Closed-loop Feedback Trojan Detection Technique Based on Hierarchical Model

Wu,Jinlong[1,a *], Gu,Haidong[2,b] and Xie,Yixin[3,c]

[1] Jiangnan Institution of computing technology, Wuxi, Jiangsu Province, China

[2] Jiangnan Institution of computing technology, Wuxi, Jiangsu Province, China

[3] Jiangnan Institution of computing technology, Wuxi, Jiangsu Province, China

[a]wjinlong@mail.ustc.edu.cn, [b]joysmith852@gmail.com, [c]mkroy892@gmail.com

Keywords: Trojan Detection; hierarchical model; closed-loop feedback

Abstract. In recent years, as the representative of APT attack incidents continues to increase, which steal the confidential information. It's a serious threat to network security and user privacy. The key to prevent such attacks is how to detect Trojan behavior from the network traffic of APT attacks. In this paper, we propose a novel approach based closed-loop feedback model and hierarchical detection model. The result we present improved the detection rate and reduced false positives in detecting Trojan behavior from network traffic. Thus, we can protect user privacy effectively and maintain secure of network environment relatively.

Introduction

Trojan control is the most important aspect in implementation of APT (Advanced Persistent Threat) attacks which are highly targeted, small attack range and difficult to obtain samples in a timely manner [1]. According to a well-known anti-APT company's (FireEye) investigation concluded that more than 95% of the corporate network hosts suffered Trojan invasion, and the Trojan sample detection rate of less than 10% [2]. So, the traditional detecting approach is difficult to play an effective role.

This paper aims to propose an intelligent Trojan detection approach based on closed-loop feedback model. So that our approach is not only valid to learn unsupervised, but also remove a part of false alarms and reduce the false alarm rate. Meanwhile, we improved the Trojan detection rates based on the hierarchical model.

Proposed Detection Methodology

There must be some of the fixed features during the network traffic of communication process of Trojan to achieve its full functionality, whatever Trojans hide themselves and avoid killing by antivirus software. These features contains upload and download traffic feature [1], the keep-alive feature called "heartbeat" [3], the connection time of communications feature [4] and so on. All of which can be extracted from the network traffic of Trojans. Therefore, to detect Trojan of APT attack from the network traffic is effectively. We use machine learning algorithms to train the detection model. In order to reduce the false positive rate and achieve unsupervised learning, the detected results will be as the input of the data mining module to mine the false positive feature. The false positive feature corresponding data will be fed back to the input as a part of training set to adjust training model and form self-feedback closed loop.

Detection Model Design

Trojan Communication Behavior Description. We use D to represent all network traffic, C means normal network traffic, and T represents Trojans communication data. Then, network traffic behavior feature vector F is defined as follows [5].

$$F(D) = F_1(D) \times F_2(D) \times \ldots \times F_n(D)$$
$$D = C \bigcup T$$

Random variables $F_i(i = 1,2,\ldots,n)$ represent a single behavioral characteristic property of the network traffic, and is a polynomial time computable.

Additionally, we define polynomial time computable function S_F which is consistent with the classification function $F(D)$, and S_F is defined as follows.

$$S_F : F_1 \times F_2 \times \ldots \times F_n \to \{0,1\}$$

We mark 0 as the normal network communications, and 1 refers to a Trojan communications. We call (F, S_F) as the detection model of Trojan communication behavior.

According to the communication behavior of Trojan, the whole process can be divided into keeping connection stage and command control stage, i.e., the detection can be divided in two stages corresponding. The detection process is as follows.

Firstly, we need to extract features of communication behavior according to the different stage of Trojan's. We use F_{ki} and F_{oi} to represent the attributes of the feature during each two stages. Wherein, $F_{ki}(i = 1,2,\ldots,n)$ represents the attributes of the feature during keeping connection, and $F_{oi}(i = 1,2,\ldots,n)$ represents the other.

Secondly, we need to select the algorithm to build detect classification model S_{Fk} and S_{Fo} based on the behavioral features and properties of the data types marked as F_{ki} and F_{oi}, where S_{Fk} represents the classifier of Trojan keeping connection stage, S_{Fo} represents the classifier of Trojan command control stage classifier.

$$S_{Fk} : F_{k1} \times F_{k2} \times \ldots \times F_{kn} \to \{0,1\}, S_{Fo} : F_{o1} \times F_{o2} \times \ldots \times F_{on} \to \{0,1\};$$

Finally, the result is the output of $S_{Fk} \bigcup S_{Fo}$, which means if we detected an alarm at any stage, it considers that the presence of Trojans communication behavior.

Hierarchical Detection Model. According to the features of each stage of the Trojan communication, their behavior features mainly distributed in the network layer, transport layer and application layer.

Network layer features e.g., many sub-connections during primary connection, standard deviation of packet interval.

Transport Layer features, e.g., communications time, "heartbeat" packet to keep-alive, upload and download traffic[6].

Application Layer features e.g., packet entropy, specific port in communication.

Thus, the authors extracted features from network traffic at three layers which were the features of network layer based on IP protocol, transport layer based on TCP protocol and application layer based on HTTP protocol. So we need to build three classifiers based on machine learning for different layers recorded as C_{ip}, C_{tcp} and C_{http}. If it targets some feature in any layer, the traffic will be marked as Trojan data, i.e., the alarms are generated by the output of $C_{ip} \bigcup C_{tcp} \bigcup C_{http}$.

Closed-loop Feedback Model. In General, there are some false positive results for the detection of any machine learning models. We designed a closed-loop feedback model who has a data mining module to remove some false alarms from the detected results. Finally, we feed the false positive results back as the input for the training sets of machine learning model. Then, it can modify training model and reduce false positive rate. The model is designed as follows:

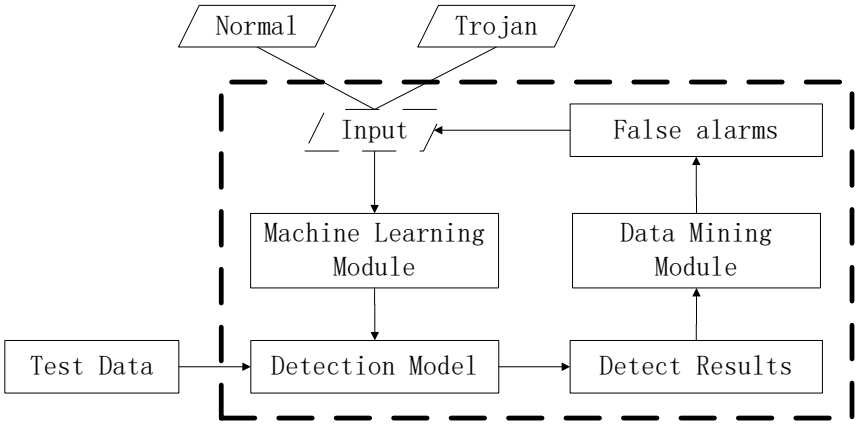

Figure 1: Closed-loop feedback model

In the model, Figure dash parts formed a self-feedback closed-loop, which is used to modify training model and reduce false positive rate.

Experimental Analysis

In order to verify the proposed approach, we collect 65 types of Trojans traffic which contain 224 data flows. Meanwhile, we select MSN, QQ and other normal traffic from 10 normal applications which contain 276 data flows, i.e., a total of 500 data flows for our experiment from the Internet. Two experiments as follows.

The first experiment, we detected 219 data flows from all 224 Trojan flows by our hierarchical detection model, the detect rate is 97.7%. We choose 10 data flows from 10 Trojans from detected. According to the number of features that Trojans targeted in each layer, we can view the features distribution by hierarchical detection model.

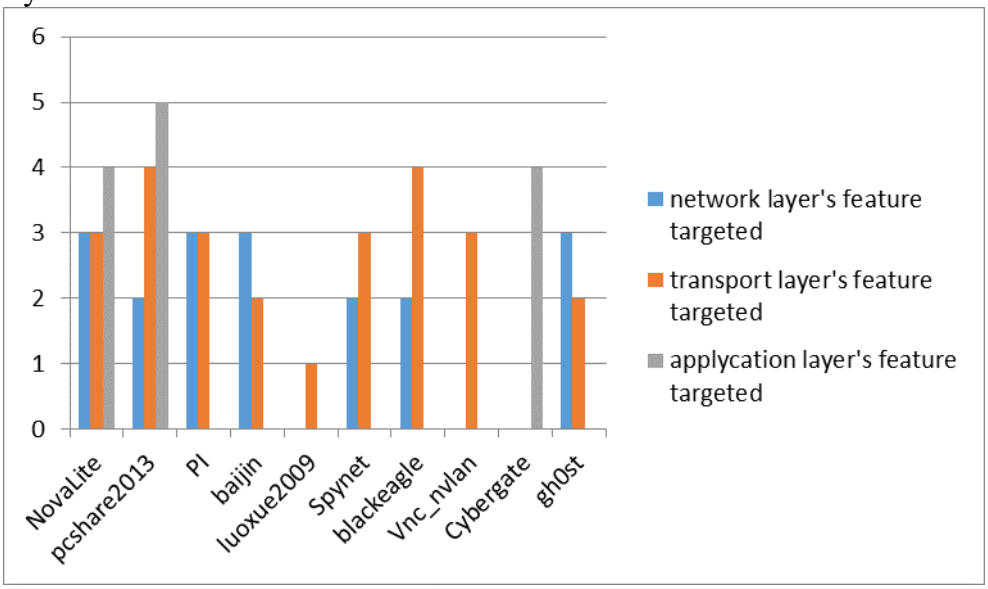

Figure 2: Feature distribution in each layer

The results show that some of detected Trojan data flows only target features in one or two layers, i.e., there are not all targeted features evenly distributing in every layer, some layers may not be targeted. Then we verify the hierarchical detection model can improve the detection rate during whole communication stage.

In the second experiment, we compare the false positive rate between common model and feedback model based on the hierarchical detection model, the experimental data are all normal traffic form 10 normal applications.

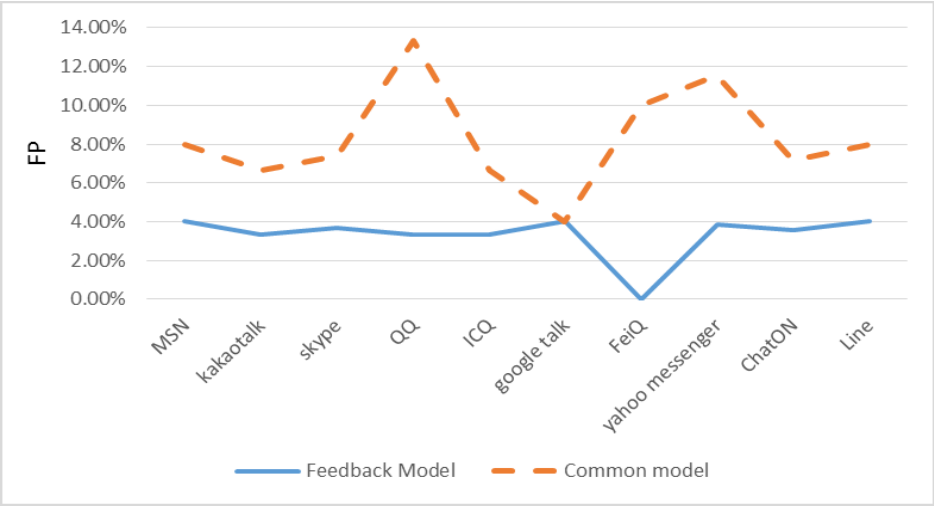

Figure3: Contrast feedback model and common model

The results indicate that our feedback model can reduce the false positive rate, i.e., our approach can get more accurate results.

Summary

In this paper, we presented the idea of detecting Trojan which may be applied in APT attack. Our approach based on the hierarchical detection model and feedback detection model. The two experiments verified our detection model can improve the detection rate and remove part of false alarm, reduce the false positive rate. Further, we can detect Trojans from attacks as APT more effective and provide a more secure network environment to user.

References

[1] PENG Guo-jun, WANG Tai-ge,SHAO Yu-ru,LIU Meng-leng ,Technology and Implementation to Detect Unknown Trojan based on Network Flow Characteristics

[2] Information on http://www2.fireeye.com/rs/fireye/images/fireeye-advanced-targeted-attacks.pdf

[3] MENG Lei,LIU Sheng-li,LIU Long,CHEN Jia-yong,SUN Hai-tao, Trojan Rapid Detection Method Based on Heartbeat Behavior Analysis

[4] Li Shicong, Yun Xiaochun, and Zhang Yongzheng, A Model of Trojan Communication Behavior Detection Based on Hierarchical Clustering Technique

[5] Research on Trojan Horse Detection Technology Based on Communication Behavior Analysis

[6] Dan Jiang, Kazumasa Omote, Approach to Detect Remote Access Trojan in the Early Stage of Communication. 2015 IEEE 29th International Conference on Advanced Information Networking and Applications