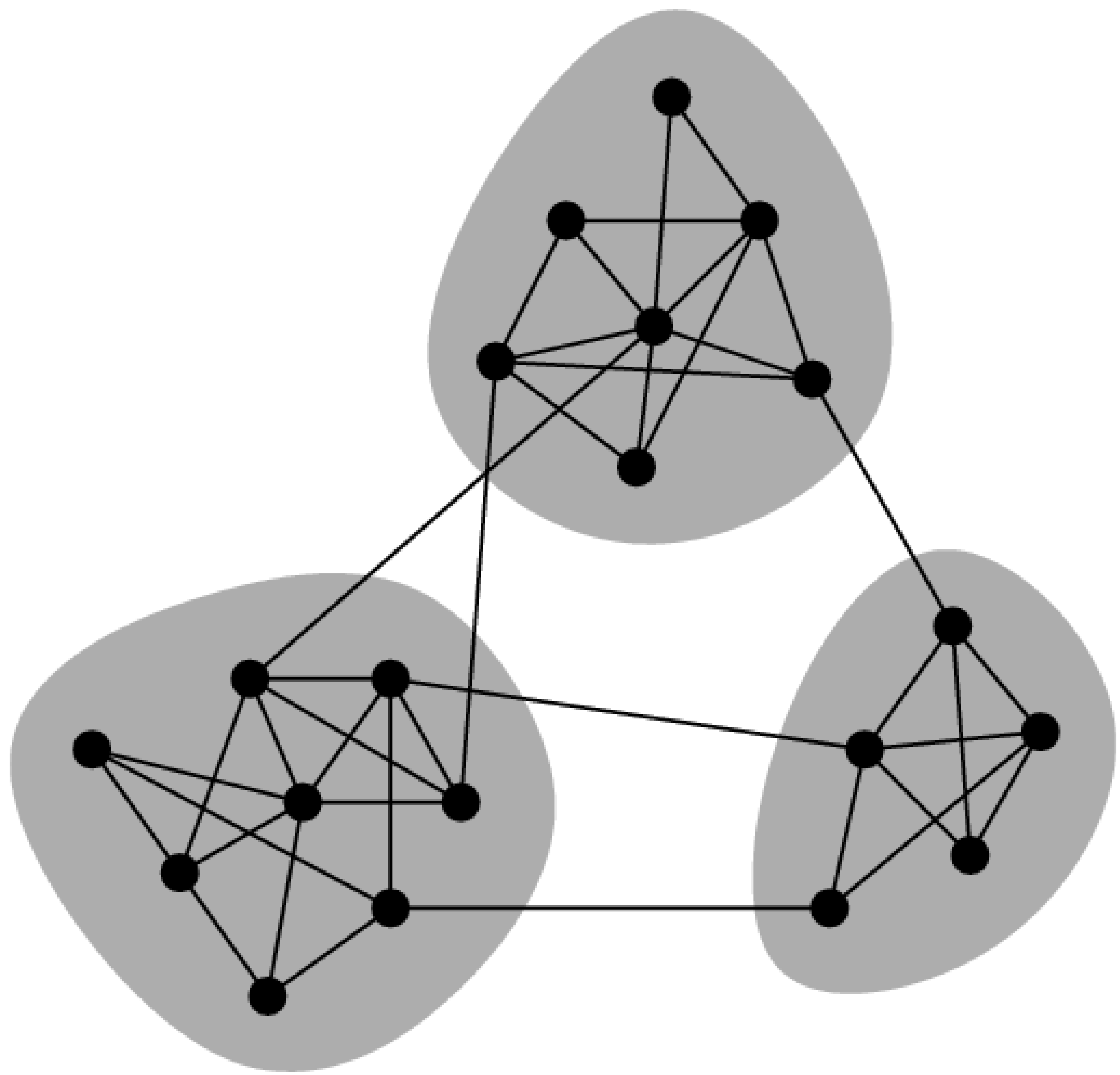
Modularity, Community Structure, and Spectral Properties of Networks

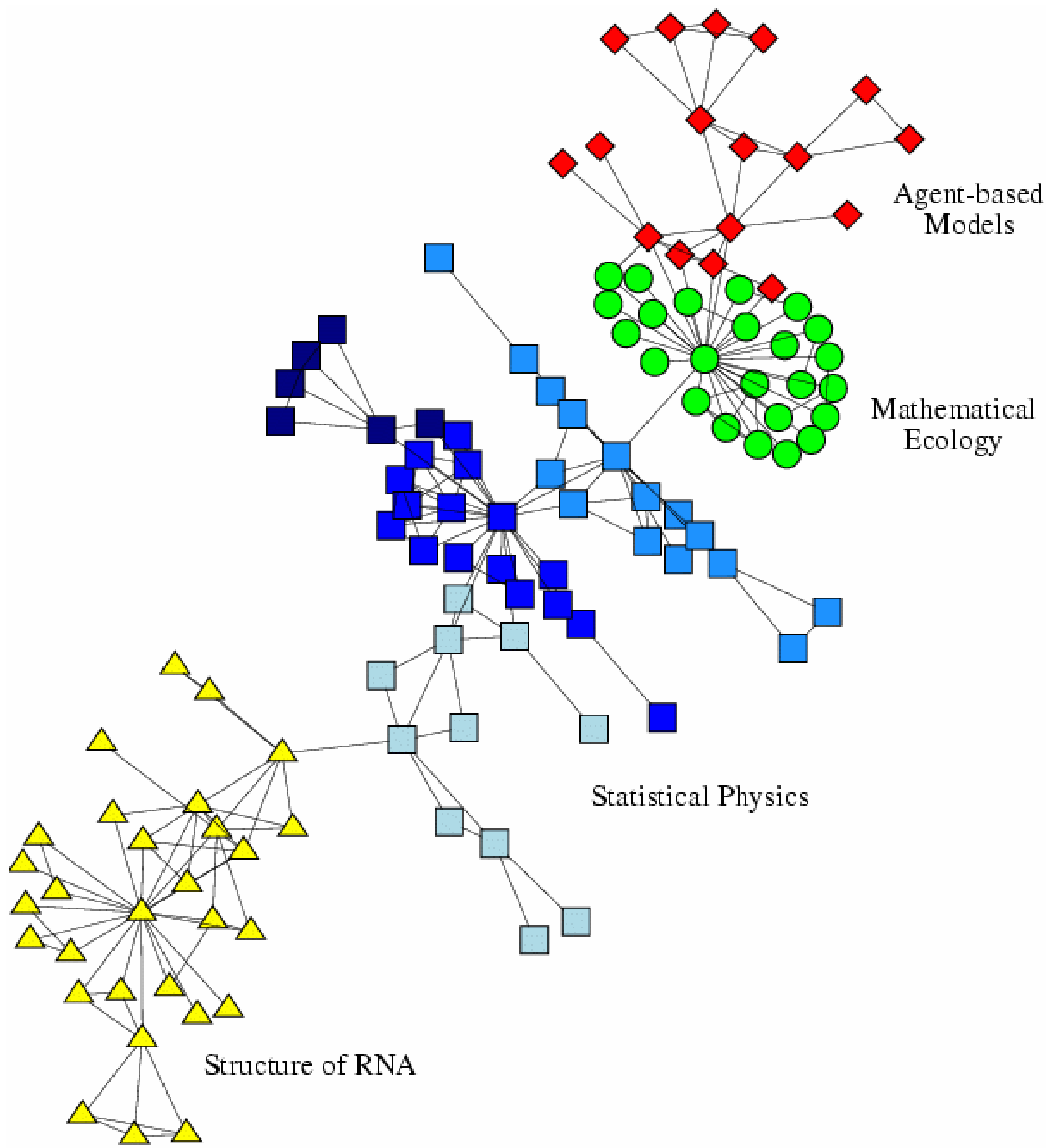
Mark Newman
University of Michigan

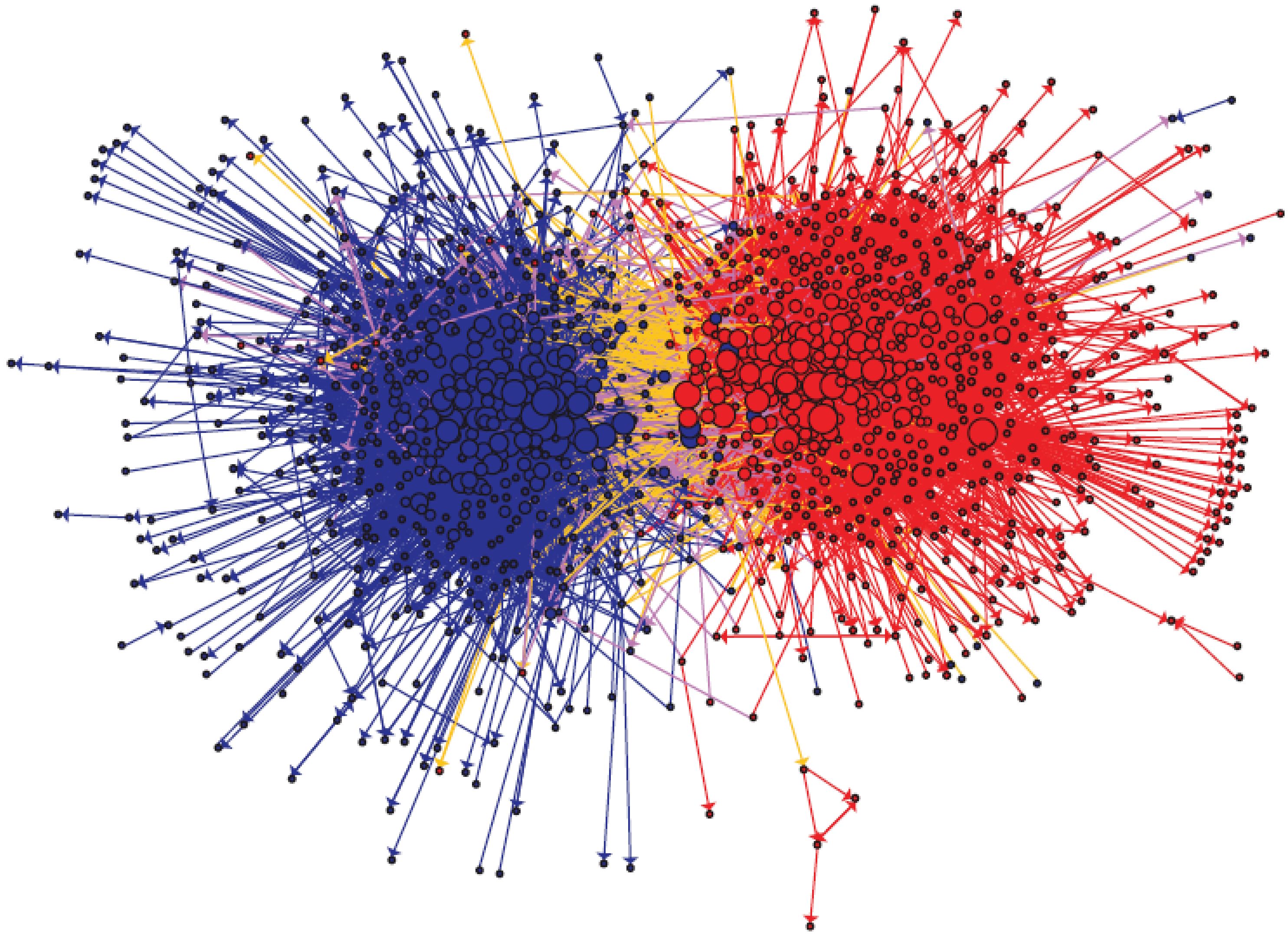
Community structure



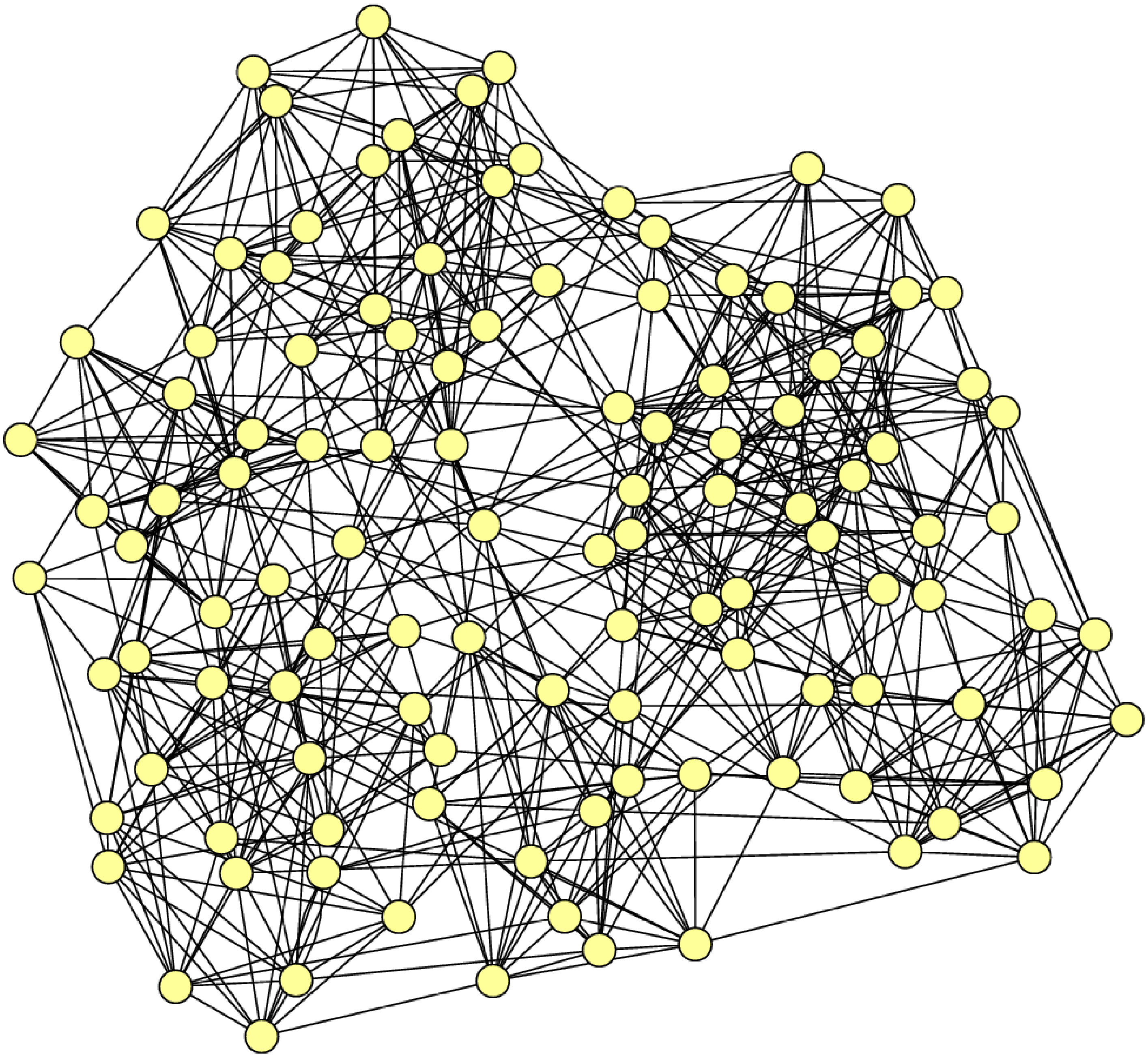
Community structure

- Communities are of interest in many cases:
  - World Wide Web
  - Citation networks
  - Social networks
  - Metabolic networks
- Properties of communities may be quite different from average properties of a network


Agent-based
Models
Mathematical
Ecology
Statistical Physics
Structure of RNA



Adamic & Glance 2005



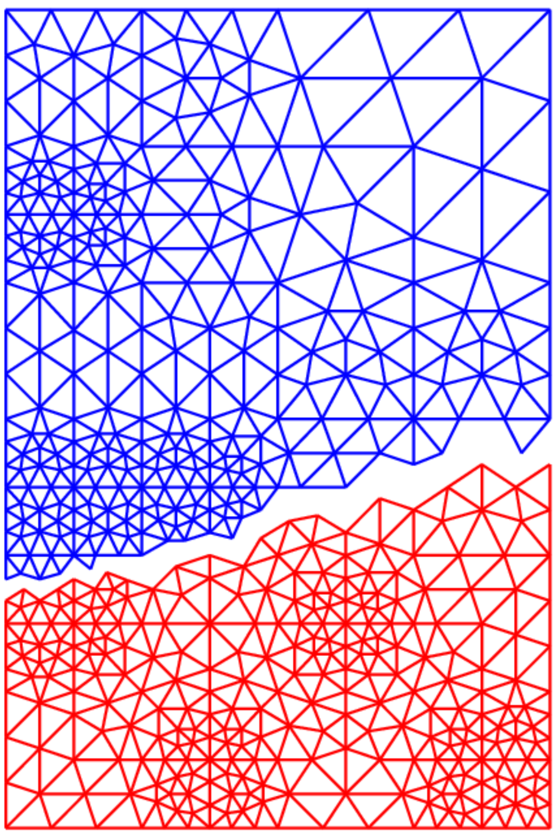
Computer science approaches

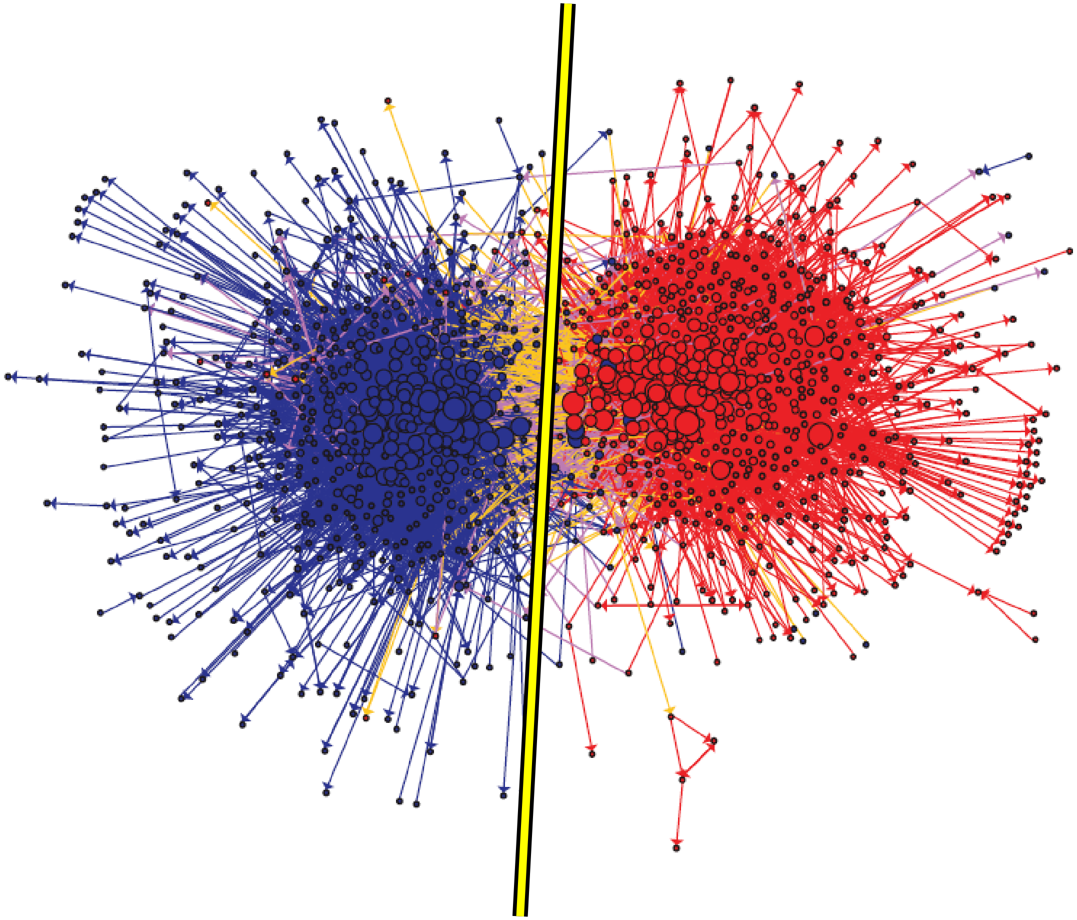
- Graph partitioning
  - Minimum cuts
  - Ratio cuts and variants
- Algorithms include:
  - Kernighan-Lin method and variants
  - Geometric methods
  - Spectral partitioning
- Applications:
  - Parallel computing
  - VLSI design and other CAD applications

Social science approaches

- Clustering
  - Block modeling
  - Community structure
- Algorithms include:
  - Best fits to stochastic block models
  - Hierarchical clustering based on single- or average-linkage clustering of similarity indices
  - Betweenness-based methods

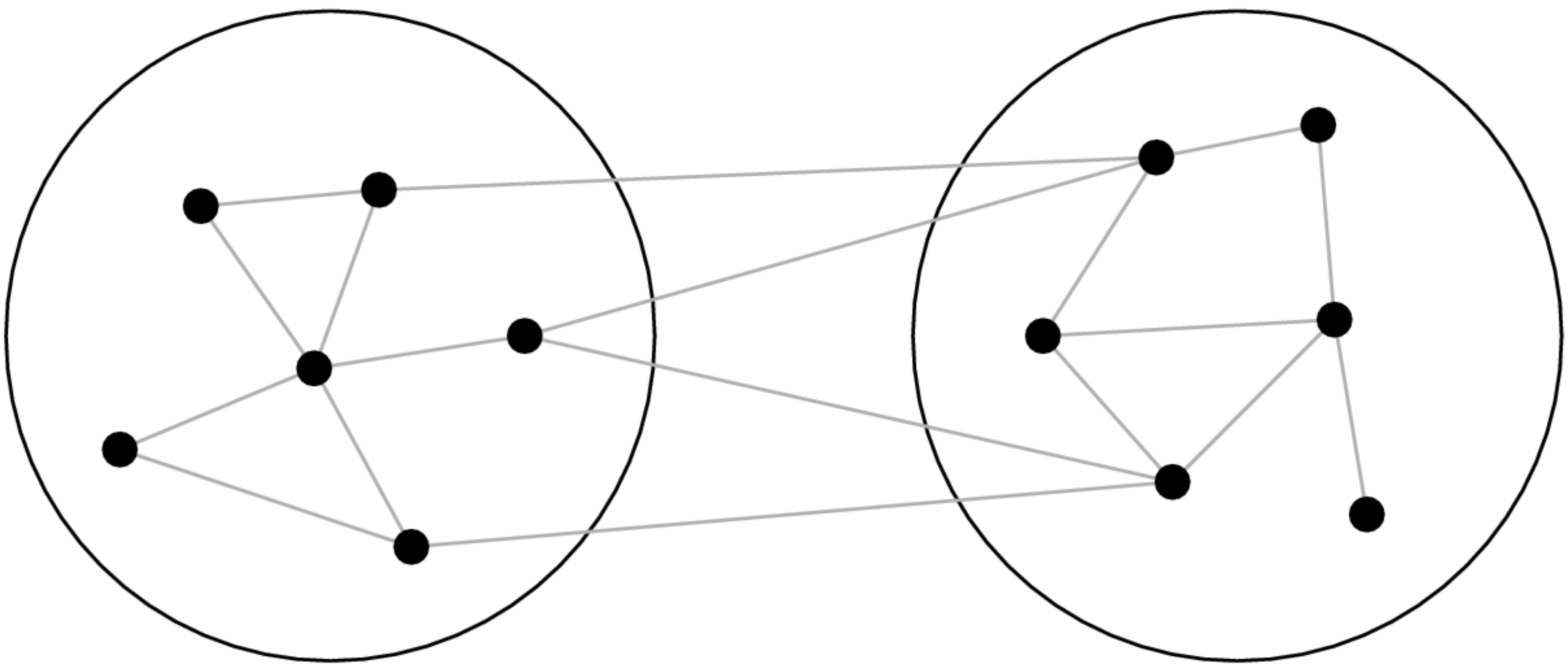
- It is tempting to suggest that graph partitioning and clustering are really the same thing, but there are important differences between the problems they address:





Graph partitioning

- Graph partitioning algorithms are typically based on *minimum cut* approaches or similar:



Graph partitioning

- Minimum cut partitioning breaks down when we don't know the sizes of the groups
  - Optimizing the cut size with the groups sizes free puts all vertices in the same group

Graph partitioning

- Minimum cut partitioning breaks down when we don't know the sizes of the groups
  - Optimizing the cut size with the groups sizes free puts all vertices in the same group
- *Cut size is the wrong thing to optimize*
  - A good division into communities is not just one where there are a small number of edges between groups
  - There must be a *smaller than expected* number

Modularity

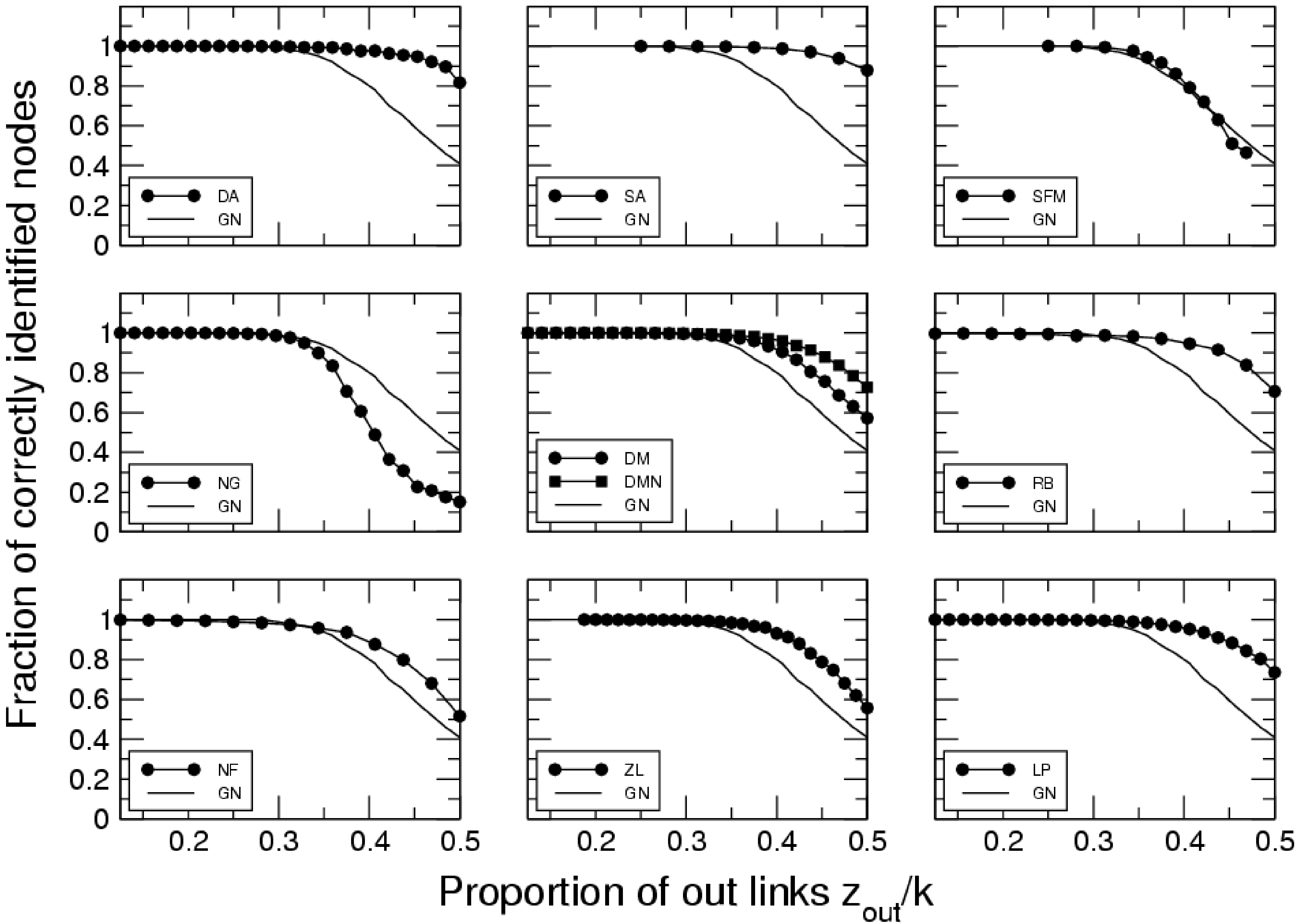
(Newman and Girvan 2004)

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")

Danon *et al.* (2005)



- There are similarities to social networks:
  - It has long been understood that communities in social networks define roles
  - A central node in a community may have a lot of power within the community, but little elsewhere
  - A “broker” node that lies between communities may have influence as a carrier of information between otherwise unconnected people
  - See Granovetter (1973), Burt (1976)

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

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

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

We call **B** the modularity matrix

- Now we write $\mathbf{s}$ as a linear combination of the eigenvectors $\mathbf{u}_i$ of the modularity matrix:

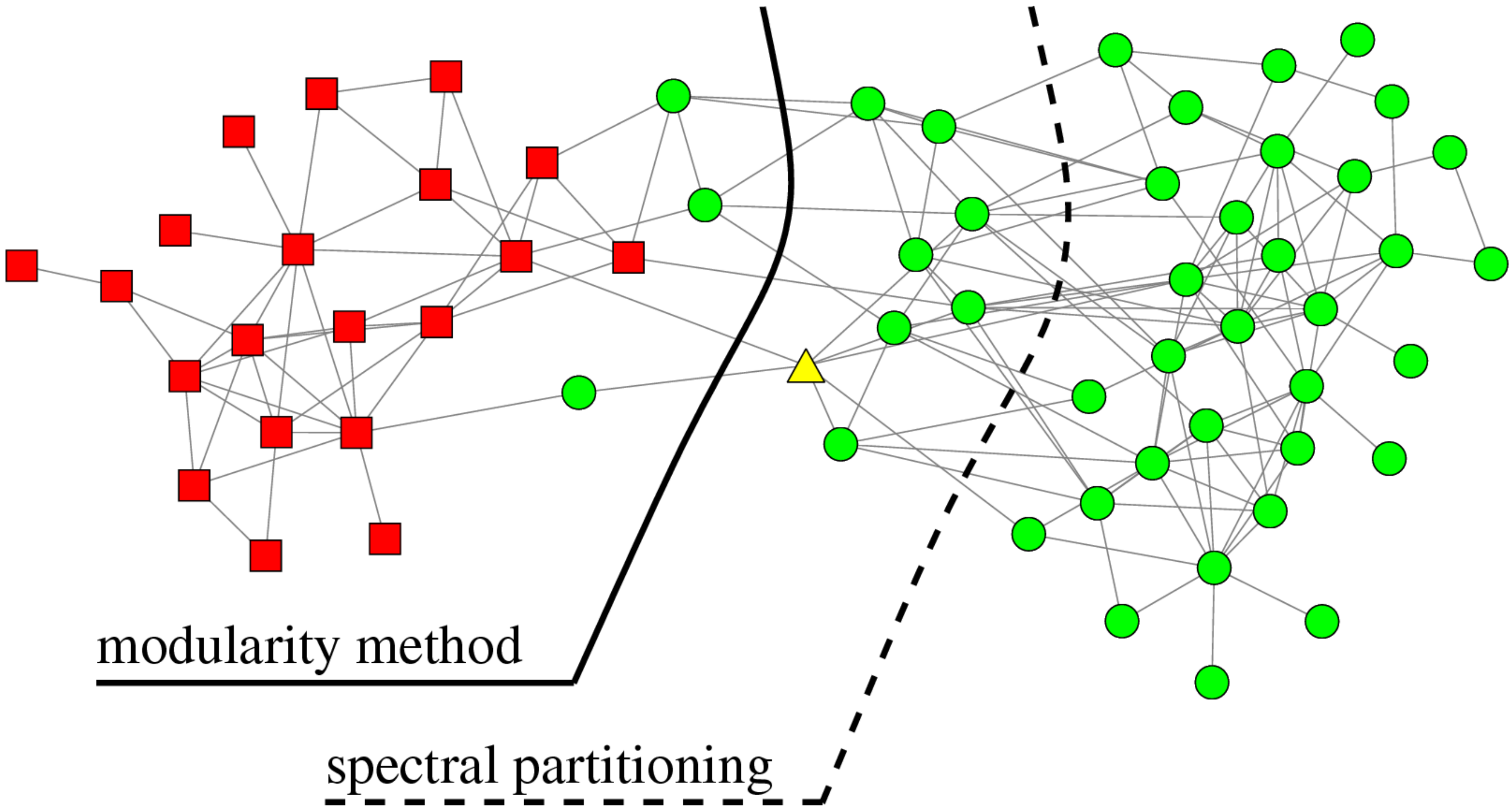
$$\mathbf{s} = \sum_{i=1}^{n} a_i \mathbf{u}_i, \qquad \text{with} \quad a_i = \mathbf{u}_i^T \mathbf{s}$$

$$Q = \frac{1}{4m} \mathbf{s}^T \mathbf{B} \mathbf{s} = \frac{1}{4m} \sum_i a_i^2 \beta_i$$

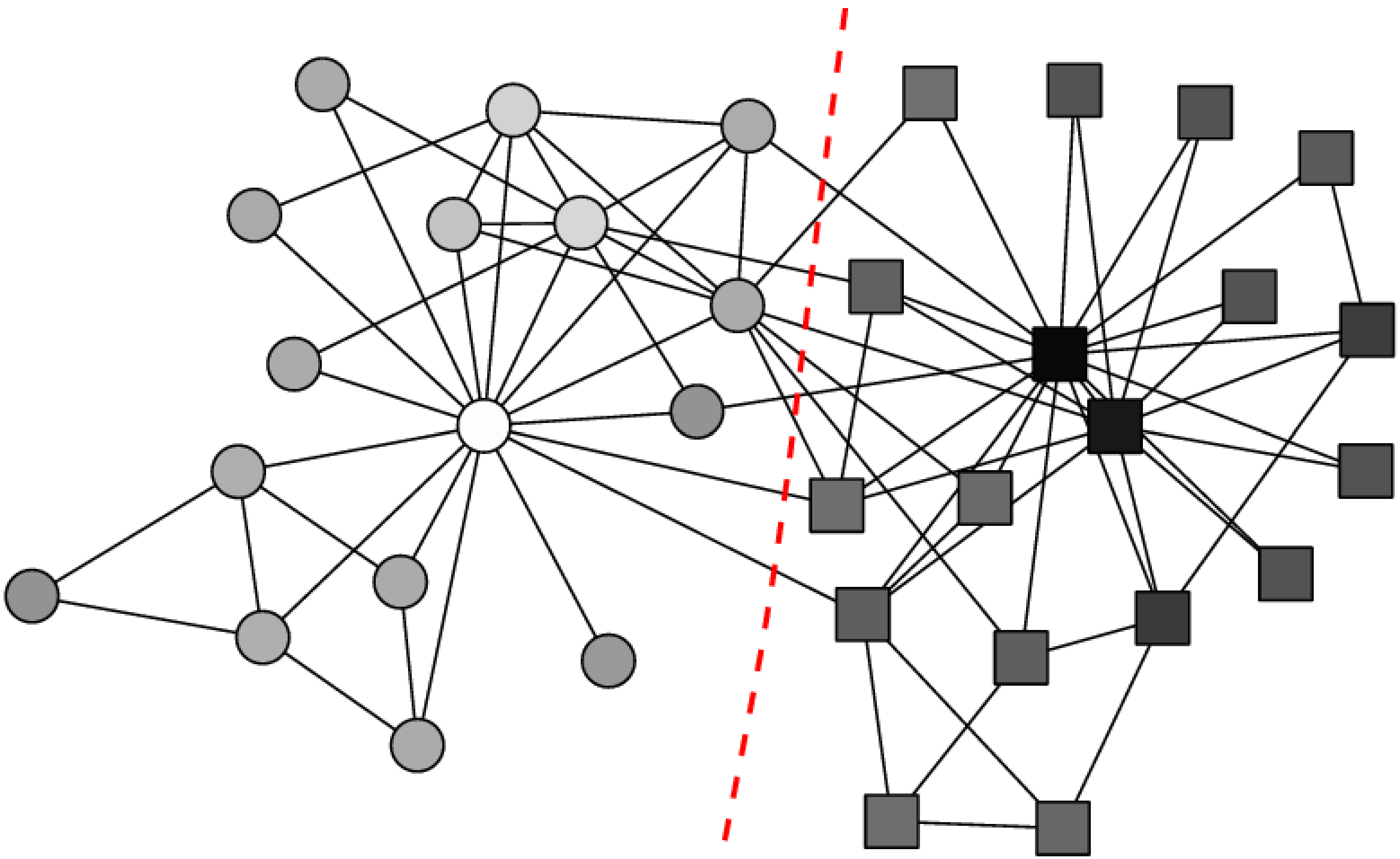
- We want to put as much weight as possible in the term corresponding to the largest eigenvalue

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

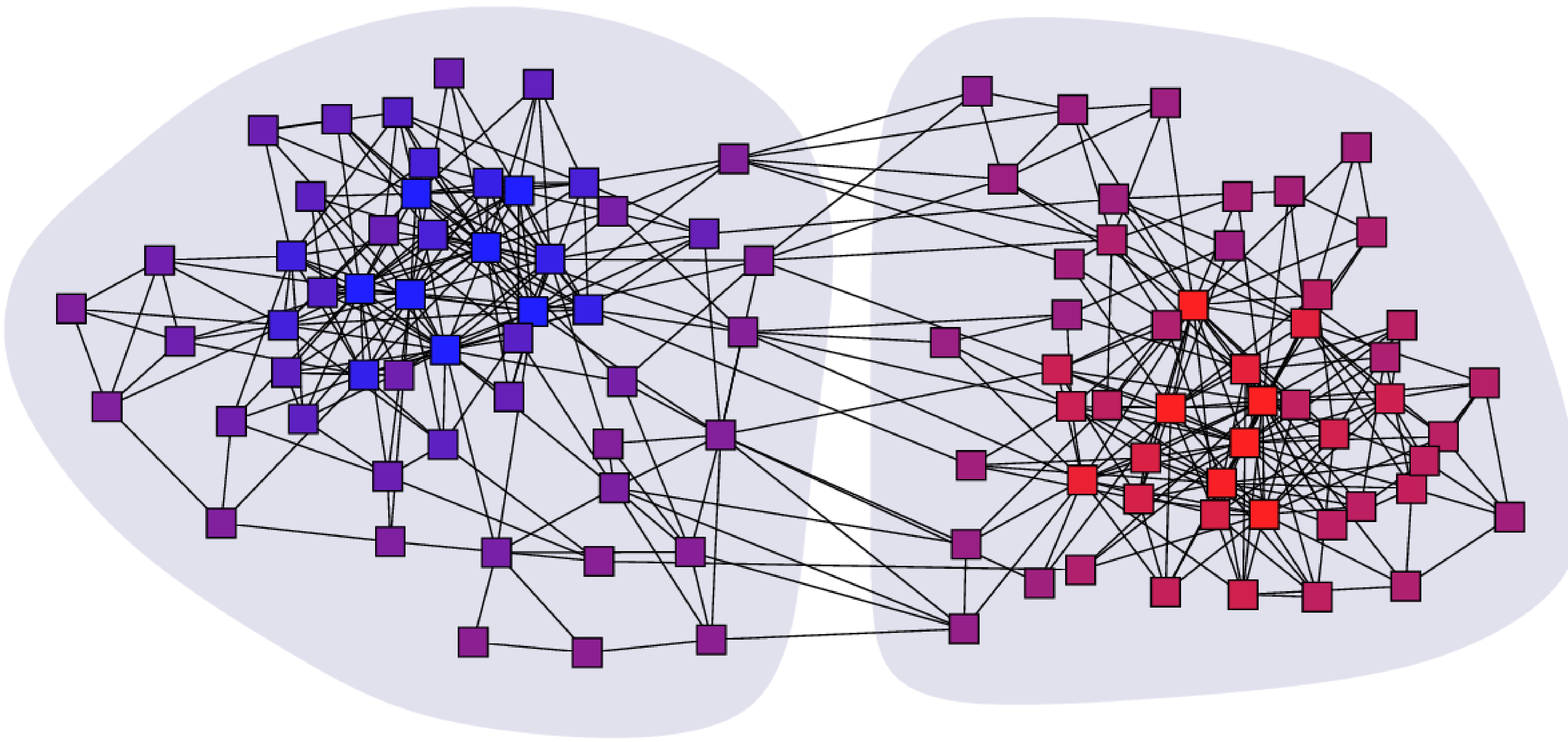
Example: animal network



Social network



Books about politics

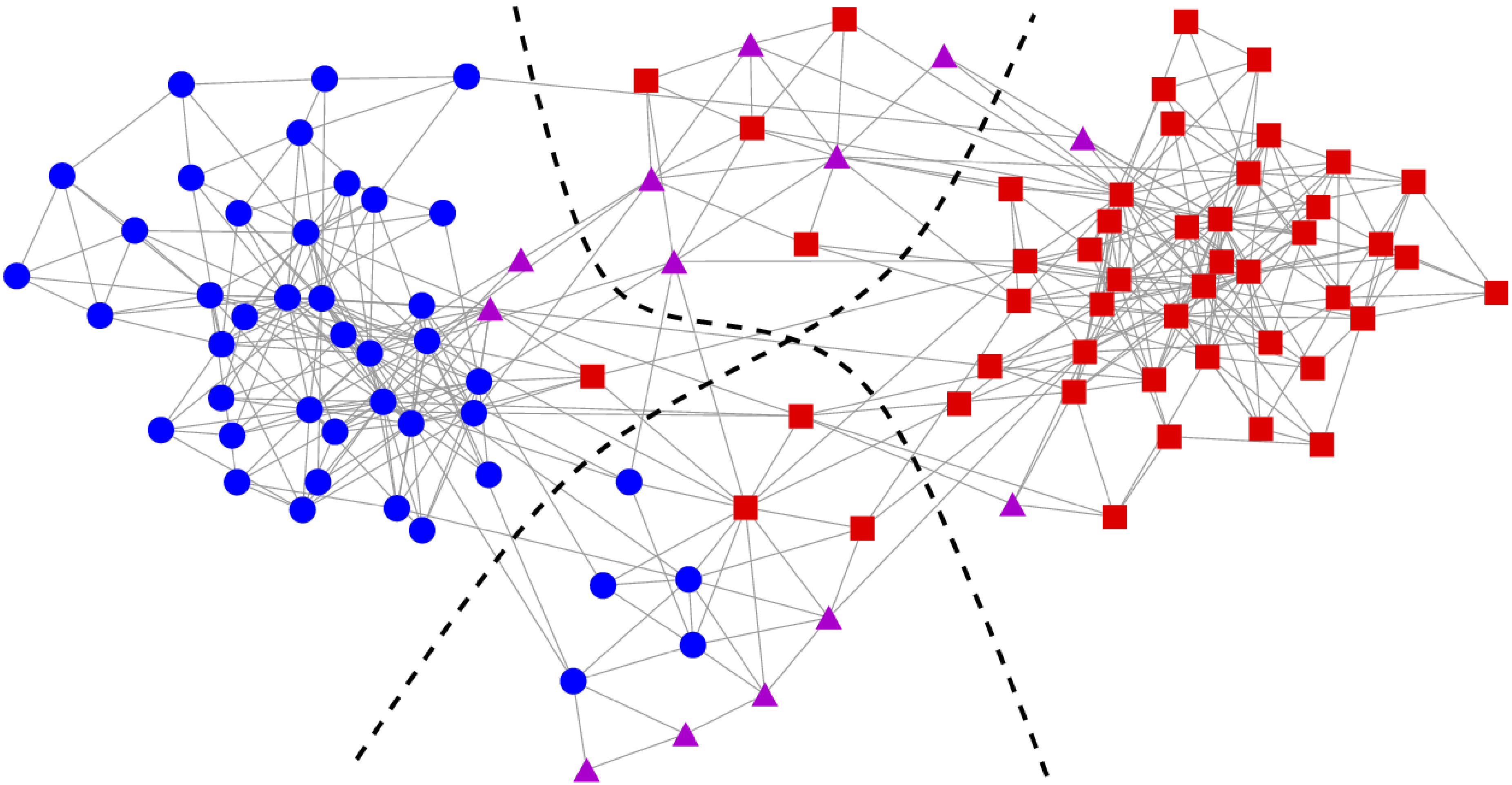


Spectral properties of modularity matrix

- Vector (1, 1, 1, . . .) is always an eigenvector of **B** with eigenvalue zero
- Eigenvalues can be either positive or negative
  - So long as there is *any* positive eigenvalue we will never put all vertices in the same group
- But there may be no positive eigenvalues
  - All vertices in same group gives highest modularity
  - We call such networks *indivisible*

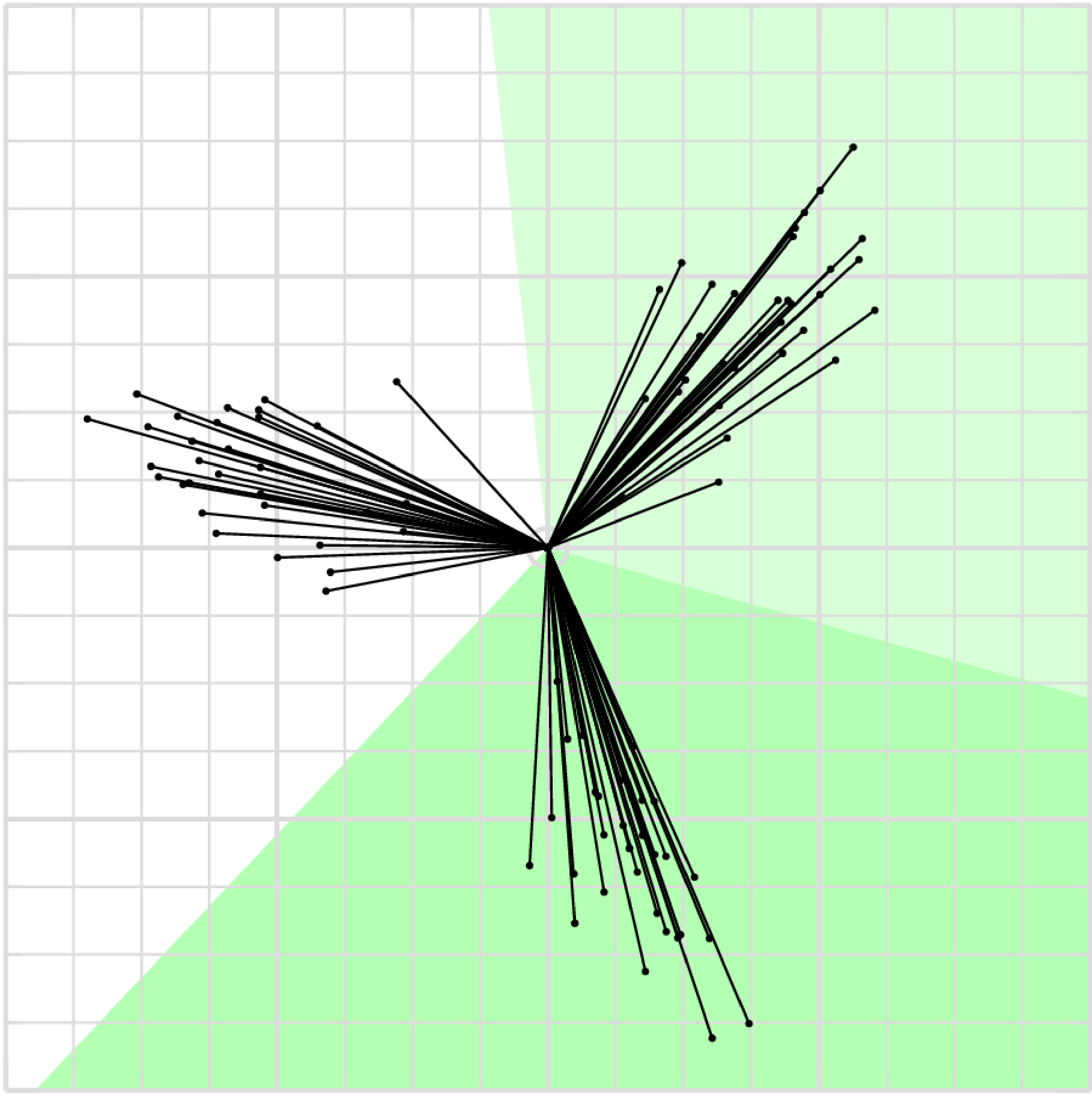
Dividing into more than two groups

- Simplest approach is repeated division into two groups
  - Divide in two, then divide those parts in two, etc.
- Stop when there is no division that will increase the modularity
  - But this is precisely when the subgraph is indivisible
  - Stop when there are no positive eigenvalues of the modularity matrix



Other eigenvectors

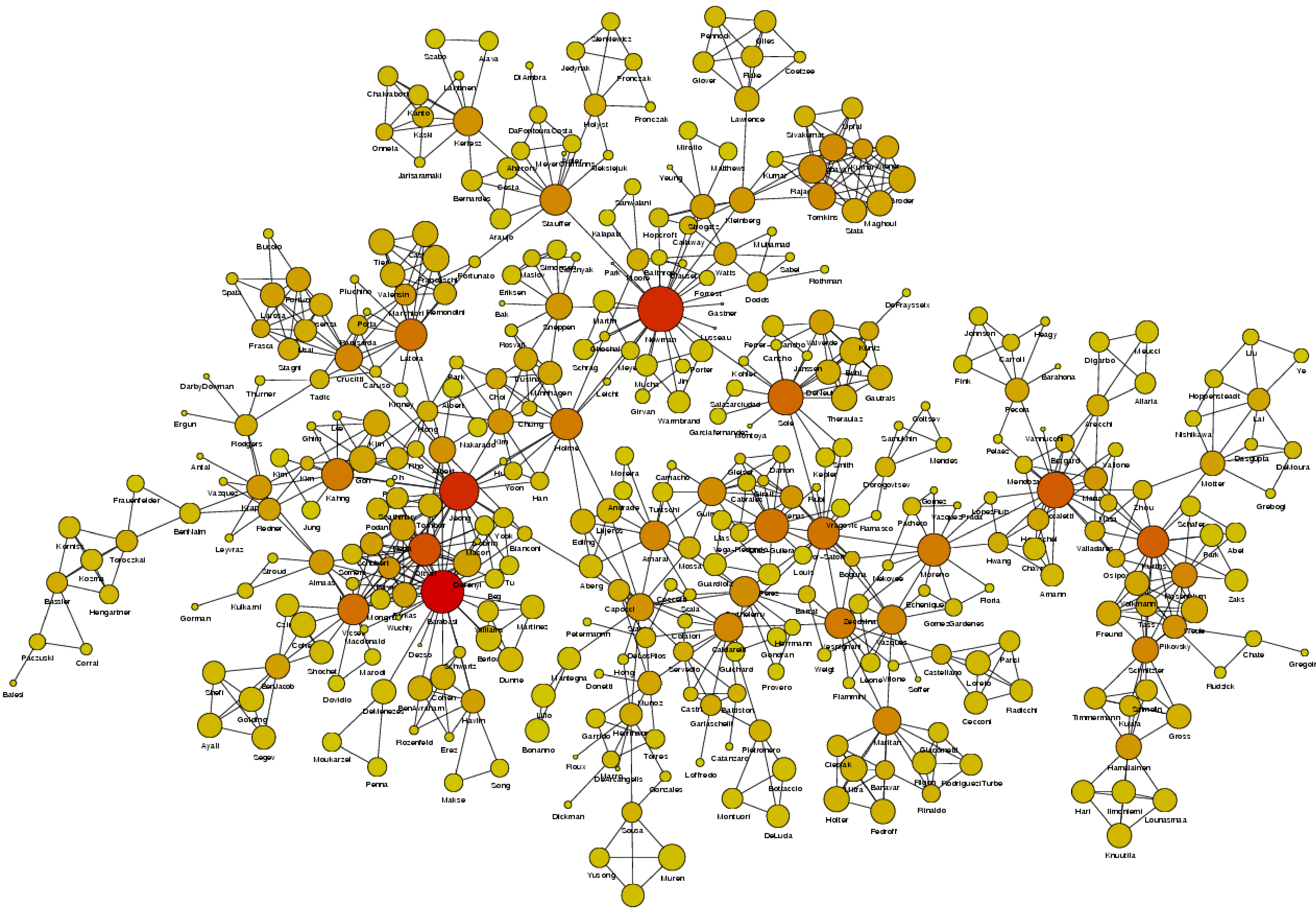
$$
\begin{aligned}
Q &= n\alpha + \mathrm{Tr}[\mathbf{S}^T\mathbf{U}(\mathbf{D}-\alpha\mathbf{I})\mathbf{U}^T\mathbf{S}] \\
&= n\alpha + \sum_{j=1}^{n}\sum_{k=1}^{c}(\beta_j-\alpha)\left[\sum_{i=1}^{n} U_{ij}S_{ik}\right]^2 \\
&= n\alpha + \sum_{k=1}^{c}\sum_{j=1}^{p}\left[\sum_{i=1}^{n}\sqrt{\beta_j-\alpha}\,U_{ij}S_{ik}\right]^2 \\
&= n\alpha + \sum_{k=1}^{c}\sum_{j=1}^{p}\left[\sum_{i\in G_k}[\mathbf{r}_i]_j\right]^2 \\
&= n\alpha + \sum_{k=1}^{c}|\mathbf{R}_k|^2,
\end{aligned}
$$



Centrality

- This also leads to some new measures:
  - The magnitude of the vertex vectors measures how large a contribution they *can* make to the modularity. We call this *community centrality*.
  - The angle that the vertex vectors make to the community vectors is a measure of *core/periphery structure*, i.e., how close a vertex is to the middle of the group it belongs to

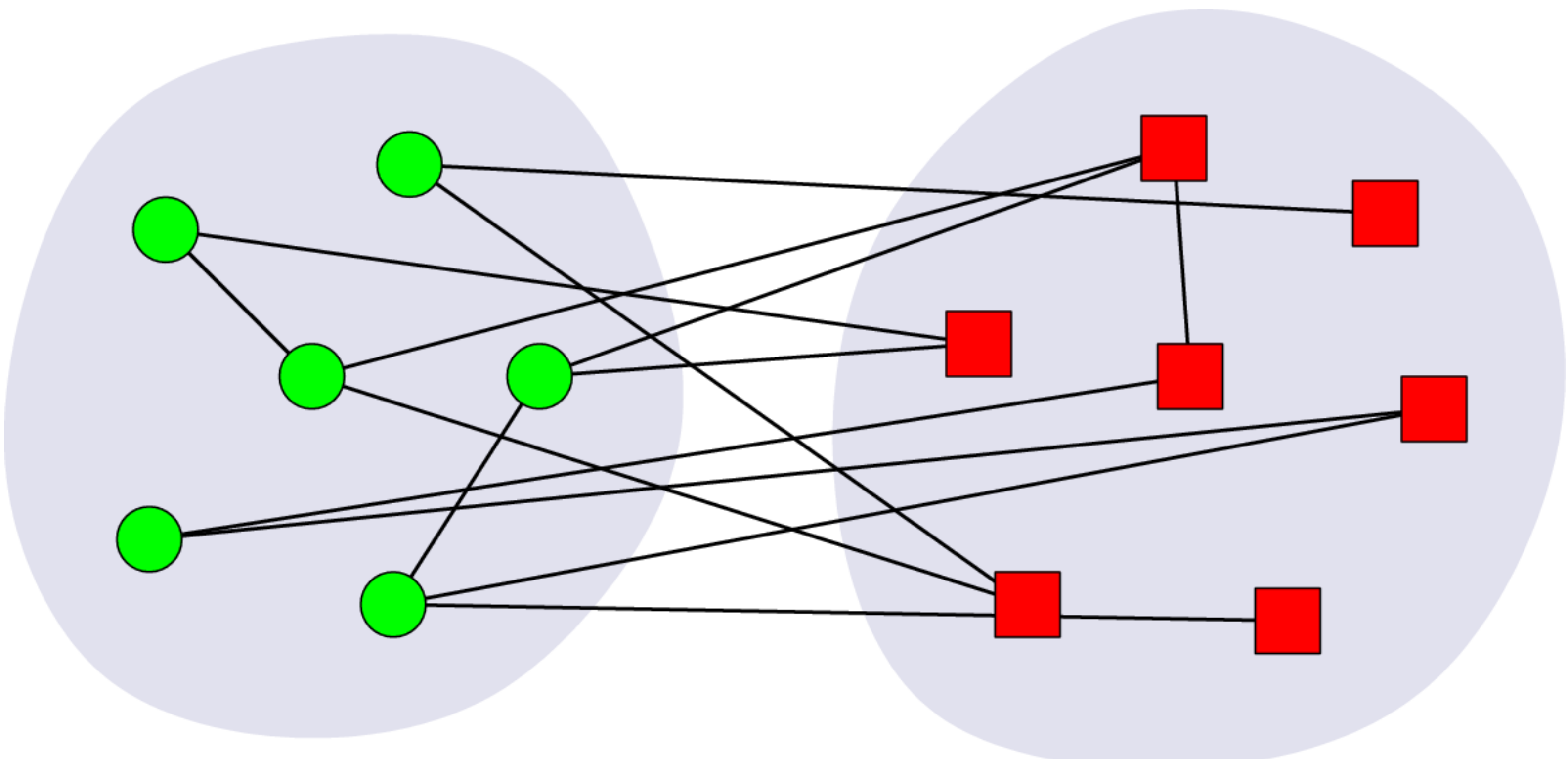
Example: Collaboration network



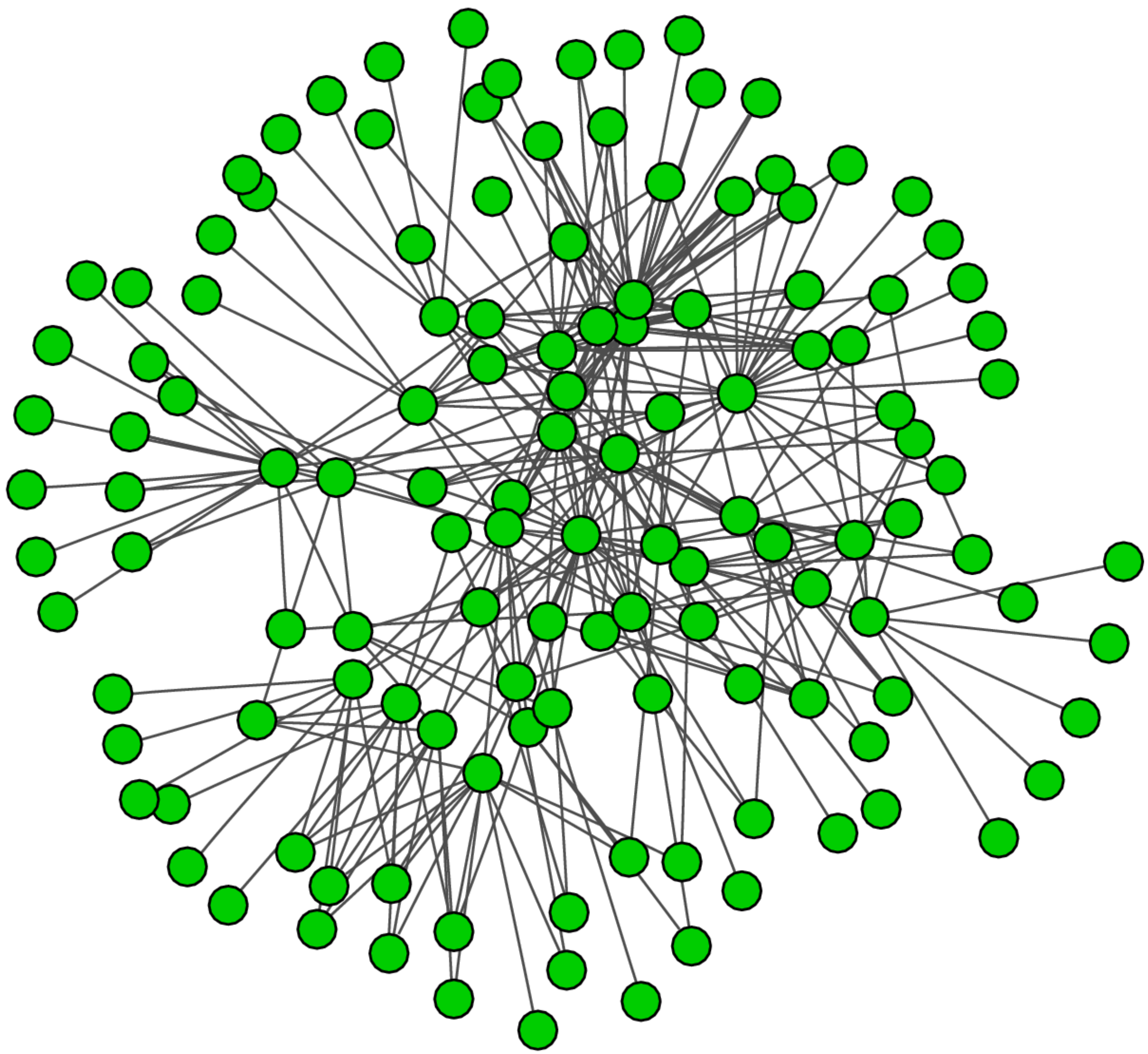
Negative eigenvalues

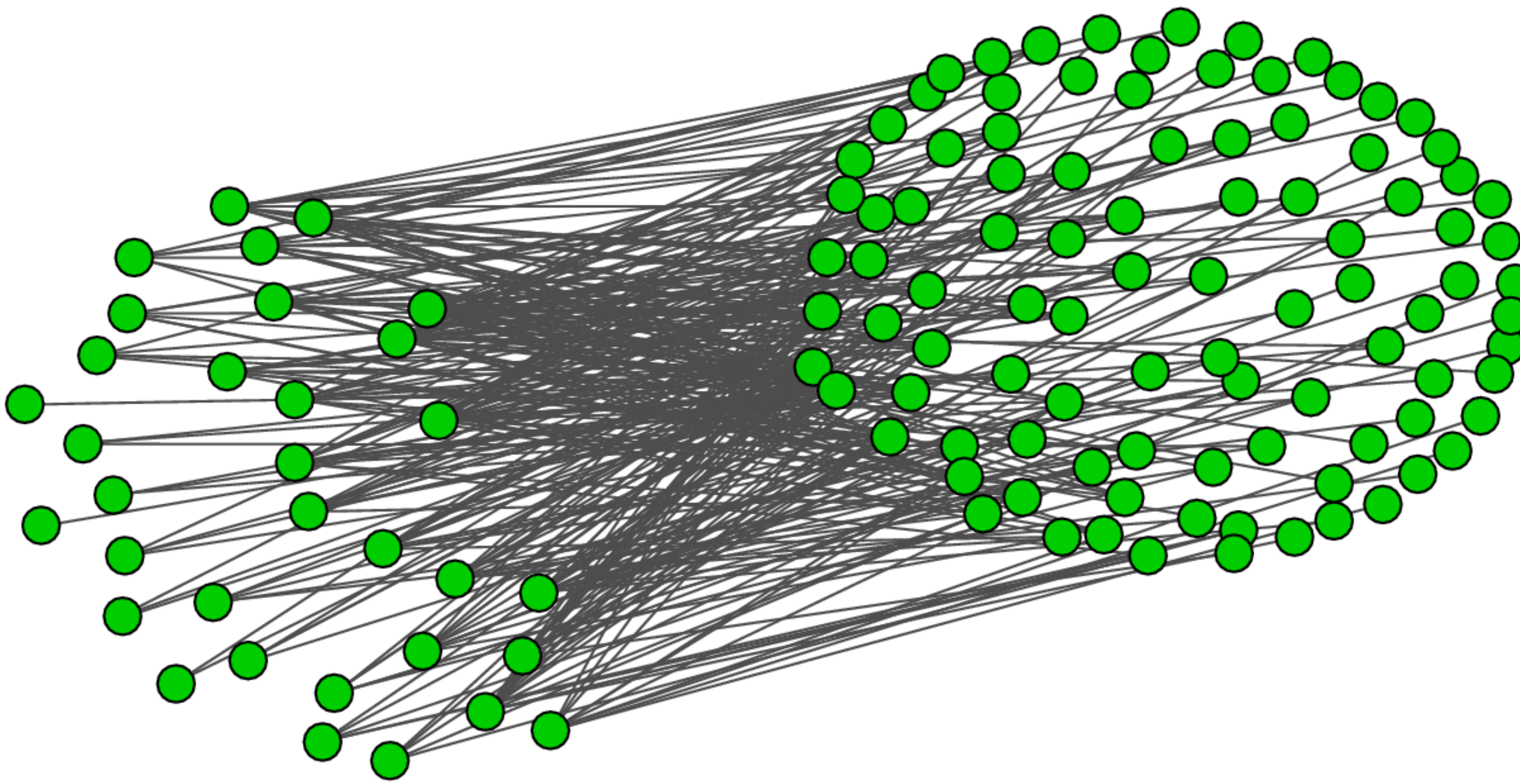
- Unlike the Laplacian, the modularity matrix has negative eigenvalues
- These tell us about *minimization* of the modularity
- A division with negative modularity has *fewer* edges than expected within communities (or more than expected between communities)

- This corresponds to a network with bipartite structure
- Or *k*-partite in the general case

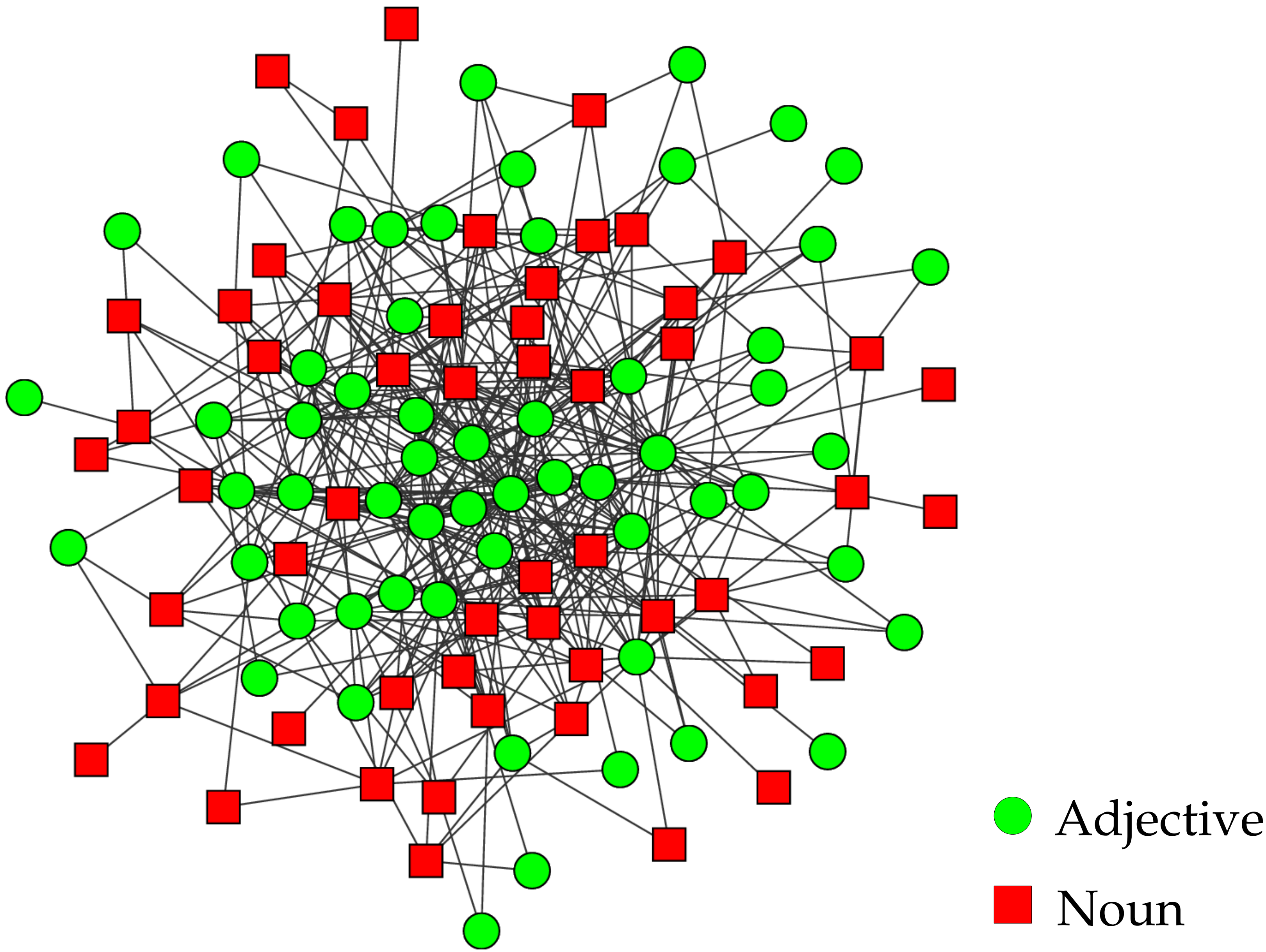


Example: Food web of hosts and parasites

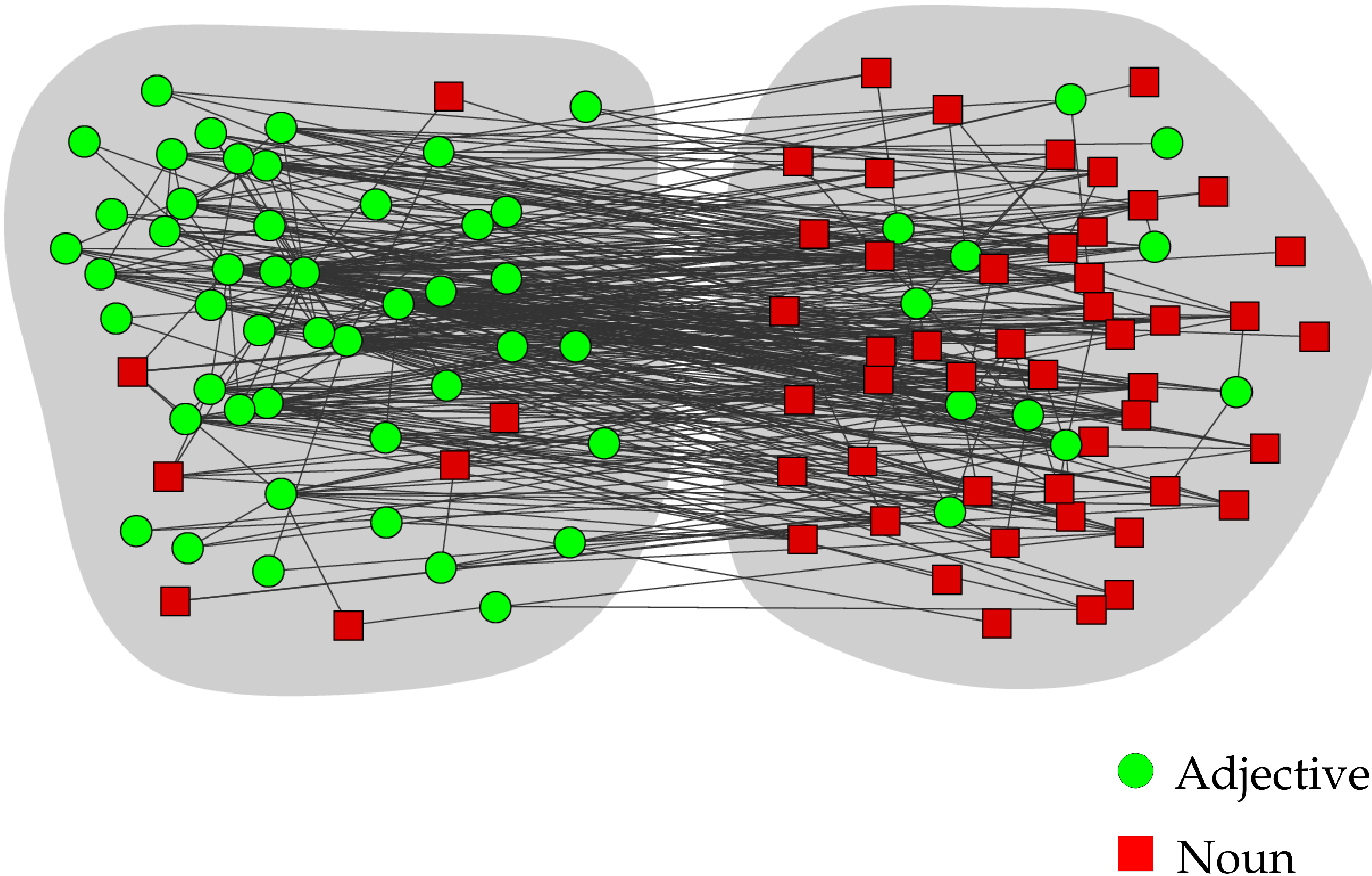




Network of word adjacencies



Network of word adjacencies



Summary

- Modularity maximization appears to be a highly competitive approach to community detection in networks
- It can be formulated as a spectral optimization problem, which leads to fast and accurate algorithms
- There are close connections between the spectrum of the modularity matrix and the community structure
- These can lead to some useful measures, such as centrality measures and methods for detecting bipartite and k-partite structure

References

- This work:
  - M. E. J. Newman, preprint physics/0602124 (*PNAS*, in press)
  - M. E. J. Newman, preprint physics/0605087
- Other relevant papers:
  - M. E. J. Newman, *Phys. Rev. E* **67**, 026126 (2003)
  - M. E. J. Newman and M. Girvan, *Phys. Rev E* **69**, 026113 (2004)
  - L. Danon *et al.*, *J. Stat. Mech.* P09008 (2005)
  - R. Guimerà and L. A. N. Amaral, *Nature* **433**, 895-900 (2005)
  - M. Huss and P. Holme, preprint q-bio/0603038