AIN SHAMS UNIVERSITY
Faculty of Computer
& Information Sciences
Information Systems
Department

A Privacy Preservation Publishing Approach in Big Data Streams

**Thesis submitted to the Department of Information Systems
Faculty of Computer and Information Sciences
Ain Shams University, Egypt**

**In fulfillment of the requirements for the Master of Science Degree
(MSc) in Computer and Information Sciences**

BY

Saad Abd Elhameed Saad Abd Elhameed

B.Sc. of Computer & Information Sciences, Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University
Teaching Assistant at Software Engineering and Information Technology
Department, Faculty of Engineering and Technology, Egyptian Chinese
University

Under the Supervision of

Prof. Dr. Mohamed Essam Khalifa

Professor at Basic Sciences Department,
Faculty of Computer and Information Sciences,
Ain Shams University
Vice President for Graduate Studies and Research,
Egyptian Chinese University

Dr. Sherin Mohamed Mahmoud Moussa

Associate Professor,
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

December 2017

Acknowledgement

First and foremost, I am grateful to Almighty Allah for His immense blessings and graciously helping me to complete this thesis.

This thesis owes its existence to the help, support, and inspiration of many people. In the first place, I owe my deepest gratitude to my main supervisor Prof. Dr. Mohamed Essam Khalifa whose great knowledge, experience and sharp sense of research direction have provided invaluable feedback to improve the quality of this thesis. This thesis would not have been possible without his sound advice and encouragement. I would like to express my sincere appreciation and gratitude to my associate supervisor, Dr. Sherin Mohamed Moussa for her tremendous amount of support, guidance, insightful comments, and invaluable collaboration. It was my great pleasure and honor to have such professors as my supervisors for their endless support and cooperation during my study and research, in addition to their final revision of the thesis.

Last, but not least with all my appreciation and love that no words can express, I would like to thank all my friends and family members for their endless love and support. I offer my love and heartfelt thanks to my parents for their lifelong support in all my endeavors. I dedicate this thesis to my dear parents, my beloved wife, and my son who always provided me with love, prayers, blessings, advice and care. They are behind any success in my life.

Abstract

With the recent remarkable and fast evolution in telecommunication and computing technologies, great amounts of individuals' data are collected and used by several organizations in the society. This includes diverse data sources, often for data of high dimensionality. At the same time, most of these data are stored in tabular format and can include sensitive content.

In some cases, some organizations need to share these gathered data to be used in business analysis, decision making or scientific research purposes, which can involve sensitive information about individuals. However, these data cannot be published in their original form to other third parties, due to the associated privacy concerns. Consequently, preserving data privacy represents an essential task in order to allow such data to be published with the guarantee of preserving individuals' privacy when sharing their included private data. This protects the identity of individuals from being discovered, and their sensitive information from being disclosed by any intruder through the published data. The data required to be published can be static or dynamic from data streams, including Single Sensitive Attribute (SSA) or Multiple Sensitive Attributes (MSA). In this context, Privacy-Preserving Tabular Data Publishing (PPTDP) has drawn considerable attention from the research community, where different anonymization approaches have been proposed to preserve the privacy of individuals' tabular data.

This thesis introduces a comparative study to analyze and evaluate the main different data anonymization approaches that have been introduced in PPTDP. The study investigates the three broad areas of research: SSA, MSA and data streams. A detailed criticism is presented to highlight the strengths and the weaknesses of each approach, supported by detailed comparison tables. In addition, the presented study investigates the deployment of the data anonymization approaches in the cloud and Internet of Things (IoT) environments. Besides, a research gap analysis is discussed, with a focus on capturing the current state of art in this field in order to highlight the future directions that can be considered. In addition, we consider the area of privacy-

preserving of static data publishing by proposing an Enhanced Additive Noise (EAN) approach for privacy-preserving microdata with SSA publishing. The EAN approach enforces a newly-proposed privacy constraint on the value of Sensitive Attribute (SA) in the published data, whereas the original values of the other attributes are published to preserve better data utilization and attributes' distribution. Hence, the proposed approach maintains better published data utility to allow more accurate mining and analytical results from the published data, where more robust privacy protection against privacy attacks is provided.

On the other hand, data streams have become a widely-adopted data representation format in many real-world domains and applications. Similarly, this data streaming may be needed to be published for different scientific research, mining, or analysis purposes. However, such streams may also contain personal-specific data that could be considered as sensitive data about individuals. When sharing these streams, these sensitive data should be well-protected against many privacy disclosure attacks to preserve individuals' privacy. This makes the privacy preserving of data streams, while maintaining their utilization, is a real challenge. Consequently, in this thesis, the area of privacy-preserving of data stream publishing is investigated, where some research studies have started to consider different ways of privacy-preserving to publish such data streams. However, the investigated approaches consider data streams with only SSA. In addition, they do not protect the published streams from all possible privacy attacks. Thus, this thesis proposes a novel Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing, in which stream tuples are anonymized sequentially. Besides, two new privacy restrictions are introduced to restrict the published Sensitive Attributes (SAs) values: Semantic-diversity and Sensitivity-diversity. Thereby, RSA-SA approach can protect the sensitive values of the published data streams against the related privacy attacks, which are the attribute disclosure, skewness, similarity, and sensitivity attacks. In addition, RSA-SA approach handles data streams that have either single or multiple sensitive attributes with a minimum information loss and delay time. Therefore, the data utility of the published data streams is efficiently maintained to provide more accurate mining and analytical results out of such streams, where robust invulnerability to privacy attacks is sustained.

Table of Contents

List of Abbreviations

ABE	Attribute-Based Encryption
ABS	Attribute Based Signatures
AD	Attribute Disclosure
ADM	Anonymous Data Sets Management
AN	Additive Noise
ANGELMS	Anatomy and Generalization on Multiple Sensitive attributes
APE	Average Protecting Expectation
APTT	Average Processing Time per Tuple
BUG	Bottom-Up Generalization
CASTLE	Continuously Anonymizing STreaming data via adaptive ClustEring
CAVG	Normalized Average EC size
CSV	Comma Separated Values
DA	Data Anonymization
DF	Delay-Free
DGM	Domain Generalization Hierarchy
DM	Discernibility Metric
DU	Data Update
EAN	Enhanced Additive Noise
EC	Equivalence Class
EI	Explicit Identifier attributes
EMD	Earth Mover Distance
HDFS	Hadoop Distributed File System
ID	Identity Disclosure

ILT	Information Loss per Tuple
IoT	Internet of Things
KNN	K-Nearest Neighbor
LBS	Location-Based Services
MD	Membership Disclosure
MSA	Multiple Sensitive Attributes
NCP	Normalized Certainty Penalty
NLP	Normalized correlation Loss Penalty
NSAs	Non-Sensitive Attributes
PA	Permutation Anonymization
PPDP	Privacy-Preserving Data Publishing
PPTDP	Privacy-Preserving Tabular Data Publishing
PSI	Privacy Specification Interface
QIDs	Quasi Identifier attributes
QIT	Quasi-Identifier Table
QIT	Quasi-Identifier Tuple
RSA-SA	Restricted Sensitive Attributes-based Sequential Anonymization
SA	Sensitive Attribute
SANATOMY	Stream ANATOMY
SAs	Sensitive Attributes
SeA	Sensitivity Attack
SiA	Similarity Attack
SkA	Skewness Attack
SKY	Stream K-anonYmity
SSA	Single Sensitive Attribute
ST	Sensitive Table

ST	Sensitive Tuple
Sw	Sliding window
SWAF	Sliding Window Anonymization Framework
TDS	Top-Down Specialization
TPTDS	Two-Phase Top-Down Specialization
WCDSA	Weak Clustering-based Data Streams k-Anonymity

List of Figures

List of Tables

List of Publications

[1] Saad A. Abdelhameed, Sherin M. Moussa, and Mohamed E. Khalifa, "Privacy-Preserving Tabular Data Publishing: A Comprehensive Evaluation from Web to Cloud", Computers & Security, Elsevier, IF 2.849, pp. 74-95, 2018, DOI: 10.1016/j.cose.2017.09.002, 72.

[2] Saad A. Abdelhameed, Sherin M. Moussa, and Mohamed E. Khalifa, "Enhanced Additive Noise Approach For Privacy-Preserving Tabular Data Publishing", In Intelligent Computing and Information Systems (ICICIS), 2017 IEEE Eighth International Conference on, pp. 284-291. IEEE, 2017.

[3] Saad A. Abdelhameed, Sherin M. Moussa, and Mohamed E. Khalifa, "Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) Approach for Privacy-Preserving Data Stream Publishing", Submitted to Information Sciences, Elsevier, IF 4.832, 2017.