Research Article

A Universal High-Performance Correlation Analysis Detection Model and Algorithm for Network Intrusion Detection System

Hongliang Zhu,1 Wenhan Liu,1 Maohua Sun,2 and Yang Xin1

1Beijing University of Posts and Telecommunications, Beijing, China
2Information School, Capital University of Economics and Business, Beijing, China

Correspondence should be addressed to Hongliang Zhu; zhuhongliang@bupt.edu.cn

Received 2 February 2017; Accepted 3 May 2017; Published 23 May 2017

Academic Editor: Zonghua Zhang

Copyright © 2017 Hongliang Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In big data era, the single detection techniques have already not met the demand of complex network attacks and advanced persistent threats, but there is no uniform standard to make different correlation analysis detection be performed efficiently and accurately. In this paper, we put forward a universal correlation analysis detection model and algorithm by introducing state transition diagram. Based on analyzing and comparing the current correlation detection modes, we formalize the correlation patterns and propose a framework according to data packet timing and behavior qualities and then design a new universal algorithm to implement the method. Finally, experiment, which sets up a lightweight intrusion detection system using KDD1999 dataset, shows that the correlation detection model and algorithm can improve the performance and guarantee high detection rates.

1. Introduction

(A) Background. Intrusion detection is a kind of technology which recognizes the intrusion by collecting and analyzing the protected system information [1]. The crucial functions are monitoring Internet and computer system, discovering and distinguishing the intrusion behaviors or attempts, and generating intrusion alarm in real time [2]. Intrusion detection can be thought as a binary technology that distinguishes whether the system state is “normal” or “attack” [3]. The requirements of the intrusion detection system are the detection rate, that is, detection accuracy, followed by real time. Only in high detection speed, it can deal with massive data transmitted in Internet in time, get rid of missing information for low speed [4], cause false negatives and false positives, and minimize the losses brought by the intrusion. However, with the diversification in kind and increasing in number of network attack means, there is a key issue of low detection rate for intrusion detection system [5]. In addition, the traditional intrusion detection system detects slowly and consumes large amounts of resources. With the quick development of network speed, it can not process the massive data transmitted in real time, resulting in a large increase of false positives rate and false negatives rate [6]. Also these problems are becoming more and more serious. Detection rate and detection speed have become important indicators of intrusion detection system real-time requirements [7]. How to build a high detection rate and detection speed intrusion system has become the focus of current research. Figure 1 gives an overview of a universal network intrusion detection framework. A key point in this figure is the use of deep analysis modules to process the associated events.

The deep analysis module plays an important role in intrusion detection system. We can see that the data of deep analysis modules come from two parts, one part is the result of detection on the upper layer and the other part is the raw data. Various data packets or events are processed by correlation algorithm, such as a correlation detection of event frequency and correlation detection for multiple parallel events and so on. The performance of correlation detection influences directly the detection rate and detection speed of intrusion detection system. However, the factors which influence the correlation detection result are various, so it is difficult to extract a unified correlation detection algorithm.
Therefore, it is very necessary to find a universal correlation detection algorithm to increase detection rate and speed.

(B) Related Works. In order to detect anomalies in network, correlate parameters from different layers should be combined [8]. Some papers focus on building a new hierarchical framework for intrusion detection as well as data processing based on the feature classification and selection [9–11]. Intrusion detection system has been studied by means of machine learning, and the detection rate has got improvements [12–19]. In addition, intrusion detection has been performed by using feature association technique, and the data set has been used for analysis [20–25].

(C) Contribution. In this paper, we propose a novel method to increase the detection rate of intrusion detection system and improve the detection speed. This method is a correlation analysis detection model based on data packet timing and behavior quality, aiming to solve the problem of versatility, consistency, and the integrity of packet detection. This method enables us to overcome the disadvantage of traditional intrusion detection system.

The rest of this paper is organized as follows. In Section 2, we analyze and compare the current common data packet correlation detection modes briefly. Section 3 presents the generating process of the algorithm in detail. In Section 4, we present the detection process for intrusion detection system and make some experiments. In the end, we conclude the paper in Section 5.

2. System Overview

In intrusion detection system, for single session, there will be false positives when describing threatening events only by single feature, in order to reduce the behavior features of attraction events accurately. However, some papers have pointed out that there is relevance among different attack events [26]. If every session is analyzed separately, we can not identify the attack behavior exactly. While when we consider the related sessions correctly, we can identify an attack event completely. Nowadays, the majority of correlation detecting methods of intrusion detection system are as follows: correlation detection of event frequency, correlation detection for multiple parallel events, correlation detection for multiple serial events, correlation detection based on source IP of the event, correlation detection based on destination IP of the event, correlation detection based on resource of events and destination IP, and correlation of session [27, 28]. There are more and more weaknesses in traditional correlation detection, such as low detection rate and poor accuracy. Thus, we put forward a unified correlation detection algorithm and build a data packet correlation detection model based on the data packet timing and behavior quality, aiming to solve the problem of the versatility, consistency, and integrity of intrusion detection. Figure 2 gives an overview of a data packet correlation detection mode.

According to the behavior features, there are two kinds of intrusion events: one is unconditional trigger and the other is conditional trigger.

(i) Unconditional trigger: correlated detecting based on the order of event occurring, including correlation detection of event frequency, correlation detection for multiple parallel events, and correlation detection for multiple serial events.
(ii) Conditional trigger: correlated detection based on behavior feature, including single behavior feature and complex behavior feature.

3. Correlation Analysis Detection Model

In this section, we present the correlation analysis detection model as follows.

3.1. Concept Definition

Definition 1 (distributed packet flow). The time series is given as \( T = \{ t_1, t_2, \ldots, t_i, \ldots \} \). The number of nodes is \( n \). A distributed packet flow is defined as \( S = \{ S_1, S_2, \ldots, S_n \} \); each item of \( S \) is \( S_k \) (\( k = 1, 2, \ldots, n \)), a single data packet, which is the original event collection on \( T \).

Definition 2. The primitive event \( E \) is two-tuple \((T, A)\).

(i) \( T \) is the timestamp of the original event, that is, the time node of the event on the time series.

(ii) \( A \) is the behavioral characteristics of the original event.

Definition 3. The LAMBDA syntax is used to define the different relationships between primitive events: \( e_0, e_1, \ldots, e_n \).

(i) Existence: \( \exists \) indicates whether the event \( e_i \) exists or not.

(ii) Parallel: \( e_1 | e_2 \) indicates that the events \( e_1 \) and \( e_2 \) are parallel relations.

(iii) Serial: \( e_1; e_2 \) indicates that the events \( e_1 \) and \( e_2 \) are serial relations.

Definition 4. The initial state of the system in the state diagram is \( S_0 \), the intermediate state is \( S_i \), and \( N \) is the termination status. Each node in the following graph represents the current state of the event, and if there is only one transition condition, an arrow arc exists between the two nodes to indicate the transition of the event state. The mark on the arc represents the transition condition.

3.2. Formal Expression of Behavior Detection

3.2.1. Unconditional Trigger Type

(i) A Correlation Detection of Event Frequency. This method detects that the original event that contains the threat behavior feature directly, and then it performs response processing.

The state transition relationship is

\[ N = e_i. \quad (1) \]

The state diagram is shown in Figure 3.

(ii) Correlation Detection for Multiple Parallel Events. Some threat behaviors can be detected when multiple events occur at the same time.

The state transition relationship is

\[ N = e_0 | e_1 | e_2 | \cdots | e_n. \quad (2) \]

The state diagram is shown in Figure 4.

(iii) Correlation Detection for Multiple Serial Events. Some threat behaviors can be detected when multiple events occur in sequence.

The state transition relationship is

\[ N = e_0; e_1; e_2; \ldots; e_n. \quad (3) \]

The state transition diagram is shown in Figure 5.

3.2.2. Conditional Trigger Type

(i) According to the Single-Event Feature of the Event Correlation Detection. Some threat behaviors can be detected when multiple events simultaneously satisfy a certain behavioral characteristic.

The state transition relationship is

\[ N = A_i \quad (i = 0, 1, 2, \ldots, n). \quad (4) \]

The state transition diagram is shown in Figure 6.
3.3. Detection Algorithm Generation. According to the data packet correlation detection mode and state diagram analysis, this paper proposes the data packet correlation detection model, which can be used to detect anomaly or original data packet in intrusion detection system to improve system detection rate and reduce detection time.

3.3.1. Correlation Detection Formula

Definition 1. \( S_0 \) indicates the initial state of the detection system, \( S_x \) indicates that the system state is detected at any time, \( A_x \) represents the behavioral characteristics of the event, and \( N_x \) indicates the termination status of the detection system.

Definition 2. \( N_x(S_x \mid A_x) \) indicates that the input behavior attributes \( A_x \) resulting in changes in system status, and the detection system termination status is \( N_x \).

According to the above formula definition and state diagram, packet correlation detection formula can be as follows:

\[
\begin{align*}
\forall A_{n_1} \neq \forall A_{n_2}, \\
\forall A_{n_1} = \forall A_{n_2}, & \quad (a, z = 1, 2, 3, \ldots), \\
& \quad (a, x = 1, 2, 3, \ldots), \\
& \quad (i, x = 1, 2, 3, \ldots), \\
& \quad (i, n \in \mathbb{N}^+). \\
\end{align*}
\]

3.3.2. Formula Proof

(1) \( S_0 \mid A_i \), that is,

\[
\forall S_{ai} = \forall S_{ax} = \forall S_{zi} \quad (a, i, x, z = 0, 1, 2, \ldots),
\]

indicates that the detection system starts from \( S_0 \) and ends at final state \( N_x \) after a single behavior of \( A_i \).

(2) \( S_0 \mid S_1 \mid S_2 \mid \cdots \mid S_{n} \mid A_i \), that is,

\[
\forall S_{ai} \neq \forall S_{zi},
\]

\[
\forall A_{n_1} \neq \forall A_{n_2},
\]

\[
(a, z = 1, 2, 3, \ldots), \\
(a, z = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, n \in \mathbb{N}^+).
\]

indicates that the detection system is parallel to multiple events \( S_0, S_1, \ldots, S_n \) and ends at final state \( N_x \) after a single behavior of \( A_i \).

(3) \( S_0 \mid S_1 \mid S_2 \mid \cdots \mid S_{n} \mid (A_0, A_1, A_2, \ldots, A_n) \), that is,

\[
\forall S_{ai} \neq \forall S_{zi},
\]

\[
\forall A_{n_1} \neq \forall A_{n_2},
\]

\[
(a, z = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, n \in \mathbb{N}^+).
\]

indicates that the detection system is parallel to multiple events \( S_0, S_1, \ldots, S_n \) and ends at final state \( N_x \) after the composite behavior of \( A_0, A_1, A_2, \ldots, A_n \).

(4) \( S_0 \mid S_1 \mid \cdots \mid S_{n} \mid A_i \), that is,

\[
\forall S_{ai} = \forall S_{zi},
\]

\[
\forall A_{n_1} = \forall A_{n_2},
\]

\[
(a, z = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, n \in \mathbb{N}^+).
\]

indicates that the detection system is parallel to multiple events \( S_0, S_1, \ldots, S_n \) and ends at final state \( N_x \) after a single behavior of \( A_i \).

(5) \( S_0 \mid S_1 \mid S_2 \mid \cdots \mid S_{n} \mid (A_0, A_1, A_2, \ldots, A_n) \), that is,

\[
\forall S_{ai} = \forall S_{zi},
\]

\[
\forall A_{n_1} = \forall A_{n_2},
\]

\[
(a, z = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, n \in \mathbb{N}^+).
\]

indicates that the detection system is serial to multiple events \( S_0, S_1, \ldots, S_n \) and ends at final state \( N_x \) after the composite behavior of \( A_0, A_1, A_2, \ldots, A_n \).

(6) \( S_0 \mid S_1 \mid \cdots \mid S_{n} \mid A_i \mid A_j \), that is,

\[
\forall S_{ai} = \forall S_{zi},
\]

\[
\forall A_{n_1} = \forall A_{n_2},
\]

\[
(a, z = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, x = 1, 2, 3, \ldots), \\
(i, n \in \mathbb{N}^+).
\]

indicates that the detection system is serial to multiple events \( S_0, S_1, \ldots, S_n \) and ends at final state \( N_x \) after the composite behavior of \( A_0, A_1, A_2, \ldots, A_n \).
3.3.3. Detection Algorithm Proposed. Based on the detection algorithm and the formula proposed above, this paper proposes the data packet correlation detection model. Figure 8 gives a formal representation of the data packet correlation detection model by state diagram. A key point in this figure is the use of deep analysis modules to process the associated events.

When the anomaly detection results or original data packets flow into deep analysis module in intrusion detection system, the data packet correlation detection model based on timing and behavior features can detect the events pointedly and thoroughly with the existing detection modes. Compared with traditional and single detection model, this algorithm increases detection speed and precision. Figure 9 is the detection framework of this algorithm.

In the first layer of this model, it detects timing characteristic (e.g., sniffing attacks in sequence) and behavior characteristic (e.g., as attacking continuously certain IP address). Deep analysis is in the second layer, which detects correlation detects combined with timing characteristic and behavior characteristic and aims at the detection of various persistent concealed attack behaviors. There is no need to detect data flows according to behavior feature in order and it can detect attack behaviors roundly, simplifying traditional detection modes.

4. Lightweight Intrusion Detection System Based on Correlation Analysis Detection Model

The flow diagram of deep analysis in lightweight intrusion detection system based on correlation analysis detection model is shown in Figure 10. The crucial part of the diagram is correlation detection. Firstly, it verifies ports and finds flows correlation table. Secondly, it uses correlation analysis detection model to detect and makes DPI and DFI identification. Finally, timing and behavior features are written into correlation table and these results are returned.

In this part of our paper, we will analyze and compare the traditional intrusion detection system and intrusion detection system based on correlation analysis detection model by detecting the data set that consists of 41 features in KDD1999. Then we compare the results of detection rate and detection time.

All verification work of this paper is based on KDD1999 data set. Before the experiment, we preprocess KDD1999 data set to meet experiment requirements. The environment of experiment is Window 7 operation system, and the hardware parameters are Quad-Core Intel Core i7 processor 3.2GHz, 4096 MB RAM.

4.1. Preprocessing of Data Set. KDD1999 is a standard data set used for intrusion detection test. The KDD1999 dataset consists of a total of 5 million records, and it also provides a 10% training data set and a test data set. There are 494021 instances in training data set and 41 features in each instance, while there are 311029 instances in test data set. We divide the training data set into 5 parts: DOS attack, PROBE attack, R21 attack, U2R attack, and NORMAL. And NORMAL means normal data, excluding attack. There are 6 different kinds of attacks in DOS, 4 kinds of attacks in PROBE, 8 kinds of attacks in R21, and 4 kinds of attacks in U2R. The number of NORMAL instances is 97278 in the whole training data.
set; the number of DOS instances is 391458; the number of PROBE instances is 4107; the number of R21 instances is 1126, while the number of U2R instances is 52 and the sample category distribution and apart of attack types are shown in Table 1.

According to Table 1, there are 22 kinds of various attack types. As for KDD1999 test data set, we divide every type of attack into 2 parts according to the same principle: known attack and unknown attack. The known attack means the attack types that have appeared in the training data set, while the unknown attack means the attack types that have not appeared in training data set. There are 39 kinds of attack types totally in test data set: 10 types in DOS, 4 are unknown attacks; 6 types in PROBE, 2 are unknown attacks; 15 types in R21, 7 are unknown attacks; 8 types in U2R, 4 are unknown attacks. These attack types are listed in Table 1. There are 4166 instances in PROBE attack, and 2377 instances are known attack instances, accounting for 57.1% PROBE instances; 1789
instances are known attack instances, representing 42.9%. The figures for instances of known and unknown attacks are in Table 1.

4.2. Experimental Program. After the analysis and process of KDD 1999 data set in the section above, the number of instances in training set is much larger. We know that, in order to make experiment operation more convenient, we should select data to reduce the number of instances. We sample the DOS, PROBE, R2L, U2R, and NORMAL in training set randomly and respectively and ensure the consistency of these samples and the original sample. Then we combine these 5 new samples and form a new training data set. These 5 samples are the combination of NORMAL and DOS, the combination of NORMAL and PROBE, the combination of NORMAL and R2L, the combination of NORMAL and U2R, and the combination of NORMAL, DOS, PROBE, R2L, and U2R. The instance of 5 new training sets is 98630. These 5 training sets are flowed into the traditional intrusion detection system and intrusion detection system based on data packet correlation detection model. We compare the performances based on the same data resource by comparing detection rate, detection time, and so on.

4.3. Experimental Results and Analysis. The experiment result is shown in Table 2; it is obvious to get that the detection rate rises sharply in intrusion detection system based on data packet correlation detection model and the detection time decreases, promoting the efficiency of detection system. Therefore, for intrusion detection system based on the correlation analysis detection model in this paper, the detection rate of known and unknown attacks is high, improving the performance of intrusion detection system.

Table 1: KDD1999 sample category distribution and partly attack type statistics.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Attack type</th>
<th>The training set</th>
<th>The test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMAL</td>
<td>/</td>
<td>97278</td>
<td>60593</td>
</tr>
<tr>
<td>DOS</td>
<td>/</td>
<td>794</td>
<td></td>
</tr>
<tr>
<td></td>
<td>apache2</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td></td>
<td>back</td>
<td>97278</td>
<td></td>
</tr>
<tr>
<td></td>
<td>land</td>
<td>60593</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mailbomb</td>
<td>1098</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neptune</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pod</td>
<td>5000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>processstable</td>
<td>58001</td>
<td></td>
</tr>
<tr>
<td>PROBE</td>
<td>/</td>
<td>306</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ipsweep</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mscan</td>
<td>1053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nmap</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>portswep</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>saint</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>satan</td>
<td>736</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>1633</td>
<td></td>
</tr>
<tr>
<td>R2L</td>
<td>/</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ftp_write</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>imap</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>named</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>phf</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sendmail</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>U2R</td>
<td>/</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td></td>
<td>http_tunnel</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>loadmodule</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>perl</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ps</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rootkit</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sqlattack</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>494021</td>
<td>311029</td>
</tr>
</tbody>
</table>

Table 2: The experiment result.

<table>
<thead>
<tr>
<th></th>
<th>Detection rate (%)</th>
<th>Detection time (s)</th>
<th>Known attack detection rate (%)</th>
<th>Unknown attack detection rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional IDS</td>
<td>79</td>
<td>182</td>
<td>87</td>
<td>71</td>
</tr>
<tr>
<td>IDS based on the correlation model</td>
<td>92</td>
<td>156</td>
<td>95</td>
<td>89</td>
</tr>
</tbody>
</table>
5. Conclusions

In this paper, we build a high-performance correlation analysis detection model, which aims to resolve the low detection rate and slow detection speed.

For the intrusion detection system, we put forward a kind of universal network intrusion detection framework. Meanwhile, we analyze and compare the current common correlation intrusion detection modes. Finally, we propose a data packet correlation detection model and algorithm based on the data packet timing and behavior characteristics. In the experiments, this kind of correlation detection model has improvement in performance than former. In this paper the present popular intrusion detection system has good practical value.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by National Natural Science Foundation of China Project (61302087).

References


Submit your manuscripts at
https://www.hindawi.com