An Improved Link Prediction Algorithm Based on Degrees and Similarities of Nodes

Qingshuang Sun, Rongjing Hu, Zhao Yang, Yabing Yao, Fan Yang
School of Information Science & Engineering
Lanzhou University
Lanzhou, Gansu Province, China
{sunqsh2015, hurj, yangzhao14, yaoyb14,yangfan2014 }@lzu.edu.cn

***Abstract*—Link prediction is to calculate the probability of a potential link between a pair of unlinked nodes in the future. It has significance value in both theoretical and practical. The similarity of two nodes in the networks is an essential factor to determine the probability of a potential link between them. One of the important methods with the similarity of two nodes is to consider common neighbors of two nodes. However, the number of common neighbors only describes a kind of quantitative relationship without taking into account the topology of given networks and the information of local structure which consist of a pair of nodes and their common neighbors. Therefore, we introduce the concept of the degrees of nodes and the idea of community structure and propose a new similarity index, namely, local affinity structure(LAS). The LAS method describes the closeness of a pair of nodes and their common neighbors. We evaluated LAS on twelve different networks compared with other three similarity based indexes which consider the degree of nodes. From the experimental results, our method shows obvious superiority in improving the accuracy of link prediction.**

Keywords—link prediction, common neighbors, degrees of nodes, community structure

I. Introduction

Link prediction builds a bridge between complex networks and information science which deals with the prediction and restoration of missing information. In the study of complex networks, many of the networks which can be observed are incomplete, and there are still parts of networks' properties unknown. Link prediction is researched to fill the missing links and identify the spurious link, etc. by calculating the possibility of missing links between a pair of nodes using the information of nodes and the structure of networks. There are various methods to solve this predicting problem and the data in the networks can be depicted in different ways. In computer science, the main approach based on the Markov chains and machine learning. Sarukkai applied Markov chains to predict the links and analysis paths of the networks [1]. Then Zhu extended the link prediction method based on Markov chains to the prediction of WWW networks [2]. The accuracy of link prediction by using the information of nodes properties and other external information, nevertheless, it is difficult to get these information, sometimes, impossible. Compared with properties of nodes, it is easier to obtain more reliable information of the structure of networks. Recently, more and more attention has been paid to the structure of networks to predict the unknown links. The methods of link prediction which based on the topological structure of networks can be classified into two categories: similarity-based algorithms and maximum likelihood estimation based methods. Similarity-based algorithms can achieve high precision in most networks and have a linear time complexity. The main drawbacks are that this kind of algorithms cannot make a fully use of the information of networks and deeper structures of networks cannot be found. The maximum likelihood estimation based methods can discover the hidden structures of networks such as hierarchical structure model and stochastic block model. However, the computational complexity prevents this kind of methods from being widely used.

The similarity-based algorithms include various indices, Local similarity indices, such as CN, Salton Index, Jaccard Index, Sørensen Index, RA [3- 7], Global similarity indices, such as Katz, SimRank [8, 9], and Quasi-local indices, such as Local Path and Local Random Walk [10, 11]. Local similarity indices only consider direct neighbor information. These

978-1-5090-5507-4/17/$31.00 ©2017 IEEE
ICIS 2017, May 24-26, 2017, Wuhan, China

indices can achieve high precision in most networks and have a linear time complexity when calculating the similarity score for every possible link. So, local methods are still good candidates for solving link prediction problem in large networks. However, in this kind of algorithms, the original method only consider the quantitative relationship among a pair of nodes and their common neighbors but not the structural relationship among them.

In this paper, considering the deficiency of common neighbor algorithm, we proposed an approach that combines the similarity of two nodes and local information. We used the neighborhood information of a pair of nodes to predict the potential link between them and then combined this link prediction approach with local information of the two given nodes. We tested our method with real-world data sets. The results showed that this approach could improve the accuracy of predicting unknown links.

II. LINK PREDICTION METHODS

A. *Link prediction*

In the real world, we would abstract some entities and some sorts of phenomenon into objects (nodes).Their relationships can be described as connected links. Therefore, a complex network is one of complex systems which contain many objects and relationships between them. For any real-world network, we can get observed network. However, most of the complex networks are sparse, which means there are many missing links in the observed networks. These missing links can be divided into two categories: 1) links exist but not observed, 2) links should exist or will occur in the future. So, the goal of predicting missing links is to calculate the probability of potential links between two unlinked nodes in the future by analyzing the information of nodes and links in the observed networks [12, 13].

Here we provide some description of networks. Any given networks can be denoted by G = (V, E), V and E represents the set of nodes and links in the networks, respectively. For any nodes *x* and *y* belonging to V, there are

$\Gamma(x)$: The set of neighbor of node x

$\Gamma(x) \cap \Gamma(y)$: Common neighbors of nodes x and y.

k_x: The degree of node x.

B. *Similarity index considering the degrees of nodes*

Salton index [4] - Salton index is defined as:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x k_y}} \tag{1}$$

Jaccard index [5] - This index was proposed by Jaccard over a hundred years ago, which is defined as:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2}$$

Sørensen index [6] - This index is mainly used for ecological community data, which is defined as:

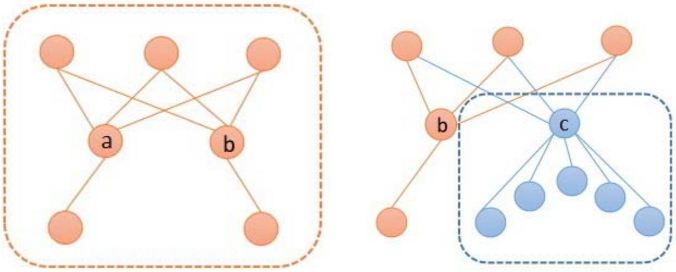
$$S_{xy} = \frac{2 \times |\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \tag{3}$$

III. LOCAL AFFINITY STRUCTURE INDEX

Our method motivated by the algorithms illustrated above only consider the quantitative relationship among a pair of nodes and their common neighbors but not the structural relationship among them. Community structure is one of the topological characteristics of complex networks. The characteristics of community structure represent that there are more links within one community than the amount of links among different communities. A given network can be made up of some different communities. By analyzing the structure of the community, it can reveal the structure of function modules of the networks and characterize some important properties of the networks. Accordingly, we introduce the idea of community structure into link prediction and improve the similarity index between two nodes. Based on the idea of community structure, the similarity of any pair of nodes in the networks is related to the density of the local structure. The more affinity of two nodes and their common neighbor, the higher probability of potential link between the two nodes. By analyzing the affinity relationship of local structure which includes any two nodes and their common neighbors, we define a new link prediction method.

Here, we provide an example to introduce the idea of our algorithm. Suppose we are analyzing social networks and predicting potential friendship between two individuals. Assume that individuals *a*, *b* and *c* are newcomer of a football club. *a* and *b* has three common neighbors in this club, as well as *b* and *c*. *a* is reserved, so he have limited friends. On the contrary, *c* loves sports, and not only joins the football club but also joins the basketball club and swimming club. So we can say that *a* and *b* has a more affinity relationship than *b* and *c*, no matter that the number of common neighbors of *a* and *b* and *b* and *c* is same, or not. The possibility of potential friendship between *a* and *b* is greater than *b* and *c*.

Fig.1. A simple network to illustrate LAS index



As shown in Fig.1, the degree of nodes *a* and *b* is 4, and the degree of node *c* is 8. Also, the number of the common neighbors of nodes *a* and *b* and nodes *b* and *c* are 3. We need to determine which link, between nodes *a* and *b* or between nodes *b* and *c*, is more likely to appear in the future. Hereby, we use the concept of degree of nodes and the idea of community structure. The ratio of the number of common neighbors and degree of nodes *a* is 3/4, correspondingly, this ratio is 3/8 for node *c*. The ratio for node *a* is greater than the ratio for node *c*. So, the relationships between nodes *a* and *b* and their common neighbors are closer; they have a more affinity local structure, furthermore, nodes *a* and *b* may locate in the same community structure. The neighbor nodes of node *c* except the common neighbors for nodes *b* and *c* may have a greater attractiveness for node *c*, the relationship for them may be closer. Based on the analysis above, it is more likely that the node *a* and node *b* will be connected with each other in the future.

Based on the idea above, any two nodes *x*, *y* and their common neighbors of one network, we define the following local affinity structure (LAS) index as:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x} + \frac{|\Gamma(x) \cap \Gamma(y)|}{k_y} \qquad (4)$$

Obviously, LAS index depicts the affinity relationship of a pair of nodes and their common neighbors.

IV. EXPERIMENT

A. Evaluation method

To verify the algorithm, we should divide the set E into two parts: training set E^T and test set E^P. It is clearly that, $E = E^T \cup E^P$ and $E^P \cap E^T = \emptyset$. In this paper, AUC (area under the receiver operating characteristic curve) is adopted to measure the accuracy of link prediction algorithm [14]. When verifying a given prediction algorithm, the unknown links of the network were divided into two parts: non-existed links and the links used for test. The algorithm calculates a score for each link. Therefore, AUC can be interpreted as the probability that the score of a link in the test set is higher than a non-existed link. These two links are randomly selected. That is, each randomly selected link in the test set is compared to a randomly selected non-existed link. The AUC is defined as:

$$\text{AUC} = \frac{n' + 0.5n''}{n} \qquad (5)$$

Where n' denotes the times that a link in the test set has a higher score than a non-existing link and n'' denotes the times that both have the same score. For the definition of AUC, the value of AUC ranges from 0.5 to 1. Higher of the value of the AUC, better the method performed.

B. Data sets of real-world networks

a) *Celegan* [15]: the neural networks of C. elegans.

b) *Karate* [16]: the social relationship networks of 34 members of a karate club.

c) *Lesmis* [17]: the relationship networks of characters in Les Miserable

d) *Netscience* [18]: the networks of co-authorship of the study of complex networks.

e) *Dolphins*[19]:a social network of frequent associations between 62 dolphins.

f) *Polbooks* [20]: a network of books about US politics published in 2004.

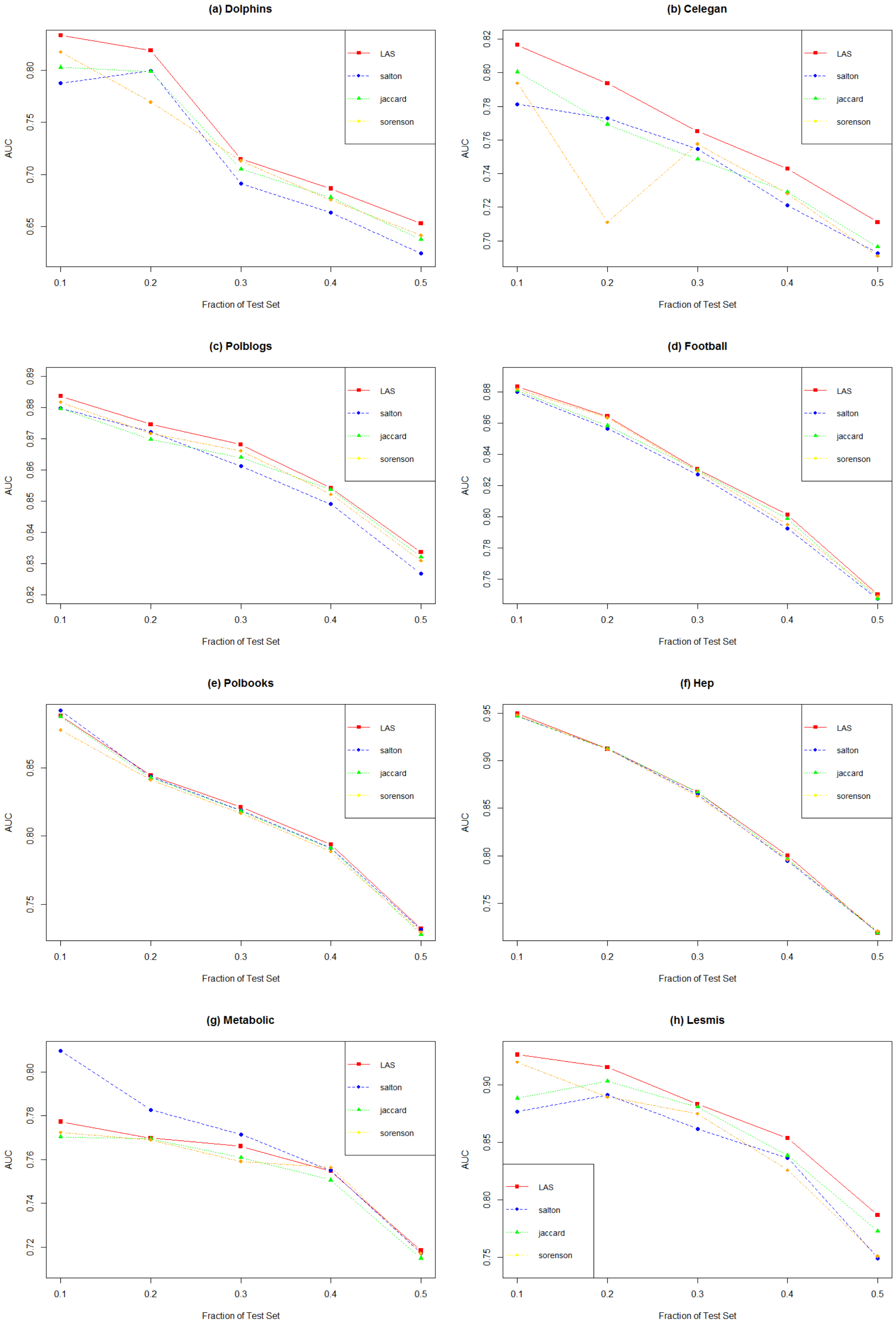
g) *Metabolic* [21]: metabolic network of C.elegans.

h) *Polblogs* [22]: a network of hyperlinks between weblogs on US politics, recorded in 2005.

i) *Football* [23]: a network of American football games.

j) *Hep* [24]: network of coauthor ships between scientists posting preprints on the High-Energy Theory E-Print.

Fig.2. Comparison of AUC for different methods with each networks



(a) Dolphins

(b) Celegan

(c) Polblogs

(d) Football

(e) Polbooks

(f) Hep

(g) Metabolic

(h) Lesmis

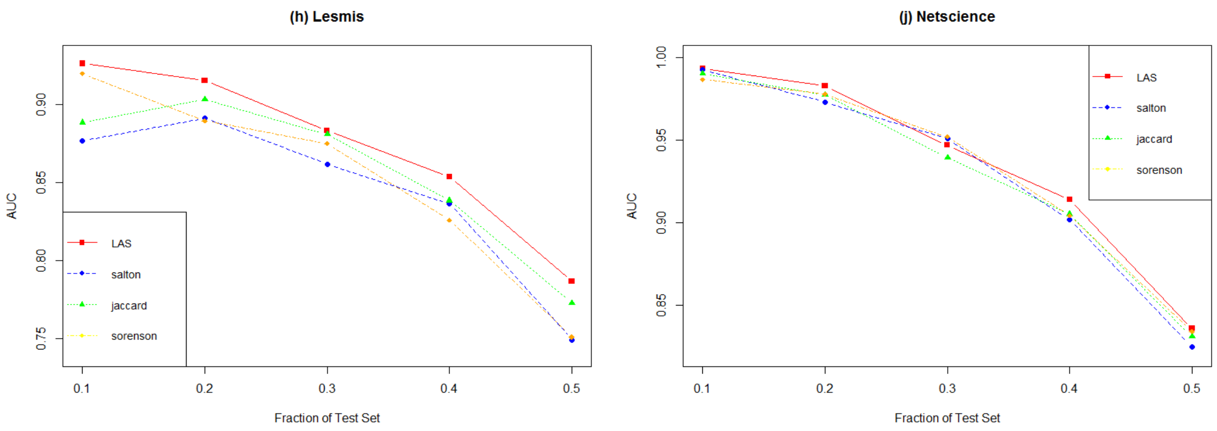


TABLE I. THE TOPOLOGICAL FEATURES OF THE TEN NETWORKS

Network	V	E	Average Degree	Average Clustering coefficient
Adjnoun	112	425	7.589	0.190
Celegan	297	2345	14.465	0.308
Polblogs	1490	19090	25.642	0.360
Football	115	613	10.661	0.403
Polbooks	105	441	8.4	0.488
Karate	34	78	4.588	0.588
Hep	8361	15751	3.768	0.636
Metabolic	453	4596	20.291	0.655
Usair	332	2126	12.807	0.749
Netscience	1589	2742	3.451	0.878

C. Experiments on real-world networks

In this part, twelve real-world networks were used to evaluate our method. In the experiments the test set fraction f ranges from 0.5 to 0.9, and the interval was set to be 0.1. The results are shown in Fig. 2 by AUC values with each data set. In each graph in Fig. 2, we compared the AUC of our method with other three similarity indexes which are considered the degrees of nodes in the networks. It is obvious that the accuracy of our method is higher in all these networks when applied to predicting missing links. The reason is that it is more likely to have a link between two nodes in a more compact local structure. Consequently, the accuracy of link prediction has improved to some extent by our method.

Compared the results of experiments on twelve networks, the networks which has a higher average clustering coefficient, the performance of AUC on it is better. Clustering coefficient can reflect the density of triangles within a local network environment. The basic idea of our method is that more affinity of two nodes, the more likely there has a link between them. The higher the clustering coefficient of the networks, the higher the accuracy of link prediction.

Also, we find that for some networks that the clustering coefficient are not very large, our method still has a good performance. For instance, the clustering coefficient of Hep is only 0.636, however, during the experiment, we found that the modularity of Hep is high which means the Hep has clear community structure. As a result, there had a good accuracy of prediction on Hep. The reason for this phenomenon is that a more compact local structure compensates for the loosening of the global structure in some extent. Also proves the validity of the idea of community structure that we introduced to our method.

Moreover, for a large network with more nodes and links, even if the clustering coefficient is not large and it does not have clear community structure, it has a good prediction accuracy. That is because more nodes and links can provide us with more information, in contrast to the sparse network, and a better performance.

Finally, from Fig.2, the value of AUC decreases fast with the increase of the fraction of missing links, which means that the ratio of training set decreased. With the increase of the fraction of missing links, the part of the networks which used to predict shows limited properties of the networks, such as common neighbors of a pair of nodes and the existed relationship between nodes. So the less the data of training set, the lower the accuracy of link prediction.

V. CONCLUSION

Link prediction is an important research area in the complex networks; it has a strong application foreground. The community structure of networks is which internal structure is high compacted and external connection is relatively sparse. The higher clustering coefficient of networks, the higher the clustering degree of the networks. LAS index depicts the affinity of nodes *x* and *y* and their common neighbors, so LAS index has a better performance in the networks which has a higher clustering coefficient. Simulation experiments were carried on twelve real-world networks with different clustering coefficient showed that our analysis was reasonable.

ACKNOWLEDGE

Rongjing Hu is the corresponding author. This work was partially supported by National Natural Science Foundation of China under grants No. 21503101 and National Science Foundation of Gansu Province of China No. 1506RJZA223.

REFERENCE

[1] Saruukkai B R, “Link prediction and path analysis using markov chains,” Computer Networks, 2010, pp. 377-386.

[2] Zhu J, “Using Markov Chains for Structural Link Prediction in Adaptive Web Sites,” Lecture Notes in Computer Science, 2004, pp. 60-73.

[3] Reichardt J, White D R, “Role models for complex networks,” The European Physical Journal B, 2007, pp. 217-224.

[4] Salton G, Mcgill M J, “Introduction to modern information retrieval,” McGraw-Hill, 1983.

[5] Jaccard P, “Etude de la distribution florale dans une portion des Alpes et du Jura,” Bulletin De La Societe Vaudoise Des Sciences Naturelles, 1901, pp. 547-579.

[6] Sørensen, T, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons,” Biol Skr , 1948, pp. 1-34.

[7] T Zhou, L Lü, Y C Zhang, “Predicting missing links via local information,” The European Physical Journal B, 2009, pp. 623-630.

[8] Katz L, “A new status index derived from sociometric analysis,” Psychometrika, 1953, pp. 39-43.

[9] Jeh G, Widom J, “SimRank: a measure of structural-context similarity,” Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 538-543.

[10] L Lü, C H Jin, T Zhou, “Similarity index based on local paths for link prediction of complex networks,” Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, pp. 593-598.

[11] W Liu, L Lu, “Link Prediction Based on Local Random Walk,” Epl, 2010, pp. 58007-58012.

[12] L Lü, and T Zhou, “Link prediction in complex networks: A survey,” Physica A Statistical Mechanics & Its Applications, pp. 1150-1170.

[13] Getoor, Lise, and C. P. Diehl, “Link mining: a survey,” Acm Sigkdd Explorations Newsletter, pp.3-12.

[14] N Wu, J Li, E Dong, “A prediction method enhanced by the degree of nodes,” International Conference on Natural Computation, pp. 1515-1519.

[15] Watts D J, Strogatz S H, “Collective dynamics of 'small-world' networks,” Nature, 1998, pp. 440-2.

[16] Zachary W W, “An Information Flow Model for Conflict and Fission in Small Groups1,” Journal of Anthropological Research, 1977, p. 473.

[17] Knuth D E, “The Stanford GraphBase: a platform for combinatorial algorithms,” DBLP, 1993.

[18] Newman M E, “Finding community structure in networks using the eigenvectors of matrices,” Physical Review E Statistical Nonlinear & Soft Matter Physics, 2006, pp. 92-100.

[19] Lusseau D, Schneider K, Boisseau O J, et al, “The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations,” Behavioral Ecology & Sociobiology, 2003, pp. 396-405..

[20] V. Krebs, unpublished, http://www.orgnet.com/.

[21] Duch J, Arenas A, “Community detection in complex networks using extremal optimization,” Physical Review E Statistical Nonlinear & Soft Matter Physics, 2005, pp. 986-1023.

[22] Adamic L A, Glance N, “The political blogosphere and the 2004 U.S. election: divided they blog,” International Workshop on Link Discovery. ACM, pp. 36--43.

[23] Girvan M, Newman M E J, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences of the United States of America, 2002, pp. 7821-6.

[24] Newman M E J, “The structure of scientific collaboration networks,” Proceedings of the National Academy of Sciences of the United States of America, 2001, pp. 404-9.