COMMUNITY STRUCTURE DETECTION IN NETWORKS BASED ON BEE COLONY

BILAL SAOUD

Electrical Engineering department, University of Bouira, 10000, Bouira, Algeria
E-mail: Bilal340@gmail.com

Abstract - In this paper we propose a new method to find the community structure in networks. Our method is based on bee colony and the maximization of modularity to find the community structure. We use a bee colony algorithm to find the first community structure that has a good value of modularity. To improve the community structure, that was found, we merge communities until we get a community structure that has a high value of modularity. We provide a general framework for implementing our approach. We tested our method on computer-generated and real-world networks with comparison to very known community detection methods. The obtained results show the effectiveness of our proposition.

Index Terms - Networks, Bee Colony, Modularity, Normalized mutual information.

I. INTRODUCTION

Finding community structure for networks is very important. Many studies have been done about finding community structure in network in order to give extra information about networks and to help the understanding of the network and the relationships between nodes. For example, in social networks a community structure can be defined as a group of people who have common features and/or interests. If we have information about people, we will be able to classify them in groups. However, we do not have this kind of information, but we have a network (graph) which represents the relationship between them. Networks can represent many systems shush as biology, computer science, linguistics, etc. A network has a set of vertex (or node) and a set of edge. The vertexes represent the individual's members of the system and the edges represent the links between nodes in a defined relationship of the system. For example, in a social network, a person is represented by a node, and the social interaction of relationship between two people is represented by an edge. In the following paper we consider a network as a graph $G(V;E)$, such as V is the set of nodes ($|V| = n$) and E is the set of edges ($|E| = m$). Lots of studies have done around networks and how to find the community structure among them. In generally, a network has parts that are more densely connected than other parts. In other words, the nodes in these parts are highly connected between them. These parts (nodes and edges) are called communities. Some scientists have defined a community like a group of nodes much more strongly connected to each other than other nodes. Furthermore, there are different definitions of community. In our study, we have based on the first definition of a community. The set of communities in network is called a community structure. Many community structure detection methods have been developed. Some methods can find the community structure for uniparitie and unweighted networks, others for weighted networks. And some methods find only disjoint communities (the set of intersection between communities is empty) others can find overlapping communities. Furthermore, some methods are hierarchical (merging or splitting methods); others are based on maximization of an objective function etc. Therefore, these methods can be classified according to different criteria. More information about community detection methods can be found in [1] – [2].

In this paper, we propose a new method for unweighted and undirected networks to find their community structure. Our method is based on Bee colony [3], [4]. Our approach is detailed in section II. Experimental results are shown in section III. Finally, we conclude in section IV.

II. OUR METHOD

In this section we are going to describe our contribution to find a community structure in networks. Our study focuses on networks with only a single type of vertex and a single type of undirected, unweighted edge. However, it is possible to extend our method to more complicated network types.

Our method uses the function of modularity. Modularity [5] can measure the strength of a community. The value of Modularity Q is based on the observed fraction $e(c_i)$ of edges within communities and the expected fraction $a(c_i)$ of edges within the same communities, $Q = \sum_{c_i} e(c_i)a(c_i)^2$. For an undirected and unweighted graph G, with n vertices and m edges, whose vertices are grouped in communities $(c_1, \dots, c_k)$. Let i be a vertex, d_i its degree, and c_i its community. $A_{n,n}$ is the adjacency matrix of the graph, and $P_{n,n}$ is the adjacency matrix of the corresponding null model. The detailed expression of the modularity is given as follow:

$$Q = \frac{1}{2m}\sum_i \sum_j (A[i,j] - P[i,j])\delta(c_i, c_j) \qquad (1)$$

And

$$\delta(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For an undirected and unweighted network $G(V;E)$, where $V = (v_1, v_2, \dots, v_n)$ is the set of vertices, $E = (e_1, e_2, \dots, e_m)$ is the set of edges. Our method searches to find a partition of the network G with k communities (clusters) $\pi = c_1, c_2, \dots, c_k$, where $c_i = \emptyset$, $c_i \cap c_j = \emptyset$, $(i = 1:k; j = 1:k)$ and $V = \cup_{i=1}^{k} c_i$. We try to find the community structure π that gives the highest value of modularity Q. We used a Bee colony algorithm to split the network G into k communities (clusters), where $k = \sqrt{n}$.

After we got the first community structure we merge communities in order to find the best community structure. We apply a process of merging between communities. We use a strategy based on the number of intra-community and inter-community. Two communities should merge if they had more connection (intra-community) to other communities. The function of similarity between communities S_{ij} is as follow:

$$\frac{\text{Intra} - \text{compmunity links betweenc}_i; c_j}{\sqrt{d_{c_i} d_{c_j}}} \quad (3)$$

And

$$d_{c_i} = \sum_{j=1}^{|c_i|} \text{degree}(v_j) \quad (4)$$

Algorithm 1 describes our method.

Data: $G(V;E)$
Result: $\pi = c_1, \dots, c_k$
1 $n = |V|$;
2 $k = \sqrt{n}$;
3 π = BeeColonyAlg();
4 $\Delta = 1$;
5 repeat
6 Q = modularity($G;\pi$);
7 $\pi = \pi'$;
8 Find two communities (c_i, c_j) to merge according to Eq(3);
9 $\pi' = \pi - (c_i, c_j) \cup c_{ij}$ and $c_{ij} = c_i \cup c_j$;
10 $Q' =$ modularity($G; \pi'$);
11 $\Delta = Q' - Q$;
12 until $\Delta < 0$;

Algorithm 1: The algorithm of the proposed method.

III. RESULTS

In this section, we evaluate the effectiveness and efficiency of our method. Our experiment is to demonstrate the performance of our approach to find disjoint communities on computer-generated and several real networks (Zackary's Karate Club, American College Football, Dolphins and Books about US Politics).

The results of our approach are compared with results obtained by well-known methods: Spinglass method [6], Infomap method [7], Louvain method [8], Label Propagation method [9], Fast Greedy method [5].

A. Performance measure

To compare our approach to other existing works, we used the metric of normalized mutual information [10]. This metric allows us to compare the community structure that was found by methods with the real community structure. It is based on defining a confusion matrix N, where the rows correspond to the real communities, and the columns correspond to the found communities. The element of N, N_{ij} is the number of nodes in the real community that appear in the found community j. The formula based on information theory of similarity between two partitions A and B, is:

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij} \log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_A} N_{i.} \log\left(\frac{N_i}{N}\right) + \sum_{j=1}^{c_B} N_{.j} \log\left(\frac{N_{.j}}{N}\right)} \quad (5)$$

c_A represents the number of real communities, c_B represents the number of found communities.

B. Computer-generated networks

The proposed method is tested on computer-generated networks benchmark proposed by Lancichinetti et al. [11].

The benchmark parameters are the number of nodes N; the exponents γ and β of the degree and community size distribution respectively (both distributions are power laws); the number of average degree $\langle k \rangle$; number of communities N_c; and the mixing parameter μ.

Each node shares a fraction $1 - \mu$ of its links with other nodes of its community and a fraction μ with the other nodes of the network.

Figure 1 shows the variation of the normalized mutual information obtained by the proposed method and Spinglass method [6], Infomap method [7], Louvain method [8], Label Propagation method [9] and Fast Greedy method [5] on the benchmarks networks, with the parameters: mixing parameter μ between 0.1 and 0.9, $\langle k \rangle = 16$, $\gamma = 3$, $\beta = 2$, $N = 128$ and $N_c = 4$.

The value of NMI is good for all the methods when μ changes from 0 to 0.4, which means that all the methods can find the true community structure correctly.

When μ is between 0.4 and 0.6 some methods can classify nearly all the nodes correctly. Our method is good in this part. It is difficult for all the methods to find the true community structure when μ is greater than0.6, but our method is still more accurate than the other methods. In real-world networks the community structure is not clear. From figure 1 we see that our

approach could discover community structure better than the other methods.

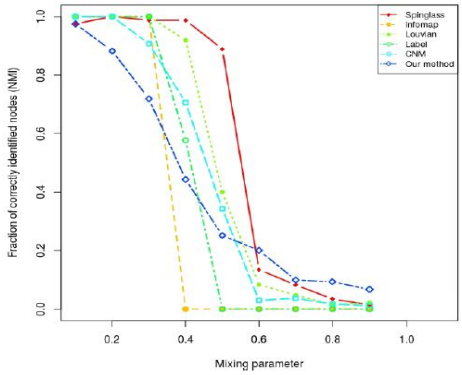


Figure 1: NMI vs. mixing parameter μ.

C. Real networks

In this section, we give the simulation results of our approach, Spinglass method [6], Infomap method [7], Louvain method [8], Label Propagation method [9] and Fast Greedy method [5] on real networks. We have chosen some real networks drawn from disparate fields (Zachary, Dolphins, Football and Books) with a know community structure. With these networks we will be able to compare the found and the real community structure for each method.

Here is some information about these real networks:

1) Zachary's club network [12] is very known network. It is a real network that represents a social network about friendships between 34 members of a karate club at a university in the United States in the 1970 ($n = 34$ and $m = 78$). The network has two clusters.
2) Dolphins Network [13] has a community structure with two communities. It is an undirected and it shows the frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand. This network has $n = 62$ nodes and $m = 159$ links.
3) College football network [14] describes the schedule of Division I Games for the year 2000 season. This network is made of 115 teams (each team is represented by a node) and 613 edges. It is community structure has 12 communities.
4) Books about US politics Network [15] is a book network about US politics that were published around 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent purchasing of books by the same buyers. Compiled by Valdis Krebs. Books network has three communities.

Table 1 gives obtained results on networks. As can be seen from Tables 1, our method can regroup the most nodes in the correct communities on Zachary's karate club, the dolphin social network, American college football and books about US politics. Figure 2 and figure 3 show the community structure that was found by our method on football network and dolphin network.

We evaluated the performance of our method with other different real networks without a known community structure. A brief description of these networks is given below.

1) Jazz network is a collaborative network [16], which represents the association between jazz musicians. The jazz musicians are represented by nodes and edge exists between nodes just if two musicians played together. The network has $n = 198$ nodes and $m = 2742$ edges.
2) Word adjacencies network represents the adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens [17]. It has $n = 112$ nodes and $m = 425$ edges.
3) Les Miserables network is co-appearance network of characters in the novel Les Miserables [18]. The network has $n = 77$ nodes and $m = 254$ edges.

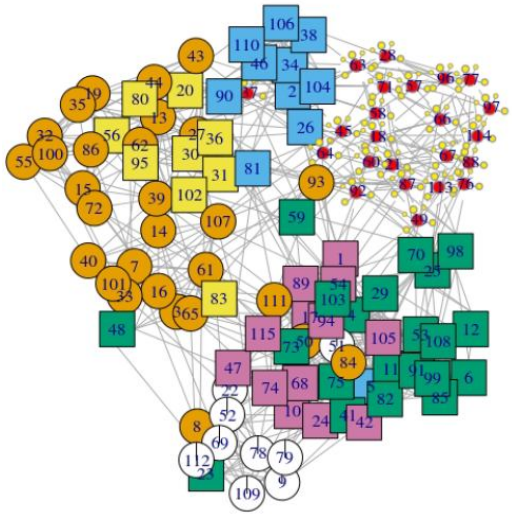


Figure 2: The football network structure is detected by the proposed method and represented by different colors.

Table.1 Performance results on real networks (M1:Spinglass, M2:Infomap, M3:Louvain, M4:Label propagation, M5:Fast greedy, M6: Our method)

Methods	Karate		Dolphins		Football		Books	
	NMI	Q	NMI	Q	NMI	Q	NMI	Q
M1	0.68	0.41	0.55	0.52	0.89	0.60	0.47	0.52
M2	0.50	0.40	0.53	0.52	0.91	0.60	0.49	0.52
M3	0.58	0.41	0.46	0.51	0.87	0.60	0.50	0.52
M4	1	0.37	0.86	0.37	0.85	0.59	0.56	0.51
M5	0.69	0.38	0.55	0.49	0.70	0.54	0.53	0.50
M6	0.68	0.41	0.54	0.52	0.68	0.48	0.54	0.49

Table.2 Performance results on real networks with unknown community structure (M1:Spinglass, M2:Infomap, M3:Louvain, M4:Label propagation, M5:Fast greedy, M6: Our method)

Methods	Jazz		Word adjacencies		Miserables	
	|c|	Q	|c|	Q	|c|	Q
M1	5	0.44	8	0.31	7	0.55
M2	7	0.28	2	0.009	9	0.54
M3	4	0.44	7	0.28	6	0.55
M4	2	0.28	1	0	6	0.54
M5	4	0.43	7	0.29	5	0.50
M6	7	0.37	10	0.23	6	0.52

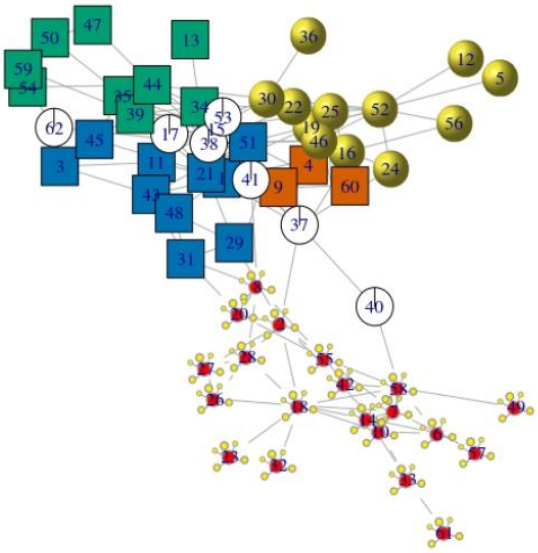


Figure 3: The dolphin network structure is detected by the proposed method and represented by different colors.

From table 2 we can see that our method finds community structures with a good value of modularity. It is difficult to compare the methods between them because we do not have a reference (a known community structure). Some methods can find community structure with a high value of modularity on some networks but on other networks they could not like infomap and label propagation method.

CONCLUSION

In this paper we have presented a new method in order to find community structure in complex networks. The method is based on Bee colony to split a network (graph) into communities. We merged communities to improve the community structure. The method can find communities for undirected and unweighted networks. Results obtained on computer generated networks and real benchmark social networks show the capacity of the approach to detect real communities. Future developments are required like testing the method on large scale networks, detect overlapping communities.

REFERENCES

[1] S. Fortunato, "Community detection in graphs", Phys Rep, 486:75-174, 2010.

[2] S. Fortunato and D. Hric, "Community detection in networks: A user guide", Phys. Rep. 659:1-44, 2016.

[3] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony (ABC) algorithm and applications", Artif Intell Rev. 42(1):21-57, 2014.

[4] D. Karaboga, and B. Akay, "A survey: algorithms simulating bee swarm intelligence", Artif Intell Rev. 31:61, 2009.

[5] A. Clauset, MEJ. Newman and C. Moore, "Finding community structure in very large networks", Phys Rev, 70: 66111, 2004.

[6] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection", Phys Rev E. 74: 016110, 2006.

[7] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure", Proceedings of the National Academy of Sciences of the United States of America, 105:1118-1123, 2008.

[8] VD. Blondel, JL. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks", J stat Mech-Theory E, P10008, 2008.

[9] U. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Phys Rev E, 76: 036106,2007.

[10] L. Danon, A. Diaz-guilera, J. Duch and A. Arenas, "Comparing community structure identification", J Stat Mech, P09008, 2005.

[11] A. Lancichinetti, S. Fortunato and F. Radicchi, "Benchmark graphs for testing community detection algorithms", Phys. Rev. E 78: 046110, 2008.

[12] WW. Zachary, "An information flow model for conflict and fission in small groups", J Anthropol Res. 33:452-473, 1977.

[13] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten and S.M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations", Behav Ecol Sociobiol. 54:396-405, 2003.

[14] M. Girvan and MEJ. Newman, "Community structure in social and biological networks", Proc Natl Acad Sci USA, 99:7821-7826, 2002.

[15] MEJ. Newman: Mark Newmans Network Data Collection, http://wwwpersonal.umich.edu/ mejn/netdata/ (2018)

[16] P.M. Gleiser and L. Danon, "Community structure in jazz", Advances in complex systems, 6(4):565-573, 2003.

[17] MEJ. Newman, "Finding community structure in networks using the eigenvectors of matrices", Phys. Rev. E 74:036104, 2006.

[18] D.E. Knuth, "The Stanford GraphBase: a platform for combinatorial computing", Addison-Wesley Reading, 1993.

★★★