# Intelligent Systems Reference Library

Volume 177

**Series Editors**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
Lakhmi C. Jain, Faculty of Engineering and Information Technology, Centre for Artificial Intelligence, University of Technology, Sydney, NSW, Australia;
KES International, Shoreham-by-Sea, UK;
Liverpool Hope University, Liverpool, UK

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

** Indexing: The books of this series are submitted to ISI Web of Science, SCOPUS, DBLP and Springerlink.

More information about this series at http://www.springer.com/series/8578

Leslie F. Sikos · Kim-Kwang Raymond Choo
Editors

# Data Science in Cybersecurity and Cyberthreat Intelligence

Springer

*Editors*
Leslie F. Sikos
School of Science
Edith Cowan University
Joondalup, WA, Australia

Kim-Kwang Raymond Choo
Department of Information Systems
and Security
University of Texas at San Antonio
San Antonio, TX, USA

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-030-38787-7 ISBN 978-3-030-38788-4 (eBook)
https://doi.org/10.1007/978-3-030-38788-4

© Springer Nature Switzerland AG 2020
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.
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.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Obtaining accurate information about online activities in near-real time is becoming increasingly difficult, particularly because of the constantly increasing data volume, complexity, variety, and veracity, as well as the unscalability of traditional data processing algorithms. Data science is an emerging research field that can provide solutions to these challenges and addresses some of the issues related to them, such as data heterogeneity, ambiguity, trustworthiness, and reliability. Aggregating data from diverse sources for security professionals' and decision-makers' dashboards and fusing data derived from different, often proprietary formats became commonplace; yet challenges in terms of preventive cybersecurity measures, the automation of complex behavior matching for detecting unknown threats, efficiently processing ever-growing signature databases, and many more remain. In addition, user expectations are higher and legislation is stricter than ever before—think of the Algorithmic Accountability Act of 2019 in the US, for example, which requires the evaluation of automated systems relying on machine learning for any potential bias.

Chapter 1 provides an overview of formal knowledge representation, a branch of AI, and how it can be utilized in cyberthreat intelligence for automatically classifying cyberthreats based on the attack technique or the threat impact, and using hybrid models. Standards that can be used for the formal representation and efficient exchange of cyberthreat intelligence are also discussed, including purpose-built taxonomies and ontologies, and how automated reasoning can be performed on datasets that utilize these. In Chap. 2, a state-of-the-art logic programming approach is presented, which was designed to predict enterprise-targeted cyberattacks. This approach aggregates hacker discussions trends from multiple hacker community websites to predict future cyberattack incidents, with a promising precision–recall trade-off. Chapter 3 explains how malicious and DGA-generated URLs can be identified using machine learning techniques and rule-based approaches, and reviews tools and data sources utilized by these approaches. In addition, a lightweight framework is proposed to identify previously unknown malicious URLs while minimizing the need for manual label creation in the training dataset. Chapter 4 demonstrates the implementation of PCA, SVM, kNN, linear regression,

two-layer perceptron, decision tree, and Gaussian naïve Bayes in intrusion detection systems. In this context, dataset analysis is considered object classification, which is described formally. Chapter 5 looks into the standards and protocols used for securing wearable mHealth devices. It systematically reviews possible attack types by network layer, and details the security requirements of mHealth devices. It justifies the need for data inference not only because of the dramatic reduction of data transmission frequency (thereby maximizing battery power), but also because reducing the volume of transmitted data reduces the number of potential attacks. Based on the industry experience of the authors, Chap. 6 discusses the main pitfalls of utilizing data science in cybersecurity, covering data source issues, feature engineering challenges, the importance of the data source, metric selection, considerations for choosing an algorithm, and algorithm convergence.

While data science is no longer limited to business intelligence and analysis, one of its most common application areas is extracting valuable insights from business data and market trends. However, what is less known but can be learned from this book is that data science is already well-utilized and has a huge potential in security applications, such as for enhancing predictive measures for vulnerability exploitation and intrusion detection, identifying nontrivial network traffic patterns, and performing automated reasoning over security datasets. These can be useful when designing new security standards, frameworks, and protocols, and when implementing data science approaches in cybersecurity applications.

Perth, WA, Australia Leslie F. Sikos, Ph.D.

# Contents

**1 The Formal Representation of Cyberthreats for Automated Reasoning** . . . . . 1
Leslie F. Sikos
1.1 Introduction to Knowledge Organization in and Modeling of Cyberthreat Intelligence . . . . . 1
1.2 Threat Classification . . . . . 2
1.2.1 Attack Technique-Based Threat Classification . . . . . 2
1.2.2 Threat Models for Threat Impact-Based Classification . . . . 3
1.2.3 Hybrid Models . . . . . 3
1.3 Representing and Exchanging Cyberthreat Intelligence . . . . . 5
1.3.1 Cyberthreat Taxonomies . . . . . 5
1.3.2 Cyberthreat Ontologies . . . . . 8
1.3.3 Utilizing the Formal Representation of Information Traversing Communication Networks in Cyberthreat Intelligence . . . . . 8
1.4 Automated Reasoning over Formally Represented Threat Knowledge . . . . . 9
1.5 Summary . . . . . 10
References . . . . . 10

**2 A Logic Programming Approach to Predict Enterprise-Targeted Cyberattacks** . . . . . 13
Mohammed Almukaynizi, Ericsson Marin, Malay Shah, Eric Nunes, Gerardo I. Simari and Paulo Shakarian
2.1 Introduction . . . . . 14
2.2 Related Works . . . . . 15
2.3 Preliminaries . . . . . 17
2.3.1 Syntax . . . . . 17
2.3.2 Semantics . . . . . 18

2.4 Desired Technical Properties . . . . . 19
2.5 A Novel Logic Programming-Based Cyberthreat Prediction System . . . . . 20
2.5.1 Learner . . . . . 20
2.5.2 Predictor . . . . . 21
2.6 Data Description . . . . . 22
2.6.1 Ground Truth . . . . . 22
2.6.2 Hacker Community Discussions . . . . . 22
2.7 Extracting Indicators of Cyberthreat . . . . . 23
2.7.1 CVE to CPE Mappings . . . . . 23
2.7.2 Extracting Entity Tags . . . . . 24
2.8 Predicting Enterprise-Targeted Attacks . . . . . 25
2.8.1 Setup . . . . . 25
2.8.2 Evaluation . . . . . 25
2.8.3 Results . . . . . 26
2.9 Conclusion . . . . . 28
References . . . . . 30

**3 Discovering Malicious URLs Using Machine Learning Techniques** . . . . . 33
Bo Sun, Takeshi Takahashi, Lei Zhu and Tatsuya Mori
3.1 Introduction . . . . . 33
3.2 Related Works . . . . . 35
3.2.1 Malicious URL Detection . . . . . 35
3.2.2 DGA Domain Detection . . . . . 37
3.3 Tools and Data Sources . . . . . 38
3.3.1 Web Client Honeypots . . . . . 38
3.3.2 Web Crawlers . . . . . 39
3.3.3 URL Datasets . . . . . 39
3.3.4 Passive DNS Database . . . . . 40
3.3.5 Search Engines . . . . . 40
3.4 Machine Learning Techniques . . . . . 41
3.4.1 Bayesian Sets . . . . . 41
3.4.2 Other Machine Learning Algorithms . . . . . 43
3.5 AutoBLG Framework . . . . . 43
3.5.1 High-Level Overview . . . . . 43
3.5.2 URL Expansion . . . . . 44
3.5.3 URL Filtering . . . . . 46
3.5.4 URL Verification . . . . . 48
3.6 Evaluation . . . . . 48
3.6.1 Preliminary Experiment . . . . . 48
3.6.2 Performance of the AutoBLG Framework . . . . . 50
3.6.3 Comparisons . . . . . 52

3.7 Discussion . . . . . 52
3.7.1 URL Expansion . . . . . 53
3.7.2 Limitation of Query Patterns . . . . . 54
3.7.3 URL Verification . . . . . 54
3.7.4 Online Operation . . . . . 54
3.8 Conclusion . . . . . 55
3.9 Appendix . . . . . 55
References . . . . . 59

**4 Machine Learning and Big Data Processing for Cybersecurity Data Analysis** . . . . . 61
Igor Kotenko, Igor Saenko and Alexander Branitskiy
4.1 Introduction . . . . . 61
4.2 Related Works . . . . . 63
4.3 Machine Learning Methods . . . . . 65
4.4 Datasets, Architectures, and Experiments . . . . . 69
4.4.1 Detection of Attacks Against IoT Structure . . . . . 69
4.4.2 Detection of Host Scanning and DDoS Attacks . . . . . 75
4.5 Conclusion . . . . . 82
References . . . . . 83

**5 Systematic Analysis of Security Implementation for Internet of Health Things in Mobile Health Networks** . . . . . 87
James Jin Kang
5.1 Introduction . . . . . 87
5.1.1 WSN . . . . . 89
5.1.2 WBAN . . . . . 89
5.1.3 PT . . . . . 89
5.1.4 MC . . . . . 90
5.1.5 CT . . . . . 90
5.2 Threats and Attacks . . . . . 90
5.2.1 Layer 1 (Physical Layer) . . . . . 92
5.2.2 Layer 2 (Data Link Layer) . . . . . 93
5.2.3 Layer 3 (Network Layer) . . . . . 93
5.2.4 Layer 4 (Transport Layer) . . . . . 94
5.2.5 Layers 5, 6, and 7 (Session, Presentation, and Application Layers) . . . . . 94
5.3 Security Requirements for mHealth Devices . . . . . 94
5.3.1 Confidentiality . . . . . 94
5.3.2 Data Integrity . . . . . 95
5.3.3 Availability . . . . . 95
5.3.4 Privacy Policy . . . . . 95
5.3.5 Data Inference . . . . . 96

5.4 Optimization of Security Mechanisms for mHealth . . . . . 98
5.4.1 Authentication . . . . . 98
5.4.2 Authorization . . . . . 99
5.4.3 IEEE 802.15.6 WBAN Security Protocol . . . . . 99
5.4.4 Key Management Protocols . . . . . 100
5.4.5 Route Discovery Protocols . . . . . 102
5.4.6 NSA Suite B and E Algorithm . . . . . 102
5.4.7 Application-Specific Security Mechanisms . . . . . 104
5.5 Communication Protocols of WBAN . . . . . 104
5.5.1 ANT/ANT+ . . . . . 104
5.5.2 ZigBee . . . . . 105
5.5.3 Bluetooth/(BT-LE) . . . . . 105
5.5.4 IEEE P11073 PHD Protocol . . . . . 107
5.6 Future Areas of Study . . . . . 108
5.6.1 Security and Quality of Sensor Devices . . . . . 109
5.6.2 Privacy . . . . . 109
5.6.3 Security Measures in PHD . . . . . 109
5.6.4 Compatibility and Standardization of Security Protocols Versus Application Protocols . . . . . 110
5.6.5 Unique Identifiers as an Authentication Mechanism . . . . . 110
5.6.6 User Centeredness . . . . . 110
5.6.7 Intelligent Sensors with QoS . . . . . 111
5.7 Conclusions . . . . . 111
References . . . . . 112

**6 Seven Pitfalls of Using Data Science in Cybersecurity** . . . . . 115
Mike Johnstone and Matt Peacock
6.1 Introduction . . . . . 115
6.2 Data Source . . . . . 118
6.2.1 Synthetic Data . . . . . 119
6.2.2 Real-World Data . . . . . 120
6.3 Feature Engineering . . . . . 120
6.4 Evaluation Metrics . . . . . 122
6.5 Algorithm Selection . . . . . 124
6.6 Algorithm Convergence . . . . . 125
6.7 Algorithm Poisoning . . . . . 126
6.8 Conclusion . . . . . 128
References . . . . . 128

# About the Editors

**Leslie F. Sikos, Ph.D.** is a computer scientist specializing in network forensics and cybersecurity applications powered by artificial intelligence and data science. He has worked in both academia and the industry, and acquired hands-on skills in datacenter and cloud infrastructures, cyberthreat management, and firewall configuration. He regularly contributes to major cybersecurity projects in collaboration with the Defence Science and Technology Group of the Australian Government, CSIRO's Data61, and the CyberCRC. He is a reviewer of journals such as Computers & Security and Crime Science, and chairs sessions at international conferences on AI in cybersecurity. Dr. Sikos holds professional certificates and is a member of industry-leading organizations, such as the ACM, the IEEE Special Interest Group on Big Data for Cyber Security and Privacy, and the IEEE Computer Society Technical Committee on Security and Privacy.
https://www.lesliesikos.com

**Kim-Kwang Raymond Choo** received Ph.D. in Information Security in 2006 from Queensland University of Technology, Australia. He currently holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio (UTSA), and has a courtesy appointment at the University of South Australia. In 2016, he was named the Cybersecurity Educator of the Year—APAC (Cybersecurity Excellence Awards are produced in cooperation with the Information Security Community on LinkedIn), and in 2015 he and his team won the Digital Forensics Research Challenge organized by Germany's University of Erlangen-Nuremberg. He is the recipient of the 2019 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence in Scalable Computing (Middle Career Researcher), 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty, Outstanding Associate Editor of 2018 for IEEE Access, British Computer Society's 2019 Wilkes Award Runner-up, 2019 EURASIP Journal on Wireless Communications and Networking (JWCN) Best Paper Award, Korea Information Processing Society's Journal of Information Processing Systems (JIPS) Survey Paper Award (Gold) 2019, IEEE Blockchain 2019 Outstanding Paper Award, IEEE TrustCom 2018 Best Paper Award, ESORICS 2015 Best Research Paper Award, 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency, Fulbright Scholarship in 2009, 2008 Australia Day Achievement Medallion, and British Computer Society's Wilkes Award in 2008. He is also a Fellow of the Australian Computer Society, an IEEE Senior Member, and Co-Chair of IEEE Multimedia Communications Technical Committee's Digital Rights Management for Multimedia Interest Group.