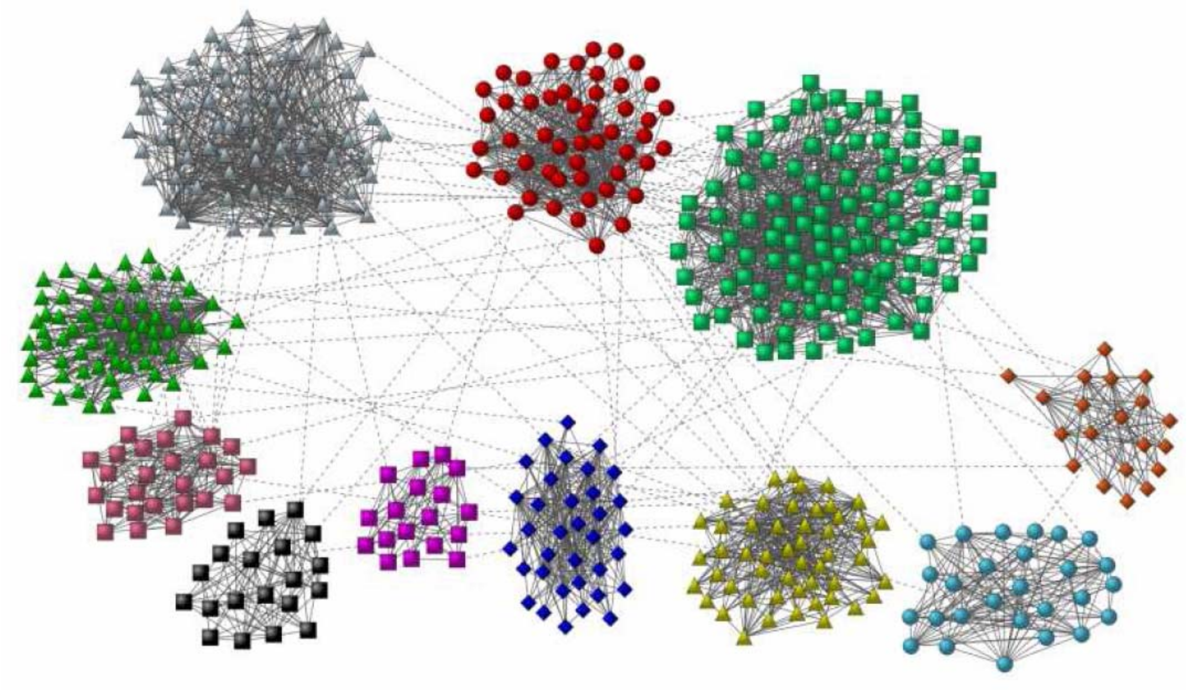
Significance of Community Structure

胡延庆

北京师范大学管理学院

2009年12月21日

Definition of Community



In communities, the link density is comparatively high and among communities the link density is comparatively low.

Many complex systems can be represented as networks and separating a network into communities could simplify the functional analysis considerably.

Why We Pay Attention to Significance of Community Structure

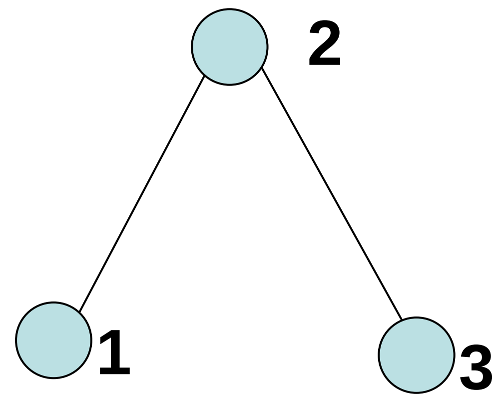
1. Some networks contain error links.

2. Many detecting algorithms have random factors.

So we should evaluate the sensitivity of community structure

Quantitive Definition of Multi-Communities Structure

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$



$$0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_n$$

$$L = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

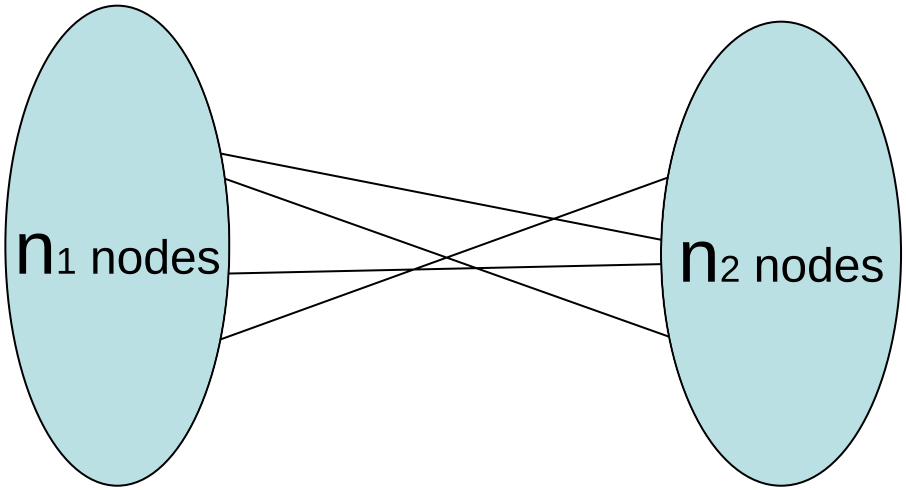
v_i is the corresponding orthogonal and normalized eigenvector

$$span\{v_1, v_2, \cdots, v_n\} = R^n$$

$$\forall s \in R^n, s = a_1 v_1 + \cdots + a_n v_n$$

$$s = [1, -1, -1, 1, \cdots]$$

Bi-community



denotes the bi-community structure

Objective Function

$$Min Z = s^T L s = \sum a_i^2 \lambda_i$$

$$st. \quad a_1^2 + a_2^2 + \cdots + a_n^2 = n$$

$$\text{where } a_i = v_i^T s$$

Minimum number of links between the two communities.

$$Min Z \approx Max \hat{Z} \; a_2^2 \lambda_2$$ simple

For multi-communities structure, we define s_j as community j's community structure vector. If node i belongs to community j we let $S_{ji} = 1$ otherwise $S_{ji} = -1$

The optimal partition is defined as:

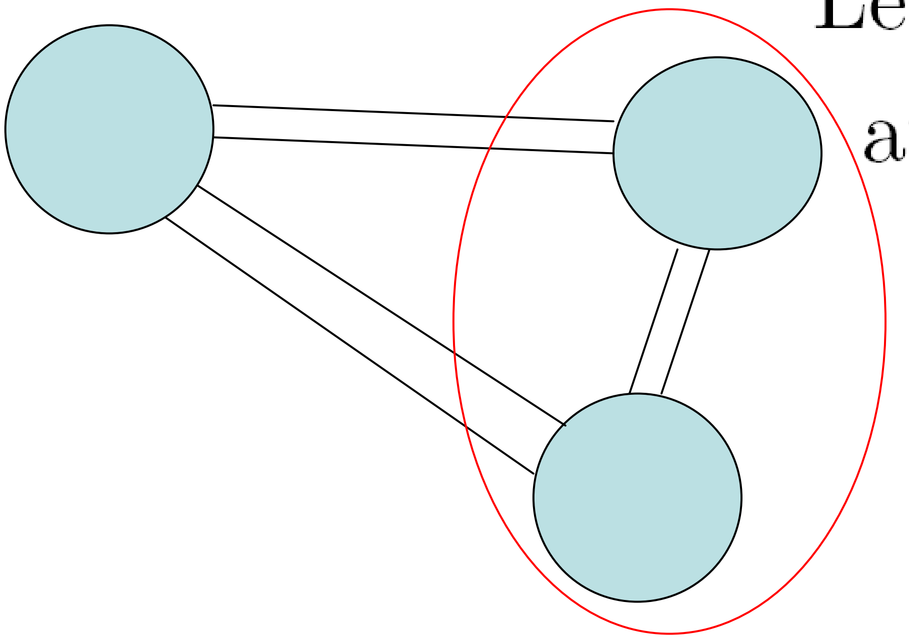
$$MinZ = \sum_{i=1}^{c} S_i^T L S_i$$

Let $S = (S_1^T, S_2^T, \cdots, S_c^T)^T$

and $\hat{L} = diag(L, L, \cdots, L)$

thus we have

$$MinZ = S^T \hat{L} S$$



We can obtain all orthogonal and normalized eigenvectors u_q and the corresponding eigenvalues τ_q of $\hat{L}$, where $q = 1, 2, \cdots, n \times c$. Obviously, each eigenvalue of L will be $\hat{L}$'s eigenvalue and each them will repeat c time. Without losing any generality, we can let $\tau_{ci-c+j} = \lambda_i, j = 1, 2, \cdots, c$. Let SU as the eigenvectors set of the eigenvalues of $\lambda_2, \lambda_3, \cdots, \lambda_c$ of matrix $\hat{L}$. SU can be written as $SU = \{(v_2^T, 0 \cdots, 0), \cdots, (v_c^T, 0, \cdots, 0), \cdots, (0, 0, \cdots, v_c^T)\}$, where each 0 denote a n-dimensional vector and SU has $c \times (c-1)$ elements. We can expand SU as a space SSU in which each point is the liner combination of the elements in set SU. As the bi-partition problem, the multi-partition problem can be written as:

$$MinZ = \sum_{q=1}^{n \times c} b_q^2 \tau_q \approx Max\hat{Z} = \sum_{u_q \in SSU} b_q^2 \tau_q \approx \bar{\lambda} \sum_{u_q \in SSU} b_q^2$$

where $b_q = S^T u_q$ and $\overline{\lambda}$ is the average value of λ_2 to λ_c. $\sum_{u_q \in SSU} b_q^2$ denotes the length of vector S projection in space SSU. Obviously, the longer of the projection, the S is more optimal.

Robustness of Space SSU

The space SSU is expanded by the simple combination of $v_2, v_3, \cdots, v_c$, therefore, the robustness of space SSU is the robustness of the eigenvalues $\lambda_2, \lambda_3, \cdots, \lambda_c$ and eigenvectors $v_2, v_3, \cdots, v_c$.

$$(\delta L + L)(\delta v_i + v_i) = (\delta \lambda_i + \lambda_i)(\delta v_i + v_i)$$

deleting the second-order small quantities we have

$$\delta L v_i + L \delta v_i = \lambda_i \delta v_i + \delta \lambda_i v_i$$

after some deductions we obtain:

$$\delta \lambda_i = \frac{v_i^T \delta L v_i}{v_i^T v_i}$$

$$\delta v_i = \sum_{j=1}^{n} h_{ij} v_j$$

$$h_{ij} = \frac{v_j^T \delta L v_i}{v_j^T v_j (\lambda_i - \lambda_j)}, (i \neq j)$$

$$|\delta\lambda_i| \leq \|\delta L\|$$

Which implies that eigenvalues are always not related to robustness of community structure for unweighted network.

Without losing any generality, for any $i \neq 1$ we can let $a_{i1} = a_{ii} = 0$.

Then the comparative error of v_i can be denoted as

$$\frac{|\delta v_i|}{|v_i|} \leq \| \delta L \| \sum_{j \neq i, j=2}^{n} \frac{1}{|\lambda_i - \lambda_j|}$$

$\| \delta L \|$ is the perturbation strength

$\sum_{j \neq i, j=2}^{n} \frac{1}{|\lambda_i - \lambda_j|}$ is the robustness

Integrating the robustness of λ_2 to λ_c

we define R as the robustness of space SSU

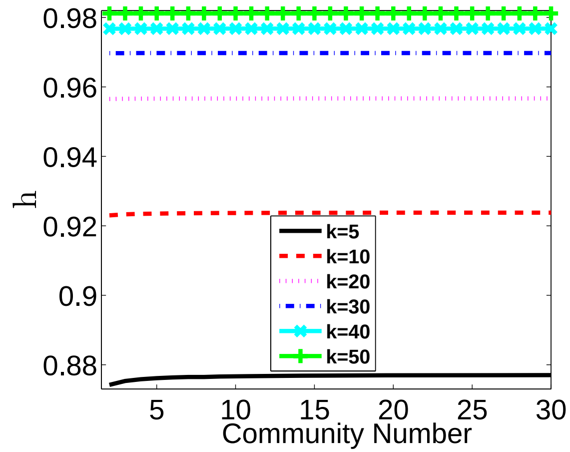
$$R = \sum_{j=c+1}^{n} \frac{1}{|\bar{\lambda} - \lambda_j|}$$

Index of Significance

the most significant community structure

$$MinR = \sum_{i=c+1}^{n} \frac{1}{\lambda_{c+1} - \bar{\lambda}}$$

$$\sum_{i=1}^{n} \lambda_i = nk$$



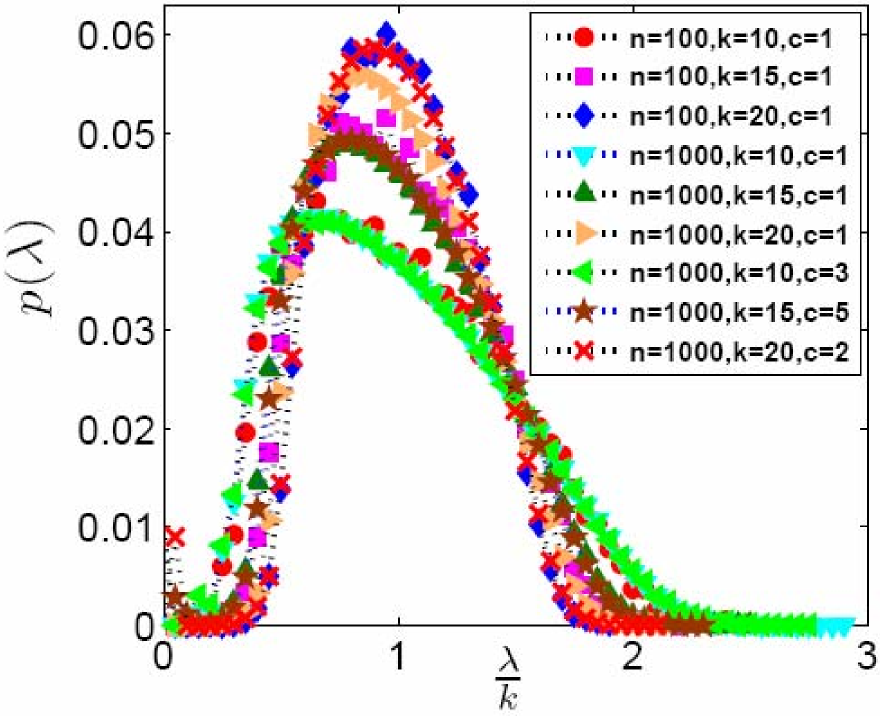
Employ lagrange multiplier method

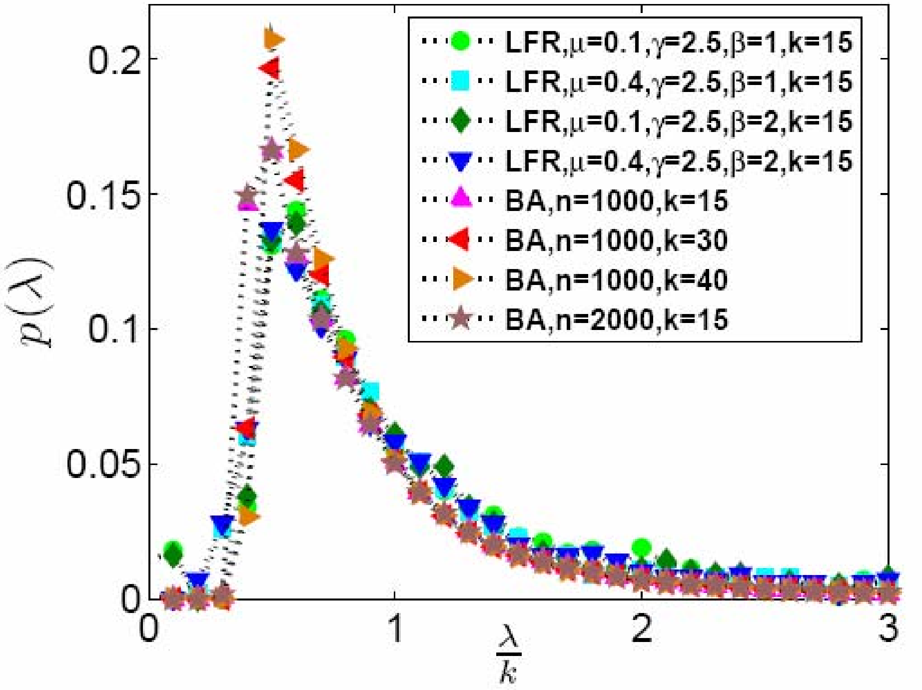
R will achieve it's global minimum

$$R = \frac{(n-c)^2}{nk} \approx \frac{n}{k}$$

when $\bar{\lambda} = 0, \lambda_{c+1}, \lambda_{c+2}, \cdots, \lambda_n = \frac{nk}{n-c},$

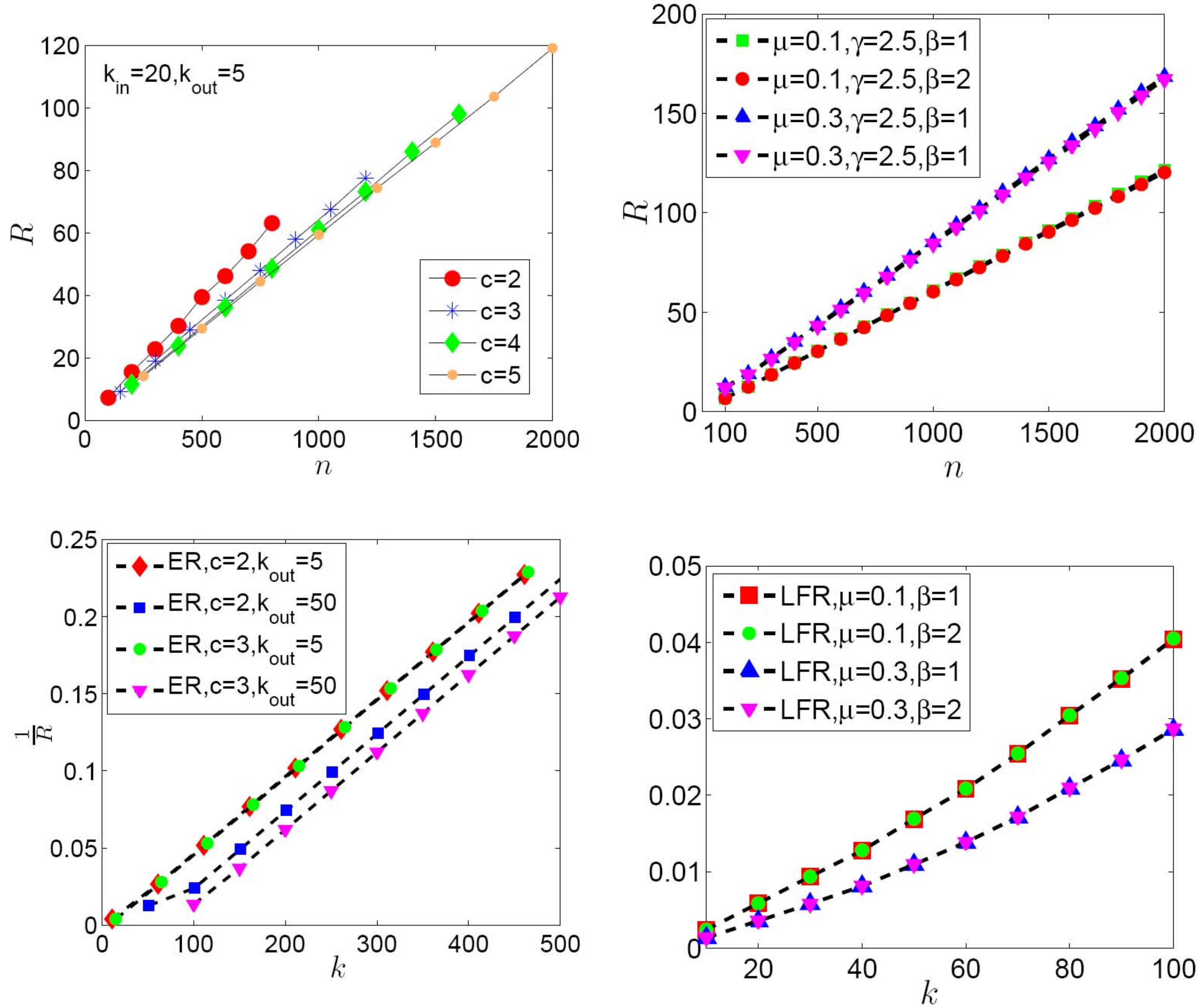
$\bar{\lambda} = 0$ implies that there are no any connections among communities.





R holds $R = h\frac{n}{k}$ $R \propto n, \frac{1}{R} \propto k$

Define significance index as $H = \frac{1}{h}$ $H \in (0,1)$



k_in=20,k_out=5
c=2
c=3
c=4
c=5
R
n
μ=0.1,γ=2.5,β=1
μ=0.1,γ=2.5,β=2
μ=0.3,γ=2.5,β=1
μ=0.3,γ=2.5,β=1
R
n
ER,c=2,k_out=5
ER,c=2,k_out=50
ER,c=3,k_out=5
ER,c=3,k_out=50
1/R
k
LFR,μ=0.1,β=1
LFR,μ=0.1,β=2
LFR,μ=0.3,β=1
LFR,μ=0.3,β=2
k

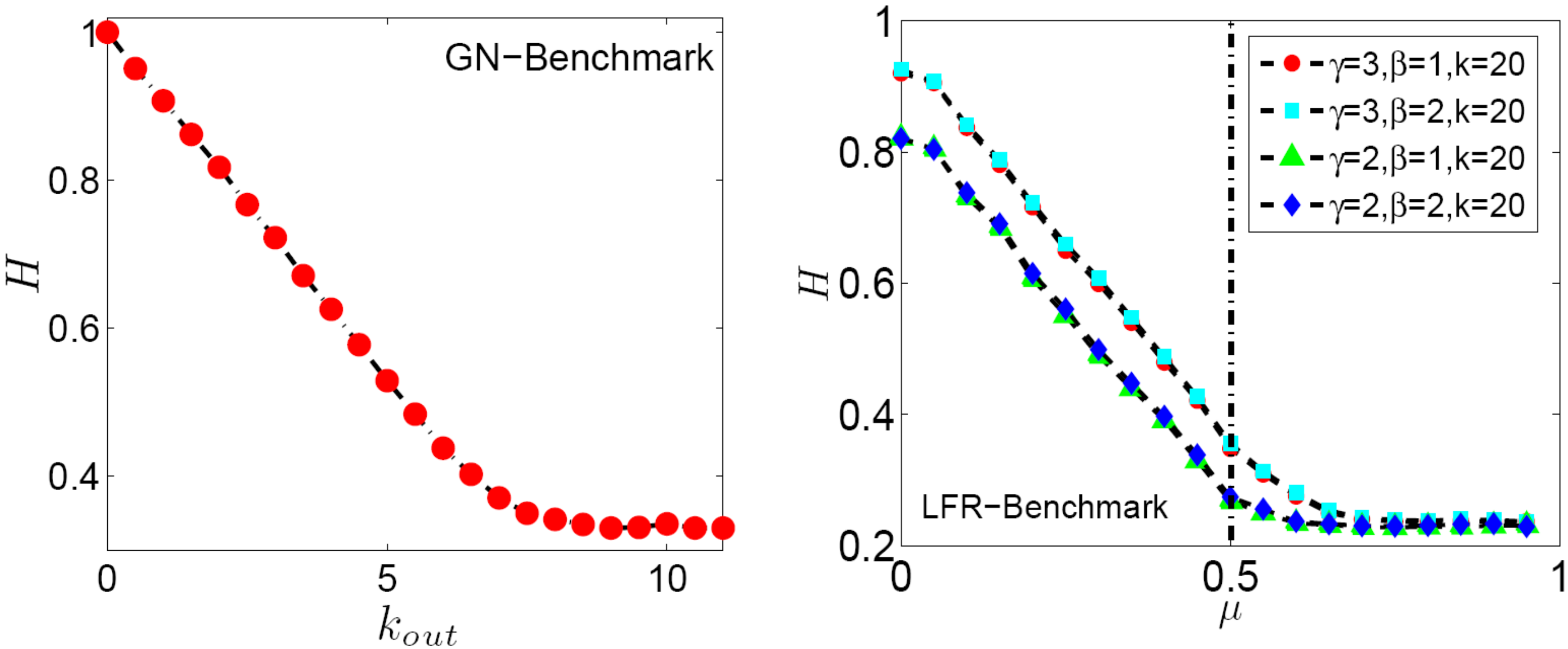
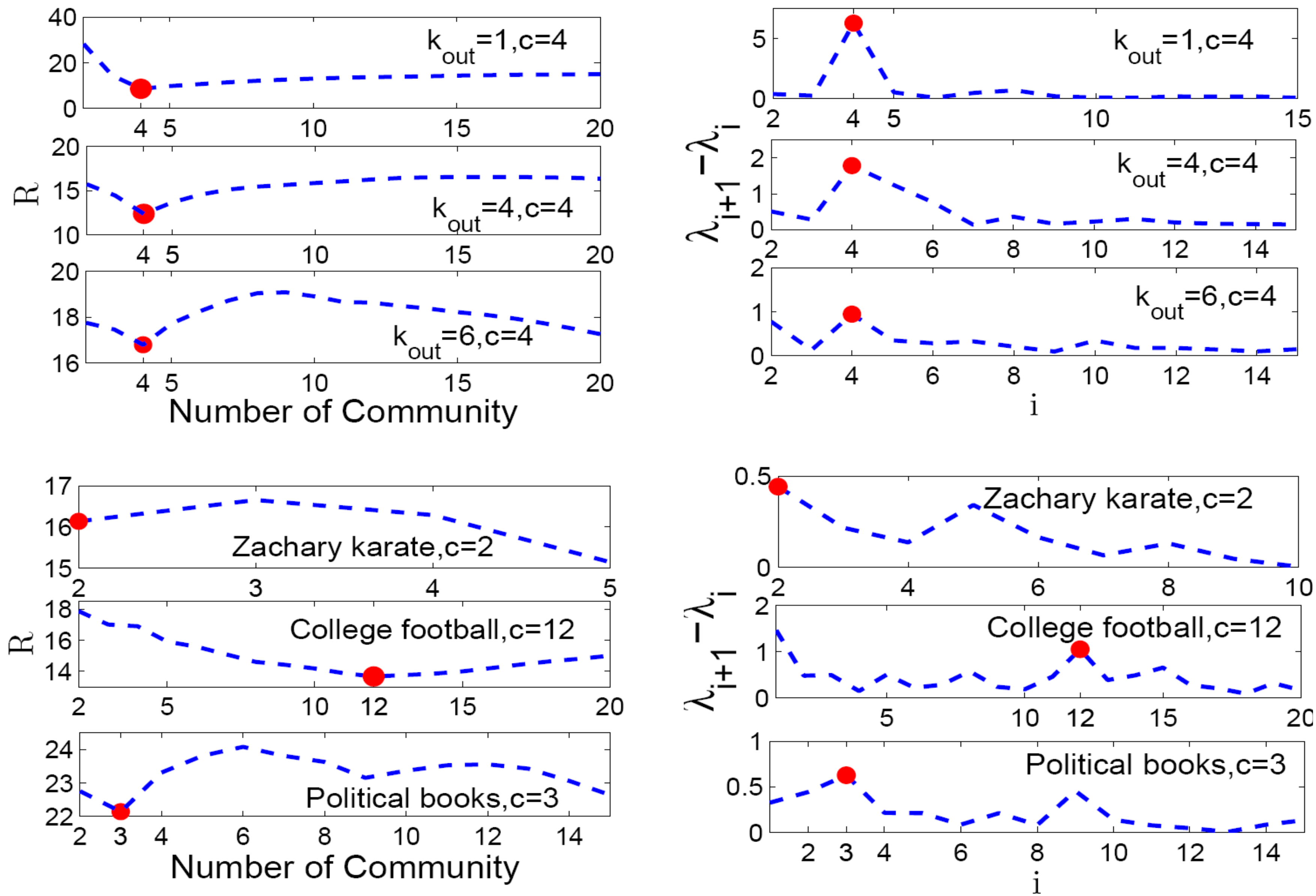
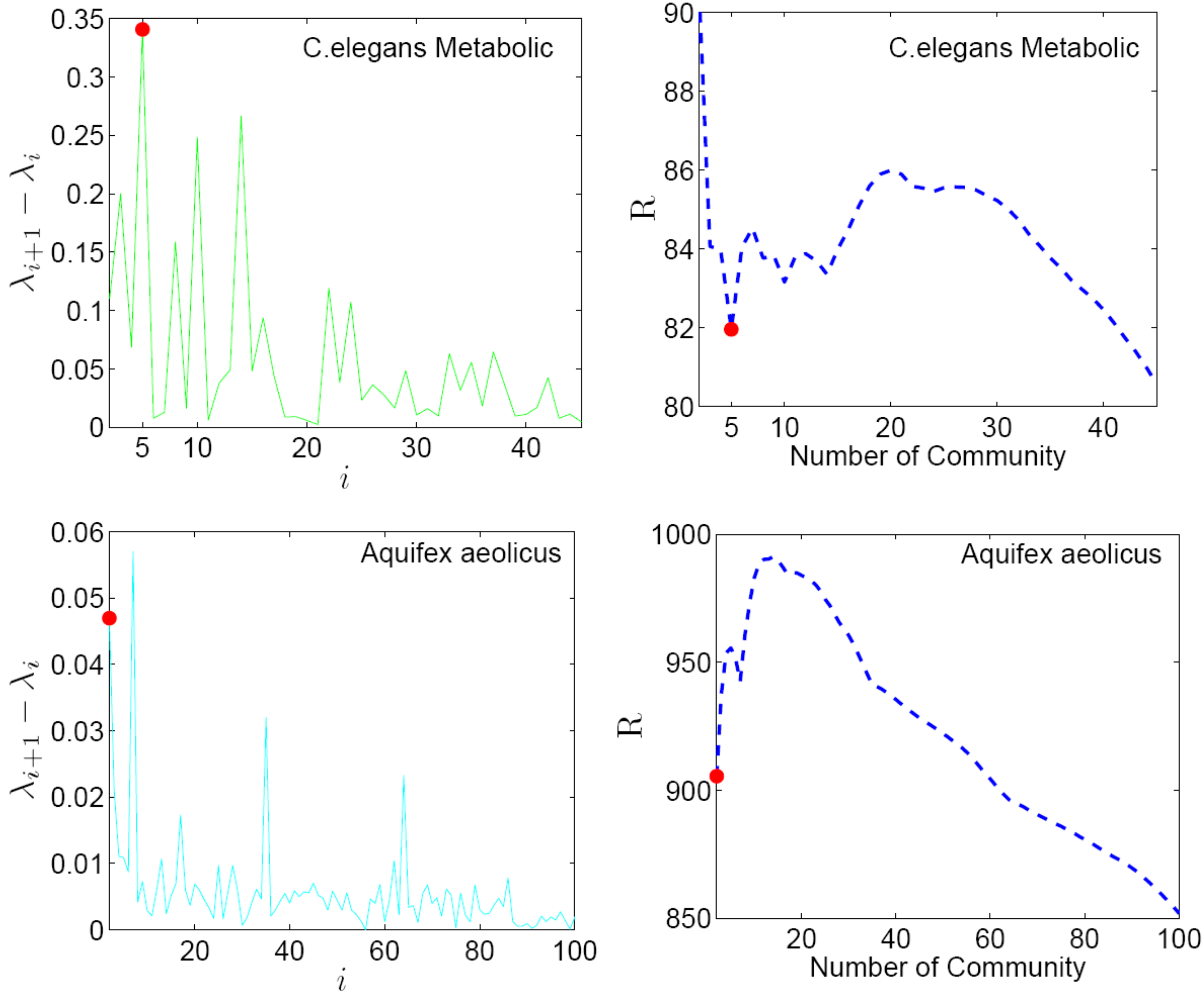


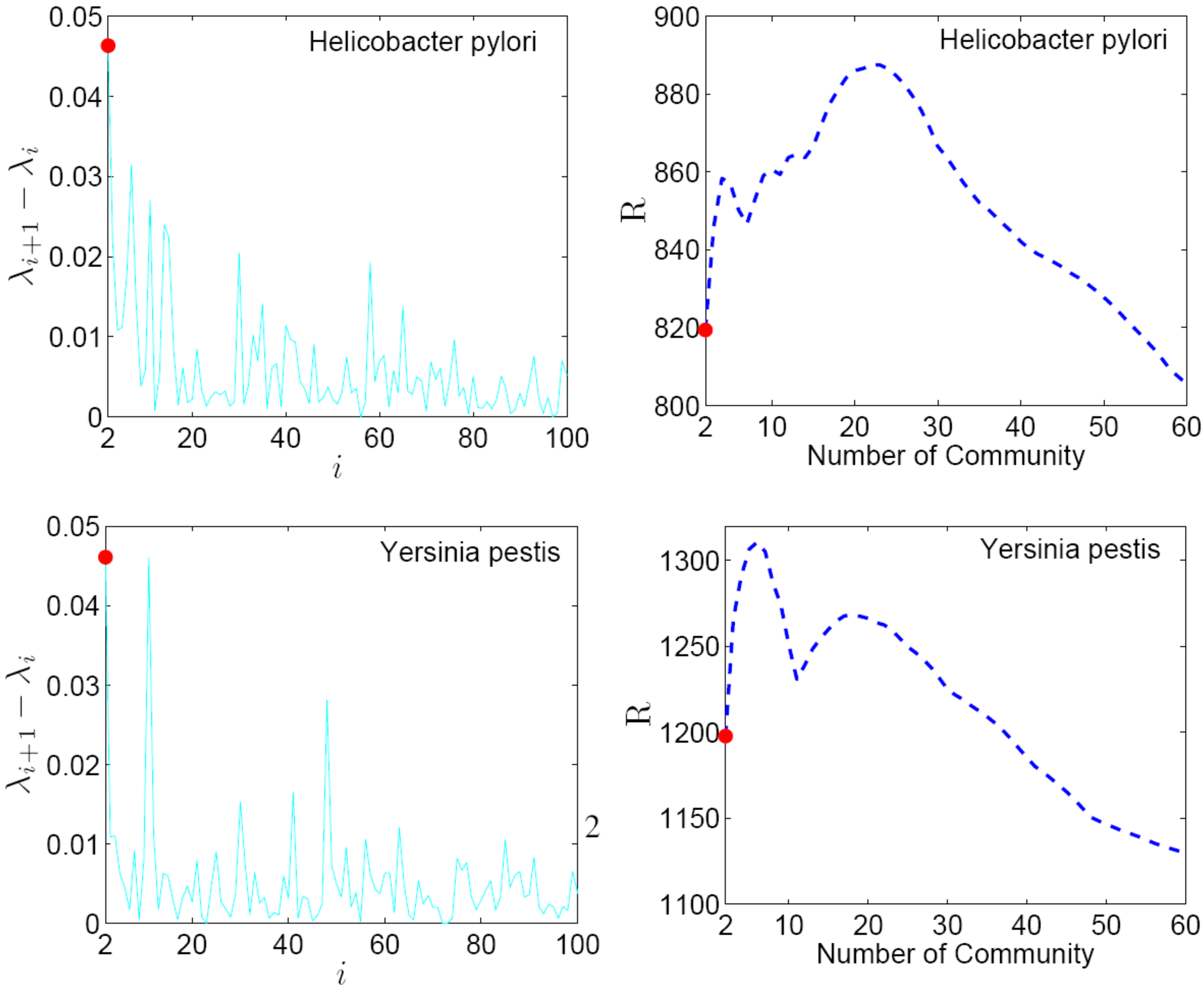
FIG. 3: The performance of H index in both GN-benchmark and LFR-benchmark. In GN-benchmark, we can see that H decrease with increasing of k_{out}. When the community structure is very clear H close to 1 very much, and the network close to no community structure network H close to 0.3 which implies that for a given network when H is less than 0.3 it is not safe to say there exit significant community structure. In LFR-benchmark, the average degree $k = 20$, maximum degree is 50 and $p(k) \propto k^{\gamma}$. Maximum and minimum community sizes are 50 and 20 respectively, more over $p(m) \propto m^{\beta}$ where, m denotes the community size. We can see that with the increase of mix parameter μ, the H index decrease. When $\mu \geq 0.5$ (no significant community) H is near 0.3 which is similar with GN-benchmark.

How to obtain the optimal community number c

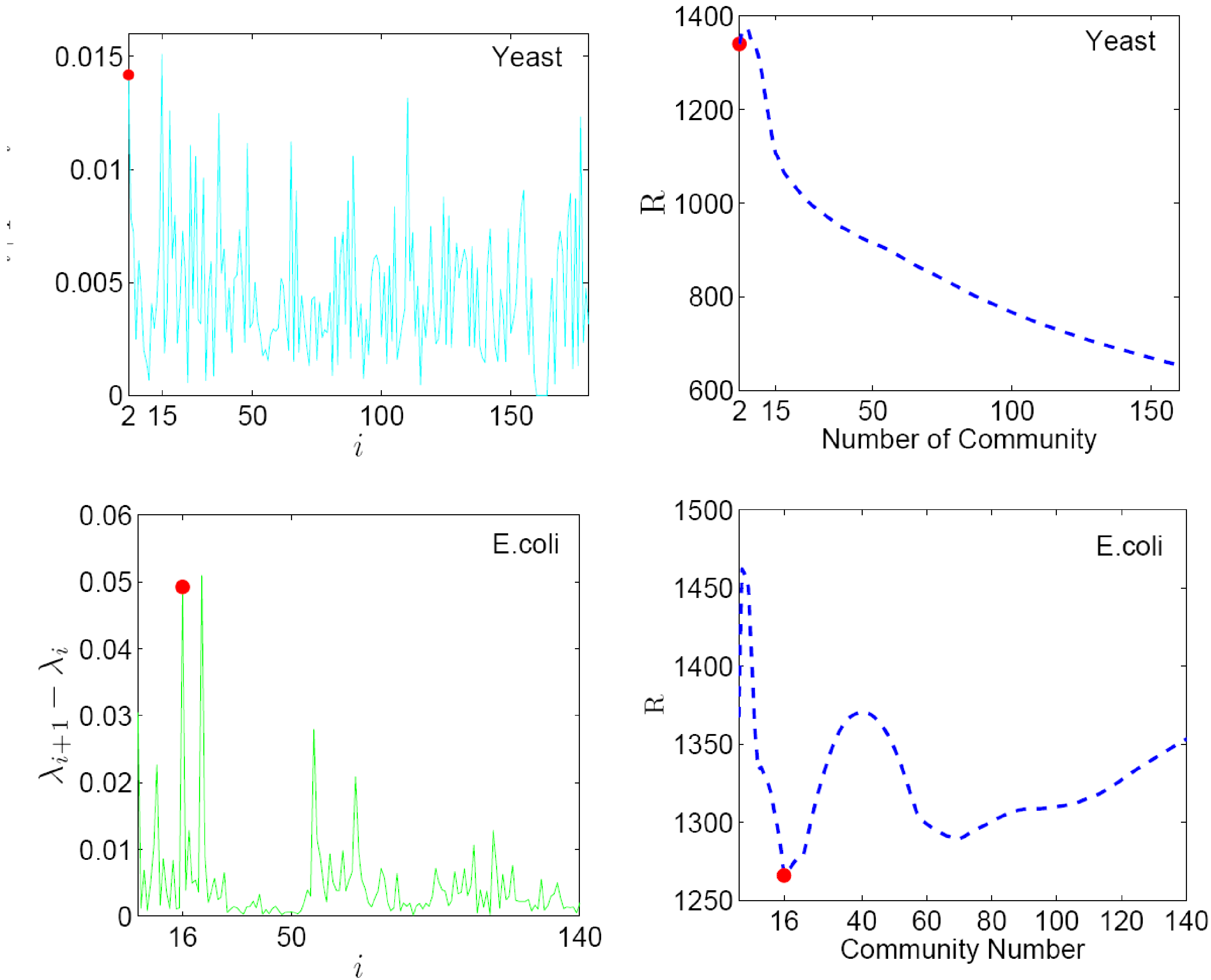


network	n	edge number	$\hat{R}$	H	type
E.coli	1442	5873	0.14	0.14	protein
Yeast	1458	1993	0.14	0.40	
H.Sapiens	693	982	0.21	0.21	
Celegans metabolic	453	4596	0.19	0.62	metabolic
Aquifex aeolicus	1437	3272	0.19	0.36	
Helicobacter pylori	1341	3087	0.19	0.36	
Yersinia pestis	1922	4383	0.18	0.36	
Celegans neural	297	2148	0.24	0.52	neural
Santa Fe scientists	260	612	0.14	0.22	social
Zachary karate	34	78	0.27	0.46	
Dolphin	62	159	0.27	0.42	
College football	115	613	0.38	0.79	
Jazz	198	2742	0.42	0.47	
Email	1133	5452	0.22	0.42	
Political blogs	1222	16716	0.29	0.22	
Political books	105	441	0.34	0.56	

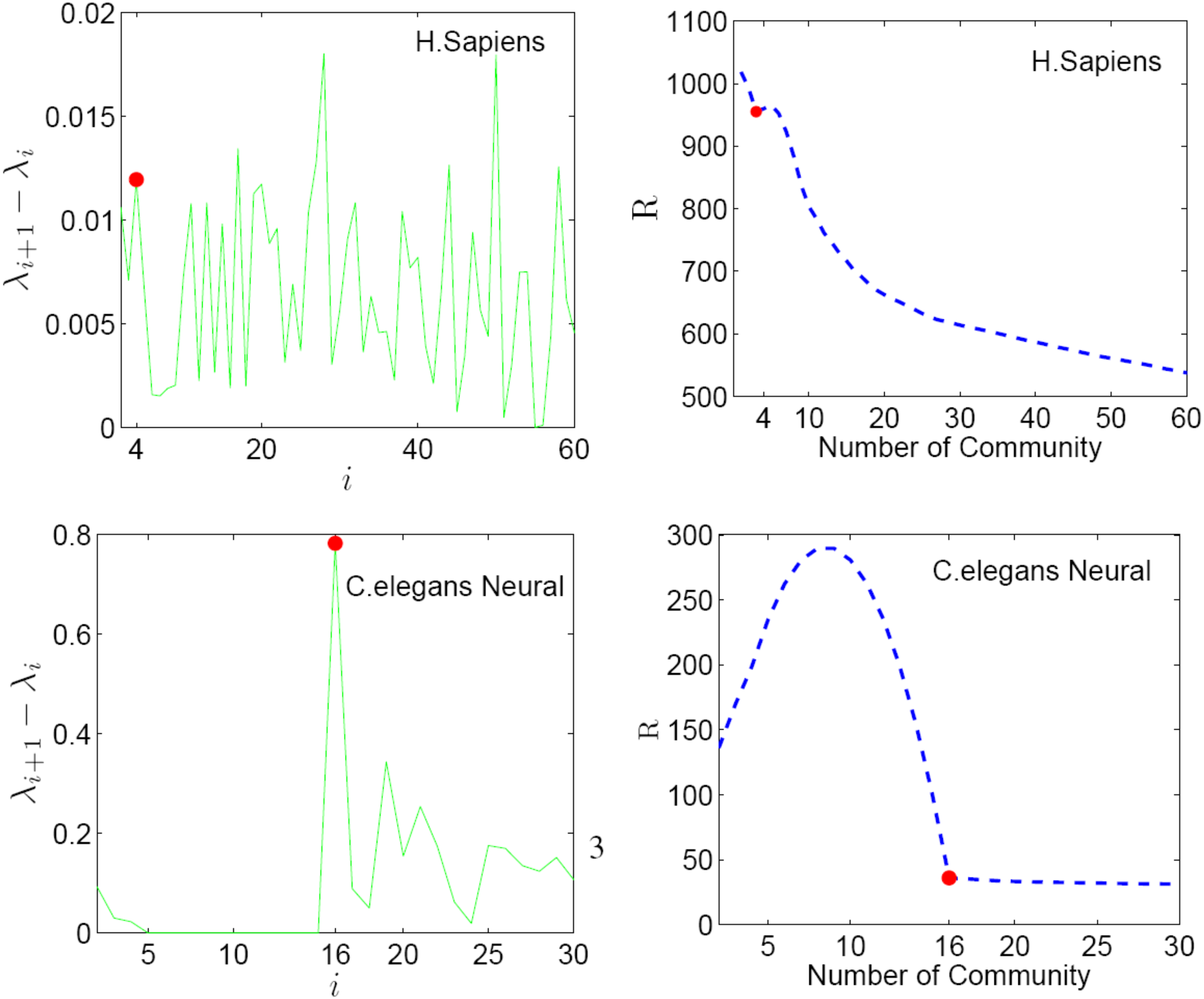

C.elegans Metabolic
$\lambda_{i+1} - \lambda_i$
i
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
20
30
40
C.elegans Metabolic
R
Number of Community
90
88
86
84
82
80
Aquifex aeolicus
$\lambda_{i+1} - \lambda_i$
i
0.06
0.05
0.04
0.03
0.02
0.01
0
20
40
60
80
100
Aquifex aeolicus
R
Number of Community
1000
950
900
850

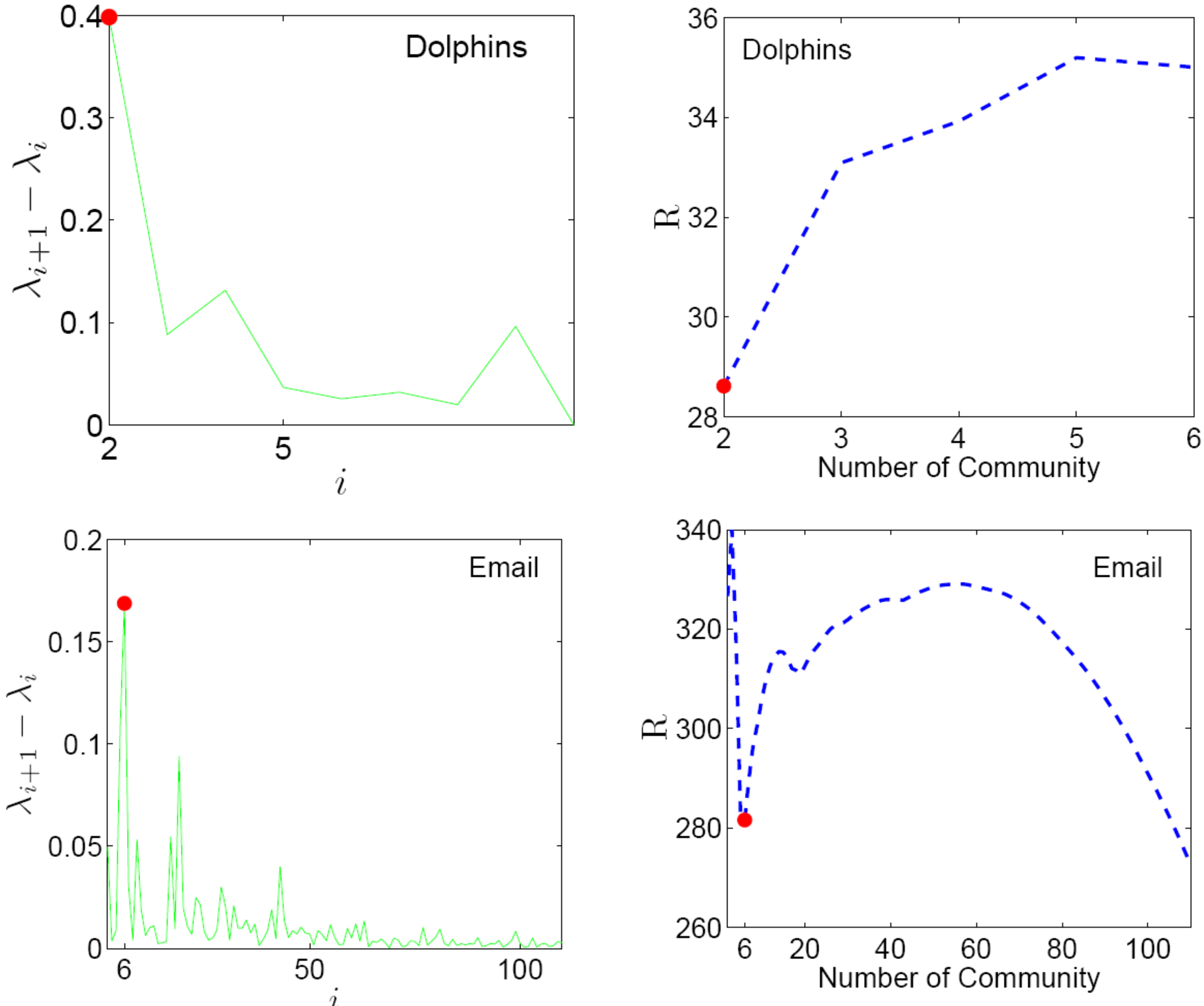


Helicobacter pylori
$\lambda_{i+1} - \lambda_i$
0.05
0.04
0.03
0.02
0.01
0
2
20
40
60
80
100
i
Helicobacter pylori
R
900
880
860
840
820
800
2
10
20
30
40
50
60
Number of Community
Yersinia pestis
$\lambda_{i+1} - \lambda_i$
0.05
0.04
0.03
0.02
0.01
0
2
20
40
60
80
100
i
2
Yersinia pestis
R
1300
1250
1200
1150
1100
2
10
20
30
40
50
60
Number of Community

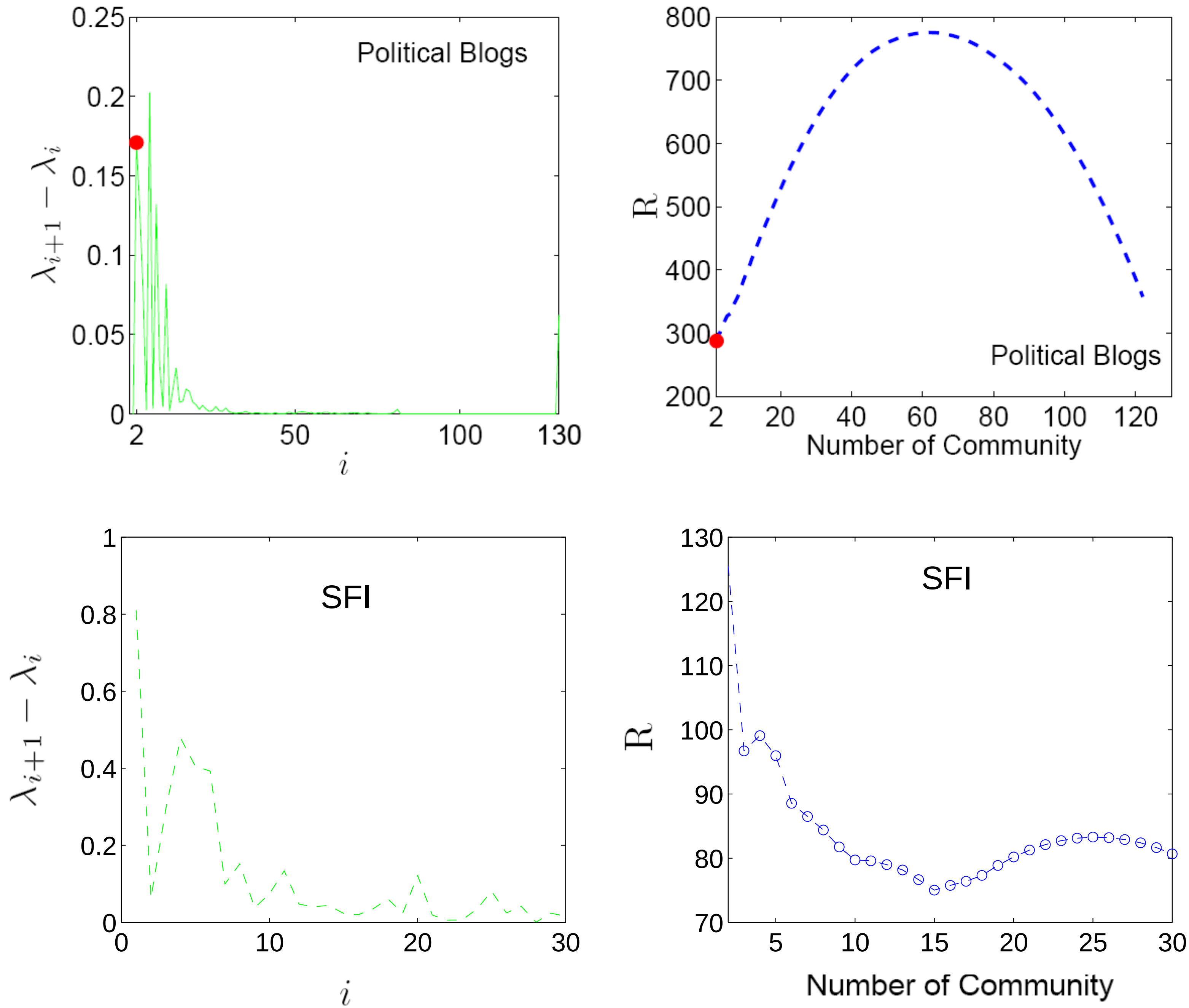


0.015
0.01
0.005
0
2 15 50 100 150
i
Yeast
1400
1200
1000
800
600
R
2 15 50 100 150
Number of Community
Yeast
0.06
0.05
0.04
0.03
0.02
0.01
0
$\lambda_{i+1} - \lambda_i$
16 50 140
i
E.coli
1500
1450
1400
1350
1300
1250
R
16 40 60 80 100 120 140
Community Number
E.coli


H.Sapiens
$\lambda_{i+1} - \lambda_i$
0.02
0.015
0.01
0.005
0
4
20
40
60
i
H.Sapiens
R
1100
1000
900
800
700
600
500
4
10
20
30
40
50
60
Number of Community
C.elegans Neural
$\lambda_{i+1} - \lambda_i$
0.8
0.6
0.4
0.2
0
5
10
16
20
25
30
i
3
C.elegans Neural
R
300
250
200
150
100
50
0
5
10
16
20
25
30
Number of Community



Dolphins
0.4
0.3
0.2
0.1
0
2
5
$\lambda_{i+1} - \lambda_i$
i
Dolphins
36
34
32
30
28
2
3
4
5
6
R
Number of Community
Email
0.2
0.15
0.1
0.05
0
6
50
100
$\lambda_{i+1} - \lambda_i$
i
Email
340
320
300
280
260
6
20
40
60
80
100
R
Number of Community



Political Blogs
0.25
0.2
0.15
0.1
0.05
0
2
50
100
130
$\lambda_{i+1} - \lambda_i$
i
800
700
600
500
400
300
200
2
20
40
60
80
100
120
R
Number of Community
Political Blogs
SFI
1
0.8
0.6
0.4
0.2
0
0
10
20
30
$\lambda_{i+1} - \lambda_i$
i
SFI
130
120
110
100
90
80
70
5
10
15
20
25
30
R
Number of Community

Thank you!