ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/105088, November 2016* ISSN (Online) : 0974-5645

Comparing Algorithms of Community Structure in Networks

Vikas Kumar*, Anubhav Sisodia, Umesh Maini, Pankaj and Abhineet Anand

Centre for Information Technology, University of Petroleum and Energy Studies, Dehradun - 248007, Uttrakhand, India;
vk02724@gmail.com, anubhavsisodia710@gmail.com, umeshmaini2013@gmail.com, dhullpankaj21@gmail.com, aanand@ddn.upes.ac.in

Abstract

In this work a brief overview of several existing approaches to find community structure and their comparison has been done. The best suited algorithm is selected on various grounds such as space and time complexities but sometimes they are not the only factors to decide whether the algorithm is perfect or not. So many factors are responsible for choosing the algorithm which are uncertain. So which algorithm is the best is still an open question? However various changes are suggested to the approaches to optimize them from time to time. Among various community structure, social network, ecological network and neural network are the most popular and currently it is being implemented on several social networking websites like twitter, LinkedIn, Face book, Amazon etc.

Keywords: Between's, Cliques, Community Structure, Modularity Function

1. Introduction

Community Structures can be defined as the division of network into various modules (groups, clusters, communities) which are connected to each other. The modules comprises of nodes and edges which have dense connections between the nodes within the same modules but have sparse connections between nodes in the different modules and these modules are formed using Graph Theory. In[1] the article emphasizes on finding communities in a network using different algorithms and optimizing the solution. There are number of approaches and tools available to generate Community Structure. This context proposed some set of algorithm for discovering community structure in networks like Minimum Cut Method, Girvan-Newman algorithm, Modularity maximization, Clique based method. The various algorithms are designed accordingly for different situations. The algorithms designed work on different aspects of real time system. These approaches have their application in both the sociological and biological networks. The base of the entire social network is based on the networks clustering and community detection. On the other hand the neural networks in the biological networks are also derived from clusters comprising of nodes and edges. Moreover, many algorithms for community detection also require some prior knowledge about the community structure, e.g., the number of the communities, which is very difficult to be obtained in real-world networks. The method of Greedy, Brute Force and Divide & Conquer is also applicable in many algorithms depending on the nature of the communities[2].

2. Existing approaches to Find Community Structure

To form a community analysis of network of known structure, several methods have been developed and implemented with varying success rate. Some of them which are popular and widely used are defined in this document.

* *Author for correspondence*

2.1 Max Flow Min Cut Algorithm

Max Flow and Minimum Cut Algorithm is the most primitive and the simplest approach to detect the communities in directed graphs. The approach consists of two algorithms i.e. Maximum Flow and the other is Minimum Cut algorithm.

Minimum Cut- The minimum cut approach is designed basically for the abstraction of the materials which are flowing through the edges in a particular network. However it is applicable on both set of directed or undirected graphs. In this, the network is divided into certain number of parts which are approximately of equal size. The method basically explains about a partition of the graph vertex set into two parts such that the sums of the weights of the edges connecting the two parts are minimum.

Maximum Flow- The above method which is explained works in accordance with the Maximum Flow Problem. It states that usually in a flow network we have to maximize the amount of flow from source (starting node) to the sink (last node). However the conservation of flow is followed on every intermediate node i.e. Incoming Flow=Outgoing Flow.

The Max-Flow Min-Cut method by Ford and Fulkerson in 1956 showed the combination of both maximum flow and the minimum s-t cut method. The Ford and Fulkerson method consists of following constraints-

- Flow of an edge < = Capacity of an edge
- The conservation of flow on every intermediate node is followed[3].

A BFS or DFS is followed from the starting node (source) to find out the augmenting path. An augmenting path is a path free of cycles from source to sink. The process is repeated till the maximum flow is calculated. The Ford and Fulkerson method only works on weighted graphs. The time taken to find an augmenting path is of the order of O (n) where n is the no. of edges. The time taken to complete the full algorithm is O (nf) where f is the flow in terms of integers. However in some cases it may worsen more. According to the recent studies in 2002 the Maximum Flow is used to identify the communities in the World Wide Web. The Web graph is directed but for the purpose of calculation it is considered as undirected. It is also used in the scheduling of the two processes and context switching. The algorithm can be applicable in Travel Salesman Problem. The disadvantage or flaws are not as such found in the Minimum Cut Maximum Flow approach but it gives an average solution for finding the community structure in the general networks .The complexity of the approach becomes high in many situations. Also the overlapping communities cannot be solved through this approach.

2.2 Newman-Girvan Algorithm

This algorithm iteratively removes edges from a network to split it into distinct communities. The edges are removed on the basis of between's measure assigned to them; the edges with higher between's are removed first. After removal of each edge between's are re-evaluated for the remaining edges.

The general form of Girvan-Newman algorithm is as follows:

- Calculate between's scores for all edges in the network.
- Find the edge with the highest score and remove it from the network.
- Recalculate between's for all remaining edges.
- Repeat from step 2

Between's calculation is the most important part of this algorithm. The between's measure of an edge can be defined as the total number of shortest paths between all node pairs of a network/graph passing through it[4]. We can calculate between's using three different measures: geodesic edge between's or shortest path between's, random-walk edge between's and current-flow edge between's.

Shortest Path Between's:-

Calculation of between's using Shortest Path Between's is most efficient way to calculate it. In this method breadth-first search algorithm is used to find out the shortest path between two nodes and weight age is assigned to each node. Also weight age is assigned to the edges for each shortest path between all node pairs. Finally between's for an edge is calculated as the total sum of weight age assigned to it during calculation of each shortest path. Using this method it takes O (mn) time to calculate between's for one pair and O (m^2n) or O (n^3) time for recalculations, in worst-case on a sparse graph. While the other two methods take O (n^3) time to calculate between's for one pair and O (n^4) for recalculations. Also a modified version of this algorithm has been published to reduce its complexity; it says that the leaf nodes can be removed

from the graph without calculating between's, as the weight age for an edge connected to leaf will be minimum always. This algorithm also includes "recalculation". In each cycle, first all the leaf nodes should be removed and the between's measure has to be recalculated. The general form of the algorithm is as follows:

- Remove all the leaf nodes
- Find between's measure of all edges
- Removes the edge with highest between's
- Repeat the steps from step1[4,5].

2.3 Modularity

Modularity is a mathematical function which measures the structure of networks. It is designed for the measurement of strength of division of a network into communities (also known as groups, clusters or modules). Communities with high value of modularity have dense connections within that community while communities with low value of modularity have sparse or scattered connections between nodes in different clusters. Modularity is defined as the difference of value obtained by the fraction of the edges that are within the module to the expected such fraction if edges were randomly distributed. The modularity value lies in the range [−0.5, 1) and it is represented by Q. The various techniques for modularity optimization are Greedy Techniques (hierarchical Clustering), Simulated annealing, Extremal Optimization, Spectral optimization. Newman-Girvan is the most popular Modularity function used, in which each partition of a network is divided into n disjoint modules which is denoted by Q, known as Modularity which is shown in Equation 1.

Modularity Function

$$Q = \sum_{s=1}^{n} \left[\frac{e_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right]$$

Where e_s is the number of edges in module s, d_s is the total degree of nodes in module s and m is the total number of edges in the network.[6] A good partition have Q closer to 1, which have groups with more internal connections than expected whereas bad partition have Q closer to 0, and have groups with less internal connections than we expected. Another modularity maximization approach is the Louvain Method in which communities are repeatedly optimized until global value of modularity is maximized. This method uses greedy optimization to form communities from large networks which has run time of O (n log n). This method is more efficient for identifying communities in large networks. The method has been proved successful for networks of sizes up to million nodes and billions of links. However, it has been found that modularity maximization suffers a resolution limit and fails to identify modules forsmall scale communities. This random approach assumes that each node in a network can get attached to any other node. This assumption is however not reasonable if the network is very large. For this reason, optimization of modularity in large networks would fail to form small network communities. **Studies that Use the Louvain method**: Twitter network (2.4M nodes 38M links, Twitter), LinkedIn social network (21M nodes, LinkedIn), Mobile phone networks (4M nodes, 100M links) etc.

2.4 Clique based Method

The methods which were discussed in the previous section aims at identifying those partitions in which no overlapping of subgroups between two cluster. However, in real time scenario it is often that these kinds of problems can be seen, which is also shown in Figure 1. This article is based on the methods to find out those overlapping (common) communities. One approach is to just make a duplicate copy of the overlapping nodes and share that among to their corresponding communities. It is a simplest approach to solve overlapping community, but the problem arises in finding out the nodes which are to be duplicated, to find out such nodes it is preferred to use some sort of algorithms which are also helpful in simple graph partitioning.

Another approach is to find **Maximal Cliques**; a clique is a subset of vertices in which every two vertices are adjacent to each other. Its algorithm is given.

- Find all maximal cliques where a maximal clique is a subgroup whose all nodes are connected to each other.
- Create clique overlap matrix.
- Corresponding to overlap matrix create threshold matrix at value k-1 where k is no. of nodes in a clique.
- Form communities on the basis of threshold matrix[7].

Maximum clique method is not suitable for dense graph as it makes a large matrix, but this can be efficiently used in sparse graph.

Clique Percolation Method (CPM) is another method which has been widely used to find overlapping communities. This method is given by in[7,8]. Its algorithm is given.

- K-clique is a clique with k nodes where a clique is a complete sub graph.
- Several K-cliques communities are formed from complete network
- From the network K-cliques community forms a union of all k-clique
- Particularly that type of union is formed which can be reached from each other through a series of adjacent k-cliques.
- If and only if two k-cliques are sharing k-1 nodes only than they are said to be adjacent k-cliques.

Each connected components in the clique graph form a community. The algorithm has an application in data mining, network analysis, informatics etc. and also claimed to be efficient on real systems.

Figure 1. Community or cluster of people in certain Network.

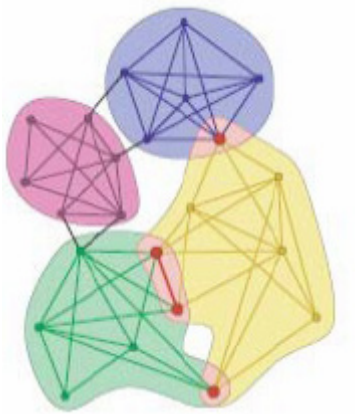

Figure 2. Overlapping of Communities is marked in Red Color.

3. Comparing Community Structure Identification

Max Flow Min Cut approach is quite different from the other algorithms is that it follows a cohesive or a divisive approach in which the graph is divided by Cuts and then solved through Max Flow technique. The algorithm can only be used for a sparse network. The complexity of Minimum Cut using Ford and Fulkerson method is less in case of integers[9]. The drawback of this approach is that the nodes which are common among two clusters fails to belong to one community, which creates overlapping therefore this problem is overcome further by the Clique based method for detecting communities. Unlike the earlier approaches, Girvan Newman algorithm provides result of reasonable quality, due to which it has been implemented in a number of standard software packages. Although it has been widely used, but it also has its own limitations as it repeatedly calculates the between's for all edges of the graph, because after removal of an edge with higher between's, between's for others also get affected[10]. Its run-time complexity increases up to $O(m^2n)$ on a sparse graph having m edges and n nodes. Because of its high time complexity, it is not beneficial to use this algorithm for a large scale network, can be implemented on the networks having a few thousand nodes or less than that. The modularity optimization has been described by Newman-Girvan which focused on having modularity as the objective function and has runtime ranging from $O(m^2n)$ to $O(n^5)$, where m is edges and n is the number of vertices in the network. This becomes very challenging for large scale social networks where there exist millions of nodes and possibly billions of edges between them and only be used for a maximum of 10,000 nodes[11]. Another greedy optimization "Louvain method" is used which is an efficient and easy-to-implement. The method has been used with success for networks for sizes up to 100 million nodes and billions of links. This algorithm has run time complexity of O(n log n) which is far better than "Newman-Girvan" modularity. In today's world, "Newman-Girvan" is the most widely used methods for detecting communities in extremely large networks[12]. In order to find overlapping community structure of complex networks, many researchers have proposed methods. Maximal clique is a graph based technique

which is used to find overlapping (common subgroup) between communities. For sparse graphs it has been very effectively used technique as it requires forming an Overlap and Threshold matrix which is suitable only for sparse graph. Moreover duplicity and Clique Percolation are the simplest among them and more often used in[13]. All methods are having some major impact under certain scenario. Some of them are very simple and effective but fails in dense graph. So, selection of algorithm can be done on the basis of situation which is best suitable for it. To evaluate a problem, several methods have been proposed but which method is best suitable is still an open question. The algorithm best suitable for a problem depends upon many factors and the above comparison can also help the readers which gives an idea to choose the appropriate algorithm for a given problem.

4. Conclusion and Closing Thoughts

In this work a brief overview of several existing approaches to find community structure and their comparison has been done. At present O (n log n) is the fastest complexity which is used to find an unknown number of communities. For analysis of extremely large network it does not guarantee that the communities found are the best possible one or not. Other algorithms which are more computationally expensive have other merits, such as accuracy or the ability to identify overlapping communities. Questions related to the best suitable method still remain open and also to search for faster and more accurate method further study is suggested so that more satisfactory results would emerge.

5. References

1. Finding local community structure in networks. Available from: https://arxiv.org/pdf/physics/0503036v1.pdf. Date Accessed: 04/03/2005.
2. Aaron C, Cristopher M, Mark N. Hierarchical structure and the prediction of missing links in networks. Nature. 2008 May; 453:98–101.
3. Maximal flow through a network. Available from: http://www.cs.yale.edu/homes/lans/readings/routing/ford-max_flow-1956.pdf. Date Accessed: 20/09/1999.
4. Community Structure based on Node Traffic in Networks. Available from: http://ieeexplore.ieee.org/document/6804502/. Date Accessed: 01/03/2014.
5. Newman MEJ. Detecting community structure in networks. The European Physical Journal B. 2004 Mar; 38(2):321–30.
6. Performance of modularity maximization in practical contexts. Available from: https://arxiv.org/pdf/0910.0165v2.pdf. Date Accessed: 1/04/2010.
7. Community detection in graphs. Available from: https://arxiv.org/pdf/0906.0612v2.pdf. Date Accessed: 25/01/2010.
8. Gergely P, Imre D, Illes F, Tamas V. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005 Jun; 435:814–8.
9. Ravindra A, James O, Robert T. Improved Time Bounds for the Maximum Flow Problem. Society for Industrial and Applied Mathematics Journal on Computing. 1989 Oct; 18(5):939–54.
10. Fast algorithm for detecting community structure in networks. Available from: http://arxiv.org/pdf/cond-mat/0309508v1.pdf . Date Accessed: 22/09/2003.
11. Finding community structure in very large networks. Available from: https://arxiv.org/pdf/cond-mat/0408187v2.pdf. Date Accessed: 30/08/2004.
12. Vincent B, Jean-Loup G, Renaud L, Etienne L. Fast unfolding of communities in large networks. arXiv:0803.0476v2. 2008 Jul; 1–12.
13. Frederic C, Karande C. A note on the problem of reporting maximal cliques. Theoretical Computer Science. 2008 Nov; 407(1-3):564–8.