A parameter free similarity index based on clustering ability for link prediction in complex networks

Zhihao Wu, Youfang Lin, and Yao Zhao

Beijing Key Lab of Traffic Data Analysis and Mining,
School of Computer and Information Technology,
Beijing Jiaotong University

Link prediction in complex networks based on solely topological information is a challenging problem. In this paper, we propose a novel similarity index, which is efficient and parameter free, based on clustering ability. Here clustering ability is defined as average clustering coefficient of nodes with the same degree. The motivation of our idea is that common-neighbors are able to contribute to the likelihood of forming a link because they own some ability of clustering their neighbors together, and then clustering ability defined here is a measure for this capacity. Experimental numerical simulations on both real-world networks and modeled networks demonstrated the high accuracy and high efficiency of the new similarity index compared with three well-known common-neighbor based similarity indices: CN, AA and RA.

PACS numbers: 89.75.Hc, 89.20.Hh

I. INTRODUCTION

Many complex systems can be modeled using complex networks, such as social, biological and information systems, and the study of complex networks has attracted increasing attention and becomes a popular tool in many different branches of science [1-8]. Link prediction in complex networks aims at estimating the likelihood of the existence of a link between two nodes, and it has many applications in different fields. For example, predicting whether two users know each other can be used to recommend new friends in social networking sites, and in the field of biology, accurate prediction of protein-protein interaction has great value to sharply reduce the experimental costs, especially when our knowledge is very limited. For example, 80% of the molecular interactions in cells of Yeast [8] and 99.7% of human [9] are still unobserved. Large amount of missing links make the observed networks sparser, which of course leads to more difficulty to predict.

The problem of link prediction can be defined in different backgrounds considering different information. In this paper, we focus on link prediction relying on solely topological information. There are three main kinds of methods: local, global and quasi-local methods [10]. Local methods [10-13] are always very efficient, while global ones [14-17] are more accurate. Quasi-local methods, such as Ref. [18], can be designed by adding some constraints on global ones, but always bring some parameters.

CN [10] is the most well-known local similarity index, which regards pair of nodes with more common neighbors is more likely connected by a link. The drawback of CN is that all common neighbors are treated as the same, so that many pairs of nodes are given the same likelihood. AA [11] and RA [12] indices solve this problem by distinguishing common-neighbors by node-degree in different way. Some prediction results have shown AA and RA can predict missing links more accurately than CN for better resolution. But in some cases, degree is still limited. For example, the max degree of some large sparse networks is not very big, so degree is not able to provide sufficient resolution in these cases. In this paper, we will focus on this problem and give a novel similarity index with better discriminative resolution.

Besides, there are also some sophisticated models to solve the problem of link prediction. Clauset et al. proposed an algorithm based on the hierarchical network structure, which gives good predictions for the networks with hierarchical structures [19, 20]. Guimera et al. solved this problem using stochastic block model [21]. Recently, Linyuan Lü et al. propose a concept of structural consistency, which can reflect the inherent link predictability of a network, and they also propose a structural perturbation method for link prediction, which is more accurate and robust than the state-of-the-art methods [22]. Not only these methods can give good prediction results in some networks, another significance of these methods is to give insights into the mechanism of link formation, network evolution, and even the link predictability [22]. However, there should be more effort to make these methods efficient enough.

In this paper, we will propose a simple, efficient and parameter free similarity index based on clustering ability, which is defined by average clustering coefficient of nodes with the same degree. The study of Liben-Nowell and Kleinberg showed that CN and AA perform better than other seven well-known local similarity indices. Some later literature showed that RA perform even better than CN and AA. So we compare our method with the three state-of-the-art common-neighbor based indices in exper-
ments. Experimental results on 12 real-world networks drawn from 6 different fields show that our method outperforms other compared well-known common-neighbor based methods. Especially, we find the new similarity index can always give impressive promotion in sparse networks with low average clustering coefficient. Further, we verify this point employing a tunable network model.

II. METHODS

Considering an unweighted undirected simple network $G(V,E)$, where $V$ is the set of nodes and $E$ is the set of links. For each pair of nodes, $x, y \in V$, we assign a score $s_{xy}$. Since $G$ is undirected, the score is symmetry. All the nonexistent links are sorted in decreasing order according to their scores, and the links in the top are most likely to exist. The common-used framework always sets the similarity to the score, so the higher score means the higher similarity, and vice versa.

A. Compared similarity indices

In this paper, we will compare our method with three well-known similarity indices: CN, AA and RA. Their definitions and relevant motivations are introduced as follows:

1) CN (common neighbors) For a pair of nodes, CN counts the number of common neighbors. In common sense, more common neighbors indicates larger probability to form/exist a link between two nodes. The definition of CN is given in equation (1), in which $\Gamma(x)$ denotes the set of neighbors of node $x$.

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|$$

(1)

2) AA (Adamic-Adar). AA refines the simple counting of common neighbors by assigning the less-connected nodes more weight, and is defined as equation (2).

$$s_{xy}^{AA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\log(k_z)}$$

(2)

3) RA (Resource allocation). RA is also an index based on common neighbors, and the motivation comes from resource allocation dynamics on complex systems. For a pair of unconnected nodes, $x$ and $y$, the node $x$ can send some resource to $y$, with their common neighbors playing the role of transmitters. In the simplest case, assume that each transmitter has a unit of resource, and will equally distribute the resource to all its neighbors. The similarity can be defined as given by equation (3), which measures the amount of resource $y$ received from $x$. Comparing with AA (Adamic-Adar), which only simply replaces $k_z$ by $\log(k_z)$, the little difference only makes the results different significantly when the degrees of common neighbors are comparatively high.

$$s_{xy}^{RA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{k_z}$$

(3)

In respect to CN, most other CN-based similarity indices are design by weakening high-degree nodes or bring some other link or structural information into the definition of measures. In one word, all these motivations can be summed up in offering more discriminative resolution.

B. The new similarity index

The new index is called CA (clustering ability) and its definition is given in equation (4).

$$s_{xy}^{CA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} C(k_z)$$

(4)

where $C(k_z)$ is the average clustering coefficient of nodes with degree equal to $k_z$. The clustering coefficient of a node is defined in equation (5).

$$C_i = \frac{2t_i}{k_i(k_i - 1)}$$

(5)

where $t_i$ is the number of triangles passing through node $i$ and $k_i$ is the degree of node $i$.

The motivation of the CA index comes from the assumption that common-neighbors can contribute likelihood to a pair of nodes, because they have some ability to cluster other nodes together. Please note clustering ability has some difference with the current clustering situation of a node. For example, we cannot say a node with only two neighbors that are not connected has no ability to cluster its neighbor together. On the contrary, if the link between its neighbors is unobserved or missing, we will totally make a wrong judgement. Thus, we use the average clustering coefficient of nodes with the same degree to estimate the clustering ability of a node, and we think it is a more robust way.

One may ask whether there is a big difference between $C(k)$, $1/k$ and $1/\log(k)$. The answer is definitely yes. We plot $C(k)$ versus $k$ for all tested networks in Figure 1 and Figure 2. It is easy to find in most cases their distributions on $k$ are very different. There are two main aspects: 1) when degrees are relatively small, $C(k)$ not only does not decrease as fast as $1/k$ or $1/\log(k)$, but even increases a bit for some networks, such as PPI1 and PPI2 networks. In most cases, $C(k)$ distributes a little flat or decreases very slow when $k$ is small; 2) when degrees are large, the distribution of $C(k)$ can be very broad, instead of a straight line (RA) or a curve (AA). That means nodes with similar degrees can provide very different contributions in forming a link.
C. Estimation

To estimate the predicted results comprehensively, here we employ two estimators: AUC and precision, which are commonly used in other related literatures [18, 22]. The basic preparation for calculation of the two methods is the same. To test the precision of a prediction algorithm, the observed links \( E \) is randomly divided into two parts: the training set \( E_t \) is treated as known information, while the probe set \( E_p \) is used for testing and no information in the probe set is allowed to be used for prediction. Obviously, \( E = E_t \cup E_p \) and \( E_t \cap E_p = \text{null} \). In this paper, we consider 10% of links as test links.

AUC is a standard metric, the area under the receiver operating characteristic (ROC) curve, to quantify the accuracy of the prediction algorithms. In this situation, it can be interpreted as the probability that a randomly chosen missing link (belonging to \( E_p \)) is given a higher score than a randomly chosen nonexistent link (which belongs to \( U - E \), where \( U \) denotes the set of all node-pairs). In practice, the calculation of AUC is given as defined by equation (6), where \( n \) is the times of independent comparisons, \( n' \) denotes the times the missing links having a higher score, and \( n'' \) counts situations they have the same score. A higher value of AUC indicates better results.

\[
AUC = \frac{n' + 0.5n''}{n} \quad (6)
\]

Given the ranking of the non-observed links, the precision is defined as the ratio of relevant items selected to the number of items selected. That means if we take the top-\( L \) links as the predicted ones, among which \( L_r \) links are right, then the precision can be defined as equation (7). Higher precision indicates higher prediction accuracy.

\[
\text{precision} = \frac{L_r}{L} \quad (7)
\]

III. EMPIRICAL ANALYSIS

A. Tests on real-world networks

In this paper, we will compare CA index with three other well-known common-neighbor based similarity indices on 12 real-world networks drawn from various fields. PPI1 [23] and PPI2 [23, 24] are two protein-protein interaction networks. Food [25] and Grassland [26] are two food web networks. Dolphins [27] and Jazz [28] are dolphins and musician social networks. MacNeu [29] and MouseNeu [30] are two neural networks. PB [31] and Email [32] are two social networks from electronic information systems. Grid [33] and INT [34] are two artificial infrastructure networks. The basic topological features of these networks are given in Table I.

In Figure 4, we plot the corresponding measure of common-neighbor’s contribution for CA, RA and AA in the twelve real-world networks. It shows that there are big differences among the three indices, whether for small-degree nodes or large-degree nodes. For nodes with small degree, \( C(k) \), which indicates average clustering coefficient of nodes with degree \( k \), always decreases very slowly with the increase of degree, and for PPI networks the trend is even in the opposite direction. For nodes with large degree, \( C(k) \) always distributes more broad, rather than as a line or a curve.

First, we show the experimental results of the four efficient methods evaluated by AUC in Table III with those entries corresponding to the highest accuracies being emphasized by black. All experimental results in this paper are average of 100 runs. RA performs the best among all methods. CA performs the second-best and is slightly worse than RA. On five networks, AA, RA and CA get the same results. On the rest of networks, RA gets four best results and CA gets another three best results.

In Table III shows the precision results of the compared
indices on the 12 networks. Clearly, CA performs better than all other three indices with a distinct advantage. Among all tested networks, CA can predict more precisely than AA and RA. Only on PB network, CN gets a better result than CA. At the same time, big improvements are attained in many networks, such as PPI1, PPI2, Food, Grid and INT networks. One may note that all the above five networks have two common statistical features: low average clustering coefficient and relatively low average degree. If we look back at these networks in Figure 1, we can find an interesting phenomenon: these networks are exactly the ones in which the distributions of $C(k)$ are very different from $1/k$ and $1/\log(k)$. We figure that this goes to show the capacity of CA in better estimating the contribution of common-neighbors. The large improvements may have relations with both low average clustering coefficient and low average degree, although it is hard to give a definite theoretical analysis. We will give some more evidence on this issue employing a network model in the following section.

### B. Tests with PS model

To demonstrate the relationship between the advantages of CA index and network features, we test the above four similarity indices on artificial networks generated by Popularity versus Similarity (PS) model [35]. PS model considers the factor of popularity and similarity at the same time in the growing procedure of a network, so that it can generate networks with more similar features with those of real-world networks than some other well-known models, such as WS model [33], BA model [36] and etc.

Two parameters of PS model are tuned in our analysis: one is the temperature parameter $T$, which can be used to tune the average clustering coefficient of the generated network, and the other one is $m$, which is a parameter controlling the average node degree $<k> = 2m$. The parameter $T$ ranges from 0 to 1 and a higher value corresponds to lower clustering coefficient. We set two groups of networks with $m$ equal to 3 and 9, indicating spare

| Network | Prec | CN  | AA  | RA  | CA  |
|---------|------|-----|-----|-----|-----|
| PPI1    | 0.184| 0.145| 0.080| 0.202|
| PPI2    | 0.236| 0.196| 0.109| 0.240|
| Food    | 0.007| 0.006| 0.008| 0.023|
| Grassland | 0.064| 0.126| 0.126| 0.126|
| Dolphins | 0.128| 0.114| 0.095| 0.130|
| Jazz    | 0.821| 0.898| 0.824| 0.850|
| MacNeu  | 0.574| 0.578| 0.554| 0.604|
| MouseNeu| 0.046| 0.047| 0.047| 0.052|
| PB      | 0.418| 0.380| 0.252| 0.400|
| Email   | 0.203| 0.320| 0.256| 0.328|
| Grid    | 0.120| 0.098| 0.080| 0.131|
| INT     | 0.105| 0.104| 0.083| 0.110|

FIG. 1. Common-neighbor’s contribution versus degree for CA, RA and AA in 12 real-world networks. $C(k)$ indicates the average clustering coefficient of nodes with degree $k$.  

TABLE III. Link prediction accuracy measured by precision on 12 real-world networks.
and dense networks respectively. For each group, we give three different values of temperature parameter $T$, as 0.1, 0.5 and 0.9, to generate networks with different average clustering coefficient. For each combination of the two parameters we generate 10 networks, and link prediction algorithms run 10 times for each realization. The average statistical features of their giant connected components are given in Table IV.

We also plot $C(k)$, $1/k$ and $1/log(k)$ versus degree $k$ for the artificial networks in Figure 2. Among these networks, the one with $m$ equal to 3 and $T$ equal to 0.9 is our most interested network, which is sparse and with low average clustering coefficient. More importantly, we find that the distribution of $C(k)$ versus degree is very similar to what we see in Figure 1, i.e. for nodes with small degree, the trend of $C(k)$ is a bit ascending and for nodes with large degree, $C(k)$ distributes more broadly with the growth of degree $k$.

Table IV and Table V show the link prediction results of four similarity indices on these artificial networks with different features. For results estimated under AUC, RA outperforms AA and RA by a large rate of 40% and 61.5%, respectively. While on dense networks, differences among the four indices are very small. Thus what we find on the real-world networks are well verified by the test results on artificial networks generated by PS model.

| parameters | $N$ | $M$ | $<k>$ | $<d>$ | $<C>$ |
|------------|-----|-----|-------|-------|-------|
| m=3 T=0.1 | 946.1 | 1960.8 | 6.384 | 7.099 | 0.783 |
| m=3 T=0.5 | 969.7 | 3040.2 | 6.271 | 7.321 | 0.416 |
| m=3 T=0.9 | 848.3 | 1459.9 | 3.442 | 4.554 | 0.071 |
| m=9 T=0.1 | 1000 | 9720 | 19.440 | 1.982 | 0.851 |
| m=9 T=0.5 | 1000 | 8912.9 | 17.826 | 2.196 | 0.532 |
| m=9 T=0.9 | 993.9 | 4026.9 | 8.103 | 3.029 | 0.160 |

TABLE V. Link prediction accuracy measured by AUC on artificial networks generated by PS model.

| AUC | CN | AA | RA | CA |
|-----|----|----|----|----|
| m=3 T=0.1 | 0.961 | 0.978 | 0.98 | 0.98 |
| m=3 T=0.5 | 0.859 | 0.872 | 0.873 | 0.873 |
| m=3 T=0.9 | 0.606 | 0.608 | 0.608 | 0.607 |
| m=9 T=0.1 | 0.986 | 0.994 | 0.996 | 0.996 |
| m=9 T=0.5 | 0.924 | 0.943 | 0.948 | 0.948 |
| m=9 T=0.9 | 0.712 | 0.72 | 0.72 | 0.72 |

TABLE VI. Link prediction accuracy measured by precision on artificial networks generated by PS model.

| Prec | CN | AA | RA | CA |
|------|----|----|----|----|
| m=3 T=0.1 | 0.482 | 0.663 | 0.687 | 0.657 |
| m=3 T=0.5 | 0.174 | 0.2 | 0.193 | 0.211 |
| m=3 T=0.9 | 0.036 | 0.03 | 0.028 | 0.041 |
| m=9 T=0.1 | 0.962 | 0.984 | 0.986 | 0.985 |
| m=9 T=0.5 | 0.352 | 0.363 | 0.369 | 0.37 |
| m=9 T=0.9 | 0.101 | 0.101 | 0.091 | 0.099 |

C. Runtime

At last, we show the efficiency of CA index in Table VII. Since the rest predicting procedures are the same for different similarity indices in the link prediction framework we used if similarity matrix is prepared, we only show the time cost of calculating the similarity matrix for the four indices. Clearly, CN runs fastest among the four indices. In most cases, CA can run competitively fast comparing with AA and RA. The most complex part of CA is the calculation of clustering coefficient, which has a computational complexity of $O(Nd_{max}^2)$, where $d_{max}$ is the max degree of a network. Therefore, CA is very efficient, especially for sparse networks.

| networks | CA | CN | AA | RA |
|----------|----|----|----|----|
| PPI1 | 890.14 | 69.68 | 1223.93 | 906.19 |
| PPI2 | 1083.52 | 81.74 | 1398.48 | 1060.61 |
| Food | 5.54 | 0.14 | 0.45 | 0.40 |
| Grassland | 5.99 | 0.11 | 0.47 | 0.48 |
| Dolphins | 6.51 | 0.13 | 0.49 | 0.43 |
| Jazz | 32.14 | 2.24 | 4.62 | 4.17 |
| MacNeu | 15.33 | 1.03 | 1.70 | 1.65 |
| MouseNeu | 2.08 | 0.08 | 0.25 | 0.24 |
| PB | 405.91 | 44.25 | 89.83 | 70.74 |
| Email | 182.46 | 10.13 | 84.57 | 79.04 |
| Grid | 1410.01 | 64.26 | 1185.65 | 377.06 |
| INT | 833.25 | 65.36 | 2039.75 | 393.49 |

IV. CONCLUSION

In this paper, we present a novel efficient and parameter free common-neighbor based similarity index, called CA. The main difference from other well-known common-neighbor based similarity indices lies in the way of evaluating common-neighbor’s contribution. CA index assumes common-neighbor with higher clustering ability contributes more to the likelihood of forming a link between a pair of nodes, and here average clustering coefficient of nodes with the same degree is used to mea-
FIG. 2. Common-neighbor’s contribution versus degree for CA, RA and AA in artificial networks. $C(k)$ indicates the average clustering coefficient of nodes with degree $k$.

sure the clustering ability of nodes with this degree. In both real-world networks and artificial networks, we find our measure of common-neighbor’s contribution is very different with those of AA and RA, and the bigger the differences are, the better CA performs than AA and RA.

Experimental results on both real-world networks and artificial networks show that CA index outperforms state-of-the-art common-neighbor based similarity indices in precision, especially on sparse networks with low average clustering coefficient. Although the calculation of clustering coefficient in CA index needs more time than AA and RA, the time costs of CA on real-world networks show that it is still a very efficient index. The computational complexity of clustering coefficient is $O(N d_{max}^2)$, where $d_{max}$ is the max node degree of a network. Thus, the computational complexity of CA is nearly $O(N)$ on sparse networks.

**ACKNOWLEDGMENTS**

The authors acknowledges Linyuan Lü, Tao Zhou and Carlo Vittorio Cannistraci for supporting some source codes and datasets used in this paper. This work is supported by the National Natural Science Foundation of China (Grants No. 61403023)

[1] R. Albert and A.-L. Barabási, Reviews of modern physics 74, 47 (2002).
[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Physics reports 424, 175 (2006).
[3] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Advances in Physics 56, 167 (2007).
[4] G. Bianconi, P. Pin, and M. Marsili, Proceedings of the National Academy of Sciences 106, 11433 (2009).
[5] H.-W. Shen and A.-L. Barabási, Proceedings of the National Academy of Sciences 111, 12325 (2014).
[6] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gobrech, N. Li, N. Simonis, et al., Science 322, 104 (2008).
[7] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wuf, Proceedings of the National Academy of Sciences 105, 6959 (2008).
[8] D. Liben-Nowell and J. Kleinberg, Journal of the American society for information science and technology 58, 1019 (2007).
[9] L. Lü and T. Zhou, Physica A: Statistical Mechanics and its Applications 390, 1150 (2011).
[10] F. Lorrain and H. C. White, The Journal of mathematical sociology 1, 49 (1971).
[11] L. A. Adamic and E. Adar, Social networks 25, 211 (2003).
[12] T. Zhou, L. Lü, and Y.-C. Zhang, The European Physical Journal B-Condensed Matter and Complex Systems 71, 623 (2009).
[13] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, Scientific reports 3 (2013).
[14] L. Katz, Psychometrika 18, 39 (1953).
[15] G. Jeh and J. Widom, in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, 2002) pp. 538–543.
[16] F. Göbel and A. Jagers, Stochastic processes and their applications 2, 311 (1974).
[17] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, Knowledge and data engineering, IEEE Transactions on 19, 355 (2007).
[18] L. Lü, C.-H. Jin, and T. Zhou, Physical Review E 80, 046122 (2009).
[19] A. Clauset, C. Moore, and M. E. Newman, Nature 453, 98 (2008).
[20] S. Redner, Nature 453, 47 (2008).
[21] R. Guimerà and M. Sales-Pardo, Proceedings of the National Academy of Sciences 106, 22073 (2009).
[22] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, Proceedings of the National Academy of Sciences 112, 2325 (2015).
[23] A. Ben-Hur and W. S. Noble, Bioinformatics 21, i38 (2005).
[24] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, Bioinformatics 22, 1998 (2006).
[25] J. E. Cohen, D. N. Schittler, D. G. Raffaelli, and D. C. Reuman, Proceedings of the National Academy of Sciences 106, 22335 (2009).
[26] H. A. Dawah, B. A. Hawkins, and M. F. Claridge, Journal of animal ecology, 708 (1995).
[27] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology 54, 396 (2003).
[28] P. M. Gleiser and L. Danon, Advances in complex systems 6, 565 (2003).
[29] R. Köter, Neuroinformatics 2, 127 (2004).
[30] D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, and R. C. Reid, Nature 471, 177 (2011).
[31] L. A. Adamic and N. Glance, in Proceedings of the 3rd international workshop on Link discovery (ACM, 2005) pp. 36–43.
[32] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, Physical review E 68, 065103 (2003).
[33] D. J. Watts and S. H. Strogatz, nature 393, 440 (1998).
[34] N. Spring, R. Mahajan, and D. Wetherall, in ACM SIGCOMM Computer Communication Review, Vol. 32 (ACM, 2002) pp. 133–145.
[35] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguná, and D. Krioukov, Nature 489, 537 (2012).
[36] A.-L. Barabási and R. Albert, science 286, 509 (1999).