Limitation and applicability of modularity measures in community detection

Supplementary Materials 3

Xiang-Sun Zhang [a]*, Rui-Sheng Wang [b], Yong Wang [a], Ji-Guang Wang [a], Yu-Qing Qiu [a], Lin Wang [a], and Luonan Chen [c]

[a]Academy of Mathematics and Systems Science, CAS, Beijing 100080, China, [b]Department of Physics, The Pennsylvania State University, University Park, PA 16801, USA, [c]Department of Electronics, Information, and Communication Engineering, Osaka Sangyo University, Osaka 574-8530, Japan.

Relationship between the modularity measures Q, D, and the weak definition of community

Given a graph $G = (V, E)$, and partitions $P_K = \{G_1, G_2, \cdots, G_K\}$=$\{(V_1, E_1)$,$(V_2, E_2)$,$\cdots$,$(V_K, E_K)\}$, $K = 1, \cdots, |V|$, we denote $Q(P), D(P)$ as the values of Q, D on the partition P.

Relationship between D and the weak definition of community

For modularity measure D, we have the following proposition.

Proposition 2 Let us denote $D(P) = \sum_{i=1}^{K} D_i$. If for $\forall\ i$, G_i satisfies the weak definition [1], then we have $D(P) > 0$.
Proof. If G_i satisfies the weak definition, then $D_i > 0$ is valid. It is easy to see that $D(P) = \sum_{i=1}^{K} D_i > 0$ is valid.

We note that the reverse of **Proposition 2** is not correct, i.e., if $D(P) > 0$ then it is not necessary all G_i ($\forall\ i$) satisfies the weak definition. An example is shown in Figure 1(b) in the main text. Suppose we have a network which is a 15-clique connected with another two nodes by three edges. In our experimental the optimization of D partition the network into two communities (Shown in Figure 1(b) in the main text). The first one is the 15-clique with D_i value 13.8. The second one is the community with two nodes with D_i value -0.5. So $D(P) = 13.3 > 0$. However we can simply check that the second community does not satisfy the weak definition.

Relationship between Q and the weak definition of community

For modularity measure Q, we have the following proposition.

Proposition 3 Let us denote $Q(P) = \sum_{i=1}^{K} Q_i$. If for $\forall\ i$, G_i satisfies the weak definition [1], then there exists a constant B, and we have $Q_i > B \geq 0$, i.e., Q has a positive lower bound.
Proof
(1) In the simplest situation, $P = \{G_1, G_2\}$, we use $S_1 = L(V_1, V_1), S_2 = L(V_2, V_2), S_{12} = L(V_1, V_2)$ to denote the edges in G_1, G_2, and the edges between them respectively (as shown in figure S1).

*To whom correspondence should be addressed at: E-mails: zxs@amt.ac.cn

$$
\begin{aligned}
Q_1 &= \frac{S_1}{L} - \left(\frac{2S_1 + S_{12}}{2L}\right)^2 \\
4L^2 Q_1 &= 4S_1 L - (2S_1 + S_{12})^2 \\
&= 4S_1(S_1 + S_2 + S_{12}) - 4S_1^2 - 4S_1 S_{12} - S_{12}^2 \\
&= 4S_1 S_2 - S_{12}^2,
\end{aligned}
$$

where L represents the number of edges in the whole network. By the weak definition of module,

$$
\begin{cases} 2S_1 > S_{12} \\ 2S_2 > S_{12}, \end{cases}
$$

we have

$$
Q_1 = \frac{4S_1 S_2 - S_{12}^2}{4L^2} > 0.
$$

Similarly we have

$$
Q_2 = \frac{4S_1 S_2 - S_{12}^2}{4L^2} > 0.
$$

Thus let

$$
B = \frac{4S_1 S_2 - S_{12}^2}{2L^2} \geq 0.
$$

We have $Q = Q_1 + Q_2 > B \geq 0$

(2) Accordingly, for $K = 3$, we have

$$
Q_1 > \frac{S_{23}(S_{12} + S_{13})}{L^2} > 0
$$

$$
Q_2 > \frac{S_{13}(S_{12} + S_{23})}{L^2} > 0
$$

$$
Q_3 > \frac{S_{12}(S_{13} + S_{23})}{L^2} > 0
$$

Thus let

$$
B = \frac{S_{23}(S_{12} + S_{13}) + S_{13}(S_{12} + S_{23}) + S_{12}(S_{13} + S_{23})}{L^2} > 0.
$$

and we have $Q = Q_1 + Q_2 + Q_3 > B > 0$

(3)In a general case,

$$
\begin{aligned}
Q_1 &= \frac{S_1}{L} - \left(\frac{2S_1 + \sum_{j=2}^{K} S_{1j}}{2L}\right)^2 \\
4L^2 Q_1 &= 4S_1 L - (2S_1 + \sum_{j=2}^{K} S_{1j})^2 \\
&= 4S_1(\sum_{i=1}^{K} S_i + \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} S_{ij}) - 4S_1^2 - 4S_1 \sum_{j=2}^{K} S_{1j} - \left(\sum_{j=2}^{K} S_{1j}\right)^2 \\
&= 4S_1 \sum_{i=2}^{K} S_i + 4S_1 \sum_{i=2}^{K-1} \sum_{j=i+1}^{K} S_{ij} - \left(\sum_{j=2}^{K} S_{1j}\right)^2 = (*).
\end{aligned}
$$

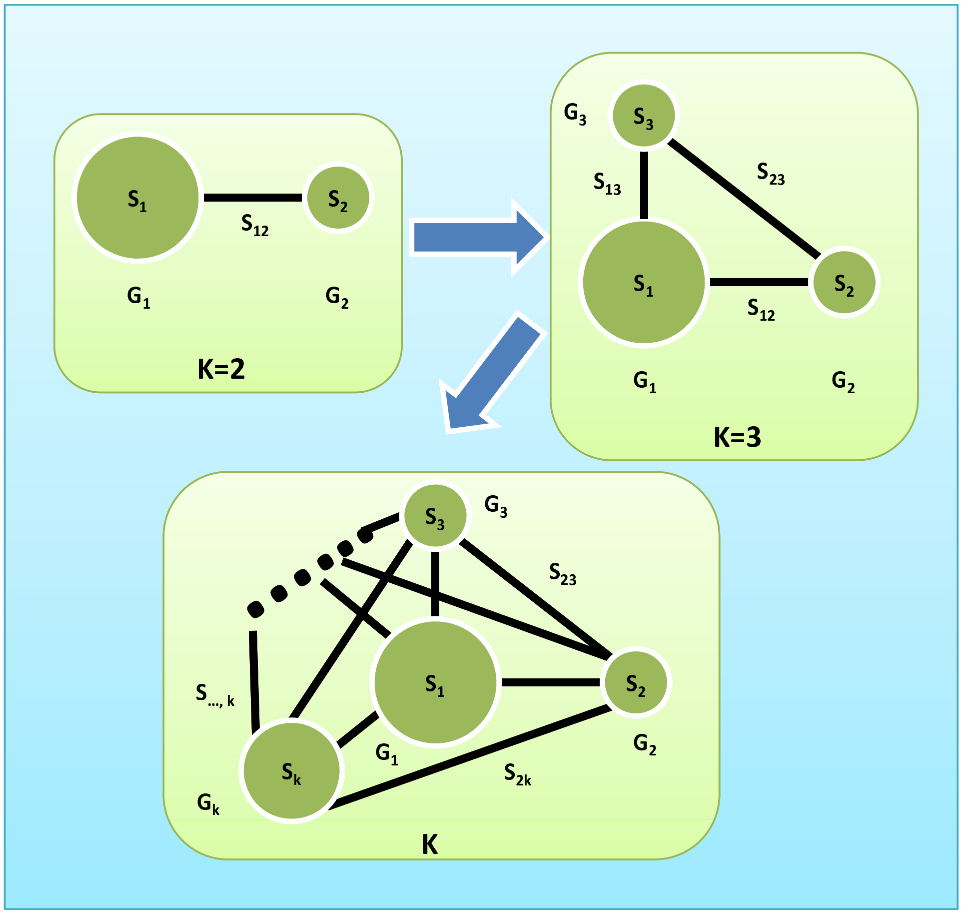

Figure S1: Illustration of different partitions for graph G.

Since all the modules satisfy weak definition, we have

$$\begin{cases} 2S_1 > \sum_{j=2}^{K} S_{1j} \\ 2S_2 > \sum_{j\neq 2} S_{2j} \\ \cdots \\ 2S_K > \sum_{j\neq K} S_{Kj}, \end{cases}$$

so

$$\begin{aligned}
(*) \quad &> \left(\sum_{j=2}^{K} S_{1j}\right)\sum_{i=2}^{K}\sum_{j\neq i} S_{ij} + 2\left(\sum_{j=2}^{K} S_{1j}\right)\sum_{i=2}^{K-1}\sum_{j=i+1}^{K} S_{ij} - \left(\sum_{j=2}^{K} S_{1j}\right)^2 \\
&= \left(\sum_{j=2}^{K} S_{1j}\right)\left(\sum_{i=2}^{K}\sum_{j\neq i} S_{ij} + 2\sum_{i=2}^{K-1}\sum_{j=i+1}^{K} S_{ij} - \sum_{j=2}^{K} S_{1j}\right) \\
&= \left(\sum_{j=2}^{K} S_{1j}\right) 4\sum_{i=2}^{K-1}\sum_{j=i+1}^{K} S_{ij},
\end{aligned}$$

and further

$$\begin{aligned}
Q_1 &> \frac{4\left(\sum_{j=2}^{K} S_{1j}\right)\sum_{i=2}^{K-1}\sum_{j=i+1}^{K} S_{ij}}{4L^2} \\
&= \frac{\left(\sum_{j=2}^{K} S_{1j}\right)\sum_{i=2}^{K-1}\sum_{j=i+1}^{K} S_{ij}}{L^2} > 0,
\end{aligned}$$

then we can get a lower bound B for Q as

$$B = \sum_{l=1}^{K} \frac{\left(\sum_{j\neq l} S_{lj}\right)\sum_{i\neq l}\sum_{j=i+1}^{K} S_{ij}}{L^2}$$

Thus we have $Q = \sum_{l=1}^{K} Q_l > B > 0$.

Similarly the reverse of **Proposition 3** is not correct, i.e., if $Q(P) > 0$ then it is not necessary all G_i ($\forall\ i$) satisfies the weak definition. A simple example is given in Figure 1(a) in main text, where there are five 6-cliques, any two of which are connected by eight links. Experimental result shows that the optimization of Q partition the network into five communities and identify every 6-clique as a separate community (Shown in Figure 1(a) in the main text). The optimal value of $Q(P) = 0.280$ is larger than zero. However these five communities all have 15 inner-links and 32 out-links and do not satisfy the weak community definition. We further calculate the lower bound B in **Proposition 3** which has a value 0.32 and is larger than the optimal value of $Q(P)$.

References

[1] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proc Natl Acad Sci U S A*, 101(9):2658–2663, March 2004.