# Modularity

CMSC 858L

## Module-detection for Function Prediction

- Biological networks generally modular (Hartwell+, 1999)
- We can try to find the modules within a network.
- Once we find modules, we can look at over-represented functions within a module, e.g.:
  - If a majority of the proteins within a module have annotation A, predict annotation A for the other proteins in the module.
- $\Rightarrow$ Graph clustering methods
  - Min Multiway Cut, Graph Summarization, VI-Cut: examples we've already seen.
  - Methods often borrowed from other "community detection" applications.

# *Modularity*

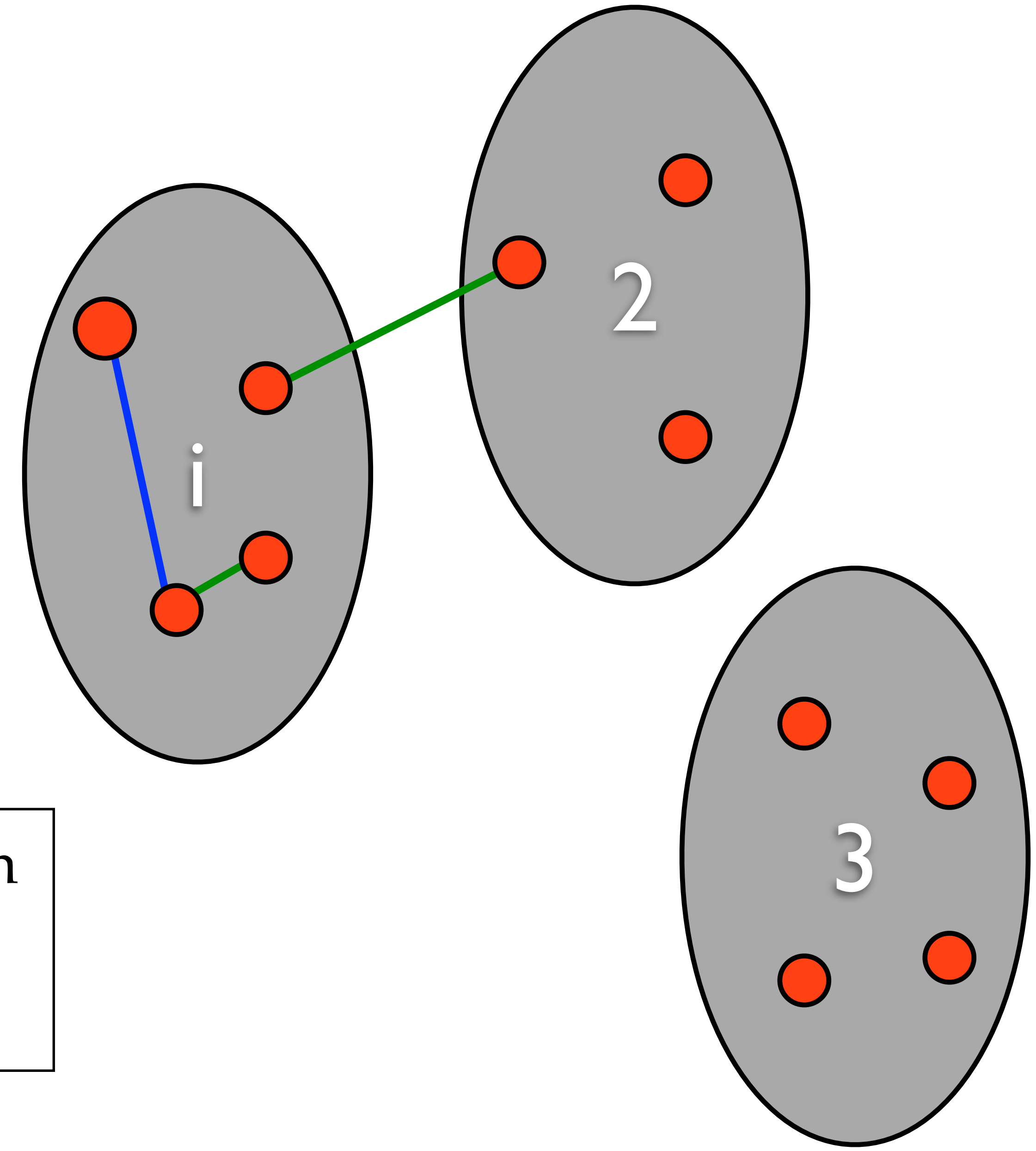
# Modularity

$e_{ii}$ = % edges in module i

$e_{ii}$=|{(u,v) : u ∈ $V_i$, v ∈ $V_i$, (u,v)∈E}| / |E|

$a_i$ = % edges with at least 1 end in module i

$a_i$=|{(u,v) : u ∈ $V_i$, (u,v) ∈ E}| / |E|




Modularity is:

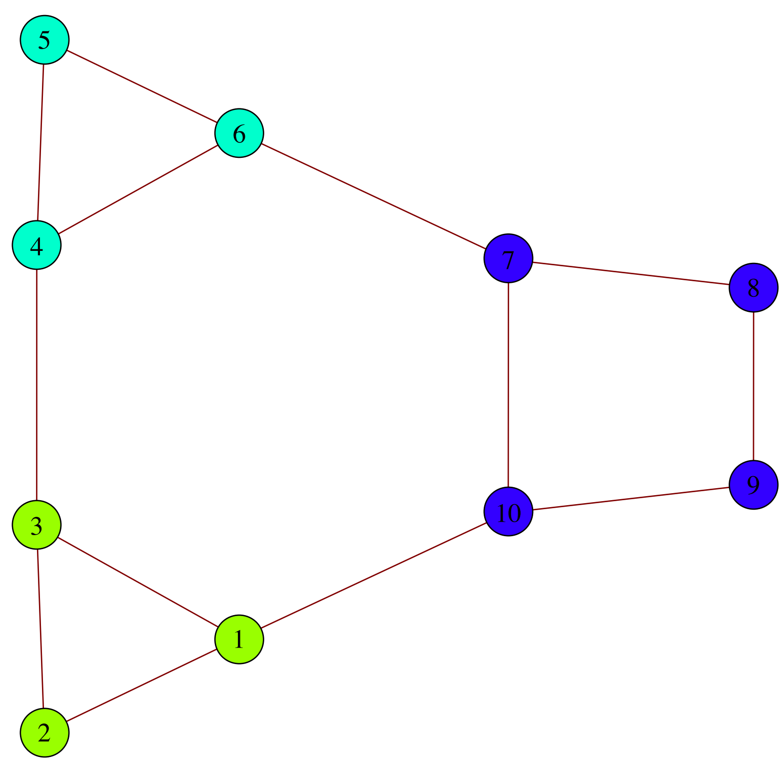
$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2)$$

probability a random edge would fall into module i

probability edge is in module i

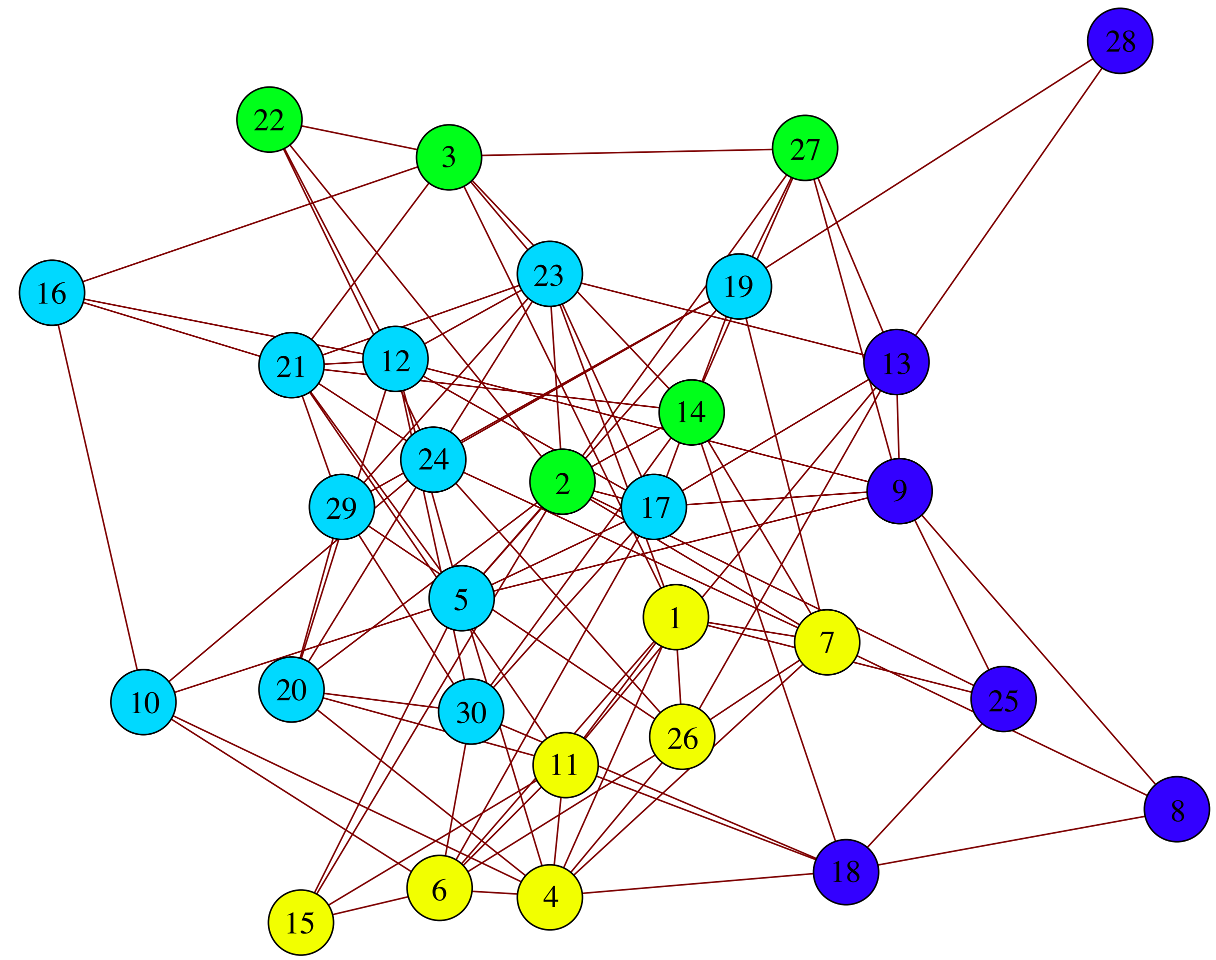
High modularity ⇒ more edges within the module that you expect by chance.

# Examples




Communities Assigned to a small graph

Note: maximizing modularity will find it's own # of clusters




Communities assigned to a random graph

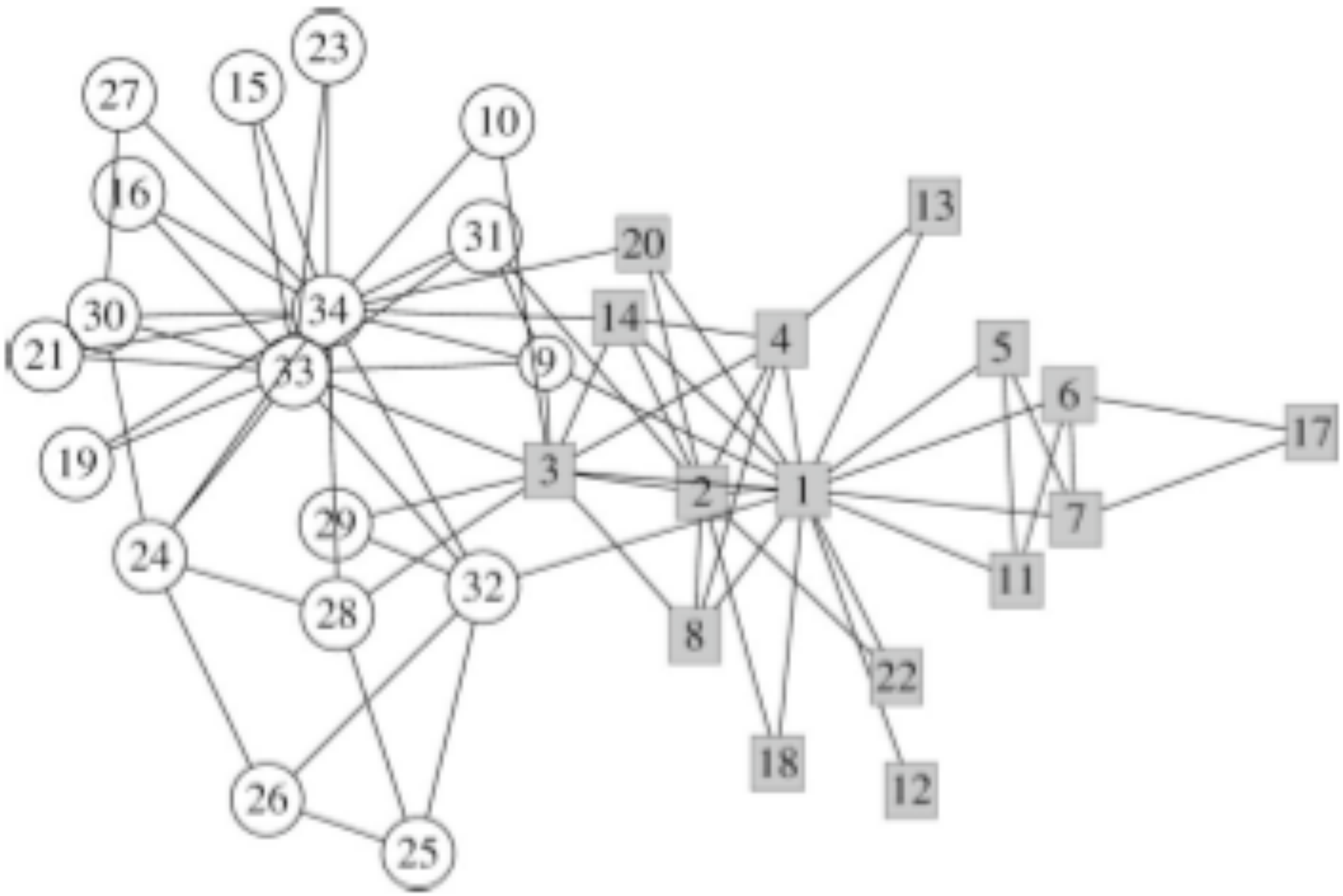
# Modularity Algorithm #1

- Modularity is NP-hard to optimize (Brandes, 2007)

- Greedy Heuristic: (Newman, 2003)
  - C = trivial clustering with each node in its own cluster
  - Repeat:
    - Merge the two clusters that will increase the modularity by the largest amount
    - Stop when all merges would reduce the modularity.

# Karate Club (again)

Newman-Girvan, 2004




Only 3 is in the “wrong” community.

# *Maximizing Modularity via a Spectral Technique*

# Another View of Modularity

normalization

adjacency matrix

probability a random edge would go between i and j

$$Q = \frac{1}{4m} \sum_{\substack{i,j \\ \text{in same module}}} \left( A_{ij} - \frac{k_i k_j}{2m} \right)$$

m = # edges in graph
$k_i$ = degree(i)

Consider the case of only 2 modules.

Let $s_i$ = 1 if node i is in module 1; -1 if node i is in module 2

$$\begin{aligned} Q &= \frac{1}{4m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j + 1) \\ &= \frac{1}{4m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j \end{aligned}$$

# Goal: Maximize modularity

- Try to find $\pm 1$ vector $\mathbf{s}$ that maximizes the modularity.
- Start with the case above: only two groups.
- Then show how to extend to $\geq 2$ groups.
- Will use some ideas from linear algebra.

$$Q = \frac{1}{4m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j$$

$$= \frac{1}{4m} \mathbf{s}^T B \mathbf{s}$$

s is a {-1,1} membership vector

"modularity" matrix

Let $u_i$ ($i = 1,\ldots,n$) be the eigenvectors of matrix B with eigenvalue $\beta_i$ for vector $u_i$. (Assume $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4 \geq \ldots \geq \beta_n$)

Write s as:

$$\mathbf{s} = \sum_i a_i u_i$$

where:

$$a_i = u_i^T \mathbf{s}$$

$$\mathbf{s} = \sum_i a_i u_i \qquad a_i = u_i^T \mathbf{s}$$

$$\begin{aligned} Q &= \frac{1}{4m}\mathbf{s}^T B \mathbf{s} \\ &= \left(\sum_i a_i u_i^T\right) B \left(\sum_j a_j u_j\right) \\ &= \left(\sum_i a_i u_i^T B\right) \left(\sum_j a_j u_j\right) \\ &= \sum_i \sum_j a_i a_j u_i^T B u_j \end{aligned}$$

drop the (1/4m) →

Note:

1. $Bu_j = \beta_i u_j$
2. When $i \neq j$, $u_i^{\mathrm{T}} \mathrm{B} u_j = 0$ because $u_i \perp u_j$

$$Q = \sum_i (u_i^T \mathbf{s})^2 \beta_i$$

# To Maximize Q

$$Q = \sum_i (u_i^T \mathbf{s})^2 \beta_i$$

- If we were allowed to choose any **s** we'd pick the one that is parallel to $u_1$.
- **But:** $s_i$ must be +1 or -1.
  This is a severe restriction.
- **So:** maximize $u_1 \cdot \mathbf{s}$, the projection of s along vector $u_1$.
- To do this: choose $s_i = 1$ if $u_1 > 0$, and $s_i = -1$ if $u_1 \leq 0$.

# Subsequent Splits

The modularity if this module was split according to s

The modularity of module g as it stands now

$$Q = \frac{1}{2m}\left[\frac{1}{2}\sum_{i,j\in g} B_{ij}(s_i s_j + 1) - \sum_{i,j\in g} B_{ij}\right]$$

$$= \frac{1}{2}\sum_{i,j\in g} B_{ij}s_i s_j + \frac{1}{2}\sum_{i,j\in g} B_{ij} - \sum_{i,j\in g} B_{ij}$$

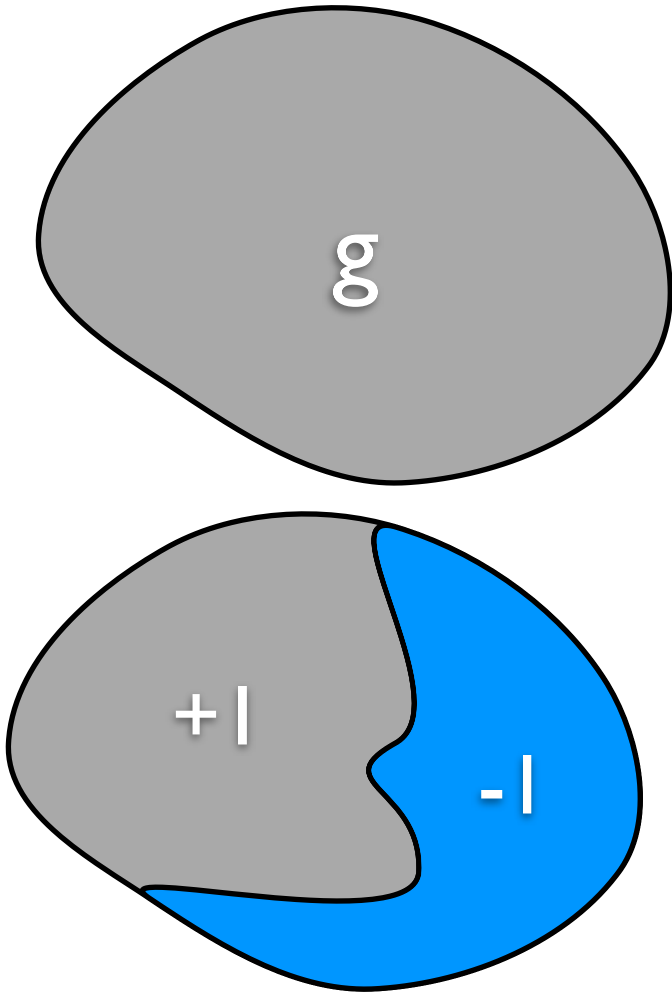
$$= \frac{1}{4m}\left[\sum_{i,j\in g} B_{ij}s_i s_j - \sum_{i,j\in g} B_{ij}\right]$$

$$\sum_{i,j\in g} B_{ij} = \sum_{i,j\in g} s_i s_j \delta_{i,j} \sum_{k\in g} B_{ik}$$

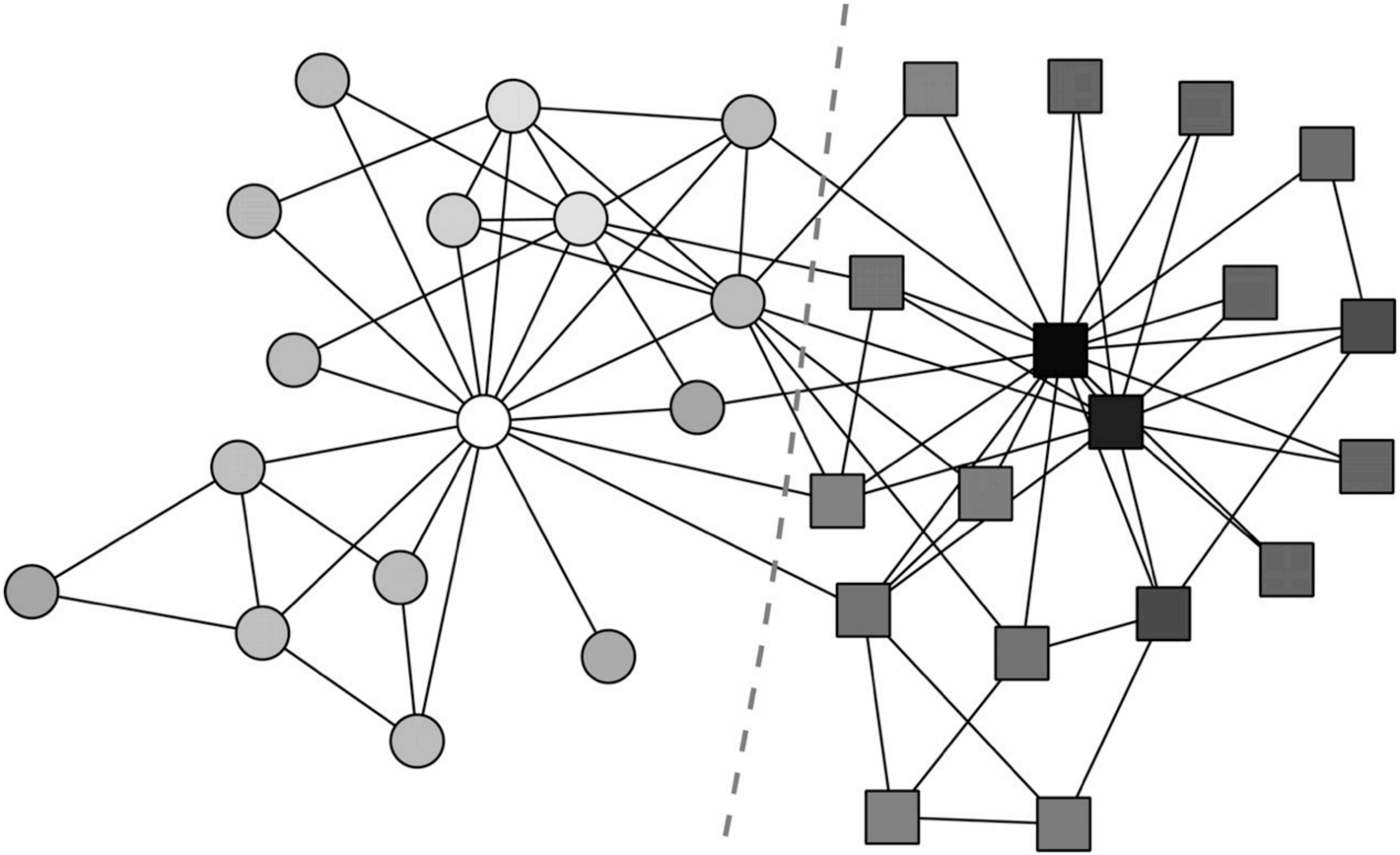
$$= \frac{1}{4m}\sum_{i,j\in g}\left[B_{ij} - \delta_{ij}\sum_{k\in g} B_{ik}\right]s_i s_j$$

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

$$= \frac{1}{4m}\mathbf{s}^T\mathbf{B}^{(g)}\mathbf{s},$$



# Karate Club Results: Exactly Right



(Newman, 2006)

# Greedy Improvement

- Given a partition of the network
- Repeat:
  - Find the vertex that would yield the largest modularity increase if it were moved into a different community AND that has not yet been moved
  - Move the vertex into that new community
- Return the best partitioning ever observed

largest increase might be negative

Similar to the Kernighan-Lin graph partitioning heuristic (details in a few slides)

# Additional Results

Girvan-Newman (betweenness)

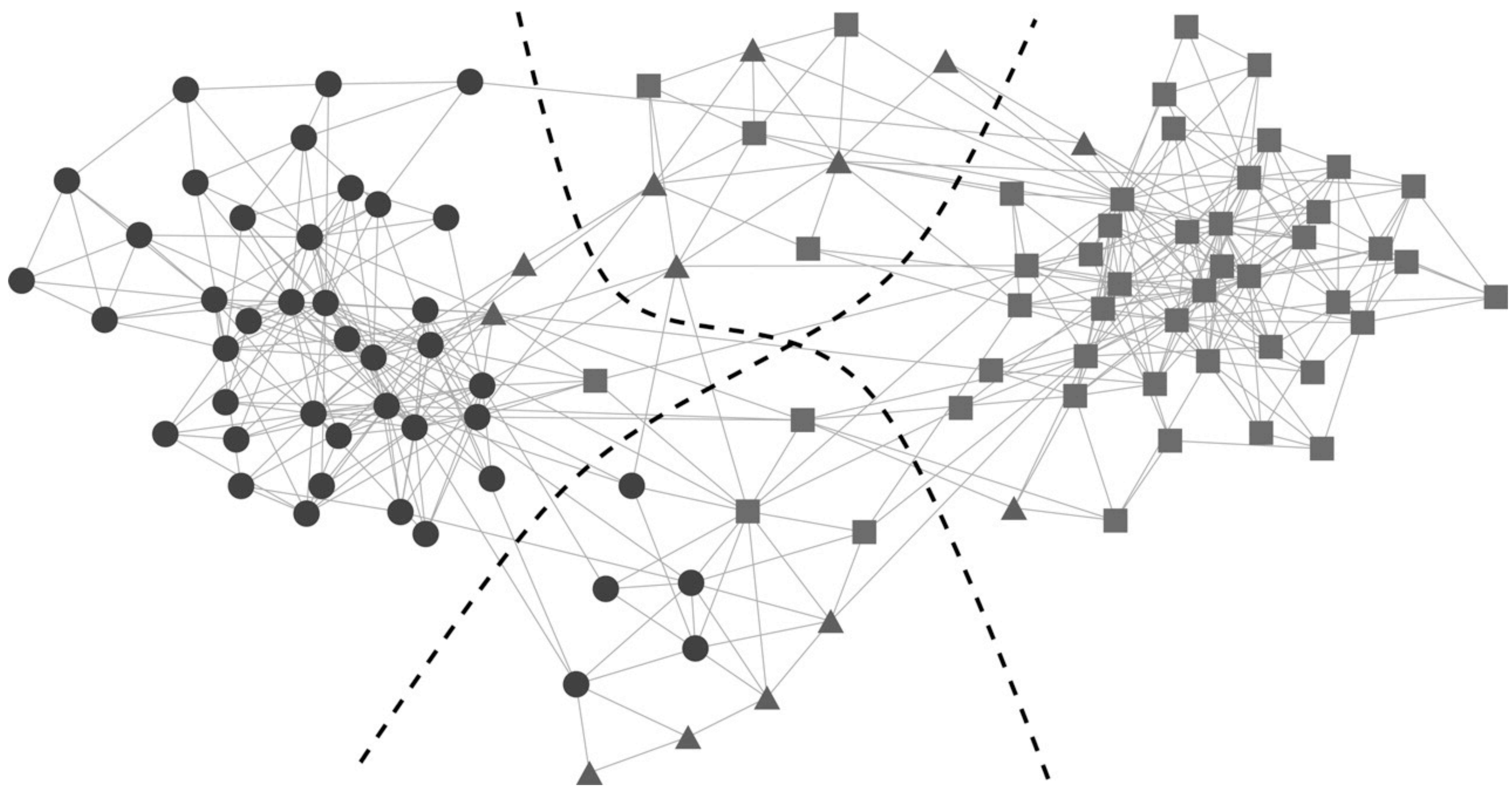
Newman Spectral

| Network | Size *n* | Modularity *Q* | | | |
|---|---|---|---|---|---|
| | | GN | CNM | DA | This article |
| Karate | 34 | 0.401 | 0.381 | 0.419 | 0.419 |
| Jazz musicians | 198 | 0.405 | 0.439 | 0.445 | 0.442 |
| Metabolic | 453 | 0.403 | 0.402 | 0.434 | 0.435 |
| E-mail | 1,133 | 0.532 | 0.494 | 0.574 | 0.572 |
| Key signing | 10,680 | 0.816 | 0.733 | 0.846 | 0.855 |
| Physicists | 27,519 | – | 0.668 | 0.679 | 0.723 |

Greedy Hierarchical

Newman, 2006

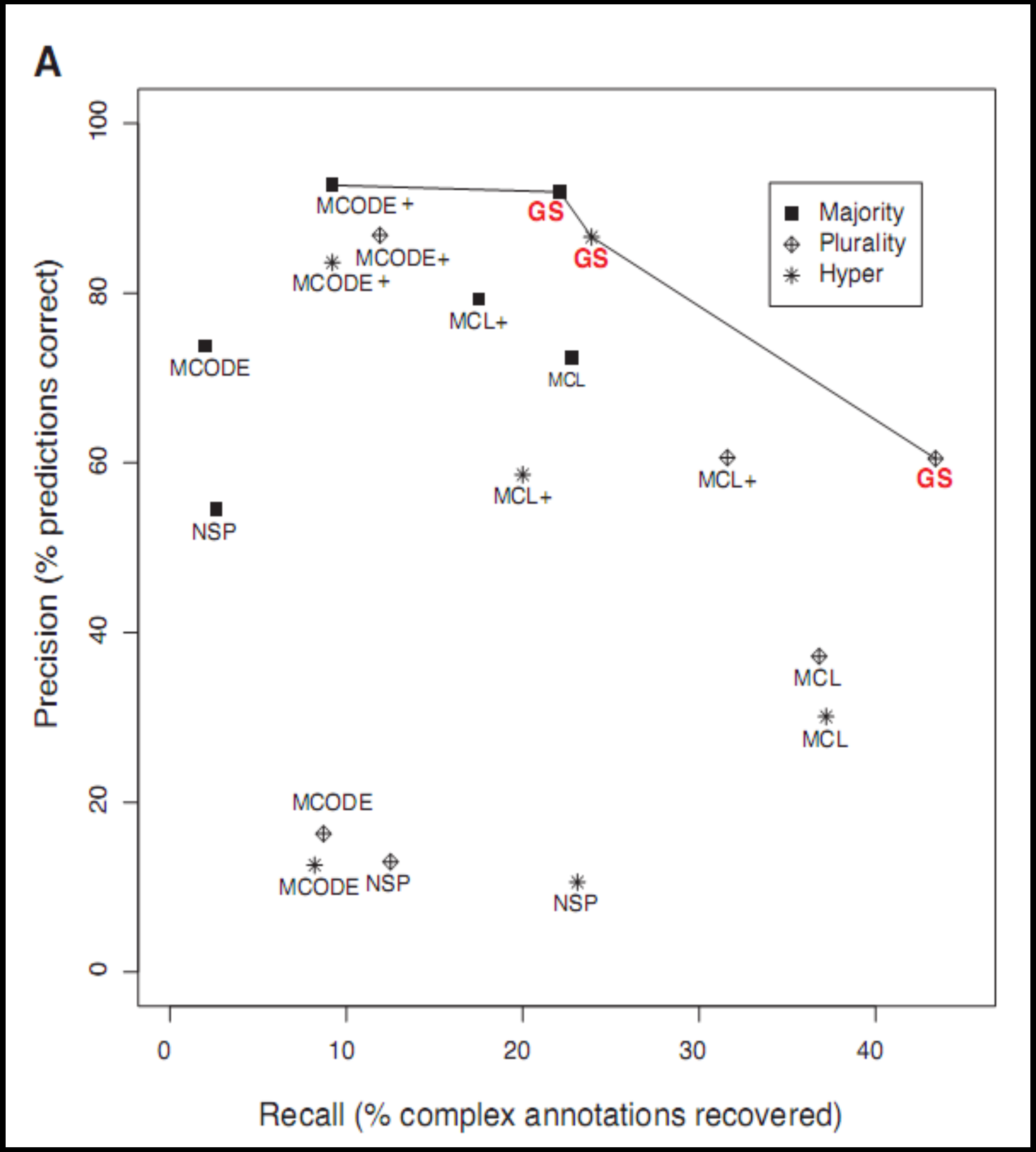
# Krebs Political Books



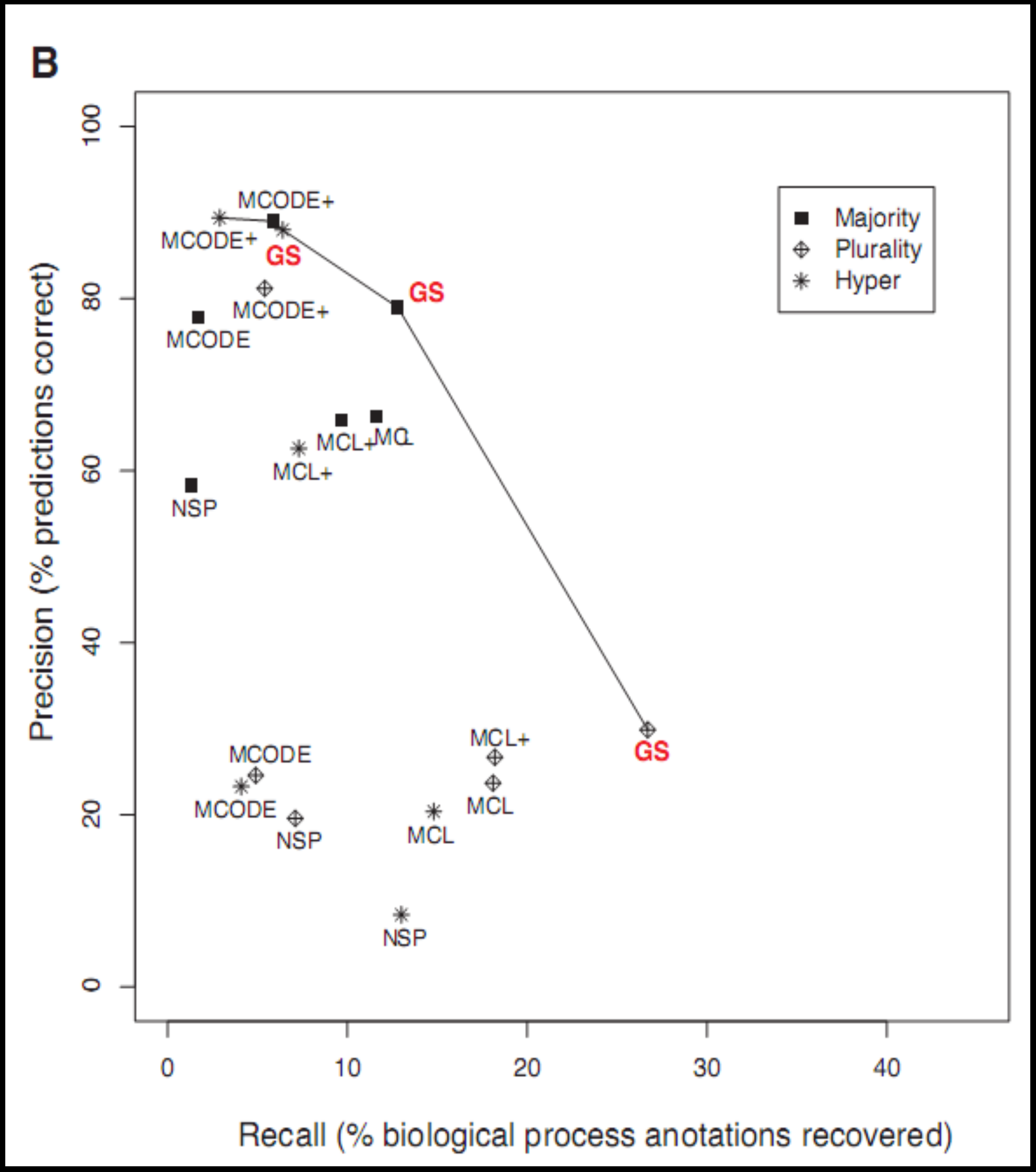
Nodes = political books; shape = conservative (squares) / liberal (circles) / "centrist" (triangles)

Edges = books frequently bought by the same readers on Amazon.com

# Complexes




# Biological Processes




"+" indicates parameters tuned to maximize precision

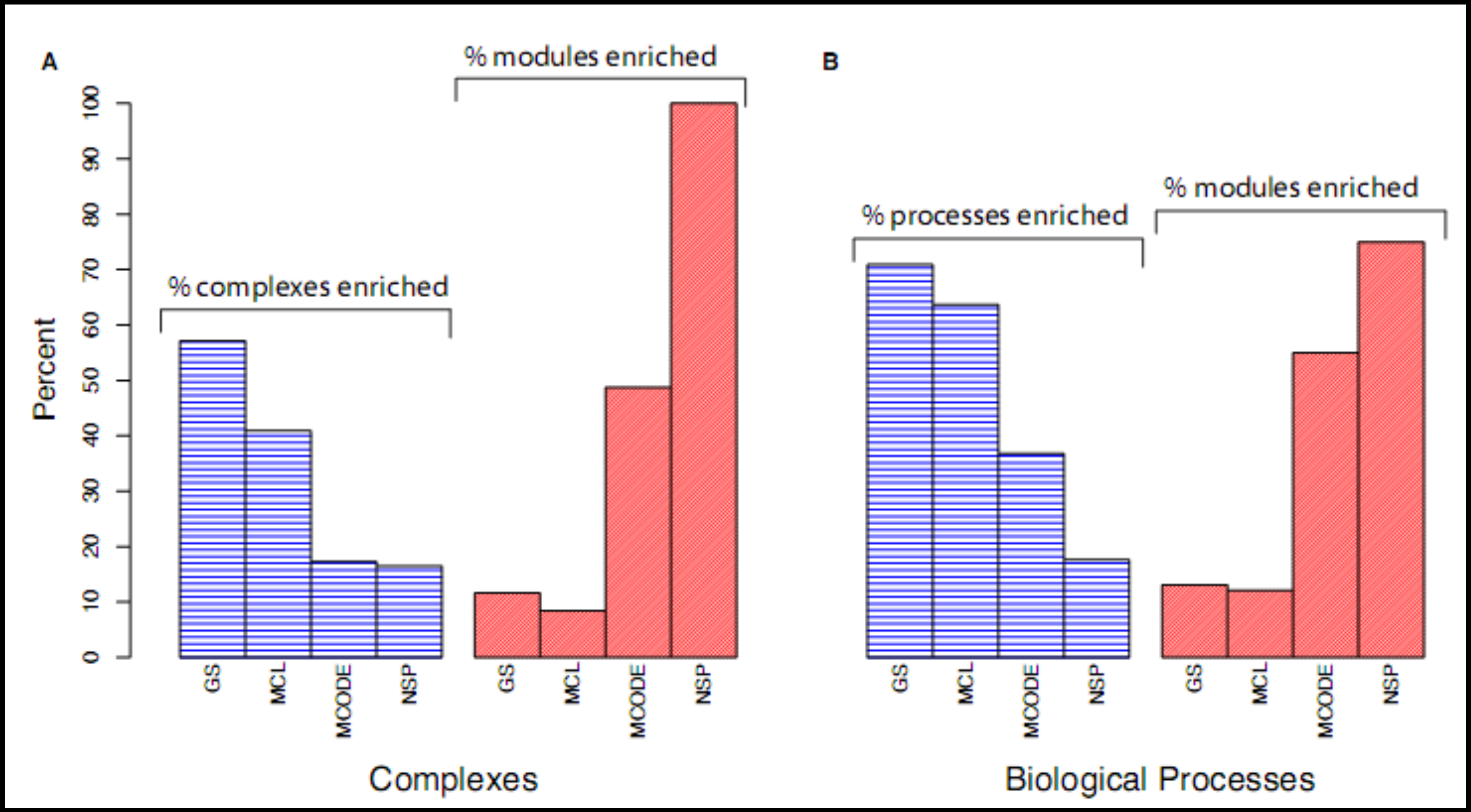
## All GS predictions are Pareto optimal

Many unique predictions made by each algorithm

# % Modules Enriched

A lower % of GS modules are enriched for some annotation, but not indicative of  predictive performance.

"Easy" to get legitimate statistical significant enrichment.



# Summary: Modularity

- Modularity is widely used as a measure for how good a clustering is.
- Particularly popular in social network analysis, but used in other contexts as well (e.g. Brain networks).
- Has a “resolution” preference: for a given network, will tend to prefer clusters of a particular size.
- Often this means the clusters are too big.
- A good example of where a spectral clustering technique can work.