Research Review on Algorithms of Community Detection in Complex Networks

Jingyi Zhang¹, Zhixin Ma¹, Qijuan Sun¹ and Jun Yan¹
¹School of Information Science and Engineering, Lanzhou University, Lanzhou, China
E-mail: 15002526259@163.com

Abstract. As a basic feature of complex networks, community structure has gradually become a research hot spot. The community detection algorithm is helpful to analyze the real structure of the network. It is of great significance for the discovery of community, the analysis of community structure and function and the evolution and development of community. In this paper, these algorithms are analyzed and summarized, and some existing evaluation indicators of community detection algorithms are introduced. Finally, we look forward to the future research directions of complex networks community detection fields with some of my own recommendations.

1. Introduction

Community identification of complex networks has become one of the hot issues in the field of big data research. It has important theoretical and behavioral significance for topology analysis, function analysis and behavior prediction of complex networks [1]. Therefore, the community structure has attracted more and more widespread attention. In 2002, the concept of community structure proposed by Newman thought that the connection between nodes in the community is more closely than the connection between nodes in the different community [2].

Community detection algorithms are divided into overlapping community detection algorithms and non-overlapping community detection algorithms. Non-overlapping community detection algorithms can detect several independent communities, each node uniquely belongs to a community, as shown in figure 1. However, in real large-scale networks, overlapping communities are more common [3], as shown in figure 2. The overlapping community detection algorithm has better practical significance than the non-overlapping community detection algorithm. Firstly, overlapping nodes are the key nodes in the network, so the communities are connected with each other. Secondly, overlapping communities can better reflect the real network structure in the real world.

![Figure 1. Network structure of non-overlapping communities.](image1)

![Figure 2. Network structure of overlapping communities.](image2)
In this paper, the existing community detection algorithms are summarized and introduced based on existing research in the second section, and the characteristics of each algorithm are analyzed. In the third section, some existing community detection algorithm evaluation indicators are introduced. The forth section concludes the work in this paper and gives some thoughts and suggestions for the future research direction in this field.

2. Introduction of Community Detection Algorithms

Nowadays, there have been a large number of community detection algorithms. According to the different starting points of algorithms, these algorithms can be roughly divided into four categories: the algorithms based on modularity, the algorithms based on clique percolation, the algorithms based on label propagation, and the algorithms based on hierarchical partitioning. The following are the analysis of the classical algorithms in these partition algorithms.

2.1. The Algorithms Based on Modularity

Newman and Girvan [4] proposed the concept of modularity in 2004, enabling researchers to have a quantitative understanding of the concept of the community. In later studies, modularity was used to evaluate the quality of community partition. The larger the value, the better the corresponding community structure.

The FastQ algorithm proposed by Newman [5] initializes each node as a community, and merged the two communities with the largest increment or the smallest decrease in modularity. Repeating the process until all the nodes in the network belong to the same community. FastQ uses the physical meaning of modularity: the greater the modularity, the better the corresponding community structure. Therefore, through this merger, the modularity of the divided community structure increases in the direction of maximization. FastQ is a greedy strategy. Its output is a dendrogram, disconnected from the maximum of modularity to obtain the final community structure. The time complexity of the algorithm is low, but it can only obtain the approximate optimal community structure.

Clauset et al. found that in the FastQ algorithm, the communities that need to be merged each time are very time-consuming. The real world network is often a sparse network, so the combined update operation can also be implemented in a more efficient manner. Therefore, Clauset et al. [6] improved and re-implemented the FastQ algorithm to form the CNM algorithm. The data structure used in this algorithm is balanced binary trees and max heaps, so as to quickly select two communities that can be merged, thereby reducing the time complexity of the original algorithm.

The Louvain algorithm proposed by Blondel et al. [7] is a heuristic algorithm based on modularity optimization. The algorithm can quickly extract hierarchical community structures from the network. Its goal of optimization is to maximize the results of modularity of the entire community division. When the Louvain algorithm calculates the changed values of modularity that moves a node from one community to another, it only involves the current two communities, and it has nothing to do with other communities in the network. So it can be calculated quickly and the number of communities is reduced during the iteration very fast. Therefore, the execution efficiency of Louvain algorithm is very high, it is an unsupervised community detection algorithm, and the algorithm itself is very simple and easy to implement.

Waltman et al. [8] proposed a smart local moving algorithm (SLM) based on the idea of Louvain algorithm. After the first phase of the Louvain, the SLM reapplies the Louvain to a series of small communities in each community divided by the first phase of the Louvain algorithm. In the second phase, each small community is used as a vertex to build a new network. The Louvain algorithm cannot further improve the community structure by merging or moving individual nodes from one community to another in some cases, but there is no such situation in the SLM algorithm. When the SLM algorithm is run iteratively, the possibility of increasing the modularity is constantly sought by segmenting the community and moving the set of nodes from one community to another. Therefore, there can always be further improvements in the community structure by performing more iterations of the algorithm in the algorithm of SLM.
2.2. The Algorithms Based on Clique Percolation

The early community detection algorithms are basically applied to the non-overlapping community structure. With the proposal of the clique percolation algorithm, the research field of community detection algorithms is moving towards the overlapping community structure. The main idea of this type of algorithm is to regard the community as a aggregation of cliques. A clique is a complete sub-graph where any two nodes are connected, and each clique is closely connected by a shared node.

The clique percolation method (CPM) proposed by Palla et al. [9] in 2005, believing that the internal edges of the community are more likely to form complete sub-graphs based on the principle of close connection between internal nodes of the community. Therefore, by identifying the cliques, the purpose of identifying the community structure in the network can be achieved. The algorithm needs to set a custom parameter k, indicating the number of nodes in the search for a clique. Firstly, searching for the k-clique in the network and establishing a new graph with the k-clique as a node. Then, if there are k-1 common nodes between the two k-cliques in the original graph, the new graph will connect the two cliques. Finally, each connected sub-graph is a community. Since k in the algorithm is a user-defined parameter, the value of k will affect the detection result of the CPM. The smaller the value of k, the larger the community that the algorithm will eventually detect, and the more sparse the community structure. At the same time, the algorithm has a defect that cannot allocate nodes outside of the complete sub-graph. For real networks, especially for sparse networks, the conditions of CPM are too strict and only a small number of overlapping communities can be found.

The sequential clique percolation algorithm (SCP) proposed by Kumpula et al. [10] abandons the idea of CPM and uses a method of serialization to implement community division. The algorithm is divided into two steps. Probing all the k cliques instead of finding the largest one, and then probing the community structure. The idea of the SCP is as follows. If two cliques of size k-1 are members of the same k clique, and the two k cliques are connected through k-1 nodes, then the two k-1 cliques can be put in the same community. In most cases, the SCP is more efficient than the CPM. However, because the SCP does not consider the maximum number of cliques, but the number of k-1 cliques, and a largest clique often contains many small k cliques, it also contains many small k-1 cliques, so the number of k-1 cliques obviously exceeds the number of the largest cliques in the SCP. This will result in more comparisons than in the circumstances of finding largest cliques. So it will seriously affect performance in large complex networks. The SCP is very efficient when k is very small, but in some networks when k is large, the time complexity is quite high.

Lee et al. [11] proposed a greedy clique expansion (GCE) algorithm, which first detects all maximal cliques with at least k nodes in the graph as seeds, and uses the fitness function to expand the current unexpanded maximum seeds. Expanding until no node is added to increase the fitness value, then selecting the largest seed in the unexpanded remaining graph to continue to expand. The fitness function used by the algorithm is shown in formula (1).

\[ F_s = \frac{k_{in}^S}{(k_{in}^S + k_{out}^S)^\alpha} \]  

\[ k_{in}^S \] represents the internal degree of the community S; \[ k_{out}^S \] represents the external degree of the community S; \[ \alpha \] represents an adjustable parameter, and it usually takes a value of 0.9-1.5. Finally, the distance between all communities is detected. The community distance measurement function is shown in formula (2).

\[ \delta_E(S, S') = 1 - \frac{|S \cap S'|}{\min(|S|, |S'|)} \]  

Where S and S' represent two different communities. The formula can be interpreted as the proportion of the nodes in the smaller community and not in large communities in the network. The algorithm of GCE should final fusion of similar communities because some community structures are similar. The algorithm thinking is easy to understand. And the algorithm proposes that GCE has achieved excellent
results in the paper co-author network, and in other networks, such as protein networks, the social network and so on, also have better effect. However, because the algorithm randomly selects seeds, the result is less stable.

2.3. The Algorithms Based on Label Propagation

Generally, in a complex network, edges between nodes represent the propagation of information between individuals. According to community characteristics, it can be known that nodes in a community share the same information, while nodes in different community share different information. So the community detection algorithms based on label propagation was generated.

The label propagation algorithm (LPA) was proposed by Zhu et al. [12] in 2002. It is a graph-based semi-supervised learning method. The basic idea is to use label information of labeled nodes to predict the label information of unlabeled nodes. In 2007, Raghavan et al. [13] applied LPA to community detection for the first time. As shown in figure 3 is the process of label propagation. The basic idea of this algorithm is to assign a unique label to each node and iteratively update the labels of all nodes. During each iteration, for the current node, the label with the most number of occurrences in the neighbor nodes is assigned to the current node; when the label with the most number of occurrences in the neighbor nodes is not unique, a label is randomly selected to be assigned to the current node. The label of the nodes in the network no longer changes and the algorithm terminates. Finally, nodes with the same label belong to one community.

![Figure 3. The process of label propagation](image)

The biggest advantage of the label propagation algorithm is that the algorithm does not require any parameters, and it has a linear time complexity, so the execution efficiency is very high. But the algorithm can only detect non-overlapping community structures.

In order to be more applicable to the real network structure, Gregory [14] proposed a community overlap propagation algorithm (COPRA) extended the LPA to enable detection of overlapping community structures. The algorithm introduces a new label structure \((c, b)\), \(c\) denotes the community identifier, \(b\) denotes the subordinate coefficient of node \(x\) in community \(c\). The value is between 0 and 1, and the sum of all the subordinate coefficients of node \(x\) is equal to 1. The rules for updating the community subordinate coefficient of a node are shown in formula (3).

\[
b_t(c, x) = \frac{\sum_{y \in N(x)} b_{t-1}(c, y)}{|N(x)|} \tag{3}
\]

\(b_t(c, x)\) represents the subordinate coefficient of the node \(x\) in the community \(c\) when iterating \(t\) times; \(N(x)\) represents the adjacent node set of the node \(x\). We know that if there are multiple labels available, the algorithm will still randomly select through experiments. It leads to the instability of the algorithm.
results and unpredictability of mining community conditions. At the same time, the algorithm is limited the number of communities to which overlapping nodes belong. On the basis of COPRA, the BMLPA proposed by Wu et al. [15] removes the limitation on the number of communities overlapping nodes belong to and further improves the original algorithm.

In 2013, the SLPA proposed by Xie et al. [16] also expanded the LPA. Each node in the algorithm has a label storage list. The labels are processed by post-processing thresholds. Nodes with multiple labels are called overlapping nodes. So LPA to SLPA also completed the leap from non-overlapping community detection algorithms to overlapping community detection algorithms.

Wang Ting et al. [17] based on SLPA’s improved HLPA creatively uses a hybrid update strategy to improve the efficiency of label propagation and avoid the swaying phenomenon that occurs when the label is propagated in a bipartite graph or an approximate bipartite graph.

In the later study, Liu et al. [18] proposed the ELPA by combining the natural advantages of the link community with the efficiency of the LPA. Jia et al. [19] proposed the SSCLPA detect the initial community based on the similarity score of the Sørensen-Dice index, using different update strategies for allocated and unallocated nodes and constraint conditions so as to obtain the community structure in the network.

2.4. The Algorithms Based on Hierarchical Partitioning

The community detection algorithms based on hierarchical partitioning includes two types: split hierarchical method and condensed hierarchical method. The former splits the entire network from top to bottom until a single node is treated as a community; the latter is opposite, seeing a single node as a community through a single link condensed and merged into a community from bottom to top.

Girvan and Newman [20] proposed the structure of a hierarchical community, namely the GN algorithm, by repeatedly deleting the edge with the highest edge betweenness. The algorithm first calculates the edge betweenness of all edges in the network. Removing the edge with the highest edge betweenness. Then calculating the edge betweenness of the remaining edges in the network, and repeats the process until there is no edge in the network. Finally a top-down hierarchical tree is obtained. The tree can be divided into hierarchical layers with the largest modularity. Because GN algorithm needs to repeatedly calculate the edge betweenness of each edge, the time complexity is very high in complex networks, but the accuracy of the algorithm is better.

Gregory [21] proposed CONGA algorithm to extend the GN. The algorithm introduces the split betweenness of nodes based on the edge betweenness. The algorithm first computes the edge betweenness of each edge and the split betweenness of each node in the network. Then it deletes the edge with the largest edge betweenness or splits the node with the largest split betweenness. Calculating the edge betweenness and split betweenness again. Iterating until there are no edges in the network. The algorithm has a large amount of network computation for a large number of nodes, thereby increasing the time complexity of the algorithm.

The local fitness measure (LFM) proposed by Lancichinetti et al. [22] can still find both overlapping and hierarchical structures. This algorithm is the earliest community detection algorithm that can discover overlapping and hierarchical structures at the same time. The complexity of the algorithm is $O(n^2)$.

3. Index of Algorithm Evaluation

Nowadays, the indicators for evaluating the performance of community detection algorithms are accuracy, normalized mutual information, modularity, Jaccard coefficient, clustering coefficient, etc. This paper focuses on three classical evaluation indicators: accuracy normalized mutual information and modularity.

In classification and cluster analysis, accuracy is a commonly used measure. It uses the real category information of the data as a benchmark to measure how many data are divided into the correct groupings. This is a relative concept with another indicator “misleading rate”. Therefore, in the community detection, the definition of accuracy is as shown in formula (4).
\[ A = \frac{\text{number of nodes assigned to the correct community}}{n} = 1 - \frac{\text{misclassified number of nodes}}{n} \quad (4) \]

Where \( A \) represents accuracy and \( n \) represents the number of nodes in the network. The value of the accuracy is \([0,1]\). The larger the value, the less the nodes that are misclassified, and the closer the community structure is to the real network structure. The accuracy of calculation needs to know the true community division structure of the network and it is a typical external evaluation index.

Normalized mutual information (NMI) is usually used to measure the difference between the community structure of the community detection algorithm and the real community structure. Then the accuracy of the community detection algorithm can be measured \([23]\). The definition of NMI is shown in formula (5).

\[
\text{NMI}(P_1, P_2) = \frac{2[H(P_1) + H(P_2) - H(P_1, P_2)]}{H(P_1) + H(P_2)}
= \frac{-2 \sum_{i=1}^{C_1} \sum_{j=1}^{C_2} N_{ij} \log \frac{N_{ij}N}{N_iN_j}}{\sum_{i=1}^{C_1} N_i \log \frac{N_i}{N} + \sum_{j=1}^{C_2} N_j \log \frac{N_j}{N}} \quad (5)
\]

\( H(P_j) \) represents the entropy value of the actual community structure partition; \( H(P_2) \) represents the entropy value of the community structure divided by the algorithm; \( H(P_1, P_2) \) represents the joint entropy value of the community structure divided by the actual community structure and algorithm; \( C_j \) represents the actual number of communities; \( C_i \) represents the number of communities divided by the community detection algorithm; \( N \) is the total number of nodes; \( N_{ij} \) represents the number of nodes of communities \( i \) and \( j \) respectively. The NMI value is in the range \([0,1]\). The larger the value, the closer the division of the community structure to the real community network, and the higher the accuracy of the algorithm. NMI is also a typical external evaluation indicator.

The use of modularity \([4]\) to evaluate the quality of community classification does not require the prior knowledge of the community real division results. The general form of modularity is shown in formula (6).

\[
Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(C_i, C_j) \quad (6)
\]

\( A \) is the connection matrix of the network; \( m \) represents the total number of edges in the network; \( P_{ij} \) represents the expected number of edges between node \( i \) and node \( j \) in the random graph; \( C_i \) and \( C_j \) represent the communities where node \( i \) and node \( j \) reside; When node \( i \) and node \( j \) belong to the same community, \( \delta \) takes a value of 1, otherwise \( \delta \) takes a value of 0. In general, the value of \( Q \) is in the range \([0,1]\). When \( Q > 0.3 \), it can be considered that there is a clear community structure in the network. In a given network, when the modularity reaches a maximum, the community is divided into the optimal segmentation of the current network at this time. The solution to the maximum modularity has been proved to be a NP-complete problem, resulting in a series of algorithms based on greedy ideas have been generated. At the same time, it can be found that the modularity of large-scale networks is also relatively large. Therefore, in the process of evaluating the quality of community classification, the quality of community divisions of networks of different sizes cannot be compared simply by the size of the modularity.

4. Conclusion
This paper analyzes and summarizes some of the commonly used classical community detection algorithms, and elaborates the representative evaluation indicators of this algorithm. In recent years, community detection has always been a key work in the field of complex networks. With the gradual complexity of networks in the real world, community detection research has become increasingly important. For the future research work, this paper proposes the following two points of thinking: (1)
the vast majority of overlapping community detection algorithms are limited to the discovery of overlapping nodes. There is no further study. For example, an overlapping node belongs to multiple communities. Nodes are more subordinate to which community is worthy of our further exploration;

(2) Since most networks in the real world will change over time, community detection algorithms should be considered more suitable for dynamic networks. Therefore, the research of community detection algorithms still has a broader research background and practical significance.

5. References

[1] Di Jin, Dayou Liu, Bo Yang, et al. Fast and complex network clustering algorithm based on local probing [J]. Journal of Electronics, 2011, 39 (11): 2540 - 2546.

[2] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99 (12): 7821.

[3] Palla G, Derényi I, Farkas I, et al. uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814.

[4] Newman M E, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69 (2 Pt 2): 026113.

[5] Newman M E J. Fast algorithm for detecting community structure in networks. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2003, 69 (6 Pt 2): 066133.

[6] Clauset A, Newman M E, Moore C. Finding community structure in very large networks. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 70(6 Pt 2):066111.

[7] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 2008 (10):155-168.

[8] Waltman L, Eck N J V. A smart local moving algorithm for large-scale modularity-based community detection [J]. European Physical Journal B, 2013, 86 (11):1-14.

[9] Palla G, Derényi I, Farkas I, et al. uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814.

[10] Kumpula J M, Kivelä M, Kaski K, et al. Sequential algorithm for fast clique percolation[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2008, 78 (2 Pt 2): 026109.

[11] Lee C, Reid F, Mcdaid A, et al. Detecting highly overlapping community structure by greedy clique expansion [J]. 2010.

[12] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation [J]. 2002.

[13] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical review E, 2007, 76(3): 036106.

[14] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2010, 12 (10): 103018.

[15] Wu Z H, Lin Y F, Gregory S, et al. Balanced multi-label propagation for overlapping community detection in social networks [J]. Journal of Computer Science and Technology, 2012, 27 (3): 468 - 479.

[16] Xie J, Szymanski B K. Towards linear time overlapping community detection in social networks[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2012: 25 -36.

[17] Wang T, Qian X, Wang X. HLPA: A hybrid label propagation algorithm to find communities in large-scale networks[C]// IEEE, International Conference on Awareness Science and Technology. IEEE, 2015:135-140.

[18] Liu W, Jiang X, Pellegrini M, et al. Discovering communities in complex networks by edge label propagation [J]. Sci Rep, 2016, 6: 22470.

[19] Jia H C, Ratnavelu K. A semi-synchronous label propagation algorithm with constraints for community detection in complex networks [J]. Scientific Reports, 2017, 7: 45836.

[20] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the national academy of sciences, 2002, 99 (12): 7821-7826.

[21] Gregory S. Finding overlapping communities using disjoint community detection algorithms [J]. Complex networks, 2009: 47-61.
[22] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 033015.
[23] Steinhaeuser K, Chawla N V. Identifying and evaluating community structure in complex networks [J]. Pattern Recognition Letters, 2010, 31(5): 413-421.