HYBRID ANONYMIZATION TECHNIQUE FOR IMPROVING THE PRIVACY IN NETWORK DATA

YAHAYA ABD RAHIM

DOCTOR OF PHILOSOPHY

2016

HYBRID ANONYMIZATION TECHNIQUE FOR IMPROVING THE PRIVACY IN NETWORK DATA

YAHAYA ABD RAHIM

A thesis submitted

in fulfillment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of Doctor of Philosophy.

Signature	:	……..
Supervisor Name	:	PROF.DATUK DR SHAHRIN BIN SAHIB
Date	:	……..……

DECLARATION

I declare that this thesis entitled "Hybrid Anonymization Technique for Improving the Privacy in Network Data" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ……..

Name : YAHAYA ABD RAHIM

Date : 1 AUGUST 2016

DEDICATION

First and foremost, I would like to thank ALLAH Almighty, for giving me excellence health, ideas and comfort environment so that I can complete this thesis as scheduled.

My greatest thank is to my wife (Siti Nadiah Bt Sh Mohammad Mustafa), my father (Abd Rahim Bin Mohammed), my mum (Esah Bt Abd Rahman), my late son (Mohamad Alif Bin Yahaya) and my siblings (Zulkiflee, Jamaluddin, Jaffri, Hairulnizam and Kamal) for their continuous understanding, motivation, encouragement and patience throughout my PhD journey.

I also dedicate this PhD thesis to my many friends who have supported me throughout the process. I will always appreciate all they have done for helping me to complete my thesis and develop my computing skills in algorithms and data processing.

ABSTRACT

There has been a considerable research over the last decades on methods for limiting disclosure in data publishing, especially for the last twenty years in the computer science field. Researchers have studied the problems of publishing microdata or network data without revealing any sensitive information that may have cause the paradigm preservation of information privacy. There are organizations that would like to publish their data for research, advertisement or prediction purposes. Nevertheless, they had the problems in information loss and lack of privacy. Hence, there are a few techniques and research that have been in highlights like the K-anonymity, l-diversity, generalization, clustering and randomization techniques, but most of these techniques is not comprehensive and the chances to lose the information is still high and may cause privacy leakage on the original data. The contribution of this research is the hybrid technique in anonymization process that will improve the protection and the privacy of data. With this better and comprehensive solution, it will decrease the loss of information. There are four major phases in this methodology as research guidance. The first phase is an overview of the entire research process and the second phase is the description of the anonymization process and techniques. It will be followed by the third phase of describing the design and module of the system, and the fourth phase is the researcher highlights on the comparison methods that are designed in this study. The researcher stated that there are two main contributions in this research. The first contribution is to introduce a new technique to anonymize the network data using the hybrid technique; and for the second contribution, the researcher creates a profile of a hybrid anonymization technique based on K-anonymity, l-diversity, generalization, clustering and randomization techniques. It is quite difficult to identify the best technique of anonymization process. Due to this, the researcher provides the details of analyzing, summarizing and profiling of the anonymization techniques. The researcher realizes that there are a few opportunities to advance this research within this domain in the near future, such as implementing a real-time based in anonymization process. Unfortunately, this type of processing needs to be revamped from the architectural design until the data processing part; and it is more thought-provoking if it were implemented in a real-time based or in the batch processing process, if the variable of the optimization is to be used in the anonymization process. Apart from that, the profiling of the anonymization processing techniques will also help the researcher to propose a generalization technique that might be implemented to anonymize data either using the micro or the network data.

ABSTRAK

Telah maklum bahawa dari dulu hinggalah ke akhir dekad ini, telah terdapat banyak penyelidikan yang agak cemerlang pada kaedah untuk menghadkan pendedahan dalam penerbitan data, terutamanya dalam bidang sains komputer. Didalam kajian ini, penyelidik telah mengkaji masalah microdata penerbitan atau data rangkaian tanpa mendedahkan apa-apa maklumat sensitif yang berkaitan dengan paradigma privasi maklumat. Terdapat banyak organisasi yang ingin menerbitkan data mereka untuk tujuan penyelidikan, pengiklanan atau ramalan; walau bagaimanapun, mereka mempunyai masalah dalam kehilangan maklumat dan kewajaran privasi. Oleh itu, terdapat beberapa teknik dan penyelidikan yang telah di bangunkan seperti k-anonymization, l-diversity, generalisasi, perkelompokan maupun teknik rawak, akan tetapi kebanyakan teknik-teknik tersebut tidak melakukan proses ini secara menyeluruh dan ianya masih lagi terdapat masalah untuk kehilangan maklumat dengan kadar yang masih tinggi dan boleh menyebabkan kebocoran privasi pada data asal. Atas keperihatinan ini, sumbangan penyelidikan ini adalah menerbitkan teknik hibrid dalam proses anonymization yang akan meningkatkan perlindungan dan privasi data. Dengan kata lain, penyelesaian ini lebih baik dan menyeluruh, dan ianya akan mengurangkan kehilangan maklumat. Perlaksanaan penyelesaian ini, akan melalui empat fasa utama untuk menjana kaedah yang mana ianya dapat digunakan sebagai panduan penyelidikan. Fasa pertama adalah gambaran keseluruhan proses penyelidikan secara keseluruhannya dan fasa kedua pula adalah memperihalkan proses anonymization dan tekniknya. Ianya akan disusuli dengan fasa ketiga di mana ianya menerangkan reka bentuk dan modul sistem dalam teknik anonymization tersebut, dan fasa keempat pula adalah acara kemuncak di mana penyelidik akan memperihalkan perbandingan yang direka dalam kajian ini. Penyelidik menyatakan bahawa terdapat dua sumbangan utama dalam kajian ini. Sumbangan pertama adalah untuk memperkenalkan teknik baru untuk anonymize data rangkaian iaitu menggunakan teknik hibrid; manakala sumbangan kedua pula, penyelidik mencipta profil teknik hibrid anonymization berdasarkan kepada pengubahansuai teknik teknik yang sediaada seperti k-anonymization, l-diversity, generalisasi, pengkelompokan dan teknik rawak. Namun demikian, ianya agak sukar untuk mengenal pasti teknik terbaik dalam proses anonymization tersebut. Oleh yang demikian, penyelidik telah menyediakan butiran penganalisaan, ringkasan dan profil teknik anonymization, dan di samping itu, penyelidik juga menyedari bahawa terdapat beberapa peluang untuk memajukan penyelidikan ini dalam domain yang sama untuk masa yang akan datang, seperti melaksanakan proses anonymization dalam masa nyata. Namun demikian, pemprosesan jenis ini memerlukan rombakan yang besar dari peringkat reka bentuk senibina sehinggalah ke bahagian pemprosesan data; dan ianya akan menjadi lebih provokasi sekiranya ianya dapat dilaksanakan dalam masa nyata atau dalam pemprosesan secara berkelompok dengan memasukkan elemen atau pembolehubah pengoptimuman untuk digunakan dalam proses anonymization ini. Selain itu, teknik pemprosesan anonymization secara pemprofilan juga akan membantu penyelidik untuk mencadangkan satu teknik generalisasi yang mungkin dapat dilaksanakan untuk proses anonymize data samada data dalam bentuk mikro atau data rangkaian.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank ALLAH Almighty, for giving me excellence health, ideas and comfort environment so that I can complete this thesis was scheduled.

I would like to express sincere appreciation to Prof Datuk Dr. Haji Shahrin bin Sahib@Sahibuddin for his excellent guidance, supervision, motivation, encouragement, patience and insight throughout the years of this PhD endless journey.

I would like to extend my thanks to the staff of FTMK, FKEKK and PPS for their time, guidance and support during my studies. Greatest appreciation goes to MHE and UTeM for their sponsorship with allowance during this study. The appreciation also goes to UTeM Computer Department especially Networking Section and other related government and private agencies for their support.

Lastly, but in no sense the least, I am thankful to all colleagues and friends especially Othman, Shahdan, Mohd Faizal, Sanusi, Robiah, Siti Rahayu, Asrul Hadi, Azuan, Radzi, Rady, Hafiz, AF Nizam, Fairuz, Azlishah, Naim, Nurhizam, Norazman, Zaki and Suhaimi for their valuable time, understanding, suggestions, comments and continuous motivation which made my PhD years a memorable and valuable experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

data in IP address representative versus Port Number Representative

LIST OF APPENDICES

LIST OF ABBREVIATIONS

CAIDA	-	The Cooperative Association for Internet Data Analysis
IHSN	-	International Household Survey Network
Dest_IP	-	Destination IP Address.
Dest_Port	-	Destination Port.
Protocol	-	TCP (Transmission Control Protocol), UDP (User Datagram Protocol)
RO	-	Research Objective.
RP	-	Research Problem.
RR	-	Randomization Response
EHR	-	Electronic health records
IT	-	Information Technology
TCPDUMP	-	File using tcpdump (with root) to capture the packets and saving them to a file to analyze
FEM	-	Feature extraction module
AM	-	Anonymization module
Libpcap	-	Software that used to capture packets travelling over a network
Wincap	-	Software that allows your network interface card to (NIC) operate in "promiscuous" mode
IPv4	-	IP version 4 formats
SPSS	-	Statistical Package for the Social Sciences system
PRAM	-	Post-randomization
Tcp / ip	-	Transmission Control Protocol (**TCP**) and the Internet Protocol (**IP**)
GPS	-	Global Positioning System

LIST OF PUBLICATIONS

Y.A.Rahim, S. Sahib, & Mastura, M. (2010). The Privacy on Spatio Temporal Data Paper presented at the *2nd International Conference on Knowledge Discovery (ICKD 2010)*

Y.A.Rahim, S.Sahib, & M.K.A. Ghani (Jan, 2011). Pseudonmization Techniques for Clinical Data: Privacy Study in Sultan Ismail Hospital Johor Bahru. *International Journal of Computing, Scientific & Academic Publishing*, 2(6), 89-98

Y.A. Rahim, S. Sahib, & M.K.A. Ghani (Jan, 2012). Pseudonmization Techniques for Privacy Study with Clinical Data. *International Journal of Information Science*, Vol. 2 No. 6, 2012, pp. 75-78.

Y.A.Rahim & S.Sahib (Jan, 2012). Randomization techniques in privacy studies. *Proceeding In International Computing and Convergence Technology (ICCCT), 2012, Index IEEE Xplore.*

Y.A.Rahim. Anonymization Clinical Data: Privacy Case Study. *Journal of Telecommunication, Electronic and Computer (JTEC)*. (Submitted).

CHAPTER 1

INTRODUCTION

1.1 Introduction

Privacy protection is an important issue in data transferring, data publishing, data management, database and data mining. The data must be protected of their privacy before the data can be published. However, in terms of protecting data, privacy is one of the big issue and an important problem with microdata or network data. The key principle that being used in all of these efforts to assure the low-risk and high-value data is that the trace of anonymization in the process of sanitizing data before it releases the information of data that potentially sensitive cannot be extracted.

However, currently, the utility of these techniques in protecting the host identities, user behaviours, network topologies, and security practices within enterprise networks has come under security. In short, several works have shown than unveiling the sensitive data in anonymize network gave traces that it may not be as difficult as initially thought. The naive solution to this problem is to address the specifics of these attacks as they are discovered. However, by doing this it could fail to address the underlying problems in its entirety. While isolated advances in network data, anonymization is important, without a holistic approach to the problem. They will simply shift the information-encoding burden to other properties of the traces, resulting in breaches of future privacy. Given the significant reliance on anonymize network traces for security research, it is clear that a more exhaustive and detailed analysis of the trace anonymization problem is in order.

The problem of anonymize publicly released data in the work of information technology (IT) is more than decades. . Over the past several decades, statisticians and computer scientists have developed approaches to anonymize the various forms of microdata, which are essentially the databases of attributes collected about individuals. One prominent example is data census, which collects information about salary, marital status, and other potentially sensitive attributes from the population of an area or a country. This microdata, much like network data, is valuable to researchers for tracking trends, and as such the anonymize microdata must provide accurate information about potentially sensitive information. At the same time, it is essential that the specifications from the data cannot be linked to the individuals. In response, several anonymization methods, privacy definitions, and utility metrics have been developed to ensure the microdata can be used for a wide spectrum of analyses while simultaneously provide principled and concrete guarantees on the privacy of those individuals within the data.

At first glance, it would seem as though the accumulated knowledge of microdata anonymization can be directly applied to network data anonymization since the two scenarios share many common characteristic, including the similarities of privacy and utility goals. However, the inherently complex nature of the network data makes direct application of these microdata methods difficult, at best. However, researcher can learn from the existing microdata that anonymization literature and glean significant insight into how to approach the problem of network data anonymization in a principled and systematic fashion.

In addition, this research points out the necessity of utility measures in quantifying the extent to which anonymization may alter the results that have obtained from analysis of the data. It is important to note that there are additional challenges that are not addressed here, such as the legal and ethical issues when collecting the network data. As a whole, the

comparison between these fields of micro data and network data anonymization serves to focus the attention of the research community on a holistic approach towards the network data anonymization that enables the type of collaboration necessary to further progress in the areas of network and computer security research.

1.2 Definition of Anonymization

Anonymization is a process and it is derived from word anonymity. Anonymity is derived from the Greek word anonymia, means "without a name" or "namelessness". Anonymity is the state to change someone or a person's name into another word to remain unknown to most people (Dankar, 2008). In colloquial use, anonymous typically refers to a person, and often ways that the personal identity, or personally identifiable information about that person is ambiguous (Hales, 2010). In other words, anonymity means protection of names and other pieces of information that can lead to identify the participants; researchers do not have to ask participants to reveal information that would aid the researchers in identifying participants' individual data (Bonneau, 2009).

Anonymization is referring to the act or process or techniques that make anonymous, of hiding or disguising identities (Brunell, 2013). Anonymization in other terms is de-identification path that removing the identities of a person from the data set and make the information of the person cannot be retrieved by the owner. It protects sensitive information in the database so it can be transferred without hacking of the information. In statistic samples survey, anonymization is typically required for the production of public use files, and to a lesser extent of generating licensed files. The reason why need to hide identities, because it is as one of the security precautions. However, anonymization has their own weaknesses in certain cases which it can be workable with this new technique that will be introduced and discussed in next subtopic later on in this thesis.

1.3 The Needed of Anonymization

In a microdata set, anonymization means removing or modifying the identifying variables contained in the set of data. Typically an identifying variable is the one that describes a characteristic of a person that is observable that is registered like identification on numbers or generally, that can be known to other persons (Duprie, 2010).

An identifying variable includes with the direct identifiers and indirect identifiers. The direct identifiers mean the variables which contain details such as names, addresses or identity card numbers. They permit direct identification of a respondent but are not needed for statistical or research purposes and should thus be removed from the published dataset. However the indirect identifiers mean the characteristics that may be shared by several respondents and whose combination could lead to the re-identification of one of them. For example, the combination of variables such as addresses, age, sex and profession would be identified if only one individual of that particular sex, age and profession lived in that particular district. Such variables are needed for statistical purposes and should thus not be removed from the published data files and this type of identifiers needs to anonymize. The process on anonymize the data will consist the information in determining which variables are potential identifiers which relies on one's personal judgement, and in modifying the level of precision of these variables to reduce the risk of re-identification to an acceptable level (Zang, 2011). The challenge is to maximize security or privacy while minimizing resulting information loss. It should be noted that the disclosure risk does not only depend on the presence of identifying variables in the dataset, but also on the existence of an intruder, which in turn depends on the potential benefit this intruder would reap from re-identification (Faizal, 2008). For some types of data such as business data, the intruder's motivation can be high. For other types of datasets, like household surveys in developing

countries, the motivation would typically be much lower as there is little to gain in re-identifying respondents.

In intruder process, the data often re-identification is done by matching data from various sources, for example matching sample survey data with administrative registers. The higher the cost, the lower the chances of an intruder can leak the information in the dataset, because that is the cost of re-identification. Generally to account for these various parameters, a disclosure scenario must be defined as a first step in the anonymization process. Scenarios can be classified into these two categories, either nosy neighbour scenarios or the external archive scenarios. In nosy neighbour scenarios, with assume the intruder has enough information on a single unit, or a few of them, and this information stems from the personal knowledge. In other words, the intruder belongs to the circle of acquaintances of a statistical unit; while the external archive scenarios, the scenarios are based on the assumption that the intruder can link records belonging to the distributed dataset to records from another available dataset (or register) which contains direct identifiers. The intruder does so by using identifying variables available in both datasets as merging keys (data matching). Conservative assumptions are often made in order to define a worst case scenario.

When producing the microdata file, the individual should always keep the user perspective in mind (Spielauer, 2011). It is fundamental that the released of the file has met this research requirement. Both information content and the choice of protection methods have to focus as much as possible on the user's needs. Knowledge of the statistical analysis that the users generally want to perform helps deciding the anonymization strategy (Benevenuto, 2010).

1.4 Problem Statement

Currently, many organizations are published their data either for research, advertising and analysing the pattern of data. However, Blundo (2011) stated that most of organizations are cautions in publishing their data. Because of that, this research studied the problem of publishing the microdata or network data without revealing sensitive information lead to the privacy preserving paradigms. K-anonymity, l-diversity, generalization, clustering and randomization are the examples of the most frequent and the current techniques that are being used in the anonymization process. Nevertheless, these five existing techniques suffer from at least one of the following drawbacks like information loss and loss on privacy with the data. However, Myra (2010) stated that most of the existing anonymization techniques suffer with information losses and the privacy data losses.

Therefore it is necessary to introduce a new technique in anonymization the data to prevent the loss of information and the privacy of data; and moreover the proposed technique must relate with the technique in the network data for adapting the existing issues and the technique that is proposed must the comprehensive. However, Martin (2010) remains states that currently, it is quite difficult to find the technique that process or generate with comprehensive. Martin (2010) states the comprehensive mean all the attributes or all the data have chances to anonymous format. Most of the current anonymization techniques using a single or a few of attributes in their anonymous process like k-anonymization or l-diversity.

Besides generating the suitable technique, selecting a few techniques that gather in one big technique as the hybrid technique is also important in developing the anonymization with the network data. It is because the success of anonymization technique depends on the achievement of the objectives that be mentioned on next subtopics. *The*

Cooperative Association for Internet Data Analysis (CAIDA, 2009) states that there are numerous features inside the anonymity data, especially with network traffic or data network that are used until today. The significant contributions of the hybrid technique can increase the published data especially for the research purposes.

The summary of the problem faced by the current anonymization process are illustrated in Table 1.1.

Table 1.1 Summary of research problems

No	Research Problem
RP 1	Chances of losses information in most anonymization process are high.
RP 2	Most anonymization process like l-diversity or k-anonymization still may cause privacy leakage with the original data from user information.
RP 3	Most of the technique likes l-diversity, k-anonymization or de-anonymization does not participate well with the payload data, and however, there be lose the integration of data.

Thus, this research will focus on introducing a new comprehensive technique to anonymize the network data. The improvement of the framework also is introduced in this research because it makes the process more comprehensive. The result of this research may be useful for those who owned the data and would like to publish their data for research, advertise or maybe for prediction or analyse the pattern or other activities.

1.5 Research Objective

Due to the lack of data privacy, the current techniques especially the privacy of the data owner was in jeopardy and this will make the data or the information in unstable and to be loss, leak or even worse still being used by a third party. This is why this research will be important because it will introduce a hybrid technique that will improve the current