

Privacy-Preserving Data Publishing

An Overview

Synthesis Lectures on Data Management

Editor

M. Tamer Özsu, *University of Waterloo*

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series will publish 50- to 125 page publications on topics pertaining to data management. The scope will largely follow the purview of premier information and computer science conferences, such as ACM SIGMOD, VLDB, ICDE, PODS, ICDT, and ACM KDD. Potential topics include, but not are limited to: query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Privacy-Preserving Data Publishing: An Overview

Raymond Chi-Wing Wong and Ada Wai-Chee Fu

2010

An Introduction to Duplicate Detection

Felix Naumann and Melanie Herschel

2010

Keyword Search in Databases

Jeffrey Xu Yu, Lu Qin, and Lijun Chang

2010

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2010

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Privacy-Preserving Data Publishing: An Overview

Raymond Chi-Wing Wong and Ada Wai-Chee Fu

ISBN: 978-3-031-00706-4 paperback

ISBN: 978-3-031-01834-3 ebook

DOI 10.1007/978-3-031-01834-3

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #3

Series Editor: M. Tamer Özsu, *University of Waterloo*

Series ISSN

Synthesis Lectures on Data Management

ISSN pending.

Privacy-Preserving Data Publishing

An Overview

Raymond Chi-Wing Wong

The Hong Kong University of Science and Technology

Ada Wai-Chee Fu

The Chinese University of Hong Kong

SYNTHESIS LECTURES ON DATA MANAGEMENT #3

ABSTRACT

Privacy preservation has become a major issue in many data analysis applications. When a data set is released to other parties for data analysis, privacy-preserving techniques are often required to reduce the possibility of identifying sensitive information about individuals. For example, in medical data, sensitive information can be the fact that a particular patient suffers from HIV. In spatial data, sensitive information can be a specific location of an individual. In web surfing data, the information that a user browses certain websites may be considered sensitive. Consider a dataset containing some sensitive information is to be released to the public. In order to protect sensitive information, the simplest solution is not to disclose the information. However, this would be an overkill since it will hinder the process of data analysis over the data from which we can find interesting patterns. Moreover, in some applications, the data must be disclosed under the government regulations. Alternatively, the data owner can first modify the data such that the modified data can guarantee privacy and, at the same time, the modified data retains sufficient utility and can be released to other parties safely. This process is usually called as privacy-preserving data publishing. In this monograph, we study how the data owner can modify the data and how the modified data can preserve privacy and protect sensitive information.

KEYWORDS

privacy preservation, data publishing, anonymity, data mining

Contents

1	Introduction	1
1.1	Data Publishing	1
1.2	Significance	3
1.3	Organization	4
2	Fundamental Concepts	7
2.1	Anonymization	9
2.2	Information Loss Metric	18
2.3	Privacy Models	19
2.4	Other Privacy Models	21
2.5	Conclusion	28
3	One-Time Data Publishing	29
3.1	Knowledge about Quasi-identifiers	29
3.2	Knowledge about the Distribution of Sensitive Values	40
3.3	Knowledge about the Linkage of Individuals to Sensitive Values	44
	3.3.1 Information That Some Individuals Do Not Have Some Sensitive Values	44
	3.3.2 Information That Some Individuals Have Some Sensitive Values	45
3.4	Knowledge about the Relationship among Individuals	46
3.5	Knowledge about Anonymization	47
3.6	Knowledge Mined from the Microdata	52
3.7	Knowledge Mined from the Published Data	58
3.8	How To Use Published Data	60
	3.8.1 Aggregate Queries	61
	3.8.2 Information Loss	62

3.8.3	Evaluation with Data Mining and Data Analysis Tools	64
3.8.4	Querying over an Uncertain Database	65
3.9	Conclusion	66
4	Multiple-Time Data Publishing	69
4.1	Individual-Based Correlation	69
4.1.1	Data publishing from Static Microdata	69
4.1.2	Data Publishing from Dynamic Microdata	73
4.2	Sensitive Value-Based Correlation	80
4.2.1	Protection for Permanent Sensitive Values	83
4.2.2	Protection for Transient Sensitive Values	85
4.3	Conclusion	87
5	Graph Data	89
5.1	Data Model	89
5.2	Adversary Attacks	90
5.2.1	Assumption of Adversary Knowledge	90
5.2.2	Active Attacks	92
5.3	Utility of the Published Data	93
5.4	k -Anonymity	94
5.4.1	Vertex Degree	94
5.4.2	1-Neighborhood	95
5.4.3	Vertex Partitioning	96
5.4.4	k -Automorphism	98
5.5	Multiple Releases of Data Graphs	100
5.6	Other Approaches	101
5.7	Future Directions	101
6	Other Data Types	103
6.1	Spatial Data	103
6.1.1	With Anonymizer	104
6.1.2	Without Anonymizer	107

6.2	Transactional Data	108
6.3	Conclusion	111
7	Future Research Directions	113
7.1	One-Time Data Publishing	113
7.2	Multiple-Time Data Publishing	114
7.3	Publishing Graph Data	114
7.4	Publishing Data of Other Forms	115
A	Definition of Entropy l-Diversity and Recursive l-Diversity	117
	Authors' Biographies	127