Conversation-based p2p botnet detection

Shaojun Zhang, Dr.Ali Ghorbani
Computer Science, University of New Brunswick

Faculty of COMPUTER SCIENCE
UNB
ISCX Information Security Centre of Excellence

Introduction

Currently, Botnet(robot network) already becomes one of the most dangerous threat to Internet security. Botnet is a collection of compromised computers called zombie or bot. Zombies are controlled by malicious machines called botmaster through C&C channel. Botnet can be used for a plenty of malicious behaviors including DDOS, Spam, steal sensitive information and etc., which could be very serious threats to the Internet. In traditional centralized botnet, there is a point of failure which makes the botnet not that robust. For example, servers are used in IRC-based botnet which is the first type of botnet and HTTP-based botnet which can be destroyed by finding out the C&C servers. To overcome the weakness, attackers began to use peer-to-peer networks for C&C communication which makes botnets more robust. And every peer in the p2p botnet can act as both client and server. As a result, detector cannot find a central point of failure in p2p botnets.

We propose a p2p botnet detection method based on conversation in a time window. This is the first time to use conversation-based features to detect p2p botnet. The features of conversation can differentiate p2p botnet conversations from normal conversations. Decision Tree and SVM are applied to the features to classify the normal conversations and the p2p botnet conversations.

Background

Varity of botnets are widely used in today's network for various purposes. In terms of the C&C communication they used, Cooke et al. classified botnets into three possible categories, namely centralized, P2P and random. Centralized and random topologies' weaknesses can be made use of by detector are pointed out by the authors. On the other hand, p2p botnet has the most complex design and lowest detectability. Researchers proposed a handful of research concentrated on botnet detection so far and there are quite a lot techniques for botnet detection. These detection methods fall into 2 categories.

Host-based: It is the most straightforward detection method. It treats the bot binaries as a virus, Trojan or some other malicious malwares and detects it the way that Anti-virus software.

Network-based: This kind of approach concentrates on the network traffic to find signatures of the content or behavior patterns of p2p botnet. Network-based detection based on signature is widely used to detect IRC-based botnet and it has high detection rate with low false positive rate. One limitation is that signature-based detection approach can only detect known botnet whose content is not encrypted. Nowadays, most of the research is based on network behavior because of the encryption. In behavior-based detection approach, people derived features from flows or packets. These features can reveal the difference between p2p botnet traffic and non-malicious traffic.

Motivation

Currently, p2p botnet has advantages over traditional botnet and it becomes the most serious malware in the wild. There are several difficulties in detecting p2p botnet.

- It is decentralized. There is no weak-point. And even some peers are down, it still can work well.
- The behavior of peers in the p2p botnet is similar with the normal p2p application like bittorrent or gnutella.
- To make the p2p botnet more robust, the authors of p2p botnet encrypt the C&C communication. So it may be useless to look at the content of the traffic.

To data, most network-based approaches depend on flow or packet. And there are some shortcomings like losing information about the traffic, consuming more memory and etc. So the motivation of this research is to overcome these drawbacks. From a higher level of view, a conversation-based detection approach is proposed. And this is the first time to detect p2p botnet based on conversation (as shown in Figure.1.).

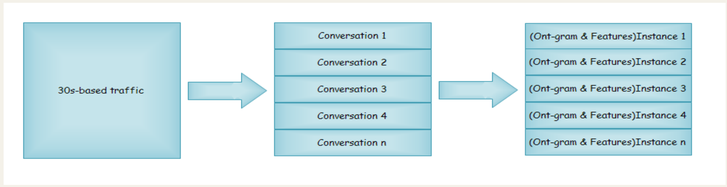

Fig. 1. Time-based Conversation

Methodology

Fig. 2 describes the proposed detection framework in a general way, including 4 main parts. .

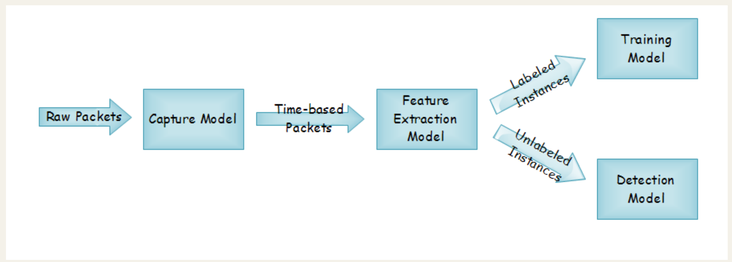

Fig. 2 Detection Framework

- Capture Model: Capture the traffic from the network.
- Feature Extraction Model: Generate conversations and extract features from every conversation.
- Training Model: Use labeled instances for training.
- Detection Model: Use the classifier generated by training model to classified unlabeled instances.

Result

We have 6 datasets and all the instance are labeled. For training the labels are used in J48 and SVM, because they are supervised learning techniques. For testing the labels are used to evaluate the performance of these method. Dataset 1,3,5 contains instances from normal and storm traffic, while dataset 2,4,6 contains instances from normal and waledac traffic. These datasets are used in the 6 experiments in table 1.

	Train Set	Test Set
1	Dataset 1	Dataset 2
2	Dataset 2	Dataset 1
3	Dataset 3	Dataset 4
4	Dataset 4	Dataset 3
5	Dataset 5	Dataset 6
6	Dataset 6	Dataset 5

Table 1 Experiments

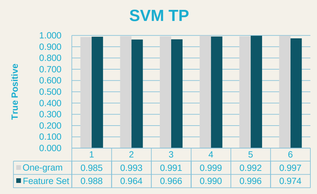

Fig. 3. SVM True Positive

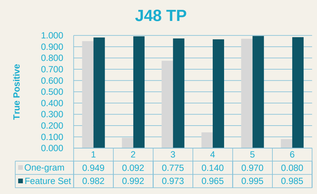

Fig. 4. J48 True Positive

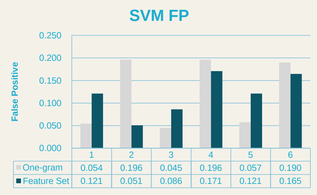

Fig. 5. SVM False Positive

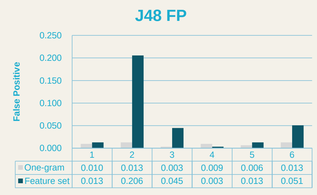

Fig. 6. J48 False Positive

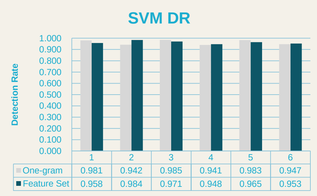

Fig. 7. SVM Detection Rate

Fig. 8. J48 Detection Rate

The results show that both one-gram and feature set give good results by applying SMV algorithm. And for J48, only feature set has a good result.

Conclusion

We proposed a p2p botnet detection approach based on the features which are derived from network conversations. By applying the powerful machine learning techniques. the results show its ability to detect p2p botnet with high true positive rate and relative low false positive rate. Further more, as our conversations are based on time window. We can apply this method for online p2p botnet detection.

Acknowledgements

I would like to thank Dr. Ali Ghorbani for his supervision throughout my researsh and I am also thankful to the students in our lab for their help and suggestions.