Robust Clustering Techniques in Bioinformatics

Rob Beverly

18.417 Fall 2004

Why Clustering?

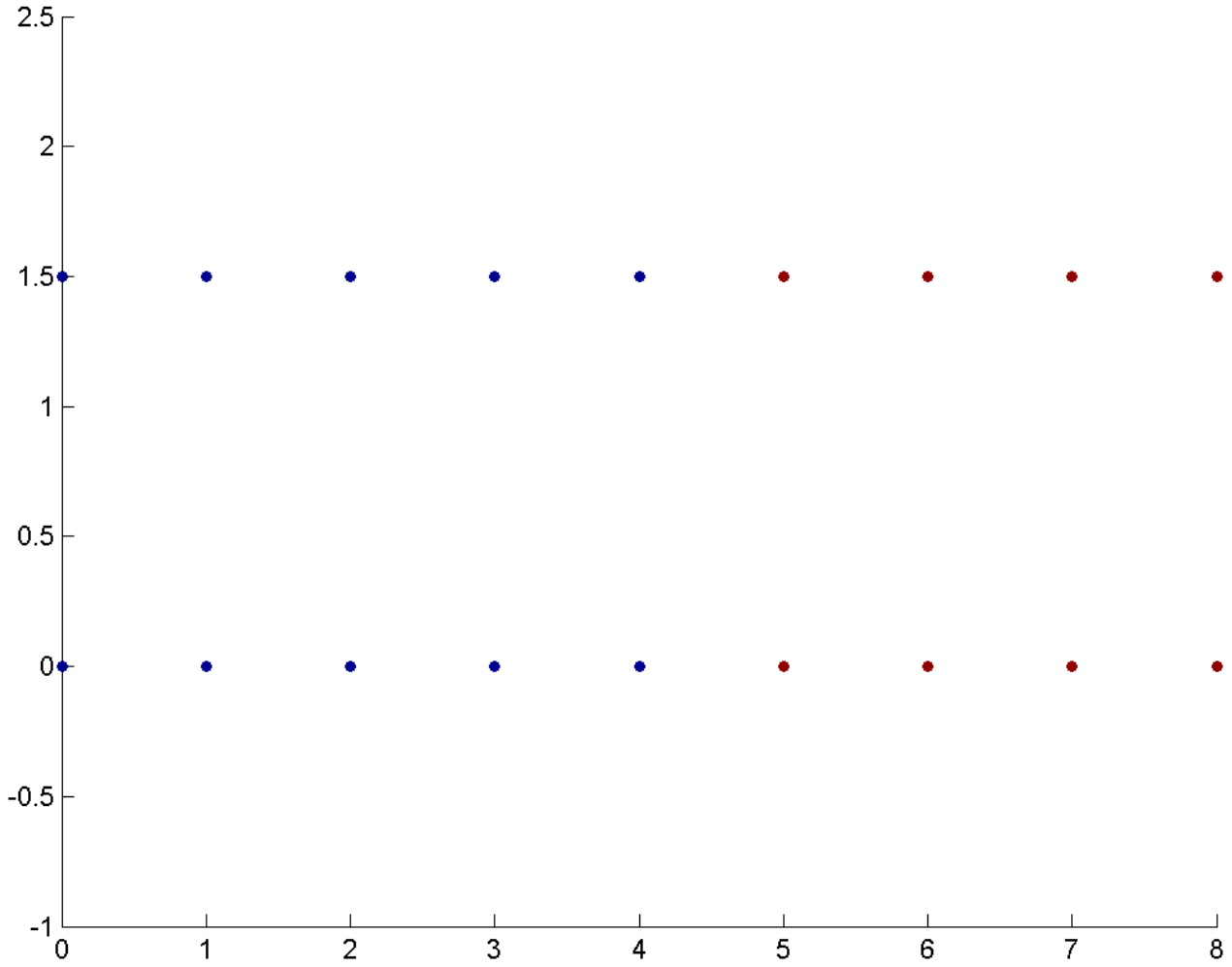
- Class Discovery
  - Given just the data, can one find inherent classes/clusters
- Class Prediction
  - Given an existing clustering, predict class of new elements

k-Means Clustering Bad

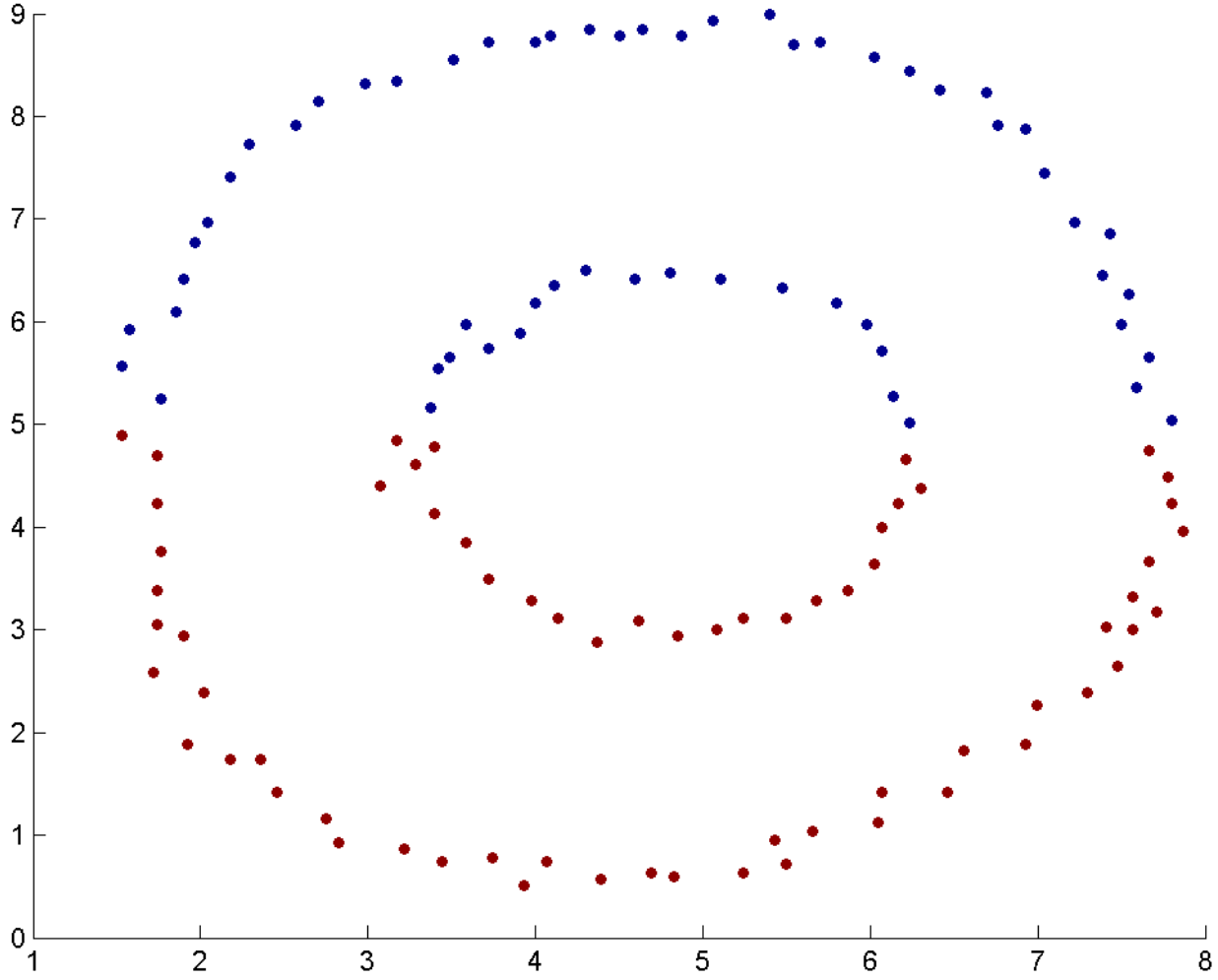
- k-Means clustering often used
- Simple
- Fast

- Centroids force spherical interpretation of the data
- Easy to construct degenerate examples:

k-2 Clustering



k-2 Clustering



Spectral Clustering

- High-level:
  - Construct a neighbor graph
    - k-nearest neighbor
    - threshold
  - Assign weights to edges
  - Define transition probability over edges
  - Cluster based on eigenvectors of probability matrix

Spectral Clustering

- Assign weights based on Euclidian distance in d-dimensional space with exponential fall-off:

  If an edge exists between vertices i and j in the graph, then assign weight:

  $$W_{ij} = \exp\{-\beta \|x_i - x_j\|\}$$

Spectral Clustering

- Define a Markov random walk over the graph by normalizing edge weights to form transition probabilities
- Let D be a diagonal matrix with elements D_{ii} equal to the sum of weights for node I
- Then:

  $P = D^{-1}W$
- And:

$$P_{ij} = \frac{W_{ij}}{\sum_j W_{ij}}$$

Spectral Clustering

- Distribution of points after t random steps converges as t increases
- If graph is connected and ergodic, the distribution becomes independent of starting point
- Recover this effect from the eigenvectors

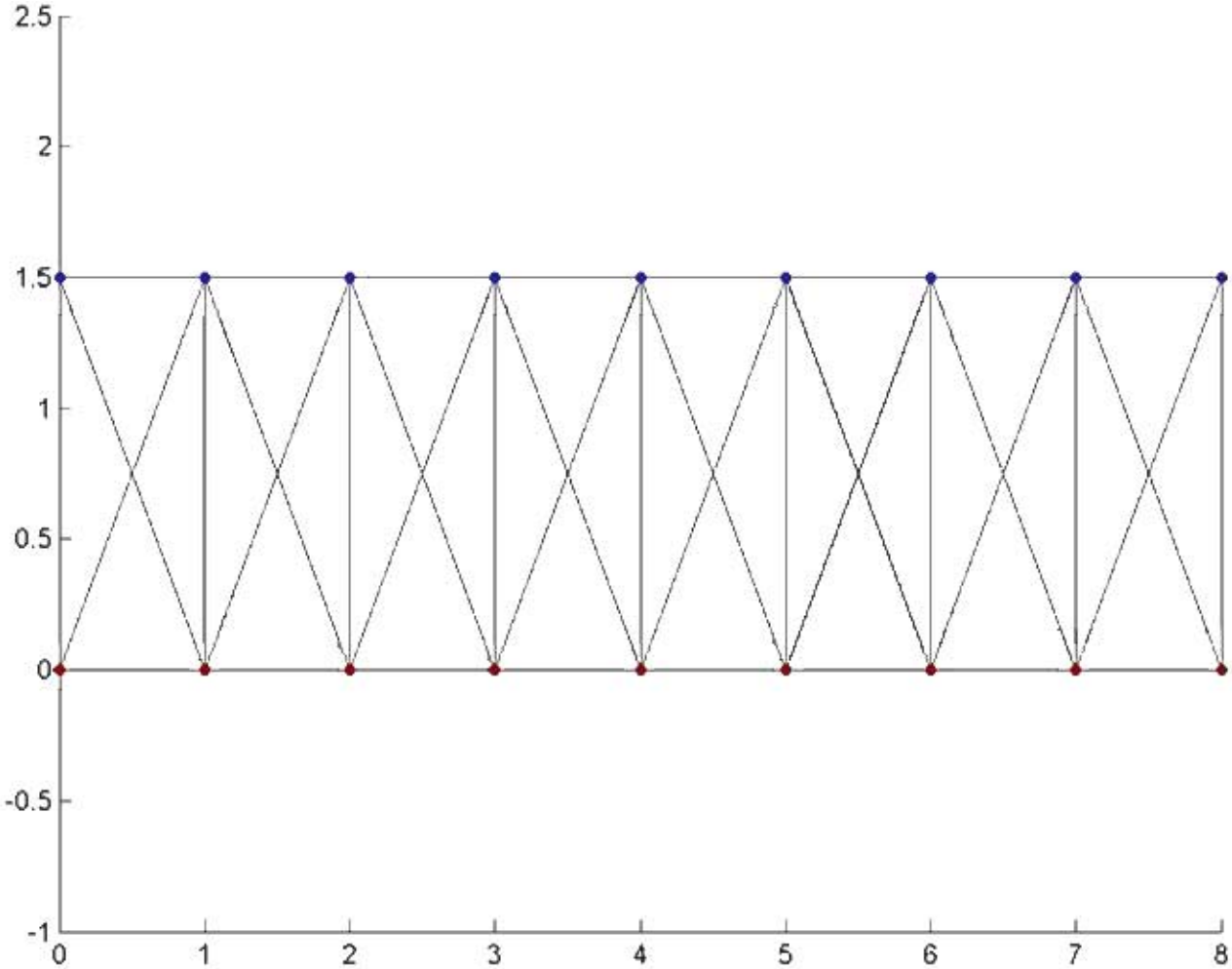
Spectral Clustering

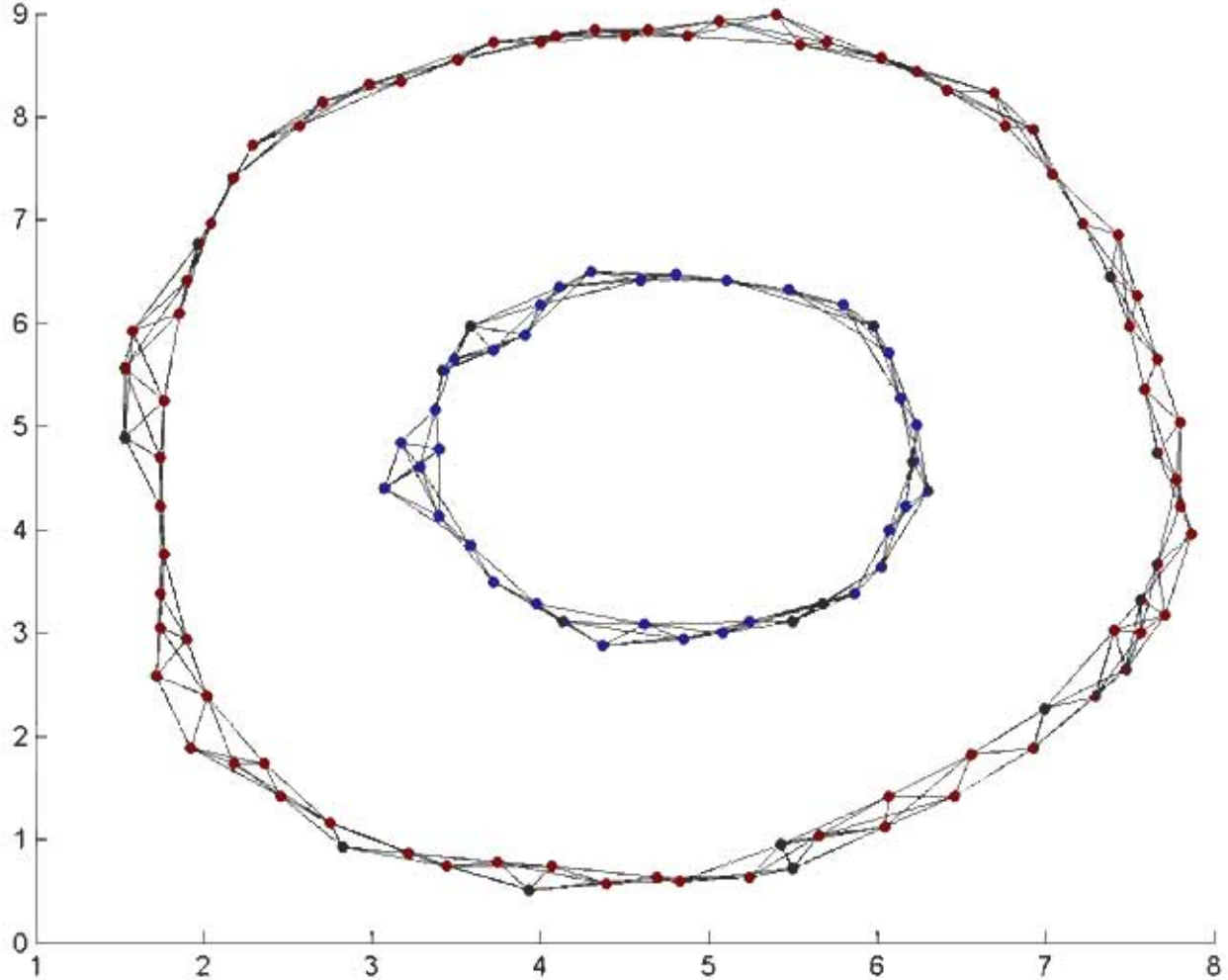
- Computing random walk:
- Find eigenvectors corresponding to second largest eigenvalue (largest correction to asymptotic limit) of either:
- Stochastic matrix:

  $P = D^{-1}W$

- Laplacian:

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = D^{\frac{1}{2}} P^t D^{\frac{1}{2}}$$


2.5
2
1.5
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6
7
8


9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8

Research Questions

- 1: Does spectral clustering outperform traditional methods on real data sets

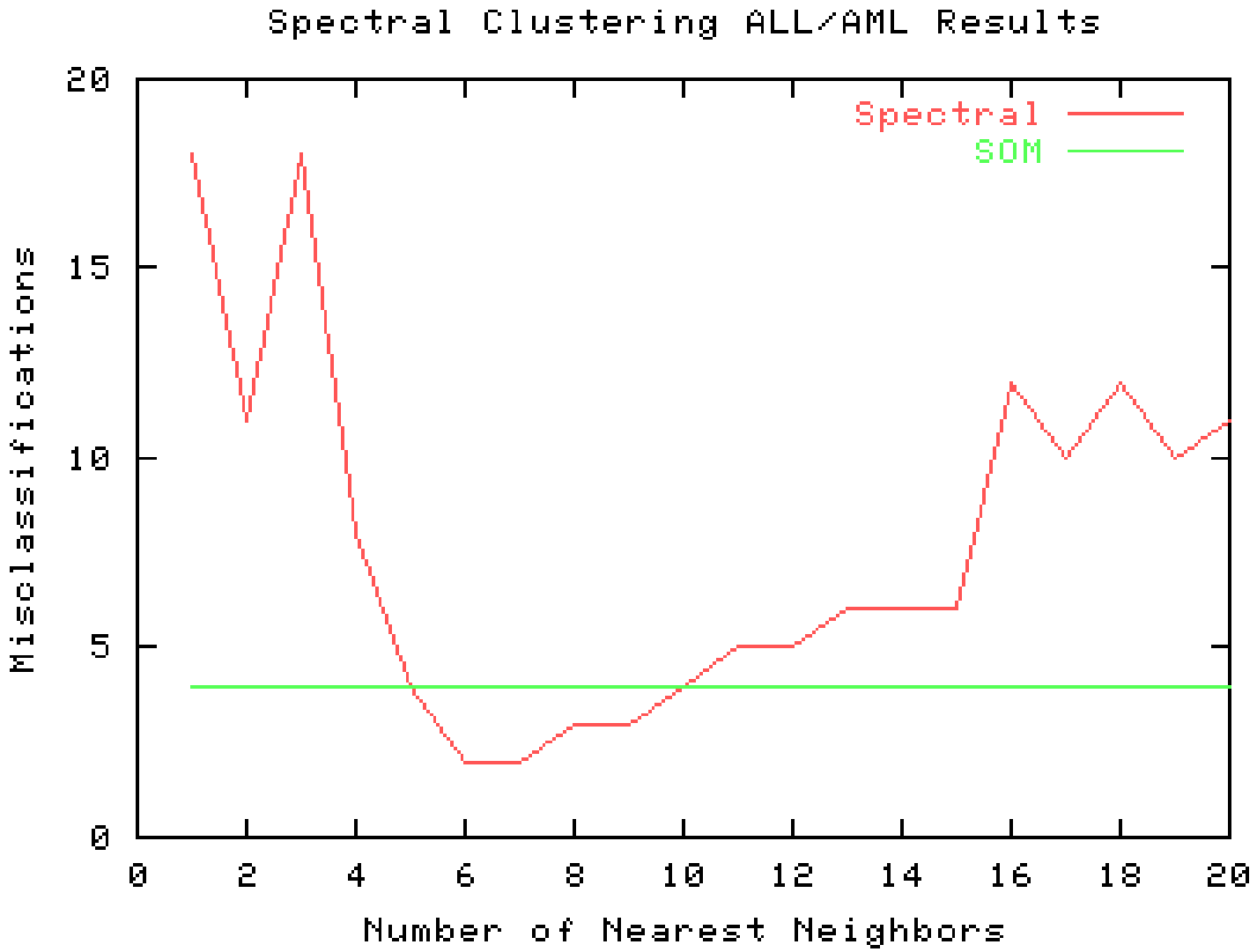
Canonical ALL/AML Dataset

- [Golub et al, 1999]
- Gene expression patterns (~7000) from Microarrays of 38 patients with leukemia
- Attractive because there are two inherent types of leukemia: ALL and AML
- Paper uses k-Means based Self-Organizing Maps (SOMs) to cluster

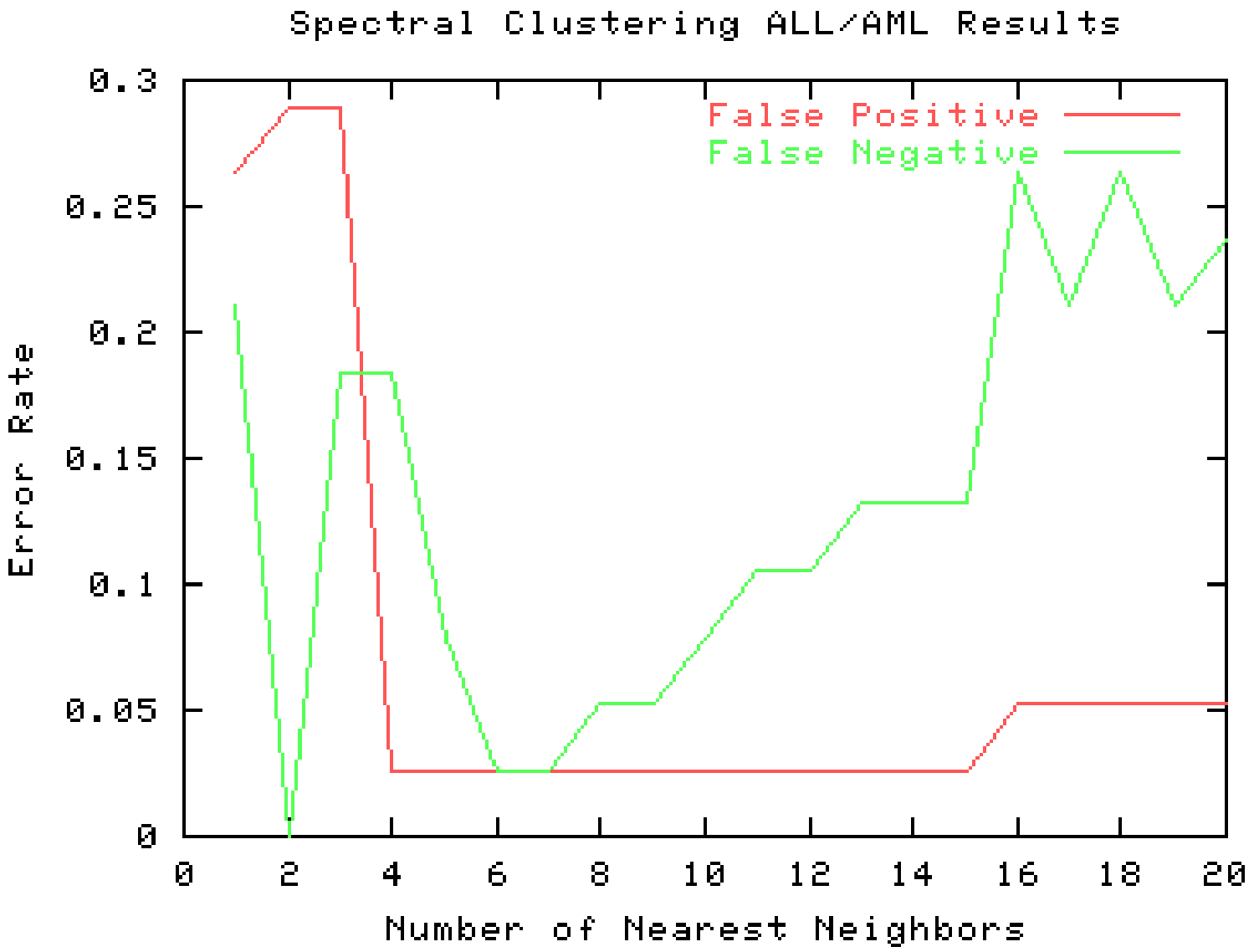
ALL/AML

- Golub Results:
  - Cluster 1: 24/25 ALL
  - Cluster 2: 10/13 AML
  - 1 False Positive, 3 False Negatives
  - Total 4 misclassifications: ~10%
- Does spectral clustering perform better?
  - Yes
  - 2 misclassifications

ALL/AML


Spectral Clustering ALL/AML Results
Spectral
SOM
Misclassifications
Number of Nearest Neighbors
0
5
10
15
20
0
2
4
6
8
10
12
14
16
18
20

ALL/AML


Spectral Clustering ALL/AML Results
False Positive
False Negative
Error Rate
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10
12
14
16
18
20
Number of Nearest Neighbors

Spectral Clustering

- Finding more than two clusters?
- Recursive
  - subdivide until correct number of clusters
- Multicut:
  - Find k eigenvectors corresponding to the k largest eigenvalues
  - Run k-means clustering on resulting matrix

Number of Clusters

- How can we know a priori the number of clusters in the data?
- Explored a divisive clustering algorithm [Newman 2003]

Divisive Clustering

- Start with k-nearest neighbors graph
- Compute all-pairs shortest paths
- Iterate until graph is empty:
  - Find edge e with largest number of SPs traversing it
  - Remove e
  - Compute modularity score Q
- Graph with highest modularity score is selected as representing the inherent clusters

Modularity Score

$$a_i = \sum_j e_{ij}$$

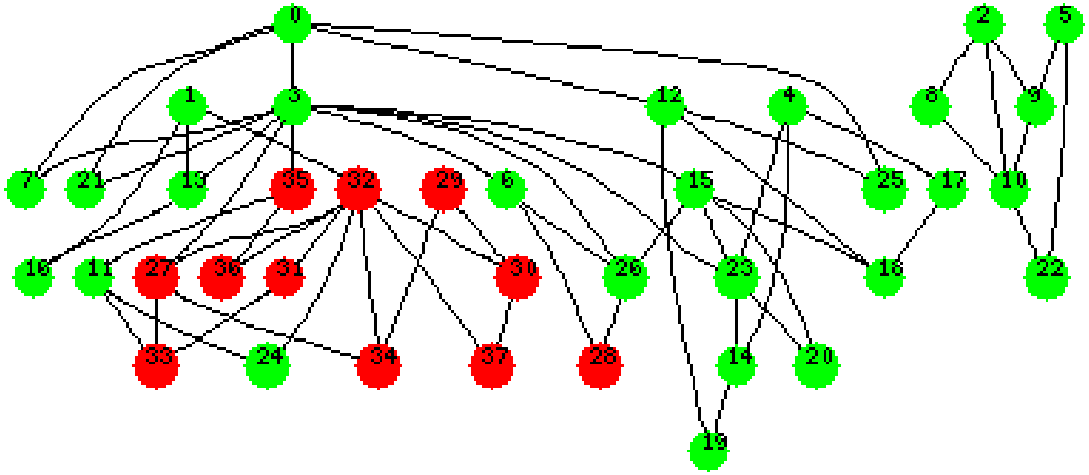
$$Q = \sum_i \left(e_{ii} - a_i^2 \right)$$

- e_{ij} is the fraction of edges from cluster i to cluster j
- Intuition: edges within a cluster minus expected value if edges fall at random
- Q=0 implies random number of within cluster edges

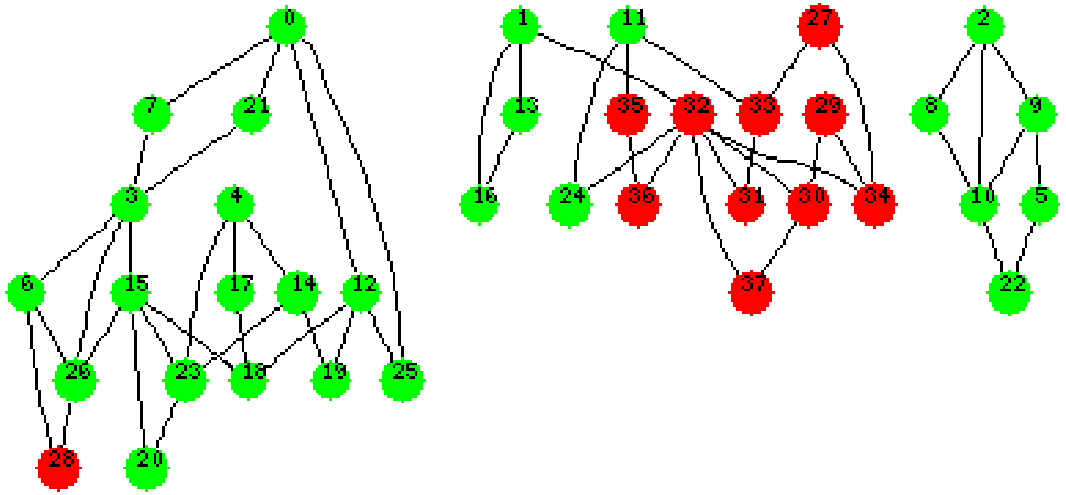
Research Questions

- 1: Does spectral clustering outperform traditional methods on real data sets
- 2: Can we infer the correct number of clusters

ALL/AML Divisive Clustering

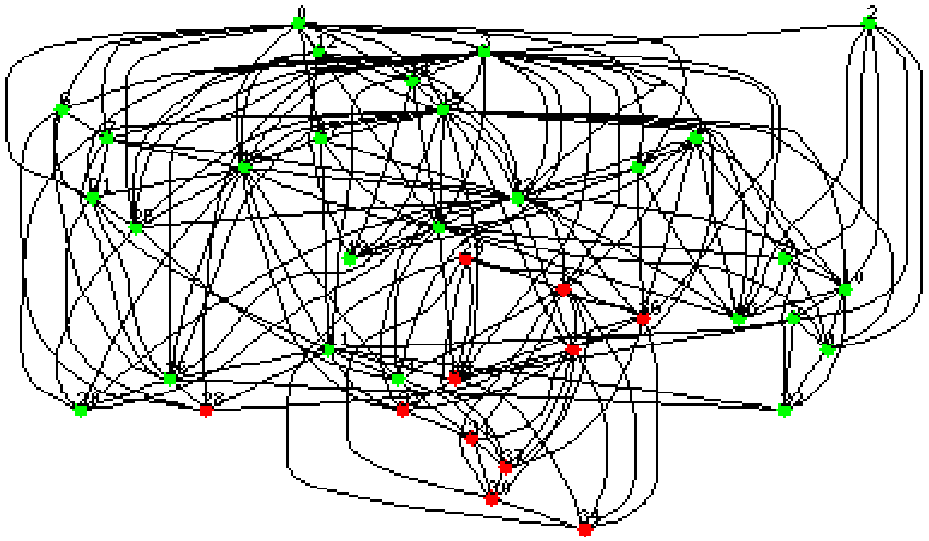


k=2 Nearest Neighbors



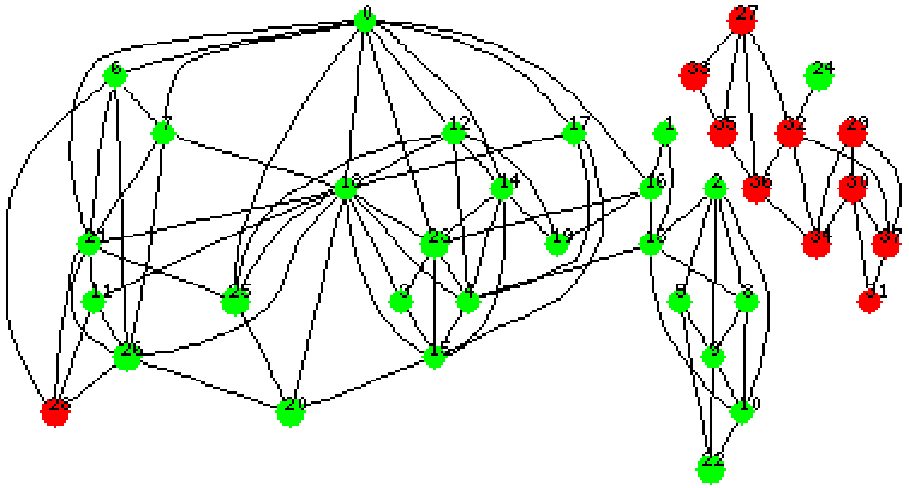
Resulting Clusters

ALL/AML Divisive Clustering



k=6 Nearest Neighbors

Resulting Clusters



Do our Results Generalize

- ALL/AML an older, well-studied data-set
- Relatively easy to do well on

- More recent:
- Gene expression-based classification of malignant gliomas [Nutt et al, 2003]

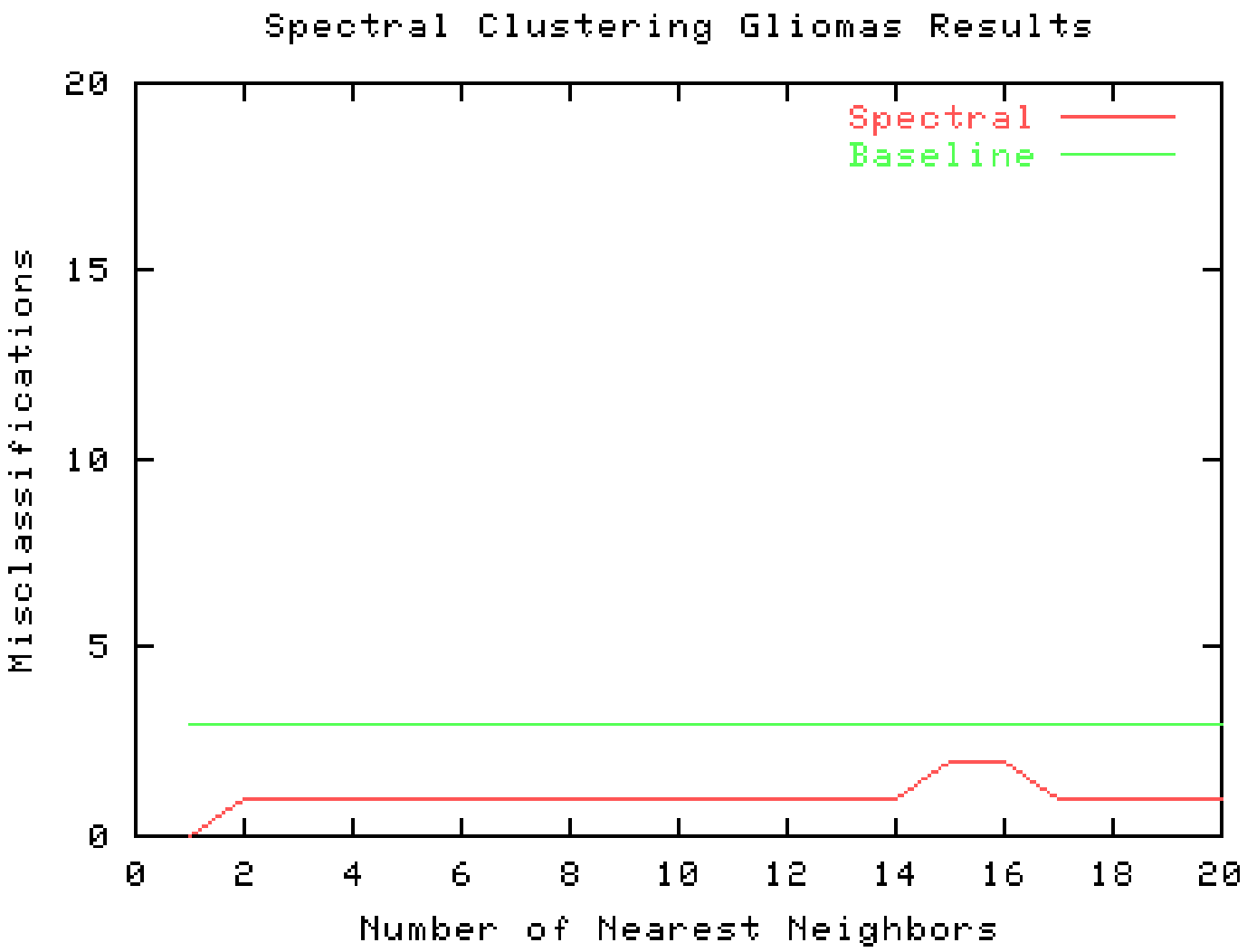
Malignant Gliomas

- Study two different brain cancers with different courses of treatment:
  - Glioblastomas
  - Anaplastic Oligodendrogliomas
- Distinguishing between them is "diagnostically challenging"
- Gene expression patterns (~12,000) from 50 gliomas

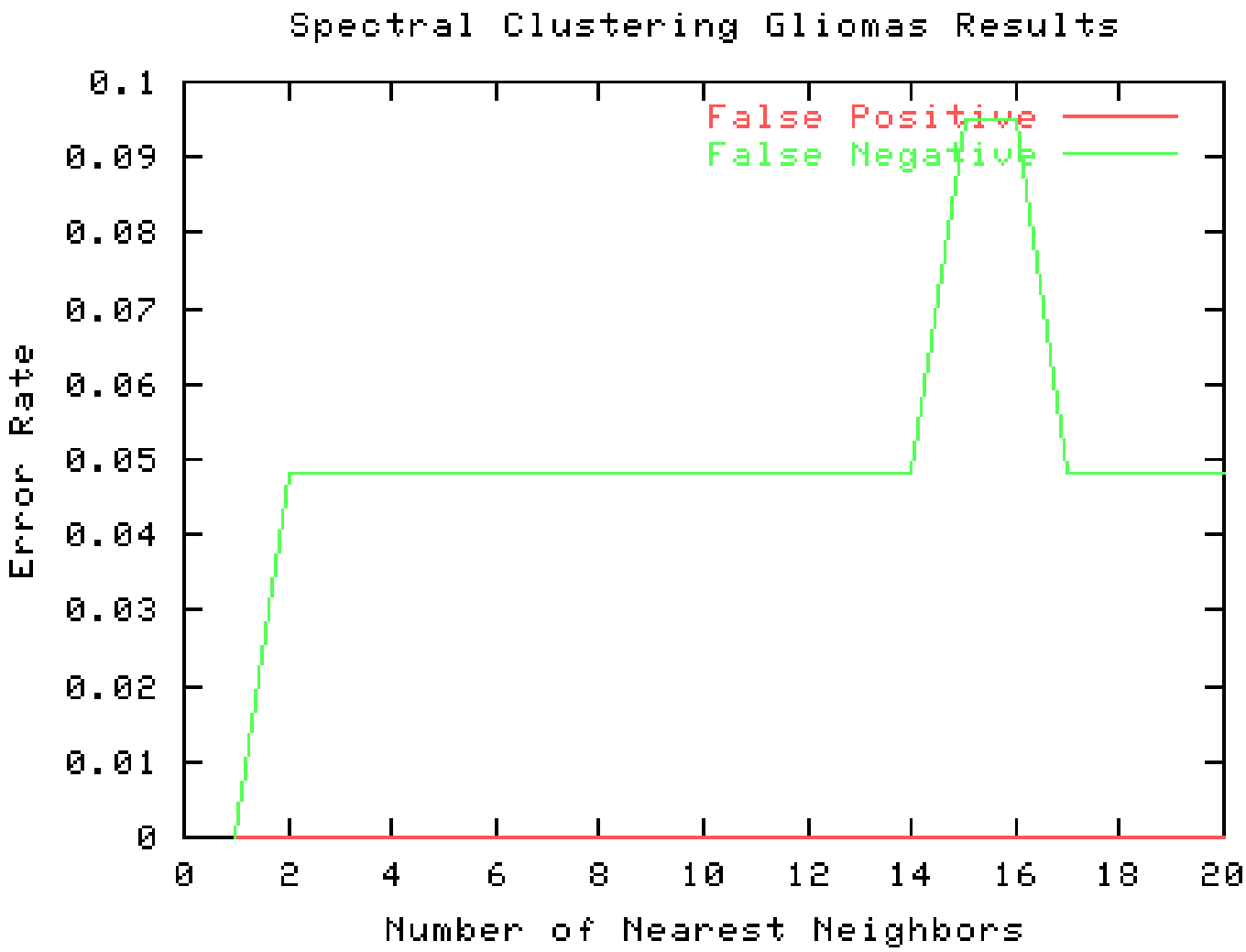
Clustering Malignant Gliomas

- First attempt: poor error rates
- Read paper more carefully:
  - Variation filtering step to reduce noise
  - Genes with less than 100 units of variation removed
- Reduced data set from ~12,000 genes to ~5,000

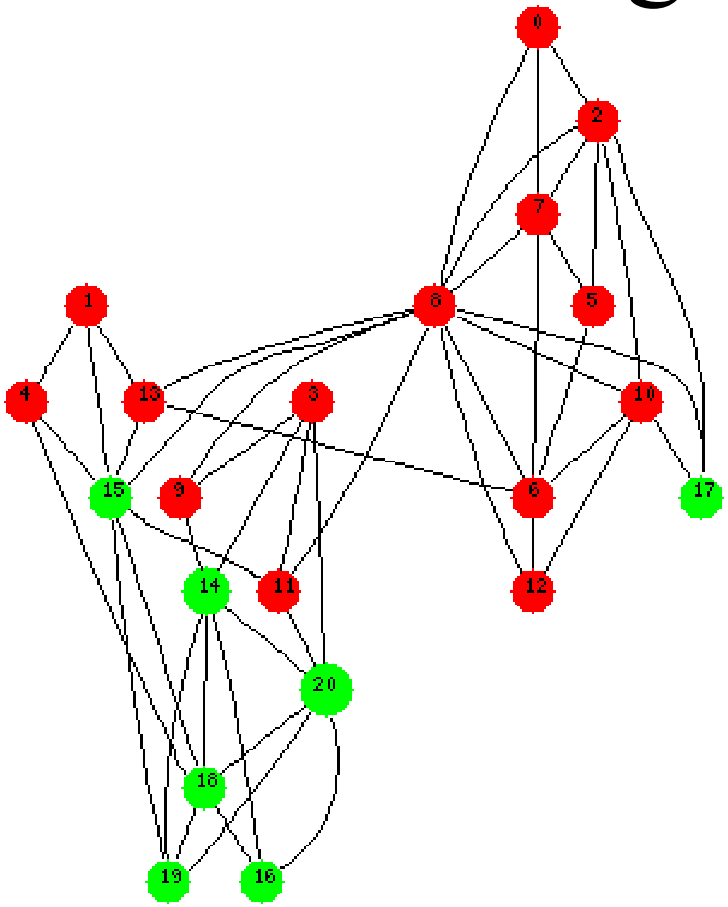
Clustering Malignant Gliomas



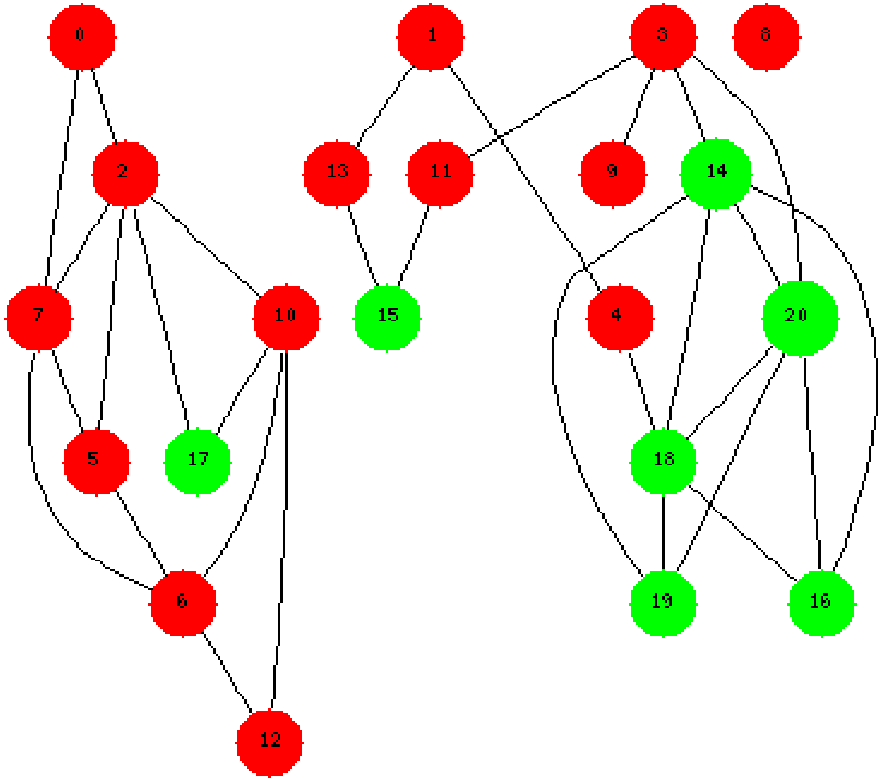
Clustering Malignant Gliomas



Clustering Malignant Gliomas

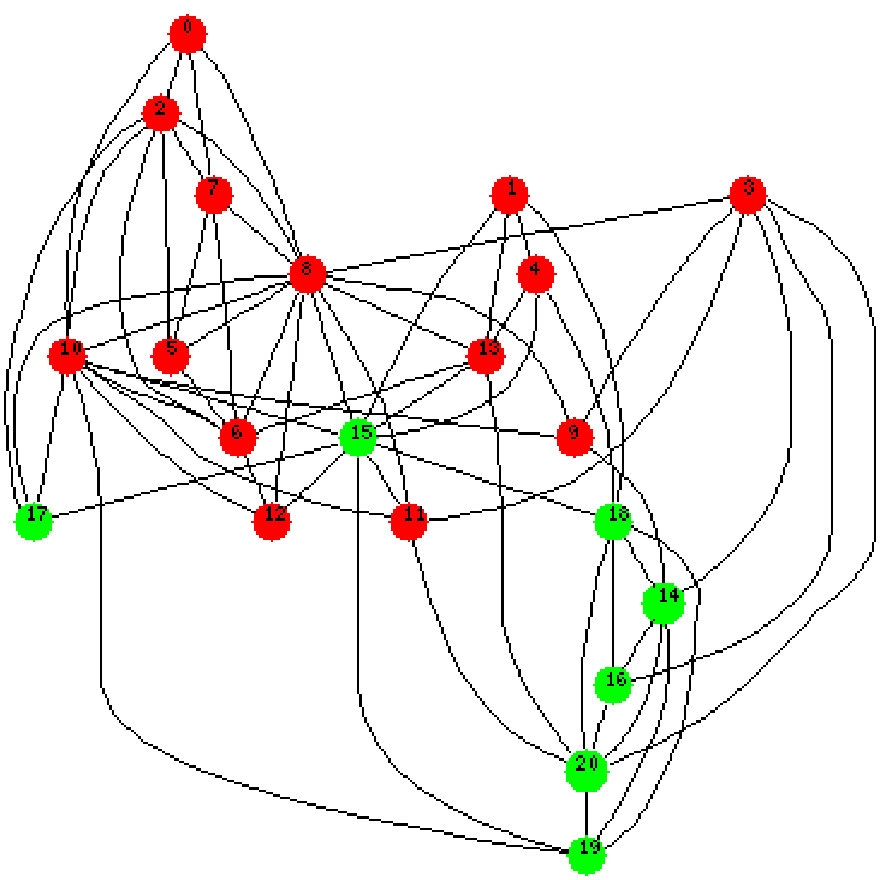


k=3 Nearest Neighbors

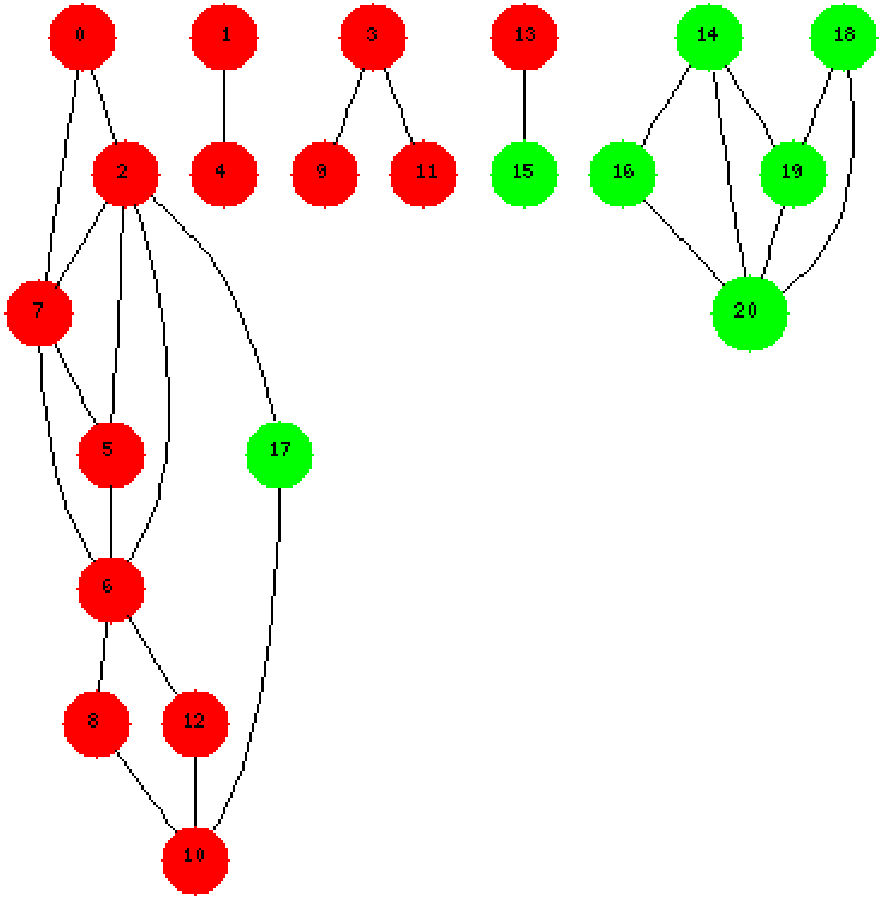


Resulting Clusters

Clustering Malignant Gliomas



k=4 Nearest Neighbors



Resulting Clusters

Conclusions

- All methods require some knowledge of underlying data to tune parameters
- Spectral clustering offers (demonstrably) better results on gene expression datasets
- No clear number of clusters in Gliomas study

Thanks!