

## 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.

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

---

# 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](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9781108415279](http://www.cambridge.org/9781108415279)  
DOI: 10.1017/9781108231954

© Vandana P. Janeja 2022

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

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.

## Contents

---

|   |             |     |
|---|-------------|-----|
| <i>Preface</i>  | <i>page</i> | ix  |
| <i>Acknowledgment</i>   |             | xii |
| <b>1 Introduction: Data Analytics for Cybersecurity</b>   |             | 1   |
| 1.1 What Is Cybersecurity?  |             | 1   |
| 1.1.1 Assets Affected   |             | 2   |
| 1.1.2 Motivation, Risks, and Attaining Security   |             | 4   |
| 1.2 Handling Cyberattacks   |             | 6   |
| 1.2.1 Subareas of Cybersecurity   |             | 6   |
| 1.3 Data Analytics  |             | 8   |
| 1.3.1 Why Is Data Analytics Important<br>for Cybersecurity? A Case Study of Understanding<br>the Anatomy of an Attack |             | 9   |
| 1.3.2 How Can Data Analytics Help?  |             | 12  |
| <b>2 Understanding Sources of Cybersecurity Data</b>  |             | 14  |
| 2.1 End-to-End Opportunities for Data Collection  |             | 14  |
| 2.2 Sources of Cybersecurity Data   |             | 16  |
| 2.2.1 Log Data  |             | 17  |
| 2.2.2 Raw Payload Data  |             | 20  |
| 2.2.3 Network Topology Data   |             | 21  |
| 2.2.4 User System Data  |             | 23  |
| 2.2.5 Other Datasets  |             | 24  |
| 2.3 Integrated Use of Multiple Datasets   |             | 26  |
| 2.4 Summary of Sources of Cybersecurity Data  |             | 27  |

|          |   |    |
|----------|---|----|
| <b>3</b> | <b>Introduction to Data Mining: Clustering, Classification, and Association Rule Mining</b>                                 | 29 |
| 3.1      | Knowledge Discovery and Data Mining Process Models  | 29 |
| 3.2      | Data Preprocessing  | 31 |
| 3.2.1    | Data Cleaning   | 33 |
| 3.2.2    | Data Transformation and Integration   | 34 |
| 3.2.3    | Data Reduction  | 35 |
| 3.3      | Data Mining   | 36 |
| 3.3.1    | Measures of Similarity  | 37 |
| 3.3.2    | Measures of Evaluation  | 40 |
| 3.3.3    | Clustering Algorithms   | 42 |
| 3.3.4    | Classification  | 49 |
| 3.3.5    | Pattern Mining: Association Rule Mining   | 55 |
| <b>4</b> | <b>Big Data Analytics and Its Need for Cybersecurity: Advanced DM and Complex Data Types from Cybersecurity Perspective</b> | 60 |
| 4.1      | What Is Big Data?   | 60 |
| 4.2      | Big Data in Cybersecurity   | 61 |
| 4.3      | Landscape of Big Data Technologies  | 63 |
| 4.4      | Mahout and Spark Comparative Analysis   | 64 |
| 4.5      | Complex Nature of Data  | 66 |
| 4.5.1    | Nature of Data: Spatial Data  | 67 |
| 4.5.2    | Nature of Data: Graph Data  | 69 |
| 4.5.3    | Nature of Data: Other Types of Complex Data   | 69 |
| 4.6      | Where Does Analytics Fit in for Cybersecurity?  | 70 |
| 4.6.1    | Change Detection in Massive Traffic Datasets  | 70 |
| 4.6.2    | Multipronged Attacks  | 73 |
| 4.7      | Privacy and Security Issues in Big Data   | 75 |
| <b>5</b> | <b>Types of Cyberattacks</b>  | 78 |
| 5.1      | Types of Attacks  | 78 |
| 5.2      | Example: Social Engineering   | 80 |
| 5.2.1    | Computational Data Model for Social Engineering   | 81 |
| 5.3      | Example: Advanced Persistent Threat   | 84 |
| 5.3.1    | Attributes of APTs  | 85 |
| 5.3.2    | Data Analytics Methods for Persistent Threats   | 88 |
| <b>6</b> | <b>Anomaly Detection for Cybersecurity</b>  | 91 |
| 6.1      | What Are Anomalies?   | 91 |
| 6.2      | Context   | 93 |

|          |   |            |
|----------|---|------------|
| 6.3      | Motivating Example: BGP Hijacking   | 95         |
| 6.4      | Challenges in Understanding Anomalies                                       | 96         |
| 6.5      | Interpretation  | 98         |
| 6.6      | Treating Anomalies  | 99         |
| <b>7</b> | <b>Anomaly Detection Methods</b>  | <b>101</b> |
| 7.1      | Statistical Outlier Detection Tests   | 102        |
| 7.1.1    | Discordancy Tests   | 103        |
| 7.1.2    | IQR-Based Outlier Detection   | 104        |
| 7.2      | Density-Based Outlier Detection:<br>OPTICS-OF Identifying Local<br>Outliers | 105        |
| 7.3      | Distance-Based Outlier Detection  | 108        |
| 7.4      | Outlier Detection through Clustering  | 110        |
| <b>8</b> | <b>Cybersecurity through Time Series and<br/>Spatial Data</b>               | <b>112</b> |
| 8.1      | Spatial and Temporal Data   | 112        |
| 8.1.1    | Spatial Data  | 112        |
| 8.1.2    | Spatial Autocorrelation   | 113        |
| 8.1.3    | Spatial Heterogeneity   | 113        |
| 8.1.4    | Temporal Data   | 115        |
| 8.1.5    | Temporal Autocorrelation  | 115        |
| 8.1.6    | Temporal Heterogeneity  | 116        |
| 8.2      | Some Key Methods for Anomaly Detection in Spatial<br>and Temporal Data      | 116        |
| 8.2.1    | Spatial Anomaly Detection   | 117        |
| 8.2.2    | Spatial Outlier Detection   | 117        |
| 8.2.3    | Spatial Neighborhood  | 119        |
| 8.2.4    | Temporal Anomaly Detection  | 121        |
| 8.2.5    | Temporal Neighborhoods  | 122        |
| 8.3      | Cybersecurity through Spatiotemporal<br>Analysis                            | 123        |
| 8.3.1    | Spatial Anomalies in Cybersecurity  | 123        |
| 8.3.2    | Neighborhood-Based Anomalies  | 124        |
| 8.3.3    | Spatio temporally Clustered<br>Anomalies                                    | 125        |
| 8.3.4    | Insights through Geo-Origination<br>of IP Addresses                         | 126        |
| 8.4      | Temporal Behaviors in Evolving Networks                                     | 127        |

|           |  |     |
|-----------|--|-----|
| <b>9</b>  | <b>Cybersecurity through Network and Graph Data</b>                          | 128 |
| 9.1       | Graph Properties   | 130 |
| 9.2       | Graphs for Cybersecurity: Understanding Evolving Network Communication       | 133 |
| 9.3       | Graphs for Cybersecurity: Similarity-Based Redundancy in Router Connectivity | 134 |
| <b>10</b> | <b>Human-Centered Data Analytics for Cybersecurity</b>                       | 137 |
| 10.1      | Human Perspective to Cybersecurity: Phishing                                 | 138 |
| 10.2      | Human Perspective to Cybersecurity: Insider Threat                           | 140 |
| 10.3      | User/Employee Viewpoint  | 142 |
| 10.4      | Attacker Viewpoint: Anomaly Detection Methods                                | 144 |
| <b>11</b> | <b>Future Directions in Data Analytics for Cybersecurity</b>                 | 147 |
| 11.1      | Data Analytics in Cyberphysical Systems                                      | 147 |
| 11.1.1    | Cyberphysical Systems  | 147 |
| 11.1.2    | Internet-of-Things (IoT)   | 149 |
| 11.2      | Multidomain Mining   | 151 |
| 11.2.1    | Integrating Multiple Heterogeneous Data                                      | 151 |
| 11.2.2    | Integrated Alerts from Multiple Sources                                      | 154 |
| 11.3      | Advanced Machine Learning Models   | 157 |
| 11.3.1    | Deep Learning  | 157 |
| 11.3.2    | Generative Adversarial Networks  | 159 |
| 11.3.3    | Model Reuse  | 161 |
| 11.4      | Ethical Thinking in the Data Analytics Process                               | 163 |
|           | <i>References</i>  | 165 |
|           | <i>Index</i>   | 189 |



## 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

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<sup>1,2</sup> and security directives for Information Assurance Training, Certification, and Workforce Management,<sup>3</sup> 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.

<sup>1</sup> 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

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

<sup>3</sup> Information Assurance Workforce Improvement Program, DoD 8570.01-M [www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/857001m.pdf](http://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 on to 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

## Acknowledgments

xiii

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.

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

---