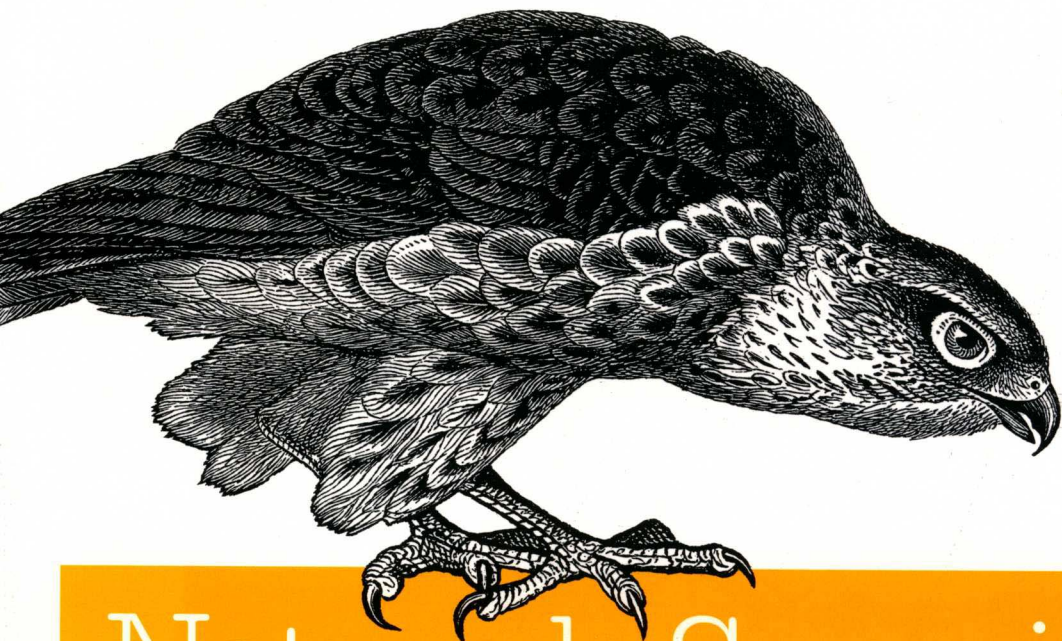


O'REILLY®



Network Security Through Data Analysis

基于数据分析的网络安全 (影印版)

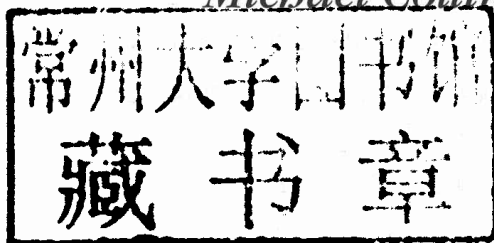
東南大學出版社

Michael Collins 著

基于数据分析的网络安全 (影印版)

Network Security Through Data Analysis

Michael Collins 著



Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'REILLY®

O'Reilly Media, Inc. 授权东南大学出版社出版

南京 东南大学出版社

图书在版编目 (CIP) 数据

基于数据分析的网络安全:英文/(美)柯林(Collins, M.)
著. —影印本. —南京:东南大学出版社, 2014.10
书名原文: Network Security through Data Analysis
ISBN 978-7-5641-5007-5

I. ①基… II. ①柯… III. ①计算机网络—安全技术
—英文 IV. ①TP393.08

中国版本图书馆 CIP 数据核字 (2014) 第 115560 号

江苏省版权局著作权合同登记

图字: 10-2013-360 号

©2014 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Press, 2014. Authorized reprint of the original English edition, 2013 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版 2014。

英文影印版由东南大学出版社出版 2014。此影印版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有, 未得书面许可, 本书的任何部分和全部不得以任何形式重制。

基于数据分析的网络安全 (影印版)

出版发行: 东南大学出版社

地 址: 南京四牌楼 2 号 邮编: 210096

出 版 人: 江建中

网 址: <http://www.seupress.com>

电子邮件: press@seupress.com

印 刷: 常州市武进第三印刷有限公司

开 本: 787 毫米 × 980 毫米 16 开本

印 张: 21.75

字 数: 426 千字

版 次: 2014 年 10 月第 1 版

印 次: 2014 年 10 月第 1 次印刷

书 号: ISBN 978-7-5641-5007-5

定 价: 66.00 元

本社图书若有印装质量问题, 请直接与营销部联系。电话 (传真): 025-83791830

Preface

This book is about networks: monitoring them, studying them, and using the results of those studies to improve them. “Improve” in this context hopefully means to make more secure, but I don’t believe we have the vocabulary or knowledge to say that confidently—at least not yet. In order to implement security, we try to achieve something more quantifiable and describable: *situational awareness*.

Situational awareness, a term largely used in military circles, is exactly what it says on the tin: an understanding of the environment you’re operating in. For our purposes, situational awareness encompasses understanding the components that make up your network and how those components are used. This awareness is often *radically* different from how the network is configured and how the network was originally designed.

To understand the importance of situational awareness in information security, I want you to think about your home, and I want you to count the number of web servers in your house. Did you include your wireless router? Your cable modem? Your printer? Did you consider the web interface to CUPS? How about your television set?

To many IT managers, several of the devices listed didn’t even register as “web servers.” However, embedded web servers speak HTTP, they have known vulnerabilities, and they are increasingly common as specialized control protocols are replaced with a web interface. Attackers will often hit embedded systems without realizing what they are—the SCADA system is a Windows server with a couple of funny additional directories, and the MRI machine is a perfectly serviceable spambot.

This book is about collecting data and looking at networks in order to understand how the network is used. The focus is on analysis, which is the process of taking security data and using it to make actionable decisions. I emphasize the word *actionable* here because effectively, security decisions are restrictions on behavior. Security policy involves telling people what they shouldn’t do (or, more onerously, telling people what they *must* do). Don’t use Dropbox to hold company data, log on using a password and an RSA dongle, and don’t copy the entire project server and sell it to the competition. When we make

security decisions, we interfere with how people work, and we'd better have good, solid reasons for doing so.

All security systems ultimately depend on users recognizing the importance of security and accepting it as a necessary evil. Security rests on people: it rests on the individual users of a system obeying the rules, and it rests on analysts and monitors identifying when rules are broken. Security is only marginally a technical problem—information security involves endlessly creative people figuring out new ways to abuse technology, and against this constantly changing threat profile, you need cooperation from both your defenders and your users. Bad security policy will result in users increasingly evading detection in order to get their jobs done or just to blow off steam, and that adds additional work for your defenders.

The emphasis on actionability and the goal of achieving security is what differentiates this book from a more general text on data science. The section on analysis proper covers statistical and data analysis techniques borrowed from multiple other disciplines, but the overall focus is on understanding the structure of a network and the decisions that can be made to protect it. To that end, I have abridged the theory as much as possible, and have also focused on mechanisms for identifying abusive behavior. Security analysis has the unique problem that the targets of observation are not only aware they're being watched, but are actively interested in stopping it if at all possible.

The MRI and the General's Laptop

Several years ago, I talked with an analyst who focused primarily on a university hospital. He informed me that the most commonly occupied machine on his network was the MRI. In retrospect, this is easy to understand.

“Think about it,” he told me. “It's medical hardware, which means its certified to use a specific version of Windows. So every week, somebody hits it with an exploit, roots it, and installs a bot on it. Spam usually starts around Wednesday.” When I asked why he didn't just block the machine from the Internet, he shrugged and told me the doctors wanted their scans. He was the first analyst I've encountered with this problem, and he wasn't the last.

We see this problem a lot in any organization with strong hierarchical figures: doctors, senior partners, generals. You can build as many protections as you want, but if the general wants to borrow the laptop over the weekend and let his granddaughter play Neopets, you've got an infected laptop to fix on Monday.

Just to pull a point I have hidden in there, I'll elaborate. I am a firm believer that the most effective way to defend networks is to secure and defend *only* what you need to secure and defend. I believe this is the case because information security will always require people to be involved in monitoring and investigation—the attacks change too

much, and when we do automate defenses, we find out that attackers can now use them to attack us.¹

I am, as a security analyst, firmly convinced that security should be inconvenient, well-defined, and constrained. Security should be an artificial behavior extended to assets that must be protected. It should be an artificial behavior because the final line of defense in any secure system is the *people* in the system—and people who are fully engaged in security will be mistrustful, paranoid, and looking for suspicious behavior. This is not a happy way to live your life, so in order to make life bearable, we have to limit security to what must be protected. By trying to watch everything, you lose the edge that helps you protect what's really important.

Because security is inconvenient, effective security analysts must be able to *convince* people that they need to change their normal operations, jump through hoops, and otherwise constrain their mission in order to prevent an abstract future attack from happening. To that end, the analysts must be able to identify the decision, produce information to back it up, and demonstrate the risk to their audience.

The process of data analysis, as described in this book, is focused on developing security knowledge in order to make effective security decisions. These decisions can be forensic: reconstructing events after the fact in order to determine why an attack happened, how it succeeded, or what damage was done. These decisions can also be proactive: developing rate limiters, intrusion detection systems, or policies that can limit the impact of an attacker on a network.

Audience

Information security analysis is a young discipline and there really is no well-defined body of knowledge I can point to and say “Know this.” This book is intended to provide a snapshot of analytic techniques that I or other people have thrown at the wall over the past 10 years and seen stick.

The target audience for this book is network administrators and operational security analysts, the personnel who work on NOC floors or who face an IDS console on a regular basis. My expectation is that you have some familiarity with TCP/IP tools such as *netstat*, and some basic statistical and mathematical skills.

In addition, I expect that you have some familiarity with scripting languages. In this book, I use Python as my go-to language for combining tools. The Python code is illustrative and might be understandable without a Python background, but it is assumed that you possess the skills to create filters or other tools in the language of your choice.

1. Consider automatically locking out accounts after x number of failed password attempts, and combine it with logins based on email addresses. Consider how many accounts you can lock out that way.

In the course of writing this book, I have incorporated techniques from a number of different disciplines. Where possible, I've included references back to original sources so that you can look through that material and find other approaches. Many of these techniques involve mathematical or statistical reasoning that I have intentionally kept at a functional level rather than going through the derivations of the approach. A basic understanding of statistics will, however, be helpful.

Contents of This Book

This book is divided into three sections: data, tools, and analytics. The data section discusses the process of collecting and organizing data. The tools section discusses a number of different tools to support analytical processes. The analytics section discusses different analytic scenarios and techniques.

Part I discusses the collection, storage, and organization of data. Data storage and logistics are a critical problem in security analysis; it's easy to collect data, but hard to search through it and find actual phenomena. Data has a footprint, and it's possible to collect so much data that you can never meaningfully search through it. This section is divided into the following chapters:

Chapter 1

This chapter discusses the general process of collecting data. It provides a framework for exploring how different sensors collect and report information and how they interact with each other.

Chapter 2

This chapter expands on the discussion in the previous chapter by focusing on sensors that collect network traffic data. These sensors, including *tcpdump* and NetFlow, provide a comprehensive view of network activity, but are often hard to interpret because of difficulties in reconstructing network traffic.

Chapter 3

This chapter discusses sensors that are located on a particular system, such as host-based intrusion detection systems and logs from services such as HTTP. Although these sensors cover much less traffic than network sensors, the information they provide is generally easier to understand and requires less interpretation and guesswork.

Chapter 4

This chapter discusses tools and mechanisms for storing traffic data, including traditional databases, big data systems such as Hadoop, and specialized tools such as graph databases and REDIS.

Part II discusses a number of different tools to use for analysis, visualization, and reporting. The tools described in this section are referenced extensively in later sections when discussing how to conduct different analytics.

Chapter 5

System for Internet-Level Knowledge (SiLK) is a flow analysis toolkit developed by Carnegie Mellon's CERT. This chapter discusses SiLK and how to use the tools to analyze NetFlow data.

Chapter 6

R is a statistical analysis and visualization environment that can be used to effectively explore almost any data source imaginable. This chapter provides a basic grounding in the R environment, and discusses how to use R for fundamental statistical analysis.

Chapter 7

Intrusion detection systems (IDSes) are automated analysis systems that examine traffic and raise alerts when they identify something suspicious. This chapter focuses on how IDSes work, the impact of detection errors on IDS alerts, and how to build better detection systems whether implementing IDS using tools such as SiLK or configuring an existing IDS such as Snort.

Chapter 8

One of the more common and frustrating tasks in analysis is figuring out where an IP address comes from, or what a signature means. This chapter focuses on tools and investigation methods that can be used to identify the ownership and provenance of addresses, names, and other tags from network traffic.

Chapter 9

This chapter is a brief walkthrough of a number of specialized tools that are useful for analysis but don't fit in the previous chapters. These include specialized visualization tools, packet generation and manipulation tools, and a number of other toolkits that an analyst should be familiar with.

The final section of the book, Part III, focuses on the goal of all this data collection: analytics. These chapters discuss various traffic phenomena and mathematical models that can be used to examine data.

Chapter 10

Exploratory Data Analysis (EDA) is the process of examining data in order to identify structure or unusual phenomena. Because security data changes so much, EDA is a necessary skill for any analyst. This chapter provides a grounding in the basic visualization and mathematical techniques used to explore data.

Chapter 11

This chapter looks at mistakes in communications and how those mistakes can be used to identify phenomena such as scanning.

Chapter 12

This chapter discusses analyses that can be done by examining traffic volume and traffic behavior over time. This includes attacks such as DDoS and database raids, as well as the impact of the work day on traffic volumes and mechanisms to filter traffic volumes to produce more effective analyses.

Chapter 13

This chapter discusses the conversion of network traffic into graph data and the use of graphs to identify significant structures in networks. Graph attributes such as centrality can be used to identify significant hosts or aberrant behavior.

Chapter 14

This chapter discusses techniques to determine which traffic is crossing service ports in a network. This includes simple lookups such as the port number, as well as banner grabbing and looking at expected packet sizes.

Chapter 15

This chapter discusses a step-by-step process for inventorying a network and identifying significant hosts within that network. Network mapping and inventory are critical steps in information security and should be done on a regular basis.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic

Indicates new terms, URLs, email addresses, filenames, and file extensions.

Constant width

Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

Constant width bold

Shows commands or other text that should be typed literally by the user.

Constant width italic

Shows text that should be replaced with user-supplied values or by values determined by context.



This icon signifies a tip, suggestion, or general note.



This icon indicates a warning or caution.

Using Code Examples


Supplemental material (code examples, exercises, etc.) is available for download at https://github.com/mpcollins/nsda_examples

This book is here to help you get your job done. In general, if example code is offered with this book, you may use it in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Network Security Through Data Analysis* by Michael Collins (O'Reilly). Copyright 2014 Michael Collins, 978-1-449-3579-0."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at permissions@oreilly.com.

Safari® Books Online

Safari[®]
Books Online  *Safari Books Online* is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of

books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at <http://oreil.ly/nstda>.

To comment or ask technical questions about this book, send email to bookquestions@oreilly.com.

For more information about our books, courses, conferences, and news, see our website at <http://www.oreilly.com>.

Find us on Facebook: <http://facebook.com/oreilly>

Follow us on Twitter: <http://twitter.com/oreillymedia>

Watch us on YouTube: <http://www.youtube.com/oreillymedia>

Acknowledgements

I need to thank my editor, Andy Oram, for his incredible support and feedback, without which I would still be rewriting commentary on network vantage over and over again. I also want to thank my assistant editors, Allyson MacDonald and Maria Gulick, for riding herd and making me get the thing finished. I also need to thank my technical reviewers: Rhiannon Weaver, Mark Thomas, Rob Thomas, André DiMino, and Henry Stern. Their comments helped me to rip out more fluff and focus on the important issues.

This book is an attempt to distill down a lot of experience on ops floors and in research labs, and I owe a debt to many people on both sides of the world. In no particular order,

this includes Tom Longstaff, Jay Kadane, Mike Reiter, John McHugh, Carrie Gates, Tim Shimeall, Markus DeShon, Jim Downey, Will Franklin, Sandy Parris, Sean McAllister, Greg Virgin, Scott Coull, Jeff Janies, and Mike Witt.

Finally, I want to thank my parents, James and Catherine Collins. Dad died during the writing of this work, but he kept asking me questions, and then since he didn't understand the answers, questions about the questions until it was done.

Table of Contents

Preface.....	ix
--------------	----

Part I. Data

1. Sensors and Detectors: An Introduction.....	3
Vantages: How Sensor Placement Affects Data Collection	4
Domains: Determining Data That Can Be Collected	7
Actions: What a Sensor Does with Data	10
Conclusion	13
2. Network Sensors.....	15
Network Layering and Its Impact on Instrumentation	16
Network Layers and Vantage	18
Network Layers and Addressing	23
Packet Data	24
Packet and Frame Formats	24
Rolling Buffers	25
Limiting the Data Captured from Each Packet	25
Filtering Specific Types of Packets	25
What If It's Not Ethernet?	29
NetFlow	30
NetFlow v5 Formats and Fields	30
NetFlow Generation and Collection	32
Further Reading	33
3. Host and Service Sensors: Logging Traffic at the Source.....	35
Accessing and Manipulating Logfiles	36
The Contents of Logfiles	38
The Characteristics of a Good Log Message	38

Existing Logfiles and How to Manipulate Them	41
Representative Logfile Formats	43
HTTP: CLF and ELF	43
SMTP	47
Microsoft Exchange: Message Tracking Logs	49
Logfile Transport: Transfers, Syslog, and Message Queues	50
Transfer and Logfile Rotation	51
Syslog	51
Further Reading	53
4. Data Storage for Analysis: Relational Databases, Big Data, and Other Options.	55
Log Data and the CRUD Paradigm	56
Creating a Well-Organized Flat File System: Lessons from SiLK	57
A Brief Introduction to NoSQL Systems	59
What Storage Approach to Use	62
Storage Hierarchy, Query Times, and Aging	64

Part II. Tools

5. The SiLK Suite.	69
What Is SiLK and How Does It Work?	69
Acquiring and Installing SiLK	70
The Datafiles	70
Choosing and Formatting Output Field Manipulation: rwcut	71
Basic Field Manipulation: rwfilt	76
Ports and Protocols	77
Size	78
IP Addresses	78
Time	80
TCP Options	80
Helper Options	82
Miscellaneous Filtering Options and Some Hacks	82
rwfileinfo and Provenance	83
Combining Information Flows: rwcount	86
rwset and IP Sets	88
rwuniq	91
rwbag	93
Advanced SiLK Facilities	93
pmaps	93
Collecting SiLK Data	95
YAF	96

rwptoflow	98
rwtuc	98
Further Reading	100
6. An Introduction to R for Security Analysts.....	101
Installation and Setup	102
Basics of the Language	102
The R Prompt	102
R Variables	104
Writing Functions	109
Conditionals and Iteration	111
Using the R Workspace	113
Data Frames	114
Visualization	117
Visualization Commands	117
Parameters to Visualization	118
Annotating a Visualization	120
Exporting Visualization	121
Analysis: Statistical Hypothesis Testing	121
Hypothesis Testing	122
Testing Data	124
Further Reading	127
7. Classification and Event Tools: IDS, AV, and SEM.....	129
How an IDS Works	130
Basic Vocabulary	130
Classifier Failure Rates: Understanding the Base-Rate Fallacy	134
Applying Classification	136
Improving IDS Performance	138
Enhancing IDS Detection	138
Enhancing IDS Response	143
Prefetching Data	144
Further Reading	145
8. Reference and Lookup: Tools for Figuring Out Who Someone Is.....	147
MAC and Hardware Addresses	147
IP Addressing	150
IPv4 Addresses, Their Structure, and Significant Addresses	150
IPv6 Addresses, Their Structure and Significant Addresses	152
Checking Connectivity: Using ping to Connect to an Address	153
Tracerouting	155
IP Intelligence: Geolocation and Demographics	157

DNS	158
DNS Name Structure	158
Forward DNS Querying Using dig	159
The DNS Reverse Lookup	167
Using whois to Find Ownership	168
Additional Reference Tools	171
DNSBLs	171
9. More Tools.....	175
Visualization	175
Graphviz	175
Communications and Probing	178
netcat	179
nmap	180
Scapy	181
Packet Inspection and Reference	184
Wireshark	184
GeoIP	185
The NVD, Malware Sites, and the C*Es	186
Search Engines, Mailing Lists, and People	187
Further Reading	188

Part III. Analytics

10. Exploratory Data Analysis and Visualization.....	191
The Goal of EDA: Applying Analysis	193
EDA Workflow	194
Variables and Visualization	196
Univariate Visualization: Histograms, QQ Plots, Boxplots, and Rank Plots	197
Histograms	198
Bar Plots (Not Pie Charts)	200
The Quantile-Quantile (QQ) Plot	201
The Five-Number Summary and the Boxplot	203
Generating a Boxplot	204
Bivariate Description	207
Scatterplots	207
Contingency Tables	210
Multivariate Visualization	211
Operationalizing Security Visualization	213

Further Reading	220
11. On Fumbling.....	221
Attack Models	221
Fumbling: Misconfiguration, Automation, and Scanning	224
Lookup Failures	224
Automation	225
Scanning	225
Identifying Fumbling	226
TCP Fumbling: The State Machine	226
ICMP Messages and Fumbling	229
Identifying UDP Fumbling	231
Fumbling at the Service Level	231
HTTP Fumbling	231
SMTP Fumbling	233
Analyzing Fumbling	233
Building Fumbling Alarms	234
Forensic Analysis of Fumbling	235
Engineering a Network to Take Advantage of Fumbling	236
Further Reading	236
12. Volume and Time Analysis.....	237
The Workday and Its Impact on Network Traffic Volume	237
Beaconing	240
File Transfers/Raiding	243
Locality	246
DDoS, Flash Crowds, and Resource Exhaustion	249
DDoS and Routing Infrastructure	250
Applying Volume and Locality Analysis	256
Data Selection	256
Using Volume as an Alarm	258
Using Beaconing as an Alarm	259
Using Locality as an Alarm	259
Engineering Solutions	260
Further Reading	260
13. Graph Analysis.....	261
Graph Attributes: What Is a Graph?	261
Labeling, Weight, and Paths	265
Components and Connectivity	270
Clustering Coefficient	271
Analyzing Graphs	273