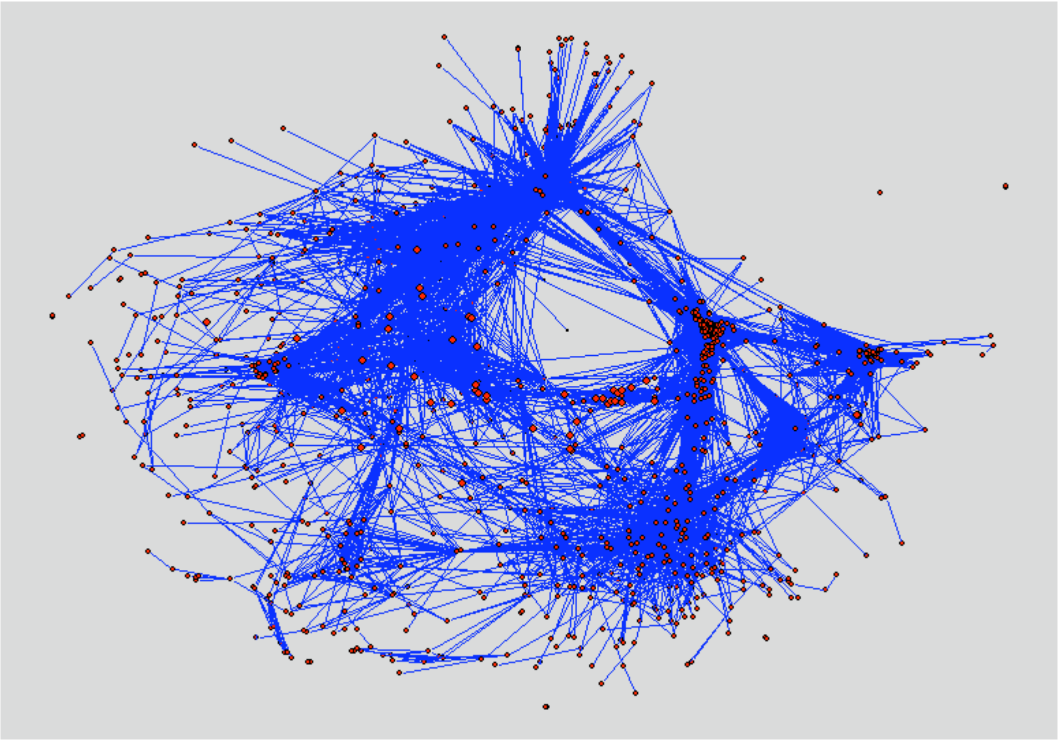
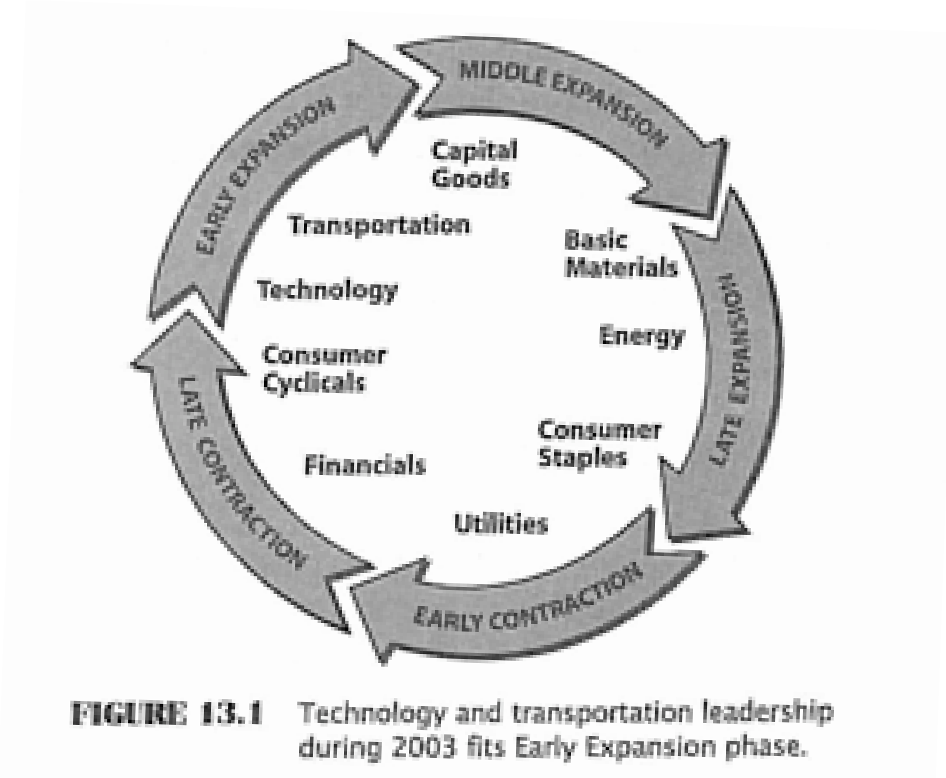
CSS07
Pattern recognition's role in the articulation and validation of Complex Structures
Part 1: Unsupervised Learning



New York Stock Exchange Network



Known cyclic structure

Gregory Leibon
Dartmouth College

Goals:

- A basic introduction to pattern recognition
  - Dimension reduction, supervised learning, unsupervised learning, graphical models (See: *Pattern Classification* by Duda, Hart, and Stork and *The Elements of Statistical Learning* by Hastie, Tibshirani, Friedman)
- Some specific tools to use in the next couple week!
  - In MATLAB: MDS, K-means, HMM, Basic Kernel Methods
  - Others: Spectral Clustering, Persistent Homology, Bayesian Networks
- Explore a Concrete Example: Chunk of U.S. Equity's Market

- The market analysis presented here is a brief look into current work in progress with Alyssa Anderson, Robert Savell, Daniel Rockmore, and Scott Pauls.

Market as a Complex System

- Long Term Goal

  - is to form a model of the market with simple local behavior out of which emerges the complex collective behavior of the market we see in the world.

- Role of Pattern Recognition

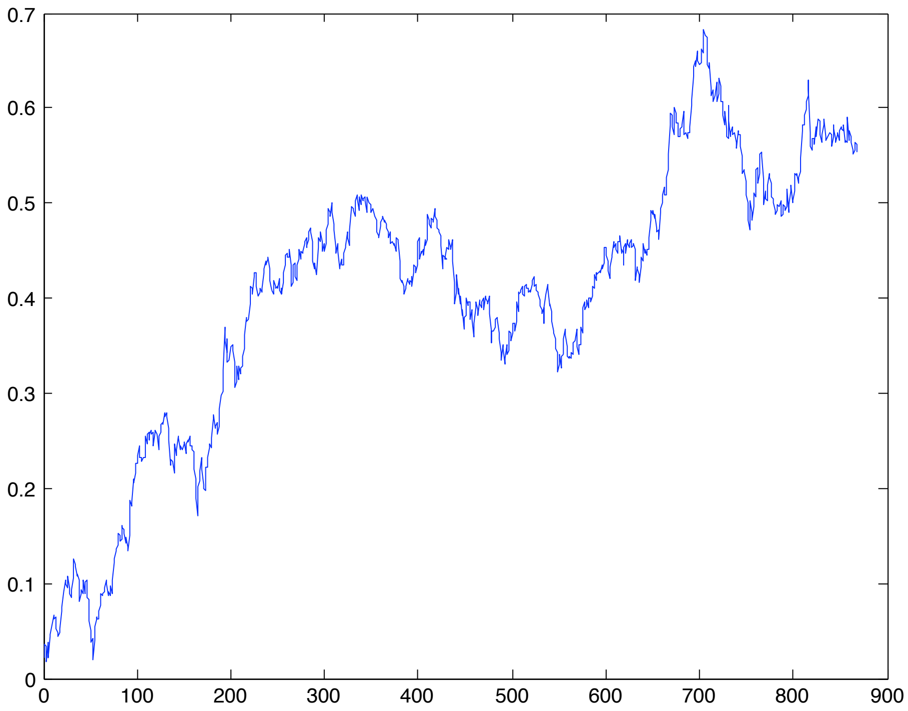
  - is that in a realistic model one must understand how to articulate the local behavior as well understand the emergent patterns that characterize the a complex system.

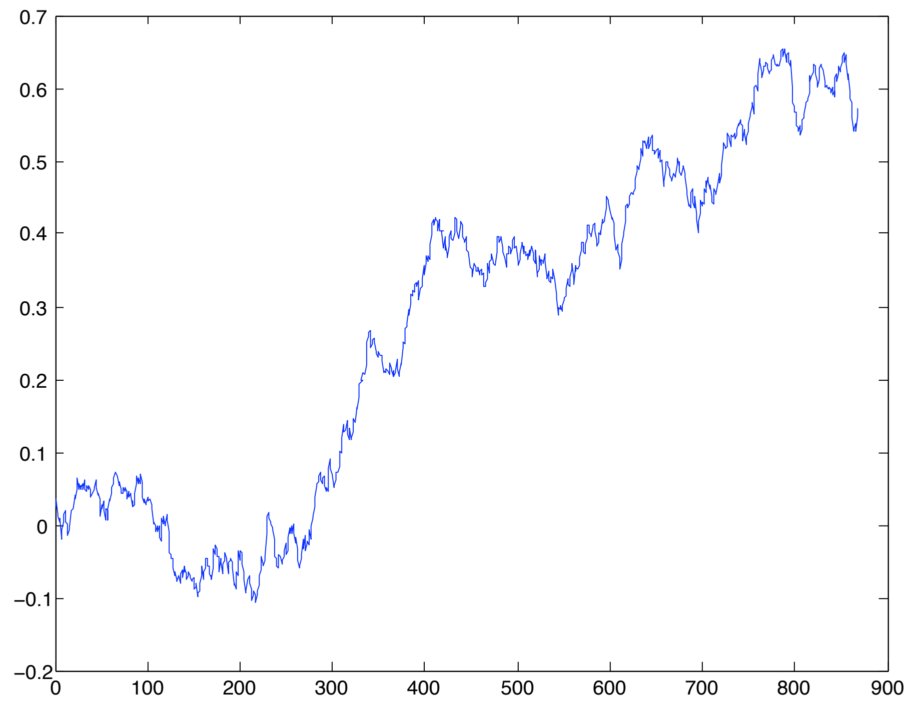
- Data Collected

  - 6000+ interacting tradable equities from NASDAQ and NYSE from *Yahoo! finance*.
  - Data collected for each equity was the Open, Close, High, and Low price as well as the Volume traded on each trading day over the last 15 years.
  - Annotation (partial) collected was the equity's names, sector, industry, and index membership.

An Equity's Times Series

$$X_t = \frac{C_t - C_{t-1}}{C_{t-1}} \approx d(\ln(S_t))$$

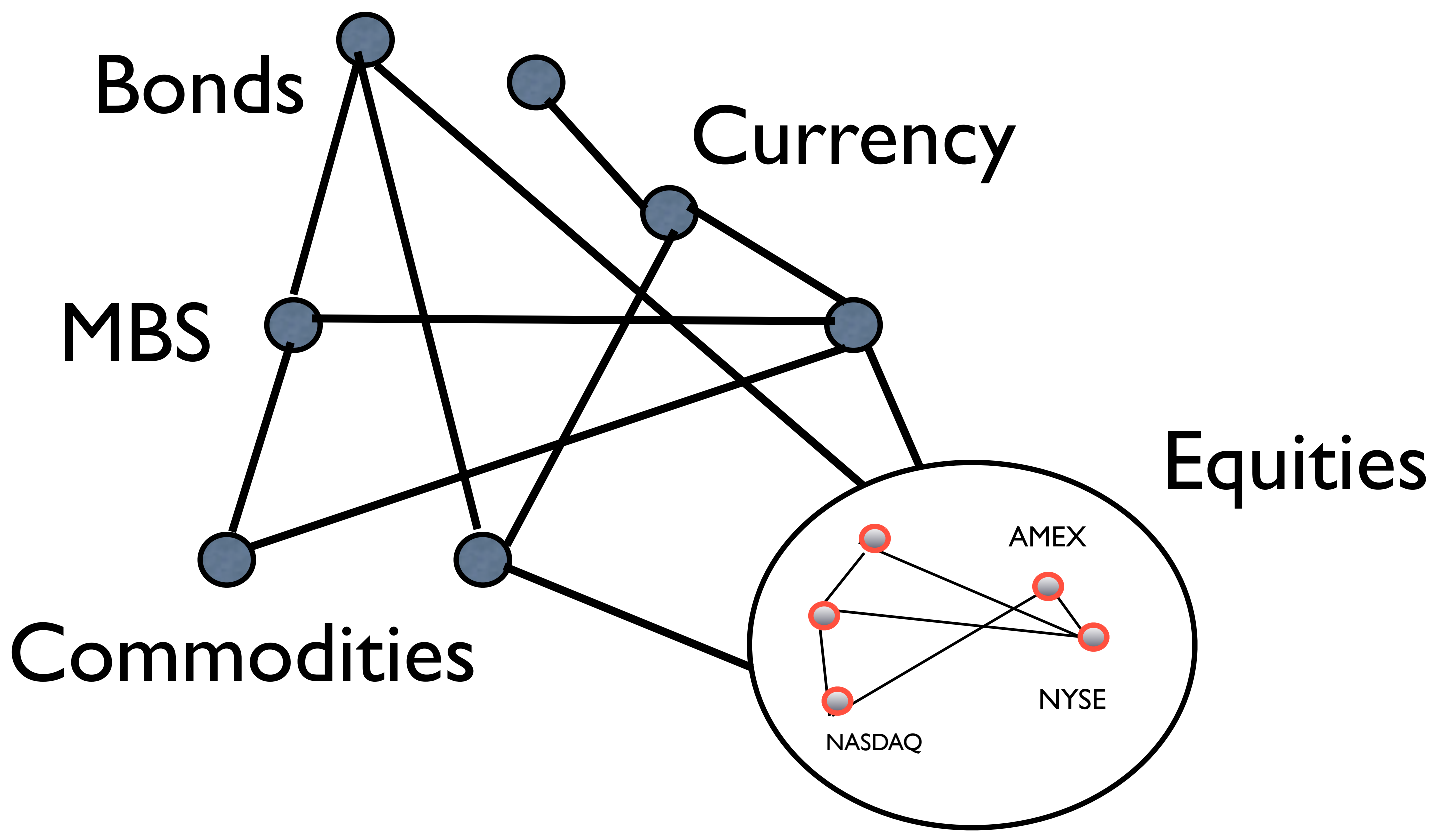




A well known rough approximation:

$$d(\ln(S_t)) = \sigma(t)dB_t + c(t)dt$$

Our Market as a Network



Correlation Metric

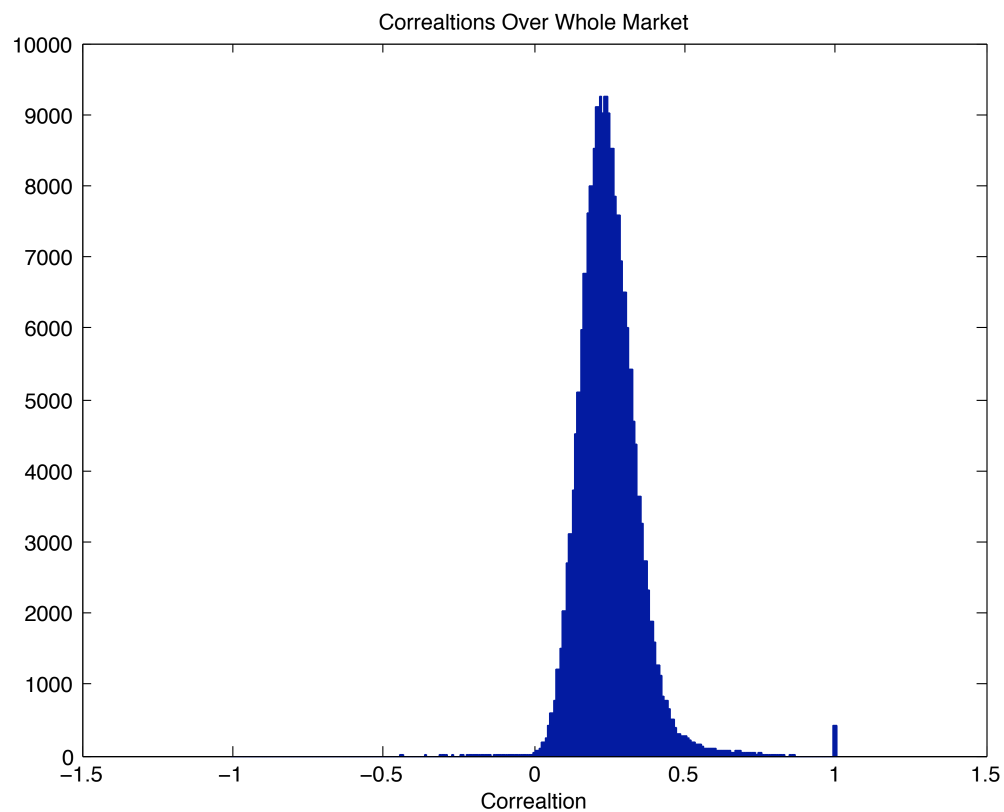
Normalize: $$\widehat{X} = \frac{X - \langle X \rangle}{\sqrt{\left\langle \left(X - \langle X \rangle \right)^2 \right\rangle}}$$

Correlation: $$\rho(X,Y) = \left\langle \widehat{X}\widehat{Y} \right\rangle$$

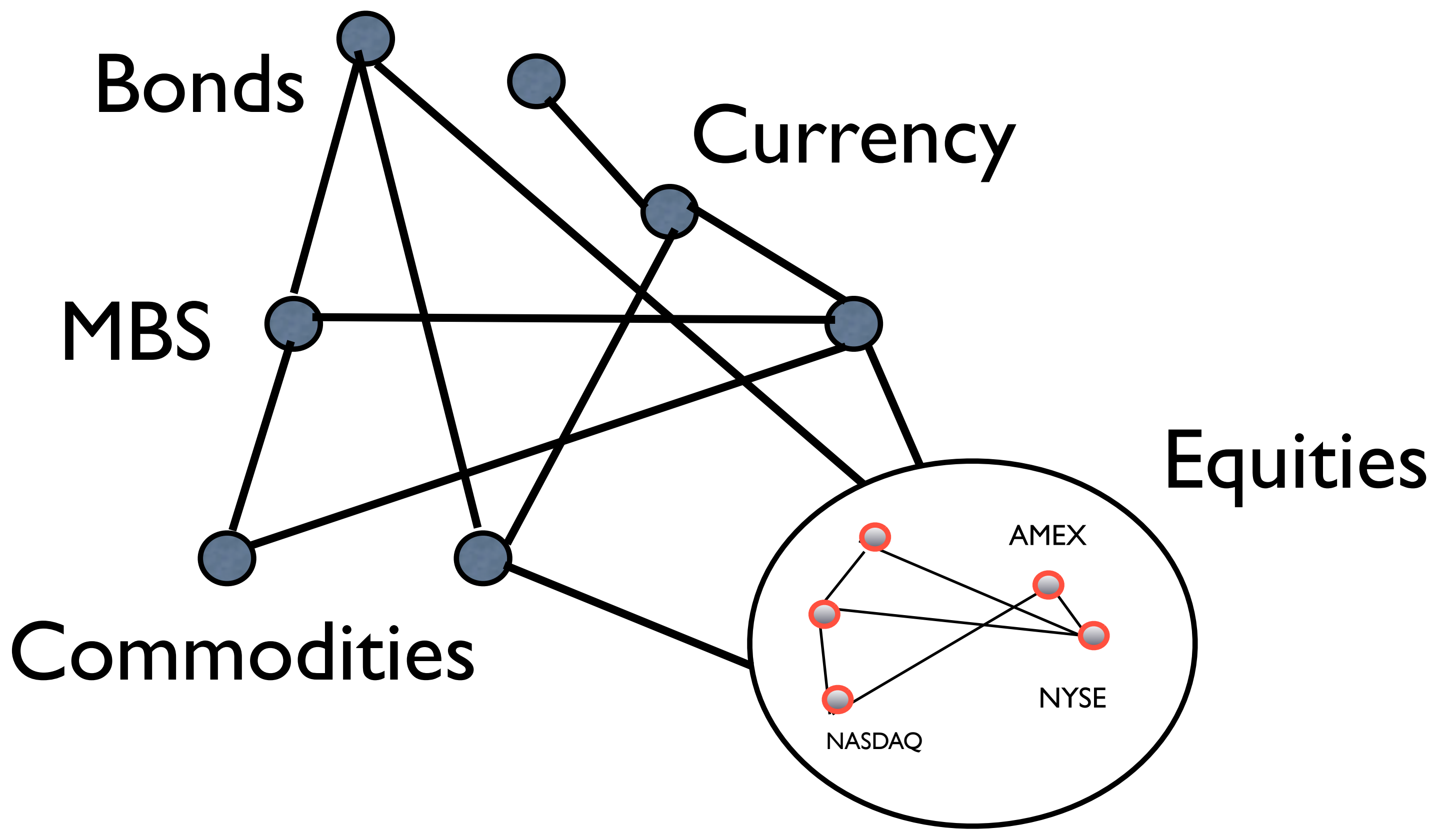
Metric: $$D(X,Y) = \arccos(\rho(X,Y))$$

Corresponds to a metric graph in $S^{1000 \pm c}$

Correlation: S&P500



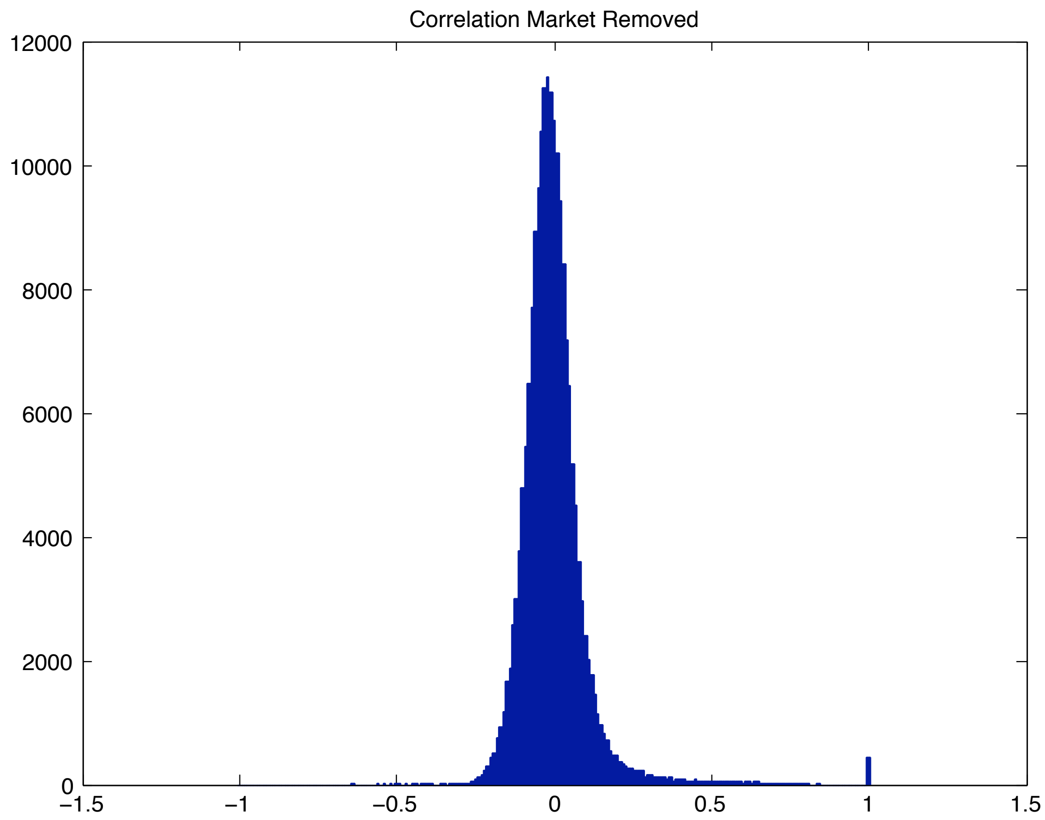
How much of this correlation is internal to our equities market and how much due to external forces from the whole economy?



$$d(\ln(S_t)) = \sigma dB_t + c dt + \tau dF_t$$

Confounding Factors

Each day there is pressure on the equities market to absorb money. We can remove this effect:



Network Structure

- First step: Dimension Reduction
- Our network is embedded in 1000+ dimensions. To see and work with our network, it is useful to attempt to embedded it a lower dimensional space.
- One great simple tool for doing this is Multi Dimensional Scaling, MDS.
- With MDS: For each dimension, we can produce an approximate lower dimensional embedding, call it f(X).
- Might use the Eulcidean: $d'(X,Y) = \sin\left(\frac{d(X,Y)}{2}\right)$

MDS algorithm

Simply, attempt to minimize a positive loss function that would be zero if for all X,Y

d(X,Y)=d(f(X),f(Y)).

Example Raw Stress: $L = \sum \left(d(X,Y) - d(f(X), f(Y)) \right)^2$

Minimization Techniques: Gradient Decent, Newton Raphson, Iterative Majorization, Tabu Search, Genetic Algorithms, Simulated Annealing....

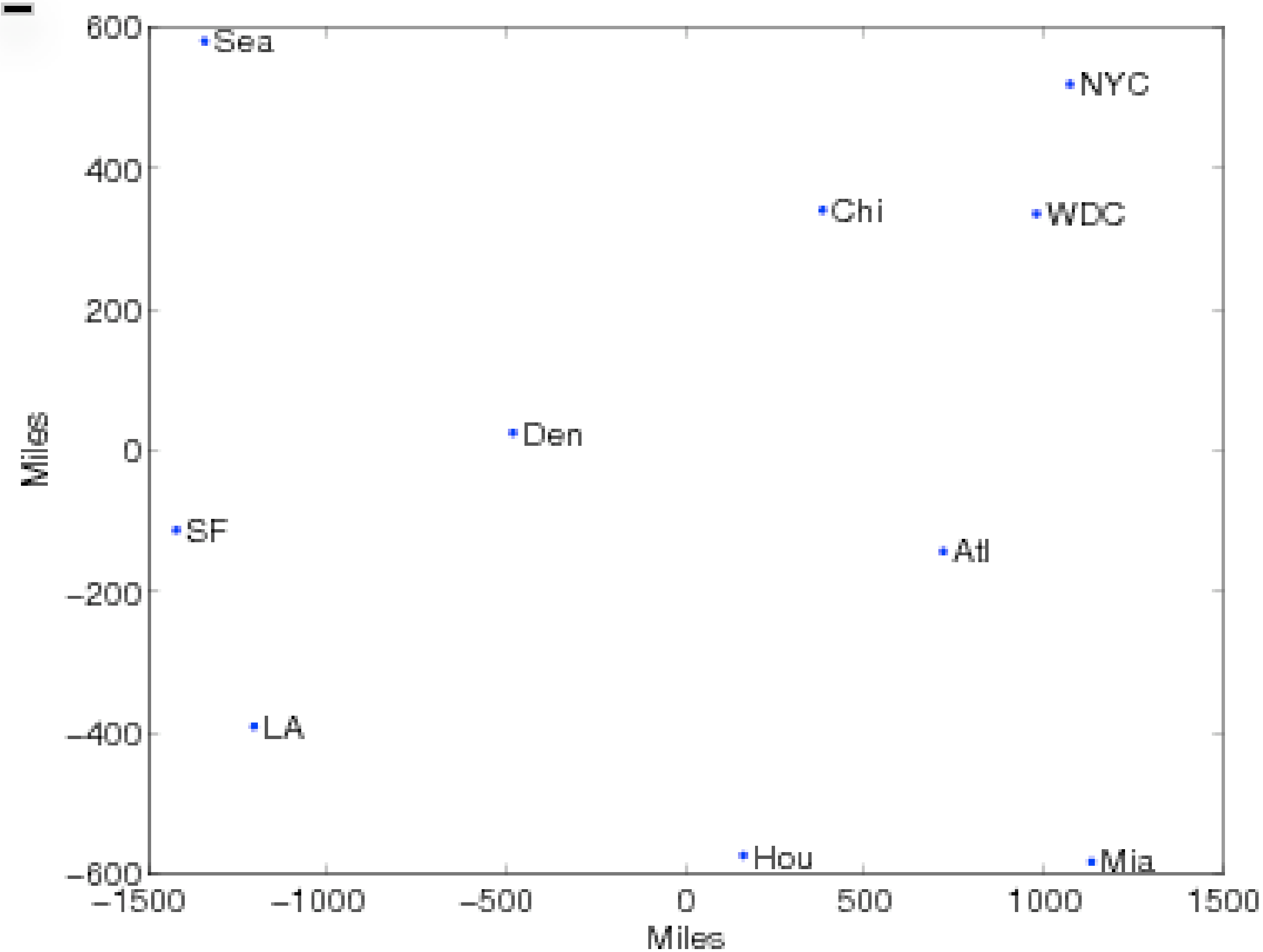
Himalayas

2-d Example

```
cities =
{'Atl','Chi','Den','Hou','LA','Mia','NYC','SF','Sea','WDC'};
D = [    0  587 1212  701 1936  604  748 2139 2182   543;
       587    0  920  940 1745 1188  713 1858 1737   597;
      1212  920    0  879  831 1726 1631  949 1021  1494;
       701  940  879    0 1374  968 1420 1645 1891  1220;
      1936 1745  831 1374    0 2339 2451  347  959  2300;
       604 1188 1726  968 2339    0 1092 2594 2734   923;
       748  713 1631 1420 2451 1092    0 2571 2408   205;
      2139 1858  949 1645  347 2594 2571    0  678  2442;
      2182 1737 1021 1891  959 2734 2408  678    0  2329;
       543  597 1494 1220 2300  923  205 2442 2329     0];
```

Example taken from MatLab's help

MDS 2-d

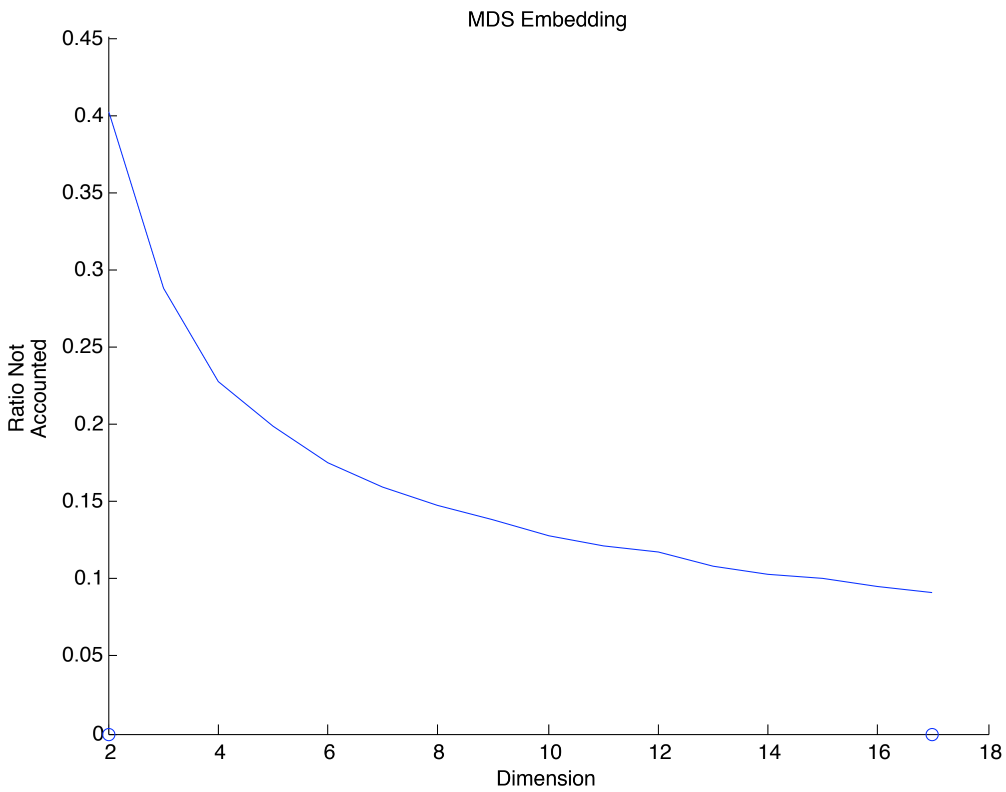


Sea
NYC
Chi
WDC
Den
SF
Atl
LA
Hou
Mia
600
400
200
0
−200
−400
−600
−1500
−1000
−500
0
500
1000
1500
Miles
Miles

For S&P500

Call the embedding f and use the Euclidean distance d(f(X),f(Y))

$$S = \frac{\sum \left| d(X,Y) - d(f(X),f(Y)) \right|}{\sum \left| d(X,Y) \right|}$$



MATLAB code:

Here Closes is the 1000 by 500 matrix with columns corresponding to each equity's normalized daily Close.

```
Cor=corrcoef(Closes,'rows','pairwise')

Dist=sin(acos(Cor/2));

opt=statset('MaxIter',5000);

Dim15= mdscale(Dist,15,'Options',opt);
```

% Here Y15 is a fifteen dimension embedding of vertices.

Network Shape

- Re-scale network
- Clustering to find nodes at new scale.
- Unsupervised learning.
- Clustering techniques: kmeans, spectral clustering, hierarchical clustering, *-Linkage Clustering, Delaunay Complex Exploitation Algorithms,...

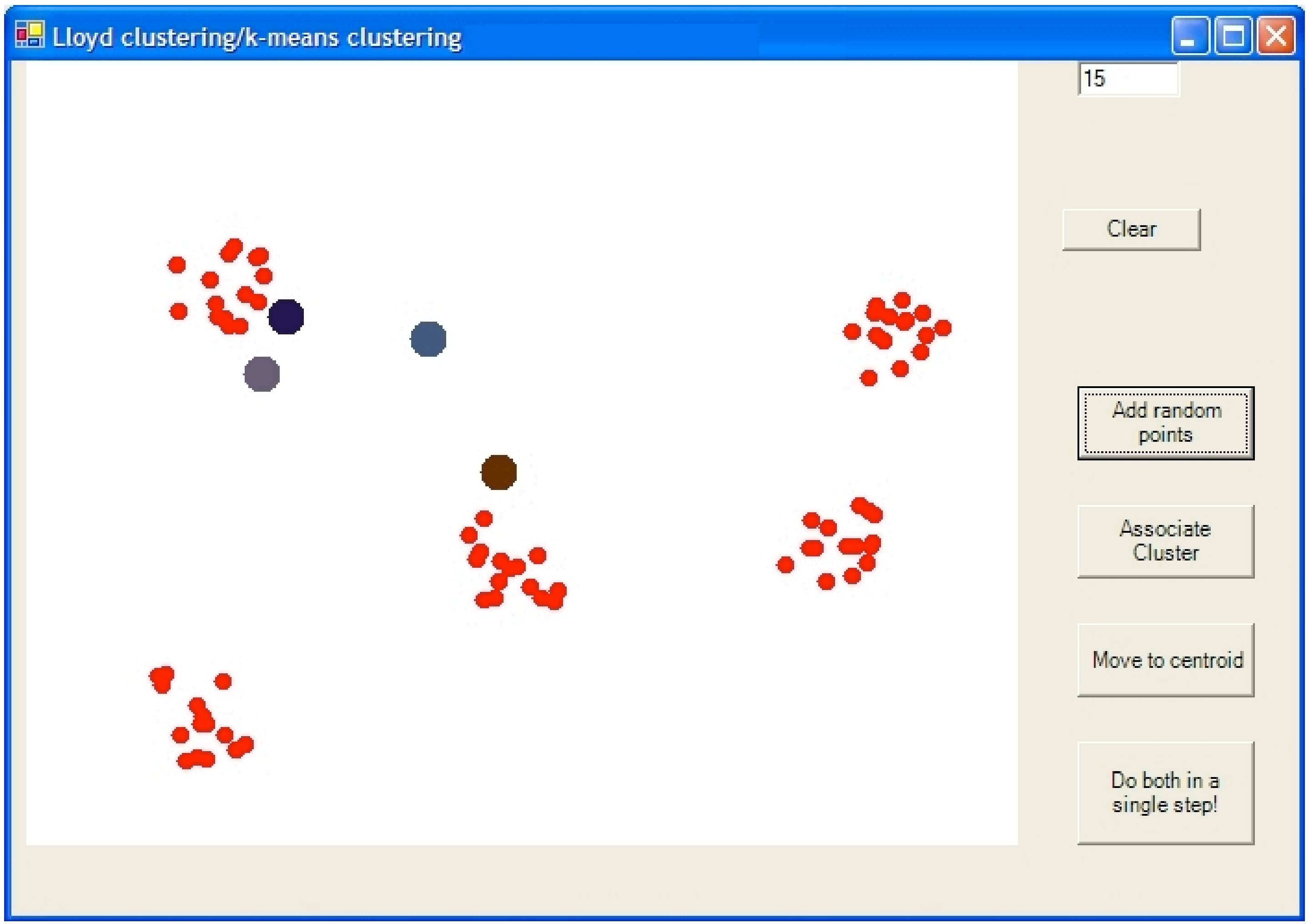
K-means

- Simplest clustering algorithm is k-means
- To run requires fixing N=#(Clusters)
- Requires an Euclidean type embedding (we have one via the MDS)
- Once again, we are essentially minimizing a loss function:

$$L = \sum_{k=1}^{N} \sum_{x_i \in C_k} \left(x_i - \mu_k \right)^2$$

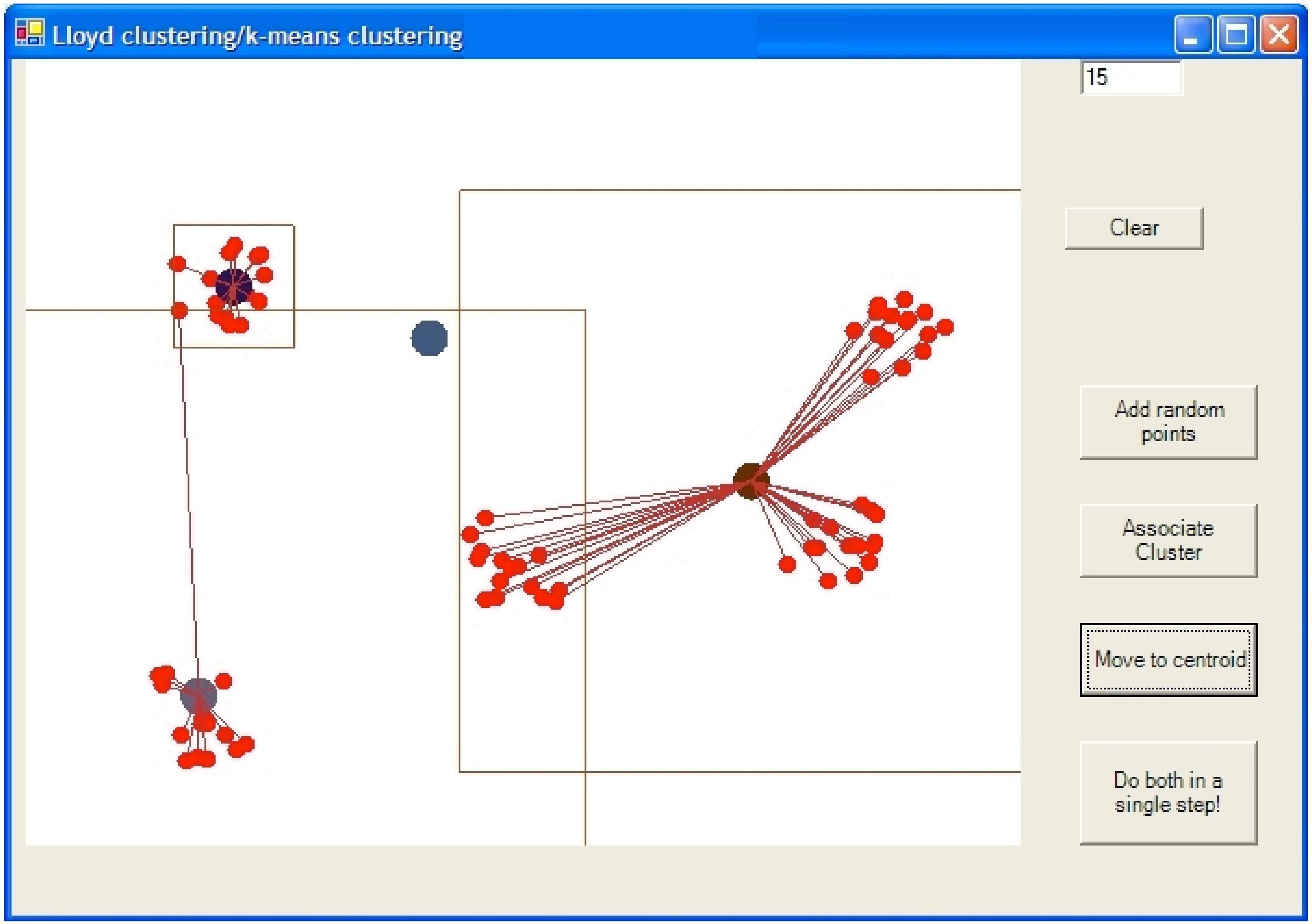
K-means algorithm:

1. Randomly choose points in each cluster and compute centroids.

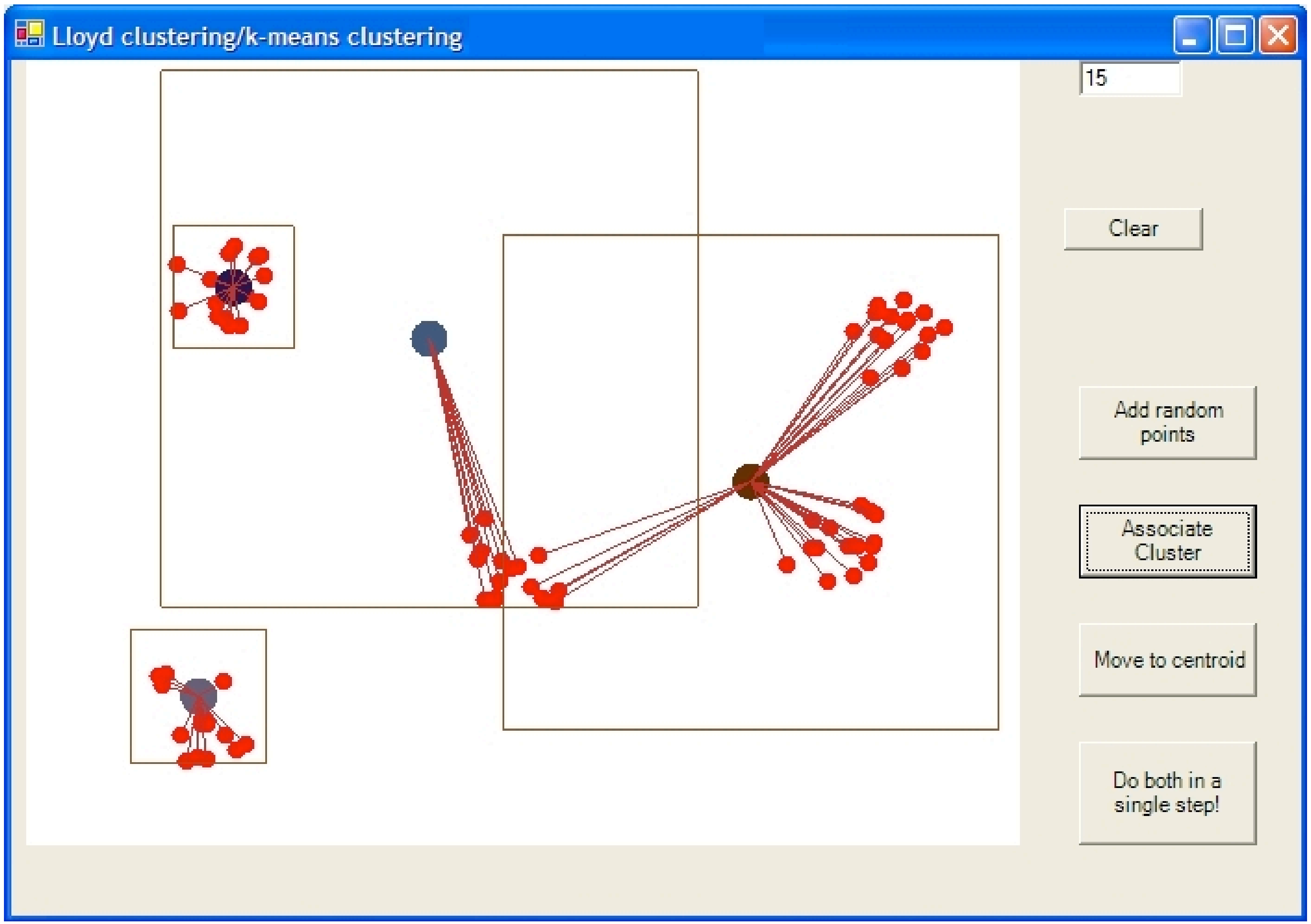


Example from: http://en.wikipedia.org/wiki/K-means_algorithm

2. Organize points by distance to the centroids.
3. Update centroids

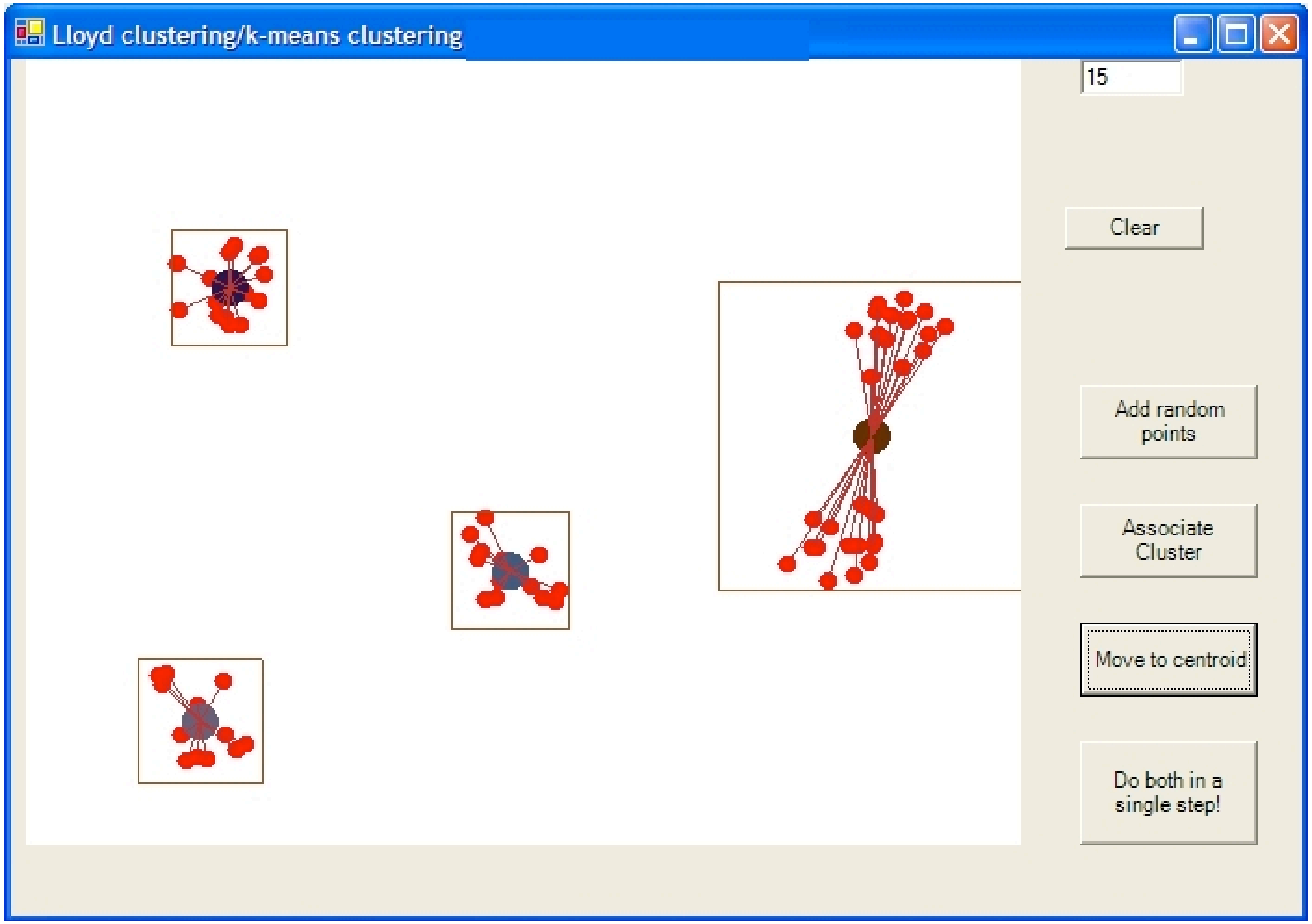


4. Repeat...



Lloyd clustering/k-means clustering
15
Clear
Add random points
Associate Cluster
Move to centroid
Do both in a single step!

...until stable.



MATLAB

Recall Dim15 was the 15 dimensional Euclidean realization of of correlation information.

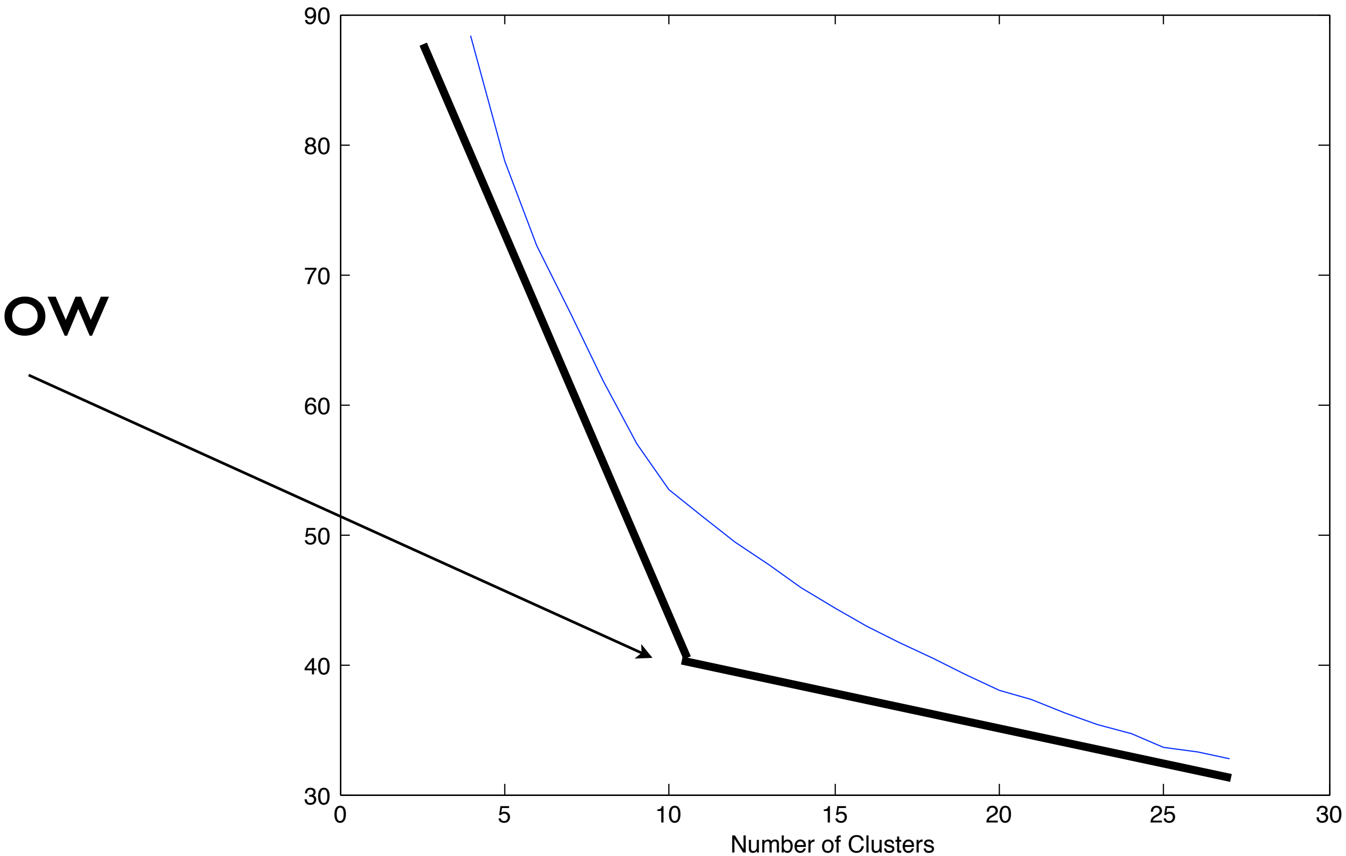
```
Clusters = kmeans(Dim15,20);
```

Clusters is a list of 500 labels between 1 and 20, one for each point in Dim15, indicating the cluster that each sample has been associated with.

Hardest part is choosing N=#(Clusters)

Elbowlogy:

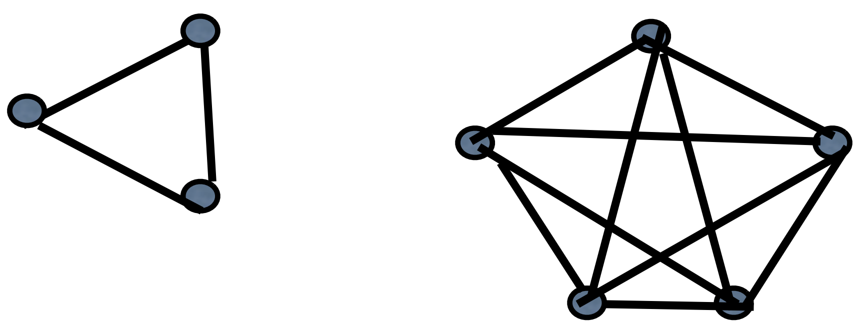
elbow



or...

Laplacian

$$L = \begin{bmatrix} 1 & -1/2 & -1/2 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & 1 & -1/2 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & -1/2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1/4 & -1/4 & -1/4 & -1/4 \\ 0 & 0 & 0 & -1/4 & 1 & -1/4 & -1/4 & -1/4 \\ 0 & 0 & 0 & -1/4 & -1/4 & 1 & -1/4 & -1/4 \\ 0 & 0 & 0 & -1/4 & -1/4 & -1/4 & 1 & -1/4 \\ 0 & 0 & 0 & -1/4 & -1/4 & -1/4 & -1/4 & 1 \end{bmatrix}$$



Laplacian

$$d^{'}(X,Y) = \sin\left(\frac{\rho(X,Y)}{2}\right)$$

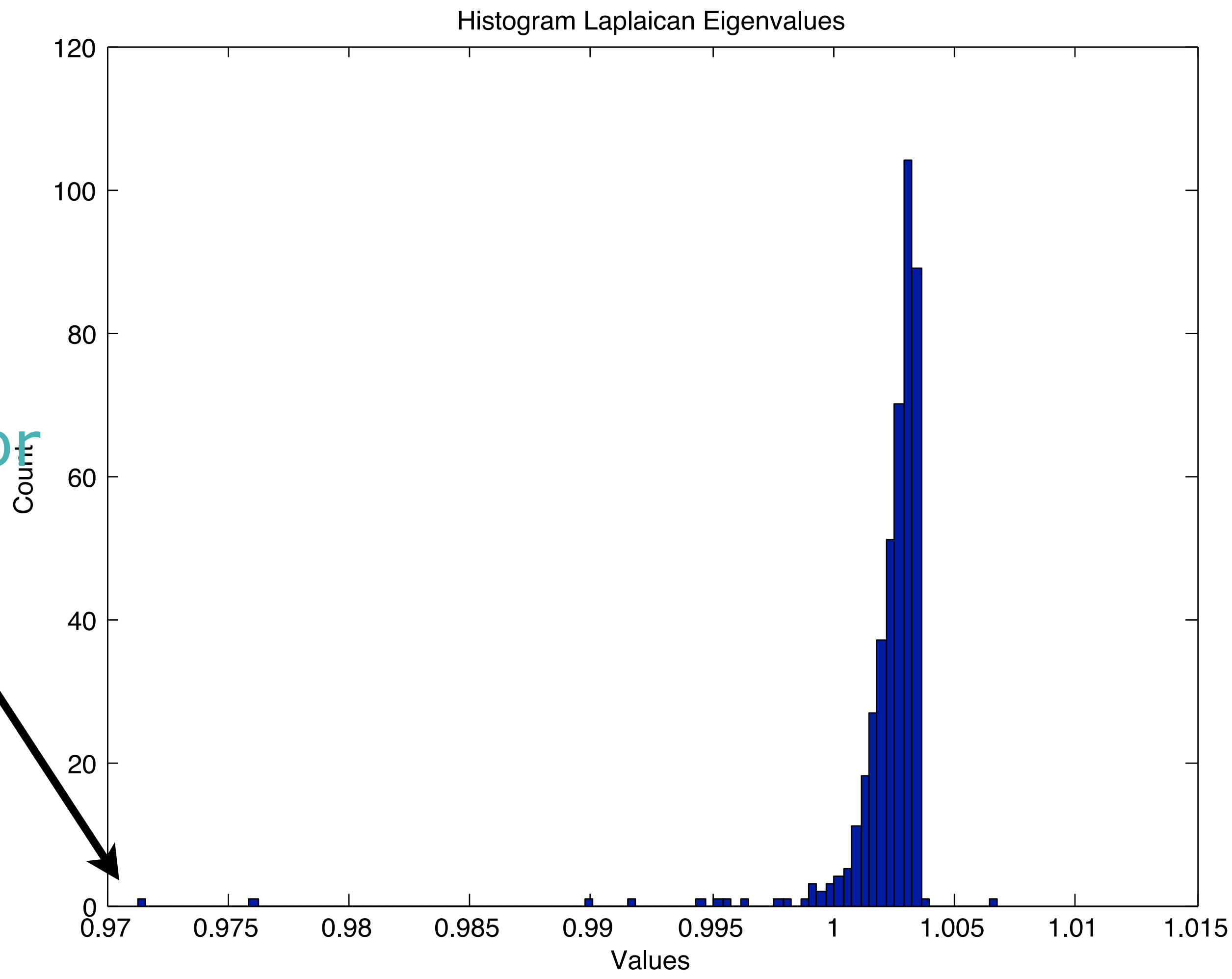
$$S = e^{-\frac{d^{'}(X,Y)^2}{\sigma^2}}$$

$$D = \sum_j S_{ij}$$

$$\Delta = I - D^{-1/2} S D^{-1/2}$$

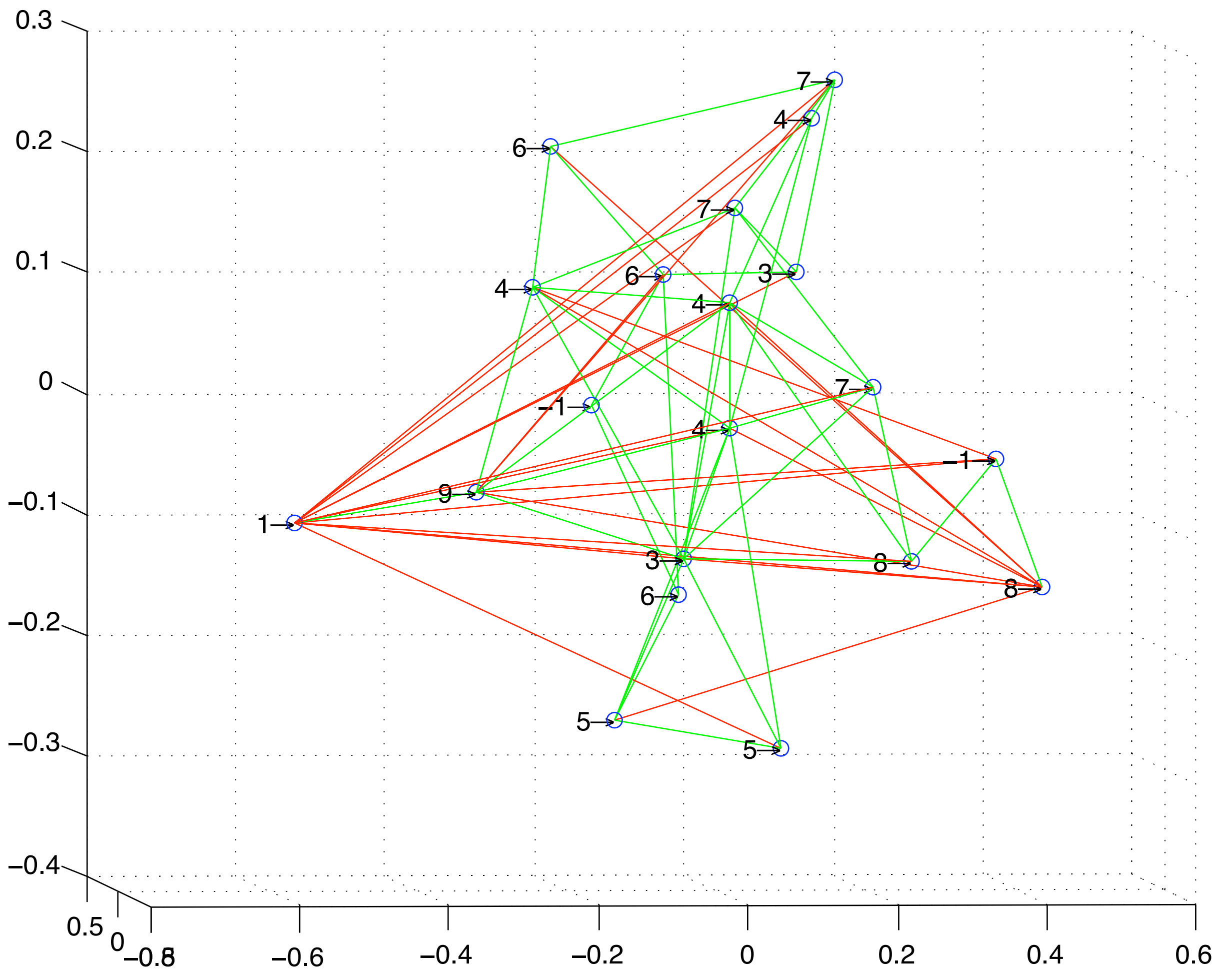
Eigenvalues>0

Fiedler Vector



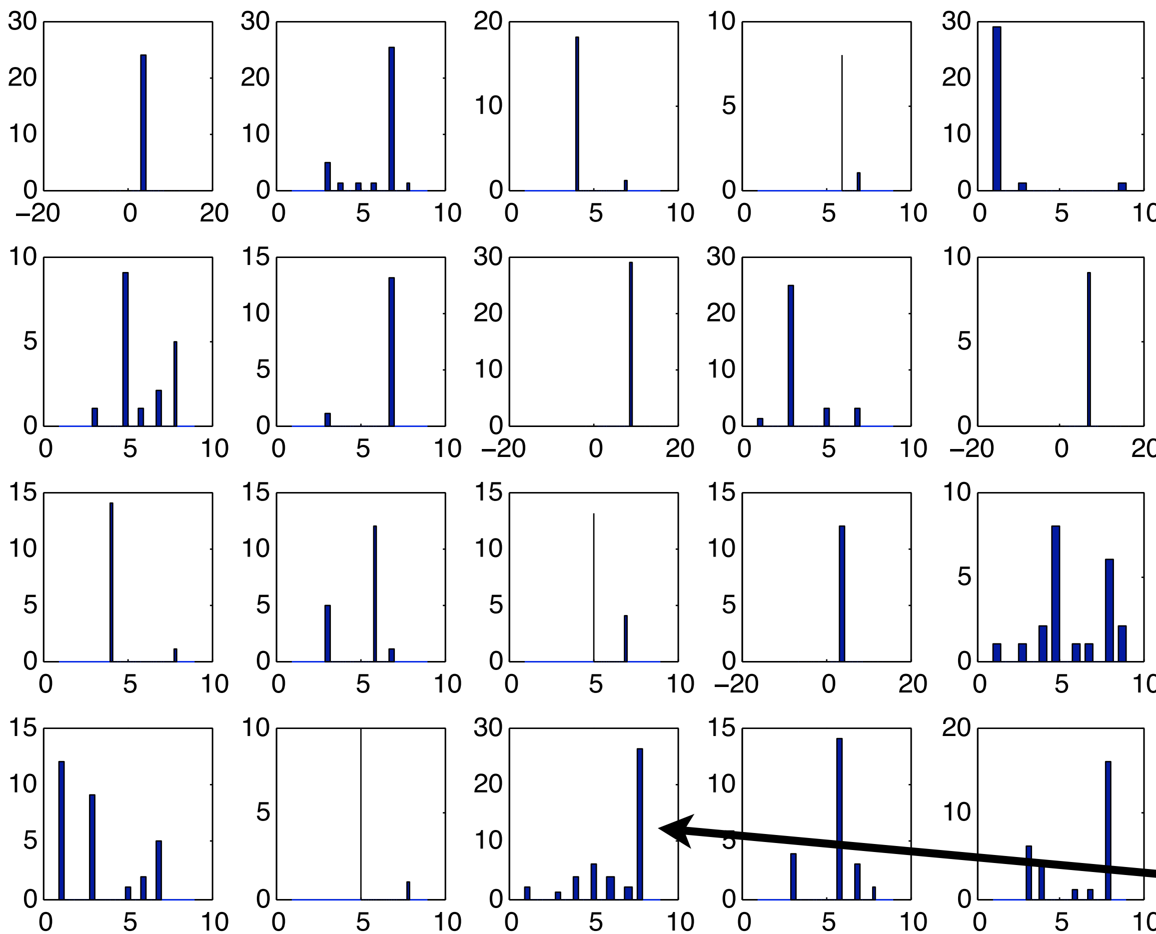
Major clusters correspond to the "outlier" eigenvalues.

#(Clusters)=20



Take a guess!

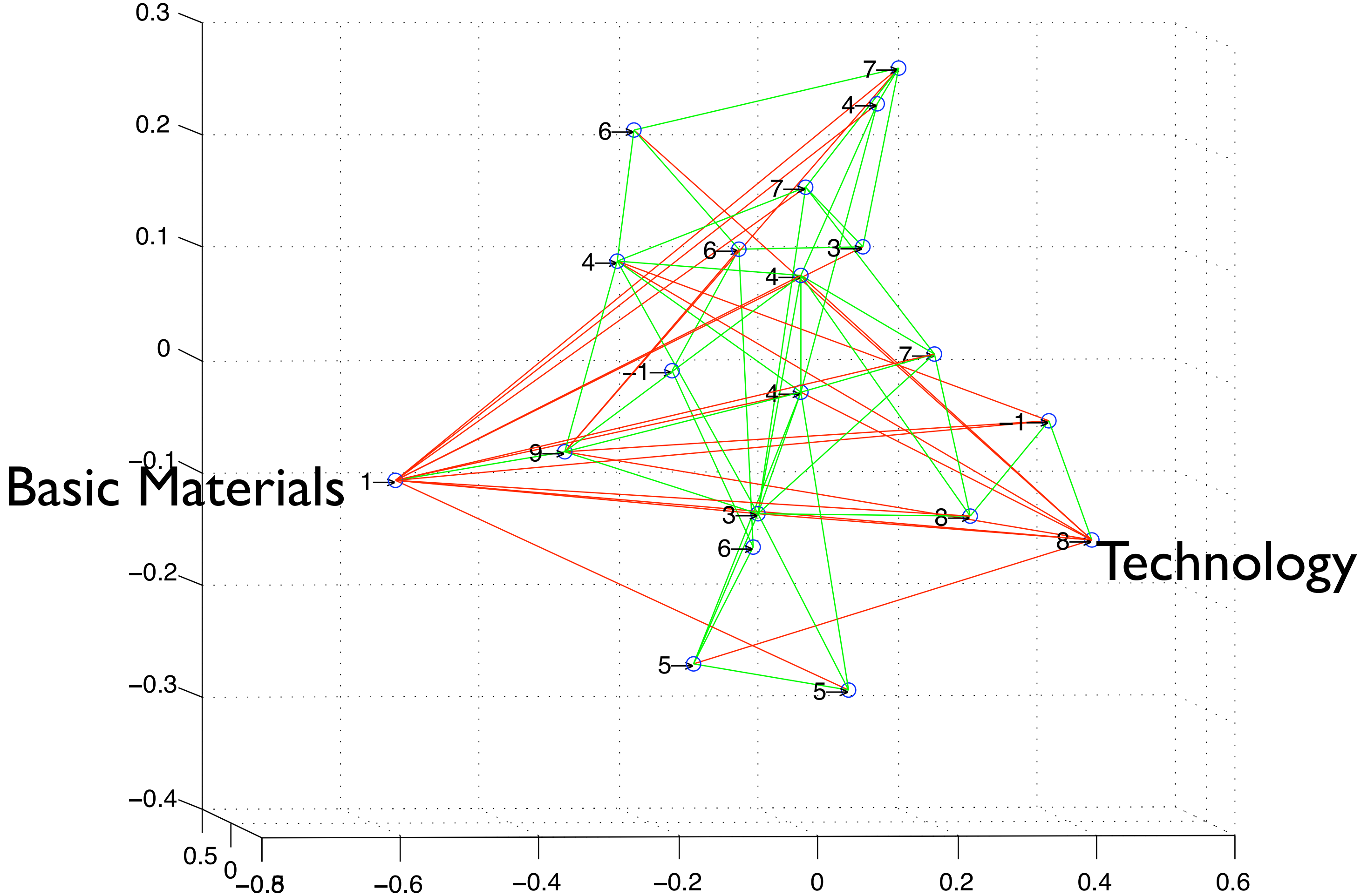
Sectors

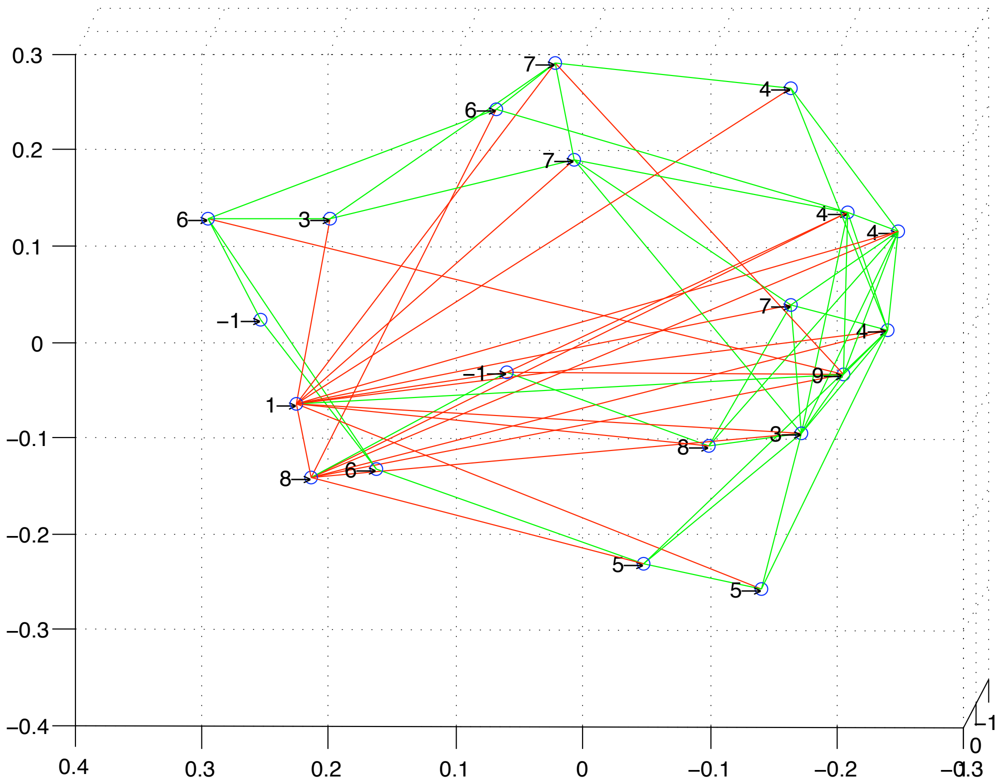


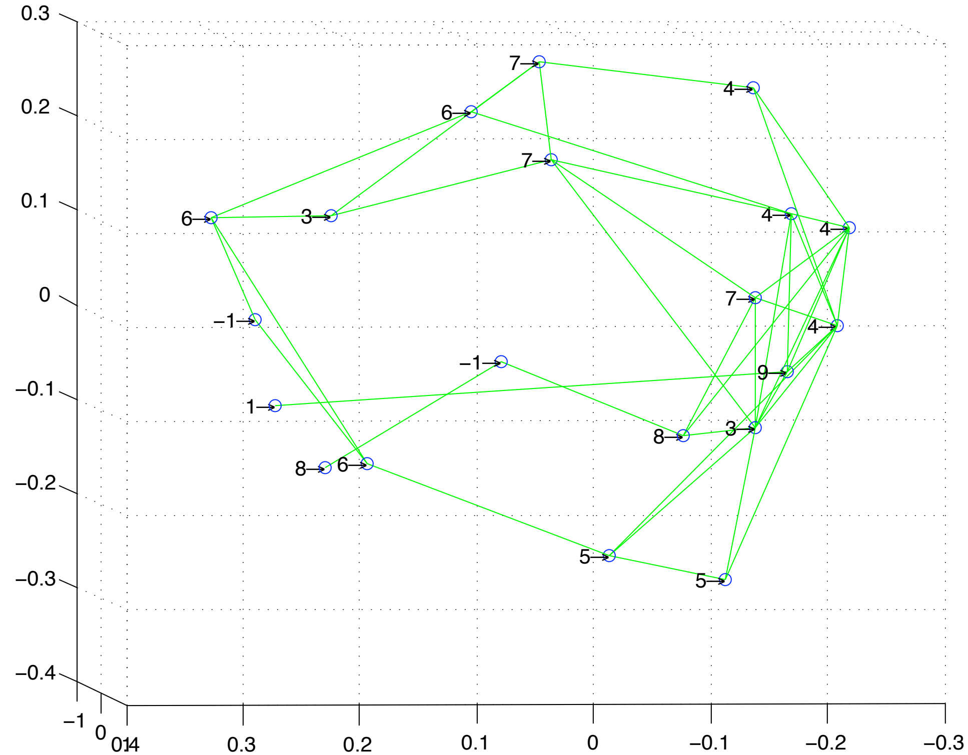
1 Basic Materials
2 Conglomerates
3 Consumer Goods
4 Financial
5 Healthcare
6 Industrial Goods
7 Services
8 Technology
9 Utilities

Guess?

The Recent "Battle"

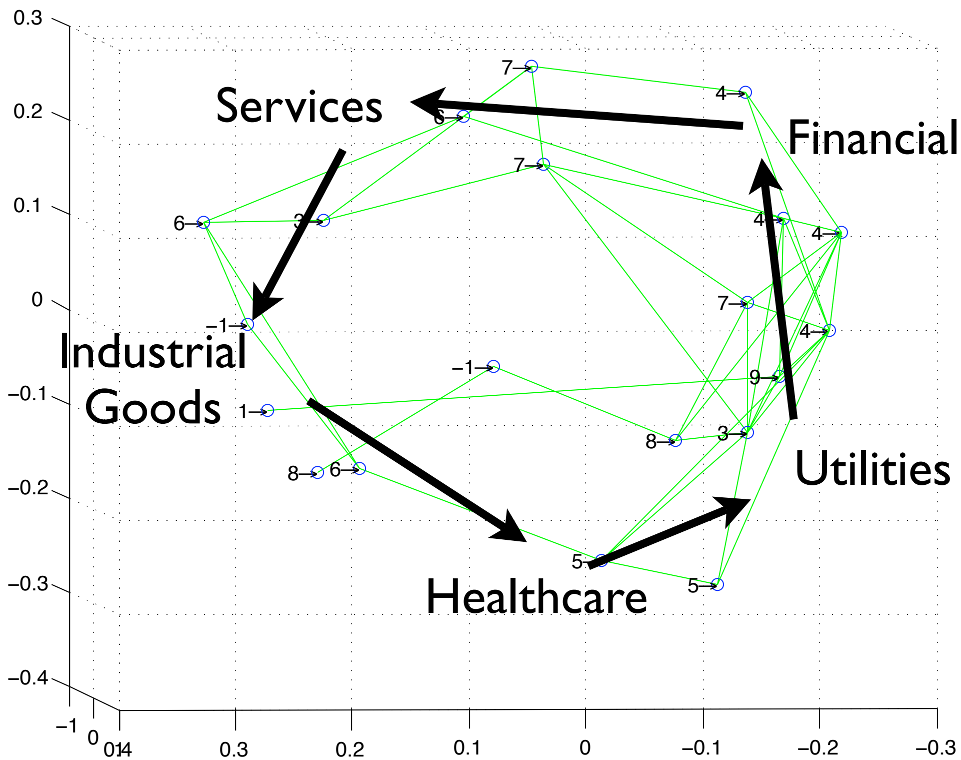




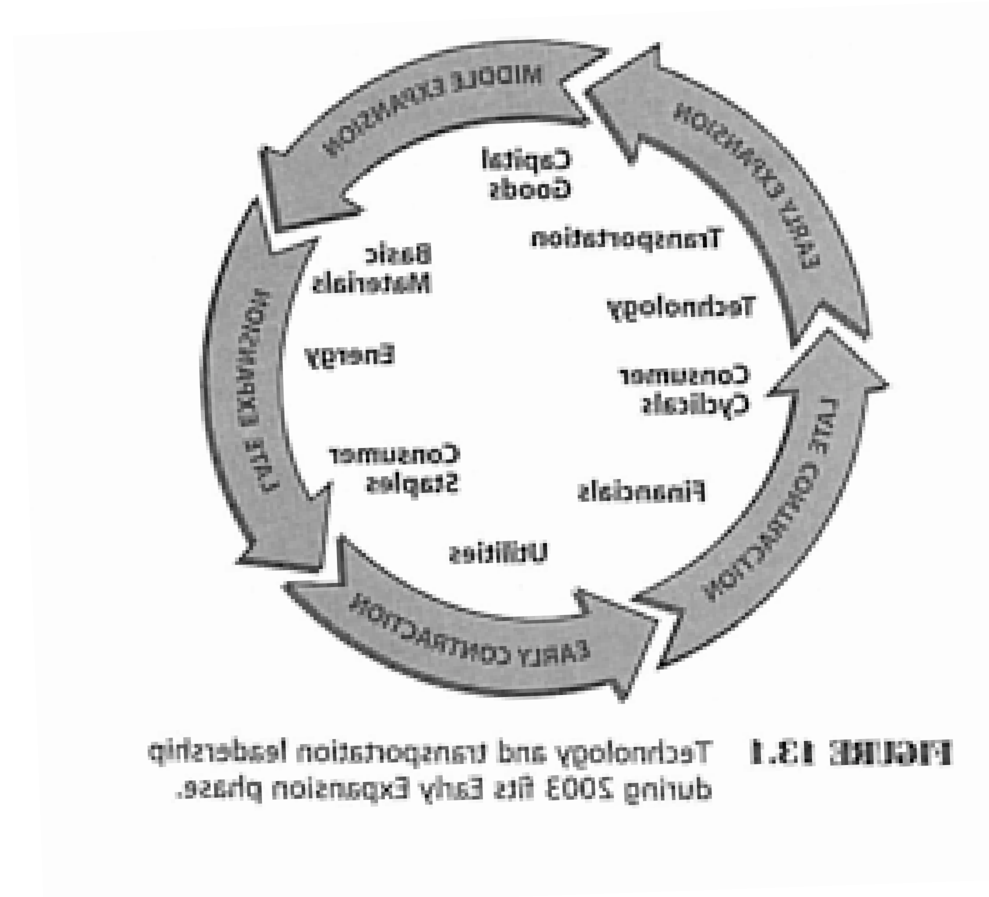


Notice the circle: A manifold, not a “ball” cluster.

The Cycle

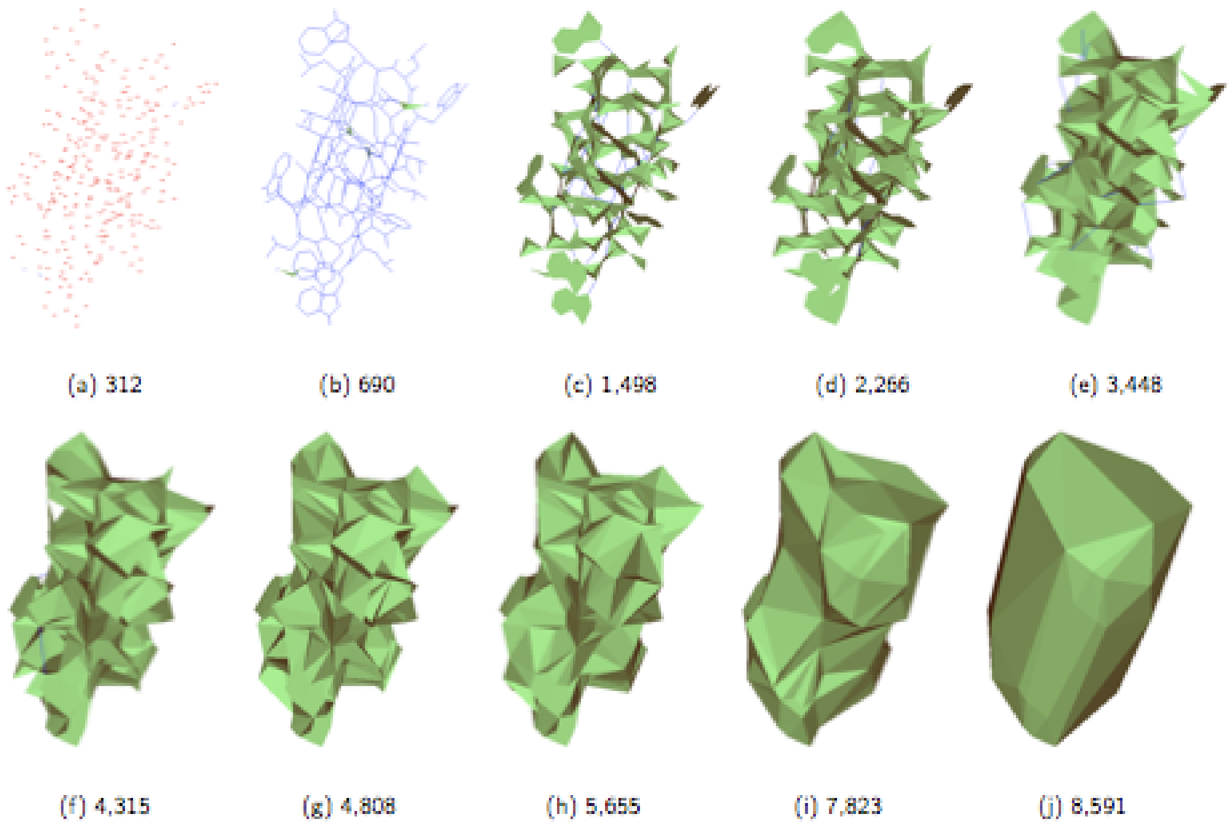


Services
Financial
Industrial Goods
Utilities
Healthcare



MIDDLE EXPANSION
EARLY EXPANSION
LATE CONTRACTION
EARLY CONTRACTION
LATE EXPANSION
Capital Goods
Transportation
Technology
Consumer Cyclicals
Financials
Utilities
Consumer Staples
Energy
Basic Materials
FIGURE 13.1 Technology and transportation leadership during 2003 fits Early Expansion phase.

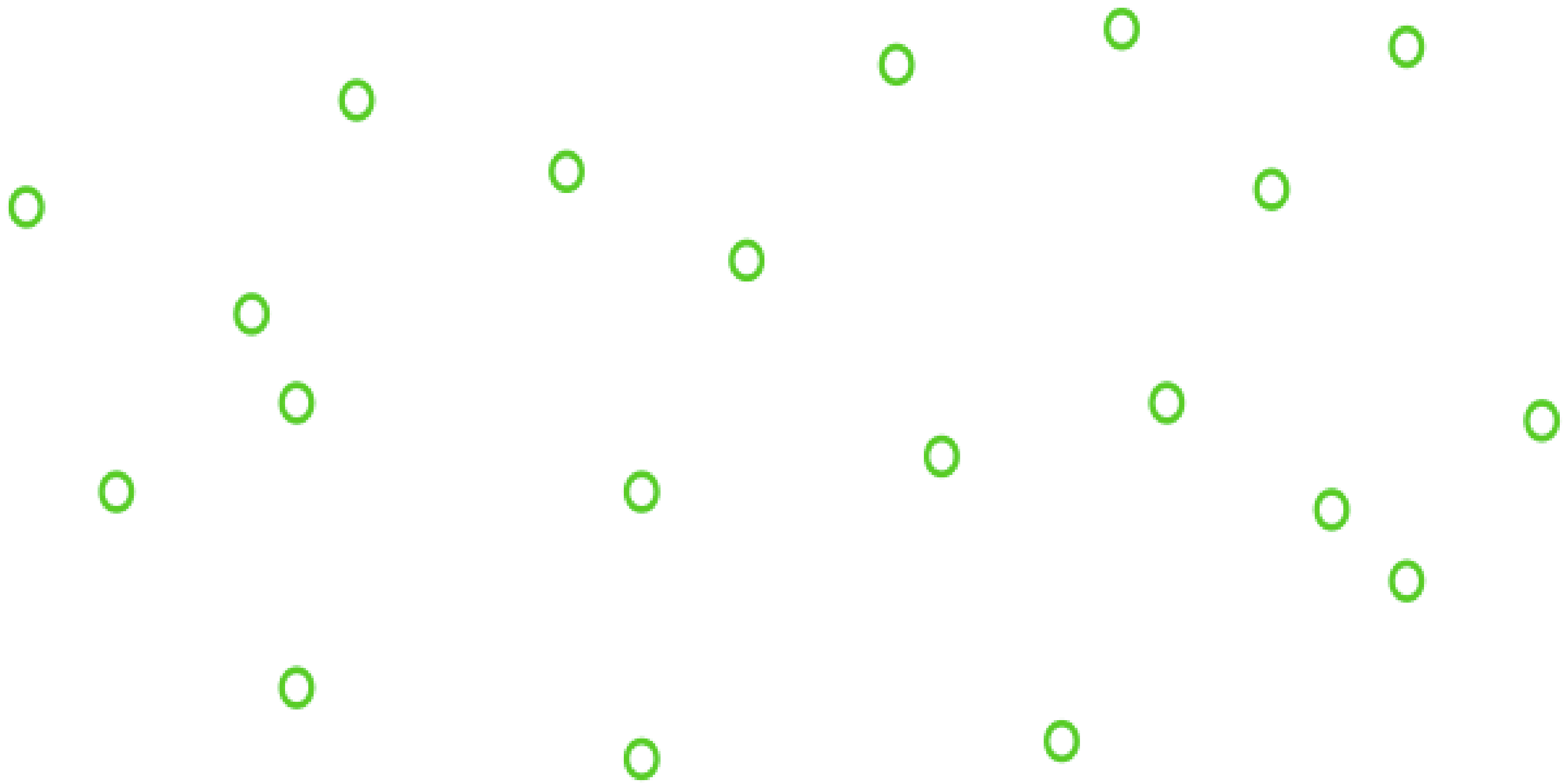
- Manifold Learning: Loads of current research!
- "Most Fun"=persistent homology. Detects the emergence of topology. (See: *Topology for Computing* by Zomorodian)



"Bottom up" method

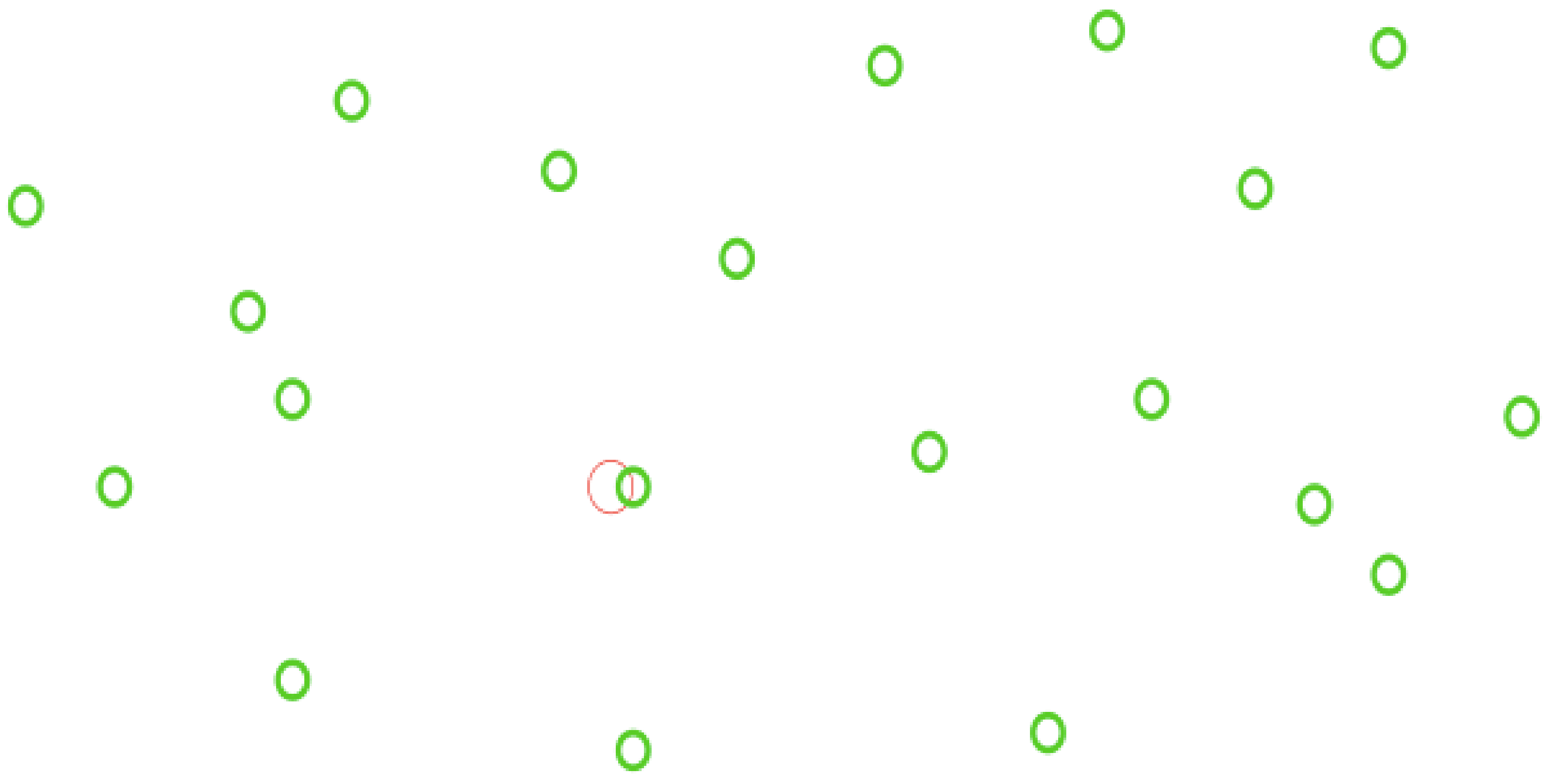
Empty Sphere Method

Take a bunch of point...



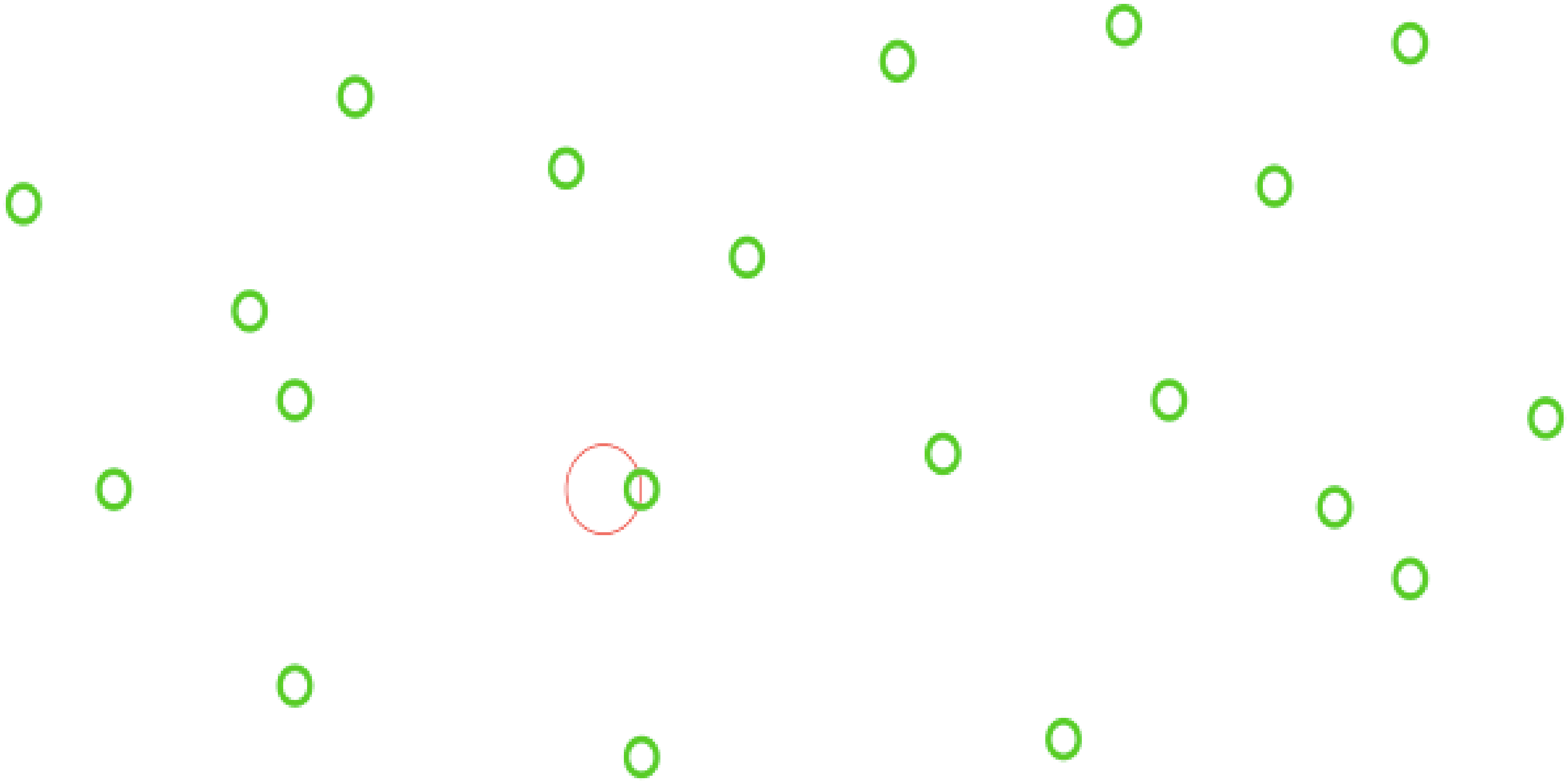
Empty Sphere Method

Search for empty spheres...



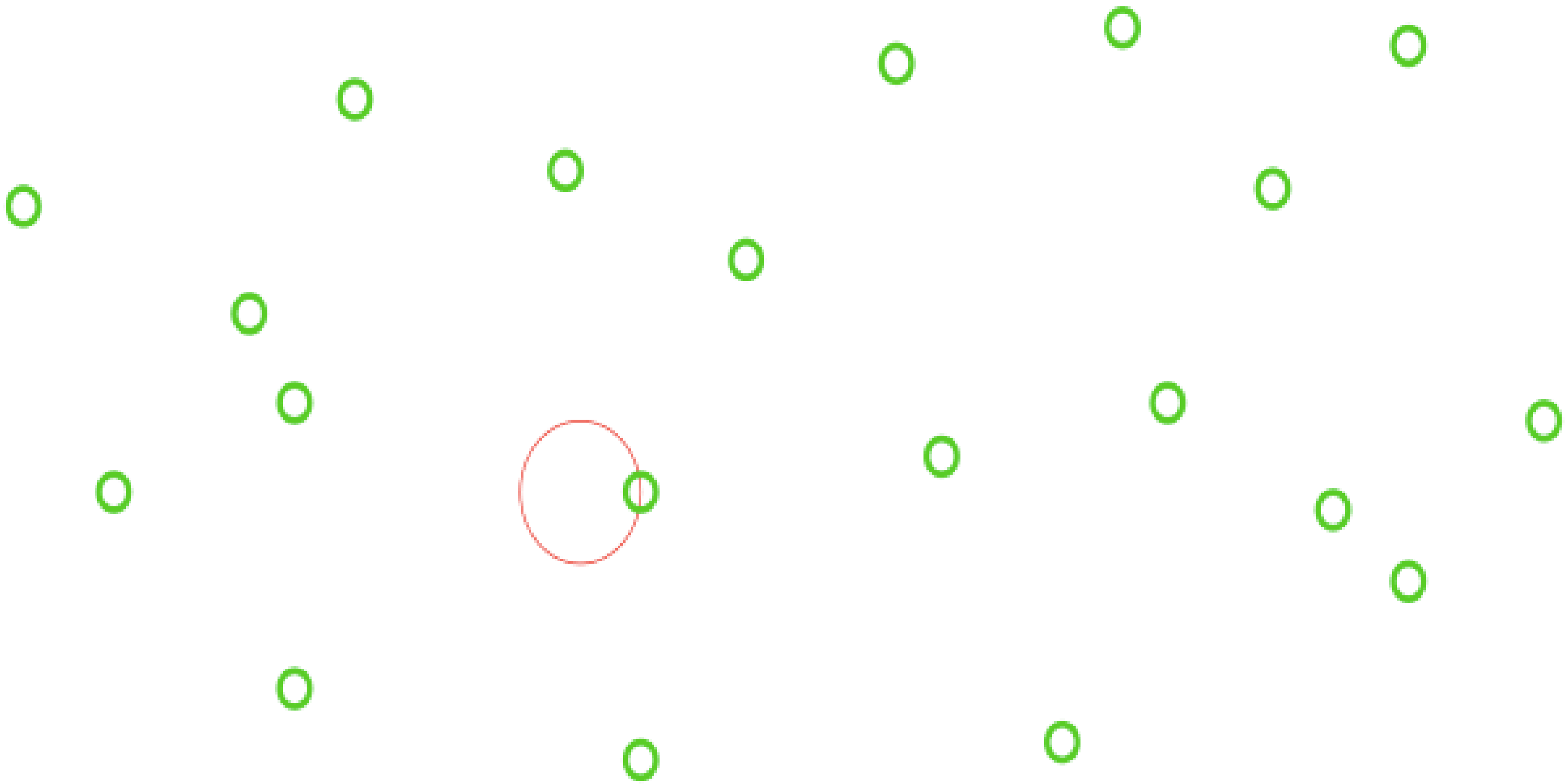
Empty Sphere Method

Search for empty spheres...



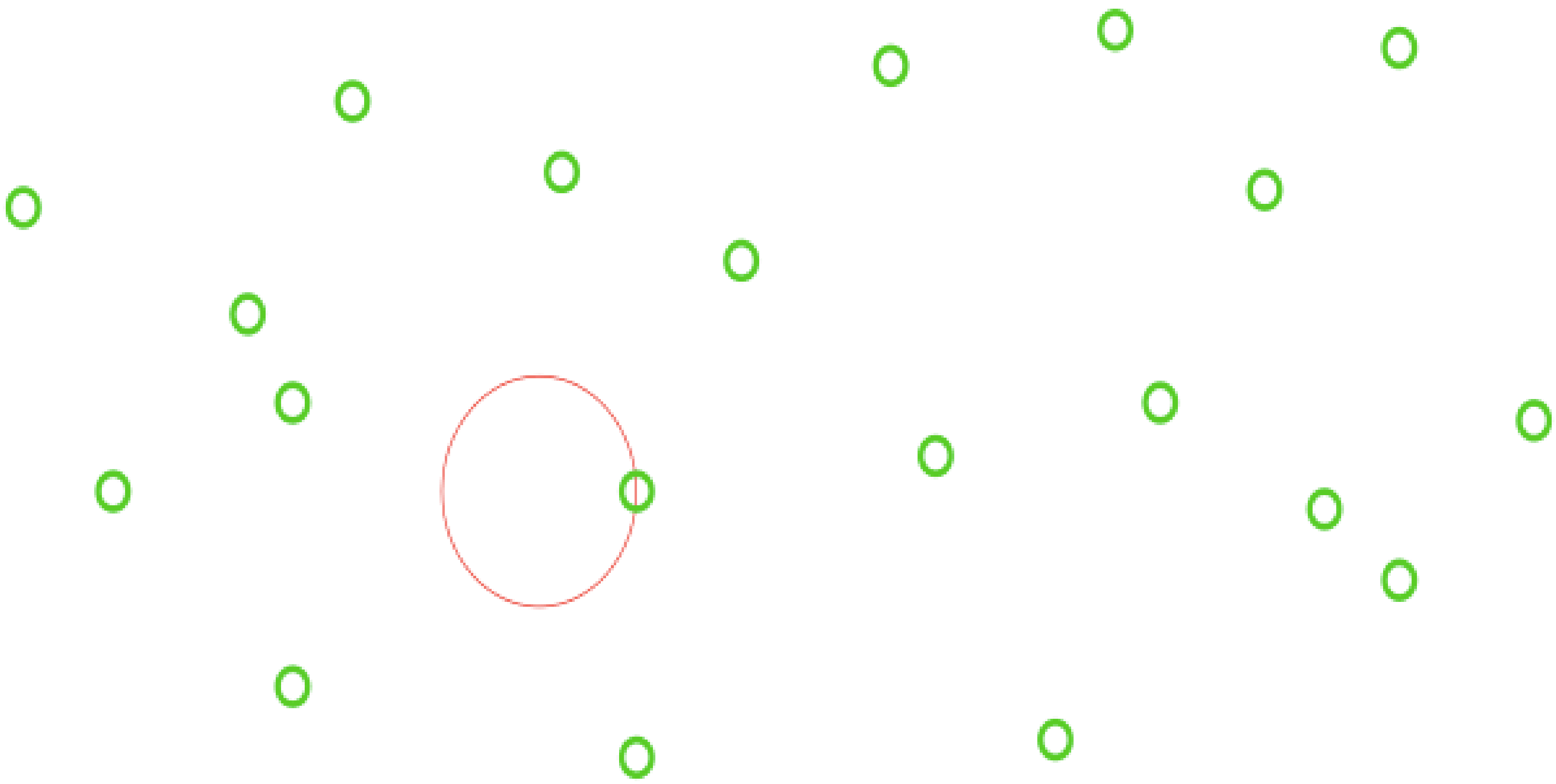
Empty Sphere Method

Search for empty spheres...



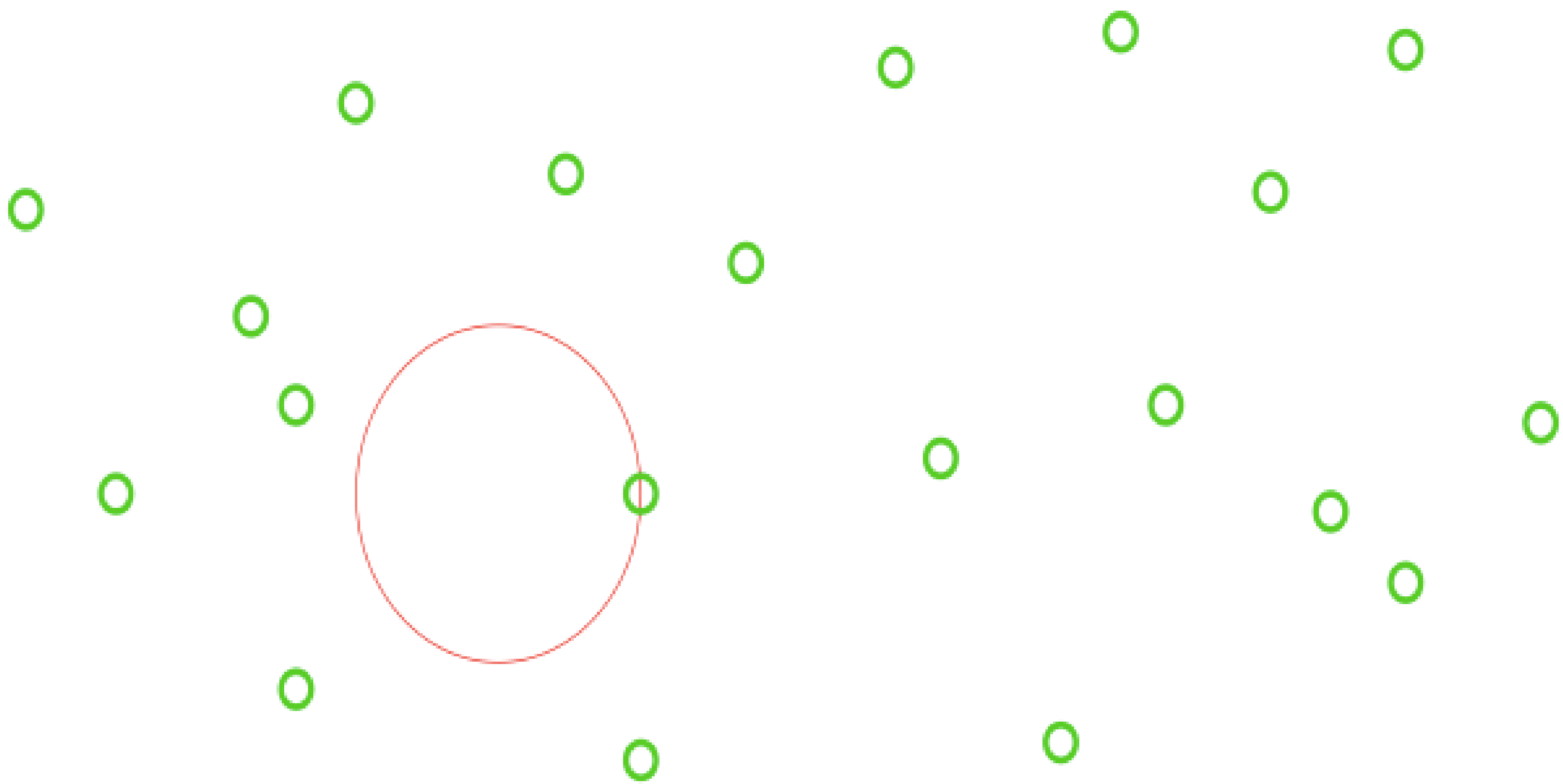
Empty Sphere Method

Search for empty spheres...



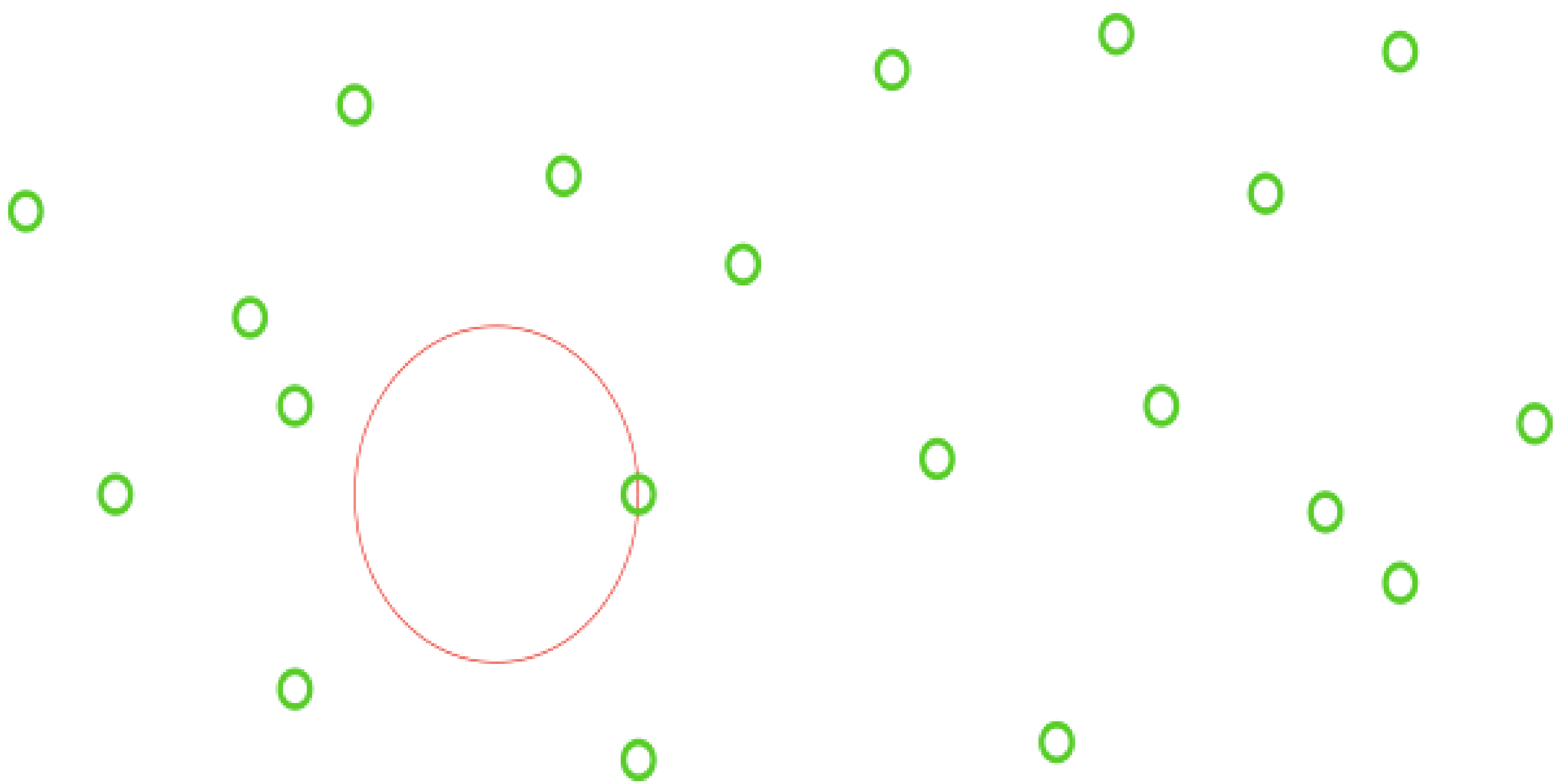
Empty Sphere Method

Search for empty spheres...



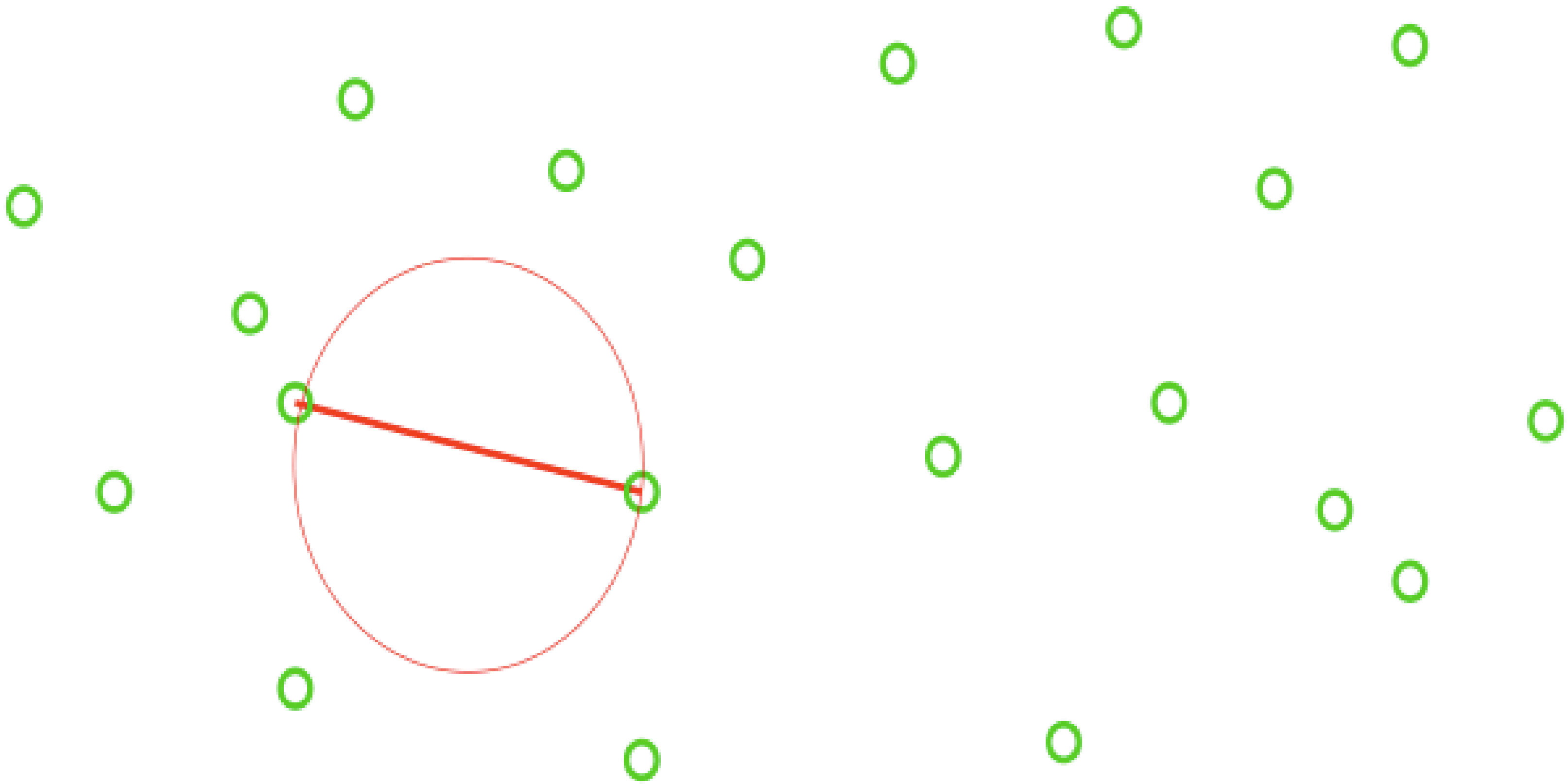
Empty Sphere Method

Search for empty spheres...



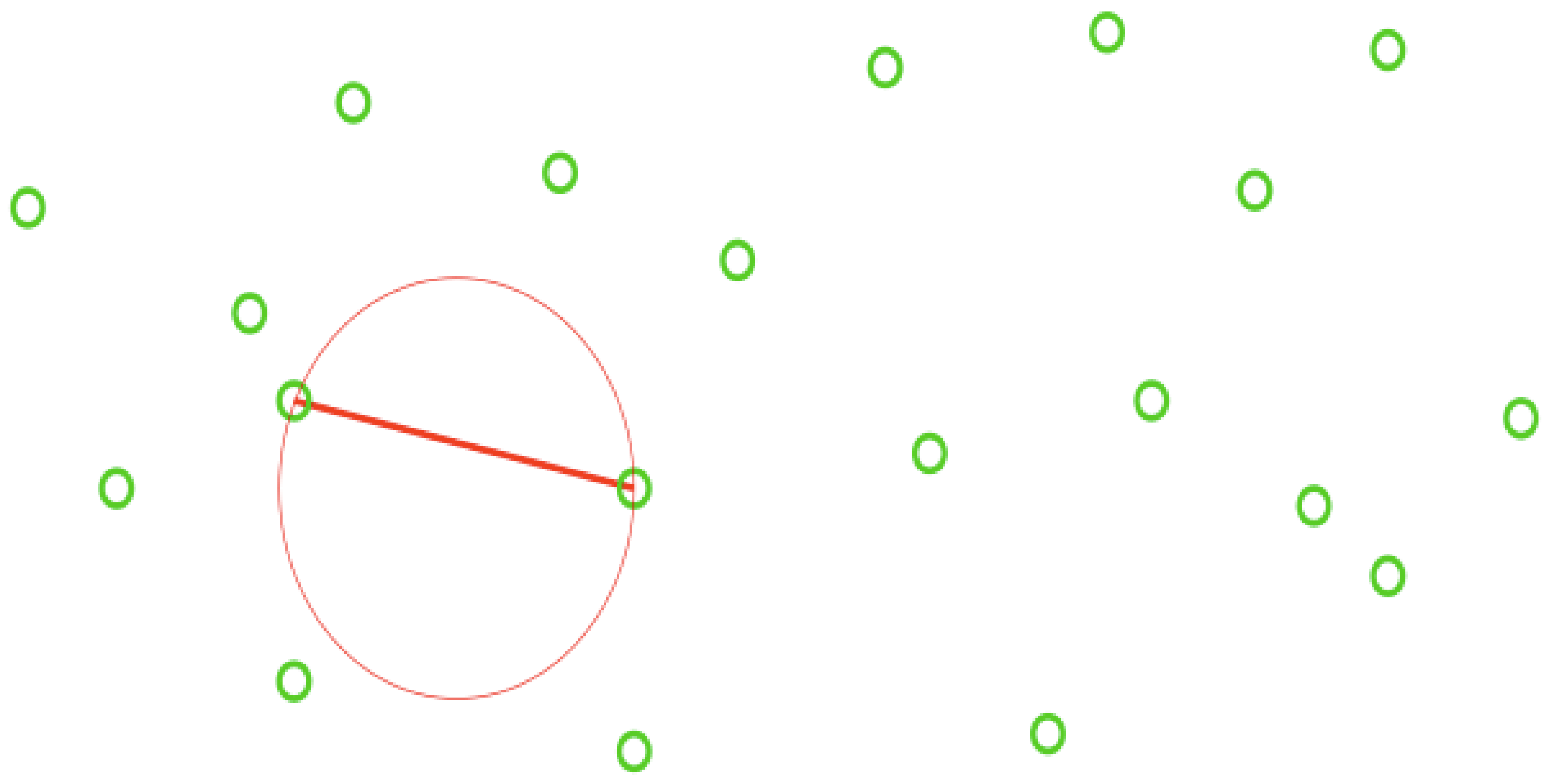
Empty Sphere Method

Search for empty spheres...



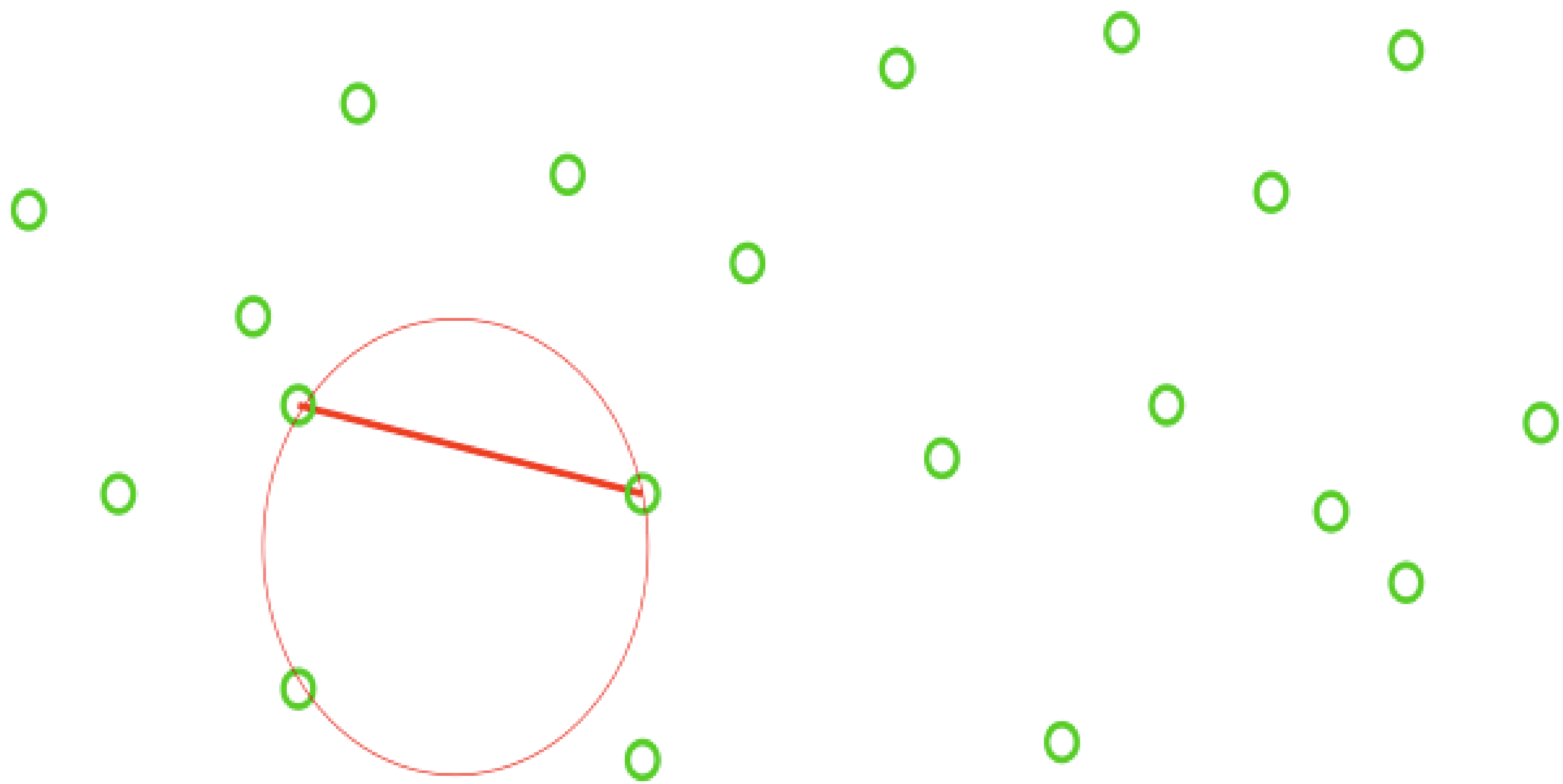
Empty Sphere Method

Search for empty spheres...



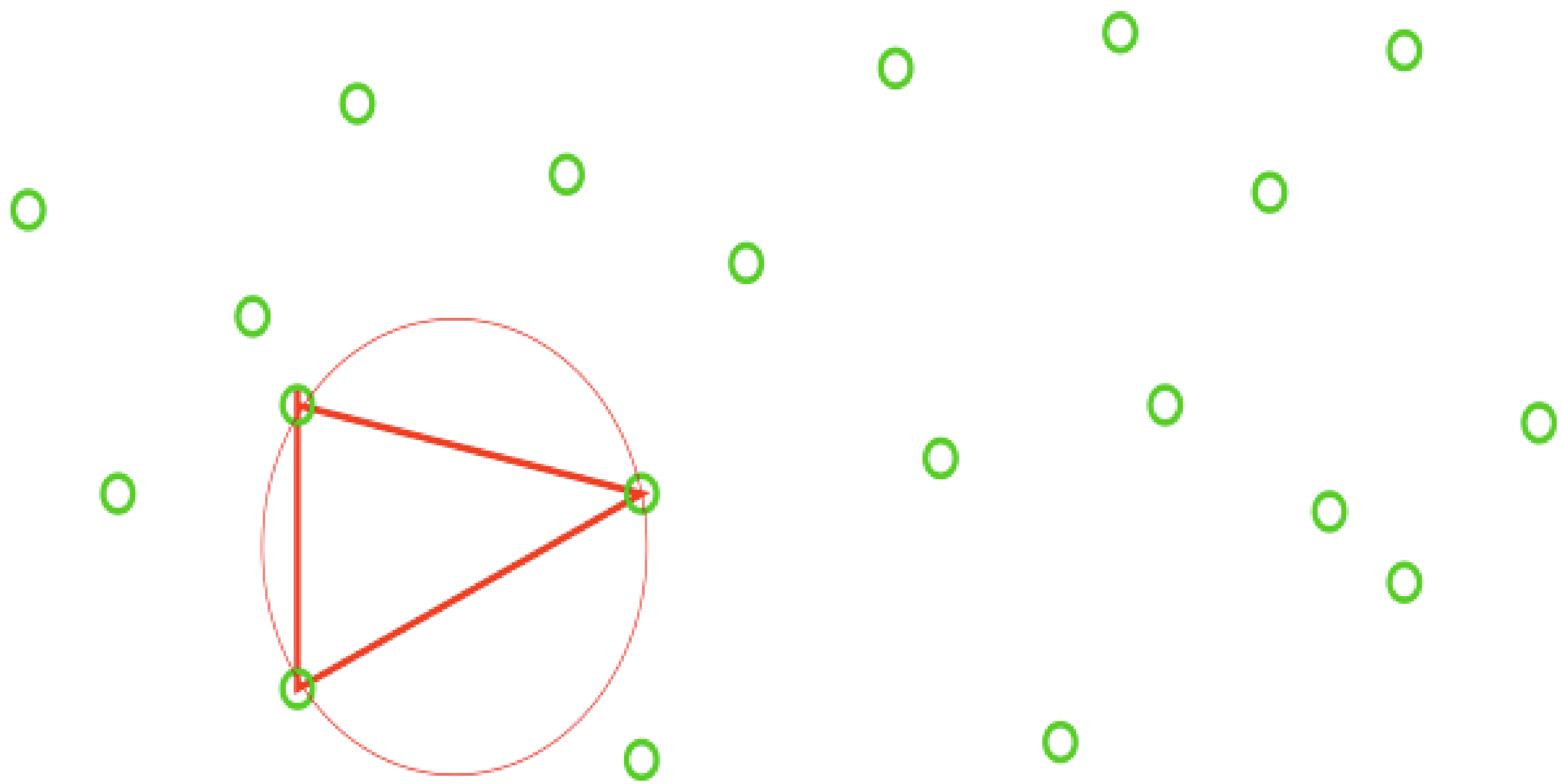
Empty Sphere Method

Search for empty spheres...



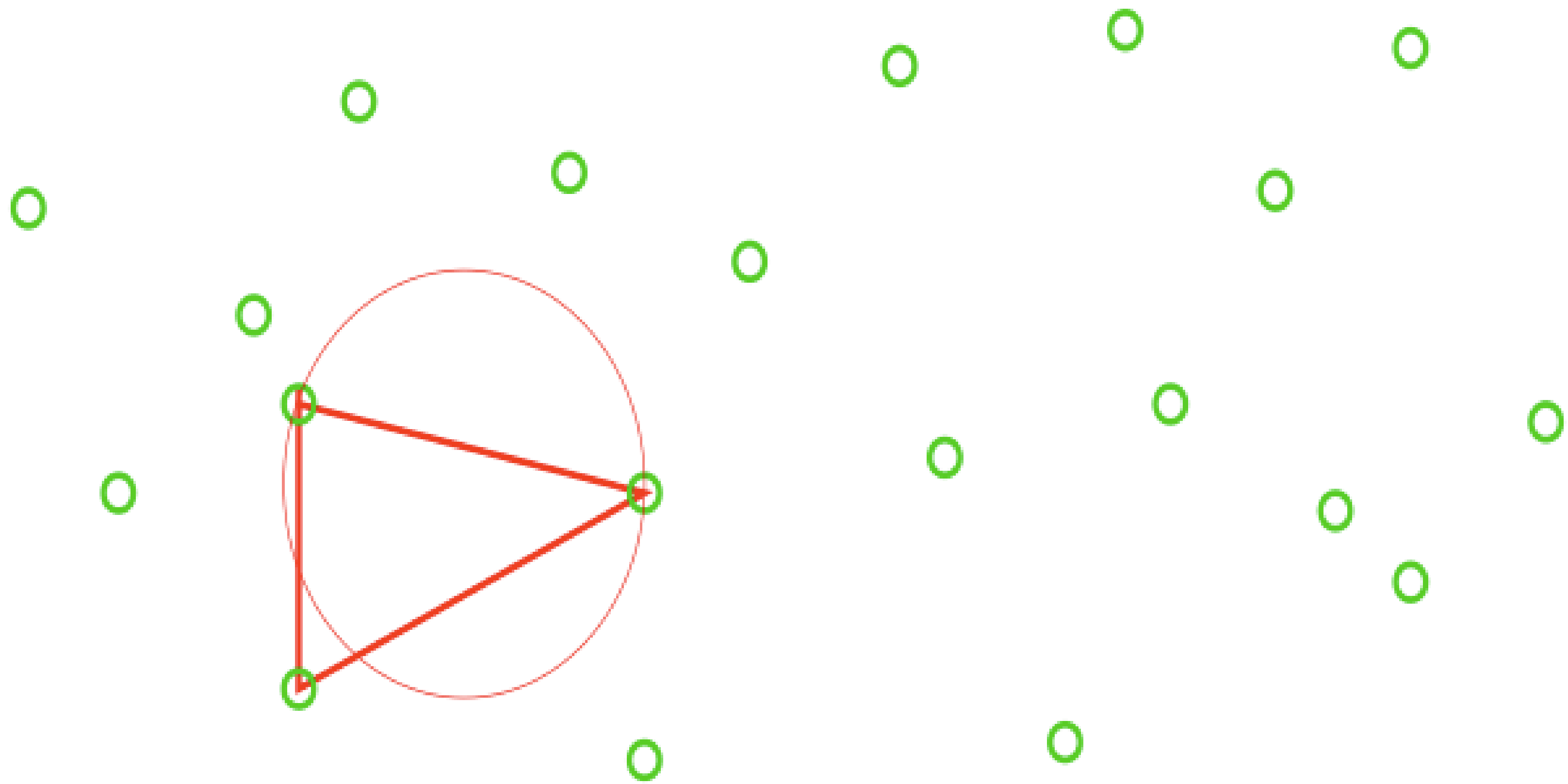
Empty Sphere Method

Search for empty spheres...



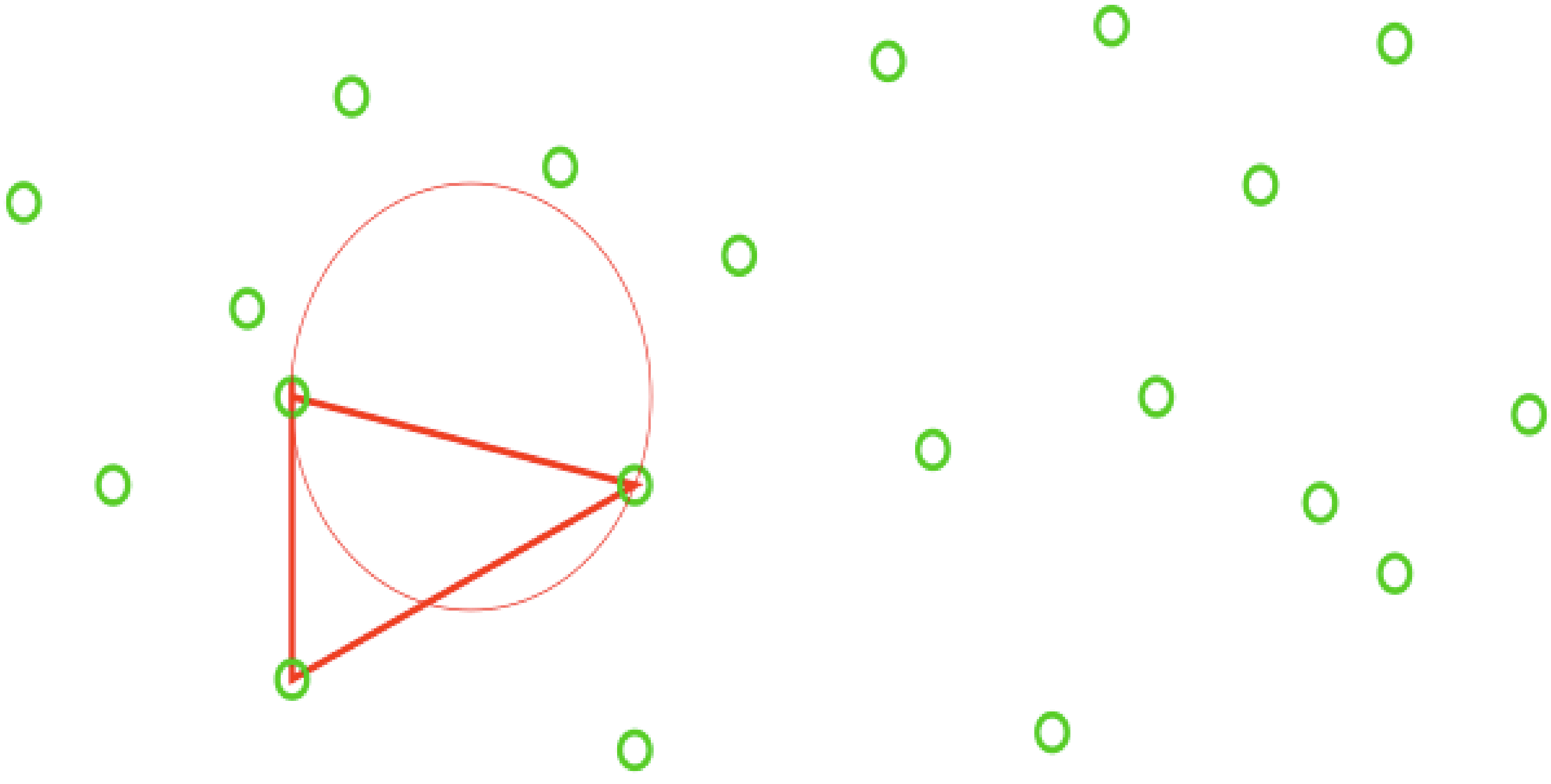
Empty Sphere Method

Search for empty spheres...



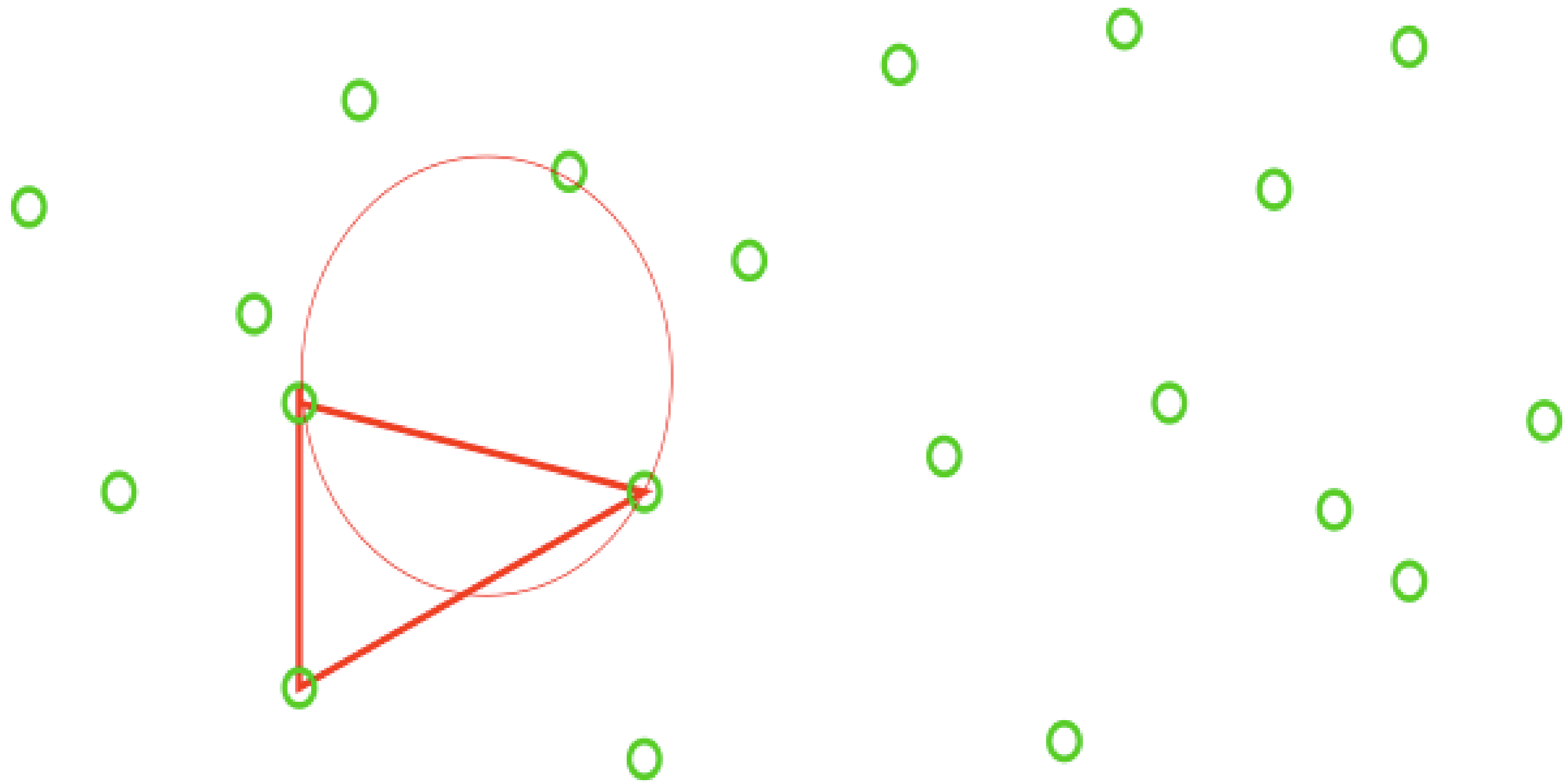
Empty Sphere Method

Search for empty spheres...



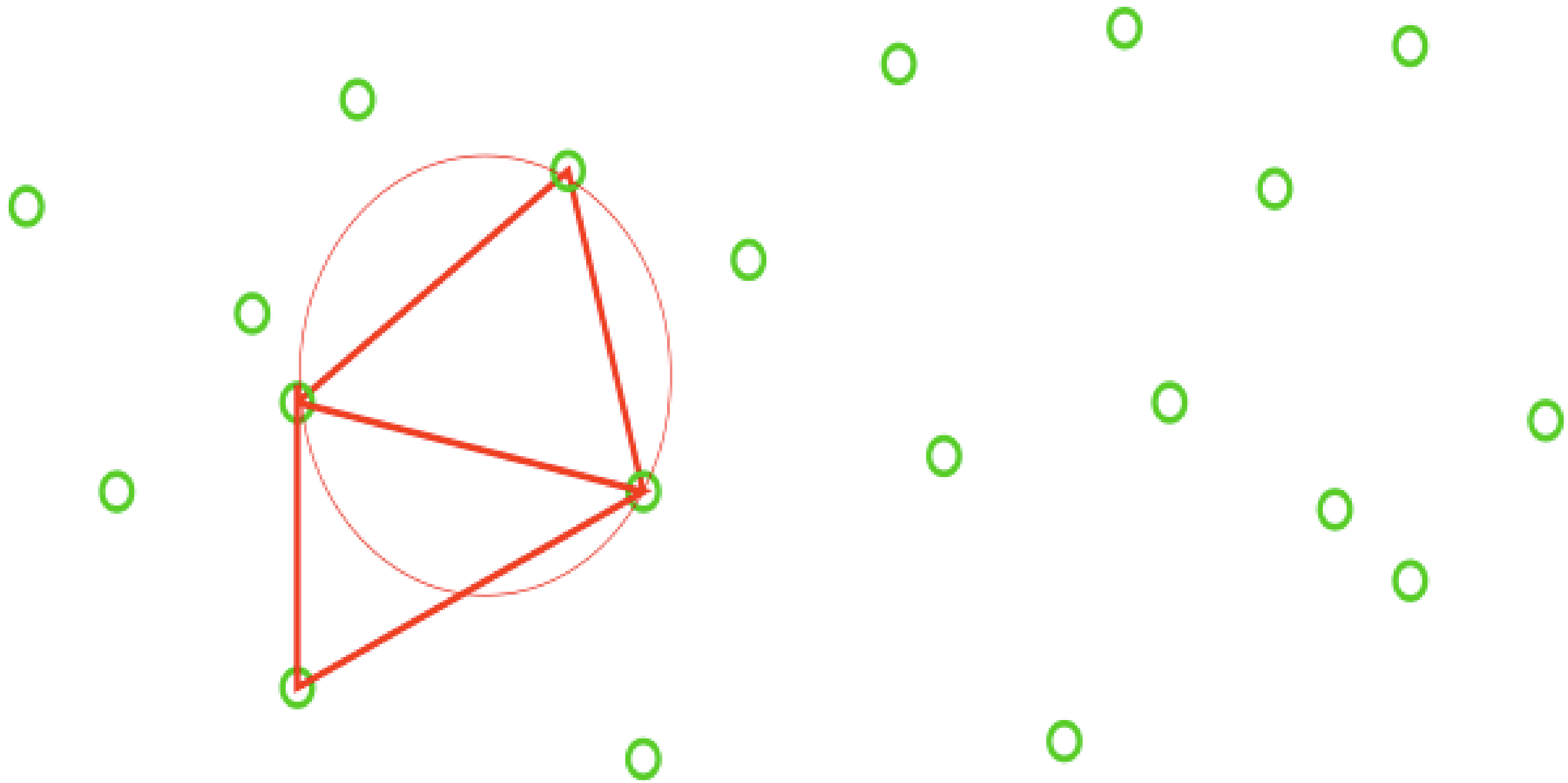
Empty Sphere Method

Search for empty spheres...



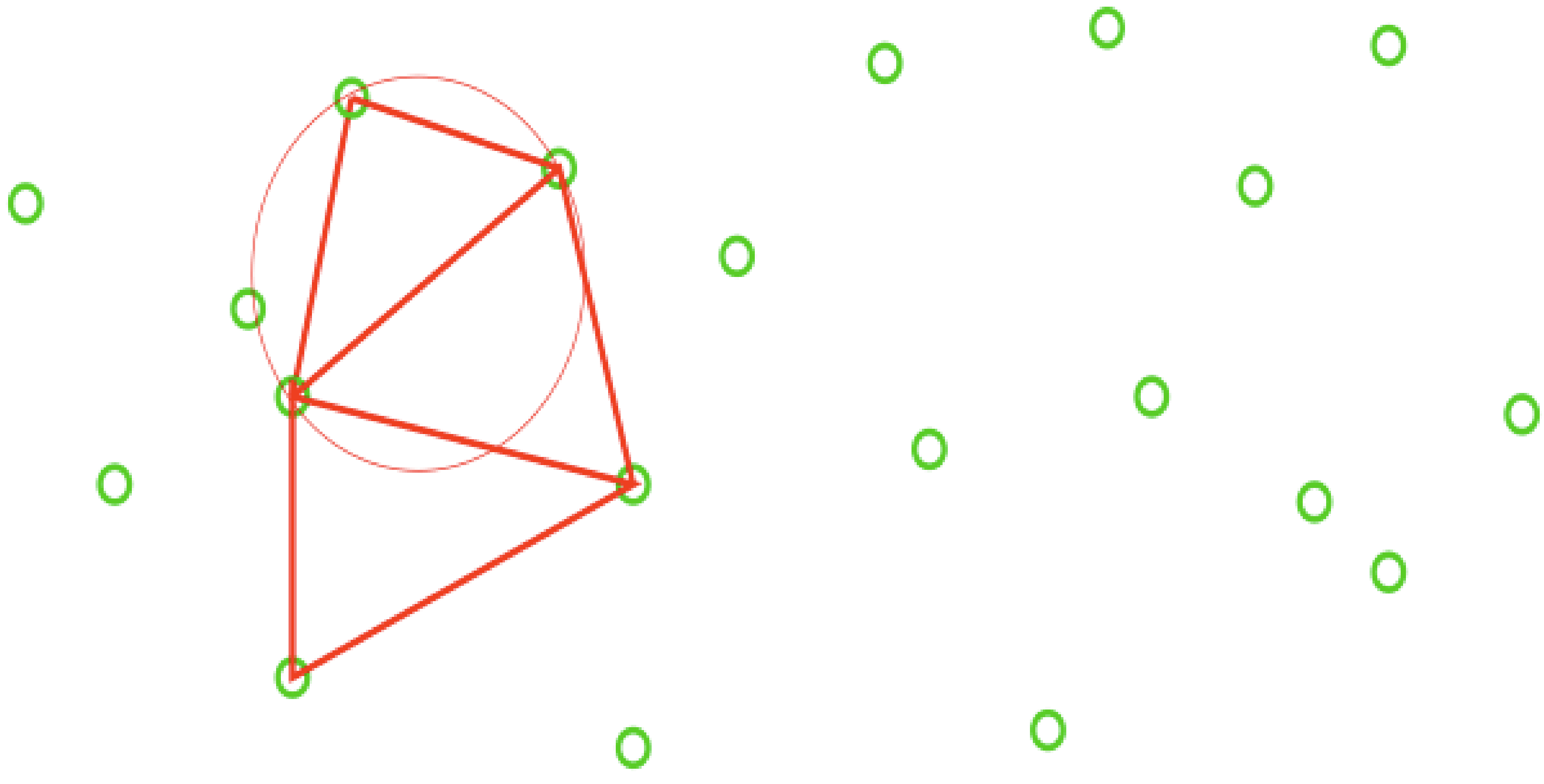
Empty Sphere Method

Search for empty spheres...



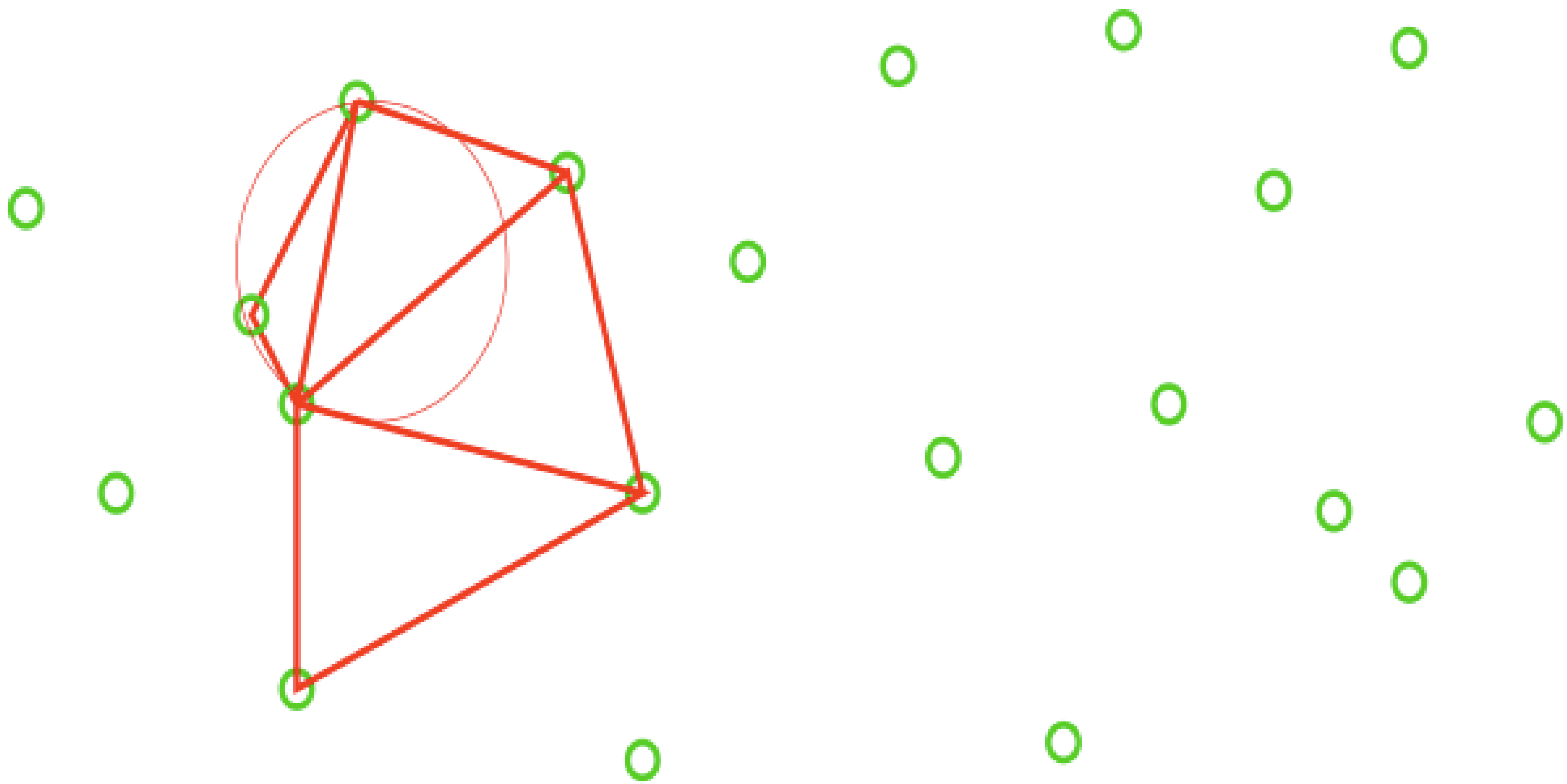
Empty Sphere Method

Search for empty spheres...



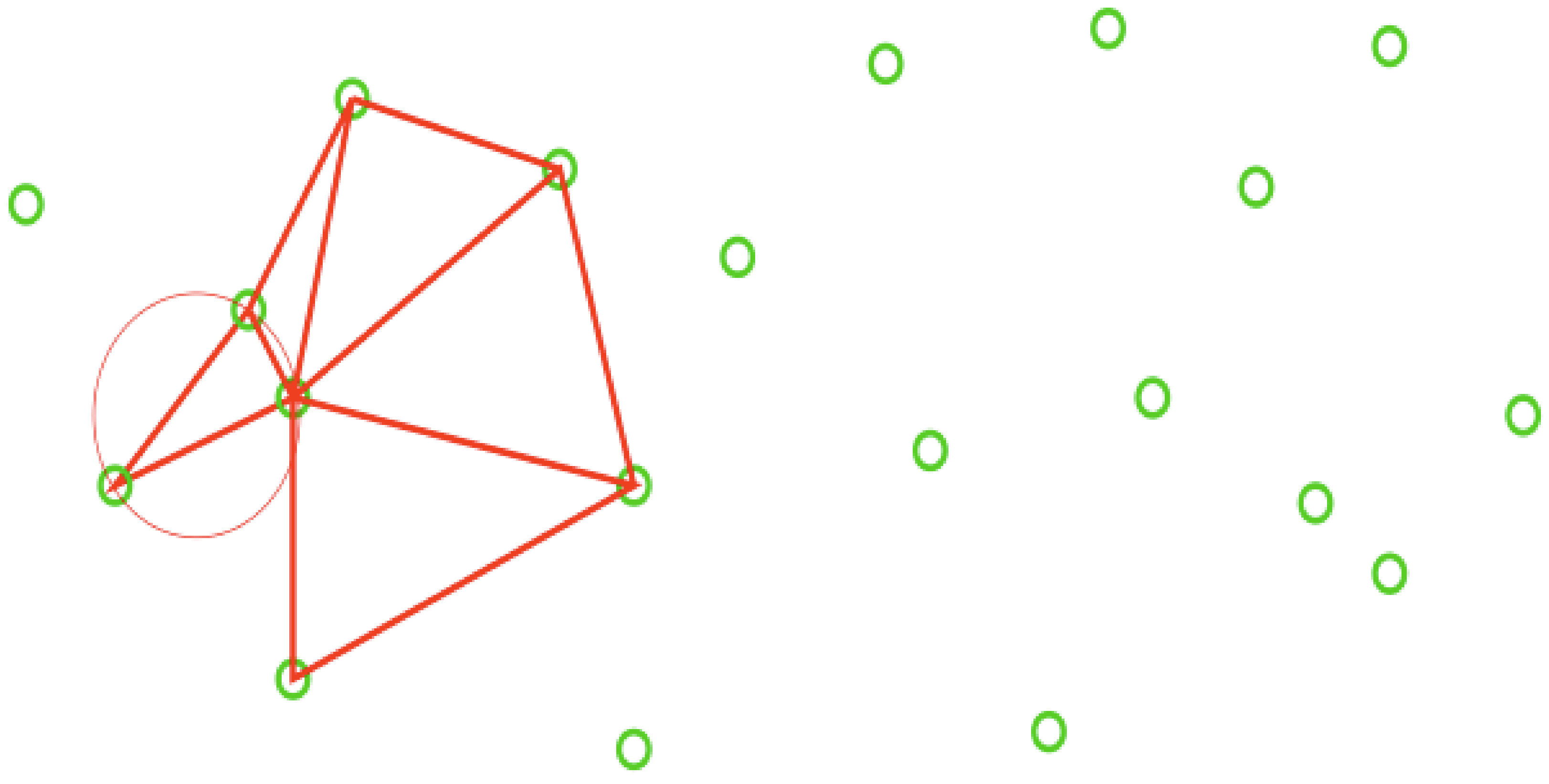
Empty Sphere Method

Search for empty spheres...



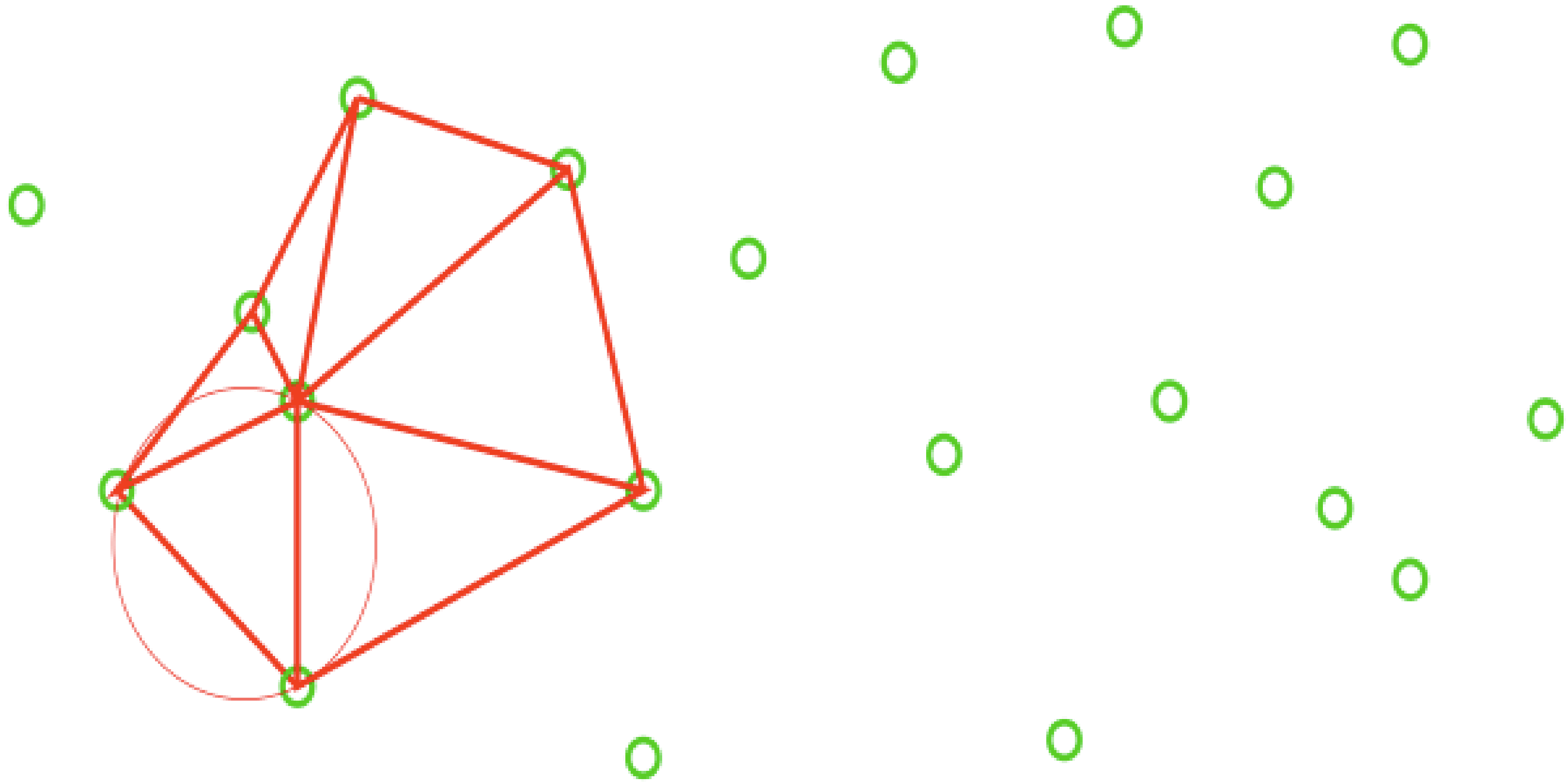
Empty Sphere Method

Search for empty spheres...



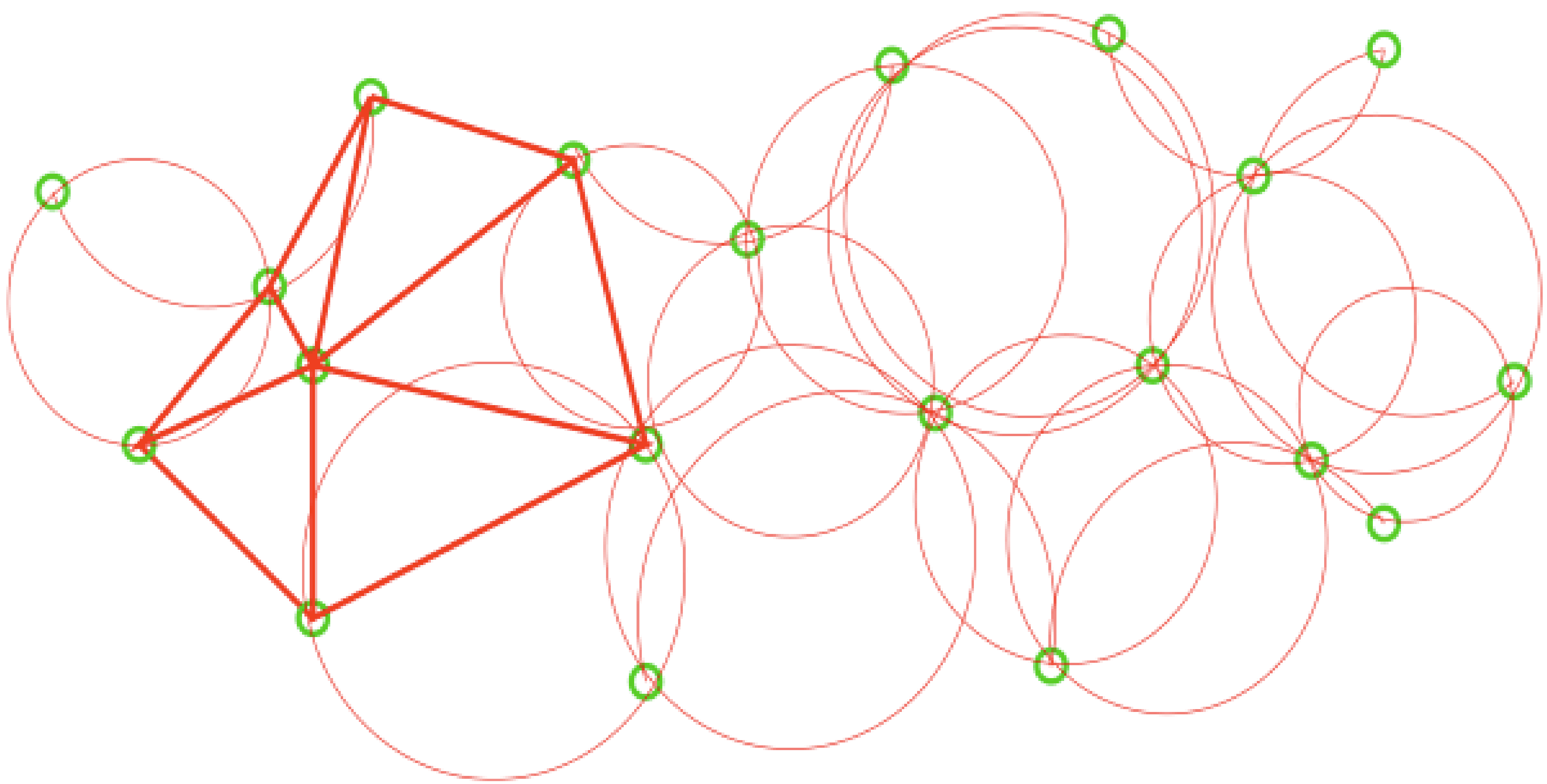
Empty Sphere Method

Search for empty spheres...



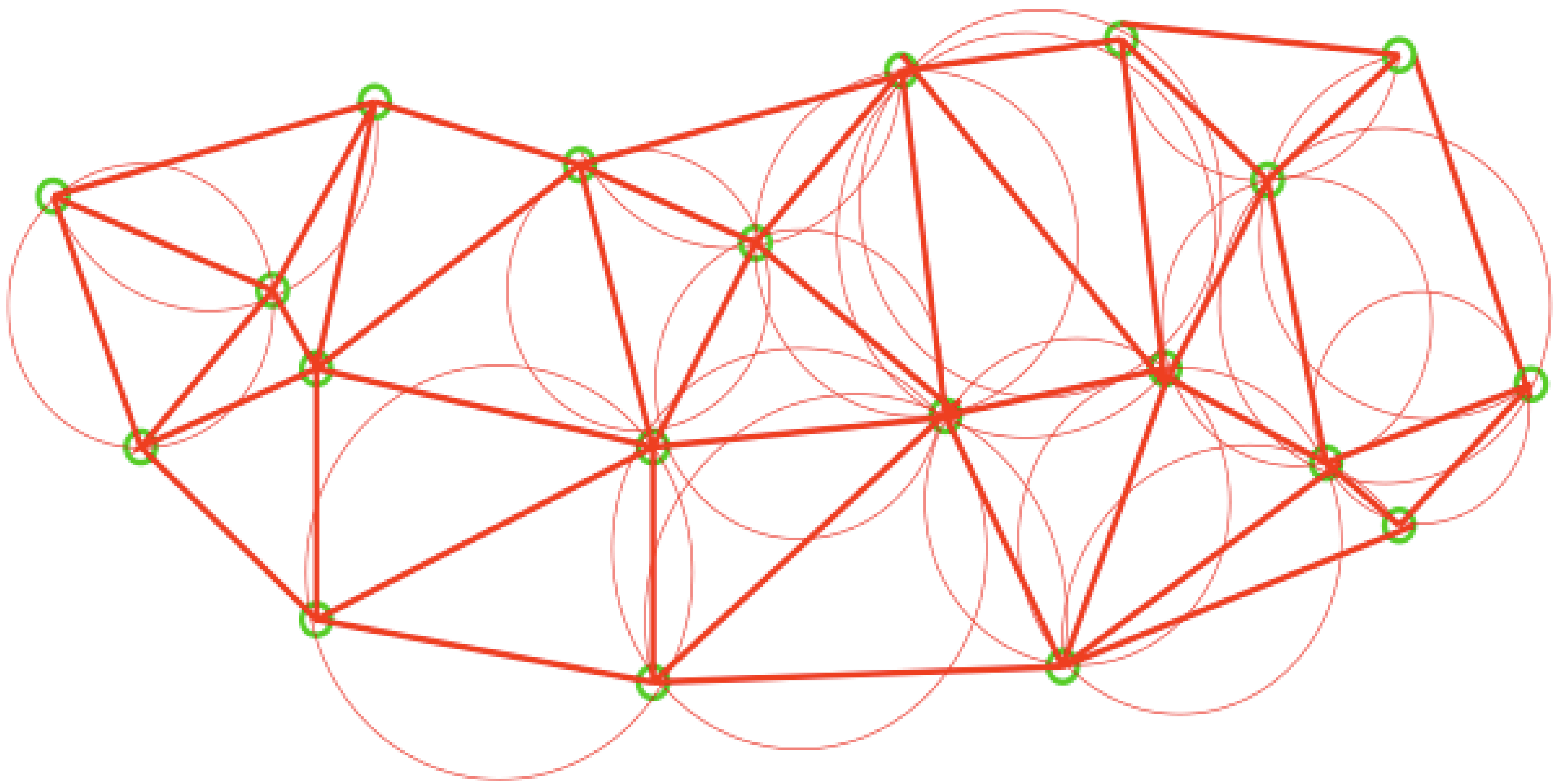
Empty Sphere Method

Search for empty spheres...

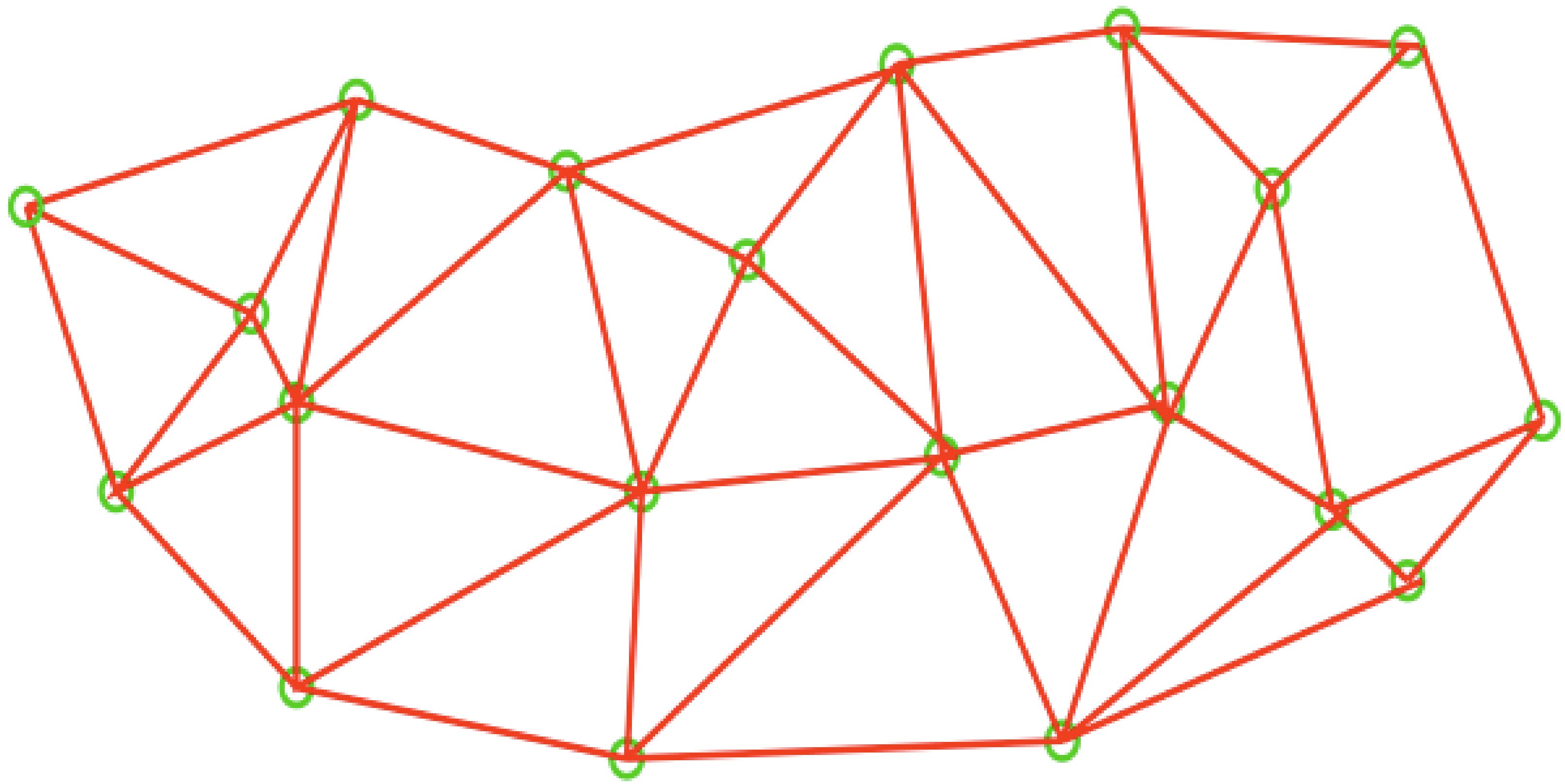


Empty Sphere Method

Search for empty spheres...



Viola, le *Delaunay Decomposition*.



“Top Down”: Spectral Clustering

1. Form the Laplacian

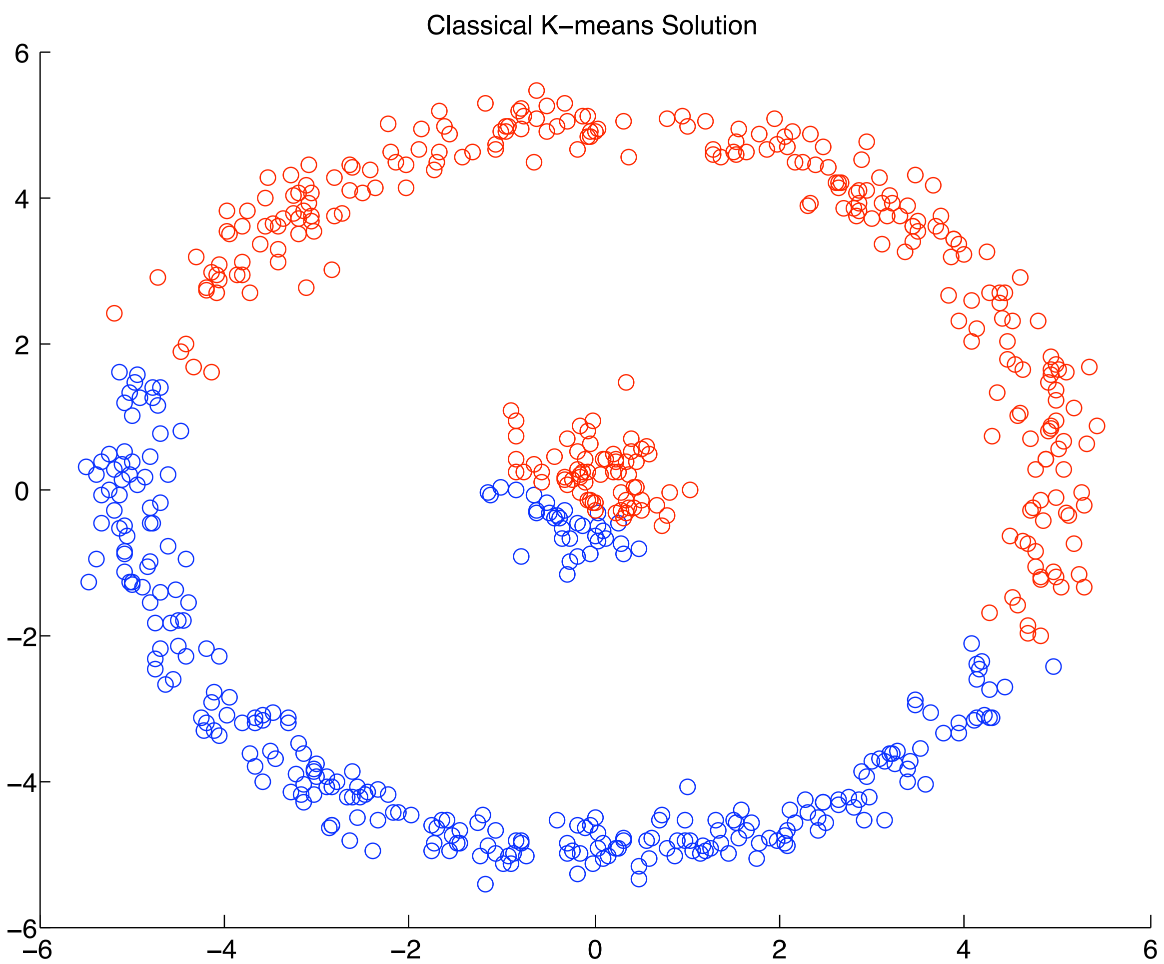
2. List the K eigenvectors corresponding to the K smallest non-zero eigenvalues. (This is our dimension reduction into K-dimensional Euclidean space).

3. Normalize each row so all our points lie on a sphere.

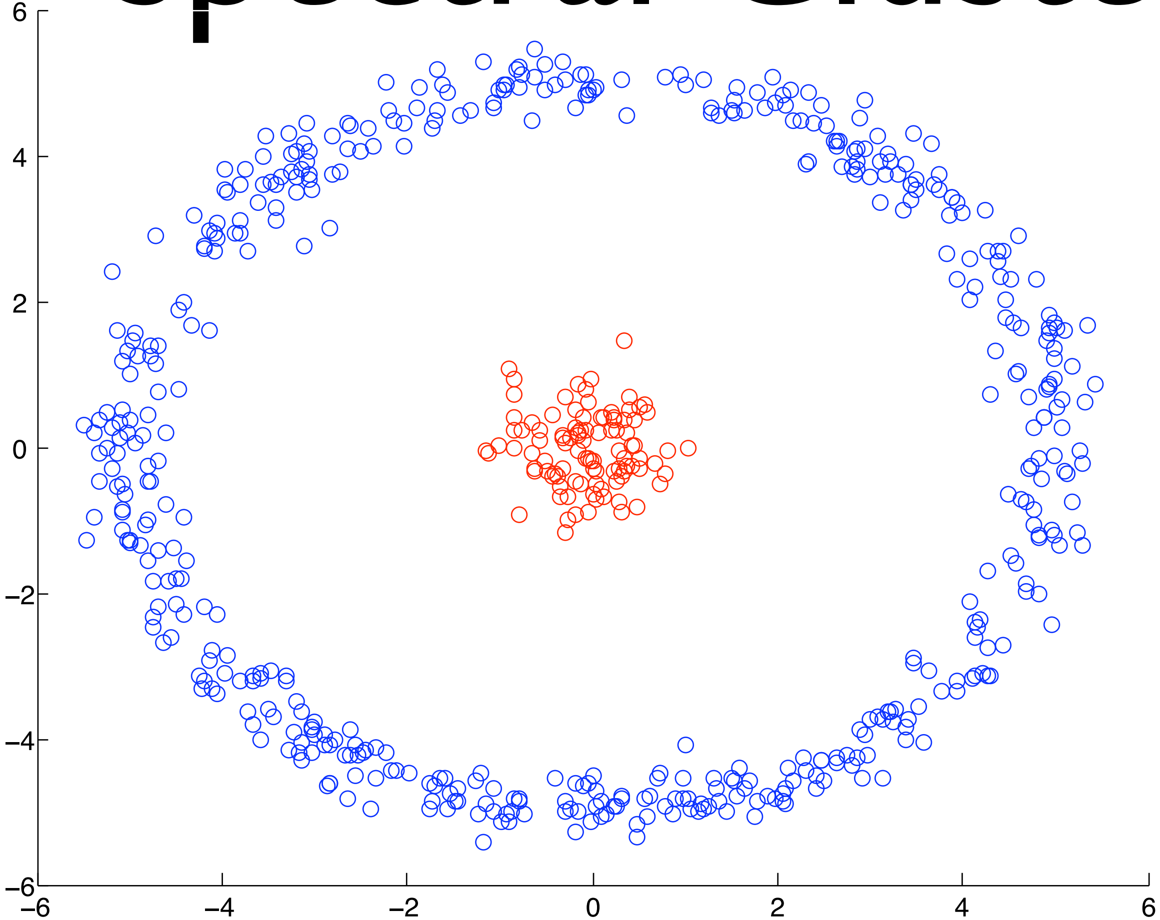
4. Apply k-means.

See: On spectral clustering: Analysis and an algorithm. A. Y. Ng, M. I. Jordan, and Y. Weiss. In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*, 2002.

Classical K-means



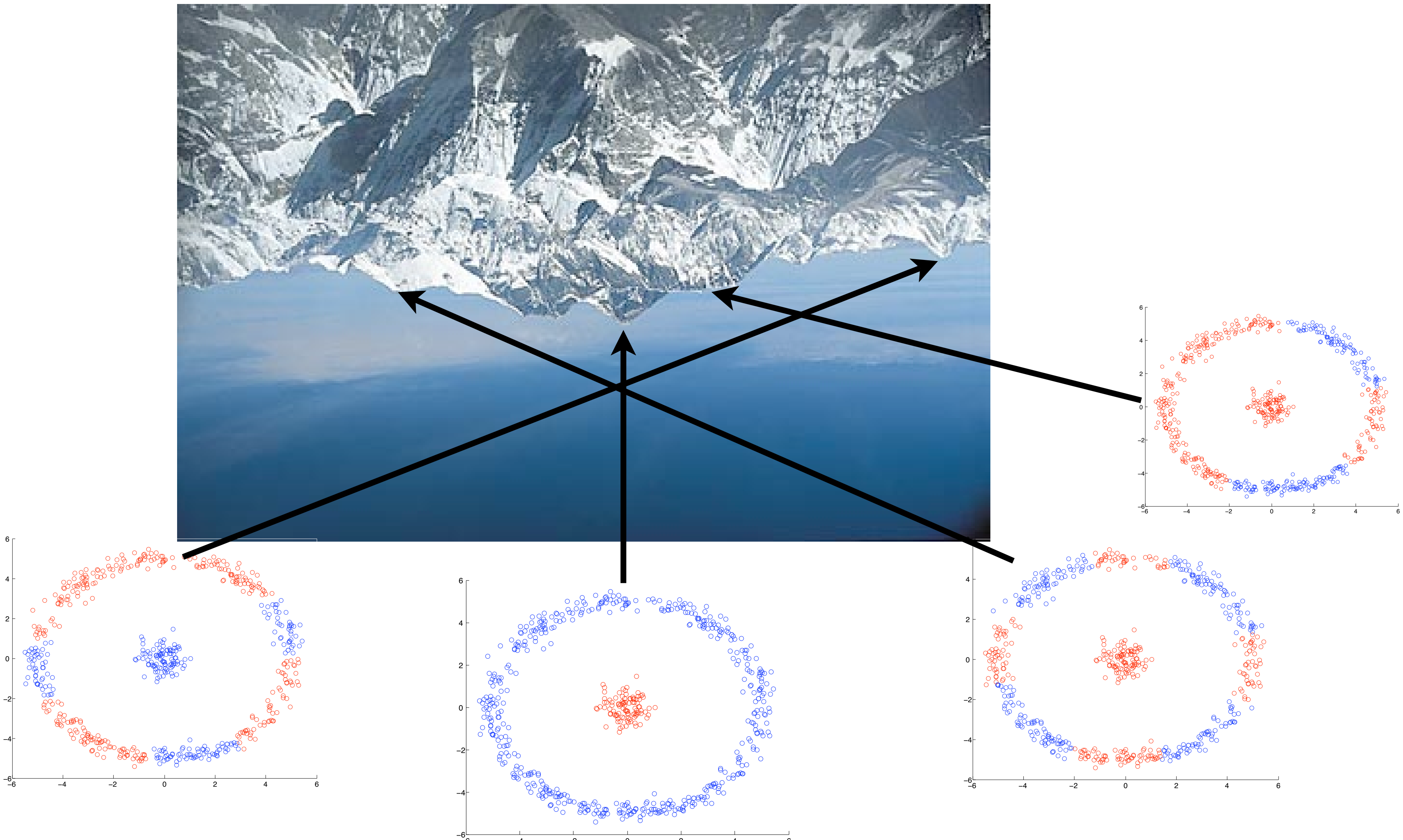
Spectral Clustering



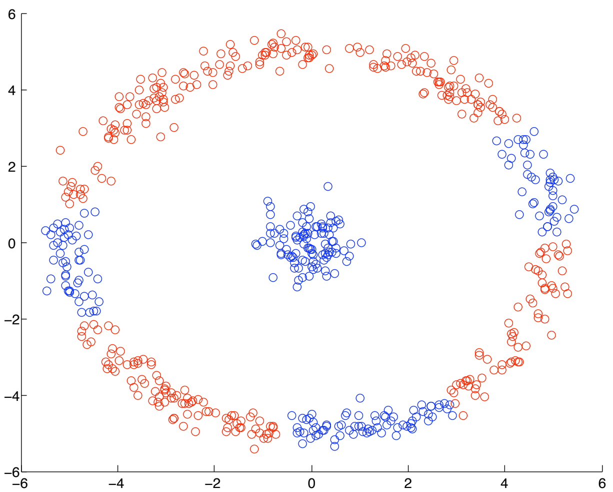
Manifold Learning

$$L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 488$$

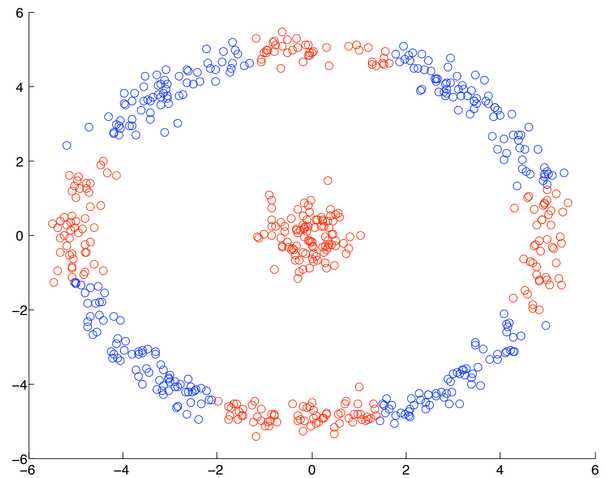
Himalayas



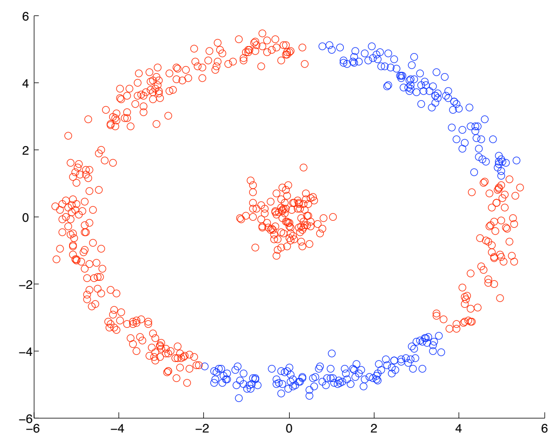
Watch Out! 100 convergent runs of the k-means algorithm were performed.



$$L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 536$$

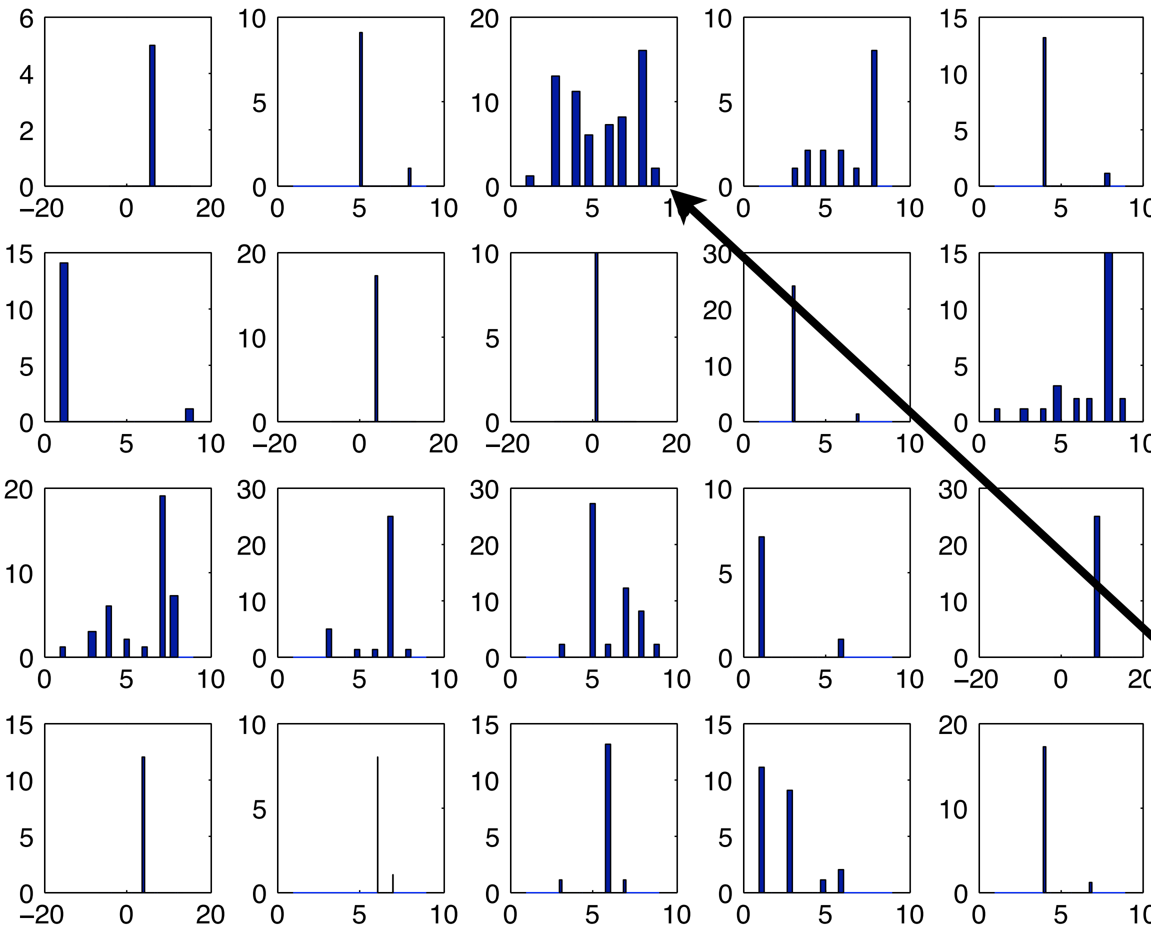


$$L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 538$$



$$L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 534$$

The Spectral Market



1 Basic Materials
2 Conglomerates
3 Consumer Goods
4 Financial
5 Healthcare
6 Industrial Goods
7 Services
8 Technology
9 Utilities

Guess who that is?

Another popular "bottom up" method

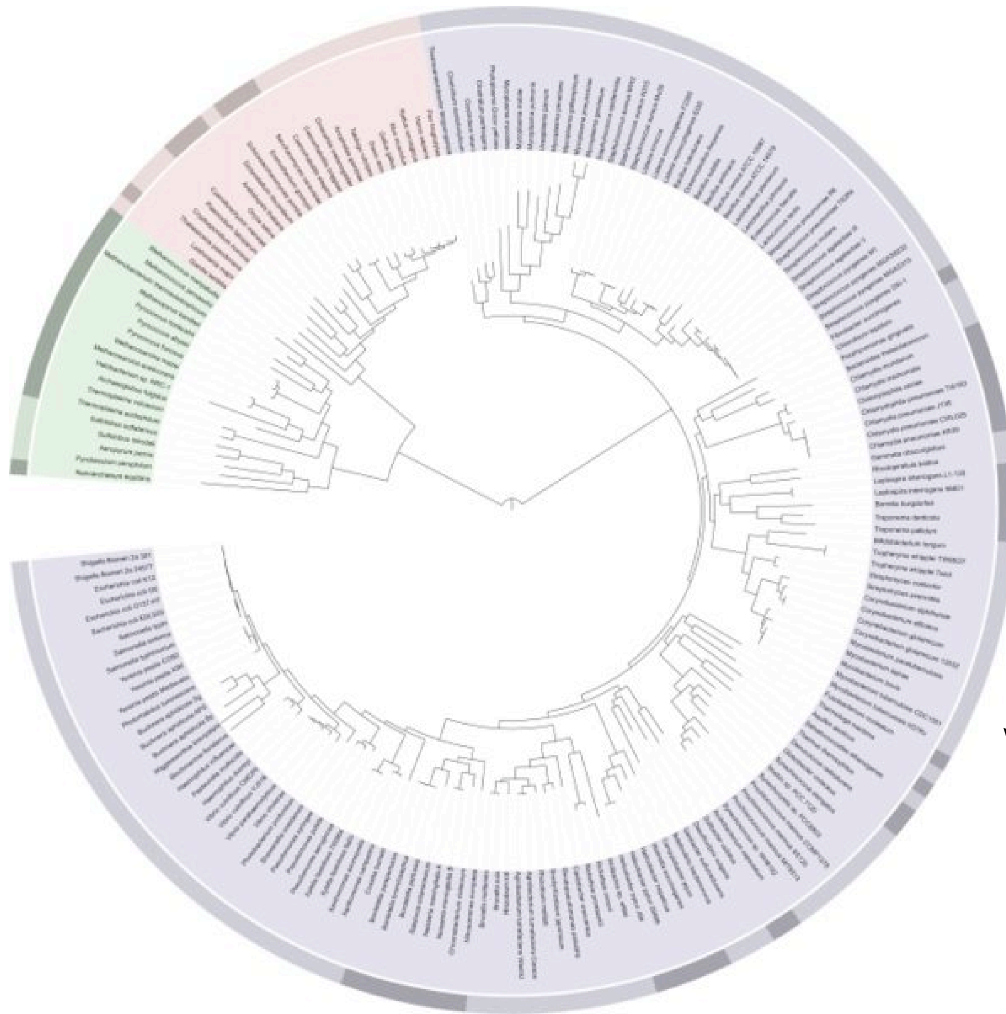
*-Linkage Clustering

*=*single-linkage*, *complete-linkage*, *average-linkage* clustering....

Why are we looking all these different types of clustering procedures!....

No Free Lunch Theorem: Each reasonable cluster algorithm will be optimal/sub-optimal under the appropriate conditions.   (KNOW YOUR SYSTEM!)

Tree of Life



A dendrogram

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

This discussion was taken from the very nice online tutorial:

http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html

2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

We need to choose a measure of closest!

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

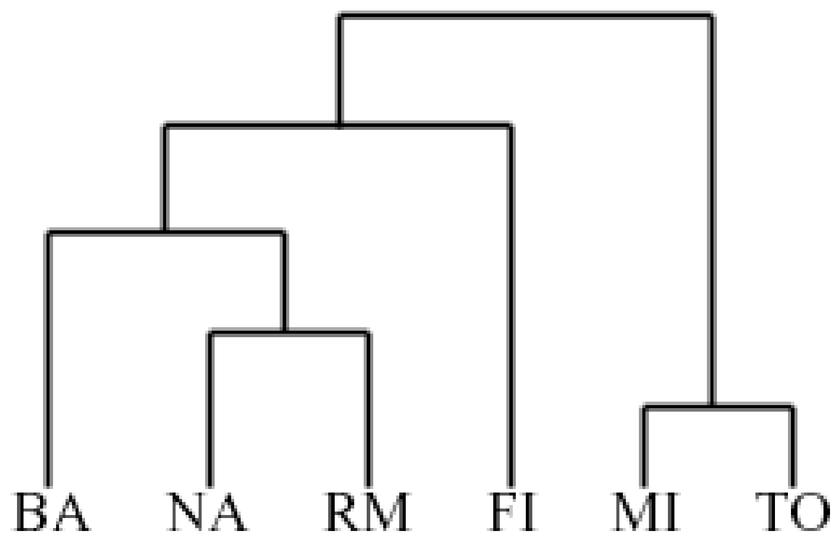
	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

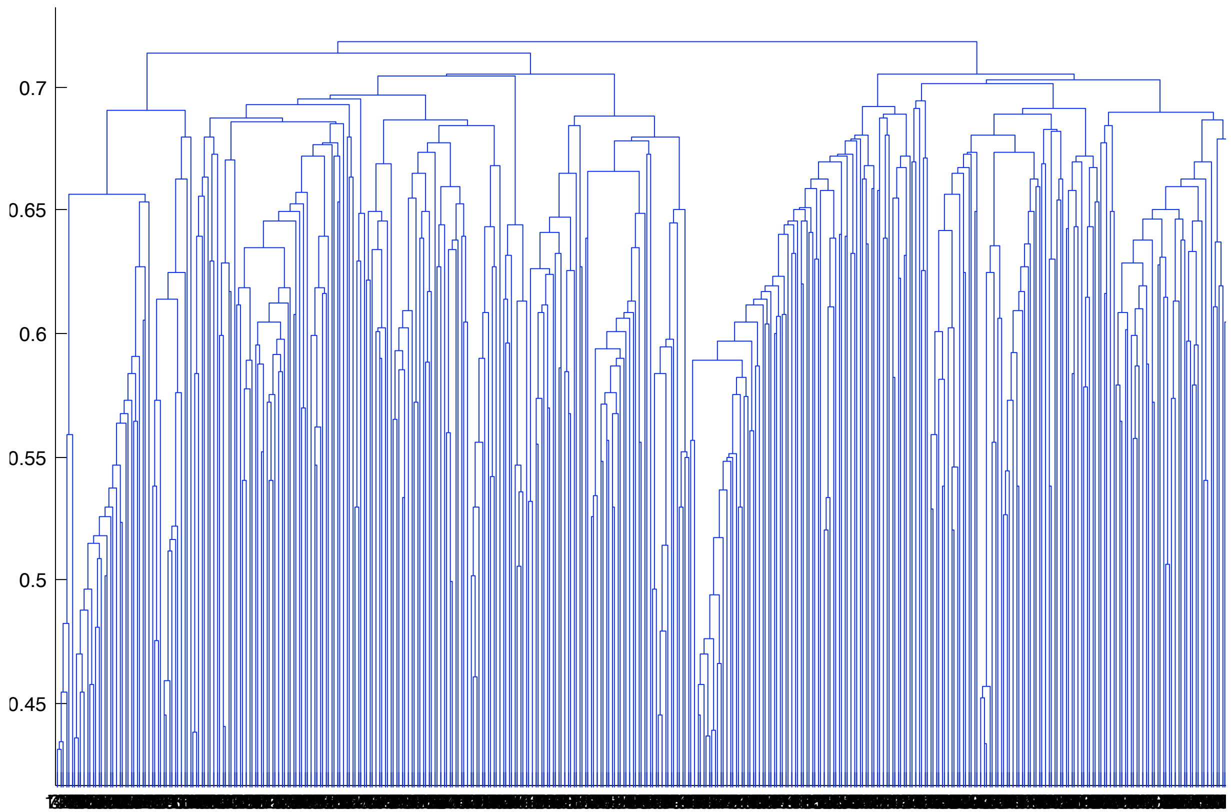
	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

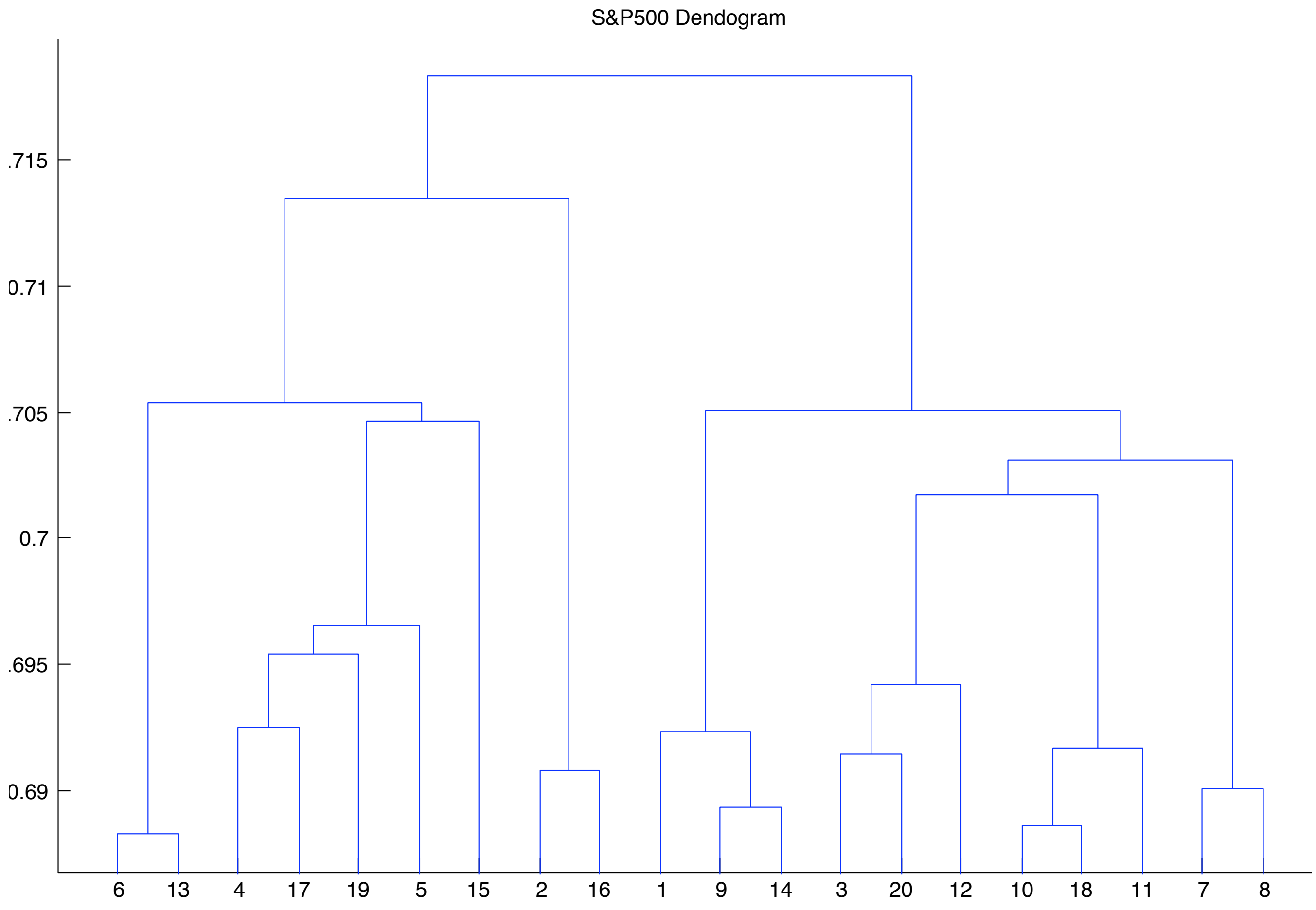
	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0



Our Market...

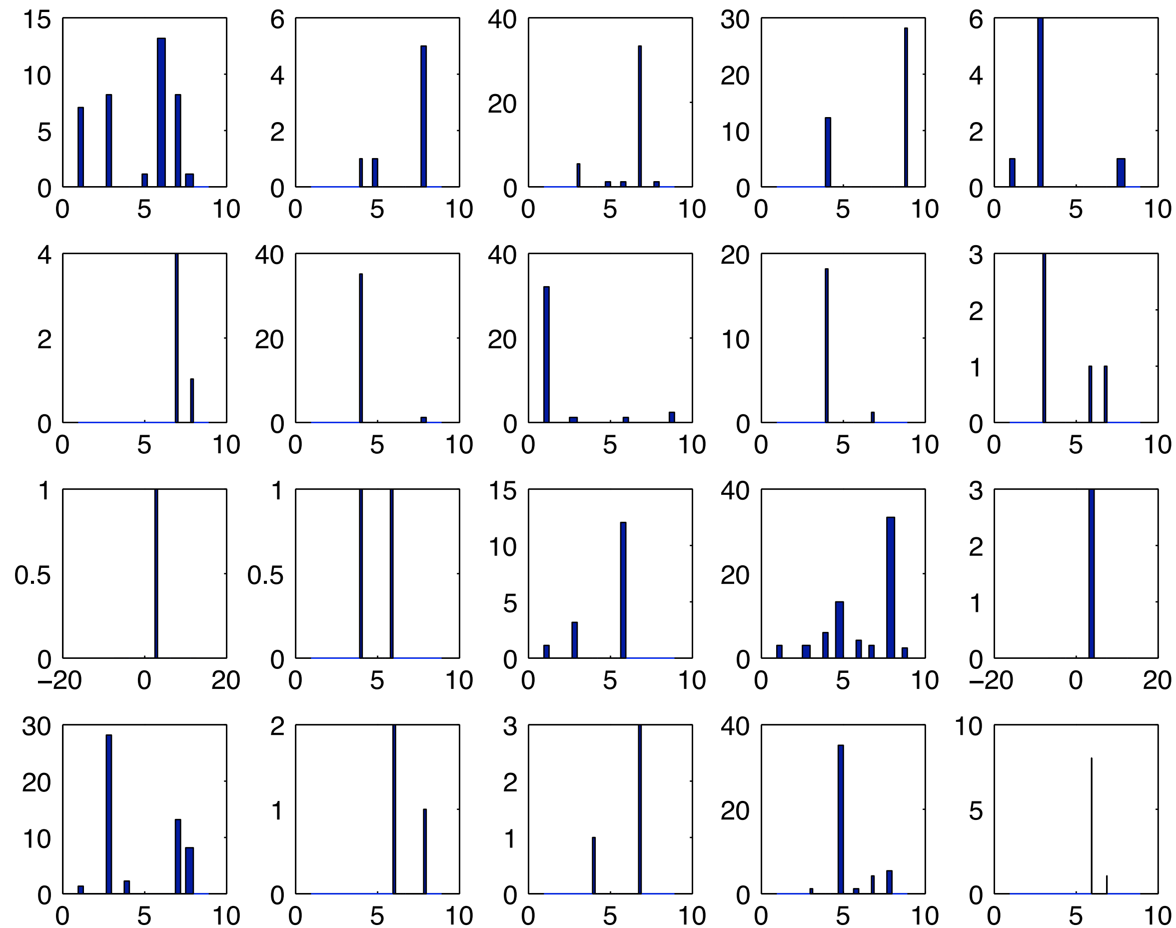


The S&P500

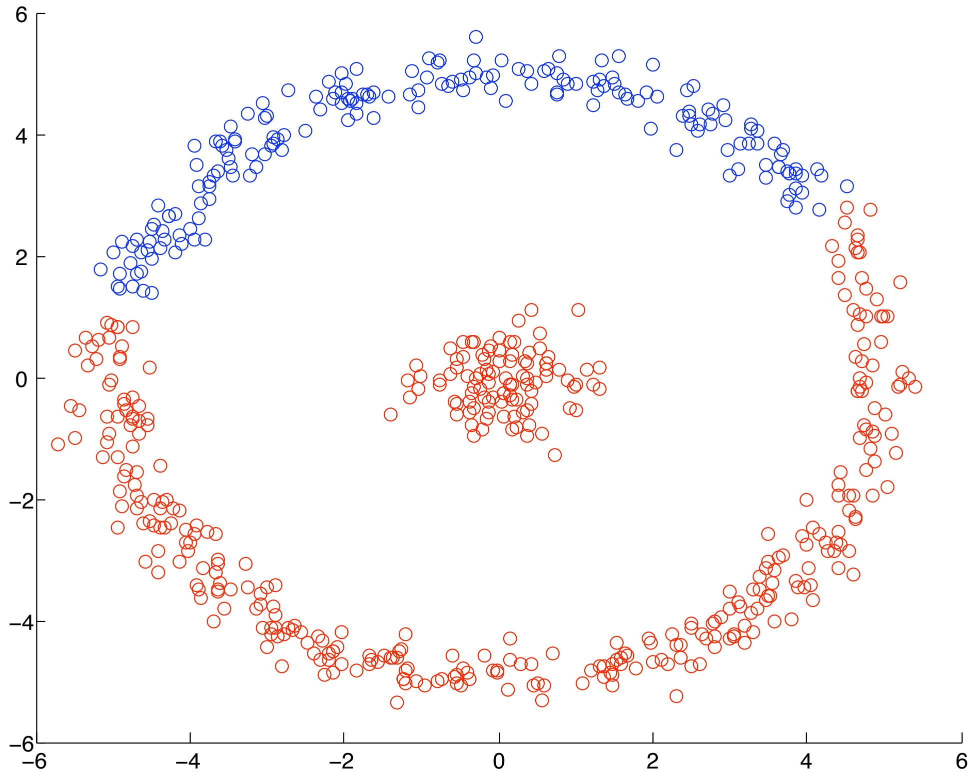


S&P500 Dendogram
.715
0.71
.705
0.7
.695
0.69
6
13
4
17
19
5
15
2
16
1
9
14
3
20
12
10
18
11
7
8

Sector break down of linkage clusters



If we apply it to our "manifold"



$$d_{Average}(C_k, C_l) = \frac{\sum_{x \in C_k y \in C_l} d(x,y)}{|C_k||C_l|}$$

Mat Lab

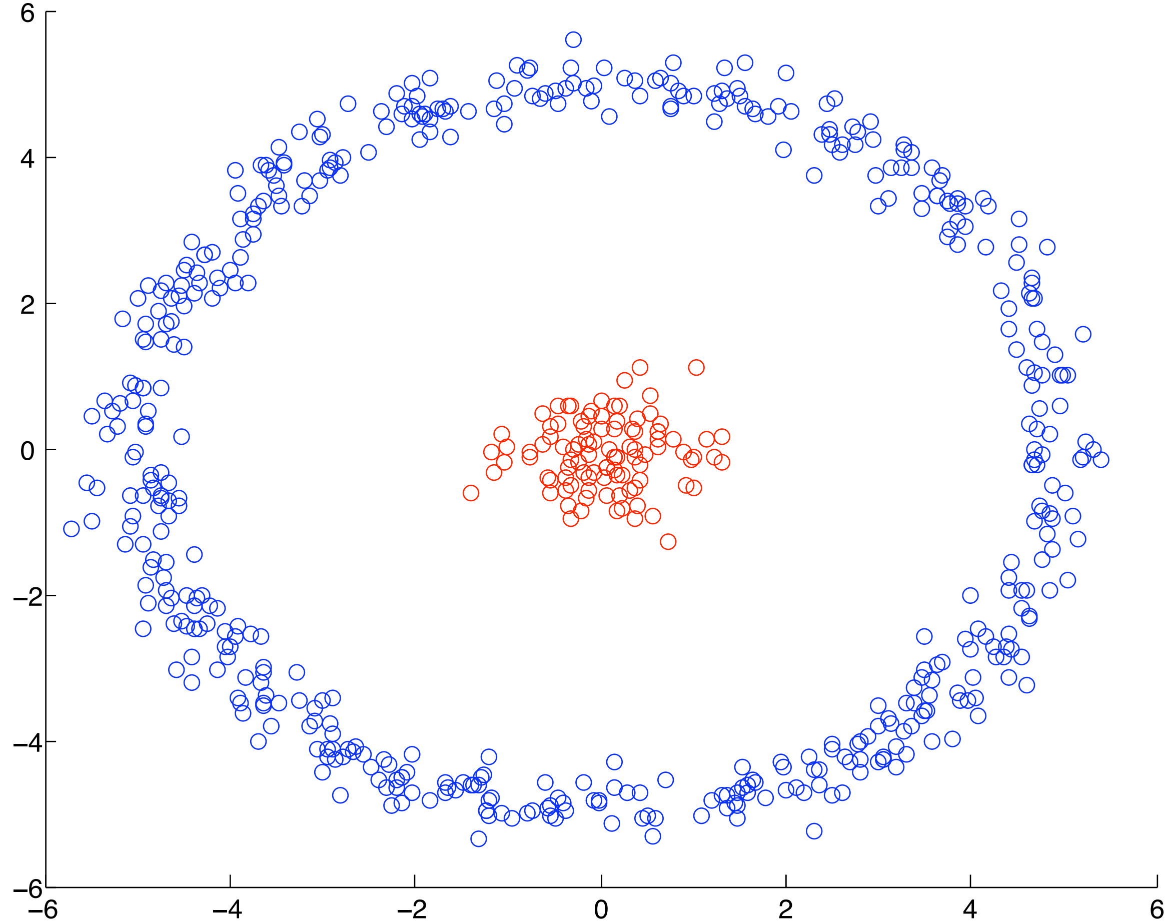
```
Dist=sin(acos(Cor)/2);

Z= linkage(squareform(Dist),'average');

[H,T] = dendrogram(Z,20);

Clusters=cluster(Z,20);
```

Text

The choice of "closeness" is given by the average distance, called *average-linkage clustering*.

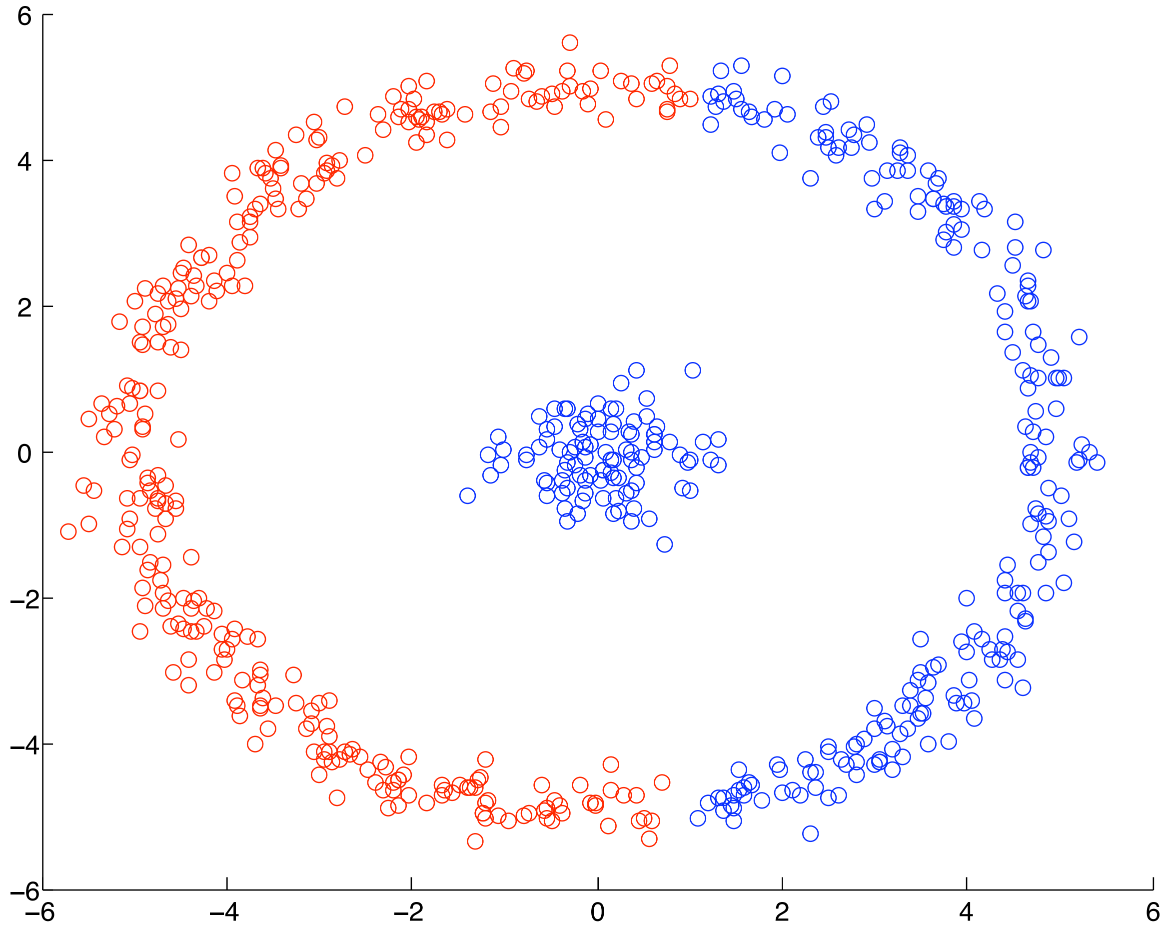
SIngle-linkage clustering uses the minimal distance....

Single-Linkage Clustering



$$d_{Single}(C_k, C_l) = \min_{x \in C_k y \in C_l} d(x, y)$$

Complete-Linkage Clustering



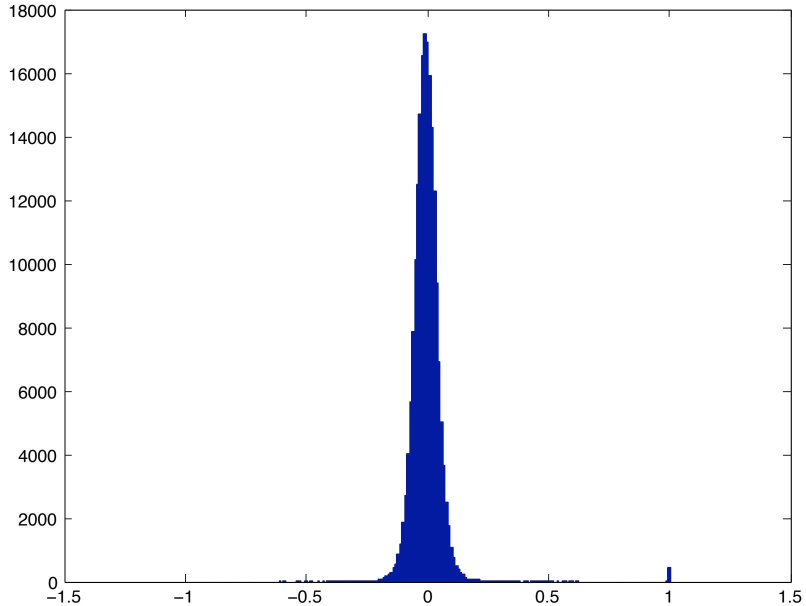
$$d_{Complete}(C_k, C_l) = \max_{x \in C_k y \in C_l} d(x, y)$$

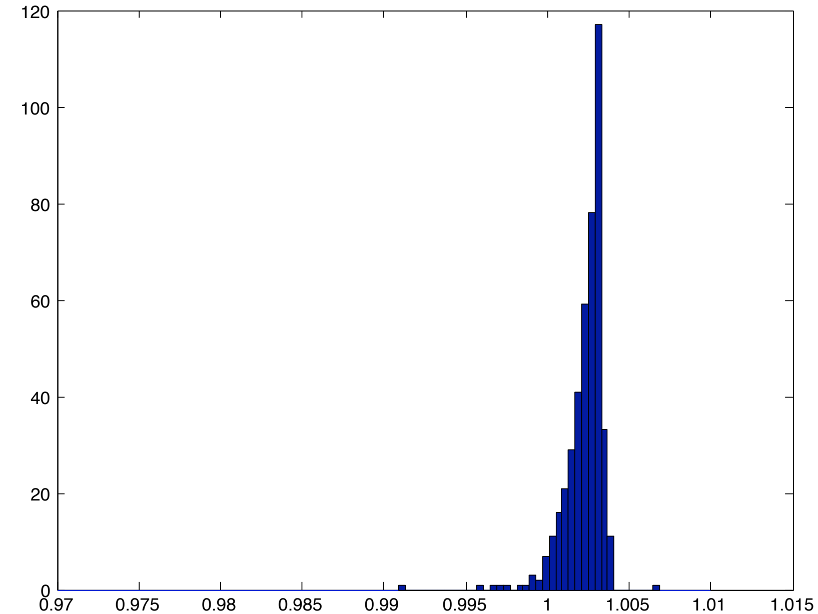
Combinations and Validation

- Validate clusters
- Use distinct clustering tools and explore if and how they disagree
- Combining Clustering: K-means for initial, then hierarchical methods to clean them up. This will work on very large data sets.

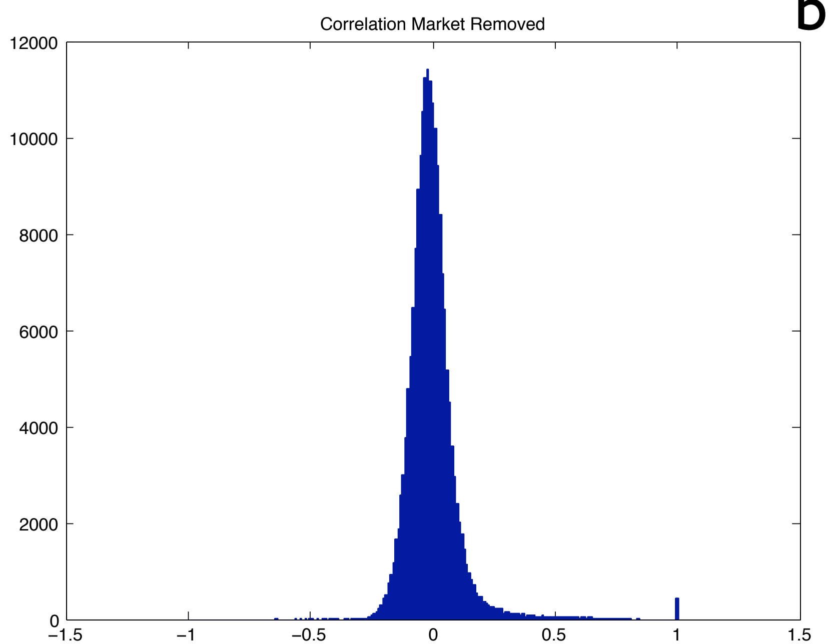
When we remove the clusters....

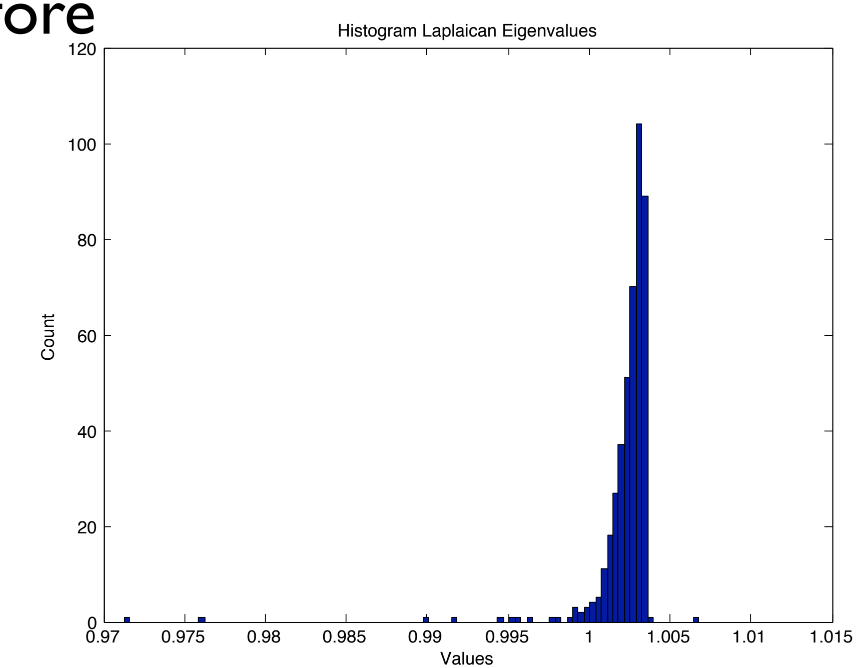
after





before





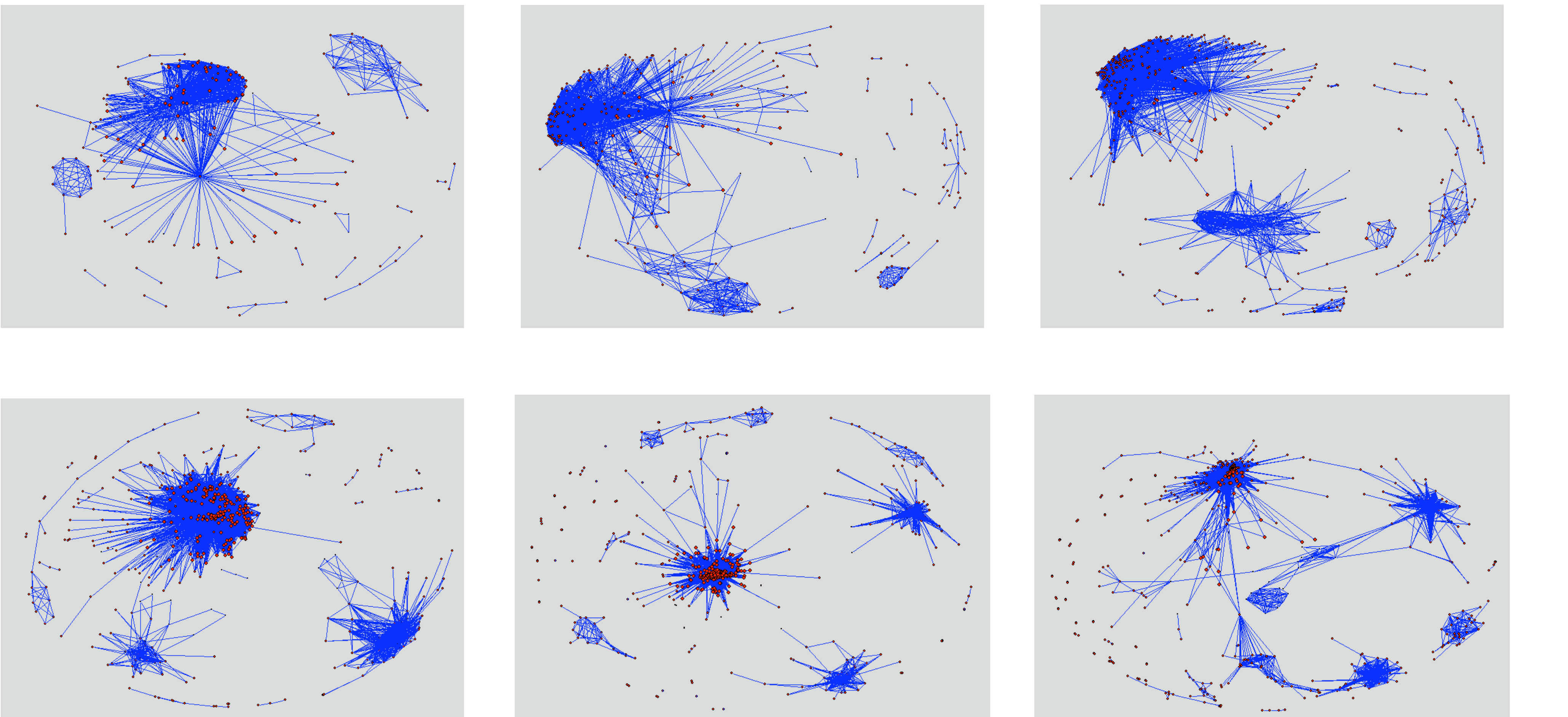
Fun Project!

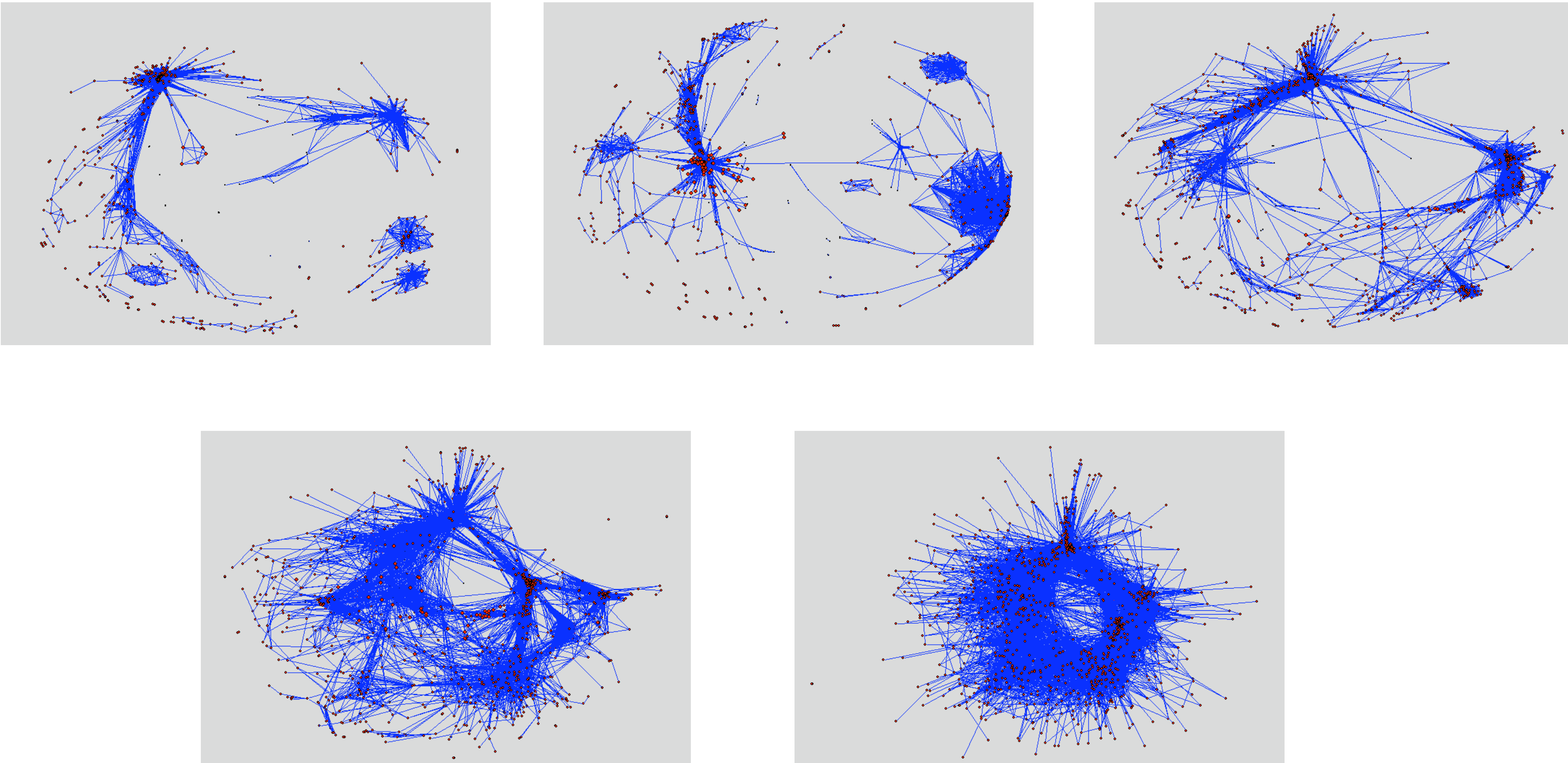
- Explain the next scale's "partition".
- Devise ways to study discretely over lapping clusters. (Standard ways are Gaussian mixture models and fuzzy clustering type techniques).

Other Methods to study the network

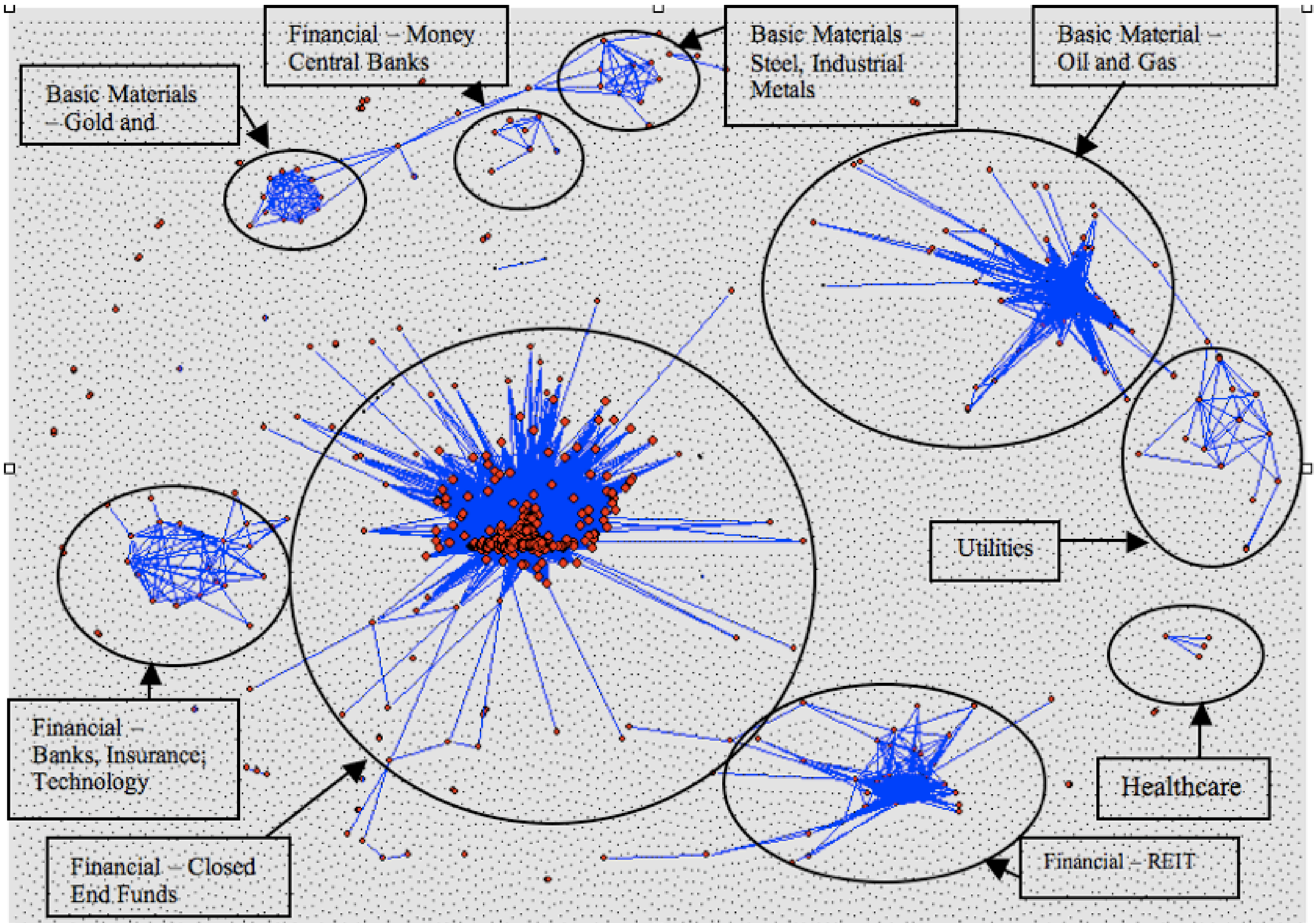
- Fix a correlation level, retain edges with at least this much correlation, and study the resulting network.
- Small World: Clustering Coefficient and mean shortest path length.
- Degree distribution and power laws.

NYSE: (0.7:0.5:0.2)





NYSE



Financial – Money Central Banks
Basic Materials – Steel, Industrial Metals
Basic Material – Oil and Gas
Basic Materials – Gold and
Utilities
Healthcare
Financial – Banks, Insurance, Technology
Financial – Closed End Funds
Financial – REIT

NASDAQ