Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 985-990
doi:10.4028/www.scientific.net/AEF.6-7.985

A Local Overlapping Community Detection Method in Complex Networks

Yan Peng[1,a], Yanmin Li[1,b], Lan Huang [1,c]
Longju Wu[1], Guishen Wang[1] and Chao Zhang[1]

[1]College of Computer Science and Technology, Jilin University, Changchun, 130012, China

[a]pengyan10@mails.jlu.edu.cn, [b]yanmin10@mails.jlu.edu.cn, [c]huanglan@jlu.edu.cn

Keywords: overlapping community detection, community center, local metric.

Abstract. Community structure detection has great importance in finding the relationships of elements in complex networks. This paper presents a method of simultaneously taking into account the weak community structure definition and community subgraph density, based on the greedy strategy for community expansion. The results are compared with several previous methods on artificial networks and real world networks. And experimental results verify the feasibility and effectiveness of our approach.

Introduction

In recent years, community structure detection has become one of the most important research fields in the complex network analysis. In real world networks, there will always be some overlapping nodes not only belonging to one community. Therefore, researchers have proposed a lot of approaches for finding overlapping communities. For large-scale complex networks, we usually do not know the ground truth of communities. In addition, if we consider detecting communities in global views, it may lead to high complexity. Thus, local method for community detection has become an important idea. Papers [1-5] present methods for discovering communities from the view of locality. But some of these methods have to specify the parameters unpredictable, and some cannot find the outliers in networks even if they have good results. Our paper presents a new local metric that combines the weak community structure definition with community subgraph density. The method can effectively exclude outliers in networks.

Local Community Detection

The basic idea of local community detection is first to select a community center as the current community, and then to add or remove nodes from the community according to certain strategy.

The Selection of Community Center. The main several methods of selecting centers are as follows.

Random Node [1]. Take a random node from the network graph that has not yet been assigned to any community as a center. However, if each time select a different node, the results are probably different, which is unstable.

Rank Removal (RaRe) [2]. Nodes are first ranked by some measure of importance, for example, PapeRank. Highly ranked nodes are then removed in groups until small connected components are formed.

Link Aggregate(LA) [3]. Nodes are first ranked, typically using PageRank. Then a node is added to any cluster if adding it improves the value of its metric. If the node is not added to any cluster, it creates a new cluster.

Clique Percolation [4]. That is, using the results of CPM algorithm as the centers.

Maximal Cliques [5]. First finding all maximal cliques in the network graph. Then use clique coverage heuristic to remove those high coverage cliques. That is, centers are a set of the maximal cliques which are not near-duplicate.

In many complex networks, especially social networks, the elements in the center of a community are easy to form clique structures. Lee et al. [6] have demonstrated the higher efficiency and accuracy of maximal cliques than other methods. So our method also use maximal cliques as community centers.

The Selection of Local Metric. The typical two local metric are as follows.

According to the characteristic of a community in a network, the internal edges are more than the external edges. Paper [2] used the ratio of the internal edges against total edges of a community as the local metric,

$$F = \frac{E_{in}}{E_{in} + E_{out}}, \tag{1}$$

where E_{in} and E_{out} respectively means the internal and external edges of a community.

According to the weak community structure definition, the internal degree is greater than the external degree. Paper [1,5] used the ratio of the internal degree against the total degree of a community as the local metric.

$$F = \frac{K_{in}}{K_{in} + K_{out}}, \tag{2}$$

where K_{in} and K_{out} respectively means the internal and external degree of a community.

In these metrics, Eq.2. presents better characteristic. However, its results usually include many outliers. Although the author added an index α to adjust the size of communities, it was not intended to exclude outliers. Fig.1 illustrates this problem. We have a local community C, nodes n_1,…, n_6, which may be outliers since they are barely connected to C. Let us assume that all nodes in A, A means the set of adjacent nodes of C, n_1 maximizes the Eq.3, then n_1 to n_4 will be added into C, one by one. The reason is that every addition of that kind of nodes do not affect the external degree number but will increase the internal degree number by two. However, such addition would be inaccurate since nodes on the chain, especially those in the chain back, are distant and can be presumed to be outliers. So it would be better if the local metric can handle sparse chains of nodes for some networks. We propose a new local metric that can solve this problem.

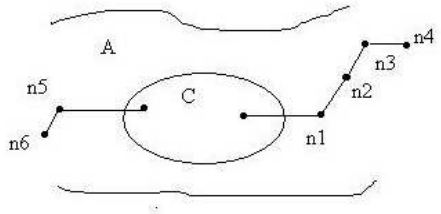

Fig.1. A local community with outlier chains

Our function simultaneously takes into account the weak community structure definition and community subgraph density. In a subgraph, the more the links are, the greater the density is. So we define the community subgraph density as the ratio of the number of internal edges of the subgraph and the number of edges of the corresponding complete graph. The density function is as follows,

$$d(C) = \frac{E_{in}}{\frac{|C| \times (|C| - 1)}{2}} = \frac{K_{in}}{|C| \times (|C| - 1)}. \tag{3}$$

Our metric is proposed as,

$$Q = (1 - \lambda) \times \frac{K_{in}}{K_{in} + K_{out}} + \lambda \times \frac{K_{in}}{|C| \times (|C| - 1)}, \tag{4}$$

$\lambda (0 \le \lambda \le 1)$is a parameter allowing the results to be fine-tuned. Setting $\lambda = 0$ will produce the same results as Eq.2, while larger values will make the community density more significant than before. This also can produce smaller groups for larger values of λ which allows communities to be produced across a variety of resolutions. In fact, the value of λ is better in near 0.1, thus, we can exclude those chains of outliers effectively, and can avoid too constricted on forming community.

Local Community Detection Method based on a Metric Combining Weak Community Structure and Community Density: WDLCD

The specific steps are as follows.

(1) Find all maximal cliques in network graph G with at least k nodes (k=3 or 4).

We preprocess these cliques as does in paper [5]. First, we order the maximal cliques, largest first. Then a clique will be discarded if at least Φ percent of its nodes have already been covered twice by other larger, accepted cliques. Φ is called the clique coverage heuristic,

$$\phi = \frac{|C' \cap C_i|}{|C'|}, \tag{5}$$

where C' means the clique in question and |C'| means the number of nodes in it. C_i means certain clique that has been accepted. $|C' \cap C_i|$ means the number of the same nodes of the two cliques. We also choose a value of 0.75 for Φ.

(2) Choose the largest unexpanded clique as a community center (ComCenter).

The candidate nodes set composes of all the adjacent nodes of the current community, denote A.

Algorithm: Local Community Detection
Input: A clique and a network G.
Output: A local community.

1. Add $n_i \in$ ComCenter to C.
   Compute Q.
2. for each $n_i \in A$ do
   Compute Q' when n_i is included;
   end for
3. Select n_i with the maximum Q'.
4. if Q' > Q
   Add n_i to C;
   Q = Q';
   end if
5. for each $n_i \in C$ && $n_i \notin$ ComCenter
   Compute Q' when n_i is excluded;
   if Q' > Q
   Remove n_i from C;
   Q = Q';
   end if
   end for
6. if 4 occurs, update A, repeat from 2; otherwise, end.

(3) Dealing with high-overlapped communities.

The overlap heuristic about two communities is ε [5], similar to Φ,

$$\varepsilon = \frac{|C' \cap C_i|}{\min(|C'|, |C_i|)}, \tag{6}$$

where C' may be a clique or a community we just get. If the overlap heuristic that C' with any already accepted community C_i is no smaller than ε, then C' and C_i are high-overlapped communities, so discard C'. Otherwise, accept C'. We set a value of 0.6 for ε. We realize that before we expand a selected clique, if the overlap heuristic of it with certain C_i accord with the situation above, the expansion can be omitted.

(4) Repeat from (2), until no cliques remain.

Experiment Results

Since the ground truth of communities in many large networks are hard to define. So we apply our algorithm on artificial networks and real networks. All networks are undirected and unweighted. For artificial networks, their real community structures are known, so we can measure the accuracy of our results by comparing the ground truth of communities. NMI (Normalized Mutual Information) is an evaluation index like this. And Lancichinetti et al. [1] modified it to be able to handle overlapping communities. Therefore, we use NMI as the evaluation criterion on artificial networks. For real world networks, we use function EQ [7] to measure the results, ignoring of its limitations. We use other algorithms as comparing, LFM [1], which also uses local greedy optimization strategy, where we set α to 1.1 in order to make a suitable comparison; CPM [8], which has been described above; COPRA [9], which utilizes a label propagation technique, where we set v to 3; CONGO [10], which uses node split method, can be applied to the situation that the number of communities are known, where we set h to 2. And all implementations of them we used were from the authors.

Experiments on Artificial Networks. *LFR Overlapping Benchmark*. It generates networks have no outliers, however, we still can verify our algorithm's effectiveness than several classical overlapping community detection algorithms. We set the number of nodes to 1000, the average nodes degree of the network to 20, the maximum nodes degree to 50, the minimum and maximum size of a community to 10 and 50, the index of the power-law distribution of nodes degree and community size to 2 and 1, mixing parameter(mp) to 0.4~0.9. The greater value of mp means the community structures of the network are less obvious. The results are shown in Table 2.

Table 2. NMI of algorithms on LFR benchmarks

NMI Al / mp	WDLCD(k=4)	LFM	CPM(k=4)	COPRA	CONGO
0.4	0.774359	0.682941	0.828691	0.766439	0.601413
0.5	0.744079	0.474787	0.595566	0.713293	0.401882
0.6	0.534451	0.201068	0.350144	0	0.254304
0.7	0.351203	0.037085	0.250781	0	0.15737
0.8	0.1003	0	0.102625	0	0.015425

From the above artificial datasets, we can demonstrate WDLCD performs competitively against several known overlapping community detection algorithms.

Experiments on Real World Networks. *Yeast* [11]. Protein-protein interaction networks in yeast, communities are likely to group proteins having the same specific function within the cell. The dataset after processing has 2284 nodes and 6646 edges. The results are shown in Table 3.

Table 3. EQ of algorithms on Yeast data

Algorithm	EQ
WDLCD (k=3)	0.1638
CPM (k=3)	0.1311
LFM	0.1707
COPRA	0.0215

NetScience [12]. Coauthorship network of scientists working on network theory and experiment. The dataset after processing has 1461 nodes and 2742 edges. The results are shown in Table 4.

Table 4. EQ of algorithms on NetScience data

Algorithm	EQ
WDLCD (k=4)	0.3461
CPM (k=4)	0.2957
COPRA	0.3434
LFM	0.1120

In the two experiments, WDLCD also performs better than other algorithms, ignoring LFM acquire a better EQ in yeast.

Conclusion

We introduced a new local metric for detecting overlapping communities which combines the weak community structure definition with the density of a community, can solve the problem of outlier chains in networks. We demonstrate that our algorithm WDLCD is better than several classical algorithms in artificial networks and real world networks.

Although our algorithm can get better results than several classical algorithms, its effectiveness and efficiency are poorer than the excellent local overlapping community detection algorithm GCE [5]. So further work should try to continue improving the local metric and try to make the algorithm parallel. On the other hand, the modularity function used to evaluate the results bases on the theory that when the number of links within the community are much more than the number of links between the community, the better of the community structure of the network. Owning to the exclusion of outlier chains, we need to propose a more effective function to measure the results. Furthermore, we also need to find better and larger real networks to experiment.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61175023, National High-Tech Program (2009AA02Z307), project of science and technology innovation platform of computing and software science (985 engineering), Project Research of Clustering Algorithm Aim at the Data with Multidimensional Mixed Attributes under Grant No. 20121102 Supported by Graduate Innovation Fund of Jilin University, Fundamental Research Funds for the Central Universities (Grant No. 201103195) and the Key Laboratory for Symbolic Computation and Knowledge Engineering, Ministry of Education, China.

References

[1] A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the Overlapping and Hierarchical Community Structure in Complex Networks, New J. Phys. 11, 033015 (2009).

[2] J. Baumes, M. Goldberg, M. Krishnamoorthy, Finding Communities by Clustering a Graph into Overlapping Subgraphs, IADIS International Conference on Applied Computing 2005.

[3] J. Baumes, M. Goldberg, M.M. Ismail, Efficient Identification of Overlapping Communities, Intelligence and Security Informatics, LNCS 3495 (2005).

[4] M.S. Shang, D.B. Chen, T. Zhou, Detecting Overlapping Communities based on Community Cores in Complex Networks, Chin. Phys. Lett. Vol.27, No. 5 (2010) 058901.

[5] C. Lee, F. Reid, A. McDaid and N. Hurley, Detecting Highly Overlapping Community Structure by Greedy Clique Expansion, The 4th SNA-KDD Workshop'10 (2010).

[6] C. Lee, F. Reid, A. McDaid and N. Hurley, Seeding for Pervasively Overlapping Communities, Phys. Rev. E 83, 066107 (2011).

[7] H.W. Shen, X.Q. Cheng, K.Cai, Detecting Overlapping and Hierarchical Community Structure in Networks, Phys. A: Statistical Mechanics and its Applications, vol. 388, Issue 8 (2009), pp 1706-1712.

[8] G. Palla, I. Derenyi, I. Farkas and I. Vicsek, Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, Nature 435, 814-819 (2005).

[9] S. Gregory, Finding Overlapping Communities in Networks by Label Propagation, New J. Phys. 12, 103018 (2010).

[10] S. Gregory, A Fast Algorithm to Find Overlapping Communities in Networks, Knowledge Discovery in Databases (PKDD 2008), LNCS 5211, pp 408-423 (2008).

[11] D.B. Bu, Y. Zhao, L. Cai, Topological Structure Analysis of the Protein-Protein Interaction Network in Budding Yeast, Nucleic Acids Research, 2003, Vol.31, No.9, 2443-2450.

[12] M.E.J. Newman, Finding Community Structure in Networks Using the Eigenvectors of Matrices, Phys. Rev. E 74, 036104 (2006).