Scale-free Characteristics and Link Prediction in Complex Railway Network

ZHANG Zengping¹*, HAN Yu², GAO Kuanyun¹, LI Bo¹ and ZHOU Yinghao¹

¹School of Computer Information and Management, Inner Mongolia University of Finance and Economics, Hohhot 010070, China
²School of Business Administration, Inner Mongolia University of Finance and Economics, Hohhot 010070, China
*Corresponding author’s e-mail: 100002170@imufe.edu.cn

Abstract: The link-prediction of complex networks is a new method of network structure mining, which is different from the traditional methods based on machine learning. It reveals those existing relations and predicts possible ones, by using the similarity index of network structure. This paper studied the topology of the railway and link-prediction. The railway dataset was firstly obtained by sampling and pre-processing. Its degree obeys the power-law distribution with the scale-free characteristics. Then, we give some predictions for railway networks. These tests reached good prediction accuracy. The research result has demonstrated that the topological structure of network could be better choice to the link-prediction of a real complex networks.

1. Introduction
With the development of network information technology, various complex systems continue to emerge. Therefore, it is of important theoretical and application values to recognize and characterize these complex systems [1-3]. At present, complex network theory is becoming a powerful tool for the study of various complex systems [4-10], such as social networks, technology networks, etc. The link-prediction in complex networks aims to reveal the unknown relations and predict the possible ones. Early researches are mainly based on Markov chain and machine learning. Among them, Sarukkai [11] proposed a link-prediction method, which uses Markov chain to realize link-prediction and path analysis. O'Madadhain et al. [12] reported a link-prediction method based on machine learning. The method uses the observed network topology information and node attributes. They then established a conditional probability model with machine learning to predict the link. The effect of this method depends on the extraction of feature information. However, in many cases, the extraction of feature information is difficult. With increasing of network scale, the network becomes extremely complex. So, the feature extraction becomes impossible. In addition, the attributes of nodes always have noise, and the extracted feature vectors cannot express the exact information. Considering the sparsity of the real network, the dataset for machine learning is seriously unbalanced, which makes the training unable to get a good model. Compared with feature information extraction, network structure information is easier to obtain and more reliable. Therefore, link-prediction, based on complex network theory, has become a research focus. To sum up, it mainly includes three technologies. They are local similarity, global similarity and maximum likelihood estimation. Local similarity is to use the similarity index of the common neighbor to predict the link. For the network with high clustering
coefficient, it has good prediction effect. However, due to the loss of most of the information of the network structure, there are some limitations. The global similarity is mainly based on the random walk process. The core of maximum likelihood estimation method is to predict the possibility of link existence from the perspective of network structure likelihood. Because of likelihood function and estimator, the computational complexity is larger than that of the former two methods. In this paper, the global similarity index is used for link-prediction in the railway network. Among them, the global similarity indexes include matrix forest index (MFI), average commuting time (ACT), cosine similarity (COS+), and restart random walk (RWR). The average AUC metrics are used to evaluate their prediction effectiveness.

2. Structure and its characteristics for complex railway network

2.1. Network data acquisition and preprocessing

The data used in the study are mainly train numbers. The observation data are shown in figure1.

![Figure1](image1.png)  
**Figure1. Observation data of railway network**

![Figure2](image2.png)  
**Figure2. The network lists**

The observed data include five attributes such as train number, serial number, train departure time, departure station and terminal station. In order to simplify, we only consider the simple undirected graph. Therefore, it is necessary to carry out data preprocessing. We take the city as the node of the network, and the link between two cities as an edge, regardless of self-loop and repeated link. Then, a simple undirected network is formed by using undirected network processing technology. The network file data is shown in figure2. The network size are 555 nodes and 1702 edges. In order to calculate, we code the city node in decimal system (see figure3(a) and (b)).

![Figure3](image3.png)  
**Figure3. Decimal coding and labeling of undirected networks.**

The observed data include five attributes such as train number, serial number, train departure time, departure station and terminal station. In order to simplify, we only consider the simple undirected graph. Therefore, it is necessary to carry out data preprocessing. We take the city as the node of the network, and the link between two cities as an edge, regardless of self-loop and repeated link. Then, a simple undirected network is formed by using undirected network processing technology. The network file data is shown in figure2. The network size are 555 nodes and 1702 edges. In order to calculate, we code the city node in decimal system (see figure3(a) and (b)).

2.2.5 Degree distribution characteristics of railway network

Network is a kind of geometric skeleton describing the connection between urban nodes of railway network, as shown in figure4.
According to the topology, we find that the network has obvious heterogeneity, namely, there are Hub nodes. We plot degree distribution curve on the double logarithmic coordinates. As shown in figure 5, the black scatter is the statistical probability of degree, and the red one is the fitting curve. The statistical results show that the degree distribution of railway network obeys power-law and the degree index is close to 2.18. The network has scale-free characteristics. Its typical feature is that most nodes tend to connect few nodes with large degree. Meanwhile, considering the actual situation of railway construction, we use mean field theory to analyze the degree distribution of railway network.

It is assumed that there are only isolated nodes in the network at the initial time. The network evolves with discrete time, and each time step \( \Delta t \) adds \( m \) edges, so the network has \( 2mn \) edges after \( n \) step, namely

\[
\sum_{j=1}^{n} k_j = 2mn
\]

Where \( k_j \) denotes the degree of node \( v_j \). the importance of Hub is expressed as probability

\[
\Pi(k_i) = \frac{k_i}{\sum_{j} k_j} = \frac{k_i}{2mn}
\]

The new \( m \) edges tend to connect the important nodes, so its probability is

\[
p(v_i) = m \left[ \prod (k_j) \right] \left[ 1 - \prod (k_j) \right]^{m-1}
\]

\[
= m \left[ \prod (k_j) \right]
\]

According to the continuity theory of mean field, the degree \( k_i \) of node \( v_i \) evolves as

\[
\frac{\partial k_i}{\partial t} = m \Pi(k_i) = \frac{k}{2t}
\]

If the urban node \( v_i \) is added to the railway network in step \( i \), the initial condition is \( k_i(i) = m \), and the solution of equation (4) is

\[
k_i(t) = m \left( \frac{t}{i} \right)^{\frac{1}{2}}
\]

In general, after time \( t \), we randomly select a city node \( v_i \) and the selection probability follows \( \rho(v_i) = 1/t \). According to the solution of the dynamic equation (5), the instantaneous degree distribution and probability density function of the network are

\[
P(k, i) = \frac{dp(k, i)}{dk} = \frac{k^{m-1} e^{-k}}{m^{m} \Gamma(m)}
\]

Equation (6) shows that if the degree distribution of the network is independent of time, then the
stable degree distribution of the network is \( P(k) = 2m^2k^{-3} \), where the scale index is \( \gamma = 3 \). This analytical result is good agreement with the statistical results.

3. Link prediction in railway network

3.1. Global similarity index

3.1.1. Matrix forest index (MFI).
Based on matrix forest theory, the similarity index of nodes \( v_x \) and \( v_y \) is defined as
\[
S_{xy} = (I + \alpha L)_{xy}^{-1}
\]  
(7)
Where \( \alpha > 0 \) and \( L \) is Laplacian matrix.

3.1.2. Average commuting time (ACT).
The average number of steps of a random particle in the network from node \( v_x \) to node \( v_y \) is set to \( m(x, y) \), then the average commute time of node \( v_x \) and node \( v_y \) can be defined as
\[
n(x, y) = m(x, y) + m(y, x),
\]
and its numerical solution can be obtained by the pseudo inverse matrix \( L^* \) of Laplacian matrix. The smaller the ACT of two nodes, the closer the two nodes are, and the similarity index is defined as
\[
s_{xy}^{ACT} = (l_{xx}^* + l_{yy}^* - 2l_{xy}^*)^{-1}
\]  
(8)

3.1.3. Cosine similarity of random walk (COS+).
A node \( v_i \) in the network is transformed into a vector \( v_i \) in Euclidean space by matrix \( \frac{1}{\lambda} U^T e_i \) (where \( A \) is the diagonal matrix of the pseudo inverse eigenvalue of Laplace matrix, \( U \) is the normalized orthogonal eigenvector matrix, and \( e_i \) is the unit column vector). The cosine similarity of the vectors corresponding to node \( v_x \) and node \( v_y \) is
\[
s_{xy}^{COS+} = \frac{v_x^T v_y}{||v_x|| ||v_y||}
\]  
(9)

3.1.4 Restart random walk (RWR).
Restart random walk index belongs to the application of PageRank. At time \( t \), the probability vector of the random walk particle at node \( v_x \) to other nodes in the network in the next step is defined as
\[
\mathcal{G}_x(t+1) = \rho P^T \mathcal{G}_x(t) + (1-\rho) e_x
\]  
(10)
Where \( P \) is the Markov transition probability matrix, \( \rho \) is the probability of the particle random walk, and \( e_x \) is the unit column vector of the particle starting point. The steady state solution after iteration is
\[
\mathcal{G}_x(t) = (I - \rho P^T)^{-1} e_x
\]  
(11)
Then the probability of the particle random walk \( \mathcal{G}_{xy} \) from node \( v_x \) to node \( v_y \) can be defined as
\[
s_{xy}^{RWR} = \mathcal{G}_{xy} + \mathcal{G}_{yx}
\]  
(12)

3.2. AUC evaluation metrics
AUC is the area under the ROC curve in radar signal detection theory, which is often used to evaluate
the overall accuracy of the classifier. Therefore, we use AUC as the evaluation of railway network link-prediction. We can simplify the calculation process of AUC. Each time, an edge and a non-existent edge are randomly selected from the test set for comparison. If the predicted value of the existing edge is greater than that of the no existing edge, 1 is added; If the predicted values are equal, 0.5 is added. \(n\) tests were conducted independently. If there are \(n_1\) times that the predicted value with an edge is greater than that without an edge, and \(n_2\) times that the predicted values are equal, then AUC can be defined as

\[
AUC = \frac{n_1 + 0.5n_2}{n}
\]

(13)

Obviously, if all the scores are randomly generated, then AUC = 0.5. Therefore, the proportion of AUC greater than 0.5 shows how accurate the algorithm is compared with the random selection method, and the closer to 1, the more accurate it is.

4. Experimental test
Firstly, the observed railway network dataset is divided into training set and testing set. Then, four global similarity indexes are trained and tested.

4.1. Dataset partition
The random walk method is used to divide the network data into training set and testing set. The ratio of training set and testing set is 9:1. The random walk partition method includes the following three steps.

1) Initially, a random node is selected in the network and a wandering particle is released.
2) The particle randomly walks to the neighbor node and puts the connected edge into the testing set.
3) Repeat step 2, until 10% of the connected edges are included in the testing set, and the remaining 90% of the edges in the network are used as the training set.

4.2. Experimental test and measurement results
Table 1. The experimental results of the simulation.

|        | MFI     | ACT     | Cos+    | RWR     |
|--------|---------|---------|---------|---------|
| AUC1   | 0.9226  | 0.8698  | 0.8617  | 0.9556  |
| AUC2   | 0.9292  | 0.8939  | 0.8549  | 0.965   |
| Mean value | 0.9259  | 0.8819  | 0.8583  | 0.9603  |
| Mean square error | 1.07×10^{-5} | 1.45×10^{-4} | 1.16×10^{-4} | 2.21×10^{-5} |

According to the theoretical model of global similarity index, the simulation experiments are carried out on matrix forest index (MFI), average commuting time (ACT), cosine similarity (COS+) and restart random walk (RWR), respectively. The experimental results are shown in table 1.

From table 1, we see that the index of restart random walk (RWR) is the best, of which AUC = 0.9603, and the mean square error is 2.21×10^{-5}. At the same time, we also found that MFI and RWR indicators are better than the other two models, which shows that the railway network has scale-free characteristics.

5. Conclusion
By analyzing whether there is a direct link between any two cities in the railway network, this paper explores the network topology characteristics and link-prediction of the railway network. According to the degree distribution characteristics, we find that the topological structure of railway network belongs to scale-free complex network, and the degree distribution obeys power-law distribution. This result shows that the construction and evolution of railway network tend to connect important first and second tier city nodes. In link prediction, we compare four global similarity indexes and find that RWR is the best index, AUC = 0.9603 and MSE is 2.21×10^{-5}. We also find that the MFI and RWR
indexes with clustering characteristics have better test results than the other two models, which shows that the railway network has not only scale-free characteristics, but also large aggregation characteristics. To sum up, exploring the railway network and predicting the probability of railway construction between cities could be an important reference value for the planning and construction of railway network.

Acknowledgements
This research is funded by National Natural Science Foundation of China under Grant 61563038, Natural Science Foundation of Inner Mongolia Autonomous Region under Grant 2020MS06021, and Talent Development Foundation of Inner Mongolia Autonomous Region.

Reference
[1] Hu L, Chan K C C. (2016) Fuzzy Clustering in a Complex Network Based on Content Relevance and Link Structures. J. IEEE. Transactions On Fuzzy Systems., 24(2): 456-470.
[2] Coquide, C., Georgeot, B., Giraud, O. (2017) Distinguishing humans from computers in the game of go: A complex network approach. J. Europhysics Letters., 119(4): 48001.
[3] Urena, R., Kou, G., Dong, Y., et al. (2019) A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. J. Information Sciences., 478: 461-475.
[4] Kuhlman C J, Kumar V S A, Marathe M V, et al. (2015) Inhibiting diffusion of complex contagions in social networks: theoretical and experimental results. J. Data Mining And Knowledge Discovery., 29(2): 423-465.
[5] Bodaghi, A., Goliaei, S., Salehi, M. (2019) The number of followings as an influential factor in rumor spreading. J. Applied Mathematics and Computation., 357: 167-184.
[6] Albieri Barone Cantero S V, Goncalves D N, Dos Santos Scabini L F, et al. (2020) Importance of Vertices in Complex Networks Applied to Texture Analysis. J. IEEE. Transactions on Cybernetics., 50(2): 777-786.
[7] Sahel Sahhaf, Wouter Tavernier, Dimitri Papadimitriou, et al. (2017) Routing at Large Scale: Advances and Challenges for Complex Networks. J. IEEE. Network., 31(4):108-118.
[8] X P Wu, H Y Yang, S C Han. (2016) Analysis on network properties of multivariate mixed air traffic management technical support system based on complex network theory. J. Acta Physica Sinica., 65(14):140203.
[9] Piovesan, D., Minervini, G., Tosatto S C E. (2016) The RING 2.0 web server for high quality residue interaction networks. J. Nucleic Acids Research., 44(W1): W367-W374.
[10] Wu Minggong, Ye Zelong, Wen Xiangxi, Jiang Xurui. (2020) Air traffic complexity recognition method based on complex networks. J. Journal of Beijing University of Aeronautics and Astronautics., 46(5): 839-850.
[11] Sarukkai R R. (2000) Link prediction and path analysis using Markov chains. J. Computer Networks., 33(1-6):377-386.
[12] O'Madadhain, J., Hutchins, J., Smyth, P. (2005) Prediction and ranking algorithms for event-based network data. In: Proceedings of the ACM SIGKDD 2005. New York. pp. 23-30.