Overview of Algorithms for Detecting Community Structure in Complex Networks

Man Gao [a], Yinghong Ma [b]

School of management science and engineering, Shandong Normal University, Jinan, 250014, China

[a]386635483@qq.com, [b]yinghongma71@163.com

Abstract

Community structure is a very important property of complex networks. Detecting communities in networks is of great importance in biology, computer science, sociology and so on. In recent years, a lot of community discovery algorithms have been proposed aiming at different kinds of large scale complex networks. In this paper, we review some latest representative algorithms, focusing on the improved methods based on the modularity function, the algorithms which can detect overlapping and hierarchical community structure in networks, and the benchmark in detecting communities. Finally, some future directions are pointed out.

Keywords

Complex Network, Community Structure, Modularity Function, Overlapping Communities.

1. Introduction

As the physical meaning and mathematical characteristics of the network properties of in study, it was found that many real networks all have a common nature, namely the community structure. In recent years, the community structure analysis in biology, computer graphics and has wide application in sociology [1]. Current research on community structure has a lot of algorithm in complex networks, this article introduces several representative methods.

2. The traditional community discovery algorithm

The traditional community found is divided into two categories, the main methods in the computer graphics division and hierarchical clustering in sociology.

Figure segmentation is usually through iterative segmentation will be divided into sub network: first the network is divided into two sub networks according to the optimal division, and then repeat to the optimal binary of these sub network, until into a given number of subnet. Two famous algorithm is Kernighan-Lin [2] and spectrum split method based on Laplace matrix [3].

Kernighan-Lin algorithm is a heuristic optimization method. It is based on greedy algorithm principle could be divided into two scale known network community. Its basic idea is to introduce a gain for the division of the network function of Q, defined as the number of edges of two corporate internal minus the connection between the two clubs number of edges, and then look for ways to make the Q value of the largest division. Spectrum split method is based on the second smallest Eigen value of the corresponding eigenvector for binary network could be divided into two scale known network community. Due to network corresponding Laplace matrix is real symmetric matrix, the feature vector are orthogonal to each other, therefore, the second smallest Eigen value corresponds to the feature vector is composed of positive and negative elements of the vector, spectrum split method is according to vector elements of positive and negative characteristics of the network node is divided into two parts.

Kernighan-Lin algorithm and Spectrum split method main drawback is that cannot guarantee iterative binary can get the correct classification (for example, there are many results will be divided into three sub network), and the lack of effective to determine when to stop binary code. For Kernighan-Lin

algorithm, it before the start of another major problem is that the algorithm needs to know the size of the two sub networks, otherwise the algorithm will get accurate results.

3. Hierarchical clustering method

Hierarchical clustering is based on the similarity between nodes of the network into a tree form, cut in a certain place, get the corresponding community structure. There are many kinds of definition method of node similarities, commonly used with two main node number of adjacent nodes at the same time, the Euclidean distance between nodes, the correlation coefficient, etc. Hierarchical clustering method does not need to specify the number of network community and the community size, but the hierarchical clustering method is not sure which is divided into the network of the optimal partition in addition, the hierarchical clustering method relies on the similarity of nodes selection, easy to network of some nodes into separate organizations or divide the outside of the network nodes can not correctly.

4. Splitting method and condensation method

According to add edges and remove edges, the algorithms can be divided into two small categories: condensation method and splitting method.

Condensation method of thought is to use a method to calculate the similarity between the nodes, then node to start from the top of the similarity, add edges of the original space network. On the contrary, the partition algorithm, usually from the focus of network, try to find a minimum similarity of the connected nodes , and then connect them in turn. Repeat this process, the whole network is gradually divided into smaller parts.

4.1 Split method

The traditional network partitioning method can not effectively deal with network partitioning when community size and number of the community to determine or such problems as how to determine the optimal partition, and community structure of the network analysis often require algorithm can accurately and automatically determine the number of community network and give the corresponding corporate division of "natural". Therefore, in recent years a large number of effective algorithm is more suitable for the analysis of the network society is put forward.

Splitting method is the most typical algorithms in Girvan and Newman was proposed in 2002 by a through edge removed according to the hierarchical decomposition of network community analysis method (The GN algorithm [4]). The GN algorithm is to find the most likely in the association between the sides, through continuous removing the side get the hierarchy of the network. The basic idea is through continuously remove edge from the network through Betweenness (the largest handing the Betweenness). Among them, the edge betweenness is defined as a network after the number of the edges of the shortest path. For network with community structure, edge tend to have high betweenness between community, by removing these numerical boundary, the potential of community structure can be divided into network. The GN algorithm makes up for the shortcomings of some traditional methods, however, not sure which step to stop . To solve this problem, therefore, Girvan and Newman proposed module degrees [5] as the criteria for measuring corporate division, with the highest degree of module value into the community structure of most likely. In addition, based on the clustering algorithm [6]and [7] based on the information center degree algorithm is also divided method of the common algorithm.

4.2 The condensation method

Because the GN algorithm complexity is higher, Newman presents a fast algorithm (NFA) [8]through the way of constant consolidation nodes directly optimal values for the network community structure, starting from nodes group consisting of a single, each time with edge connected and combined makes value increase or reduce the largest at least two groups of nodes, until all the nodes are combined as a collection of nodes. On the basis of the fast algorithm, the algorithm CNM [9] adopt heap data structure to calculate and update the network module, greatly improving the calculation speed.

5. Based on the module optimization algorithm

In recent years, based on the different understanding of community structure, many styles are derived from the week of the new algorithm. Can be divided into division algorithm, based on module degree method, the dynamic algorithm and the method based on information theory.

Girvan and Newman was originally used for the GN algorithm proposed module degrees stop indicators, but quickly become a vital element of other algorithm, formed a kind of algorithm based on the module. The typical based on module algorithm with simulated annealing algorithm (SA) [10]and extremal optimization algorithm (EO) [11]. Similarly, the genetic algorithm (GA) [12] and Tabu algorithm (TA) [13] also is often used to optimize module Q values.

6. Based on the label propagation algorithm

Module optimization methods can't found that less than a certain size of community. In actual network, especially in large-scale network, corporate size, the problem is particularly prominent. Dynamic method arises at the historic moment, the most representative method of dynamic label propagation algorithm (algorithm) [14][15] and random walk algorithm.

Based on label propagation algorithm is a kind of heuristic algorithm. 2007 Raghavan first put forward using the label transmission technologies such as identify the community structure of the network. When the initial, every node in the network is assigned to the only digital label. In the process of each iteration of the algorithm, each node USES most of its neighbor nodes have labels. When there are multiple target number of nodes neighbors most tags, are randomly chosen in these labels as a target node identification. As the label on the network transmission in accordance with the above way, dense link node group will be in reach a consensus on the label. At the end of the algorithm, node group with the same label will gather together to form a club. The algorithm is simple in concept, easy to realize and the characteristics of fast and efficient.

7. The method based on markov dynamics

Markov random walk theory has been applied in the field, Random walk on the network is a common network dynamics, wandering from one node in network, in accordance with a certain probability randomly select one of the nodes adjacent nodes as the next step of starting point, and repeat this process. Assume a network consists of n nodes, first thinks that the network has n community divided according to the two clubs in the choice of similarity index, consolidated community for a new club and establish a new division: update the community similarity between after step, end of the algorithm and get every step to determine a corresponding divided into a community structure of complex, the tree is a hierarchical structure of the community system.

In 2000, Dongen put forward famous markov clustering algorithm MCL [16]. The method for the probability of random walk, to simulate the flow in the network by adjusting the corresponding transition probability calculation method of markov process, strengthen the strong flow, reduced weak stream, and make the network become clear community structure. According to the similarity between data points finish clustering [17]. Orson, after put forward the FEC algorithm based on the random walk model [18].

8. Overlapping community discovery algorithm

Previously introduced non-overlapping community discovery method divide each node strictly to a particular community, and the real world, this kind of non overlapping community structure sometimes does not reflect the real network structure, so overlapping community become a new hotspot in the research of the community found in recent years. Currently has many different types of overlapping community discovery method is put forward, according to their basic theory and technology adopted by the different, the main can be divided into: mass of seepage theory method, local extension and optimization method, the fuzzy association method, etc. By Palla according to high density, on the edge of society to make corporate internal easy to form a circle (clique) and the

association between edge is unlikely to form the fact that this circle, this paper proposes a Percolation Method called circle (clique Percolation Method, CPM) [19]. First you need to specify a k value as a priori (4); Then find out the network in all size greater than k group; After the group as vertices, when between the two groups have no less than k - 1 node overlapping formed when an edge, to build a mass of figure; Mass of each component in the figure, the collection of the corresponding group is to a club. In addition, Kumpula and others contain only a given network node according to the vast network of sequentially inserted into the network edge, this paper proposes a fast and kind of CPM algorithm can analyze the weighted network community [20]. Zhang and others to use the fuzzy k-means for network community structure of the fuzzy partition [21]. The method through the club membership degree to each node (value within the range of [0, 1]) to expand module degree to measure the evaluation index of fuzzy community divided into good or bad. The determination of k value as well as the feature vector to solve the efficiency of the algorithm is based on the analysis of large-scale network community structure is lower.

In addition, there are many algorithms [22-24]not introduce on here.

9. Conclusion

The community structure of complex networks has become a very challenging today and future research field. This paper introduces the find in the network community structure of several famous algorithms. Mainly includes the computer graphics and the sociology of two kinds of hierarchical clustering algorithm. People take advantage of these algorithms are analyzed many actual network, divided into their community structure, and the nature of these associations are analyzed. How to according to the characteristics of the heterogeneous network, find a fast and reliable network community structure analysis algorithm, is still the main problems need to be solved in the future.

References

[1] Girvan M, Newman M E J. Community structure in social and biological networks. Proc Nail Acad Sci USA, 2001, 99(12):7821-7826
[2] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49(2):291–307.
[3] POTHEN A, SIMON H D, LIOU K P. Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM J. Matrix Anal. Appl., 1990, 11(3):430–452.
[4] 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.
[5] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J].Physical Review E, 2004, 69(2):026113.
[6] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9):2658.
[7] FORTUNATO S, LATORA V, MARCHIORI M. Method to find community structures based on information centrality [J]. Phys. Rev. E, 2004, 70.
[8] NEWMAN M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6):066133.
[9] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks [J]. Physical Review E, 2004, 70(6):066111.
[10] KIRKPATRICK S, JR. D G, VECCHI M P. Optimization by simmulated annealing [J]. science, 1983, 220(4598):671–680.
[11] GUIMERA R, AMARAL L A N. Functional cartography of complex metabolic networks [J].Nature, 2005, 433(7028):895–900.
[12] HOLLAND J H. Adaptation in natural and artificial systems [J]. 1975.

[13] GLOVER F. Future paths for integer programming and links to artificial intelligence [J]. Computers & Operations Research, 1986, 13(5):533–549.
[14] RAGHAVAN U, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E, 2007, 76(3):036106.
[15] NOH J D, RIEGER H. Random walks on complex networks [J]. Phys. Rev. Lett., 2004, 92(11):118701.
[16] VAN DONGEN S. Graph clustering by flow simulation [J]. PhD Thesis University of Utrecht, 2000.
[17] FREY B J, Dueck D. Clustering by passing messages between data points [J]. Science. 2007, 315(5814): 972-976.
[18] YANG B, CHEUNG W K, Liu J M. Community mining from signed social networks [J].IEEE Transactions on Knowledge and Data Engineering. 2007, 19(10): 1333-1348.
[19] PALLA G, DER ˝ENYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043):814–818.
[20] KUMPULA J M, KIVEL ¨A M, KASKI K, et al. Sequential algorithm for fast clique percolation [J]. Physical Review E, 2008, 78(2):026109.
[21] ZHANG S H, WANG R S, ZHANG X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering [J]. Physica A: Statistical Mechanics and its Applications, 2007, 374(1):483–490.
[22] WANG XY, Li JQ Detecting communities by the core-vertex and intimate degree in complex networks [J], Physical A .2013. 392407, 2555-2563.
[23] SHI C, Cai YN, FuD, etal. A Link Clustering based overlapping community detection algorithm [J].Data knowledge Engineering, 2013, 87: 394-404.
[24] CONG M, Liu J, Ma L, etal. A efficient agent-based algorithm for overlapping community detection in networks [J], Physica A, 2014, 403:71-84.