A new method for detecting communities in network based on the affinity between nodes

Jin Zeng, Chenxi Shao*, Xingfu Wang and Fuyou Miao
University of Science and Technology of China, China

*E-mail: cxshao@ustc.edu.cn

Abstract. It has been found that many networks display community structure, analysis of the community structure in the network is meaningful for understanding the real-world network. Recently, many community discovery methods have been proposed, most of which are based on the assumption that nodes in the network can only belong to a community at most. The proposed method in the paper is based on the affinity between nodes and it does not agree with the assumption, i.e., the method is able to discover overlapping community structures in networks. We apply the method to various real-world networks and artificial networks, the experimental results confirm that it has better performance on networks with overlapping community structures and non-overlapping community structures.

1. Introduction
Complex networks are abstracted from complex systems, the vertex in the network represents the individual in the system and the edge represents the relationship between individuals. Analysis of network community structure is helpful to understand the system. For example, communities may be related to functional modules in biochemical networks or individuals with common interests in social networks [1]. Previous researches demonstrate that there are community structures in most complex networks, that are group nodes densely connected within and loosely connected with the rest of the network [2,3]. The existed community discovery methods are roughly divided into the following categories: Modularity optimization method [4,5,6], because of the flaw of the definition of Modularity, such methods are exposed to resolution limits [7]; Methods based on propagation dynamics [8,9], this type of method holds the idea that the information remains in the same community with high probability after several times dissemination, they are fast but poor robustness; Methods based on graph bisection which requires prior-information. The methods mentioned above are effective for detecting non-overlapping community but rarely able to detect network overlapping community which is ubiquitous in real-world networks. For example, individuals could belong to different social circles in social networks. Another, a large fraction of proteins belong to several protein complexes simultaneously in protein networks [10].

In this paper, we propose a localized community detection method based on the affinity between nodes. In order to measure the quality of community structures, the stability of community structure has been defined. When node joined or was remove from a community, the stability of the community would change. Group of nodes with the largest local stability is considered as a suggested community.
The proposed method does not limit the number of communities to which the node belongs, i.e., it is effective for detecting overlapping community. The method has been applied to various real-world networks and artificial networks. The community structures identified by the method are in agreement with the actual communities present in real-world networks. Moreover, it achieves greater score in artificial networks and is effective for networks with overlapping and non-overlapping community structures.

2. Definition

The relationship between nodes is close or distant in complex networks, i.e., it is easy for information propagates from the node to some nodes but difficult to the rest. For example, it is easy to exchange information with friends but difficult with strangers in social networks. We proposed a hypothesis that there is affinity between individuals in systems. We proposed a hypothesis that there is affinity between individuals in systems. Under the affinity, the relationship between individuals becomes close or alienated that explains the existence of the community structures in the network. Intuitively, the greater the affinity, the closer the relationship between the nodes. For two nodes in the network, the greater the ratio of the number of paths shorter than k to the product of their degrees, the closer they are. For simplicity, we set k=2. Therefore, the affinity factor f(i,j) between two nodes is defined as

\[
f(i,j) = \frac{\sum_{k=1}^{n} A_{ik}A_{jk}}{(\sum_{k=1}^{n} A_{ik}^{2})(\sum_{k=1}^{n} A_{jk}^{2})^{\frac{N}{2}}} \tag{1}
\]

where A is adjacency matrix of the network, \(\theta\) controls the strength of influence of nodes degree. The affinity factor \(f(i,j)\) is positively related to the number of common neighbors (paths with length 2) and negatively correlated with the product of degrees of the two nodes. Since most nodes are not sharing any neighbor node, we construct a virtual node connected with each node in the network. Therefore, the Eq. (1) is changed to

\[
f(i,j) = \frac{\sum_{k=1}^{n} A_{ik}A_{jk+1}}{(\sum_{k=1}^{n} A_{ik}^{2}+1)(\sum_{k=1}^{n} A_{jk}^{2}+1))^{\frac{N}{2}}} \tag{2}
\]

The affinity \(A(i,j)\) is positively related to the affinity factor and negatively related to the shortest distance \(d\) between the two nodes. It is defined as

\[
A(i,j) = \frac{f(i,j)}{d^{\ast}} \tag{3}
\]

As mentioned above, the definition of community is that community is a group of nodes which connected densely within and loosely connected with the rest, but it is fuzzy not a quantitative measure. However, a large room has been given to researchers. Based on the definition, the stability of community is proposed. The community is considered as stable structure, when the community expands to a certain scale, the stability would be degraded if it continues to expand. In the process of community expansion, the new node that joins the community will be the frontier, which causes a change of the affinity within and outside the community. Therefore, the stability is positively related to the sum of affinity of the inside for the frontiers of the community and is negatively related to the sum of affinity of the outside. For simplicity, we just take account of the affinity between neighbor nodes. Consequently, the affinity of the inside for the frontiers of the community \(g\) Fin(g) is formulated as

\[
\text{Fin}(g) = \Sigma_{i \in f} \Sigma_{j \in \{n \mid g\}} A(i,j) \tag{4}
\]

The affinity of the outside for the frontiers is formulated as

\[
\text{Fout}(g) = \Sigma_{i \in f} \Sigma_{j \in \{n \mid \sim g\}} A(i,j) \tag{5}
\]

where \(f\) is the set of frontiers of the community, \(n\) is the set of neighbor nodes of node \(i\). Consequently, the stability \(S(g)\) of community \(g\) is formulated as
3. Method
Initially, each node in the network is considered as an initial community. Firstly, select node as seed community according a certain rule. In the process of seed community expansion, select the best node as a candidate of the member of the community at each step. In order to determine whether the candidate should join the community, the fitness of node i is been defined as
\[
\text{fitness}(g, i) = s(g \cup i) - s(g)
\] (7)

Join the node that is connected to the community and whose fitness is the largest and greater than zero into the community at each step until the fitness of all nodes connected to the community is less than 0. As a consequence the stability of the community obtained by the expansion of the seed is the local maximum. It is not necessary to pursue global optimal value of the stability, because the whole network as a community, the stability \(S(g)\) will be the largest, but such a solution is obviously meaningless. Then re-select the seed and repeat the process until all nodes in the network have their own community. As for the strategy of selection of the seed, it is a solution to choose randomly a node that have not yet belonged to any community as the seed, but it is low-efficiency. Since there are often leader nodes in communities, the nodes are ranked by the sum of affinity for their neighbor nodes (The greater the affinity for other node, the greater the probability that the node is leader), then select node that is not affiliated with any community as the seed in turn. In this way, it is high-efficiency to select as few nodes as possible to achieve the goal of dividing the whole network into communities. In the process of the expansion of the seed, some nodes may be divided into multiple communities. When the expansion of the seed is over, if there is a community obtained by other seed is similar to the community (we agree that two communities are similar when the jaccard coefficient of them exceeds 0.5), the two communities should be merged. We can describe our proposed method in the following steps.

1. Rank the nodes in descending order according to the sum of affinity for their neighbors.
2. Select the top ranked nodes that does not belong to any community as the seed.
3. Join the neighbor node whose fitness is the greatest in the seed community and mark the node as belonging, then update the seed community. Repeat the process until the fitness of all nodes connected to the seed community less than 0
4. Compare the community with each of existed community, merge them into a single community if the jaccard coefficient of them exceeds 0.5
5. Repeat step (2) to step (4) until all nodes in the network have their own community.

The algorithm is an agglomerative algorithm, i.e., the community is gradually expanded from the seed. In the process of expansion, it costs \(O(\mu s^2 < k >^2)\) time to compute the fitness of a node, the candidate should be selected from \(\mu s\) neighbor nodes. Therefore, each expansion costs \(O(\mu s^2 < k >^2)\), where \(\mu\) is the ratio of the number of frontiers to the size of the community, \(<k>\) is the average degree of the nodes in the network and \(s\) represents the average scale of the communities. Because the seed community need to be updated after joining new node, the time complexity is \(O(\mu s^3 < k >^2)\). Although the time complexity of the algorithm is proportional to the cube the average scale of the communities, compare with the whole network, the average scale of the communities is much smaller.

4. Experiment
We apply our proposed method and other representative methods to the real-world networks (section 4.1) and artificial networks with different scales (section 4.2). The normalized mutual information(NMI) and modularity \(Q\) are used as quantitative measure to evaluate the performances of various methods.

4.1. Real-world network
The first is the Zachary karate network which captures 34 members of a karate club, documenting 78 pairwise links between members who interacted outside the club. Because of the conflict between the president and the coach, the club is divided into two communities. The division is shown in Fig. 1. The performances of the proposed method and other methods are shown in Table 1. As Table 1 shows, none of these methods divides the network into two communities.

**Figure 1.** The natural division of the Zachary karate network that includes two community structures.

It does not indicate that these methods are inaccurate, contrarily, most mainstream methods divide it into four communities. The reason that the communities identified by the methods are different from the natural communities is that community structure is hierarchical. Most methods, in pursuit of accuracy, tend to divide the natural communities further. The community structures identified by our method is shown in figure 2. It is worth mentioning that the node 10 is wavering, i.e., the node 10 belong to two communities. Because the classification of node 10 differs from the standard, the NMI between identified communities and natural communities is less than LPA and Fast-unfolding.

**Table 1.** Some indexes of communities identified by the methods.

| Method       | NMI   | Modularity | The number of communities |
|--------------|-------|------------|---------------------------|
| Our method   | 0.633 | 0.418      | 4                         |
| Fast-unfolding | 0.677 | 0.393      | 4                         |
| LPA          | 0.699 | 0.385      | 3                         |
| CNM          | 0.524 | 0.277      | 5                         |

It is not a mistake that the proposed method put the node 10 into two communities, many methods that can detect overlapping community, also consider the node 10 is a wavering node. As shown in figure 2, the number of connections between node 10 and two communities is equal.

**Figure 2.** The Zachary karate network is divide into four communities by our proposed method.

To further validate the method, it is applied to the Risk network, the Football network and the Dolphins network. The experimental results can been seen in table 2, table 3, table 4.
Table 2. The normalized mutual information (NMI) between natural communities and communities identified by the methods.

| Network  | Our method | LPA  | Fast-Unfolding | CNM  |
|----------|------------|------|----------------|------|
| Risk     | 0.908      | 0.778| 0.869          | 0.839|
| Dolphins | 0.467      | 0.888| 0.474          | 0.403|
| Football | 0.911      | 0.830| 0.892          | 0.861|

Table 3. The modularity Q of communities identified by the methods.

| Network  | Our method | LPA  | Fast-Unfolding | CNM  |
|----------|------------|------|----------------|------|
| Risk     | 0.667      | 0.572| 0.631          | 0.486|
| Dolphins | 0.521      | 0.378| 0.500          | 0.439|
| Football | 0.581      | 0.553| 0.604          | 0.538|

Table 4. The number of communities identified by the methods, the figure in parentheses represents the number of natural communities.

| Network  | Our method | LPA  | Fast-Unfolding | CNM  |
|----------|------------|------|----------------|------|
| Risk(6)  | 7          | 5    | 7              | 6    |
| Dolphins(2) | 13      | 2    | 7              | 7    |
| Football(12)| 13      | 9    | 9              | 9    |

As tables show, In Risk network and Football network, the results using our method are comparable with the standard, but the performance on Dolphins network is a disappointment. It can be found that Fast-unfolding and CNM also perform poorly, while LPA is excellent which is related to the principle of the LPA. The LPA assigns a label to every node in networks, and adopts the label that a maximum number of its neighbors have at every iteration. Therefore, LPA is disadvantageous to the small community in the network, i.e., small community is easy to be merged with large community adjacent to it. The standard division of Dolphins network is two communities with comparable size, although there are many connections between the two communities. From table 2, the modularity of the communities identified by LPA is obviously less than the other three. The other methods divide the network into more fine-grained communities, so the communities identified with lower NMI and greater modularity.

Through the above experiments, we found that what we proposed in this paper is a method of pursuing fine-grained community. In order to achieve the greatest local maximum of the stability of community, as far as possible to divide the whole network into small communities with higher stability. As a consequence, it would reduce the time complexity of the method.

4.2. Artificial network

We analyzed the proposed method on a class synthetic benchmark networks with planted partition [11]. The significance of community structure is controlled by a mixing parameter \( \mu \in [0,1] \), where smaller values give clearer community structure. Compare with GN-Network, the size of community and the degree of the node are no longer fixed, but obey the power law distribution in the benchmark. Using the method proposed in [11], the synthetic networks with different \( \mu \), different scales and overlapping communities are generated. The experimental results are shown in figure 3. Figure 3 shows that the mixing parameter \( \mu = 0.5 \) is a critical point. The community structure is no longer defined in the strong sense when \( \mu \) is greater than 0.5. The communities identified by our method are comparable with the standard division, while the other methods are also 90% similar to the standard partition when \( \mu < 0.5 \). When \( \mu > 0.5 \), the NMI between the standard partition and communities
identified by the various methods drops rapidly, where the LPA is most sensitive, it is already not working when \( \mu = 0.6 \) (dividing the whole network into a community). Compared with the other method, the performance of our method decreases slowly.

![Figure 3](image)

**Figure 3.** The normalized mutual information between planted communities and communities identified by the various methods in the benchmark with different size.

It is still about 75% similar to the standard when \( \mu = 0.7 \), which indicates that the proposed method is more robust. Furthermore, as the scale of network grows, the running time of the other methods significantly increased, while the proposed method does not. It confirms that the time complexity of the method is related to the average scale of communities, not the network's. We test the proposed method on the benchmark networks with overlapping community, and the OSLOM [12], COPRA [13] are used as the contrast. The performance of the three methods is shown in Fig.4. As the figure shows, as the fraction of overlapping nodes increases, there is no significant fluctuation in the performance curve of our method and the OSLOM, while the COPRA suddenly fails when the fraction of overlapping nodes reaches about 0.7. The reason is that The COPRA is the same as the LPA, based on label propagation, so it has lower running time than the proposed method and OSLOM, but is unable to deal with networks with a high proportion of overlapping nodes. The performance of the three methods is comparable when the mixing parameter \( \mu <= 0.5 \), all of which can divide the network currently. When \( \mu > 0.5 \), the COPRA fails suddenly, the performance of the proposed method and the OSLOM also decreased, but the decline trend of the method proposed is slower. Comparing with the OSLOM, our method is slightly backward in accuracy, but has lower running time.

![Figure 4](image)

**Figure 4.** The number of node is 1000. The left is the change curve of performance corresponding to the fraction of the number of overlapping nodes and the right corresponds to the mixing parameter.

5. Conclusion
We have introduced a community detection method, which based on the affinity between nodes and the stability of community. The proposed method is effective for networks with overlapping community structures and non-overlapping community structures, and the time complexity is related to the average scale of community structures. Since the scale of communities and the size of whole network are not in an order of magnitude, the method could be applied to large-scale networks.

Acknowledgement
The work is supported by the NNSFC No.61472381, No.61174144, No.61472382, No.61772490, No.61572454

References
[1] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Nature 435, 814 (2005)
[2] Fortunato, Santo. Community detection in graphs[J]. Physics Reports, 2009, 486(3):75-174.
[3] Newman M E J. The Structure and Function of Complex Networks[J]. Siam Review, 2003, 45(2):167-256.
[4] Newman M E. Fast algorithm for detecting community structure in networks.[J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 69(6 Pt 2):066133.
[5] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics, 2008, 2008(10):155-168.
[6] Clauset A, Newman M E, Moore C. Finding community structure in very large net-works.[J]. Physical Review E, 2004, 70(2):066111.
[7] Fortunato S, Barthlemy M. Resolution limit in community detection.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(1):36-41.
[8] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(3 Pt 2):036106.
[9] Pons P., Latapy M. (2005) Computing Communities in Large Networks Using Random Walks. In: Yolum .,, Gngr T., Grgen F., zturan C. (eds) Computer and Information Sci-ences - ISCIIS 2005. ISCIS 2005. Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg.
[10] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society.[J]. Nature, 2005, 435(7043):814.
[11] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2008, 78(4 Pt 2):046110.
[12] Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks[J]. Plos One, 2011, 6(4):e18961.
[13] Gregory S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2009, 12(10):2011-2024.