Preserving Privacy in On-Line Analytical Processing (OLAP)

Advances in Information Security

Sushil Jajodia
Consulting Editor
Center for Secure Information Systems
George Mason University
Fairfax, VA 22030-4444
email: jajodia@gmu.edu

The goals of the Springer International Series on ADVANCES IN INFORMATION SECURITY are, one, to establish the state of the art of, and set the course for future research in information security and, two, to serve as a central reference source for advanced and timely topics in information security research and development. The scope of this series includes all aspects of computer and network security and related areas such as fault tolerance and software assurance.

ADVANCES IN INFORMATION SECURITY aims to publish thorough and cohesive overviews of specific topics in information security, as well as works that are larger in scope or that contain more detailed background information than can be accommodated in shorter survey articles. The series also serves as a forum for topics that may not have reached a level of maturity to warrant a comprehensive textbook treatment.

Researchers, as well as developers, are encouraged to contact Professor Sushil Jajodia with ideas for books under this series.

Additional titles in the series:

SECURITY FOR WIRELESS SENSOR NETWORKS by Donggang Liu and Peng Ning; ISBN: 978-0-387-32723-5
MALWARE DETECTION edited by Somesh Jha, Cliff Wang, Mihai Christodorescu, Dawn Song, and Douglas Maughan; ISBN: 978-0-387-32720-4
ELECTRONIC POSTAGE SYSTEMS: Technology, Security, Economics by Gerrit Bleumer; ISBN: 978-0-387-29313-2
MULTIVARIATE PUBLIC KEY CRYPTOSYSTEMS by Jintai Ding, Jason E. Gower and Dieter Schmidt; ISBN-13: 978-0-378-32229-2
UNDERSTANDING INTRUSION DETECTION THROUGH VISUALIZATION by Stefan Axelsson; ISBN-10: 0-387-27634-3
QUALITY OF PROTECTION: Security Measurements and Metrics by Dieter Gollmann, Fabio Massacci and Artsiom Yautsiukhin; ISBN-10: 0-387-29016-8
COMPUTER VIRUSES AND MALWARE by John Aycock; ISBN-10: 0-387-30236-0
HOP INTEGRITY IN THE INTERNET by Chin-Tser Huang and Mohamed G. Gouda; ISBN-10: 0-387-22426-3
CRYPTOGRAPHICS: Exploiting Graphics Cards For Security by Debra Cook and Angelos Keromytis; ISBN: 0-387-34189-7
PRIVACY PRESERVING DATA MINING by Jaideep Vaidya, Chris Clifton and Michael Zhu; ISBN-10: 0-387- 25886-8
BIOMETRIC USER AUTHENTICATION FOR IT SECURITY: From Fundamentals to Handwriting by Claus Vielhauer; ISBN-10: 0-387-26194-X

Additional information about this series can be obtained from
http://www.springer.com

Preserving Privacy in On-Line Analytical Processing (OLAP)

by

Lingyu Wang
Concordia University
Canada

Sushil Jajodia
George Mason University
USA

Duminda Wijesekera
George Mason University
USA

Springer

Lingyu Wang
Concordia University
Concordia Institute for Information
Systems Engineering
Montreal, QC
Canada H3G 1T7
wang@ciise.concordia.ca

Sushil Jajodia
George Mason University
Mail Stop 5B5
Suite 417, Research I Bldg.
4400 University Drive
Fairfax, VA 22030-4444
jajodia@gmu.edu

Duminda Wijesekera
George Mason University
Mail Stop 5B5
Suite 417, Research I Bldg.
4400 University Drive
Fairfax, VA 22030-4444
dwijesek@gmu.edu

Library of Congress Control Number:

Preserving Privacy in On-Line Analytical Processing (OLAP)
by Lingyu Wang, Sushil Jajodia, Duminda Wijesekera

ISBN-10: 0-387-46273-2
ISBN-13: 978-0-387-46273-8
e-ISBN-10: 0-387-46274-0
e-ISBN-13: 978-0-387-46274-5

Printed on acid-free paper.

© 2007 Springer Science+Business Media, LLC.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now know or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks and similar terms, even if the are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

To my wife Quan, with love.
- Lingyu

To my wife, Kamal.
- Sushil

To Devika and Olindi, for the past and hopes beyond.
- Duminda

Preface

Information is of strategic and operational importance to any organization. With rapid advancements in computer and network technology, it is possible for an organization to collect, store, and retrieve vast amounts of data of all kinds quickly and efficiently. Databases of today no longer contain only data used for day–to–day data processing; they have become information systems that store everything—vital or not—to an organization.

Aggregation and derivation are important and desirable effects of collecting data and storing them in databases. Aggregation is the combining of information from various sources. Derivation goes one step further; it uses different pieces of data to deduce or create new or previously unavailable information from the aggregates.

Unfortunately, aggregates and derived data pose serious problems since new information can be derived from available information in several different ways. If the concerns related to security are not properly resolved, security violations may lead to losses of information that may translate into financial losses or losses whose values are obviously high but difficult to quantify (e.g., national security). At the same time, these large information systems represent a threat to the privacy of individuals since they contain a great amount of detail about individuals. Admittedly, the information collection by an organization is essential for conducting its business; however, indiscriminate collection and retention of data represents an extraordinary intrusion on privacy of individuals.

Privacy protection is both a personal and fundamental right of all individuals. Individuals have a right to expect that organizations will keep personal information confidential. One way to ensure this is to require that organizations collect, maintain, use, and disseminate identifiable personal information and data only as necessary to carry out their functions.

This book addresses the problem of disclosure in on-line analytical processing (OLAP) systems. These systems are designed to help analysts in gaining insights to the large amount of data stored in enterprise data warehouses. OLAP systems heavily depend on aggregates of data in order to hide in-

significant details and hence to accentuate global patterns and trends. As the underlying data model, a data cube can nicely organize multi-dimensional aggregates formulated using dimension hierarchies inherent to given data.

It is critical that data cubes be analyzed for possible inference or derivation problems. This book first provides a good understanding of the ways such inference problems may arise. Armed with this understanding, the book then takes steps to eliminate these problems. Following a review of past research in statistical databases, the book shows how existing methodologies can be borrowed for developing security measures in OLAP data cubes. These measures produce improved results than by their counterparts in statistical databases, but they also inherit certain limitations from the latter. Taken such limitations as a lesson, the book discusses several novel methods that finally demonstrate a promising way to completely eliminate threats to privacy in OLAP systems.

Fairfax, VA, *Lingyu Wang, Sushil Jajodia, Duminda Wijesekera*

Contents

1 Introduction 1
1.1 Background 1
1.2 Problem Statement 4
1.3 Overview 6
1.3.1 Overview of Basic Concepts 7
1.3.2 Adapting Previous Approaches to Data Cubes 8
1.3.3 A Lattice-based Solution 10

2 OLAP and Data Cubes 13
2.1 OLAP 13
2.2 Data Cube 15

3 Inference Control in Statistical Databases 21
3.1 Query Set Size Control and Trackers 23
3.2 The Star Query Model 25
3.3 Key-Specified Queries 26
3.4 Linear System Attack and Audit Expert 28
3.5 Intractbility of Inference Control 32

4 Inferences in Data Cubes 37
4.1 Introduction 37
4.2 Preliminaries 38
4.3 Arbitrary Range Queries 41
4.4 Restricted Range Queries 45
4.4.1 Even Range Query Attack 45
4.4.2 Indirect Even Range Query Attack 48
4.4.3 Skeleton Query Attack 49
4.5 Conclusion 51

5 Cardinality-based Inference Control 53
5.1 Introduction 53
5.2 Preliminaries 57
5.2.1 Data Cube 57
5.2.2 Compromisability 60
5.2.3 Formalization Rationale 63
5.3 Cardinality-based Sufficient Conditions 66
5.3.1 Trivial Compromisability 66
5.3.2 Non-trivial Compromisability 68
5.4 A Three-Tier Inference Control Model 76
5.5 Cardinality-based Inference Control for Data Cubes 80
5.5.1 Inference Control Algorithm 80
5.5.2 Correctness and Time Complexity 81
5.5.3 Implementation Issues 83
5.5.3.1 Integrating Inference Control into OLAP 83
5.5.3.2 Re-ordering Tuples in Unordered Dimensions 84
5.5.3.3 Update Operations 85
5.5.3.4 Aggregation Operators Other Than Sum 86
5.6 Conclusions 86

6 Parity-based Inference Control for Range Queries 91
6.1 Introduction 91
6.2 Preliminaries 93
6.2.1 Motivating Examples 93
6.2.2 Definitions 94
6.3 Applying Existing Methods to MDR Queries 97
6.3.1 Query Set Size Control, Overlap Size Control and Audit Expert 97
6.3.2 Finding Maximal Safe Subsets of Unsafe MDR Queries 99
6.4 Parity-Based Inference Control 102
6.4.1 Even MDR Queries 103
6.4.2 Characterizing the QDT Graph 107
6.4.3 Beyond Even MDR Queries 110
6.4.4 Unsafe Even MDR Queries 112
6.5 Discussion 114
6.6 Conclusion 116

7 Lattice-based Inference Control in Data Cubes 119
7.1 Introduction 119
7.2 The Basic Model 120
7.3 Specifying Authorization Objects in Data Cubes 123
7.4 Controlling Inferences in Data Cubes 126
7.4.1 Inferences in Data Cubes 127
7.4.2 Preventing Multi-Dimensional Inferences 130
7.4.2.1 Assumptions 130

7.4.2.2 A Special Case 131
7.4.2.3 The General Case 136
7.4.3 Eliminating One-Dimensional Inferences 140
7.5 Implementation Options and Complexity 143
7.6 Summary 145

8 Query-driven Inference Control in Data Cubes 147
8.1 Introduction 147
8.2 Authorization Objects and Queries in Data Cubes 148
8.3 The Static Approach and Its Impact on Availability 149
8.4 Query-Driven Prevention of Multi-Dimensional Inferences 151
8.4.1 A Special Case 152
8.4.2 The General Case 156
8.4.3 Authorizing Queries 160
8.4.4 Complexity Analysis 165
8.5 Summary 167

9 Conclusion and Future Direction 169

References 173

Index 179