Modularity Optimization for Clustering in Social Networks

Narges.Azizifard, Dr.Mehregan. Mahdavi, and Dr.Babak. Nasersharif

Abstract—By considering the increase of datasets related to social networks, we need scalable algorithms to analyze these networks. In this paper we propose an algorithm which goals in finding communities, which optimize modularity. This algorithm uses random local search agent. Experimental results show that our algorithm gives better modularity in comparison with other methods.

Keywords—Community, Modularity, Random Local Search Agent, Social Networks.

I. INTRODUCTION

Many datasets can be represented as graphs or networks that network nodes can be seen as individuals and edges represent relationships between pairs of individuals. For example, in a telecommunication network, nodes are phone numbers and edges show that two nodes communicate[1].

The reason of social networks popularity is that they are easy to use. These networks make people of all over the world able to communicate with each other[2]. One of the common feature of these networks is called community structure which represent connected groups(clusters) that there should be many edges within each group and few between the groups. Resulted groups are fraction of individuals that have similar features or connected via relations[1].

The problem of finding communities in social networks has been revealed recently and several metrices for evaluating community structure have been proposed[3], [4], [5]. Among them modularity $Q$ is the most accurate[6]. Modularity is a criterion for evaluating the quality of partitioning a network into clusters[7].

$Q$ is proposed by Newman and Girvan[3]. Suppose a particular division of network to $k$ communities, this can be represented by a $k\times k$ symmetric matrix $e$ which each element $e_{ij}$ is the fraction of all network edges that link vertices in group $i$ to group $j$. Trace matrix $Tr(e)=\Sigma_i e_{ii}$ represents a fraction of network edges that connect the vertices in a group and obviously a good division has a high value of $Tr(e)$. Although this value alone is not a good measure of the quality, because placing all vertices in a single group would give the maximal value 1 whereas no information of community structure is provided.

The row sum $a_i=\Sigma_j e_{ij}$ shows the fraction of edges that has at least one vertice in community $i$. $a_i^2$ is the expected fraction of edges within community $i$ when the edges were distributed randomly on the network. So the modularity measure can be defined as follow:

$$Q = \sum_i (e_{ii} - a_i^2)$$

Where $||x||$ is the sum of matrix $x$ elements. Values of $Q$ that are close to 1 represent better community structure. $Q$ usually falls in the range from 0.3 to 0.7[5].

The remainder of this paper is organized as follow. Section II discusses the related works. In section III proposed method is elaborated and in section IV the experimental results are shown.

II. RELATED LITERATURE

Finding communities in complex networks is revealed recently by many authors. In this section we noted some of previous methods.

A spectral clustering method for finding communities in social network is presented in [8]. In this method for completely use of network features, core members are used for mining communities, the authors utilized page rank method for community detection and proved that their method is better in terms of time and accuracy.

An accurate review of some community detection algorithms is proposed in [9] that gives the description about the algorithms and their results in detail.

One of the most popular algorithms is presented by Neman and Girvan (denoted GN) [3], [10] which is a divisive hierarchical clustering algorithm. Edge removal divides network to communities, the edges to remove are chosen by using betweenness measure. The idea is that if two groups are linked by some edges between them, then all the paths between vertices in one group to vertices in other groups include these edges. Paths give scores to edges betweenness, by accounting all the paths passing through each edge and removing the edge with maximal score, links within network are broken. This process is repeated and is divided to smaller paths until a stop criterion is reached, this criterion is modularity. A hybrid model of this approach in [11] and a faster version based on same strategy in [12] is proposed.

Approaches to community detection based on genetic algorithm are available in [13], [14],[15]. In [16] genetic method is proposed that algorithm uses a fitness function which make able to identify groups of vertices in the network.
that have dense intra connections and sparser inter connections.

In [17], [18] authors proposed a genetic algorithm that uses Newman and Girvan fitness function for measuring network modularity. An individual is included of N genes that N is the nodes number. The i-th gene corresponds to i-th node, and its value is the identifier of node i. authors use a non standard one-way crossover in which, given two individuals A and B, a community identifier j is chosen randomly, and the identifier j of nodes j1,…,jh of A is transferred to the same nodes of B.

A different approach is described in [19] which distance criterion between groups for social networks clustering in genetic algorithm is based on random walks, the representation they use is the k-medoids where each cluster center is represented by one of the nodes of the network and the number of clusters k should be known in advance. Fitness function attempts to minimize sum of all the pair-wise distance between nodes.

A random graph is produced by some random processes and the features like number of nodes and edges and links between them are identified randomly. This method in [20] is used for community detection in networks.

In data mining, community detection is a clustering problem. Members between clusters can place in one or more clusters which is called community overlapping. Identifying of this overlapping is done in [21]. Authors proposed a new algorithm for identifying the community overlapping in complex networks using fuzzy c-means clustering approach. The concept of modularity matrix for community detection is introduced in [22].

Random walks have important advantages such as they detect community structure. This approach is used in [23], the authors proposed a criterion for detecting nodes similarity based on random walks for clustering.

Extremal Optimization (EO) method is proposed in [24] for finding communities which is a divisive algorithm for graph partitioning. In this method modularity is optimized using a heuristic search based on EO algorithm. Authors produced results using real and simulated computer networks and compare with other approaches. A part of our proposed method is based on fitness function that used in EO method.

III. PROPOSED METHOD

Our approach is a combination of the method proposed in [24] based on EO and random local search agent[25]. First we discuss about these approaches then propose our method.

A. EO Method

The proposed approach in [24] is a divisive algorithm that using a heuristic search based on EO algorithm proposed by Bottcher and Percus[26], [27], optimizes modularity. This approach uses fitness function λi for partition the network. In fact Q is the global variable for optimizing that represented in equation (1) and λi is a local variable that described a node.

\[ \lambda_i = \frac{k_i(i)}{k_i} - a_i(i) \]

\( k_i(j) \) is the number of links that node i in community r has with the nodes in same community and \( k_i \) is the node i degree. \( a_i \) is the fraction of links that have one or two ends in community r to all links. The steps of this heuristic search for modularity optimization are as follow:

At first all nodes in network are divided in two random partitions that the number of nodes in each partition is equal. Each community has its own connected elements.

In each iteration, the node with low fitness is moved to another partition. Indeed for each movement the fitness of some nodes is recalculated.

The process is repeated until an optimal state with a maximum value of Q is reached, then all the links between both partitions are deleted and proceeded recursively with every resultant connected component. The process finishes when the modularity Q could not be improved[24].

B. Random Local Search Agent

In last decade different agent-based solutions were proposed to solve optimization problems. One of the successful approaches to agent-based optimization is the concept of A-Teams. An A-Team is composed of simple agents that represent complex collective behavior. The A-Team architecture first proposed by Talukdar[28] as a set of objects including multiple agents and memories which through interactions produce solutions of optimization problems. Random local search agent is used in [25] to solve distributed and non-distributed clustering problems. In fact to cope with these problems it is proposed to use a set of agents cooperating within the A-Team. A middleware environment developed by authors in [25] and referred to as JABAT (JADE-Based A-Team) is used to implement clustering problem.

As we mentioned before, in fact communities in social networks are clusters. The global process of random local search agent is as follow:

Two nodes are selected randomly that belong to different clusters(communities).

Their community ID is exchanged.

The value of fitness function is calculated and process finishes when it could not be improved.

C. Our Approach

We use fitness function in [24] for partitioning the network into communities and at end we implement random local search agent for modularity optimization. Our approach is as follow:

At first all the nodes in network have been seen as one community.

In each iteration the node with low fitness (\( A_i \)) is moved to another partition until Q could not be improved.

Then by using random local search agents two nodes are selected randomly that belong to different communities and their community ID is exchanged.

The process is repeated until an optimal state with maximum modularity Q is reached, then all links between both partitions are removed.

This process is proceeded recursively with every resultant connected component until modularity could not be improved.

The empirical results on three social network datasets showed that our method optimizes the modularity in comparison with other approaches.
IV. EVALUATION

In this section we tested our approach that has been written in MATLAB on three social network datasets, Zachary Karate Club[29], Dolphins[30], and Jazz Musicians of Gleiser and Danon[31] and compare the results with other algorithms. These networks are undirected and connected, so no transformation has been conducted.

A. Zachary Karate Club Network

This dataset describes the personal relations between members of a karate club and was created by Zachary[29], who studied the friendship of 34 members of a karate club over a period of two years and analyzed how the club divide into two new clubs after an internal conflict. Zachary could show that the personal relations where a good indicator for the prediction of which member joined which of the new founded clubs. This dataset has been used by several authors to evaluate the quality of clustering methods. It has 34 nodes and 78 edges. In table I we present the results for the maximum modularity achieved by our algorithm compared to the modularity obtained by GN[10],Newman[11], and DA[24] that shows our method gives better modularity. The results partition of our method consisted of 4 communities.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newman</td>
<td>0.381</td>
</tr>
<tr>
<td>DA</td>
<td>0.419</td>
</tr>
<tr>
<td>GN</td>
<td>0.401</td>
</tr>
<tr>
<td>Our method</td>
<td>0.424</td>
</tr>
</tbody>
</table>

B. Dolphins Network

This network of 62 bottlenose dolphins (nodes) living in Doubtful Sound, New Zealand, that represented by Lasseau from seven years of dolphins behavior. This network has 159 edges. Each node is a dolphin and each edge shows consequence relations between dolphins. Table II shows the result by our algorithm compare to the modularity obtained by GN[10] and Newman[11]. The results partition of our method consisted of 3 communities.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newman</td>
<td>0.52±/−0.03</td>
</tr>
<tr>
<td>GN</td>
<td>0.52±/−0.01</td>
</tr>
<tr>
<td>Our method</td>
<td>0.53</td>
</tr>
</tbody>
</table>

C. Jazz Musicians Network

The network of collaborations between early jazz musicians of Gleiser and Danon[31] from The Red Hat Jazz Archive has 196 nodes and 2742 edges. A link between two nodes means that they have at least one musicians in common. Table III compares the modularity results obtained by our algorithm and GN[10],Newman[11], and DA[24] that shows our algorithm gives better modularity. The results partition of our method on this network consisted of 4 communities.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newman</td>
<td>0.438</td>
</tr>
<tr>
<td>DA</td>
<td>0.445</td>
</tr>
<tr>
<td>GN</td>
<td>0.405</td>
</tr>
<tr>
<td>Our method</td>
<td>0.468</td>
</tr>
</tbody>
</table>

REFERENCES


