Local degree blocking model for missing link prediction in complex networks

Zhen Liu,1,a) Weike Dong,1 and Yan Fu1

1 Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China

(Dated: 10 June 2014)

Recovering and reconstructing incomplete networks by accurately identifying missing links is a vital task in the domain of network analysis and mining. In this article, by studying a specific local structure, namely a degree block having a node and its all immediate neighbors, we find it contains important statistical features of link formation in complex networks. We therefore propose a local blocking (LB) predictor to quantitatively identify missing links in given networks via local link density calculations. The promising experimental results performed on six real-world networks suggest that the new index can substantially outperform other traditional local similarity-based methods. After further analyzing the correlations between the LB scores and those given by two other methods, we find that the features of LB index are analogous to those of both PA index and short path based index, which empirically verify that the two aspects simultaneously captured by the LB index are jointly driving link formation in complex networks.

PACS numbers: PACS 89.75.Hc Networks and genealogical trees; PACS 89.20.Ff Computer science and technology
Keywords: Local degree block; Link prediction; Accuracy; Complex networks

Link prediction in complex networks is an open issue of data mining and knowledge discovery and studied wildly by researchers from disparate scientific communities in recent years. Studying link prediction problem has a profound scientific meaning in understanding networks growth and evolution. On the other hand, excellent link predictors also have a broad spectrum of applications in a wide range of domains, such as unveiling protein-protein interactions in biological networks, finding latent friendships between users in social networks, and understanding user’s purchase preference in E-commerce systems, etc. In this article, we proposed a novel link predictor which can accurately identify missing links in six real-world networks while has low time complexity in terms of our assessments. On the theoretical side, to the best of our knowledge, the new model is a firstly proposed multi-mechanism driven link predictor in the literatures which has provided a new sight into uncovering the subtle principle of link formation.

I. INTRODUCTION

Studying link formation mechanism is of significance in understanding the network growth and evolution. Conversely, uncovered link formation mechanism can also help and guide us to develop some new link prediction methods. For example, common neighbors method is originated from social balance mechanism and some machine learning based link prediction methods are developed by homophily mechanism. To fulfill the task of missing link prediction, two types of information are utilized widely including the entity or node’s property information and the network’s topological information. Compared with the network’s topological information, the entity’s property information such as users’ personal information in social networks or proteins’ functional attributes in protein-protein interaction networks may not be available for the reasons such as privacy preservation and absent or unreliable prior biological knowledge. Therefore, the network’s topological information is more preferable in most cases. Recent studies have revealed that some network’s topological properties can be used to fit the link formation likelihood. Aaron Clauset et al. analyzed the hierarchical structure of the networks and proposed a HRG model to estimate the link probability in a dendrogram. Camisano et al took into account the local community and proposed an efficient paradigm called LCP to calculate the link likelihood between pairs of nodes. Therefore, it has great potentials for us to further explore and study the correlations between network’s topological information and link formation.

So far, the network structure based link prediction methods can be mainly divided into two categories. One is using local network information to make a prediction whereas the other is using global network information to fulfill link prediction task. The link prediction methods using network’s global information are commonly more accurate but very expensive in computation and therefore hard to be applied to large-sized networks. Roger Guimerà et al. proposed a Stochastic Block Model (SBM) which is a typical link prediction model using global network information and is able to give very accurate link predictions on

a)Author to whom correspondence should be addressed. Electronic mail: quake.liu0625@gmail.com.
various kinds of networks. Liu et al. recently proposed a Fast Blocking probabilistic Model (FBM) based on a greedy strategy which can significantly improve the computational efficiency and has slightly better link prediction accuracy compared with the SBM. However, despite the significant reduction of implementation time compared with the SBM model, the network partitioning procedure used by the FBM still relies on the network’s global information, which means that the time consumption of computations in massive networks would be remaining a bottleneck for the FBM. The second type of structure based link prediction method is called local similarity based index (or proximity index) which has a big family including common neighbors, Adamic Adar, resource allocation etc. Such kind of method commonly has weaker link prediction performance but lower time complexity than global network information based methods for their simple computational paradigms by merely using local information of the network and is more suitable to implement link prediction in massive networks. According to the opposite characteristics of the two categories of link prediction methods, we can summarize that an ultimate goal to design an excellent link predictor is to well handle the dilemma between computational efficiency and link prediction accuracy. Therefore, to study a simple yet superior link prediction method is of profound theoretical and practical interests in this domain.

In this article, we inherit the computing framework of the FBM and design a novel link prediction index by using the local structural information of the network. According to our assessments, the new index performs better than the traditional local indices while has the same time complexity compared with them, therefore, it can easily implement the task of link prediction in massive networks. After a deeper analysis, we also found that the new index can capture two aspects of link formation in complex networks simultaneously, including large degree principle and short path principle.

II. METHOD

Analyses on diverse link distributions in real-world networks are commonly able to inspire us to find underlying link formation mechanisms. For example, the study of community structure enables us to understand that links are more likely to cluster in the communities while less likely to occur between the communities. In particular, we notice that, for a given network, each node and its immediate neighbors can be naturally treated together as a specific local structure. In a sense, a node owning a larger degree will be more likely to be connected by other nodes during the network evolution. This can be interpreted as a preferential attachment mechanism which was first introduced by Barabási et al. for addressing their well-known BA network model and has drawn a lot of attention by researchers from disparate scientific fields. Therefore, we assume that such a special structure which can be called a degree block in this article would carry some useful information to reflect the trend of link formation. In more detail, we’d like to investigate the statistical features of degree block structure in complex networks.

To quantitatively calculate the connecting likelihood for pairs of nodes in a degree block $x$, we introduce a simple measure called link density, which can be defined as

$$D_x = \frac{|E_x|}{|V_x||V_x| - 1} / 2,$$  \hspace{1cm} (1)

where $|E_x|$ and $|V_x|$ are the number of links and the number of nodes in the block $x$ respectively. The denominator of Eq. (1) denotes the maximal feasible number of links in the block. Moreover, for pairwise blocks, we can also calculate the link density between the two blocks $x$ and $y$, which can be defined as

$$D_{xy} = \frac{|E_{xy}|}{|V_x||V_y|},$$  \hspace{1cm} (2)

where $|E_{xy}|$ and $|V_x|$ denote the maximal feasible number of links between blocks $x$ and $y$. As it has high possibility that block $x$ and block $y$ are overlapped, the number of overlapped nodes will both count towards $|V_x|$ and $|V_y|$ illustrated in Fig. 1. Note that the Eq. (1) and Eq. (2) represent a general computational framework inherited from the literature.

For a network containing nodes with number $|V|$, it obviously has $|V|$ degree blocks. After obtained the link
density within and between all blocks by using Eq. (1) and Eq. (2) together in a given network, the score of link similarity for a non-adjacent nodes pair can be calculated as

\[ LB(u, v) = \sum_{u \in b_x, v \in b_y, x \neq y} D_{b_xb_y} + \sum_{u \in b_x, v \in b_x} D_{b_x}, \] (3)

where \( b_x \) and \( b_y \) denote all possible pairs of blocks which contain node \( u \) and node \( v \). According to Eq. (3), the link similarity can be calculated locally by merely utilizing node \( u \) and node \( v \)'s neighbor information. Thus, we call this proximity measure Local Blocking (LB) index. The whole procedure of similarity calculation for an observed network can be described in TABLE II.

According the descriptions of TABLE II, we can easily deduce that the complexity of the algorithm is \( O(|V|^2) \) which is identical to other proximity indices like CN, AA, etc.

### III. RESULTS

In this article, six real-world networks are considered to evaluate our new link prediction index. (1) Karate\[17\]. Social network of friendships between 34 members of a karate club at a US university in the 1970s. (2) CN Air\[18\]. The network of China air transportation system, which contains 121 airports and 733 airlines. (3) Infectious\[19\]. This network describes the face-to-face behavior of people during the exhibition INFECTIOUS: STAY AWAY in 2009 at the Science Gallery in Dublin. Nodes represent exhibition visitors; edges represent face-to-face contacts that were active for at least 20 seconds. Multiple edges between two nodes are possible and denote multiple contacts. The network contains the data from the day with the most interactions. After multiple edges between a pair of nodes are incorporated to one edge, the network finally contains 410 visitors and 2396 edges. (4) C. elegans\[20\]. The neural network of the nematode worm C. elegans, in which an edge joins two neurons if they are connected by either a synapse or a gap junction. (5) H. friends\[21\]. This social network contains friendships between users of the website hamsterster.com. (6) Wikivote\[22\]. Wikipedia is a free encyclopedia written collaboratively by volunteers around the world. Active users can be nominated to be administrator. A public voting begins after some users are nominated. Other users can express their positive, negative or neutral idea towards all the candidates. The most voted candidate will be promoted to administrator status. This process implies a social network in which users are nodes and the action of voting from someone to another demonstrates a directed link. In this article, we treat it as an undirected network. The basic topology statistics of the six networks are summarized in TABLE III.

Here, we apply two widely adopted accuracy measures to evaluate the missing link prediction performance of algorithms on six networks including AUC (area under the receiver operating characteristic curve)\[22\] and precision. AUC can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. In the implementation, among \( n \) times of independent comparisons, if there are \( n_1 \) times the missing link having higher score and \( n_2 \) times that they have the same score, the AUC value is

\[ AUC = \frac{n_1 + 0.5n_2}{n}. \] (4)

Different from AUC, precision only focuses on the \( L \) links with the highest scores. Among the top-\( L \) links, if \( L_r \) links are accurately predicted (there are \( L_r \) links in the probe set), then the precision equals \( L_r/L \).

The links randomly removed from the network constitute the probe set of missing links while the rest of the network constitutes the training set. In our tests, the fraction of links removed ranges from 10% to 90% (The interval is 10%). To ensure the results are of statistical significance, each value of AUC or precision is achieved by averaging over 100 implementations with independent random divisions of training set and probe set.

Here, four traditional proximity measures are considered for performance comparison including PA, CN, AA and RA. The PA measure\[22\] assumes that the probability of a future link between two nodes is proportional to their degrees’ product. Hence, it is defined as:

\[ PA(u, v) = |\Gamma(u)| \cdot |\Gamma(v)|, \] (5)

where \( |\Gamma(u)| \) is the number of node \( u \)'s neighbors. The CN measure is the application of the social balance theory stating that two nodes who share more neighbors tend to be connected. Formally, the measure is defined as:

\[ CN(u, v) = |\Gamma(u) \cap \Gamma(v)|. \] (6)

The AA measure\[22\] refines the CN by increasing the scores of pairs of nodes in which the more neighbors in common between them lead to less connections. It’s formally defined as:

\[ AA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log(|\Gamma(z)|)}. \] (7)

The RA measure\[21\] is a tiny revised version of AA, which is defined as:

\[ RA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|}. \] (8)

The accuracy comparison results are plotted in Fig. 2 and Fig. 3. The indices of CN, AA and RA can be
TABLE I. Description of the algorithm of local blocking

(a) Algorithm of local blocking

Input: an observed network formalized as a $|V| \times |V|$ matrix.
Output: link similarity matrix for all non-adjacent pairs of nodes.

- If there are $|V|$ nodes in the network, for each node $v_i$, treat it and its immediate neighbors as a degree block.
- Using Eq. (1) to calculate the link density for all the $|V|$ blocks. The number of operations is $|V|$.
- Using Eq. (2) to calculate the link density of pairwise blocks. The number of operations is $|V|(||V| - 1)/2$.
- Using Eq. (3) to calculate the scores of link similarity for non-adjacent pairs of nodes.

TABLE II. Topology statistics of six real-world networks

|       | Karate | CN Air | Infectious | C.elegans | H. friends | Wikivote |
|-------|--------|--------|------------|------------|------------|----------|
| $|V|$  | 34     | 121    | 410        | 297        | 1858       | 7115     |
| $|E|$  | 78     | 733    | 2396       | 2148       | 12533      | 103689   |
| $C$   | 0.588  | 0.788  | 0.385      | 0.308      | 0.167      | 0.209    |
| $D$   | 0.139  | 0.101  | 0.029      | 0.049      | 0.007      | 0.004    |
| $\langle d \rangle$ | 2.408 | 2.214 | 3.773 | 2.946 | 3.453 | 3.248 |
| $\langle k \rangle$ | 4.588 | 12.116 | 11.803 | 14.465 | 13.491 | 28.324 |

$|V|$ is the number of nodes and $|E|$ is the number of links between them. $C$ and $D$ are the average clustering coefficient and the density of the network, respectively. If a vertex $v_i$ has $k_i$ neighbours, $k_i(k_i - 1)/2$ edges could exist among the vertices within the neighbourhood. Thus, the local clustering coefficient for a network can be defined as $C_i = \frac{2|\{e_{v_i,:}, v_j\in N_i\}|}{k_i(k_i - 1)}$, where $N_i$ denotes the neighbours of $v_i$. $C$ is defined as $\frac{1}{|V|} \sum_{i=1}^{|V|} C_i$. $D$ denotes the density of the network which is defined as $2|E|/(|V|(|V| - 1)/2)$. ($\langle d \rangle$) and ($\langle k \rangle$) denote the average degree and the average shortest path distance, respectively.

regarded as one type of index for their AUC curves and precision curves are nearly the same in most networks. According to the AUC results, except for the comparable accuracy curves given by PA index in Karate and CN Air networks, LB index performs significantly better than other traditional proximity indices. Different to AUC measure which is originated from the statistical perspective, precision can be used to indicate the capacity of predicting more directly which degrees tending to be connected in real-world networks.

Due to the remarkable prediction performance shown by the LB index, we believe that the LB index has captured some latent topology features which impact on the link formation in complex networks. By analyzing the correlations between the LB similarity scores and those scores obtained by some other local similarity based methods such as CN, AA, etc., to our surprise, we find that LB similarity scores are strongly correlated with those scores given by PA index shown in Fig. 4. This indicates that the scores of LB index, to some extent, are analogous to those of PA index, and also demonstrates that the feature of pairs of nodes owning higher degrees tending to be connected in real-world networks can be captured by the LB index. But according to the comparison results shown in Fig. 2 and Fig. 3, LB index obviously performs better than the PA index in most tested networks. After further investigating the Fig 4, we notice that those pairs of nodes having identical PA scores usually have non-identical LB scores. This implies that LB index may have captured some other important features in the networks.

IV. ANALYSIS
FIG. 2. AUC comparisons for missing link prediction between LB, CN, AA, RA and PA approaches on six networks. Each of AUC values is averaged over 100 implementations and the error bar represents the standard deviation.

FIG. 3. Precision comparisons for missing link prediction between LB, CN, AA, RA and PA approaches on six networks. Each of precision value is averaged over 100 implementations and the error bar represents the standard deviation.

link formation factors. By observing the topology difference among these pairs of nodes having identical PA scores, we find that the main difference for them is the shortest path distance. The correlations between shortest path distance and LB scores for all the node pairs in the six networks illustrated in Fig. 5 indicate that pairs of nodes having longer shortest path distance would obtain lower LB scores. In fact, the short path feature is a basic idea of path-based proximity predictors, such as Katz and Local path (LP), which means that the LB index would be similar to these predictors as well. Qualitatively speaking, a score given by the LB index demonstrates that a pair of nodes having larger degrees and shorter shortest distance between them will be more similar to each other, thus, they are more likely to create a link.

According to our correlation analysis, we conclude that the link formation in real-world networks has two typical scenarios: i) The neighbor numbers of a node commonly represent its activity. More neighbors denote the node
is more active. Therefore, two active entities in a given network are very likely to form a link; ii) Long topology distance would hinder two nodes to form a link. The shorter the shortest path distance between two vertices is, the higher likelihood a link has to be established between them. Of course, the two factors may play different roles in different networks. For example, in social networks, large degree nodes may play a more important role because popular nodes or active nodes may attract much attention from other nodes, which is easier to be captured by the PA index. However, in Infectious networks, interactive infections would be the results of neighborhood influence in a short distance (commonly face to face). Therefore, in social networks, link formation procedure would be large degree first, whereas in infectious networks, link formation procedure would be short distance first. This is able to explain why PA index performs pretty bad in infectious network shown in Fig. 2 and Fig. 3 and the correlation distribution between LB scores and PA scores in Infectious network is more scattered than other results shown in the rest networks. In most cases, link formation would be the consequence of joint influence by multiple mechanisms. For example, beside the impact of large degree nodes, Leskovec et al’s study also shows that most new links in social networks span very short distances, typically closing triangles. Because of the two aspects of link formation simultaneously captured by the LB index in a balanced way, it has the capacity of performing better missing link prediction than other traditional proximity indices.

V. CONCLUSIONS

In this article, we proposed a degree blocking model by using network’s local topology information to predict missing links in various real-world networks including social networks, a physical network, a biological network and an epidemic contact network. Experimental results validate that, in most real-world networks, the new index outperformed other local similarity-based indices by two accuracy measures, namely AUC and Precision. According to our analysis, the LB index is essentially a balanced hybrid version of PA index and short path based index. In our opinion, to design a new link predictor by simply combining the two types of indices together would not be adapted to a wide range of networks because the two aspects usually play different roles in different networks and the binding or coupling pattern would be complicated (commonly nonlinear). However, a significant merit of the LB index is that it is able to fit the two aspects, without parameter tuning, to various kinds of networks properly in terms of our experiments. To the best of our knowledge, the LB index is the firstly proposed multi-mechanism driven link predictor which has presented some new evidences to support a widely accepted viewpoint that the process of link formation in complex networks would be a result of the joint influence of several mechanisms. Our work has provided new insights for researchers in developing some simple yet effective link prediction methods in the future.

ACKNOWLEDGMENTS

Z. L. wish to thank the funds of National Natural Science Foundation of China under grant Nos. 60903073, 61103109, the research funds for central universities under grant No. ZYGX2012J085 and Special Project of Sichuan Youth Science and Technology Innovation Research Team (2013TD0006).

1Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, “Link prediction in complex networks: A local naive bayes model,” EPL (Europhysics Letters) 96, 48007 (2011).

2Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, “Link prediction and recommendation across heterogeneous social networks,” in Data Mining (ICDM), 2012 IEEE 12th International Conference on (IEEE, 2012) pp. 181–190.

3A. Popescul and L. H. Ungar, “Statistical relational learning for link prediction,” in IJCAI workshop on learning statistical models from relational data, Vol. 2003 (Citeseer, 2003) pp. 81–87.

4M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using supervised learning,” in SDM06: Workshop on Link Analysis, Counter-terrorism and Security (2006).

5D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” Journal of the American society for information science and technology 58, 1019–1031 (2007).

6A. Clauset, C. Moore, and M. E. Newman, “Hierarchical structure and the prediction of missing links in networks,” Nature 453, 98–101 (2008).

7C. V. Cannistraci, G. Alainis-Lobato, and T. Ravasi, “From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks,” Scientific reports 3 (2013).

8R. Guimerà and M. Sales-Pardo, “Missing and spurious interactions and the reconstruction of complex networks,” Proceedings of the National Academy of Sciences 106, 22073–22078 (2009).
FIG. 4. Correlation distribution between PA scores and LB scores and the red lines are the trend line fitted by Generalized Additive Model (GAM) plot smoothing. The distribution of PA scores and LB scores is plotted in double-logarithmic scale.

FIG. 5. Correlation distribution between Shortest path distances and LB scores for all node pairs and the red lines are the trend line fitted by Generalized Additive Model (GAM) plot smoothing. The distribution of shortest path distances and LB scores is plotted in double-logarithmic scale.

9Z. Liu, J.-L. He, K. Kapoor, and J. Srivastava, “Correlations between community structure and link formation in complex networks,” PloS one 8, e72908 (2013).

10L. A. Adamic and E. Adar, “Friends and neighbors on the web,” Social networks 25, 211–230 (2003).

11T. Zhou, L. Lü, and Y.-C. Zhang, “Predicting missing links via local information,” The European Physical Journal B 71, 623–630 (2009).

12A. Clauset, “Finding local community structure in networks,” Physical review E 72, 026132 (2005).

13M. Newman, “Communities, modules and large-scale structure in networks,” Nature Physics 8, 25–31 (2012).

14G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” Nature 435, 814–818 (2005).

15F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, “Defining and identifying communities in networks,” Proceedings of the National Academy of Sciences of the United States of America 101, 2658–2663 (2004).

16A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” science 286, 509–512 (1999).
17 W. Zachary, “An information flow model for conflict and fission in small groups,” Journal of anthropological research 33, 452–473 (1977).

18 H.-K. Liu, L.-Y. Lü, and T. Zhou, “Uncovering the network evolution mechanism by link prediction (in Chinese),” Sci Sin Phys Mech Astron 41, 816–823 (2011).

19 L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, “What’s in a crowd? analysis of face-to-face behavioral networks,” Journal of theoretical biology 271, 166–180 (2011).

20 J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, “The structure of the nervous system of the nematode caenorhabditis elegans,” Philosophical Transactions of the Royal Society of London. B, Biological Sciences 314, 1–340 (1986).

21 “Hamsterster friendships network dataset,” (2014).

22 J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” in Proceedings of the 19th international conference on World wide web (ACM, 2010) pp. 641–650.

23 J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Signed networks in social media,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM, 2010) pp. 1361–1370.

24 D. J. Watts and S. H. Strogatz, “Collective dynamics of small-world networks,” nature 393, 440–442 (1998).

25 J. A. Hanley, B. J. McNeil, et al., “A method of comparing the areas under receiver operating characteristic curves derived from the same cases,” Radiology 148, 839–843 (1983).

26 M. E. Newman, “Clustering and preferential attachment in growing networks,” Physical Review E 64, 026102 (2001).

27 C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, “Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding,” Bioinformatics 29, i199–i209 (2013).

28 L. Katz, “A new status index derived from sociometric analysis,” Psychometrika 18, 39–43 (1953).

29 P. Holme and B. J. Kim, “Growing scale-free networks with tunable clustering,” Physical review E 65, 026107 (2002).

30 J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, “Microscopic evolution of social networks,” in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, 2008) pp. 462–470.

31 P. Symeonidis, E. Tiakas, and Y. Manolopoulos, “Transitive node similarity for link prediction in social networks with positive and negative links,” in Proceedings of the fourth ACM conference on Recommender systems (ACM, 2010) pp. 183–190.

32 C. Wang, V. Satuluri, and S. Parthasarathy, “7th IEEE international conference on data mining,” in 7th IEEE International Conference on Data Mining (IEEE Press, 2007) pp. 1–10.