COMPARATIVE STUDY OF K-ANONYMITY ALGORITHMS FOR PRIVACY PRESERVING DATAMINING

Osman Abbas Elsheikh Idris Abd Elrahman

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Computer Science (Information security)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2013

This project is dedicated to my family for their endless support and encouragement.

ACKNOWLEDGEMENT

All praise be to Allah, the Most Merciful, for His Love and Guidance Salutations on the Prophet Muhammad (PBUH), his family, and fellow companions. May I express my appreciation to ALLAH, the beneficent, the merciful, for making me a Muslim and blessing me with the privilege of acquiring a higher degree.

My heartfelt gratitude goes to my parents for bearing with me weakness upon weakness from cradle to date. I would like to thank my supervisor, **Assoc. Prof. Dr. Subariah Ibrahim** for his direction time and motivation throughout the project. I wish to thank my brother Mohammed for support and encouragement. May ALLAH reward you all the relentless efforts to see through this academic pursuit.

ABSTRACT

Nowadays, privacy issue becomes one of the main concerns of persons among their raw data. This happens at a time, when more and more historically public information is also electronically available. When these data are linked together, they provide an electronic shadow of a person or organization that is as identifying and personal as a fingerprint even when the information contains no explicit identifiers, such as name and phone number. Other distinctive data, such as birth date and ZIP code, often combine uniquely and can be linked to publicly available information to re-identify individuals. However, there are several k-anonymity algorithms available in the literature to solve that problem such as Datafly and Incognito. Nevertheless, their study of performances in terms of efficiency and accuracy is lacking. In this study, we compare these two k-anonymity algorithms. So that users can select which algorithm is more suitable for their data mining. The finding shows that Datafly gives higher overall efficiency. Comparing with Incognito which gives high accuracy. Consistent good performance of Incognito in k-anonymity has made a promising k-anonymity techniques to be used in the Privacy Preserving Technique.

ABSTRAK

Pada masa kini, isu privasi menjadi salah satu keprihatinan seseorang tentang data Asal. Ini terjadi apabila semakin banyak sejarah maklumat umum diperolehi secara elektronik. Apabila data ini berhubungan antara satu sama lain, data tersebut memberikan bayangan elektronik tentang seseorang atau organisasi yang dikenalpasti dan sulit seperti cap jari walaupun maklumat tersebut tiada pengenalan yang nyata seperti nama dan nombor telefon.Data yang berkaitan lain seperti tarikh lahir dan kod ZIP, sering digabungkan dan dihubungkan untuk dipaparkan pada umum untuk mengenalpasti semula seseorang individu. Walaubagaimanapun, terdapat banyak algoritma k-anonimiti di dalam penyelidikan untuk menyelesaikan masalah tersebut seperti Datafly dan Incognito.Tambahan lagi, kajian tentang prestasi daripada segi kadar efisyensi dan ketepatan adalah lemah. Dalam kajian ini, kami membandingkan kedua-dua algoritma k-anonimiti ini. Supaya pengguna dapat pilih algoritma mana yang lebih sesuai untuk kenalpasti data.Penemuan ini menunjukkan bahawa Datafly memberikan kadar efisyensi yang lebih tinggi secara keseluruhannya.berbanding dengan Incognito yang memberikan lebih tinggi kadar ketepatannya. Kestabilan prestasi incognito yang baik dalam k-anonimiti membuatkan teknik k-anonomiti yang meyakinkan untuk digunakan di dalam TEknik Pengekalan Privasi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ABBREVIATION	DEFINITION
DM	Data Mining
PPDM	Privacy Preserving Data Mining
HIPAA	Health Insurance Portability and Accountability Act
IL	Information Loss
k	The anonymization level
PT	Private Table
QI	Quasi-Identifier
SSN	SSN Social Security Number
DGH	Domain Generalization Hierarchy
MGT	Modified Generalization Table
VGH	Value of Generalization Hierarchy

CHAPTER 1

INTRODUCTION

1.1 Introduction

Nowadays the size of the data, which is collected daily by the public and private institutions are increasing dramatically. The process of extracting enormous of datasets by data mining has become very important to help decision making processes. In contrast, contain explicitly data sets for data mining can be used to imply to information consist of the original data that might not be intended to release for the public. Therefore, there is a privacy violation for those who refer to whom these data. Data mining can be prevented these data sets only if there are safeguards from compromising the privacy. It has been suggested the concept of privacy preserving data mining (PPDM) In response to these concerns privacy (Ciriani, 2008).

The privacy is considered one of the most critical characteristics of information systems which should be offered. Thus, there are several efforts have been suggested to integrating many techniques to maintain the privacy in order to safeguard obtains sensitive information through the extract the knowledge. It can be classified existing conservation techniques of data mining and according to the dimensions of the following five different (Verykios, 2004):

(i) Data distribution (central or distributed).

(ii) The amendment applied to the data (encryption, disorder, generalization, and so on) in order to cleanse them.

(iii) Algorithm to extract the data that has been designed this technique to save the privacy.

(iv) Type of data which need it for the protection from disclosure.

(v) The method adopted for privacy preservation.

There are several techniques such as K-anonymity and randomization (Rakesh Agrawal, 2000, Samarati, 1998, Dakshi Agrawal, 2001) that have been proposed in the last years for the performance and privacy of data mining. Moreover, the problem has been explained in different communities, such as a database group, and community statistical disclosure control and encryption community (Charu, 2007).

The privacy of the individual to whom the data belongs to, released data were at first "de-identified" by deleting explicit identifiers for instance names, addresses, and phone numbers. However this de-identified data could still have other identifying characteristics such as birth date, postal code, race and sex, when they are considered all together, almost uniquely relate to specific individuals. These sets of characteristics are often called *quasi-identifiers*. For instance, in one statistic, Sweeney finds out that 87.1% of the US population can be uniquely identified by the combination of their 5-digit zip code, gender, and date of birth because such records can be linked to publicly available databases such as voter lists and driving records (Sharow 2007).

1.2 Problem Background

K-anonymization is a technique that prevents joining attacks by generalizing and suppressing portions of the released microdata so that no individual can be uniquely distinguished from a group of size k. The real-world algorithms Datafly and μ-Argus are compared to MinGen. Both Datafly and μ-Argus use heuristics to make approximations, and so, they do not always yield optimal results (Sweeney, 2002). It is shown that Datafly can over distort data and μ-Argus can additionally fail to provide adequate protection. One of the problems is that Datafly makes crude decisions–generalizing all values associated with an attribute and suppressing all values within a tuple.

The view of k-anonymization problem from the perspective of inference attacks over all possible combinations of attributes. (Ciriani, 2007) showed that when the data contains a large number of attributes which may be considered quasi-identifiers; it becomes difficult to anonymize the data without an unacceptably high amount of information loss. This is because an exponential number of combinations of dimensions can be used to make precise inference attacks, even when individual attributes are partially specified within a range (Vijayarani, 2010). The provided analysis of the effect of dimensionality on k-anonymity methods, conclude that when a data set contains a large number of attributes which are open to inference attacks, are faced with a choice of either completely suppressing most of the data or losing the desired level of anonymity.

There are many algorithms under the k-anonymity technique have been proposed to preserve the privacy in data mining such as Bayardo-Agrawal (Bayardo, 2005), Mondrian (Samarati, 2001) and Approximation Algorithms (Gagan Aggarwal, 2005)all these algorithms use different standards to measure the quality of output and try to improve production against those standards. However, in the context of data, the trade-off between speed and optimality is not possible since researchers need to work on an anonymized data set with minimal information loss. Moreover, as opposed to the heuristic-based approaches, by insuring an optimal solution that can be located efficiently, researchers will benefit immensely, for the better the quality of the anonymized data the more valuable that data is for the research. Among these algorithms, the Datafly (Sweeney, 1997) and Incognito (David, 2005) are the most popular approaches in the privacy preserving data mining. Several studies have evaluated the results of k-anonymization algorithms based on a particular data mining task, such as information loss and efficiency (Issa, 2009),(Nurul H, 2012) and (Khaled Elemam, 2009). However, the evaluation of information loss has not previously been explored by using the global metric (Dissimilarity Metric). Specially, to evaluate Datafly and Incognito algorithms

1.3 Problem Statement

In the midst of the vast amount of information available electronically led to the disclosure of individual privacy. When these data are linked together, they provide an electronic shadow of a person or organization that is as identifying and personal as a fingerprint even when the information contains no explicit identifiers, such as name and phone number. There are several k-anonymity algorithms available in the literature, however their study on performances in terms of efficiency and accuracy is lacking. In this study, we are going to compare these two k-anonymity algorithms. So that users can select which algorithm is more suitable for their data mining.

1.4 Purpose of Study

In this research the performance of K-anonymity algorithms namely Incognito and Datafly were compared in terms of efficiency and accuracy (Information loss). At the end of this comparison, an analysis of their performances was discussed and the algorithm that shows better performance is highlighted and recommended.

1.5 Objective of Study

This research has the following objectives:

i. Studying the Privacy Preserving Data Mining.

ii. Evaluating the performance of Datafly and Incognito algorithms in Privacy Preserving Data Mining.

1.6 Scope of Study

The scope of the project is listed below:

i. The data being used in this study were from the University of California, Irvine (UCI) Cup 1996 Census dataset. Its size is 30000 rows.
ii. Weka 3.6.9 being used to implement the evaluation.
iii. Performance was evaluated based on efficiency and accuracy (Information loss) metrics.

1.7 Significant of Study

This study evaluates the performance of two k-anonymity algorithms namely Datafly and Incognito for privacy preserving data mining in terms of efficiency and accuracy. By studying each one and investigate them to show which one is more suitable to be used in privacy preserving in data mining.

1.8 Organization of Report

The thesis consists of 4 chapters. Chapter one describes the introduction, background of the study, research objectives and questions, the scope of the study and its primary objectives. The second chapter reviews available and related literature on Privacy Preserving in data mining, K-anonymity approaches. Chapter three describes the study methodology along with the appropriate framework for the study. The fourth chapter provides the analysis of the preliminary results of the study.

REFERENCES

Ciriani, S. Foresti, and P. Samarati (2008). "K-ANONYMOUS DATA MINING A SURVEY." Springer US.

Dakshi Agrawal, C. C. A. (2001). On the Design and Quantification of Privacy Preserving Data Mining Algorithms. ACM PODS Conference.

David, K. L. (2005). "Incognito: Efficient full-domain k-anonymity." ACM SIGMOD Int'l Conf. on Management of Data, Baltimore, MD.

Ercan Nergiz, C. C. (2006). "Thoughts on k-Anonymization." Department of Computer Sciences, Purdue University.

Issa, R. (2009). Satisfying K-Anonymity:New Algorithm and Empirical Evaluation. Under the auspices of the Ottawa-Carleton Institute for Computer Science. Ottawa, Ontario, Canada, Carleton. Master of Computer Science.

Rakesh Agrawal, R. S. (2000). Privacy-Preserving Data Mining. ACM SIGMOD Conference.

Samarati, P. (2001). "Protecting respondents identities in microdata release. " IEEE Transactions on Knowledge and Data Engineering **13**: 1010–1027.

Samarati, S. L. (1998). "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression." IEEE Symp. on Security and Privacy.

Samarati, V. (2007). "k-Anonymity." Springer US, Advances in Information Security.

Sweeney, L. (1997). "Guaranteeing anonymity when sharing medical data, the datafly system." Journal of the American Medical Informatics Association. Hanley&Belfus,Inc.

Sweeney, L. (2001). " Information Explosion." Washington, DC: Urban Institute.

Verykios, V. S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., Theodoridis, Y.: (2004).

(OCR)., O. F. C. R. 2003. Summary of the HIPAA privacy rule. *In:* SERVICES, U. S. D. O. H. A. H. (ed.).

EVFIMEVSKI, J. G., AND R. SRIKANT. Limiting Privacy Breaches in Privacy Preserving Data Mining. PODS Conf, 2003.

MACHANAVAJJHALA, D. GEHRKE, AND M. VENKITASUBRAMANIAM. 2006. diversity: Privacy beyond k-anonymity. *22nd IEEE International Conference on Data Engineering.*

ALEXANDRE EVFIMIEVSKI, T. G. 2009. Privacy-Preserving Data Mining. *IGI Global*.

ALI, E. S. M. 2009. *THE ENCRYPTION – DECRYPTION OF THE COLUMN LEVEL FOR ACOMMERCIAL DBMS WITH SUPPORTING USER INTERFACE.* Master of Computer Science (Information Security), UNIVERSITI TEKNOLOGI MALAYSIA.

AMIRBEKYAN A, E.-C. V. 2007. Privacy-preserving k-NN for small and large data sets. *The 7th IEEE International Conference Data Mining.*

APARNA KOLLI, A., LAKSHMI GANESH 2010. Methods for Database Security.

BAYARDO, R., AGRAWAL, R 2005. Data privacy through optimal k-anonymization. *Conf. on Data Engineering.*

BERTINO, E., FOVINO, I.N., PROVENZA, L.P. 2005. A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery.

CHARU, A. A. P. 2004. A condensation approach to privacy preserving data mining. *International Conference on Extending Database Technology (EDBT).*

CIRIANI, V., DE CAPITANI DI VIMERCATI, S., FORESTI, S., AND SAMARATI, P 2007. k-Anonymity. Secure Data Management in Decentralized Systems. Advances in Information Security. *Springer US*.

CLIFTON, J. V. A. C. Privacy preserving association rule mining in vertically partitioned data. International Conference on Knowledge Discovery and data maining, 2002 Edmonton, Alberta, Canada.

CLIFTON, W. J. A. C. 2005. Privacy-preserving distributed k-anonymity. *Data and Application Security.* CTIFIP WG: Storrs,.

MARTIN, D. K., A. MACHANAVAJJHALA, J. GEHRKE, AND J. HALPERN. 2007. Worst-case background knowledge for privacy-preserving data publishing. *International Conference on Data Engineering.*

DAKSHI AGRAWAL, C. C. A. 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. *ACM PODS Conference.*

EL EMAM, K., JABBOURI, S., SAMS, S., DROUET, Y., AND POWER 2006. Evaluating common deidentification heuristics for personal health information. *Journal of Medical Internet Research*.

ELISA BERTINO, D. L., AND WEI JIANG 2007. A Survey of Quantification of Privacy Preserving Data Mining Algorithms.

FINKENZELLER, K. 2003. *RFID handbook: Fundamentals and applications in contactless smart cards and identification*, John Wiley & Sons.

GARFINKEL, R., B 2005. *RFID: Applications, security, and privacy*, Addison-Wesley.

GAYATRI NAYAK, S. D. 2011. A SURVEY ON PRIVACY PRESERVING DATA MINING APPROACHES AND TECHNIQUES. *International Journal of Engineering Science and Technology (IJEST)*.

GE WEIPING, W. W., ZHOU HAOFENG 2006. Classification Mining Base on Privacy Protect. *Computer Research and Development.*

GOLDREICH., O. 1998. *Secure multi-party computation* [Online]. http://www.wisdom.weizmann.ac.il/home/oded/publichtml/foc.html.

KARGUPTA, S. D., Q. WANG, AND K. SIVAKUMAR 2003. On the Privacy Preserving Properties of Random Data Perturbation Techniques
In: CONF, P. I. I. L. (ed.) *Data Mining.*

IYENGAR, V. 2002. Transforming data to satisfy privacy constraints. *Conf. on Knowledge Discovery and Data Mining.*

KE WANG, B. C. M. F. A. P. S. Y. 2005. Template based privacy preservation in classification problems. *In ICDM.*

KHALED ELEMAM, F. K. D., ROMEO ISSA,ELIZABETH JONKER, DANIEL AMYOT, ELISE COGO,JEAN-PIERRE CORRIVEAU,MARK WALKER, SADRUL
CHOWDHURY,REGISVAILANCOURT,BPHARM,PHARAD,PHARMD, TYSONROFFEY,JIM BOTTOMLEY 2009. A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. J Am Med Inform Assoc.

KRISTEN LEFEVRE, D. J. D., RAGHU RAMAKRISHNAN 2005. Incognito: Efficient FullDomain KAnonymity. *In Proc. of the ACM SIGMOD Conference on* Management of Data. Baltimore, Maryland.

Haisheng L. 2010. Study of Privacy Preserving Data Mining",. Third International Symposium on Intelligent Information Technology and Security Informatics.

ZHANG, S. J., AND A. BRODSKY. 2007. Information disclosure under realistic assumptions: privacy versus optimality. *Conference on Computer and Communications Security*

LAUR, H. L., AND T. MIELI' AINEN 2006. Cryptographically private support vector machines. *In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

LINDELL, Y. A., B 2002. Privacy preserving data mining. *Induction of decision trees.* QUINLAN, J. R: Mach. Learn.

NERGIZ, M. A., AND C. CLIFTON 2007. Hiding the presence of individuals from shared databases. *In Proc. 27th ACM International Conference onManagement of Data (SIGMOD).*

OLIVEIRA, S. R. M., ZAIANE, O.R 2002. Privacy preserving frequent itemset mining*In:* ICDM, I. (ed.) *Security and Data Mining.*

PAILLIER, P. 1999. Public-key cryptosystems based on composite degree residuosity classes. *In Advances in Cryptography—EUROCRYPT'99.* Prague, Czech Republic.

RIVEST, L. A., AND M. DERTOUZOS 1978. On data banks and privacy homomorphisms. *In Foundations of Secure Computation.* DeMillo.

RAKESH AGRAWAL, R. S. 2000. Privacy-Preserving Data Mining. *ACM SIGMOD Conference.*

RANDOMIZED, W. 1965. A survey technique for eliminating evasive answer bias. *In: The American Statistical Association*.

CHAWLA, C. D., AND F. MCSHERRY. Toward Privacy in Public Databases. Second Theory of Cryptography, 2005.

VIJAYARANI, A. T., M.SAMPOORNA 2010. Analysis of Privacy Preserving K-Anonymity Methods and Techniques. *Proceedings of the International Conference on Communication and Computational Intelligence* Kongu Engineering College, Perundurai, Erode, T.N.,India.

SAMARATI, P. 2001. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering,* 13**,** 1010–1027.

SAMARATI, S. L. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. *IEEE Symp. on Security and Privacy*.

SHARMA, D. 2012. A Survey on Maintaining Privacy in Data Mining. *International Journal of Engineering Research and Technology,* 1.

SHIPRA AGRAWAL, J. R. H. 2005. A Framework forHigh-Accuracy Privacy-Preserving Mining. *The 21st International Conference on Data Engineering.*

SRIKANT, R. A. A. R. 2000. Privacy Preserving Data Mining. *In:* CONF, P. A. S. (ed.) *Management of Data.*

STANLEY R. M. OLIVEIRA, O. R. Z. 2002. Privacy Preserving Frequent Itemset Mining. *In:* MIN-, D. & ING WORKSHOP ON PRIVACY, S., AND DATA MINING, (eds.) *Data Mining Workshop on Privacy, Security, and Data Mining,.* Maebashi City, Japan: IEEE International Conference.

SWEENEY, L. 1997. Guaranteeing anonymity when sharing medical data, the datafly system. *Journal of the American Medical Informatics Association. Hanley & Belfus,Inc*.

SWEENEY, L. 2002. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty*.

SWEENEY, L. 2008. *Computational Disclosure Control A Primer on Data Privacy Protection.* Massachusetts Institute of Technology.

TAO, X. 2007. M-invariance: towards privacy preserving republication of dynamic datasets. *International Conference on Management of Data.*

TRUTA, T. M., CAMPAN, A., ABRINICA, M., AND MILLER, J. 2008. A Comparison between Local and Global Recoding Algorithms for Achieving Microdata PSensitive K-Anonymity. *Acta Universitatis Apulensis.* Alba Iulia, Romania.

VAIDYA, J. A., C. Privacy-Preserving association rule mining in vertically partitioned data. International Conference on Knowledge Discovery and Data Mining, 2002 Edmonton, Alberta, Canada. ACM Press, New York.

VALENTINA CIRIANI, S. D. C. D. V., SARA FORESTI, AND PIERANGELA SAMARATI 2007. K-anonymity. *In T. Yu and S. Jajodia, editors, Security in Decentralized Data Management. Springer, Berlin Heidelberg*.

WALTERS, G. J. 2001. Human Rights in an Information. *University of Toronto Press.*

WINKLER, J. J. K. A. W. E. 2003. Multiplicative Noise for Masking Continuous Data. *In:* 2003-01 (ed.) *Statistical.* Washington D.C: Research Division,US Bureau of the Census.

WU XIAO-DAN, Y. D.-M., LIU FENG-LIL, WANG YUN-FENG, CHU CHAO-HSIEN 2007. Privacy Preserving Data Mining Algorithms by Data Distortion. *School of Management, Hebei University of technology, P.R.China,.*

XIAOLIN ZHANG, H. B. 2010. Research on Privacy Preserving Classification Data Mining Based on Random Perturbation. *International Conference on Information, Networking and Automation (ICINA).*

YANG, R. W. A. Z. 2004. Privacy-preserving bayesian network structure computation on distributed heterogeneous data. *In:* SIGKDD, A. (ed.) *International Conference on Knowledge Discovery and Data Mining (KDD).*

YAO, A. C. Protocols for secure computations,". Symposium on Foundations of Computer Science, 1982. IEEE.

ZHAN, S. M., AND L. CHANG 2005. Privacy-preserving collaborative association rule mining. *Working Conference on Data and Applications Security.* University of Connecticut, Storrs, CT, U.S.A.

ZHAN, W. D. A. Z. 2002. Building decision tree classifier on private data. *Privacy, Security, and Data Mining.* Maebashi City, Japan: IEEE.

ZHAN, W. D. A. Z. Using randomized response techniques for privacy-preserving data mining. International Conference on Knowledge Discovery and Data Mining, 2003 Washington, DC, USA.

ZHANG GUORONG, Y. J. 2007. Clustering algorithm of maintaining privacy under distribute environment. *Computer Application and Engineering.*