Chapter 4: Analyzing “Badness”

“Sometimes, bad *is* bad.”

Huey Lewis & The News, Sports, Chrysallis Records, 1983

This chapter will continue the exploration of the AlienVault IP Reputation database that began in Chapter 3 and assumes the reader is familiar with the description of the data set and has at least followed along with the preliminary analyses. As stated in the introduction of the book, no attempt has been made to incorporate consideration of or conduct analyses on Internet Protocol (IP) version 6 (IPv6) addresses and all the examples found in this chapter will be based on IPv4. Given the slow adoption and migration to IPv6 plus the plethora of “badness” still on IPv4 networks, this should not be a practical limitation in any way, shape or form.

The struggle to protect, defend and understand our modern networks begins and ends—more often than not—with an IP Address. IP addresses are defined in RFC 791, the “Internet Protocol / DARPA Internet Program / Protocol Specification” (http://tools.ietf.org/html/rfc791), which has an elegant and succinct way of describing them:

“A *name* indicates what we seek. An *address* indicates where it is. A *route* indicates how to get there.”

Global entities slice and dice them for public and private use; devices, systems and applications log them for future reference; network management systems test, group, display and report on them; and, security tools make critical decisions based upon them. But, what—exactly—*is* an IP address and what part do they play in the quest for finding and mitigating “badness”?

Dissecting The “IP Address”

Unless you write operating system level code for a living or create and/or contribute to network-level tools such as tcpdump (<http://www.tcpdump.org/>) or Wireshark (<http://www.wireshark.org/>) on a regular basis, you may not have given IP addresses more of thought than the strings you use with a ping, nessus or nmap command. To perform security-oriented analyses of your system and network data, you must fully understand as much as you can about security domain data elements, just as those who perform data analyses in financial, agricultural or bio-medial disciplines must understand the underpinnings of the data elements in those fields. Since IP addresses are one of the most fundamental security domain data elements, let’s dig a bit deeper into them so we can fully integrate them into our analytics endeavors.

IPv4 addresses are comprised of four bytes, known as octets, and we usually come across them in a form called dotted decimal notation (e.g. “192.168.1.1”). Since we know an 8-bit byte can range in value from 0 through 255, that means the dotted decimal range is 0.0.0.0 through 255.255.255.255 for a total of 4,294,967,296 possible addresses (the maximum value of a 32-bit integer). This is an important fact to remember: *any IP address can be converted to/from a 32-bit integer value*. The first reason why this is important has to do with saving time and space. If you are writing or using a tool that only perceives an IP address as a character string or set of character strings, then you are potentially wasting space by trading a 4-byte/32-bit representation for a (worst case) 15-byte/120-bit representation. Furthermore, you are also choosing to use less efficient string comparison code versus integer arithmetic and comparison plus bitwise operations to accomplish the same tasks. While this may have little-to-no impact in some scenarios, the repercussions grow significant when dealing with large volumes of IP addresses (and become worse in the IPv6 world).

Due to the way TCP/IP was designed and how IPv4 networks are implemented there are numerous ways to segment/group them to make it easier to manage individual networks and interoperate in the global network.

32-bit integer (“how does your computer see an IP address?”) + machine info

Part of a subnet / logical layout / MAC addresses, perhaps has a hostname (DNS)

Larger context: part of a global network organized by ASNs (BGP)

Lager context: Has a physical location

Mapping Outside the Continents

USE CASE: Visualizing AlienVault ASN data (force-directed network graphs of malhost ASN groupings)

Augmenting Data

In an analyst’s dream world, every data set you are asked to crunch through would be error-free and have all the attributes necessary for thorough and robust analyses. Sadly, information security is no different from other disciplines (i.e. “we aren’t special”) when it comes to imperfect data sets and highly distributed referential data or more metadata sources. This *can* pose challenges to effective data analyses, but it is usually possible to find and use the data you need.

Even though we have geographic information in our AlienVault data set, the internet has both physical and logical groupings, which we will cover more in the next chapter. It might be interesting to see how this data looks through a different lens, and for this example we’ll augment our data set with additional data from the IANA IPv4 Address Space Registry (https://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml). This data is a very high level grouping of IPv4 address space registry allocations and most of the registrants are not responsible for the malicious activity of individual nodes. So, while we cannot use this information to cast blame, it will give us one view of where badness is clustered, enabling us to perform additional investigations which we’ll cover in Chapter 4.

type="tip"

IANA provides a handy link to the CSV version of the IPv4 address space allocations as well as a link to the traditional annotated text file. If you run the example code, you may see some strange behavior at times due to the CSV file being incomplete. You can either practice your data munging skills and convert the fixed-width version in the text file to CSV or use the version of the CSV that’s on our companion web site if you encounter any issues.

The data frame foundational data structure in *R* and *pandas* makes it very straightforward to reference and incorporate new data into our analyses and your own projects will follow something close to this basic pattern:

* downloading (if necessary) of new data
* parsing and converting the new data into a data frame
* validating the contents and structure of the new data
* performing any necessary munging of the new data to make it easier to process/incorporate
* performing any necessary munging of the existing data to make it easier to incorporate the new data
* extracting or computing relevant information from the new data source
* creating one or more new columns in our existing data frame
* running new analyses

For this example, we process the IANA data to see which registry allocations have the most malicious nodes.

R code to incorporate IANA IPv4 Allocations

# retrieve IANA prefix list

ianaURL <- "http://www.iana.org/assignments/\

ipv4-address-space/ipv4-address-space.csv"

ianaData <- "data/ipv4-address-space.csv"

if (file.access(ianaData)) {

download.file(ianaURL,ianaData)

}

iana <- read.csv(ianaData,stringsAsFactors=FALSE)

str(iana) # examine it(now shown below)

# clean up the iana prefix

iana$Prefix <- sub("^(00|0)","",iana$Prefix,perl=TRUE)

iana$Prefix <- sub("/8$","",iana$Prefix,perl=TRUE)

head(iana$Prefix) # not shown

# extract just the prefix from the AlienVault list

av.IP.prefix <- sapply(strsplit(av$IP,'.',fixed=TRUE),"[",1)

av$Designation <- sapply(av.IP.prefix,function(ip) {

iana[iana$Prefix == ip,]$Designation

})

desig <- summary(factor(av$Designation))

desig <- desig[order(-desig)]

desig

APNIC RIPE NCC

93775 74789

ARIN LACNIC

42358 18914

Administered by ARIN Administered by RIPE NCC

17974 5893

Administered by APNIC AFRINIC

2615 1896

Administered by AFRINIC Level 3 Communications, Inc.

322 31

PSINet, Inc. AT&T Bell Laboratories

30 24

Hewlett-Packard Company Digital Equipment Corporation

3 1

Python code to incorporate IANA IPv4 Allocations

# retrieve IANA prefix list

ianaURL = "http://www.iana.org/assignments/\

ipv4-address-space/ipv4-address-space.csv"

ianaData = "data/ipv4-address-space.csv"

if not os.path.isfile(ianaData):

urllib.urlretrieve(ianaURL, filename=ianaData)

iana = pd.read\_csv(ianaData)

iana # examine it (not shown below)

# clean up the iana prefix

iana['Prefix'] = iana['Prefix'].map(lambda x:

str(int(x.rstrip("8").rstrip("/"))))

iana['Prefix'] # (not shown)

# extract just the prefix from the AlienVault list

avPrefix = [ octet[0] for octet in av['IP'].str.split('.') ]

av['Designation'] = [ iana[(iana['Prefix'] == prefix)].Designation

for prefix in avPrefix ]

pd.value\_counts(av['Designation'])

[APNIC] 93775

**[RIPE NCC] 74789**

[ARIN] 42358

**[LACNIC] 18914**

**[Administered by ARIN] 17974**

[Administered by RIPE NCC] 5893

[Administered by APNIC] 2615

[AFRINIC] 1896

[Administered by AFRINIC] 322

[Level 3 Communications, Inc.] 31

[PSINet, Inc.] 30

[AT&T Bell Laboratories] 24

[Hewlett-Packard Company] 3

[Digital Equipment Corporation] 1

dtype: int64

We can do a quick check against the main IANA allocation table to see if this matches overall block assignments:

R code (no Python example) to extract IANA block assignments

df <- data.frame(table(iana$Designation),stringsAsFactors=FALSE)

colnames(df) <- c("reg","ct")

av.reg <- df[df$reg %in% names(desig),]

av.reg[with(av.reg, order(-ct)),]

reg ct

8 APNIC 45

**3 Administered by ARIN 44**

10 ARIN 36

**39 RIPE NCC 35**

**33 LACNIC 9**

2 Administered by APNIC 6

5 Administered by RIPE NCC 4

6 AFRINIC 4

1 Administered by AFRINIC 2

34 Level 3 Communications, Inc. 2

12 AT&T Bell Laboratories 1

18 Digital Equipment Corporation 1

28 Hewlett-Packard Company 1

38 PSINet, Inc. 1

There is some variation, but overall—as expected—the larger blocks contribute the majority of malicious hosts. We’ve highlighted “RIPE NCC”, “Administered by ARIN” and “LACNIC” in the output since “RIPE NCC” has a significantly larger number of malicious hosts than it’s assignment count might imply (nearly double that of it’s very close neighbor “ARIN”) and “LACNIC” and “Administered by ARIN” both have a similar number of malicious hosts yet have very different allocation counts. Delving into why might be an interesting exercise (for the intrepid reader).