CLOUD BASED PRIVACY PRESERVING DATA MINING MODEL USING HYBRID K-ANONYMITY AND PARTIAL HOMOMORPHIC ENCRYPTION

HUDA OSMAN MANSOUR OSMAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JUNE 2022

DEDICATION

This thesis is dedicated to my lovely husband Mohammed Shokai for his patience, continuous support, and abundant generosity.

To my dear father Osman, my beloved mother (may Allah have mercy on her) Thuraya, my second granule mother Najwa and my dears' brothers, Montaser and Ahmed; to their encouragement, support, and prayers for me.

To my dears and lovely children Mohja and Mojtaba for their patience and motivation for me.

To my beloved country "Sudan".

ACKNOWLEDGEMENT

So many thanks to Allah S.W.T for granting me the patient, success, and strength I needed to finish this thesis. I am grateful to any researcher, academician, and friend who helps me during preparing this thesis.

I am very thankful to my main supervisor Dr. Maheyzah Md Siraj for her guidance, advice, and motivation. Without her continued support and interest, this thesis would not have been the same as presented here.

I wish to express my sincere appreciation to my previous main supervisor Professor Dr. Mohd. Aizaini Maarof for encouragement, guidance. I also thank my co-supervisor Dr. Fuad Abdulgaleel Abdoh Ghaleb to advice and support.

I am indebted to Universiti Teknologi Malaysia - Malaysia, University of Kassala - Sudan, and Ministry of Higher Education - Sudan for funding my study.

ABSTRACT

The evolution of information and communication technologies have encourage numerous organizations to outsource their business and data to cloud computing to perform data mining and other data processing operations. Despite the great benefits of the cloud, it has a real problem in the security and privacy of data. Many studies explained that attackers often reveal the information from third-party services or third-party clouds. When a data owners outsource their data to the cloud, especially the SaaS cloud model, it is difficult to preserve the confidentiality and integrity of the data. Privacy-Preserving Data Mining (PPDM) aims to accomplish data mining operations while protecting the owner's data from violation. The current models of PPDM have some limitations. That is, they suffer from data disclosure caused by identity and attributes disclosure where some private information is revealed which causes the success of different types of attacks. Besides, existing solutions have poor data utility and high computational performance overhead. Therefore, this research aims to design and develop Hybrid Anonymization Cryptography PPDM (HAC-PPDM) model to improve the privacy-preserving level by reducing data disclosure before outsourcing data for mining over the cloud while maintaining data utility. The proposed HAC-PPDM model is further aimed reducing the computational performance overhead to improve efficiency. The Quasi-Identifiers Recognition algorithm (QIR) is defined and designed depending on attributes classification and Quasi-Identifiers dimension determine to overcome the identity disclosure caused by Quasi-Identifiers linking to reduce privacy leakage. An Enhanced Homomorphic Scheme is designed based on hybridizing Cloud-RSA encryption scheme, Extended Euclidean algorithm (EE), Fast Modular Exponentiation algorithm (FME), and Chinese Remainder Theorem (CRT) to minimize the computational time complexity while reducing the attribute disclosure. The proposed QIR, Enhanced Homomorphic Scheme and k-anonymity privacy model have been hybridized to obtain optimal data privacy-preservation before outsourced it on the cloud while maintaining the utility of data that meets the needs of mining with good efficiency. Real-world datasets have been used to evaluate the proposed algorithms and model. The experimental results show that the proposed QIR algorithm improved the data privacy-preserving percentage by 23% while maintaining the same or slightly better data utility. Meanwhile, the proposed Enhanced Homomorphic Scheme is more efficient comparing to the related works in terms of time complexity as represented by Big O notation. Moreover, it reduced the computational time of the encryption, decryption, and key generation time. Finally, the proposed HAC-PPDM model successfully reduced the data disclosures and improved the privacy-preserving level while preserved the data utility as it reduced the information loss. In short, it achieved improvement of privacy preserving and data mining (classification) accuracy by 7.59 % and 0.11 % respectively.

ABSTRAK

Evolusi teknologi maklumat dan komunikasi telah mendorong banyak organisasi menggunakan sumber luar untuk perniagaan dan data mereka atas komputeran awan bagi melaksanakan perlombongan dan operasi pemprosesan data. Walaupun komputeran awan terdapat banyak kelebihan, ia mempunyai masalah dari segi keselamatan dan privasi data. Banyak kajian menjelaskan bahawa penyerang sering mendedahkan maklumat dari perkhidmatan atau komputeran awan pihak ketiga. Apabila pemilik data menyimpan data mereka atas awan, terutama model awan SaaS, sukar untuk menjaga kerahsiaan dan integriti data. Pelombongan Data Pemelihara Privasi (PPDM) bertujuan untuk menyelesaikan operasi perlombongan data sambil melindungi data pemilik dari pencerobohan. Model PPDM terdahulu mempunyai beberapa kelemahan. Antaranya pendedahan sebahagian maklumat peribadi yang mengundang kejayaan pelbagai jenis serangan. Selain itu, mempunyai utiliti data yang teruk dan masalah prestasi pengiraan yang tinggi. Oleh yang demikian, penyelidikan ini bertujuan untuk merekabentuk dan membangunkan model Hibrid Anonimisasi Kriptografi PPDM (HAC-PPDM) untuk meminimumkan kelemahan tersebut dari segi meningkatkan tahap pemeliharaan privasi sebelum penyumberan luar data untuk penambangan melalui awan sambil mengekalkan utiliti data. Model HAC-PPDM bertujuan untuk mengurangkan masalah prestasi pengiraan bagi meningkatkan kecekapan. Algoritma Quasi-Identifiers Recognition (QIR) ditakrifkan dan direka bentuk bergantung pada klasifikasi atribut dan dimensi Quasi-Identifiers menentukan untuk mengatasi pendedahan identiti yang disebabkan oleh Quasi-Identifiers memaut untuk mengurangkan kebocoran privasi. Skim Homomorfik Dipertingkat direka bentuk berdasarkan penghibridan skim penyulitan Cloud-RSA, algoritma Euclidean Lanjutan (EE), algoritma Eksponensiasi Modular Pantas (FME) dan Teorem Baki Cina (CRT) untuk meminimumkan kerumitan masa pengiraan sambil mengurangkan pendedahan atribut. QIR yang dicadangkan, Skim Homomorphic Dipertingkat dan model privasi k-tanpa nama telah dihibridkan untuk mendapatkan pemeliharaan privasi data yang optimum sebelum menyumber luarnya pada awan sambil mengekalkan utiliti data yang memenuhi keperluan perlombongan dengan kecekapan yang baik. Set data dunia nyata telah digunakan untuk menilai algoritma dan model yang dicadangkan. Hasil eksperimen menunjukkan bahawa algoritma QIR yang dicadangkan menaikkan peratusan pemeliharaan privasi data sebanyak 23% sambil mengekalkan utiliti data yang sama atau sedikit lebih baik. Sementara itu, Skema Homomorfik yang disempurnakan adalah lebih cekap berbanding penyelidikan terdahulu dari segi kompleksiti masa yang diwakilkan dengan notasi Big O. Tambahan pula, ia mengurangkan masa pengiraan bagi penyulitan, penyahsulitan dan masa penjanaan kekunci. Akhir sekali, pengurangan masa pengiraan untuk masa keseluruhan , lebih-lebih lagi, model HAC-PPDM yang dicadangkan berjaya mengurangkan pendedahan data dan meningkatkan tahap pemeliharaan privasi sambil mengekalkan utiliti data kerana ia dapat mengurangkan kehilangan maklumat. Secara ringkas, ia mencapai peningkatan dari segi pemeliharaan privasi dan ketepatan pelombongan data (klasifikasi) masing-masing sebanyak 7.59 % dan 0.11 %.

.

TABLE OF CONTENTS

TITLE **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CRT	-	Chinese Remainder Theorem
DM	-	Data Mining
EC	-	Equivalent Class
EEA	-	Extended Euclidean Algorithm
EHS	-	Enhanced Homomorphic Scheme
EIs	-	Explicit Identifiers
FHE	-	Fully Homomorphic Encryption
FME	-	Fast Modular Exponential
GI	-	Generalization intensity
HAC-PPDM	-	Hybrid Anonymization Cryptography for Privacy Preserving Data Mining
HE	-	HE Homomorphic Encryption
IaaS	-	Infrastructure as a Service
MIT	-	Massachusetts Institute of Technology
NSs	-	Non-Sensitive Attributes
PaaS	-	Platform as a Service
PG	-	Privacy Gain
PHE	-	Partial Homomorphic Encryption
PP	-	Privacy Preserving
PPDM	-	Privacy Preserving Data Mining
QIDs	-	Quasi-Identifiers
QIR	-	Quasi-Identifiers Recognition Algorithm
SaaS	-	Software as a Service
SAs	-	Sensitive Attributes
SMC	-	Secure Multi-party Computation
SQI	-	Selective Quasi-Identifiers

LIST OF SYMBOLS

k	-	Value for k-anonymity
l	-	Value for l-diversity
α	-	SAs threshold of QIR
β	-	QIDs threshold of QIR
µ	-	NSs threshold of QIR
t	-	Value for t-closeness

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

In the last few years, many fields of knowledge have turned to cloud computing, to perform data mining and other operations. Outsourcing data to perform mining operations is very useful for data owners who do not have adequate computing resources, or do not have sufficient experience to apply data mining techniques (Wang et al., 2018). In general, data mining is the process of discovering interesting patterns and knowledge within large amount of data stored in databases, data warehouses or other information repositories. Patterns and knowledge learned from data mining, is useful, especially in the business for prediction or decision-making process.

In recent times, cloud computing has become a dominant technology in most fields of studies. According to a latest cloud report, 94% of major companies use at least one cloud service (Flexera, 2021). Moreover, the cloud plays a big role in improving and developing smart cities. Cloud computing has revolutionized companies and organizations regarding their data-processing mechanism, especially in methods of data storage, access, and processing, including data mining and analysis (Samanthula et al., 2015). Extra advancements in cloud computing support scalable information technology services, which are characterized by customized price model. Multi-tenant feature of cloud computing environment act as attractive portal for academicians, as it is a convenient way for a user to share data and collaborate with other users. Despite various advantages of cloud computing it faced real problems of security and privacy of data (Alenizi et al., 2021; Dagher et al., 2019).

Data owners get worried to outsource their business over cloud network, despite its great benefits. Concerns related to privacy in cloud computing emerged from diverse factors, for example, loss of control, multi-tenancy, wide distribution, and lack of trust. When owner's data is released into cloud network, the owner might lose the control to manage these data. In addition, services providers who analyse the data can misuse these data, or can disclosure to beneficiaries, due to financial motives. This creates a major challenge facing data privacy. Modern direction of data analysis has been based, mainly on statistical analysis of data. Data mining relates significantly to members of this class (Aldeen et al., 2016; Zigomitros et al., 2020).

Privacy preservation of data mining is a significant issue over the cloud. Therefore, accomplishing data mining objectives without sacrificing individual's privacy, is not only important, but is compulsory to the success of data mining. The privacy preserving data mining (PPDM) approaches aim to extract useful knowledge from huge data volumes, simultaneously preserving privacy and utility of the data. Hence, maintaining data privacy in the cloud has become one of the most important issues in recent years (Wang et al., 2018).

PPDM aims to protect the privacy of individual data or sensitive knowledge, without losing the utility of the data. The basic idea of PPDM is to modify the data in such a way as to perform data mining algorithms effectively, without compromising the confidentiality of sensitive information contained in the data. Private and sensitive data of individuals must be protected and maintained before being outsourced to cloud. Privacy preservation is regarded as a major pre-requisite to perform data mining operations over cloud. Challenges facing privacy of sensitive information over cloud computing are fast growing. In various applications of cloud, the data sets frequently grow and/or alter over time; thus the requirement for effective techniques to preserve the privacy of data regularly (Aldeen et al., 2016; Puri et al., 2019).

Outsourced data, frequently includes private and sensitive data about persons, usually outsourced through non-government agencies or/and government institutions. The private and sensitive information has to do, significantly, with resource for medical research, further research, direction analysis, and public funds allocation for these non-government agencies or/and government institutions (Domingo-Ferrer et al., 2019).

1.2 Problem Background

Generally, there are three major issues which are required to be addressed in PPDM over cloud computing. First, the privacy of owner's data, which is outsourced to cloud for mining, should be protected from the leakage. Outsourced data may contain private and sensitive data; like business financial records or banking datasets, patients' illnesses or symptoms in medical datasets, and similar ones (Zhang, et al., 2018). There are many ways and opportunities to misuse sensitive data when exposed to the public (Zhang et al., 2018), especially since the cloud still suffers from fundamental privacy issues (Alenizi et al., 2021). With regards to this, there are three causes that can lead to privacy leakage and violation: (Abdelhameed et al., 2018; Fung et al., 2010)

i. Attribute disclosure: ability to infer sensitive/private information of some individuals from released dataset. Main cause of attribute disclosure is values homogeneity in which the values of the sensitive attributes (SAs) in the one equivalence class are similar (Abdelhameed et al., 2018; Aldeen & Salleh, 2019b).

ii. Identity disclosure: ability to match pair of records in two separate tables with the assist of some attribute's values (quasi-identifiers attributes), which can lead to identity conformity of an individual's private information. From the key reasons of identity disclosure is linking of quasi-identifiers attributes (QIDs) that resulted from not identifying the significant QIDs accurately (Yan et al., 2018).

iii. Membership disclosure: Ability to know whether a victim's record exist in outsourced dataset or not.

Second issue in PPDM over cloud is utility of outsourced data. Accomplishing successful privacy preservation on outsourced data requires changing of data values by PPDM methods, which may negatively affect utility of the data. In PPDM it is so important to outsource dataset with great utility (Henriksen-Bulmer & Jeary, 2016; Sudhakar & Rao, 2020). Utility here refers to the data remaining truthful, accurate and not containing any big loss of information. If a poor utility dataset is outsourced, it makes it difficult to use the results of the data mining, because false-positive and false-negative results may be obtained (Lee et al., 2017). This issue is affected by the following parameters:

i. Accuracy: This determines the proximity of the sanitized value to the primary value (Agrawal et al., 2008).

ii. Completeness or data loss: This investigates the degree of the missed data within the sanitized database (Lee et al., 2017; Agrawal et al., 2008).

iii. Truthfulness: means that every sanitized record corresponds to a single primary record (Lee et al., 2017).

iv. Consistency: This is related to all internal constraints, i.e., the relationship present within the different fields of the data item or amongst various data items in the database (Agrawal et al., 2008).

Third issue is complexity of PPDM models that developed for privacy preserving of outsourced dataset. In the cloud and its applications, the data is growing and changes a lot; massive volume of data is added within short periods of time in addition to continuous modification of the data in the cloud. These reasons make the fulfilment of privacy standards and conditions to increase computation time and negatively affect performance (Reddy at el., 2018). Cryptography-based methods are often used to improve data privacy and data utility. However, these methods require high computational overhead.

The current PPDM models over cloud can classified into three wide groups. The first group is anonymization-based techniques; for example, K-anonymity, L-diversity, T- closeness. The second group is cryptography-based techniques; for example, homomorphic encryption and oblivious-transfer. The third group is hybrid-based techniques, in which more than one anonymization model is combined, or anonymization model are combined with other methods/models. Detailed information about these methods is discussed in Chapter 2.

i. Anonymization-based Models

The anonymization-based models are ranked first and outperform the rest of the PPDM solutions in protecting privacy disclosures. This method, being more practical, has several algorithms for implementation (Aldeen & Mazleena, 2018). The anonymization-based models concealing the identity of the individual and/or the sensitive data by applying some operations (e.g., suppression, generalization, and perturbation).

Some solutions offered based on anonymization include, improved K-anonymity to protect attribute disclosure (Zhang at el., 2016), heuristic indistinguishable group anonymization (HIGA) scheme to prevent identity disclosure (Brown, 2017), and employing suppression and splitting operations to protect privacy disclosures (Terrovitis at el., 2017). Sei et al. (2019) proposed a new privacy model dependent on l-diversity and t-closeness, with a method that addressed sensitive Quasi-Identifiers (QIDs). Victor & Lopez (2020) proposed an approach for sensitive outsourced data using graph theoretic algorithms based on k-anonymity.

Some examples of anonymization-based models developed to improve data utility are two top-down anonymization algorithms to preserve data utility where threat to information loss exists, developed by Gong at el., (2017). Aldeen & Salleh, (2019a) hybrid K-anonymity and data relocation algorithm to improve utility in terms of truthfulness and data loss. Lee et al., (2017) proposed a method based on restriction of generalization to improve accuracy and degrade data loss. Venkata et al., (2020) present an efficient index based quasi-identifier strategy to ensure privacy preservation and achieve high data utility over incremental and distributed data sets.

The current solutions based on anonymization have been unsuccessful in reaching optimal privacy preservation in term of protecting attribute and identity disclosures together. Hence, there still revealing some private/sensitive information lead to the success of several types of PPDM attacks (Abdelhameed et al., 2018; Agarwal & Sachdeva, 2018; Domingo-Ferrer et al., 2019a). Besides, The anonymization-based models suffer from big data loss, if compared with cryptography-based methods, due to operations of generalization and suppression (Abdelhameed et al., 2018; Zigomitros et al., 2020).

ii. Cryptography-based Models

The solutions based on cryptography encrypt the data to preserve data privacy and confidentiality it offer the best level of protection for privacy disclosures and data utility (Taric & Poovammal, 2017). Examples of some methods and protocols based on cryptography address the privacy leakage issue include encoding of attributes with the random key to each participating to prevent attribute disclosure, by Sharma and Shukla (2017). Another one was, privacy-preserving method for data mining classifier using homomorphic encryption for smart city applications, proposed by Amma and Dhanaseelan (2018). Chandravathi and Lakshmi (2019) proposed a technique to improve security of the standard RSA, based on using Extended Euclidean Algorithm (EEA) in key generation, increasing the complexity in private key. Furthermore, El Makkaoui et al. (2019) proposed encryption scheme for preserving data confidentiality in the cloud, based on RSA cryptosystem for accomplishing privacy-preserving, while reducing the computational time complexity. Shukla et al. (2020) present a novel encryption method for systems based on cloud computing. An example of a study that used cryptography to preserve data utility was by Li et al. (2017), who provided a Cryptographic Data Publishing System (CDPS).

The cryptography-based models provide optimal level of privacy preserving and data utility (Taric & Poovammal, 2017). However, they have low efficiency due to high performance overhead (Zhang et al., 2018; Zigomitros et al., 2020).

iii. Hybrid-based Models

The hybrid-based Models hybrid more than one anonymization model, more the one encryption scheme, or anonymization model are combined with encryption scheme/s. Recently some PPDM solutions use hybridization for a achieve higher level of privacy preserving for example Yang et al. (2015) combine between cryptography, statistical analysis, and anonymization. Li et al. (2016a) hybridize homomorphic encryption scheme and a secure comparison scheme. Aldeen and Salleh, (2019a) merge K-anonymity with data relocation method. Aldeen and Salleh, (2019b) hybridize K-anonymity, L-diversity, and (a, k)-anonymity. Most of the current hybrid-based models inherit the weakness of anonymization & cryptography.

According to the weakness in the current PPDM models and solutions the research addresses attribute disclosure and identity disclosure together to minimizing privacy leakage and improve privacy preservation of outsourced over cloud. The research also focusses to reduce the data loss that faced the current PPDM models for maintain the data utility. Besides, it aimed to reduce the computational time consumed to improve the efficiency.

The above three major issues of PPDM over cloud computing can summaries in Figure 1.1 along with the problems causing them, in addition to solutions often used to overcome each issue/problem. The shaded boxes in the Figure 1.1 illustrate the problems in each issue which will be addressed by this research. The research focuses identity and attribute disclosure to reduce privacy leakage because identity disclosure is one of the serious forms of confidentiality violation (Zhang & Nayak, 2020). Minimizing identity disclosure alone does not protect privacy, ensuring real anonymity protection requires addressing identity disclosure and attribute disclosure (Omer & Mohamad, 2016). Completeness or data loss is from main the issues of data utility that can cover most quality of anonymized data (Lee et al., 2017), therefore, the research focused on it mainly to maintain the data utility. Great efficiency of PPDM model make it more practical for privacy preserving of the outsourced data.

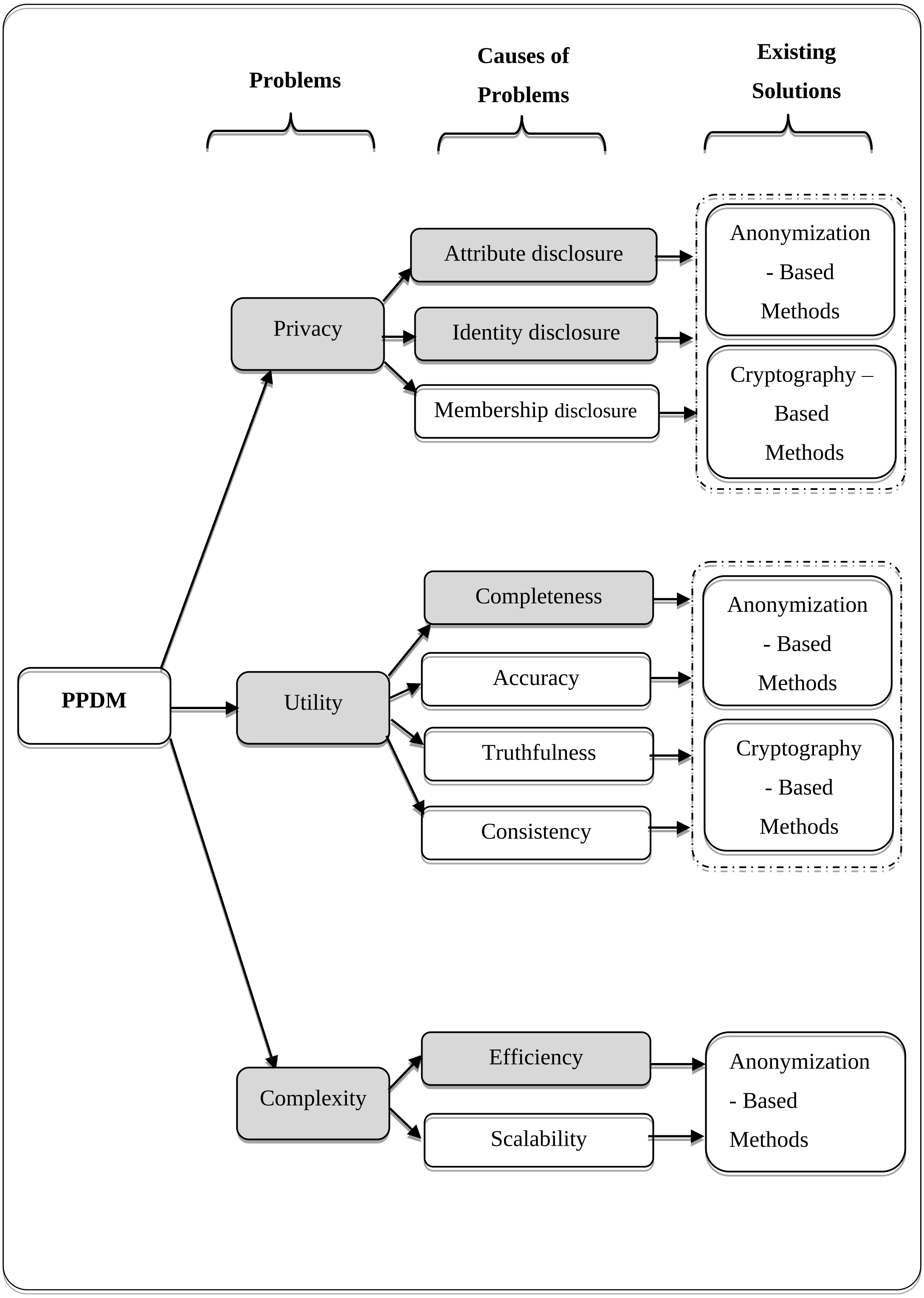

Figure 1.1 The research problem background

Most of the existing PPDM models have utility problem; when the privacy-preserving is improved the utility of the data go down (see Figure 1.2) especially with the anonymization-based models (Abdelhameed et al., 2018; Nayahi & Kavitha, 2017; Fovino & Masera, 2016). This is due to anonymization operations like generalization and suppression. However, the anonymization-based techniques are most popularly used among researchers because of its simplicity and ease of implementation.

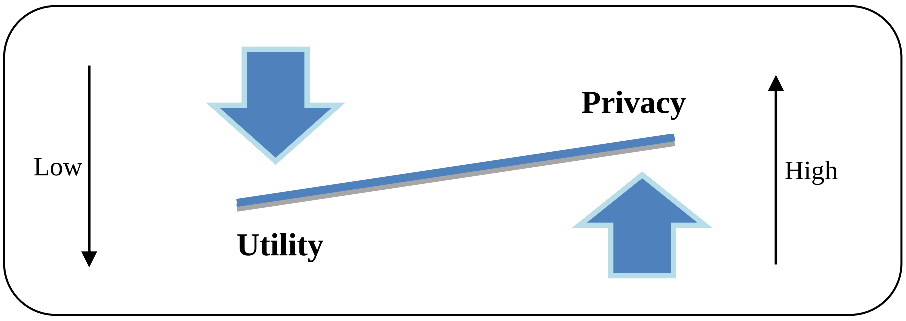

Figure 1.2 Privacy versus data utility

Cryptography-based models provided good utility and good privacy-preserving at same time but achieved low efficiency in terms of computational time complexity, compared to anonymization-based models because of encryption and decryption operations.

1.3 Problem Statement

The current PPDM models over the cloud have data privacy leakage resulted from identity disclosure and attribute disclosure. The identity disclosure can be caused by QID linking because of not identifying the significant QIDs precisely, while the attribute disclosure can be caused by values homogeneity. Modern methods of privacy-preserving outsourced data, seek to prevent identity and attribute disclosures that lead to privacy leakage and then its subsequent violation. However, the recent solutions still face some problems that lead to failure in achieving optimal privacy-preserving for outsourced data, as the data is still vulnerable to breach by some types of attacks. Therefore, designing an enhanced PPDM model to prevent attribute and identity disclosures is significant to keeping data private and confidential.

Most of the techniques and models currently used to maintain privacy-preserving such as models based on anonymization, modify the data by performing some operations to meet privacy requirements. The data modification, in turn, leads to loss of data, which negatively affects the general utility of the data. Data utility is important to execute the mining operations or further analysis processes in the future. Low data utility may give false and unhelpful mining results. Therefore, the utility of data must be considered when designing a privacy-preserving model.

Furthermore, the cryptography-based models have low efficiency due to high computational complexity. The low efficiency of a privacy-preservation model makes it less practical for use in privacy-preserving of the outsourced data. This is especially true since the outsourcing process of the data assumes more computational complexity in processing in the cloud and the burden of connection.

1.4 Research Questions

The main research question:

How to improve the privacy-preserving of outsourced data for mining over the cloud while maintaining data utility with the best efficiency?

The support research questions are:

i. How to identify the significant QIDs accurately, to reduce QIDs linking?

ii. How to prevent the values homogeneity to reduce attribute disclosure while minimizing the computational complexity to improve the efficiency?

iii. How to reduce the data disclosures to reduce the privacy leakage before outsourcing data to the cloud while maintaining the data utility?

The research questions can be solved by this hypothesis:

The privacy-preservation of the outsourced data can be improved by reducing the privacy leakage caused by identity and attribute disclosures (data disclosure). Identifying the significant QIDs accurately can reduce the QIDs linking, thus, reduce the identity disclosure. While the attribute disclosure can reduce by preventing the values homogeneity. Also, reducing the loss of data, which its privacy is preserved, maintains the utility of the data. The efficiency can be improved by reducing the computational time consumed.

1.5 Research Aim

The research aims to design an enhanced PPDM model over the cloud for improving the privacy-preserving of outsourced data by reducing privacy leakage while maintaining the efficiency of data utility.

1.6 Research Objectives

The objectives of this research that lead to achieving the research aim are:

i. To design Quasi-Identifiers Recognition (QIR) algorithm based on re-identification of risks for identifying significant QID attributes to reduce QIDs linking.

ii. To enhance the homomorphic scheme based on partial homomorphic encryption to prevent the values homogeneity and reduce attribute disclosure while minimizing the computational time complexity to improve the efficiency.

iii. To hybridize the K-anonymity model, QIR algorithm, and enhanced homomorphic scheme for reducing data disclosures to improving privacy-preserving while maintaining data utility.

1.7 Research Scopes

This study rests on the following scopes and limitations:

i. To evaluate and validate the introduced model, two datasets have been used; the first one is a dataset of bank direct marketing (2014) while the second is for adult census (1996 – updated 2016). Two real datasets were from the machine learning repository in University of California, Irvine, which were widely used by other researchers in PPDM studies. For example, bank direct marketing was used by Abdul et al. (2016), Aldeen et al. (2016), Aldeen & Salleh (2019a, 2019b) and Yousra & Mazleena (2018). Adult dataset was used by Dagher et al. (2019), Gong et al. (2017), Kaur & Agrawal (2019), Lee et al. (2017), Nayahi & Kavitha (2017), Prasser et al. (2020), Reddy et al. (201), Sei et al. (2019) and Simi et al. (2017).

ii. This research was focused on privacy-preserving of the dataset before it is outsourced to mining over the cloud. The reason being that, it is most comprehensive to limit violation of privacy in the first place, before starting mining process, and reduces re-identification attacks (Henriksen-Bulmer & Jeary, 2016).

iii. The maintaining of data utility was relied on by calculating the data loss resulting from generalization use, that has been done using generalization intensity measure.

iv. The data mining technique that was used to verify the correctness and accuracy of the data, which privacy was preserved by the proposed model, was the classification.

1.8 Significance of the Research

The research is significant to the field of PPDM over cloud, and its applications, due to the following motivations:

i. The process of outsourcing data on the cloud to implement mining operations is greatly beneficial, and gets a lot of interests, especially from data owners who do not have sufficient experience with data mining techniques, or do not have sufficient resources.

ii. The privacy concerns of cloud computing mainly motivated this research. Another motivating factor related to the different types of attack that aim to breach and violate data privacy. Such attacks hinder exploitation of the amazing benefits of the cloud in data mining operations. As such, it required effective models to maintain the privacy of data before outsourcing it to the cloud.

iii. Some of the main reasons for the leakage of data privacy are identity and attribute disclosures. Reducing these disclosures helps reduce privacy leakage and achieves a higher level of privacy preservation of outsourced data.

iv. Utility of data is significant to the mining process, most of the operations that are carried out on data to meet the requirements of privacy, negatively affect the utility. This needs to be considered when designing privacy-preservation models.

v. Low computational complexity helps improve the efficiency of privacy preservation models. High performance of privacy preservation processes is important, especially since the outsourcing data process requires additional complexity in data processing on the cloud, and in communication.

1.9 Definition of Terms

i. Privacy-preserving

The ability to prevent information from being disclosed to unauthorized entities using, mechanisms and methods that limit the leakage of data privacy, and thus prevent its violation (Agrawal et al., 2008; Fung et al., 2010).

ii. Data Utility

It is from the main issues related to data quality, where it based on the context of the data usage. It is evaluated by data loss, accuracy, truthfulness, and consistency (Agrawal et al., 2008; Lee et al., 2017).

iii. Complexity

It measures the scalability and efficiency of a specific PPDM method, where efficiency indicates execution of algorithm with optimal performance. It is usually estimated by the amount of space and time consumed, and scalability represents the efficiency directions of algorithm if the data size is increased (Agrawal et al., 2008).

iv. Data outsourcing

It is a data model in which the data owner authorizes other parties to manage and process the data (Carminati, 2009).

v. Quasi-Identifiers

The attributes which can identify an individual’s identities through linkages between attributes, like gender, age, ZIP, etc. (Zarezadeh et al., 2020).

vi. Anonymization

Anonymization is often used to conceal identity of data owners and/or sensitive information. It makes the data or individual's information vague and unknown, so as to maintain privacy using one or more of data sanitizing operations; generalization, suppression, anatomization, perturbation (Mendes & Ao, 2017; Reddy et al., 2018).

vii. Cryptography

Using the encryption techniques to preserve data privacy and confidentiality while preserving the utility (Mendes & Ao, 2017).

1.10 Thesis Outline

This thesis is constituted of 7 chapters organized as shown in Figure 1.5. This Chapter 1 is an introduction of the whole research. Chapter 2 surveys the research area of PPDM over cloud computing, through a review of some basic concepts, definitions, and current PPDM solutions. Chapter 3 presents the research methodology, while Chapter 4 describes design of the Quasi-Identifiers Recognition algorithm. Chapter 5 explains the design of an Enhanced Homomorphic Scheme. Chapter 6 demonstrates the design and development of the Hybrid Anonymization Cryptography PPDM model, and finally, Chapter 7 concludes the thesis, also giving some future work suggestions.

REFERENCES

Abdelhameed, S. A., Moussa, S. M., & Khalifa, M. E. (2018). Privacy-preserving tabular data publishing: A comprehensive evaluation from web to cloud. *Computers and Security*, *72*, 74–95.

Abdelhameed, S. A., Moussa, S. M., & Khalifa, M. E. (2019). Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing. *Knowledge-Based Systems*, *164*, 1–20.

Abdul, Y., Aldeen, A. S., Salleh, M., & Razzaque, M. A. (2016). *A Technique of Data Privacy Preservation in Deploying Third Party Mining Tools over the Cloud Using SVD and LSA*. *11 (2)*, 27–34.

Agarwal, S., & Sachdeva, S. (2018). An Enhanced Method for Privacy-Preserving Data Publishing. *Studies in Computational Intelligence*, *713*, 61–75.

Agrawal, R., Srikant, R., Agrawal, R., & Srikant, R. (2008). Privacy-preserving data mining Models and Algorithms. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 29(2), 439–450.

Ahmad, I., & Khandekar, A. (2014). Homomorphic Encryption Method Applied to Cloud Computing. *International Journal of Information & Computation Technology*, *4*(15), 1519–1530.

Al-Hamami, A. H., & Aldariseh, I. A. (2012). Enhanced method for RSA cryptosystem algorithm. *Proceedings of International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2012*, 402–408.

Al-Lawati, A., & Al-Badi, A. H. (2016). The impact of cloud computing IT departments: A case study of Oman’s financial institutions. *Proceeding of 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016*, 62–71.

Al-sit, W. T., Al-zoubi, H., & Al-Jubouri, Q. (2019). Cloud Security based on the Homomorphic Encryption. *International Journal of Advanced Computer Science and Applications*, *10*(8), 300–307.

Alabdulkarim, A., Al-Rodhaan, M., Tian, Y., & Al-Dhelaan, A. (2019). A privacy-preserving algorithm for clinical decision-support systems using random forest. *Computers, Materials and Continua*, *58*(2), 585–601.

Aldeen, Y. A. A. S., & Salleh, M. (2019a). Privacy Preserving Data Utility Mining Architecture. In *Smart Cities Cybersecurity and Privacy*, 253–268. Elsevier Inc.

Aldeen, Y. A. A. S., & Salleh, M. (2019b). Techniques for Privacy Preserving Data Publication in the Cloud for Smart City Applications. In *Smart Cities Cybersecurity and Privacy*, 129–145. Elsevier Inc.

Aldeen, Y. A. A. S., Salleh, M., & Aljeroudi, Y. (2016). An innovative privacy preserving technique for incremental datasets on cloud computing. *Journal of Biomedical Informatics*, *62*, 107–116.

Aldeen Yousra, S., & Mazleena, S. (2018). A New Heuristic Anonymization Technique for Privacy Preserved Datasets Publication on Cloud Computing. *Journal of Physics: Conference Series*, *1003*(1), 1–15.

Alenizi, B. A., Humayun, M., & Jhanjhi, N. (2021). Security and Privacy Issues in Cloud Computing. *Journal of Physics: Conference Series*, *1979*(1), 1–12.

Ali, K., Akhtar, F., Memon, S. A., Shakeel, A., Ali, A., & Raheem, A. (2020). Performance of Cryptographic Algorithms based on Time Complexity. *Proceedings of 3rd International Conference on Computing, Mathematics and Engineering Technologie, ICoMET 2020*, 4–8.

Alwatban, I. S., & Emam, A. Z. (2014). Comprehensive Survey on Privacy Preserving Association Rule Mining: Models, Approaches, Techniques and Algorithms. *International Journal on Artificial Intelligence Tools*, *23*(05), 1–28.

Babu, K. S. (2013). *Utility-Based Privacy Preserving Data Publishing*. National Institute of Technology Rourkela, India.

Bampoulidis, A., Markopoulos, I., & Lupu, M. (2019). PrioPrivacy: A local recoding K-anonymity tool for prioritised Qasi-identifiers. *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence Workshops, WI 2019 Companion*, 314–317.

Carminati, B. (2009). Secure Data Outsourcing. In *Encyclopedia of Database Systems*, 61–66. Springer, Boston, MA.

Barua, H. B., & Mondal, K. C. (2019). A comprehensive survey on cloud data mining (CDM) frameworks and algorithms. *ACM Computing Surveys*, *52*(5), 1–62.

Bayardo, R. J. (2005). Data Privacy Through Optimal k-Anonymization. *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)*, 217–228.

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, *17*(2), 169–177.

Brown, E. E. (2017). Improving privacy preserving methods to enhance data mining for correlation research. *Proceedings of IEEE SoutheastCon 2017*, 1–4.

Chandrakar, I., & Hulipalled, V. R. (2020). Articulation point based quasi identifier detection for privacy preserving in distributed environment. *International Journal of Communication Networks and Information Security*, *12*(1), 77–82.

Chandravathi, D., & Lakshmi, P. V. (2019). Privacy preserving using extended euclidean algorithm applied to RSA-homomorphic encryption technique. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 3175–3179.

Chen B, Cheung P, Cheung P, & Kwok Y. (2015). Cypherdb: A novel architecture for outsourcing secure database processing. *IEEE Transactions on Cloud Computing*, *6*(2), 372–386.

Çiğşar, B., & Ünal, D. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Scientific Programming*, *2019*, 1–8.

Contel Bradford. (2020). *7 Most Infamous Cloud Security Breaches.* Available at: https://blog.storagecraft.com/7-infamous-cloud-security-breaches/ (Accessed: 18 October 2020).

Dagher, G. G., Benjamin, ·, Fung, C. M., Mohammed, N., & Clark, J. (2019). SecDM: privacy-preserving data outsourcing framework with differential privacy. *Knowledge and Information Systems*, *62*, 1923–1960.

Dasgupta, D., Akhtar, Z., & Sen, S. (2020). Machine learning in cybersecurity: a comprehensive survey. *Journal of Defense Modeling and Simulation*, 19(1), 57-106.

Deshmukh, M., Tijare, P., & Sawalkar, S. (2016). A Survey on Privacy Preserving Data Mining Techniques for Clinical Decision Support System. *International Research Journal of Engineering and Technology*, *3*(5), 6054–6056.

Domingo-Ferrer, J., Farràs, O., Ribes-González, J., & Sánchez, D. (2019). Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges. In *Computer Communications*,140, 38–60. Elsevier B.V.

Domingo-Ferrer, J., & Soria-Comas, J. (2015). From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, *74*, 151–158.

El Makkaoui, K., Beni-Hssane, A., & Ezzati, A. (2019). Speedy Cloud-RSA homomorphic scheme for preserving data confidentiality in cloud computing. *Journal of Ambient Intelligence and Humanized Computing*, *10*(12), 4629–4640.

El Makkaoui, K., Beni-Hssane, A., Ezzati, A., & El-Ansari, A. (2017). Fast Cloud-RSA Scheme for Promoting Data Confidentiality in the Cloud Computing. *Procedia Computer Science*, *113*, 33–40.

Fovino, I. N., & Masera, M. (2016). Privacy Preserving Data Mining—"A State of the Art." *In 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2108–2112.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, *42*(4), 1–53.

Gan, W., Lin, J. C. W., Chao, H. C., & Zhan, J. (2017). Data mining in distributed environment: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), 1–19.

Gentry, C. (2009). *A fully Homomorphic Encryption Scheme*. PhD Thesis, Stanford University, Stanford, California.

Gong, Q., Yang, M., Chen, Z., Wu, W., & Luo, J. (2017). A framework for utility enhanced incomplete microdata anonymization. *Cluster Computing*, *20*(2), 1749–1764.

Guo, N., Yang, M., Gong, Q., Chen, Z., & Luo, J. (2019). Data anonymization based on natural equivalent class. *Proceedings of the IEEE 23rd International Conference on Computer Supported Cooperative Work in Design, CSCWD 2019*, 22–27.

He, Z., Cai, Z., Sun, Y., Li, Y., & Cheng, X. (2017). Customized privacy preserving for inherent data and latent data. *Personal and Ubiquitous Computing*, *21*(1), 43–54.

Henriksen-Bulmer, J., & Jeary, S. (2016). Re-identification attacks—A systematic literature review. *International Journal of Information Management*, *36*(6), 1184–1192.

Islam, M. A., Islam, M. A., Islam, N., & Shabnam, B. (2018). A Modified and Secured RSA Public Key Cryptosystem Based on "n" Prime Numbers. *Journal* *of Computer and Communications*, *06*(03), 78–90.

Jain, P., Gyanchandani, M., & Khare, N. (2019). Improved k-anonymity privacy-preserving algorithm using Madhya Pradesh State election commission big data. In *Studies in Computational Intelligence*, 771, 1–10. Springer Verlag.

K.Saranya, K.Premalatha, S. R. (2015). A Survey on Privacy Preserving Data Mining. *Proceedings of the 2nd International Conference on Electronics and Communication Systems (ICECS)*, 1740–1744.

Kaaniche, N., & Laurent, M. (2017). Data security and privacy preservation in cloud storage environments based on cryptographic mechanisms. *Computer Communications*, *111*, 120–141.

Kabir, E., Mahmood, A. N., Wang, H., & Mustafa, A. K. (2020). Microaggregation Sorting Framework for K-Anonymity Statistical Disclosure Control in Cloud Computing. *IEEE Transactions on Cloud Computing*, *8*(2), 408–417.

Kaklamanis, M. M., & Filippakis, M. (2019). A comparative survey of machine learning classification algorithms for breast cancer detection. *Proceedings of the 23rd Pan-Hellenic Conference on Informatics*, 97–103.

Kaur, A., & Sofat, S. (2016). A proposed hybrid approach for Privacy Preserving Data Mining. *Proceeding of International Conference on Inventive Computation Technologies (ICICT)*, 1–6.

Kaur, G., & Agrawal, S. (2019). Differential Privacy Framework : Impact of Quasi-identifiers on Anonymization. *Proceedings of 2nd International Conference on Communication, Computing and Networking*, 35–42.

Khan, A. (2020). *Fast modular exponentiation*. Available at: https://www.khanacademy.org/computing/computer-science/cryptography/modarithmetic/a/fast-modular-exponentiation (Accessed: 29 December 2020).

Kim, S., Sung, M. K., & Chung, Y. D. (2014). A framework to preserve the privacy of electronic health data streams. *Journal of Biomedical Informatics*, *50*, 95–106.

Kohlmayer, F., Prasser, F., Eckert, C., & Kuhn, K. A. (2014). A flexible approach to distributed data anonymization. *Journal of Biomedical Informatics*, *50*, 62–76.

Kumar, A., & Tiwari, N. (2012). Effective implementation and avalanche effect of AES. *International Journal of Security, Privacy and Trust Management (IJSPTM)*, *1*(3), 31–35.

Kumar, R., Pattnaik, P. K., & Sharma, Y. (2016). Privacy Preservation in Distributed Environment Using RSA-CRT. *Proceedings of the Second International Conference on Computer and Communication Technologies*, 29–34.

Kumaraswamy, S., Manjula, H., & Venugopal, R. (2017). Secure cloud based privacy preserving dataminning platform. *Indonesian Journal of Electrical Engineering and Computer Science*, *7*(3), 830–838.

Le, J., Liao, X., & Yang, B. (2017). Full autonomy: A novel individualized anonymity model for privacy preserving. *Computers and Security*, *66*, 204–217.

Lee, H., Kim, S., Kim, J. W., & Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC Medical Informatics and Decision Making*, *17*(1), 1–12.

Lee, Y., & Lee, K. (2017). Re-identification of medical records by optimum quasi-identifiers. *Proceeding of 19th International Conference on Advanced Communication Technology*, 428–435.

Li, H., Yu, J., Zhang, H., Yang, M., & Wang, H. (2020). Privacy-Preserving and Distributed Algorithms for Modular Exponentiation in IoT with Edge Computing Assistance. *IEEE Internet of Things Journal*, *7*(9), 8769–8779.

Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016a). Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1547–1861.

Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016b). Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1547–1861.

Li, P., Li, J., Huang, Z., Gao, C. Z., Chen, W. Bin, & Chen, K. (2017). Privacy-preserving outsourced classification in cloud computing. *Cluster Computing*, *21*(1), 277–286.

Li, T., Liu, Z., Li, J., Jia, C., & Li, K. C. (2017). CDPS: A cryptographic data publishing system. *Journal of Computer and System Sciences*, *89*, 80–91.

Lin, koo N. (2017). *Privacy preserving data mining based on k-anonymity l-diversity and SVM*. MSc Thesis, Universiti Teknologi Malaysia, Malaysia.

Liu, X., Deng, R. H., Yang, Y., Tran, H. N., & Zhong, S. (2018). Hybrid privacy-preserving clinical decision support system in fog–cloud computing. *Future Generation Computer Systems*, *78*, 825–837.

Machanavajjhala, A., Gehrke, J., & Kifer, D. (2006). ℓ -Diversity : Privacy Beyond k-anonymity. *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, 1–12.

Makkaoui, K. El, Ezzati, A., & Beni-hssane, A. (2017). Cloud-RSA: An Enhanced Homomorphic Encryption Scheme. *Advances in Intelligent Systems and Computing*, *520*, 471–480.

Maqsood, F., Ali Shah, M., Ahmed, M., & Mumtaz Ali, M. (2017). Cryptography: A Comparative Analysis for Modern Techniques. *International Journal of Advanced Computer Science and Applications*, *8*(6), 442–448.

Mehmood, R., & Selwal, A. (2020). Recent Innovations in Computing. *Proceedings of ICRIC 2020*, 455–467.

Mendes, R., & Ao, J. O. (2017). Privacy-Preserving Data Mining : Methods , Metrics , and Applications. *IEEE Access*, *5*, 10562–10582.

Mortazavi, R., & Jalili, S. (2015). Preference-based anonymization of numerical datasets by multi-objective microaggregation. *Information Fusion*, *25*, 85–104.

Nageswari Amma, N. G., & Ramesh Dhanaseelan, F. (2018). Privacy Preserving Data Mining Classifier for Smart City Applications. *Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES 2018*, *Icces*, 645–648.

Nayahi, J. J. V., & Kavitha, V. (2017). Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop. *Future Generation Computer Systems*, *74*, 393–408.

Okeyinka, A. E. (2015). Computational Speeds Analysis of RSA and ElGamal Algorithms on Text Data. *Proceedings of the World Congress on Engineering and Computer Science*, 21–24.

Omer, A. M., & Mohamad, M. M. Bin. (2016). Simple and effective method for selecting quasi-identifier. *Journal of Theoretical and Applied Information Technology*, *89*(2), 512–517.

Omer, M. Z., Gao, H., & Mustafa, N. (2017). Privacy-preserving of SVM over vertically partitioned with imputing missing data. *Distributed and Parallel Databases*, *35*, 363–382.

Otgonbayar, A., Pervez, Z., Dahal, K., & Eager, S. (2018). K-VARP: K-anonymity for varied data streams via partitioning. *Information Sciences*, *467*, 238–255.

Parmar, K., & Shah, V. (2016). A Review on Data Anonymization in Privacy Preserving Data Mining. *International Journal of Advanced Research in*

Computer and Communication Engineering, *5*(2), 75–79.

Petre, R.-Ş. (2012). Data mining in Cloud Computing. In *Database Systems Journal*, *3*(3), 67–71.

Prasser, F., Bild, R., & Kuhn, K. A. (2016). A Generic method for assessing the quality of De-Identified health data. *Studies in Health Technology and Informatics*, *228*, 312–316.

Prasser, F., Kuhn, K. A., & Eicher, J. (2020). Flexible data anonymization using ARX — Current status and challenges ahead. *Software: Practice and Experience*, *50*(7), 1277–1304.

Puri, V., Sachdeva, S., & Kaur, P. (2019). Privacy preserving publication of relational and transaction data: Survey on the anonymization of patient data. *Computer Science Review*, *32*, 45–61.

Rebollo-Monedero, D., Forné, J., Soriano, M., & Puiggalí Allepuz, J. (2017). p-Probabilistic k-anonymous microaggregation for the anonymization of surveys with uncertain participation. *Information Sciences*, *382*, 388–414.

Reddy, S. R. ., Raju, K. V. S. V. ., & Valli Kumari, V. (2018). Personalized privacy preserving incremental data dissemination through optimal generalization. *Journal of Engineering and Applied Sciences*, *13*(11), 4205–4216.

Rivest, R., Adleman, L., & Dertouzos, L. (1978). On data banks and privacy homomorphisms. *Foundations of Secure Computation*, *4*(11), 169–180.

Kohavi, R., & Becker, B. (2016). *Adult Census Income*. Available at:
https://www.kaggle.com/uciml/adult-census-income

S. Moro, P. Cortez, & P. Rita. (2014). *UCI Machine Learning Repository: Bank Marketing Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

S.Aldeen, Y. A. A. (2016). *Heuristic Based Privacy Preservation Technique for Data Publication in Cloud Computing Environment*. PhD Thesis, Universiti Teknologi Malaysia, Malaysia.

S.Aldeen, Y. A. A., & Salleh, M. (2016). A Hybrid K-anonymity Data Relocation Technique for Privacy Preserved Data Mining in Cloud Computing. *Journal of Internet Computing and Services*, *17*(5), 51–58.

Samanthula, B. K., Elmehdwi, Y., & Jiang, W. (2015). k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data. *IEEE Transactions on Knowledge and Data Engineering*, *27*(5), 1261–1273.

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. *PODS*, *98*(188), 10–1145.

Sei, Y., Okumura, H., Takenouchi, T., & Ohsuga, A. (2019). Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness. *IEEE Transactions on Dependable and Secure Computing*, *16*(4), 580–593.

Selvaraj, B., & Periyasamy, S. (2016). A review of recent advances in Privacy preservation in health care data publishing. *International Journal of Pharma and Bio Sciences*, *7*(4), 33–41.

Shah, A., & Gulati, R. (2016). Privacy Preserving Data Mining: Techniques, Classification and Implications -A Survey. *International Journal of Computer Applications*, *137*(12), 975–8887.

Sharma, S., & Shukla, D. (2017). Efficient multi-party privacy preserving data mining for vertically partitioned data. *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*, 1–7.

Shukla, D., Dwivedi, V., & Munesh, T. (2021). Encryption algorithm in cloud computing. *Materials Today: Proceedings*, *37*, 1869–1875.

Shyma Mogtaba, & Kambal, E. (2016). Association Rule Hiding for Privacy Preserving Data Mining. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6171*, 505–517.

Simi, M. S., Nayaki, K. S., & Elayidom, M. S. (2017). An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity. *IOP Conference Series: Materials Science and Engineering*, 1–9.

Singh, S., Singh, P., Garg, R., & Mishra, P. K. (2016). Mining association rules in various computing environments: A survey. *International Journal of Applied Engineering Research*, *11*(8), 1–12.

Song, F., Ma, T., Tian, Y., & Al-Rodhaan, M. (2019). A New Method of Privacy Protection: Random k-Anonymous. *IEEE Access*, *7*, 75434–75445.

Srijayanthi, S., Sethukarasi, T., & Thilagavathy, A. (2019). Efficient anonymization algorithm for multiple sensitive attributes. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 4961–4963.

Sudhakar, R. V., & Rao, T. C. M. (2020). Security aware index based quasi–identifier approach for privacy preservation of data sets for cloud applications. *Cluster Computing*, *23*(4), 2579–2589.

Suresh, S., Huang, H., & Kim, H. J. (2015). Scheduling in compute cloud with multiple data banks using divisible load paradigm. *IEEE Transactions on Aerospace and Electronic Systems*, *51*(2), 1288–1297.

Sweeny, L. (2002). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Puzziness and Knowledge-Based Systems*, *10*(5), 557–570.

Taric, G. J., & Poovammal, E. (2017). A Survey on privacy preserving data mining Techniques. *IOSR Journal of Computer Engineering*, *17*(5), 2278–2661.

Telikani, A., & Shahbahrami, A. (2017). Optimizing association rule hiding using combination of border and heuristic approaches. *Applied Intelligence*, *47*, 544–557.

Uttarwar, N., & Pradhan, M. A. (2017). K-NN data classification technique using semantic search on encrypted relational database. *Proceedings of 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*, 1–6.

Verma, R., & Sharma, A. K. (2020). Cryptography : Avalanche effect of AES and RSA. *International Journal of Scientific and Research Publications*, *10*(4), 119–125.

Victor, N., & Lopez, D. (2020). Privacy preserving sensitive data publishing using (k,n,m) anonymity approach. *Journal of Communications Software and Systems*, *16*(1), 46–56.

Wang, B., Zhan, Y., & Zhang, Z. (2018). Cryptanalysis of a Symmetric Fully Homomorphic Encryption Scheme. *IEEE Transactions on Information Forensics and Security*, *13*(6), 1460–1467.

Wang, K., Zhao, W., Cui, J., Cui, Y., & Hu, J. (2019). A K-anonymous clustering algorithm based on the analytic hierarchy process. *Journal of Visual Communication and Image Representation*, *59*, 76–83.

Wang, R., Zhu, Y., Chen, T.-S., & Chang, C.-C. (2018a). An Authentication Method Based on the Turtle Shell Algorithm for Privacy-Preserving Data Mining. *The Computer Journal*, *61*(8), 1123–1132.

Wang, R., Zhu, Y., Chen, T.-S., & Chang, C.-C. (2018b). Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness. *Journal of Computer Science and Technology*, *33*(6), 1231–1242.

Wang, W., Chen, L., & Zhang, Q. (2015). Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation. *Computer Networks*, *88*, 136–148.

Wei, D., Natesan Ramamurthy, K., & Varshney, K. R. (2018). Distribution-preserving k-anonymity. *Statistical Analysis and Data Mining*, *11*(6), 253–270.

Widodo, Budiardjo, E. K., Wibowo, W. C., & Achsan, H. T. Y. (2019). An Approach for Distributing Sensitive Values in k-Anonymity. *International Workshop on Big Data and Information Security, IWBIS 2019*, 109–114.

Wong, K. S., Tu, N. A., Bui, D. M., Ooi, S. Y., & Kim, M. H. (2019). Privacy-Preserving Collaborative Data Anonymization with Sensitive Quasi-Identifiers. *Poceeding of 12th CMI Conference on Cybersecurity and Privacy, CMI 2019*, 1–6.

Wu, W., Parampalli, U., Liu, J., & Xian, M. (2019). Privacy preserving k-nearest neighbor classification over encrypted database in outsourced cloud environments. *World Wide Web*, *22*(1), 101–123.

Xun, T. Z. (2017). *Comparative Study on Randomization Techniques in Privacy Preserve Data Mining on Diabetes Dataset*. MSc Thesis, Universiti Teknologi Malaysia, Malaysia.

Yan Y, Wang W, Hao X, Z. L. (2018). Finding quasi-identifiers for k-anonymity model by the set of cut-vertex. *Engineering Letters*, *26*(1), 1–11.

Yang, G., Ye, X., Fang, X., Wu, R., & Wang, L. (2020). Associated attribute-aware differentially private data publishing via microaggregation. *IEEE Access*, *8*, 79158–79168.

Yang, J.-J., Li, J.-Q., & Niu, Y. (2015). A hybrid solution for privacy preserving medical data sharing in the cloud environment. *Future Generation Computer Systems*, *43–44*, 74–86.

Yang, P. A. N., Xiong, N. N., Member, S., & Ren, J. (2020). Data Security and Privacy Protection for Cloud Storage : A Survey. *IEEE Access*, *8*, 131723–131740.

Yi, X., Paulet, R., & Bertino, E. (2014). Homomorphic Encryption. In *Homomorphic Encryption and Applications.* 27–46. Springer, Cham.

Zarezadeh, M., Mala, H., & Khajeh, H. (2020). Preserving Privacy of Software-Defined Networking Policies by Secure Multi-Party Computation. *Journal of Computer Science and Technology*, *35*(4), 863–874.

Zhang, H., Zhou, Z., Ye, L., & Du, X. (2018). Towards privacy preserving publishing of set-valued data on hybrid cloud. *IEEE Transactions on Cloud Computing*, *6*(2), 316–329.

Zhang, L., Xuan, J., Si, R., & Wang, R. (2017). An Improved Algorithm of Individuation K-Anonymity for Multiple Sensitive Attributes. *Wireless Personal Communications*, *95*(3), 2003–2020.

Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., & Chen, J. (2014). A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. *Journal of Computer and System Sciences*, *80*(5), 1008–1020.

Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., & Chen, J. (2013). Combining Top-Down and Bottom-Up: Scalable Sub-tree Anonymization over Big Data Using MapReduce on Cloud. *Proceeding of 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 501–508.

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.

APPENDIX A List of Related Publications

LIST OF PUBLICATIONS

Journal with Impact Factor

i. **Osman, H**., Siraj, M., Ghaleb, F., Saeed, F., & Alkhammash, E., Maarof M. (2021). Quasi-Identifiers Recognition Algorithm for Privacy Preservation of Cloud Data Based on Risk Re-Identification. *Wireless Communications and Mobile Computing, 1*, 1–23. (**Indexed by WOS and Scopus Impact Factor= 2.336, Q3).**

Indexed Conference Proceedings

ii. **Osman H**., Maarof, M., Siraj, M. (2020). Hybrid Solution for Privacy-Preserving Data Mining on the Cloud Computing. *Proceeding of International Conference of Reliable Information and Communication Technology*. 748-758. Springer, Cham. **(Indexed by WOS and SCOPUS).**

iii. **Osman**, **H**., Siraj, M., & Maarof, M. (2021). HAC: Model for Privacy-Preserving Outsourced Data Over Cloud. *Proceeding of 3rd International Cyber Resilience Conference (CRC)*, 1-4 **(Indexed by SCOPUS)**.