Problem Set 4

CSCI 5352

Ben Braun

3.17.2025

Problem 1

e_{rr} is the diagonal for region r

a_r is the row sum for region r

$Q_r = e_{rr} - a_r^2$

After transforming $\mathcal{M}$ into an undirected matrix, we can directly calculate e_{rr}, a_r, and Q_r from the matrix:

$\mathcal{M}$	Northeast	Midwest	South	West	Canada	e_{rr}	a_r
Northeast	0.238	0.084	0.099	0.104	0.028	0.238	0.553
Midwest	0.084	0.134	0.088	0.059	0.016	0.134	0.381
South	0.099	0.088	0.166	0.067	0.011	0.166	0.262
West	0.104	0.059	0.067	0.146	0.016	0.146	0.392
Canada	0.028	0.016	0.011	0.016	0.170	0.170	0.241

$Q = \sum_r Q_r = 0.122.$

The overall assortativity of the system is somewhat low, with a weak community structure. Faculty are slightly more likely to remain within their region after their PhD. According to the Q_r values, the South and Canada have the strongest community structures and drive the overall assortativity of the system, while the other regions are mildly disassortative and prefer to leave their region.

Problem 2

Part A

For the divided network, we can find the values of e_r and a_r in terms of n and g.

Node Group	e_r	a_r
1	$\frac{g-1}{n-1}$	$\frac{2g-1}{2(n-1)}$
2	$\frac{n-g-1}{n-1}$	$\frac{2(n-g)-1}{2(n-1)}$

Similar to problem 1, we can then use the equation for Q in terms of e_r and a_r to find the modularity in terms of n and g.

Below is my derivation, with work shown (slightly condensed for brevity):

$$\begin{aligned}
Q &= \sum_r (e_r - a_r^2) \\
&= [\frac{g-1}{n-1} - (\frac{2g-1}{2(n-1)})^2] + [\frac{n-g-1}{n-1} - (\frac{2(n-g)-1}{2(n-1)})^2] \\
&= \frac{4(g-1)(n-1) - 4g^2 + 4g - 1}{4(n-1)^2} + \frac{4(n-g-1)(n-1) - 4(n-g)^2 + 4(n-g) - 1}{4(n-1)^2} \\
&= \frac{4gn - 4n - 4g^2 + 3 - 4n + 4ng - 4g^2 + 3}{4(n-1)^2} \\
&= \frac{-8g^2 + 8ng - 8n + 6}{4(n-1)^2} \\
&= \frac{-4g^2 + 4ng - 4n + 3}{2(n-1)^2}
\end{aligned}$$

Part B

A split exactly down the middle gives us $g = \frac{n}{2}$.

To find the optimal modularity, we can treat Q as a function of g:

$$Q(g) = \frac{-4g^2 + 4ng - 4n + 3}{2(n-1)^2}$$

We can then take the derivative of Q with respect to g:

$$\begin{aligned}
\frac{dQ}{dg} &= \frac{-8g + 4n}{2(n-1)^2} \\
&= \frac{4n - 8g}{2(n-1)^2} \\
&= \frac{2n - 4g}{(n-1)^2}
\end{aligned}$$

Setting $\frac{dQ}{dg} = 0$ gives us the critical point:

$$\begin{aligned}
2n - 4g &= 0 \\
g &= \frac{n}{2}
\end{aligned}$$

We can confirm this is a maximum by taking the second derivative:

$$\frac{d^2Q}{dg^2} = \frac{-4}{(n-1)^2} < 0$$

Since the second derivative is negative, the function is concave and $g = \frac{n}{2}$ is a maximum. Therefore, the optimal modularity is achieved when the network is split exactly down the middle.

Note that this applies specifically to even n because odd n would make a perfect split down the middle impossible.

Problem 3

Part A

We find e_{rs} by counting up the edges between groups r and s. We then count the total possible edges between r and s to find n_{rs}. Their ratio gives us the maximum likelihood mixing matrix:

e_{rs}/n_{rs}	Orange	Teal
Orange	$4/10$	$2/20$
Teal	------	$4/6$

We can then use the formula for the log likelihood, plugging in the values of e_{rs} and n_{rs} from the table:

$$\begin{aligned} \ln \mathcal{L} &= \sum_{r,s} e_{rs} \ln \frac{e_{rs}}{n_{rs}} + (n_{rs} - e_{rs}) \ln(\frac{n_{rs} - e_{rs}}{n_{rs}}) \\ &= (4 \ln \frac{4}{10} + 6 \ln \frac{6}{10}) + (2 \ln \frac{2}{20} + 18 \ln \frac{18}{20}) + (4 \ln \frac{4}{6} + 2 \ln \frac{2}{6}) \\ &\approx -17.0509 \end{aligned}$$

Part B

For the DC-SBM mixing matrix, we need to count the stubs rather than the edges within and between groups. Then we calculate κ_r for each group by summing the stubs for that group.

ω_{rs}	Orange	Teal	κ_r
Orange	8	2	10
Teal	2	8	10

We then apply the given formula for log likelihood:

$$\begin{aligned}\ln \mathcal{L} &= \sum_{r,s} \omega_{rs} \ln \frac{\omega_{rs}}{\kappa_r \kappa_s} \\ &= (8 \ln \frac{8}{10 \times 10}) + (2 \ln \frac{2}{10 \times 10}) + (2 \ln \frac{2}{10 \times 10}) + (8 \ln \frac{8}{10 \times 10}) \\ &\approx -56.0598\end{aligned}$$

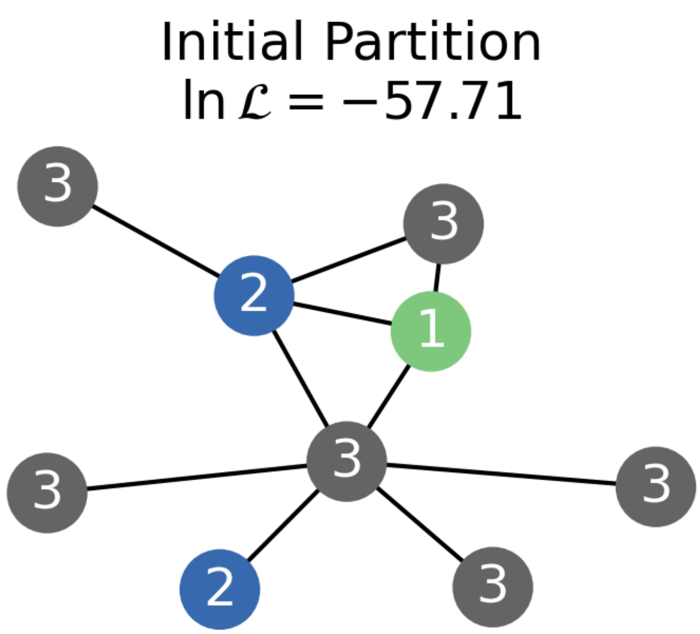
Part C

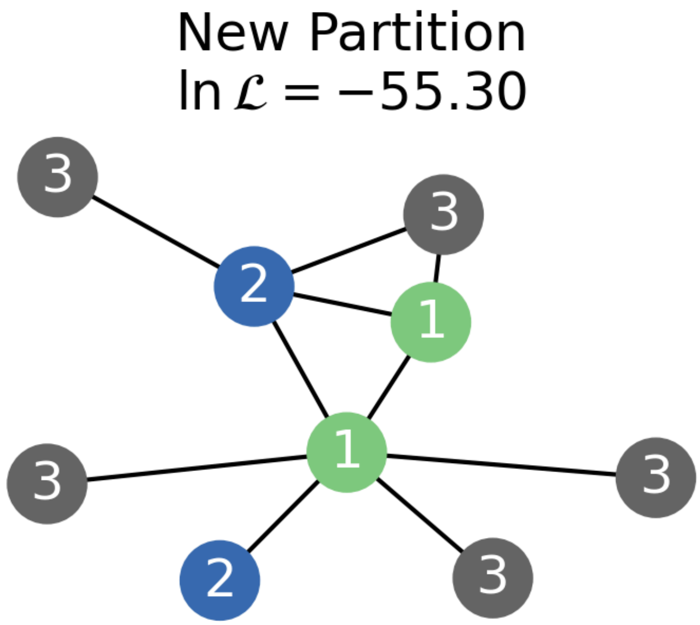
The SBM is $\exp(56.0598 - 17.0509) \approx 8 \times 10^{16}$ times more likely than the DC-SBM model to generate the observed network.

Problem 4

Part A

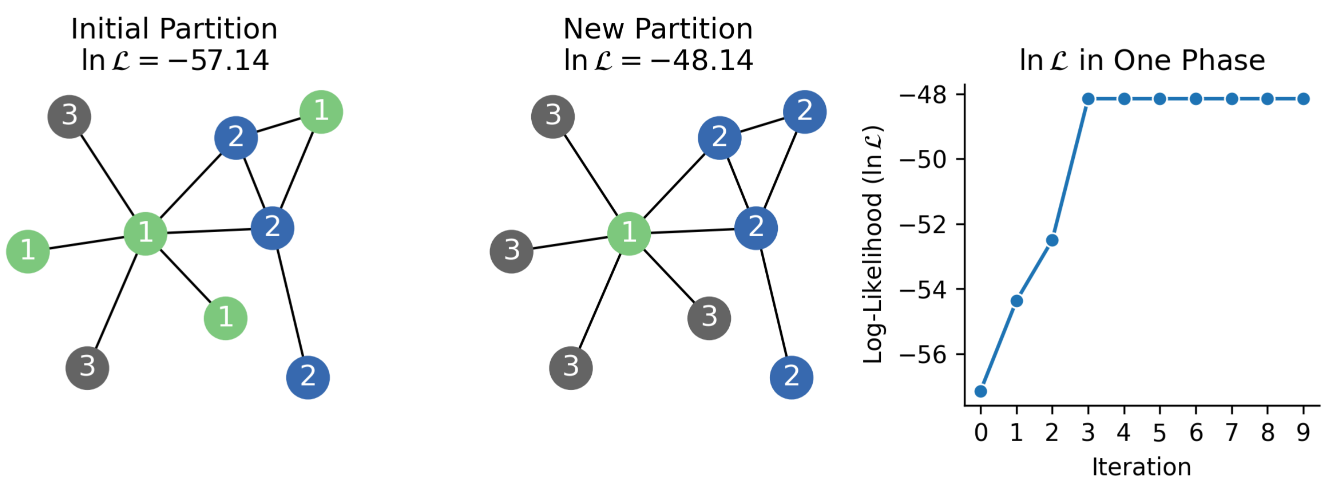
One Move:





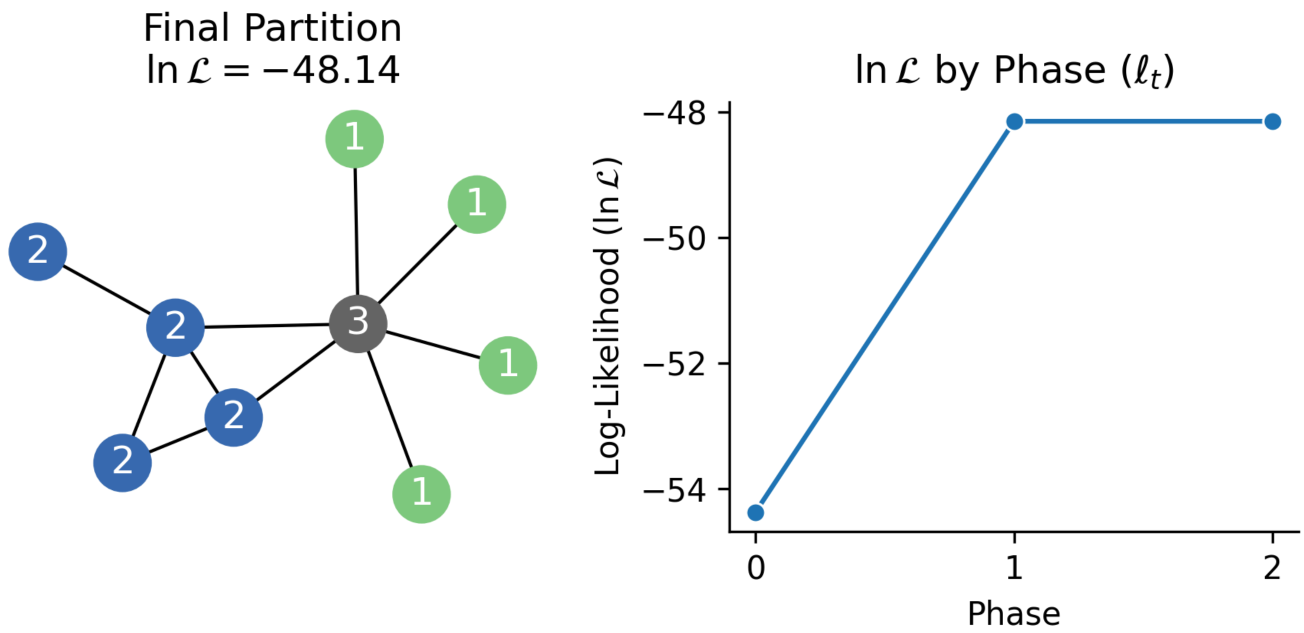
Part B

One Phase:



Part C

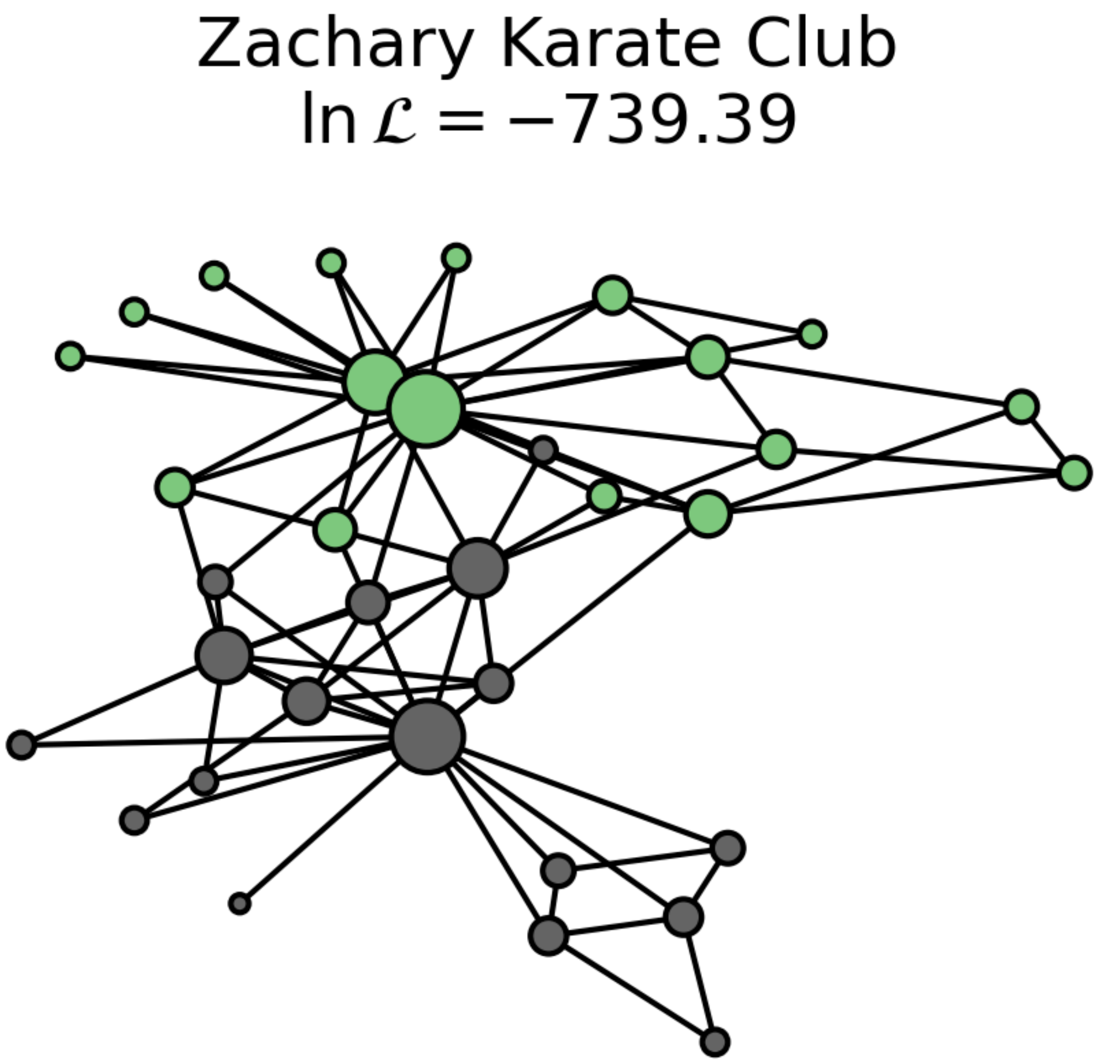
One Full DC-SBM:



ω	**Group 1**	**Group 2**	**Group 3**	κ
Group 1	0	4	0	4
Group 2	4	0	2	6
Group 3	0	2	8	10

For this relatively small graph, we achieve the maximum log-likelihood very fast - usually within 1-2 phases, as shown by the ℓ_t plot. My guess is that this is mostly a function of how few nodes need to be changed. The partition itself is pretty good, separating out groups 2 and 3 well. However, this graph is probably better suited to a 2-group partition, so groups 1 and 3 end up being a bit awkward here.

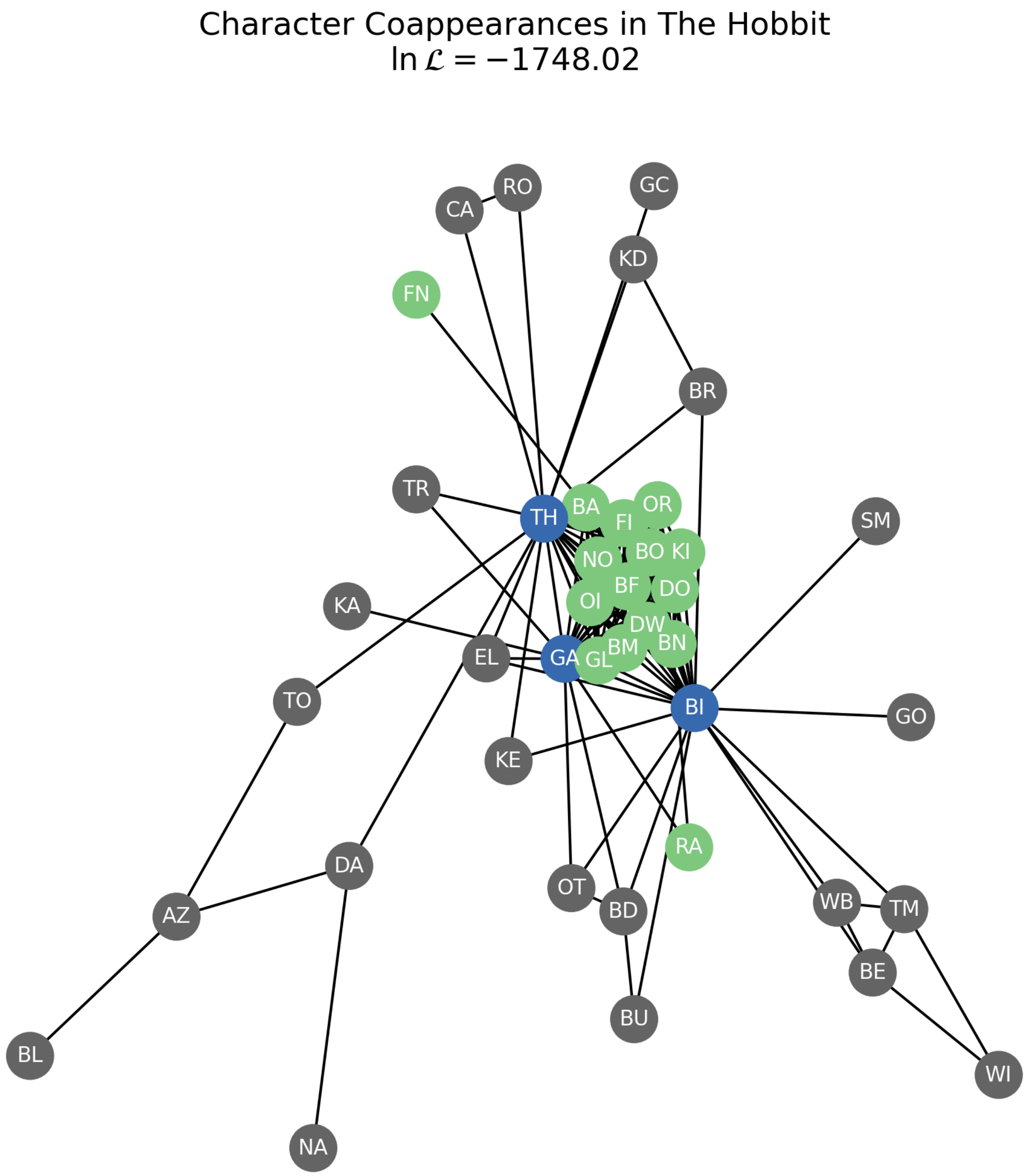
Part D



This algorithm worked very well on the Karate Club example, taking only around 20-25 runs to reliably produce the optimal partition. This is a lot fewer than I expected, although I suspect that larger graphs or graphs with more unusual structure may need more runs. Needing only a few runs suggests that the algorithm can converge on the optimal partition from a wide variety of random partitions.

Part E

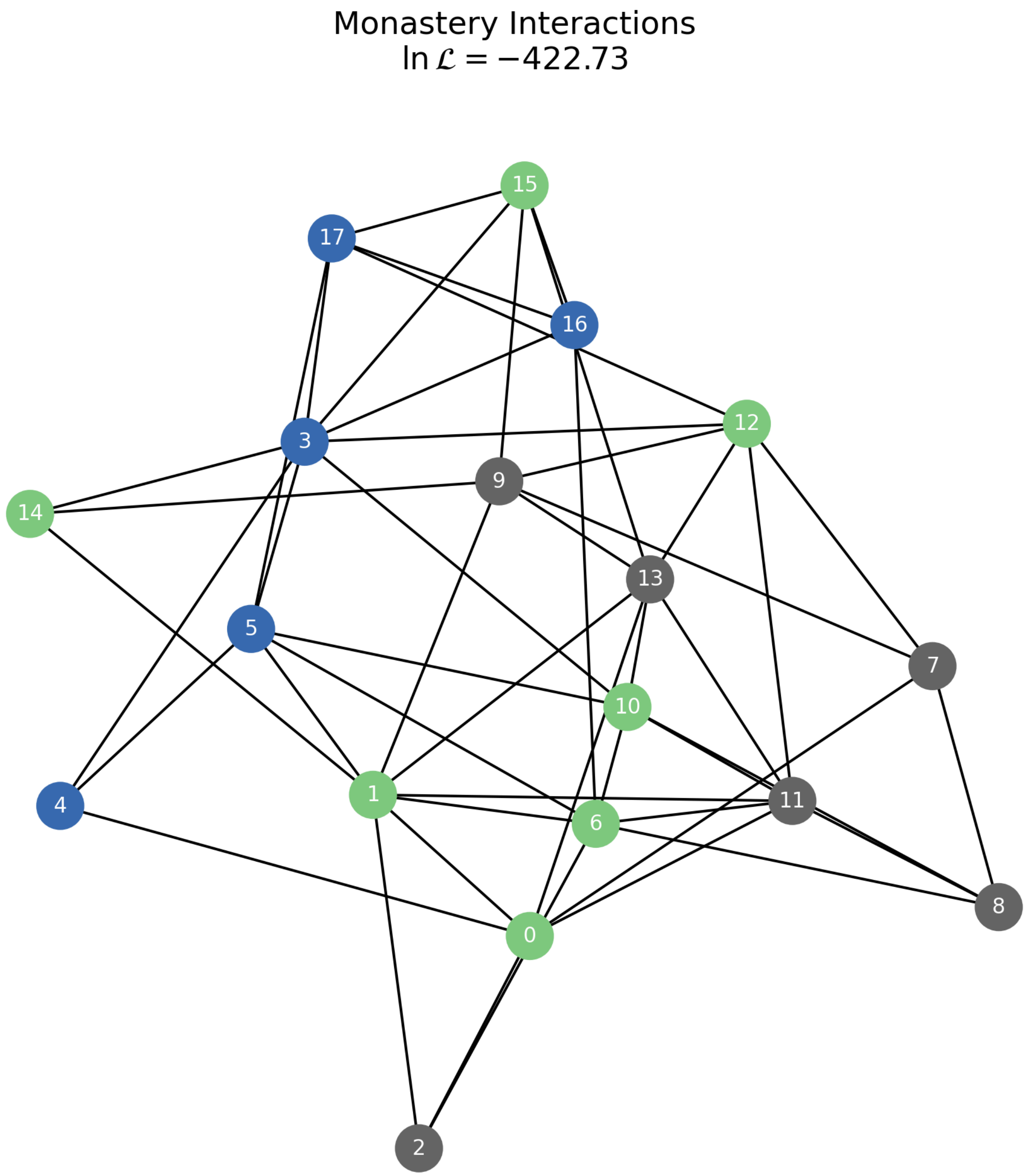
Coappearances of Characters in ***The Hobbit*** by J.R.R. Tolkien:



ω	Group 1	Group 2	Group 3	κ
Group 1	156	0	40	196
Group 2	0	26	26	52
Group 3	40	26	6	72

This network clearly has a core-periphery structure, with a dense core formed by the ensemble cast of dwarves. This core is surrounded by the main characters, which are themselves surrounded by supporting characters.

Monastery Interactions



ω	Group 1	Group 2	Group 3	κ
Group 1	14	16	15	45
Group 2	16	0	11	27

ω	**Group 1**	**Group 2**	**Group 3**	κ
Group 3	15	11	0	26

This graph appears to be largely disassortative, which may reflect the underlying social structure/hierarchy of the monastery. Notably, one group has many internal interactions while the other groups have none at all. I was not able to find any information about the dataset itself, but I would guess that the groups might loosely represent different roles or ranks within the monastery.