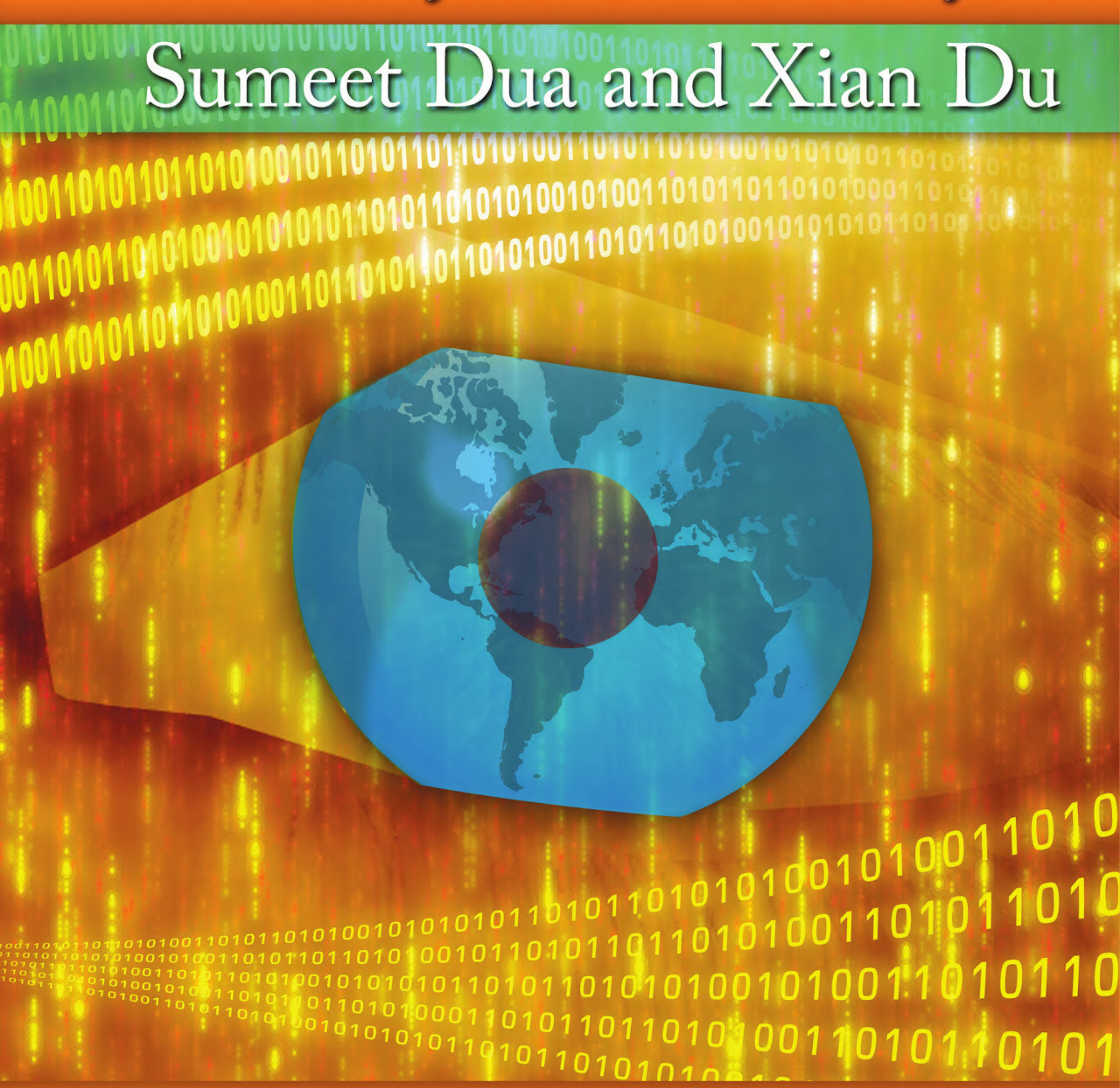
Data Mining and Machine Learning in Cybersecurity

Sumeet Dua and Xian Du



CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

Data Mining and Machine Learning in Cybersecurity

Sumeet Dua and Xian Du

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

Auerbach Publications
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Auerbach Publications is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4398-3943-0 (Ebook-PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Auerbach Web site at
http://www.auerbach-publications.com

Contents

List of Figures

List of Tables

Preface

In the emerging era of Web 3.0, securing cyberspace has gradually evolved into a critical organizational and national research agenda inviting interest from a multidisciplinary scientific workforce. There are many avenues into this area, and, in recent research, machine-learning and data-mining techniques have been applied to design, develop, and improve algorithms and frameworks for cybersecurity system design. Intellectual products in this domain have appeared under various topics, including machine learning, data mining, cybersecurity, data management and modeling, and privacy preservation. Several conferences, workshops, and journals focus on the fragmented research topics in this area. However, transcendent and interdisciplinary assessment of past and current works in the field and possible paths for future research in the area are essential for consistent research and development.

This interdisciplinary assessment is especially useful for students, who typically learn cybersecurity, machine learning, and data mining in independent courses. Machine learning and data mining play significant roles in cybersecurity, especially as more challenges appear with the rapid development of information discovery techniques, such as those originating from the sheer dimensionality and heterogeneous nature of the network data, the dynamic change of threats, and the severe imbalanced classes of normal and anomalous behaviors. In this book, we attempt to combine all the above knowledge for a single advanced course.

This book surveys cybersecurity problems and state-of-the-art machine-learning and data-mining solutions that address the overarching research problems, and it is designed for students and researchers studying or working on machine learning and data mining in cybersecurity applications. The inclusion of cybersecurity in machine-learning research is important for academic research. Such an inclusion inspires fundamental research in machine learning and data mining, such as research in the subfields of imbalanced learning, feature extraction for data with evolving characteristics, and privacy-preserving data mining.

Organization

In Chapter 1, we introduce the vulnerabilities of cyberinfrastructure and the conventional approaches to cyber defense. Then, we present the vulnerabilities of these conventional cyber protection methods and introduce higher-level methodologies that use advanced machine learning and data mining to build more reliable cyber defense systems. We review the cybersecurity solutions that use machine-learning and data-mining techniques, including privacy-preservation data mining, misuse detection, anomaly detection, hybrid detection, scan detection, and profiling detection. In addition, we list a number of references that address cybersecurity issues using machine-learning and data-mining technology to help readers access the related material easily.

In Chapter 2, we introduce machine-learning paradigms and cybersecurity along with a brief overview of machine-learning formulations and the application of machine-learning methods and data mining/management in cybersecurity. We discuss challenging problems and future research directions that are possible when machine-learning methods are applied to the huge amount of temporal and unbalanced network data.

In Chapter 3, we address misuse/signature detection. We introduce fundamental knowledge, key issues, and challenges in misuse/signature detection systems, such as building efficient rule-based algorithms, feature selection for rule matching and accuracy improvement, and supervised machine-learning classification of attack patterns. We investigate several supervised learning methods in misuse detection. We explore the limitations and difficulties of using these machine-learning methods in misuse detection systems and outline possible problems, such as the inadequate ability to detect a novel attack, irregular performance for different attack types, and requirements of the intelligent feature selection. We guide readers to questions and resources that will help them learn more about the use of advanced machine-learning techniques to solve these problems.

In Chapter 4, we provide an overview of anomaly detection techniques. We investigate and classify a large number of machine-learning methods in anomaly detection. In this chapter, we briefly describe the applications of machine-learning methods in anomaly detection. We focus on the limitations and difficulties that encumber machine-learning methods in anomaly detection systems. Such problems include an inadequate ability to maintain a high detection rate and a low false-alarm rate. As anomaly detection is the most concentrative application area of machine-learning methods, we perform in-depth studies to explain the appropriate learning procedures, e.g., feature selection, in detail.

In Chapter 5, we address hybrid intrusion detection techniques. We describe how hybrid detection methods are designed and employed to detect unknown intrusions and anomaly detection with a lower false-positive rate. We categorize the hybrid intrusion detection techniques into three groups based on combinational methods. We demonstrate several machine-learning hybrids that raise detection accuracies in

the intrusion detection system, including correlation techniques, artificial neural networks, association rules, and random forest classifiers.

In Chapter 6, we address scan detection techniques using machine-learning methods. We explain the dynamics of scan attacks and focus on solving scan detection problems in applications. We provide several examples of machine-learning methods used for scan detection, including the rule-based methods, threshold random walk, association memory learning techniques, and expert knowledge-rule-based learning model. This chapter addresses the issues pertaining to the high percentage of false alarms and the evaluation of efficiency and effectiveness of scan detection.

In Chapter 7, we address machine-learning techniques for profiling network traffic. We illustrate a number of profiling modules that profile normal or anomalous behaviors in cyberinfrastructure for intrusion detection. We introduce and investigate a number of new concepts for clustering methods in intrusion detection systems, including association rules, shared nearest neighbor clustering, EM-based clustering, subspace, and informatics theoretic techniques. In this chapter, we address the difficulties of mining the huge amount of streaming data and the necessity of interpreting the profiling results in an understandable way.

In Chapter 8, we provide a comprehensive overview of available machine-learning technologies in privacy-preserving data mining. In this chapter, we concentrate on how data-mining techniques lead to privacy breach and how privacy-preserving data mining achieves data protection via machine-learning methods. Privacy-preserving data mining is a new area, and we hope to inspire research beyond the foundations of data mining and privacy-preserving data mining.

In Chapter 9, we describe the emerging challenges in fixed computing or mobile applications and existing and potential countermeasures using machine-learning methods in cybersecurity. We also explore how the emerging cyber threats may evolve in the future and what corresponding strategies can combat threats. We describe the emerging issues in network monitoring, profiling, and privacy preservation and the emerging challenges in intrusion detection, especially those challenges for anomaly detection systems.

Authors

Dr. Sumeet Dua is currently an Upchurch endowed associate professor and the coordinator of IT research at Louisiana Tech University, Ruston, Louisiana. He received his PhD in computer science from Louisiana State University, Baton Rouge, Louisiana.

His areas of expertise include data mining, image processing and computational decision support, pattern recognition, data warehousing, biomedical informatics, and heterogeneous distributed data integration. The National Science Foundation (NSF), the National Institutes of Health (NIH), the Air Force Research Laboratory (AFRL), the Air Force Office of Sponsored Research (AFOSR), the National Aeronautics and Space Administration (NASA), and the Louisiana Board of Regents (LA-BoR) have funded his research with over $2.8 million. He frequently serves as a study section member (expert panelist) for the National Institutes of Health (NIH) and panelist for the National Science Foundation (NSF)/CISE Directorate. Dr. Dua has chaired several conference sessions in the area of data mining and is the program chair for the *Fifth International Conference on Information Systems, Technology, and Management* (ICISTM-2011). He has given more than 26 invited talks on data mining and its applications at international academic and industry arenas, has advised more than 25 graduate theses, and currently advises several graduate students in the discipline. Dr. Dua is a coinventor of two issued U.S. patents, has (co-)authored more than 50 publications and book chapters, and has authored or edited four books. Dr. Dua has received the Engineering and Science Foundation Award for Faculty Excellence (2006) and the Faculty Research Recognition Award (2007), has been recognized as a distinguished researcher (2004–2010) by the Louisiana Biomedical Research Network (NIH-sponsored), and has won the Outstanding Poster Award at the NIH/NCI caBIG—NCRI Informatics Joint Conference; Biomedical Informatics without Borders: From Collaboration to Implementation. Dr. Dua is a senior member of the IEEE Computer Society, a senior member of the ACM, and a member of SPIE and the American Association for Advancement of Science.

Dr. Xian Du is a research associate and postdoctoral fellow at the Louisiana Tech University, Ruston, Louisiana. He worked as a postdoctoral researcher at the Centre National de la Recherche Scientifique (CNRS) in the CREATIS Lab, Lyon, France, from 2007 to 2008 and served as a software engineer in Kikuze Solutions Pte. Ltd., Singapore, in 2006. He received his PhD from the Singapore–MIT Alliance (SMA) Programme at the National University of Singapore in 2006.

Dr. Xian Du's current research focus is on high-performance computing using machine-learning and data-mining technologies, data-mining applications for cybersecurity, software in multiple computer operational environments, and clustering theoretical research. He has broad experience in machine-learning applications in industry and academic research at high-level research institutes. During his work in the CREATIS Lab in France, he developed a 3D smooth active contour technology for knee cartilage MRI image segmentation. He led a small research and development group to develop color control plug-ins for an RGB color printer to connect to the Windows® system through image processing GDI functions for Kikuze Solutions. He helped to build an intelligent e-diagnostics system for reducing mean time to repair wire-bonding machines at National Semiconductor Ltd., Singapore (NSC). During his PhD dissertation research at the SMA, he developed an intelligent color print process control system for color printers. Dr. Du's major research interests are machine-learning and data-mining applications, heterogeneous data integration and visualization, cybersecurity, and clustering theoretical research.

Chapter 1

Introduction

> Many of the nation's essential and emergency services, as well as our critical infrastructure, rely on the uninterrupted use of the Internet and the communications systems, data, monitoring, and control systems that comprise our cyber infrastructure. A cyber attack could be debilitating to our highly interdependent Critical Infrastructure and Key Resources (CIKR) and ultimately to our economy and national security.
>
> **Homeland Security Council**
> *National Strategy for Homeland Security, 2007*

The ubiquity of cyberinfrastructure facilitates beneficial activities through rapid information sharing and utilization, while its vulnerabilities generate opportunities for our adversaries to perform malicious activities within the infrastructure.* Because of these opportunities for malicious activities, nearly every aspect of cyberinfrastructure needs protection (Homeland Security Council, 2007).

Vulnerabilities in cyberinfrastructure can be attacked horizontally or vertically. Hence, cyber threats can be evaluated horizontally from the perspective of the attacker(s) or vertically from the perspective of the victims. First, we look at cyber threats vertically, from the perspective of the victims. A variety of adversarial agents such as nation-states, criminal organizations, terrorists, hackers, and other malicious users can compromise governmental homeland security through networks.

---

* Cyberinfrastructure consists of digital data, data flows, and the supportive hardware and software. The infrastructure is responsible for data collection, data transformation, traffic flow, data processing, privacy protection, and the supervision, administration, and control of working environments. For example, in our daily activities in cyberspace, we use health Supervisory Control and Data Acquisition (SCADA) systems and the Internet (Chandola et al., 2009).

For example, hackers may utilize personal computers remotely to conspire, proselytize, recruit accomplices, raise funds, and collude during ongoing attacks. Adversarial governments and agencies can launch cyber attacks on the hardware and software of the opponents' cyberinfrastructures by supporting financially and technically malicious network exploitations.

Cyber criminals threaten financial infrastructures, and they could pose threats to national economies if recruited by the adversarial agents or terrorist organizations. Similarly, private organizations, e.g., banks, must protect confidential business or private information from such hackers. For example, the disclosure of business or private financial data to cyber criminals can lead to financial loss via Internet banking and related online resources. In the pharmaceutical industry, disclosure of protected company information can benefit competitors and lead to market-share loss. Individuals must also be vigilant against cyber crimes and malicious use of Internet technology.

As technology has improved, users have become more tech savvy. People communicate and cooperate efficiently through networks, such as the Internet, which are facilitated by the rapid development of digital information technologies, such as personal computers and personal digital assistants (PDAs). Through these digital devices linked by the Internet, hackers also attack personal privacy using a variety of weapons, such as viruses, Trojans, worms, botnet attacks, rootkits, adware, spam, and social engineering platforms.

Next, we look at cyber threats horizontally from the perspective of the victims. We consider any malicious activity in cyberspace as a cyber threat. A cyber threat may result in the loss of or damage to cyber components or physical resources. Most cyber threats are categorized into one of three groups according to the intruder's purpose: stealing confidential information, manipulating the components of cyberinfrastructure, and/or denying the functions of the infrastructure. If we evaluate cyber threats horizontally, we can investigate cyber threats and the subsequent problems. We will focus on intentional cyber crimes and will not address breaches caused by normal users through unintentional operations, such as errors and omissions, since education and proper habits could help to avoid these threats.* We also will not explain cyber threats caused by natural disasters, such as accidental breaches caused by earthquakes, storms, or hurricanes, as these threats happen suddenly and are beyond our control.

1.1 Cybersecurity

To secure cyberinfrastructure against intentional and potentially malicious threats, a growing collaborative effort between cybersecurity professionals and researchers from institutions, private industries, academia, and government agencies has engaged in

---

* We define a normal cyber user as an individual or group of individuals who do not intend to intrude on the cybersecurity of other individuals.

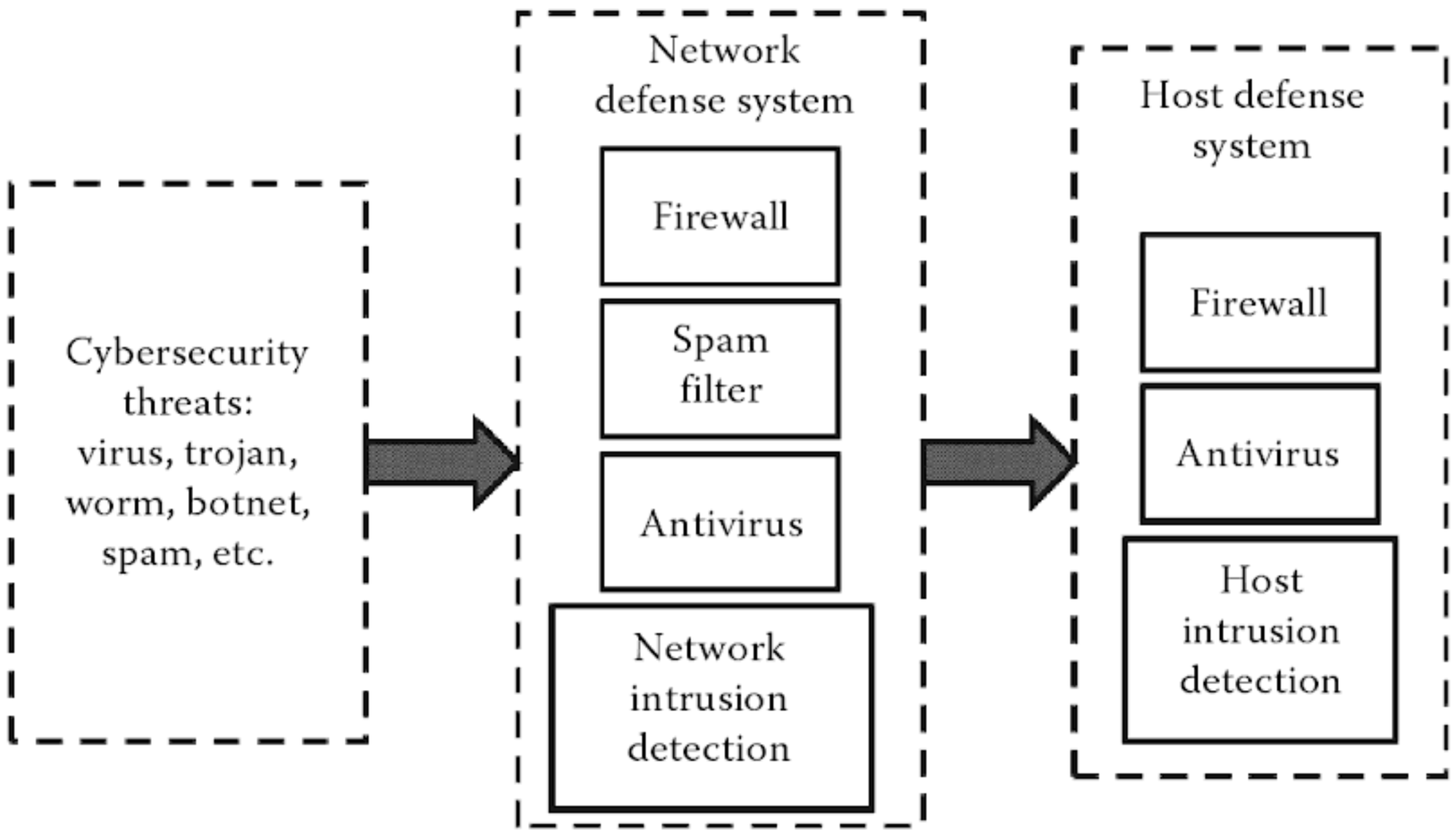


Figure 1.1 Conventional cybersecurity system.

exploiting and designing a variety of cyber defense systems. Cybersecurity researchers and designers aim to maintain the confidentiality, integrity, and availability of information and information management systems through various cyber defense systems that protect computers and networks from hackers who may want to intrude on a system or steal financial, medical, or other identity-based information.*

As shown in Figure 1.1, conventional cybersecurity systems address various cybersecurity threats, including viruses, Trojans, worms, spam, and botnets. These cybersecurity systems combat cybersecurity threats at two levels and provide network- and host-based defenses. Network-based defense systems control network flow by network firewall, spam filter, antivirus, and network intrusion detection techniques. Host-based defense systems control upcoming data in a workstation by firewall, antivirus, and intrusion detection techniques installed in hosts.

Conventional approaches to cyber defense are mechanisms designed in firewalls, authentication tools, and network servers that monitor, track, and block viruses and other malicious cyber attacks. For example, the Microsoft Windows® operating system has a built-in Kerberos cryptography system that protects user information. Antivirus software is designed and installed in personal computers and cyberinfrastructures to ensure customer information is not used maliciously. These approaches create a protective shield for cyberinfrastructure.

However, the vulnerabilities of these methods are ubiquitous in applications because of the flawed design and implementation of software and network

* The three requirements of cybersecurity correspond to the three types of intentional threats: confidentiality signifies the ability to prevent sensitive data from being disclosed to third parties; integrity ensures the infrastructure is complete and accurate, and availability refers to the accessibility of the normal operations of cyberinfrastructures, such as delivering and storing data.

infrastructure. Patches have been developed to protect the cyber systems, but attackers continuously exploit newly discovered flaws. Because of the constantly evolving cyber threats, building defense systems for discovered attacks is not enough to protect users. Higher-level methodologies are also required to discover the embedded and lurking cyber intrusions and cyber intrusion techniques, so that a more reliable security cyberinfrastructure can be utilized.

Many higher-level adaptive cyber defense systems can be partitioned into components as shown in Figure 1.2. Figure 1.2 outlines the five-step process for those defense systems. We discuss each step below.

Data-capturing tools, such as Libpcap for Linux®, Solaris BSM for SUN®, and Winpcap for Windows®, capture events from the audit trails of resource information sources (e.g., network). Events can be host-based or network-based depending on where they originate. If an event originates with log files, then it is categorized as a host-based event. If it originates with network traffic, then it is categorized as a network-based event. A host-based event includes a sequence of commands executed by a user and a sequence of system calls launched by an application, e.g., send mail. A network-based event includes network traffic data, e.g., a sequence of internet protocol (IP) or transmission control protocol (TCP) network packets. The data-preprocessing module filters out the attacks for which good signatures have been learned.

A feature extractor derives basic features that are useful in event analysis engines, including a sequence of system calls, start time, duration of a network flow, source IP and source port, destination IP and destination port, protocol,

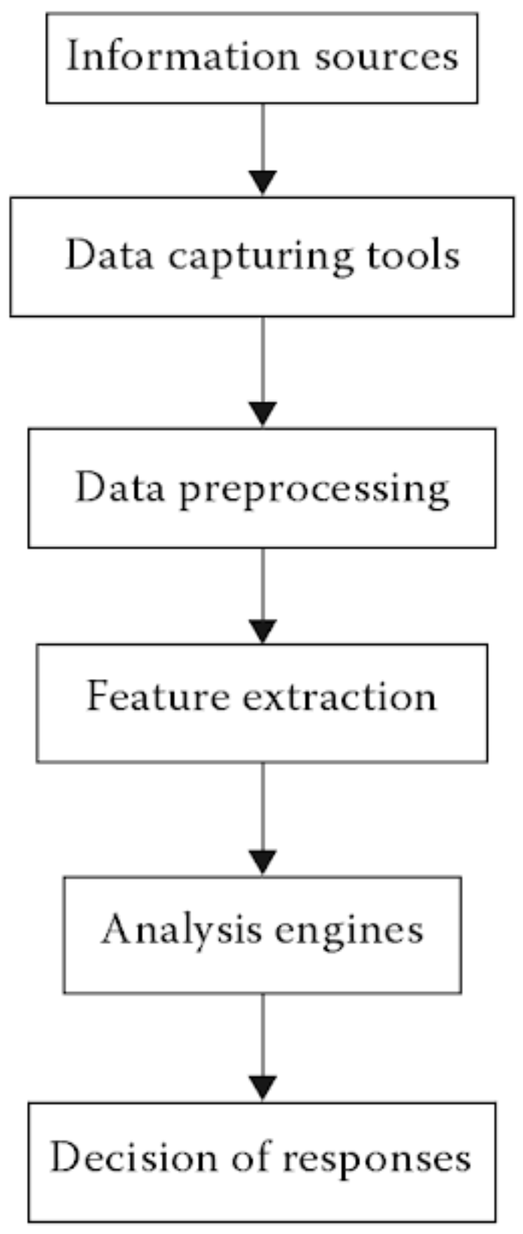


Figure 1.2 Adaptive defense system for cybersecurity.

number of bytes, and number of packets. In an analysis engine, various intrusion detection methods are implemented to investigate the behavior of the cyberinfrastructure, which may or may not have appeared before in the record, e.g., to detect anomalous traffic. The decision of responses is deployed once a cyber attack is identified. As shown in Figure 1.2, analysis engines are the core technologies for the generation of the adaptation ability of the cyber defense system. As discussed above, the solutions to cybersecurity problems include proactive and reactive security solutions.

Proactive approaches anticipate and eliminate vulnerabilities in the cyber system, while remaining prepared to defend effectively and rapidly against attacks. To function correctly, proactive security solutions require user authentication (e.g., user password and biometrics), a system capable of avoiding programming errors, and information protection [e.g., privacy-preserving data mining (PPDM)]. PPDM protects data from being explored by data-mining techniques in cybersecurity applications. We will discuss this technique in detail in Chapter 8. Proactive approaches have been used as the first line of defense against cybersecurity breaches. It is not possible to build a system that has no security vulnerabilities. Vulnerabilities in common security components, such as firewalls, are inevitable due to design and programming errors.

The second line of cyber defense is composed of reactive security solutions, such as intrusion detection systems (IDSs). IDSs detect intrusions based on the information from log files and network flow, so that the extent of damage can be determined, hackers can be tracked down, and similar attacks can be prevented in the future.

1.2 Data Mining

Due to the availability of large amounts of data in cyberinfrastructure and the number of cyber criminals attempting to gain access to the data, data mining, machine learning, statistics, and other interdisciplinary capabilities are needed to address the challenges of cybersecurity. Because IDSs use data mining and machine learning, we will focus on these areas. Data mining is the extraction, or "mining," of knowledge from a large amount of data. The strong patterns or rules detected by data-mining techniques can be used for the nontrivial prediction of new data. In nontrivial prediction, information that is implicitly presented in the data, but was previously unknown is discovered. Data-mining techniques use statistics, artificial intelligence, and pattern recognition of data in order to group or extract behaviors or entities. Thus, data mining is an interdisciplinary field that employs the use of analysis tools from statistical models, mathematical algorithms, and machine-learning methods to discover previously unknown, valid patterns and relationships in large data sets, which are useful for finding hackers and preserving privacy in cybersecurity.

Data mining is used in many domains, including finance, engineering, biomedicine, and cybersecurity. There are two categories of data-mining methods: supervised and unsupervised. Supervised data-mining techniques predict a hidden function using training data. The training data have pairs of input variables and output labels or classes. The output of the method can predict a class label of the input variables. Examples of supervised mining are classification and prediction. Unsupervised data mining is an attempt to identify hidden patterns from given data without introducing training data (i.e., pairs of input and class labels). Typical examples of unsupervised mining are clustering and associative rule mining.

Data mining is also an integral part of knowledge discovery in databases (KDDs), an iterative process of the nontrivial extraction of information from data and can be applied to developing secure cyberinfrastructures. KDD includes several steps from the collection of raw data to the creation of new knowledge. The iterative process consists of the following steps: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation, as described below.

Step 1. During data cleaning, which is also known as data cleansing, noise and irrelevant data are removed from the collection.
Step 2. Data integration combines data from multiple and heterogeneous sources into one database.
Step 3. Data-selection techniques allow the user to obtain a reduced representation of the data set to keep the integrity of the original data set in a reduced volume.
Step 4. In data transformation, the selected data is transformed into suitable formats.
Step 5. Data mining is the stage in which analysis tools are applied to discover potentially useful patterns.
Step 6. Pattern evaluation identifies interesting and useful patterns using given validation measures.
Step 7. In knowledge representation, the final phase of the knowledge-discovery process, discovered knowledge is presented to the users in visual forms.

Data-mining techniques are used to aid in the development of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes, which include real-time data sampling, selection, analysis and query, and mining peta-scale data to classify and detect attacks and intrusions on a computer network (Denning, 1987; Lee and Stolfo, 1998; Axelsson, 2000; Chandola et al., 2006; Homeland Security Council, 2007). Learning user patterns and/or behaviors is critical for intrusion detection and attack predictions. Learning these behaviors is important, as they can identify and describe structural patterns in the data automatically and theoretically explain data and predict patterns. Automatic and theoretic learning require complex computation that calls for abundant machine-learning algorithms. We will discuss the concept of machine learning in Section 1.3.