# Modularity optimization in community identification of complex networks

# Supplementary Material 2

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.

## Discrete convex analysis on Q and D

For an arbitrary partition of a network $P = \{G_1, G_2, \cdots, G_K\} = \{(V_1, E_1), (V_2, E_2), \cdots, (V_K, E_K)\}$, we discuss the two-stage optimization problems:

$$Q_{II} : \max_K Q_I(k) = \max_K \max_{P_k} \sum_{s=1}^{K} Q_s; \tag{1}$$

and

$$D_{II} : \max_K D_I(K) = \max_K \max_{P_k} \sum_{s=1}^{K} D_s; \tag{2}$$

where $Q_I(K)$ and $D_I(K)$ are the solutions from the first-step optimization problems. And

$$Q_I : \max_K Q_I(K) \;\text{ and }\; D_I : \max_K D_I(K) \tag{3}$$

are the second-step optimization problems.

Two exemplary modular networks are used here. One is a ring of dense lumps which consist of $N$ $(N \geq 8$ *and* $N = 2^k, k = \{3, 4, 5 \cdots\})$ dense lumps each with $m$ nodes. There are $l_{bw}$ links between adjacent lumps. Let $A_s, s = 1, 2, \cdots, N$ denote the $m \times m$ adjacency matrix of the $s$th lump $G_s = (V_s, E_s)$. Thus, the adjacent matrix $A$ of the whole network is $Nm \times Nm$. Here we assume that all lumps have the same number of links $l_{in}$. The second exemplary network is a special version of the ad hoc network, which also consist of $N$ dense subgraphs, but there are $l_{bw}$ link between each pair of dense subgraphs. So the total number of links in this network is $L = Nl_{in} + N(N-1)l_{bw}/2$. When $L$ is fixed, the larger $l_{bw}$ is, the more ambiguous the lumps $G_s$ become; the larger $l_{in}$ is, the more loosely connection between the lumps.

### The ring network of lumps

(1) Modularity function $Q$

Suppose that we partition the whole network into $K$ communities with each community containing $N_i$ lumps, $N_1 + \cdots + N_K = N$. When $K = 1$, $Q_P = 0$, then we discuss the situation of $K \geq 2$ as follows:

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

$$
\begin{aligned}
&\max_{K\geq 2}\max_{P_k}\sum_{i=1}^{K} Q_i \\
= &\max_{K\geq 2}\max_{\sum_{i=1}^{K} N_i=N}\sum_{i=1}^{K}\left[\frac{N_i 2l_{in}+2(N_i-1)l_{bw}}{N2l_{in}+2Nl_{bw}}-\left(\frac{N_i 2l_{in}+2(N_i-1)l_{bw}+2l_{bw}}{N2l_{in}+2Nl_{bw}}\right)^2\right] \\
= &\max_{K\geq 2}\max_{\sum_{i=1}^{K} N_i=N}\sum_{i=1}^{K}\frac{-1}{(N2l_{in}+2Nl_{bw})^2}[(2l_{in}+2l_{bw})^2N_i^2-N(2l_{in}+2l_{bw})^2N_i \\
&\quad +2Nl_{bw}(2l_{in}+2l_{bw})] \\
= &\max_{K\geq 2}\max_{\sum_{i=1}^{K} N_i=N}\sum_{i=1}^{K}\frac{1}{N^2}\left(-N_i^2+NN_i-\frac{2Nl_{bw}}{2l_{in}+2l_{bw}}\right) \\
= &\max_{K\geq 2}\max_{\sum_{i=1}^{K} N_i=N}\left\{1-\frac{Kl_{bw}}{N(l_{in}+l_{bw})}-\sum_{i=1}^{K}\frac{N_i^2}{N^2}\right\}
\end{aligned}
$$

Note that the first-step optimization problem is a discrete convex program in the feasible region $F=\{1,2,4,\cdots,N/2^{s+1},N/2^s,N/2^{s-1},\cdots,N\}$. A function (or a programming) whose variables take discrete values (or, say, the sample values) is called as discrete convex (concave) function (or programming) if they can be embedded into a continuous convex (concave) function (or programming). Solving the K-K-T equation of the above first-step optimization problem leads to $N_1=\cdots=N_K=\frac{N}{K}$, then

$$
\max_{K\geq 2} Q_I(K) \quad = \quad \max_{K\geq 2}\left\{1-\frac{1}{K}-\frac{l_{bw}}{N(l_{in}+l_{bw})}K\right\}.
$$

So

$$
Q_I(K)=\begin{cases} 1-\frac{1}{K}-\frac{l_{bw}}{N(l_{in}+l_{bw})}K & K\geq 2 \\ 0 & K=1 \end{cases}
$$

It is easy to see that $Q_I(K)$ is a discrete concave function, then the solution is given by the derivative of $Q_I(K)$ at zero. we have solution

$$
K^*=\langle\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}}\sqrt{N}\rangle_F \tag{4}
$$

where $\langle\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}}\sqrt{N}\rangle_F$ means the integer in $F$ nearest to $\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}}\sqrt{N}$. The solution is either on the boundary of $F$ or an interior point of $F$ depending on the values of $l_{bw}$ and $l_{in}$: $Q_I(1)\leq\cdots\leq Q_I(N/2^s)\leq\cdots\leq Q_I(N)$, when $l_{bw}\leq\frac{l_{in}}{N-1}$; $Q_I(\langle\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}}\sqrt{N}\rangle_F)\geq\max\{Q_I(N),Q_I(1)\}$, when $l_{bw}>\frac{l_{in}}{N-1}$.

When $\langle\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}}\sqrt{N}\rangle_F=N$, ( $l_{bw}<\frac{l_{in}}{9N/16-1}$), $Q$ identifies each lump as a qualified community. As $l_{bw}$ becomes larger, the optimal $K$ will be less than $N$ so that $Q$ fails to identify qualified communities, i.e., it suffers from resolution limit until the value of $l_{bw}$ reaches to $l_{in}$. When $l_{bw}>l_{in}$, the single lump will not satisfy the weak definition of community anymore.

(2) Modularity density $D$

$$
\begin{aligned}
&\max_{K\geq 2}\max_{P_k}\sum_{i=1}^{K} D_i \\
= &\max_{K\geq 2}\max_{\sum_{i=1}^{K} N_i=N}\left\{\sum_{i=1}^{K}\left(\frac{N_i 2l_{in}+2(N_i-1)l_{bw}}{N_i m}-\frac{2l_{bw}}{N_i m}\right)\right\} \\
= &\max_{K\geq 2}\max_{\sum_{i=1}^{K} N_i=N}\sum_{i=1}^{K}\left(\frac{-4l_{bw}}{N_i m}+\frac{2l_{in}+2l_{bw}}{m}\right)
\end{aligned}
$$

where $m$ is the number of nodes in $A_s$. The first-step optimization is a convex programming problem with solution $N_1=\cdots=N_K=\frac{N}{K}$, then

$$\max_{K\geq 2} D_I(K)=\max_{K\geq 2}\left\{-\frac{4l_{bw}}{m}\frac{K^2}{N}+K\frac{2l_{in}+2l_{bw}}{m}.\right\} \tag{5}$$

So

$$D_I(K)=\begin{cases} -\frac{4l_{bw}}{m}\frac{K^2}{N}+K\frac{2l_{in}+2l_{bw}}{m} & K\geq 2 \\ \frac{2(l_{in}+l_{bw})}{m} & K=1 \end{cases}$$

The solution is $K^*=\langle\frac{(l_{in}+l_{bw})N}{4l_{bw}}\rangle_F$. With the same reasoning for $Q$, we can easily get $D_I(1)\leq\cdots\leq D_I(N/2^s)\leq\cdots\leq D_I(N)$, when $l_{bw}\leq\frac{l_{in}}{3}$; $D_I(\langle\frac{l_{in}+l_{bw}}{4l_{bw}}N\rangle_F)\geq\max\{D_I(N),D_I(1)\}$ , when $l_{bw}>\frac{l_{in}}{3}$.

When $\langle\frac{l_{in}+l_{bw}}{4l_{bw}}N\rangle_F=N$ ( $l_{bw}<\frac{l_{in}}{2}$), $D$ identifies each lump as a community satisfying the weak definition. But when $l_{bw}$ becomes larger, the optimal $K$ will be less than $N$, and $D$ fails to identify qualified communities, i.e., it suffers from resolution limit until $l_{bw}=l_{in}$, from where single lump does not satisfy the weak definition of community anymore.

## The *ad hoc* network

(1) Modularity function $Q$

$$
\begin{aligned}
&\max_{K}\max_{P_k}\sum_{i=1}^{K} Q_i \\
= &\max_{K}\max_{\sum_{i=1}^{k} N_i=N}\sum_{i=1}^{K}\left[\frac{N_i 2l_{in}+N_i(N_i-1)l_{bw}}{N2l_{in}+N(N-1)l_{bw}}-\left(\frac{N_i 2l_{in}+N_i(N_i-1)l_{bw}+N_i(N-N_i)l_{bw}}{N2l_{in}+N(N-1)l_{bw}}\right)^2\right] \\
= &\max_{K}\max_{\sum_{i=1}^{k} N_i=N}\sum_{i=1}^{k}\left\{\frac{2l_{in}-l_{bw}}{N^2[2l_{in}+l_{bw}(N-1)]}(-N_i^2+NN_i)\right\}
\end{aligned}
$$

Note that the first-step optimization is a convex programming if $l_{bw}<2l_{in}$, then it has solution $N_1=\cdots=N_K=\frac{N}{K}$. We further have

$$Q_{II}:\ \max_{K}\left\{\frac{2l_{in}-l_{bw}}{2l_{in}+l_{bw}(N-1)}(1-\frac{1}{K})\right\} \tag{6}$$

as a convex problem and the solution is $K^*=N$.

When $2l_{in}<l_{bw}$, $Q_I$ is a concave programming, the solution is reached at the boundary. Note that $Q_I(K)$ is a monotonously decreasing function, then $K^*=1$.

Since each dump in the *ad hoc* network will not satisfy the weak definition when $l_{bw}\geq\frac{2l_{in}}{N-1}$, $Q$ suffers misidentification when $\frac{2l_{in}}{N-1}<l_{bw}<2l_{in}$.

Table S2-1: The properties of optimization model to maximize the modularity measures $Q$ and $D$ on two exemplary networks.

| | The ring of lumps | The Ad hoc network |
|---|---|---|
| $Q$ | $Q_I(K)$ is a discrete concave function, thus $Q_{II}$ is a discrete convex programming | $Q_I(K)$ is a discrete concave function, thus $Q_{II}$ is a discrete convex programming when $l_{bw} < 2l_{in}$. and $Q_I(K)$ is a discrete convex function, thus $Q_{II}$ is a discrete concave programming when $l_{bw} \geq 2l_{in}$ |
| | $Q_I$ is a discrete convex programming | $Q_I$ is a discrete convex programming when $l_{bw} < 2l_{in}$, and a discrete concave programming when $l_{bw} \geq 2l_{in}$ |
| $D$ | $D_I(K)$ is a discrete concave function, thus $D_{II}$ is a discrete convex programming | $D_I(K)$ is a linear function, thus $D_{II}$ is a linear programming, then is a discrete convex programming |
| | $D_I$ is a discrete convex programming | $D_I$ is a linear programming |

(2) Modularity density $D$

$$\begin{aligned}
& \max_K \max_{P_k} \sum_{i=1}^{K} D_i \\
= & \max_K \max_{\sum_{i=1}^{K} N_i = N} \sum_{i=1}^{K} \left\{ \frac{N_i 2l_{in} + N_i(N_i - 1)l_{bw}}{N_i m} - \frac{N_i(L - N_i)l_{bw}}{N_i m} \right\} \\
= & \max_K \max_{\sum_{i=1}^{K} N_i = N} \frac{1}{m} \sum_{i=1}^{K} \{2l_{in} + 2N_i l_{bw} - (N+1)l_{bw}\}
\end{aligned}$$

Now the first-step optimization is a simple linear programming problem with any feasible solution as the optimal solution. Then

$$D_{II} := \max_K \{K(2l_{in} - (N+1)l_{bw}) + 2N l_{bw}\} \tag{7}$$

is also a linear function, then

$$K^* = \begin{cases} N & if \;\; l_{bw} < 2l_{in}/(N+1), \\ 1 & if \;\; l_{bw} > 2l_{in}/(N+1). \end{cases} \tag{8}$$

and when $l_{bw} = \frac{2l_{in}}{N+1}$, any $K$ is a solution.

Note that each lump in the *ad hoc* network will not satisfy the weak definition when $l_{bw} \geq \frac{2l_{in}}{N-1}$, then $D$ suffers resolution limit for $\frac{2l_{in}}{N+1} < l_{bw} < \frac{2l_{in}}{N-1}$.

The above analysis are summarized in two tables S2-1 and S2-2.

Table S2-2: The result of community partitions of two exemplary networks using different modularity measures.

| | The ring of lumps | The ad hoc network |
|---|---|---|
| $Q$ | $Q_I(1) \leq \cdots \leq Q_I(N/2^s) \leq \cdots \leq Q_I(N)$, when $l_{bw} \leq \frac{l_{in}}{N-1}$<br>$Q_I(\langle\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}}\sqrt{N}\rangle_F) \geq \max\{Q_I(N), Q_I(1)\}$, when $l_{bw} > \frac{l_{in}}{N-1}$ | $Q_I(1) \leq \cdots \leq Q_I(N/2^s) \leq \cdots \leq Q_I(N)$, when $l_{bw} \leq 2l_{in}$<br>$Q_I(N) \leq \cdots \leq Q_I(N/2^s) \leq \cdots \leq Q_I(1)$, when $l_{bw} > 2l_{in}$ |
| $D$ | $D_I(1) \leq \cdots \leq D_I(N/2^s) \leq \cdots \leq D_I(N)$, when $l_{bw} \leq \frac{l_{in}}{3}$<br>$D_I(\langle\frac{l_{in}+l_{bw}}{4l_{bw}}N\rangle_F) \geq \max\{D_I(N), D_I(1)\}$ , when $l_{bw} > \frac{l_{in}}{3}$ | $D_I(1) \leq \cdots \leq D_I(N/2^s) \leq \cdots \leq D_I(N)$, when $l_{bw} \leq \frac{2l_{in}}{N+1}$<br>$D_I(N) \leq \cdots \leq D_I(N/2^s) \leq \cdots \leq D_I(1)$, when $l_{bw} > \frac{2l_{in}}{N+1}$ |