Compound Attack Prediction Method Based on Improved Algorithm of Hidden Markov Model

Dongmei Zhao[1,2], Hongbin Wang[1,*] and Shixun Geng[1]

[1]*College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang, China*
[2]*Hebei Key Laboratory of Network and Information Security, Shijiazhuang, China*
E-mail: yunpeng@hebtu.edu.cn
[*]*Corresponding Author*

Received 16 September 2020; Accepted 27 October 2020;
Publication 26 December 2020

Abstract

Network attacks are developing in the direction of concealment, complexity, multi-step, etc., making it difficult to identify and predict. In order to solve the problems such as the difficulty of determining the matching degree of the network attack, the difficulty of predicting the attack intention, and the incorrect calculation of the alarm intent sequence due to the incorrect alarm information, a hidden Markov model based on improved algorithm composite attack prediction is proposed. Firstly, in order to improve the learning ability and adaptability of the algorithm, an improved Baum-Welch algorithm is proposed to train the hidden Markov model (HMM) and generate new HMMs. Then use the Forward algorithm to calculate the HMM with the maximum probability of generating a pre-processed alarm message sequence. When the alarm message sequence is misreported, the attack intent sequence obtained by the classic Viterbi algorithm may be biased. This paper improves the Viterbi algorithm to make the extracted attack intention sequence more accurate. Finally, simulation results show that the model can effectively

Journal of Web Engineering, Vol. 19_7–8, 1213–1238.
doi: 10.13052/jwe1540-9589.197813

extract attack intention sequence and improve the accuracy of compound attack prediction.

Keywords: Network security, hidden Markov model, compound attack prediction, attack intention, Baum-Welch algorithm, forward algorithm, Viterbi algorithm.

1 Introduction

With the development of science and technology, computer networks have been widely used, bringing great convenience to production and life, but the network is a double-edged sword. While people enjoy its convenience and speed, the network threats they are facing are also increasing. Because composite attacks are difficult to find and defend, this type of attack has continued to emerge in recent years, and has gradually become the mainstream attack method for network attacks. This makes passive defense technologies that only focus on unilateral detection can no longer meet the current network security needs [1], an active defense technology that can automatically adjust in time according to the changes and development of attacks to achieve the purpose of identification and prediction seems particularly important. As a kind of active defense technology, the compound attack prediction method has become a hot and difficult point in research today.

In 1999, Huang M-Y et al. Of the United States proposed that: we can extract the intention in the attack behavior first, and then predict the compound attack based on Intention Modeling [2]. After that, experiments show that this method can predict but the effect is not very good [3]. The research of "compound attack prediction method" started relatively late. In 2005, Bao Xuhua and others put forward a compound attack prediction algorithm based on attack intention, which uses the extensible digraph to represent the attack category and its logical relationship, associates the alarm information in two ways of backward matching and missing item matching, and calculates the possibility according to the accumulated weight of the generated attack chain and the weight of each branch in the attack logic graph, which can effectively reduce false positives and false negatives and improve the prediction accuracy, but the selection of thresholds depends on experience [4]. In 2006, Yan Fen et al. Improved the traditional Petri net description of composite attack methods. Based on CTPN, the compound attack scenario was modeled, and the model was used to correlate alarms. This method is simpler and more practical than the traditional method, but

the threshold value of Selection also relies on expert knowledge [5]. In 2011, Chen Can et al. Proposed a composite attack prediction method based on attack utility. The attack intention is used to describe the composite attack process, a composite attack logic diagram based on attack intention is established, and the concept of attack utility is introduced to improve prediction accuracy. However, the setting of certain parameters of the attack utility lacks certain standards [6]. In 2017, Pilar Holgado and others proposed a new intrusion detection early warning method based on Hidden Markov model. In this method, the hidden state is regarded as a similar stage of a specific attack, and then according to the state probability of each step of the current multi-step attack, Viterbi and forward algorithm are used to calculate the best state sequence to achieve real-time prediction of composite attacks [7]. In 2018, Wang Hui and others proposed an intrusion prediction algorithm based on improved attack graph. Firstly, Bayesian inference was used to calculate the probability of single-step attack, and then the designed association relationship quantification algorithm was used to associate single-step attack, so as to realize the dynamic prediction of the potential intrusion intention in the network [8]. In 2019, Ju Ankang et al. Proposed a composite attack detection method based on network communication anomaly recognition, using graph-based anomaly analysis and wavelet analysis to identify abnormal behaviors during network communication, and associated abnormal behaviors to detect multiple steps. Attack, but the time complexity of this method is relatively high [9]. In 2020, Holgado P et al. Proposed an intrusion detection and early warning method based on Hidden Markov model. The method trained the observation results offline, then trained the model through supervised and unsupervised methods, and calculated the probability matrix. Finally, according to the state probability of each attack, the prediction model uses Viterbi and forward backward algorithm to calculate the best state sequence to achieve prediction [21]. In 2020, Mohammad Samar Ansari et al. Proposed a prediction method based on deep learning technology. This method gives a shallow learning intrusion detection method by estimating the repeated attack probability of malicious sources in a given future time interval, points out the limitations of shallow technology in prediction, and then proposes a deep method based on recurrent neural network which is more suitable for alarm prediction task [22]. In 2020, Liu Kun et al. Proposed a network attack profit graph model and attack profit path prediction algorithm. This method proposes an attack profit matrix generation algorithm to generate attack profit matrix, and then proposes a path profit feasibility analysis method, which proves the feasibility of network attack profit analysis. Finally, the selection

and prediction method is designed to realize the attack path prediction [23]. After research, it is found that the research on compound attack prediction methods is summarized into the following four categories [10]:

(1) Compound attack prediction method based on antecedents and consequences: This prediction method is based on the detected attack behavior and the relationship between the attack and the successor to predict the attack steps that the attacker will continue to perform at subsequent times. Due to the variety of attack behaviors and their varied forms, it is difficult to implement a composite attack prediction method based on antecedents and consequences; (2) Compound attack early warning method based on CTPN: Petri net is a network information flow model driven by events according to rules, which can effectively describe the series, parallel and concurrent relationships between attack behaviors. It has not only strict mathematical expression, but also intuitive graphic expression method, which can provide rich system analysis and intuitive system description, and then effectively express the transformation process of attack intention and predict the attack behaviour [11]. Based on the theory of traditional Petri net, Coloured and Time Based Petri net represents the transformation process of attack intention and attack intention as two kinds of nodes: repository and transition. By adding a kind of "threshold" to model the attack scenario, the correlation of alarm information is analyzed, the attack scene graph is established, and the prediction result is deduced. The problem with this method is that prediction is only an addition, and its real intention is to identify, not predict. Therefore, the combined attack warning method based on ctpn can only predict part of the attack scenario, but can not achieve the effect of accurate prediction; (3) Composite attack prediction method based on Bayes offense and defense: This method uses the Bayesian network to map out the causal relationship between the alarm information, and uses Bayes rule to constantly modify the probability value on the attack behavior node, but this method is only suitable for one-to-one attack and defense mode, which obviously has certain limitations; (4) Compound attack prediction method based on attack intention [6]: This method models attack intention as a key factor, which is mainly based on the compound attack logic relationship of attack intention [3], but the parameter setting involved in this method is not standardized, which will have a certain impact on the prediction results, and it is difficult to match the compound attack.

In order to predict compound attacks more effectively, this paper proposes a compound attack prediction method based on improved algorithm of hidden

Markov model. This method first trains HMMs through the improved Baum-Welch algorithm based on the pre-processed alarm message sequence, using all the trained HMMs as a model library for compound attack discrimination, and then using the Forward algorithm to determine the probability value of generating the alarm information sequence in each model, and the hidden Markov model with the highest probability is the most likely compound attack. Then, the improved Viterbi algorithm with strong anti-noise is used to obtain the hidden best attack intention sequence, so as to predict the next most likely attack by the attacker. Finally, simulation experiments verify the effectiveness of the algorithm.

2 Hidden Markov Model

2.1 Hidden Markov Model Related Concepts

Hidden Markov Model (HMM) is a probabilistic statistical model about time series. It generates a non-observable random state sequence (state sequence) from the hidden Markov chain. Each state generates an observation to generate an observable random state sequence (observation sequence), and each state position of the sequence can be regarded as a moment.

Hidden Markov model is a tuple composed of 5 items, namely $\lambda = (Q, V, A, B, \pi)$: Q is the finite set of HMM hidden states, V is the finite set of HMM observations, Q and V are $Q = \{q_1, q_2, q_3, \ldots q_n\}$, $V = \{v_1, v_2, v_3, \ldots v_M\}$, where N is the total number of all possible states and M is the total number of all possible observations. State q is invisible and state v is visible; A is the state transition probability matrix, $A = [a_{ij}, i, j = 1, 2, 3, \ldots N]_{N*N}$, where $a_{ij} = P(i_{t+1} = q_j | i_t = q_i)$ is the probability of state q_i at time t transferring to state q_j at time $t + 1$; B is the observation probability matrix, $B = [b_j(k), j = 1, 2, 3, \ldots N, k = 1, 2, 3, \ldots M]_{M*N}$, where $b_j(k) = P(o_t = v_k | i_t = q_j)$ represents the probability that time t is in state qj and generates observation v_k; π is the initial state probability vector, $\pi = (\pi_i)$, where $\pi_i = P(i_1 = q_j), i = 1, 2, 3 \ldots N$ represents the probability of being in state q_j at time $t = 1$, Since the state sequence Q is determined by π and A, and the observation sequence V is determined by B, the hidden Markov model is generally expressed as $\lambda = (A, B, \pi)$.

2.2 Hidden Markov Model Example

Take a casino dice game as an example. In order to increase the probability of winning, the casino often resorts to cheating. For example, the dice game

Table 1 Differences between normal dice and lead-filled dice

	Normal Dice a	Lead-filled Dice b
Probability of throwing 1 points	1/6	0
Probability of throwing 2 points	1/6	1/8
Probability of throwing 3 points	1/6	1/8
Probability of throwing 4 points	1/6	3/16
Probability of throwing 5 points	1/6	3/16
Probability of throwing 6 points	1/6	3/8

is based on the number of points rolled to determine the outcome. The casino will use the modified dice to increase the chance of winning. In the process of rolling the dice multiple times in a row, the normal dice is used in most cases. During the period, the lead-filled dice are randomly mixed to increase the probability of the desired number of points being rolled. The rule is to use normal dice first. If the normal dice was just used, then the probability of the next dice being lead dice is 10%; if the lead dice was just used, then the probability of the next dice being normal dice is 80%. The difference between normal dice and lead-filled dice is shown in Table 1.

In the above example, there are two random sequences, one is the invisible dice type sequence (state sequence), and the other is the visible roll point sequence (observation sequence). The three elements of the hidden Markov model can be determined according to the rules.

The dice corresponds to the state, then the state set $Q = \{\text{normal a}, \text{lead b}\}$, where $N = 2$; the number of points rolled corresponds to observation, then the observation set $V = \{1, 2, 3, 4, 5, 6\}$, where $M = 6$.

The initial state probability is: $\pi_1 = 1, \pi_2 = 0$.

The state transition probability matrix A is: $\begin{bmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{bmatrix}$.

The observation probability matrix B is:

$$\begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 1/8 & 1/8 & 3/16 & 3/16 & 3/8 \end{bmatrix}.$$

Looking at the three basic problems in the HMM from the above examples [12]:

1. Evaluation problem: Given the known observation sequence O and HMM $\lambda = (A, B, \pi)$, determine the probability that O is generated by λ (i.e. the probability that the sequence of points thrown is thrown by the cheating model).

Table 2 Correspondence between problems and algorithms

Problems	Solving Algorithms
Evaluation problem	Forward algorithm
Decoding problem	Viterbi algorithm
Learning problem	Baum-Welch algorithm

2. Decoding problem: Under the condition that the observation sequence O and HMM $\lambda = (A, B, \pi)$ are known, calculate the state sequence corresponding to the observation sequence O (that is, in the sequence of points to be rolled, which points are to be rolled by the lead dice).
3. Learning problem: Under the condition that the observation sequence O is known, calculate the HMM $\lambda = (A, B, \pi)$ that can produce the observation sequence with the maximum probability (that is to say, the cheating model is calculated from the sequence samples of the number of points thrown).

The solutions of these three problems are the three algorithms in HMM, and the corresponding relationship is shown in Table 2.

2.3 Hidden Markov Model Algorithms

Forward Algorithm:
Definition 1 Forward variable: In HMM λ, at time t, all observation sequences before time t are obtained, and the state at time t is Si. The probability of this event is recorded as $\alpha_t(i) = P(o_1, o_2, o_3, \ldots o_t, q_t = S_i|\lambda)$, then

- $\alpha_1(i) = P(o_1, q_1 = S_i|\lambda) = \pi_i \cdot b_i(o_1)$
- $\alpha_t(i) = P(o_1, o_2, o_3, \ldots o_t, q_t = S_i|\lambda)$
- $P(O|\lambda) = \sum_i^N \alpha_t(i)$

The algorithm steps are as follows:

1. Initialization:
$$\alpha_1(i) = \pi_i \cdot b_i(o_1) \tag{1}$$
2. Recursion:
$$\alpha_{t+1}(j) = \left[\sum_i^N \alpha_t(i)\alpha_{ij}\right] b_j(o_{t+1}) \tag{2}$$
3. Termination:
$$P(O|\lambda) = \sum_i^N \alpha_t(i) \tag{3}$$

Viterbi Algorithm:
The Viterbi algorithm is also a grid structure similar to the Forward algorithm, which is used to solve the second decoding problem. Under the condition that the observation sequence O and HMM $\lambda = (A, B, \pi)$ are known, the state sequence Q corresponding to the observation sequence O is calculated. The calculated Q should be "optimal" under a certain criterion , So Q is also called the optimal path.

Definition 2 $\delta_t(i)$ is the maximum probability that the state sequence at time t is $\{q_1, q_2, q_3, \ldots q_t\}$ and $q_t = S_i$ produces the observation sequence $O = \{o_1, o_2, o_3, \ldots o_t\}$. $\delta_t(i) = \max_{q_1,\ldots q_{t-1}} P(q_1, q_2, q_3, \ldots q_{t-1}, q_t = S_i, o_1, o_2, o_3, \ldots o_t|\lambda)$.

- "Optimal": the probability is the highest

$$Q^* = \arg\max P(Q|O, \lambda) \tag{4}$$

- Viterbi variable:

$$\delta_t(i) = \max_{q_1,\ldots q_{t-1}} P(q_1, q_2, q_3, \ldots q_{t-1}, q_t = S_i, o_1, o_2, o_3, \ldots o_t|\lambda) \tag{5}$$

- Recursive relationship:

$$\delta_{t+1}(j) = [\max_{q_1,\ldots q_{t-1}} \delta_t(i) * \alpha_{ij}] * b_j(o_{t+1}) \tag{6}$$

- Memory variable: $\Psi_t(i)$ is the previous state of the current state of the optimal path.

The algorithm steps are as follows:

1. Initialization:

$$\delta_1(i) = \pi_i * b_i(o_1), \Psi_1(i) = 0, 1 \le i \le N \tag{7}$$

2. Recursion:

$$\delta_t(j) = [\max_1 \le i \le N \delta_{t-1}(i) * \alpha_{ij}] * b_j(o_t), 2 \le t \le T, 1 \le j \le N \tag{8}$$

$$\Psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) * \alpha_{ij}] \tag{9}$$

3. Termination:

$$P* = \max_{1 \le i \le N} [\delta_T(i)], q*_T = \arg\max_{1 \le i \le N} [\delta_T(i)] \tag{10}$$

4. Trace back:

$$q*_t = \Psi_{t+1}(q*_{t+1}), t = T-1, T-2, T-3, \ldots 1 \tag{11}$$

Baum-Welch Algorithm:
The Baum-Welch algorithm is used to solve the third learning problem. Under the condition that the observation sequence O is known, calculate the HMM $\lambda = (A, B, \pi)$ that can produce the observation sequence with the maximum probability, and satisfy some optimal criterion, that is, $P(O|\lambda)$ is the largest. The algorithm steps are as follows:

1. Initialization: Randomly assign $\pi_i, a_{ij}, b_j(k)$ (to satisfy the probability condition), get HMM λ_0, set $i = 0$;
2. EM steps:
   Step E: Calculate the expected values of $\xi_t(i,j)$ and $\gamma_t(i)$ according to formulas (12) and (13).
   $\xi_t(i,j)$ is the probability of being in state i at time t and being in state j at time $t+1$ under the conditions that the HMM and the observation sequence O are known [12]:

$$\begin{aligned}\xi_t(i,j) &= P(q_t = S_i, q_{t+1} = S_j|O, \lambda) \\ &= \frac{\alpha_t(i)\alpha_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\alpha_{ij}b_j(o_{t+1})\beta_{t+1}(j)}\end{aligned} \tag{12}$$

   $\gamma_t(i)$ is the probability of being in state i at time t:

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{13}$$

   Step M: According to the $\xi_t(i,j)$ and $\gamma_t(i)$ calculated in step E, the formula (14), (15), (16) is used to re-estimate π_i, a_{ij} and $b_j(k)$ respectively to get the HMM λ_{i+1}.

$$\pi_i = \gamma_1(i) \tag{14}$$

$$\alpha_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{15}$$

$$b_j(k) = \frac{\sum_{t=1}^{T}\gamma_t(j)\delta(o_t, v_k)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{16}$$

Termination condition: Assign $i = i + 1$, and loop the EM step until the values of π_i, a_{ij} and $b_j(k)$ converge. The HMM at this time is the required model.

3 Improvement of HHM Modeling Algorithms

3.1 Improvement of Viterbi Algorithm

The classic Viterbi algorithm can infer the implicit attack intention sequence based on the alarm message sequence, but this method ignores an important problem, that is, the alarm message sequence still has false alarms after preprocessing. In actual situations, the alarm information generated by the intrusion detection system will have a certain deviation, and false alarms often occur. Once this happens, it may affect the accuracy of Viterbi algorithm, thus affecting the accuracy of acquiring attack intention sequence. In this paper, we improve the Viterbi algorithm to reduce the impact on the calculation of the attack intent sequence when the alarm information is misreported. A comparison factor is obtained by training data to determine whether there is false alarm in the alarm information. When the probability of the alarm information sequence in state S_j at time t is lower than the comparison factor, it indicates that there is likely to be false alarm in the alarm information. At the same time, discard the alarm information and continue to analyze the next alarm information. The specific calculation steps are as follows:

Known observation sequence $O = \{o_1, o_2, o_3, \ldots o_t\}$ and HMM $\lambda = (A, B, \pi)$, respectively select $1 - t$ correct alarm information sequences to calculate the minimum probability value to determine the contrast factor, which are $\{\min_j \delta_1(j), \min_j \delta_2(j), \min_j \delta_3(j), \min_j \delta_t(j)\}$, set to $C = \{c_1, c_2, c_3, \ldots c_t\}$.

1. Initialization:

$$\delta_1(i) = \pi_i * b_i(o_1), \Psi_1(i) = 0, 1 \le i \le N \tag{17}$$

2. Recursion:

$$\delta_t(j) = [\max_{1 \le i \le N} \delta_{t-1}(i) * \alpha_{ij}] * b_j(o_t), 2 \le t \le T, 1 \le j \le N \tag{18}$$

$$\Psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) * \alpha_{ij}] \tag{19}$$

$$\lambda_{\max} = \arg \max_{1 \le i \le N} [\delta_t(i)] \tag{20}$$

If $\lambda_{\max} < c_t$ and $t < T$, then

$$\delta_t(j) = [\max_{1 \le i \le N} \delta_{t-1}(i) * \alpha_{ij}] * b_j(o_{t+1}), 2 \le t \le T-1, 1 \le j \le N \tag{21}$$

$$\Psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) * \alpha_{ij}] \tag{22}$$

If $\lambda_{\max} < c_t$ and $t \ge T$, then

$$P^* = \max_{1 \le i \le N} [\delta_{T-1}(i)], q_T^* = \arg \max_{1 \le i \le N} [\delta_{T-1}(i)] \tag{23}$$

Otherwise go to step (3).

3. Termination:

$$P* = \max_{1 \le i \le N} [\delta_T(i)], q*_T = \arg \max_{1 \le i \le N} [\delta_T(i)] \tag{24}$$

4. Trace back:

$$q*_t = \Psi_{t+1}(q*_{t+1}), t = T-1, T-2, T-3, \ldots 1 \tag{25}$$

3.2 Improvement of Baum-Welch Algorithm

The Baum-Welch algorithm can effectively calculate the HMM corresponding to the observation sequence, but because the attack characteristics will vary greatly at different times, the training algorithm needs to have a strong adaptability and learning ability. For example, an HMM is trained based on the observation sequence from a previous period as a training data set, and the characteristics of the new observation sequence received recently must also be reflected in the HMM [13]. To this end, this paper has improved the Baum-Welch algorithm and modified the calculation method of the state transition probability distribution matrix.

Set the current time period as the nth time period, and then use the previous observation sequence of the $n-1$ time periods as part of the data, set as $O = \{o_1, o_2, o_3, \ldots o_{n-1}\}$, in order to highlight the characteristics that the closer the observation sequence to the current time has the greater influence on the calculated hidden Markov model, a forgetting factor is added to the formula. Then the state transition probability $a_{ij}(n)$ of the current time period is:

$$a_{ij}(n) = \frac{\sum_{t=0}^{n-1} 2^{-r(n-1-l)} trans - counts(i,j,l)}{\sum_{t=0}^{n-1} 2^{-r(n-1-l)} counts(i,l)} \tag{26}$$

Among them, $trans - counts(i, j, l)$ is the expected number of times of observation sequence O_t from state i to j, $counts(i, l)$ is the expected number of times of O_t in state i, in $2^{-r(n-1-l)}$, $r > 0$ is the forgetting factor, and the specific value is determined according to the actual situation.

3.3 Establishment of Hidden Markov Composite Attack Prediction Model Based on Improved Algorithms

The compound attack has its distinctive characteristics. Because it is composed of multiple attack steps, the ultimate attack target is the same, so there's a connection between the attack steps, each attack step depends on the previous attack step, and each alarm information can show its corresponding attack step [14]. The relationship between attack steps and alarm information can be well reflected in the hidden Markov model [15]. Using hidden Markov model to model compound attack can solve the problem that attack intention is difficult to identify.

The complete process of compound attack prediction based on the improved algorithm of hidden Markov model is: (1) Collect the alarm information sent by the intrusion detection system, and then preprocess the collected alarm information set to generate the alarm information sequence. According to the generated alarm information sequence, the improved Baum-Welch algorithm is used to train the hidden Markov model, and all the trained hidden Markov models are used as a model library for compound attack discrimination. Then according to the current preprocessing observation sequence and the Forward algorithm of the hidden Markov model, the hidden Markov model with the maximum probability to generate the preprocessing alarm information sequence is calculated. The hidden Markov model with the maximum probability is the most likely compound attack. And with the execution of the attack steps, the alarm message sequence will be longer and longer, and the probability value will be larger and larger. Therefore, as the attacker approaches the goal of the attack, the hidden Markov model of the compound attack approaches the compound attack being performed by the attacker. Finally, the pre-processed alarm information sequence is input into the hidden Markov model obtained through compound attack discrimination. The improved Viterbi algorithm in this paper is used to obtain the hidden best intent attack sequence to achieve the prediction purpose. The specific flowchart is shown in Figures 1–3.

In this paper, the observation probability matrix B of the hidden Markov model is established according to the calculation method provided in

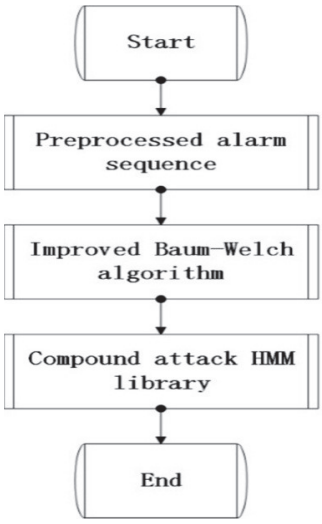


Figure 1 Flow chart of HMM establishment.

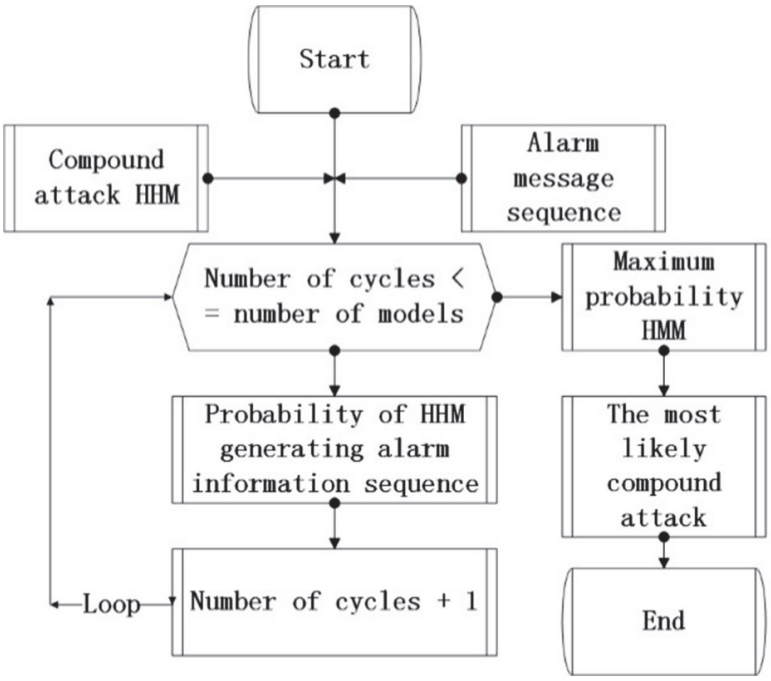


Figure 2 Flow chart for identifying a compound attack.

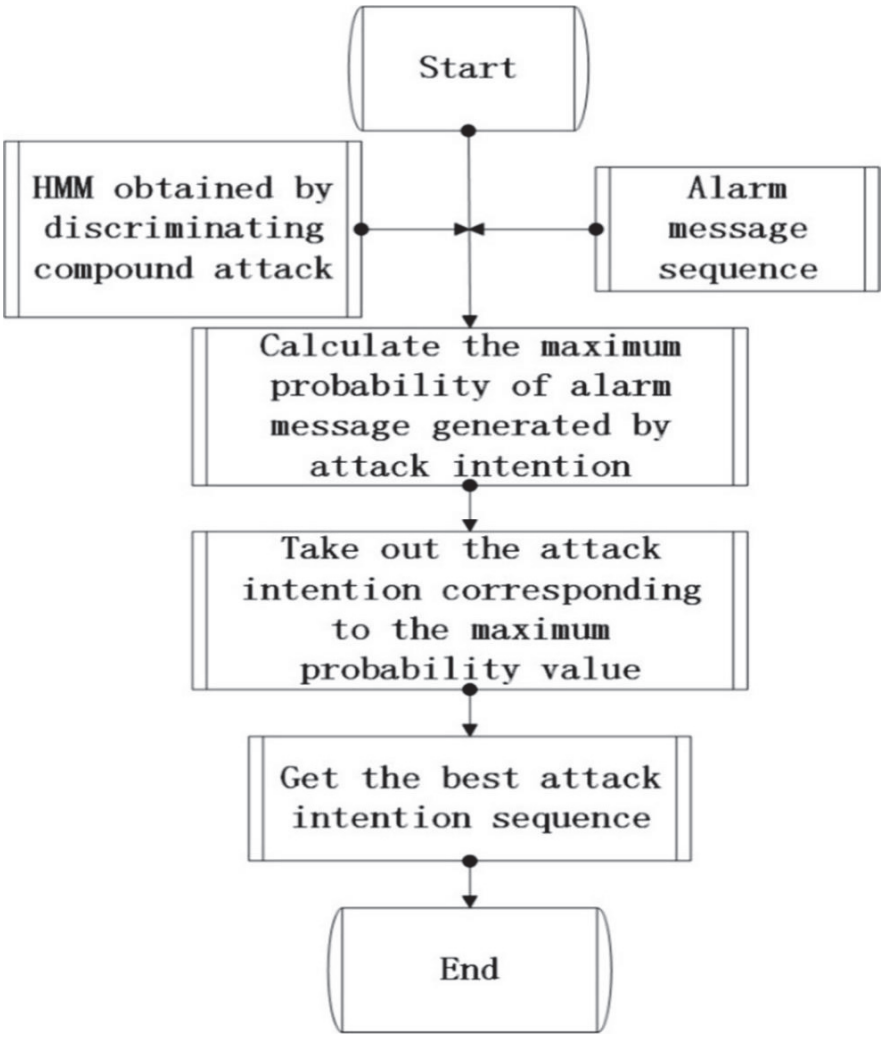


Figure 3 Flow chart of identifying attack intention and prediction.

reference [10], and the state transition probability a_{ij} is calculated according to the association rules proposed in reference, then the state transition probability matrix A in the hidden Markov model can be calculated.

The compound attack recognition and prediction model based on the improved algorithm of the hidden Markov model is shown in Figure 4. The attack intention sequence of the attacker is hidden and cannot be observed

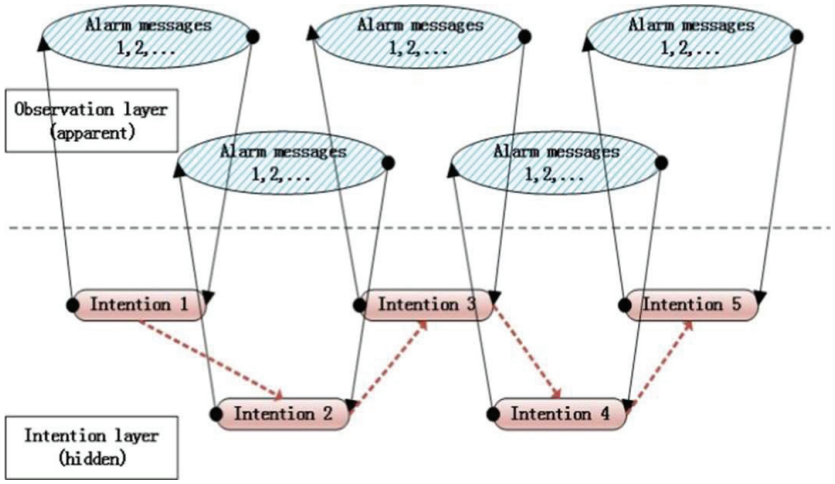


Figure 4 Compound attack recognition and prediction model based on improved algorithm of hidden Markov model.

directly, but only the alarm information sent by the intrusion detection system can be observed directly, so the invisible attack intention can only be judged according to the visible alarm information sequence. The black arrow in the figure indicates that the attack intention can generate alarm information, and the corresponding attack intention can be inferred based on the alarm information [16]. The red dotted arrow in the figure indicates the transfer of attack intention, and each transfer has a certain probability. In the compound attack, in order to achieve an attack intention, the attacker can take many different attack means, and the intrusion detection system will send many different alarm information to many different attack means, so the same attack intention may produce many different alarm information.

3.4 Establishment of Early Warning System

The hidden Markov model early warning system based on improved algorithm mainly includes two parts: intrusion detection system and early warning analyzer. As the name implies, intrusion detection system is used to detect whether there is intrusion, which is an indispensable part of early warning system. It mainly collects information through multiple sensor agents in the network that needs to be protected, and analyzes it to detect whether there is intrusion or abnormal behavior. Once the signs of intrusion are found, it will send out alarm information. The main function of the early-warning

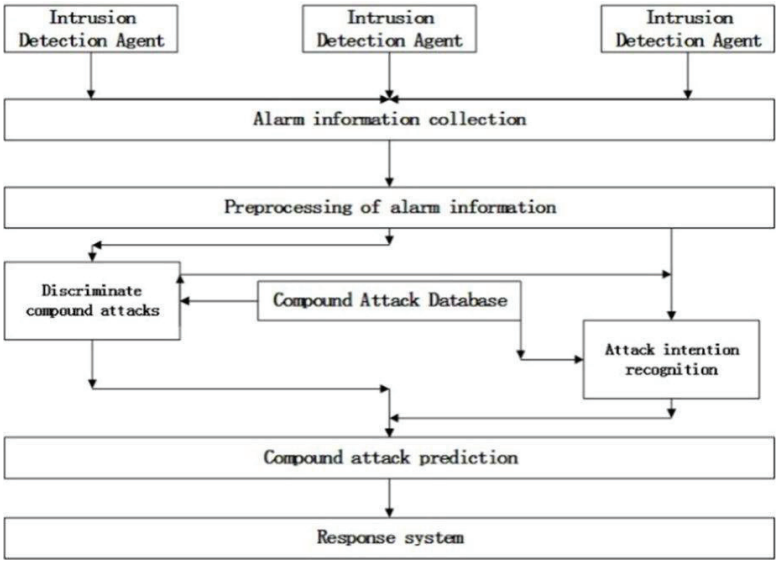


Figure 5 Early warning system architecture.

analyzer is to process and analyze the alarm information reported by the intrusion detection system, first remove false positives and duplicate alarm information, and analyze which type of composite attack is currently being received, identify the real attack intention of the attacker and predict the attack behavior that the attacker is most likely to take in the next step. The architecture of the early warning system is shown in Figure 5.

As can be seen from Figure 5, the early warning system architecture is divided into alarm information collection, pre-processing of alarm information, identification of attackers' attack intentions, identification of corresponding compound attacks, compound attack prediction, and compound attack databases.

1. Alarm information collection: It mainly collects the original alarm information sent by the intrusion detection system, which is the beginning of the entire system operation.
2. Alarm information preprocessing: because there are some false alarms and repeated alarms in the original alarm information sent by the intrusion detection system, it is necessary to preprocess the alarm information, remove the false alarm information, merge the repeated alarm information, and form a sequence of complete alarm information [17], providing data basis for the hidden Markov model.

3. Discrimination of compound attack: analyze the pre-processing alarm information sequence and find out the corresponding hidden Markov model of compound attack.
4. Recognition of attack intentions: The true attack intentions of the attackers are identified according to the alarm message sequence and the composite attack hidden Markov model that has been determined.
5. Compound attack prediction: The hidden Markov model is used to predict the compound attack to determine the most likely attack behavior of the attacker.
6. Compound attack database: a database in which the known compound attack behaviors are stored in the network.

4 Simulation Experiment and Analysis

4.1 Alarm Information Preprocessing

The error or repetition of the alarm information is generally caused by several different intrusion detection systems triggered by an attack intent in a composite attack or the same intrusion detection system is triggered multiple times simultaneously. For example, a port scanning attack often targets the port numbers of multiple hosts, or attacks on multiple hosts on different networks will trigger multiple different intrusion detection systems to issue multiple alarm messages. The alarm information pre-processing mechanism is mainly based on the redundant relationship between alarm messages to filter. The main pre-processing mechanism can be described as follows: first of all, delete the non alarm information. Every alarm message generated by the intrusion detection system has a MessageClassID. Judge whether the value of the MessageClassID of the alarm message is 0. If it is not, it is a non alarm message. Then delete the illegal alarm information, to determine the format of the alarm information, if the alarm information is lack of key parameters, it is illegal alarm information. Finally, according to D_S evidence theory fuses alarm information and generates alarm information sequence.

4.2 Experimental Results and Analysis

In this experiment, the attack test data set LLDOS1.0 (inside) provided by DARPA of the Massachusetts Institute of Technology's Lincoln Laboratory in 2000 was used, the DDOS compound attack data is extracted from it to verify the ability of the hidden Markov model to predict the compound attack and whether the prediction effect is improved after the improved method.

Table 3 Attack intention and corresponding alarm information

Attack Intention	Alarm Message Number	Alarm Message
IPSweep(State1)	Alert1	ICMP PING NMAP
	Alert2	ICMP Echo Reply
Sadmind Ping(State2)	Alert3	RPC portmap request sadmind UDP
	Alert4	RPC sadmind UDP PING
Sadmind Exploir(State3)	Alert5	NETMGT_PROC_SERVICE
Daemon Installed(State4)	Alert6	RPC sadmind query with root creden tials attempt UDP

Table 4 Initial probability distribution matrix

State1	State2	State3	State4
0.2	0.4	0.2	0.2

Table 5 State transition probability distribution matrix

	State1	State2	State3	State4
State1	0.5	0.5	0	0
State2	0	0.5	0.5	0
State3	0	0	0	1
State4	0	1	0	0

There are five phases in the LLDOS 1.0 DDOS compound attack scenario: (1) Scan the active host from a remote site through IPSweep; (2) Scan through the service port in the active IP to find the host with the Sadmind remote execution command vulnerability; (3) Through the Sadmind remote command execution vulnerability on the host, obtain the root permission; (4) Install DDoS attack Trojan software on the host with root authority; (5) Start DDoS and launch DDos attack on the target. The attack intention and corresponding alarm information of the hidden Markov model of the DDoS attack are shown in Table 3.

The initial probability distribution matrix π of DDoS-HMM is shown in Table 4.

The state transition probability distribution matrix A of DDoS-HMM is shown in Table 5.

The observation probability distribution matrix B of DDoS-HMM is shown in Table 6.

Table 6 Observation probability distribution matrix

	Alert1	Alert2	Alert3	Alert4	Alert5	Alert6
State1	0.4979	0.4979	0.0010	0.0010	0.0010	0.0010
State2	0.0010	0.0010	0.7479	0.2479	0.0010	0.0010
State3	0.0008	0.0008	0.0008	0.0008	0.9958	0.0008
State4	0.0008	0.0008	0.0008	0.0008	0.0008	0.9958

Table 7 Contrast factors

C1	C2	C3	C4
0.05	0.01	0.00005	0.00001

In the process of using the improved Viterbi algorithm to identify the attack intention, it is necessary to set a contrast factor to delete the false alarm information. After processing and analyzing the data, select 1–4 error free alarm information sequences to obtain the minimum probability value to determine the contrast factors $C = \{c_1, c_2, c_3, c_4\}$, among which the error free alarm information sequence must include all possible attack methods in the attack process The calculated contrast factor is shown in Table 7.

Before the complex attack is completed, the received alarm information sequence is {alert1, Alert2, alert3}. At this time, by using the improved Viterbi algorithm, the attack intention sequence can be identified as {state1, State2} [18], and the attacker's next attack intention can be predicted as state3, which shows that the composite attack prediction method based on the improved hidden Markov model is feasible.

When the alarm information sequence is correct, experiments show that the compound attack prediction method based on the hidden Markov model can accurately identify the attack intention sequence and predict the next attack intention. In the following, some false alarm information sequences are randomly introduced into the alarm information sequence to verify the accuracy of the Viterbi algorithm's recognition of attack intent before and after improvement. Firstly preprocess the introduced alarm information, and then use the optimized Baum-Welch algorithm to train a hidden Markov model for the DDoS attack, and then compare the effects before and after the improvement of the Viterbi algorithm. As shown in Figure 6.

It can be seen from Figure 6 that when the false alarm rate of the alarm information sequence is between 0% and 30%, there is a certain gap between the identification of attack intention before and after the improvement of Viterbi method. The accuracy of the identification of attack intention after the

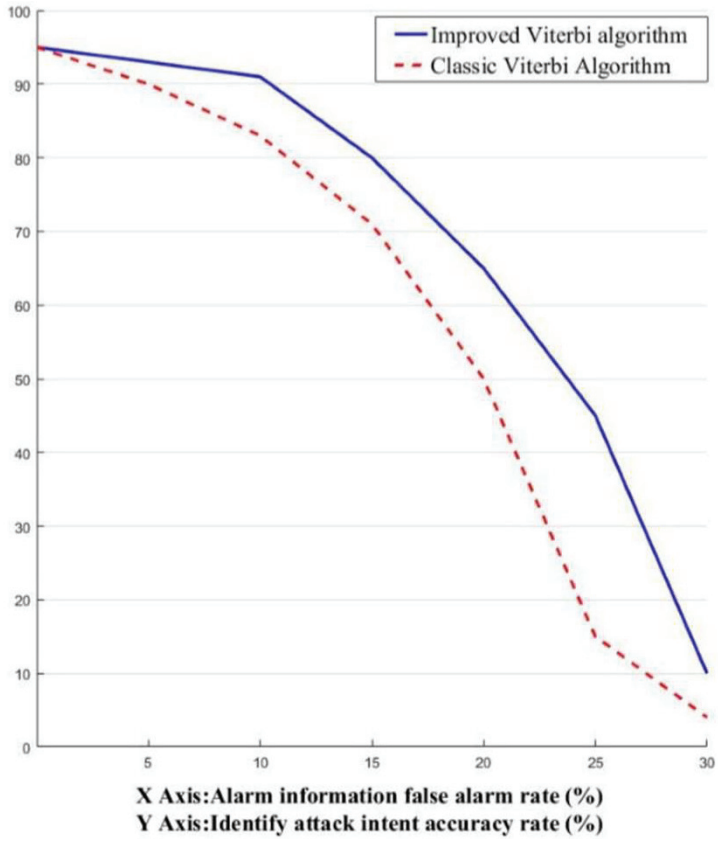


Figure 6 Comparison of attack accuracy recognition accuracy.

improvement of Viterbi method decreases slowly compared with that before the improvement of Viterbi method when the false alarm rate increases, so the prediction accuracy of composite attack is also higher, so it can be proved that the method proposed in this paper has better effect on compound attack prediction.

4.3 Comparison with Four Mainstream Methods of Compound Attack Prediction

Compared with the current four mainstream composite attack prediction methods, the performance evaluation of the method proposed in this paper is the best in the aspects of whether it needs prior knowledge, whether it can

Table 8 Performance comparison

Method of Prediction	Prior-knowledge	Connect Online	Multiple Information Fusion	Anticipating New Compound Attacks	Objectivity
Prediction method of composite attack based on cause and effect	√	√	×	√	√
CTPN-based composite attack early warning method	√	√	√	×	√
Bayes attack and defense-based compound attack prediction method	√	√	√	√	×
Composite attack prediction method based on attack intention	√	√	×	√	√
Method of this paper	×	√	√	√	√

carry out online association, whether it can achieve multi alarm information fusion, whether it can predict new composite attacks and objectivity, such as Table 8 shows.

5 Summary

In this paper, an improved Baum Welch algorithm is proposed to train hidden Markov model (HMM) and generate new HMMs, and then forward algorithm is used to calculate the HMM with maximum probability to generate pre-processed alarm message sequence, which improves the learning ability and generalization ability of the model. Aiming at the problem of false alarm in alarm information sequence, Viterbi algorithm is improved to make the extracted attack intention sequence more accurate.

Traditional intrusion detection technology is difficult to identify multi-step attacks [19]. In the context of compound attack becoming the mainstream attack method, the prediction method of compound attack based on the improved algorithm of hidden Markov model proposed in this paper can better solve the problem of predicting attack intention in compound attack

prediction research, and can effectively improve the training effect of HMM, so as to improve the accuracy of compound attack prediction. The method in this paper has certain reference significance in the research of compound attack prediction, but the accuracy of its prediction is affected by the attack intention library. The more comprehensive the attack intention database is, the higher the prediction accuracy of the method is, and the less comprehensive the attack intention base is, the lower the prediction accuracy is.

In order to better identify composite attacks and conduct risk assessment [20], we will study the following work:

1. Continue to extract attack intent by combining a large number of complex attack instances to make the attack intent of compound attacks more reasonable and complete.
2. In order to give full play to the ability of attack prediction, a more abundant hidden Markov model library for compound attacks is established.

Acknowledgements

The authors acknowledge the National Natural Science Foundation of China (Grant: 61672206),The Key Research and Development Program of Hebei (No.20310701D).

References

[1] Geng N (2015). Approach to Forecasting Multi-Step Attack Using Hidden Markov Model Based on Particle Swarm Optimization. Telecom Power Technologies, 000(003), 69–71.

[2] Langley, Pat, Simon, et al. (1995). Applications of machine learning and rule induction. Communications of the ACM, 38(11), 54–64.

[3] Ming-Yuh Huang, Robert J. Jasper, et al (1999). A large scale distributed intrusion detection framework based on attack strategy analysis. Computer Networks, 31(23/24), 2465–2475.

[4] Bao X H, Dai Y X , Feng P H, et al. (2005). A Detection and Forecast Algorithm for Multi-Step Attack Based on Intrusion Intention. Journal of Software, 16(12), 2132–2138.

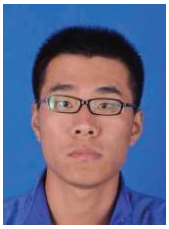
[5] Yan F, Huang H, et al. (2006). A Detection Algorithm for Multi-Step Attack Based on CTPN. Chinese Journal of Computers, 029(008), 1383–1391.

[6] Chen C, Yan B P, Li J. (2011). Forecast Algorithm for Multi-step Attack Based on Attack Utility. Microellectronics & Computer, 028(003), 81–84.
[7] Pilar Holgado, Victor A. Villagra, Luis Vazquez (2017). Real-time multistep attack prediction based on Hidden Markov Models. IEEE Transactions on Dependable and Secure Computing, (99), 1–1.
[8] Wang H, et al. (2018). Intrusion Prediction Algorithm Based on Correlation Attack Graph. Computer Engineering, 044(007), 131–138.
[9] Ju A K, Guo Y B, et al. (2019). Multi-step attack detection method based on network communication anomaly recognition. Journal on Communications, 040(007), 57–66.
[10] Zhang Y X, Zhao D M, Liu J X. (2013). Approach to Forecasting Multi-step Attack Based on Fuzzy Hidden Markov Model. Journal of Applied sciences, 13(22), 955–4960.
[11] Li C Y, Qi Y D, Wang X H, et al. (2019). DDoS Attack and Defense Confrontation Evaluation Based on Attack and Defense Game and Stochastic Petri Net. Computer Systems & Applications, 28(01), 27–33.
[12] Juan J. Flores, Felix Calderon, Anastacio Antolino, et al. (2015). Network anomaly detection by continuous hidden markov models: An evolutionary programming approach. Intelligent Data Analysis, 19(2), 391–412.
[13] Apurva S, Deepak R. (2015). Post-Attack Intrusion Detection using Log Files Analysis. International Journal of Computer Applications, 127(18), 19–21.
[14] Yang Y, Jin S, Fang B. (2015). Security risk assessment based on bayesian multi-step attack graphs. Journal of Computational Information Systems, 11(11), 3911–3918.
[15] Zhang Y X, Zhao D M, Liu J X. (2014). The Application of Baum-Welch Algorithm in Multistep Attack. The Scientific World Journal, 5(1), 1–7.
[16] Qiu H, Wang K. (2016). Real-time Network Attack Intention Recognition Algorithm. International Journal of Security & Its Applications, 10(4), 51–62.
[17] Anna Buczak, Erhan Guven. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 18(2), 1–1.
[18] Alqurashi S, Batarfi O. (2016). A Comparison of Malware Detection Techniques Based on Hidden Markov Model. Journal of Information Security, 07(3), 215–223.S

[19] Rathore D, Jain A. (2019). Design Hybrid method for intrusion detection using Ensemble cluster classification and SOM network. International Journal of Advanced Computer Research, 2(3), 181–186.
[20] Yang Y , Jin S , Fang B. (2015). Security risk assessment based on Bayesian multi-step attack graphs. Journal of Computational Information Systems, 11(11), 3911–3918.
[21] Holgado P, Víctor A. Villagrá, Luis Vázquez. (2020). Real-Time Multistep Attack Prediction Based on Hidden Markov Models. IEEE Transactions on Dependable and Secure Computing, 17(1), 134–147.
[22] Mohammad Samar Ansari, Vaclav Bartos, Brian Lee. (2020). Shallow and Deep Learning Approaches for Network Intrusion Alert Prediction, 171, 644–653.
[23] Liu K, Wang H, Shen Z H. (2020). Prediction of network attack profit path based on NAPG model. The Journal of China Universities of Posts and Telecommunications, 0021, 1005–8885.

Biographies

Dongmei Zhao, Doctor of Engineering (Master of Network Security), Professor. Graduated from the Xidian University in 2007. Worked in Hebei normal university. Her research interests include network security situation estimation and prediction.



Hongbin Wang, studying in Computer Science and Technology, College of Computer and Cyber Security, Hebei Normal University. His research interests is network security.

Shixun Geng, master of applied software technology, graduated from Hebei Normal University in 2018. Her research interest is network security situation prediction.