2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)
IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V51. 117

A Fast Method for Community Detection Based on the Contribution of Nodes

Xiangtao Chen [a], Houwang Xing[a]

[a]School of Information Science and Engineering, Hunan University, Changsha,China

Abstract. Communities exist in complex network of different areas, and there have been some methods to identify the communities in networks. However, the high computationally demanding limits their applications. In this paper, we propose a measurement of relationship between node and community, and then give a new comparative definition for community in complex networks. Based on these definitions we put forward the corresponding detecting algorithm. The effectiveness of our algorithm is demonstrated by extensive experiments on lots of computer generated graphs and public available real-world graphs. The result shows our algorithm is fast and accuracy in detecting communities.

Keywords: Comparative Definition, Community Structure, Contribution, Complex Network

1. Main text

Complex network can be used to describe various types of complex networks, such as social networks, communication networks, Web links, biological networks and many more. However, complex networks have been no strict definition precise. This network has the small-world effect and scale-free properties[1], between the regular networks and random networks. Along with the physical meaning and mathematical properties of complex networks in-depth exploration, researchers have found that many real networks have a common characteristic: community structure.

He problem of community structure partition in complex network is become a hotspot in recent years. In order to identification the communities, a lot of different algorithms are proposed, such as spectral bisection method[2], GN algorithm[3], Rapidly Newman Algorithm[4] and so on.

Generally, a community in a network is a subgraph whose nodes are densely connected within it but sparsely connected with the rest of the network [5]. There are also other community definitions based on the topology of networks, for instance strong, weak [6, 7] and LS-set [8] community definition. Hu et al [9] also given an improved community definition, and proposed a measurement the relationship of node and community with attractive force. But the measurement of attractive force just calculates the links between node and community. However, if there are the same numbers of links between nodes with two different communities, it will not identify the node belong to. We confirm that the relationship between node and community have two sides. The one is the edges between node and community, the other one is the denseness of the community. There is another question about the algorithm of Hu et al [9] is that the time complexity is $O(n^2)$, which is costly.

In this paper, we proposed a measurement of the relationship between node and community, which called contribution(node to community). Then, we given a definition of community based on contribution, which defined as: a community is a set of nodes, each node's contribution inside the community should be larger than or at least equal to its contribution to any other community. Meanwhile, we given an algorithm to detect communities, the method is efficient and adapt to large networks.

E-mail address: 101211615@qq.com. xinghouwang@126.com.

The rest of this paper is organized as follow. Section II gives our comparative definition for communities in networks. Section III presents the corresponding method in details. To verify the validity and utility of our algorithm, we run detailed experiments on a lot of public networks in Section IV. Section V conclusions and future work are summarized

2. Community Definition

2.1. Definitions and Concepts

In recent years, the study of complex networks of a major discovery is that many complex networks, including Internet, WWW, and connectivity of metabolic networks whose degree function with power law distribution. Barabasi A L and Albert [1] put forward that a lot of real-network with two traits. First, growth characteristics of community. Second, priority connection features, in other words, new nodes tend to link with those with a high degree of 'big' nodes. This phenomenon is also known as 'rich get richer' or 'Matthew effect'.

We confirm that the component units of the network are communities, therefore communities also have these characters in the process of coming into being societies. Nodes will link to denseness cluster when partitioned to a community. In addition, nodes will link to the communities which have more links with itself.

Definition of denseness community: if C is a community of G, $D(C)$ can be formulated by follows:

$$D(C)=\frac{2|e|}{|V|(|V|-1)} \tag{1}$$

Where $|e|=\sum_{i,j\in C} A_{ij}$, $|V|=\sum_{i\in C} v_i$ |.

Denseness of society $D(C)\in[0,1]$, if $D(C)$ =1 illuminated that there is a link between any two nodes in C, and the community C is forming globally coupled network.

Definition of node contribution: node i contribution to the community C can be formulated as follows:

Ensure that you return to the 'Els-body-text' style, the style that you will mainly be using for large blocks of text, when you have completed your bulleted list.

$$Con(i)=\alpha D(C)+\beta\ (\sum_{j\in C} A_{ij}/d(i)) \tag{2}$$

Where parameter α,β adjust the proportion between $D(C)$ and links ratio of node i , and should satisfy $\alpha+\beta=1$. $\sum_{j\in C} A_{ij}$ is the number of links of node i in C, and $d(i)$ is the degree of node i .

Obviously, arbitrariness community C is the subgraph of G($C\subset G$). For each node in C, the node contribution to its community is bigger than any other community.

Definition of strong Community: if $C_1,C_2,C_3,\cdots,C_k$ are k communities of G, C_k ,k=1,2,…,k should satisfy that

$$\bigcup_{k=1}^{k=m} C_k = G \tag{3}$$

and

$$\forall j\in C_k,\ \ Con(C_{kj})\geq \max\left\{Con(C_{lj}), l=1,2,.....,k\right\}$$

Definition of Weak Community: if $C_1,C_2,C_3,\cdots,C_k$ are k communities of G, C_k ,k=1,2,…,k should satisfy that

$$\bigcup_{k=1}^{k=m} C_k = G \tag{4}$$

and

$$\sum_{i\in C} Con(i)\geq \max\left\{\sum_{i\in C_t} Con(i), t=1,2,.....,k\right\}$$

Clearly, a community in a strong sense is also a community in a weak sense, whereas the converse is not true.

2.2. Example of the definition

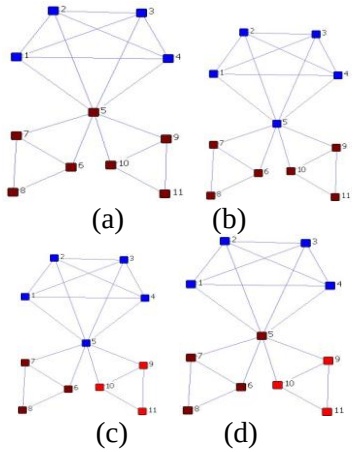


Fig.1. The formation of unconnected communities during the process of the community detection

As shown in Fig.1, we use a toy network with 3 obvious communities to show the formation of special nodes during the community detection process. As shown in Fig.1(a), there are two communities linked to node 5, and node 5 has equal community attractive force from both of these two communities. So it is impossibility to partition node 5 into a precise community according to algorithm [9, 10]. But employing the definition of contribution in this paper, this case will be avoid, as the node 5 contribution to blue cluster is 0.7 while contribution to the brown cluster is 0.46. In Fig.1(b), the node 5 moves from brown cluster to blue cluster. However, node 5 is a bridge in brown community and his departure will cause brown community unconnected. To avoid this case, it is more reasonable to regard brown community as two separated small communities as shown in Fig.1(c). In addition, why not partition node 5 to the one of two small clusters like Fig.1(d)? Through calculate the node 5 contribution to the small cluster is 0.55(<0.7), so the case is avoided.

3. Detection Algorithm

3.1. Algorithm

According to community definition [9], Hu et al. gave the definition of attractive force $F_{C,i} = \sum A_{ij}$ ($j \in C$)of community C to node i As $F_{C,i}$,the attractive force is just built upon the number of neighbors of i in community C. As referred in the algorithm proposed by Hu et al. whose time complexity is $O(n^2)$, each node will be shifted into the community or communities with the biggest attractive force, respectively. Qi Ye et al [10] extends Hu et al.'s method, improving the efficiency of the community detection algorithm. According to the definitions in section II, our method is based on the algorithm framework proposed by Qi Ye et al. [10]. First, we initially set each node into a unique cluster, then employing contribution based on self-organize process, for every node, move it into the community with the largest contribution. We now sum up our improved version ground on their algorithm is shown as follows:

Algorithm 1：The framework

Input： A graph G(V,E);
Output： A list of communities CmtyV
Declare： NIdCmtyH//store the information between node and the communities which have the node's neighbor;
Begin
// initial the beginning communities
Do while $V \neq NULL$
Select node $i \in$ V
Add node i and half of its neighbors randomly to CmtyV as a community;
Update NIdCmtyH,V;
End while
L=NumCycle;//
While L>0
For each node v∈V

Get the largest contribution of node v to the community C in CmtyV;
Moving the node v to the community C or staying in the original community;
End for
L++;
End while
End

Calculating the contribution of nodes to communities is the core of the algorithm. In the process, need to get the number of edges between node and the community, and the number of nodes. In addition, the edges in the community.

Algorithm 2：GetCon(i,c,currCon)//Get the node's contribution

Input: node i ,community c
Output: currCon
Begin
numCount = 0;
For node j in community c
If exit edge link node i and j
numCount++;
End if
End for
currCon= α *2*c(e)/[c(v)*(c(v)-1)]+
β *numCount/D(i);
//c(e) and c(v) is the number of edges and nodes in community c respectively;
end

3.2. Algorithm analysis

The time complexity of our algorithm is $O(Lnd^2/2)$. The step of initial process runs in time $O(dn)$. Deciding a node belong to the community is the core process of the method, which have to calculate the node's contribution to the adjacent communities. According to the initial process, a node has $d/2$ neighbor communities at most. So the core process runs in time $O(nd^2/2)$ and the repeated times is uncertain, where d is the average degree and n is the number of nodes in network. According to the numerical experiments in real networks, around 10 repeating steps, the division will be fixed. Now assume that the algorithm will repeat L times, so we think the time complexity is $O(Lnd^2/2)$. It is lower than many algorithms for detection community structure. The space complexity of our algorithm is $O(nd/2+e)$. As we have to store sparse network G in an adjacency-list representation, the desirable property that the amount of memory it requires is $O(n+e)$. We will have to maintain the communities, and the nodes adjacent communities, and the amount of space it requires in the hash tables is $O(nd/2)$. Comparing with the original algorithm proposed by Hu et al.[14] is faster. In addition, and the parameter α =0.4 and β =0.6 will get a better outcome.

4. Experiments

In this section, we carry out experiments on several types of network, including computer-generated networks and real-world networks to verify the effectiveness of our algorithm. The experiments are performed on a ordinary PC(CPU=Intel（R）Core（TM）2 Duo P7450 2.13GHz，RAM=2G，L2 Cache = 3072kB) running a windows XP operating system, program language C++.

4.1. Benchmarks on Artificial Networks

To test the performance of the algorithm for hard clustering we use first computer-generated graph with a known community structure which is defined as RN(C,s,d,pin) [3,11]. Where C is the number of communities, s is the number of the nodes in a community, d is the degree of node; and pin is the density of the links in inter-community. Obviously, the bigger of pin, the community structure is clearer. In this paper, we employ the general artificial network RN(4,32,16,pin). As shown in Fig.2, x-axis representation the parameter pin, y-axis representation clustering accuracy.

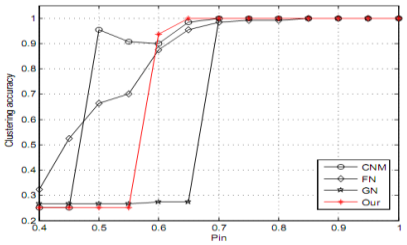


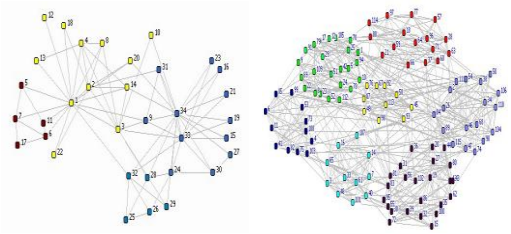
Fig.2. performances of community detection algorithms in benchmark graphs.

As shown in Fig.2, every algorithm can precise divide the communities when pin>0.7. Our method will get the accuracy communities begin pin=0.6. And the node will be in a right community while pin=0.6 and the nodes accuracy above 90%.

4.2. Typical social networks

1) Karate Club Network

The famous Karate Club Network analyzed by Zachary [13] is widely used as a test example for network clustering methods. The network consists of 34 members as nodes and 78 edges representing friendship between members of club. Owing to a conflict between the club's administrator and the instructor, the club split into two smaller groups. As shown in Fig.3(a), we get a partition of 3 communities, including 3 weak communities and 1 strong one. We can find the modularity of the communities found by our method is Q = 0.402 while modularity got by the GN partition is Q=0.401. As shown in Fig.3(a), we can find that the two communities of the administrator and the instructor are separated into 2 smaller communities, respectively.



（a）Zachary karate club（b）college football

Fig.3. The communities extracted by our algorithm in Zachary karate club and USA college football network.

2) College Football Network

American college football network [3] represents the schedule of Division I games for the 2000 season. It consists of 115 vertices and 616 edges which are the representations of football teams and regular season games among them respectively. During the 2000 season games among them are divided into 12 conference containing around 8 to 12 teams each. Games are more frequent between members of the same conference than between members of different conference. Apparently, each conference can be considered as one community of the network. As shown in Fig.3(b) , we can divide the football network into 7 communities with 5 strong communities and 6 weak communities. The modularity Q of the partition is 0.579 by our method. The GN algorithm find 10 communities whose modularity is Q=0.599, and the CNM algorithm find 5 communities whose modularity is Q=0.547.

4.3. Comparison with other methods

GN [3] is the most famous community detection methods, which based on edge betweenness. CNM [12] is a fast community detection algorithm which improved on Newman greedy algorithm. Here we compare the time consumption and the modularity of our method with that of these two methods on six real-world networks listed in Table 1. The introduction about these networks can be found from their references.

For small network, such as "karate", the time consumption of the three methods doesn't differ much. While for the network "Jazz", which has a high density, our algorithm has the largest modularity. For network "Ca-CondMat", which has a middle scale and density, when GN can not give a result and CNM needs 2694s, our method gives an excellent performance. It just needs 132s totally to finish the task. From the above results, one can see that our method is suitable for networks with both high densities and large scales.

Table1: Comparisons of the community detection algorithm GN, CNM and ours.

Networks	ref	Vertices	Edges	GN		CNM		ours	
				Q	t(s)	Q	t(s)	Q	T(s)
Karate Club	[13]	34	78	0.401	0.91	0.381	**0**	**0.402**	0.02
Football	[5]	115	613	**0.599**	77	0.547	**0.08**	0.588	0.27
Jazz	[14]	198	2742	0.405	238	0.439	**0.28**	**0.443**	0.73
E-mail	[15]	1133	5451	**0.532**	6713	0.511	**2.72**	0.512	3.72
Ca-GrQc	[16]	5242	28980	**0.849**	57394	0.816	18.19	0.790	**10.08**
Ca-CondMat	[17]	23133	93468	-	-	0.650	2694	0.617	**132**

5. Discussion And Future Work

In this paper, we focus on the problem of detecting non-overlapping communities in network. Aim at the relationship between node and communities, we proposed the definition of the node contribution, then given the definition of the community and the corresponding detection community algorithm. Our method is fast and easy to implement. To evaluate the effectiveness or our algorithm, we give lots of experiments. The result shows our algorithm is very efficient and the time complexity is $O(Lnd^2/2)$, and when d<<n, the time complexity of algorithm is linear. In future work, we will research how to get a fast overlapping community algorithm to get soft partitions for massive graphs based on this algorithm.

6. References

[1] Barabasi A L,Albert R.Emergence of scaling in random networks[J]. Science, 1999, 286:509-512

[2] Capocci A,Servedio V D P,Caldarelli G,et a1.Detecting communities in large networks[J]. Physica A, 2005, 352:669—676

[3] Newman M E J,Girvan M.Finding and evaluating community structure in networks[J].J,Phys Rev E, 2004,69:026113

[4] Newman M E J. Fast algorithm for detecting community structure in networks[J].Phys Rev E,2004,69:066133

[5] Newman M E J．Detecting community structure in networks[J]．Eur Phys J B，2004,38：321—330

[6] Radicchi F,Castellano C,Cecconi F,et a1.Defining and identifying communities in networks[J]. Proc Natl Acad Sci USA,2004,101:2658-2663

[7] Castellano C,Radicchi F.Self-contained algorithms to detect communities in networks[J].Eur Phys J B, 2004, 38:311—319．

[8] Wasserman S,Faust K.SociM Network Analysis[M],Cambridge UK:Cambridge Univ Press, 1994.

[9] Y. Hu, H. Chen, and et al., “Comparative definition of community and corresponding identifying algorithm,” Phys. Rev. E, vol. 78, p. 026121,2008.

[10] Qi Ye, Bin Wu et al. Detecting Communities in Massive Networks based on Local Cmommunity Attractive Force Optimization, Proc. ASONAM 2010:291-295

[11] Girvan M，Newman MEJ．Community structure in social and biological networks．Proc．of the National Academy of Science.2002，9(12)：7821-7826．

[12] Clauset A,Newman M E J,Moore C. Finding community structure in very large networks[J], Phys Rev E, 2004,70(6): 066111.

[13] W. W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33, pp. 452-473 1977

[14] P.Gleiser and L. Danon , Adv. Complex Syst.6, 565 (2003).

[15] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt and A. Arenas, Physical Review E , vol. 68, 065103(R), (2003).

[16] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

[17] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.