Multi-scale View Reveals Easily Detectable Community in Complex Networks

Qingju Jiao*, Yuanyuan Jin

School of Computer & Information Engineering, Anyang Normal University, Anyang 455000, China

Corresponding Author Email: qjjiao@aynu.edu.cn

ABSTRACT

Community is a prominent feature of complex networks, and many methods have emerged for community detection. However, the existing methods have many intrinsic drawbacks and cannot work effectively. To solve the problems, this paper probes deep into the effects of multi- and single-scale community detection methods. Firstly, a typical multi-scale method was adopted to detect the communities in synthetic and real-world networks. Then, five single-scale methods were employed for community detection in the same networks. The detection results of both types of methods were analyzed in details. The results show that, for a given network, the communities must be prominent and easily detectable by single-scale methods, if these communities are both strongly synchronous and if the synchronous partition falls within the stable partitions detected by multi-scale methods. If these communities only satisfy the first condition, then they can be detected by some special single-scale methods. If these communities are asynchronous, then they cannot be effectively detected by single-scale methods, and should be treated with multi-scale approaches. The research findings shed new light on the design of community detection methods.

1. INTRODUCTION

Network is an effective tool to describe and analyze complex systems. Being a vital feature of network, community structure has been successfully applied to various fields. Many community detection methods have been developed for different types of networks. These approaches are based on either global or local topologies [1].

The modularity-based method is a global community detection approach. The modularity [2, 3] can detect communities, and reflect the quality of the method (the higher the modularity, the better the method). Many techniques, ranging from simulated annealing (SA) [4], the greedy algorithm [3] to spectral optimization [5, 6], have been adopted to maximize the modularity. However, the global topological structure-based methods may suffer under partition. The local methods detect community based on the neighborhood information of a node, clustering the nodes in the same neighborhood into a community [7]. The drawback of the local topological structure-based methods is that the divisions they generate may consist of smaller (which may even include one or several) nodes and dense cores [8]. In general, the existing community detection methods mainly focus on the matching between detected and actual communities. Nonetheless, there is no report that identifies the type of community detected by single-scale methods.

This paper probes deep into the effects of multi- and single-scale community detection methods. Experiments on synthetic and real-world networks demonstrate that single-scale methods can easily detect highly synchronous communities in stable partitions, but not asynchronous communities. The research results provide reference for the design of community detection strategies.

2. MULTI-SCALE COMMUNITY DETECTION

In this section, a multi-scale method (ISIMB) [9] was adopted to detect the communities in synthetic and real-world networks. In the ISIMB, the first step is to define the transition probability from node i to node j as follows, based on an improved random walk on network:

$$P_{ij}(t+1) = (1-\alpha)\frac{1}{SP(i,j)} + \alpha\sum_{k=1}^{|U_i|}\frac{1}{d_i}P_{kj}(t) \tag{1}$$

where, $P_{ij}(t+1)$ is the transition probability from node i to node j at the $(t+1)th$ time, α is a variable regulating the weight between global and local topologies. $SP(i,j)$ is the shortest path between node i and node j, d_i is the degree of node i, U_i denotes the set of neighbours of node i, and k is a member of U_i. Taking the average of P_{ij} and P_{ji} computed by formula (1), the similarity between node i and node j can be obtained as:

$$S_{ij} = \frac{P_{ij} + P_{ji}}{2} \tag{2}$$

Next, the value of α was adjusted from 0 to 1, and the corresponding similarities between the two nodes were computed by formula (2), forming a series of similarity matrices. Based on the similarity matrices, the linkage hierarchical clustering was performed to create a community tree, in which each leaf is a node from the original network and each branch is a module. The number of modules was set to a proper level to cut the tree. The number of communities was configured based on the number of eigenvalues of the similarity matrix significantly greater than an empirical

threshold (0.2). Since there are 60 regulators, a total of 60 community trees were set up, and the network was divided into 60 partitions on different scales.

Afterwards, a stable partition was obtained from the results of the multi-scale method. The stable partition was defined as $\eta=\xi/\omega$, where ξ is the number of regulators which the network partition spans, and ω is the total number of regulators. The network partition is stable if η is greater than one-tenth.

The SIMIB method was adopted to detect the multi-scale communities in two synthetic networks, namely, the H15-2 network [10, 11] and the LFR network [12]. The H15-2 is a homogenous in-degree network with two preset hierarchical thresholds. This 256-node network has 15 edges linking up 16 nodes in the most internal community. In addition, there are two edges between the most internal community and the most external community (64 nodes), and one edge between the former with any other random node in the network. The H 15-2 network has two predefined hierarchical levels including 16 communities and 4 communities, which means that the H 15-2 network has two real community structures. The 16-way real community structure is more stable than the 4-way real community structure because it captures stronger local network topology.

The LFR network has many user-defined parameters: the number of nodes N, the mean degree k, the maximum degree $maxk$, the minimum micro-community size $minc$, the maximum micro-community size $maxc$, the number of overlapping nodes in micro-community on, the membership of overlapping nodes in micro-community om, the minimum macro-community size $minC$, the maximum macro-community size $maxC$, the mixing parameter for macro-communities $m\mu1$, and the mixing parameter for micro-communities $m\mu2$. To set up the synthetic network, these parameters were configured as $N=1,000$, $k=maxk=16$, $minc=maxc=10$, $on=om=0$, $minC=maxC=50$, $m\mu1=0.03$ and $m\mu2=0.08$. According to the LFR, the generated synthetic network will have two hierarchical community structures with 40 communities and 20 communities each. The first level will have stronger community structures than the second level because its communities have stronger local features set by the parameters.

Figure 1 shows the multi-scale communities in the H15-2 network and the LFR network obtained by the ISIMB method. It can be seen that the H15-2 network has two stable partitions H16 ($\eta=29/30$) and H4 (6/60). The η values indicate that H16 is the more stable partition. In the LFR network, the predefined hierarchical community structures (named L40 and L20) are mined by the ISIMB method, and the values η of the L40 and L20 partitions are 14/60 and 7/60.

Next, the ISIMB method was applied to detect the multi-scale communities in four real-world networks, namely, Zachary's karate club [13] (karate network), American College Football Network [14] (football network), Krebs' books network on American politics (http://www.orgnet.com/) (polbook network), and Bottlenose dolphin network [15] (dolphins network).
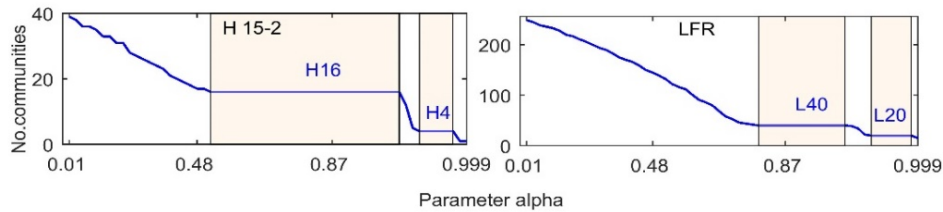

Figure 1. Multi-scale communities of H15-2 and LFR networks, the "Parameter alpha" represents the parameter α in the ISIMB method
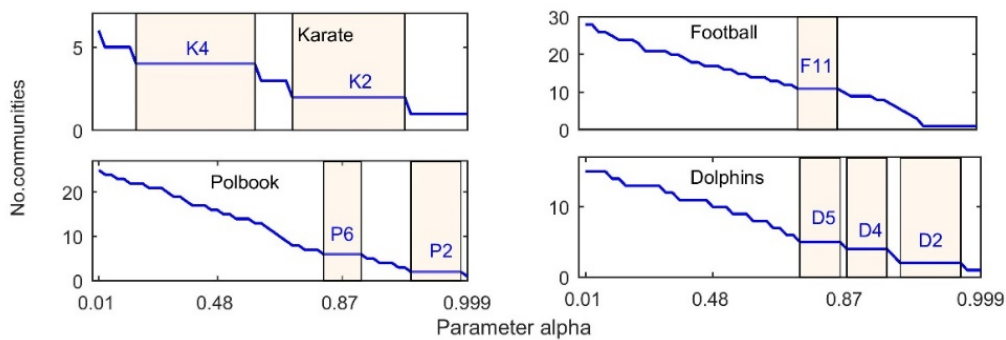

Figure 2. Multi-scale communities in four real-world networks, the "Parameter alpha" represents the parameter α in the ISIMB method

Table 1. The stable partitions and η values in four real-world networks

Network	Karate		Football	Polbook		Dolphins		
Stable partition	K4	K2	F11	P6	P2	D5	D4	D2
η	20/60	19/60	7/60	7/60	9/60	7/60	7/60	10/60

The karate network is a small network that has 34 nodes and 78 edges. The network is divided into two clusters, and is widely used to test various community detection methods. The football network is a network of American football games between Division I-A colleges during regular season fall 2000, and it has 115 nodes and 615 edges. The network has 12 real communities. The third real network is the polbook network, whose nodes represent books about U.S. politics published around the 2004 presidential election and sold by the online bookseller Amazon.com, and edges between books represent the frequent co-purchasing of books by the same buyers. The polbook network has 105 nodes and 441 edges. The last network is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, it has 159 edges. The polbook and dolphins networks have 3 and 2 real communities, respectively.

Figure 2 presents the multi-scale communities detected by the ISIMB from the four networks. The stable partitions and η values are listed in Table 1. In the Karate network, the partition with 4 communities (K4) and that with 2 communities (K2) covered more regulators than the other partitions, and appeared to be more stable. Hence, the Karate network has two stable partitions K4 and K2. In the football network, there is only one stable partition with 11 communities (F11), which means that the real community structure may not be a stable partition. In the polbook network, the stable partitions include one with 6 communities (P6) and another with 2 communities (P2). In the dolphins network, the ISIMB discovered three

stable partitions (D5, D4 and D2), with D2 being the most stable one.

3. SINGLE-SCALE COMMUNITY DETECTION

At the same time, five single-scale community detection methods are introduced to detect the communities in the above two synthetic and four real-world networks, including the improved subspace iteration method (ISIM) [1], the Infomap [16], greedy modularity optimization method [3], Louvain [17] and OSLOM [18]. The detection results were compared with the metadata by normalized mutual information [19] (NMI):

$$NMI(\chi,\gamma) = \frac{-2\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} n_{ij}^{XY} \log(\frac{n_{ij}^{XY} \cdot N}{n_i^X \cdot n_j^Y})}{\sum_{i=1}^{n_X} n_i^X \log(\frac{n_i^X}{N}) + \sum_{j=1}^{n_Y} n_j^Y \log(\frac{n_j^Y}{N})}$$

(3)

where, $\chi=(X_1,X_2,\ldots,X_{nX})$ and $\gamma=(Y_1,Y_2,\ldots,Y_{nY})$ are two partitions of a network; n_X and n_Y are the number of communities in the two partitions, respectively; N is the number of nodes in a network; n_i^X and n_j^Y are the number of nodes in communities X_i and Y_j, respectively; n_{ij}^{XY} is the number of nodes shared by communities X_i and Y_j: $n_{ij}^{XY} =| X_i \cap Y_j|$. The greater the NMI, the better the partition results.
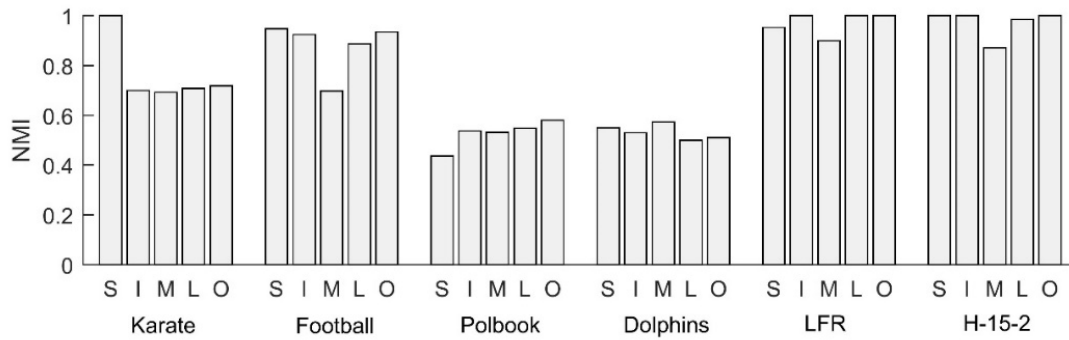

Figure 3. The detection results on the six networks, S, I, M, L and O refers to the ISIM, the Infomap, the greedy modularity optimization, the Louvain modularity and the OSLOM, respectively
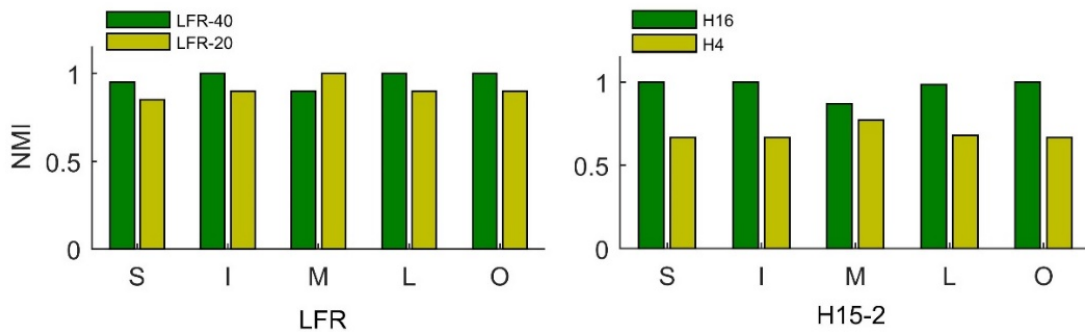

Figure 4. The detection results on the two synthetized networks

The detection results on the six networks are displayed in Figure 3. The real communities in the karate and football networks were easily detected by the five single-scale methods. In particular, all real communities in the karate network were

discovered by the ISIM method. By contrast, the test methods had some difficulty in mining the real communities of the polbook and dolphins networks, with the NMIs fluctuating about 0.5. For the two synthetic networks, the five single-scale

methods show good performance, and the values of NMI of Infomap, Louvain and OSLOM are equal to one (see Figure 3 or green pillars in Figure 4) when we regard 40 communities and 16 communities as standard communities for LFR and H 15-2 respectively. Using 20 communities and 4 communities as standard communities for LFR and H 15-2 respectively, the real communities in the LFR network are basically discovered by five single-scale methods (see pale yellow pillars in Figure 4), but the case is worse in the H 15-2 network (see pale yellow pillars in Figure 4).

4. STABLE PARTITION REVEALS EASILY DETECTABLE COMMUNITY

To disclose the relationship between multi- and single-scale community detection methods, this section analyzes the α value(s) or partition(s) that ensures the matching between detected and actual communities. The matching degree is measured by the Jaccard similarity coefficient (see formula 4).

$$ J_{XY} = \frac{|X \cap Y|}{|X \cup Y|} \tag{4} $$

where, X and Y are two communities, $|X \cap Y|$ is the number of nodes shared by communities X and Y, and $|X \cap Y|$ is the number of nodes merged by communities X and Y.

Taking the karate network for example, the two real communities were detected simultaneously with a high Jaccard similarity, when the α value fell within [0.78, 0.96]. The two communities are highly synchronous if the Jaccard similarity is 1 (i.e. the real communities are completely identified), weakly synchronous if the Jaccard similarity is greater than 0.7. On the contrary, if the real communities are not discovered simultaneously with a certain Jaccard similarity when the α is equal to one (or some) values, the real communities are asynchronous. For the two synthetic networks, both 40 and 20 real communities in the LFR network are strong community synchronism, both 16 and 4 real communities are also strong community synchronism. Similarly, the real communities in the dolphins network also have the attribute of weak community synchronism (the α are 0.93 and 0.94).

Figure 5 describes the synchronous/asynchronous relations between communities in the polbook and football networks.

As shown in Figure 5(a), the two real communities (1 and 3) exhibited the highest Jaccard similarity when the α value was between 0.86 and 0.92 (blue lines), while the α value corresponding to the peak Jaccard similarity of the other community 2 fell between 0.93 and 0.997. Hence, the real communities in the polbook network are asynchronous. As shown in Figure 5(b), the real communities in the football network were weakly synchronous, for all of them were identified clearly when the α value changed from 0.58 to 0.66 (blue lines).

Furthermore, we examined whether the synchronous partition is in the stable partitions. For instance, the synchronous partition in the karate network was obtained when the α value was 0.85. In the two synthetic networks, all the synchronous partitions were found within the stable partitions, indicating that the real communities are both synchronous and stable. In other words, the real communities can be completely identified by the single-scale methods.

For the four real-world networks, the communities in the karate network were easily detected by all five single-scale methods, for the strongly synchronous partitions correspond to the stable partition K2 in Figure 2. Of course, the greedy modularity optimization did not completely disclose the real communities. But its results were very close to the stable partition K4. In the football network, the weakly synchronous partition (α: 0.58~0.66) corresponds to the stable partition F11, revealing that the real communities are both synchronous and stable. Therefore, the communities in the football network can be well detected by the five single-scale community detection methods. The synchronous partition in the dolphins network was not located in any stable partitions (D5, D4, and D2). This means the five single-scale methods have difficulty in detecting the real communities in this network. For the polbook network, the five single-scale methods performed poorly because the real communities are asynchronous.

To sum up, for a given network, the communities must be prominent and easily detectable by single-scale methods, if these communities are both strongly synchronous and if the synchronous partition falls within the stable partitions detected by multi-scale methods. If these communities only satisfy the first condition, then they can be detected by some special single-scale methods. If these communities are asynchronous, then they cannot be effectively detected by single-scale methods, and should be treated with multi-scale approaches.
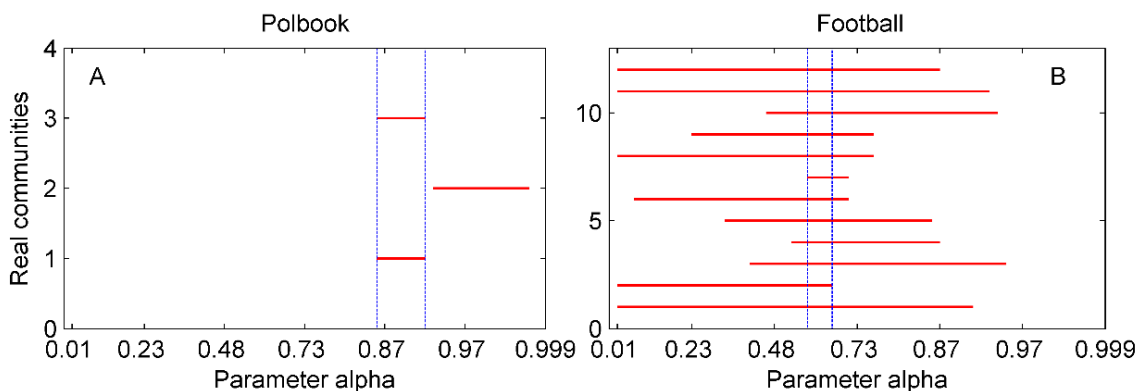

Figure 5. The synchronous/asynchronous relations between communities in the polbook and football networks, the "Parameter alpha" represents the parameter α in the ISIMB method

5. CONCLUSIONS

Community structure is a significant feature in complex networks. The characteristic of community in which nodes have the same attribute is widely used in various fields, and abundant results have been obtained. Therefore, many community detection methods are proposed. But these methods have intrinsic drawbacks, and it is hard to design effective methods. In order to solve this problem, we employ multi-scale view to reveal easily detectable community structure in complex networks. Experiments on both synthetic and real networks show that, for a given network, the communities must be prominent and easily detectable by single-scale methods, if these communities are strongly synchronous and if the synchronous partition falls within the stable partitions detected by multi-scale methods. The results can provide one with a detailed guide for designing community structure mining methods.

ACKNOWLEDGMENT

REFERENCES

[1] Jiao, Q.J., Huang, Y., Shen, H.B. (2015). Community mining with new node similarity by incorporating both global and local topological knowledge in a constrained random walk. Phys. A: Statistical Mechanics and its Applications, 424: 363-371. https://doi.org/10.1016/j.physa.2015.01.022

[2] Newman, M.E.J., Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 69(2): 026113. https://doi.org/10.1103/PhysRevE.69.026113

[3] Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review. E, 69: 066133. https://doi.org/10.1103/PhysRevE.69.066133

[4] Guimera, R., Sales-Pardo, M., Amaral, L.A.N. (2004). Modularity from fluctuations in random graphs and complex networks. Physical Review. E, 70: 025101. https://doi.org/10.1103/PhysRevE.70.025101

[5] Richardson, T., Mucha, P.J., Porter, M.A. (2009). Spectral tripartitioning of networks. Physical Review. E, 80: 036111. https://doi.org/10.1103/PhysRevE.80.036111

[6] Newman, M.E.J. (2006). Modularity and community structure in networks. Proc. Proceedings of the National Academy of Sciences of the United States of America, 103(23): 8577-8582. https://doi.org/10.1073/pnas.0601602103

[7] Ahn, Y.Y., Bagrow, J.P., Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. Nature, 466: 761-764. https://doi.org/10.1038/nature09182

[8] Newman, M.E.J. (2012). Communities, modules and large-scale structure in networks. Nature Physics, 8: 25-31. https://doi.org/10.1038/nphys2162

[9] Jiao, Q.J., Huang, Y., Shen, H.B. (2016). A new multi-scale method to reveal hierarchical modular structures in biological networks. Molecular BioSystems, 12: 3724-3733. https://doi.org/10.1039/C6MB00617E

[10] Arenas, A., Fernandez, A., Gomez, S. (2008). Analysis of the structure of complex networks at different resolution levels. New Journal of Physics, 10(5): 053039. https://doi.org/10.1088/1367-2630/10/5/053039

[11] Arenas, A., Díaz-Guilera, A., Pérez-Vicente, C.J. (2006). Synchronization reveals topological scales in complex networks. Physical Review Letters, 96(11): 114102. https://doi.org/10.1103/PhysRevLett.96.114102

[12] Lancichinetti, A., Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 80: 016118. https://doi.org/10.1103/PhysRevE.80.016118

[13] Zachary, W.W. (1977). An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33(4): 452-473. https://doi.org/10.1086/jar.33.4.3629752

[14] Girvan, M., Newman, M.E.J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12): 7821-7826. https://doi.org/10.1073/pnas.122653799

[15] Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M. (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 54: 396-405. https://doi.org/10.1007/s00265-003-0651-y

[16] Rosvall, M., Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America, 105(4): 1118-1123. https://doi.org/10.1073/pnas.0706851105

[17] Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008: P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

[18] Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S. (2011). Finding statistically significant communities in networks. PLoS One, 6: e18961. https://doi.org/10.1371/journal.pone.0018961

[19] Strehl, A., Ghosh, J. (2003). Cluster ensembles-a knowledge reuse framework for combining multiple partition. The Journal of Machine Learning Research, 3: 583-617. https://doi.org/10.1162/153244303321897735