Network Information Security Status Assessment Based on Big Data Analysis

Gong Dejing

Information Management Department, Zhongnan University of Economics and law, Wuhan, China

Keywords: Network attack, Dynamic Bayesian network, Network nodes, Comprehensive evaluation

Abstract: When formulating network attack strategy, there is an uncertainty in target network information, and the attacker lacks comprehensive, reliable and real-time attack basis, so it is difficult to achieve the attack effect. Therefore, a scientific and complex network attack method was proposed. The earnings, loss, cost and risks of the attacker were analyzed, and the index system was established, and then a dynamic Bayesian network was used to comprehensively evaluate the attack effect of network nodes, thus overcoming the shortcoming that traditional node importance degree evaluation methods conducted static evaluation by relying on the single index of network topology or target node. Therefore, the simulation experiments indicated that this method integrated more node relations and observation information when attacking, avoiding the gap between the actual attack effect and the theoretical expectation when carrying out attack by means of static evaluation, and at the same time, the attack precision was more accurate and the attack efficiency was higher.

1. Introduction

The situational awareness system, monitoring and control system, information hinge center and various power units in network space information confrontation are made up of highly-connected complex network. In the situational awareness system, first attacking the network system of the other party can directly damage or break down its information defense system. With the continuous development of information technology, complex network is more and more widely applied in military and economic field, the organizational structure of the network application is more and more collaborative. At the same time, the application levels develop towards many directions, and network attack behavior is more uncertain. Therefore, the attack strategy presents a complicated and diversified tendency. How to use the limited power to attack the numerous nodes of the target network most significantly is related to the attack efficiency, and the formulation of network attack strategy requires to precisely evaluate the network attack effect, which has been reached a consensus. How to make a qualitative and quantitative evaluation of the network attack effect, test the effectiveness of attack behavior and the security of network system in the complex network environment has become a research focus in the related fields.

2. Construction and theoretical foundation of evaluation index system of network nodes

Because there are more than ten measurement methods of network node importance, covering multiple aspects such as the topology structure, dynamics process and so on, objectively evaluating the attack effect of the nodes in complex networks must establish a reasonable and comprehensive evaluation index system. According to general principles of constructing the index system, all kinds of factors are comprehensively considered and analyzed, and the index system from the perspective of local and global properties, cascading failure of the network is constructed. Then through a series of key measurement indexes reflecting the structure information of network system, the properties of topological structure are analyzed, and the change in the structure attribute values of the evolution network is quantified, and later the connection relationship changes of the nodes in dynamical evolution process when attacking complex network systems are explained, so as to judge the damage of the network function after the attack and acquire the comprehensive evaluation conclusion of attack effect, which provides a strong goal-oriented strategy for the subsequent

DOI: 10.25236/etmhs.2019.355

selective attack.

Attack benefit refers to the effect of an anticipated action before network attack. In the network attack process, it mainly refers to the impact of the network node being attacked on the other party's network after it is paralyzed, importance of the deliberately attacked node in the enemy's network as well as the possible impact on the overall situation after it is paralyzed by the attack, so as to determine the attack strategy.

It is easy to quantify the importance index of local network attributes of the nodes, and the attribute information of neighboring nodes only needs to be considered, and it is only suitable for analyzing the importance of local network nodes in large-scale networks.

Definition 1 Node degree

The degree of node i in the network is defined as the number of nodes adjacent to it, and it is expressed as follows:

$$K(i) = \sum_{j \in G} a_{ij} \tag{1}$$

Where a_{ij} =1 represents the direct connection between node i, j （$i \neq j$）, otherwise, $a_{ij} = 0$. This attribute is the degree that a single node in the local network affects the functional characteristics of other nodes. At the same time, the importance of the network nodes not only depends on their own attribute information, but also the degree of adjacent nodes also has a certain influence on their importance. Based on the information of adjacent nodes and agglomeration factor, the node importance can be defined as $L(i)$, specifically as follows:

$$L(i) = \sum_{j \in \Gamma(i)} \sum_{u \in \Gamma(j)} N(u) \tag{2}$$

Where $\Gamma(i)$ represents the set of adjacent nodes of node i; $\Gamma(j)$ is the set of nodes nearest to node j; and $N(u)$ is the sum of the number of nearest neighbor nodes and the number of next-closest neighbors of node u.

Definition 2 Closeness centrality

The proximity of nodes represents the reciprocal of the sum of the shortest path distances between node i and other nodes in the network. If d_{ij} is the shortest distance from node i to node j, its expression is as follows:

$$CC_i = N / \sum_{j=1}^{N} d_{ij} \tag{3}$$

The larger the closeness centrality of the node is, the degree that it is in the center of the network will be, and the more important the node will be.

Definition 3 Betweenness centrality

If $g_{jk}(i)$ is the number of shortest paths passing between node j and node k going through node i, and g_{jk} is the number of all shortest paths between node j and node k, then the expression of betweenness centrality is as follows:

$$BC_i = \sum_{i \neq j \neq k \in V} g_{jk}(i) \Big/ g_{jk} \tag{4}$$

If a node is the only way which must be passed by other nodes in the network, the more important its position is, the greater its influence on network communication will be.

Definition 4 Clustering coefficient

The connection degree between all nodes connected to a node in the network can be defined as node clustering coefficient and network clustering coefficient.

The node clustering coefficient C_i is defined as follows:

$$C_i = \frac{2e_i}{u_i(u_i - 1)} \tag{5}$$

Where u_i is the number of nodes connected to node i, and e_i is the number of edges that may exist between the nodes connected to node i.

The network clustering coefficient is defined as the average value of each node clustering coefficient in the network, as shown below:

$$C = 1/N \sum_{i=1}^{N} C_i \tag{6}$$

The closeness degree of the relation between nodes in the network is proportional to the network clustering coefficient. When the coefficient value is 1, the network is a complete graph and there are edge connections between any nodes. If the coefficient is 0, all nodes in the network are isolated ones and there are no edges between nodes.

In addition, after comprehensively considering the number of adjacent nodes of the node and the connection degree between between them, the importance of nodes can be more objectively determined by using a method based on the information of adjacent nodes and clustering coefficient. The specific definition is as follows:

$$P(i) = \frac{fi}{\sqrt{\sum_{j=1}^{N} f_i^2}} + \frac{g_i}{\sqrt{\sum_{j=1}^{N} g_i^2}} \tag{7}$$

Where f_i is the degree of node i and the sum of degrees of adjacent nodes $f_i = k(i) + \sum_{u \in \Gamma(i)} k(u)$; $k(u)$ is the degree of node $k(u)$; and g_i is expressed as follows:

$$g_i = \frac{\max_{j=1}^{N}\left\{\frac{c_j}{f_j}\right\} - \frac{c_i}{f_i}}{\max_{j=1}^{N}\left\{\frac{c_j}{f_j}\right\} - \min_{j=1}^{N}\left\{\frac{c_j}{f_j}\right\}} \tag{8}$$

Where c_i is the node clustering coefficient.

Based on the importance index of the global network attribute, the topological information of the whole network system is mainly considered. The selection of the index requires to comprehensively and accurately reflect the functional structure of the network, and the computational complexity is relatively high, so it is suitable for the index parameters which are considered when conducting a parallel attack on larger region or multiple level and multiple subnet in the complex network.

Definition 5 Feature vector

When using the degree index to evaluate node important degree, the adjacent nodes are regarded as the same important. This consideration is impractical, and there is no equality between nodes. Therefore, when judging the importance of the node, the influence of adjacent nodes on it shall be considered. If a node is more strongly influenced by adjacent nodes, then the node importance is likely to be higher. If it is affected by the adjacent nodes weakly, even if the number of adjacent nodes is large, it isn’t necessarily important, which is regarded as the feedback of the importance of the adjacent nodes.

The feature vector was adopted in this paper to measure this property of the nodes, that is, the feature vector of the maximum eigenvalue corresponding to the network adjacency matrix, which is specifically defined as follows:

$$C_e(i) = \lambda^{-1}\sum_{j=1}^{N} a_{ij}\varepsilon_j \tag{9}$$

Where λ is the maximum eigenvalue of the adjacency matrix, and $\varepsilon = (\varepsilon_1, \varepsilon_2 ... \varepsilon_n)^T$ is the feature vector corresponding to the maximum eigenvalue of the adjacency matrix. The reputation of a single node in the network can be regarded as a linear combination of the reputation of other nodes, and then a linear equation set can be obtained. The feature vector corresponding to the maximum eigenvalue of the equation set can measure the importance of each node.

Definition 6 Compactness

Compactness can measure the ability that a node in the network exerts influence on other nodes. The stronger the compactness of a node is, the more important it is for the function realization of the network system will be, and the more central it is in the network topology will be, as defined below:

$$V(i) = \frac{N-1}{\sum_{j=1}^{N} d_{ij}} \tag{10}$$

Where d_{ij} is the shortest distance between node i, j , and the index of compactness largely depends on the topology structure of the network, so the time complexity of calculation should be taken into account.

3. Node importance of complex load network based on cascade failure

The evaluation of the importance of network nodes in the previous two sections mainly considered from the static perspective, and in reality, there are specific or abstract loads on most of the networks [7, 8,12,13]. Their distribution can be determined by multiple factors, and network topology structure is one of the main factors. The loads determined by the topological structure can be defined as "structure load", and when specific physical loads cannot be judged in the network, "structure load" can be used to estimate the destroy-resistance of complex network and node importance degree. Here, the number of shortest paths is used to measure the load size, that is, the more shortest paths going through the node are, the higher the load on the node will be [8]. The specific definition is as follows:

$$L_i = \frac{\sum_{i \neq j} \frac{s_{ij}(k)}{s_{ij}}}{n(n-1)} \tag{11}$$

Where s_{ij} is the number of all shortest paths between node i and node j, and $s_{ij}(k)$ is the number of shortest paths between node i and node j going through node k.

Attack loss refers to the resource consumption spent by attack. In network attack, all kinds of attack means (equipment) have their own resource cost after performance evaluation, which can be extracted as the corresponding indexes. It mainly refers to Party B's computer resource consumption when using Trojan, viruses and other means to attack, and it can measure the occupation and attack time indicators including bandwidth, CPU, internal storage, etc.

The expression of the occupation rate of CPU is as follows:

$$\bar{R}_{cpu} = \frac{\sum_{i=1}^{n} R_{i_{cpu}}^{t} - R_{i_{cpu}}^{0}}{n} \tag{12}$$

Where $\bar{R}_{cpu}$ is the average CPU utilization rate of host group under attack after network attack, and $R_{i_{cpu}}^{t}$ and $R_{i_{cpu}}^{0}$ are the CPU utilization rate of single attack terminal after and before network attack, respectively; Similarly, the expression of internal storage occupancy rate and bandwidth occupancy rate is formula (6) and (7) below:

Internal storage occupancy rate:

$$\bar{R}_{mem} = \frac{\sum_{i=1}^{n} R_{i_{mem}}^{t} - R_{i_{mem}}^{0}}{n} \tag{13}$$

Bandwidth occupancy rate:

$$\bar{R}_{band} = \frac{\sum_{i=1}^{n} R_{i_{band}}^{t} - R_{i_{band}}^{0}}{n} \tag{14}$$

Most selective attack strategies of traditional complex networks did not consider cost factor during the attack, and under such a premise, the attack cost is not the factors considered when removing nodes or edges in the network. According to this condition, the network is very weak when conducting selective attack to scale-free network, but in reality, the scale-free network can show a robustness which isn’t consistent with the assumption when receiving attack, so comprehensively measuring attack strategy needs to consider the cost factor.

The network G containing N nodes and E edges can be defined as a set $G=(N,E)$. When the network is attacked for every time, it needs to remove M nodes, and then $U(M)$ is denoted as the attack cost. Its definition is as follows:

$$U(M)=\sum_{i\subset\Gamma(M)} H(i) \tag{15}$$

$H(i)$ is defined as the function of node degree x, and $H(i)=x$. At this moment, under the same attack strategy, the cost of removing a node with node degree of x is x, and the nodes with a large node degree have a larger attack cost. In reality, there is a upper limit value in the attack cost of attack behavior, representing that removing all the nodes N in the network G needs the cost N. Therefore, to facilitate quantification, the normalization processing of $U(M)$ is as follows:

$$\bar{U}(M) = \frac{U(M)}{U(N)} \tag{16}$$

The main risk in network attack lies in the protection degree of the other party’s system as well the difficulty of repairing its own loophole after being discovered by the other party, mainly including security of operating system, firewall, intrusion detection, router security, interchanger security. The index of loophole repair difficulty of target network nodes is the loophole risk rate of network nodes.

Single loophole adopts risk rate $P(V_i)$ to quantify, and it is determined by the loophole’s popularity P_x, easiness P_y, and influence P_z, and $P(V) = P_x * P_y * P_z$. The multi-stage attack chain formed by the attacker is composed of N loopholes, and the attack condition of M loopholes needs to be met to realize the attack, i.e. $V = V_1 \wedge V_2 \wedge V_3 \wedge ... \wedge V_m$, so the attack risk degree can be defined as:

$$R\left(A\right) = P(V)_1 \wedge (V)_2 \wedge (V_3) \wedge ... \wedge (V_m) \tag{17}$$

The change in structure and function of the target network before and after the attack reflects the change in its operating efficiency, and it can reflect the effect of a single attack. Here, the maximum connected subgraph after the attack O(M) is used to quantify and normalize the network efficiency, and it is expressed with E（M）as follows:

$$E（M）=\frac{\mathrm{O(M)}}{|N|} \tag{18}$$

Where $|N|$ is the total number of nodes containing the network. According to the above definition, calculating the attack effect of target network under certain attack strategies shall comprehensively consider attack cost and attack loss, so as to effectively control the attack gains and loss, cost and cost-effectiveness ratio. The more significant the inverse relation between cost, consumption and network efficiency is, the more effective the attack strategy will be, while on the contrary, the lower the effectiveness of the attack strategy will be.

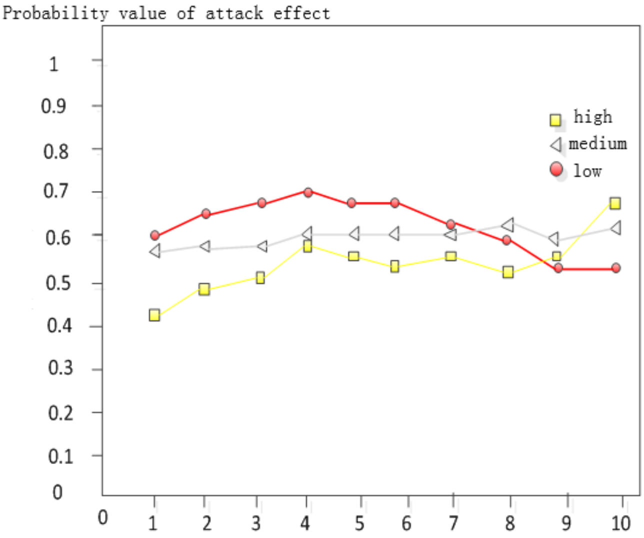

Fig. 1 Probability distribution of attack effects of node 3 with at high, medium and low time

4. Conclusion

When formulating attack strategy, there is uncertainty and uncontrollability in the attacked network structure, defensive disposition as well as important nodes, and at the same time, the index data evaluating attack effect is often not comprehensive. Therefore, using static method to evaluate the target network has a strong passivity, and there is often a gap with the expected effect. Based on this, the dynamic Bayesian network method was adopted to comprehensively analyze various factors in network attack, establish evaluation index system by using dynamic Bayesian network method, and dynamically assess network attack effect. Considering the impact of network nodes on the global and local after being attacked, attack cost and loss were regarded as the measurement factors of decision-making basis when attacking, and then a new attack method was proposed, so as to provide more scientific, independent and controllable auxiliary decision-making means when formulating the network attack plan and implementing attack, so it can greatly improve the effect of the complex network attack.

References

[1] Backes M, Berrang P, Manoharan P. How well do you blend into the crowd? - d-convergence: A novel paradigm for quantifying privacy in the age of Big-Data[J]. 2015.

[2] Hueneman D. Privacy on Federal Civilian Computer Networks: A Fourth Amendment Analysis of the Federal Intrusion Detection Network, 18 J. Marshall J. Computer & Info. L. 1049 (2000)[J]. Jitpl, 2000(4):1049.

[3] Swayne D E, Pavade G, Hamilton K, et al. Assessment of national strategies for control of high-pathogenicity avian influenza and low-pathogenicity notifiable avian influenza in poultry, with emphasis on vaccines and vaccination.[J]. Rev Sci Tech, 2011, 30(3):839-870.

[4] Tarman T D, Witzke E L, Kellogg B R, et al. Final Report for the Intrusion Detection for Asynchronous Transfer Mode (ATM) Networks Laboratory Directed Research and Development Project [J]. Office of Scientific & Technical Information Technical Reports, 2001.

[5] De Blaeij A T, Nunes P A L D, van den Bergh, Jeroen C. J. M. Modeling 'No-choice' Responses in Attribute Based Valuation Surveys [J]. Working Papers, 2005.