

# A Normalized and a Hybrid Modularity

Haifeng Du<sup>1</sup>, Douglas R. White<sup>2</sup>, Yike Ren<sup>3</sup>, Shuzhuo Li<sup>1,4</sup>

<sup>1</sup>School of Public Policy and Administration, Xi'an Jiaotong University,  
Xi'an, Shaanxi Province, 710049, China

<sup>2</sup>School of Social Sciences, University of California, Irvine, Irvine, CA 92697 USA

<sup>3</sup>School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi Province, 710049

<sup>4</sup>Institute for Population and Development Studies,  
Xi'an Jiaotong University, Xi'an, Shaanxi Province, 710049

**Abstract:** The study of community structures is a hot spot for many inhomogeneous networks. Modularity plays an important role in this area, because it is a criterion for community detection, and a basis for community detection algorithms. Although commonly used in papers concerning community structures, modularity is seldom fully studied. In this paper, we investigate problems with the properties of modularity as defined by Newman and we propose a modularity normalized for number of groups as well as a hybrid modularity that improves on properties that reflect the interactions among communities. We also illustrate the basic flowchart of a “bottom-up merging” community detection strategy based on the properties of modularity, and explore a detection algorithm inspired by hybrid modularity.

## 1 . Introduction

The central idea of “Community Structure” is widely used in the study of social, biological and technical networks, among others. It represents an important future direction of complex networks research.<sup>[1-8]</sup> In order to evaluate community structure, several measurements based on network density have been developed.<sup>[9]</sup> Modularity is the characteristic of a system that has been partitioned into smaller subsystems which interact with each other. The “modularity” index  $Q$  proposed by Newman et al.<sup>[10-14]</sup> is a popular method for evaluating how good is a particular division of networks into communities, assuming they interact but do not overlap. It is not only a criterion for community detection, but also provides basic insights needed to explore other algorithms for community detection.<sup>[1,10,14]</sup> There has been a recent proliferation of community structure studies and algorithms for fast computation of community structure in large networks.<sup>[14-15]</sup> Modularity usually acts as an objective function in these algorithms and thus community structure detection is basically an optimization process for the objective function  $Q$ . There are several other recent algorithms to detect community structure that are representative of the modularity approach<sup>[16-18]</sup>.

A network can be represented as  $G(V, A)$ , where  $V = \{v_1, v_2, \dots, v_n\}$  represents the

set of nodes in the network and  $A$  is the set of connections. Detecting community structure consists of dividing a network into  $m$  sub-networks  $V_p \neq \emptyset, p = 1, 2, \dots, m$ ,

where  $\bigcup_{p=1}^m V_p = V, V_p \cap V_q = \emptyset$  for  $p \neq q$ , and modularity may be defined as<sup>[1]</sup>:

$$Q = \sum_{p=1}^m \left[ e_{pp} - \left( \sum_{q=1}^m e_{pq} \right)^2 \right] \quad (1)$$

Here  $e_{pq}$  is the fraction of edges in the original network that connect vertices in community  $p$  to those in community  $q$ , with  $e_{pp} = e_{pq}$  for  $p = q$ . The definition of modularity  $Q$  can be interpreted in two parts: the first term reflects the connections within the community and the second term describes the connections between any two communities. According to Newman, the larger the value of  $Q$ , the stronger the community structure of the network.<sup>[10]</sup> In practice, values of  $Q$  typically fall in the range of 0.3 to 0.7.<sup>[10-12]</sup>

One problem of modularity  $Q$ , however, is that it is not normalized to deal with networks that are not homogeneous in the sizes of “community structure” clusters. Fig. 1 shows three examples of community structure that are problematic in terms of comparisons.

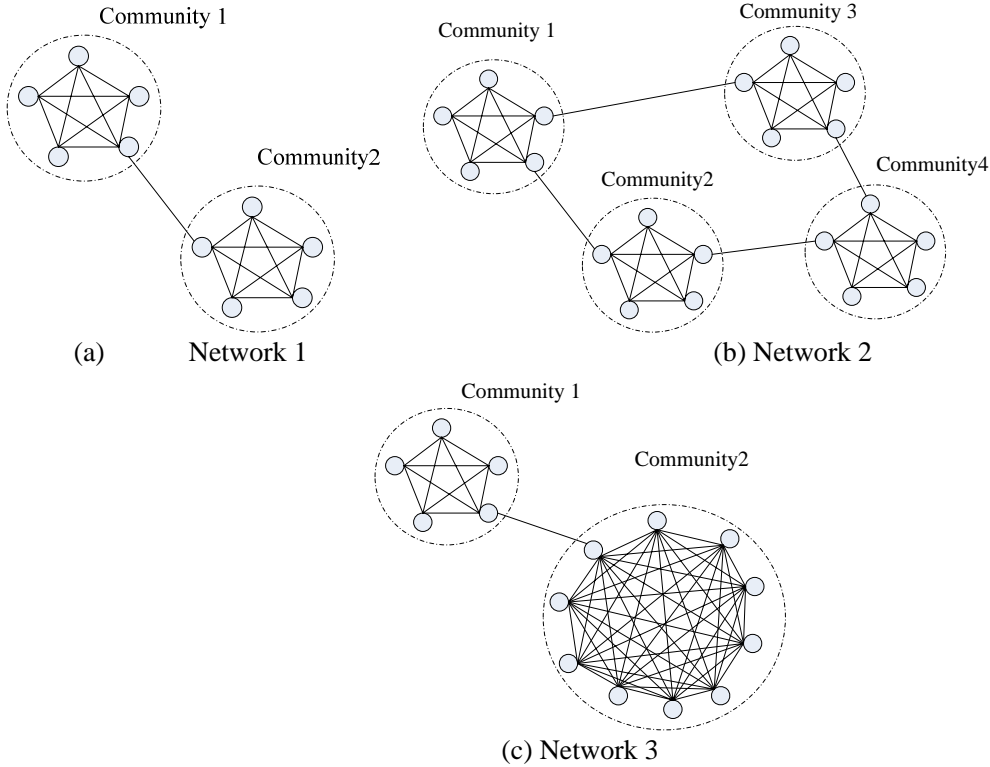


Fig. 1. Three illustrative situations of community structure

**Situations 1 and 2.** Network 1 and 2 shown in Fig. 1 (a) and (b) are very similar to each other. Both of them have very clear community structure and the communities are the same, namely, each sub-network has five fully connected nodes. The only difference is that network 1 has 2 communities while network 2 has 4 communities. According to Eq. (1), we get the modularity  $Q$  values, 0.4036 and 0.6591, for network 1 and network 2, respectively. Why is the modularity value so different for similar community structures?

**Situation 3.** Two communities in network 3 are shown in Fig.1 (c). There are 5 fully connected nodes in community 1, 10 fully connected nodes in community 2 and only one tie between those two communities. We get  $Q= 0.2688$  for network 3.

Intuitively, these comparisons do not agree with Newman’s declaration that the larger the value of  $Q$ , the stronger the community structure of the network. The properties of the modularity  $Q$  need further investigation. In this paper, we first discuss how the network density affects this measure of modularity, and we develop a modularity normalized for number of communities. Second, based on this normalized modularity, we summarize the general steps for the agglomerative or “bottom up” algorithm to detect community structure. Third, we further normalize  $Q$  for consistency with the agglomerative algorithm. Fourth, we show results for various test networks of the effectiveness of the new modularity for community detection and comparison of these modularity values across different networks.

## 2. Normalized Modularity

Network structures can vary widely and there are many different types of community structures for a network. Community detection is an NP hard problem.<sup>1 [19]</sup> Applying a detection algorithm to obtain modularity values for all possible community structures is not feasible. Theoretical analysis of the properties of community structure and modularity is also challenging. However, the calculation for regular networks with a fixed number of edges per node is relatively simple. Fig. 2 shows a regular network<sup>[21]</sup>: we start with a ring of  $n$  nodes, each connects to its  $k$  nearest neighbors by undirected edges.

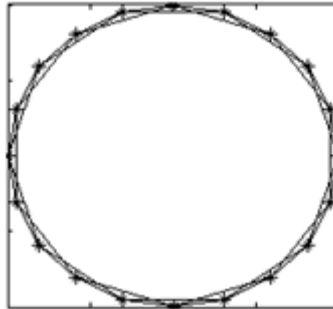


Fig. 2. Regular lattice network

---

<sup>1</sup> A network with  $n$  nodes has  $\sum_{k=1}^n \frac{1}{k!} \sum_{j=1}^k \binom{k}{j} j^n$  different possible community structures.<sup>[20]</sup>

If the regular lattice network has  $m$  communities, the modularity  $Q$  is

$$Q = \frac{nk - mk'(k' + 1)}{nk} - \sum_{i=1}^m (e_i)^2, \quad (2)$$

where  $\sum_{i=1}^m e_i = 1$ ,  $k' = \frac{k}{2}$ , and when  $e_1 = e_2 = \dots = e_m = \frac{1}{m}$ ,

$$Q = 1 - \left( \frac{k+2}{4n} m + \frac{1}{m} \right). \quad (3)$$

If  $\frac{k}{n}$  is small,  $Q$  becomes large. The density of the network  $d = \frac{kn}{0.5n(n-1)} \propto \frac{k}{n}$ .

The more sparse the network, the larger the modularity  $Q$ . For  $e_{pq} = 0$  and  $p \neq q$ , there is no connection between any two communities and the community structure is perfectly partitioned and  $Q$  reaches its maximum. But community structure does not invariably become clearer with larger  $Q$ . The other factor affecting the  $Q$  value in Eq. (3) is  $m$ . And because

$$\begin{aligned} Q &= \sum_{p=1}^m [e_{pp} - (e_{pp})^2] \\ &= \sum_{p=1}^m e_{pp} - \sum_{p=1}^m (e_{pp})^2 \\ &= 1 - \sum_{p=1}^m (e_{pp})^2 \end{aligned} \quad (4)$$

then if  $e_{11} = e_{22} = \dots = e_{mm}$ , and  $Q_{\max} = 1 - \frac{1}{m}$ .

Hence, as the number of communities  $m$  becomes larger, the maximum value of  $Q$  will increase but will never reach 1. Further, for  $Q$  even to reach  $Q_{\max}$  the size of the communities must be identical, as implied by Eq. (4). The definition of  $Q$  thus asserts that communities are most distinct when there is a balance in size among them. As a first step to allow comparison among measures of the extent to which communities structure achieves modularity given differences in community size, we define a *normalized*

*modularity*  $\bar{Q}$  for  $m$  communities that can achieve  $\bar{Q}_{\max} = 1$  if  $m \geq 1$ :

$$\bar{Q} = \frac{m}{m-1} Q = \frac{m}{m-1} \sum_{p=1}^m \left[ e_{pp} - \left( \sum_{q=1}^m e_{pq} \right)^2 \right] \quad (5)$$

We calculate  $\bar{Q}$  for the three networks shown Fig. 1 and get 0.8072, 0.8788 and 0.5376,

respectively. Thus  $\bar{Q}$  is better than  $Q$  in representing the degree of community structure in a network in a way that is independent of the size of the subgroups.

$\bar{Q}$  cannot be used for comparisons across networks, however, unless the network density and total size are equal. Fig.3 shows how the modularity  $Q$  of random networks changes with the density  $d$  (e.g. the connection probability) and the network size  $n$ . The size of the random networks changes from 30 to 400 with a step 5 and the density changes from 0.1 to 0.9 with a step 0.05. Using Newman’s Algorithm, <sup>[12]</sup> we compute the modularity for these random networks and plot  $Q$  and  $\bar{Q}$ , respectively. Each value in Fig. 3 is the average of 100 random simulation results. Generally, the modularity of random networks is  $Q < 0.3$  and  $\bar{Q} < 0.4$ . According to Fig.3, we can see that  $Q$  and  $\bar{Q}$  have similar properties: the value of each increases with decreasing network density, and decreasing network size.

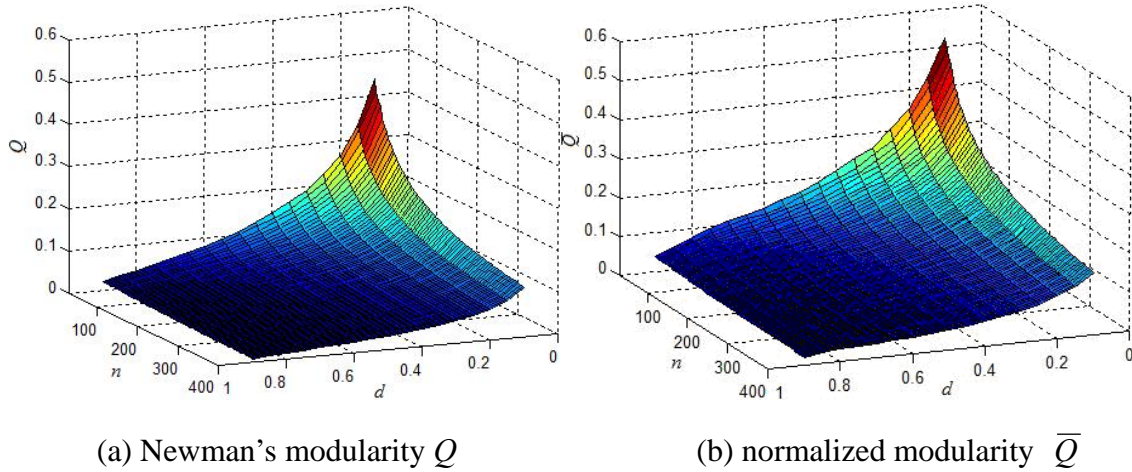


Fig. 3.  $Q$  and  $\bar{Q}$  for the networks with different size and density

### 3. Algorithm

Three basic strategies based on modularity have been proposed to detect community structure by optimizing the modularity function  $Q$ . One is a “top-down” dividing strategy, for example, the *Newman-Girvan algorithm* <sup>[1,10]</sup> based on the “edge betweenness”, and S.White and Smyth’s *spectral clustering approach*, <sup>[2]</sup> which takes the whole network as a single community and then divides a big community into two small subgroups repetively. Another is a genetic algorithm based on mixing, mutation, and evolutionary selection. <sup>[22]</sup> A third is a “bottom-up” agglomerative or merging strategy, as for example Newman’s fast algorithm <sup>[12]</sup> and Clauset et al.’s algorithm for large networks <sup>[14]</sup>, which treats each node as a community and then combines two small communities to form a big

one. The main steps of these merging strategies are:

**Step1:** Take each node in the network as a community;

**Step2:** Calculate  $\varphi_{pq}$  for each pair of communities;  $\varphi_{pq}$  is the change in  $Q$  when combining community  $p$  and  $q$  to form a new community;

**Step3:** Combine the two communities which have the maximum change ( $\varphi_{pq}$ );

**Step4:** Repeat step 2 and 3 until  $\varphi_{pq} \leq 0$

The performance of the algorithm is dependent on  $\varphi_{pq}$  which in turn is determined by the definition of modularity. If we change the definition of  $Q$  we get a different  $\varphi_{pq}$  and the relevant merging algorithm. We consider  $\bar{Q}$ . Because  $\bar{Q}$  and  $Q$  differ only in  $m$ , the normalized modularity  $\bar{Q}$  will yield the same algorithm as the one based on  $Q$  (see Appendix).

More fundamentally, however, there is a particular assumption built into Newman's definition of  $Q$  in Eq. (1), where the term for interactions among communities,

$$\sum_{p=1}^m \left( \sum_{q=1}^m e_{pq} \right)^2 = \sum_{p=1}^m \left( e_{pp} + \sum_{q=1, p \neq q}^m e_{pq} \right) \left( e_{pp} + \sum_{q=1, p \neq q}^m e_{pq} \right) = \sum_{p=1}^m e_{pp}^2 + \sum_{p=1}^m \left( 2e_{pp} \sum_{q=1, p \neq q}^m e_{pq} \right) + \sum_{p=1}^m \left( \sum_{q=1, p \neq q}^m e_{pq} \right)^2$$

is evaluated as the sum of densities both within and between communities and with sums that have equal weight, as if the communities were all of equal size.  $\bar{Q}$  corrects for this only in part. If we want to include in a modularity measure the differences in community sizes we need to consider this interaction term. If we do so, however, we need to evaluate within- and between- community densities not by total network density but by each of the respective row/column expected densities for each  $p, q$  pair which will differ according to their marginals.<sup>2</sup> For this purpose, the middle term in the last expression,

$e_{pp} \sum_{q=1, p \neq q}^m e_{pq}$  is then the key factor for measuring the interactions between community  $p$  and

others. The  $\left( \sum_{q=1}^m e_{pq} \right)^2$  term in the value of  $Q$  as the intercommunity component of

---

<sup>2</sup> Newer community detection algorithms by Reichardt and S. Bornholdt recognize and adjust this problem.<sup>[4-7]</sup>

modularity, then, is exaggerated. To shrink this term to its proper contribution to

modularity, we add  $e_{pp} - e_{pp} \sum_{q=1}^m e_{pq}$  in  $Q$  to define a hybrid modularity  $\tilde{Q}$ :<sup>3</sup>

$$\tilde{Q} = \frac{1}{2} \sum_{p=1}^m \left[ 2e_{pp} - e_{pp} \left( \sum_{q=1}^m e_{pq} \right) - \left( \sum_{q=1}^m e_{pq} \right)^2 \right] \quad (6)$$

As before,  $\tilde{Q}$  will yield the same algorithm as one normalized for differences in number of communities (see Appendix). This means we can normalize  $\tilde{Q}$  while using a  $\varphi_{pq}$  consistent with an unnormalized  $\tilde{Q}$ . Thus we define

$$\bar{\tilde{Q}} = \frac{m}{m-1} \tilde{Q} \quad (7)$$

Fig 4. is the same simulation experiment as shown in Fig. 3.  $\tilde{Q}$  and  $\bar{\tilde{Q}}$  have similar features to  $Q$  and  $\bar{Q}$ , for example,  $\tilde{Q}_{\max} = 1 - \frac{1}{m}$ , but  $\tilde{Q} \geq Q$ ,  $\bar{\tilde{Q}}_{\max} = 1$ , and  $\bar{\tilde{Q}}_{\max} = 1$ .

---

<sup>3</sup> We use  $e_{pp} - e_{pp} \sum_{q=1}^m e_{pq}$  to instead  $e_{pp} - \left( \sum_{q=1}^m e_{pq} \right)^2$  in Newman's modularity  $Q$ , and then get a

simplified definition of the modularity  $\hat{Q} = \sum_{p=1}^m \left[ e_{pp} - e_{pp} \sum_{q=1}^m e_{pq} \right]$ . We have then

$$\varphi_{pq} = 2e_{pq} - 2e_{pq} \times \left( \sum_{i=1}^m e_{pi} + \sum_{i=1}^m e_{qi} \right) - e_{pp} \times \sum_{i=1}^m e_{qi} - e_{qq} \times \sum_{i=1}^m e_{pi} .$$

According to the  $\varphi_{p,q}$ , we detected the community of the networks shown in table 1. The results indicate that the algorithm based on

$\hat{Q}$  is not extremely better than Clauset et al.'s algorithm. So we combined  $\hat{Q}$  and  $Q$  together to put

forward hybrid modularity  $\tilde{Q}$  to improve the detection results.

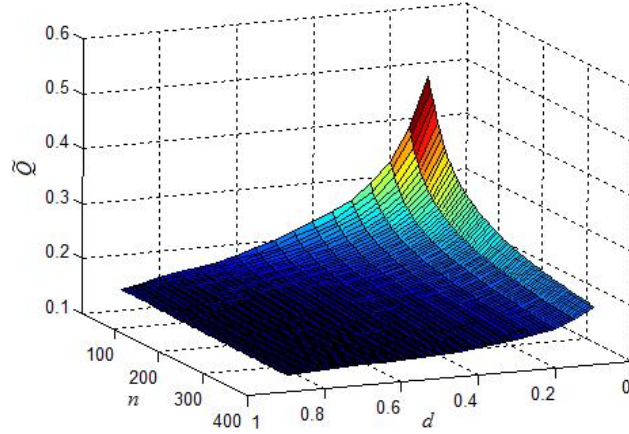


Fig. 4.  $\tilde{Q}$  for the networks with different size and density

For an algorithm based on  $\tilde{Q}$  and the basic framework of “bottom-up” merging strategy,  $\varphi_{pq}$  is calculated by Eq. (8):

$$\varphi_{pq} = 2e_{pq} - e_{pq} \times \left( \sum_{i=1}^m e_{pi} + \sum_{i=1}^m e_{qi} \right) - \frac{1}{2} e_{pp} \times \sum_{i=1}^m e_{qi} - \frac{1}{2} e_{qq} \times \sum_{i=1}^m e_{pi} - \sum_{j=1}^m e_{qj} \sum_{i=1}^m e_{pi} \quad (8)$$

## 4 Applications

Ucinet<sup>[23]</sup> and Pajek<sup>[24]</sup> networks are used as platforms to compare our algorithm based on  $\tilde{Q}$  with Clauset et al.’s algorithm<sup>[14]</sup> for detecting the community structure in stable networks.

The networks in table 1 are a few examples in the Ucinet or Pajek test networks. We list the main characteristics of these networks. All of them are 0-1 symmetric.

We apply the algorithm based on  $\tilde{Q}$  and Clauset et al.’s algorithm to detect the community structure for each network and calculate Newman’s modularity  $Q$  and Normalized modularity  $\bar{Q}$ , respectively. The two modularities for Clauset et al.’s algorithm are denoted as  $Q^c$  and  $\bar{Q}^c$  and for the algorithm based on  $\tilde{Q}$  are denoted as  $Q^m$  and  $\bar{Q}^m$ . We define:

$$\text{gap} = \frac{Q^c - Q^m}{Q^c} \times 100\% \quad (9)$$

$$\overline{\text{gap}} = \frac{\overline{Q}^c - \overline{Q}^m}{\overline{Q}^c} \times 100\% \quad (10)$$

The smaller the gap, the closer the result found by the algorithm based on  $\tilde{Q}$  compared to Clauset et al.'s algorithm. Especially, if  $\text{gap} < 0$ , the algorithm based on  $\tilde{Q}$  gives a higher score than Clauset et al.'s algorithm.

Table 1 The parameters for the networks in Ucinet or Pajek.

Networks	Size	Average degree	Density	Resource
Drugnet	293	0.969	0.013	Ucinet
Zachary	34	2.294	0.278	Ucinet
lcrn	327	2.061	0.013	Pajek
ADF073	262	2.046	0.016	Pajek
BKHAM	44	4.046	0.188	Pajek
BKOFF	40	6.150	0.315	Pajek
c	65	3.846	0.120	Pajek
cc	62	4.645	0.152	Pajek
CENPROD	131	4.817	0.074	Pajek
dnet	180	1.272	0.014	Pajek
GR3_53	144	5.000	0.070	Pajek
GR3_60	120	3.000	0.050	Pajek
KAPTAIL	39	8.103	0.426	Pajek
MREZA3	144	3.653	0.051	Pajek
nooy	85	31.741	0.756	Pajek

Table 2 shows the detection results. According to the value of the gap, we can see that the algorithm based on  $\tilde{Q}$  gives a higher community detection score than Clauset et al.'s algorithm for 40% of the networks shown in table 1, and for 26.6% of the networks, the algorithm based on  $\tilde{Q}$  and Clauset et al.'s algorithm obtain the same detection results. For the remaining networks, the detection results of the algorithm based on  $\tilde{Q}$  algorithm is very close to those of Clauset et al.'s algorithm. Generally, the sign of the gap agrees with that of  $\overline{\text{gap}}$ , because the two algorithms find similar number of the communities for these networks. Due to the effect of the smaller numbers of communities (3-4 versus 5-20) in some of these results, the signs of the gaps for the Zachary, BKHAM and KAPTAIL networks are not the same.

Table 2 Detecting results of the algorithm based on  $\tilde{Q}$  and Clauset et al.'s algorithm

Networks	Clauset et al.'s algorithm			Algorithm based on $\tilde{Q}$			gap ( % )	$\overline{\text{gap}}$ ( % )
	$Q^C$	$\overline{Q}^C$	$m$	$Q^m$	$\overline{Q}^m$	$m$		
Drugnet	0.7447	0.7839	20	0.7478	0.7893	19	<b>-0.4163</b>	<b>-0.6889</b>
Zachary	0.3807	0.5710	3	0.3942	0.5256	4	<b>-3.5461</b>	7.9510
lcrn	0.8831	0.9321	19	0.8827	0.9317	19	0.0453	0.04291
ADF073	0.8810	0.9397	16	0.8810	0.9397	16	<b>0.0000</b>	<b>0.0000</b>
BKHAM	0.1800	0.2700	3	0.1952	0.2603	4	<b>-8.4444</b>	3.5926
BKOFF	0.3413	0.4551	4	0.3367	0.4490	4	1.3478	1.3404
C	0.5778	0.6934	6	0.5813	0.7267	5	<b>-0.6058</b>	<b>-4.8024</b>
Cc	0.6726	0.8408	5	0.6726	0.8408	5	<b>0.0000</b>	<b>0.0000</b>
CENPROD	0.0360	0.0393	12	0.0970	0.1051	13	<b>-169.4400</b>	<b>-167.4300</b>
Dnet	0.6369	0.7643	6	0.6359	0.7631	6	0.1570	0.1570
GR3_53	0.6519	0.8148	5	0.6340	0.7925	5	2.7458	2.7369
GR3_60	0.6739	0.7581	9	0.6739	0.7581	9	<b>0.0000</b>	<b>0.0000</b>
KAPTAIL	0.2910	0.3881	4	0.2716	0.4074	3	6.6667	<b>-4.9729</b>
MREZA3	0.6748	0.7712	8	0.6775	0.7904	7	<b>-0.4001</b>	<b>-2.4896</b>
Nooy	0.7932	0.9065	8	0.7932	0.9065	8	<b>0.0000</b>	<b>0.0000</b>

Fig.5 to Fig. 7 show the community structure detection results of the Zachary, BKHAM and KAPTAIL networks by Clauset et al.'s and the  $\tilde{Q}$ -based algorithms. The first two have a negative gap where  $Q^m$  values surpass those of  $Q^C$ , and the third a negative  $\overline{\text{gap}}$  where  $\overline{Q}^m$  values are lower than  $\overline{Q}^C$ . This allows a comparison of types of differences for discrepancies that occur with fewer numbers of communities. In the first two cases, in Fig.5 and Fig.6, the  $Q^C$  underperformance is associated with not finding an extra community division compared as with  $Q^m$ ; similarly,  $\overline{Q}^m$  underperformance in Figure 7 is associated with not finding an extra community division compared with  $\overline{Q}^C$ .

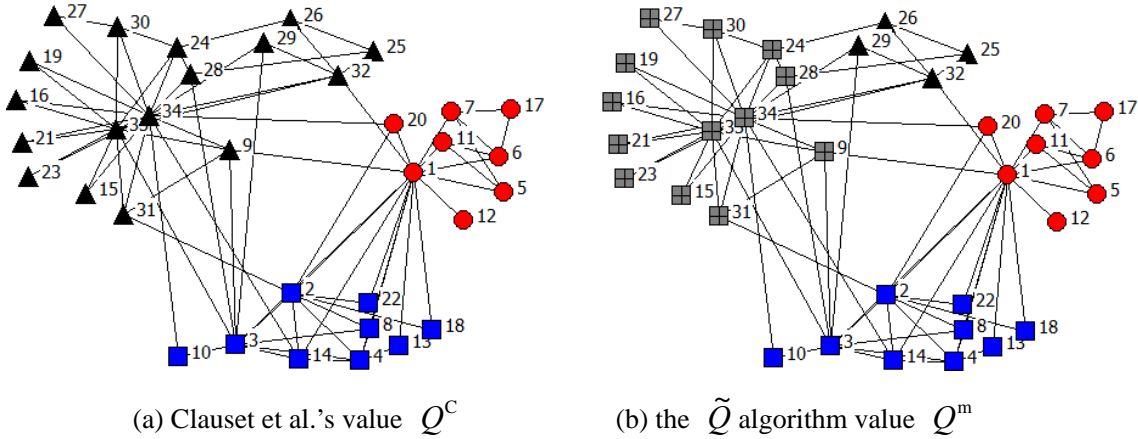


Fig.5 Community structure for Zachary

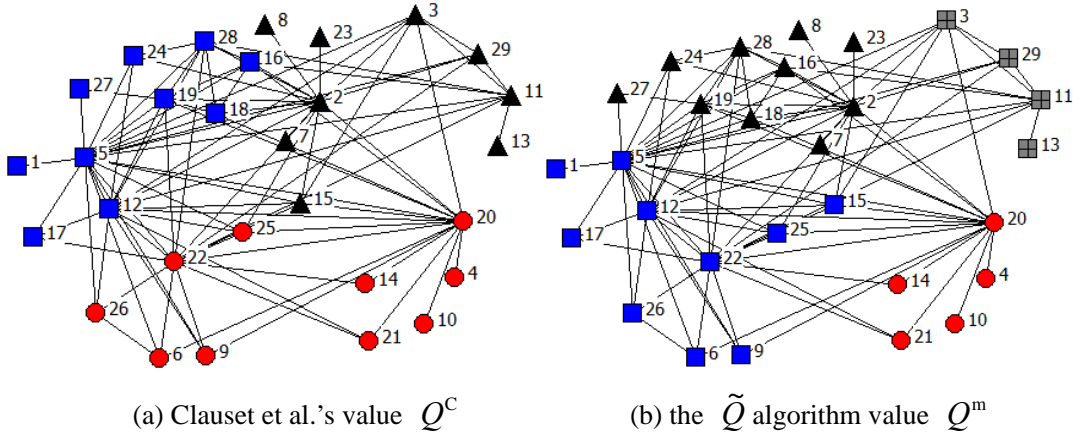


Fig.6 Community structure for BKHAM

Are there “errors” in Fig.5 that would suggest underperformance of  $Q^m$  with respect for example to the black triangle nodes as a distinct community? Nodes “25-26-32” are a complete clique and so constitute a valid community, and given that, node 29 belongs with them. Are there “errors” in Fig.6 that would suggest underperformance of  $Q^m$  with respect for example to the gray square nodes as a distinct set? Node “13” connects only to “11” so they must be in the same community. But “3-29-11” are a complete graph, and so can constitute a separate community, which they do. There are very similar small-size community “penalties” in both cases that suggest that  $Q^m$  is not underperforming but is penalized for possibly identifying very small communities. Here and in Figure 5, Cluset et al.’s  $Q^C$  algorithm gains advantages from having more equal-sized communities in each case, but this does not hold for Figure 6.

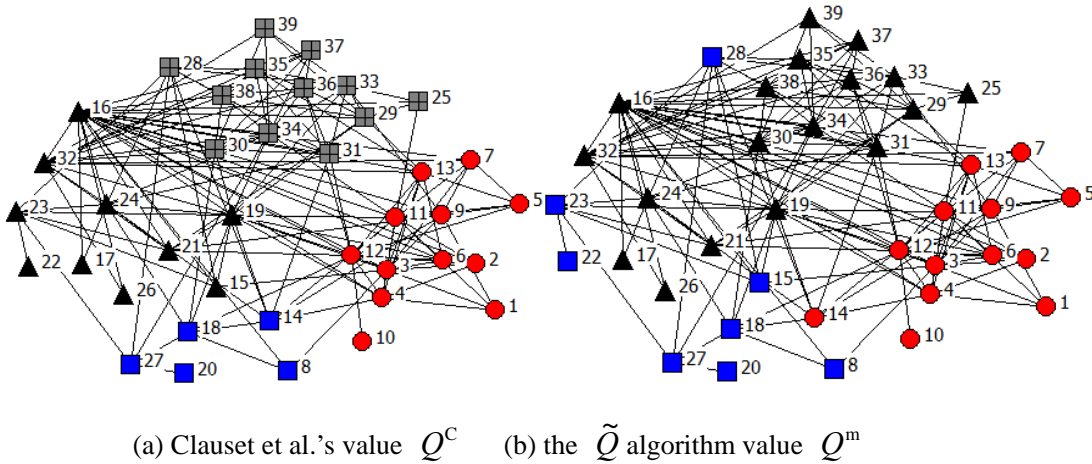


Fig.7 Community structure for KAPTAL

Intuitively, the edge profile of the blue square node labeled “28” in Fig.7(b) should belong to the triangle community, because of its 9 ties, 3 of them interact with the seven other blue squares (3/7), only 1 of them connects to the large red circle community (1/13) and 5 of them connect to the black triangles (5/18). Because of  $3/7 > 5/17 > 1/5$ , relative density should assign it to the blue square community. The  $Q^C$  4-community structure of

Fig.7(b), however, reduces the size of the blue square community and eliminates its only cycle in (b), and thus one of the links of “28” to the blue squares, forcing it into the upper group, where it has more alignments in (a) with the triangles, so Clauset et al.’s algorithm clearly does not reflect relative community density. The error, then, in Clauset et al.’s algorithm, seems to be in having too many clusters.

The “mistake” in each type of case seems to be finding too few communities and ones that are too large rather than too many and too small. Contrary to Newman’s  $Q$ , being too unequal does not discriminate the mistakes in every case.

## 5 Discussion and Conclusion

Basically, the detection procedure of community structure is to divide a network into non-overlapping subgroups. There are two important aspects for community structure. One is how to detect communities, and the other is how to evaluate the detection result. However, the recent study in this area has overwhelmingly focused on the detecting strategies,<sup>[25-37]</sup> few discussions concerning the measurement index. We lay out the problem of modularity, and

(1) Explore the basic characteristics of Newman’s modularity  $Q$ . The definition of  $Q$  indicates the meaning of community structure is clear: the connections within a community are dense and the connections between any two communities are sparser. But  $Q$  value is affected by the density of the network and the number of the communities. Usually, a sparser network with more communities results in a larger  $Q$ .

(2) Find modularity for the community detection using maximization of  $\varphi_{pq}$ , the increase in modularity when combine community  $p$  and  $q$  to form a new community. This is an index for the “bottom-up” merging strategy based on modularity. Calculating  $\varphi_{pq}$  depends totally on the modularity measure. Hence, different definitions of modularity lead to different calculation of  $\varphi_{pq}$  and a different detection processes.

We propose the *normalized modularity*  $\bar{Q}$  which includes to some extent the effect of the number of communities on modularity, and then explore a hybrid modularity and its detection algorithm. The simulation results indicate the effectiveness of the algorithm. Future studies would benefit from a focus on evaluating detection results, the costs and benefits of different modularities and their agglomerative index  $\varphi_{pq}$  in relation to the design, comparison and theoretical analysis of new modularities.

## Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (NSFC) (70671083), Program for New Century Excellent Talents in Universities (NCET) of the Ministry of Education of China (NCET-04-0931, NCET-07-0668), Morrison Institute for Population and Resource Studies at Stanford, Santa Fe Institute International Program and the 2<sup>nd</sup> period of the National 985 Project of the Ministry of Education and Treasury Department of China (07200701). Part of this work was carried out while H. Du was hosted by Anthropology department in School of Social Sciences, University of California, Irvine.

## Appendix

According to the definition of  $\bar{Q}$ , communities  $p$  and  $q$  are combined to form a new community with a new modularity  $\bar{Q}'$  and  $\phi_{pq} = \bar{Q}' - \bar{Q}$ . Further, the number of communities  $m$  will be reduced by 1. Hence for  $\bar{Q}'$ , we have

$$\begin{aligned}\bar{Q}' &= \frac{m-1}{m-2} \left( \sum_{p=1}^m e_{pp} + e_{pq} + e_{qp} - \sum_{p=1}^m \left( \sum_{q=1}^m e_{pq} \right)^2 - 2 \sum_{p=1}^m e_{qp} \sum_{q=1}^m e_{pq} \right) \\ &= \frac{(m-1)^2}{m(m-2)} \bar{Q} + \frac{m-1}{m-2} \left( 2e_{pq} - 2 \sum_{p=1}^m e_{qp} \sum_{q=1}^m e_{pq} \right)\end{aligned}\tag{A1}$$

There are two terms in Eq. (A1):  $\frac{(m-1)^2}{m(m-2)} \bar{Q}$  represents the change in  $\bar{Q}$  due to the decrease in the number of communities and  $\frac{m-1}{m-2} \left( 2e_{pq} - 2 \sum_{p=1}^m e_{qp} \sum_{q=1}^m e_{pq} \right)$  reflects the effects of community emergence. Because  $\frac{(m-1)^2}{m(m-2)} \bar{Q}$  is the same when any two communities combine together, only the second term should be considered to design a detection algorithm. Obviously, when  $m$  is pre-defined,  $\frac{m-1}{m-2} \left( 2e_{pq} - 2 \sum_{p=1}^m e_{qp} \sum_{q=1}^m e_{pq} \right)$  is proportional to  $2e_{pq} - 2 \sum_{p=1}^m e_{qp} \sum_{q=1}^m e_{pq}$ . Hence, to detect community structure, the algorithm

based on the normalized modularity  $\bar{Q}$  will be the same as the one based on  $Q$ .

## References

- [ 1 ] Girvan, M, Newman M E J. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 2002: 7821–7826.
- [ 2 ] White S., Smyth P. A Spectral Clustering Approach to Finding Communities in Graphs. SIAM International Conference on Data Mining. California: Newport Beach, 2005.
- [ 3 ] CRS, Marcelo F. Camperi and Kristina Lisa Klinkner, Discovering Functional Communities in Dynamical Networks, q-bio.NC/0609008
- [ 4 ] J. Reichardt and S. Bornholdt. Clustering of sparse data via network communities - A prototype study of a large online market, J. Stat. Mech. (2007) P06016,
- [ 5 ] J. Reichardt and S. Bornholdt. Statistical Mechanics of Community Detection Phys. Rev. E **74** (2006) 016110
- [ 6 ] J. Reichardt and S. Bornholdt .When are networks truly modular? Physica D **224(1-2)** (2006) 20-26
- [ 7 ] J. Reichardt and S. Bornholdt. Partitioning and modularity of graphs with arbitrary degree distribution Physical Review E, **76(1)** (2007) 015102
- [ 8 ] Newman M E J, Barabási A L, Watts D J. The Structure and Dynamic of Networks. Princeton University Press, 2006.
- [ 9 ] Wasserman S., Faust K. Social network analysis: methods and applications. New York : Cambridge University Press : 1994.
- [ 10 ] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Phys. Rev, 2004, E **69**, 026113.
- [ 11 ] Newman M E J. Detecting community structure in networks. Eur. Phys. J. B, 2004, (**38**): 321-330.
- [ 12 ] Newman M E J. Fast algorithm for detecting community structure in networks. Phys. Rev,

- 2004, E **69**, 066133.
- [ 13 ] Newman M E J, Modularity and community structure in networks, M. E. J. Newman, Proc. Natl. Acad. Sci. USA, 2006, **103**, 8577-8582.
  - [ 14 ] Clauset C, Newman M E J, Moore C. Finding community structure in very large networks. Phys. Rev, 2004, E **70**, 066111.
  - [ 15 ] Xutao Wang, Guanrong Chen, Hongtao Lu. A very fast algorithm for detecting community structures in complex networks. Physica A **384** (2007) 667-674
  - [ 16 ] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino and A. Rapisarda, Dynamical clustering methods to find community structures, physics/0607179
  - [ 17 ] Claire P. Massen, Jonathan P. K. Doye, Thermodynamics of Community Structure, cond-mat/0610077
  - [ 18 ] Chayant Tantipathananandh, Tanya Berger-Wolf and David Kempe, A Framework For Community Identification in Dynamic Social Networks. Proceedings of the 13<sup>th</sup> ACM SIGKDD international conference on Knowledge discovery and data mining 717-736. <http://portal.acm.org/citation.cfm?id=1281192.1281269>
  - [ 19 ] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, Maximizing Modularity is hard, physics/0608255
  - [ 20 ] Richard O D, Peter E H, David G S. Pattern Classification, Second Edition. New York: John Wiley & Sons, Inc., 2001.
  - [ 21 ] Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature 1998, **393**, 440-442.
  - [ 22 ] Tasgin, M. Community Detection Model using Genetic Algorithm in Complex Networks and Its Application in Real-Life Networks, MS Thesis, Graduate Program in Computer Engineering, Bogazici University, 2005.
  - [ 23 ] Download UCINET for Windows, Version 6 Software for Social Network Analysis. <http://www.analytictech.com/downloaduc6.htm>, 2006-04-26.
  - [ 24 ] Pajek Program for Large Network Analysis. <http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm>, 2006-04-26.
  - [ 25 ] Radicchi F, et al. Defining and identifying communities in networks. Preprint

- cond-mat/0309488, 2003.
- [ 26 ] Jake M. Hofman, Chris H. Wiggins, A Bayesian Approach to Network Modularity, [arxiv:0709.3512](#)
  - [ 27 ] Leonardo Angelini, Stefano Boccaletti, Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia, Fast identification of network modules by optimization of ratio association, [cond-mat/0610182](#)
  - [ 28 ] L. Angelini, D. Marinazzo, M. Pellicoro and S. Stramaglia, Natural clustering: the modularity approach, [cond-mat/0607643](#)
  - [ 29 ] Alex Arenas, Alberto Fernandez, Santo Fortunato, Sergio Gomez, Motif-based communities in complex networks, [arxiv:0710.0059](#)
  - [ 30 ] Jim Bagrow and Erik Boltt, A Local Method for Detecting Communities, [cond-mat/0412482](#)
  - [ 31 ] Andrea Capocci, Vito D. P. Servedio, Guido Caldarelli, Francesca Colaiori, Detecting communities in large networks, [cond-mat/0402499](#)
  - [ 32 ] Leon Danon, Albert Díaz-Guilera, Jordi Duch and Alex Arenas, Comparing community structure identification, *Journal of Statistical Mechanics: Theory and Experiment* (2005): **P09008** = [cond-mat/0505245](#)
  - [ 33 ] Jordi Duch and Alex Arenas, Community detection in complex networks using extremal optimization, *Physical Review E* **72** (2005): 027104
  - [ 34 ] G. W. Flake, S. R. Lawrence, C. L. Giles and F. M. Coetzee, Self-organization and identification of Web communities, *IEEE Computer* **36** (2002): 66--71
  - [ 35 ] Santo Fortunato and Marc Bathélemy, Resolution limit in community detection, [Proceedings of the National Academy of Sciences \(USA\)](#) **104** (2007): 36-41
  - [ 36 ] Santo Fortunato, Vito Latora and Massimo Marchiori, A Method to Find Community Structures Based on Information Centrality, [cond-mat/0402522](#)
  - [ 37 ] David Gfeller, Jean-Cédric Chappelier, and Paolo De Los Rios, Finding instabilities in the community structure of complex networks, *Physical Review E* **72** (2005): 056135