Community detection with consideration of non-topological information*

Zou Sheng-Rong(邹盛荣)[a], Peng Yu-Jing(彭昱静)[a], Liu Ai-Fen(刘爱芬)[b], Xu Xiu-Lian(徐秀莲)[b], and He Da-Ren(何大韧)[b]†

[a] College of Information Engineering, Yangzhou University, Yangzhou 225009, China
[b] College of Physics Science and Technology, Yangzhou University, Yangzhou 225009, China

(Received 7 June 2010; revised manuscript received 21 July 2010)

In a network described by a graph, only topological structure information is considered to determine how the nodes are connected by edges. Non-topological information denotes that which cannot be determined directly from topological information. This paper shows, by a simple example where scientists in three research groups and one external group form four communities, that in some real world networks non-topological information (in this example, the research group affiliation) dominates community division. If the information has some influence on the network topological structure, the question arises as to how to find a suitable algorithm to identify the communities based only on the network topology. We show that weighted Newman algorithm may be the best choice for this example. We believe that this idea is general for real-world complex networks.

Keywords: community division, algorithm, topological structure, weighted network

PACS: 89.75.Hc, 89.75.Fb **DOI:** 10.1088/1674-1056/20/1/018902

1. Introduction

It has become well known in the last decade that descriptions in terms of complex networks are usually suitable for complex systems and that the networks usually show three basic characteristics: small-world, scale-free, and community structure.[1,2] Detecting communities is of great importance in many scientific and technological areas, such as sociology, biology, computer science, and so on. Many community detecting algorithms have already been published, [3] the most influential of them by Newman.[3−7] He adopts a traditional qualitative definition of community, i.e., 'the division of network nodes into groups within which the network connections are dense, but between which they are sparser.'[4] His famous community partitioning algorithms are based on a widely accepted quantity, 'modularity'. Newman and Girvan proposed its definition as[4]

$$Q = \sum_i \left[e_{ij} - \left(\sum_j e_{ij} \right)^2 \right] = \mathrm{Tr}\boldsymbol{E} - \parallel \boldsymbol{E}^2 \parallel, \quad (1)$$

where e_{ij}, the element of matrix $\boldsymbol{E}$, denotes the fraction of edges between the divided communities i and j, and Q denotes the ratio of the edge number inside the communities to the total edge number. In the so-called Clauset–Newman–Moore (CNM) algorithm,[7] in every division step (e.g., joining two possible communities by an edge), the authors suggested calculating the 'change' of the modularity. The best division of communities corresponds to the largest modularity change.[7] This community partitioning algorithm maps a network onto a dendrogram. However, Palla *et al.* noticed that, in many real-world networks, communities strongly overlap each other, and therefore mapping onto a dendrogram is sometimes impossible. They defined a so-called 'l-clique community' as 'a union of all l-cliques (complete subgraphs of size l) that could be reached from each other through a series of adjacent l-cliques (where adjacency means sharing l-1 nodes) and proposed a corresponding 'C-Finder' community partitioning algorithm.[8] Using this algorithm, one can identify all the l-clique communities by inputting the number l. These two algorithms, CNM and C-Finder, are famous and representative. In addition to these, we will mention another community partitioning algorithm, which was proposed by our research group.[9] The so-called 'Chang Hui (CH) algorithm'[9] is applicable only in affiliation bipartite networks.[10] Such a graph contains two types of nodes. One is called 'acts', which can be organiza-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 70671089 and 10635040).
†Corresponding author. E-mail: darendo10@yahoo.com.cn
© 2011 Chinese Physical Society and IOP Publishing Ltd

tions, events, or activities, and the other is named 'actors', which are participants. Some acts include common actors and are described as 'overlapping'. The authors of Ref. [9] define the 'hierarchy level', h, as the least overlapping actor numbers of the acts. An h-hierarchy level community is defined as the union of the overlapping acts, which share at least h common actors. In higher hierarchy levels the communities are composed of stronger overlapping acts. Using the CH algorithm one can distinguish all the hierarchy level communities directly based on the empirical data.

In all the three mentioned algorithms, and many others, the communities are identified only according to the topological structure information specifying which nodes are connected by edges.[3] However, often the community identification depends on non-topological information, which is not directly included in topological structures. For example, the fellows in a research group usually communicate very often in their common workplace and in many other instances. They form a dense idea communication community; however, often they do not publish papers together (e.g., each of them only publishes papers with his/her common supervisor). If we set up a co-authorship bipartite network in an ordinary way by defining the authors as actors and an edge between two authors as the co-authorship of a paper, the communities identified only by topological information are often very different from the research groups. Some group members may publish papers with some scientists outside the group. These group members may belong to another denser part of the network if their out-group papers contain more authors and thus induce more edges. So, one can understand that in a traditional co-authorship bipartite network the topological community identification is sometimes strongly influenced by the author number in papers instead of being fully determined by scientific communication relationships. This is inappropriate if we want to show scientific cooperation by communities. A question arises: is it possible to correctly identify scientific research communities only by topological information (supposing that the research group affiliation is unknown)? Our answer is: although one cannot directly know which authors belong to a common research group only by their co-authorships (the topological structure), we may identify the group communities because the group affiliation strongly influences the topological structure. Often the group members only publish papers with their common supervisor; this makes each group resemble a star-like subgraph. There can be an optimum way to identify such communities.

To the best of our knowledge, there are only a few reported investigations relating to this topic. One example is that community identification in the networks of committee and subcommittee assignments in the United States House of Representatives[11] or in the networks between Members of the United States Congress[12] should depend on some political factors, which are not shown in topological structures. The authors of Refs. [11] and [12] successfully determined the 'hidden' political information by network topological analyses.

In this paper, we will discuss the question using the aforementioned example. The rest of the article is as follows. In Section 2 we build a representative co-authorship network. The research group affiliation information is actually known but the community identification will be made based only on the topological structure. Then, in Section 3, we report the topological community identification results of the network using three partition algorithms introduced previously and compare the results with the research group affiliation communities. In Section 4, the community identification is be performed using a weighted topological method proposed by Newman[13] with three different weight definitions. In Section 5, a summary will be presented.

2. A representative co-authorship network

We collected all the published co-author information of a well known scientist, A-L Barabasi, between 1999 and 2006 and built a co-authorship affiliation bipartite network.[14] The authors are defined as nodes. Two nodes are connected by an edge if they are authors of at least one common paper. The topological structure denotes the information about which authors (nodes) are connected by edges (co-authorships). The author research group affiliation information is non-topological. In order to clearly show the author research group affiliation information, a few authors, who were affiliated to more than two institutions in at least one paper, were eliminated. There are 75 authors in the network who were co-authors of 52 published papers. The small data number permits us to clearly illustrate the network with community structure. As can be seen in Fig. 1, the authors mainly belong to

three larger research groups directed by Barabasi, Oltvai and Vicsek, respectively. In addition, there are authors who belong to some other smaller groups. In Fig. 1 the authors in the Barabasi group are denoted by solid circles, those in the Oltvai group by solid diamonds, those in the Vicsek group solid triangles, and those in other smaller groups by solid squares. If one agrees that each research group should be identified as a scientific community, this network should be clearly partitioned to three communities. The rest of the authors may be considered as forming a 'remanent' community. One sees in Fig. 1 that, by the network definition, the densest subgraph at the lower-right only represents a paper with a very large author number.

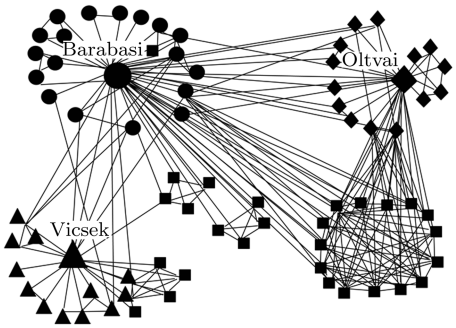


Fig. 1. The representative co-authorship network showing three scientific communities and a remanent community.

3. The topological community identification results

We performed the topological community identification of the network shown in Fig. 1 using the three partition algorithms introduced in Section 1. Figure 2 shows the partitioning result using the CNM algorithm.

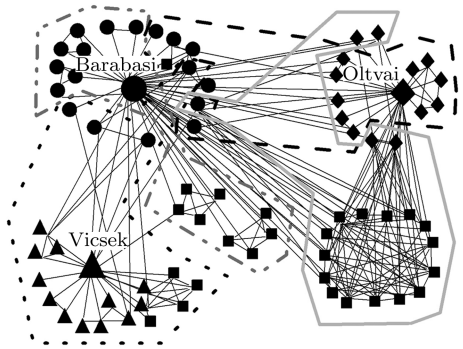


Fig. 2. Community partitioning result using the CNM algorithm.

In Fig. 2, the three research group communities and the remanent community nodes are shown by the same solid symbols as mentioned in the last section; each CNM algorithm partitioning community is shown by a wider surrounding line. These are shown as solid, dashed, dotted and dash-dotted lines. One sees that the best CNM algorithm community division (corresponding to the largest change of the modularity), gives rise to four communities. Although there are some 'mistakes', the division basically shows the three research group communities and the remanent community.

Figure 3 shows the partitioning result using the C-Finder algorithm. With all the possible l numbers, we can only find three identified communities, which are indicated by the wider solid surrounding lines in Fig. 3 and show the best possible consistency with the three research group communities although the consistency is still poor. The C-Finder algorithm partitioning communities are too small. Also, Professor Vicsek belongs to two communities using this division due to the different l number selection.

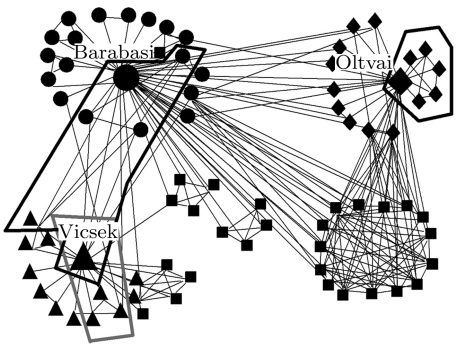


Fig. 3. Community partitioning result using the C-Finder algorithm.

Figure 4 shows the partitioning result using the CH algorithm. Again, each CH algorithm partitioning community is shown by a wider solid surrounding line. One sees that, with all the possible h numbers, we can only find two communities, which show the 'best' consistency with the three research group communities. The consistency may be even worse. Many authors belong to both the communities using this division due to the different h number selection.

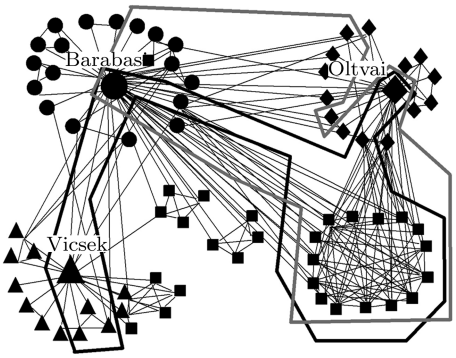


Fig. 4. Community partitioning result using the CH algorithm.

In order to quantitatively compare the three partitioning algorithm results, we define a 'consistency coefficient' as

$$c = 2i/(a+b), \tag{2}$$

where i denotes the node numbers in an algorithm partitioning community, which are superpositioned with the nodes in a research group community, a denotes the total node number in the algorithm partitioning community, and b denotes the total node number in the research group community. If all the nodes are superpositioned, $c = 1$. If there are no superpositioned nodes, $c = 0$. Table 1 gives the consistency coefficient results.

It is clear that the CNM algorithm gives the best results as can be seen in the figures. The C-Finder results are much worse, and the CH algorithm gives the worst results.

Table 1. Consistency coefficient results.

algorithm	edge weight	Barabasi group	Vicsek group	Oltvai group
CH	1	0.28	0.23	0.40
C-Finder	1	0.34	0.35	0.57
CNM	1	0.50	0.77	0.65
W–N	$k_i \times k_j$	0.52	0.41	0.64
W–N	$\max(k_i,k_j)$	0.67	0.79	0.79
W–N	k_i/k_j	0.67	0.79	0.84

4. Community identification by weighted Newman algorithm

One may ask why the CNM algorithm performs much better than the C-Finder and CH algorithms. Is it absolutely better than the other two? The answer may be that the CNM algorithm is much better only for this example. In this case, in each research group an author mainly connects only with the common supervisor, therefore the supervisor shows a much larger degree, which is defined as its total neighbouring edge number, than the other ordinary group members. The 'correct' communities partitioned by non-topological information (the research group affiliation) are highly centralized (near to star-subgraphs) and far from complete subgraphs. According to Section 1, the C-Finder and CH algorithms show an advantage only for communities formed by strongly overlapping complete subgraphs. This is why they perform badly here.

This discussion provides an idea with which we may even improve the consistency coefficient results. Since the key point in the relatively successful identification of research group communities by the CNM algorithm is the large degree differences between the supervisor and the ordinary group members in each research group, even better consistency coefficient results should be obtained if the 'position differences' between the supervisor and other members can be magnified. In a 'simple graph' an edge between two nodes only indicates the existence of their interaction. In most real-world networks, the interaction should be explained by some other information, named 'edge weight', such as strength or frequency of the interaction. In a weighted network, node strength is defined as the sum of neighbouring edge weights of the node. It reduces to degree if the total edge weight equals 1 in an unweighted network. If we define an edge weight according to the topological structure of the network to magnify the node strength value differences between the supervisor and the ordinary group members, we should get even better consistency coefficient results using the weighted–Newman (W–N) algorithm.[13]

The W–N algorithm also calculates the modular-

ity change in every division step and the best division of communities corresponds to the largest modularity change. However, in a weighted network, the modularity is defined as

$$Q = \frac{1}{2m}\sum_{ij}\Big[W_{ij} - \frac{s_i s_j}{2m}\Big]\delta(c_i, c_j), \qquad (3)$$

where W_{ij} denotes the edge weight between nodes i and j, if i and j are not connected, $W_{ij} = 0$, s_i denotes the node strength of i and $m = \sum_{ij} W_{ij}$ is the total edge weight. Let c_i be the community to which node i is assigned; one has $\delta(c_i, c_j) = 1$ if i and j belong to a common community, $\delta(c_i, c_j) = 0$ otherwise.

Table 1 lists the consistency coefficient results of W–N algorithm partitioning with three different edge weight definitions: (1) $W_{ij} = k_i \times k_j$; (2) $W_{ij} = \max(k_i, k_j)$; (3) $W_{ij} = k_i/k_j$, where k_i denotes the degree of node i. As can be seen in Table 1, both the definitions (2) and (3) give rise to better consistency coefficients than the CNM algorithm where the edge weight equals 1. Among the three definitions, (1) is the worst and (3) is the best. It is easy to analytically show that definition (1) induces the smallest magnification of the node strength differences between the supervisor and the ordinary group members and definition (3) gives the largest. This is in agreement with the results shown in Table 1.

Figure 5 shows the community partitioning result using the W–N algorithm with the edge weight definition k_i/k_j. The four identified communities are indicated by solid, dashed, dotted and dash-dotted lines, respectively.

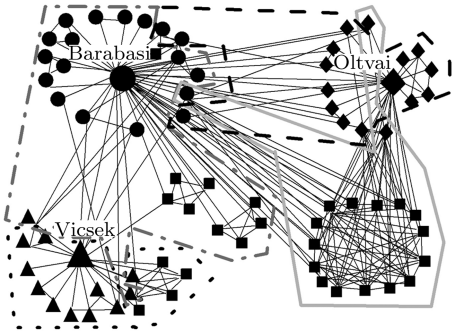


Fig. 5. Community partitioning result using the W–N algorithm with the edge weight definition k_i/k_j.

5. Summary

We argue that in some real-world networks some unknown non-topological information dominates community division. If the non-topological information strongly influences the topological structure, our task is to find the most suitable algorithm to identify the communities based only on network topological information. In this paper a very simple example is investigated where the correct communities are star-like. For this example we show that the W–N algorithm is the best choice. We believe that, for other examples, different algorithms may be chosen for this task.

References

[1] Albert R and Barabasi A L 2002 *Rev. Mod. Phys.* **74** 47
[2] Newman M E J 2003 *SIAM Rev.* **45** 167
[3] Fortunato S 2010 *Phys. Rep.* **486** 75
[4] Newman M E J and Girvan M 2004 *Phys. Rev.* E **69** 026113
[5] Newman M E J 2004 *Phys. Rev.* E **69** 066133
[6] Newman M E J 2003 *Soc. Networks* **25** 83
[7] Clauset A, Newman M E J and Moore C 2004 *Phys. Rev.* E **70** 066111
[8] Palla G, Derenyi I, Farkas I and Vicsek T 2005 *Nature* **435** 814
[9] Chang H, Su B B, Liu C P, Gao M, Di Z R and He D R 2008 *Int. J. Mod. Phys.* C **19** 1537
[10] Xu X L, Fu C H, Liu C P and He D R 2010 *Chin. Phys.* B **19** 060501R
[11] Porter M A, Mucha P J, Newman M E J and Friend A J 2007 *Physica* A **386** 414
[12] Zhang Y, Friend A J, Traud A L, Porter M A, Fowler J H and Mucha P J 2008 *Physica* A **387** 1705
[13] Newman M E J 2004 *Phys. Rev.* E **70** 056131
[14] http://www.nd.edu/∼alb/html/publications.html; http://angel.elte.hu/∼vicsek/