Data Analytics for Cybersecurity

As the world becomes increasingly connected, it is also more exposed to a myriad of cyber threats. We need to use multiple types of tools and techniques to learn and understand the evolving threat landscape. Data are a common thread linking various types of devices and end users. Analyzing data across different segments of cybersecurity domains, particularly data generated during cyber–attacks, can help us understand threats better, prevent future cyber–attacks, and provide insights into the evolving cyber threat landscape. This book takes a data oriented approach to studying cyber threats, showing in depth how traditional methods such as anomaly detection can be extended using data analytics, and also applies data analytics to non–traditional views of cybersecurity, such as multi domain analysis, time series and spatial data analysis, and human-centered cybersecurity.

VANDANA P. JANEJA is Professor and Chair of the Information Systems department at the University of Maryland, Baltimore County. Most recently, she also served as an expert at the National Science Foundation supporting data science activities in the Directorate for Computer and Information Science and Engineering (CISE) (2018–2021). Her research interests include discovering knowledge in presence of data heterogeneity. Her research projects include anomaly detection in network communication data, human behavior analytics in heterogeneous device environments, geospatial context for IP reputation scoring, spatiotemporal analysis across heterogeneous data, and ethical thinking in data science. She has been funded through state, federal, and private organizations.

Data Analytics for Cybersecurity

VANDANA P. JANEJA
University of Maryland, Baltimore County (UMBC)

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108415279
DOI: 10.1017/9781108231954

First published 2022

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-41527-9 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Cambridge University Press & Assessment
978-1-108-41527-9 — Data Analytics for Cybersecurity
Vandana P. Janeja
Frontmatter
More Information

Contents

Cambridge University Press & Assessment
978-1-108-41527-9 — Data Analytics for Cybersecurity
Vandana P. Janeja
Frontmatter
More Information

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-41527-9 — Data Analytics for Cybersecurity
Vandana P. Janeja
Frontmatter
More Information

Preface

Cybersecurity is a pervasive need in our connected world to counter threats affecting individuals, organizations, and governments. The acceptance and adoption of technology on multiple types of nontraditional devices force cybersecurity solutions to address challenges emerging in the areas of not only computer networks but also sensor networks, industrial control systems, and user devices. Data are the common thread across all these types of devices and end users, especially data generated during cyberattacks. Increasingly, the focus of cybersecurity is shifting to analyzing data in not only a retrospective manner but also a prospective manner across different cybersecurity domains. This data-driven understanding of attacks can potentially prevent future cyberattacks and provide insights into the evolving cyber threats.

Data analytics pushes beyond the traditional themes of security to seamlessly weave the analysis of threats across several applications. This book applies data analytics concepts and techniques to the domain of cybersecurity, discusses methods to evaluate data sources in cyberattacks, and provides insights into data mining methods that can be utilized for cybersecurity. Finally, this book also looks at nontraditional views of data analytics for cybersecurity in time series and spatial data, discussing the need and application of big data analytics and offering a human-centered analytics perspective to cybersecurity.

Although there are several books on network security, information assurance, and forensics and other books on data mining, there is a need to address both cybersecurity and data analytics in a synergistic manner. In addition, I have developed a graduate and undergraduate course on data analytics for cybersecurity. I taught this course for several semesters without a textbook. While developing this course, I reviewed several very good books. However, there was a gap in the material available for the data analytics perspective of cybersecurity in depth. This motivated me to write a book that looks at the

Cambridge University Press & Assessment
978-1-108-41527-9 — Data Analytics for Cybersecurity
Vandana P. Janeja
Frontmatter
More Information

different domains affected by cyberattacks, spanning computer networks, industrial control systems, sensor networks, drones, and other connected devices. Data analytics provides a window of learning into such systems by looking at the massive amounts of data being generated that may go untapped. For instance, this book addresses the human-centered perspective to cyberattacks, and it addresses the multiple facets of data analytics for cybersecurity, such as anomaly detection across spatial and temporal data.

I expect the book will be primarily useful for teaching graduate and undergraduate cybersecurity courses that take a data analytics perspective. It should also be relevant for the industry and government as it discusses the potential avenues of understanding and discovering cyberattacks and additional knowledge about them to better inform future decision-making.

The need for data analytics is also evident from Information Assurance requirements such as those stated in federal frameworks[1,2] and security directives for Information Assurance Training, Certification, and Workforce Management,[3] where various technical cybersecurity positions and their functions are outlined in detail. For instance, the Computer Network Defense (CND) Service Provider Incident Responder (IR) position has functions such as (CND-IR.2) – collect, analyze, intrusion artifacts – and (CND-IR.7) – correlate incident data and perform CND trend analysis. Similarly, the CND-Analyst (A) position has functions such as (CND-A.4) – perform analysis of log files – (CND-A.5) – characterize and analyze network traffic to identify anomalous activity and potential threats to network resources – and (CND-A.8.) – perform event correlation. These and many other security functions, such as forensics, threat hunting, and such clearly indicate the need to incorporate an analytics perspective to cybersecurity.

What Does the Book Cover?

The book spans from introductory concepts of cybersecurity, foundations of data analytics, and applications of data analytics concepts to cybersecurity applications.

[1] Workforce Framework for Cybersecurity(NICE Framework), NIST Special Publication 800-181 Revision 1, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-181r1.pdf, Last Accessed May 2021

[2] Workforce Framework for Cybersecurity (NICE Framework), https://niccs.cisa.gov/workforce-development/cyber-security-workforce-framework , Last Accessed May 2021

[3] Information Assurance Workforce Improvement Program, DoD 8570.01-M www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/857001m.pdf; 8140.01 reissues and renumbers DoDD 8570.01, Last Accessed May 2021

Chapter 1 introduces the basic concepts of cybersecurity and the data analytics perspective to cybersecurity. It lays out the areas of study and how data analytics should be a key part of the spectrum of cybersecurity solutions.

Chapter 2 focuses on understanding sources of cybersecurity data and the end-to-end opportunities for data collection. It goes onto discuss the sources of cybersecurity data and how multiple datasets can be leveraged in understanding cyber threats.

Chapter 3 gets into the techniques of data analytics focusing on the three pillars of data mining, namely clustering, classification, and association rule mining, and how each can be used for cybersecurity. This chapter can be seen as a crash course in data mining. It begins with an understanding of the overall knowledge discovery and data mining process models and follows the elements of the data life cycle. This chapter outlines foundational elements such as measures of similarity and of evaluation. It outlines the landscape of various algorithms in clustering and classification and frequent and rare patterns.

Chapter 4 focuses on the big data elements of cybersecurity, looking at the landscape of the big data technologies and the complexities of the different types of data, including spatial and graph data. It outlines examples in these complex data types and how they can be evaluated using data analytics. Chapter 5 highlights the various types of cyberattacks and how data analytics methods can potentially be used to analyze these attacks.

Chapter 6 and 7 holistically focus on anomaly detection. Chapter 6 focuses on what anomalies are and more specifically what anomalies are in the cybersecurity domain, and what some of the features of anomalies are. Chapter 7, on the other hand, focuses on techniques of detecting anomalies starting with some of the basic statistical techniques, going into data analytics techniques.

While Chapter 4 introduces the complex types of data, Chapter 8 delves into the specifics of spatial and temporal analytics with topics such as spatial neighborhood and temporal evolution of large amounts of network traffic data. Chapter 9 extends the ideas of complex data by looking into cybersecurity through network and graph data. Chapter 10 brings in the human-centered data analytics perspective to cybersecurity. Finally, Chapter 11 discusses several key directions, such as data analytics in cyberphysical systems, multidomain mining, machine Learning concepts such as deep Learning, generative adversarial networks, and challenges of model reuse. Last but not the least, the chapter closes with thoughts on ethical thinking in the data analytics process.

Acknowledgments

The views expressed in this book are my own and do not reflect those of the organizations I am affiliated with.

I want to thank Carole Sargent, who planted the seed for this book and helped me learn the ropes of the process. Many thanks to my editor at Cambridge University Press, Lauren Cowles, and the Cambridge University Press team, Amy He, Adam Kratoska, Johnathan Fuentes, Rebecca Grainger, Andy Saff, and Neena S. Maheen, who were patient with the many deadlines and generous in their time reviewing the book and its many facets. Many thanks to the reviewers of this book who helped improve it with their detailed feedback.

The work in this book is built on the foundation of my work with many of my students, and the many conversations with them have inspired the writing of this book. I would like to thank my PhD advisees, Josephine Namayanja, Chhaya Kulkarni, Sara Khanjani, Ira Winkler, Faisal Quader, Mohammad Alodadi, Ali Azari , Yanan Sun, Lei Shi, and Mike McGuire; my master's advisees, Henanksha Sainani, Prathamesh Walkikar, Suraksha Shukla, Vasundhara Misal, Anuja Kench, Akshay Grover, Jay Gholap, Javad Zabihi, Kundala Das, Prerna Mohod, Abdulrahman Alothaim, and Revathi Palanisamy; my graduate students who worked on research projects with me, Yuanyuan Feng, Ruman Tambe, Song Chen, Sandipan Dey, Tania Lobo, Monish Advani, Ahmed Aleroud, Abu Zaher Md Faridee, Sai Pallaprolu, and Justin Stauffer; and my undergraduate research students, Gabrielle Watson, Olisaemeka Okolo, Aminat Alabi, Mackenzie Harwood, Adrian Reyes, Jarrett Early, Alfonso Delayney, David Lewis, and Brian Lewis.

My thought process and contributions are also framed by work with advisors, collaborators, colleagues, and industry affiliates who have contributed to my research and thinking in the area of security and data analytics, including Lucy Erickson, Carolyn Seaman, Aryya Gangopadhyay, Anupam

Joshi, Susan Sterrett, Krishnan Chellakarai, Yatish Joshi, and Raghu Chintalapati. I also want to thank my advisors and mentors: Vijay Atluri, Nabil Adam, Jaideep Vaidya, Erwin Gianchandani, Keith Bowman, Philip Rous, Patrice McDermott, Erin Lavik, and Chaitan Baru. This work is also built through the support of Dr. Freeman A. Hrabowski III, who has always been generous with his time as my mentor.

Last but not the least, my heartfelt thanks go to my biggest allies and pillars of support: Neelu, Suhasini, Harish, Deiptii, Nisha, Viraj, and Mukul, and my single most avid cheerleader, Vihan, who kept encouraging me with timeline prompts to complete this book.