Botnet Event Analysis

1. BOTNET SCANNING EVENT ANALYSIS

In our one year honeynet traffic, we found 43 botnet global scan events. We first analyzed the overall sender (bot) characteristics of the all the senders. Then, we analyzed each event individually and compare the characteristics among different events.

In this book chapter, we focused on the following characteristics of botnet scanning behavior.

- Bot IP distribution and AS distribution
- Bot operating system characteristics
- Botnet scan arrival behavior
- Bot arrival and departure process observed in the scanning events
- Bot observed local scan rate behavior
- Botnet scanning source and destination relationship

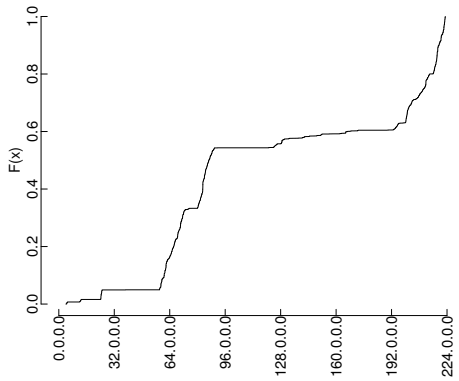

Figure 1: The number of all unique source IP addresses, as a function of IP address space. On the x-axis, IP address space is binned by /24.

1.1 Source Characteristics of Bots

We observed thousands of senders in most of the events. In 43 events, we totally observed 63,851 unique senders. Figure 1 shows the number of senders (bots) observed over all the events, as a function of IP address space. The overall trend is very similar to the spamming IP distribution in [2]. From the figure we knew, most bots are from 60.* – 90.* and 193.* – 222.* and some are from 24.* (cable modem provider). The figure illustrated that the bots mostly come from quite concentrated IP ranges. This result confirmed the result from the bot spamming behavior study [2].

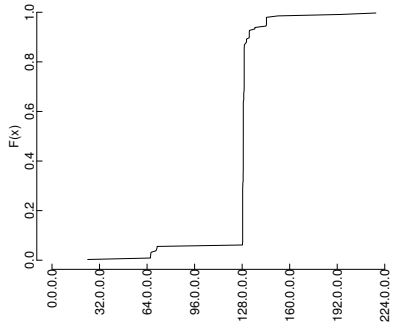

Figure 2: The number of all unique source IP addresses for the event on TCP port 2967 on 2006-11-26, as a function of IP address space. On the x-axis, IP address space is binned by /24.

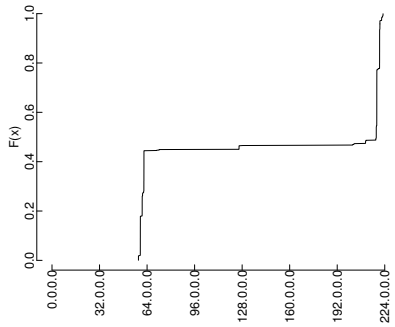

Figure 3: The number of all unique source IP addresses for the event on TCP port 5000 on 2006-8-24, as a function of IP address space. On the x-axis, IP address space is binned by /24.

We also analyzed the IP space distribution for every event. We found for most events we got the similar IP space distribution as figure 1. However, there are some events whose IP space distributions are far from the total distribution. Figure 2 and Figure 3 shows a few such examples. Since different events might be corresponding to different botnets, this implies the IP space distributions of different botnets can be quite different. Therefore, the coarse grain IP range based botnet filtering or detect might not work well in practice.

In our study, we found most bots are from a relative small number of ASes. More than 22% of bots are from the five ASes, and 41% of the bots from 20 ASes. In Table 1, we showed the top 20 ASes and the corresponding number of bots for each AS. From the analysis of the top 20 ASes, we found about 21% of the bots are Asia, mainly from China,

AS number	#Source	AS Name	Primary Country
4134	4449	CHINANET-BACKBONE	China
9318	2988	Hanaro Telecom Inc	Korea
3462	2712	Data Communication Business Group	Taiwan
4837	2091	CHINA169-BACKBONE	China
5617	1849	Polish Telecom's commercial IP network	Poland
7132	1660	SBC Internet Services	United States
6327	1545	Shaw Communications Inc.	Canada
19262	1441	Verizon Internet Serv	United States
3320	1060	Deutsche Telekom AG	Germany
3352	855	Internet Access Network of TDE	Spain
7738	744	Telecomunicacoes da Bahia S.A	Brazil
20961	675	Autonomous System	Poland
577	619	Bell Canada	Canada
3269	609	Telecom ITALIA	Italy
9394	541	CHINA RAILWAY Internet(CRNET)	China
12322	533	PROXAD AS for Proxad/Free ISP	France
8167	498	Telecomunicacoes de Santa Catarina SA	Brazil
3356	493	Level 3 Communications	United States
25310	469	Cable and Wireless Access LTD	United Kingdom
4766	429	Korea Telcom	Korea

Table 1: Amount of scan received from botnet scanning in the top 20 ASes.

Operating System	Clients
Windows	58797 (92%)
-Windows 2000 or XP	58028 (90.8%)
-Windows 98	404 (0.63%)
-Windows NT	329 (0.51%)
-Windows 2003	25 (<.1%)
-Windows 95	11 (<.1%)
Linux	9 (<.1%)
Novell	23 (<.1%)
HP-UX	1 (<.1%)
Unidentified	5021 (7.8%)
Total	63851

Table 2: The operating system distribution for unique senders of received scan, as determined by passive OS fingerprinting.

Korea and Taiwan. Europe and North America (Unite States and Canada) have similar amount of bots 9.5% and 9% respectively. Surprisingly there are also about 2% bots coming from Brazil. The bot population is from 2860 ASes in total. Although our honeynet detection sensor is in Unite States but the bots indeed come from all over the world. The overall result are similar to the result from [2]. The difference between our result and the result from [2] is mainly that we observed more hosts from Europe than them.

1.2 Operating Systems of Bots

We also investigated the prevalence of operating system among the bots. We used p0f [3] tool to identified the operating system versions. P0f is a passive OS fingerprinting tool which mainly uses the TCP options within the TCP SYN packets to identify the operating system versions. For each bot, we might observe multiple SYN packets. Sometimes, the different SYN packets from a bot might be given different OS results by p0f. We used the following priorities to solve the potential conflict. We think the other OS types have higher priority than Windows, and Windows has higher priority than Unknown. The rule is to favorite the non-Windows operating systems and to try to avoid assigning Unknown. Table 2 shows the operating system distribution we found. We found 92% of the bots are identified as Windows machines by p0f [3]. And among the Windows machines, 90.8% of the bots are Windows 2000 or XP. This result supported the conventional wisdom that botnet army are mainly comprised Windows machines.

We also did the similar analysis at per event level. We found for all the 43 events the dominated operating system are Windows. We did not observe any events which mainly consist of other types of machines. Although, there are some rumors that some botnets are Linux or Unix based, based on our finding, we believe the percentage of non-Windows based botnets in the botnet population are really low.

1.3 Scan Arrival Characteristics

For all the botnet events, we analyzed how the scan sessions arrive in time. We found for most events the very beginning and the very end of the events have complex arrival

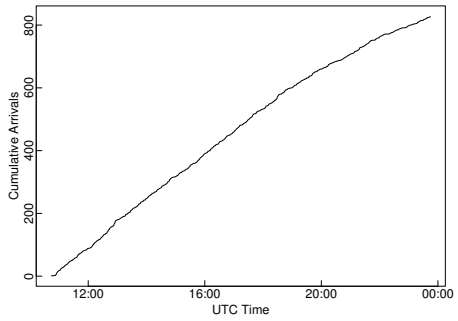

Figure 4: The cumulative scan session arrival process of the event on TCP port 8888 on 2006-02-06, which corresponding to a backdoor shell.

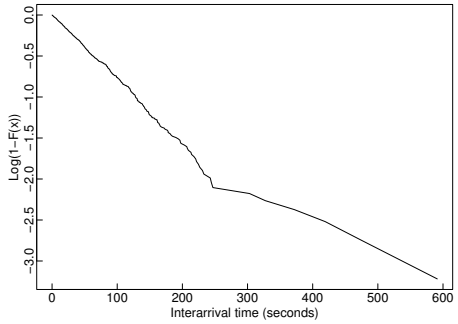

Figure 5: The inter-arrival time log scale CDF of the event on TCP port 8888 on 2006-02-06, which corresponding to a backdoor shell.

behavior. However, for most events in the middle part, the scan arrival speeds are quite constant, and the more than half of the events' inter-arrival time follows exponential distributions. This suggested that the scan arrivals follow a Poisson distribution. One plausible explanation for this is based on the law of rare events. Usually the botnet scans a large IP scope, and the sensor is only a tiny portion of it. If the botnet uses random scanning, for each scan session there is a small probability p to arrive the honeynet detection sensor. According the law of rare events, the observed scan sessions in a given time interval will follow a Poisson distribution and the inter arrival time will follow an exponential distribution. Among the 43 events, about 25 (58%) events the inter-arrival time follows an exponential distribution. This suggested most botnets indeed use a random scan strategy. An example of the scan arrival and scan inter arrival time is shown in Figure 4 and Figure 5 respectively.

1.4 Source Arrival and Departure

We also investigated for each event when the bots are observed. We defined, for a given bot, the time it begins to scan as its true source arrival time, and the time it stops to scan as its true source departure time. We cannot measure

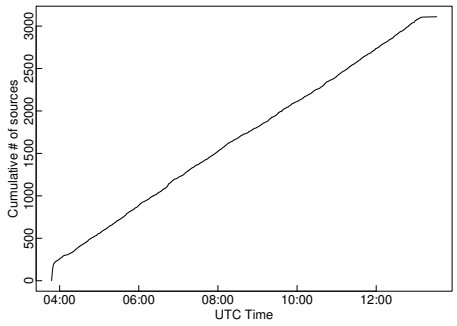

Figure 6: The arrival process of the event on TCP port 1433 on 2006-01-22 (from 2006-01-22 21:00 to 2006-01-23 07:00), which corresponding to a MS SQL Server vulnerability.

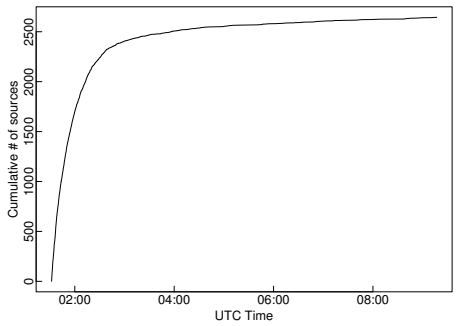

Figure 7: The arrival process of the event on TCP port 1433 on 2006-08-24, which corresponding to a MS SQL Server vulnerability.

the true arrival time and departure time of the bots, since the botnet might scan a large range and the honeynet sensor can only observe a small sample of the scans. Instead, we defined the time of the first scan we seen from a given bot as its observed arrival time, and the time of the last scan we seen from the same bot as its observed departure time. For random scanning, we can assume the scans we observed are a random sampling from the total scan population. Certainly the sampling errors will influence the results. The number of scan between the first scan sent out by a bot and the first scan we observed from that given bot follows a geometry distribution. If we assume the scan speed is close to constant, the time difference of the first scan sent out by a bot and the first scan we observed from that bot will also follow a geometry distribution. We can make the similar argument to the true departure time of the bot and the departure time we observed. For the long lived events usually we can use the observed arrival and departure time as good approximation of the true arrival and departure time. For the short lived event the observed arrival and departure time might not be able to present true arrival and departure time.

For the long lived events, we found there are two types of source arrival processes. In some events, most bots arrived at the beginning part of the events, but on some other events bots arrivals distributed over the whole period of the event duration. Figure 6 and Figure 7 showed such two representative cases respectively.

In Figure 6, most bots arrived at the beginning part of the events. This might correspond to the case that after the botmaster typed the scan command in the command and control channel, immediately the bots in the channel received the scan command and began to scan. The true source arrival times of bots are same, so the observed source arrival time follows a geometry distribution.

In Figure 7, the bot arrive uniformly in the event duration, which indicate the true source arrival time of different bots are different and also should be uniformly distributed in time. There are two possibilities to make this happen. One possibility is that every bot defer to execute the scan command by random seconds uniformly. The other possibility is that the scan command is the default channel topic [1]. Therefore, after a bot join the channel, it will get the scan command and start scanning. From the data we cannot separate these two cases.

In the departure process, we found, in all the long-lived events, many bots depart before the events end.

For the events most bots arrived at the beginning part of the events, we observed at the end of event, the bot departure rate increased sharply. We analyzed several botnet source code genres and found in most case the botmaster asks the bot to scan a fixed amount of time. If that is the case, it makes sense that at the end of the time specified by the botmaster all the remaining bots end the scanning.

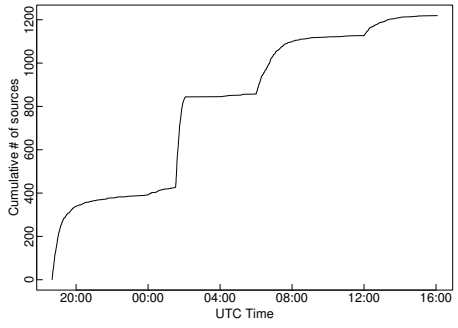

Figure 8: The bot arrival process of event on TCP port 139 from 2006-08-24 13:40 to 2006-08-25 11:04, which corresponding to a Netbios-SSN scan.

There is one event different from other events, in which the bots arrived in groups, but the total scan arrivals are still linear in time. In Figure 8 we can see there are four major groups of bots arrived in batch. But in Figure 9 the number of scan arrivals is still linear in time. Through further analysis we found, after the first group of bots departed, the

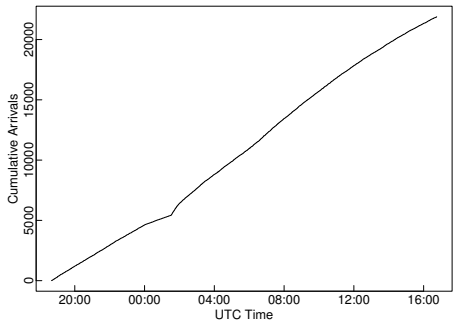

Figure 9: The scan arrival of event on TCP port 139 from 2006-08-24 13:40 to 2006-08-25 11:04, which corresponding to a Netbios-SSN scan.

second group of bots arrived immediately. This is also true for other consecutive groups of bots. Obviously, the botmaster intentionally divide the bots in four groups to do the scanning one after another.

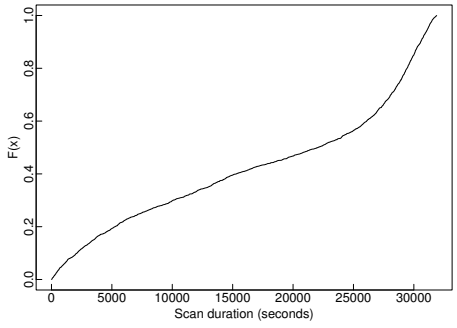

Figure 10: The CDF of observed scan duration of bots of event on TCP port 1433 on 2006-08-24, which corresponding to a MS SQL Server vulnerability.

We also studied the observed bot scan duration, *i.e.*, the time between the first scan observed from a given bot and the last scan observed from the given bot. An example CDF of the scan duration is shown in Figure 10. However, we found the scan duration varies from events to events. There is no clear pattern can be found.

1.5 Observed Local Scan Rate

We calculated the local scan rate of a given bot as the number of scans we observed minus one over its observed scan duration. The idea behind is that we can think after the first scan arrives we started the timer, and in the observed scan duration we will observed the scans except the first one. We will not define the local scan rate for the senders from which only one scan is observed.

We first looked at the CDF of local scan rate of different events. In four cases, the numbers of bots which send more than one scans are very small, so the CDF is not very repre-

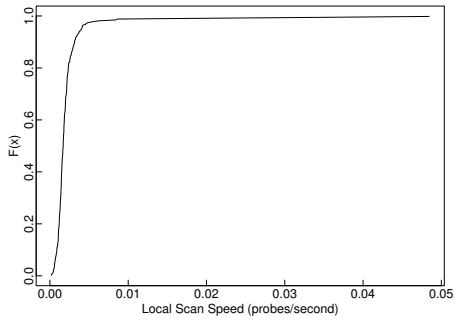

Figure 11: The CDF of local scan rate distribution of the event on TCP port 5900 on 2006-09-26, which corresponding to a VNC vulnerabilty.

sentative. For the remaining cases, we found most bots have similar local scan rate with a few bot with very high local scan rate. We further analyzed the bots with very high local scan rate, and find they are not necessarily the bots which send most scans. Many of such cases are due to they have very short observed scan duration. Figure 11 shows an example of such a CDF distribution.

We further investigated whether the local scan speed have any correlation with the bot arrival and departure time. We did not find any obvious trend. We believe in most case, the bot arrival and departure time might not have strong correlation with their local scan speed. However they might have certain weak correlation and which can be buried into the random noises in the data. Figure 12 and Figure 13 show an example of this analysis.

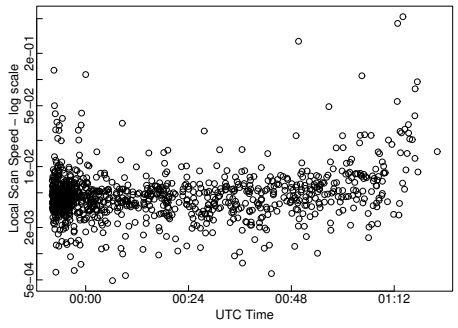

Figure 12: The scatter plot of the source observed arrival times and their corresponding observed scan rate of the event on TCP port 1025 on 2006-09-19.

1.6 Scan Source Destination Relationship

We also analyzed source destination relationships. We mainly studied two distributions: how many sources target a destination address in the honeynet sensor, and how many destinations are contacted by a source.

We found in all the events, the distribution of how many

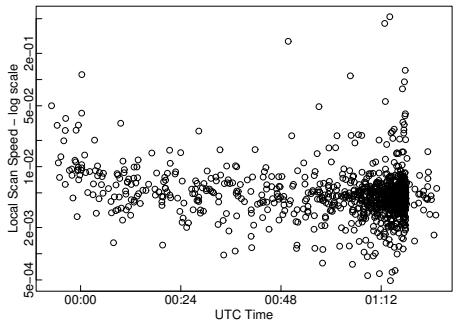

Figure 13: The scatter plot of the source observed departure times and their corresponding observed scan rates of the event on TCP port 1025 on 2006-09-19.

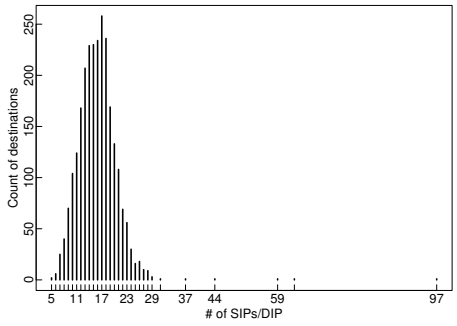

Figure 14: The distribution of number of sources a destination contacted of the event on TCP port 1433 on 2006-08-24, which corresponding to a MS SQL Server vulnerability.

sources a destination contacts is close to the binomial distribution with only very few exceptions. This implies that the source usually choose the destination uniform randomly. Figure 14 is such an example.

The distribution of how many destinations a source targets is more complex. Sometimes it has multiple modes. The conjecture is that it can be explained as a multiplex of multiple binomial distributions, due to different bots might have different scan speeds and durations. In Figure 15 we showed an example which clearly has this pattern.

2. REFERENCES

[1] RAJAB, M. A., ZARFOSS, J., MONROSE, F., AND TERZIS, A. A multifaceted approach to understanding the botnet phenomenon. In *Proc. of ACM/USENIX IMC* (2006).

[2] RAMACHANDRAN, A., AND FEAMSTER, N. Understanding the network-level behavior of spammers. In *Proceedings of ACM SIGCOMM '06* (September 2006).

[3] ZALEWSKI, M. the new p0f. `http://lcamtuf.coredump.cx/p0f.shtml`.

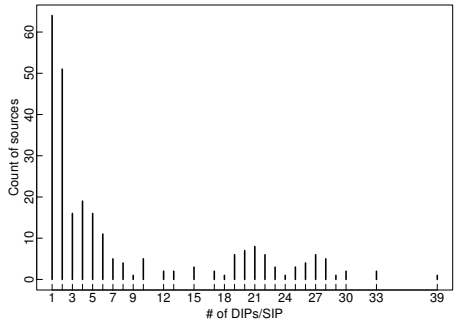

Figure 15: The distribution of number of destinations a source touched of the event on TCP port 2967 on 2006-11-27, which corresponding to a backdoor shell.