Hierarchical Overlapping Community Discovery Algorithm Based on Node purity

Guoyong Cai, Ruili Wang, and Guobin Liu

Guilin University of Electronic Technology, Guilin, Guangxi, China
ccgycai@guet.edu.cn, wangruili1207.1@163.com

Abstract. A hierarchical overlapping community discovery algorithm based on node purity (OCFN-PN) is proposed in the paper. This algorithm chooses the maximal relative centrality as the initial community, which solves the problem of inconsistent results of the community discovery algorithm based on fitness resulting from randomly choosing nodes. Before optimizing and merging communities, the community overlapping degree and the joint-union should be calculated so that the problems of twice merging can be solved. Research results show that this algorithm has lower time complexity and the communities obtained by this algorithm are more suitable for real world networks.

Keywords: hierarchical overlapping community discovery algorithm; node purity; relative centrality; overlapping degree; the joint-union

1 Introduction

Community structure is a key characteristic of many networks, that is, nodes tend to aggregate into communities and nodes within communities are tightly connected while nodes between communities are loosely connected. Recently most of community discovery algorithms transform community discovery problem into hierarchical segmentation problem[1-5,7,9,10-16], which assumes that individuals of networks only belong to one community and communities of a network are isolated groups, such as the classical GN algorithm[1], Kernighan-Lin algorithm[16], and Fast Newman algorithm[2]. These isolated groups become the subgroups of larger groups until all groups become subgroups of a group, thus, hierarchical structure of the whole network is formed. However, this assumption is only suitable for some networks, for instance, the organization system network or taxonomy networks, but not suitable for most of real world networks. Many researches show that social networks not only have hierarchical structure but also have overlapping communities, in other word, a community is not the subcommunity of another community and its individuals usually belong to many different communities.

Take Facebook, one of online social network sites, as an example, every Facebook user has average 130 friends and these friends may belong to different social groups, such as high school, university, and family[6]. Marlow et al. find that groups appearing in the ego network are corresponding to acquaintances circles of different life stages[8]. An overlapping community discovery algorithm is firstly proposed by Palla et al. in 2005, and after that there are many other overlapping community discovery algorithms be proposed[10-15]. However, all of these algorithms have certain limitations, for instance, the inconsistent results problem caused by choosing node randomly and the twice merging problem. In order to deal with these problems, a new hierarchical overlapping community discovery algorithm is presented in this paper, which not only can discover hierarchical structure and high overlapping communities but also has lower time complexity.

2 Problems description

- Inconsistent results problem

Recently many methods about how to select the initial community are proposed. Assuming that every node belongs to at least one community, Lancichinetti et al. use the way of randomly choosing nodes which have assigned to none of communities as the initial community in the fitness algorithm and the ending condition is that each node belongs to at least one community[14]. Baumes et al. randomly chooses edge as the initial community[15]. Because of the difference of the initial community, the final obtained communities may be different.

- Twice merging problem

The merging process of overlapping community discovery algorithms has the repeating merging problem, for example, there are three communities C1, C2, and C3, in which C1 , C3 and C2, C3 cannot be merged, while community A is merged by C1 and C2, and then community A can be also merged with C3.

- One-sideness community discovery

Complex network researchers find that most social networks not only have hierarchical structure but also may be overlaps between communities in the same level. Many of the existing community discovery algorithms define a community as a k-clique, and use the clique filter algorithm to discover the communities, but, this method cannot discover the hierarchical structure of networks. In order to solve this problem, a fitness algorithm is proposed, which can both find the hierarchical structure of networks and the overlapping communities. However, the fitness algorithm may lead to inconsistent results problem.

3 OCFN-PN algorithm

3.1 Basic concepts

Node relative centrality equals to node absolute centrality divides the maximum possible degree of network nodes. Besides, the node having the largest relative centrality is called a core node.

Community fitness is defined by equation $f_C = \frac{k_{in}^C}{(k_{in}^C + k_{out}^C)^{\partial}}$, in which k_{in}^C and k_{out}^C are inter-degree and outer-degree separately, k_{in}^C equals to the double of the number of edges whose two endpoints are both in community C, k_{out}^{C} equals to the number of edges that only one endpoint in community C, ∂ is a adjustable parameter and its value directly controls the size of communities and the hierarchical structure of networks, moreover, the bigger the value of ∂, the less the size of the community.

Node purity is defined by equation $P = F_2 / F_1$ in which F1 and F2 are the fitness of community C1 and C1*(C1*is the new community formed after a new node A join into community C1), and $F_1 = f_{C_1}$, $F_2 = f_{C_1^*}$. If P>1, the new node A will be available, and vice versa. Thus, the value of node purity can be used to determine whether put the new node into community or not.

The existing overlapping degree calculating methods only consider the number of overlapping nodes but ignore the degree of overlapping nodes which is an important factor to judge how much the degree of overlapping communities is. For instance, in Fig.1.a and Fig.1.b the number of overlapping nodes of the two communities is both 3, but the degree of overlapping nodes of Fig.1.b is obviously bigger than that of Fig.1.a. In the existing overlapping degree calculating methods Fig.1.a is regarded as Fig.1.b, which would lead to the same overlapping degree even Fig.1.a is quite different with Fig.1.b.

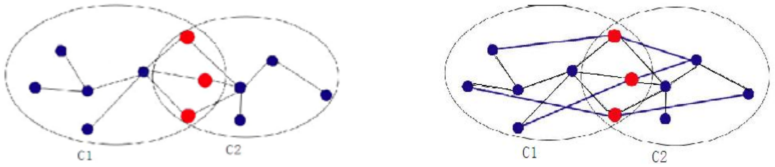

Fig. 1. a (left) and 1.b (right) two cases of overlapping communities

In order to settle this problem, a new overlapping degree calculating method is proposed in this paper. This formula contains the direct overlapping nodes, the indirect overlapping nodes, and another affecting the overlapping degree factor, which is the overlapping node degree.

Let the common nodes of community C1 and C2 be the direct overlapping and the common neighborhood nodes of community C1 and C2 be the indirect overlapping, so the overlapping degree formula can be described as

$$Doverlap(C_i, C_j) = \beta * \frac{|C_i \cap C_j| + \sum_k d(v_k^{ovlp})}{|C_i \cup C_j| + \sum_k d(v_k^{ovlp})} + (1-\beta) * \frac{|N(C_i) \cap N(C_j)|}{|N(C_i) \cup N(C_j)|} \qquad (1)$$

in which $|C_i \cap C_j|$ is the number of overlapping nodes, $\sum_k d(v_k^{ovlap})$ is the total degree of overlapping nodes, $|C_i \cup C_j|$ is the total number of nodes of community Ci and Cj, $|N(C_i) \cap N(C_j)|$ is the common neighborhood nodes of community Ci and Cj,

$|N(C_i) \cup N(C_j)|$ is the total number of neighborhood nodes of community Ci and Cj, β is an adjustable parameter whose value can control the ratio of direct overlapping. The bigger the value of β, the smaller the proportion of direct overlapping, and vice versa. Usually let the value of β be bigger than 0.5 and smaller than 0.8.

Joint-union is defined by equation $Q = \frac{|C_i \cap C_j|}{\min(C_i, C_j)}$, in which $\min(C_i, C_j)$ means the number of nodes of the community whose nodes are smaller. In this paper if Q of community Ci and Cj is bigger than K (after many test find that it is better when K is 0.7358), the twice merging problem can be solved.

3.2 Basic idea

Choosing the node with biggest relative centrality as the initial community, using node purity to determine whether put the new node into the initial community or not, and using the fitness algorithm to find the communities of networks. Because the parameter ∂ of the fitness calculating formula and one node is traveled more than one time, the hierarchical structure and the overlapping communities of the given network can be discovered. Besides there is overlap between communities so the discovered communities need to be merged. Before merging communities, the overlapping degree and the joint-union should be calculated. If the overlapping degree is bigger than K and the joint-union is bigger than γ, the two communities should be merged.

OCFN-PN algorithm.

```
Input: network graph G(V, E), in which V is the set of nodes, E is the set of edges.
Output: the set of communities{ C1,…,Cm }
For i=1 To n-1
  Calculating the degree of every node in network;
  Using the relative centrality formula calculates the relative centrality of every node;
Using bubble sort algorithm orders the relative centrality from the biggest to the smallest and stores the result in list DL;
Choosing the first data of list DL-the core node as the initial community C1 ;
```

```
6    For i=1 To the number of neighborhood nodes of C1 subtracting 1
7      Using fitness calculating formula calculates all of the fitness of neighborhood nodes of C1 and stores the results in list FL;
8      Choosing the first data of list FL as a candidate node;
9      Calculating the purity P of this candidate node;
10     If P> 1
11       Let this candidate node join into C1 and forms a new community C1*;
12       The number of community C1* pulses 1;
13     else
14       Regarding this candidate node as an unavailable node;
15   Calculating the relative centrality of the remaining nodes and repeating step 4 to step 13;
16   Until all nodes in network belong to at least one community;
17   Return {C1,…,Cm} and the number of every community
```

Optimizing and merging algorithm .

```
Input: the communities obtained by OCFN-PN algorithm
Output: the set of communities {C1,…,Cl}, (l⩽m)
1   For i=1 To m
2      Using bubble sort algorithm orders the relative centrality from the biggest to the smallest and stores the result in list NL;
3      Choosing the first data in NL;
4     Using overlapping degree calculating formula calculates the overlapping degree OD between this community and the other communities;
```

```
5        If OD> 0
6            Calculating the joint-union Q between the two communities;
7            If Q >K And OD>$\gamma$
8                Merging these two communities;
9      Choosing the second community in list NL and repeat step 4 to 8;
10   Repeat 2 to 9;
11   Until no community can merge anymore;
12   Return {C1,…,Cl}
```

3.3 Time complexity

The time complexity of OCFN-PN algorithm is O(n2) and the time complexity of optimizing and merging algorithm is O(m2) separately, so the time complexity of the hierarchical overlapping community discovery algorithm is O(n2).

However, the time complexity of this algorithm is a little bigger than that of the algorithm based on fitness, the reasons are as follow:

1. In order to avoid the problem of inconsistent results, we choose the node with the biggest relative centrality as the initial community.
2. In order to obtain better nature communities, we calculate every node’s purity to determine whether the candidate node is available or not.
3. In order to obtain more reasonable communities, we consider the number of overlapping nodes and their degree during the optimization and merge algorithm.

4 The experimental result and its analysis

The experimental data in this paper comes from Zachary's karate club, the classical data set of social network. This network is described by sociologist Zachary, who use two years to observe the members and social friendships of this club. This network is an undirected graph containing 34 nodes and 78 edges, see Fig.2.

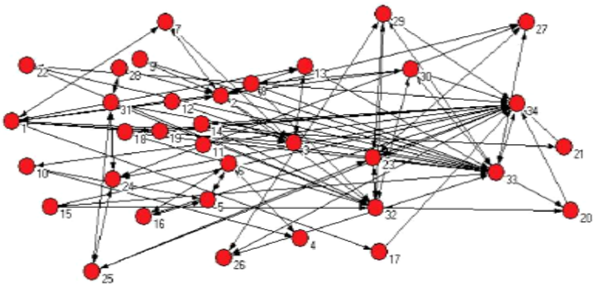

Fig. 2. Zachary's karate club network

Because Zachary's karate club itself is an overlapping community, the algorithm proposed in this paper can be used. Tab.1 describes the different community structure of Zachary's karate club obtained by this algorithm when the value of ∂ is different and Fig.3 and Fig.4 show the hierarchical structure of Zachary's karate club network when ∂=1.7.

Table 1. The communities of Zachary's karate club obtain by OCFN-PN algorithm

∂	β	γ	the communities of Zachary's karate club
0.3	0.7	0.6	{34,9,31,33,15,16,19,21,23,30,24,27,28,10,3,29,32,25,26 ,14,9,13,11,22,18,8,1,20,5,6,7,12,2,17,4}
0.7	0.7	0.6	{34,9,31,33,15,16,19,21,23,30,24,27,28,10,3,29,32,25,26 ,14},{1,2,18,22,20,8,4,14,3,13,12,10,9,31,5,6,12,7,11,17}
1	0.7	0.6	{34,9,31,33,15,16,19,21,23,9,3,30,24,27,28,10,14}, {1,2,18,22,20.31,8,4,14,3,13,12,10,9,31,5,6 }, {12,7,11,17,6,3,29,9,20,32,25,26,14}
1.3	0.7	0.6	{34,9,31,33,15,16,19,21, 9,23,30,27,24,28,10}, {1,2,18,22,20,8,4,14,3,13,12,10}{7,1,5,11,6,17,12,13,18, 22}, {26,24,25,28,32,29)
1.5	0.7	0.6	{34,9,31,33,15,16,19,21, 9,23,30,27,24,28,3,10}, {1,2,18,22,20,8,4,14,3,13,9,12,10},{7,5,11,6,17}, {26,24,25,28,32,29)
1.7	0.7	0.6	{34,23,15,19,31,30,21,31, 9,29,10,16,27,33,10}, {26,24,25,28,32,29),{1,2,13,3,4,18,8,20,9,14,13,22}, {3,14,9,20,10,8,31,29},{7, 5,11,17,6 },{12}

From Tab.1, when ∂=0.3 all nodes of the network form a community, which makes no sense to communities discovery. And when ∂=1.7 this algorithm generates six communities, one of which only contains one node. So the value of ∂ is usually bigger than 0.6 but smaller than 1.5.

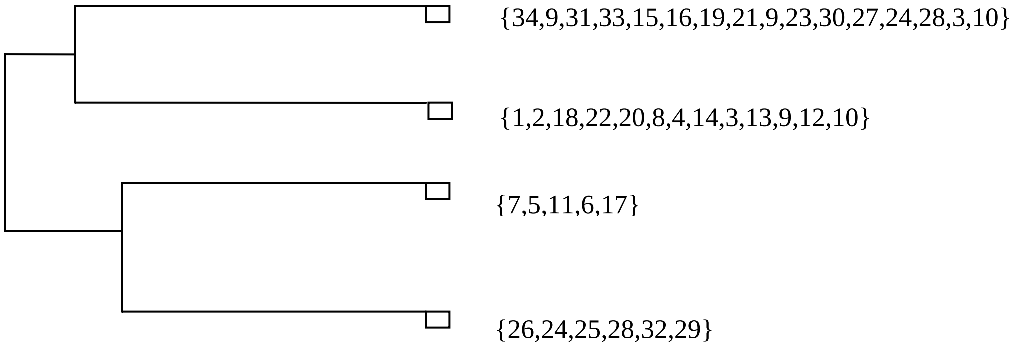

Fig. 3. the hierarchical structure of Zachary's karate club network when ∂ is 1.5.

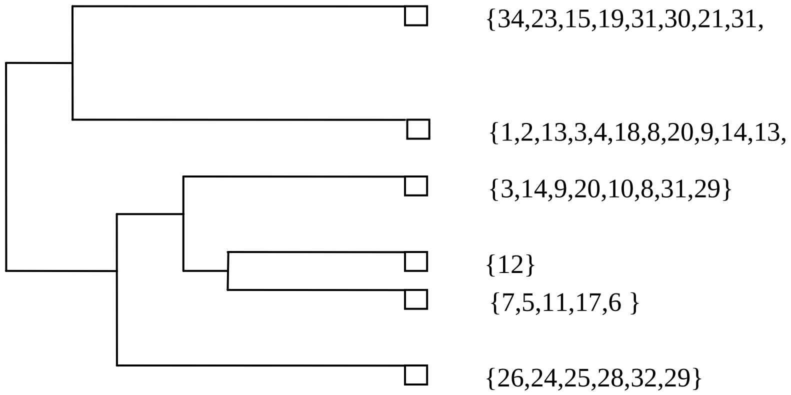

Fig. 4. the hierarchical structure of Zachary's karate club network when ∂ is 1.7.

From Tab.1, Fig.3, and Fig.4, we can conclude that different ∂, β and γ lead to different communities and different hierarchical structure. The adjustable parameter ∂ and β have be introduced above, so here we mainly introduce parameter γ. The function of parameter γ is the same as that of parameter β which is used to determine whether merge communities or not. γ =0 means the nodes of two communities are the same, which makes no sense to the optimizing and merging algorithm. γ =1 means the communities do not need to be merged.

From comparing the time complexity and the obtained communities' accuracy rate of several classical algorithms(see Tab.2), the communities obtained by OCFN-PN algorithm is better for Zachary's karate club.

Table 2. The time complexity and the obtained communities' accuracy rate when ∂ = 1.5, β =0.7, γ =0.6

Algorithm	Time complexity	Communities' accuracy rate
GN	O(n2)	97%
Kernighan-Lin	O(n2 logn)	100%
Fast Newman	O(n2)	97%
Fitness Algorithm	O(n2)	97%
OCFN-PN	O(n2)	100%

5 Conclusions.

The hierarchical overlapping community discovery algorithm based on node purity is proposed in this paper to discover the hierarchical structure and overlapping communities, which improves the fitness algorithm. This algorithm is more efficient, because it not only solves the problem of inconsistent results, twice merging and false subset merging when its time complex is the same of the fitness algorithm, but also has the higher accuracy rate when ∂ is 1.5, β is 0.7, and γ is 0.6.

Acknowledgements

This work is supported in part by Chinese National Science Foundation(Grant No. 61063039) and 2011 graduate student scientific research innovation projects in Guangxi(Project Numbers: 2011105950812M20) . We extend our gratitude to the anonymous reviewers for their insightful comments.

References

1. M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA, vol. 99: 7821-7826(2002).
2. M. E. J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E, vol. 69, 066133(2004).
3. G. Palla, I. Derényi, Farkas I and Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society, Nature 435: 814-818(2005).
4. J. M. Hofman and C. H. Wiggins, Bayesian approach to network modularity, Phys. Rev. Lett. vol. 100, 258701: 409-418(2008).
5. A. Clauset, M. E. J. Newman, and C. Moore, Finding community structure in very large networks, Physical Review E, vol. 70, 066111(2004).

6. Facebook Press Room, Facebook statistics, February. URL http://www.facebookcom/press/info.php.statistics, 2009.
7. J. Duch and A.Arenas, Community detection in complex networks using extremal optimization. Physical Review E, vol. 72, 027104(2005).
8. Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn, Maintained relationships on facebook, March 2009. URL http://overstated.net/2009/03/09/maintained-relationships-on-facebook.
9. M. Girvan and M. E. J. Newman, Community structure in social and biological networks, PNAS 99(12):7821–7826(2002).
10. A. Clauset, Finding local community structure in networks, Phys. Rev. E, 72(2): 26132-26137(2005).
11. S. Gregory, An algorithm to find overlapping community structure in networks, Lecture Notes in Computer Science, 4702: 91-99(2007).
12. S. Gregory, Finding Overlapping Communities Using Disjoint Community Detection Algorithms, In Complex Networks: CompleNet 2009, pages 47–61. Springer, May 2009.
13. Nina Mishral, Robert Schreiber, Isabelle Stanton, and Robert E, Tarjan, Clustering social networks, Lecture notes in computer science, 4863: 56-57(2007).
14. Andrea Lancichinetti, Santo Fortunato, and Janos Kertesz. Detecting the overlapping and hierarchical community structure of complex networks. New Journal of Physics 11, 033015 (2009).
15. J. Baumes, M. Goldberg, M. Krishnamoorthy, M. MagdonIsmail, and N. Preston. Finding communities by clustering a graph into overlapping subgraphs. In International Conference on Applied Computing (IADIS 2005).
16. F. Radicchi, C. Castellano. Defining and Identifying Communities in Networks [J]. PNAS, 2004, 101 (9): 2658-2663.