Marcus A. Maloof (Ed.)

Machine Learning and Data Mining for Computer Security

Methods and Applications

With 23 Figures

Springer

Marcus A. Maloof, BS, MS, PhD
Department of Computer Science
Georgetown University
Washington DC 20057-1232
USA

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2005928487

Advanced Information and Knowledge Processing ISSN 1610-3947
ISBN-10: 1-84628-029-X
ISBN-13: 978-1-84628-029-0

Printed on acid-free paper

Printed in the United States of America (MVY)

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springeronline.com

To my mom and dad, Ann and Ferris

Foreword

When I first got into information security in the early 1970s, the little research that existed was focused on mechanisms for preventing attacks. The goal was airtight security, and much of the research by the end of decade and into the next focused on building systems that were provably secure. Although there was widespread recognition that insiders with legitimate access could always exploit their privileges to cause harm, the prevailing sentiment was that we could at least design systems that were not inherently faulty and vulnerable to trivial attacks by outsiders.

We were wrong. This became rapidly apparent to me as I witnessed the rapid evolution of information technology relative to progress in information security. The quest to design the perfect system could not keep up with market demands and developments in personal computers and computer networks. A few Herculean efforts in industry did in fact produce highly secure systems, but potential customers paid more attention to applications, performance, and price. They bought systems that were rich in functionality, but riddled with holes. The security on the Internet was aptly compared to "Swiss cheese."

Today, it is widely recognized that our computers and networks are unlikely to ever be capable of preventing all attacks. They are just way too complex. Thousands of new vulnerabilities are reported to the Computer Emergency Response Team Coordination Center (CERT/CC) annually. We might significantly reduce the security flaws through good software development practices, but we cannot expect foolproof security as technology continues to advance at breakneck speeds. Further, the problems do not reside solely with the vendors; networks must also be properly configured and managed. This can be a daunting task given the vast and growing number of products that can be networked together and interact in unpredictable ways.

In the middle 1980s, a small group of us at SRI International began investigating an alternative approach to security. Recognizing the limitations of a strategy based solely on prevention, we began to design a system that could detect intrusions and insider abuse in real time as they occurred. Our research and that of others led to the development of intrusion detection systems. Also

in the 1980s, computer viruses and worms emerged as a threat, leading to software tools for detecting their presence. These two types of detection technologies have been largely separate but complementary. Intrusion detection systems focus on detecting malicious computer and network activity, while antiviral tools focus on detecting malicious code in files and messages.

To succeed, a detection system must know what to look for. This has been easier to achieve with viral detection than intrusion detection. Most antiviral tools work off a list containing the "signatures" of known viruses, worms, and Trojan horses. If any of the signatures are detected during a scan, the file or message is flagged. The main limitation of these tools is that they cannot detect new forms of malicious code that do match the existing signatures. Vendors mitigate the exposure of their customers by frequently updating and distributing their signature files, but there remains a period of vulnerability that has yet to be closed.

With intrusion detection, it is more difficult to know what to look for, as unauthorized activity on a system can take so many forms and even resemble legitimate activity. In an attempt to not miss something that is potentially malicious, many of the existing systems sound far too many false or inconsequential alarms (often thousands per day), substantially reducing their effectiveness. Without a means of breaking through the false-alarm barrier, intrusion detection will fail to meet its promise.

This brings me to this book. The authors have made significant progress in our ability to distinguish malicious activity and code from that which is not. This progress has come from bringing machine learning and data mining to the detection task. These technologies offer a way past the false-alarm barrier and towards more effective detection systems.

The papers in this book address one of the most exciting areas of research in information security today. They make an important contribution to that area and will help pave the way towards more secure systems.

Monterey, CA
January 2005

Dorothy E. Denning

Preface

In the mid-1990s, when I was a graduate student studying machine learning, someone broke into a dean's computer account and behaved in a way that most deans never would: There was heavy use of system resources very early in the morning. I wondered why there was not some process monitoring everyone's activity and detecting abnormal behavior. At least in the case of the dean, it should not have been difficult to detect that the person using the account was probably not the dean.

About the same time, I taught a class on artificial intelligence at Georgetown University. At that time, Dorothy Denning was the chairperson. I knew she worked in security, but I knew little about the field and her research; after all, I was studying rule learning. When I told her about my idea of learning profiles of user behavior, she remarked, "Oh, there's been lots of work on that." I made copies of the papers she gave me, and I started reading.

In the meantime, I managed to convince my lab's system administrator to let me use some of our audit data for machine learning experiments. It was not a lot of data—about three weeks of activity for seven users—but it was enough for a section in my dissertation, which was not about machine learning approaches to computer security.

After graduating, I thought little about the application of machine learning to computer security until recently, when Jeremy Kolter and I began investigating approaches for detecting malicious executables. This time, I started with the literature review, and I was amazed at how widespread the research had become. (Of course, the Internet today is not the same as it was in 1994.)

Ten years ago, it seemed that most of the articles were in computer security journals and proceedings and few were in the proceedings of artificial intelligence and machine learning conferences. Today, there are many publications in all of these forums, and we now have the new field of data mining. Many interesting papers appear in its literature. There are also publications in literatures on statistics, industrial engineering, and information systems. This description does not take into account recent work on fraud detection, which is relevant to applications in computer security, even though it does

not involve network traffic or audit data. Indeed, many issues are common to both endeavors.

Perhaps I am a little better at doing literature searches, but in retrospect, this "discovery" should not have been too surprising since there is overlap among these areas and disciplines. However, what I needed and wanted was a book that brought this work together. In addition to research contributions, I also wanted chapters that described relevant concepts of computer security. Ideally, it would be part textbook, part monograph, and part special issue of a journal.

At the time, Jeremy Kolter and I were preparing a paper for the Third IEEE International Conference on Data Mining. Xindong Wu of the University of Vermont was the program co-chair, and during a visit to his Web site, I noticed that he was an editor of Springer's series on Advanced Information and Knowledge Processing. After a few e-mails and words of encouragement, I submitted a proposal for this book. After peer review, Springer accepted it.

Intended Audience

The intended audience for this book consists of three groups. The first group consists of researchers and practitioners working in this interesting intersection of machine learning, data mining, and computer security. People in this group will undoubtedly recognize the contributors and the connection of the chapters to their past work.

The second group consists of people who know about one field, but would like to learn more about the other. It is for people who know about machine learning and data mining, but would like to learn more about computer security. These people have a dual in computer security, and so the book is also for people who know this field, but would like to learn more about machine learning and data mining.

Finally, I hope graduate students, who constitute the third group, will find this volume attractive, whether they are studying machine learning, data mining, statistics, or information assurance. I would be delighted if a professor used this book for a graduate seminar on machine learning and data mining approaches to computer security.

Acknowledgements

As the editor, I would like to begin by thanking Xindong Wu for his early encouragement. Also early on, I consulted with Ryszard Michalski, Ophir Frieder, and Dorothy Denning; they, too, provided important, early encouragement and support for the project. In particular, I would like to thank Dorothy for also taking the time to write the foreword to this volume.

Obviously, the contributors played the most important role in the production of this book. I want to thank them for participating, for submitting high-quality chapters, and for making my job as editor easy.

Of the contributors, I consulted with Terran Lane and Clay Shields the most. From the beginning, Terran helped identify potential contributors, gave advice on the background chapters I should consider, and suggested that, ideally, the person writing the introductory chapter on computer security would work closely with the person writing the introductory chapter on machine learning. Clay Shields, whose office is next to mine, accepted a fairly late invitation to write an introductory chapter on information assurance. Even before he accepted, he was a valued and close source for papers, books, and ideas.

Catherine Drury, my editor at Springer, was a tremendous help. I really have appreciated her patience, advice, and quick responses to e-mails. Finally, I would like to thank the Graduate School at Georgetown University. They provided funds for production expenses associated with this project.

Bloedorn, Talbot, and DeBarr would like to thank Alan Christiansen, Bill Hill, Zohreh Nazeri, Clem Skorupka, and Jonathan Tivel for their many contributions to their work.

Early and Brodley's chapter is based upon work supported by the National Science Foundation under Grant No. 0335574, and the Air Force Research Lab under Grant No. F30602-02-2-0217.

Kolter and Maloof thank William Asmond and Thomas Ervin of the MITRE Corporation for providing their expertise, advice, and collection of malicious executables. They also thank Ophir Frieder of IIT for help with the vector space model, Abdur Chowdhury of AOL for advice on the scalability of the vector space model, Bob Wagner of the FDA for assistance with ROC analysis, Eric Bloedorn of MITRE for general guidance on our approach, and Matthew Krause of Georgetown for helpful comments on an earlier draft of the chapter. Finally, they thank Richard Squier of Georgetown for supplying much of the additional computational resources needed for this study through Grant No. DAAD19-00-1-0165 from the U.S. Army Research Office. They conducted their research in the Department of Computer Science at Georgetown University, and it was supported by the MITRE Corporation under contract 53271.

Lane would like to thank Matt Schonlau for providing the data employed in the study as well as the results of his comprehensive study of user-level anomaly detection techniques. Lane also thanks Amy McGovern and Kiri Wagstaff for their invaluable comments on draft versions of his chapter.

Washington, DC
March 2005

Mark Maloof

List of Contributors

Eric E. Bloedorn
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508, USA
bloedorn@mitre.org

Carla E. Brodley
Department of Computer Science
Tufts University
Medford, MA 02155, USA
brodley@cs.tufts.edu

Philip Chan
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
pkc@cs.fit.edu

David D. DeBarr
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508, USA
debarr@mitre.org

James P. Early
CERIAS
Purdue University
West Lafayette, IN 47907-2086, USA
earlyjp@cerias.purdue.edu

Wei Fan
IBM T. J. Watson Research Center
Hawthorne, NY 10532, USA
weifan@us.ibm.com

Klaus Julisch
IBM Zurich Research Laboratory
Saeumerstrasse 4
8803 Rueschlikon, Switzerland
kju@zurich.ibm.com

Jeremy Z. Kolter
Department of Computer Science
Georgetown University
Washington, DC 20057-1232, USA
jzk@cs.georgetown.edu

Terran Lane
Department of Computer Science
The University of New Mexico
Albuquerque, NM 87131-1386, USA
terran@cs.unm.edu

Wenke Lee
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
wenke@cc.gatech.edu

Marcus A. Maloof
Department of Computer Science
Georgetown University
Washington, DC 20057-1232, USA
maloof@cs.georgetown.edu

Matthew Miller
Computer Science Department
Columbia University
New York, NY 10027, USA
mmiller@cs.columbia.edu

Debasis Mitra
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
dmitra@cs.fit.edu

Clay Shields
Department of Computer Science
Georgetown University
Washington, DC 20057-1232, USA
clay@cs.georgetown.edu

Salvatore J. Stolfo
Computer Science Department
Columbia University
New York, NY 10027, USA
sal@cs.columbia.edu

Lisa M. Talbot
Simplex, LLC
410 Wingate Place, SW
Leesburg, VA 20175, USA
talbotlm@ieee.org

Gaurav Tandon
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
gtandon@cs.fit.edu

Contents