ISSN: 1813-4890

The State of Community Discovery Research

Wenzhang Wang

Tianjin University of Technology and Education, Tianjin 300222, China

Abstract

Community discovery is the process of finding out the community structure given a network graph. The subgraphs corresponding to the sub-sets of nodes with close internal connections are called communities. The node sets of each community that have no intersection with each other are called non-overlapping communities, and those that have intersections are called overlapping communities. There are many different community discovery algorithms, traditional algorithms include hierarchical clustering algorithms, spectral methods, graph segmentation, and label propagation. The optimization methods include splitting method, spectral method, dynamic algorithm and so on. There are pros and cons to each method. Under different circumstances, making selections can more accurately and quickly discover community structures in complex networks.

Keywords

Community Discovery; Overlapping Communities; Hierarchical Clustering; Spectral Methods; Dynamic Algorithms.

1. Introduction

With the development of society and the advancement of science and technology, the connection between people has become more frequent and the relationship has become closer, and this complex connection has formed a complex social network. Similar to protein interaction network, e-mail network, gene association network, metabolism network, transportation network and so on. This type of network is called complex network because of its complex structure, network evolution, diversity of connections and nodes, and multi-complexity fusion [1]. The study of complex networks has always been a research hotspot in many fields. Community structure is a common feature in complex networks, and the whole network is composed of many communities. Points within a community are tightly connected, while those between communities are sparsely connected. Finding the correct social structure in a network is crucial to understanding the structure and function of the entire network.

At present, the main applications of community discovery: in the biological field, metabolic network analysis, gene regulation network analysis, master gene identification, etc. By analyzing the virus transmission network, identify key communities and susceptible groups of infectious diseases, strengthen protection, cut off the transmission path, and control the spread of the virus. In e-commerce, community discovery is used for more accurate advertisement placement, so as to establish a more reliable recommendation system and realize personalized interest recommendation. In addition, the use of community discovery to analyze criminal activities can effectively combat criminal networks and maintain social stability.

2. Related Research

In recent years, intensive research has been carried out on mining the community structure of complex networks. Traditional community discovery algorithms include hierarchical clustering algorithms, spectral methods, graph segmentation, label propagation, etc. The advantage of

ISSN: 1813-4890

these algorithms is that they can better discover the network community structure, but the disadvantage is that when the network is large or incomplete, it will be subject to some constraints [3].

2.1. Hierarchical Clustering Algorithms

Hierarchical clustering algorithms assume that there is a hierarchy of communities. Calculate the similarity between nodes by a certain similarity measure, and sort the nodes according to the similarity from high to low, and gradually reconnect each node. Among them, there are two kinds of aggregation method and split method:

1. Cohesion method: According to the similarity from strong to weak, the corresponding node pairs are connected to form a dendrogram, and the dendrogram is cross-cut according to the needs to obtain the community structure. 2. Splitting method: find out the weakest interconnected nodes and delete the edges between them, and divide the network into smaller and smaller components through such repeated operations, and the connected network constitutes a community.

The advantage of the hierarchical clustering algorithm is that it can help us interpret the clustering results in a visual way by drawing a dendrogram. Another advantage of hierarchical clustering is that it does not require prior specification of the number of clusters.

2.2. Graph Segmentation

The graph segmentation method is to regard the community as a dense subgraph structure, and divide the nodes in the graph into n groups of predetermined size, and the number of edges between these groups is the lowest. The early segmentations were all bipartite graphs, and in the case of multiple divisions, one of the subgraphs was subdivided. The KL algorithm decomposes the network into 2 communities of known size through a heuristic process based on greedy optimization. The algorithm introduces a gain function for the division of the network, which is defined as the difference between the number of edges in the two communities and the number of edges in the two communities, and seeks the maximum division method of Q. The disadvantage of the KL algorithm is that the size of the two subgraphs must be specified first, otherwise the correct result will not be obtained, and the practical application is of little significance.

The algorithm based on max flow was proposed by G.W.Flake. He added virtual source nodes and end nodes to the network, and proved that after the maximum flow algorithm, the community containing the source node just satisfies the property that the nodes in the community have more links than the links outside the community.

2.3. Graph Clustering Algorithms

The graph clustering algorithm is derived from the graph partition theory, and the core is to regard the clustering problem as a graph segmentation problem. The clustering process is actually an optimization of the graph partitioning process. The purpose of optimization is to make the similarity between subgraphs smaller and the similarity within subgraphs larger. The GN algorithm is a split-type hierarchical graph clustering method. The FMM, CNM, and BGLL algorithms all belong to the agglomerative hierarchical graph clustering method based on the maximization of modularity. Iteratively selects clusters that increase the modularity of the current cluster to merge. Until the cluster structure with higher modularity can no longer be divided [6]. Maximizing modularity may fail to identify many small-scale clusters that actually exist. At the beginning of Louvain algorithm, each node is regarded as an independent community, and then the node with the largest modularity gain is selected from the neighbor nodes to join, and it is gradually merged until all nodes in the network are traversed. The algorithm performs well in terms of efficiency and effect, and can discover hierarchical

ISSN: 1813-4890

community structures, and its optimization goal is to maximize the modularity of the entire graph attribute structure (community network).

2.4. Label Propagation Algorithm

Label propagation algorithm is a kind of heuristic algorithm. The main process is to initially assign a unique label to each node, and then propagate the label according to the similarity of each node and its neighbors, so that nodes with the same label are finally divided into the same community [7]. Common algorithms include LPA, SLPA, LPPB, and LPANNI algorithms.

The main idea of the LPA algorithm is that at first each node has an independent label, then there are n different labels in the network, and in each iteration, for each node, change its label to the label that appears most frequently in its neighbors, if such a label If there are more than one, select one at random. Through iteration, until the label of each node is the same as the label that appears most frequently among its neighbors, a stable state is reached, and the algorithm ends. At this point, nodes with the same label belong to the same community.

Because the calculation process of the LPA algorithm is relatively simple, it does not need to optimize any function, so the algorithm is faster, and the number of communities in the network can be determined by itself. The time required to assign labels to nodes is O(n), and the time spent to update labels during label propagation is O(m), so the complexity of the algorithm is O(m+n) [8].

The SLPA algorithm divides the complex network into communities through the historical label information of nodes in the process of label information propagation. The time complexity of the SLPA algorithm is O(tm), which is related to the number of edges in the network and the number of iterations. Therefore, the SLPA algorithm is not suitable for dealing with denser networks, and it is not easy to determine the number of iterations.

3. Evaluation Standard

Evaluating the quality of a community discovery algorithm is usually considered in terms of modularity, standard mutual information, adjusted Rand coefficient, accuracy, and separation. When we know the real community division results, the standard mutual information NMI and the adjusted Rand coefficient ARI are often used as evaluation indicators. When we do not know the result of the real community division, the modularity Q is often used as an evaluation metric.

3.1. Normalized Mutual Information

The Normalized mutual information NMI is used as the evaluation standard of artificially generated network division. The closer the NMI value is to 1, the better the effect of the community divided by the algorithm. The definition of standardized mutual information NMI is as follows:

$$NMI = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(\frac{N_{ij}\times N}{N_{i\cdot}\times N_{\cdot j}}\right)}{\sum_{i=1}^{C_A} N_{i\cdot}\log\left(\frac{N_{i\cdot}}{N}\right)+\sum_{j=1}^{C_B} N_{\cdot j}\log\left(\frac{N_{\cdot j}}{N}\right)} \quad (1)$$

Among them, C_A represents the standard community division result, C_B represents the community division result obtained by the algorithm. The row of matrix N corresponds to the standard community division result. The column of matrix N corresponds to the community division result obtained by the algorithm. The sum of the i-th row is recorded as $N_{i\cdot}$, the sum of the j-th column is denoted as $N_{\cdot j}$.

ISSN: 1813-4890

3.2. Modularity Q

The difference between the actual number of edges in the community and the expected number of edges in the community in the case of random connections. The larger the modularity value, the closer the result is to the real community structure. The formula is as follows:

$$Q = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta\left(c_i, c_j\right) \tag{2}$$

A is the adjacency matrix of the network, m is the total number of edges in the network, represents the degree of node i, and represents the community label where node i is located. If i=j, then (i,j)=1, else (i,j)=0

4. Summarize

Community discovery research has great theoretical and applied value. In the continuous development, the focus and focus of community discovery research have undergone some changes. In view of some characteristics of social network endpoint network topology and community structure in the current Internet environment, community discovery research research still faces many challenges.

The first is the overlap of communities. Traditional community discovery generally identifies a situation where a node belongs to only one community. However, in real life, a person may belong to multiple different communities at the same time, and it is the key to information transmission and social interaction. The essential. Therefore, research on overlapping communities deserves our attention.

The second point is the locality of the community. With the continuous improvement of the degree of informatization, the scale of the social network is getting larger and larger, and it becomes very difficult to obtain the global information of the network. Many calculations are required under the network data, which will cause the algorithm to be inefficient.

Finally, it is the dynamic nature of the network. The traditional community discovery studies are all static networks, which cannot effectively deal with and discover the heterogeneous network community structure containing multi-dimensional relationships. The purpose of studying the dynamics of a network is to reveal the influence of the network topology on the dynamic processes that occur on it, and whether these dynamic processes can reflect the topology characteristics of its "carrying network". Dynamic community discovery has always been a challenging problem, and only a few studies have attempted it using methods that incorporate multidimensional data.

References

[1] Qiao Shaojie, Han Nan, Zhang Kaifeng, Zou Lei, Wang Hongzhi, Louis Alberto GUTIERREZ. Detection algorithm of overlapping communities in complex network big data [J]. Journal of Software, 2017,28(03):631-647.

[2] Chen Jie, Li Rui, Zhao Shu, Zhang Yanping. A new clustering coverage algorithm for community detection based on graph representation [J]. Journal of Electronics, 2020, 48(09): 1680-1687.

[3] Yu Zhiyong, Chen Jijie, Guo Kun, Chen Yuzhong, Xu Qian. Discovery of Overlapping Communities Based on Influence and Seed Expansion [J]. Journal of Electronics, 2019, 47(01): 153-160.

[4] Li He, Yin Ying, Li Yuan, Zhao Yuhai, Wang Guoren. Large-scale dynamic network community detection based on multi-objective evolutionary clustering [J]. Computer Research and Development, 2019, 56(02): 281-292 .

[5] Yang Gui, Zheng Wenping, Wang Wenjian, Zhang Haojie. A Weighted Dense Subgraph Community Discovery Algorithm [J]. Journal of Software, 2017, 28(11): 3103-3114.

[6] Zheng Wenping, Che Chenhao, Qian Yuhua, Wang Jie, Yang Gui. A Graph Clustering Algorithm Based on Path Metrics Between Nodes [J]. Chinese Journal of Computers, 2020, 43(07): 1312-1327.

[7] Gong Weihua, Shen Song, Pei Xiaobing, Yang Xuhua. Clustering and Association Methods of Double Heterogeneous Communities in Location-Based Social Networks [J]. Chinese Journal of Computers, 2020, 43(10): 1909-1923.

[8] Zhang Baocong. Comparative Analysis of Classical Community Discovery Algorithms [D]. Shanxi University, 2017.