Node-based Sampling P2P Bot Detection

Chunyong Yin1,2*, Ruxia Sun1, Lei Yang1, Darius Iko1

1School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, P.R.China
2Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science & Technology, Nanjing 210044, P.R.China
e-mail: yinchunyong@hotmail.com

Abstract

The concept of using node-based sampling for the treatment of packet capture mechanism based on Libpcap of network-based detecting Peer-to-Peer botnet process was tested, and its effect on the time window of feature extracting and sampling time interval was explored. Node-based sampling treatment resulted in significant increase in the detection performance due to node profile of the novel behaviors to the detected computer in Peer-to-Peer bot detection, and the degradation of false positive. At relatively right time window (e.g., about 180s), precision was completely maximized, while the false positive decreased by 10% to 15%. The detection rate can be significantly increased due to the false positive degradation. A new performance index called Comprehensive Evaluation Index is proposed for more clearly represent the effectiveness. Sampling can reduce more than 60% input raw packet traces and achieve a high detection rate (about 99%) and a low false positive rates (0-2%).

Keywords: Node-based, detection, network behavior, peer-to-peer, botnet

1. Introduction

The protection concept of detecting potential threat for the large scale of malicious software, sometimes referred to as the botnet, would be of strategic significance since such threats are serious and threatening. Botnets, networks of malware-infected machines (bots) controlled by botmaster, usually carry out their nefarious tasks, such as sending spam, launching denial of service attacks, and even stealing personal data [1], [2].

In this process, how to detect botnets and remove them has become a basic problem during researching botnets. Botnets also has a variety of type, including P2P botnets, IRC botnets, and HTTP botnets etc [3-7]. Especially, P2P bot which is distributed and uncentralized, if detected in a reasonable time, will be able to improve network security greatly [8].

In the literature, the approach of bot detection using signature technique has been widely addressed, [9-15] and such an approach has been found to be effective to find known bots as Phatbot etc. C. Kolbisch et al. [16] proposed a signature-based malware detection system which used special graphs generated to determine the bot. The method needs to be trained before being used and the detection rate is only 64%, although it is possible to detect various kinds of bot. Signature-based method is not capable enough to detect unknown bots even a variant of known bots. Therefore, with the increasing number of new bots/bot variants, the detection rate may decrease very quickly.

In the literature, flow-based technique for bot detection can increase the detection rate, and the mechanism was proposed to represent more general bot behavior than the signature technique [17-19]. Generally speaking, the relevant available research reports on bot detection have been focused on the flow-based technique. C. Livadas et al. [20] also developed a system to detect C&C traffic of botnets based on flow. Similar to that of reported earlier, the system contains two stages: one is extracting several per-flow traffic attributes including flow duration, maximum initial congestion window, and average byte counts per packet; another is using a Bayesian network classifier to train and detect bot. However, the false positive rate is still high (~15.04%). H. Choi et al. [21] proposed a botnet detection mechanism solely based on monitoring of DNS traffic in the connection stage of bot, meanwhile the botnet can easily evade this algorithm when the botnet rarely used DNS at initializing and had never used it again since then. B. Wang et al. [22] reported a detection approach of P2P botnet by observing the stability of control flows in initial time intervals of 10 minutes. This differs from the usage of the protocol
by a normal user which may fluctuate greatly with user behavior. J. Kang and Y. Song [23] proposed a novel real-time detecting model named Multi-Stream Fused Model, in which they deal with different types of packets in different methods, while this model could not reach a desirable detecting precision when operated in a large-scale network environment and it could also generate extra harm to the Internet. D. Liu et al. [24] presented a general P2P botnets detection model based on macroscopic feature of the network streams by utilizing cluster technique. However, the proposed method was unreliable or non-functional if only a single infected machine is present on the network. Until now, there has been no report regarding the application of the node-based concept to bot detection, a more effective and high-efficiency method in finding bot.

This study applied the concept of node-based to bot detection, which has higher level than flow and packets. It was assumed that, treating bot detection with node-based might result in better detection performance than that with flow-based, meanwhile the node-based has broader adaptability than the flow-based which may be sensitive to new behaviors from bots implementing highly varied protocols. Based on such an assumption, the effect on the time window of feature extracting and sampling time interval was explored.

2. Experimental Section
2.1. Experimental Dataset
Computing intelligence algorithm was utilized from Weka which is learning framework tool. The experimental dataset produced from a combination of the French chapter of the honeynet project, the Traffic Lab at Ericsson Research in Hungary, and the Lawrence Berkeley National Lab [25] was collected and merged in virtual environment. The merging process was conducted by using the Wireshark technology for evaluation, and the feature matrix was produced via extracting at different time window for 1 day. To get the key discriminating attributes in real traffic, the node-based attribute was first thoroughly analyzed using general bot behavior which was mentioned in the literature before and then generated by correlation based attribute evaluator. By utilizing node behavior, some key attributes with relatively important feature can be obtained from the traffic dataset.

2.2. Bot Detection Treatment
In one set of experiments, the node-based method was implemented on bot detection in the datasets mentioned above and node feature is defined as 6 attributes (Number of protocols used for time interval, Number of flows used for time interval, Number of packets sent for time interval, Ratio of number of packets sent to number of packets received for time interval, Average length of packets sent, and Ratio of average sending packets length to average receiving packets length for time interval) and time window is 10s, 60s, 180s for sampling respectively. In another set of experiments, to conduct the bot detection treatment in the principle of flow, four attributes (Variance of payload packet length for time interval, Number of packets exchanged for time interval, The size of the first packet in the flow and flows per address / total flows) are generated by correlation algorithm based attribute evaluator with higher discriminatory power and time window is 300s. During this process, the detection was achieved by decision tree classifier only, and no extra classifier (e.g., neural network or Bayesian network) was applied. The time windows reported throughout this paper were all based on the unit in second.

2.3. Sampling Measurement
Generally, when carrying out flow-based detection or node-based detection, each of packets will be processed one by one. However, that situation is unsuitable for real time detection in high-speed network and produces high packet loss rate. Sampling measurement is introduced to decrease the number of packet processing while keeping the higher detection rate.

The effect of sampling on bot detection is shown in Figure 1. Two kinds of different detection way, i.e., normal detection and sampling detection, were analyzed in Figure 1, respectively. It can be seen from Figure 1 that when time was at a certain moment, for example, around t1, the normal detection way can possibly detect more bots than sampling detection, for example, 2 bots for normal detection and 1 bot for sampling detection at time t1; however, with
the increasing amount of time, the number of finding bots will have similar effect on these two methods, for example, 2 bots for normal detection and 2 bots for sampling detection at time t2. The asymptotic same result on detection at certain moment is due to the cycle limit of the found bots from the real world.

![Figure 1. Sampling detection model](image)

2.4. Proposed Evaluation Indexes on P2P Bot Detection

Generally speaking, the relevant available research reports on detection performance has been focused on the detection rate (i.e., accuracy), although precision is also available. However, there have been no reports regarding the correlation application of both detection rate (DR) and precision, a critical comprehensive evaluation index. Due to the two-class nature of the detection (i.e., bot detection and normal detection), there are usually four evaluation indexes, i.e., True positive (TP), False positive (FP), True Negative (TN), and False Negative (FN). Accordingly, there are three kinds of rates as follow:

\[
Detection\ rate = \frac{TP}{TP + FN}\quad (1)
\]

\[
FP\ rate = \frac{FP}{FP + TN}\quad (2)
\]

\[
Precision = \frac{TP}{TP + FP}\quad (3)
\]

However, it can be seen from the following proof that the trend of FPR can be reached by Precision.

\[
\text{precision} \to 1 \iff \frac{TP}{TP + FP} \to 1 \\
\Rightarrow FP \to 0 \\
\Rightarrow \frac{FP}{FP + TN} \to 0 \iff FPR = 0
\]

Also, the detection rate and the precision both have the equivalent importance in the detection system. So, the following evaluation index proposed - which is called Comprehensive Evaluation Index (CEI) - may have a strategic significance for the evaluation of detection performance.

\[
CEI = DR \times 50\% + Precision \times 50\%
\quad (4)
\]
3. Results and Discussion

3.1. Effect of Detection Performance on Period of 10 Seconds

The detection rate increased very slowly with increasing the amount of training data and reach the higher value 0.667 for 50, while precision reach 0.545. However, with the further increasing amount of training data, both the detection rate and precision decreased very quickly and reach 0.333 and 0.333, respectively. When the amount of training data is 10, both detection rate and precision are momentarily higher, which can be due to the fact that training data is close to bot behavior. Furthermore, the less data may also result in the increased detection rate and precision.

3.2. Effect of Detection Performance on Period of 60 Seconds

For node-based detection on a period of 60 seconds, the maximum detection performance has been achieved, i.e., detection rate is for 1 and precision is for 1. When the amount of training data is 10, detection rate has been the maximum, while the detection rate decreased very quickly with increasing the amount of training data and reached the lower value 0.306 at 50. When the amount of training data is 10, both detection rate and precision are momentarily higher, which can also be due to the fact that training data is close to bot behavior. Furthermore, the less data may also result in the increased detection rate and precision. But at 60, the maximum detection performance is reached again, presumably due to the period of 60 seconds is similar to real bot cycle.

3.3. Effect of Detection Performance on Period of 180 Seconds

With increasing the amount of training data, detection rate always kept the maximum 1 and false positive decreased to the minimum 0, while precision reached the maximum 1 too. These results showed that the right time window was obtained for bot detection.

3.4. Effect of Sampling Detection on Bot Detection

When time window was 10s, 60s, and 180s, the sampling detection experiment was carried out with time interval of 10s, 30s, 60, and 180s, respectively. Our results are shown in Table 1, 2, 3. It can be seen that when time window was 180s, increasing time interval did not make any effect to detection performance.

| Evaluation indexes | Time interval (s) | | Table 3. Effect of time intervals on time window for 180s |
|--------------------|------------------|------------------|
| Detection rate     | 0                | 10               | 1                | 1                | 1                | 1                | 1                |
| False positive     | 0                | 0                | 0                | 0                | 0                | 0                | 0                |
| Precision          | 1                | 0                | 1                | 0                | 1                | 1                | 1                |

3.5. Effect of Time Window Addition on Detection

The effect of time window addition on the CEI of P2P bot detection algorithm is shown in Figure 2. Various amounts of time window were set for detection algorithm at same dataset before the CEI measurements. It can be seen from Figure 2 that when time window was below a certain level, for example, around 180s(4*45s), the CEI increased very quickly with the increasing amount of time; a further increase in the time charge had only a small effect on the CEI.

3.6. P2P Botnet Detection in the Principle of Flow

The above results showed that significant increase in the detection performance of the P2P botnet detection occurred during the time window treatment, mostly due to be close to real bot cycle time. For verifying performance of node-based in the P2P botnet detection, it would be desirable to utilize another detection method. For this purpose, the authors designed another set of experiments so that the bot detection was carried out in the principle of flow by extracting flow attributes. Interestingly, it was found that during the flow-based bot detection treatment...
process using time window with 300s, detection performance cannot effectively increase (see Figure 3). Also, extracting flow attributes did not significantly change the false positive. It is noted that the node-based detection treatment resulted in a significant increase in detection rate and precision, which may be explained by 1) node-based method may not be sensitive to new behaviors from bots implementing highly varied protocols; 2) attribute selection. On the other hand, the flow-based bot detection need longer time window than the flow-node bot detection for similar detection performance (Figure 3) presumably due to the fact the compatibility of attribute generated from the flow is poor. Furthermore, Figure 3 showed that the flow attribute extracting did not significantly improve the performance of detection rate and false positive of bot detection in the long time window.

Figure 2. The CEI of P2P bot detection algorithm as a function of time window (measure time by the second)

Figure 3. Detection rate VERSUS false positive for both detection with time window (time window size is 300s for flow-based, and 180s for node-based)

The detection performance of flow-based method can be effectively improved when the attribute treatment is carried out in proper measurement. However, based on the compatibility of detection at new bots, it was decided that the node-based was used. Based on the results shown in Figure 9 and Figure 11, it can be concluded that under the conditions studied, the time window treatment in node-based algorithm would significantly increase the detection performance of bots. The detection of system as a result of the introduction of node-based method can also possibly contribute to the detection of unknown bots using different protocols. Further work on more comprehensive understanding of P2Pbot detection process, e.g., group-based, node profiling, and attribute processing, may still be needed in the future.

4. Conclusion

The treatment of packet capture mechanism based on Libpcap of network-based detecting P2P botnet process with using node-based sampling resulted in increased detection rate due to node profile of the novel behaviors as well as the degradation of the amount of traffic processed under the sampling condition. When time window was relatively right, for example, about 180s, the detection rate was more than 90%. In the sampling process, the false positive and the amount of traffic processed can be decreased by 30% and 50-60%, respectively. Precision could be significantly increased at right time window due to novel node profile; however, when performing flow based detection was effectively decreased.

Acknowledgement

This project was funded by Jiangsu Key Laboratory of Meteorological Observation and Information Processing (KDX1105) to Dr. Chunyong Yin. The authors wish to thank the
reviewers for their valuable comments and suggestions which helped in improving the quality of the paper.

References