Data Analysis for Anomaly Detection to Secure Rail Network

Huaqun Guo  
Institute for Infocomm Research  
Agency for Science Technology and Research  
Singapore  
guohq@i2r.a-star.edu.sg

Xiaoyi Shen, Wang Ling Goh  
School of Electrical and Electronics Engineering  
Nanyang Technological University  
Singapore  
xiaoyi002@e.ntu.edu.sg;  
ewlgoh@ntu.edu.sg

Luying Zhou  
Institute for Infocomm Research  
Agency for Science Technology and Research  
Singapore  
lzhou@i2r.a-star.edu.sg

Abstract— The security, safety and reliability of rail systems are of the utmost importance. In order to better detect and prevent anomalies, it is necessary to accurately study and analyze the network traffic and abnormal behaviors, as well as to detect and alert any anomalies if happened. This paper focuses on data analysis for anomaly detection with Wireshark and packet analysis system. An alert function is also developed to provide an alert when abnormality happens. Rail network traffic data have been captured and analyzed so that their network features are obtained and used to detect the abnormality. To improve efficiency, a packet analysis system is introduced to receive the network flow and analyze data automatically. The provision of two detection methods, i.e., the Wireshark detection and the packet analysis system together with the alert function will facilitate the timely detection of abnormality and triggering of alert in the rail network.

Keywords—Rail Network Security; Data Analysis; Anomaly Detection; Packet Analysis System;

I. INTRODUCTION

The security, safety, and reliability of urban transportation systems (i.e. rail systems) are crucial, especially for city with dense population. However, with urban transportation systems increasingly reliant on their “cyber” components, cyber-attacks have posed great threat in catastrophic disruption and damages. In 2003, the train signalling and dispatch system had been brought down by a computer virus for more than 4 hours in Jacksonville, Florida [1]. In 2007, a tram derailment occurred due to a wireless remote controller played by a Polish teenager and 12 people were injured [2]. In 2012, in Shenzhen, China, the wireless channels used for the train-to-infrastructure communications were interfered and jammed by passengers’ WiFi enabled mobile phones [3]. Thus, coping with the increasing risk of cyber-attacks has become a major concern for transit operators.

A typical rail network system is schematically shown in Fig. 1 [4]. If a malicious intrusion occurs, it is more likely to cause rail network damage and data leakage. What is worse, it can directly lead to property damage or personal injury of passengers. Thus, this paper aims to analyze data collected from the rail network, so as to understand the characteristics of different attribute fields such as network protocols, time-based network traffic statistics, host-based network traffic statistics, and so on. Using Wireshark combined with multiple theoretical methods in statistical analysis as a technical support and method to detect the abnormality. Furthermore, a packet analysis system together with alert function is then proposed to facilitate real-time anomaly detection.

The remainder of this paper is organized as follows. A literature review on the detection methods is introduced in Section II. Section III provides the rail network data analysis. Section IV describes the data analysis for anomaly detection with Wireshark, while Section V presents a packet analysis system together with alert function. Finally, Section VI concludes this work.

![Figure 1. A typical rail network system](image-url)
II. LITERATURE REVIEW

A. Statistical Detection Method

Statistical-based abnormal detection is the earliest and most mature application of abnormal detection technology. It identifies abnormalities by analyzing behaviors with large statistical deviations from normal activities. The statistical analysis method samples target behavior characteristics, such as distribution of audit data, intensity of measured behavior, distribution of different audit behaviors, etc. A series of parameters describing the behavior of the target are calculated to form a detection behavior profile. Since network events (such as packet arrivals) occur dynamically, the system merges each acquired behavior profile and target behavior profile to obtain a normal behavior profile. During detection, an incoming packet is judged by comparing with the normal behavior profile. When the threshold value is exceeded, an abnormal alarm is issued. Hyperstatistical Theory-Based Detection Method [5] and Probability Distribution and Stochastic Process Method are examples of a statistical detection method [6].

The advantages of detection method based on statistical analysis can be summarized as follow:

- Statistical detection method obtains a normal traffic model mainly through historical traffic, finding abnormalities by analyzing whether it meets normal model.
- Statistical detection technology is mainly based on the macro characteristics of the flow to detect abnormalities, which can better meet the real-time nature.
- It requires no prior knowledge of security vulnerabilities or offensive behavior to detect unknown anomalous behaviors.
- It can accurately detect denial of service attacks (DDoS) and periodic anomalies (such as port scans).

B. Machine Learning Detection Method

A Machine Learning (ML) algorithm model is used to detect the abnormality through establishing a range of spatial values through training data. Under normal circumstances, the data values of multidimensional normal network behavior will fall into the normal range. However, when abnormal behavior occurs, the value will fall outside the normal range [7]. Based on this, we can basically judge that an abnormality has occurred. Unlike traditional algorithms, ML algorithms do not set a fixed filter. It is based on changes in the state of the network and dynamic, periodic training and update filter conditions. The Artificial Neural Network (ANN) [8], Support Vector Machine (SVM) [9] and Decision Tree [10] are the examples of Machine Learning techniques that can be used in anomaly detection.

For network intrusion detection system with stringent real-time requirements, the corresponding response will be detected and stimulated before the destructive effect occurs. The output of ANN can be a value of the possibility of invasion or the level of data hazard. Through nonlinear analysis, this value is learned rather than mechanically matched in a rule-based system. This gives the ANN the ability to detect new intrusions.

SVM based intrusion detection models mainly include data preprocessors, classifiers, decision-response, and SVM training. The intrusion detection scheme is far superior to the neural network-based attack recognition in terms of performance and recognition rate. This shows that the SVM method is much better than the neural network in dealing with small sample. SVM can also well overcome the neural network convergence.

The decision tree algorithm relies on training data to automatically generate a decision tree classification model. It can use the generated decision tree to classify unknown data to attain the intended prediction. The abnormality detection process is used as a process of detecting and predicting data packets, and thereafter, creating a decision tree algorithm and model to predict whether it is an abnormality. The decision-making process of the predictive decision tree is the process of matching a path from root node to a leaf node, which reduces the number of matched operations as well as the complexity of calculation.

C. Entropy-Based Detection Method

Entropy can be used to determine randomness. The main reason for using entropy for abnormality detection is that it can measure the randomness of incoming network traffic. The higher the entropy value, the higher the randomness of the traffic. On the contrary, the lower the entropy value, the higher the certainty of the traffic. This feature makes it possible to judge network attacks by using entropy. There are three main methods used nowadays, namely the Alive Entropy-Based Detection Method [11], Relative Entropy-Based Detection Method [12], and Cross Entropy-Based Detection Method [13].

III. RAIL NETWORK DATA ANALYSIS

A. Data Analysis

In the rail network data analysis, 287 data samples are collected (each sample have almost 12000 packet information) from the rail network. We analyze the number of packets from each IP address and calculate the percentage of each IP address.

The IP protocol is used to connect multiple packet-switched networks. It transports data packets between the source and destination addresses. It also provides data reassembly to accommodate the different size of the request in the networks. From the sample data, the main IP protocol used in the rail network are summarized through analyzing the packets. According to the result, the frequently used protocols in the rail network are: HSRP (76%), TCP (17%), ARP (2.3%), and Intel ANS probe (2%). Figure 2 shows the distribution of these protocols in the rail network. The right side and left side legends shown in the figure are two redundant links in case one failed.

In the targeted rail network, there is a special protocol called Modbus under TCP. When communicating over the same Modbus network, this protocol determines that each
controller needs to know its device address, recognize messages sent by address, and decide what action to execute. The controller also generates a response message and sends it using the Modbus protocol in the event that a response is required.

Specific Modbus messages are sent only at a specific 2 minutes time period every day. From the sampled data, source/destination IP addresses and port are analyzed and the time interval is shown in Fig. 3.

From the data analysis, a trust list is generated. If the information contained in the packet do not match the data in the trust list, an abnormality may have occurred.

B. Firewall Rules

The Firewall rules have been set up to allow or block/deny packets after data analysis from the rail network. The firewall rules are listed using the format below:

<table>
<thead>
<tr>
<th>Source</th>
<th>Destination</th>
<th>Direction</th>
<th>Protocol</th>
<th>Source / Destination Port Number (if TCP or UDP)</th>
<th>Action</th>
</tr>
</thead>
</table>

Transport / Application protocols includes but not limited to:

- TCP port 1025, 1206 and 5000
- 172.24.50.1 | 172.24.50.2 | TCP | destination port 5000 | Allow
- 172.24.50.1 | 172.24.50.2 | TCP | destination port 1025 | Allow

MAC-level protocols

1. ARP (address resolution protocol)
   - Any | Any | Bi-direction | ARP | Allow

![Protocol used in network](image)

Figure 2. Protocol used in network.

Figure 3. Time interval in Modbus.

2. Intel ANS
   - Any | Any | Bi-direction | Intel ANS | Allow

3. Cisco’s LOOP
   - Any | Any | Bi-direction | LOOP | Allow

4. STP (Spanning Tree Protocol)
   - Any | Any | Bi-direction | STP | Allow

Network-level / routing protocols

1. HSRP (Cisco’s Hot Standby Router Protocol)
   - 172.24.1.1 | 172.24.50.2 | HSRP | Allow

2. ICMP (Internet Control Message Protocol)
   - Includes but not limited to:
     - 172.24.2.1 | 172.24.61.1 | Bi-direction | ICMP | Allow
     - 172.50.42.1 | 172.24.50.1 | Bi-direction | ICMP | Allow

3. NBNS (NetBIOS Name Service)
   - Includes but not limited to:
     - 172.24.61.1 | 172.24.255.255 | NBNS | Allow
     - 172.24.61.2 | 172.24.255.255 | NBNS | Allow

4. NTP (Network Time Protocol)
   - Includes but not limited to:
     - 172.24.1.1 | 172.24.50.1 | Bi-direction | NTP | Allow
     - 172.24.2.1 | 172.24.50.1 | Bi-direction | NTP | Allow

5. OSPF (Open Shortest Path First)
   - 172.24.1.1 | 224.0.0.5 | OSPF | Allow
   - 172.24.1.2 | 224.0.0.5 | OSPF | Allow

6. RSL (Radio Signalling Link)
   - 172.24.50.1 | 172.24.50.2 | Bi-directional | RSL | Allow

IV. DATA ANALYSIS WITH WIRESHARK

Wireshark is a very popular freeware used in network packet analysis [14]. It can intercept a variety of network packets and show details. It can be used in different ways too. According to these characteristic, a few methodologies can be used via Wireshark to detect abnormality.

(1) Number of packets

The number of packets is shown in Table I. If the percentage of the packet is increasing or decreasing out of threshold we set (concluded from the data analysis), there is abnormality happened and need to do the further action.

![Figure 2](image)

![Figure 3](image)

TABLE I. THE NUMBER OF PACKETS

<table>
<thead>
<tr>
<th>Topic/Item</th>
<th>Count</th>
<th>Ave.</th>
<th>Min Val</th>
<th>Max Val</th>
<th>Rate (ms)</th>
<th>Percent</th>
<th>Burst rate</th>
<th>Burst start</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Addresses</td>
<td>57405</td>
<td>0.0958</td>
<td>100%</td>
<td>0.2400</td>
<td>102.278</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>255.255.255.255</td>
<td>2</td>
<td>0.0000</td>
<td>0.00%</td>
<td>0.0100</td>
<td>176.550</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>224.0.0.5</td>
<td>992</td>
<td>0.003</td>
<td>0.33%</td>
<td>0.0600</td>
<td>592.369</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.24.50.1</td>
<td>11028</td>
<td>0.0184</td>
<td>19.19%</td>
<td>0.1500</td>
<td>102.248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.24.50.2</td>
<td>5324</td>
<td>0.0089</td>
<td>9.27%</td>
<td>0.0600</td>
<td>2.567</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.24.61.1</td>
<td>23103</td>
<td>0.0386</td>
<td>40.25%</td>
<td>0.0700</td>
<td>6.258</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.24.2.1</td>
<td>118</td>
<td>0.002</td>
<td>0.21%</td>
<td>0.0300</td>
<td>580.330</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.24.1.1</td>
<td>352</td>
<td>0.0094</td>
<td>19.83%</td>
<td>0.1500</td>
<td>102.248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.24.1.2</td>
<td>2303</td>
<td>0.0386</td>
<td>40.27%</td>
<td>0.0700</td>
<td>592.361</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.26.4.5</td>
<td>4</td>
<td>0.0000</td>
<td>0.00%</td>
<td>0.0100</td>
<td>178.512</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.26.4.5</td>
<td>1</td>
<td>0.0000</td>
<td>0.00%</td>
<td>0.0100</td>
<td>178.512</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(2) Appearance of new packets

If there are new IP addresses, IP protocols and port numbers that are not in the trust list (conclude from the data analysis) appears, signifying that abnormality may have occurred. For example, if IP address (172.14.63.1) is not in the trust list but appears in Table II, the abnormal packets are detected.

(3) Frequency of packets

The frequency fluctuations for every source and destination to send or receive the packets are fixed within a certain range. If it exceeds the range, abnormality may exist. For example, the frequency of sending and receiving packets between 172.24.10.1 and 172.24.50.1 is 0.5s (per four packets); if the frequency changes significantly, abnormality is occurred.

For anomaly detection with Wireshark, its advantage is that it is easy to implement. The first step is to download the software and configure all the parameters. Next, the filter condition is to be set in advance, followed by visualizing of results and data analysis. Lastly, the result is to be compared with the trust list to determine if an abnormality has happened.

There are disadvantages with Wireshark. First, analysis needs to be done by users after receiving the packets, and it is not able to conduct real-time analysis. Second, it requires the operator to constantly monitor the interface to see if there is an abnormality, thereby draining time and human resources.

V. PACKET ANALYSIS SYSTEM DESIGN

As mentioned in the previous session, although Wireshark is easy to implement, it requires extra time and human resource. A system that can analyze the packets captured from the rail system network to automatically provide real-time result is thus proposed in this Section. The process to analyze the packets is shown in Fig. 4.

Firstly, when receiving the file transported from the rail network, the system will read all packets and ensure that it is operational.

Thereafter the arrival time need to be analyzed. For some specific packet data (for example, the Modbus packets are only sent at a fixed period of 2 minutes), the reception at a wrong time period can be concluded as an abnormality.

Other information contained in the packets ought to be analyzed as well. The designed software code is used to examine the data frames to acquire the source and destination MAC addresses and the total size of the packet.

After getting the source/destination MAC addresses and size of packet, the next step is to evaluate the protocols used in the network. If there appears a new protocol that is not able to be recognized by the program (the protocols written in the software code are the same as those in the trust list), it may be abnormal.

Lastly, the source/destination IP addresses and the port number are to be determined. If there appears different IP addresses that are not listed in our trust list, an alert is triggered.

<table>
<thead>
<tr>
<th>Address A</th>
<th>Address B</th>
<th>Packets</th>
<th>Bytes</th>
<th>Packets A -&gt; B</th>
<th>Bytes A -&gt; B</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.26.4.51</td>
<td>255.255.255.255</td>
<td>1</td>
<td>82</td>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>100.26.4.52</td>
<td>255.255.255.255</td>
<td>1</td>
<td>82</td>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>172.24.1.1</td>
<td>224.0.0.2</td>
<td>23,012</td>
<td>1426k</td>
<td>23,012</td>
<td>1426k</td>
</tr>
<tr>
<td>172.24.1.1</td>
<td>224.0.0.5</td>
<td>105</td>
<td>11k</td>
<td>105</td>
<td>11k</td>
</tr>
<tr>
<td>172.24.1.1</td>
<td>172.14.50.1</td>
<td>2</td>
<td>180</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>172.24.1.2</td>
<td>224.0.0.2</td>
<td>23,014</td>
<td>1426k</td>
<td>23,014</td>
<td>1426k</td>
</tr>
<tr>
<td>172.24.1.2</td>
<td>224.0.0.5</td>
<td>87</td>
<td>9,438</td>
<td>87</td>
<td>9,438</td>
</tr>
<tr>
<td>172.24.1.2</td>
<td>172.24.50.1</td>
<td>2</td>
<td>180</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>172.24.2.1</td>
<td>172.24.61.1</td>
<td>116</td>
<td>8,584</td>
<td>116</td>
<td>8,584</td>
</tr>
<tr>
<td>172.24.2.1</td>
<td>172.24.50.1</td>
<td>2</td>
<td>771</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>172.24.10.1</td>
<td>172.24.50.1</td>
<td>5,644</td>
<td>340k</td>
<td>2,852</td>
<td>174k</td>
</tr>
<tr>
<td>172.24.20.1</td>
<td>172.24.255.255</td>
<td>5</td>
<td>771</td>
<td>5</td>
<td>771</td>
</tr>
<tr>
<td>172.24.50.1</td>
<td>172.24.50.2</td>
<td>5,324</td>
<td>340k</td>
<td>2,662</td>
<td>171k</td>
</tr>
<tr>
<td>172.24.30.1</td>
<td>172.50.42.1</td>
<td>44</td>
<td>3256</td>
<td>22</td>
<td>1,628</td>
</tr>
<tr>
<td>172.24.61.1</td>
<td>172.24.255.255</td>
<td>19</td>
<td>2,271</td>
<td>19</td>
<td>2,271</td>
</tr>
<tr>
<td>172.24.61.2</td>
<td>172.24.255.255</td>
<td>19</td>
<td>2,271</td>
<td>19</td>
<td>2,271</td>
</tr>
<tr>
<td>100.26.63.1</td>
<td>172.24.255.255</td>
<td>8</td>
<td>1,045</td>
<td>8</td>
<td>1,045</td>
</tr>
</tbody>
</table>

Figure 4. Process of analyzing packets.

When the system detects an abnormality, an alert is triggered to print the alert information on the screen, play the alert music and pop up the alert window (Fig. 5).

The packet analysis system can analyze the rail network packet automatically when uploading the network flow files in real time and the classification is detailed. The format of the output is easy to recognize. When there is an abnormality, an alarm notification can be performed in various ways so that abnormal conditions can be processed in real time to ensure network security.
Figure 5. Alert is triggered.

The disadvantage of the packet system is that it can only perform automatic analysis for the anomaly detection, but require human to identify the abnormality according to the normal network features at the current design. In our future work, more advanced anomaly detection methods will be designed and developed so as to identify the abnormality automatically.

VI. CONCLUSION

Abnormality detection is an essential step of building secure network. In this paper, rail network data are collected and analyzed. Two methods of the abnormality detection are presented. From data analysis, the main features of the rail network can be obtained, and a trust list can be generated. With these characteristics and the information concluded from the Wireshark detecting, the abnormality can be easily found. There is no requirement to perform complex operations when receiving packets, but it does require the operators to click on the screen and analyze the data one by one. There are also some time delays. To solve the above mentioned problems with Wireshark, a packet analysis system is designed to analyze the data for anomaly detection. It can receive the network flow and perform automatic analysis. After detecting abnormality, an alert is triggered to inform the person in charge to resolve the problem.

ACKNOWLEDGMENT

This work is supported by the National Research Foundation (NRF), Prime Minister’s Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-31) and administered by the National Cybersecurity R&D Directorate. Special thanks are also given to SMRT Trains Ltd to provide domain knowledge and technical support.

REFERENCES