A Probabilistic Method
for Detecting Anomalous Program Behavior

Kohei Tatara\textsuperscript{1}, Toshihiro Tabata\textsuperscript{2}, and Kouichi Sakurai\textsuperscript{2}

\textsuperscript{1} Graduate School of Information Science and Electrical Engineering,
Kyushu University
tatara@itslab.csce.kyushu-u.ac.jp
http://itslab.csce.kyushu-u.ac.jp/

\textsuperscript{2} Faculty of Information Science and Electrical Engineering,
Kyushu University, Japan
\{tabata,sakurai\}@csce.kyushu-u.ac.jp

Abstract. In this paper, we, as well as Eskin, Lee, Stolfo \cite{7} propose
a method of prediction model. In their method, the program was char-
acterized with both the order and the kind of system calls. We focus
on a non-sequential feature of system calls given from a program. We
apply a Bayesian network to predicting the $N$-th system call from the
sequence of system calls of the length $N - 1$. In addition, we show that
a correlation between several kinds of system calls can be expressed by
using our method, and can characterize a program behavior.

Keywords: Intrusion detection, Anomaly detection, System call, Bayes-
ian network

1 Introduction

The increase of computers connected to a network with the rapid spread of the
Internet is being enhanced. Therefore, many cases of an unauthorized access to
the computer by the malicious user are reported, and intrusion detection system
which detects such intrusions is getting more and more important.

Many intrusions exploit a vulnerability which is inherent in a program, called
buffer overflow. An attacker rewrites the return address of a function by over-
flowing an internal buffer. Thereby, the attacker can take over the control of the
program and it is possible to execute arbitrary codes \cite{1}. On the other hand,
there are researches of anomaly detection system which detects the intrusion
using buffer overflow by monitoring the control flow of a program \cite{2–11,14}.

An anomaly detection consists of a learning period which learns behavior
during the normal operation of a program, and a monitoring period which sup-
ervises the action of a program by comparing with the records of normal op-
erations. Such an anomaly detection system has an advantage of possibility to
detect an novel intrusion compared with the system which only detects a known
intrusion based on the signature. However, the overhead of anomaly detection
which supervises the execution of a program becomes comparatively large. So it
is important to chose the data which characterize the operation of a program.
Forrest et al. [2] showed that it is possible for normal operation of a program to be characterized by the history of the sequences of system calls which a program emits. If an intrusion using buffer overflow happens, since some sequences of system calls which are not seen during the normal operation will be observed, it is detectable.

Using the method based on N-gram, we can break a sequence of system calls into the sub-sequences of fixed length $N$, and judge whether the operation of a program is normal or not by comparing with those. In order to make only $N$-gram applicable to comparison, it is greatly dependent on $N$ whether incorrect detection takes place. Forrest et al. refered to the optimal value of $N$, namely 6. The 6 is experimentally optimal value on the trade-off between detection capability and efficiency. That is, although efficiency improves when the $N$ is set as less value, detection capability declines a little. They did not touch on the reason why the value of 6 was drawn as optimal value, but was treated as the “magic number” for some years ahead. In order to abolish the necessity of asking for the optimal value of $N$ from the training data, some researchers made attempts to generate a finite state machine [6]. And, other researchers also set $N$ from the normal operation data automatically [7]. Lee et al. [14] did research which explains the mechanism of $N = 6$, from a standpoint of information theory using conditional entropy.

In recent years, the research which clarified the reason was reported by analyzing the data which Forrest et al. exhibited [12]. Kymie et al. proved that the condition in which at least one “Minimal foreign length” exists in the sequences of system calls collected when an intrusion happened is that the value of $N$ is six or more. Minimal foreign length is the minimum value of $N$ in which at least one unique sub-sequence exists, when the sequences of system calls are divided into the sub-sequences of length $N$ [12]. Clearly from their report, in order to use system call as normal data of an anomaly detection, the features from the sequences of system calls used for learning need to differ from those observed from anomalous processes. Although the more than 200 kinds of system calls exist in the current version of Linux, about dozens of kinds of the system calls are actually emitted by the program at most. The total number of possible N-grams is too small to characterize the behavior of program by the system calls in the case of $N = 1$. If $\Sigma$ kinds of system calls are published in the program, then there are $\Sigma^N$ possible sequences of length $N$. Therefore, as $\Sigma$ is large, the probability under which Minimal foreign length is found increases by leaps and bounds. Kymie et al. showed the method of calculating the optimal value of $N$, when the stide [2] and Hofmyer et al.’s method [3] were used for an anomaly detection system. However, when these methods are applied to other data sets, we do not affirm whether $N = 6$ may become the “magic number.” When the optimal value of $N$ is 7 in a certain data set, performing anomaly detection using the value 6 contains a possibility of overlooking abnormal sequences.

On the other hand, there is a research using the rule learning program, called RIPPER, which focuses on correlation of system calls [13]. In this paper, Lee et al. input those sequences which are attached the labels, “normal” or “abnormal”