SpringerBriefs in Computer Science

Series Editors
Stan Zdonik
Peng Ning
Shashi Shekhar
Jonathan Katz
Xindong Wu
Lakhmi C. Jain
David Padua
Xuemin Shen
Borko Furht
VS Subrahmanian

For further volumes
http://www.springer.com/series/10028

Asaf Shabtai • Yuval Elovici • Lior Rokach

A Survey of Data Leakage Detection and Prevention Solutions

Springer

Asaf Shabtai
Department of Information Systems
Engineering
Ben-Gurion University
Beer-Sheva, Israel

Lior Rokach
Department of Information Systems
Engineering
Ben-Gurion University
Beer-Sheva, Israel

Yuval Elovici
Department of Information Systems
Engineering
Telekom Innovation Laboratories
Ben-Gurion University
Beer-Sheva, Israel

ISSN 2191-5768 ISSN 2191-5776 (electronic)
ISBN 978-1-4614-2052-1 ISBN 978-1-4614-2053-8 (eBook)
DOI 10.1007/978-1-4614-2053-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012932272

© The Author(s) 2012
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Information and data leakage pose a serious threat to companies and organizations as the number of leakage incidents and the cost they inflict continues to increase. Whether caused by malicious intent or by an inadvertent mistake, data loss can diminish a company's brand, reduce shareholder value, and damage the company's goodwill and reputation. Data leakage prevention (DLP) has been studied both in academic research areas and in practical application domains. This book aims to provide a structural and comprehensive overview of current research and practical solutions in the DLP domain. Existing solutions have been grouped into different categories based on a taxonomy described in the book. The taxonomy presented characterizes DLP solutions according to various aspects such as leakage source, data state, leakage channel, deployment scheme, prevention and detection approaches, and action taken upon leakage. In the commercial section solutions offered by the leading DLP market players are reviewed based on professional research reports and material obtained from vendor Web sites. In the academic section available academic studies have been clustered into various categories according to the nature of the leakage and the protection provided. Next, the main data leakage scenarios are described, each with the most relevant and applicable solution or approach that will mitigate and reduce the likelihood or impact of data leakage. In addition, several case studies of data leakage and data misuse are presented. Finally, the related research areas of privacy, data anonymization, and secure data publishing are discussed.

We would like to express our gratitude for all colleagues and graduate students that generously gave comments on drafts or counsel otherwise. We would like to express our special thanks to Jennifer Evans, Jennifer Maurer, Courtney Clark, and the staff members of Springer for their kind cooperation throughout the production of this book. We would like to thank to Prof. Zdonik, S., Prof. Ning, P., Prof. Shekhar, S., Prof. Katz, J., Prof. Wu, X., Prof. Jain, L.C., Prof. Padua, D., Prof. Shen, X., Prof. Furht, B. and Prof. Subrahmanian, V. for including our book in their important series (SpringerBriefs in Computer Science).

Beer-Sheva, Israel

Asaf Shabtai
Yuval Elovici
Lior Rokach

Contents

1 Introduction to Information Security 1

2 Data Leakage 5

3 A Taxonomy of Data Leakage Prevention Solutions 11

3.1 What to protect? (data-state) 11

3.2 Where to protect? (deployment scheme) 12

3.3 How to protect? (leakage handling approach) 13

4 Data Leakage Detection/Prevention Solutions 17

4.1 A review of commercial DLP solutions 17

4.1.1 Market overview 17

4.1.2 Technological offerings of market leaders 17

4.1.3 Conclusions, remarks, and problems with the state of the art in industrial DLP 19

4.2 Academic research in the DLP domain 21

4.2.1 Misuse detection in information retrieval (IR) systems 24

4.2.2 Misuse detection in databases 26

4.2.3 Email leakage protection 28

4.2.4 Network/web-based protection 30

4.2.5 Encryption and access control 31

4.2.6 Hidden data in files 33

4.2.7 Honeypots for detecting malicious insiders 34

5 Data Leakage/Misuse Scenarios 39

5.1 Classification of data leakage/misuse scenarios 39

5.1.1 Where did the leakage occur? 39

5.1.2 Who caused the leakage? 39

5.1.3 What was leaked? 40

5.1.4 How was access to the data gained? 40

5.1.5 How did the data leak? 41

5.2 Description of main data leakage/misuse scenarios 41

5.3 Discussion 46

6 Privacy, Data Anonymization, and Secure Data Publishing ... 47
6.1 Introduction to data anonymization ... 47
6.2 Elementary anonymization operations ... 48
6.2.1 Generalization ... 48
6.2.2 Suppression ... 51
6.2.3 Permutation ... 51
6.2.4 Perturbation ... 51
6.3 Privacy models ... 52
6.3.1 Basic concepts ... 52
6.3.2 *k*-Anonymity ... 52
6.3.3 *L*-Diversity ... 54
6.3.4 *K*-Uncertainty ... 55
6.3.5 (X,Y)-Privacy ... 56
6.3.6 (X,Y)-Anonymity ... 56
6.3.7 (X,Y)-Linkability ... 57
6.4 Metrics ... 58
6.4.1 Information metrics ... 58
6.4.2 Search metrics ... 59
6.5 Standard anonymization algorithms ... 60
6.6 Multiple-release publishing ... 62
6.6.1 Single vs. multiple-release publishing ... 63
6.6.2 Publishing releases in parallel ... 63
6.6.3 Publishing releases in sequence ... 64
6.6.4 Anonymizing sequential releases ... 64

7 Case studies ... 69
7.1 Misuse detection in database systems ... 69
7.1.1 Applying unsupervised context-based analysis ... 70
7.1.2 Calculating a misusability score for tabular data ... 74
7.2 Using honeytokens ... 76
7.3 Email leakage ... 79

8 Future Trends in Data Leakage ... 83

References ... 87