The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)
Chengdu-Jiuzhaigou, China, August 19–23, 2010

On Resolution Limit of the Modularity in Community Detection

Junhua Zhang[1,2,*] Xiang-Sun Zhang[1]

[1]Academy of Mathematics and Systems Science, CAS, Beijing 100190, PR China
[2]Key Laboratory of Random Complex Structures and Data Science,
Academy of Mathematics and Systems Science, CAS, Beijing 100190, PR China

Abstract Modularity Q has been broadly used as a valid measure for community detection in complex networks. Fortunato and Barthélemy later proposed that modularity contains an intrinsic scale that depends on the total number of links in the network. But some extra restrictions on some parameters are needed for their analysis. In this paper we further study this problem. Here we give general analysis and more details to show that the resolution limit of Q depends not only on the total links but also on the degree of interconnectedness between pairs of communities. Without any constraint imposed on the parameters, there exists a proper area of the validity (or invalidity) of Q, which is deduced by the definitions of community structure and modularity Q, all these make the analysis here more reasonable.

Keywords Community detection; limited resolution; complex networks

1 Introduction

The research on complex networks has attracted a great deal of attention in recent years. One main reason is that a wide variety of complex systems can be described by networks. Among them, some are the world wide web [1], social networks [2], biological networks [3, 4], the food webs [5] and so on.

One of the most relevant features of complex networks is community structure, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Detecting communities is fundamental for uncovering the links between structure and function in complex networks, and has a lot of applications in many different disciplines such as sociology, biology and computer science. So a huge number of approaches or algorithms have been proposed for community detection in recent years [6, 7, 8, 9, 10, 11, 12, 13, 14, 15], and a great many of methods were reviewed and evaluated in Refs. [16, 17].

In [18] Newman and Girvan introduced a modularity function, which has been broadly used as a valid measure for community structure. Specifically, the modularity Q is defined as follows:

$$Q = \sum_{c=1}^{k} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right] \tag{1}$$

*Corresponding author: zjh@amt.ac.cn

where L is the total number of links in the network, l_c is the number of links within community c, d_c is the sum of degrees of all of the nodes inside community c, and the sum is over all of the k communities of the partition. Actually, the first term of each addend in Eq. (1) is the fraction of edges of the graph inside the community, whereas the second term represents the expected fraction of edges under the configuration model.

Usually it is thought that the modularity function provides a way to determine if a partition is valid to decipher the community structure in a network. So based on maximization of the modularity Q over all the possible partitions of a network many methods have been developed for community detection [19, 20, 21, 22, 23, 24].

However, in [25] Fortunato and Barthélemy pointed out that modularity optimization may fail to identify communities smaller than a scale which depends on the total links of the network, even in cases where communities are unambiguously defined. This reminds the researchers should pay more scrupulousness in using Q. Specially, they determined some resolution thresholds for internal links in some situations. We notice that their analysis proceeds with extra restrictions on some parameters. We think that these restrictions are not proper in some cases, and are a little unreasonable somewhere.

In this paper we further study the problem of resolution limit of modularity Q. unlike the discussion in [25] we don't impose any constraint on the parameters, in fact a proper area for them is deduced by the definitions of community structure and modularity Q. Based on the investigation we give general analysis and more details to show that the resolution limit of Q depends not only on the total links but also on the degree of interconnectedness between pairs of communities.

2 The limited resolution of modularity Q in community detection

According to Eq. (1), for a subgraph $\mathbb{S}$, if the corresponding addend is positive, which means there are more links inside $\mathbb{S}$ than one would expect by random chance, then one can say that $\mathbb{S}$ is indeed a community. That is to say, a subgraph $\mathbb{S}$ with l_c internal links and total degree d_c is a community if

$$\frac{l_c}{L} - \left(\frac{d_c}{2L}\right)^2 > 0. \tag{2}$$

Let $d_c = 2l_c + t_c$, where t_c denotes the number of links joining nodes of community c to the rest of the network. Similar to [25] we express t_c as $t_c = \alpha l_c$, where $\alpha \geq 0$. Then (2) becomes

$$\frac{l_c}{L} - \left(\frac{(2+\alpha)l_c}{2L}\right)^2 > 0. \tag{3}$$

Therefore the authors in [25] conclude that to make the subgraph to be a community there must be an upper limit for the internal links which is

$$l_c < \frac{4L}{(2+\alpha)^2}. \tag{4}$$

Based on the assumption $l_c < L/4$ and $\alpha < 2$ the authors studied the resolution limit of modularity Q. And for some special cases they determined the thresholds for l_c.

But we think that the assumption as well as the discussion somewhere in [25] is not proper, because in the expression $t_c = \alpha l_c$ the parameter α is indeed related to l_c. So something more needs to be clarified about the limited resolution of modularity Q. In the following we'll give some general analysis and more details about this problem.

3 Which factors affect the resolution limit of Q

In fact, from the discussion above we know that

$$0 \le \frac{l_c}{L} \le 1 \quad \text{and} \quad 0 \le \frac{2l_c + t_c}{2L} \le 1.$$

And in order to make a subgraph $\mathbb{S}$ being a community we must have

$$\frac{l_c}{L} - \left(\frac{2l_c + t_c}{2L}\right)^2 > 0, \tag{5}$$

That is,

$$\frac{2l_c + t_c}{2L} < \sqrt{\frac{l_c}{L}}.$$

So we have

$$\begin{aligned} t_c &< 2\sqrt{Ll_c} - 2l_c \\ &= 2\left(\sqrt{\frac{L}{l_c}} - 1\right) l_c. \end{aligned} \tag{6}$$

That is to say, modularity Q directly implies that the outer links must be smaller than a suitable multiple of the inner links for a community.

Furthermore, from $t_c = \alpha l_c$ we obtain

$$\alpha < 2\left(\sqrt{\frac{L}{l_c}} - 1\right). \tag{7}$$

In (6), if $l_c = (1/4)L$, then $t_c < 2l_c$, which means that the total internal degree of the subgraph is larger than its external degree. In this case, the subgraph $\mathbb{S}$ would be a community according to the "weak" definition given by Radicchi et al. [7]. From (7) we also get that $\alpha < 2$.

Actually, whenever the inner link $l_c \in (0, L)$ is given we can get a suitable interval in which α must take values for the subgraph to be a community. For example, when $l_c = (1/2)L$, we get $\alpha < 2(\sqrt{2} - 1)$; when $l_c = (1/9)L$, we get $\alpha < 2(3 - 1) = 4$; and so on. Denote $f(x) = \sqrt{L/x} - 1$, we know that $f(x)$ monotonously decreases in the interval $(0, L)$. So the internal for α will become larger and larger when l_c gets smaller and smaller. However, the width of the corresponding interval for $t_c/(2L)$ will get its maximum at $l_c = (1/4)L$.

In fact, from (6) we know that in order the subgraph to be a community $t_c/(2L)$ must satisfy the following inequality:

$$\frac{t_c}{2L} < \sqrt{\frac{l_c}{L}} - \frac{l_c}{L}. \tag{8}$$

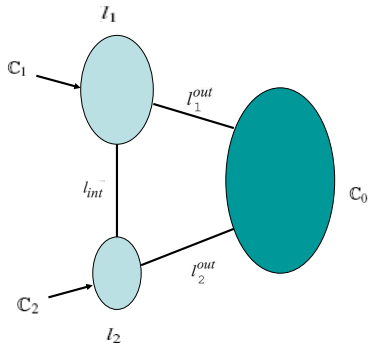

Figure 1: Scheme of a network partition into three or more modules [25].

Let $g(y) = \sqrt{y} - y$ $(0 \le y \le 1)$. Then

$$g''(y) = -\frac{1}{4}\frac{1}{y\sqrt{y}} < 0 \quad (0 < y \le 1).$$

So

$$g(y) = \sqrt{y} - y \quad \begin{cases} \text{increase}, & 0 \le y < \frac{1}{4}, \\ \text{decrease}, & \frac{1}{4} < y \le 1. \end{cases} \tag{9}$$

That is to say, $g(y)$ gets its maximum at $y = 1/4$. Therefore, according to modularity Q, a subgraph $\mathbb{S}$ can be a community only when the inner and the outer links satisfy some conditions. And when $l_c = (1/4)L$ the condition for t_c will be the most loosest which is $t_c/(2L) < \sqrt{l_c/L} - l_c/L = 1/4$. Some more special cases are listed in Table 1.

Table 1 - For some special l_c, the internals in which α and $t_c/(2L)$ must locate for a community

l_c	internal for α	internal for $t_c/(2L)$
L	$\alpha = 0$	$t_c/(2L) = 0$
$(1/2)L$	$\alpha < 2(\sqrt{2}-1) \approx 0.8284$	$t_c/(2L) = \alpha/4 < 0.2071$
$(1/3)L$	$\alpha < 2(\sqrt{3}-1) \approx 1.4641$	$t_c/(2L) = \alpha/6 < 0.2440$
(1/4)L	$\alpha < 2(2-1) = 2$	$t_c/(2L) = \alpha/8 < 0.25$
$(1/5)L$	$\alpha < 2(\sqrt{5}-1) \approx 2.4721$	$t_c/(2L) = \alpha/10 < 0.2472$
$(1/6)L$	$\alpha < 2(\sqrt{6}-1) \approx 2.8990$	$t_c/(2L) = \alpha/12 < 0.2416$
$(1/9)L$	$\alpha < 2(3-1) = 4$	$t_c/(2L) = \alpha/18 < 0.2222$
$(1/16)L$	$\alpha < 2(4-1) = 6$	$t_c/(2L) = \alpha/32 < 0.1875$

In [25], the authors studied the resolution limit of modularity Q, but they only consider the case where $l_c < (1/4)L$ and $\alpha < 2$. In the following we'll further study this problem in a more general framework.

The network analyzed in [25] is also investigated here (see Fig. 1). This network with total L links consists of at least three communities. We focus on a pair of communities, $\mathbb{C}_1$

and $\mathbb{C}_2$, and distinguish three types of links: those internal to each of the two communities (l_1 and l_2, respectively), between $\mathbb{C}_1$ and $\mathbb{C}_2$ (l_{int}) and between the two communities and the rest of the network $\mathbb{C}_0$ (l_1^{out} and l_2^{out}). Just like the discussion above, we express the numbers of external links in terms of those in the communities $\mathbb{C}_1$ and $\mathbb{C}_2$: $l_{int} = \alpha_1 l_1 = \alpha_2 l_2$, $l_1^{out} = \beta_1 l_1$ and $l_2^{out} = \beta_2 l_2$, with $\alpha_1, \alpha_2, \beta_1, \beta_2 \geq 0$.

For the network in Fig. 1 two partitions P_A and P_B are considered. In partition P_A, $\mathbb{C}_1$ and $\mathbb{C}_2$ are taken as separate communities, and in partition P_B they are considered as a single community. The subdivision of the rest of the network, $\mathbb{C}_0$, is arbitrary but identical in both partitions. We want to compare the modularity values Q_{P_A} and Q_{P_B} of the two partitions and, because modularity is a sum over the communities, the contribution of $\mathbb{C}_0$ is the same in both partitions and is denoted by Q_0. By the modularity in (1) we have

$$Q_{P_A} = Q_0 + \frac{l_1}{L} - \left[\frac{2l_1 + l_{int} + l_1^{out}}{2L}\right]^2 + \frac{l_2}{L} - \left[\frac{2l_2 + l_{int} + l_2^{out}}{2L}\right]^2,$$
$$Q_{P_B} = Q_0 + \frac{l_1 + l_2 + l_{int}}{L} - \left[\frac{2l_1 + 2l_2 + 2l_{int} + l_1^{out} + l_2^{out}}{2L}\right]^2.$$

So

$$\begin{aligned}\Delta Q = Q_{P_B} - Q_{P_A} &= \frac{l_{int}}{L} + \left[\frac{2l_1 + l_{int} + l_1^{out}}{2L} + \frac{2l_2 + l_{int} + l_2^{out}}{2L}\right]^2 \\ &\quad - 2 \cdot \frac{2l_1 + l_{int} + l_1^{out}}{2L} \cdot \frac{2l_2 + l_{int} + l_2^{out}}{2L} - \left[\frac{2l_1 + 2l_2 + 2l_{int} + l_1^{out} + l_2^{out}}{2L}\right]^2 \\ &= \frac{l_{int}}{L} - \frac{(2l_1 + l_{int} + l_1^{out})(2l_2 + l_{int} + l_2^{out})}{2L^2}.\end{aligned} \tag{10}$$

Because $\mathbb{C}_1$ and $\mathbb{C}_2$ are both communities by construction, a larger modularity is expected for the partition where the two communities are separated, i.e. $Q_{P_A} > Q_{P_B}$, which in turn implies $\Delta Q < 0$. To this end, from (10) we obtain

$$2Ll_{int} < (2l_1 + l_{int} + l_1^{out})(2l_2 + l_{int} + l_2^{out}).$$

i.e.,

$$2L\alpha_1 l_1 < (2 + \alpha_1 + \beta_1) l_1 \cdot (2 + \alpha_2 + \beta_2) l_2.$$

So we have

$$l_2 > \frac{2L\alpha_1}{(\alpha_1 + \beta_1 + 2)(\alpha_2 + \beta_2 + 2)}. \tag{11}$$

Similarly we have

$$l_1 > \frac{2L\alpha_2}{(\alpha_1 + \beta_1 + 2)(\alpha_2 + \beta_2 + 2)}. \tag{12}$$

Because $\mathbb{C}_1$ and $\mathbb{C}_2$ are both communities, and the outer links of them are $(\alpha_1 + \beta_1) l_1$ and $(\alpha_2 + \beta_2) l_2$, respectively. By (7) we have

$$\alpha_1 + \beta_1 < 2\left(\sqrt{\frac{L}{l_1}} - 1\right), \quad \alpha_2 + \beta_2 < 2\left(\sqrt{\frac{L}{l_2}} - 1\right). \tag{13}$$

Therefore, from (11) we obtain

$$\begin{aligned} l_2 &> \frac{2L\alpha_1}{\left[2\left(\sqrt{\frac{L}{l_1}}-1\right)+2\right]\left[2\left(\sqrt{\frac{L}{l_2}}-1\right)+2\right]} \\ &= \frac{L\alpha_1}{2\sqrt{\frac{L}{l_1}}\cdot\sqrt{\frac{L}{l_2}}} \\ &= \frac{\alpha_1\sqrt{l_1 l_2}}{2}. \end{aligned}$$

i.e.,

$$l_2^2 > \frac{\alpha_1^2 l_1 l_2}{4}.$$

Furthermore we can get

$$l_2 > \frac{\alpha_1^2}{4} l_1 = \frac{\alpha_1}{4}\alpha_1 l_1 = \frac{\alpha_1}{4} l_{int} = \frac{l_{int}^2}{4l_1}.$$

So the condition for $\mathbb{C}_1$ and $\mathbb{C}_2$ separated apart is

$$l_{int}^2 < 4l_1 l_2. \tag{14}$$

On the other hand, we use $\mathbb{C}$ to denote the network generated by merging $\mathbb{C}_1$ and $\mathbb{C}_2$. Then its internal and external links are $l_1 + l_2 + \alpha_1 l_1$ and $\beta_1 l_1 + \beta_2 l_2$, respectively. Because $\mathbb{C}_1$ and $\mathbb{C}_2$ are both communities, by (6) we have $\alpha_1 l_1 + \beta_1 l_1 < 2(\sqrt{Ll_1} - l_1)$ and $\alpha_2 l_2 + \beta_2 l_2 < 2(\sqrt{Ll_2} - l_2)$. So

$$\begin{aligned} \beta_1 l_1 + \beta_2 l_2 &< 2(\sqrt{Ll_1} - l_1) + 2(\sqrt{Ll_2} - l_2) - \alpha_1 l_1 - \alpha_2 l_2 \\ &= 2(\sqrt{L}(\sqrt{l_1} + \sqrt{l_2}) - (l_1 + l_2 + \alpha_1 l_1)). \end{aligned}$$

If $4l_1 l_2 < l_{int}^2$, i.e., $2\sqrt{l_1 l_2} < l_{int}$, then $\sqrt{l_1} + \sqrt{l_2} < \sqrt{l_1 + l_2 + l_{int}}$. So we have

$$\begin{aligned} \beta_1 l_1 + \beta_2 l_2 &< 2(\sqrt{L(l_1 + l_2 + l_{int})} - (l_1 + l_2 + \alpha_1 l_1)) \\ &= 2(\sqrt{L(l_1 + l_2 + \alpha_1 l_1)} - (l_1 + l_2 + \alpha_1 l_1)). \end{aligned}$$

From (6) we know that it is possible for $\mathbb{C}$ to be a community.

It is worth noticing that our discussion is of some difference from the argument in [25], where the authors only demonstrated two special cases with $l_1 = l_2 = l$: when $\alpha_1 = \alpha_2 = 2$ and $\beta_1 \approx 0, \beta_2 \approx 0$, the two communities may be merged if $l < L/4$; and when $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1/l$, the partition may fail to separate the two communities if $l < \sqrt{L/2}$. Here we give some more intuitionistic and more comprehensive understanding about the resolution limit of modularity Q. A more explicit relationship is given for the internal and external links of two communities which are separated by a partition.

Moreover, from (14) we can also get

$$\alpha_1 \alpha_2 < 4, \tag{15}$$

which means that the number of the links between the communities cannot be too large comparative to that of the internal links of each one.

For example, if $\alpha_1 = 3$ and $\alpha_2 = 1$, according to $l_{int} = \alpha_1 l_1 = \alpha_2 l_2$, we have $l_{int} = l_2 = 3l_1$, so $L \geq l_1 + 3l_1 + 3l_1 = 7l_1$ and $L \geq l_2 + (1/3)l_2 + l_2 = (7/3)l_2$. Because $\mathbb{C}_1$ and $\mathbb{C}_2$ are both communities, from (13) we have

$$\beta_1 < 2\left(\sqrt{\frac{L}{l_1}} - 1\right) - 3 \triangleq \Lambda_1, \quad \beta_2 < 2\left(\sqrt{\frac{L}{l_2}} - 1\right) - 1 \triangleq \Lambda_2.$$

Because $\Lambda_1 \geq 2(\sqrt{7}-1)-3 = 2\sqrt{7}-5 \approx 0.2915$ and $\Lambda_2 \geq 2(\sqrt{7/3}-1)-1 = (2/3)\sqrt{21}-3 \approx 0.0551$, we might as well set $\beta_1 = 0.2 = 1/5$ and $\beta_2 = 0.05 = 1/20$. From (11) and (12) we can obtain

$$l_1 > \frac{100}{793}L \approx 0.1261L,$$
$$l_2 > \frac{300}{793}L \approx 0.3783L.$$

Specially, if we set $\beta_1 = \beta_2 = 0$, then

$$l_1 > \frac{2}{15}L \approx 0.1333L,$$
$$l_2 > \frac{2}{5}L = 0.4L,$$

which implies that under this situation the two communities $\mathbb{C}_1$ and $\mathbb{C}_2$ may be merged by optimizing modularity Q if $l_1 < (2/15)L$ and $l_2 < (2/5)L$.

We also notice that in [25] the authors considered the network with $l_1 = l_2 = l$, under $\alpha_1 = \alpha_2 = 2$ they got that if $l < l_R^{max} = L/4$ the two communities may not be resolved. In fact $\alpha_1 = \alpha_2 = 2$ doesn't satisfy the relationship (15).

4 Conclusion

In the past several years the modularity Q has been broadly used as a valid measure for community detection, and usually one thinks that reasonable partitions can be obtained by optimizing Q for networks considered. Nevertheless, Fortunato and Barthélemy [25] recently proposed that modularity optimization can result in incorrect community divisions due to an inherent resolution limit which relates to the total number of links in the network. But the discussion in [25] is based on the pre-establishment of some parameters, for example, $l_c < (1/4)L$ and $\alpha < 2$. This looks constrained and inappropriate, because the proper area can be deduced by the definitions of community structure and modularity Q. In this paper we don't impose any constraint on the parameters, and we give general analysis and more details to show that the resolution limit of Q depends not only on the total links but also on the degree of interconnectedness between pairs of communities.

References

[1] R. Albert, H. Jeong and A.-L. Barabasi, Nature 401,130 (1999).

[2] M. E. J. Newman and J. Park, Phys. Rev. E 68, 036122 (2003).
[3] E. Almaas, B. Kov*á*cs, T. Vicsek, Z. N. Oltvai and A.-L. Barab*á*si, Nature 427, 839-843 (2004).
[4] F. Rao and A. Caflisch, J. Mol. Biol. 342, 299-306 (2004).
[5] J. A. Dunne, R. J. Williams and N. D. Martinez, Proc. Natl. Acad. Sci. USA 99, 12917-12922 (2002).
[6] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).
[7] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Proc. Natl Acad. Sci. USA 101, 2658-2663 (2004).
[8] G. Palla, I. Der*é*nyi, I. Farkas and T. Vicsek, Nature 435, 814 (2005).
[9] J. M. Kumpula, M. Kivel*ä*, K. Kaski and J. Saram*ä*ki, Phys. Rev. E 78, 026109 (2008).
[10] F. Wu and B. A. Huberman, Eur. Phys. J. B 38, 331-338 (2004).
[11] M. E. J. Newman, Proc. Natl Acad. Sci. USA 103, 8577-8582 (2006).
[12] A. Clauset, C. Moore and M. E. J. Newman, Nature 453, 98-101 (2008).
[13] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. 93, 218701 (2004).
[14] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006).
[15] P. Ronhovde and Z. Nussinov, Phys. Rev. E 81, 046114 (2010).
[16] S. Fortunato, Physics Reports 486, 75-174 (2010).
[17] M. A. Porter, J.-P. Onnela and P. J. Mucha, Notices of the AMS 56, 1082-1097 (2009).
[18] M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).
[19] R. Guimer*à* and L. A. N. Amaral, Nature 433, 895-900 (2005).
[20] R. Guimer*à*, M. Sales-Pardo and L. A. N. Amaral, Nature Physics 3, 63-69 (2007).
[21] M.E.J. Newman, Eur. Phys. J. B 38, 321-330 (2004).
[22] M. E. J. Newman, Phys. Rev. E 74, 036104 (2006).
[23] J. Duch and A. Arenas, Phys. Rev. E 72, 027104 (2005).
[24] J. Zhang, S. Zhang and X.-S. Zhang, Physica A 387, 1675-1682 (2008).
[25] S. Fortunato and M. Barth*é*lemy, Proc. Natl. Acad. Sci. USA 104, 36-41 (2007).