International Journal of Digital Crime and Forensics
Volume 13 • Issue 2 • March-April 2021

Research on Threat Information Network Based on Link Prediction

Jin Du, Institute of Software, Chinese Academy of Science, Beijing, China & Yunnan Police College, Kunming, China

Feng Yuan, Institute of Software Application Technology, Guangzhou, China & Chinese Academy of Sciences, Guangzhou, China

Liping Ding, Institute of Software Application Technology, Guangzhou, China & Chinese Academy of Sciences, Guangzhou, China

Guangxuan Chen, Institute of Software, Chinese Academy of Science, Beijing, China & Zhejiang Police College, Hangzhou, China

Xuehua Liu, Institute of Software, Chinese Academy of Sciences, Beijing, China & University of Chinese Academy of Sciences, China

ABSTRACT

The study of complex networks is to discover the characteristics of these connections and to discover the nature of the system between them. Link prediction method is a classic in the study of complex networks. It ca not only reflect the relationship between the node similarity. More can be estimated through the edge, which reveals the intrinsic factors of network evolution, namely the network evolution mechanism. Threat information network is the evolution and development of the network. The introduction of such a complex network of interdisciplinary approach is an innovative research perspective to observe that the threat intelligence occurs. The characteristics of the network show, at the same time, also can predict what will happen. The evolution of the network for network security situational awareness of the research provides a new approach.

KEYWORDS

Complex Network, Interdisciplinary Approach, Link Prediction, Network Evolution Mechanism, Situational Awareness, Threat Information Network, Threat Intelligence Complex System Science Method

1. LINK PREDICTION AND THREAT INTELLIGENCE

With the Internet as the representative of the rapid development of network information technology, human society has entered a complex network era. Human life and production activities are increasingly dependent on complex systems. As an interdisciplinary emerging field, network science and engineering have been gradually formed and developed rapidly.

The network topology has expanded people's understanding of complex systems, and complex networks are more in-depth to describe the essence of complex systems. Network science is not only an extension of classical graph theory and stochastic graph theory in mathematics, but also an innovative development of system science and complexity science. Scholars through the complex network involved in the economics, biology, physics and other disciplines of observation and research, the use of network nodes between the topology to find unknown or future will be generated links,

DOI: 10.4018/IJDCF.2021030106

the problem becomes more important research points, this is the link prediction problem. The link prediction problem of complex networks refers not only to the prediction of future links, but also to the predictions of links that already exist but not yet found.

1.1 Link Prediction

In nature, there are numerous complex systems that cover ecosystems, social networks, economic networks, and so on. Many complex shapes, such as social networks, political networks, and so on, are closely related to our lives. Complex networks are abstractly described by these methods in a scientific way, and the nature of these systems is discovered. And with the complex network of in-depth study, scholars have found that many do not look the same network, but surprisingly has a lot of similar characteristics. The foundation of complex network construction is the connection between individual and individual. However, when the network is built, the incompleteness and uncertainty of the collected information will cause many of the edges that should have existed to disappear, and many errors occur. To a large extent affected the network attributes and the integrity of the network, the study of complex network interference.

In order to solve this problem, scholars in various fields began to study the network link prediction. There are two reasons for the network link prediction to be concerned in each field. First, from the theoretical point of view, the link prediction algorithm can not only accurately describe the "node similarity", but also in the complex network Prediction, at the same time will reveal the inherent factors of network evolution, that is, the evolution of the network mechanism. In the near future, link prediction is likely to provide a fair and unified platform for the evolution mechanism of the network, and in essence, to promote the study of network evolution mechanism. Second, it is reflected in its practical application value. Network link prediction can not only be used to predict some of the interaction in the biological field, thus saving research time and money, but also applied to the economic network, traffic network research, can bring more intuitive economic benefits and national resources savings. In an evolving online social network, link predictions can be used to determine similarity through the user's historical behavioral attributes, thus determining the likelihood that two users who have never had a relationship become friends (Signoretto et al., 2011; Leskovec et al., 2009; Viswanath et al., 2009; Bader & Kolda, 2007; Chatfleld, 2013; Sharan & Neville, 2008; Bringmann et al., 2010; Juszczyszyn et al., 2011; R˜ummele et al., 2015; Davis et al., 2013).

1.2 Threat Intelligence

In recent years, the concept of "threat intelligence" has rapidly emerged in various fields, especially in the field of information security, and many security vendors have launched threat-related services. As the threat of intelligence is not consistent with the definition of the various areas of security focused on the current, mainly for the definition of data security. Network threat information refers to any information that helps organizations identify, evaluate, monitor, and respond to network threats. Such information includes offensive indicators

(Indicator of compromise, also referred to as "attack indicator", "intrusion indicator"), threatens the use of tactics, techniques and processes (TTP), detection, control or protective attack As well as security event analysis results. The sharing of network threat information can improve the security status of sharing organizations and other organizations at the same time.

Therefore, using the observation information known to the target network, it is possible to evaluate the possibility that there may be a possible edge between nodes in the network. It can be realized by link prediction method. Conducting professional research can predict the presence of but not yet found in the threat intelligence network, and predict the possible presence of a future that may not be present. This link forecast opens up a new situation for theoretical research, and also shows the actual use value of the research on network security science. So, link prediction research has a very important theoretical value and application value, which is a very important work. In this way, we can find the classification of nodes in the threat information network, so that the association between

nodes in the same set is relatively close, and the correlation between different sets is relatively sparse. And can further analyze the network topology, a better understanding and explain the function of the network, which can be more easily found in the network of some hidden laws, predict the behavior of the network.

2. LINK FORECASTING METHOD AND EVALUATION INDEX

There are three common indicators for measuring the accuracy of link prediction algorithms: AUC (Area Under the Receiver Operating Characteristic Curve), Precision and Ranking Score. The AUC is the measure of the accuracy of the algorithm as a whole, and Precision only considers whether the edge of the first L bit is accurate and the Ranking Score takes into account the predicted edge order. (Liu, 2011)

AUC is the classical method of link prediction. Randomly select a test set of edges, calculate the fractional value, than the random selection of a non-existent edge of the high value of the score. When a numerical estimate is made, each time a random selection from the test set is selected, an edge is compared with the fractional value of the randomly selected edge that does not exist. If the score of the edge of the test set is greater than the fractional value of the edge that does not exist, add 1 point; if the two scores are equal, add 0.5 points. Independently compare n 'times, if there are n' test scores of the edge of the score is greater than the edge of the score, there are nn times the two points are equal, then the AUC is defined as:

$$\text{AUC}=\frac{n'+0.5nn}{n} \tag{1}$$

Constructs a simple network G (V, E), V is the set of nodes, and E is the set of edges. The total number of nodes in the network is N (N = | V |), and the number of edges is M (M = | E |). The network has a total of N (N-1) / 2 node pairs, ie, the complete set of U. Given a method of link prediction, give a fractional value of Sxy for each pair of nodes without edges. Since G is invariant, the score is symmetric, ie Sxy = Syx. And then all the nodes are not connected in accordance with the scores from large to small sort, ranked in the front of the node, that is, the algorithm that appears to have the largest probability of even the edge of the node.

Precision is defined as the exact proportion of the score that is ranked in the front L bit (not including the edges in the training set). Precision is more accurate. If there are m m in the test set in the front L bit, then Precision is defined as:

$$\text{Precision}=\frac{m}{L} \tag{2}$$

Ranking Score mainly considers the position of the edges in the test set in the final order. Let H = U-ET be a collection of unknown edges, the unknown edge contains the actual but not yet known edge (the edge of the test set) and the non-existent edge (neither the training set nor the node pair in the test set) ri represents the rank of the unknown edge i ∈EP in the sort. Then the Ranking Score value for this unknown edge is RankSi = ri / | H |, and the Ranking Score value of the system is obtained by traversing all edges in the test set:

$$\text{RankS}=\frac{1}{\text{EP}}\sum_{i\in\text{EP}}\text{RankSi}=\frac{1}{\text{EP}}\sum_{i\in\text{EP}}\frac{ri}{\left|\text{H}\right|} \tag{3}$$

Obviously, the smaller the RankS value indicates that the edge of the test set is in the front position, which means that the probability of success is predicted, so the higher the algorithm accuracy. (Liu, 2011)

In some smaller networks, a more accurate data set partitioning method is leave-one-out, that is, each time from the network to select an edge as a test edge, predict the possibility of this edge, and then the application of Ranking Score evaluates the prediction of this edge. For all the edges of the network, this is a prediction (a total of M times), the average value of RankS, that is, the accuracy of the entire network prediction. It is worth noting that this approach is not suitable for large-scale networks because it recalculates the proximity each time.

The nature of link prediction is the cause of mining leads, which is also a concern of the network evolution model. An evolutionary model, in principle, can correspond to a link prediction algorithm. Therefore, we can use the framework of link prediction and evaluation methods to quantitatively evaluate the algorithms corresponding to different evolution models, so as to indirectly compare and evaluate the evolution model.

2.1 Implementation

We will monitor the threat of intelligence network, for example, the network contains 287 nodes 3167 edge, belonging to the smaller network, the prediction algorithm evaluation(Peng, 2015).

In order to detect the accuracy of the algorithm, the set E of known edges is randomly divided into two parts: one is the training set ET, the known information is used to calculate the fractional value; the other part is the test set EP, The information in this collection cannot be used for forecasting. Obviously, E = ET ∪EP, and ET ∩EP =Æ. For example, 10% of all edges of the network are removed from the network as a test edge to predict which edges are deleted based on the remaining 90% of the information. Here, the edge belonging to U but not E is called the edge that does not exist.

A, define the number of nodes in the network N, the network size is defined as M, divided training set ET and test set EP. Introduce the edge set E, and count the number of real networks. According to the test set, the number of edges of the test set is H = [| E | × Percent]. Where [x] represents the largest integer less than x, so [x + 0.5] is the integer closest to x.

B, randomly selected from the edge set E edge H as the test set EP. The remaining edge is the training set ET, and is the observed edge set of the network. A0ij = 1 represents a connection between node i and node j, and A0ij = 0 represents no edge between node i and node j.

C, to build an M-scale network, if the nodes in the network are not fully connected, then remove the smallest node and separate as a network until the nodes within the network are connected to each other. Define the detection parameter λ for the range [0 1]. Assign the objective function value Dλ to calculate the fitness function f.

$$f = 2r / (M (M + 1)) \tag{4}$$

where r refers to the position of the metric calculated by the metric function, M is the size of the network, and the magnitude of the value of f determines the size of the prediction probability. The larger the value of the metric function, the greater the sorting, the greater the corresponding r value, ensuring that the maximum value of the metric function has converged, and the final optimal individual is the sample at that resolution. Until λ [0.1 ~ 0.9], get 9 samples.

D, calculate the test set EP and the probability that there is no concentration of all edges.

E, the edge of the order and give the need to predict the edge.

Calculate the AUC evaluation of this algorithm. In order to evaluate the performance of this algorithm, the AUC indicator is used as an evaluation function. AUC can be seen as a probability that it is a probability that a probability R is randomly chosen from a test set to be greater than a probability R that is randomly absent. In this part, we pick out the R value of each edge from the test

Figure 1.

CN	0.941	O.913	0.860	0.928				
AA	0.971	0.962	0.851	0.951				
RA	0.981	0.959	0.863	0.967				
DCCN	0.932	0.911	0.861	0.933	0.430	0.482	0.114	0.212%
DCAA	0.959	0.962	0.860	0.958	0.012	0.061	0.011	0.488%
DCRA	0.973	0.971	0.868	0.968	0.027	0.169	0.075	0.979%
CCCN	0.952	0.905	0.868	0.950	0.021	0.001	0.671	0.009%
CCAA	0.971	0.949	0.861	0.959	0.031	0.409	0.011	0.110%
CCRA	0.958	0.943	0.873	0.967	0.008	0.221	0.173	0.289%
BCCN	0.911	0.945	0.843	0.942	3.298	0.176	0.009	0.0289%
BCAA	0.972	0.951	0.853	0.958	0.021	1.421	0.171	0.098%
BCRA	0.980	0.971	0.882	0.963	0.231	1.173	0.079	0.373%

set, and if the former is large, n1 = n1 + 1, if both are equal, n2 = n2 + 1. Repeat the operation until all edges in the test set are compared(Peng, 2015).

2.2 Simulation

In order to evaluate the effectiveness of the algorithm, this paper carries out experiments on four typical real threat network datasets. The nodes in the network represent the attack points. The links between the nodes represent the link between the threat information, the network contains 287 Node, 3167 link relationship. In this paper, the AUC value is used as the measure of accuracy, and the CN index, AA index and RA index based on the local similarity link prediction algorithm are compared. The improved link forecasting method is applied to the network real data set, The accuracy of several similarity indicators when the training set ratio is [0.1 ~ 0.9] is recorded in Figure 1, and Figure 2 shows the improvement of the AUC index of the improved algorithm relative to the contrast algorithm(Jiaying, 2016):

It can be seen from figure 1 that, on the four data sets, the CN algorithm of the importance of the node is considered: the prediction accuracy of DCCN, CCCN and BCCN algorithm is better than that of the CN algorithm. The algorithm of DCCN, CCCN and BCCN is compared with CN algorithm in 4 data sets, and its prediction accuracy is increased by 0.2123%, 0.4887% and 0.9792% respectively. Consider AA algorithm of node importance: DCAA, CCAA, overall precision of BCAA algorithm is better than the AA prediction accuracy of the algorithm and the above three kinds of algorithms on the four data sets compared to AA algorithm, the prediction accuracy were improved 0.0098%, 0.1102% and 0.2899%. Consider RA algorithm of node importance: DCRA, CCRA, BCRA algorithm overall better than the prediction accuracy of prediction accuracy of RA algorithm, the above three kinds of algorithms on the four data sets compared to RA algorithm, the forecast accuracy of 0.02891%, 0.0981% and 0.3731%. Figure 1 shows the comparison algorithm and the prediction precision of the improved algorithm of the AUC value, it can be seen that 77.9% of the prediction precision of the improved algorithm is higher than the comparison algorithm accuracy, but also a phenomenon of individual accuracy prediction accuracy is lower than the contrast algorithm, it has to do with the size of the data set and its accuracy(Jiaying, 2016)(Liben-Nowell & Kleinberg, 2007).

In the algorithm efficiency, N for the node number of CN algorithm firstly to find each pair was predicted node in the network, and then look for common neighbor nodes in the two nodes, thus CN algorithm's time complexity is O (N2). The AA and RA algorithms are only calculated based

Figure 2.

on the number of nodes in the common neighbor node, so the time complexity is the same as the CN algorithm. Consider node degree centrality algorithm, each node is calculated after the common neighbor nodes to find the DC value, the compute nodes DC time complexity is O (N), therefore, in CN, AA, RA index on the basis of considering the node degrees after centricity, improved algorithm time complexity doesn't change. Near centricity, similarly, considering node betweenness centrality algorithm, on the basis of common neighbor nodes according to the CC value of node, the BC value calculation, the computational complexity of computing nodes CC value, BC respectively O (N2), O (N3), therefore, consider node close to centricity, betweenness centrality of the algorithm's time complexity is O (N2) respectively, O (N3), compared with common neighbor, AA, RA algorithm, considering the node betweenness centrality algorithm improve the time complexity of considering node degree of centricity, close to the central the time complexity of the algorithm is the same. The experimental results show that the node importance has played a positive role in the link prediction accuracy, link prediction algorithm under the AUC evaluation index on prediction accuracy than the original link prediction algorithm has a degree of improvement..

3. CONSTRUCTION OF THREAT INTELLIGENCE PLATFORM BASED ON LINK PREDICTION

Combined with the previous experiment of link prediction algorithm and the current threat intelligence features, a large number of large data processing methods are summarized and compared. Through the platform, we extracted the data from the first three months of attacks on the platform targets as a data set. These data sets are used as machine learning samples, and link prediction method is applied to new attack behavior prediction.

At the same time, compared with domestic and foreign research progress, through summarization, experimental verification and typical case application analysis guidance platform construction of in-depth development (Lichtnwalter & Chawla, 2012)(Bliss et al.,).

First, through repeated threat intelligence data collection and analysis, combining with the distributed search engine data, collecting a large number of typical case analysis of the experiment, especially the multi-source heterogeneous data sources, through the study of large data storage and search engine, comparing the results of the analysis of link prediction algorithm to discover and record the key target of internal and external network attack alarm information and traffic information, and found the problems that exist in the cyber threat intelligence data processing.

Second, according to the characteristics of the link prediction and related research methods, the above intelligence data for coarse graining processing, to build a complex network based on threat information, in order to further study the construction of complex network analysis to lay the foundation.

Third, threat to build intelligence data network topology analysis, the characteristics of complex networks that exist in the network characteristics of excavation and in-depth analysis, will be mixed and disorderly network attack threat intelligence for regularity of deduction and analysis, found that the security situation in the hidden valuable intelligence.

Fourth, on the basis of the above based on intrusion alarm and traffic intelligence analysis, the overall situation of safety assessment to the network, and use the link prediction and related algorithm, the security situation forecast and emergency response strategy analysis, provide scientific strategic and tactical level of senior intelligence.

FOUNDATION

This work has been supported by the National Key Research and Development Program of China under grant 2016QY01W0200, Science and Technology Planning Project of Guangdong Province, China(2017B050506002), Collaborative Innovation Center for Economics crime investigation and prevention technology,Jiangxi Province(No.JXJZXTCX-007, No.JXJZXTCX-009), Science and Technology Planning Project of Guangzhou Municipality,China(201802020015), Support Scheme of Guangzhou for Leading Talents in Innovation and Entrepreneurship(No. 2016008).

REFERENCES

Bader, B. W., & Kolda, T. G. (2007). E–cient Matlab computations with sparse and factored tensors. *SIAM Journal on Scientific Computing*, *30*(1), 205–231. doi:10.1137/060676489

Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science, 5*(5), 750-767.

Bringmann, B., Berlingerio, M., Bonchi, F., & Gionis, A. (2010). Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, *25*(4), 26–35. doi:10.1109/MIS.2010.91

Chatfleld, C. (2013). *The Analysis of Time Series: an Introduction* (6th ed.). Chapman and Hall/CRC Press.

Davis, D., Lichtenwalter, R., & Chawla, N. V. (2013). Supervised methods for multi-relational link prediction. *Social Network Analysis and Mining*, *3*(2), 127–141. doi:10.1007/s13278-012-0068-6

Jiaying, C. (2016). A link prediction algorithm based on the importance of complex network nodes. *Journal of Computer Applications*.

Juszczyszyn, K., Musial, K., & Budka, M. (2011). Link prediction based on subgraph evolution in dynamic social networks. In *Proceedings of the 3rd IEEE International Conference on Social Computing, the 3rd IEEE International Conference on Privacy, Security, Risk and Trust.* Boston: IEEE. doi:10.1109/PASSAT/SocialCom.2011.15

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2009). Community structure in large networks: Natural cluster sizes andthe absence of large well-deflned clusters. *Internet Mathematics*, *6*(1), 29–123. doi:10.1080/15427951.2009.10129177

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. doi:10.1002/asi.20591

Lichtnwalter, R., & Chawla, N. V. (2012). Link prediction: fair and effective evaluation. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Istanbul, Turkey: IEEE. doi:10.1109/ASONAM.2012.68

Liu, H. (2011). *Using link prediction to predict network evolution mechanism*. SCIENTIA SINICA Phys, Mech & Astron.

Peng, X. (2015). Community network link forecast. Xi ‘an electronic science and technology major. Academic Press.

R˜ummele, N., Ichise, R., & Werthner, H. (2015). Exploring supervised methods for temporal link prediction in heterogeneous social networks. In *Proceedings of the 24th International Conference on World Wide Web*. Florence, Italy: International World Wide Web Conferences Steering Committee. doi:10.1145/2740908.2741697

Sharan, U., & Neville, J. (2008). Temporal-relational classiflers for prediction in evolving domains. In *Proceedings of the 8th IEEE International Conference on Data Mining*. Pisa, Italy: IEEE.

Signoretto, M., Van de Plas, R., De Moor, B., & Suykens, J A K. (2011). Tensor versus matrix completion: A comparison with application to spectral data. *IEEE Signal Processing Letters*, *18*(7), 403–406. doi:10.1109/LSP.2011.2151856

Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*. Barcelona, Spain: ACM. doi:10.1145/1592665.1592675

Jin Du is a Post-doctoral of Institute of Software, Chinese Academy of Sciences, and Digital Forensics Lab of Institute of Software Application Technology, Guangzhou &CAS. She has focused on digital forensics, system security and complex network, and she has undertaken a dozen national research programs and authored dozens of articles and several books.

Feng Yuan is the director of Institute of Software Application Technology, Guangzhou & Chinese Academy of Sciences. He has focused on smart city and undertaken a dozen national research programs and authored dozens of articles and several books.

Liping Ding is one of the Chinese foremost forensics scientists. She is currently the director of Digital Forensics Lab of Institute of Software Application Technology, Guangzhou &CAS, the director of Guangdong Real Data Judicial Expertise Institute, the Chairman of Computer Forensics expert committees, Chinese Institute of Electronics. She is a professor of Institute of Software, Chinese Academy of Sciences, and an adjunct professor of Law School of Renmin University of China, Central University of Finance and Economics, the Third Research Institute of the Ministry of Public Security and so on. She has focused on digital forensics, system security and trusted computing, and she has undertaken a dozen national research programs and authored dozens of articles and several books.

Guangxuan Chen is a post doctoral student of University of Chinese Academy of Sciences, Beijing and an teacher of Zhejiang Police College, Hangzhou. He has focused on digital forensics, system security and cloud computing. He has participated several national research programs and authored several articles and patents.

Xuehua Liu is a doctoral student of University of Chinese Academy of Sciences, Beijing and an engineer of Institute of Software, Chinese Academy of Sciences, Beijing. She has focused on digital forensics, system security and compiler optimization. She has been the director of an open project, participated several national research programs and authored several articles and patents.