Malicious node detection in wireless sensor networks using time series analysis on node reputation

Ouyang Xi, Li Dan, Zhang Jianyi, Liu Hongwei, Zhu Hongliang, Xin Yang,

Information Security Center, Beijing University of Posts and Telecommunications, ouyangxi@bupt.edu.cn
Information Security Center, Beijing University of Posts and Telecommunications, bylolo@bupt.edu.cn, nese@bupt.edu.cn
61741troops, liuhongwei198171@163.com
Beijing University of Posts and Telecommunications, Beijing Safe-Code Technology Co., Ltd, Zhuhongliang@bupt.edu.cn, yangxin@bupt.edu.cn

Abstract

Because of wireless sensor networks are usually deployed in hostile environment without continuous supervision, node compromise is the most critical threats of security issues of wireless sensor networks. In this paper, we proposed a method to identify malicious nodes in wireless sensor networks which have deployed with a reputation system through using time series analysis on node reputation. We focus a tricky type of attack nodes in reputation sensor network. These nodes can launch attacks or abnormal behaviors abstemious against the detection of reputation system. We give a detail of definition of these nodes in the paper and call them sub-aggressiveness malicious nodes. Our scheme to identify these nodes combined time series analysis with k-means clustering. With a series simulation, the results reveal that the proposed scheme has good performance and effectiveness to identify sub-aggressiveness malicious nodes which are hardly to be detected by reputation threshold mechanism of reputation network.

Keywords: Wireless Sensor Network, Reputation, Time Series Analysis, Malicious Nodes Detection

1. Introduction

A wireless sensor network consists of numerous compact and automated devices which are highly constrained in resources and called sensor nodes. But wireless sensor networks are very vulnerable to malicious attacks due to their characteristics, such as harsh deployment environments, sensor nodes are easy to be compromised, open wireless communication tunnels and so on.

Node compromise is one of the most detrimental attacks to a wireless sensor network [1]. When a WSN deployed in a hostile environment and without continuous supervision, sensor nodes of the network are easy to be capture by an attacker. And many secrets in security protocol of the WSN will be exposed and be extracted by adversary. By using the compromised node, the attacker can launch many other malicious attacks against WSNs. So, it is necessary to identify and isolate the malicious nodes in order to guarantee the security of a wireless sensor network. Deploy a Reputation system [2] is an effective way to identify inner malicious nodes through count the abnormal behavior of nodes and evaluates node’s reputation.

In this work, we focus a tricky type of malicious nodes. These nodes launch attacks or abnormal behaviors abstemious that help them keep their reputation above the threshold of reputation system and avoid to be identified as malicious nodes. Attacker employs numerous malicious nodes of this type can also cause considerable damage of network communication and without detection. In order to describe this type of malicious nodes conveniently, we defined their feature in attack behavior as Sub-aggressiveness and named them as Sub-aggressiveness nodes.

In this paper, we propose a method of malicious node detection bases on the condition that the network has deployed a reputation system and the reputation of every node can be acquired. We employ time series analysis (TSA) [3] into malicious nodes detection. We extract the deep-features of sensor nodes according the time series of node’s reputation and use the method of clustering to distinguish the malicious nodes from normal sensor nodes.
The organization of the rest of this paper is following. In section 2, we introduce the existing approaches of malicious node detection briefly. In section 3, we present a detailed description of our proposed method. The experiment setup and simulation results are given in section 4. And section 5 is the conclusion.

2. Related Work

With the further development of wireless sensor network applications, security issues of WSNs are facing many threats and challenges [1]. The biggest threat is vulnerable to be under attacks from malicious nodes due to characteristics of wireless sensor networks. Thus identify malicious nodes from network is an important part to ensure the security of network.

Sliva and Martins proposed a statistics-based scheme for malicious node detection of wireless sensor networks in [4]. In such a paper, authors give a series predefined regulations to describe the normal behaviors and further judge the anomaly behaviors of nodes. According to those regulations, preventive mechanisms were established and can be applied to protect WSNs against attacks from inner malicious nodes. But because lacked consideration of interactions among nodes, this scheme caused high rate of false alarming.

Bo Sun and Kii Wi in [5] proposed an anomaly detection algorithm based on Markov Chain to identify malicious nodes in MANET. In this paper, authors focus on the protection of MANET routing protocols and present details regarding feature selection, data collection, data preprocess, Markov Chain construction, classifier construction and parameter tuning. But the proposed algorithm only aims at detecting a single type attack that is route spoofing and identifying the malicious nodes where the attacks launch from. And large amounts of status information about routing table need to be preserved that render the algorithm unsuitable for the resource-constrained wireless sensor networks.

W. Junior and T. Figueriredo etc. in [6] proposed a method to detect malicious nodes using signal strength of datagram. In such an idea, detection of anomaly behaviors depends on neighborhood monitoring of the nodes. If there is an abnormal of signal strength of transmitting node is not accord with the originator node’s geographical position, an alarm may be launched by monitoring nodes. But, there is a large overhead required for data transmitting. It may cause inefficient in utilization of energy since all nodes are monitoring and processing data all the time.

ACK based anomaly detection algorithm is proposed by D.Tian and N.D.Georganas in [7]. In this algorithm, a next-hop ACK feedback technology is used to identify the unreliable communication links.

W. Du, L. Fang, and P. Ning proposed a solution for locationized anomaly detection in a group of nodes in [8]. In this method, node gets the localization information from neighboring nodes and computes the localization information itself and compares these two values. If the results show a small difference, that may indicates there is no adversary around causing the localization problem in its location.

A rule-based malicious node detection scheme in Ad Hoc is proposed by Chin-Yang Tseng etc. in [9].This scheme uses the monitoring points distributing in the network to monitor nodes whether operate in accordance with the routing norms in the process of AODV route query phase, then a finite state machine formed by the norms is used to identify nodes as normal state, suspected state, and intrusion state.

Saurabh Ganeriwal and Mani B. Srivastava proposed a reputation evaluation model for the resource-constrained wireless sensor networks BRSN [10]. In BRSN, every node will be evaluated their reputation according to their behaviors and malicious nodes can be identified by the threshold mechanism which has a low grade of reputation. But, this solution of detecting a malicious node in wireless sensor networks cannot identify tricky malicious node which have self-awareness to attack abstemious to keep itself above the threshold of reputation. Aims to solute this problem, we proposed a scheme to detect malicious nodes in wireless sensor networks using time series analysis on node’s reputation. The details of our scheme are present in the rest of the paper.
3. Approach design

3.1. Attack model and Sub-aggressiveness

There is an even trickier attack scenario in a wireless sensor network. Multiple inner malicious nodes collaborate to launch attacks against the network and with clear assignment of responsibility. In this case, there are two type nodes: attack nodes and observe nodes. Observe nodes are responsible for monitor the status of security mechanism of the network and send reminder to attack nodes. Attack nodes acquire some reminder from observe nodes to make their attacks be more subtle. In a WSN deployed with reputation system, malicious nodes are identified by a threshold of reputation. But, the attack nodes mentioned in that scenario can launch their attacks abstemious and keep their reputation above the threshold. So, we defined that characteristic of attacks as sub-aggressiveness, those attack nodes are sub-aggressiveness nodes. Our scheme is focus to how to identify those sub-aggressiveness nodes. Figure 1 depicts the sub-aggressiveness nodes in a reputation system network.

![Sub-aggressiveness nodes in a reputation network](image)

**Figure 1.** Sub-aggressiveness nodes in a reputation network

3.2. Time series analysis and related definitions

Time series is a sequence of data points, measured typically at successive time instants spaced at uniform time intervals. A common time series \( X \) is denoted as following:

\[
X = \{x_1, x_2, \cdots, x_m\}
\]

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

At recent, time series analysis has been widely applied to computer, network communication, signals processing, medical and financial sectors. Through the time series analysis, prediction and classification, we can solve a series of problem in the fields that mentioned above.

The main idea of our proposed scheme is that using time series analysis on reputation values to extract the characteristics of each node. According the characteristics, nodes will be clustered. Meanwhile, we set a standard time series to compare with time series of each node. Nodes in the
cluster that with maximum similarity of standard in time series characteristics will be identified as malicious nodes.

In order to present the proposed scheme well, we give some related definitions and notations firstly as follow.

**Definition 1. Reputation time series:** It is a sequence of node’s reputation value. There is a sensor node of a WSN denoted as K. At M consecutive sampling time points \( t_1, t_2, \ldots, t_m \), the reputation value of node K is denoted as \( R^k_n = R^k_{tn} \), where \( n = 1, 2, \ldots, m \). So we use symbol \( A^k_k \) to represent the reputation time series of node K, and it is denoted as:

\[
A^k_k = \{ R^k_1, R^k_2, \ldots, R^k_m \} \tag{2}
\]

**Definition 2. Standard sequence:** It is a benchmark data of assessing the similarity of a node’s reputation time series. It is a data sequence with the same length as each node’s reputation time series. And the value of elements in the sequence depends on the threshold of reputation system. If the length of time series is \( m \), and the threshold of reputation to decide a malicious node is \( \lambda \), then standard sequence is denoted as:

\[
\{ \lambda, \lambda, \lambda, \ldots, \lambda \} \tag{3}
\]

**Definition 3. Similarity of time series:** As for similarity evaluation of data, Gaussian radial basis function [11] is used, which can effectively reduce the influence of the measurement due to the noise and has a good performance on measure the similarity between two data. Using Gaussian radial basis function, the similarity of two data can be described as a distance \( d \). According to the function, the distance of two data A and B is denotes as follow.

\[
d = \exp \left( -\frac{(A - B)^2}{\sigma^2} \right) \tag{4}
\]

Where \( \sigma \) is corresponding width, and it is used to regulate the radial range of the function. In general, the value of it usually is five percent of the average of benchmark data.

Assume there are two time series \( X_i \) and \( Y_i \) denoted respectively as \( \{ X_{i1}, X_{i2}, \ldots, X_{in} \} \) and \( \{ Y_{i1}, Y_{i2}, \ldots, Y_{im} \} \). At the same time point \( t_i \), the similarity between data in \( X_i \) and \( Y_i \) can be assessed by the equation:

\[
s(t_i) = \exp \left( -\frac{(X_{ti} - Y_{ti})^2}{\sigma^2} \right) \tag{5}
\]

Following this logic, we can use the following function to assess the similarity of time series \( X_i \) and \( Y_i \):

\[
Sim(X_i, Y_i) = E[s(t_i)] = E \left[ \exp \left( -\frac{(X_{ti} - Y_{ti})^2}{\sigma^2} \right) \right] \tag{6}
\]
The value of $\text{Sim}(X_i, Y_i)$ is between 0 and 1. When the value of $\text{Sim}(X_i, Y_i)$ closer to 1, the higher similarity between $X_i$ and $Y_i$ is considered. Conversely, the value closer to 0 the lower similarity they have.

### 3.3. Procedure of approach

There are 4 steps of procedure in our proposed approach for malicious nodes detection.

**Step 1 Data preprocessing** : In a wireless sensor network deployed with reputation system, the reputation data of sensor nodes are the major part of our analysis. Suppose there are $n$ sensor nodes in the network except the nodes identified as malicious nodes by reputation system. We can get their reputation time series as objects of analysis through record their reputation value at every sampling time point. Then we can get a set of time series data denoted as $\{A_1, A_2, \cdots, A_k\}$.

**Step 2 Calculation of the similarity of reputation time series** : According to the threshold reputation value of reputation system for identify malicious node, a standard sequence $O$ is generated as a benchmark data to assess the similarity of reputation time series of every target node. Assume the threshold of reputation is $\lambda$ and the length of sampling time points is $M$, the standard sequence is denotes as $O = \{\lambda, \lambda, \cdots, \lambda\}$.

In this work, we employed the function (6) to calculate the similarity of reputation time series. As shown in Figure 2. The result output from the calculation part is a data set of similarity $M = \{S_1, S_2, \cdots, S_k\}$, where $S_k$ presents the similarity between $k$th time series and the standard sequence.

**Step 3 Clustering the elements in data set of similarity $M$** : In order to distinguish sub-aggressiveness nodes from normal sensor nodes, here we employed a method of clustering technology K-means [12]. After finished the step 2, a data set of similarity will be output. Each element of this set has a one-to-one relationship of every target node. By using K-means analysis we can cluster elements of data set into several groups upon prior setting.

Before the work of this step beginning, there is one point need to be confirmed that is set a value of $k$ of K-means analysis. The value of $k$ determines the result in how many groups would be output from K-means analysis. It is the key of whether output result in clustering meets the objective. In the following section, we set simulations in different value of $k$ and discuss the output results.
Step 4 Identify malicious nodes base on the results of clustering analysis: According to the outcome from K-means analysis, the cluster which center is most closest to 1 would be considered as a data set corresponding to sub-aggressiveness nodes. Base on this logic and follow the one-to-one relationship between data of similarity and target nodes, we can identify the malicious nodes in the set of target nodes.

The pseudo-code of our proposed approach is described in table 1.

### Table 1. Pseudo-code of reputation time series analysis

<table>
<thead>
<tr>
<th>Description of Algorithm</th>
<th>Reputation time series analysis based on clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INPUT</strong></td>
<td>( \text{ASS} ) (set of reputation time series ( { A_1, A_2, \cdots, A_k } )), standard sequence ( O ), value of ( K );</td>
</tr>
<tr>
<td><strong>OUTPUT</strong></td>
<td>( \text{ACS} ) (results of clustering)</td>
</tr>
<tr>
<td>InitProcess (( \text{ASS}, O )) ; //data initiate</td>
<td></td>
</tr>
<tr>
<td>For each ( A_i \in \text{ASS} ) Do ( ID_i \leftarrow \text{tag}(A_i) ) ; //tag corresponding ID to each time series</td>
<td></td>
</tr>
<tr>
<td>End For</td>
<td></td>
</tr>
<tr>
<td>while ( ( \text{ASS} ) is not null )</td>
<td></td>
</tr>
<tr>
<td>{ ( A_i \leftarrow \text{next}(\text{ASS}) ) ; //take out a time series from ( \text{ASS} )</td>
<td></td>
</tr>
<tr>
<td>For each ( R_i \in A_i ) Do ( s_i \leftarrow \text{GrF}(R_i, O) ) ; //assess the similarity of data</td>
<td></td>
</tr>
<tr>
<td>( S_i \leftarrow E(\sum s_i) ) ; //calculate the similarity of time series ( A_i )</td>
<td></td>
</tr>
<tr>
<td>End For</td>
<td></td>
</tr>
<tr>
<td>( M \leftarrow \text{collect}(S_i) ) ;</td>
<td></td>
</tr>
<tr>
<td>{ //output the data set of similarity ( M )</td>
<td></td>
</tr>
<tr>
<td>Do K-means (( M )) ; //clustering the elements in data set ( M )</td>
<td></td>
</tr>
<tr>
<td>Output ( k ) clusters ; //output the result of clustering</td>
<td></td>
</tr>
</tbody>
</table>

### 4. Simulation and analysis

In this section, we deploy a simulated environment for the proposed scheme to test and verify the effectiveness. And the simulation is consists of two stages. In the first stage, we simulated a reputation wireless sensor network as mentioned in [13], and we supposed there are several sub-aggressiveness malicious nodes were tagged previously. After the first stage, all reputation time series of sensor nodes will be produced. Then, using these data we start the second stage of the simulation which is an identification of malicious nodes according the proposed scheme.

In this simulation, we focus two performance indexes in different simulated network condition, rate of discrimination and rate of false discrimination in identification of malicious nodes. The computational methods of indexes are denoted as following:

\[
\text{Rate of discrimination} = \frac{N_c}{N_p}
\]

\[
\text{Rate of false discrimination} = \frac{N_{\text{total}} - N_c}{N_{\text{total}}}
\]

Here \( N_p \) denotes the actual number of sub-aggressiveness malicious nodes assumed in the simulated wireless sensor network, \( N_c \) denotes the number of correct identification of malicious nodes, and \( N_{\text{total}} \) represents the total number of nodes which are identified as malicious nodes by our scheme.

In order to test the effectiveness of the scheme with different condition, we set several adjustable parameters with simulated network: network nodes number \( n \), sub-aggressiveness malicious nodes number \( p \), the length of sequence \( m \) and the \( k \) value of k-means analysis. As shown in the Figure 3, curves represent the discrimination rate of our proposed scheme on identifying the sub-aggressiveness...
malicious nodes which tagged and deployed previously in the simulated network. And in Figure 4, curves represent the rate of false discrimination that identifying normal sensor nodes as malicious nodes. Here k=2, 3, 4 respectively, network nodes number n and malicious nodes number remain unchanged. As shown in these two figures, the proposed scheme achieve good results in identify sub-aggressiveness malicious with high discrimination rate and low false discrimination rate. Specifically, the poorest discrimination rate also above sixty percent and it is raising along with the length of sequence m increases. The worst false discrimination rate is lower than twenty-five percent and reduces with the length of sequence m increases. In our supposed situation of wireless sensor network, where the situation that k value of k-means analysis equal 2 can obtains the best performance in identification.
We adjust the number of malicious nodes \( p \) and keep the network nodes number \( n \) fixed to simulation. Figure 5 shows the comparison of malicious nodes discrimination rate in the different \( k \) value, here the value of \( k \) also respectively equals 2, 3, and 4.

\[ \text{Figure 5. Discrimination rate vs. malicious node number} \]

5. Conclusion

In this paper, we proposed a scheme for identify the malicious nodes from a reputation wireless sensor network. We focus a trickier attack scenario in a wireless sensor network which deployed a reputation system, and define a type of attack nodes which able to against identification with a threshold of reputation through collaborates with each other. We call this type of nodes as sub-aggressiveness malicious nodes, and using a method based on reputation time series analysis to identify them. This is an effective complement to detect malicious nodes in reputation network. The simulation results it has a good performance and effectiveness in identify sub-aggressiveness malicious nodes which are hardly to be detected by reputation threshold mechanism of reputation network.

6. Acknowledgement

This work is supported by the Fundamental Research Funds for the Central Universities (2012RC0215, 2012RC0216), National S&T Major Program (2011ZX03002-005-01) and National Natural Science Foundation of China (No. 61121061).

7. References


