A Multi-resolution Community Partition Method based on Node Degree

Quan Xia, Fan Wu and Kebin Chen*
College of Information and Communication, National University of Defense Technology, Wuhan 430019, China
*chenkebin17@nudt.edu.cn

Abstract. Community detection is one of the significant methods to study the properties of complex networks. Community resolution, as an important index of detection algorithm, is often used to evaluate the performance of the algorithm, but many classical algorithms have shortcomings in this respect. Based on this, this paper proposes a multi-resolution community partition method based on node degree. By introducing similarity coefficient and propagation coefficient of degree, the resolution of network is effectively controlled. Experiments on a scale-free network with 100 nodes show that the proposed algorithm is stable and efficient, with low time complexity. The number of communities can vary from 3 to 27 with different combinations of two coefficients, which can well discover the rules and characteristics of the network.

1. Introduction
In recent years, community structure detection of complex networks has been a hot research topic [1]. Community structure, usually refers to the set of nodes or edges that are closely connected in a complex network. Although there is no clear definition, it reflects the function of the network and has an important impact on the operation of the network. Community detection algorithms for complex networks have been applied to biology and social sciences, which has deepened people's understanding of complex systems. For example, using community detection algorithm, we can find the structure of functional modules in the metabolic network and the laws of biological metabolism [2]. By studying the community structure in social network, we can find social groups with common interests or goals [3]. In a word, it is of great practical significance to study the community structure of complex networks.

In some important community detection algorithms, resolution is usually a difficult problem to solve. The so-called community resolution refers to the ability of the algorithm to find small communities. When the resolution of the algorithm is not high, it may cause that the detected community is the superposition of multiple communities, which leads to the wrong understanding of the network information. For example, the most classic modularity algorithm has the main disadvantage of resolution problem [4]. The communities found by this algorithm are usually similar in size and cannot find small communities. The minimum unit of clique percolation method can be regarded as a small community, but the algorithm is only suitable for sparse networks. Although the label propagation algorithm can find small-scale communities, the algorithm is not stable and the community results obtained are not convincing [5]. Seed diffusion algorithm can achieve multi-resolution community detection results according to different seed selection, but common diffusion
methods are difficult to ensure the locality of the algorithm so that the selected communities may not be in line with the reality [6]. Recently, Li et al. [7] proposed an overlapping community detection method combining seed diffusion and label propagation algorithm, which avoids the uncertainty of label propagation algorithm. Moreover, by using the label propagation of nodes in the network, the seed diffusion algorithm can solve the locality problem. In a word, the algorithm can find the community stably, but it can't control the scale of the community effectively.

Based on algorithm of Li et al. [7], this paper proposes the resolution coefficient and the label propagation coefficient by analyzing the different node degrees. By controlling the two coefficients, we can find the community structure of different scales. The algorithm inherits the advantages of the algorithm of Li et al. [7], overcomes the disadvantage of randomness of label propagation, and can fully consider the localization of nodes.

2. Basic concepts

2.1. Label propagation algorithm
Label propagation algorithm is an efficient semi supervised community detection algorithm [8]. The basic idea of the algorithm is to randomly add a same label to each node in the network. When two nodes are adjacent, the similarity of the two nodes is analyzed. When the similarity is greater than a certain value, the two nodes update their own labels to make the two nodes have the same label. With the propagation of labels, the labels of similar nodes tend to be consistent at the end of iteration. Nodes with the same label are divided into the same community. The idea of label propagation algorithm is simple and its time complexity is low. But the algorithm also has the problem of poor stability. The selection of the initial propagation node of the label propagation algorithm is random. When different nodes are used as the starting point, the community structure detected is different.

2.2. Seed diffusion algorithm
Seed diffusion algorithm is a classic algorithm of overlapping detection algorithm [3]. The basic idea of the algorithm is: select seed node or seed sub network according to certain index, so as to make seeds spread continuously. At the same time, the evaluation function of a certain network community is judged. If the evaluation function increases, the diffused nodes are retained; if the evaluation function decreases, some nodes are eliminated. The seed process is iterated until the evaluation function in the network meets the requirements. The algorithm is very suitable for the detection of overlapping communities, and the selection of seed and evaluation function has many combinations, which makes it more flexible. However, there are also the following problems. Generally, the evaluation function is global, which will make the algorithm lose its locality, and the formation of many network communities mostly comes from the local behaviour of nodes.

2.3. Seed diffusion algorithm based on label propagation
In seed diffusion, if the label propagation algorithm is introduced, the seed will absorb the surrounding nodes according to its own influence. This process fully considers the local characteristics of the network, which is in line with the formation of most networks. For example, Li et al. [7] proposed an algorithm that can distinguish community levels, which not only improves the stability of the label propagation algorithm, but also can verify the detected communities. The algorithm selects seed nodes by the degree of nodes, and then absorbs surrounding nodes by comparing the degrees of neighbouring nodes. However, when dealing with some networks, the phenomenon of community flooding will appear. That is to say, when multiple large degree hubs are connected, the hub with the largest degree will absorb these hubs as members of their own community, and with the continuous propagation of labels, huge communities will be generated. Therefore, the algorithm has a certain resolution problem.

3. Multi-resolution community detection algorithm
In view of the above discussion, this paper proposes a multi-resolution community detection algorithm based on the existing seed diffusion algorithm and the label propagation process. The algorithm avoids the community flooding phenomenon in the literature. Next, we will introduce the algorithm in detail.

3.1. Resolution coefficient and seed determination

The seed selection criteria are as follows. For a complex network $G = (V, E)$, the degree $d_i$ of all nodes is calculated, and the seed nodes are judged by degree centrality. In the literature [7], it is considered that if and only if $d_i > \forall d_{i,nei}$, then the $i$-th node is the seed node, where $d_{i,nei}$ is the neighbour node of the seed. However, there is a problem with this definition. If a node with the largest degree in the network is connected with other nodes with higher degree, a giant community will appear, even including the whole network. Therefore, the resolution coefficient $r$ is introduced in this paper. If and only if the degree of the $i$-th node satisfies equation (1), the node can be used as a seed node.

\[ d_i > \forall (r \cdot d_{i,nei}) \]  

Where, $d$ is the degree of the node, and its expression is:

\[ d_k = \sum_{i}^{n} \sum_{j}^{n} e_g (\delta(i,k) + \delta(j,k)) \]  

Where, $k \in n$ represents the $k$ node in the topology; and

\[ \delta(i,k) = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases} \]  

The recommended value of $r$ is (1, 10]. The larger the value of $r$, the more seed nodes, the higher the resolution of the algorithm. In other words, the degree of seed nodes is not necessarily a local degree peak, thus avoiding the phenomenon of community inundation.

3.2. Propagation coefficient and label propagation

When the seed node is selected, the label propagation algorithm is used to spread the seeds. Starting from the seed node, a label is set for each seed node. At this time, the seed node is the source node, and the node adjacent to the seed node is the target node. When the target node obtains the label, it becomes the next source node, and its label continues to spread to the adjacent target nodes. In the reference [7], the diffusion rule is to compare the degree between the source node and the target node. When the degree of the source node is greater than the degree of the target node, the label of the source node is passed to the target node. Based on this rule, the propagation coefficient $p$ is introduced. When the degree of the source node and the target node satisfy the following equation, the label can be propagated:

\[ d_s > p \cdot d_t \]  

Where, $d_s$ is the degree of the source node and $d_t$ is the degree of the target node. The recommended value of $p$ is (1,3]. The existence of $p$ limits the propagation of labels. Only when the degree of source node is greater than a certain multiple of target node, can labels propagate. The introduction of $p$ is also in line with the reality, that is to say, usually two people with similar influence cannot control each other. Only when the influence difference reaches a certain degree, can the label spread.

3.3. Algorithm flow

According to the discussion in 3.1 and 3.2, the multi-resolution overlapping community detection algorithm proposed in this paper is shown in Figure 1.
Determine the Hub nodes

Take Hubs as sources and their neighbors as targets

Calculate the similarity between sources and targets

If similarity high enough, take these neighbors as new sources

Label propagation to neighbors

If labels do not change, end

Figure 1. The procedures of our algorithms

Step 1. Calculate the degrees of all nodes, and those satisfying equation (1) are named as seed nodes;

Step 2. Using synchronous diffusion rule, label diffusion is carried out. Diffusion starts from the seed node, and the similarity coefficient between the source node and the target node should be determined when the similarity coefficient is greater than a certain value.

Step 3. When the label diffusion does not change the label of seed nodes in the graph, the algorithm ends.

3.4. Time complexity analysis

The algorithm proposed in this paper only adds two parameters $r$ and $p$ to the algorithm in reference [7]. Therefore, the complexity of the algorithm is basically the same as that of the label propagation algorithm. However, if the value of $r$ is too large, the number of source nodes for the first time is too small, and new source nodes will be selected iteratively; similarly, if the value of $p$ is too large, the label propagation of the source node is blocked, and new source nodes will be generated by iteration.

In the worst case, the time complexity of the proposed algorithm is $n$ times of algorithm in reference [7], where $n$ is the number of nodes in the network.

4. Case Study

In this section, we will study the effect of $r$ and $p$ on the results of community detection. In this paper, a self-made scale-free network is used to study. The number of nodes $n$ in the network is 100 and the number of edges $m$ is 165, which conforms to the characteristics of scale-free network, that is, only a small number of nodes have large degree, and other nodes have few degrees. Experimental environment: Windows operating system, processor intel(R) core (TM) i5-8300, CPU @ 2.30 GHz, memory 8 GB. In order to prove that the algorithm can achieve multi-resolution clustering of nodes after introducing $r$ and $p$ parameters, that is to detect communities with different sizes. We adjust different combinations of $r$ and $p$ parameters, the community changes of the network are observed. Figure 2(a) shows the community division in the network when $r = 1$ and $p = 1$. Different colours of nodes in the graph indicate that nodes belong to different communities. We can see that the network is divided into four communities, including three large communities and one small community. At this time, the detection results of our algorithm are the same as those in reference [7]. In Figure 2(b), we increase the value of $p$ to 2, and $r$ remains at 1. It is obvious that the number of communities has changed from 4 to 17. Among these 17 communities, there are five isolated nodes. In Figure 2(c), the number of communities increases from 4 to 13 by keeping $p$ value unchanged at 1 and increasing $r$ value to 2. Moreover, there is no isolated node in the network, and the smallest community is composed of 2 nodes. It can be seen that increasing the value of $r$ and $p$ can improve the resolution of
community detection and then increase the number of community detection. However, with the increase of $p$, there will be isolated nodes in the network, which is mainly because the value of $p$ limits the rules of label propagation and increases the threshold of label propagation. In Figure 2(d), the $r$ and $p$ values are increased to 2 at the same time, and the number of communities increases to 24, among which the number of isolated node communities is 6. Therefore, by adjusting the value of $r$ and $p$, the number of communities in the network is obviously increased, which proves that the algorithm can improve the resolution of the algorithm.

![Network community partition topological graph with different parameters](image)

**Figure 2.** Network community partition topological graph with different parameters

In order to further study the influence of $r$ value and $p$ value on the algorithm resolution, we change the value of $r$ and $p$, and observe the change of the number of communities in the network after the two values change continuously. Figure 3 shows the change of community in the network when $r$ value changes under different $p$ values. It can be seen that with the increase of $r$, the number of communities in the network is on the rise. But there are several stages that need to be addressed. Firstly, when $r$ increases from 1 to 1.1, the number of communities in the network decreases slightly. This shows that in scale-free networks, the node with local peak degree is probably not the best seed node. These nodes are likely to cause the label not to be further propagated, resulting in the emergence of small communities, such as the yellow node in Figure 2 (a). When $r$ is 1.1, the number of seed nodes may not increase, but the candidates of seed nodes change. The newly selected seed nodes can include the yellow areas in their own communities. However, as the $r$ value continues to increase, more and more seed nodes are selected, so the number of communities in the network increases obviously, and the resolution of the algorithm is improved. When the $r$ value exceeds 1.7, the number of communities does not increase, which indicates that for the network studied, there is a peak resolution. When the peak value is exceeded, the number of communities in the network will not increase. Figure 4 shows the change of the number of communities in the network with different $r$ values. It can be seen that the influence of $p$ on the number of communities is more moderate than that of $r$. At the beginning, the number of communities increases slowly until the number of communities increases. The main reason for this kind of curve is that the network is scale-free. When $p$ just
increases, new seed nodes will be quickly found out, and the degree of these new seed nodes is very large. However, when $p$ increases again, because most of the nodes in the network are of small degree, it is difficult to produce new seeds with a slight increase of $p$. In short, it can be seen from Figure 3 and Figure 4 that the new algorithm can indeed adjust the network resolution through $r$ and $p$.

![Figure 3. Relationship between resolution coefficient $r$ and community number](image1)

![Figure 4. The relationship between the propagation coefficient $p$ and community number](image2)

5. Conclusion
This paper proposes a multi-resolution clustering method based on label propagation process. Based on the label propagation algorithm proposed by Li et al.[7], the resolution coefficient and propagation coefficient are introduced. By changing the two coefficients, the resolution of the algorithm is effectively controlled. In addition, the algorithm overcomes the shortcomings of traditional label propagation algorithm, and ensures that the results are stable when the parameters are fixed. In the scale-free network with 100 nodes, through the effective combination of two coefficients, the number of communities can be obtained from 3 to 27, which can well find the rules and characteristics of the network.

References
[1] Garza, S.E., and Schaeffer, S.E.: ‘Community detection with the Label Propagation Algorithm: A survey’, Physica A Stat. Mech. Appl., 2019, 534, pp. 122058
[2] Mattar, M.G., Cole, M.W., Thompsonschill, S.L., and Bassett, D.S.: ‘A functional cartography of cognitive systems’, Plos Computational Biology, 2015, 11, (12), pp. e1004533
[3] Huang, M., Zou, G., Zhang, B., Liu, Y., Gu, Y., and Jiang, K.: ‘Overlapping community detection in heterogeneous social networks via the user model’, Inf. Sci., 2018, 432, pp. 164-184
[4] Newman, M., and Girvan, M.: ‘Finding and evaluating community structure in networks’, Physical Review E, 2004, 69, (2), pp. 26113-26110
[5] Xie, J., Szymanski, B.K., and Liu, X.: ‘SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process’, in Editor (Ed.)’(Eds.): ‘Book
SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process’ (2011, edn.), pp. 344-349

[6] Lu, M., Qu, Z., Wang, Z., and Zhang, Z.: ‘Hete_MESE: Multi-Dimensional Community Detection Algorithm Based on Multiplex Network Extraction and Seed Expansion for Heterogeneous Information Networks’, IEEE Access, 2018, 6, pp. 73965-73983

[7] Li, T., and Zhang, P.: ‘Self-falsifiable hierarchical detection of overlapping communities on social networks’, New Journal of Physics, 2020, 22, (3), pp. 033014

[8] Lu, M., Zhang, Z., Qu, Z., and Kang, Y.: ‘LPANNI: Overlapping Community Detection Using Label Propagation in Large-Scale Complex Networks’, IEEE Transactions on Knowledge and Data Engineering, 2019, 31, (9), pp. 1736-1749