Botnet Detection Method Based on Survival Analysis

Jiajia Wang¹,a and Yu Chen¹,b
¹Taizhou Polytechnic College, Jiangsu Taizhou, China, 225300
aWangjiajia_99@163.com bxiaoyueryuchen@163.com

Keywords: botnet, survival analysis, life table, Mantel-COX

Abstract. Botnets communicate with legal protocols, it is difficult to directly capture botnet packets in the context of large traffic, high-speed network. Therefore, this paper analyzed the data stream to judge the botnets’ activities. First of all, we set mirror port in the network core equipment to view the network data flow information, through the capture tool we could get and format data packet relevant information. Secondly, for the data flow of different unit time, different survival rates could be obtained by life table analysis. Mantel-COX analysis was then used to test the survival rates. Because botnets must inform all bots before launching attack, we could detect the presence of abnormal condition before a cyber attack taken place.

1. Introduction
Botnet has become one of the most serious threats to network security in our country. In May 2017, a new IoT botnet called http81 appeared in China. Data showed that http81 botnet had infected more than 50 thousand network cameras in China. How to control the trend of the rapid expansion of botnets, detect the existence of botnets and eliminate their adverse effects as soon as possible, is the focus of the current research work.

2. Related Work
Currently, there are three main ways to detect botnets.

1) Detection method based on honeypot. Botnet detection method based on honeypot[1,2], able to detect unknown attacks, however, it can only capture the packets interacted with honeypot directly, can not detect attacks against other systems, and easily to be exploited by hackers.

2) Detection method based on features. Botnet detection method based on features[3,4], according to the characteristics of communication data, the attack data stream can be quickly detected, but it is helpless to the attack with unknown features. Because the communication data stream of botnet is scarce, how to extract features is also a difficult problem.

3) Detection method based on abnormal behavior. Botnet detection method based on abnormal behavior[5,6], whether online or off-line, anomaly detection has higher accuracy, but it is easy to be restricted by training samples.

Botnets have diversified forms, but in order to maintain the scale, each type of botnet has periodic communications between the bots and the servers. This paper presented a method which collected network data streams and then performed botnet survivability analysis. First we used capture tools to collect and analyze net flow, then used the life table analysis method to describe the survival rate, finally selected the Mantel-Cox to conduct multivariate analysis.

3. Test Method

3.1 Survival analysis
Survival analysis is originally used in medical and biological research to measure individual survival, and this analysis can also be used for the study of network traffic lifetime. For example, two hosts
communicate, the interval between the start and end times of a packet is the lifetime of the current data. Survival rate, also known as survival probability or survival function, can be expressed as follows: 

$$S(t) = P(T > t)$$

$t$ stands for arbitrary survival time, and $S(t)$ represents the survival probability of $T$ moments, $0 \leq S(t) \leq 1$. The steeper the slope of $S(t)$ decline, the lower the survival rate.

Network traffic has the characteristic of burstiness, it is difficult to obey the Gaussian Distribution, the statistical methods used is not suitable for net flow, survival analysis has no such requirements on data, so this paper used the method of survival analysis to detect network traffic.

3.2 Censored data

Censored data can be divided into left truncated and right truncated. When the statistical unit survival time has arrived and the data is still transmitting, the right truncation is needed. If the data is already transmitted before the statistical unit survival time begins, and then the left truncation is needed. When censored data is generated, the current network traffic has been the subject of statistical research. Most of the current detection methods are based on the right truncated data, as shown in Figure 1.

![Figure 1. Sketch map of right censored data.](image1)

If the data under study are still transmitte at the end of unit time, it need to be censored. The truncated index is represented by $\delta$, 

$$\delta = \begin{cases} 
0 & \text{The current object has been truncated} \\
1 & \text{The current object has not been truncated} 
\end{cases}$$

3.3 Life table analysis

The life table method is a nonparametric survival analysis, that is, the survival rate of the network data stream is not need be studied beforehand, and training error of the sample is avoided. When the number of samples is large and the type of distribution is uncertain, life table analysis can be used. The survival rate can be estimated by the correction of censored data [7].

The life table method calculates steps are as following:

1)The data collected from the network core device is divided into four tuples by attributes{ source address, destination address, packet delivery duration, censored or not }

2) Select the observation time and count $n_i$, $n_i$ are the number of packets sented in the observation time, if data is truncated in the observation time range, count $w_i$, $w_i$ are the number of censored packets and $d_i$, $d_i$ are the number of data packets that have not been successfully transmitted.

3) The actual number of packets observed during the observation period is calculated $n_i = n_i - w_i/2$

4) Conditional survival rate is calculated $\hat{S}(t_i) = \frac{n_i - d_i}{n_i}$, cumulative survival rate

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \times \hat{S}(t_i/t_{i-1}).$$

3.4 Mantel-Cox analysis

After calculating survival rate with life table analysis, there are many methods to compare the survival rate, including Logrank, generalized Wilcoxon and Mantel-Cox. The most widely used method is the
Mantel-Cox\cite{8,9}, this method has little effect on the use of censored data, it can be used for the comparison of two or more survival rates. In this paper, the Mantel-Cox method is used for the next step. Assuming a total of two sets of data, the chi square test is used to represent 
\[ \chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \]
\( O_1 \) is the sum of the events in the first set of data, \( E_1 \) is the expectation of the first set of data, \( O_2 \) is the sum of the events in the second set of data, and \( E_2 \) is the expectation of the second sets of data. If there is \( n(n\geq3) \) group data, the above formula can be extended to

4. Experiment and Result

4.1 Experimental environment

In this paper, virtual machines are used to simulate the actual network environment, and the specific topology was shown in Figure 2. The experiment employed 2 C & C servers, 3 PCs, 1 switches and 1 test hosts. 2 C&C servers had installed Black Energy, the switch had set up a mirror port and connected to the detection host, detection host had installed capture tool.

4.2 Data acquisition

The experimental data used in this paper were derived from analog switch mirroring ports, including data that was normally communicated between hosts, communicated between zombie hosts and communicated between bots and servers. After pruning in the four tuple format, we could proceed to the next step. Some of the raw data was shown in Figure 3. Some data formatted by four tuples was shown in Table 1.

<table>
<thead>
<tr>
<th>Source Address</th>
<th>Dest Address</th>
<th>Summary</th>
<th>Loss</th>
<th>Fail Time</th>
<th>Delta Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.11.138.18</td>
<td>61.155.222.122</td>
<td>TCF: D=443 S=49927 FTR ACK=3145625675 SEQ=127</td>
<td>60</td>
<td>0:00:00 0:00</td>
<td>0:00:00 0:00</td>
</tr>
<tr>
<td>10.11.138.18</td>
<td>61.155.221.217</td>
<td>TCP: D=443 S=49920 FTR ACK=4793788784 SEQ=411</td>
<td>60</td>
<td>0:00:00 0:00</td>
<td>0:00:00 0:00</td>
</tr>
<tr>
<td>10.11.138.18</td>
<td>61.155.221.217</td>
<td>TCP: D=443 S=49920 FTR ACK=4793788784 SEQ=411</td>
<td>60</td>
<td>0:00:00 0:00</td>
<td>0:00:00 0:00</td>
</tr>
<tr>
<td>10.11.138.18</td>
<td>61.155.221.217</td>
<td>TCP: D=443 S=49920 FTR ACK=4793788784 SEQ=411</td>
<td>60</td>
<td>0:00:00 0:00</td>
<td>0:00:00 0:00</td>
</tr>
<tr>
<td>10.11.138.18</td>
<td>61.155.221.217</td>
<td>TCP: D=443 S=49920 FTR ACK=4793788784 SEQ=411</td>
<td>60</td>
<td>0:00:00 0:00</td>
<td>0:00:00 0:00</td>
</tr>
</tbody>
</table>

Figure 3. Partial raw data.

Table 1. Part of experimental data.

<table>
<thead>
<tr>
<th>NO</th>
<th>IP(_{source})</th>
<th>IP(_{dest})</th>
<th>Duration time(ms)</th>
<th>( \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.11.138.18</td>
<td>61.155.222.122</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>10.11.138.18</td>
<td>61.155.221.217</td>
<td>272</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>10.11.138.18</td>
<td>58.218.203.232</td>
<td>151</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>10.11.138.18</td>
<td>61.155.221.219</td>
<td>141</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>10.11.138.18</td>
<td>61.147.223.132</td>
<td>161</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>10.11.138.18</td>
<td>61.147.223.132</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>10.11.138.18</td>
<td>61.155.221.232</td>
<td>197</td>
<td>1</td>
</tr>
</tbody>
</table>

In the experimental data, the source and destination addresses were the same with the original data. The packet delivery duration was determined by the departure and start time of the packet sent from the original acquisition data. The last column represented truncated attribute, if the attribute value was 1, this indicated that the packet delivery had ended at the end of the statistic time, and if the attribute value was 0, indicated that the packet was truncated. As we can see, all packets in Table 1 had been sent out before the end of the unit time.

First, the number of packets captured over time in the experiment was calculated, as shown in Figure 4. Figure 4 showed that the number of packets captured varies little over time, and the botnet did not become awareed only from the number of packets.
4.3 Experimental result
Since Mantel-Cox requires multiple groups of data to be analyzed, this paper used normal net flow to compare with experimental data with the same time interval but not at the same time.

Now, the life table analysis method and the Mantel-Cox method were used to detect the captured data, as shown in Figure 5.

![Figure 4: Captured packets over time.](image1)
![Figure 5: Detection result.](image2)

Red line was the normal data information, black line was the data collected in this experiment, the collection time of the two sets of information was not related. As seen from Figure 6, under normal circumstances, the packet survival rate was relatively low, in unit time of the experiment, the survival rate was only 0.017903 on average, which indicated that the independence between data packets was good, connection between any two hosts within a reasonable time. But to the data collected in this experiment, with the emergence of botnet, the survival rate would change greatly, the average value was about 0.161935, the survival rate difference between the two cases reached 0.14, which showed that there were significant differences between the two communication data stream, the experimental data showed that some hosts were related more closely and lasted longer. According to the normal data stream, we set the survival rate threshold 0.05 in unit time, if less than the threshold, the contact between different hosts in the network was normal, if exceeded the threshold, we should have to pay attention to the hosts in the network because of close connection. Therefore, this method can be used to detect botnet activity quickly.

5. Summarization and Prospect
This paper proposed a method of botnet detection based on survival analysis. In this method, data packets are collected and formatted, botnet could be detected only by 4 attributes sent by the packet, didn’t need to extract the features of packets, avoid capture phase error. Secondly, we used the nonparameter life table analysis to analyze the survival rate of data, this method did not need to take the data as a fixed model, more in line with the characteristics of the network, without parameter, the training error of sample was avoided. Then the Mantel-Cox method was used to detect the survival rate, which could detect the botnets’ activities more accurately and quickly. In the future work, we should distinguish the normal data and the abnormal data as soon as possible, and filter the abnormal data stream. How to identify and filter the botnet data more quickly is the focus of the work in the future.

References


