Liming Pan, Lei Gao, Jian Gao. Link prediction in weighted networks via structural perturbations. In: 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE Press, 2017, pp. 5-8.

LINK PREDICTION IN WEIGHTED NETWORKS VIA STRUCTURAL PERTURBATIONS

LIMING PAN, LEI GAO, JIAN GAO

CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, China
E-MAIL: panlm99@gmail.com, gaoleiqr@126.com, gaojian08@hotmail.com

Abstract:

Link prediction aims at revealing missing and unknown information from observed network data, or predicting possible evolutions in near future. In recent years, extensive studies of link prediction algorithms have been performed on unweighted networks. However most empirical systems are necessarily to be described as weighted networks rather than solely the topology. In this paper we extend the structural perturbation method to weighted networks. We found that by including weight information the prediction accuracy can be significantly improved on networks with homogeneous weight distributions, meanwhile less improvements for heterogeneous weighted networks. Also we compared the weighted structural perturbation method to some benchmark algorithms, both weighted and unweighted, and found generally better performance in accuracy.

Keywords:

Link prediction; Weighted networks; Matrix perturbation

1. Introduction

Networks are natural descriptions and abstractions of systems consisting with large amounts of interacting individuals [1]. Unweighted networks, which generally give the interaction relations of these individuals, are usually only partial descriptions of focal systems. Thus weighted networks are introduced to capture the strength of the interactions in addition to the topology [2]. The interpretations of link weights depend on specific networks. Typically, it can be interaction frequencies in social networks, or trophic factors in food webs. Empirically for many networks which have been studied yet, it can be understood as the interaction strength of a pair of nodes.

Link prediction (LP) algorithms [3] are designed to find missing or unknown links in networks. LP algorithms can find rich applications in different scenarios. Examples include finding experimental errors in biological systems, or predicting latent future friendships in social networks, or revealing unaccessible hidden links in technical networks, to name just a few.

In earlier studies, most LP algorithms focus on predicting missing links solely based on topology. In principle if further including the weight information, the prediction accuracies should be improved, or at least no worse. However how to incorporate the weight information into the algorithm properly is an unsolved problem and has attracted lots of attention in recent studies [4].

In this paper we propose a LP algorithm based on structural perturbations of the weighted adjacency matrices [5]. Through extensive experiments we found that by including weights, the predicting accuracy can be significantly improved in networks with weights homogeneously distributed. However, for networks with heterogeneous weight distributions, the improvement is less apparent. Possible origins of this phenomenon are discussed in this paper. We also compare the weighted structural perturbation method (WSPM) to many other benchmark algorithms. In general, WSPM has superior performance or very close to the highest.

2. Method

WSPM is based on matrix perturbation techniques, and is an extension of the unweighted version proposed in [5]. An observed weighted network can be described by an matrix W^o, with W^o_{ij} the weight of link between nodes i and j. For each W^o, we define a new matrix called the rescaled weight matrix with elements $\widetilde{W}^o_{ij} = g(W^o_{ij})$. Motivations for doing this rescaling also the explicit form of the function $g(\cdot)$ will be found in next section.

To generate predictions, we first divide $\widetilde{W}^o$ into two independent parts with $\widetilde{W}^o = \widetilde{W}^u + \widetilde{W}^p$, where $\widetilde{W}^u$ is the unperturbed matrix and $\widetilde{W}^p$ the perturbation matrix. For $\widetilde{W}^u$ without degenerate eigenvalues, we perform the following matrix perturbation:

$$S^{WSPM} = \sum_{1}^{N} (\lambda_k + \delta\lambda_k) x_k x_k^T , \quad (1)$$

where λ_k and x_k are respectively the eigenvalues and the eigenvectors of $\widetilde{W}^u$, and

$$\delta\lambda_k = x_k^T \widetilde{W}^p x_k . \quad (2)$$

Then the matrix S^{WSPM} is applied for link prediction, with S_{ij}^{WSPM} the score of the link between nodes i and j to exist. We repeat the divide and perturbation procedure for 20 times and then use the average score for prediction. The intuition behind this operation is that, the principle or mechanism of generating the missing links should in some sense consist with the perturbation links $\widetilde{W}^p$. Note that for networks with decimal link weights, usually the degeneracies of eigenvalues are broken, thus we do not discuss the perturbation of matrices with degenerate eigenvalues. The necessary details can be found in [5].

3. On the rescaling of link weights

As discussed before, the interpretation of link weights depends on specific networks. Weights of links usually have complex interplays with the function and structure of networks. Thus we do not have a simple linear way to relate the weights of networks and the likelihood for a pair of nodes to be connected. Consider social networks and suppose we try to measure the strength of social ties. Clearly one cannot simply assume that the strength is linearly proportional to the interactions frequencies between friends. But one reasonable assumption is that, the strength of friendship is a non-decreasing function of the interaction frequencies. Thus to employ link weights to predict missing links, we need to do a proper rescaling. The rescaling function will change relative magnitudes of different link weights, but keeping the order of weights fixed. With the rescaling we wish to have the new weights that better fit our problem.

In this paper we choose the power function for rescaling:

$$\widetilde{W}_{ij} \equiv g(w_{ij}) = w_{ij}^{1/\alpha} . \quad (3)$$

The reason for the choice is that the power function does not have a typical scale. Distributions of weights can be roughly divided into two categories, which are homogeneous and heterogeneous. Typical distributions of link weights are shown in Fig.1. For homogeneous weighted networks, the weights have a typical scale, while for heterogeneous weighted networks, the largest and smallest values of weights can differ in magnitudes.

For other scaling functions like exponential function or sigmoid function, they do have a characteristic length scale. Thus they perform well in networks with homogeneous weight distributions. However, for heterogeneous networks, it's unreasonable to assume a typical scale. The results of other rescaling function have been verified through extensive experiments, and they are not present in the paper.

The rescaling function contains an undetermined parameter α and it will be discussed in Section 5.

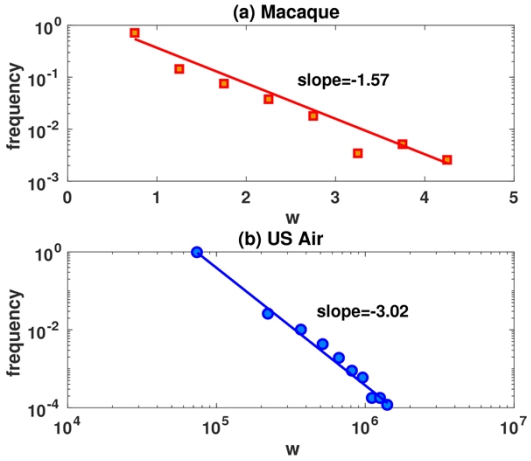

Fig.1 The weight distributions of two typical networks. (a) Macaque [6-7]. The weight distribution has a typical scale. The straight line is a linear fit in linear-log plot. (b) US Air [14]. The weight distribution is scale free. The straight line is a linear fit in log-log scale.

4. Results

To verify the performance of the algorithm, we divide a network W into two parts, namely the training network W^o and the probe network W^p with $W = W^o + W^p$. The cardinality of the probe network and the full network is related by $|W^p| = f|W|$ with f being a parameter controls the size of the probe set. The training network is assumed to be the known data and the probe network is for testing the algorithm. The predicting accuracy is then measured by the *Precision* metric, which defined as the normalized overlap of the highest score links and probe network links.

We perform the algorithm on seven networks, which are (1) Macaque [6-7], (2) Lesmis [8], (3) Highschool [7][9], (4) Residence [7][10], (5) Food Bay [11], (6) Food Mang [12] and (7) Neural [13]. The description of the networks can be found in the datasets and references therein.

We also compare WSPM to three similarity based methods CN, AA, RA, as well as their weighted versions. The definitions of these algorithms are given as follows:

$$S_{xy}^{CN} = |C(x,y)| , \quad (4)$$

$$S_{xy}^{WCN} = \sum_{z \in C(x,y)} W_{xz} + W_{yz} \,, \tag{5}$$

$$S_{xy}^{AA} = \sum_{z \in C(x,y)} 1/\log k_z \,, \tag{6}$$

$$S_{xy}^{WAA} = \sum_{z \in C(x,y)} (W_{xz} + W_{yz})/\log(1 + W_z) \,, \tag{7}$$

$$S_{xy}^{RA} = \sum_{z \in C(x,y)} 1/k_z \,, \tag{8}$$

$$S_{xy}^{WRA} = \sum_{z \in C(x,y)} (W_{xz} + W_{yz})/W_z \,. \tag{9}$$

Here, $C(x,y)$ is the set of common neighbors, k_z the degree of node z and W_z the node strength.

The predicting accuracies measured by precision are shown in Fig.2. The size of the probe network varies from 0.05 to 0.45 with 0.05 each step. To lighten the plots, we only show the accuracies of WSPM, SPM and the highest among all other six algorithms.

As we can see, WSPM is of highest accuracy for all six networks other than Lesmis. And for Lesmis, the accuracy still is very close to the highest (RA). This reveal that WSPM can successfully predict missing links in weighted network.

When comparing WSPM to SPM, we found that for the above four networks, the accuracy of WSPM is significantly improved. While for the below three networks, the accuracies are very close to SPM with only slight improvements, and the differences are smaller than the symbol size in the plot. A common feature of the above (below) networks is the weight distribution is homogeneous (heterogeneous), as shown in Fig.1(a) (Fig.1(b)).

There might be two possible reasons for this phenomenon. The first is that the rescaling function has not incorporate the weight information properly for heterogeneously distributed networks. The second is that, for heterogeneously distributed networks, the weight provide less new information in addition to the topology. Since for heterogeneously weighted networks, weights of several important links dominates all other links, and for those high weights links, their importance might be over emphasized. We still haven't got a definite answer to this question and leave it for future studies.

5. Sensitivity to the rescaling parameter

Next we study the dependence of the predicting accuracies on the rescaling parameter α. We plot the precision versus the rescaling parameterαin Fig.3. The blue line correspond to the precision with different rescaling parameters, red horizontal line the Precision of unweighted

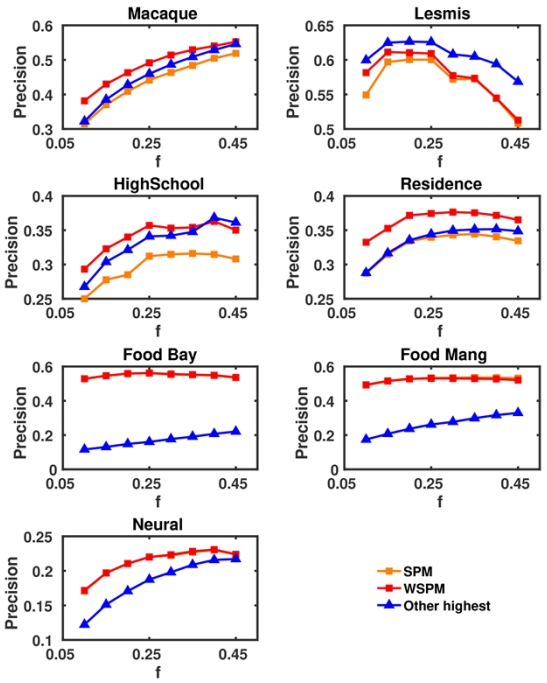

Fig.2 The predicting accuracy measured by precision for the seven networks.

SPM and black vertical line the optimal value of α. It can be seen that, for networks with homogeneous weight distributions, the optimal α is usually a small positive constant. Also the performance around the optimal value is relatively close to optimal. Thus for homogeneously distributed weighted networks, a reasonable way of finding a good rescaling parameter is to do a liner search in a region like [14]. A possible quicker convergence method is to do a linear search with finer and finer search regions. Since the algorithm is not sensitive to α around its optimal value, several steps of line search will result a reasonable value of the parameter.

For networks with heterogeneous weight distributions, the optimal parameter is relatively of a larger value. This indicates that the rescaled weighted are necessarily to become more homogeneous compared to the original weights to make good predictions. Since when α goes to infinity, the weights become identically equal to 1, and the WSPM reduces to the unweighted SPM. As discussed in Section 4, how to use the weight information in this case requires further studies.

6. Conclusions

In the paper we proposed a LP algorithm for weighted networks by using matrix perturbation techniques. The

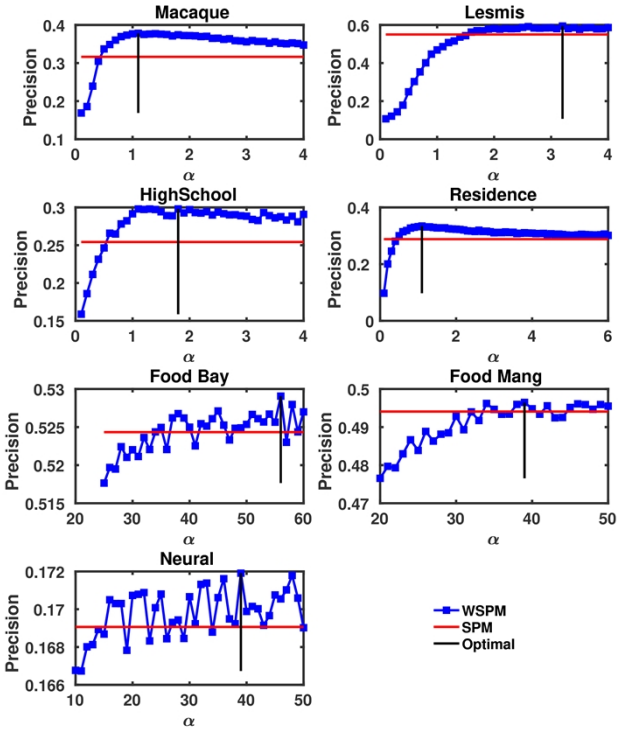

Fig.3 The precision of WSPM versus the rescaling parameter for the seven networks.

algorithm in general out performs six other methods, as well as the unweighted SPM method. Note that the WSPM method can also be applied to predict the weights of the links, or determine the existence and value of weights simultaneously.

Still there's one problem remaining to be solved when comparing the performance of WSPM and SPM. Under the current settings of weight rescaling, the improvement of predicting accuracy is much significant for homogeneous weighted networks. This is not the case for WCN, WAA, WRA compared to their corresponding unweighted version. One way to understand this is that, methods like CN are local information based methods, and weights does not essentially require a global rescaling. WSPM is based on the spectrum of networks, thus is a global method and require more reasonable pretreatment of the weights. How to incorporate link weights better for LP, especially for networks with heterogeneous weight distributions, demands further studies.

Acknowledgements

This work is supported in part by China Scholarship Council (CSC) under the Grant CSC No. 201606070053.

References

[1] A. Réka and A.-L. Barabási, "Statistical mechanics of complex networks", Reviews of Modern Physics, Vol 74, No. 1 pp. 47, 2002.

[2] B. Alain, et al, "The architecture of complex weighted networks", Proceedings of the National Academy of Sciences, USA, Vol 101, No. 11, pp. 3747-3752, 2004.

[3] L. Lü and T. Zhou, "Link prediction in complex networks: A survey", Physica A, Vol. 390, No. 6, pp. 1150-1170, 2011.

[4] J. Zhao et al, "Prediction of links and weights in networks by reliable routes", Scientific Reports, Vol. 5, pp. 12261, 2015.

[5] L. Lü et al, "Toward link predictability of complex networks", Proceedings of the National Academy of Sciences, USA, Vol. 112, No. 8, pp. 2325-2330, 2015.

[6] Macaques network dataset -- KONECT, April 2017.

[7] Y. Takahata, The Monkeys of Arashiyama. State University of New York Press, Albany, 1991.

[8] D. E. Knuth, The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley, MA, 1993.

[9] J. S. Coleman, Introduction to mathematical sociology. London Free Press Glencoe, London, 1964.

[10] L. C. Freeman, C. M. Webster, and D. M Kirke, "Exploring social structure using dynamic three-dimensional color images", Social Networks, Vol. 20, No. 2, pp. 109--118, 1998.

[11] D. Baird, J. Luczkovich, and R. R. Christian, "Assessment of Spatial and Temporal Variability in Ecosystem Attributes of the St Marks National Wildlife Refuge Apalachee Bay Florida", Estuarine, Coastal and Shelf Science. Vol. 47, pp.329, 1998.

[12] R. E. Ulanowicz, C. Bondavalli and M. S. Egnotovich, "Network Analysis of Trophic Dynamics in South Florida Ecosystem, FY 97: The Florida Bay Ecosystem", Tech. Rep. CBL 98, 1998.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of' small-world networks", Nature, Vol. 393 No. 6684, pp.440, 1998.

[14] V. Batageli. and A. Mrvar, Pajek datasets. Available at: http://vlado.fmf.unilj.si/pub/networks/data/mix/USAir97.net, 2006.