

# Anomaly Detection as a Service

## Challenges, Advances, and Opportunities

# Synthesis Lectures on Information Security, Privacy, and Trust

## Editors

**Elisa Bertino**, *Purdue University*

**Ravi Sandhu**, *University of Texas, San Antonio*

The Synthesis Lectures Series on Information Security, Privacy, and Trust publishes 50- to 100-page publications on topics pertaining to all aspects of the theory and practice of Information Security, Privacy, and Trust. The scope largely follows the purview of premier computer security research journals such as ACM Transactions on Information and System Security, IEEE Transactions on Dependable and Secure Computing and Journal of Cryptology, and premier research conferences, such as ACM CCS, ACM SACMAT, ACM AsiaCCS, ACM CODASPY, IEEE Security and Privacy, IEEE Computer Security Foundations, ACSAC, ESORICS, Crypto, EuroCrypt and AsiaCrypt. In addition to the research topics typically covered in such journals and conferences, the series also solicits lectures on legal, policy, social, business, and economic issues addressed to a technical audience of scientists and engineers. Lectures on significant industry developments by leading practitioners are also solicited.

## Anomaly Detection as a Service: Challenges, Advances, and Opportunities

Danfeng (Daphne) Yao, Xiaokui Shu, Long Cheng, and Salvatore J. Stolfo  
2017

## Cyber-Physical Security and Privacy in the Electric Smart Grid

Bruce McMillin and Thomas Roth  
2017

## Blocks and Chains: Introduction to Bitcoin, Cryptocurrencies, and Their Consensus Mechanisms

Aljosha Judmayer, Nicholas Stifter, Katharina Krombholz, and Edgar Weippl  
2017

## Digital Forensic Science: Issues, Methods, and Challenges

Vassil Roussev  
2016

### Differential Privacy: From Theory to Practice

Ninghui Li, Min Lyu, Dong Su, and Weining Yang  
2016

### Privacy Risk Analysis

Sourya Joyee De and Daniel Le Métayer  
2016

### Introduction to Secure Outsourcing Computation

Xiaofeng Chen  
2016

### Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections

Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas  
2016

### Automated Software Diversity

Per Larsen, Stefan Brunthaler, Lucas Davi, Ahmad-Reza Sadeghi, and Michael Franz  
2015

### Trust in Social Media

Jiliang Tang and Huan Liu  
2015

### Physically Unclonable Functions (PUFs): Applications, Models, and Future Directions

Christian Wachsmann and Ahmad-Reza Sadeghi  
2014

### Usable Security: History, Themes, and Challenges

Simson Garfinkel and Heather Richter Lipford  
2014

### Reversible Digital Watermarking: Theory and Practices

Ruchira Naskar and Rajat Subhra Chakraborty  
2014

### Mobile Platform Security

N. Asokan, Lucas Davi, Alexandra Dmitrienko, Stephan Heuser, Kari Kostiaainen, Elena Reshetova, and Ahmad-Reza Sadeghi  
2013

### Security and Trust in Online Social Networks

Barbara Carminati, Elena Ferrari, and Marco Viviani  
2013

### RFID Security and Privacy

Yingjiu Li, Robert H. Deng, and Elisa Bertino  
2013

### Hardware Malware

Christian Krieg, Adrian Dabrowski, Heidelinde Hobel, Katharina Krombholz, and Edgar Weippl  
2013

### Private Information Retrieval

Xun Yi, Russell Paulet, and Elisa Bertino  
2013

### Privacy for Location-based Services

Gabriel Ghinita  
2013

### Enhancing Information Security and Privacy by Combining Biometrics with Cryptography

Sanjay G. Kanade, Dijana Petrovska-Delacrétaz, and Bernadette Dorizzi  
2012

### Analysis Techniques for Information Security

Anupam Datta, Somesh Jha, Ninghui Li, David Melski, and Thomas Reps  
2010

### Operating System Security

Trent Jaeger  
2008

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2018

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.

Anomaly Detection as a Service: Challenges, Advances, and Opportunities

Danfeng (Daphne) Yao, Xiaokui Shu, Long Cheng, and Salvatore J. Stolfo

ISBN: 978-3-031-01226-6 paperback

ISBN: 978-3-031-02354-5 ebook

DOI 10.1007/978-3-031-02354-5

A Publication in the Springer series

*SYNTHESIS LECTURES ON INFORMATION SECURITY, PRIVACY, AND TRUST*

Lecture #22

Series Editors: Elisa Bertino, *Purdue University*

Ravi Sandhu, *University of Texas, San Antonio*

Series ISSN

Print 1945-9742 Electronic 1945-9750

Parts of this book are based on the PhD theses of Xiaokui Shu [244], Kui Xu [300], Gabriela F. Cretu-Ciocarlie [65], Hao Zhang [317], Karim O. Elish [77], and the paper [53].

# Anomaly Detection as a Service

## Challenges, Advances, and Opportunities

Danfeng (Daphne) Yao  
Virginia Tech

Xiaokui Shu  
IBM Research

Long Cheng  
Virginia Tech

Salvatore J. Stolfo  
Columbia University

*SYNTHESIS LECTURES ON INFORMATION SECURITY, PRIVACY, AND  
TRUST #22*

## ABSTRACT

Anomaly detection has been a long-standing security approach with versatile applications, ranging from securing server programs in critical environments, to detecting insider threats in enterprises, to anti-abuse detection for online social networks. Despite the seemingly diverse application domains, anomaly detection solutions share similar technical challenges, such as how to accurately recognize various normal patterns, how to reduce false alarms, how to adapt to concept drifts, and how to minimize performance impact. They also share similar detection approaches and evaluation methods, such as feature extraction, dimension reduction, and experimental evaluation.

The main purpose of this book is to help advance the real-world adoption and deployment anomaly detection technologies, by systematizing the body of existing knowledge on anomaly detection. This book is focused on data-driven anomaly detection for software, systems, and networks against advanced exploits and attacks, but also touches on a number of applications, including fraud detection and insider threats. We explain the key technical components in anomaly detection workflows, give in-depth description of the state-of-the-art data-driven anomaly-based security solutions, and more importantly, point out promising new research directions. This book emphasizes on the need and challenges for deploying service-oriented anomaly detection in practice, where clients can outsource the detection to dedicated security providers and enjoy the protection without tending to the intricate details.

## KEYWORDS

anomaly detection, data driven, proactive defense, program and software security, system and network security, outsource, anomaly detection as a service, deployment, data science, classification, machine learning, novelty detection, program analysis, control flow, data flow, semantic gap, inference and reasoning, code-reuse attack, data-oriented attack, advanced persistent threat, zero-day exploit, system tracing, hardware tracing, false negative, false positive, performance, usability, insider threat, fraud detection, cyber intelligence, automation, democratization of technology, Linux, Android, x86, ARM

# Contents

<b>Preface</b> .....	<b>xiii</b>
<b>Acknowledgments</b> .....	<b>xv</b>
<b>1 Introduction</b> .....	<b>1</b>
1.1 Applications of Anomaly Detection .....	1
1.2 Cohen's Impossibility Results .....	3
1.3 Zero-day Exploits and APT .....	4
1.4 Challenges of Democratizing Anomaly Detection Technologies .....	6
1.5 Major Developments on Program Anomaly Detection .....	7
1.6 New Opportunities .....	9
<b>2 Threat Models</b> .....	<b>13</b>
2.1 Faults vs. Attacks and Safety vs. Security .....	13
2.2 Data-oriented Attacks .....	14
2.3 Insider Threats and Inadvertent Data Leaks .....	16
2.4 Attacks on Control Flows .....	17
2.5 Mimicry Attacks .....	18
2.6 Segment Length and Mimicry Attack Difficulty .....	19
<b>3 Local vs. Global Program Anomaly Detection</b> .....	<b>21</b>
3.1 One Big Model vs. Multiple Small Models .....	21
3.1.1 Modeling Byte Distributions .....	21
3.1.2 Multiple Clusters for Multiple Behaviors .....	22
3.1.3 Suitability Test .....	24
3.2 Local Anomaly Detection .....	24
3.2.1 $n$ -gram .....	25
3.2.2 Hidden Markov Model (HMM) .....	26
3.2.3 Finite-state Automaton (FSA) .....	27
3.3 Global Anomaly Detection .....	27
3.3.1 Examples of Global Anomalies and Detection Attempts .....	28



	3.3.2 Segmentation and Representing Infinite Traces .....	30
	3.3.3 Inter-cluster and Intra-cluster Anomalies .....	32
<b>4</b>	<b>Program Analysis in Data-driven Anomaly Detection .....</b>	<b>37</b>
4.1	Security Impact of Incomplete Training Data .....	37
4.2	Program Analysis For Guiding Classifiers .....	39
4.2.1	Quantifying Control-flow Graph .....	40
4.2.2	Interfacing with Markov Model .....	42
4.2.3	Improving Context Sensitivity .....	43
4.3	Program Analysis For Android Malware Detection .....	44
4.3.1	Android Threat Model and National Security .....	45
4.3.2	Data-dependence Graph and Android Malware Examples .....	45
4.3.3	User-trigger Dependence-based Detection .....	47
4.4	Formal Language Model for Anomaly Detection .....	48
<b>5</b>	<b>Anomaly Detection in Cyber-Physical Systems .....</b>	<b>51</b>
5.1	CPS Security Challenges .....	51
5.1.1	Background on CPS .....	51
5.1.2	Security and the Physical World .....	52
5.2	Overview of CPS Anomaly Detection .....	54
5.3	Event-aware Anomaly Detection (EAD) Framework .....	57
5.3.1	Data-oriented Attacks on CPS .....	57
5.3.2	Reasoning Cyber-Physical Execution Semantics .....	59
5.4	Event-aware Finite-state Automaton for CPS .....	62
5.4.1	Definition of eFSA .....	62
5.4.2	Event-aware Detection in eFSA .....	63
5.5	Evaluation of Control-branch and Control-intensity Detection .....	63
5.6	Deployment of CPS Anomaly Detection .....	64
<b>6</b>	<b>Anomaly Detection on Network Traffic .....</b>	<b>67</b>
6.1	Threats of Clandestine Network Activities .....	67
6.2	Sensemaking of Network Traffic For Anomaly Detection .....	68
6.2.1	Extrusion Detection in BINDER and its Generalization .....	68
6.2.2	Multi-host Causality and Reasoning .....	69
6.2.3	Collaborative Sensemaking .....	70
6.3	Definition of Triggering-relation Discovery .....	71
6.4	Discovery of Triggering-relation Graphs for Host Security .....	73
6.5	Sparsity of Triggering Relations and Cost Matrix .....	76

<b>7</b>	<b>Automation and Evaluation for Anomaly Detection Deployment</b>	<b>79</b>
7.1	Model Drift and Adapting Anomaly Detection to Changes	79
7.2	Sanitizing Training Data	81
7.2.1	Overview of Sanitization Approaches	82
7.2.2	Impact of Basic Sanitization	82
7.2.3	Impact of Collaborative Sanitization	83
7.3	Self-calibration and Gradual Retraining	85
7.3.1	Automatic Training Optimization	85
7.3.2	Automatic Threshold Selection	86
7.3.3	Performance Under Self-calibration	87
7.3.4	Gradual Retraining	89
7.4	Tracing Overhead and Intel PT	91
7.5	Experimental Evaluation for Data-driven Anomaly Detection	94
<b>8</b>	<b>Anomaly Detection from the Industry's Perspective</b>	<b>99</b>
8.1	Anomaly Detection in Payment Card Industry	99
8.2	Security Operation Centers (SOC)	100
8.3	Anomaly Detection in the Pyramid	102
8.4	Building Your Own Anomaly Detection Toolkit	104
8.5	Leveraging External Knowledge in Cyber Security Pyramid	106
<b>9</b>	<b>Exciting New Problems and Opportunities</b>	<b>109</b>
9.1	Deep Learning and Instruction-level Anomaly Detection	109
9.2	Post-detection Forensic, Repair, and Recovery	111
9.3	Anomaly Detection of Concurrency Attacks	111
9.4	Mimicry Generation, Insider Threat Detection, Automation, and Knowledge Base	113
	<b>Bibliography</b>	<b>117</b>
	<b>Authors' Biographies</b>	<b>149</b>
	<b>Index</b>	<b>151</b>

# Preface

Anomaly detection is one of the few proactive defense approaches. This book is intended to provide an introduction to anomaly-based security defense techniques with a focus on data-science based approaches. The book summarizes the history and the landscape of anomaly detection research, systematizes and contextualizes the existing solutions, explains how various components and techniques are connected and related to each other, and more importantly, points out the exciting and promising new research and development opportunities in data-driven anomaly detection. The book focuses on the anomaly detection in program executions and computer networks. It can be used as a textbook for advanced graduate courses, or undergraduate senior elective courses.

As the need for security is becoming an integral part of the society, we intend to make this book useful and accessible for a large audience, including cybersecurity professionals at all levels, data scientists, usability engineers, and various application-domain experts. Achieving cyber security depends on inter-disciplinary research and development efforts.

The book is titled *Anomaly Detection as a Service*. It is a grand and ambitious vision that has yet to become reality. Throughout the book, we discuss how current technologies could be extended to achieve anomaly detection as a service and the gaps to be filled. With the unprecedented advances on data science and growing interests from both the academia and industry on anomaly detection, the timing for pushing for this vision could not be any better. We hope this book can encourage and engage researchers, practitioners, and vendors in anomaly-detection related innovations.

The book is organized as follows. The first three chapters introduce the anomaly detection fundamentals. The next four chapters dive into key technical areas, including program analysis, cyber-physical systems, sensemaking, and automation. The last two chapters describe industry development and future opportunities.

A brief summary of each chapter is as follows. In Chapter 1, we give an overview of the field of anomaly detection with a focus on the past, present, and future of program anomaly detection. We also describe the vision of anomaly detection as a service. In Chapter 2, we point out the importance of defining threat models in anomaly detection, and introduce major attack categories against programs, as well as the attacks against detection systems. In Chapter 3, we describe basic techniques for modeling program behaviors, and explain the differences between local anomaly detection and global anomaly detection.

In Chapter 4, we show various ways that insights from code analysis can substantially improve data-driven anomaly detection, including Android malware detection. In Chapter 5, we show how to reason about control-program semantics with respect to the physical environment,

which is important for protecting cyber-physical systems. In Chapter 6, we describe several network anomaly detection methods that are all based on making sense of massive amounts of network traffic. In Chapter 7, we show the technical advances in automating  $n$ -gram based detection, including automatic calibration, adjustment, and maintenance. In addition, we point out the key requirements for conducting rigorous experimental evaluation of data-driven anomaly detection.

In Chapter 8, we give an overview of anomaly detection technologies in the security industry and point out the anomaly-detection components in various commercial products. In the last Chapter 9, we point out several exciting new research and development opportunities that will help realize the vision of the anomaly detection as a service.

Danfeng (Daphne) Yao, Xiaokui Shu, Long Cheng, and Salvatore J. Stolfo  
October 2017

# Acknowledgments

I would like to thank Elisa Bertino for giving us this wonderful opportunity to write a book on a topic that I love. We are honored to become authors contributing to Elisa and Ravi Sandhu's series *Synthesis Lectures on Information Security, Privacy, and Trust*.

I would also like to thank my husband Chang, my daughter Radia, and my mother Yuzhu for being so supportive for me while I take on this book project and work on it almost endlessly.

I would like to thank my coauthors for being extremely generous with their time and effort on the book. I also thank Diane Cerra, our executive editor at Morgan & Claypool, for her patience and help. Finally, I would like to thank Office of Naval Research (Grant N00014-17-1-2498), National Science Foundation, and Army Research Office for their funding support.

Danfeng (Daphne) Yao  
October 2017