

# Jay and Bob Strike Back: Data-Driven Security

By Russ McRee – ISSA Senior Member, Puget Sound (Seattle) Chapter



## Prerequisites

—*Data-Driven Security: Analysis, Visualization, and Dashboards*

—R<sup>1</sup> and RStudio<sup>2</sup> as we'll only focus on the R side of the discussion

—All other dependencies for full interactive use of the book's content are found in "Tools You Will Need" in the book's introduction.

When last I referred you to a book as a tool, we discussed TJ O'Connor's *Violent Python*.<sup>3</sup> I've since been knee deep in learning R and quickly discovered *Data-Driven Security: Analysis, Visualization and Dashboards* from Jay Jacobs and Bob Rudis, hereafter referred to



**Figure 1 – Jay and Silent Bob Strike Back**

Just so you know whose company you're actually keeping here, Jay is a coauthor of *Verizon Data Breach Investigation Reports* and

Bob Rudis was named one of the "Top 25 Influencers in Information Security" by Tripwire.

I was looking to make quick use of R as specific to my threat intelligence and engineering practice as it so capably helps make sense of excessive and oft confusing data. I will not torment you with another flagrant misuse of big data vendor marketing spew; yes, data is big, we get it, enough already. Thank goodness, the Internet of Things (IoT) is now the most abused, overhyped FUD-fest term. Yet, the reality is, when dealing with a lot of data, tools such as R and Python are indispensable, particularly when trying to quantify the data and make sense of it. Most of you are likely familiar with Python, but if you haven't heard of R, it's a scripting language for statistical data manipulation and analysis. There are a number of excellent books on R, but nowhere will you find a useful blending of R and Python to directly support your informa-

tion security analysis practice as seen in Jay and Bob's book. I pinged Jay and Bob for their perspective and Bob provided optimally:

"Believe it or not, we (and our readers) actually have ZeroAccess to thank for the existence of data-driven security (the book, blog, and podcast). We started collaborating on security data analysis and visualization projects just about a year before we began writing the book, and one of the more engaging efforts was when we went from a boatload of ZeroAccess latitude and longitude pairs (and only those pairs) to maps, statistics, and even graph analyses. We kept getting feedback (both from observation and direct interaction) that there was a real lack of practical data analysis and visualization materials out there for security practitioners and the domain-specific, vendor-provided tools were and are still quite lacking. It was our hope that we could help significantly enhance the capabilities and effectiveness of organizations by producing a security-centric guide to using modern, vendor-agnostic tools for analytics, a basic introduction to statistics and machine learning, the science behind effective visual communications, and a look at how to build a great security data science team.

"One area we discussed in the book, but is worth expanding on, is how essential it is for information security professionals to get plugged-in to the broader "data science" community. Watching "breaker-oriented" RSS feeds/channels is great, but it's equally as important to see what other disciplines are successfully using to gain new insights into tough problems and regularly tap into the wealth of detailed advice on how to communicate your messages as effectively as possible. There's no need to reinvent the wheel or use yesterday's techniques when trying to stop tomorrow's threats."

Well said. I'm a major advocate for the premise of moving threat intelligence beyond data brokering as Bob mentions. This book endeavors and provides the means with which to conduct security data science. According to Booz Allen's *The Field Guide to Data Science*, "data science is a team sport." While I'm biased, nowhere is that more true than the information security field. As you embark on the journey *Data-Driven Security: Analysis, Visualization and Dashboards* (referred to hereafter as DDSecBook) intends to take you on, you'll be provided with direction on all the tools you need, so we'll not spend much time there and instead focus on the applied use of this rich content. I will be focusing solely on

1 <http://cran.rstudio.com/>.

2 <http://www.rstudio.com/>.

3 <http://holisticinfosec.blogspot.com/2013/01/toolsmith-violent-python-book-review.html>.

the R side of the discussion though as that is an area of heavy focus for me at present. DDsecBook is described with the byline, **Uncover hidden patterns of data and respond with countermeasures.** Awesome, let's do just that.

## Data-driven security

DDsecBook is laid out in such a manner as to allow even those with only basic coding or scripting (like me; I am the quintessential R script kiddie) to follow along and grow while reading and experimenting:

1. The Journey to Data-Driven Security
2. Building Your Analytics Toolbox: A Primer on Using R and Python for Security Analysis
3. Learning the “Hello World” of Security Data Analysis
4. Performing Exploratory Security Data Analysis
5. From Maps to Regression
6. Visualizing Security Data
7. Learning from Security Breaches
8. Breaking Up with Your Relational Database
9. Demystifying Machine Learning
10. Designing Effective Security Dashboards
11. Building Interactive Security Visualizations
12. Moving Toward Data-Driven Security

For demonstrative purposes of making quick use of the capabilities described, I'll focus our attention on chapters 4 and 6. As a longtime visualization<sup>4</sup> practitioner I nearly flipped out when I realized what I'd been missing in R, so chapters 4 and 6 struck close to home for me. DDsecBook includes code downloads<sup>5</sup> for each chapter and the related data so you can and should play along as you read. Additionally, just to keep things timely and relevant, I'll apply some of the techniques described in DDsecBook to current data of interest to me so you can see how repeatable and useful these methods really are.

## Performing exploratory security data analysis

Before you make use of DDsecBook, if you're unfamiliar with R, you should read *An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics*<sup>6</sup> and run through Appendix A. This will provide at least an inkling of the power at your fingertips.

This chapter introduces concepts specific to dissecting IP addresses including their representation, conversion to and from 32-bit integers, segmenting, grouping, and locating, all of which leads to augmenting IP address data with the likes

```
## Download and read in current AlienVault reputation data
fileUrl <- "http://reputation.alienvault.com/reputation.data"
download.file(fileUrl, destfile = "./data/reputation.data")

## Read in AlienVault reputation data downloaded
avRep <- "data/reputation.data "
av.df <- read.csv(avRep, sep="#", header=FALSE)
colnames(av.df) <- c("IP", "Reliability", "Risk", "Type",
                    "Country", "Locale", "Coords", "x")

# Read in list of destination IP addresses
dest.ips <- read.csv("data/dest.ips.old", col.names= c("IP"))
# Review reliability of the IP address entries
# (you could also plot a histogram)
table(av.df[av.df$IP %in% dest.ips$IP, ]$Reliability)
## 1 2 3 4 5 6 7 8 9 10
## 16 828 831 170 1 266 92 2 23 24

# Extract only the "bad" ones, designated by presence in AlienVault
# database with a reliability greater than 6 since there seems to
# be a trailing off at that point
ips <- as.character(av.df[(av.df$IP %in% dest.ips$IP) &
                          (av.df$Reliability > 6), ]$IP)

# graph it
g.cc <- graph.cc(ips, av.df)
```

Figure 2 – R code to match bad IPs to AlienVault reputation data

of IANA data. This is invaluable when reviewing datasets such as the AlienVault reputation data, mentioned at length in chapter 3, and available as updated hourly.

We'll jump ahead here to Visualizing Your Firewall Data (listing 4-16) as it provides a great example of taking methods described in the book and applying them immediately to your data. I'm going to set you up for instant success, but you will have to work for it a bit. The script we're about to discuss takes a number of dependencies created earlier in the chapter; I'll meet them in the script for you (you can download<sup>7</sup> it from my site), but only if you promise to buy this book and work though all prior exercises for yourself. Trust me, it's well worth it. Here's the primary snippet of the script, starting at line 293 after all the dependencies are met. What I've changed most importantly is the ability to measure an IP list against the very latest AlienVault reputation data. Note, I found a bit of a bug here that you'll need to update per the DDsecBook blog.<sup>8</sup> This is otherwise all taken directly ch04.r in the code download with specific attention to listing 4-16 as seen in figure 2.

I've color coded each section to give you a quick walkthrough of what's happening.

1. Defines the URL from which to download the AlienVault reputation data and provides a specific destination to download it to.
2. Reads in the AlienVault reputation data, creates a data frame<sup>9</sup> from the data, and provides appropriate column names. If you wanted to read the top of that data from the data frame, using `head(av.df, 10)` would result in figure 3.
3. Reads in the list of destination IP addresses, from a firewall log list as an example, and compares it against matches on the reliability column from the AlienVault reputation data.

4 <http://holisticinfosec.blogspot.com/search?q=visualization>.

5 <http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118793722.html>.

6 <http://cran.r-project.org/doc/manuals/R-intro.pdf>.

7 <http://holisticinfosec.org/toolsmith/files/R/IP2AV.R>.

8 <http://datadrivensecurity.info/blog/posts/2014/Aug/book-bug-graph-cc/>.

9 <http://www.r-tutor.com/r-introduction/data-frame>.

4. Reduces the dataset down to only matches for reliability above a rating of 6 as lower tends to be noise and of less value.

5. Produces a graph with the graph.cc function created earlier in the ch04.r code listing.

```

Console C:/coding/R/Data_Driven_Security_Analysis_Visualization_and_Dashboards/ch04/
> head(av.df, 10)
  IP Reliability Risk Type Country Locale Coords x
1 37.205.198.162 4 2 Scanning Host IT 42.8333015442,12.8332996368 11
2 211.160.19.250 4 2 Scanning Host CN Hang 30.2936000824,120.161399841 11
3 203.121.165.16 6 5 C&C TH 15.0,100.0 2
4 211.151.57.196 4 2 Scanning Host CN Beijing 39.9289016724,116.388298035 11
5 108.59.1.5 4 2 Scanning Host AI 0.0,0.0 11
6 108.59.5.139 4 2 Scanning Host US 39.6734008789,-75.7052001953 11
7 108.85.139.165 4 2 Scanning Host US Los Angeles 34.043800354,-118.251197815 11
8 109.169.60.121 4 2 Scanning Host US 38.0,-97.0 11
9 109.230.232.108 4 2 Scanning Host GB 51.5,-0.129999995232 11
10 110.164.217.122 4 3 Scanning Host TH Bangkok 13.7539997101,100.501403809 11
    
```

Figure 3 – The top ten entries in the Alien Vault data frame

The results are seen in figure 4 where I mapped against the Alien Vault reputation data provided with the chapter 4 download versus brand new AlienVault data as of 25 AUG 2014.

open SMB shares populated with sensitive data) in the form of a single Excel workbook with data for specific dates created as individual worksheets (tabs). My original solution was to

What changed, you ask? The IP list provided with chapter 4 data is also a bit dated (over a year now) and has likely been cleaned up and is no longer of ill repute. When I ran a list of 6100 IPs I had that were allegedly spammers, only two were identified as bad: one a scanning host, the other for malware distribution.

Great stuff, right? You just made useful, visual sense of otherwise clunky data, in a manner that even a C-level executive could understand. :-)

Another example that follows the standard set in chapter 6 comes directly from a project I'm currently working on. It matches the principles of said chapter as built from a quote from Colin Ware regarding information visualization:

“The human visual system is a pattern seeker of enormous power and subtlety. The eye and the visual cortex of the brain form a massively parallel processor that provides the highest bandwidth channel into human cognitive centers.”

Yeah, baby, plug me into the Matrix! Jay and Bob paraphrase Colin to describe the advantages of data visualization:

- Data visualizations communicate complexity quickly
- Data visualizations enable recognition of latent patterns
- Data visualizations enable quality control on the data
- Data visualizations can serve as a muse

To that end, our example.

I was originally receiving data for a particular pet peeve of mine (excessively permissive

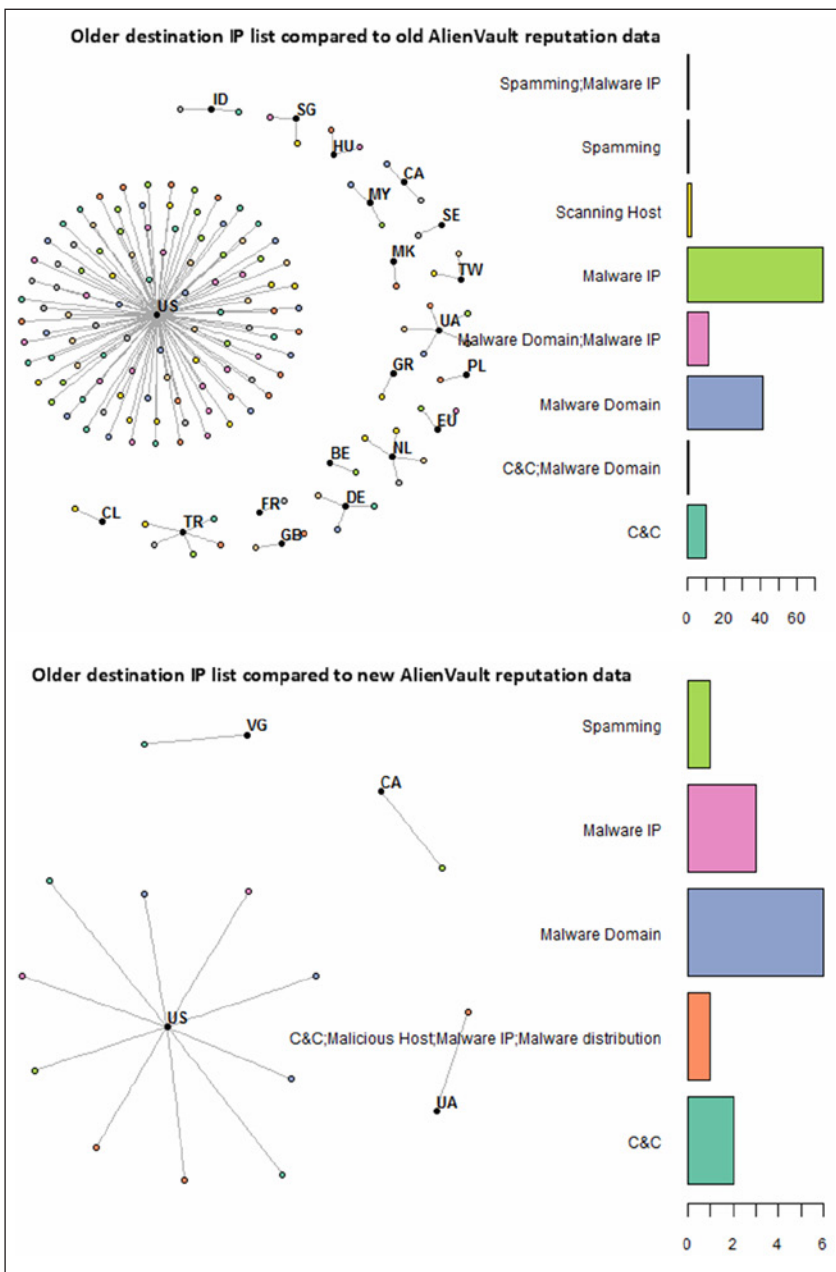


Figure 4 – Bad IPs mapped against Alien Vault reputation data by type and country

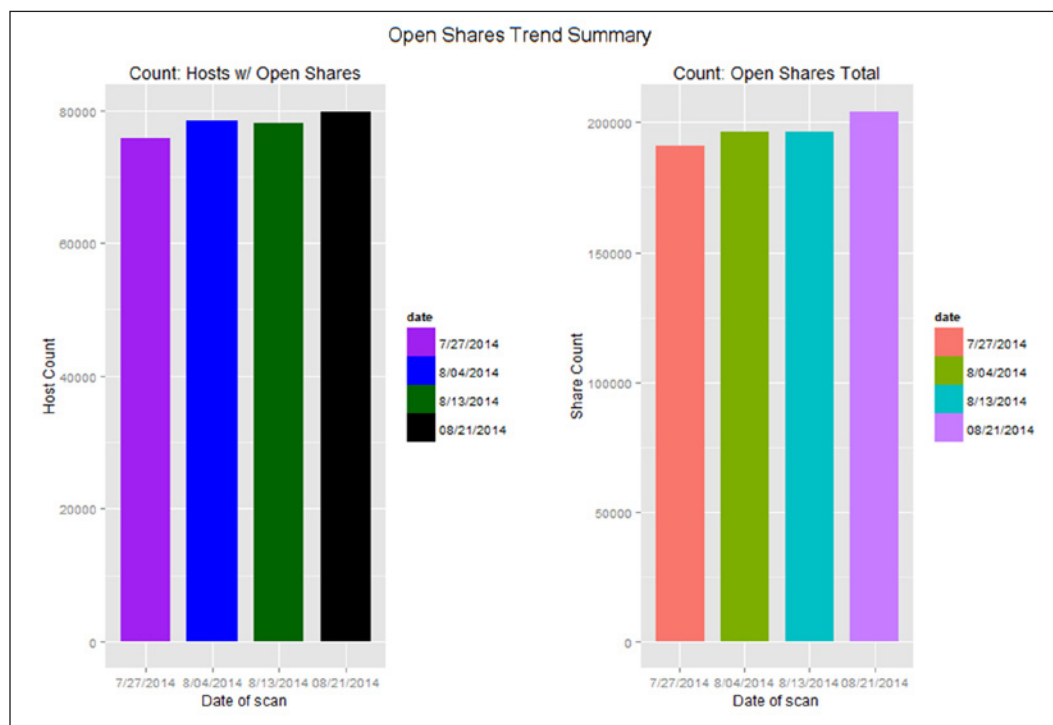


Figure 5 – Open shares host and shares counts trending over time

save each worksheet as individual CSVs then use the `read.csv` function to parse each CSV individually for R visualization. Highly inefficient given the like of the `XLConnect` library that allows you to process the workbook and its individual worksheets without manipulating the source file.

Before:

```
raw <- read.csv("./data/openshares/ShareStats0727.csv")
h <- sum(raw$HostCt)
s <- sum(raw$ShareCt)
```

After:

```
sharestats <- loadWorkbook("./data/openshares/ShareStats_8_21.xlsx")
sheet1 <- readWorksheet(sharestats, sheet = 1)
h1 <- sum(sheet1$HostCt)
s1 <- sum(sheet1$ShareCt)
```

The first column of the data represented the number of hosts with open shares specific to a business unit, the second column represented the number of shares specific to that same host. I was interested in using R to capture a total number of hosts with open shares and the total number of open shares over all and visualize in order to show trending over time. I can't share the source data with you as its proprietary, but I've hosted the R code for you.<sup>10</sup> You'll need to set your own working directory and the name and the path of the workbook you'd like to load. You'll also need to define variables based on your column names. The result of my effort is seen in figure 5.

As you can see, I clearly have a trending problem: up versus down is not good in this scenario.

<sup>10</sup> <http://holisticinfosec.org/toolsmith/files/R/OpenShares.v0.3.R>.

While this is a simple example given my terrible noob R skills, there is a vast green field of opportunity using R and Python to manipulate data in such fashion. I can't insist enough that you give it a try.

## In conclusion

Don't be intimidated by what you see in the way of code while reading *DDSecBook*. Grab R and R Studio, download the sample sets, open the book, and play along while you read. I also grabbed three other R books to help me learn including *The R Cookbook* by Paul Teeter, *R for Everyone* by Jared Land-

er, and *The Art of R Programming* by Normal Matloff. There are of course many others to choose from. Force yourself out of your comfort zone if you're not a programmer; add R to your list if you are, and above all else, as a security practitioner make immediate use of the techniques, tactics, and procedures inherent to Jay and Bob's most excellent *Data-Driven Security: Analysis, Visualization and Dashboards*.

Ping me via email if you have questions (russ at holisticinfosec dot org).

Cheers...until next month.

## Acknowledgements

—Bob Rudis, @hrbrmstr, *DDSecBook* co-author, for his contributions to this content and the quick bug fix, and Jay Jacobs, @jayjacobs, *DDSecBook* co-author.

## About the Author

Russ McRee manages the Threat Intelligence & Engineering team for Microsoft's Online Services Security & Compliance organization. In addition to *toolsmith*, he's written for numerous other publications, speaks regularly at events such as DEFCON, Black Hat, and RSA, and is a SANS Internet Storm Center handler. As an advocate for a holistic approach to the practice of information assurance Russ maintains [holisticinfosec.org](http://holisticinfosec.org). He serves in the Washington State Guard as the Cybersecurity Advisor to the Washington Military Department. Reach him at [russ at holisticinfosec dot org](mailto:russ@holisticinfosec.org) or @holisticinfosec.