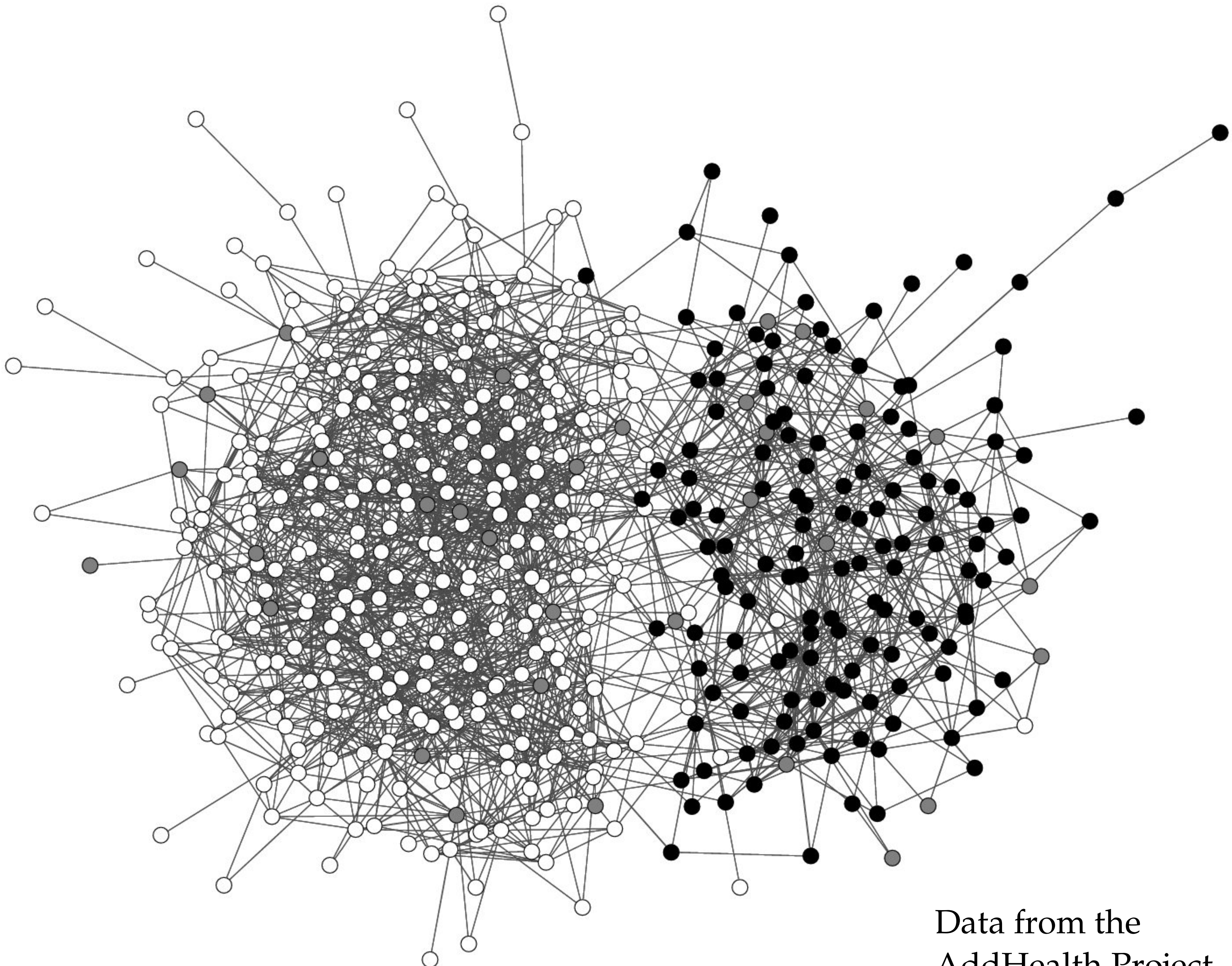
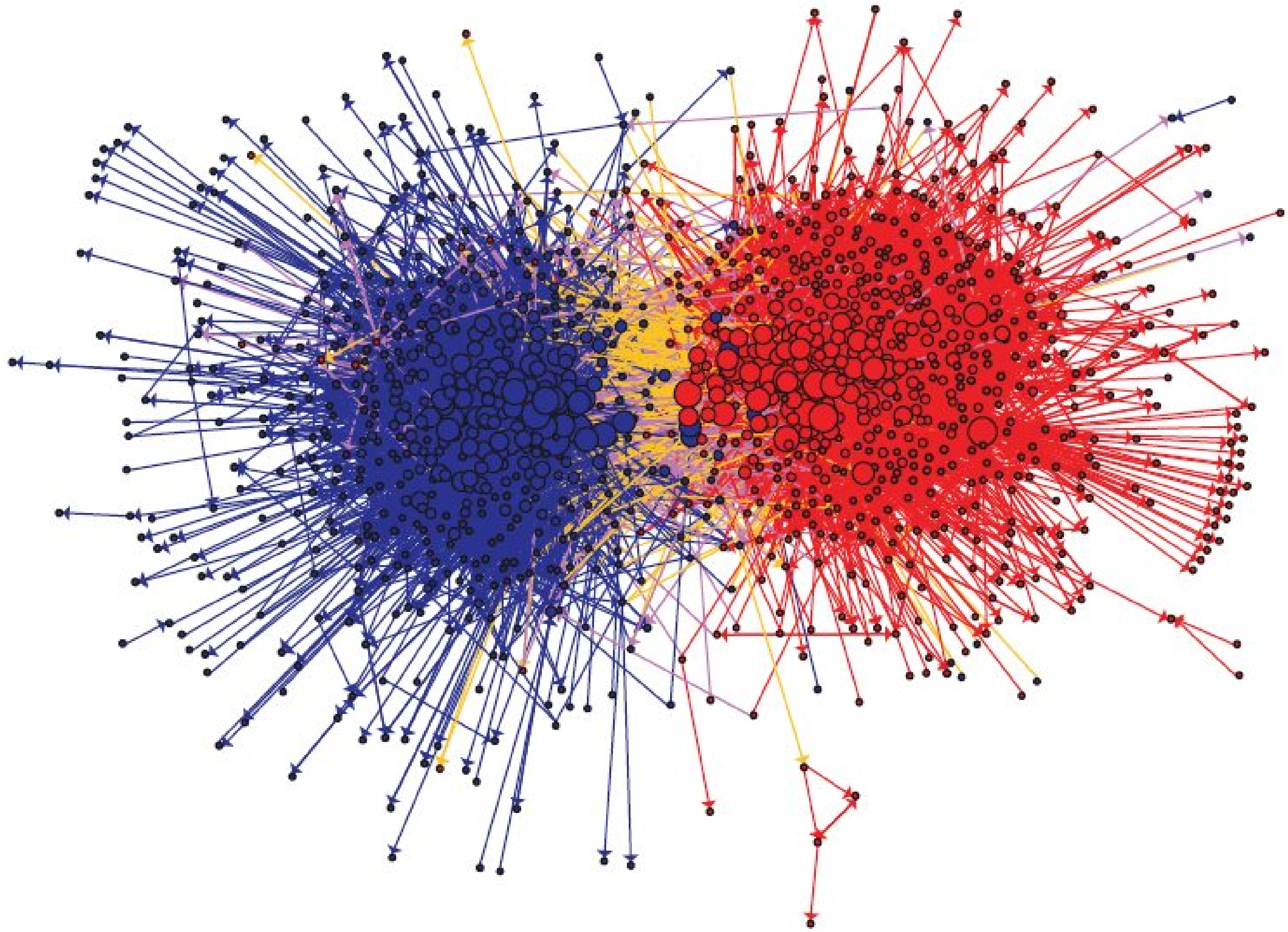
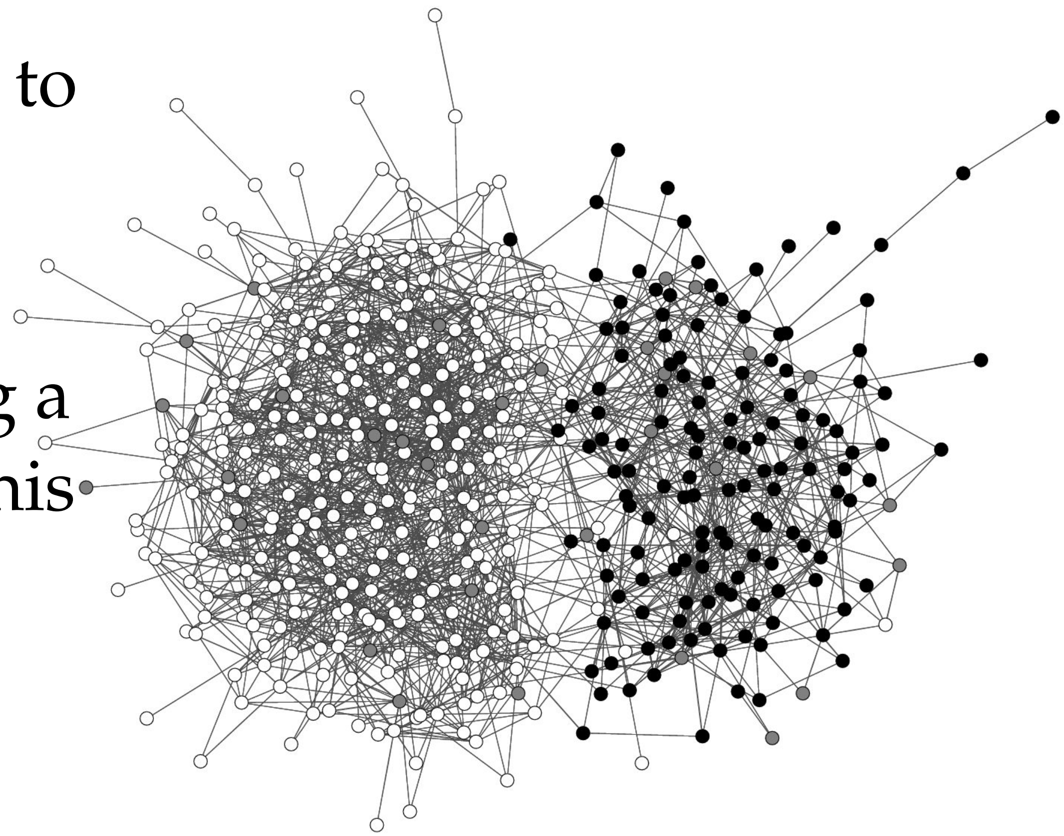


Data from the
AddHealth Project



Adamic & Glance 2005

- We can use homophily to make predictions
  - Taking a simple majority vote among a person's friends in this network predicts ethnicity with 83% accuracy
- Voting behavior:
  - On average, about 70% of your friends vote the same way as you do



- **70%** of your friends vote the same way as you do
- **79%** of people are within 5 years of their spouse's age
- In the high school, **67%** of friends were in the same grade
- **83%** of friends had the same ethnicity
- In a study in California, **72%** of people were the same ethnicity as their partner
- **91%** of Web links between political blogs are between blogs on the same side of the political aisle

Measuring homophily

- But just giving the percentage doesn't tell you much
- Some people would be the same just by chance
- Example:
  - In the high school the grades are about the same size: 25% of the students are in each grade
  - So if you made a friend at random, you would expect them to be in the same grade as you 25% of the time

Modularity

Expected percentage = 25%

Observed percentage = 67%

Modularity = 67% – 25% = 42%

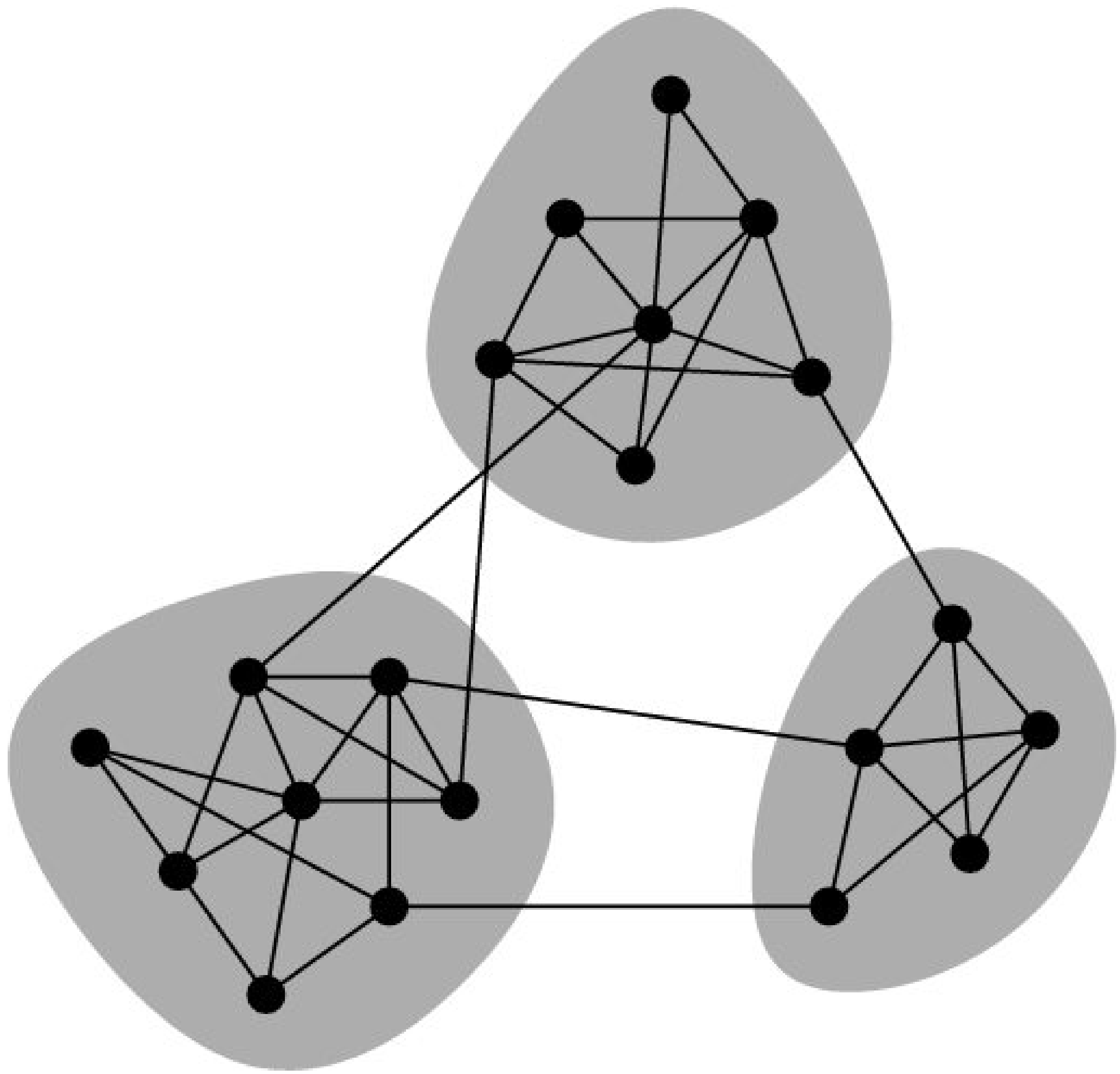
- The modularity is a measure of homophily:
  - 66% of people would be within 5 years of their spouse's age if they chose at random

Modularity = 79% – 66% = 13%

  - The equivalent figure for the California study: 29%

Modularity = 72% – 29% = 43%

Modules, groups, or communities



Modularity

Define modularity to be

$$Q = (\text{number of edges within groups}) - (\text{expected number within groups}).$$

- Modularity is measured relative to a *null model*
  - Defined by P_{ij} = probability of an edge between vertices i and j
  - Examples:
    - ➔ $P_{ij} = p$ (Erdös-Rényi random graph)
    - ➔ $P_{ij} = k_i k_j / 2m$ ("configuration model")

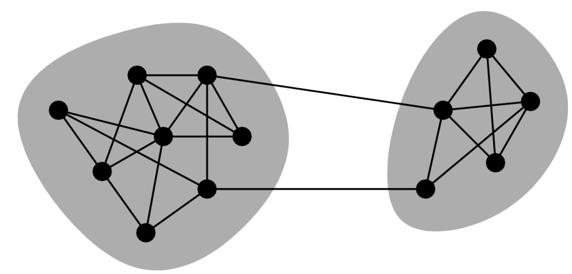
Matrix formulation

Actual number of edges between *i* and *j* is

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge } (i, j), \\ 0 & \text{otherwise.} \end{cases}$$

Expected number of edges is P_{ij}.

Modularity is sum of $A_{ij} - P_{ij}$ over all pairs of vertices (*i*,*j*) falling in the same group



Define:

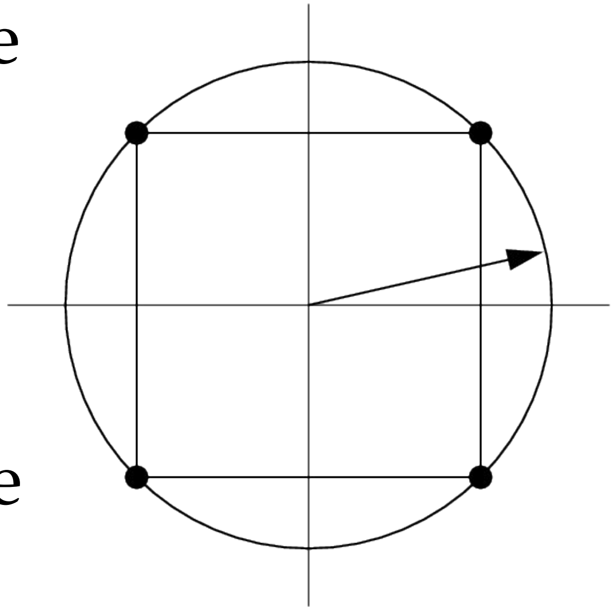
$$s_i = \begin{cases} +1 & \text{if vertex } i \text{ belongs to group 1,} \\ -1 & \text{if vertex } i \text{ belongs to group 2.} \end{cases}$$

$$
\begin{aligned}
Q &= \frac{1}{2m}\sum_{ij}\left[A_{ij}-P_{ij}\right]\delta(g_i,g_j) \\
&= \frac{1}{4m}\sum_{ij}\left[A_{ij}-P_{ij}\right](s_i s_j+1) \\
&= \frac{1}{4m}\sum_{ij}\left[A_{ij}-P_{ij}\right]s_i s_j \\
&= \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s}
\end{aligned}
$$

where $B_{ij} = A_{ij} - P_{ij}$

We call **B** the modularity matrix

- We wish to maximize $Q = \mathbf{s}^T \mathbf{B}\mathbf{s}$. True maximization is difficult, so we relax the constraint that $s_i = \pm 1$, instead enforcing only $|\mathbf{s}|^2 = n$.



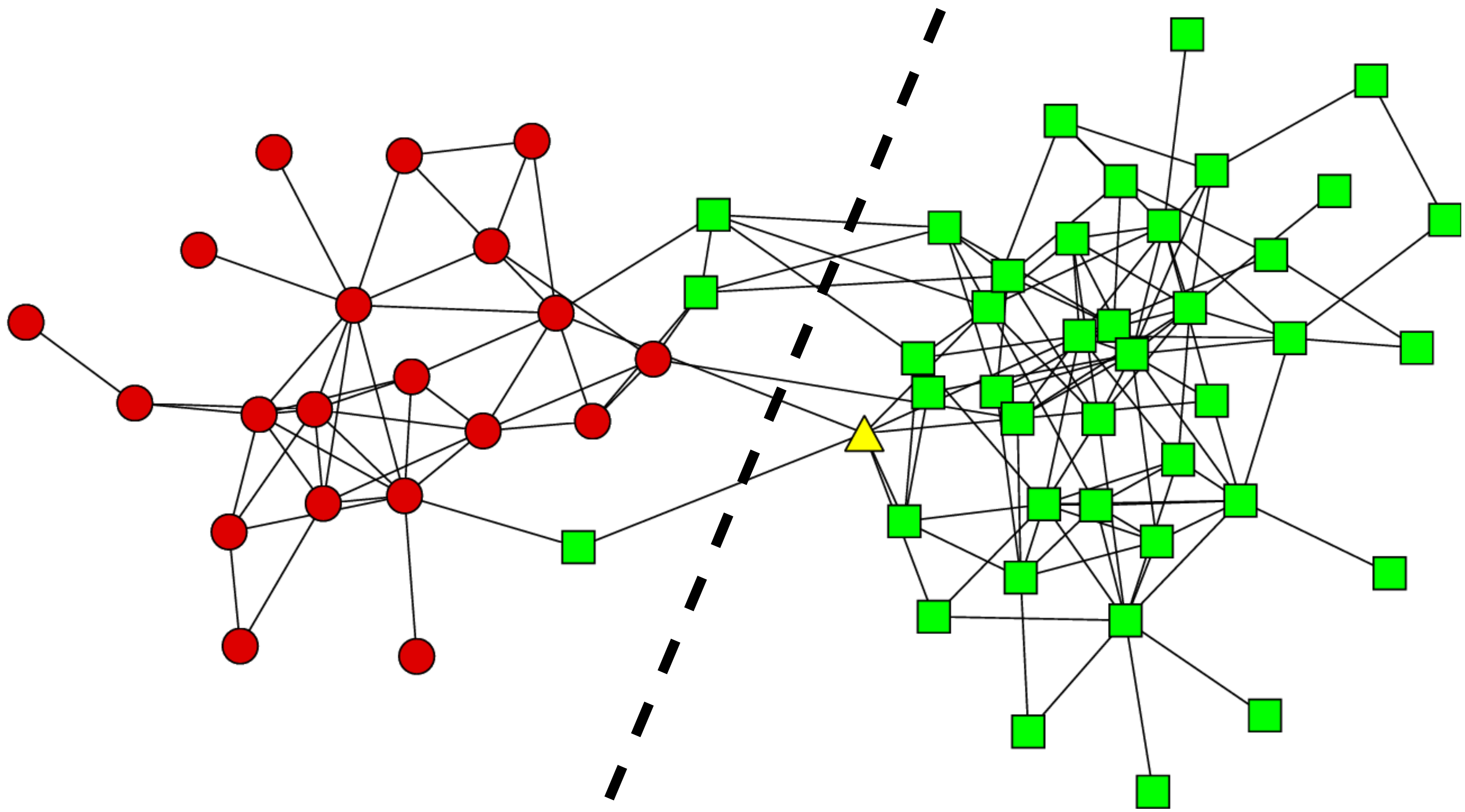
- Introducing a Lagrange multiplier we then find that Q is maximized when

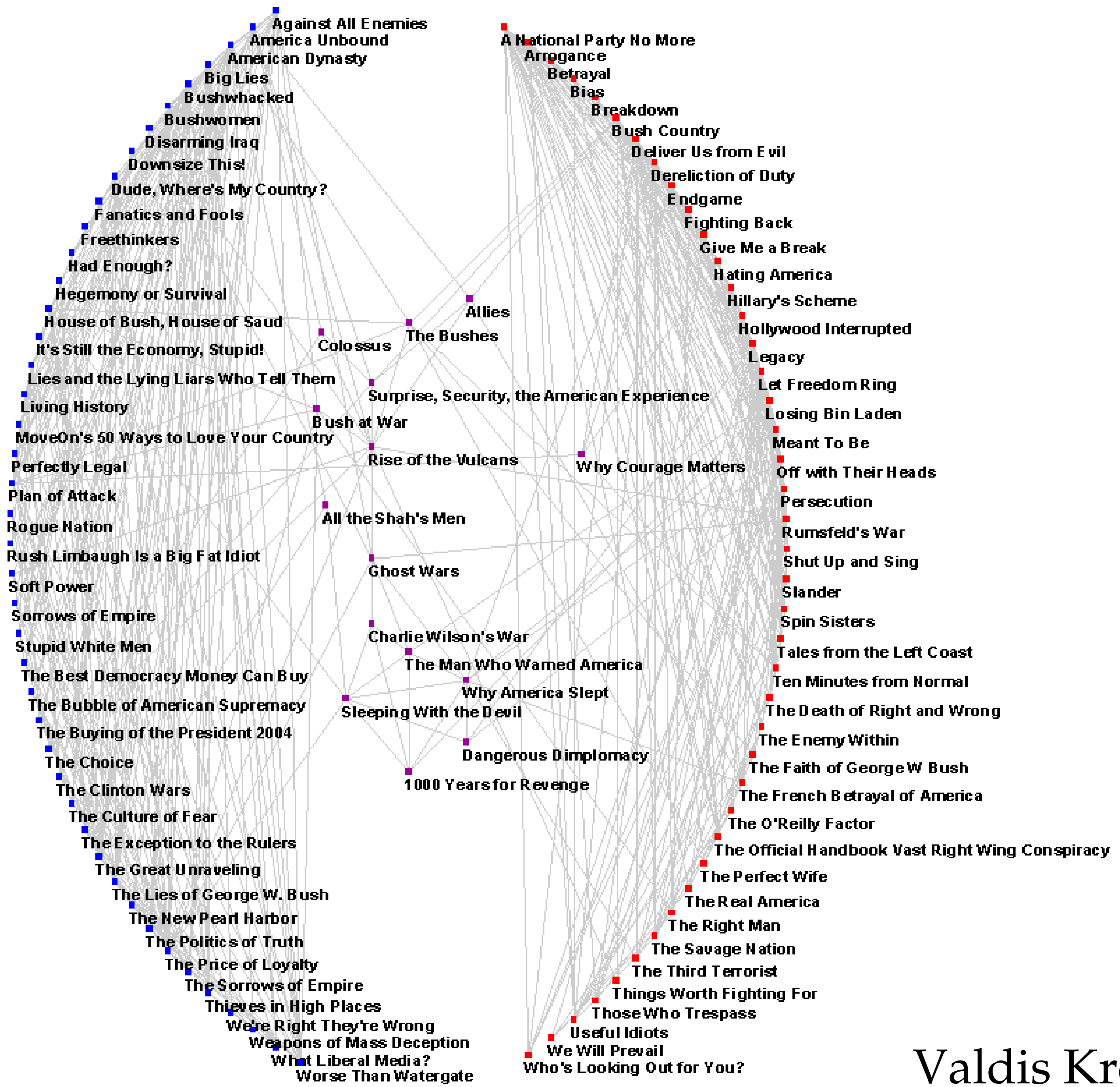
$$\mathbf{B}\mathbf{s} = \lambda \mathbf{s}.$$

- In practice we cannot achieve this maximum because of $s_i = \pm 1$, but we choose $\mathbf{s}$ as close as we can:

$$s_i = \begin{cases} +1 & \text{if } u_i^{(1)} \geq 0, \\ -1 & \text{if } u_i^{(1)} < 0. \end{cases}$$

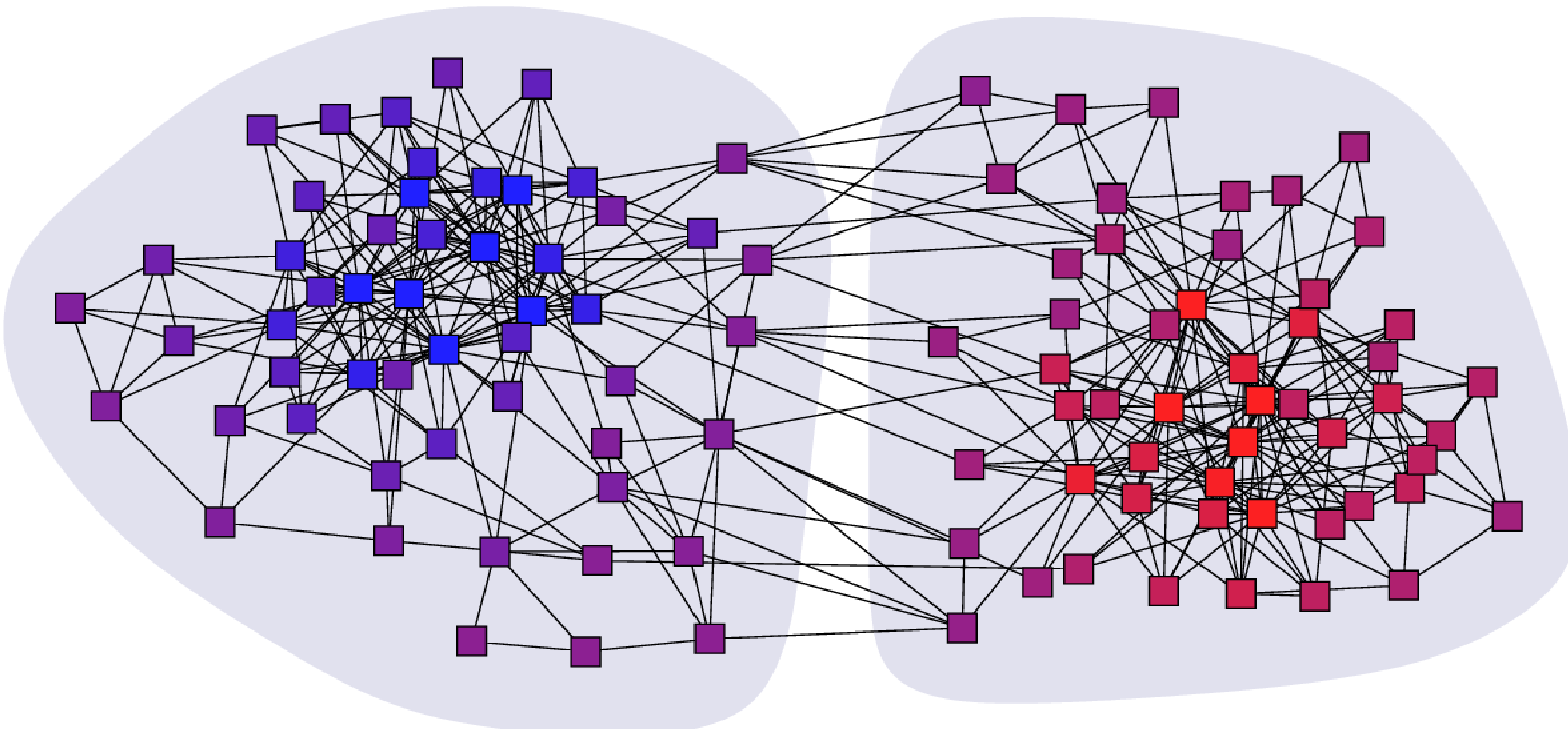
Example: animal network





Valdis Krebs

Books about politics



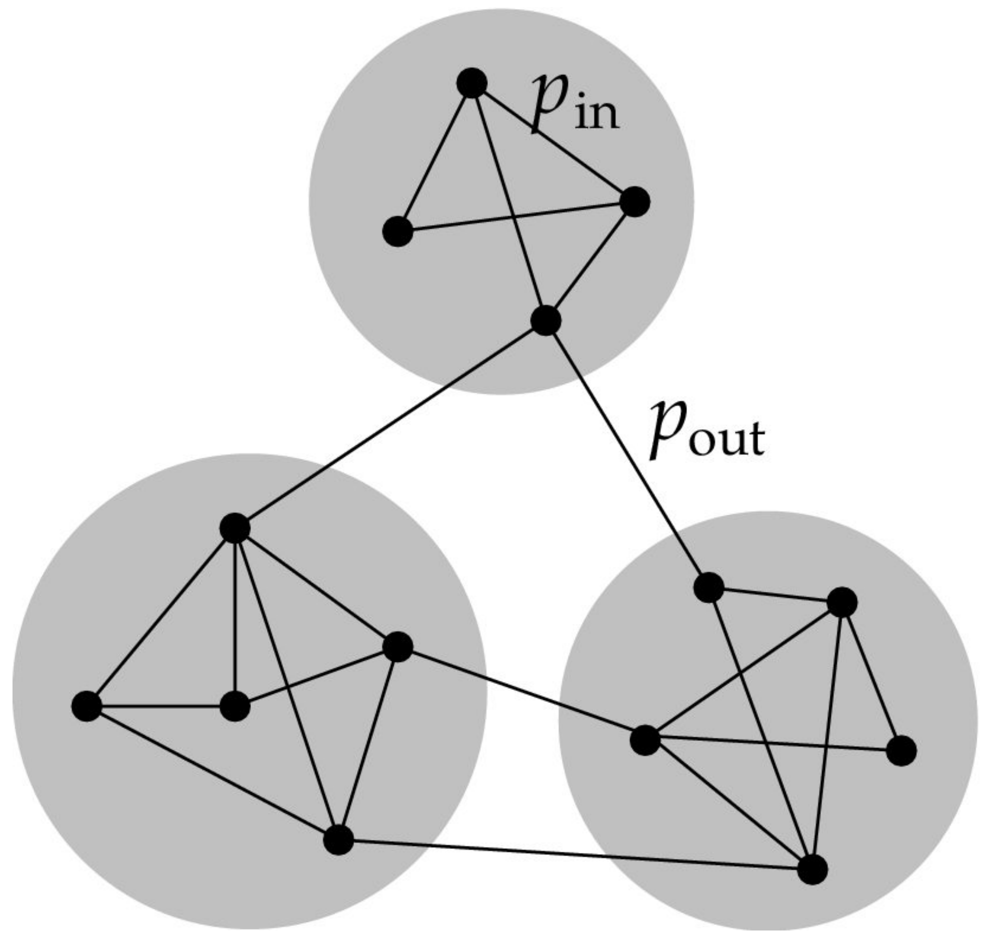
Graph spectra

- We have seen two matrix representations of the networks:
  - Adjacency matrix
  - Modularity matrix
- And we have seen that their spectra tell us useful things: **community structure**, **eigenvector centrality**
- Spectrum can be quantified by the *spectral density*:

$$\rho(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z - \lambda_i)$$

A controlled test
(Rao and Newman 2012)

- The *stochastic block model*:
- Nodes are divided into groups, with given probabilities of connection within and between them
- Often used as a benchmark or controlled test of how good our algorithms are

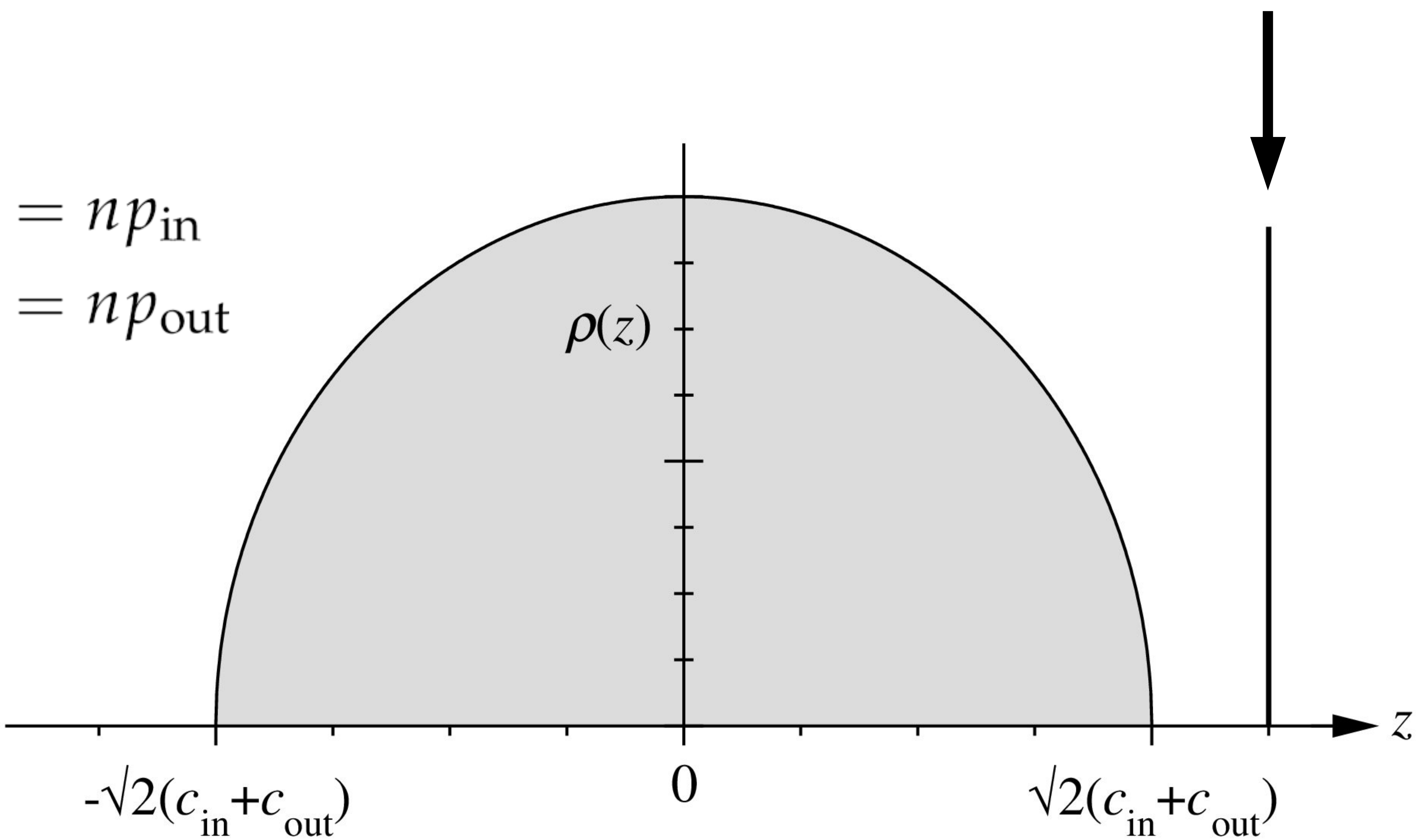


- We can calculate the spectrum of eigenvalues exactly for this model system

$$z_1 = \frac{1}{2}(c_{\text{in}} - c_{\text{out}}) + \frac{c_{\text{in}} + c_{\text{out}}}{c_{\text{in}} - c_{\text{out}}}$$

$$c_{\text{in}} = np_{\text{in}}$$

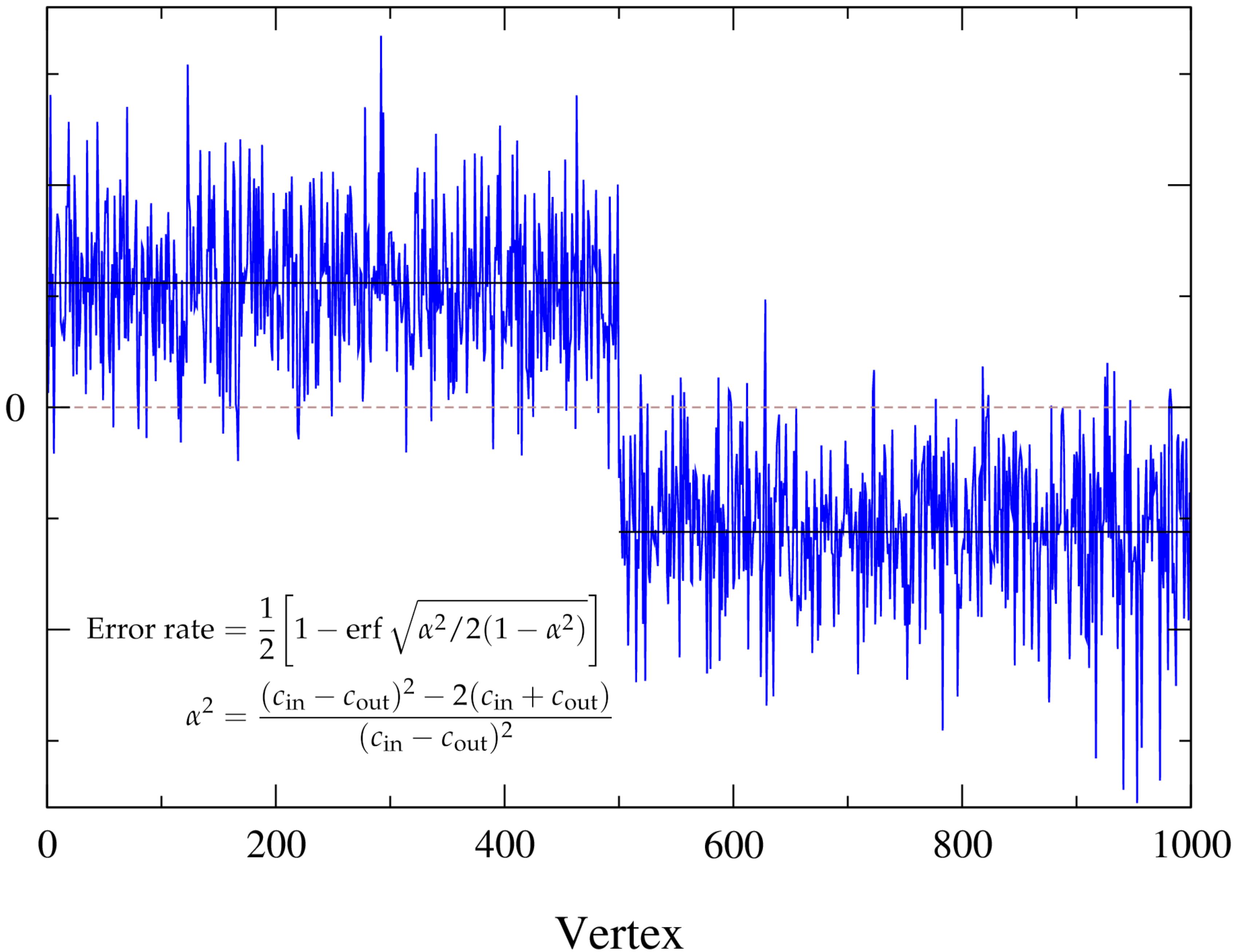
$$c_{\text{out}} = np_{\text{out}}$$

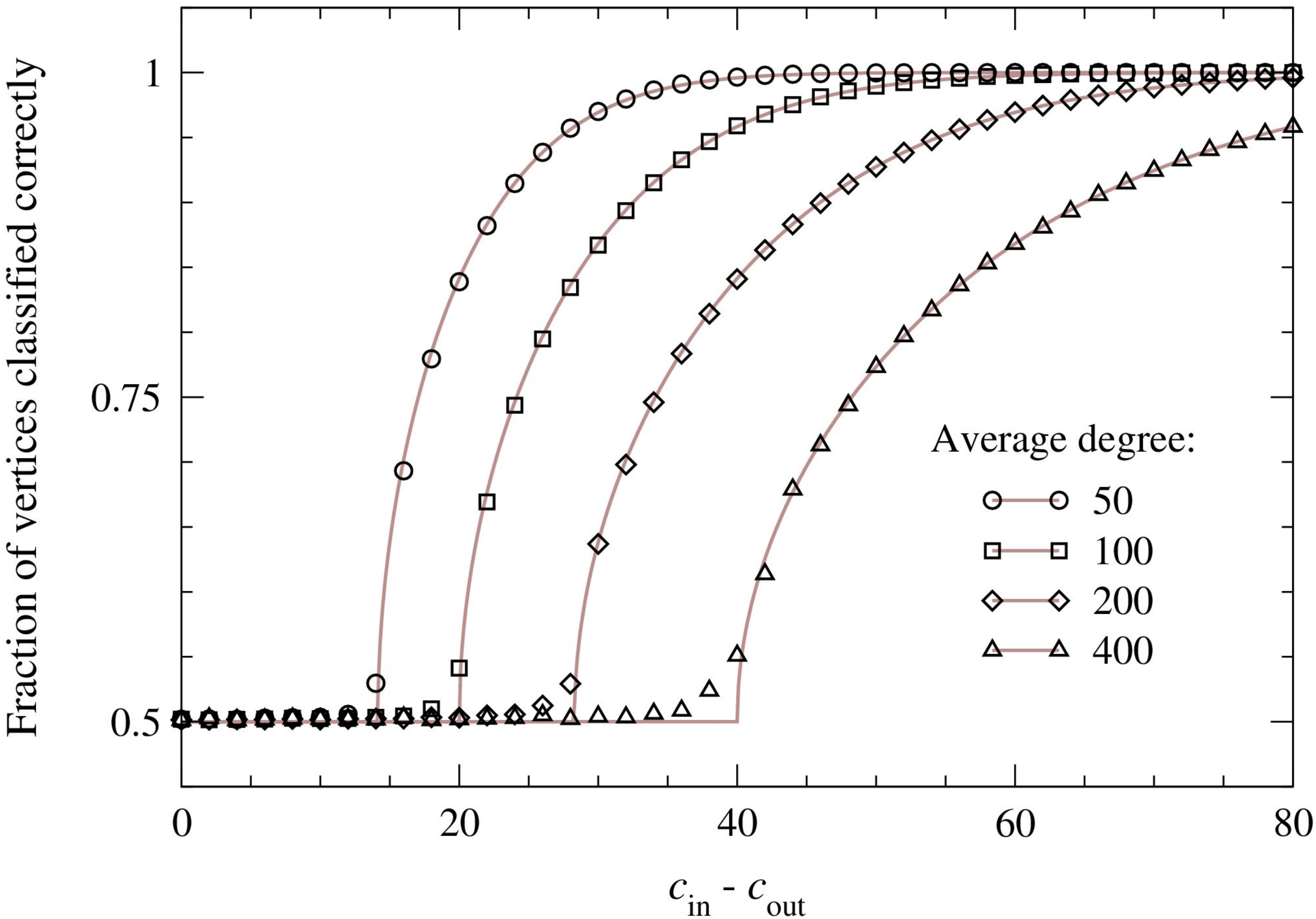


Phase transition

- The highest eigenvalue reveals the presence of community structure in the network
- But its value depends on the strength of that structure, as determined by c_{in} and c_{out}
- If this eigenvalue ever reaches the band edge, then the spectrum will become indistinguishable from that of the network with no community structure.
- This happens when

$$c_{\text{in}} - c_{\text{out}} = \sqrt{2(c_{\text{in}} + c_{\text{out}})}$$


0
0
200
400
600
800
1000
Vertex
$$\text{Error rate} = \frac{1}{2}\left[1 - \operatorname{erf}\sqrt{\alpha^2/2(1-\alpha^2)}\right]$$
$$\alpha^2 = \frac{(c_{\text{in}} - c_{\text{out}})^2 - 2(c_{\text{in}} + c_{\text{out}})}{(c_{\text{in}} - c_{\text{out}})^2}$$



Rao and MEJN, *PRL* 2012