in the birth date attribute are generalized to years, or all values in nationality are related to continents. In local recoding, values may be generalized to different generalization domains. For example, local recoding may generalize values of the age attribute into [20–39], [40–59], and [60–90]. Hence, local recoding is more similar to the original data and can preserve more information than global recoding, making the data mining operations more accurate. Additionally, overlapping intervals are unsuitable for most classification tools as they complicate classification tasks (Wong and Fu, 2010; Wen, Wu, and Castiglione, 2017; Cunha et al., 2021).

Table 2.2 serves as a means of anonymizing data through the use of the suppression way. Upon examination, it becomes apparent that the first two tuples and the last tuple share the same quasi-identifier (QI) values. These three tuples can be considered as forming an equivalence class, representing a collection of tuples in the table that share identical QI values. Similarly, the third, fourth, and fifth tuples form another equivalence class or QI-group. On the other hand, generalization is employed to transform specific values into more generalized forms. Numeric values can be changed to value ranges. For example, in Table 2.3, the value 23 can be generalized to the range 23-30. In the case of categorical values, they can be replaced with other categorical values representing broader concepts than the original values. For instance, the category "gender" can be generalized to "person."

Table 2.3, it also serves as a method for anonymizing data using the generalization approach. Upon examination, it becomes evident that the first three tuples share identical quasi-identifier (QI) values, forming an equivalence class. The remaining tuples form another equivalence class because they share identical quasi-identifier (QI) values (Wong and Fu, 2010).

Suppression and generalization ways primarily target quasi-identifier (QI) attributes, such as age, zip code, and gender, while overlooking the sensitive attributes (SA) of individuals (Wong and Fu, 2010; Jeba et al., 2022). These ways address distinct attributes separately, altering only the designated values (Wong and Fu, 2010; Jeba et al., 2022). In anonymized datasets, records are typically grouped based on shared QI values, forming equivalence classes. Within these classes, individuals exhibit similarities and are associated with sensitive values determined by the anonymization approach. Generalization provides the benefit of ensuring uniform attribute values within each group, thereby facilitating the analysis of published data (Wong and Fu, 2010). However, both suppression and generalization ways remain susceptible to indirect attacks, allowing the inference of individuals' attributes and potential identity disclosure when certain QI attributes are revealed. Moreover, both methods entail significant information loss while aiming for a heightened level of privacy (Li and He, 2023). For instance, in Table 2.2 and Table 2.3, the second equivalence class reveals a single match linked with two sensitive values for the "Gender" attribute, displaying "female" in Table 2.2 and "person" in Table 2.3. Consequently, despite attackers correctly identifying the victim's equivalence class, they may fail to accurately determine the associated sensitive value.

Table 2.2 Published by suppression

Quasi identifier table				**Sensitive table**
ID	Age	Gender	Zip code	Disease
1	*	male	*	flu
2	*	male	*	cancer
3	*	female	*	flu
4	*	female	*	heart disease
5	*	female	*	flu
6	*	male	*	heart disease

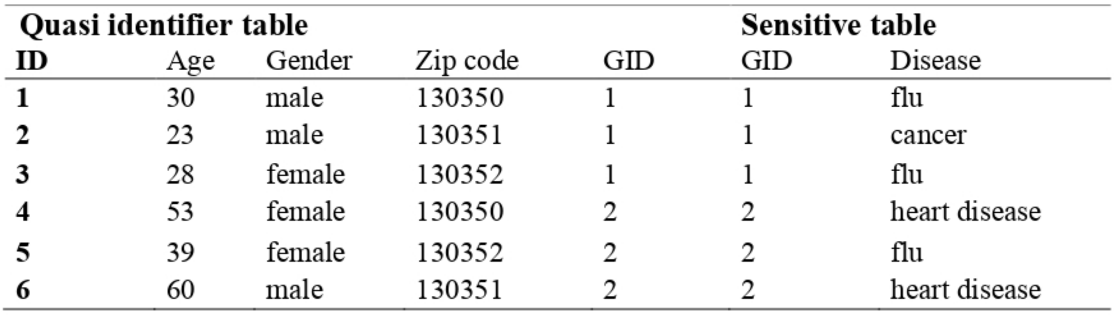
An alternative way for anonymizing data is known as bucketization, proposed by Xiao and Tao (Xiao and Tao, 2006). This method is similar to generalization in terms of creating equivalence classes but does not modify any attribute related to the quasi-identifier (QI) or the sensitive attribute (Li and He, 2023).

Table 2.3 Published by generalization

Quasi identifier table				**Sensitive table**
ID	Age	Gender	Zip code	Disease
1	23-30	person	130350	flu
2	23-30	person	130350	cancer
3	23-30	person	130350	flu
4	31-60	person	130351	heart disease
5	31-60	person	130351	flu
6	31-60	person	130351	heart disease

The bucketization process entails partitioning the original data table into non-overlapping partitions, yielding two distinct tables: the Quasi-Identifier (QI) table and the sensitive table. Each partition is assigned a unique identifier known as GID, where all tuples within a partition share the same GID value. These tuples are subsequently projected onto the QI attributes and the confidential attributes, resulting in the sensitive table. The primary objective of bucketization is to ensure that individuals within the same bucket exhibit indistinguishable values for the confidential attributes. Consequently, bucketized data complicates an observer's ability to associate specific records with individuals or infer sensitive information, as demonstrated in Table 2.4. Notably, the grouping achieved in Table 2.4 mirrors that of Table 2.3, albeit with Table 2.4 retaining all original tuple values, whereas Table 2.3 incorporates some generalized tuple values (Wong and Fu, 2010; Jayapradha et al., 2022)

Table 2.4 Published by bucketization



Quasi identifier table					Sensitive table	
ID	Age	Gender	Zip code	GID	GID	Disease
1	30	male	130350	1	1	flu
2	23	male	130351	1	1	cancer
3	28	female	130352	1	1	flu
4	53	female	130350	2	2	heart disease
5	39	female	130352	2	2	flu
6	60	male	130351	2	2	heart disease

The bucketization process entails partitioning the original data table into non-overlapping partitions, yielding two distinct tables: the Quasi-Identifier (QI) table and the sensitive table. Each partition is assigned a unique identifier known as GID, where all tuples within a partition share the same GID value. These tuples are subsequently projected onto the QI attributes and the confidential attributes, resulting in the sensitive table. The primary objective of bucketization is to ensure that individuals within the same bucket exhibit indistinguishable values for the confidential attributes. Consequently, bucketized data complicates an observer's ability to associate specific records with individuals or infer sensitive information, as demonstrated in Table 2.4. Notably, the grouping achieved in Table 2.4 mirrors that of Table 2.3, albeit with Table 2.4 retaining all original tuple values, whereas Table 2.3 incorporates some generalized tuple values (Wong and Fu, 2010; Jayapradha et al., 2022).

Based on the grouping method, some of the best approaches for ensuring data anonymity are the k-anonymity approach (Bayardo and Agrawal, 2005), l-diversity approach (Ashwin Machanavajjhala et al., 2006), T-closeness approach (Li et al., 2007) and Mondrian approach (LeFevre, DeWitt, and Ramakrishnan, 2006). The following subsections describe these approaches in further detail.

K-anonymity is a widely used and well-known privacy approach (Pawar et al., 2018). To protect individuals' privacy, Roberto and Samarati et al. (Roberto J Bayardo and Agrawal, 2005; Samarati and Sweeney, 1998) introduced the notion of k-anonymity to limit the disclosure of information. The concept of k-anonymity is based on altering the values of the quasi-identifier (QI) attributes to make it impossible for an attacker to determine the identities of individuals in a particular dataset while preserving the maximum utility of the disclosed data (see Table 2.5) (Roberto J Bayardo and Agrawal, 2005; Samarati and Sweeney, 1998).

Table 2.5 Three anonymous versions of the medical patient database relating to Table 2.1

ID	Equivalence Class	Age	Gender	Zip code	Disease
1	1	2*	Person	462***	heart disease
2		2*	Person	462***	heart disease
3		2*	Person	462***	flu
4		3*	Person	462***	heart disease
5	2	3*	Person	462***	cancer
6		3*	Person	462***	cancer
7		3*	Person	462***	cancer
8	3	≥ 40	Person	462***	flu
9		≥ 40	Person	462***	flu
10		≥ 40	Person	462***	flu

The K value serves as a privacy metric, representing the frequency of each combination of values within an equivalence class. In the given example, Table 2.5 is utilized to anonymize data through the application of suppression and generalization methods. When the value of k is set to 3, the data is considered 3-anonymous. This indicates that the tuples in the table have been generalized to the extent that there are at least three occurrences of every combination of data. The lower the K value, the lower the de-anonymization likelihood. In contrast, if the K value is greater, an attacker will have a harder time determining the identities of individuals. However, increasing the K value simultaneously reduces the data utility (Mendes and Vilela, 2017; Olatunji et al., 2022).

Although the k-anonymity approach provides some amount of privacy protection, it has a few drawbacks. First, it can be challenging for k-anonymity to identify the quality improvement qualities in external tables and determine how much data can be shared with others (Keyvanpour, 2011). Recent research has shown that 87% of the population can be identified using seemingly unimportant quasi-identifier (QI) attributes (Aldeen et al., 2015; Sweeny, 2002; Binjubeir et al., 2020). In earlier studies, researchers obtained mobility datasets of 1.5 million people using a k-anonymity approach (removing apparent identity attributes). They found that they could identify an individual with 95% accuracy using just four spatiotemporal points (De Montjoye et al., 2013; Yu, 2016). Another recent study (Yu, 2016; De Montjoye et al., 2015) that analyzed a data collection of over 1 million people's 90-day financial transactions corroborated the disadvantages of the simple k-anonymity approach. The study found that almost 90% of the subjects could be re-identified using four spatiotemporal points.

Second, Table 2.5, related to Table 2.1, illustrates three different anonymous versions of the sick individuals' database. The k-anonymity approach attempts to work on the QI attributes, such as determining a person's age, zip code, and gender, but does not invest in the sensitive attributes (SA) of the individual (Yu, 2016). As a result, the k-anonymity approach is susceptible to indirect attacks, which create the possibility of inferring a person’s attributes and disclosing their identity. Examples of this type of attack include the homogeneity attack (also known as the absence of variety in SA within an anonymized group; see the equivalence class 3 in Table 2.5) (Binjubeir et al., 2020), as well as the background knowledge attack, which is based on the following aspects: an adversary has enough background knowledge from the relationship between SAs and QI attributes to conduct probabilistic attacks (Yu, 2016; Li et al., 2014), or when the QI attributes are connected with other public data, this attack is known as a composition attack (Sweeney, 2002a; Bhaladhare and Jinwala, 2016; Yu, 2016; Keyvanpour, 2011). In addition, while trying to achieve a high level of privacy, it is impossible to avoid the loss of information when employing the k-anonymity approach (Mehmood et al., 2016). The k-anonymity approach might affect how data is used, which could cause an imprecise or even unworkable extraction of knowledge through data mining. As a result, balancing the need for privacy and the desire for utility is critical in data applications (BinJubier et al., 2022).

Machanavajjhala et al. (Machanavajjhala et al., 2007) designed an l-diversity approach to shield the identities of individuals from disclosure. It extends the k-anonymity approach. The major aim of the l-diversity approach is to protect one's privacy, which is achieved by expanding the spectrum of sensitive values. This approach treats specific attribute values the same regardless of how they are distributed in the data, resulting in an adequate representation of sensitive attributes inside each equivalence class, which guards against probabilistic inference attacks (Priyadarsini, Sivakumari, and Amudha, 2016).

Definition 1 (l-Diversity): A QI-group (or equivalence classes E) is said to satisfy l-diversity (or a QI-group is said to be l-diverse) if the probability that any tuple in this group is linked to a sensitive value is at most 1/l. The table satisfies l-diversity (or the table is said to be l-diverse) if each QI-group satisfies l-diversity. The value of l represents the level of diversity required, with higher values providing stronger privacy protection. For instance, consider Table 2.5, which is a 3-diverse table. It comprises three QI-groups. The first QI-group encompasses the first four tuples, the second QI-group consists of the fifth, sixth, and seventh tuples, and the last QI-group includes the final three tuples. In each QI-group, the probability of a tuple being associated with cancer is at most 0.3 (Pawar et al., 2018).

The distribution of values of sensitive attributes (SAs) presents a significant challenge for the I-diversity approach. This is due to the fact that various values have varied levels of sensitivity. In anonymized data, one sensitive attribute value may occur much more often than other values in an equivalence class (see equivalence class 1 in Table 2.5). This repeated occurrence of a value poses a severe privacy risk, as it allows an adversary to infer the likelihood of other entities within the equivalence class sharing the same value. This type of attack is known as a skewness attack (Pawar et al., 2018; Li et al., 2007). Because of the possibility of skewed attribute values, the construction of viable l-diverse representations is a challenging task. Additionally, this approach's ability to prevent the disclosure of an attribute through a similarity attack is insufficient (in an equivalence class, the sensitive attribute's values are different even though they are semantically related). I-diversity ensures that there is a variety of sensitive values present in each equivalence class, but it does not consider how closely together these values are semantically. This disadvantage served as the impetus for the invention of the T-closeness approach (Li et al., 2007; Priyadarsini et al., 2016; Pawar et al., 2018).

T-closeness is an extension of l-diversity-based anonymization, applied to protect the privacy of data sets. Li et al. (Li et al., 2007) proposed this approach to enhance data set privacy. In this method, the distribution of sensitive attributes in each equivalence class should resemble the distributions of those attributes in a general table; for example, the difference between the two distributions should not exceed the threshold value t (Li et al., 2014; Li ,2007).

LeFevre et al. first introduced the Mondrian approach (LeFevre et al., 2006). The Mondrian approach was initially discussed by LeFevre et al. in 2006 (LeFevre et al.). It is a greedy multidimensional technique that achieves k-anonymity across a data table's quasi-identifier (QI) attributes by recursively dividing the domain space into a number of regions, each containing at least k records. The approach starts with the entire database table as a partition, then divides it into two smaller partitions. These smaller partitions are then recursively divided into two parts until the k-anonymization principle is achieved. If so, the pre-cut partition functions as an equivalence class in the anonymized database. Using the range or set of QI attribute values present in the equivalence class itself, each equivalence class is generalized.

The Mondrian approach can be implemented using either strict or relaxed partitioning. When using strict partitioning, global recoding is the preferred method for dividing the database into non-overlapping sections. On the other hand, relaxed partitioning uses a local recording system and creates partitions that may overlap in the generalized quasi-identifier values, making it more flexible.

Overall, anonymization approaches based on grouping methods are simple and attempt to protect the privacy of individuals; however, they have an intrinsic drawback. They cannot effectively defend the sensitive values of the records from a composition attack (Li et al., 2016). Moreover, achieving optimal anonymization is an NP-Hard problem (Chambers, De Mesmay, and Ophelders, 2018; Binjubeir et al., 2020). Additionally, these techniques become ineffective in high dimensionality scenarios, as the identity of the primary record holders can be exposed by combining the data with background information or public data (Chambers et al., 2018; Hasan et al., 2018; Binjubeir et al., 2020; Andrew et al., 2019).

Given this, there is no one approach that is ideal for solving all privacy issues, as the type of privacy concern varies depending on the data used and how it is used. However, the disadvantages of one approach may be mitigated by another. This use is contextualized, and the position of the introduced approach in the literature is highlighted in Table 2.6 which lists prior works that have been discussed for privacy protection based on the grouping method.

Table 2.6 Summary of related works on preserving privacy based on grouping method.

Authors	approaches	Objectives	Method	Strength	Weakness
(Roberto J Bayardo and Agrawal, 2005) and (Samarati and Sweeney, 1998)	K-anonymization	Proposed for limiting disclosure of information and protect the privacy of persons	Modifying the values of the QI attributes enhances privacy by complicating the attacker's ability to determine individuals' identities. The K value serves as a metric for a measuring of privacy.	This approach protects an individual's identity while releasing sensitive information	K-anonymity faces indirect attacks, allowing attackers to deduce individual features, including homogeneity, background knowledge, and composition attacks. Additionally, its effectiveness diminishes in high-dimensional datasets.
(Machanavajjhala et al., 2007)	L-diversity	The primary objective of L-diversity is to uphold privacy by augmenting the diversity of sensitive values.	This approach strives to treat the values of a specific attribute uniformly, regardless of their distribution within the dataset.	This approach aims to maintain privacy by ensuring adequate representation of sensitive attributes within each E_e.	This approach is vulnerable to skewness attacks, similarity attacks, and composition attacks.
(Li et al. ,2007)	T-closeness	The primary objective of T-closeness is to uphold privacy, representing an extension of L-diversity.	The distribution of SA within any E_e should closely resemble the distribution of the attribute across the entire dataset.	This approach works to distribution SA in any E_e similar to the distribution of the attribute in an overall table which lead to preserve privacy	This approach is unable to protect the SA values of records consistently and efficiently against composition attacks.
(LeFevre et al., 2006)	Mondrian	Studying the problem of k-anonymization over the quasi- identifier attributes of a database table	Partitioning the domain space recursively into multiple regions, each containing a minimum of k records, Mondrian generalizes a set of quasi-identifier (QI) values within each equivalence class.	Getting an anonymous dataset	Local recoding closely resembles the original data and can retain more information, yet it may render the data susceptible to composition attacks. Moreover, overlapping intervals prove unsuitable for most classification tools as they introduce complexity to classification tasks.

2.4.2 Preserving Privacy Based on The Perturbation Method

The goal of perturbation is to protect sensitive information in a way that makes it challenging for an attacker to use attribute linkage attacks to identify a specific person in a published dataset or to infer a specific person's precise, sensitive value. It mainly adds uncertainty to published datasets and reduces the possibility of determining sensitive personal information (Wong and Fu, 2010; Zorarpacı and Özel, 2021). Among the most favorable methods of anonymity in perturbation is adding noise (randomization) to the data (Brand, 2002; Chawla et al. 2005; Shah and Gulati, 2016a; Li, Yan, and Zhang, 2014), creating synthetic data (Liew et al., 1985; Rubin, 1993; Domingo-Ferrer, 2002), and swapping attributes (Fienberg and McIntyre, 2004).

One of the most popular perturbation methods is randomization (adding noise) (Shah and Gulati, 2016a; Li et al., 2014; Mendes and Vilela, 2017; Zorarpacı and Özel, 2021). This method involves specific perturbation of the original data values by introducing or multiplying a randomized or stochastic number to conceal the distinct values of records. Consequently, adversaries cannot deduce the private attributes of a specific person by relating the attributes. Therefore, the perturbed data value of an individual can be significantly different from its original version, for instance, in a situation where a student's GPA is fraudulently increased from 3.45 to 3.65. Mivule (Mivule, 2013) was the first to publish work on additive noise under the general term $X + \beta$. The key notion is that instead of publishing X, the data owner releases the tuples produced from $X + \beta$, where X is the original data value and β is a random value selected from a particular distribution (Kim, 1986). The degree of privacy is determined by how well the original values of a modified attribute can be estimated (Agrawal and Aggarwal, 2001). Furthermore, experiments in (Charu and Philip, 2008; Du and Zhan, 2003; Evfimievski et al., 2002) show that some data can be maintained in randomized data mining operations. Fuller (Fuller, 1993) and Kim et al. (Kim et al.,1995) showed that the addition of random noise would not affect some simple statistical information, such as correlations and means.

Despite the simplicity and intuitive nature of the randomization method, it also has certain drawbacks, with privacy breaches being the most common (Mendes and Vilela, 2017; Binjubeir et al., 2020; Nasiri and Keyvanpour, 2020). Experimental studies (Kargupta et al., 2003; Huang et al., 2005; K. Chen and Liu, 2005; Luo and Wen, 2014;

JEBA et al., 2022) have demonstrated the limited effectiveness of randomization in preserving privacy. Private data recovery algorithms can reasonably reconstruct the original data from the perturbed data, particularly when there is a relationship or strong correlation among different attributes, which tends to be preserved even after randomization. The independent noise added to each attribute allows a private data recovery algorithm to exploit the spectral structure of the perturbed data using filtering methods, enabling accurate recovery of the original data. Furthermore, achieving optimal data privacy by adding noise significantly increases computational costs and results in the loss of some statistical data properties, rendering the dataset nearly unusable for users (Mivule, 2013). Therefore, it is essential to strike a balance between data privacy and utility (Binjubeir et al., 2020).

Data swapping was first presented by Fienberg and McIntyre (Fienberg and McIntyre, 2004) as a method for preserving data privacy, especially in datasets containing categorical attributes. The basis of this approach is to switch the original data into a distorted version that will still retain the same frequency count statistics as the original version by altering the data values of selected cells. Data swapping is useful in protecting both numerical (Reiss, Post, and Dalenius, 1982) and categorical attributes (Reiss, 1984).

Swapping allows the masking of information for all individuals, as it only needs to be performed on the sensitive attribute (SA) to break the relationship between the record and the individual, leaving the quasi-identifier (QI) attributes undisturbed. While swapping works well, it has the major disadvantage of not maintaining multidimensional relationships. Furthermore, swapping is expected to affect data mining operations (Matthews and Harel, 2011). It is also possible that swapping may cause illogical combinations, such as a record suggesting that there is a guy with ovarian cancer if the microdata database contains gender and type of cancer (Hasan et al., 2016; Murthy et al., 2019).

Rank swapping is an alternative to the swapping way (Benjamin C.M. Fung et al., 2010). The values of attribute a_i are first ranked in ascending order before swapping each of the ranked values with another randomly selected ranked value from a specified range. Rank swapping can maintain multivariate relationships more appropriately than ordinary data swapping (Matthews and Harel, 2011; Domingo-Ferrer and Torra, 2002). The main difference between rank swapping and ordinary data swapping is that the range over which the data can be swapped is restricted. The advantage here is that it limits the values that can be swapped with other values, while the difficulty lies in finding the cells for swapping that will maintain the multivariate relationships of interest (Liew et al., 1985; Lasko and Vinterbo, 2010).

Privacy in data publishing can be achieved using synthetic data (Chen et al., 2009). Synthetic data is used to produce data with similar distributional characteristics to the original information instead of altering the original dataset or using it as it is. The beauty of synthetic data stems from the fact that it comes from actual data and distributions, making it almost indistinguishable from the original data. Therefore, one of the key benefits of this approach is that an attacker cannot reveal private information by obtaining the published data. However, in practice, the identified data may lack sufficient utility (Heldal and Iancu, 2019). In addition, many statistical disclosure methods are used to generate synthetic data based on patterns found in the original dataset (Rubin, 1993). For example, condensation is used to represent synthetic data (Saita and Llirbat, 2004). The general idea is to first build a statistical model from the data by condensing the records into multiple groups based on their centers, radii, and sizes. Then, another set of data can be generated based on the statistical information.

In the last decade, various approaches for ensuring privacy in independent data publishing have been suggested, including the random rotation perturbation approach (Keke Chen and Ling Liu, 2005), random projection approach (Liu, Kargupta, and Ryan, 2006; Johnson and Lindenstrauss, 1984), probabilistic approach (Sattar et al., 2014), e-differential privacy approach $(e - DP)$ (Mohammed et al., 2011), hybrid approach (Li et al., 2016), and composition (Baig et al., 2012). The following paragrapgs describe these approaches in further detail.

Liu et al. (Keke Chen and Ling Liu, 2005) introduced a new protection method called rotating. It changes (rotates) the data in a specific way to protect private information in public datasets from composition attacks. One disadvantage of the data rotation approach is that domain-specific data attributes like Euclidean distance or the inner product are not preserved. This finding indicates that most existing modeling approaches are perturbation invariant while introducing distance inference assaults (Aggarwal and Yu, 2008; Patel, Dodiya, and Pate, 2013). Simultaneously, projection matrices have been used to anonymize mined datasets in the projection approach (Liu et al., 2006). It provides a number of random projection matrices that can be used to protect privacy from composition attacks in various data mining applications. However, identifying the actual data's approximation is possible (X. Li et al., 2014).

A probabilistic approach was designed by Sattar et al. (Sattar et al., 2014) (2014), which suggested a new approach called (d, α)-linkable. The probabilistic approach attempts to reduce the likelihood that an adversary can successfully complete a composition attack by guaranteeing that (d) sensitive values are linked with a QI-group with a likelihood of (α) by discovering the correlation between the QI and sensitive attributes.

A hybrid approach has been addressed by Li et al. (Li et al., 2016), and a composition approach by Baig et al. (Baig et al., 2012) to protect data privacy against composition attacks in many independent data publications. According to Hasan et al. (Hasan et al., 2018), composition is the first privacy approach to defend against composition attacks across multiple independent data publications. Two novel ideas were incorporated into the composition to defend against composition attacks: (ρ, α)-anonymization by sampling and composition-based generalization. Additionally, in a hybrid approach that combines sampling, generalization, and perturbation, Laplacian noise was added to the count of each sensitive attribute (SA) in each equivalence class. In these approaches, the values of the quasi-identifier (QI) are divided into equivalence classes, in which all values are identical. Consequently, members of an equivalence class are indistinguishable.

Mohammed (Mohammed et al., 2011) proposed the first generalization-based noninteractive approach, known as $e - DP$. The suggested solution creates a generalized contingency table probabilistically before introducing noise to the counts. The $e - DP$ offers a robust privacy guarantee for statistical query answering in addition to protection against composition attacks through differential privacy-based data anonymization (Ganta et al., 2008; Zorarpacı and Özel, 2021). According to (Li et al., 2016; Cormode et al., 2013; Sarathy and Muralidhar, 2011; Hasan et al., 2018), when using $e - DP$ to defend against composition attacks, a significant amount of data utility is lost during anonymization.

Table 2.7 illustrates the summary of related works on preserving privacy based on the perturbation method for multiple independent data publishing. It has been pointed out that the measurement of privacy preservation level and information loss is usually carried out through data perturbation methods (Aggarwal and Yu, 2008). The two critical concepts that should be mentioned here are privacy preservation and information loss. The privacy preservation level refers to the difficulty in estimating original data from perturbed data (Keke Chen and Ling Liu, 2005). On the other hand, information loss is a situation in which a significant portion of the information of the original dataset is lost after perturbation.

Table 2.7 Summary of related works on preserving privacy based on the perturbation method.

Authors	Approaches	Objectives	Method	Strength	Weakness
(Li et al., 2016)	a hybrid	Protecting data privacy from composition attacks and preserve data utility	They proposed a hybrid approach that integrates sampling, generalization, and perturbation, achieved by incorporating Laplacian noise into the count of every sensitive value within each equivalence class.	The proposed work reduces the risk of composition attacks and preserves data utility	There is still more data loss
(Baig et al., 2012)	composition	Protecting data privacy from composition attacks and preserve data utility	The proposed approach integrated two novel concepts: (ρ, α)- anonymization by sampling and composition-based generalization for independent datasets to protect against composition attacks.	The composition approach effectively protects privacy and preserves data utility.	There is still more data loss
(Sattar et al., 2014)	probabilistic	Reducing the likelihood an adversary can successfully complete a composition attack	They were used the probabilistic approximation to achieve the privacy principle	They were used the probabilistic approximation to achieve the privacy principle	This approach lacks improvement in several areas, including diversity, overcome of fake tuples and so on.

Table 2.7 Continued

Authors	Approaches	Objectives	Method	Strength	Weakness
(Mohammed et al., 2011)	$e - DP$	Studying the problem of anonymization for the non-interactive setting based on the generalization technique	First probabilistically generates a generalized contingency table and then adds noise to the counts	$e - DP$ provides a strong privacy guarantee for statistical query answering and protection against a composition attack by differential privacy-based data anonymization	It generates a significant amount of data utility losses during anonymization.
(Keke Chen and Ling Liu, 2005)	Random rotation	Classification of data with multiple dimensions (attributes) for privacy preserving data	Random rotation perturbation works for changing (rotation) the data in a specific manner to protect private information in public data sets	Privacy is guaranteed as long as the data values of the published data relatively differ from the data values of the original data.	The major drawback of the random rotation perturbation is that the domain-specific properties of data are not preserved.
(Liu et al., 2006)	Random projection	Random projection matrices have been utilized as tools for the preservation of the privacy of data sets	Random projection works to transfer a set (N) of the original data points from its initial high dimensional space to a lower-dimensional subspace (randomly selected).	This approach offers high-level privacy to the data	lose data and lower the risk ratio of disclosure

2.4.3 Preserving Privacy Based on Measures of Correlation (Similarity)

This method uses multiple correlated attributes (multiple dimensions) instead of a single-column distribution to achieve outstanding results for data mining operations. The data distributions are reorganized to preserve privacy based on the grouping method to conduct mining, which analyzes each dimension independently while ignoring the connections between various attributes (dimensions) (Keke Chen and Ling Liu, 2005). The perturbation method, used to protect privacy, modifies the real values of dataset D to create its anonymized version $D1$. The usability of the data is affected, or the data's distinctive attributes are lost depending on the type and amount of perturbation present (Mivule, 2013). Utilizing a correlation (similarity) metric to enhance protection and preserve more data utility is a brilliant solution to these issues (Hasan et al,. 2018; Han et al., 2012). The correlation measure aims to preserve the usefulness of the data by arranging highly correlated attributes in columns and keeping the correlations between those attributes. It also breaks the linkages between uncorrelated qualities in other columns, thereby protecting users' privacy. These associations are broken using protection methods based on anonymization approaches, such as randomly permuted data or generalization, and so on (Hasan et al., 2018; Li et al., 2012; Olatunji et al., 2022).

The strength of the association between two categorical attributes is measured using the mean square contingency coefficient (MSCC), a chi-square measure to assess the relationship's strength between two categorical attributes denoted by r. The value of this coefficient, r, falls between [0, 1]. When the value of r is greater than 0 but less than 1, the categorical attributes are related. If the value is 0, there is no correlation between the categorical attributes. A value of r = 1 indicates a perfect match between the categorical attributes (Li et al., 2012; Cramir 1946).

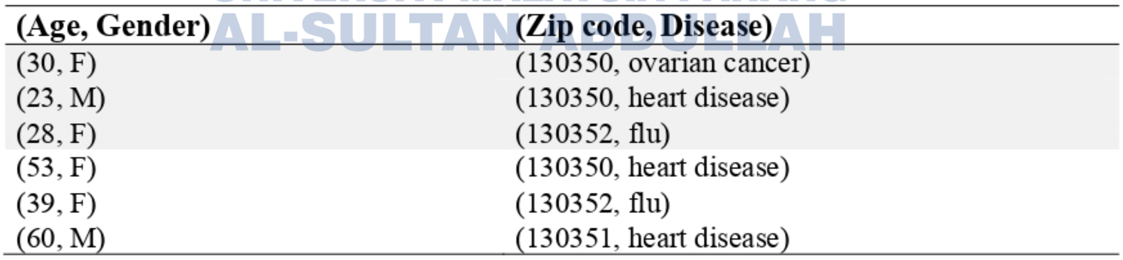
Assume attribute a_1 with value domain $\{v_{11,}\ v_{12,} \dots v_{1d1,}\}$, attribute a_2 with value domain $\{v_{21,}\ v_{22,} \dots v_{2d2,}\}$, and their domain sizes are d_1 and d_2, respectively. The MSCC between a_1 and a_2 is defined as follows (Hasan et al., 2018; Li et al., 2012):

$$\mathrm{r}^2(\mathrm{a}_1, \mathrm{a}_2) = \frac{1}{\min\{\mathrm{d}_1, \mathrm{d}_2\}} \sum_{\mathrm{i}=1}^{\mathrm{d}_1} \sum_{\mathrm{j}=1}^{\mathrm{d}_2} \frac{(\mathrm{f}_{\mathrm{ij}} - \mathrm{f}_{\mathrm{i}}.\mathrm{f}_{\mathrm{j}})^2}{\mathrm{f}_{\mathrm{i}}.\mathrm{f}_{\mathrm{j}}} \qquad 2.1$$

Where $r^2(a_1, a_2)$ is the MSCC between attributes $a_1$and a_2; $f_{i.}$ and $f_{.j}$ refer to the occurrence fractions of the v_{1i} and v_{2j} in the data, respectively; and f_{ij} is the fraction of cooccurrence of v_{1i} and v_{2j} in these data. Therefore, $f_{i.}$ and $f_{.j}$ are the marginal totals of f_{ij}: $f_{i.} = \sum_{j=1}^{d_2} f_{ij}$ and $f_{.j} = \sum_{i=1}^{d_1} f_{ij}$. $0 \leq r^2(a_1, a_2) \leq 1$.

The most recent correlation-based approaches for privacy-preserving data publishing are slicing (T. Li et al., 2012) and merging (A. Hasan et al., 2018). Slicing, in particular, has gained significant attention as a novel data anonymization approach. The authors propose a risk disclosure prevention concept that avoids generalization. In random slicing, the attribute values within a bucket are permuted to break column-wise relationships. This way effectively protects the privacy of published records by mitigating risks related to attribute and membership disclosure. Furthermore, slicing is particularly suitable for anonymizing high-dimensional data as it retains more data utility compared to attribute value generalization. Therefore, slicing ensures both data privacy and the preservation of data utility by avoiding attribute value generalization. It involves vertical partitioning (attribute grouping) and horizontal partitioning (tuple partitioning), and the resulting sliced table should undergo random permutation (see Table 2.8) (Li et al., 2012). For better understanding, slicing should be formalized, as Li et al. (2012) suggested.

Table 2.8 Published data by slicing



(Age, Gender)	(Zip code, Disease)
(30, F)	(130350, ovarian cancer)
(23, M)	(130350, heart disease)
(28, F)	(130352, flu)
(53, F)	(130350, heart disease)
(39, F)	(130352, flu)
(60, M)	(130351, heart disease)

2.4.3.1 Attribute Grouping

The microdata table T comprises of a set of t tuples, $t \in T$, and n number of a attributes, where t is a tuple of T and t is described as $t = (t[a]_1, t[a]_2 \ldots, t[a]_n)$, where $t[a]_i$ $1 \leq i \leq n$. In attribute grouping, the initial step involves separating the attributes into multiple individual attributes. Next, the related attributes are grouped together into subsets, ensuring that each attribute belongs to a distinct subset. Consequently, each subset is referred to as a cell, and the combination of these cells forms the columns. In microdata table T, there will be col columns $col_1, col_2 \dots, col_c$ satisfying $\cup_{i=1}^{c} col_i = a$ attribute and for any $1 \le i_1 \ne i_2 \le col$, $col_{i1} \cap col_{i2} = \emptyset$. When there is only one SA, the position is placed last for easy representation.

Definition 2 (cell): A cell refers to a pair of attributes, such as {(Age, Gender)}, where each cell $C_{col,E}$ is uniquely identified by the column number Col_i and the equivalence class number E_j. For instance, in Table 2.8, any cell within the column {(Age, Gender)} is identified by the values of Col_i and E_j, which $1 \le i \le col$ and $1 \le j \le E$ and the first equivalence class consists of tuples t_1= $\{t_1, t_2, t_3, t_4\}$.

2.4.3.2 Tuple Partition

The objective of tuple t partitioning is to create multiple subsets of Table T in such a way that each tuple can only belong to one of these subsets. Each subset of tuples is known as an equivalence class or bucket. Let's assume there are E_e equivalence classes, namely $E_1, E_2, \dots E_e$. In this case, the union of all these equivalence classes, from E_i to E_{e} (where $1 \le i \le e$), results in the original table T. It is important to note that for any pair of indices i_1 and i_2, where $1 \le i_1 \ne i_2 \le e$, the intersection between $E_{i=1}$ and $E_{i=2}$ is empty (denoted by $E_{i=1} \cap E_{i=2} = \emptyset$).

However, slicing can lead to data utility and privacy issues because it randomly permutes attribute values in each bucket, increasing the likelihood of erroneous tuples and negatively affecting the published microdata's utility. An attacker can use analysis of the bogus tuples to understand the notion of the implemented anonymization process, potentially violating the privacy of published data (Hasan et al., 2016; Binjubeir et al., 2020; Mendes and Vilela, 2017). For instance, the tuple t_1 for the zip code 130350 in Table 2.8 only has one matching equivalence class connected to two sensitive values. Here, l-diverse slicing can link any individual with SA values with a probability of no greater than 1/l because it has been established that slicing satisfies l-diverse slicing by being connected to the SA values by 1/2. Assume that the zip code attribute is exposed because it has a significant number of QI values (sufficient variety), and that an attacker who depends on background information is aware of this (23, M). An attacker can then define the person's sensitive attribute. Additionally, as stated in (Hasan et al., 2016), incompatibility between the QI and SA values (erroneous tuples) is may be produced if the slicing process varies the SA value (randomly) between t_1 and t_2.

Hasan et al. (Hasan et al., 2018) developed the merging approach to secure personal identification from disclosure. This been regarded as an extension of slicing. Merging's primary purpose is to preserve privacy in many separate data releases by employing cell generalisation and random attribute value permutation to seperate connection between various columns. Regarding privacy risks and data utility, the merging approach conserved data usefulness while posing minor privacy hazards because of increased false matches in the released datasets. Nonetheless, the merging approach's significant weaknesses are the randomised permutation way for the attribute values to breach the relationship between the columns and the increase in false matches for unique attributes. However, there will be a large number of matching buckets (more than the initial tuples), resulting in utility data loss, and could generate inaccurate and infeasible knowledge acquisition from data mining operations (Sharma, Singh, and Rehman, 2020;Jeba et al., 2022). As a result, the main reasons for revealing people's identities are unique attributes or the ability of some cells in the tuple to match with cells in other tuples in the same equivalence class, allowing precise extraction of a person's attributes (Li et al., 2012; Hasan et al., 2018; Binjubeir et al., 2020).

Other studies have demonstrated the rational of allowing a tuple to match more than one bucket when preventing attribute and membership disclosure (Gkoulalas-Divanis and Loukides, 2015; Li et al., 2012). When records for a single person are mapped to multiple buckets, a super bucket is created from the collection of buckets.

Table 2.9 illustrates the previous works discussed in the earlier sections for preserving privacy based on measures of correlation (similarity) between different attributes. It has been pointed out that the measurement of correlation (similarity) plays a significant role in improving the protection and keeping more data utility (Li et al,. 2012).

Table 2.9 Summary of related works on preserving privacy based on the measurement of correlation (similarity) method

Authors	Approaches	Objectives	Method	Strength	Weakness
(Li et al., 2012)	Slicing	Studying the problem of anonymization for high-dimensional data.	This approach uses vertical partitioning (attribute grouping), horizontal partitioning (tuple partition), and its sliced table should be randomly permutated.	Slicing approach provides data privacy by randomly permutated of data and preserves data utilities that is devoid of generalization.	Random permutate for attribute values are led to creating invalid tuples which will negatively affect the utility of the published microdata.
(Hasan et al., 2018)	merging	Studying the problem of anonymization for the multiple independent data publishing against composition attack.	The primary aim of merging approach is to preserve privacy by increasing the false matches in the published datasets and it uses vertical partitioning, horizontal partitioning, and its sliced table should be randomly permutated.	Getting an anonymous dataset preserves data utilities.	The major drawback of merging is the random permutation procedure and increasing the false matches in the published datasets.

2.4.4 Critical Analysis

As highlighted in the literature review, a significant challenge arises when organizations publish data that is crucial for enhancing their efficiency and aiding their future targets. Despite the anonymization approaches uses differant protection methods such as suppression and generalization, randomization, and/or combined and causing uncertainty in identity inference or sensitive value estimation (Lasko and Vinterbo, 2010). However, these approaches can still lead to the disclosure of sensitive attributes (SA) through linking attacks, where the remaining attributes (quasi-identifiers) are linked with other data sources, referred to as composition or intersection attacks (Majeed and Hwang, 2023).

Moreover, many of the existing anonymization approaches struggle to strike an optimal balance between data utility and privacy when releasing data products. Achieving a satisfactory level of privacy preservation while retaining valuable information for data mining remains a challenge (Majeed and Hwang, 2023). The evaluation criterion for assessing the effectiveness of an anonymization approach lies in its capability to protect data privacy by reducing the exposing individuals' data and ensuring that the published data remains usable (Majeed and Lee 2021; Siddique et al., 2018; Hasan et al., 2018; BinJubier et al., 2022).

The focus of this research is to get a specified level of privacy with minimum information loss for the intended data mining operations. Thus, the focus of the study is to reduce the risk of a composition attack when multiple organisations independently release anonymised data and, meanwhile, released data remain as useful as possible. Table 2.10 provides a summary of the main challenges and gaps motivating this study in the domain of anonymization approaches' performance.

Table 2.10 Critical analysis of related works on existing anonymization approaches

Approaches	**Protection methods**	**Determine the Level of Protection**	**Prevent the Disclosure of Information**	**Prevent Composition Attack**	**Data Utility Preservation**
Mondrian	The proposed approach involves recursively partitioning the domain space into multiple regions, ensuring that each region contains a minimum of k records. In each equivalence class, a set of QI values is generalized	✗	✓	✗	✓
Hybrid	A hybrid approach has been proposed that integrates sampling, generalization, and perturbation ways. It involves adding Laplacian noise to the count of each sensitive value within every equivalence class. This combination aims to enhance privacy protection.	✗	✓	✓	✓
$e - DP$	In order to provide protection, the initial step involves the probabilistic generation of a generalized contingency table, which is subsequently followed by the addition of noise to the counts.	✗	✓	✓	✗
Probabilistic	The probabilistic approach aims to lower the risk of a successful composition attack by ensuring that sensitive values are linked to a QI-group with a specific likelihood (α). This is achieved by identifying correlations between the QI and sensitive attributes.	✗	✓	✓	✓
Composition	The proposed approach combines two novel concepts, (ρ, α)-anonymization through sampling and composition-based generalization, to provide protection against composition attacks for independent datasets.	✗	✓	✓	✓
Slicing	This method incorporates vertical partitioning, horizontal partitioning, and random permutation of the sliced table. It involves attribute grouping through vertical partitioning, tuple partitioning through horizontal partitioning, and a crucial step of random permutation to enhance privacy protection	✗	✓	✓	✓
Merging	The primary goal of the merging approach is to ensure privacy by employing both vertical and horizontal partitioning of data. It aims to enhance privacy protection by increasing the occurrence of false matches for unique attributes. Additionally, it is essential to randomly permute the sliced table as part of the process.	✗	✓	✓	✓
UL	The use of lower protection level (LPL) and upper protection level (UPL) can be instrumental in determining the required level of data protection. Rank swapping is employed to protect unique and highly identical attributes, ensuring the privacy protection of data while preserving data utility.	✓	✓	✓	✓

2.5 Data Utility and Measuring Risks

The balance between data utility and privacy is often viewed as a trade-off, and it is influenced by the protection methods employed in anonymization approaches and their prioritization of these two aspects (BinJubier et al., 2022). This study describes maintaining privacy as minimizing disclosing of information on individuals. Data utility, on the other hand, refers to what extent we can use the sterile database for intensive analyses. For instance, by suppressing each Quasi-identifier (QI), a dataset can be generalized, providing maximum privacy but rendering the information obtained useless. Finding a good balance between privacy and utility is necessary because the published datasets (sanitized) must permit tasks related to data mining operations for search and analysis. As a result, the usefulness of data in published datasets is assessed by how well statistical and aggregate data are used.

The ability to protect data privacy by lowering the risk of disclosing personal information while maintaining the potential use of published data is the criterion for judging the effectiveness of the anonymization approach (Majeed and Lee 2021; Siddique et al., 2018; Hasan et al., 2018; BinJubier et al., 2022). There are two ways to evaluate the data utility, or the degree of risk disclosure, in published (sanitized) datasets. The first approach involves utilizing one of the quantified measures of information loss that have been evaluated, and the second involves using data as input into a query and assessing the accuracy of the results. Each of these measures is described in the following subsections. Readers can refer to a more thorough survey (Benjamin C M Fung et al., 2010; Anjum, 2013).

2.5.1 Measurement of risk disclosure

This section covers the measurement of disclosure risk in microdata during a composition attack. A composition attack occurs when an intrusive party, especially one knowledgeable about some of the Quasi-identifier (QI) values, attempts to identify a specific person in the microdata by linking several readily accessible records to an external database to disclose restricted information (Chen et al., 2009b). Therefore, the measurement of disclosure risk involves assessing the rarity of a cell in microdata publishing. Previous studies have proposed quantifying the risk disclosure using a certainty penalty (CP). The CP is determined by the ratio of the number of real matches, where the QI values in the original dataset are compared with the QI values in the anonymized dataset. If a record in the anonymized dataset matches the QI values in the original dataset, it is considered a disclosure of QI values. The CP is calculated as the total number of disclosed QI values (or number of real matches) divided by the total number of matches, as shown in Equation 2.2 (Hasan et al., 2018; BinJubier et al., 2022):

$$Disclosure\ risk\ ratio\ (DRR)\ = \frac{Matched\ records}{Total\ records} \times 100\% \qquad 2.2$$

2.5.2 Measurement of data quality

Data quality is evaluated based on the distortion ratio (DR). There are several methods for calculating the DR in published data (Wong and Fu, 2010) to determine how much anonymization affects data distortion. According to previous works, a suitable measure for calculating the DR is the generalized distortion ratio (GDR) (Rohilla, 2015). The swap or generalize methods are used as a protection method to break the association of the attributes. When a node p in the taxonomy tree t of two categorical attributes ($a_1^{**}, a_2^{**} \in T^{**}$) is used to swap or generalize the attributes, the DR with p is defined (BinJubier et al., 2022).

$$DR(a_1^{**}, a_2^{**}) = \begin{cases} 0, & a_1^{**} = a_2^{**} \\ \frac{|Common(a_1^{**}, a_2^{**})|}{|N|}, & a_1^{**} \neq a_2^{**} \end{cases} \qquad 2.3$$

where $|Common(a_1^{**}, a_2^{**})|$ represents the set of leaf nodes in the lowest common tree of a_1^{**} and a_2^{**} in t and $|N|$ represents the set of all leaf nodes in t.

Figure 2.5 depicts a domain of generalization hierarchies of attributes for marital status (MS). If a_i^{**} and a_j^{**} values fall into the same rank group and don't contain nonsensical combinations, their swap values are equal, and the DR is 0. In the absence of this, their generalised values are equal to $\frac{|Common(a_1^{**}, a_2^{**})|}{|N|}$, and the DR is equal to $\sum_{j=1,k=1}^{n,m} d_{j,k}$, where $d_{j,k}$ is the distortion of the attribute a_j^{**} of tuple t_k.

The distortion ratio (DR), also known as data utility, is a proportional measure that compares the amount of distortion in a generalised dataset to the amount of distortion in a fully generalised dataset. It is possible to determine the value of the data by subtracting the DR from Equation 2.4 shown below (Wong and Fu, 2010):

$$Data\ utility = \ (100 - DR)\% \qquad 2.4$$

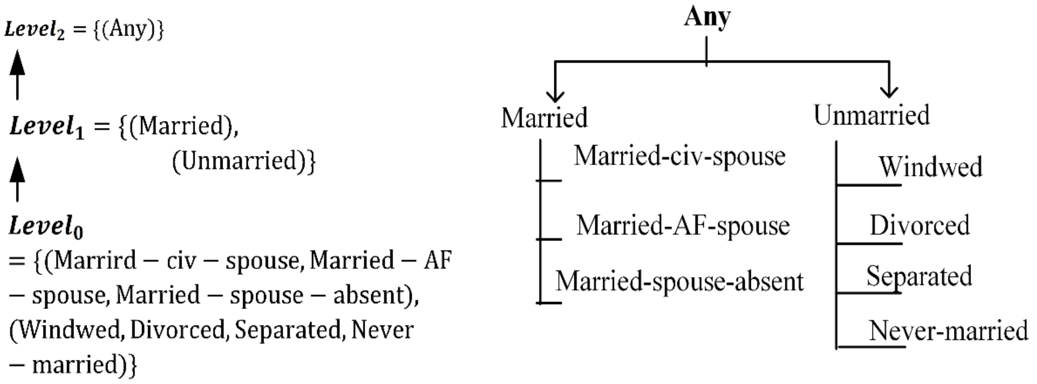

$\boldsymbol{Level_2} = \{(\text{Any})\}$
$\boldsymbol{Level_1} = \{(\text{Married}), (\text{Unmarried})\}$
$\boldsymbol{Level_0} = \{(\text{Marrird} - \text{civ} - \text{spouse}, \text{Married} - \text{AF} - \text{spouse}, \text{Married} - \text{spouse} - \text{absent}), (\text{Windwed}, \text{Divorced}, \text{Separated}, \text{Never} - \text{married})\}$
Any
Married
Unmarried
Married-civ-spouse
Married-AF-spouse
Married-spouse-absent
Windwed
Divorced
Separated
Never-married

Figure 2.5 Example of a domain of generalisation hierarchies of attributes for marital status (MS)

2.5.3 Query Workload

An aggregate query is used to measure the usefulness of data in published datasets. An arithmetic operation known as an aggregate query takes a set of values and produces a single value that expresses the importance of the data. It is common practice to provide the base numbers that represent the expected data utility to test the efficacy of the suggested approach using the aggregate query operators "COUNT," "MAX," and "AVERAGE." (Zhang et al., 2007; Jayapradha et al., 2022). A query predicate is characterized by two parameters: the dimension of the predicate d and the selectivity of the query sel. The predicate dimension d signifies the number of Quasi-identifiers (QIs) present in the predicate, while the selectivity sel indicates the number of values in each QI_{ij}, where $1 \ \leq \ j \ \leq \ d$. For instance, when responding to aggregate queries with a query predicate containing QI attributes, the "COUNT" operator is typically considered. Let T be a table containing QI quasi-identifiers, QI_1 ∷, QI_i, with $d(\mathrm{QI}_i)$ being the domain of quasi-identifier. Next, the questions are formatted as follows:

$$\textbf{SELECT COUNT}(*)\ \text{from}\ \ \mathbf{T}$$

$$\boldsymbol{Where}\ \mathrm{QI}_{i1} \in \ \mathrm{d}(\mathrm{QI}_{i1)})\ \boldsymbol{and}, \mathrm{QI}_{idim} \in \ \mathrm{d}(\mathrm{QI}_{idim})\ \boldsymbol{and}, \mathrm{s} \in \ d_s$$

Here, QI_{ij}, $1 \leq j \leq d$ represents the quasi-identifier (QI) value for attribute a_{ij}, where $\mathrm{QI}_{ij} \subseteq d(\mathrm{QI}_{ij})$ and $d(\mathrm{QI}_{ij})$ is the domain for attribute a_{ij}. s is the sensitive attribute (SA) value, where $s \subseteq d_s$ and d_s is the domain for the SA. To determine the size of each QI_{ij} ($1 \leq \mathrm{j} \leq \mathrm{dim}$), a random selection is made from the range of 0, 1, ..., sel * $|d(\mathrm{QI}_{ij})|$, where $|d(\mathrm{QI}_{ij})|$ represents the cardinality of the domain for attribute a_{ij} (Q. Zhang et al., 2007; BinJubier et al., 2022).

Each query is executed on both the original table and the anonymized table. The count obtained from the original table is referred to as $\mathrm{actual}_{\mathrm{coun}}$, while the count obtained from the anonymized table is referred to as $\mathrm{sanitized}_{\mathrm{count}}$. The difference between the result sets obtained from evaluating the query Q on raw data and sanitized data, respectively, is the normalized error for the query Q, denoted as $Error\ (Q)$. The formula used to calculate the query error is as follows (Zhang et al., 2007):

$$error(Q) = \frac{sanitized_{count} - actual_{coun}}{actual_{coun}} \qquad 2.5$$

2.6 Summary

This chapter defines the context of the thesis by providing a thorough background on the essential concepts, methodologies, and methods used in the research, along with citations to the most important publications in the literature. The presentation moves from broad, overarching ideas to more focused ones.

First, the definition and types of privacy are discussed, setting the foundation for the subsequent topics. Data publishing is the cornerstone problem in this thesis; therefore, Privacy-Preserving Data Publishing (PPDP) and the approaches used are introduced. Given that data anonymization is the main focus of this thesis, the anonymization approach is presented, where two important issues are discussed: protection methods used, and data utility and measuring risks.

Finally, several anonymization approaches are also classified based on the protection methods used, such as preserving privacy based on grouping, perturbation, and measurement correlation (similarity) methods. These approaches are analyzed in terms of their weaknesses and strengths.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The flow of the research procedure is described in this chapter, outlining the stages and methods used to accomplish the research objectives outlined in Chapter 1. It also establishes the relationship between the research questions, research objectives, and their connection to the chapters that follow. As noted in the literature review, numerous approaches have been proposed to address privacy issues, aiming to protect sensitive information from uninvited disclosure while preserving the utility of the data. However, there is still room for improving user privacy while maintaining data usefulness.

In essence, this research focuses on designing an enhanced slicing-based approach called the Upper Lower (UL) level-based protection approach, which can be used for data anonymization of published data, leading to a better balance of utility and privacy before releasing any data product. Additionally, this work introduces an improved protection method called the Lower Protection Level (LPL) and Upper Protection Level (UPL) for anonymization, which is more effective in determining the required amount of protection. UPL and LPL involve selecting specific cell values that help identify disclosure and break the link between them, enhancing the privacy of the published data while ensuring additional data utility and guaranteeing the achievement of 1-diversity in the published data.

The focus of this research is to attain a definitive degree of privacy while ensuring minimal information loss during data mining. This study aims to minimize the vulnerability of configuration assault in the event of releasing anonymous data by various independent organizations, while maintaining the integrity and functionality of the released data. This chapter is designed as follows: Section 3.2 presents the complete flow of the procedures. Section 3.3 describes the three stages of the design of the UL approach. Section 3.4 explains the methods used to evaluate the proposed approach. Section 3.5

describes the hardware and software used in this research, and the chapter is summarized in Section 3.6.

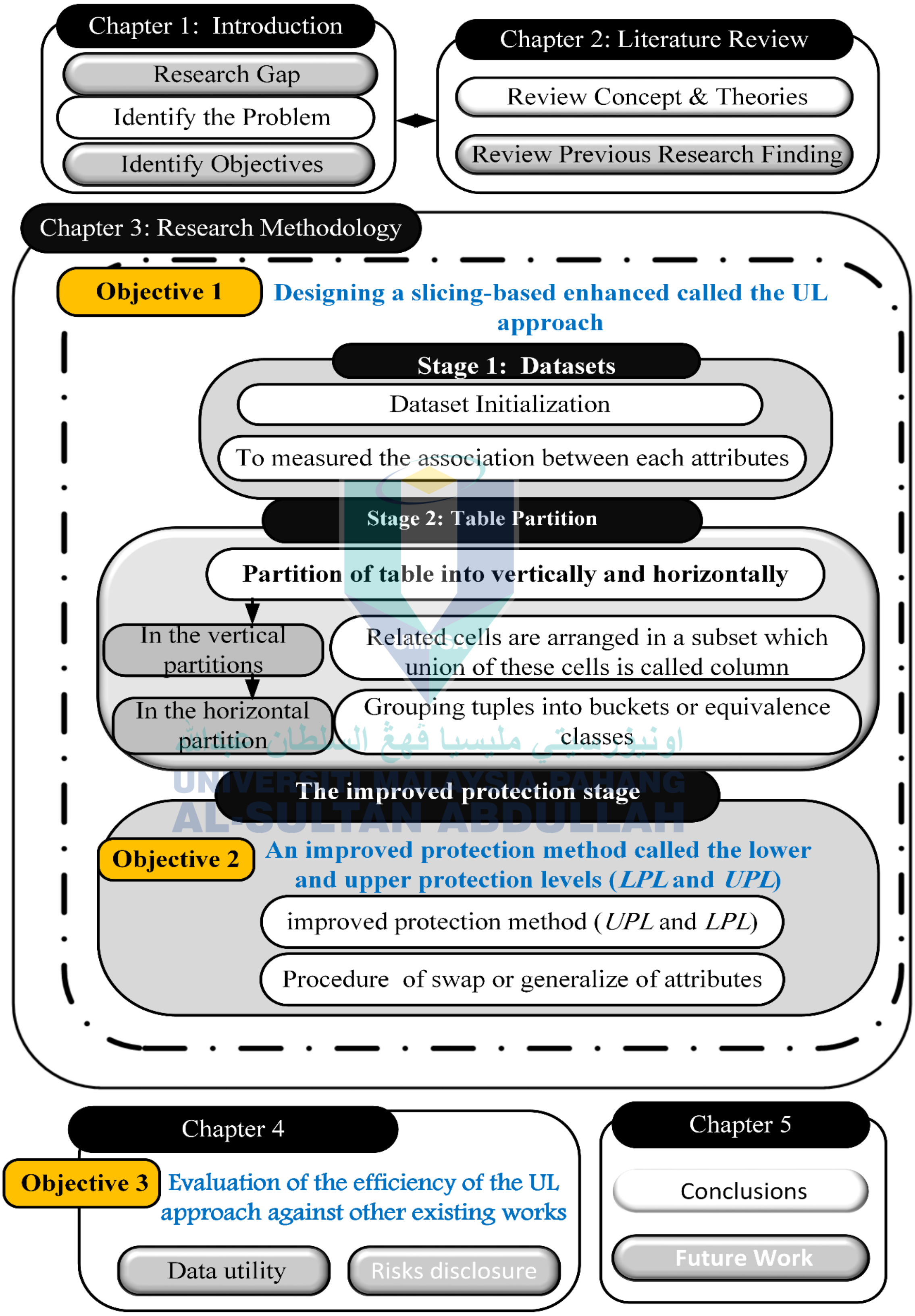

Chapter 1: Introduction
Research Gap
Identify the Problem
Identify Objectives
Chapter 2: Literature Review
Review Concept & Theories
Review Previous Research Finding
Chapter 3: Research Methodology
Objective 1
Designing a slicing-based enhanced called the UL approach
Stage 1: Datasets
Dataset Initialization
To measured the association between each attributes
Stage 2: Table Partition
Partition of table into vertically and horizontally
In the vertical partitions
Related cells are arranged in a subset which union of these cells is called column
In the horizontal partition
Grouping tuples into buckets or equivalence classes
The improved protection stage
Objective 2
An improved protection method called the lower and upper protection levels (LPL and UPL)
improved protection method (UPL and LPL)
Procedure of swap or generalize of attributes
Chapter 4
Objective 3
Evaluation of the efficiency of the UL approach against other existing works
Data utility
Risks disclosure
Chapter 5
Conclusions
Future Work

Figure 3.1 Flow of research procedures

3.2 Flow of research procedures

This section depicts the sequence of research stages taken to achieve the objectives. Figure 3.1 illustrates the overall research flow. The figure shows that the third and fourth chapters are devoted to the study's main contributions. The following are the steps in the research.

3.3 Design of UL approach

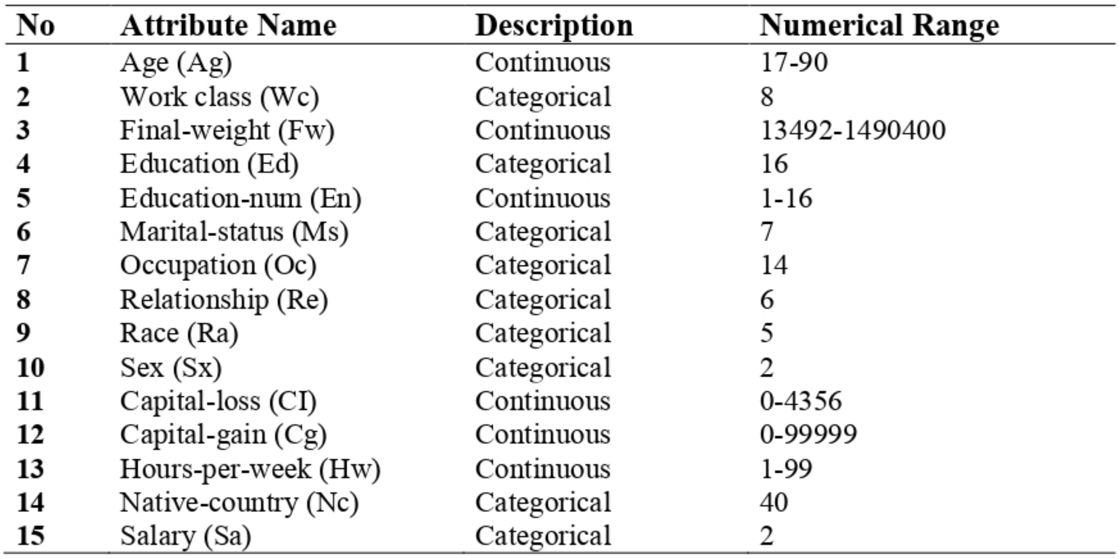
This research focuses on designing a slicing-based enhanced approach called the Upper Lower (UL) level-based protection approach that can be used for data anonymization for published data, leading to a better balance of utility and privacy before releasing any data product. Additionally, this work introduces an improved protection method called the Lower Protection Level (LPL) and Upper Protection Level (UPL) for anonymization, which is more effective in determining the amount of protection required. UPL and LPL select particular cell values that help to identify disclosure and break the link between them to enhance the privacy of the data that has been published, ensure additional utilization of data, as well as guarantee that 1-diverse is achieved in the published data (Jeba et al., 2022; Cunha et al., 2021).

Figure 3.1 illustrates the proposed UL approach to data protection, simultaneously preserving the utility of the data. This approach consists of three stages that protect the published data from unauthorized disclosure. The following is the discussion of these three stages.

3.3.1 Dataset Preparation stage

The 'Preparing the dataset' stage aims to initialize the dataset and measure the correlation between attributes. To evaluate the experiments and compare with other existing works, the 'Adult' dataset has been applied and used for the experiments. Ronny Kohavi, together with Barry Becker, extracted and assembled this dataset from the 1994 United States Census Bureau (Kohavi and Becker, 2019). Accordingly, this dataset is made up of fifteen Quasi-identifier (QI) attributes with 48,842 tuples, as depicted in Table 3.1. The classifiers for each attribute describe the type of attribute (Continuous, Categorical) and the numerical range. For example, the 'sex' attribute possesses a numerical range equal to 2 (male, female).

Table 3.1 Description of US-Census Adult dataset



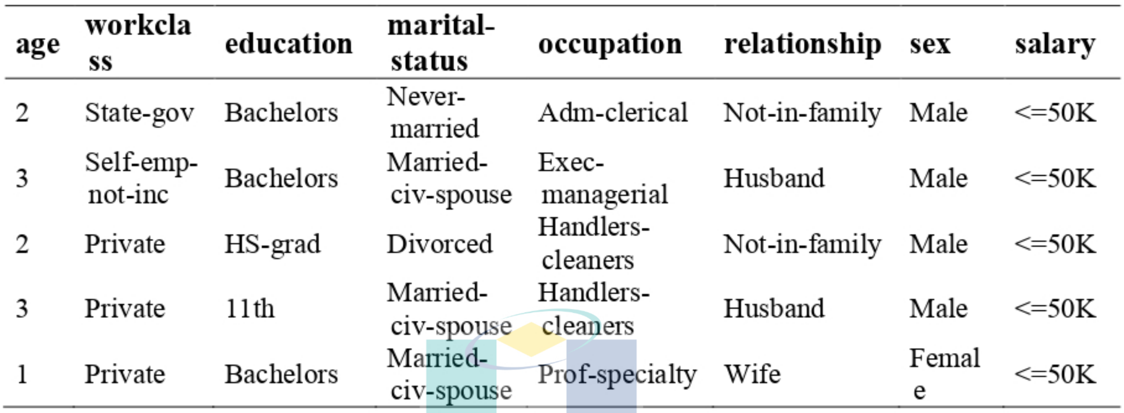
No	Attribute Name	Description	Numerical Range
1	Age (Ag)	Continuous	17-90
2	Work class (Wc)	Categorical	8
3	Final-weight (Fw)	Continuous	13492-1490400
4	Education (Ed)	Categorical	16
5	Education-num (En)	Continuous	1-16
6	Marital-status (Ms)	Categorical	7
7	Occupation (Oc)	Categorical	14
8	Relationship (Re)	Categorical	6
9	Race (Ra)	Categorical	5
10	Sex (Sx)	Categorical	2
11	Capital-loss (Cl)	Continuous	0-4356
12	Capital-gain (Cg)	Continuous	0-99999
13	Hours-per-week (Hw)	Continuous	1-99
14	Native-country (Nc)	Categorical	40
15	Salary (Sa)	Categorical	2

The Adult dataset, which includes real data, has been applied (Kohavi and Becker, 2019). In the dataset initialization process, independent datasets were required to simulate the actual scenario of independent data publishing, particularly in cases where such datasets are separately published by various organizations that have similar records. However, the pitfall of this proposition lies in the fact that an individual's data is often published by many organizations (Malin and Sweeney, 2004)· Under such conditions, any intruder can initiate a composition assault (Ganta et al., 2008; Hasan et al., 2018) on such published datasets just to alter the privacy of the dataset. More information about the case of a single publication approach was previously provided in Section 1.1 (Background).

In the dataset initialization process, five disjoint datasets of different sizes were pooled from the Adult dataset and extracted into two independent datasets called the Education and Occupation datasets. The Education dataset contains eight Quasi-identifier (QI) attribute values: work class (categorical, 8), relationship (categorical, 6), gender (categorical, 2), age (continuous, 74), marital status (categorical, 7), education (categorical, 16), and salary (categorical, 2). The Occupation dataset also contains eight QI attribute values: work class (categorical, 8), relationship (categorical, 6), gender (categorical, 2), age (continuous, 74), marital status (categorical, 7), occupation (categorical, 14), and salary (categorical, 2). The parenthesis figures indicate the attribute type and the number of classifiers for each attribute. Table 3.2 shows a sample from the 'Adult' dataset having eight QI attribute values.

In terms of attribute classification, the Education and Occupation attributes are categorical and have a vast numerical range. Consequently, the Education attribute is assumed to be a sensitive attribute (SA) in the Education dataset, while the Occupation attribute is considered an SA in the Occupation dataset. The remaining attributes are regarded as QI attributes.

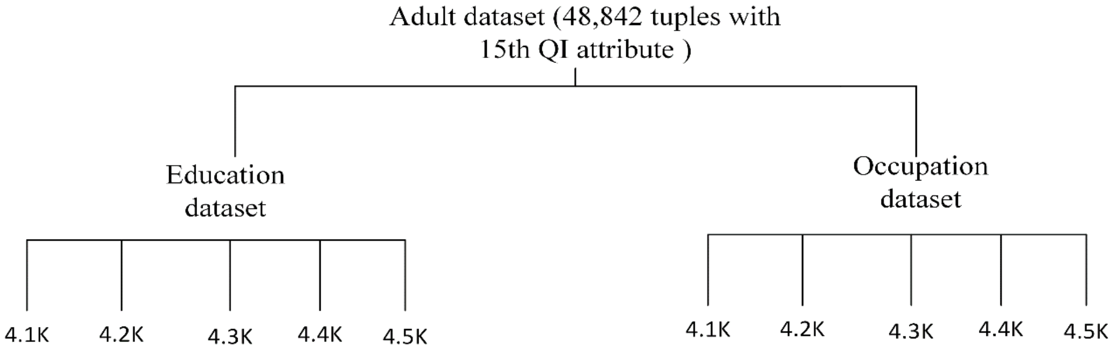
Table 3.2 Sample from "Adult" dataset having eight QI attribute values



age	workclass	education	marital-status	occupation	relationship	sex	salary
2	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	Male	<=50K
3	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	<=50K
2	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	Male	<=50K
3	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Male	<=50K
1	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Female	<=50K

The Education and Occupation datasets were extracted from the Adult dataset, which contains a total of 48,842 tuples. Randomly selecting 4000 tuples for each category, separate datasets were created. The remaining tuples were utilized to generate an overlapping tuple, which was also used to detect composition attacks.

For both the Education and Occupation categories, five duplicates were created. The overlapping tuple was constructed by gradually adding tuples in increments of 100, 200, 300, 400, and 500 to both the Occupation and Education datasets. As a result, the dataset sizes for Education and Occupation are 4.1K, 4.2K, 4.3K, 4.4K, and 4.5K (where K=1000) datasets, respectively (see Figure 3.2: An illustration of the dataset Table 3.3).



- The Education and Occupation datasets each contain 4K randomly selected tuples that do not overlap,
- The remaining 8K tuples from the Adult dataset were utilized to create an overlapping tuple pool and assess the presence of composition attacks.

Figure 3.2 An illustration of the dataset

Table 3.3 Five independent datasets of various sizes for the simulation of the actual independent data publishing scenario from Occupation and Education dataset.

No	Dataset size	# Non-Overlapping tuples	Overlapping tuples
1	4.1K	4000	100
2	4.2K	4000	200
3	4.3K	4000	300
4	4.4K	4000	400
5	4.5K	4000	500

After the initialization process, the correlation between attributes is measured. The dataset initialization generated datasets with sizes of (4.1K), (4.2K), (4.3K), (4.4K), and (4.5K) for the Education and Occupation datasets to simulate the actual independent data publishing scenario. Each dataset is treated as a microdata table T. In the case where the microdata table T has a_i attributes, where $i = 1,2, \ldots n$. the strength of the correlations between pairs of attributes can be computed using several methods (Hasan et al., 2018; Li et al., 2012; Cramir 1946). Because most attributes are categorical, the most suitable method for estimating the correlations between pairs of attributes is the mean square contingency coefficient (MSCC). The MSCC is a chi-square-based measure of the correlation between two categorical features. The value of this coefficient r ranges from [0, 1], and it is symmetric, as presented in Table 3.4 3.4. If there is a perfect relationship between the two attributes, the measure of the association will have a value of 1. Otherwise, these measures differ in their maximum value. In the case of no relationship between the two attributes, the measure of association has a value of 0. The MSCC between a_1 and a_2 is defined as follows (Hasan et al., 2018; Li et al., 2012) and as explained in Section 2.4.3:

$$r^2(a_1, a_2) = \frac{1}{min\{d_1, d_2\}} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{\left(f_{ij} - f_i . f_j\right)^2}{f_i . f_j}, \qquad 3.1$$

Table 3.4 Sample of computing the correlations between pairs of attributes

a_1 / a_2	age	workclass	education	marital-status	occupation	relationship	sex	salary
age	1	0.144699	0.145919	0.411203	0.109966	0.253803	0.043474	0.115424
workclass	0.144699	1	0.106591	0.135567	0.277626	0.113452	0.061987	0.079298
education	0.145919	0.106591	1	0.134942	0.401093	0.129015	0.043844	0.143852
marital-status	0.411203	0.135567	0.134942	1	0.209917	0.523499	0.194715	0.179016
occupation	0.109966	0.277626	0.401093	0.209917	1	0.201785	0.18627	0.134539
relationship	0.253803	0.113452	0.129015	0.523499	0.201785	1	0.288118	0.179101
sex	0.043474	0.061987	0.043844	0.194715	0.18627	0.288118	1	0.073902
salary	0.115424	0.079298	0.143852	0.179016	0.134539	0.179101	0.073902	1

The Age attribute can have an infinite number of continuous values. Therefore, this study applies discretization to partition the domain of the continuous age attribute into intervals and manage the cluster of interval values as a discrete domain. The equal-width discretization method is applied to fractionate the attribute domain into (some k) equal-sized intervals. For example, in Figure 3.3, local recoding generalizes values of the age attribute into four intervals: [17–37], [38–58], [59–79], [80-100].

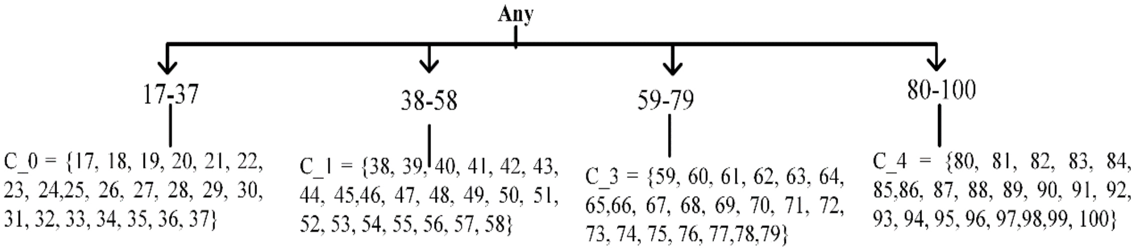

Any
17-37
38-58
59-79
80-100
C_0 = {17, 18, 19, 20, 21, 22, 23, 24,25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37}
C_1 = {38, 39, 40, 41, 42, 43, 44, 45,46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58}
C_3 = {59, 60, 61, 62, 63, 64, 65,66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77,78,79}
C_4 = {80, 81, 82, 83, 84, 85,86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,98,99, 100}

Figure 3.3 The equal-width discretization of age attribute

3.3.2 Table Partition (vertical and horizontal) Stage

After completing stage 1, stage 2 focuses on the vertical and horizontal partitionization of the table. The dataset in the table is divided both vertically and horizontally based on the correlation computation between pairs of attributes, which is known as r. This phase aims to categorize similar attributes based on the extent of their inter-attribute connections, considering both utility and privacy aspects.

Regarding data utility, closely connected attributes are grouped together to ensure that their inter-attribute relationships are preserved. However, from a privacy perspective, categorizing unrelated attributes poses a higher risk of vulnerability compared to categorizing highly connected attribute values, as it increases the likelihood of attribute identification. To enhance privacy protection, it is crucial to eliminate the connections among unrelated attributes (Li et al., 2012; BinJubier et al., 2022).

The microdata table T comprises a set of t tuples, where t belongs to T, and n number of a attributes, where each tuple t is represented as $t = (t[a]_1, t[a]_2 \dots, t[a]_n)$, with $t[a]_i$ denoting the value of attribute a for tuple t, where $1 \leq i \leq n$.

In the vertical partition, the attributes are initially separated into multiple groups. Then, similar attributes are further grouped together in subsets based on their correlation, with each attribute belonging to a specific subset. Each subset consisting of a pair of attributes is referred to as a cell, and the combination of these cells forms the columns. In microdata table T, there will be col columns $col_1, col_2 \dots, col_c$ satisfying $\cup_{i=1}^{c} col_i = a$ attributes where $1 \leq i_1 \neq i_2 \leq col$, $col_{i1} \cap col_{i2} = \emptyset$.

In addition, the quasi-identifiers (QIs), and sensitive attributes (SAs) are organized in columns $col_i, 1 \leq i \leq n$. These attributes are clustered in n columns denoted as col_n, regardless of the size of the sensitive column col^c. In certain cases, the number of attributes a_s in the sensitive column col^c may be predetermined as c.

The size of the sensitive column col^c is determined by a parameter c, which can be mathematically represented as $|col^c| = c$. If $c = 1$, then $col^c = 1$, indicating that the sensitive column contains only one attribute, denoted as $col^c = \{S\}$. In the case where $c = 2$, the process is referred to as bucketization. When $c > 1$, $|col^c| > 1$, indicating that there are multiple attributes in the sensitive column. For the purpose of this study, the sensitive attribute a_s is focused on as a single attribute.

Assuming that the sensitive attribute is located in the last column, denoted as col^c, this column is referred to as the sensitive column in Table 3.5. If there are multiple sensitive attributes in the data, their individual or collective distributions can be used (Machanavajjhala et al., 2006). In the vertical partitioning, highly related attributes (cells) are grouped together in columns, while unrelated attributes are placed in separate columns. Each individual attribute a_i is assigned to a specific subset. As shown in Table 3.5, the columns Col_i $\{col_{1,}\ col_{2,} \ldots col_{n,}\}$ contain all the attributes a_i.

Table 3.5 presents the three partitions for the columns Col_i based on the correlation calculation (r) between each pair of attributes:

1- T^* consists of columns that contain highly correlated attributes, denoted as col^*. In other words, col^* is a subset of columns in T^* where $col^* = \{col_{1,}^*,$ $col_2^*, \ldots col_i^*\}$, and col^* belongs to T^*.

2- T^{**} comprises all columns that do not have correlated attributes, represented as col^{**}. Hence, $col^{**} = \{col_{1,}^{**},\ col_2^{**}, \ldots col_i^{**}\}$.

3- T^c consists of columns that include the sensitive attribute col^c. In cases where there is a sensitive attribute present, it is positioned in the final column. It is important to note that col^c belongs to T^c, and the union of$(T^* \cup T^{**}) \cup T^c$ equals T.

Table 3.5 Example of partitions in table T

T^* contains all columns with highly correlated attributes		**T^{**} contains all columns with uncorrelated attributes**		**T^c contains a column with sensitive attributes**
$col_{1,}^*$	col_2^*	col_1^{**}	col_2^{**}	col^c
(a_1,a_2)	(a_3,a_4)	(a_5,a_6)	(a_7,a_8)	(a_s)

The K-medoid clustering algorithm, also known as the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990), is utilized to organize similar attributes into columns, ensuring that each attribute is assigned to a specific column. In Figure 3.4, this algorithm represents each attribute as a point in the cluster space, guaranteeing the resolution of every attribute. The inter-attribute disparity in the clustered space is quantified by the formula $\mathrm{d}(a_1, a_2) = 1 - r^2(a_1, a_2)$, where $r^2(a_1, a_2)$ measures the association between attributes a_1 and a_2. The resulting value ranges from 0 to 1. When two attributes are strongly correlated, their affiliated data points exhibit reduced disparity within the clustered space. For example, in Table 3.4, the measure of association $r^2(sex, \text{salary})$ is calculated as $(0.073902)^2$, resulting in 0.005461505604.

After evaluating the disparity between related data points, the k-medoid method organizes related attributes into subsets called cells, and the combination of these cells forms the columns (T^*, T^{**}, and T^c), as shown in Table 3.6 and Table 3.7 presents the categorization of the related attributes obtained from the correlation of the inter-attribute evaluation for the Educational dataset, while Table 3.8 and Table 3.9 are based on the Occupational dataset. The selection of the k-medoid method is motivated by the following reasons (Hasan et al., 2018):

i. A considerable number of available algorithms, such as the k-means calls for computation of the 'centroids.' However, the idea of 'centroids' is alien to this setting due to the fact that every attribute generates a data point in the clustering space.

ii. The k-medoid technique is considerably dynamic to outliers (data points that are relatively distant from the remaining of the data points). The clusters estimated using the k-medoid technique are not altered by the sequence of data points assessment.

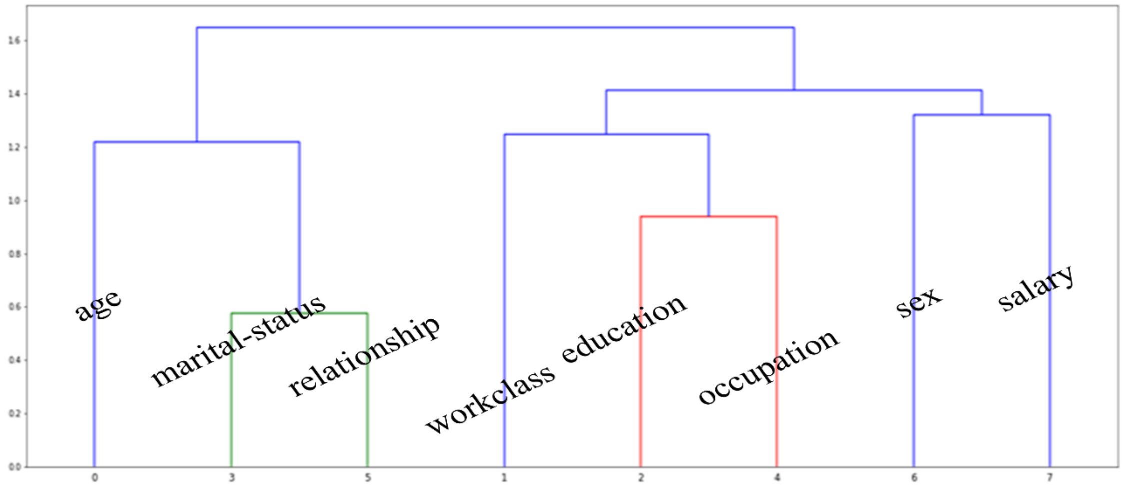


Figure 3.4 The result of the k-medoid clustering algorithm PAM.

Table 3.6 Partitions of table T into three partitions based on k-medoid algorithm PAM for Education dataset size of (4.1K, 4.2K and 4.4K)

$\boldsymbol{T}^*$ contains all columns with highly correlated attributes	$\boldsymbol{T}^{**}$ contains all columns with uncorrelated attributes	$\boldsymbol{T^c}$ contains column with sensitive attributes
(sex, salary) (age, workclass)	(marital-status, relationship) (occupation,	education)

Table 3.7 Partitions of table T into three partitions based on k-medoid algorithm PAM for Education dataset size of (4.3K and 4.5K)

$\boldsymbol{T}^*$ contains all columns with highly correlated attributes	$\boldsymbol{T}^{**}$ contains all columns with uncorrelated attributes	$\boldsymbol{T^c}$ contains column with sensitive attributes
(sex, salary) (age)	(marital-status, relationship) (workclass, occupation)	(education)

Table 3.8 Partitions of table T into three partitions based on k-medoid algorithm PAM for Occupation dataset size of (4.1K, 4.2K and 4.4K)

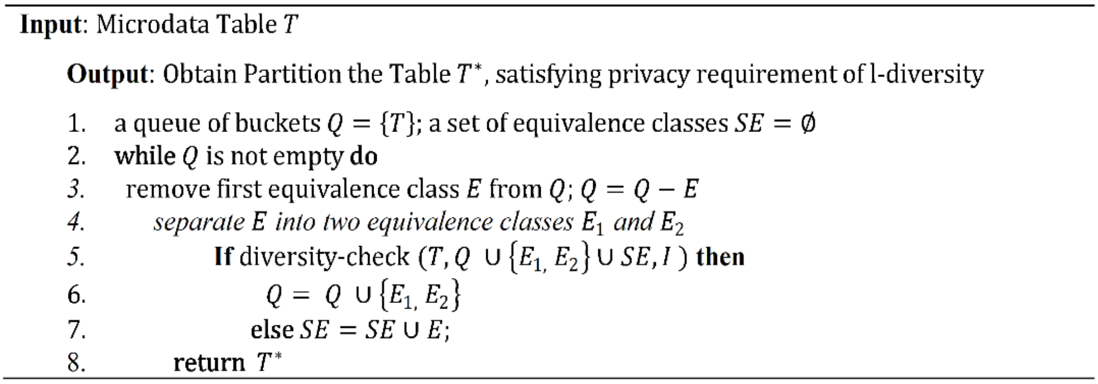
$\boldsymbol{T}^*$ conains all columns with highly correlated attributes	$\boldsymbol{T}^{**}$ contains all columns with uncorrelated attributes	$\boldsymbol{T^c}$ contains column with sensitive attributes
(sex, salary) (age, workclass)	(marital-status, relationship) (education,	occupation)

Table 3.9 Partitions of table T into three partitions based on k-medoid algorithm PAM for Occupation dataset size of (4.3K and 4.5K)

$\boldsymbol{T}^*$ contains all columns with highly correlated attributes	$\boldsymbol{T}^{**}$ contains all columns with uncorrelated attributes	$\boldsymbol{T^c}$ contains column with sensitive attributes
(sex, salary) (age)	(marital-status, relationship) (workclass, education)	(occupation)

In horizontal Partition, the table is divided into different subsets so that each tuple can only be assigned to a single subset. Every subset of these tuples is referred to as a bucket or an equivalence class. Assume there are Eequivalence classes, $E_1, E_2, \ldots E_e$ then, $\cup_{i=1}^{e} E_i = T$ for any $1 \le\ i_1\ \ne\ i_2 \le e,\ E_{i1} \cap\ E_{i2} = \emptyset.$

To achieve this partitioning, the equal-width discretization technique is applied to the age attribute domain, dividing it into k equal-sized intervals. Tuples with similar values are then categorized into buckets or equivalence classes. In this process, each individual is associated with a specific sensitive value, ensuring that an attacker cannot deduce the sensitive attribute values of an individual with a probability greater than 1/l, where l represents the number of possible sensitive values. The tuples were categorized using the Mondrian algorithm (LeFevre et al., 2006). They are separated in the equivalence classes, in the absence of generalization attributes, according to the top-down technique. Subsection 2.4.1 presented the description of tuples grouped into buckets, as seen in Figure 3.5.



```
Input: Microdata Table $T$
    Output: Obtain Partition the Table $T^*$, satisfying privacy requirement of l-diversity
    1. a queue of buckets $Q = \{T\}$; a set of equivalence classes $SE = \emptyset$
    2. while $Q$ is not empty do
    3.   remove first equivalence class $E$ from $Q$; $Q = Q - E$
    4.     separate $E$ into two equivalence classes $E_1$ and $E_2$
    5.          If diversity-check $(T, Q \cup \{E_{1,} E_2\} \cup SE, l)$ then
    6.               $Q = Q \cup \{E_{1,} E_2\}$
    7.             else $SE = SE \cup E$;
    8.       return $T^*$
```

Figure 3.5 The horizontal partition algorithm of tuples into buckets

Figure 3.5 describes the horizontal partition algorithm of tuples into buckets or equivalence classes, in which two data structures are preserved: the queue of equivalence classes and a collection of anonymized equivalence classes. Initially, the queue contains a single empty equivalence class (line 1). During each iteration (lines 2 to 7), an equivalence class E is dequeued from the queue Q and processed using the table partition horizontal algorithm. This algorithm applies the top-down technique based on the Mondrian criterion (LeFevre et al., 2006) to split the equivalence class into two. Line 5 utilizes an I-diversity checking algorithm, as shown in Figure 3.7, to ensure diversity within the equivalence class. Similarly, the two equivalence classes are attached at the bottom of the queue (line 6) to achieve a further breakdown of the equivalence class. The equivalence class is sent to line 7 by the table partition algorithm when it becomes unbreakable. The anonymized table reaches the publication stage as soon as it reaches line 8.

3.3.3 The improved protection stage

In the previous stage, the microdata table$T$ was partitioned both vertically and horizontally, organizing all attributes, including Quasi-identifiers (QIs) and Sensitive Attributes (SAs), into separate columns. This clustering serves as the foundation for the implementation of the third stage, known as the improved protection method. The primary objective is to prevent the unauthorized disclosure of individuals' identities by altering QI values in a way that conceals any potential linkages between individual values and specific attributes. This approach ensures that published data remains useful and informative.

This stage involves two steps, namely the implementation of an improved protection method using the Lower Protection Level (LPL) and Upper Protection Level (UPL), along with attribute swapping or generalization. These steps help identify specific attributes that could potentially reveal personal information and determine the appropriate level of protection required to prevent the disclosure of private data. By doing so, the UL approach effectively mitigates identity disclosure while striking a balance between preserving privacy and maintaining data utility. The goal is to safeguard personal information while still allowing the data to be useful for various purposes.

In Step 1 of the improved protection method, the strength of the relationship between attributes is primarily assessed using the correlation coefficient (r).

illustrates how UPL and LPL improve protection by identifying two types of cell values: (1) the values of unique cells and (2) highly identical values in T^{**}. LPL determines cells that possess unique values within the range of $0.0 < LPL \leq \Phi$. For such attributes, the r value typically hovers around 0 but is not exactly 0. Similarly, UPL identifies cells that have numerous similar attributes with values in the range of $\Phi \leq UPL < 1.0$. In the case of these attributes, the r value usually hovers around 1 but is not exactly 1. If such cells have a high r value in T^{**}, it indicates that these cells are likely to belong to the same equivalence class, which is referred to as matching bucket. This poses a privacy risk as an adversarial party can gain more certainty about the SA when these cells are linked to other cells in T^{*}, resulting in privacy violations. The remaining cells, which have association values distant from 1 and 0, are distinguished by their diversity, which effectively prevents attribute disclosure. It is crucial that these cells exhibit a diversity value of at least two (diversity ≥ 2), to fulfill the desired privacy objective.

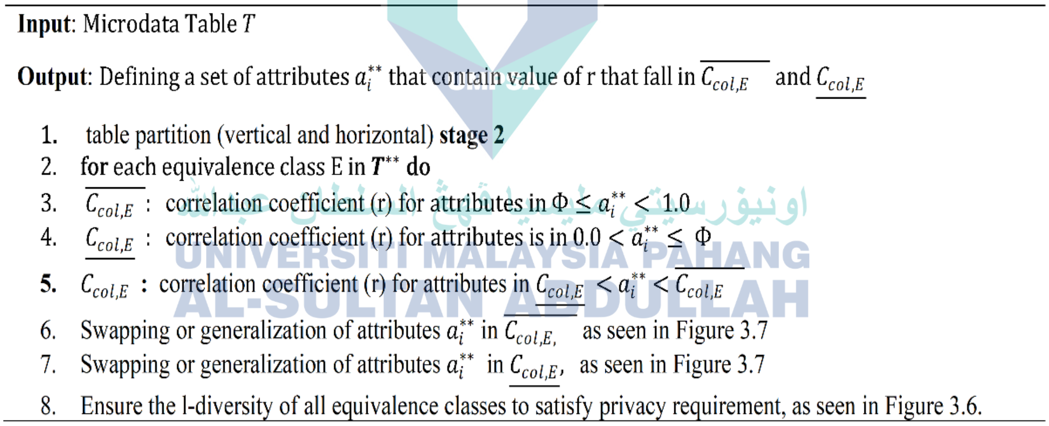
The goal of UPL and LPL is to discover the collection of unique cells and highly identical values for cells from table T^{**}, which are assumed to be known to intruders: $\overline{C_{col,E}} = \Phi \leq UPL < 1.0$, and $\underline{C_{col,E}} = 0.0 < LPL \leq \Phi$.

The attributes that are selected for swapping during this period are referred to as the swapping attributes. The number of cells that fall within this period is denoted by

$\overline{|C_{col,E}|}$ and $\underline{|C_{col,E}|}$. The values initially marked for swapping are represented by the swap rate, denoted by Φ. Usually, Φ ranges from 1% to 10%, indicating that the fraction of attributes actually swapped will be less than one.

Definition 2 (Matching Buckets): Assuming col_i^{**} represents the columns, and $col_i^{**} = \{col_1^{**}, col_2^{**}, \dots col_n^{**}\}$, where $i = 1,2,\dots n$, and and $col^{**} \in T^{**}$. Let t^{**} represent a tuple, and $t^{**}|col_i^{**}|$ represent the value of col_i^{**} in tuple t^{**}. Then, let t^{**} represent an equivalence class in the microdata table T^{**}, and $E^{**}|col_i^{**}|$ denote the multiset of values from col_i^{**} in equivalence class E^{**}. E^{**} denotes equivalence class of t^{**} if for all $1 \leq\ i \leq col^{**}$, $t^{**}|col_i^{**}| \in E^{**}|col_i^{**}|$.

Definition 3 (the Lower Protection Level (LPL) and Upper Protection Level (UPL)): LPL and UPL are correlation coefficient (r) values assigned to each cell $C_{col,E}^{**}$ in column col_i^{**}, where (LPL and $UPL \in r$).



Input: Microdata Table T

Output: Defining a set of attributes a_i^{**} that contain value of r that fall in $\overline{C_{col,E}}$ and $\underline{C_{col,E}}$

```
1. table partition (vertical and horizontal) stage 2
2. for each equivalence class E in $T^{**}$ do
3. $\overline{C_{col,E}}$ : correlation coefficient (r) for attributes in $\Phi \leq a_i^{**} < 1.0$
4. $\underline{C_{col,E}}$ : correlation coefficient (r) for attributes is in $0.0 < a_i^{**} \leq \Phi$
5. $C_{col,E}$ : correlation coefficient (r) for attributes in $\underline{C_{col,E}} < a_i^{**} < \overline{C_{col,E}}$
6. Swapping or generalization of attributes $a_i^{**}$ in $\overline{C_{col,E,}}$ as seen in Figure 3.7
7. Swapping or generalization of attributes $a_i^{**}$ in $\underline{C_{col,E}}$, as seen in Figure 3.7
8. Ensure the l-diversity of all equivalence classes to satisfy privacy requirement, as seen in Figure 3.6.
```

Figure 3.6 Protection improvement algorithm of UPL and LPL

When calculating the correlation coefficient (r) for partition Table T^{**} of attributes, the values of attribute a_i^{**} are classified into three categories, as shown in Figure 3.6.

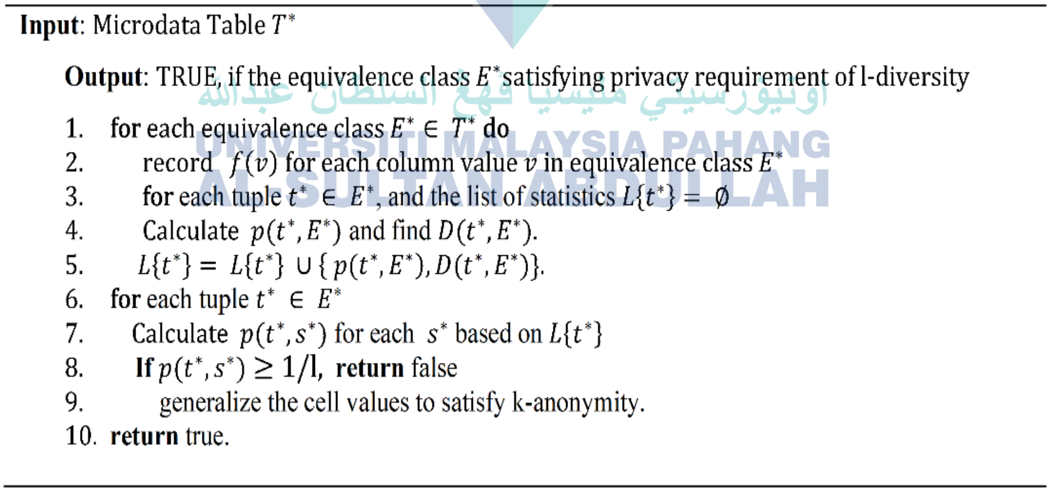
$\overline{C_{col,E}}$ includes highly identical attribute values with r within the range of $\Phi \leq a_i^{**} <\ 1.0$, as depicted in line 3.

1- $\underline{C_{col,E}}$ comprises unique attribute values with r within the range of $0.0 < a_i^{**} \leq \Phi$, as shown in line 4.

2- $C_{col,E}$ consists of the remaining cells with values of a_i^{**} that have a distant association from $\overline{C_{col,E}}$and $\underline{C_{col,E}}$, falling within the range $\underline{C_{col,E}} < a_i^{**} < \overline{C_{col,E}}$, as seen in line 5. The presence of diverse cells within equivalence classes is a crucial attribute of $C_{col,E}$ that plays a significant role in safeguarding privacy.

Line 8 serves as a check for the l-diversity privacy requirement, as illustrated in Figure 3.6.

However, it is expected that all cells have a diversity value of ≥ 2 in every equivalence class. Furthermore, attribute swapping or generalization for $\overline{C_{col,E}}$and $\underline{C_{col,E}}$ (see lines 6 and 7) enhances the veracity of information during decision-making processes. Veracity refers to the reliability of data and signifies the significance of relying on such data for data mining operations (Hasan et al., 2018).



```
Input: Microdata Table $T^*$
    Output: TRUE, if the equivalence class $E^*$satisfying privacy requirement of l-diversity
    1.  for each equivalence class $E^* \in T^*$ do
    2.      record $f(v)$ for each column value $v$ in equivalence class $E^*$
    3.      for each tuple $t^* \in E^*$, and the list of statistics $L\{t^*\} = \emptyset$
    4.      Calculate $p(t^*, E^*)$ and find $D(t^*, E^*)$.
    5.      $L\{t^*\} = L\{t^*\} \cup \{p(t^*, E^*), D(t^*, E^*)\}$.
    6.  for each tuple $t^* \in E^*$
    7.     Calculate $p(t^*, s^*)$ for each $s^*$ based on $L\{t^*\}$
    8.     If $p(t^*, s^*) \geq 1/l$, return false
    9.         generalize the cell values to satisfy k-anonymity.
    10. return true.
```

Figure 3.7 The diversity-check algorithm

The diversity-check algorithm, as illustrated in Figure 3.7, presents a comprehensive elucidation of the l-diversity checking process. For each tuple, denoted as t^*, the algorithm maintains a list named $L\{t^*\}$, which serves as a repository for statistical information pertaining to t^*'s corresponding equivalence classes or buckets. Each entry in the $L\{t^*\}$ list encompasses pertinent statistics associated with a specific matching equivalence class, represented as E^*. These statistics encompass the matching probability $p(t^*, E^*)$ and the distribution of candidate sensitive values $D(t^*, E^*)$.

The l-diversity checking algorithm commences by iteratively scanning each equivalence class denoted as E^* (lines 1 to 2) to collect the frequency $f(v)$ of occurrence for each column value v within the particular equivalence class. Subsequently, the algorithm proceeds to scan individual tuples, specifically tuple t^* within the aforementioned equivalence class E^*(lines 3 to 4) to identify all other tuples that match with t^* within E^*. During this process, the algorithm also records the matching probability denoted as $p(t^*, E^*)$ and the distribution of candidate sensitive values denoted as $D(t^*, E^*)$. These recorded values are then added to a list denoted as $L\{t^*\}$ (line 5). By the completion of line 5, the algorithm constructs the list $L\{t^*\}$ for each tuple t^*, containing crucial information about its corresponding matching equivalence class.

In the final step, a comprehensive scan is performed on the tuples within E^* to compute the $p(t^*, s^*)$ values, based on the information stored in $L\{t^*\}$. This calculation entails considering all tuples t^* associated with each sensitive value s^*. The algorithm employs these computed probabilities to establish I-diversity. Specifically, the algorithm assesses whether the condition $p(t^*, s^*) \leq 1/$l holds true for every sensitive value s^* (as outlined in lines 6 to 10). If this condition is satisfied for all sensitive values, signifying that the matching probabilities are less than or equal to 1/l, the anonymization Table T^* is considered to possess I-diversity. This criterion ensures that each sensitive value is adequately distributed across the tuples, thus promoting diversity in the anonymized dataset.

To illustrate, let's consider the anonymized table presented in Table 3.10, which introduces the concept of satisfying 2-diversity. This table consists of two equivalence classes, denoted as E^*. The first equivalence class, E_1^*, contains the first four tuples, while the second equivalence class, E_2^*, contains the last four tuples. Let's focus on a specific tuple, t_1, with the quasi-identifier (QI) values (32, F, 130352). To determine the sensitive value associated with t_1, we need to examine its matching equivalence classes.

The first step is to analyze the anonymized table's first column, which represents Age and Gender. By observing the table, we can deduce that t_1 must belong to the first equivalence class, E_1^*since there are no matches of (32, F) in the second equivalence class,

E_2^*. Consequently, we can conclude that $t_1$cannot be in the second equivalence class, E_2^*, and it must reside in the first equivalence class, E_1^*. Upon examining the zip code attribute in the second column (zip code, Disease) within the first equivalence class, E_1^*, we can determine that the column value for $t_1$must be either (130352, heart disease) or (130352, flu). It's important to note that the other two column values in the same zip code, 130352, are present. Without additional information, both heart disease and flu are equally likely to be the sensitive value associated with t_1. Consequently, the probability of correctly identifying the sensitive value for t_1 is limited to 0.5, as there is an equal chance of it being either heart disease or flu. By following a similar approach, we can verify that 2-diversity is upheld for all other tuples in Table 3.10. Let $p(t^*, E^*)$ represent the probability of tuple t_1 belonging to equivalence class E^*. For instance, in this particular example, $p(t_1^*, E_1^*)$ is equal to 1, indicating that t_1 is guaranteed to be in equivalence class E_1^*, while $p(t_1^*, E_2^*)$ is equal to 0, signifying that t_1 cannot be found in equivalence class E_2^*. A tuple t_1 can belong to multiple equivalence classes, and its overall matching degree across the entire dataset is denoted as $f(t)$ and calculated as the summation of $f(t^*, E^*)$ over all equivalence classes E^*. The probability that t_1 is a member of a specific equivalence class E^* can be expressed as follows:

$$p(t^*, E^*) = \frac{f(t^*, E^*)}{f(t^*)} \qquad 3.2$$

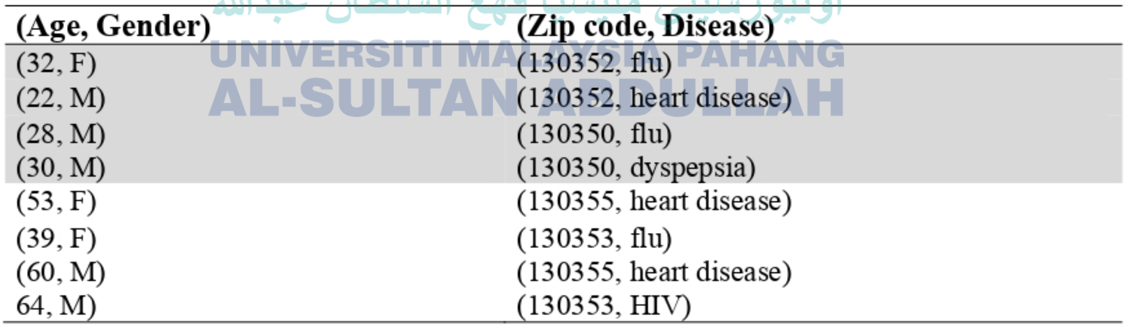
In the second step of the algorithm, the probability that a target tuple, denoted as t^*, takes a sensitive value s^*, denoted as $p(t^*, s^*)$, is computed using the law of total probability (T. Li et al., 2012). This is achieved by first calculating $p(s^*|t^*, E^*)$, which represents the probability that t^* takes sensitive value s^* given that t^* is in equivalence class E^*. The law of total probability is then employed to find the overall probability of t^* taking the sensitive value s^* as follows:

$$p(t^*, s^*) = \sum_E p(t^*, E^*) p(s^*|t^*, E^*) \qquad 3.3$$

To compute the probability $p(t^*, E^*)$, the algorithm considers the fraction of column values in tuple t^* that match the corresponding column values in equivalence class E^*. For instance, when examining Table 3.10, the column value (32, F) is one of the column values present in tuple t^*. If any column value in t^* does not apear in the corresponding column of E^*, it can be definitively concluded that t^ does not belong to E^*. Generally, equivalence class E^* may potentially match $|E^*|$ tuples, where $|E^*|$ represents the number of tuples in E^*. In the absence of additional information, the algorithm assumes independence among column values, making each of the $|E^*|$ tuples equally likely to be an original tuple. As a result, the probability of t^* being in E^* is determined by the fraction of the $|E^*|$ tuples that match t^*.

To formalize this analysis, the algorithm examines the matching between the column values of t^*, represented as $\{t^*[col_1], t^*[col_2] \dots t^*[col_c]\}$, and the column values of E^* denoted as $\{E^*[col_1], E^*[col_2] \dots E^*[col_c]\}$. To quantify the matching, the algorithm introduces $f_i(t^*, E^*)$ $(1 \leq i \leq c-1)$ as the fraction of occurrences of $t^*[col_i]$ in $E^*[col_i]$, and $f_c(t^*, E^*)$ as the fraction of occurrences of $t^*[col_c - \{s\}]$ in $E^*[col_c - \{s\}]$, where $col_c - \{s\}$ represents the set of quasi-identifier (QI) attributes in the sensitive column. For example, in Table 3.10 we have $f_1(t_1^*, E_1^*)$ = (32, F) in E_1^* =1/4 =0.25 and $f_2(t_1^*, E_1^*)$ = (130352) in E_1^* = 2/4 =0.5. Similarly, $f_1(t_1^*, E_2^*)$ = (32, F) in E_2^* = 0 and $f_2(t_1^*, E_2^*)$ = (130352) in E_2^* = 0. Intuitively, $f_i(t^*, E^*)$ measures the degree of matching on column col_i between tuple t^* and equivalence class E^*.

Table 3.10 A 2-diverse published table



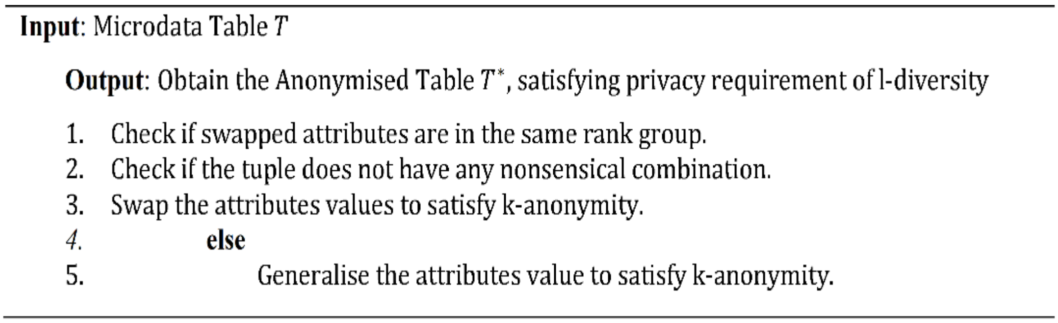
(Age, Gender)	(Zip code, Disease)
(32, F)	(130352, flu)
(22, M)	(130352, heart disease)
(28, M)	(130350, flu)
(30, M)	(130350, dyspepsia)
(53, F)	(130355, heart disease)
(39, F)	(130353, flu)
(60, M)	(130355, heart disease)
64, M)	(130353, HIV)

When computing the probability distribution $p(s^*|t^*, E^*)$ for a given pair (t^*, E^*), where t^* belongs to equivalence class E^*, the sensitive value of t^* can be determined by examining the sensitive column of E^. The sensitive column of E^ contains the quasi-identifier (QI) attributes, and only sensitive values that match t^*'s QI values are considered as candidate sensitive values for t^*. In the absence of additional knowledge, all candidate sensitive values within the equivalence class are equally likely. Let

$D(t^*, E^*)$ represent the distribution of candidate sensitive values for t^* within the equivalence class E^*.

Definition 4 ($D(t^*, E^*)$): Any sensitive value associated with t^* in equivalence class E^*, denoted by$t^*[col_c - \{s\}]$, is considered a candidate sensitive value for t^*. The total count of candidate sensitive values for t^* in E^*, including duplicates, is denoted by $f_c(t^*, E^*)$. The distribution $D(t^*, E^*)$ represents the probability distribution of the candidate sensitive values within the equivalence class E^*, and $D(t^*, E^*)(s^*)$ denotes the probability of the sensitive value s^* in this distribution. For example, in Table 3.10, the distribution for t_1^* within equivalence class E_1^* is given by $D(t_1^*, E_1^*) =$ (heart disease: 0.5, flu: 0.5). Consequently, the probability of the sensitive value "Flu" for t_1^* within E_1^* is 0.5, which can be denoted as $D(t_1^*, E_1^*)(\text{flue}) = 0.5$. Thus, the probability p$(s^*|t^*, E^*)$ is exactly equal to **(**$D(t^*, E^*)(s^*)$, i.e., p$(s^*|t^*, E^*) = (D(t^*, E^*)(s^*)$.

In step 2 of the improved protection stage, referred to as the swapping or generalization step, the protection of randomly permuted values in an equivalence class may not fully guarantee privacy from attribute or membership disclosure. This is because the permutation of values can increase the risk of attribute disclosure rather than ensuring privacy (Hasan et al., 2018). Therefore, the proposed algorithm in this study aims to satisfy the privacy requirement within each equivalence class. To optimize data utility and strengthen personal privacy within the UL approach, this step integrates rank swapping as a protection method. Rank swapping is utilized to break the association between unique attributes and cells with highly identical values. It involves exchanging attribute values between pairs of records within a subset of the original data. This procedure modifies the tuple data by swapping attribute values that are either unique or highly identical. If attribute swapping is not feasible, the protection method in this step employs attribute generalization. Generalization transforms the attributes into more generalized forms to safeguard privacy. It is important to note that rank swapping offers advantages in preserving data utility, particularly when dealing with aggregate queries, compared to the generalization approach. By implementing rank swapping, the algorithm ensures a balance between preserving privacy and maintaining data utility within the UL approach. The primary objective of attribute swapping or generalization is to generate the anonymized table T. This table guarantees that there are no nonsensical combinations (invalid tuples) within the records and satisfies the requirement of l-diverse slicing, as depicted in Figure 3.8.



```
Input: Microdata Table $T$
    Output: Obtain the Anonymised Table $T^*$, satisfying privacy requirement of l-diversity
    1. Check if swapped attributes are in the same rank group.
    2. Check if the tuple does not have any nonsensical combination.
    3. Swap the attributes values to satisfy k-anonymity.
    4.         else
    5.             Generalise the attributes value to satisfy k-anonymity.
```

Figure 3.8 Swapping or generalisation of attributes

To confirm the reliability of attribute swapping, it involves checking if the values of attribute a_i^{**} are in the same rank group (line 1). The values of attribute a_i^{**} are first ranked in ascending order before swapping each of the ranked values with another randomly selected ranked value from a specified range. For example, $Level_0$ in Figure 3.9 possesses two categories: $\{\mathrm{Federal-gov, Local-gov, State-gov}\}$ and $\{\mathrm{Self-emp-inc,\ Self-emp-not-inc}\}$. Rank swapping can maintain multivariate relationships more appropriately than ordinary data swapping (Matthews and Harel, 2011;Domingo-Ferrer and Torra, 2002)·

In line 2, the values of an attribute a_i^{**} that need to be swapped are checked to ensure that they do not contain nonsensical combinations that could adversely affect the usefulness of the published microdata.

In line 3, two attributes exchange values in cases where the two attributes belong to the same hierarchy category and nonsensical combinations are absent. However, if the attributes do not belong to the same category or if nonsensical combinations are present, attribute values are generalized to satisfy k-anonymity (line 5). During attribute generalization, the entire equivalence class is not generalized, which provides a better chance for enhancing the usefulness of the data compared to full table or column generalization. This, in turn, enhances the usefulness of the published dataset.

Definition 5 (Attribute Generalisation): Let T^{**} be a part of the microdata table T, and a_i^{**} represents a set of quasi-identifier (QI) attributes in T^{**}. The QI attribute values are replaced by their generalized model through generalization. Assume d_i^{**} and

d_j^{**} are two domains with dimensional regions $\{d_{i1,}^{**},\ d_{i2}^{**} \dots, d_{in}^{**}\}$ and $\{d_{J1,}^{**},\ d_{j2}^{**}, \dots d_{Jn}^{**}\}$ respectively, where $\cup_{d_{in,}^{**}} = d_i^{**}$ and $d_i^{**} \cap d_j^{**} = \emptyset$. The values in d_j^{**} represent the generalization of the values in domain d_i^{**}, denoted as $d_i^{**} < d_j^{**}$ (a many-to-one value generalization approach). Generalization follows a domain generalization hierarchy, which is a collection of domains ordered according to the relationship $d_i^{**} < d_j^{**}$ (see Figure 3.9).

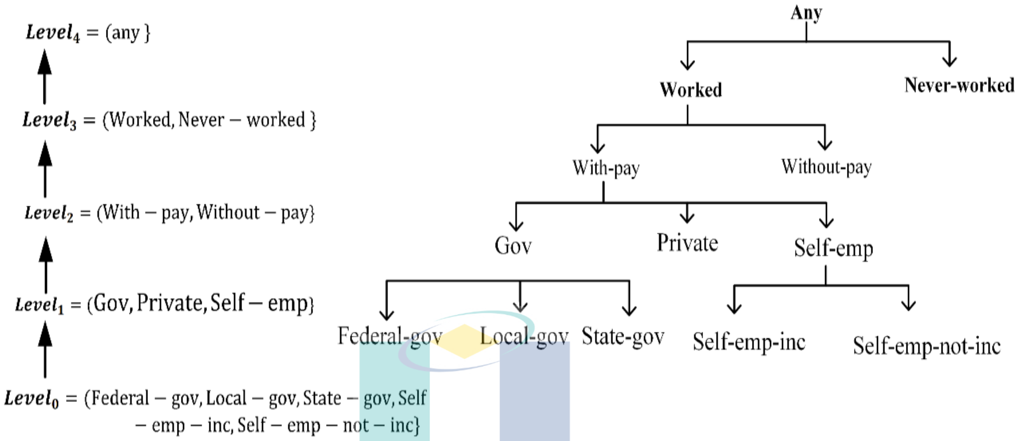

$Level_4$ = (any}
$Level_3$ = (Worked, Never − worked}
$Level_2$ = (With − pay, Without − pay}
$Level_1$ = (Gov, Private, Self − emp}
$Level_0$ = (Federal − gov, Local − gov, State − gov, Self − emp − inc, Self − emp − not − inc}
Any
Worked
Never-worked
With-pay
Without-pay
Gov
Private
Self-emp
Federal-gov
Local-gov
State-gov
Self-emp-inc
Self-emp-not-inc

Figure 3.9 of domain (left) and value (right) generalisation hierarchies for the work–class (WC) attributes

In Figure 3.9 (right), the likely domain generalization hierarchy for work-class (WC) attributes is described. At lower levels in the generalization hierarchy for WC attributes, generalization is not used. Nonetheless, at the top levels of the hierarchy, the WC tends to be more general. A singleton is a maximal domain level element that denotes the likelihood of values being generalized in every domain to a single value.

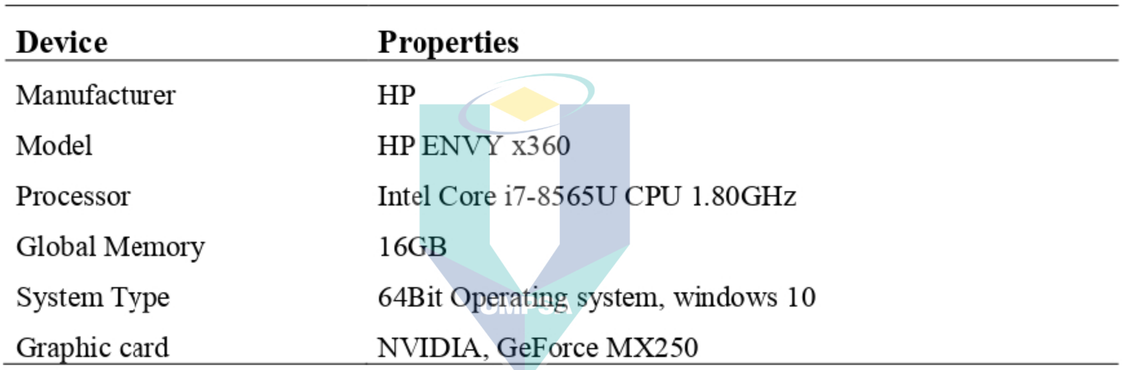
3.4 Evaluation of Performance

The experiments on real datasets are divided into two stages. The first deals with the measurement of the protection level needed, while the second is concerned with the assessment of the suggested approaches to see how well they can fight and prevent composition attack occurrences. The effectiveness of the approach was tested by comparing it to the effectiveness of similar approaches such as hybrid (Li et al., 2016), $e - DP$(Mohammed et al., 2011;Zorarpacı and Özel, 2021), merging (Hasan et al., 2018), probabilistic (Sattar et al., 2014), composition (Baig et al., 2012) and Mondrian (LeFevre et al., 2006).

3.5 Hardware and Software

A HP laptop was used to assess the approach. The features of the device with graphic device capabilities are described in Table 3.11. Furthermore, the Python language, which was invented by Guido van Rossum in 1989 (Mészárosová, 2015), was applied to implement this experiment. Python is an interpreter programming language that is created as an open-source project. It works in many operating systems, such as Microsoft Windows, Linux, and Unix systems, including MacOS X, and is fully supported by standard and third-party libraries through the mere duplication of the program’s source code (Mészárosová, 2015).

Table 3.11 Hardware and Graphic Cards Specifications



Device	Properties
Manufacturer	HP
Model	HP ENVY x360
Processor	Intel Core i7-8565U CPU 1.80GHz
Global Memory	16GB
System Type	64Bit Operating system, windows 10
Graphic card	NVIDIA, GeForce MX250

3.6 Chapter Summary

In conclusion, this chapter presented the flow of research procedures. It started by describing the problem identification, followed by an explanation of the proposed research methodology to attain the research objectives. The proposed research methodology consists of three stages, and all stages have been highlighted and put into context, from the design of the UL approach and the improved protection method for anonymization to the approach of being more effective in determining the amount of protection required to maintain sustainable data utility and achieve a higher degree of privacy protection. The datasets and methods to evaluate the proposed work have also been listed and discussed. The following chapter elaborates on the evaluation results stages presented in this chapter.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

In this section, we introduce the Upper Lower (UL) level-based protection approach for data anonymization, building upon the slicing-based approach discussed in Chapter 2. The slicing approach is recognized as a novel technique for data anonymization, offering a distinct protection method without resorting to generalization.

The fundamental concept behind the slicing approach is to prioritize data privacy while preserving data utility by leveraging correlation measurements between the attributes. Highly correlated attributes are grouped together in columns, allowing for the preservation of correlations between these attributes. This correlation-based approach safeguards privacy by disrupting associations between uncorrelated attributes in other columns through protection methods such as random permutation and generalization. However, these protection methods may not always provide a reliable defense against attribute or membership disclosure. Additionally, merging procedures can generate fake tuples, resulting in a loss of data utility and incorrect knowledge extraction.

To address the limitations of the slicing approach, the UL approach has been designed. The UL approach aims to prevent attackers from identifying individuals or disclosing sensitive information in a table while simultaneously determining the optimal balance between privacy and data utility.

This chapter evaluates the UL approach used to anonymize published data and assesses the outcomes in different sections. Section 4.2 presents the design of the UL approach, explaining the basic idea/concept, improved protection method, the dataset preparation, and parameters used while implementing the proposed work, and measuring the protection level. Then, the evaluation of the UL approach is discussed in Section 4.3 in terms of measuring risks and data utility compared to existing works such as merging, $e - DP$, Mondrian, composition, probabilistic, and hybrid approaches. Data utility is evaluated by measuring the extent of information loss and assessing the accuracy of aggregate query results obtained using the data. Finally, Section 4.4 summarizes the chapter.

4.2 Design of UL approach

The main idea behind the UL approach, introduced in this section, is to prevent attackers from identifying individuals or disclosing sensitive information in a table while simultaneously determining the optimal balance between privacy and data utility. Achieving this balance would require improved protection methods that can effectively identify specific attributes capable of detecting potential disclosure risks. By identifying these attributes, the UL approach determines the required protection level, thereby enhancing the privacy of published data without sacrificing its utility.

To conduct the experiments, independent datasets were required to simulate the realistic scenario of publishing independent data where the dataset is anonymized by the publisher without considering other published datasets. From the Adult dataset, we extracted the Education and Occupation datasets and then created five copies for each dataset. This resulted in datasets of the following sizes: 4.1K, 4.2K, 4.3K, 4.4K, and 4.5K, each containing eight QI attribute values.

In terms of the parameters, assessing the UL approach is an important step. In many cases, the Single publication model is considered a non-interactive data publishing method used for experimental analysis. The experiment was conducted in a non-interactive privacy setting. However, a significant portion of the research on differential privacy (Dwork, 2006) aligns with interactive settings. In interactive settings, the data collector performs specific functions on the data to respond to queries from the data analyzer. In this experiment, a user can access the dataset using numerical queries, as the anonymization approach adds noise to the query answers. However, the practical environment may not always support this approach, as datasets are typically published publicly. Therefore, for the experiment on differential privacy, the non-interactive setting was chosen, as highlighted in (Mohammed et al., 2011).

This chapter aims to evaluate the UL approach, focusing on achieving an optimal balance between privacy and data utility, as well as determining the required level of protection. The efficiency of the UL approach is assessed using various approaches within non-interactive privacy settings. These approaches include the hybrid approach (Li et al., 2016), merging approach (Hasan et al., 2018), $e - DP$ approach (Mohammed et al., 2011), probabilistic approach (Sattar et al., 2014), Mondrian approach (LeFevre et al., 2006), and composition approach (Baig et al., 2012). For the purpose of comparison experiments, the equivalence class was selected based on previous studies. To establish an equivalence class, we chose k = 4 and k = 6, where I-diversity is also indicated as I = 4 and I = 6. The primary goal of I-diversity is to enhance privacy preservation by increasing the diversity of sensitive values. In the context of differential privacy, Laplacian noise with $e\ =\ 0.3$ is added to the count of sensitive values within a given equivalence class (Sattar et al., 2014), representing the e-differential privacy budget. Two fundamental factors can be considered for comparison purposes: data utility and risk disclosure. Any modification or alteration of these parameters directly impacts both the preservation of privacy and the utility of the data.

4.2.1 Experimental Results for Measuring Required Protection level

In this experiment, the correlation coefficient (r) was calculated for partition Table T^{**} of attributes, with the values of attribute a_i^{**} classified into three categories, as shown in Figure 3.6.

The protection level was measured using the upper protection levels (UPL) and lower protection levels (LPL) with five changes in swap rates Φ for LPL and UPL to determine the number of cells and tuples in each variation. The summarized results of measuring the protection level are presented in Table 4.1 and Table 4.2, with a specific focus on the Educational datasets. These tables provide detailed information on the number of cells and tuples in each LPL and UPL, considering different swap rates Φ applied to the partitions T^{**}. Both the Education and Occupation datasets share the same attributes, with each attribute having an equal number of classifiers. As a result, measuring the relationship strength between attributes led to highly similar or identical results, with a specific focus on the Educational datasets.

In Table 4.1, the emphasis is on cells that contain unique attributes within the tuples. These cells are considered potentially riskier because of their uniqueness or near-uniqueness. The table provides information about the count of such cells, which helps in assessing the vulnerability of the dataset. On the other hand, Table 4.2 focuses on the number of cells in the tuples that have matching (highly identical) attributes. Cells with matching or near-matching attributes are deemed potentially riskier because most of these tuples belong to the same equivalence class. This gives the adversary confidence in inferring sensitive information by linking these attributes with other attributes or datasets.

By analyzing these tables, one can measure the protection level of the Educational datasets and gain insights into the potential risks associated with unique and matching attributes within the tuples. This information must be considered by the decision maker based on the disclosure risk and data utility, taking into account the measures of the strength of the relationship between attributes.

The strength of the association between attributes was considered due to the known strength and variety of data (see Table 3.4). Subsequently, LPL and UPL were utilized to identify specific attributes for swapping, as opposed to using a random approach to break correlations between attribute values. These steps aid in the identification of attributes that could potentially reveal personal information, determining the appropriate level of protection required to prevent the disclosure of private data. As a result, the UL approach effectively mitigates identity disclosure while striking a balance between preserving privacy and maintaining data utility. The primary goal is to safeguard personal information while still ensuring the data remains useful for various purposes.

In Table 4.1, a higher swap rate Φ indicates increased privacy, while in Table 4.2, a lower swap rate Φ indicates higher privacy but decreased data utility. These tables offer valuable insights into the trade-offs between privacy and data utility, assisting decision-makers in making informed choices.

Table 4.1 Five changes of swap rates for LPL to calculate the number of cells and tuples in each change

	$0.0 < LPL \leq \Phi$																			
Data set	$\Phi = 0.01$				$\Phi = 0.02$				$\Phi = 0.05$				$\Phi = 0.10$				$\Phi = 0.15$			
	#of cells		**#of tuples for each cell**		**# of cells**		**#of tuples for each cell**		**# of cells**		**#of tuples for each cell**		**# of cells**		**#of tuples for each cell**		**# of cell**		**#of tuples for each cell**	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
4.1K	5	16	11	156	7	28	24	654	13	44	111	1082	17	62	324	1132	21	68	542	1230
4.2K	6	35	9	488	8	59	18	1982	13	103	128	2412	17	115	342	2824	22	153	611	2910
4.3K	6	16	13	16	7	34	20	238	10	53	107	1386	17	67	398	1988	20	76	631	2088
4.4K	7	39	12	1619	7	58	12	1674	14	102	166	2590	17	146	362	2703	22	146	639	2703
4.5K	6	22	17	22	8	37	27	273	12	50	117	467	15	67	229	721	19	72	526	743

Table 4.2 Five changes of swap rates for UPL to calculate the number of cells and tuples in each change

	$\Phi \leq UPL < 1.0$																			
Data set	$\Phi = 0.99$				$\Phi = 0.98$				$\Phi = 0.95$				$\Phi = 0.90$				$\Phi = 0.85$			
	# of cells		**#of tuples for each cell**		**# of cells**		**#of tuples for each cell**		**# of cells**		**#of tuples for each cell**		**# of cells**		**#of tuples for each cell**		**# of cell**		**#of tuples for each cell**	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
4.1K	1	0	1362	0	2	0	1626	0	2	0	1626	0	2	1	1626	264	2	3	1626	302
4.2K	1	0	1412	0	2	0	1674	0	2	0	1674	0	2	0	1674	0	2	0	1674	0
4.3K	2	0	1412	0	2	0	1412	0	2	0	1412	0	2	0	1412	0	2	3	1412	34
4.4K	1	0	1501	0	2	0	1787	0	2	0	1787	0	2	0	1787	0	2	0	1787	0
4.5K	1	0	1555	0	2	0	1780	0	2	0	1780	0	2	0	1780	0	2	3	1780	36

4.3 Experimental Results for Comparison Evaluation

In recent years, numerous anonymization approaches have been proposed, showing significant superiority over older approaches. These approaches integrate various protection methods to strike a balance between privacy preservation and data usability.

For the experiment, different groups of datasets were utilized based on the initialized dataset presented in Figure 3.2, and they were used as inputs for the following approaches: $e - DP$ (Mohammed et al., 2011; Zorarpacı and Özel, 2021), Hybrid (Li et al., 2016), Probabilistic (Sattar et al., 2014), Composition (Baig et al., 2012), Mondrian (LeFevre et al., 2006), Merging (Hasan et al., 2018), and proposed approach. The objective was to calculate the privacy risks and corresponding data utility for each approach.

As a protection method against sensitive value disclosure, the hybrid approach in (Li et al., 2016) utilizes a combination of sampling, generalization, and perturbation by adding Laplacian noise to the count of every sensitive value in each equivalence class. This combination aims to protect against composition attacks. In the $e - DP$ approach, protection is achieved by initially generating a generalized contingency table, followed by adding noise to the counts. The probabilistic approach suggests a new protection method called (d, α)-linkable. It tries to limit the likelihood of an adversary completing a composition attack by ensuring that the (d) sensitive values are associated linked to a QI-group with a specific likelihood (α). This is achieved by identifying correlations between the Quasi-identifier (QI) and sensitive attributes. The Composition approach combines two novel concepts, (ρ, α)-anonymization through sampling and composition-based generalization, to provide protection against composition attacks for independent datasets. The Mondrian approach involves recursively partitioning the domain space into multiple regions, ensuring that each region contains a minimum of k records. In each equivalence class, a set of Quasi-identifier (QI) values are generalized. The primary goal of the merging approach is to ensure privacy by employing both vertical and horizontal partitioning of data. The partitioning of data aims to enhance privacy protection by increasing the occurrence of false matches for unique attributes. Additionally, it is essential to randomly permute the sliced table as part of the process.

The ability to protect data privacy by lowering the risk of disclosing personal information and maintaining the potential use of published data is the criterion for judging the effectiveness of the anonymization approach (Majeed and Lee 2021; Siddique et al., 2018; Hasan et al., 2018; BinJubier et al., 2022). The use of lower protection level (LPL) and upper protection level (UPL) as protection method in UL approach can be instrumental in determining the required level of data protection and selecting particular cells that help to identity disclosure. Then, rank swapping is employed to protect unique and highly identical attributes, ensuring the privacy protection of data while preserving data utility. Therefore, the UL approach decrease the risk of disclosing data to people and maintain possible utilization. Consequently, the experiment was carried out to compare the efficiency of the proposed UL approach with six of the approaches used. The comparison is performed based on two main factors: data utility and risk disclosure. Data utility is evaluated by measuring the extent of information loss and assessing the accuracy of query results obtained using the data. These factors are elaborated upon in the subsequent subsections.

4.3.1 Experimental Results for Data Utility Comparison

This experiment aims to evaluate the data utility achieved through the distortion ratio (DR) in published data. The DR, which quantifies the anonymization outcome on the overall distortion data, can be assessed using various methodologies, as proposed by Wong and Fu (Wong and Fu , 2010). Among these methodologies, the generalised distortion ratio (GDR) serves as a suitable measure for estimating the DR, as indicated by Rohilla (Rohilla, 2015). As explained in section 2.5 (Data Utility and Measuring Risks), the GDR is used to quantify the anonymization outcome on the overall distortion data. By employing the GDR as a metric, we can effectively assess the impact of anonymization on the data utility, allowing us to strike a balance between privacy protection and data quality.

Figure 4.1 and Figure 4.2 present the data utility outcomes resulting from data loss on the Educational dataset. In Figure 4.1, the proposed approach indicated a 2% swap rate (Φ) by Low Protection Level (LPL) and 98% by Upper Protection Level (UPL). Additionally, Figure 4.2 showed that the proposed approach had a 5% swap rate (Φ) by LPL and 95% by UPL. The selection of the swap rate (Φ) by decision makers is essential to control the required protection level, and this is achieved by referencing the modifications in swap rates, as demonstrated in Table 4.1 and Table 4.2. An increase in Φ in LPL or a decrease in Φ in UPL results in improved privacy but leads to a decrease in utility data. The evaluation of the proposed approach was conducted by comparing it with various acknowledged approaches: the hybrid approach (Li et al., 2016), merging approach (Hasan et al., 2018), $e - DP$ approach (Mohammed et al., 2011), probabilistic approach (Sattar et al., 2014), Mondrian approach (LeFevre et al., 2006), and composition approach (Baig et al., 2012). The results showed that the proposed approach achieved higher data utility than all the compared approaches. Specifically, the proposed approach (UL approach) outperformed the other approaches in terms of data utility with the Educational dataset of size 4.5K. It achieved approximately 92.47% data utility with a swap rate Φ of 2% by LPL and 98% by UPL, and approximately 92.19% data utility with a swap rate Φ of 5% by LPL and 95% by UPL. This superiority is attributed to the UL approach's reliance on value rank swapping, which ensures the preservation of more data utility compared to the merging approach. The merging approach used N fake tuples with the same Quasi-Identifier (QI) value as the original table, and the Sensitive Attribute (SA) values were assigned to them based on the main dataset's SA value distribution. Consequently, the proposed approach exhibited less data loss compared to the merging approach. The proposed approach adopted a selective generalization technique within the cell while fulfilling the privacy requirement, which played a crucial role in preserving more data utility. On the other hand, the remaining approaches employed protection methods that contributed to higher data loss. More information about the protection methods used in these approaches can be found in Section 2.4 (Protection Methods Based on Anonymization Approaches).

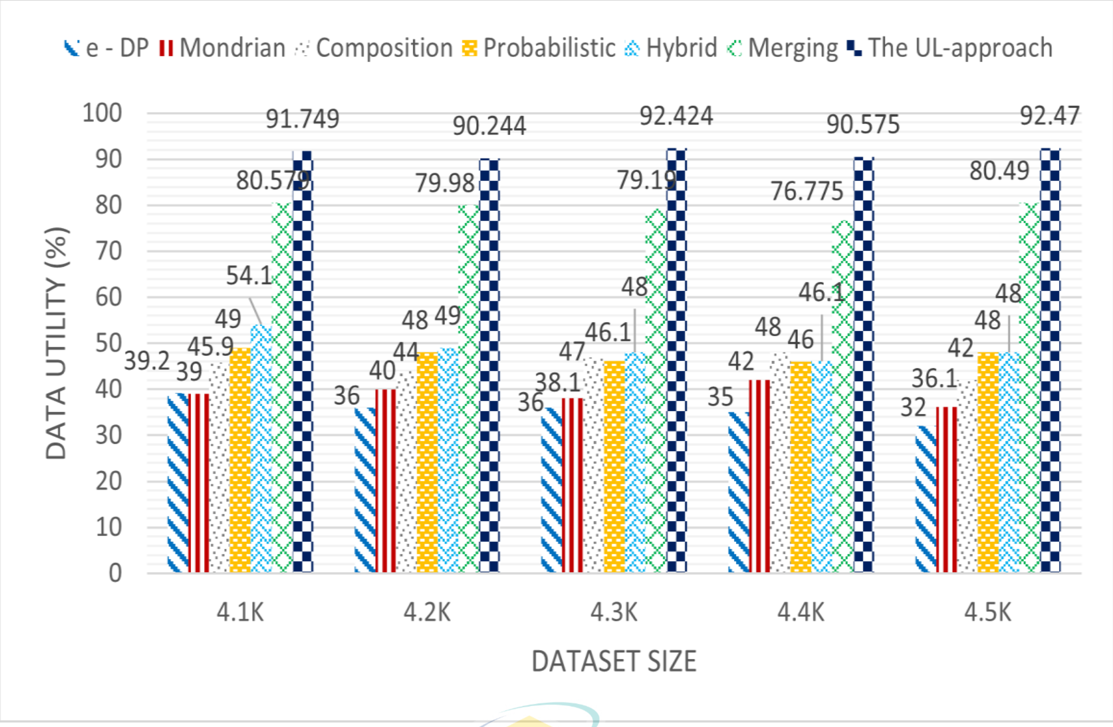


Figure 4.1 Education Dataset utility (swap rate (Φ) of 2% and 98%)

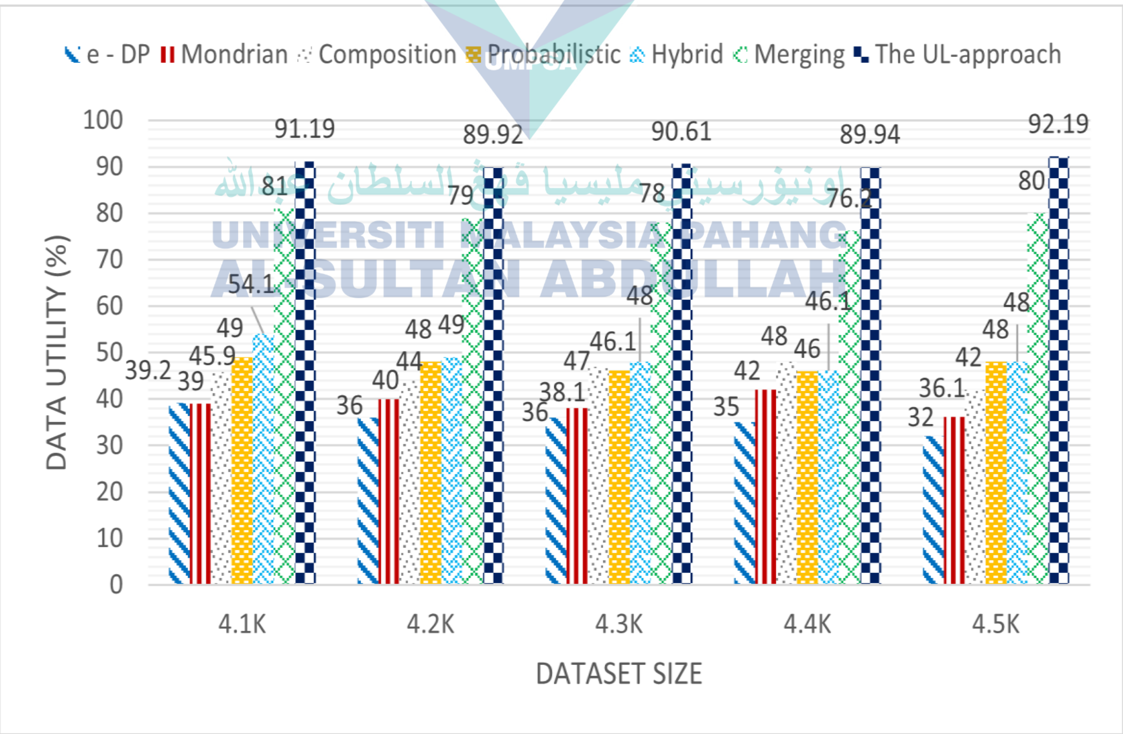


Figure 4.2 Education Dataset utility (swap rate (Φ) of 5% and 95%)

4.3.2 Experimental Results for Measuring Risks

This section focuses on measuring the disclosure risk ratio (DRR) of the anonymization approach outputs to ensure privacy preservation. Determining the appropriate level of protection is critical, and it is essential to use risk disclosure measures that are independent of the data representation method selected. Existing research has demonstrated that risk disclosure can be evaluated using a Certainty Penalty (CP) (Hasan et al., 2018; BinJubier et al., 2022), which calculates the percentage of true matches to the total matches, as explained in Section 2.5 (Data Utility and Measuring Risks).

The experimental results for the disclosure risk ratio (DRR) are presented for both the Educational dataset (Figure 4.3 and Figure 4.4) and the Occupational dataset (Figure 4.5 and Figure 4.6). DRR serves as a measure of an adversary's confidence level in inferring sensitive values within these datasets. Notably, the $e - DP$ approach (Mohammed et al., 2011) exhibited the lowest privacy risks compared to the UL approach and other existing methods. Specifically, when using k = 4, l = 4 and k = 6, l = 6 for a dataset size of 4.5K, the $e - DP$ approach achieved approximately 0.7% and 0.69% disclosure risk ratio (privacy risk) for the Education and Occupation datasets, respectively. The proposed solution (Mohammed et al., 2011) involves probabilistically generating a generalized possibility table and introducing noise to the total, providing a high level of privacy assurance and protection against composition attacks through differential privacy grounded data anonymization (Zorarpacı and Özel, 2021; A. Hasan et al., 2018), as evidenced in the results. However, previous research by (Li et al., 2016; Cormode et al., 2013; Sarathy and Muralidhar, 2011; Hasan et al., 2018) has observed that using $e - DP$ to protect against composition attacks may lead to a significant amount of data utility loss during anonymization, validating the findings discussed in Figure 4.1 and Figure 4.2.

The hybrid approach (Li et al., 2016) exhibited a lower probability of exposing the end user's private data compared to the probabilistic approach (Sattar et al., 2014), composition approach (Baig et al., 2012), Mondrian approach (LeFevre et al., 2006), and merging approach (Hasan et al., 2018). For the Educational and Occupational datasets, the hybrid approach achieved approximately 1.9% and 1.98% disclosure risk ratio (privacy risk) when using k = 4, l = 4, and approximately 1.55% and 1.6% when k = 6, l = 6, for a dataset size of 4.5K, respectively. Furthermore, compared to the probabilistic approach, the merging approach reduced the likelihood of composition attacks on the released datasets (Sattar et al., 2014), composition approach (Baig et al., 2012), and Mondrian approach (LeFevre et al., 2006).

Additionally, the proposed approach demonstrated a lower probability of exposing the user's private data compared to the hybrid approach (Li et al., 2016), merging approach (Hasan et al., 2018), probabilistic approach (Sattar et al., 2014), Mondrian approach (LeFevre et al., 2006), and composition approach (Baig et al., 2012). This achievement was accomplished by disabling unique cells and high identical cells for both the Upper Protection Level (UPL) and Lower Protection Level (LPL), as well as by enforcing the presence of numerous similar cells in every bucket. These measures effectively ensured protection against identity disclosure.

Given the subtle nuances observed in the comparative outcomes, conducting significance tests for certain approaches becomes impractical. This observation underscores the lack of statistical significance in the disparities observed between approaches. For instance, the UL approach exhibited a privacy risk of approximately 1.5% in Figure 4.4, utilizing a dataset size of 4.5K within the Education domain. Conversely, the merging, hybrid, probabilistic, and composition approaches displayed privacy risks of roughly 1.65%, 1.55%, 1.9%, and 2.2%, respectively. Consequently, due to the marginal disparities and the absence of statistical significance, conducting significance tests would not yield insights into the efficacy of the proposed approach compared to its counterparts. For further elucidation and detailed analysis, please refer to Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6.

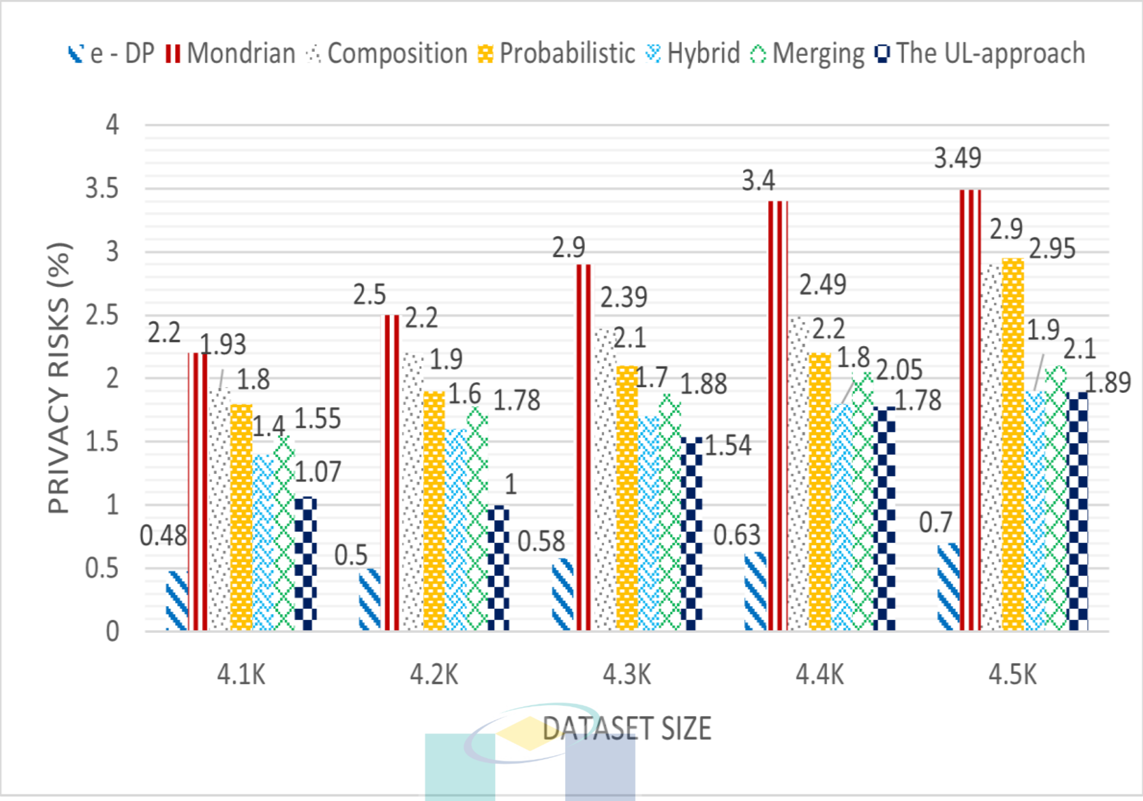


Figure 4.3 Privacy risk for Education dataset (k = 4, l = 4)

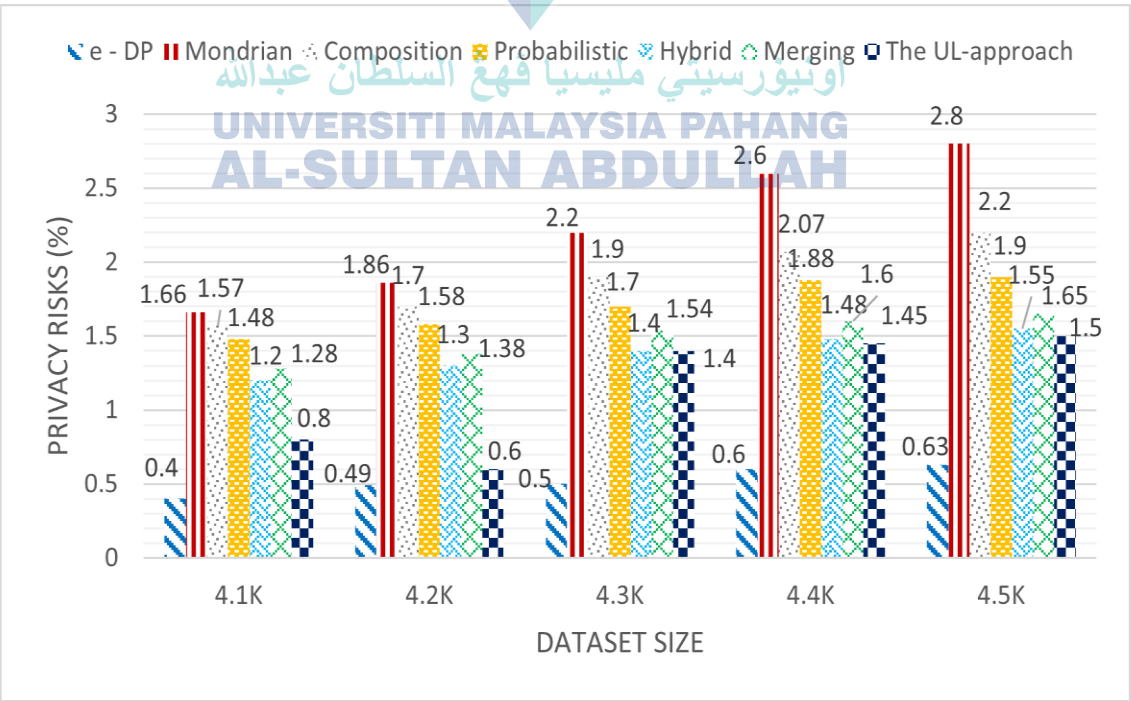


Figure 4.4 Privacy risk for Education dataset (k = 6, l = 6)

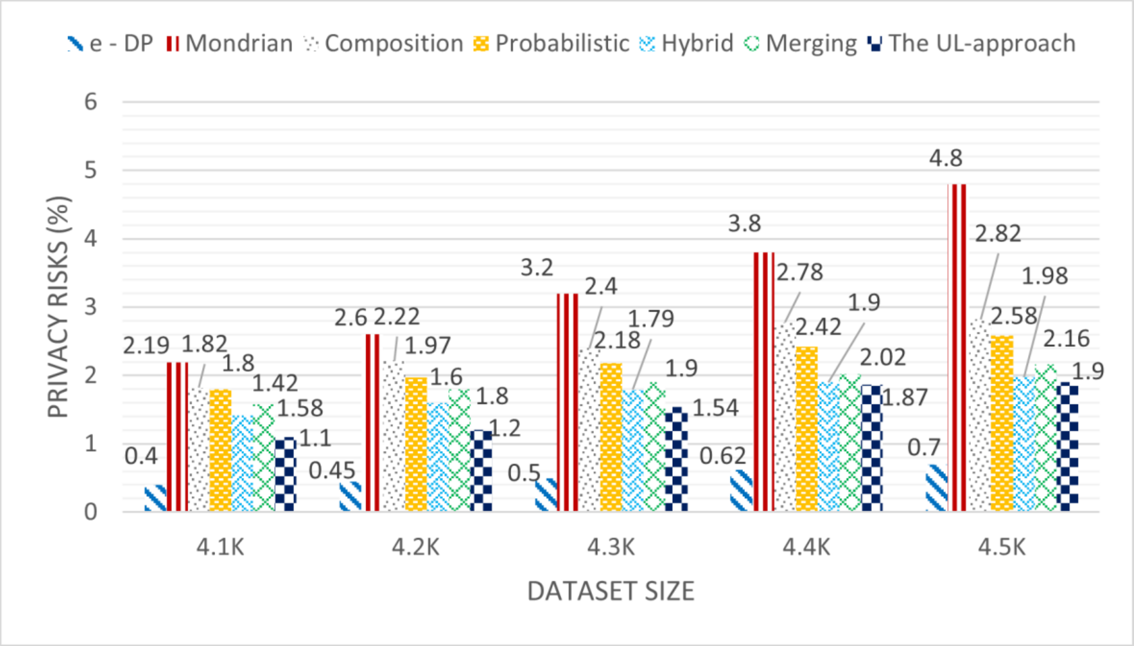


Figure 4.5 Privacy risk for Occupation dataset (k = 4, l = 4)

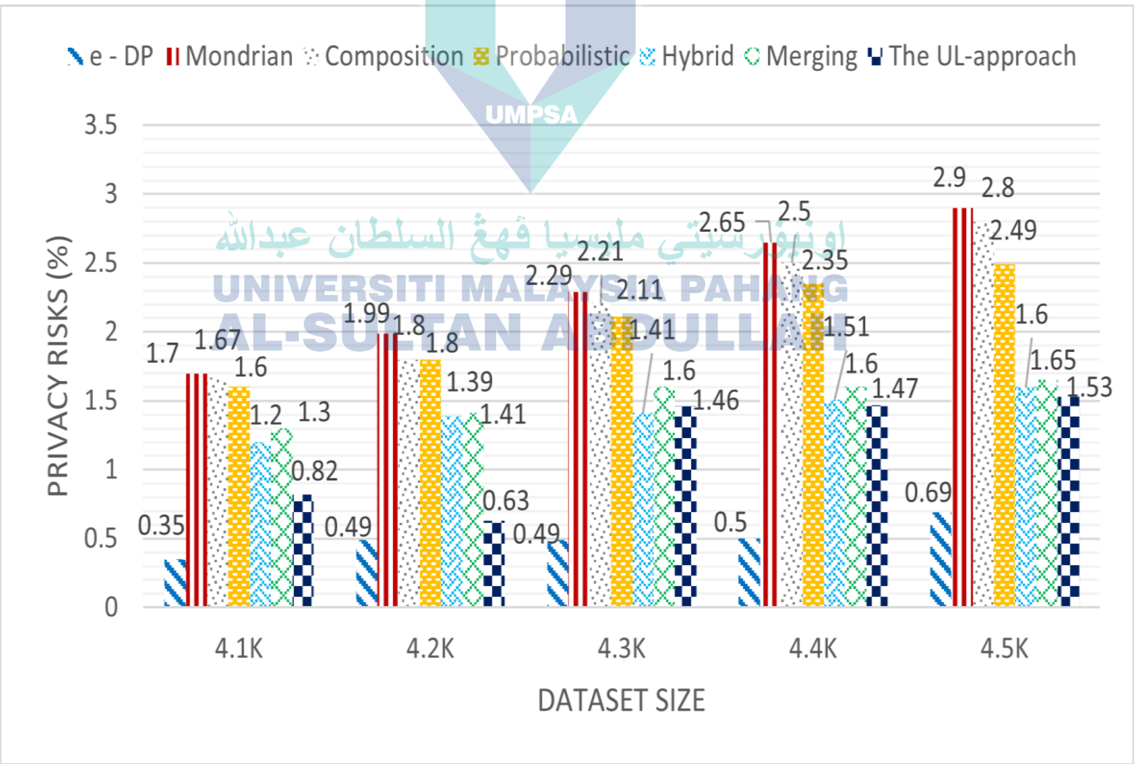


Figure 4.6 Privacy risk for Occupation dataset (k = 6, l = 6)

Figure 4.7 presents a comprehensive summary of the experimental results for the disclosure risk ratio (DRR) based on different swap rates Φ = {(1%, 99%), (2%, 98%), (5%, 95%), (10%, 90%), (15%, 85%)} for the lower protection levels (LPL) and upper protection levels (UPL) using the Educational dataset with a size of 4.5K. The results in Figure 4.7, Table 4.1, and Table 4.2 demonstrate that an increase in Φ in LPL or a reduction in Φ in UPL leads to higher privacy levels but lower data utility. Throughout this research, the composition attack was mitigated by effectively handling unique attributes and high identical attribute values through the utilization of LPL and UPL, while also introducing diversity of cells to prevent identity disclosure. Consequently, a cell is susceptible to disclosure risk if it can be distinguished from others and lacks diversity within its equivalence class (Taylor, Zhou, and Rise, 2018).

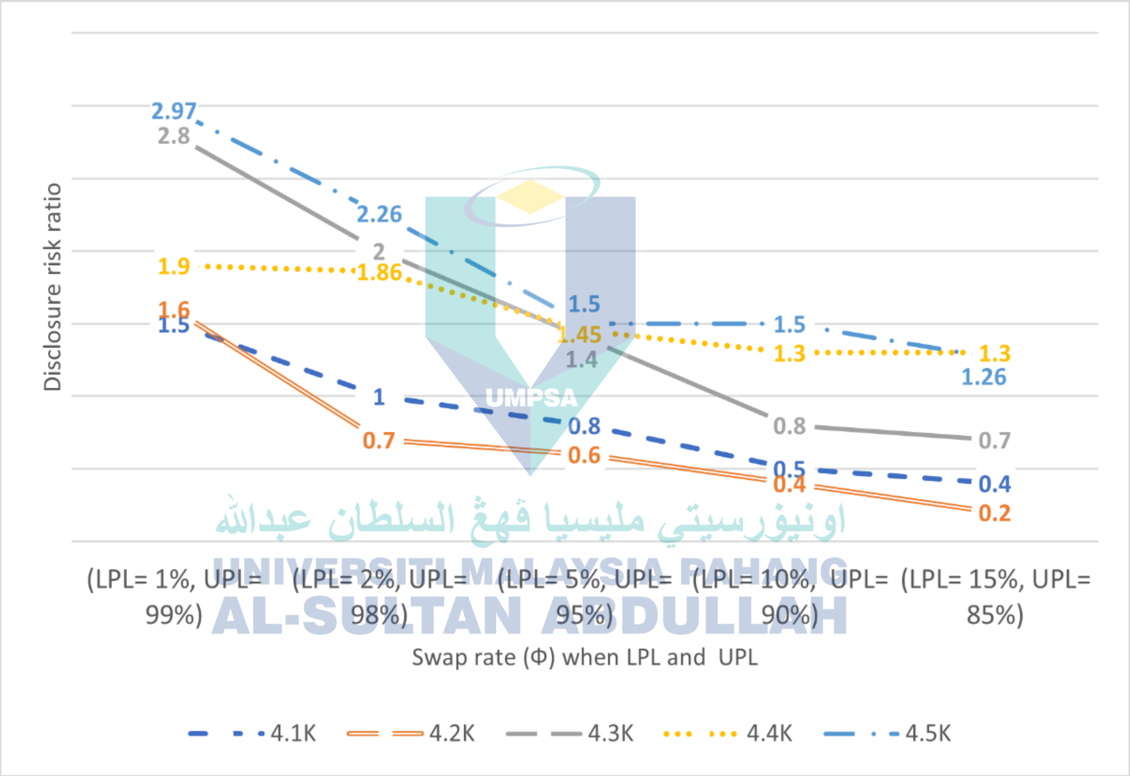


Figure 4.7 Experimental result for DRR for LPL and UPL when Φ={(1%, 99%), (2%, 98%), (5%, 95%), (10%, 90%), (15%, 85%)}

4.3.3 Experimental Results for Aggregate Query Error

In addition to assessing the utility of data through "penalty" measures derived from the distortion ratio (DR), the utility of data in anonymized form was also evaluated using a relative query error. This assessment involves using data as input for a query and evaluating the accuracy of the query results, as discussed earlier in section 2.4.4 (Data Utility and Measuring Risks). Aggregate queries, such as 'COUNT,' 'MAX,' and 'AVERAGE,' were repeatedly employed by operators to generate crucial numerical values representing the predictable data utility, thus validating the effectiveness of the approach.

To address the aggregate queries in this experiment, the "COUNT" operator was used, as elaborated in section 2.4.4 (Data Utility and Measuring Risks). Each query was applied to both the initial data and the data produced by the UL approach, along with other available approaches. The initial and anonymized data both underwent counting, where the count for the original data was denoted as org_{count}, and the count for the anonymized data was represented as anz_{count}, with anz_{count} pertaining to the proposed approach and other available approaches. The average relative error in the anonymized dataset was computed for all queries using Equation 4.1 (Zhang et al., 2007):

$$Relative\ error = \frac{org_{count} - anz_{count}}{org_{count}} * 100\% \qquad 4.2$$

In the experiment, we selected a series of quasi-identifier (QI) attributes for evaluation, including workclass; followed by sex and workclass; then sex, workclass, and marital-status; further followed by sex, workclass, marital-status, and relationship; and finally, sex, workclass, marital-status, relationship, and occupation. Figure 4.8 displays the relative query error on the y-axis based on these chosen QI attributes. The Mondrian, hybrid, $e - DP$, probabilistic, composition, merging, and proposed approach underwent evaluation with k set to 6 and l-diversity set to 6 for merging and the proposed approach. The swap rates Φ were specifically set at LPL = 5% and UPL = 95%.

To compute the relative query error, the anonymized tables generated by the proposed approach were compared with tables created by other available approaches. The comparison involved different combinations of one, two, three, four, or five QI attributes. Furthermore, for the 4.5K Occupational dataset, numerous potential query variations were formulated and executed across the anonymization tables.

Figure 4.8 illustrates the relative query error, with the y-axis representing the relative percentage error and the x-axis indicating different options for the QI attributes. The experimental results consistently demonstrate that the proposed approach, which uses the rank swapping method, outperforms generalization in terms of answering aggregate queries. Furthermore, the proposed approach exhibits a relatively minor error compared to other approaches. Thus, when attribute swapping is not feasible, the attributes are generalized to ensure privacy protection and data utility.

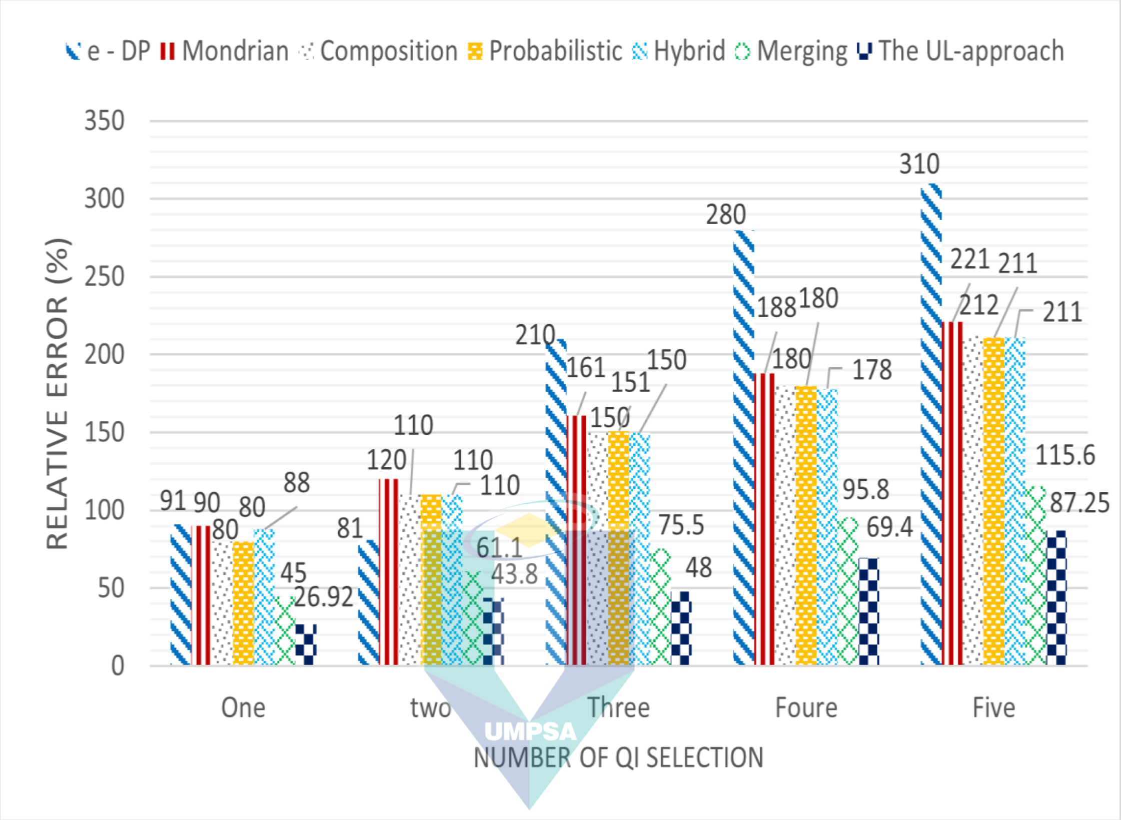


Figure 4.8 Aggregate query error

4.4 Chapter Summary

Data anonymization is a common practice to protect data privacy and facilitate knowledge extraction by creating an anonymous version of the data before its release. Numerous approaches have been developed to utilize data anonymization for safeguarding privacy, adhering to stringent regulations for protecting sensitive information or personally identifiable data while preserving data utility. In this experiments, various protection configurations were employed to address the privacy challenge. Each approach has its advantages and limitations. The performance of the proposed UL approach was evaluated using real datasets and compared with similar methods in the existing literature. The results demonstrated that the UL approach in this research achieved a superior balance between data utility and privacy.

CHAPTER 5

CONCLUSION

5.1 Introduction

This study designed and carried out many experiments before arriving at the final approach proposed in this thesis (Li et al., 2016; Hasan et al., 2018; Zorarpacı and Özel, 2021). Many anonymisation approaches were studied and investigated in the first stage, leading to the design of a slicing-based enhanced approach called the Upper Lower (UL) level-based protection approach for published data. This stage has exacerbated the shortcomings of the anonymisation approaches.

In the second stage, an improved protection method was proposed called the Lower Protection Level (LPL) and Upper Protection Level (UPL) for the anonymisation approach of being more effective in determining the amount of protection required. The goal of using UPL and LPL methods is to find the particular cell's value that helps to identify disclosure and break the link between it by value swapping to guarantee a lower risk of attribute disclosure and l-diverse slicing.

The last stage of this study involved comparing the performance of the proposed approach to that of other existing works to assess its effectiveness. The evaluation of the proposed approach revealed that this method has a high capacity to preserve more data utility and provide stronger privacy protection. The previous chapters discussed the study's design, implementation, and evaluation of all contributions. Section 5.2 of this chapter summarizes all contributions, while Section 5.3 provides recommendations for future research.

5.2 Summary of Research Contributions

This research significantly contributes to the field of Privacy-Preserving Data Publishing (PPDP) by focusing on anonymizing published data while achieving an optimal balance between privacy and data utility.

Within the realm of PPDP, several sub-contributions have been made to enhance data anonymization and protect published data while retaining its utility. One of the key sub-contributions of this study is the development of the UL design, an improved slicing-based approach that effectively reduces the risk of disclosure compared to existing approaches. The proposed approach demonstrates superior performance, empowering researchers, decision-makers, and technology experts to extract valuable knowledge from published data across diverse domains, including education and healthcare.

The second objective of this study involves investigating protective methods to determine the most effective way to prevent the disclosure of private information while preserving data utility. For this purpose, the Lower Protection Level (LPL) and Upper Protection Level (UPL) methods were employed. These methods selectively swap specific attributes, as opposed to random swapping used in existing approaches, thereby breaking correlations between attribute values. This selective swapping approach significantly enhances the privacy of published data while ensuring its utility and achieving l-diversity in the published microdata table.

To validate the proposed approach, this study utilizes existing data from related works to assess its effectiveness against composition attacks. A comprehensive evaluation is conducted to compare the proposed approach's efficacy in preserving data utility and privacy with existing approaches. The experimental results indicate that the UL approach offers superior privacy protection and is capable of preserving additional data. Specifically, the UL approach achieves approximately 92.47% greater data utility than the merging approach when the percentage of swap rate Φ is 2% using LPL and 98% using UPL, with a dataset size of 4.5K for Education. Furthermore, it achieves 92.19% data utility when the swap rate Φ percentage is 5%, using LPL and 95% using UPL, with the same dataset size for Education.

Regarding privacy risks, the proposed approach potentially reduces the risk of disclosure compared to other existing works such as the hybrid (Li et al., 2016), merging (Hasan et al., 2018), $e - DP$ (Mohammed et al., 2011), probabilistic (Sattar et al., 2014), Mondrian (LeFevre et al., 2006), and composition (Baig et al., 2012) approaches. Specifically, the UL approach achieves approximately 1.5% less privacy risk than other existing works when the percentage of swap rate Φ is 5%, using LPL and 95% using UPL, with K=6, I=6, and an Education dataset size of 4.5K.

Furthermore, the UL approach consistently provides more accurate answers to aggregate queries compared to other existing works when the UL approach uses the swapping method. The experimental results demonstrate that the swapping method consistently offers more precise answers to aggregate queries.

Additionally, this work proposes a classification of protection methods based on anonymization approaches. The primary goal of this classification is to provide satisfactory accuracy in preventing attempts to recognize the record owner's identity while preserving data utility. In this study, the protection methods based on anonymization approaches are classified into grouping methods, perturbation methods, and measurement correlation (similarity) methods.

5.3 Recommendation for Future

Similar to other scholarly research, this study leaves ample room for additional work to address its limitations and expand upon its foundation. The following is a brief list of potential future directions that could enhance or expand this work:

i. The developed UL approach is versatile and can be adapted and expanded to detect various malicious attacks, such as probabilistic attacks. Additionally, it has the capability to accommodate different types of datasets that include multiple sensitive attributes (SA),

ii. The possibility of adding or replacing another new method of the protection methods to the UL approach or extending some stages of the UL approach to increase data utility and decrease risk disclosure,

iii. The Lower Protection Level (LPL) and Upper Protection Level (UPL) methods with measures of correlation (similarity) can be applied in other areas of sciences, engineering, and technology. Using this method is worth further research as well.

REFERENCES

Aggarwal, C. C., & Yu, P. S. (2008). A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In *Privacy-preserving data mining* (pp. 11–52). Springer US. https://doi.org/10.1007/978-0-387-70992-5_2

Agrawal, D., & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 247–255.

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *29*(2), 439–450. https://doi.org/10.1145/335191.335438

Agrawal, S., & Haritsa, J. R. (2005). A framework for high-accuracy privacy-preserving mining. *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference On*, 193–204.

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*, *4*(1), 694. https://doi.org/10.1186/s40064-015-1481-x

Andrew, J., Karthikeyan, J., & Jebastin, J. (2019). Privacy Preserving Big Data Publication On Cloud Using Mondrian Anonymization Techniques and Deep Neural Networks. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 722–727. https://doi.org/10.1109/ICACCS.2019.8728384

Anjum, A. (2013). *Towards Privacy-Preserving Publication of Continuous and Dynamic Data Spatial Indexing and Bucketization Approaches*. Université de Nantes.

Baig, M. M., Li, J., Liu, J., Ding, X., & Wang, H. (2012). Data Privacy against Composition Attack. In *International Conference on Database Systems for Advanced Applications* (pp. 320–334). https://doi.org/10.1007/978-3-642-29038-1_24

Banisar, D., & Davies, S. (1999). Global trends in privacy protection : An international survey of privacy, data protection, and surveillance laws and developments. *The John Marshall Journal of Computer and Information Law*, *18*(1), 1–111.

Bayardo, R.J., & Agrawal, R. (2005). Data Privacy through Optimal k-Anonymization. *21st International Conference on Data Engineering (ICDE'05)*, 217–228. https://doi.org/10.1109/ICDE.2005.42

Bayardo, Roberto J, & Agrawal, R. (2005). Data privacy through optimal k-anonymization. *Proceedings - International Conference on Data Engineering*, 217–228. https://doi.org/10.1109/ICDE.2005.42

Bertino, E., Lin, D., & Jiang, W. (2008). A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In *Privacy-preserving data mining: Models and Algorithms* (pp. 183–205). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-70992-5_8

Bhaladhare, P. R., & Jinwala, D. C. (2016). Novel approaches for privacy preserving data mining in k-anonymity model. *Journal of Information Science and Engineering*, *32*(1), 63–78.

Binjubeir, M., Ahmed, A. A., Ismail, M. A. Bin, Sadiq, A. S., & Khurram Khan, M. (2020). Comprehensive Survey on Big Data Privacy Protection. *IEEE Access*, *8*, 20067–20079. https://doi.org/10.1109/ACCESS.2019.2962368

BinJubier, M., Arfian Ismail, M., Ali Ahmed, A., & Safaa Sadiq, A. (2022). Slicing-Based Enhanced Method for Privacy-Preserving in Publishing Big Data. *Computers, Materials & Continua*, *72*(2), 3665–3686. https://doi.org/10.32604/cmc.2022.024663

Brand, R. (2002). Microdata Protection through Noise Addition. In *Inference control in statistical databases* (pp. 97–116). Springer. https://doi.org/10.1007/3-540-47804-3_8

Cavanillas, José M, Curry, E., & Wahlster, W. (2016). *New Horizons for a Data-Driven Economy* (José María Cavanillas, E. Curry, & W. Wahlster (eds.)). Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3

Chambers, E. W., De Mesmay, A., & Ophelders, T. (2018). On the complexity of optimal homotopies. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 1121–1134. https://doi.org/10.1137/1.9781611975031.73

Charu, A., & Philip, S. Y. (2008). Privacy-Preserving Data Mining. In C. C. Aggarwal & P. S. Yu (Eds.), *ASPVU, Boston* (Vol. 34). Springer US. https://doi.org/10.1007/978-0-387-70992-5

Chawla, S., Dwork, C., McSherry, F., Smith, A., & Wee, H. (2005). Toward privacy in public databases. *Theory of Cryptography Conference*, 363–385. https://link.springer.com/content/pdf/10.1007/978-3-540-30576-7_20.pdf

Chen, B.-C., Kifer, D., LeFevre, K., & Machanavajjhala, A. (2009a). Privacy-Preserving Data Publishing. *Foundations and Trends® in Databases*, *2*(1–2), 1–167. https://doi.org/10.1561/1900000008

Chen, B.-C., Kifer, D., LeFevre, K., & Machanavajjhala, A. (2009b). Privacy-Preserving Data Publishing. *Foundations and Trends® in Databases*, *2*(1–2), 1–167. https://doi.org/10.1561/1900000008

Chen, K., & Liu, L. (2005). Privacy preserving data classification with rotation perturbation. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 589–592. https://doi.org/10.1109/ICDM.2005.121

Clifton, C., Kantarcioğlu, M., Doan, A. H., Schadow, G., Vaidya, J., Elmagarmid, A., & Suciu, D. (2004). Privacy-preserving data integration and sharing. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 19–26. https://doi.org/10.1145/1008694.1008698

Conway, R., & Strip, D. (1976). Selective partial access to a database. *Proceedings of the 1976 Annual Conference, ACM 1976*, 85–89. https://doi.org/10.1145/800191.805537

Cormode, G., Procopiuc, C. M., Entong Shen, Srivastava, D., & Ting Yu. (2013). Empirical privacy and empirical utility of anonymized data. *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 77–82. https://doi.org/10.1109/ICDEW.2013.6547431

Cramir, H. (1946). Mathematical methods of statistics. *Princeton U. Press, Princeton*, 500.

Cranor, L., Rabin, T., Shmatikov, V., Vadhan, S., & Weitzner, D. (2016). Towards a Privacy Research Roadmap for the Computing Community. *ArXiv Preprint ArXiv:1604.03160*. http://arxiv.org/abs/1604.03160

Cunha, M., Mendes, R., & Vilela, J. P. (2021). A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer Science Review*, *41*, 100403. https://doi.org/10.1016/j.cosrev.2021.100403

De Montjoye, Y.-A., Radaelli, L., Singh, V. K., & others. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1), 1376. https://doi.org/10.1038/srep01376

Ding, Y., & Klein, K. (2010). Model-driven application-level encryption for the privacy of e-health data. *ARES 2010 - 5th International Conference on Availability, Reliability, and Security*, 341–346. https://doi.org/10.1109/ARES.2010.91

Domingo-Ferrer, J. (2002). *Inference Control in Statistical Databases* (J. Domingo-Ferrer (ed.); Vol. 2316). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-47804-3

Domingo-Ferrer, J., & Torra, V. (2002). Theory and practical applications for statistical agencies. In *North-Holland: Amsterdam* (pp. 113–134). North-Holland.

Du, W., & Zhan, Z. (2003). Using randomized response techniques for privacy-preserving data mining. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03*, 505. https://doi.org/10.1145/956750.956810

Dwork, C. (2006). Differential Privacy. In *Information Security and Cryptography* (pp. 1–12). https://doi.org/10.1007/11787006_1

ElGamal, T. (1985). A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 196 LNCS*(4), 10–18. https://doi.org/10.1007/3-540-39568-7_2

Eom, C. S.-H., Lee, C. C., Lee, W., & Leung, C. K. (2020). Effective privacy preserving data publishing by vectorization. *Information Sciences*, *527*, 311–328. https://doi.org/10.1016/j.ins.2019.09.035

Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2002). Privacy preserving mining of association rules. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '02*, *29*(4), 217. https://doi.org/10.1145/775047.775080

Fienberg, S. E., & McIntyre, J. (2004). Data Swapping: Variations on a Theme by Dalenius and Reiss. In *International Workshop on Privacy in Statistical Databases* (pp. 14–29). https://doi.org/10.1007/978-3-540-25955-8_2

Fuller, W. (1993). Masking procedures for microdata disclosure. *Journal of Official Statistics*, *9*(2), 383–406.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing. *ACM Computing Surveys*, *42*(4), 1–53. https://doi.org/10.1145/1749603.1749605

Fung, B. C. M., Wang, K., Fu, A. W.-C., & Yu, P. S. (2010). *Introduction to Privacy-Preserving Data Publishing* (1st (ed.)). Chapman and Hall/CRC. https://doi.org/10.1201/9781420091502

Gambs, S., Killijian, M.-O., & del Prado Cortez, M. N. (2010). Show me how you move and I will tell you who you are. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS - SPRINGL '10*, 34. https://doi.org/10.1145/1868470.1868479

Gan, W., Chun-Wei, J., Chao, H.-C., Wang, S.-L., & Philip, S. Y. (2018). Privacy preserving utility mining: A survey. *2018 IEEE International Conference on Big Data (Big Data)*, 2617–2626. https://arxiv.org/pdf/1811.07389.pdf

Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008). Composition attacks and auxiliary information in data privacy. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, 265. https://doi.org/10.1145/1401890.1401926

Gao, S., & Zhou, C. (2020). Differential privacy data publishing in the big data platform of precise poverty alleviation. *Soft Computing*, *24*(11), 8139–8147. https://doi.org/10.1007/s00500-019-04352-1

Gkoulalas-Divanis, A., & Loukides, G. (2015). *Medical Data Privacy Handbook* (A. Gkoulalas-Divanis & G. Loukides (eds.)). Springer International Publishing. https://doi.org/10.1007/978-3-319-23633-9

Han, J., Kamber, M., & Pei, J. (2012). Introduction. In *Data Mining* (3rd ed., pp. 1–38). Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hasan, A., Jiang, Q., Chen, H., & Wang, S. (2018). A New Approach to Privacy-Preserving Multiple Independent Data Publishing. *Applied Sciences*, *8*(5), 783. https://doi.org/10.3390/app8050783

Hasan, A. S. M., Jiang, Q., & Li, C. (2017). An effective grouping method for privacy-preserving bike sharing data publishing. *Future Internet*, *9*(4), 65.

Hasan, A. S. M. T., Jiang, Q., Luo, J., Li, C., & Chen, L. (2016). An effective value swapping method for privacy preserving data publishing. *Security and Communication Networks*, *9*(16), 3219–3228. https://doi.org/10.1002/sec.1527

Heldal, J., & Iancu, D.-C. (2019). *Synthetic data generation for anonymization purposes. Application on the Norwegian Survey on living conditions/EHIS*. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S1_Norway_Heldal_Iancu_AD.pdf

Huang, Z., Du, W., & Chen, B. (2005). Deriving private information from randomized data. *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, 37–48.

Jayapradha, J., Prakash, M., Alotaibi, Y., Khalaf, O. I., & Alghamdi, S. A. (2022). Heap Bucketization Anonymity—An Efficient Privacy-Preserving Data Publishing Model for Multiple Sensitive Attributes. *IEEE Access*, *10*(21), 28773–28791. https://doi.org/10.1109/ACCESS.2022.3158312

Jeba, S., BinJubier, M., Ismail, M. A., Krishnan, R., Nair, S., & Narasimhan, G. (2022). A Hybrid Protection Method to Enhance Data Utility while Preserving the Privacy of Medical Patients Data Publishing. *International Journal of Advanced Computer Science and Applications*, *13*(11).

Jeba, S., Binjubier, M., Ismail, M. A., Krishnan, R., Nair, S., and Narasimhan, G. (2022). Classifying And Evaluating Privacy-Preserving Techniques Based On Protection Methods: A Comprehensive Study. *Journal of Theoretical and Applied Information Technology*, *100*(21).

Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, *26*(189–206), 1.

Jubeir, M. Bin, Ismail, M. A., Kasim, S., Amnur, H., & others. (2020). Big Healthcare Data: Survey of Challenges and Privacy. *JOIV: International Journal on Informatics Visualization*, *4*(4), 184–190.

Kargupta, H., Datta, S., Wang, Q., & Krishnamoorthy Sivakumar. (2003). On the privacy preserving properties of random data perturbation techniques. *Third IEEE International Conference on Data Mining*, 99–106. https://doi.org/10.1109/ICDM.2003.1250908

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data* (L. Kaufman & P. J. Rousseeuw (eds.); Vol. 344). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801

Keyvanpour, M. (2011). *Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification – based Framework*. *3*(2), 862–870.

Kim, J. J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. *Proceedings of the Section on Survey Research Methods*, 303–308.

Kim, J. J., Winkler, W. E., & others. (1995). Masking microdata files. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Kohavi, R., & Becker, B. (2019). *UMI Machine Learning Repository: Adult Data Set*. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml/datasets/Adult

Kumar, S., Tiwari, P., & Zymbler, M. (2019). Internet of Things is a revolutionary approach for future technology enhancement: a review. *Journal of Big Data*, *6*(1), 1–21.

Lasko, T. A., & Vinterbo, S. A. (2010). Spectral Anonymization of Data. *IEEE Transactions on Knowledge and Data Engineering*, *22*(3), 437–446. https://doi.org/10.1109/TKDE.2009.88

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian Multidimensional K-Anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, 25–25. https://doi.org/10.1109/ICDE.2006.101

Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, & Yong Ren. (2014). Information Security in Big Data: Privacy and Data Mining. *IEEE Access*, *2*, 1149–1176. https://doi.org/10.1109/access.2014.2362522

Li, B., & He, K. (2023). Local generalization and bucketization technique for personalized privacy preservation. *Journal of King Saud University - Computer and Information Sciences*, *35*(1), 393–404. https://doi.org/10.1016/j.jksuci.2022.12.008

Li, J., Baig, M. M., Sarowar Sattar, A. H. M., Ding, X., Liu, J., & Vincent, M. W. (2016). A hybrid approach to prevent composition attacks for independent data releases. *Information Sciences*, *367–368*, 324–336. https://doi.org/10.1016/j.ins.2016.05.009

Li, N. (2007). *t -Closeness : Privacy Beyond k -Anonymity and -Diversity*. *3*, 106–115.

Li, T., Li, N., Zhang, J., and Molloy, I. (2012). Slicing: A New Approach for Privacy Preserving Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*, *24*(3), 561–574. https://doi.org/10.1109/TKDE.2010.236

Li, X., Yan, Z., & Zhang, P. (2014). A Review on Privacy-Preserving Data Mining. *2014 IEEE International Conference on Computer and Information Technology*, 769–774. https://doi.org/10.1109/CIT.2014.135

Liew, C. K., Choi, U. J., & Liew, C. J. (1985). A data distortion by probability distribution. *ACM Transactions on Database Systems*, *10*(3), 395–411. https://doi.org/10.1145/3979.4017

Lindell, Y. (2011). Secure Multiparty Computation for Privacy Preserving Data Mining. In *Encyclopedia of Data Warehousing and Mining* (Vol. 1, Issue 1, p. 5). IGI Global. https://doi.org/10.4018/9781591405573.ch189

Liu, K., Kargupta, H., & Ryan, J. (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, *18*(1), 92–106. https://doi.org/10.1109/TKDE.2006.14

Luo, Z., & Wen, C. (2014). A chaos-based multiplicative perturbation scheme for privacy preserving data mining. *2014 IEEE 5th International Conference on Software Engineering and Service Science*, 941–944. https://doi.org/10.1109/ICSESS.2014.6933720

Machanavajjhala, A, Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). L-diversity: privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, 24–24. https://doi.org/10.1109/ICDE.2006.1

Machanavajjhala, Ashwin, Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference On*, 24.

Machanavajjhala, Ashwin, Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L -diversity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3. https://doi.org/10.1145/1217299.1217302

Majeed, A., & Hwang, S. O. (2023). A Generic Approach towards Enhancing Utility and Privacy in Person-Specific Data Publishing Based on Attribute Usefulness and Uncertainty. *Electronics*, *12*(9), 1978. https://doi.org/10.3390/electronics12091978

Majeed, A., & Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, *9*, 8512–8545. https://doi.org/10.1109/ACCESS.2020.3045700

Malin, B., & Sweeney, L. (2004). How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, *37*(3), 179–192. https://doi.org/10.1016/j.jbi.2004.04.005

Maniam, J., & Singh, D. (2020). Towards Data Privacy and Security Framework In Big Data Governanc. *International Journal of Software Engineering and Computer Systems*, *6*(1), 41–51. https://doi.org/10.15282/ijsecs.6.1.2020.5.0068

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1–29. https://doi.org/10.1214/11-SS074

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. *IEEE Access*, *4*(January), 1821–1834. https://doi.org/10.1109/ACCESS.2016.2558446

Mendes, R., & Vilela, J. P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access*, *5*, 10562–10582. https://doi.org/10.1109/ACCESS.2017.2706947

Mészárosová, E. (2015). Is Python an Appropriate Programming Language for Teaching Programming in Secondary Schools? *International Journal of Information and Communication Technologies in Education*, *4*(2), 5–14.

Mivule, K. (2013). Utilizing noise addition for data privacy, an overview. *Proceedings of The International Conference on Information and Knowledge Engineering (IKE 2012)*, 65–71.

Mohammed, N., Chen, R., Fung, B. C. M., & Yu, P. S. (2011). Differentially private data release for data mining. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, 493. https://doi.org/10.1145/2020408.2020487

Murthy, S., Abu Bakar, A., Abdul Rahim, F., & Ramli, R. (2019). A Comparative Study of Data Anonymization Techniques. *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, 306–309. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00063

Narayanan, A. (2009). *Data privacy: The non-interactive setting* [The University of Texas Austin]. https://repositories.lib.utexas.edu/bitstream/handle/2152/18424/narayanana61136.pdf?sequence=2&isAllowed=y

Nasiri, N., & Keyvanpour, M. (2020). Classification and Evaluation of Privacy Preserving Data Mining Methods. *2020 11th International Conference on Information and Knowledge Technology (IKT)*, 17–22. https://doi.org/10.1109/IKT51791.2020.9345620

Ninghui, L., Tiancheng, L., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. *Proceedings - International Conference on Data Engineering*, 106–115. https://doi.org/10.1109/ICDE.2007.367856

Olatunji, I. E., Rauch, J., Katzensteiner, M., & Khosla, M. (2022). A Review of Anonymization for Healthcare Data. *Big Data*. https://doi.org/10.1089/big.2021.0169

Patel, A., Dodiya, K., & Pate, S. (2013). A Survey On Geometric Data Perturbation In Multiplicative Data Perturbation. *International Journal*, *1*(5).

Pawar, A., Ahirrao, S., & Churi, P. P. (2018). Anonymization Techniques for Protecting Privacy: A Survey. *2018 IEEE Punecon*, 1–6. https://doi.org/10.1109/PUNECON.2018.8745425

Priyadarsini, R. P., Sivakumari, S., & Amudha, P. (2016). *Digital Connectivity – Social Impact* (S. Subramanian, R. Nadarajan, S. Rao, & S. Sheen (eds.); Vol. 679). Springer Singapore. https://doi.org/10.1007/978-981-10-3274-5

Reiss, S. P. (1984). Practical data-swapping: the first steps. *ACM Transactions on Database Systems*, *9*(1), 20–37. https://doi.org/10.1145/348.349

Reiss, S. P., Post, M. J., & Dalenius, T. (1982). Non-reversible privacy transformations. *Proceedings of the 1st ACM SIGACT-SIGMOD Symposium on Principles of Database Systems - PODS '82*, 139. https://doi.org/10.1145/588111.588134

Rohilla, S. (2015). Privacy Preserving Data Publishing through Slicing. *American Journal of Networks and Communications*, *4*(3), 45. https://doi.org/10.11648/j.ajnc.s.2015040301.18

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

Saita, C.-A., & Llirbat, F. (2004). Clustering Multidimensional Extended Objects to Speed Up Execution of Spatial Queries. In *International Conference on Extending Database Technology* (pp. 403–421). https://doi.org/10.1007/978-3-540-24741-8_24

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information (abstract). *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems - PODS '98*, *98*, 188. https://doi.org/10.1145/275487.275508

Sarathy, R., & Muralidhar, K. (2011). Evaluating Laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Priv.*, *4*(1), 1–17.

Sattar, A. H. M. S., Li, J., Liu, J., Heatherly, R., & Malin, B. (2014). A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. *Knowledge-Based Systems*, *67*, 361–372. https://doi.org/10.1016/j.knosys.2014.04.019

Senosi, A., & Sibiya, G. (2017). Classification and evaluation of Privacy Preserving Data Mining: A review. *2017 IEEE AFRICON*, 849–855. https://doi.org/10.1109/AFRCON.2017.8095593

Shah, A., & Gulati, R. (2016a). Privacy Preserving Data Mining: Techniques, Classification and Implications-A Survey. *International Journal of Computer Applications*, *137*(12).

Shah, A., & Gulati, R. (2016b). Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey. *International Journal of Computer Applications*, *137*(12), 40–46. https://doi.org/10.5120/ijca2016909006

Sharma, Amita. (2017). Literature Survey of Privacy Preserving Data Publishing (PPDP) Techniques. *International Journal Of Engineering And Computer Science*, *6*(5), 1–12. https://doi.org/10.18535/ijecs/v6i4.12

Sharma, Anil, Singh, G., & Rehman, S. (2020). A Review of Big Data Challenges and Preserving Privacy in Big Data. In *Advances in Data and Information Sciences* (pp. 57–65). Springer Nature Switzerland. https://doi.org/10.1007/978-981-15-0694-9_7

Siddique, M., Mirza, M. A., Ahmad, M., Chaudhry, J., & Islam, R. (2018). A survey of big data security solutions in healthcare. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *255*, 391–406. https://doi.org/10.1007/978-3-030-01704-0_21

Sweeney, L. (2002a). Achieving K -anonymity Privacy Protection Using Generalization and Suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 1–18.

Sweeney, L. (2002b). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Sweeny, L. (2002). k- ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Puzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Taylor, L., Zhou, X.-H., & Rise, P. (2018). A tutorial in assessing disclosure risk in microdata. *Statistics in Medicine*, *37*(25), 3693–3706. https://doi.org/10.1002/sim.7667

Vaghashia, H., & Ganatra, A. (2015). A Survey: Privacy Preservation Techniques in Data Mining. *International Journal of Computer Applications*, *119*(4), 20–26. https://doi.org/10.5120/21056-3704

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, *33*(1), 50–57. https://doi.org/10.1145/974121.974131

Wang, T., Zheng, Z., Rehmani, M. H., Yao, S., & Huo, Z. (2019). Privacy Preservation in Big Data From the Communication Perspective—A Survey. *IEEE Communications Surveys & Tutorials*, *21*(1), 753–778. https://doi.org/10.1109/COMST.2018.2865107

Wen, S., Wu, W., & Castiglione, A. (2017). *Cyberspace Safety and Security* (S. Wen, W. Wu, & A. Castiglione (eds.); Vol. 10581). Springer International Publishing. https://doi.org/10.1007/978-3-319-69471-9

Wong, R. C.-W., & Fu, A. W.-C. (2010). Privacy-Preserving Data Publishing: An Overview. *Synthesis Lectures on Data Management*, *2*(1), 1–138. https://doi.org/10.2200/S00237ED1V01Y201003DTM002

Wong, R. C.-W., Fu, A. W.-C., Liu, J., Wang, K., & Xu, Y. (2010). Global privacy guarantee in serial data publishing. *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 956–959. https://doi.org/10.1109/ICDE.2010.5447859

Xiao, X., & Tao, Y. (2006). Anatomy: Simple and effective privacy preservation. *Proceedings of the 32nd International Conference on Very Large Data Bases*, 139–150. https://www.vldb.org/conf/2006/p139-xiao.pdf

Yang, Z., Zhong, S., & Wright, R. N. (2005). Privacy-Preserving Classification of Customer Data without Loss of Accuracy. *Proceedings of the 2005 SIAM International Conference on Data Mining*, 92–102. https://doi.org/10.1137/1.9781611972757.9

Yin, C., Xi, J., Sun, R., & Wang, J. (2017). Location privacy protection based on differential privacy strategy for big data in industrial internet of things. *IEEE Transactions on Industrial Informatics*, *14*(8), 3628–3636.

Yu, J., Kuang, Z., Zhang, B., Zhang, W., Lin, D., & Fan, J. (2018). Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. *IEEE Transactions on Information Forensics and Security*, *13*(5), 1317–1332.

Yu, J., Zhang, B., Kuang, Z., Lin, D., & Fan, J. (2016). iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security*, *12*(5), 1005–1016.

Yu, S. (2016). Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access*, *4*, 2751–2763. https://doi.org/10.1109/ACCESS.2016.2577036

Yu, S., & Member, S. (2016). Big Privacy : Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access*, *4*, 2751–2763.

Zhang, N. (2006). *Privacy-preserving data mining* (Vol. 69, Issue 01) [Texas A&M University]. http://oaktrust.library.tamu.edu/bitstream/handle/1969.1/ETD-TAMU-1080/ZHANG-DISSERTATION.pdf

Zhang, N., & Zhao, W. (2007). Privacy-Preserving Data Mining Systems. *Computer*, *40*(4), 52–58. https://doi.org/10.1109/MC.2007.142

Zhang, Q., Koudas, N., Srivastava, D., & Yu, T. (2007). Aggregate query answering on anonymized tables. *2007 IEEE 23rd International Conference on Data Engineering*, 116–125.

Zigomitros, A., Casino, F., Solanas, A., & Patsakis, C. (2020). A Survey on Privacy Properties for Data Publishing of Relational Data. *IEEE Access*, *8*, 51071–51099. https://doi.org/10.1109/ACCESS.2020.2980235

Zorarpacı, E., & Özel, S. A. (2021). Privacy preserving classification over differentially private data. *WIREs Data Mining and Knowledge Discovery*, *11*(3), e1399. https://doi.org/10.1002/widm.1399