Where Are Malicious Networks Located?

Gorkem Yakin

July 13, 2012

1 Introduction

Today, we are no longer using computers that aren't connected to the Internet and we are getting more dependent each day. Having access to the Internet allows us to use e-mails to contact each other, to access our bank accounts online to pay our rents or to manage our taxes.

However, these same benefits of the Internet also cause us to be targeted by criminals. We are sent e-mails containing malware or are directed to websites that try to steal sensitive information like usernames, password and credit card numbers.

Although some of the machines used for malicious purposes are short lived, there are some Internet companies or service providers who knowingly host machines that run for a long time. If we can automatically find parts of the Internet where these malicious machines might be hosted, we can come up with better dealing mechanisms.

To this end, in the rest of the paper, we first describe four major types of networks and what their important characteristics are. Then, we look at networks that have long running malicious hosts and discuss how they are related to the networks types.

2 Network Types

There are several network types, but we focus on four major network types which are corporate, university, hosting and consumer broadband networks.

Corporate Networks. These networks belong to big corporations. These networks are usually protected by central network layer firewalls and only machines that need to be accessible from the outside are allowed to accept connections.

University Networks. Networks in this group belong to universities. Generally, network resources in these networks are controlled by a central point, but some departments can manage their own resources. Like corporate networks, these networks also might be protected by firewalls.

Hosting Networks. These networks consist of machines that are leased by customers and are used to serve content to the Internet. A hosting company manages the machines, but the served content is managed by customers of these companies.

Consumer Broadband Networks. Networks in this group can have two connections types, digital subscriber line (DSL) offered by telephone companies and cable networks offered by cable television providers, and consist of machines located in homes or small organizations. Machines in these networks are usually accessible except at certain ports like the SMTP port (25).

3 Data Collection

To understand the characteristics of different networks, we need samples for each network type. In this section, we discuss in detail how we selected and scanned the networks involved in our study.

3.1 Network Selection

To select the networks, we used the following sources:

- For corporate networks, we used the Fortune 500 list [3]. This list ranks top 500 corporations in the United States based on their gross revenues.
- For university networks, we used the national university rankings published by U.S. News and World Report [4].
- For consumer broadband networks, we used dslreports.com forums. These forums are used by U.S. based broadband subscribers to share reviews or discuss problems they are having with their connections. We also selected networks in Austria and Turkey that we know are used for broadband connections.
- For hosting networks, we did Google searches. Hosting companies are easy to find just by doing web searches, so we didn't use any particular list.

The lists we used for corporate and university networks only list the names of the organizations. To find the network ranges used by these organizations, we first searched their names on Google and found URLs of their websites. Then, we performed DNS lookups using the hostnames in the URLs and found the corresponding IP addresses. Using these IP addresses, we performed whois lookups and collected the network ranges listed in the results.

We were able to directly find the URLs of hosting companies' websites, so we applied the methodology described above starting from the DNS lookups.

On the dslreports.com forums, some users posted the public IPs used by their modems. Therefore, without doing any web searches or DNS lookups, we used these IPs to find the corresponding network ranges by doing whois lookups.

In the end, we selected 34696 /24 networks with the following breakdown: 12150 of them belong to 53 universities, 17986 of them belong to 100 corporations, 3240 of them belong to 43 hosting companies and 1320 of them belong to 6 consumer broadband providers.

3.2 Scanner

The scanner can be divided into four parts: the ASN mapper, the port scanner, the application data collector and the DNS resolver. We ran our scanner on each IP address that didn't end with either 0 or 254. We skipped .0 addresses because it is used to address networks and skipped .255 addresses because they are used as broadcast addresses.

First we found the autonomous systems associated with the IP address using the IP address to ASN mapper provided by Team Cymru Research NFP [5] and eliminated IP addresses that didn't have an ASN associated with them.

Next, we scanned the hosts within a /24 network for open ports. We wanted to collect information about ports used for different purposes. However, since it is infeasible to scan all ports, we selected the 21 most commonly used ports which we divided into 7 groups. The list of the ports we scanned can be seen in Table 1.

At the core of our scanner, we used Nmap [2] as the port scanner. With its extensive capabilities, Nmap allowed us to parallelize our scanning activities so that we could work on multiple networks concurrently.

Nmap works well for scanning TCP ports, but when it comes to UDP ports, its results may not be reliable. The UDP specification [6] doesn't require a host to respond to unknown UDP packets. When a host receives a UDP packet and an application is listening on the port specified in the UDP packet, the contents of the packet is passed to the application. If the application doesn't understand the contents of the packet, it may not send back a reply. During our scanning, we used version 5.21 of Nmap and this version generates correct UDP packets for only one of the three UDP ports, ISAKMP, that we scanned. It is supposed to also support the DNS port, but during our tests, we realized that it didn't work correctly for some hosts. The reason is that Nmap generates a DNS status request for port 53, but some hosts don't respond to this request even though there is a DNS server application running on the host. Therefore, for DNS and OpenVPN ports, we generated our own UDP packets.

After the ports were scanned, the scanner started collecting application data. With the data collector, we retrieved banners of FTP, mail and SSH server applications. Also, from the web server applications, we

Port Group	Port Number	Port Explanation
Web / FTP	21/TCP	FTP
	80/TCP	HTTP
	443/TCP	HTTP over SSL
Mail	25/TCP	SMTP
	465/TCP	SMTP over SSL
	587/TCP	Submission
	110/TCP	POP3
	995/TCP	POP3 over SSL
	143/TCP	IMAP
	993/TCP	IMAP over SSL
Remote Access	22/TCP	SSH
	3389/TCP	Remote Desktop
Samba	135/TCP	Location Service
	139/TCP	NETBIOS Session Service
	445/TCP	Microsoft Naked CIFS
LDAP	389/TCP	LDAP
	636/TCP	LDAP over SSL
DNS	53/UDP	DNS
VPN	500/UDP	ISAKMP
	1194/UDP	OpenVPN
	1723/TCP	PPTP

Table 1: Port Based Features

collected response headers sent in response to GET requests for the index page.

In parallel to the port scanner and data collector, the scanner ran a DNS resolver. For each host, this part of the system performed a reverse DNS lookup using the host's IP when the port scanner started scanning it and collected the hostname(s) associated with the IP. Using the hostnames, it also did forward DNS requests to find the IPs and mail servers associated with them.

Scanning more than limited number of ports and hosts would be disturbing from the scanned organizations' point of view if done wrong, so we took a few precautions. First of all, instead of scanning each host quickly, we used a low packet rate, 1 packet per second, but parallelized our activities to multiple /24 networks. Moreover, we served a web page on each host we used during our scanning explaining the purpose of our activities. These pages also contained our e-mail addresses in case network administrators wanted to contact us. During our scanning, we received only one complaint and we stopped our activities directed towards the organization's networks immediately.

4 Data Analysis

Before we looked at the data, we first filtered out the networks that were not being used. We marked networks that didn't have any hosts with at least one open port or one PTR record as inactive and didn't include them in our analysis.

Looking at the remaining networks, we selected eight features described below.

The first feature is the number of live hosts. We counted the number of hosts that responded to packets for FTP, HTTP and SMTP ports.

The second features is the number of hosts responding to HTTP requests. While analyzing the data we had, we saw that some hosting networks had

a high number of hosts that responded to HTTP requests.

The third feature is the number of PTR records with the most common hostname schema. Hostname schema is basically a template used by a single entity to generate hostnames in its networks. To get the schema, we replace numbers with zeroes and dashes with dots. We performed the dashed-to-dots mapping because while analyzing the data, we saw that they were used interchangeably in some networks. An example hostname schema is `dsl.0.0.0.ttnet.net.tr` which is generated from `dsl88.245-769.ttnet.net.tr`. The importance of this feature is that almost all PTR records in each /24 consumer broadband network have the same schema.

The fourth feature is the number of PTR records with the most common domain name. The importance of this feature is that some /24 networks managed by a single entity have high values for this feature.

The fifth feature is the number of distinct domain names in PTR records. The importance of this feature is that some hosting networks have high number of domain names.

The sixth feature is the ratio of letters in hostnames after the domain names are stripped from them. The ratio is calculated as (# of letters)/(# of alphanumerical characters). The importance of this feature is that some corporate and hosting networks have low values for this feature.

The seventh feature is the number of hosts that have the most common FTP banner. Machines in hosting networks are usually setup by the same network administrators or use the same sources to install software, so if multiple hosts in a network have similar FTP banners, we expect that network to be a hosting network. Our similarity criteria ignores the numerical differences between banners caused by version numbers, active user counts or time.

The eight feature is the distinct domain names in SMTP banners. The SMTP protocol says that a banner should start with a hostname. The importance of this feature is that some hosting networks have higher values for this feature.

After generating these features, we applied the Random Forest classification algorithm with 10-fold cross validation. The confusion matrix generated by the classifier is shown in Table 2.

Although we collected a lot of a lot of information, we only use only a subset of them. This is because we did the classification using different subsets of the information we have and got the best results using only these features.

Looking at the classification results, we can see that we can distinguish consumer broadband networks very well. For corporate, hosting and university networks, we can distinguish most of them well, but there are some that are incorrectly labeled as corporate or university. After looking at the data for these mislabeled networks, we saw this was due to the fact that these networks weren't very active, i.e. had few hosts with open ports and few DNS records.

5 Malicious Networks

A malicious network is defined by Stone-Gross et al. as a network that knowingly allows malicious activity to take place for extended periods of time [7]. FIRE is a system that lists such malicious networks [1]. In this section, we discuss how these networks are related to the four major network types.

The list published by the FIRE system contains IPs of malicious hosts and it is updated daily. As the malicious hosts are taken down, their IPs are removed from the list and as new malicious hosts emerge, their IPs are added. We only wanted to include long living malicious hosts in our analysis, so for each group, we found the malicious hosts that were listed on FIRE for a week starting from October 25 and grouped them according to their /24 networks. After this operation, we were left with 1411 /24 malicious networks.

Looking at Table 3, we can see that most malicious networks are labeled as hosting networks. There are some which are marked as corporate and university networks, but these are because these networks weren't very active.

	Consumer Broadband	Corporate	Hosting	University
Consumer Broadband	1244	5	12	46
Corporate	3	2871	43	758
Hosting	21	127	2554	169
University	54	562	56	8708

Table 2: Classification of Normal Networks

	Consumer Broadband	Corporate	Hosting	University
Malicious	8	115	1058	230

Table 3: Classification of Malicious Networks

References

[1] International Secure Systems Lab. FIRE: FInding RoguE Networks. `http://www.maliciousnetworks.org/`.

[2] Gordon Lyon. Nmap – Free Security Scanner For Network Exploration & Security Audits. `http://www.nmap.org`, 2010.

[3] Cable News Network. Fortune 500. `http://money.cnn.com/magazines/fortune/fortune500/2010/full_list/index.html`, 2010.

[4] U. S. News and World Report. National University Rankings. `http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities`, 2010.

[5] Team Cymru Research NFP. IP to ASN Mapping – Team Cymru. `http://www.team-cymru.org/Services/ip-to-asn.html`, 2010.

[6] J. Postel. RFC 768 - User Datagram Protocol. `http://www.ietf.org/rfc/rfc0768.txt`, 1980.

[7] Brett Stone-Gross, Christopher Kruegel, Kevin C. Almeroth, Andreas Moser, and Engin Kirda. Fire: Finding rogue networks. In *ACSAC*, pages 231–240, 2009.