Physica A 429 (2015) 177–183

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Accuracy test for link prediction in terms of similarity index: The case of WS and BA models

Min-Woo Ahn [a], Woo-Sung Jung [a,b,*]

[a] Department of Physics, Pohang University of Science and Technology, Pohang, 790-784, Republic of Korea

[b] Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, 790-784, Republic of Korea

HIGHLIGHTS

- Link prediction is applied to network model.
- Positive correlation between accuracy and mean degree is observed.
- AUC value is less dependent on network size than precision.

ARTICLE INFO

Article history:
Received 17 September 2014
Received in revised form 3 November 2014
Available online 17 February 2015

Keywords:
Link prediction
Similarity index
Accuracy metric
AUC value

ABSTRACT

Link prediction is a technique that uses the topological information in a given network to infer the missing links in it. Since past research on link prediction has primarily focused on enhancing performance for given empirical systems, negligible attention has been devoted to link prediction with regard to network models. In this paper, we thus apply link prediction to two network models: The Watts–Strogatz (WS) model and Barabási–Albert (BA) model. We attempt to gain a better understanding of the relation between accuracy and each network parameter (mean degree, the number of nodes and the rewiring probability in the WS model) through network models. Six similarity indices are used, with precision and area under the ROC curve (AUC) value as the accuracy metrics. We observe a positive correlation between mean degree and accuracy, and size independence of the AUC value.

1. Introduction

Complex networks have lately attracted considerable attention as a tool for understanding the structure of relation among agents in the system. Several systems, including social interaction network [1–4], the World-Wide Web [5–7], biological networks [8–11] and economic databases [12], have been studied to understand the structure of systems through the data. To obtain more accurate results from network analysis, we need to obtain more complete data. However, collecting data from the real world is difficult, such that the analyzed data is usually incomplete, i.e. the connection with the real world might not be represented in the data. We should check all unconnected node pairs in a network to distinguish them in a missing link (a certain relation that exists in the real world, but is not represented in the data) from a nonexistent link (when node pairs relate to each other neither in the real world nor the data). Link prediction can be used to render data collection

* Corresponding author at: Department of Physics, Pohang University of Science and Technology, Pohang, 790-784, Republic of Korea.
E-mail address: wsjung@postech.ac.kr (W.-S. Jung).

http://dx.doi.org/10.1016/j.physa.2015.01.083
0378-4371/

more efficient and the data thus collected more complete. Link prediction corresponds to a process for the calculation of the likelihood of the existence of a link based on observed links, and helps discover and restore the missing links. If the predictions are sufficiently accurate, we can efficiently identify missed connections by rechecking the data with this information.

Link prediction has been studied to find missing connections in an observed network [13–20]. Zhou et al. suggested a similarity index from the resource allocation process, and compared the results with other local similarity indices [14]. Liu and Lü suggested a few methods based on the local random walk process [15]. Similarity between two nodes was defined as the probability that one node can reach the other. This index can consider both global and local structures depending on the number of random walk steps. Clauset et al. employed a hierarchical random graph to understand the structure of real-world networks and infer missing links [13]. A hierarchical random graph is constructed from a set of nodes and the connection probabilities among the members of the set, represented by a dendrogram. Monte-Carlo simulation is used to obtain dendrograms from a given real-world network, and predicted missing links using the average connection probability calculated from the ensembles of dendrograms. Guimerà et al. developed a block model [20] that infers missing links as well as spurious connections (when nodes are connected in the data but no corresponding relation obtains in the real system). Lü and Zhou have also surveyed previous link prediction methodologies [21]. Link prediction using collaborative filtering, which has been applied to recommendation systems, can also be included into this class [16,17]. Missing link prediction in citation networks [18] and in Wikipedia [19] have also been performed.

Link prediction has also been applied to forecast emerging links from the network picture at a given point in time [22–24]. Liben-Nowell and Kleinberg applied link prediction to a collaboration network to predict future links [22], and employed and compared many kind of similarity indices. Newman discovered that the probability of a new connection in the collaboration network is proportional to the number of common neighbors [23]. D. Wang et al. used a similarity index and social network information based on human mobility pattern [24].

In spite of the above-mentioned research, many questions related to link prediction remain unanswered. Previous studies usually applied link prediction to real-world data. Although empirical tests are important to understand a given complex system and the methodology, they provide a limited perspective. As network models can provide network samples under varied conditions, understanding the methodology using basic network models is also important. In this paper, we focus on the statistical relation between network parameters and the accuracy of methodologies.

This paper is organized as follows. In Section 2, we describe link prediction in the network models. We employ two basic network models: the Watts–Strogatz (WS) model, and the Barabási–Albert (BA) model. Six indices are employed: Common Neighbor, Adamic–Adar, Resource Allocation, Jaccard, Preferential attachment, and Simrank. We use two metrics, precision and area under the ROC curve (AUC) value, as an accuracy measures. The results are detailed in Section 3 and discussed in Section 4. Finally, we summarize our findings in Section 5.

2. Methodology

The application of link prediction to network models consists of three steps: (1) missing link creation, (2) similarity calculation, and (3) accuracy calculation. The first step, the "preliminary step" of link prediction, is to create missing links in a given network. We randomly select and remove links, which are considered "missing" in later steps. The similarity calculation is the actual prediction step, where we form a list from the similarity index to find the missing links. Accuracy is calculated from the list in the final step.

2.1. Dataset

The control parameters provided are limited in real-world networks because they are difficult to control. For this reason, we employ two network models: the WS model [25] and the BA model [26]. The WS model, proposed by Watts and Strogatz, describes how real-world networks exhibit a small-world effect and high transitivity. The BA model, proposed by Barabási and Albert, provides an insight into how the degree distribution of real-world networks obeys power-law behavior. Network models provide various conditions and network ensembles for each condition, because of which we can observe the statistical relation between each network parameter and the accuracy of methodologies.

We control three parameters: The number of nodes N, the mean degree $\langle k \rangle$, and the rewiring probability p in the WS model. We set the parameters to $N_{WS} = N_{BA} = 1000$, $\langle k \rangle_{WS} = 6$ (in the WS model), $\langle k \rangle_{BA} = 10$ (in the BA model) and $p = 0.1$ as representative of default condition, and observe the correlation between each parameter and each accuracy metric. To calculate accuracy, we tested 1000 network ensembles for each network condition.

2.2. Missing link creation

To test link prediction in a static network, we applied the missing link creation process [13–15]. We can randomly divide the links into the training set (remaining link) and the probe set (missing link), where the links in the probe set are treated as the missing links. Approximately 10% of the links were removed and 100 independent creation processes were considered for each network ensemble.