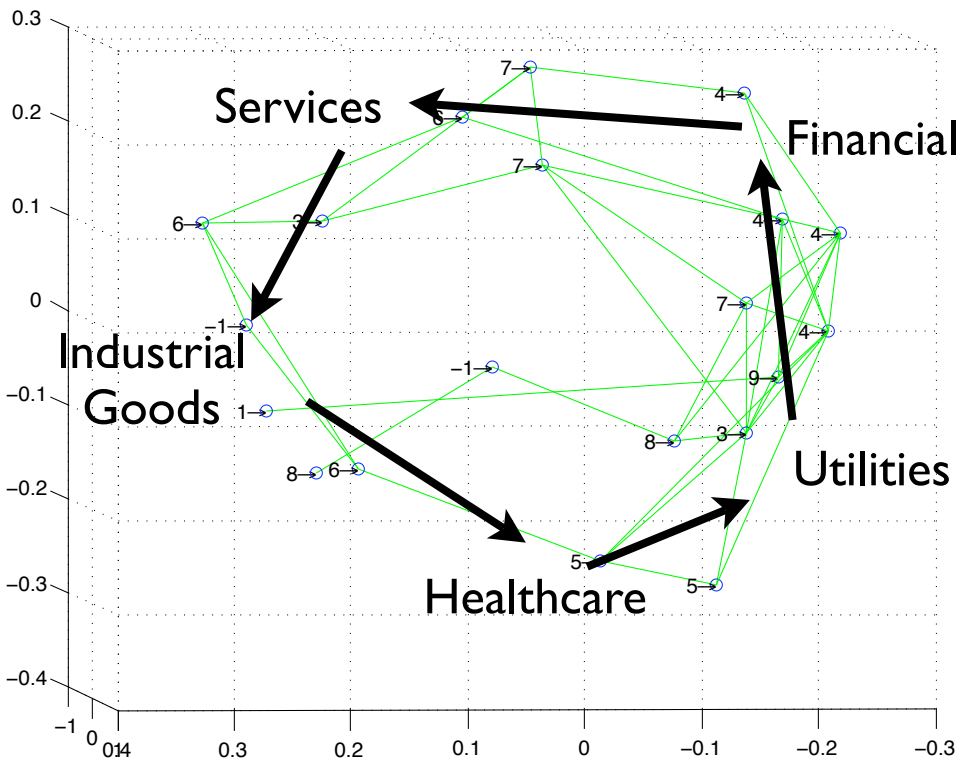
Unsupervised Learning in Complex Systems
Part A: Introduction
CSSS 2008



New York Stock Exchange Network

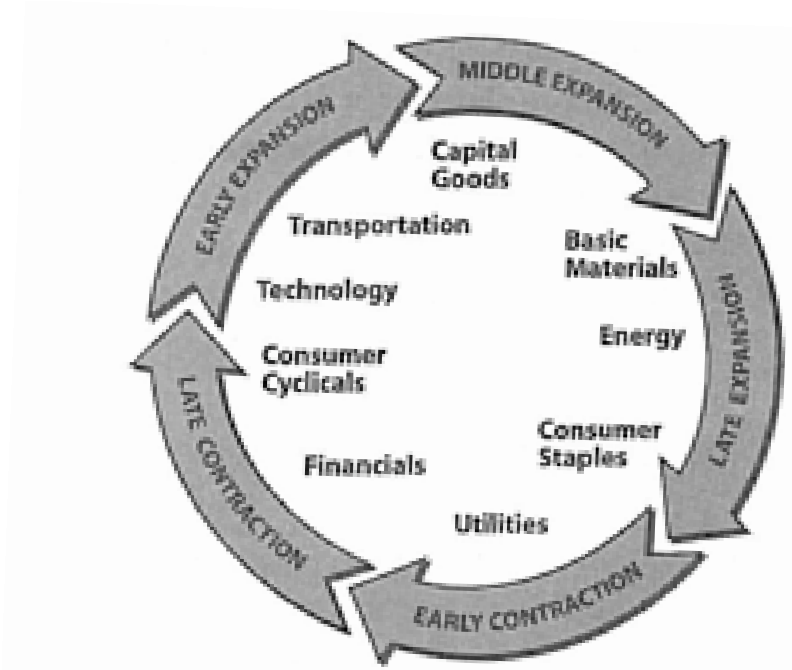


FIGURE 13.1 Technology and transportation leadership during 2003 fits Early Expansion phase.

Known cyclic structure

Gregory Leibon
Memento Security & Dartmouth College

Unsupervised learning in complex systems

Part A: Introduction (this lecture)
Part B: Can you hear the shape of the market?
Part C: Ether dipsomania in complex systems
Part D: An introduction to the Vulcan economy

Goals:

- A basic introduction to pattern recognition
  - 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 of weeks!
  - In MATLAB: MDS, K-means,...
  - Others: Random Matrix Null Models, Spectral Clustering, Partition Decoupling Method, the Green's Embedding, dimension reduction techniques...
- Explore Concrete Examples

- U.S. Equities Market
- 109th Congress
- Wikipedia
- Collaborative Reconmendation Networks: Movie Rankings

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.

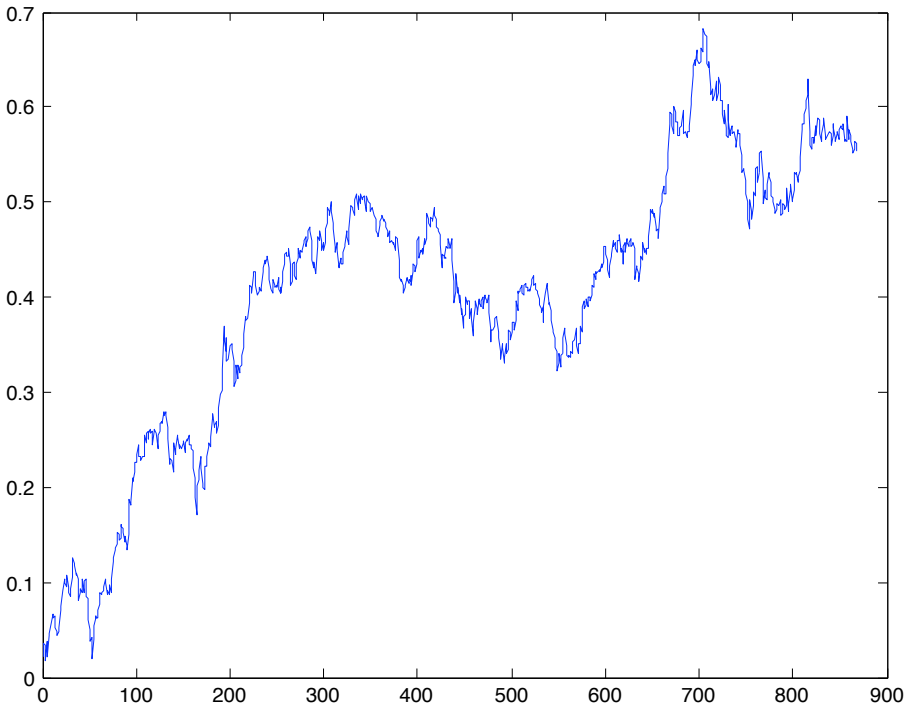
- Role of Pattern Recognition

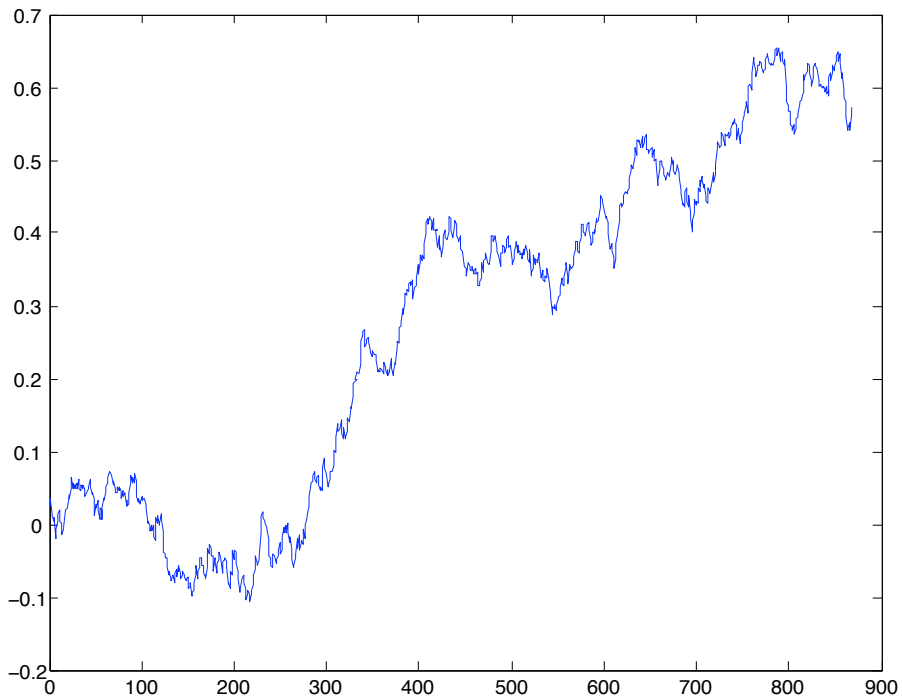
• is that in a realistic model one must understand how to articulate the local behavior as well understand the emergent patterns that characterize the a complex system.

- 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 15 years.
  - Annotation (partial) collected was the equity's names, sector, industry, and index membership.

An Equity's Times Series

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



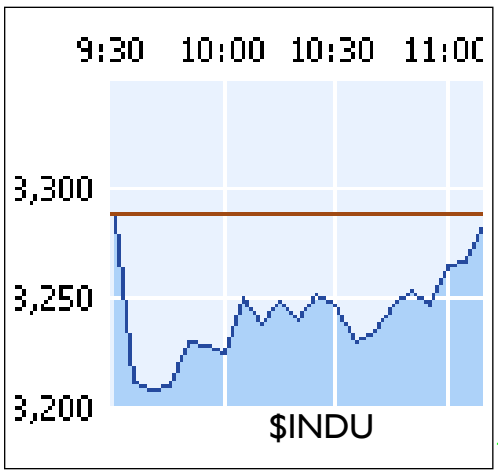


A well known rough approximation:

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

The Market as Complex System

Network



Information

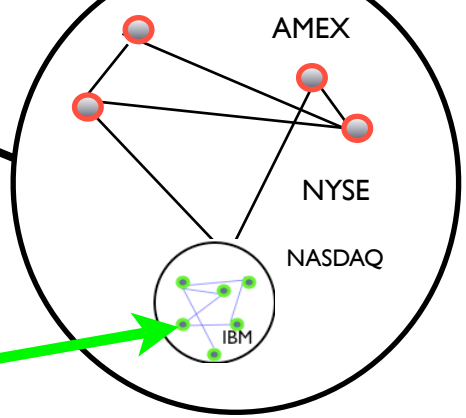
Bonds

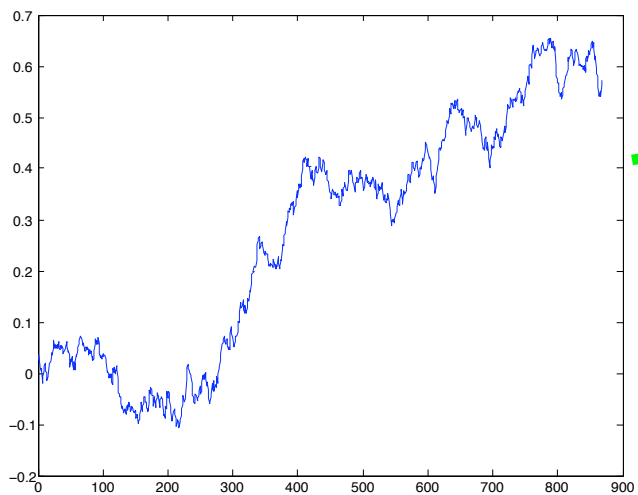
Currency

MBS

Equities

Commodities





Tradeable Atoms

Correlation Metric

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

Correlation: $$\rho(X, Y) = \hat{X} \cdot \hat{Y}$$

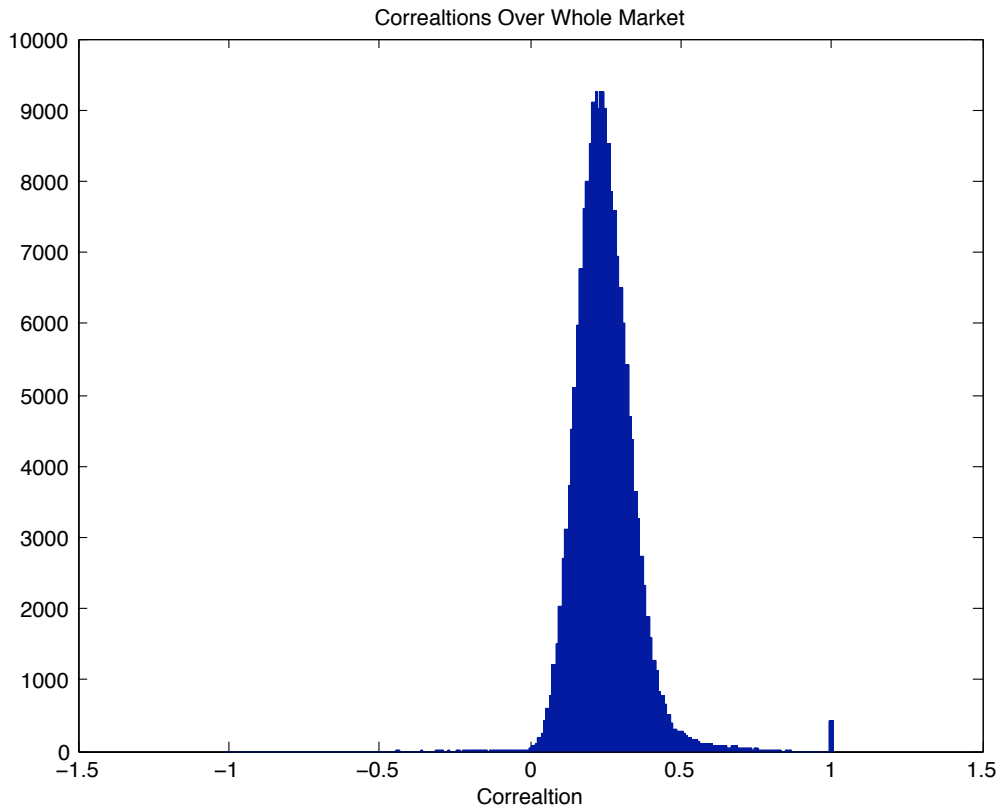
Metric: $$d(X, Y) = 2\sin(\theta/2) = \sqrt{2\,(1 - \rho(X, Y))}$$

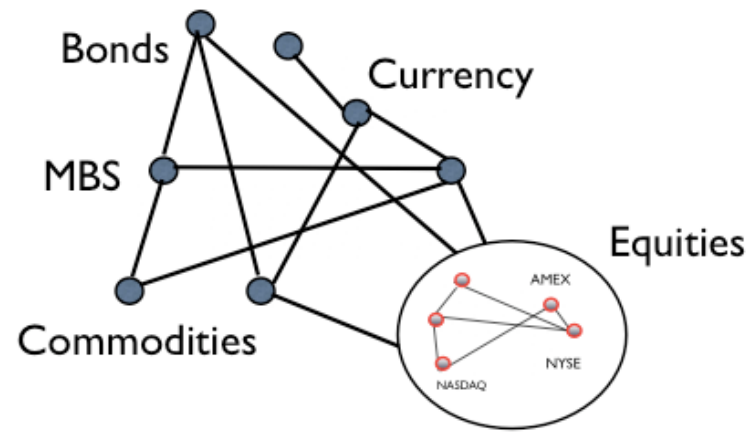
Examples: N=6,000

L=3,600 Daily Ticks

L=103,680,000 Second Ticks

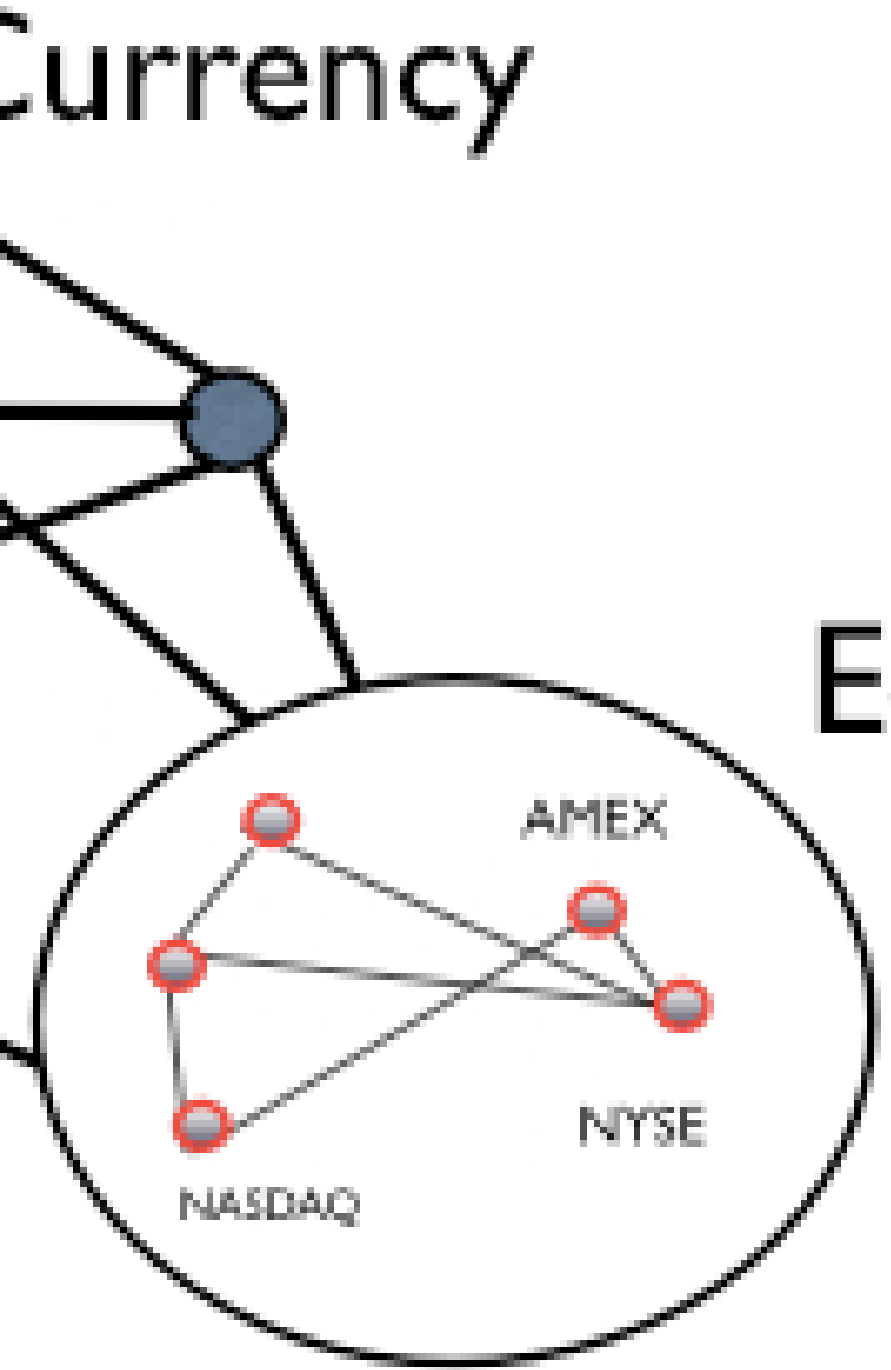
Correlation: S&P500



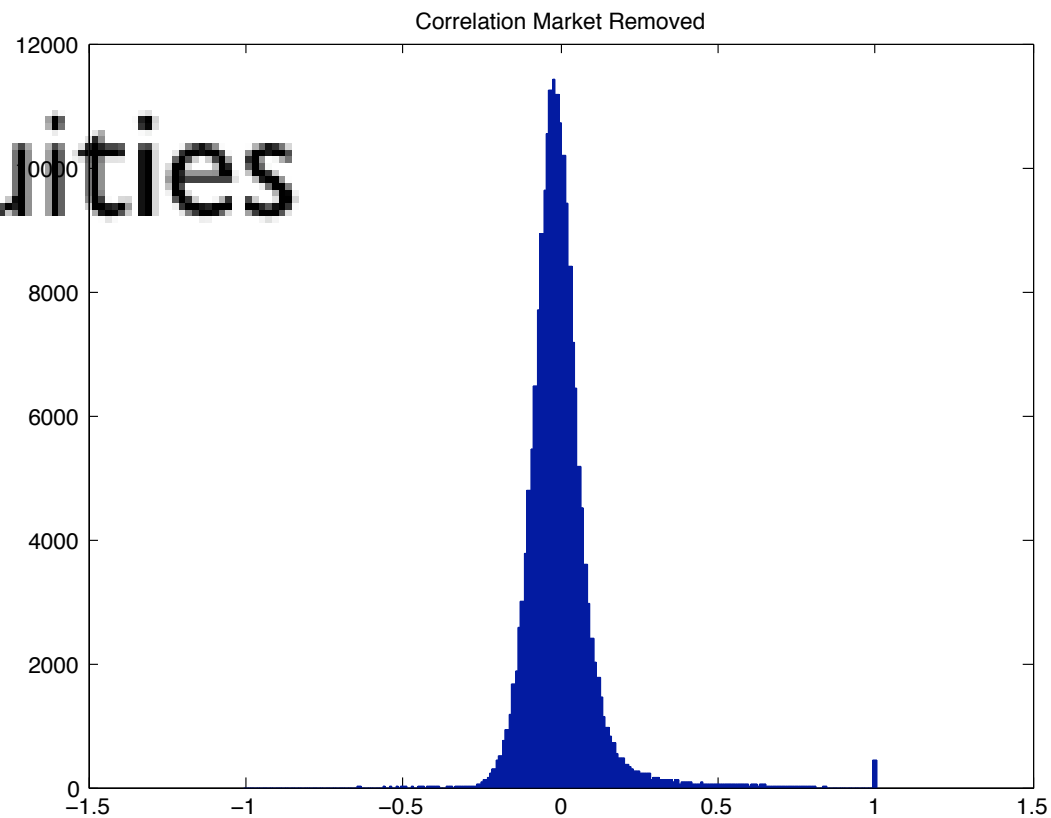


How much of this correlation is internal to our equities market and how much due to external forces from the whole economy?

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).

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....

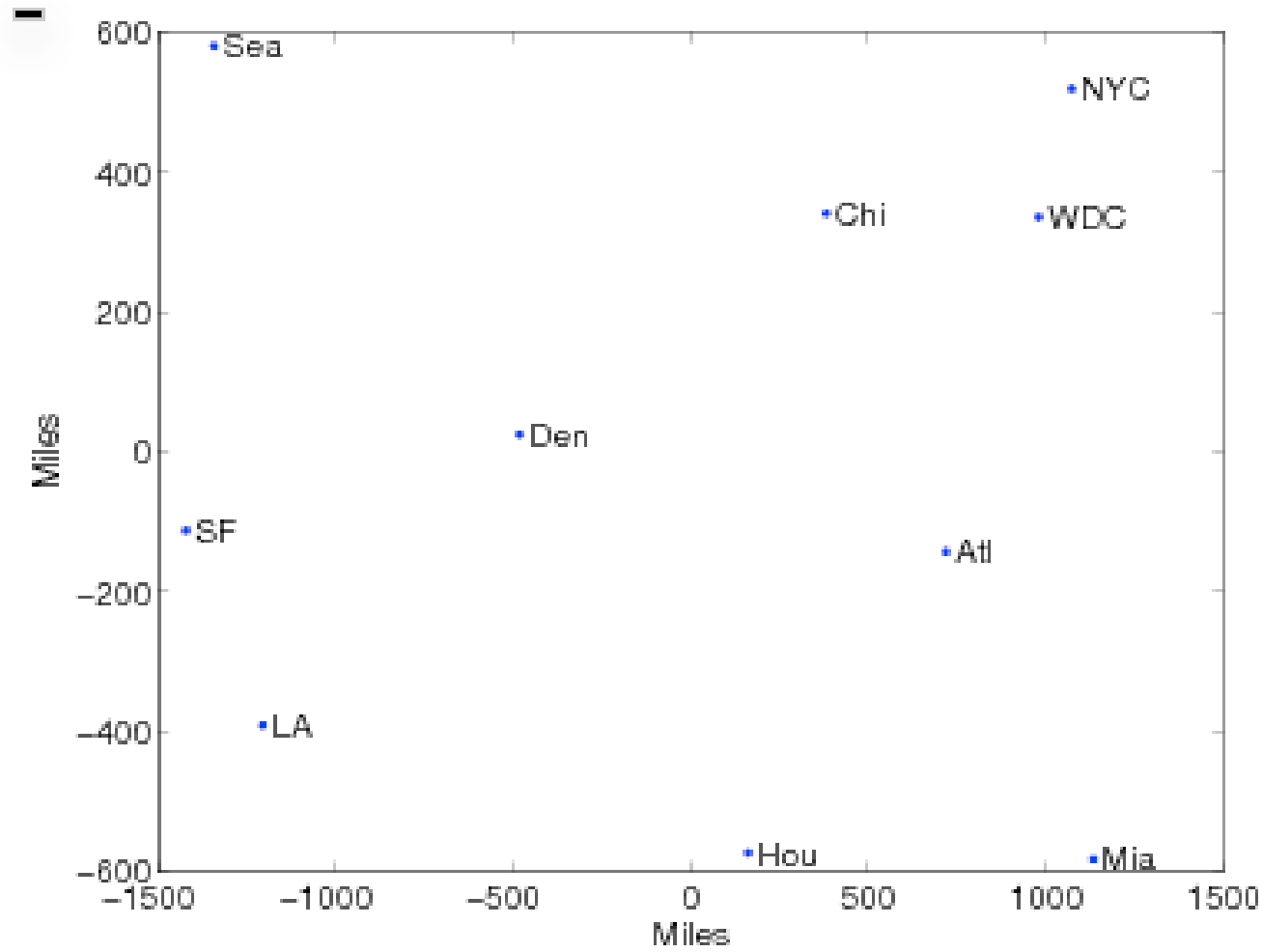
Himalayas

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

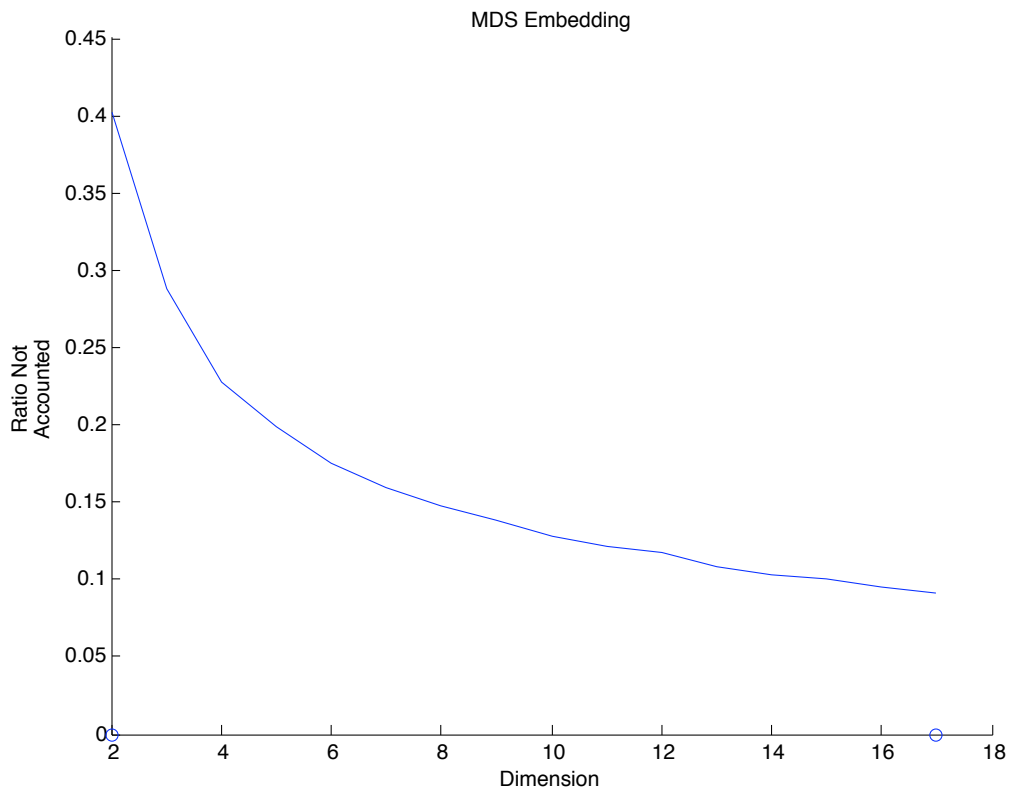
MDS 2-d


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))

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



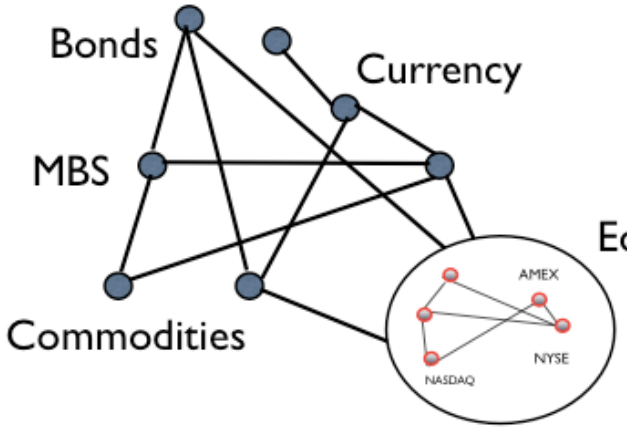
MATLAB code:

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

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

Dist=sqrt(2*(1-Cor));

opt=statset('MaxIter',5000);

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

% Here Dim15 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,...

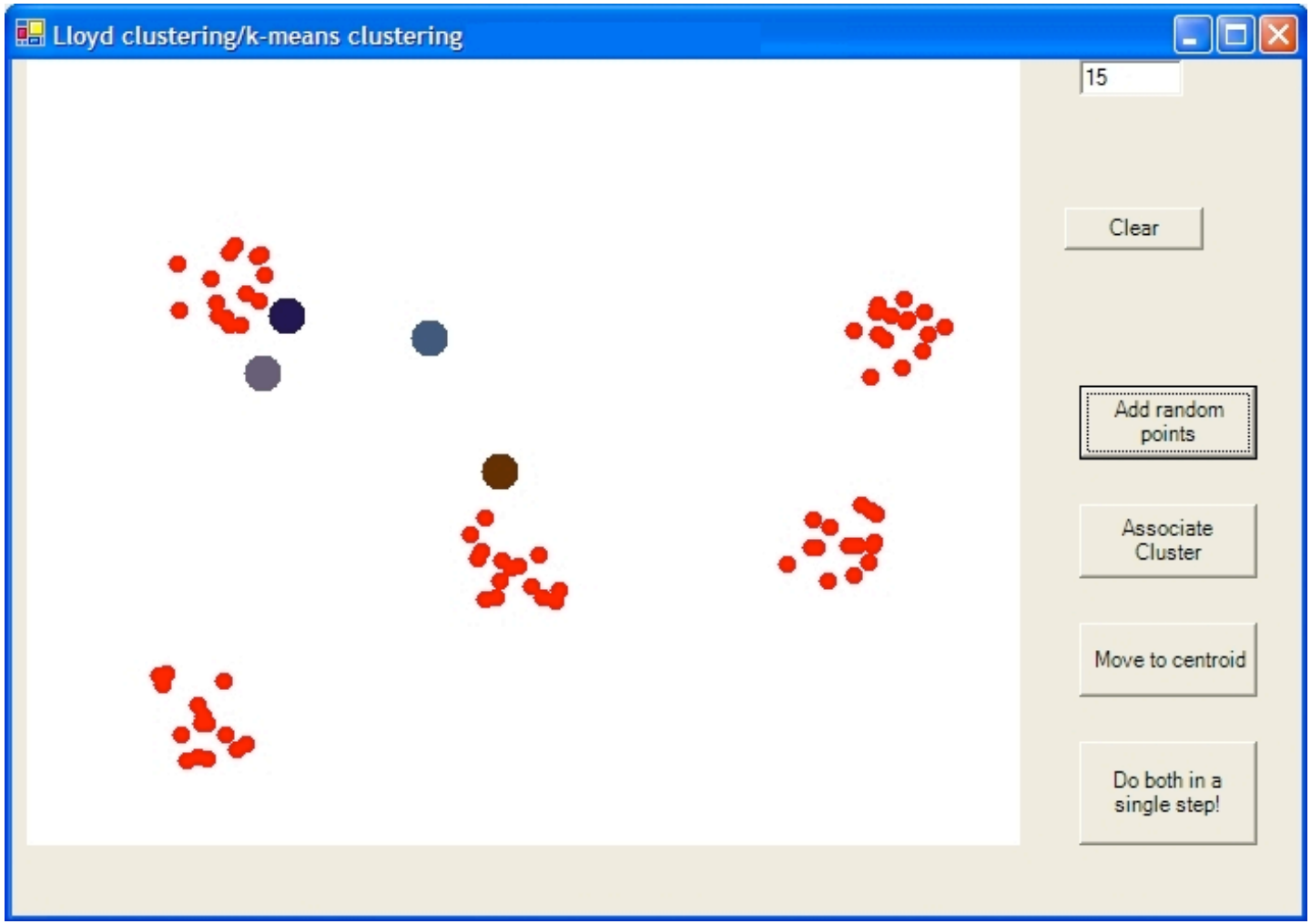
K-means

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

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

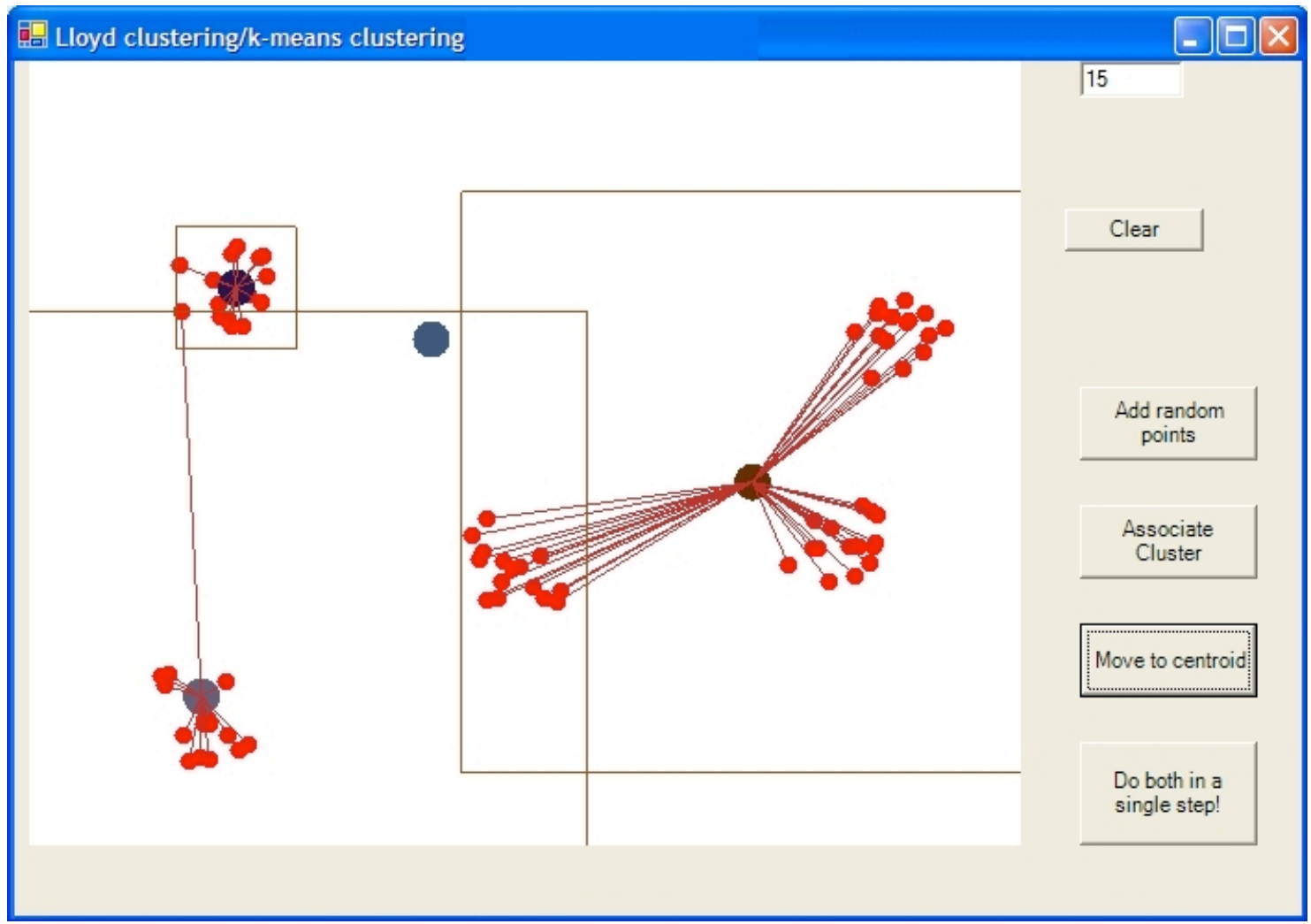
K-means algorithm:

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

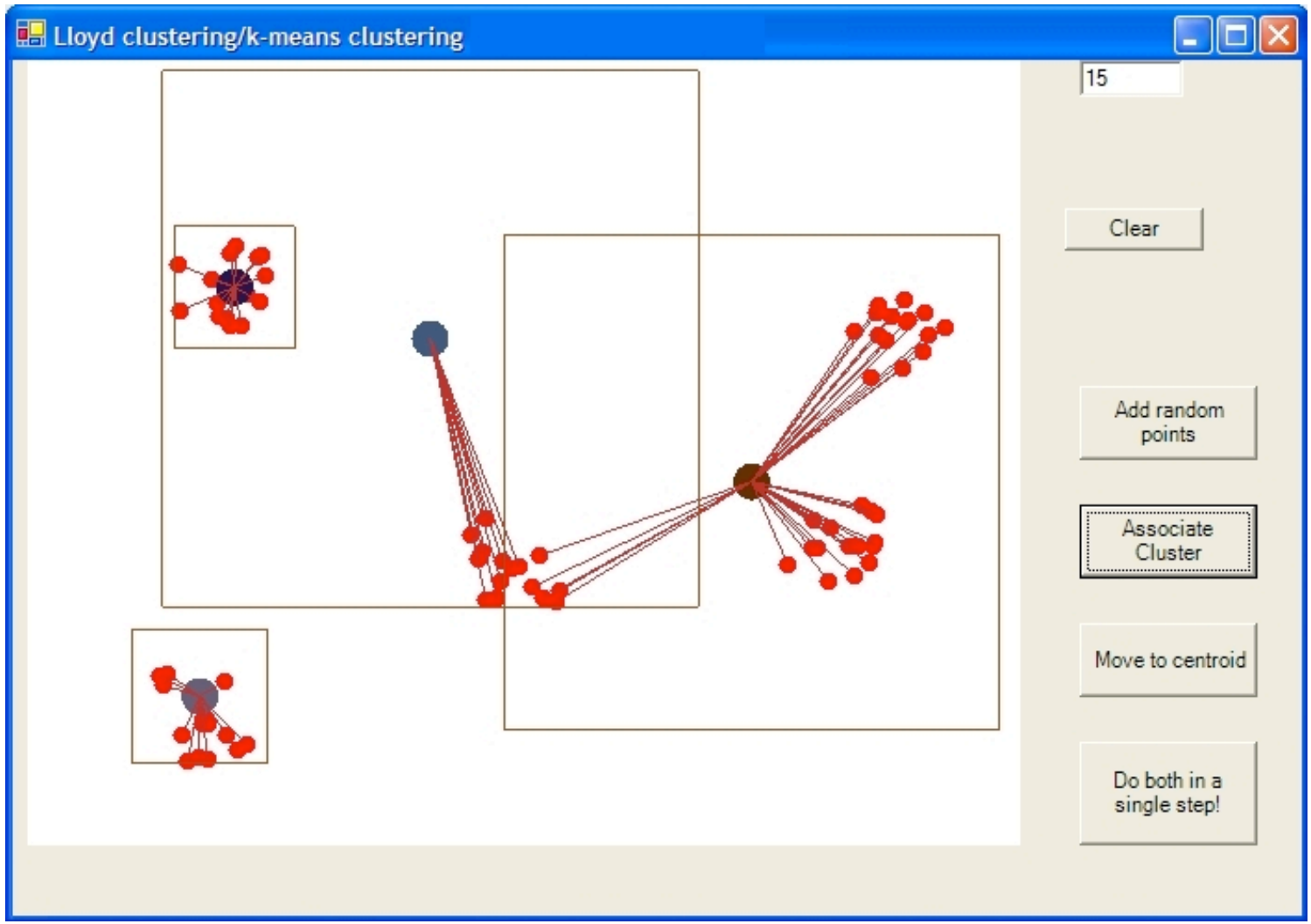


Example from: http://en.wikipedia.org/wiki/K-means_algorithm

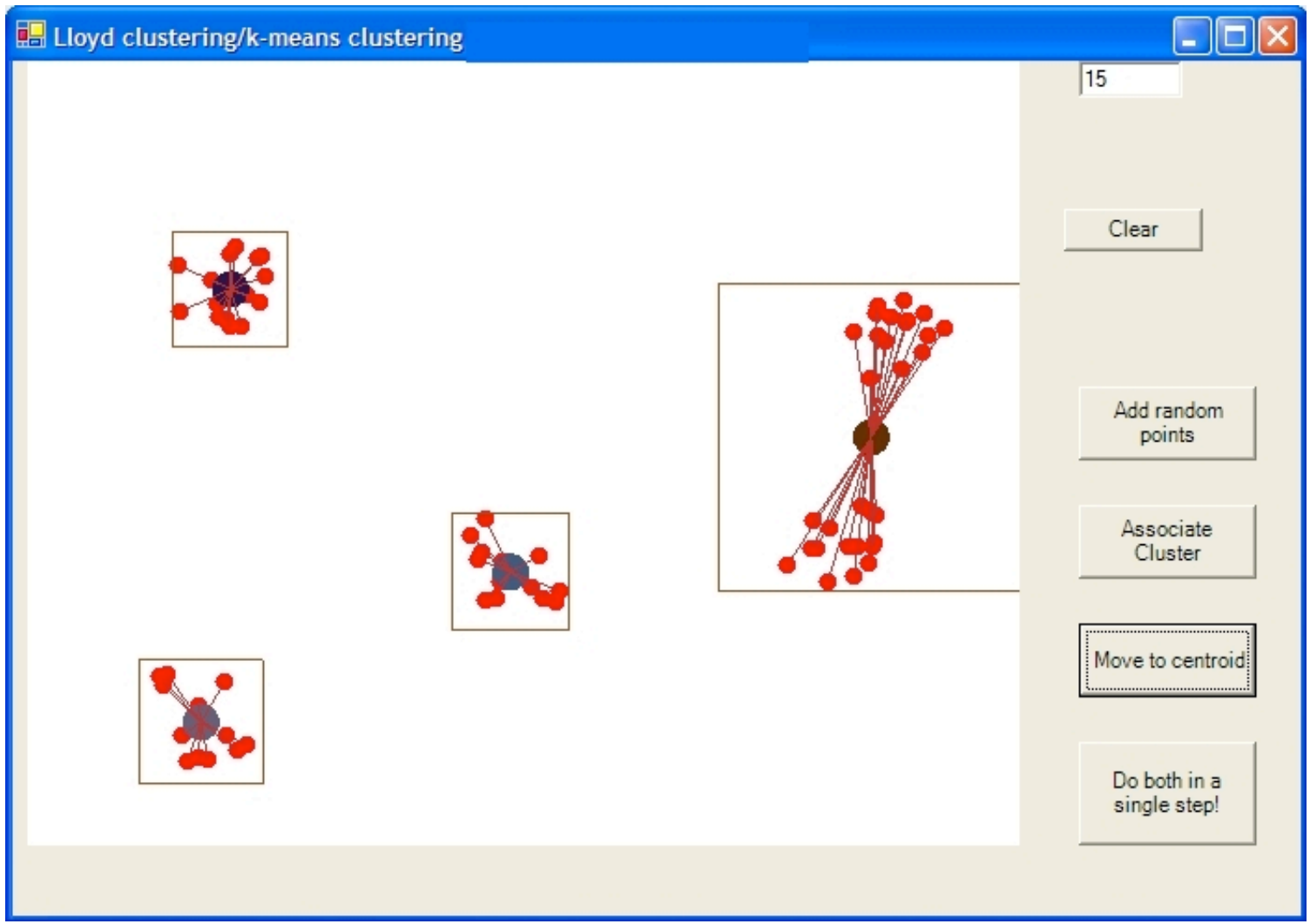
2. Organize points by distance to the centroids.
3. Update centroids



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.

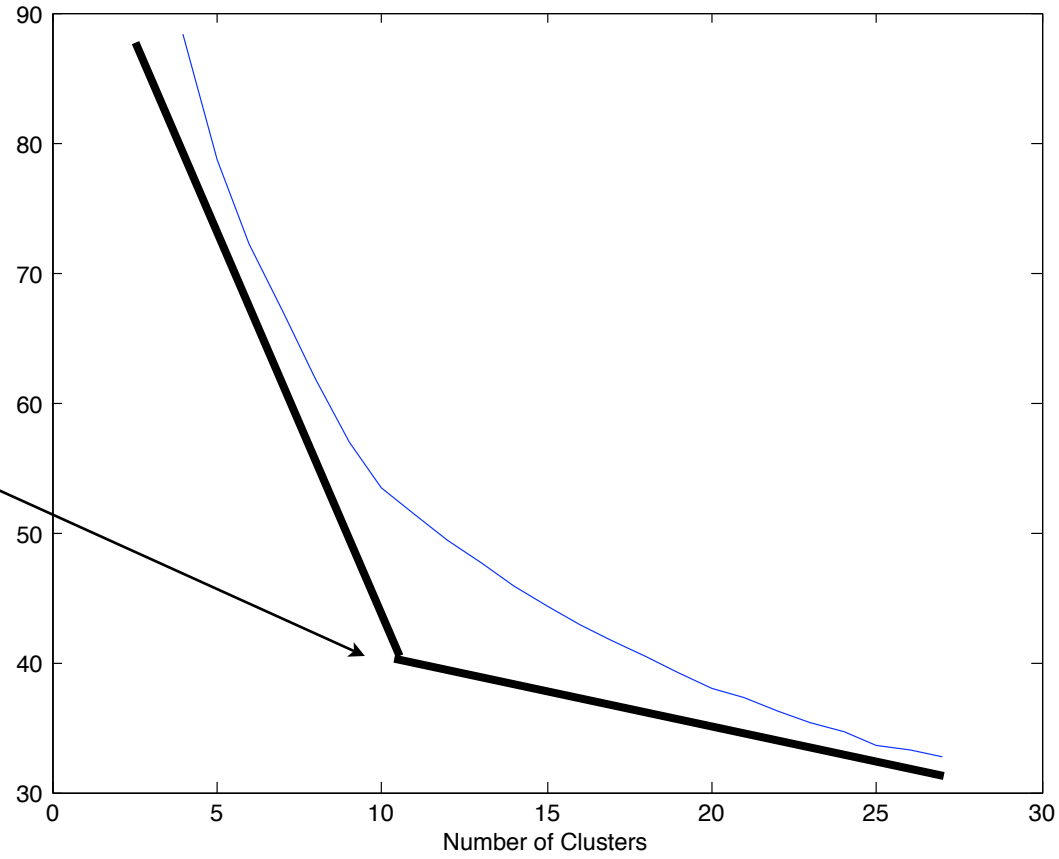
```
Clusters = kmeans(Dim15,20);
```

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 K=#(Clusters)

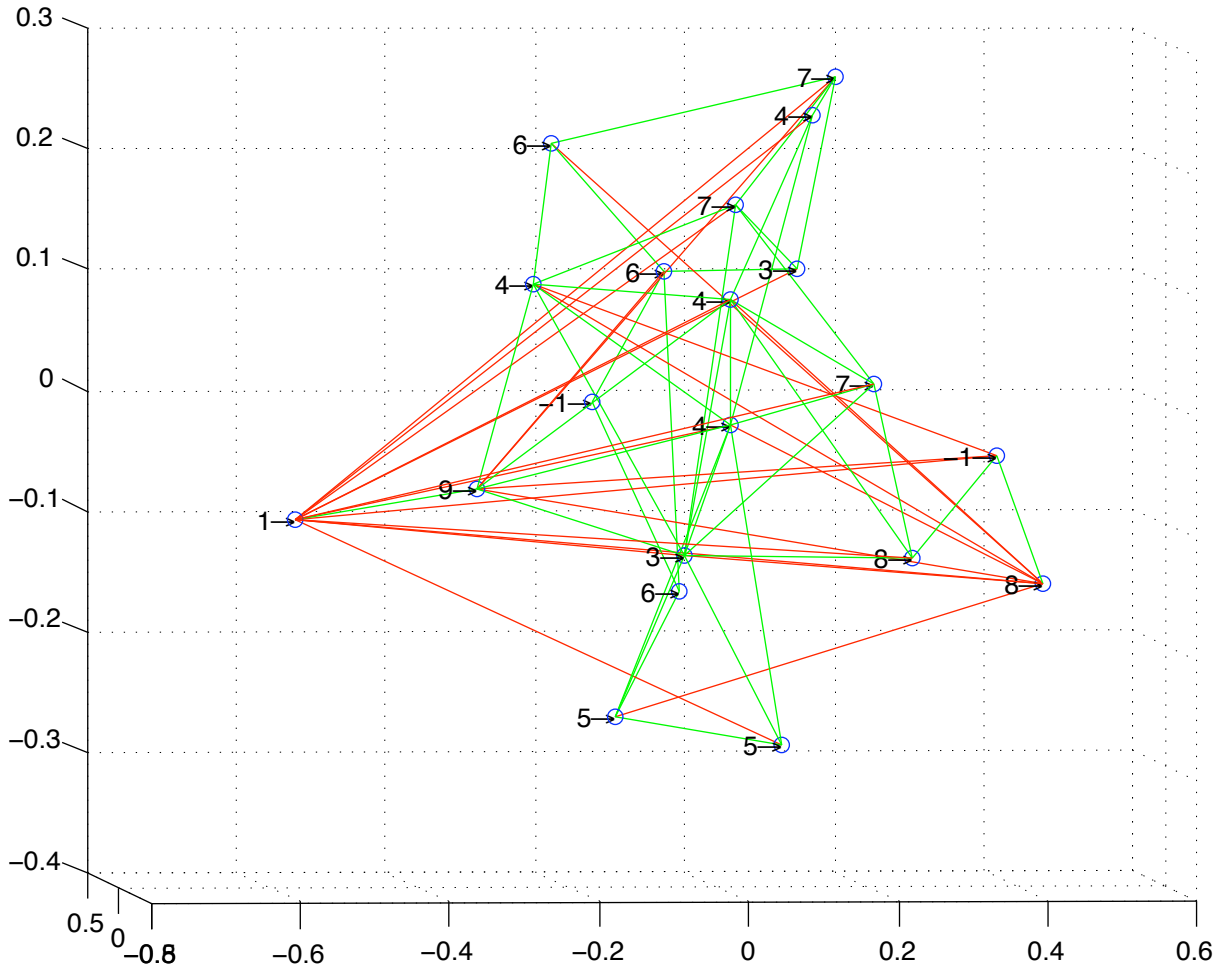
Elbowlogy:

elbow



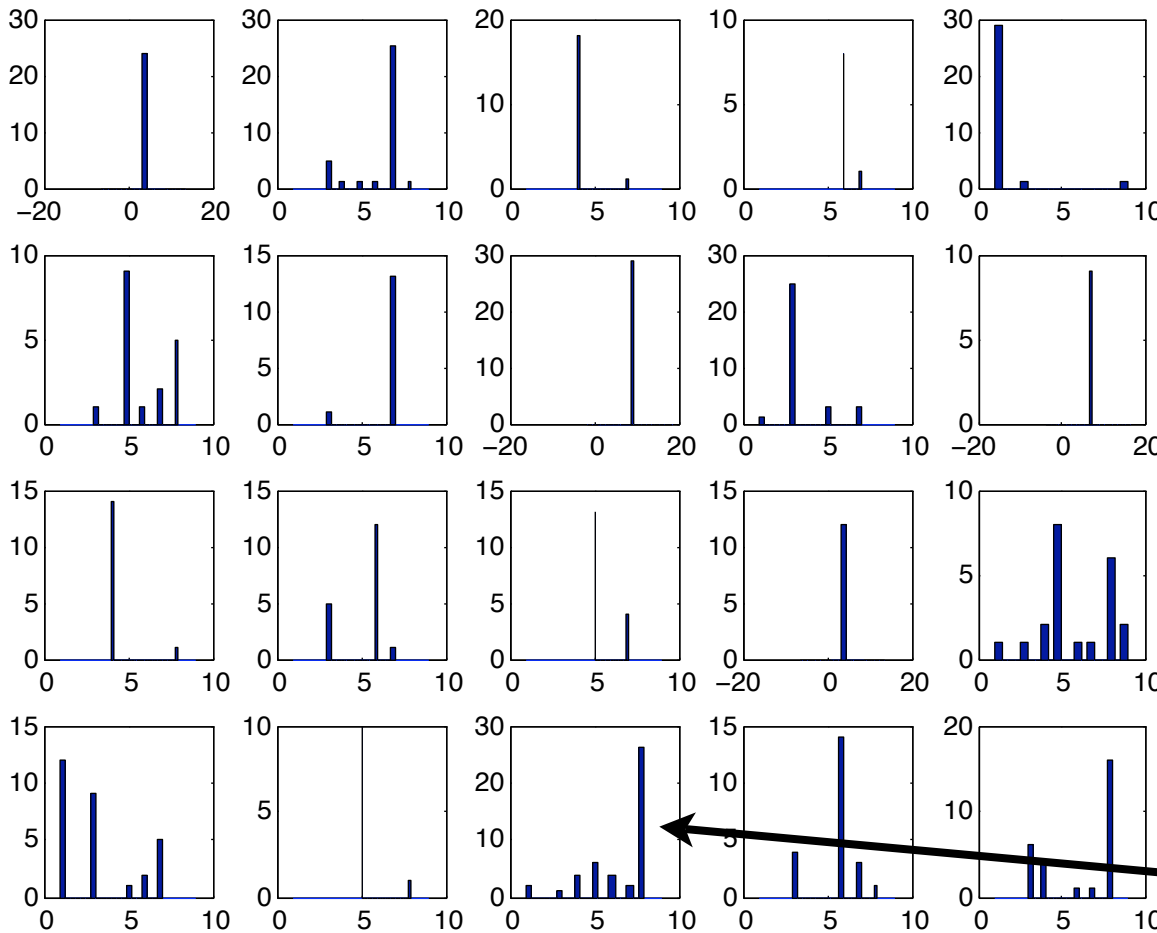
or...

#(Clusters)=20



Take a guess!

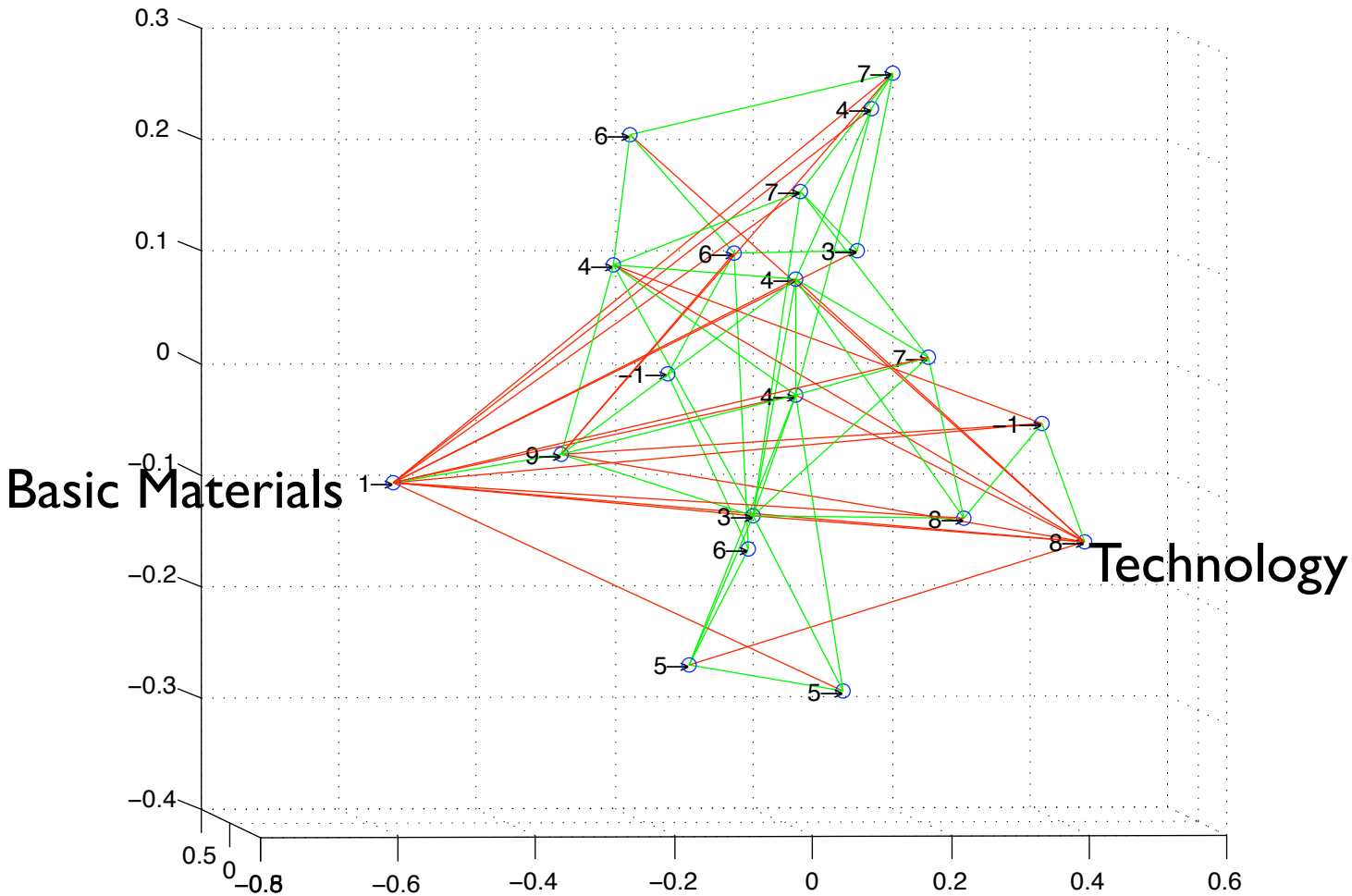
Sectors



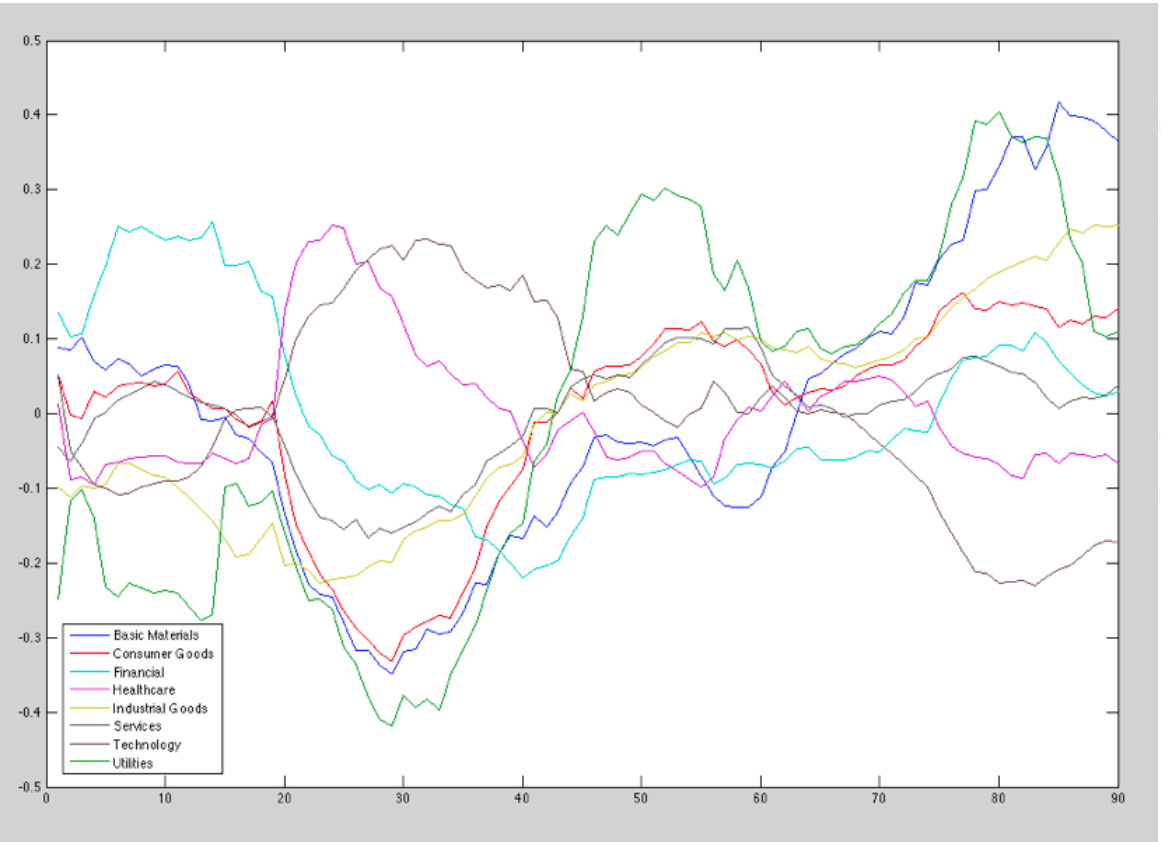
1 Basic Materials
2 Conglomerates
3 Consumer Goods
4 Financial
5 Healthcare
6 Industrial Goods
7 Services
8 Technology
9 Utilities

Guess?

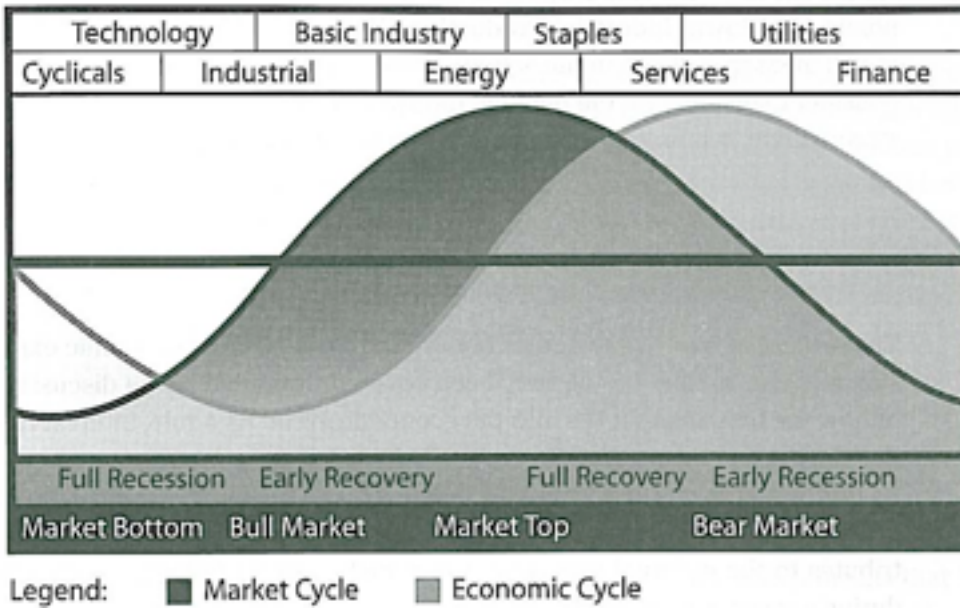
The Recent "Battle"



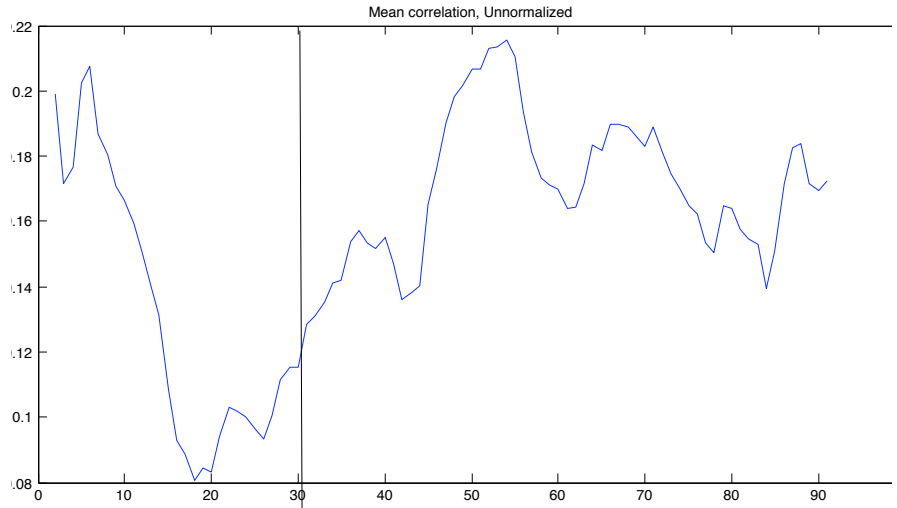
More importantly, generative models will allow us to deal with structured time series.



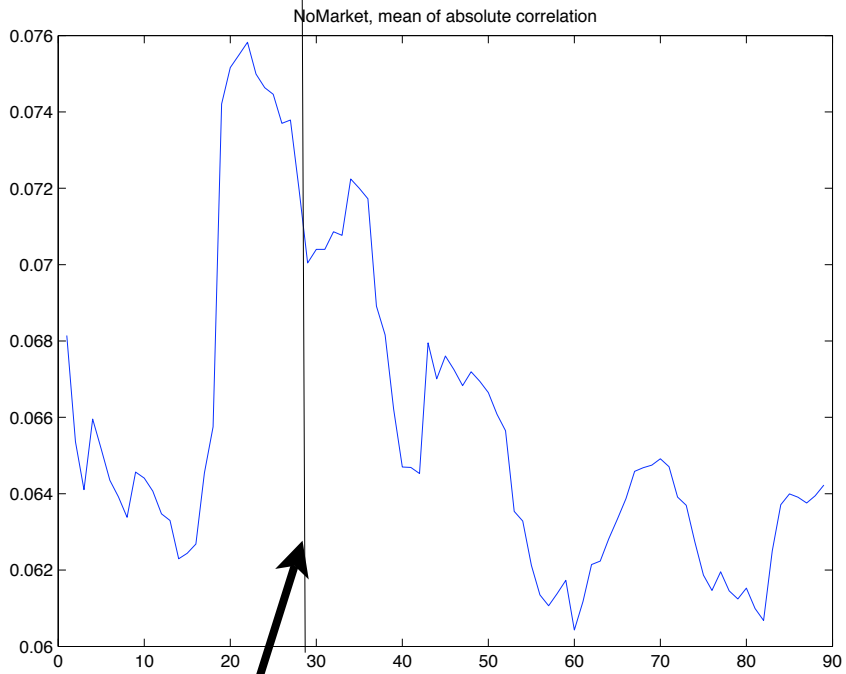
Correlation by Sector



Market Wisdom



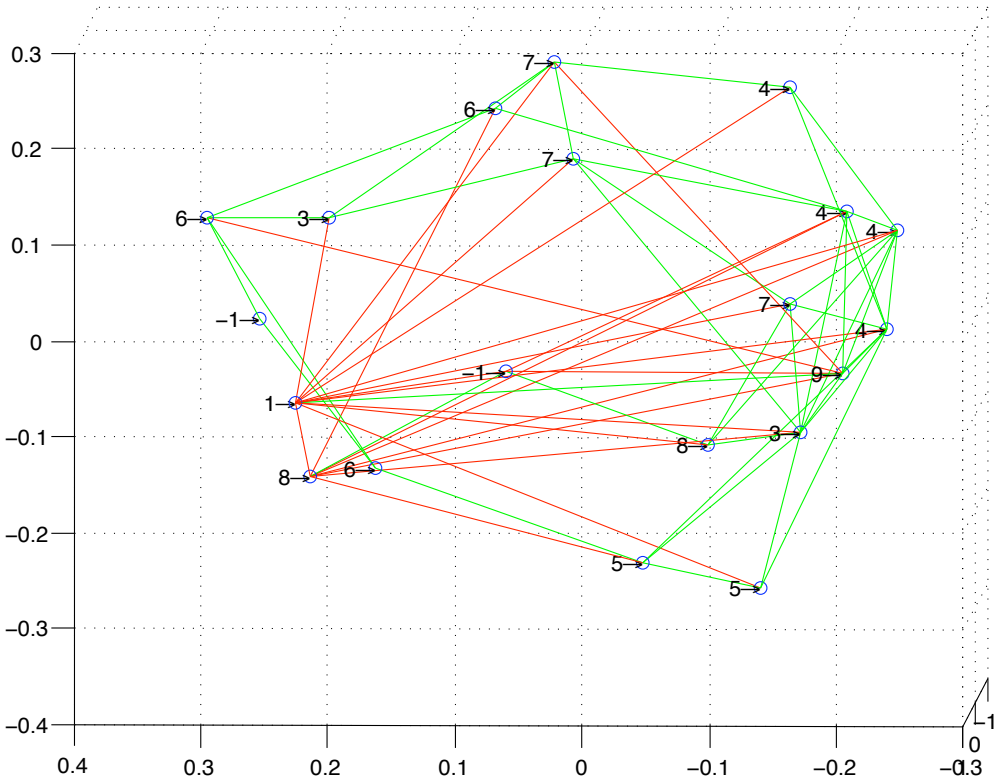
Full correlation pressure

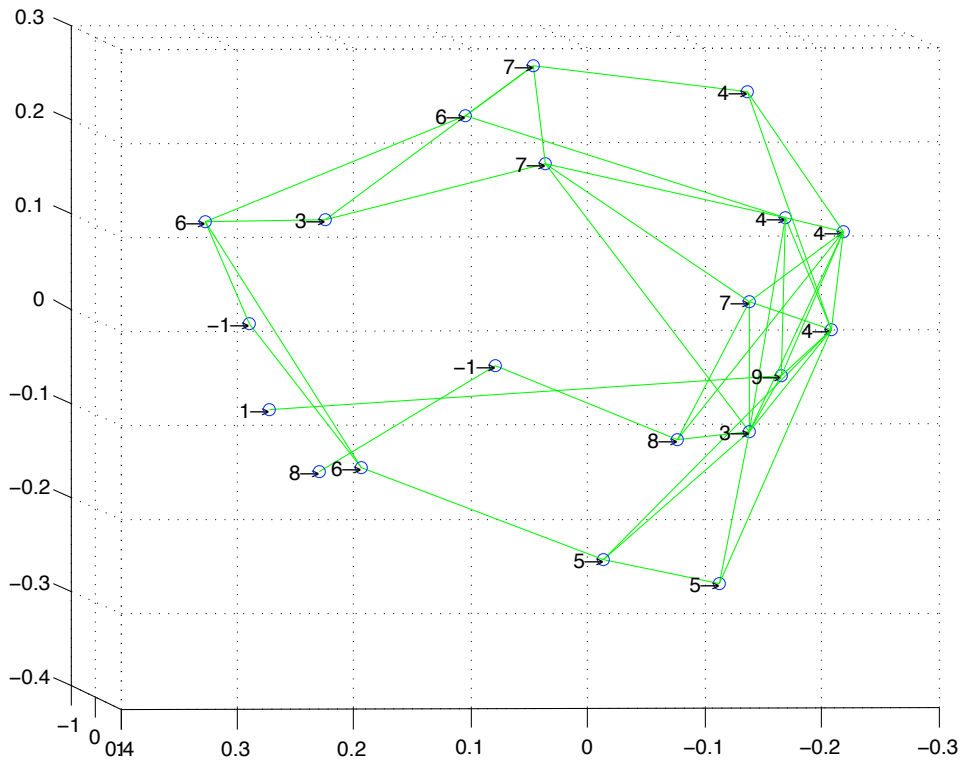


With the equities market isolated

NASDAQ Tech Stock Crash

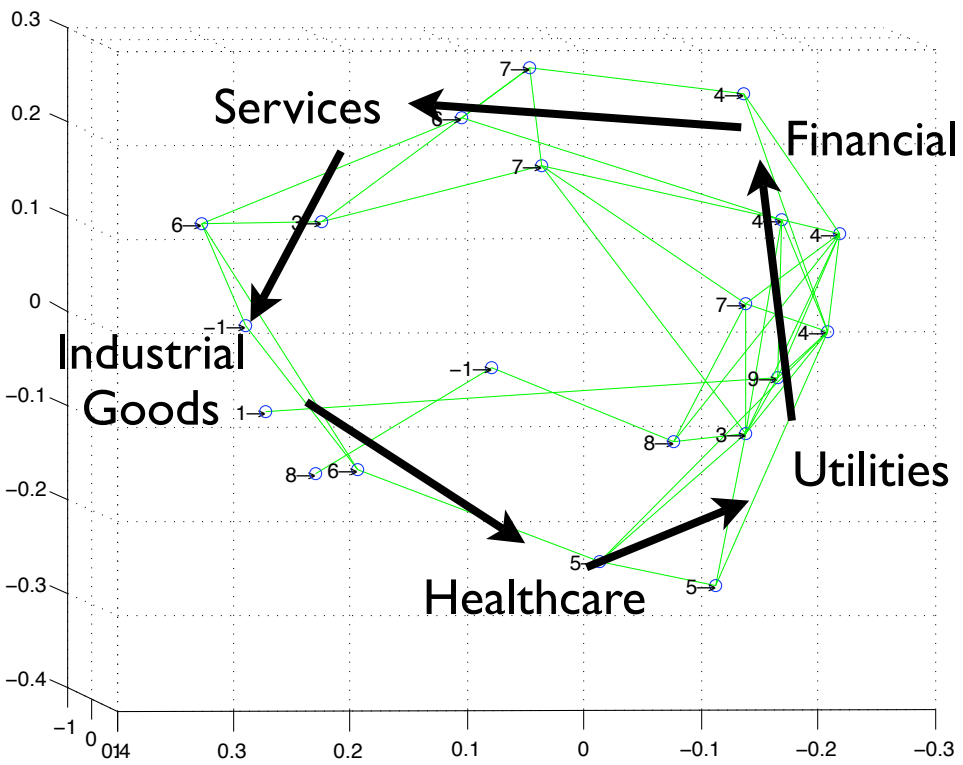
Even the static picture is interesting...

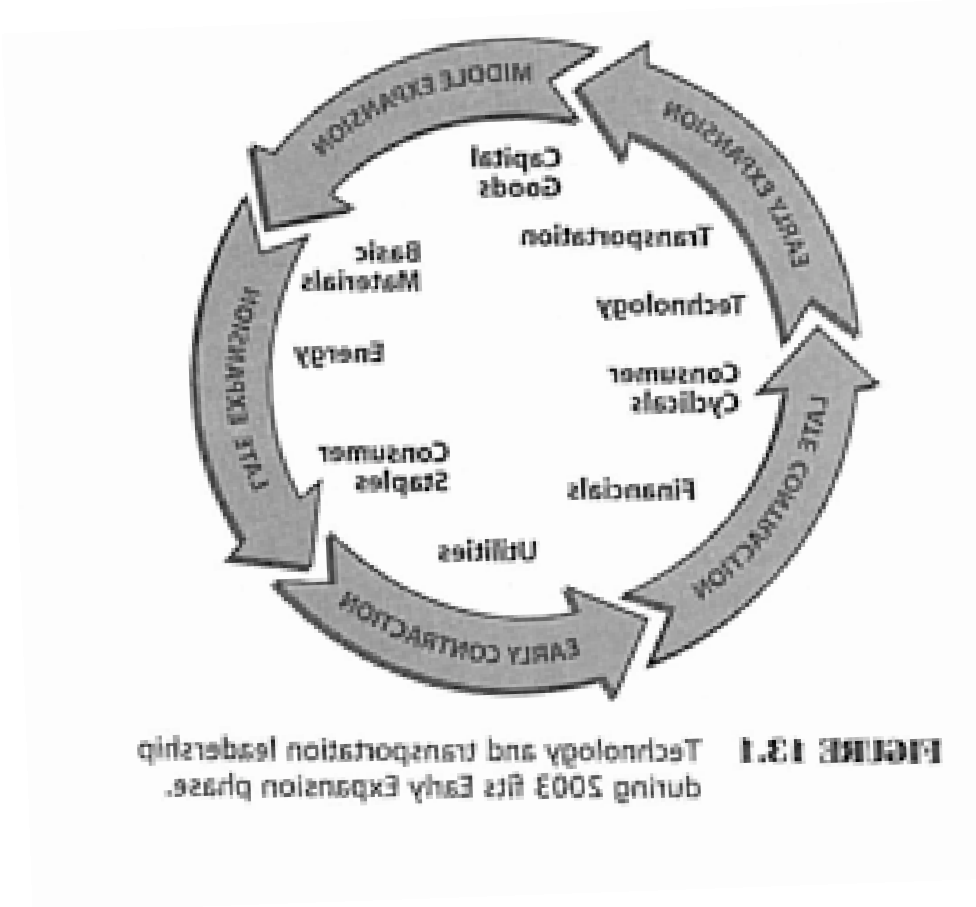




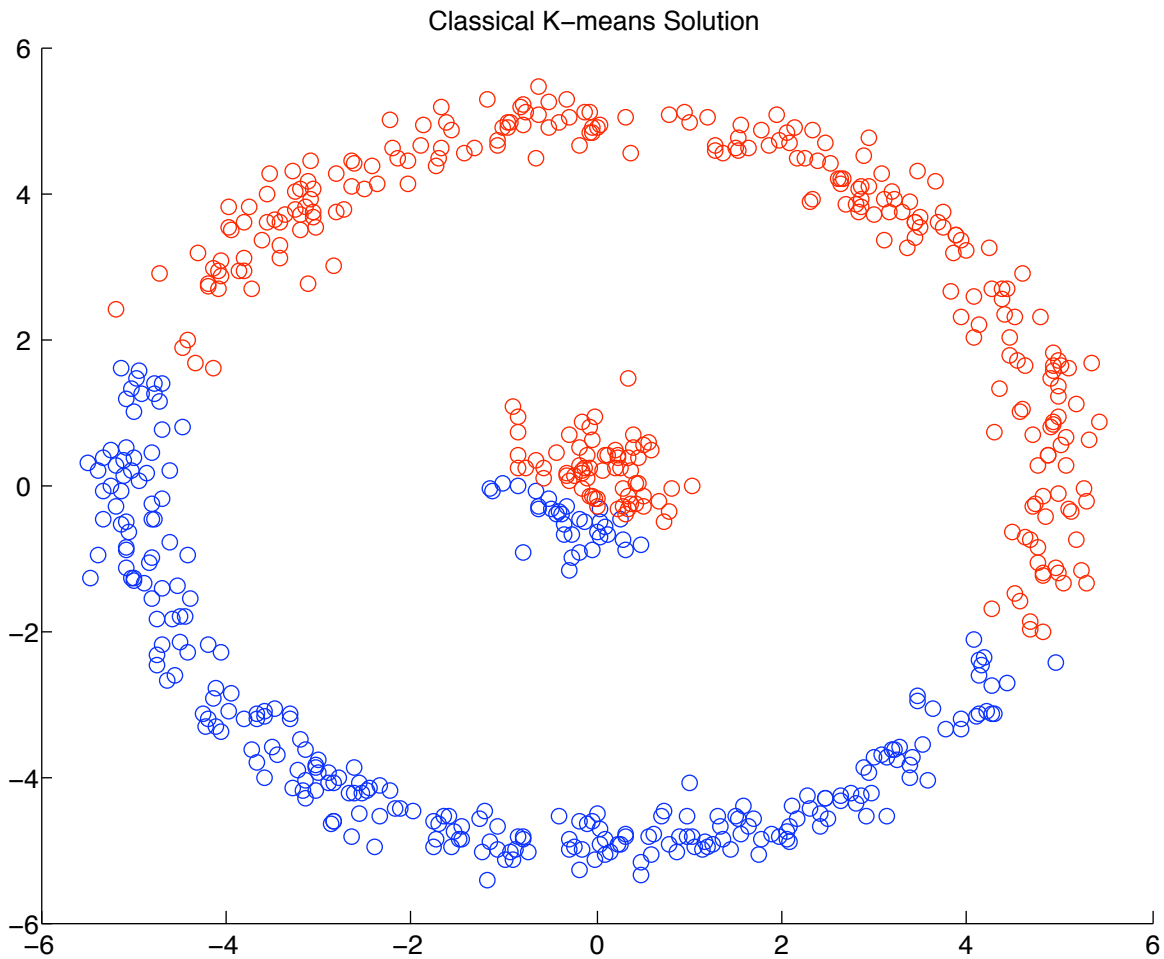
Notice the circle: A manifold, not a "ball" cluster.

The Cycle


Services
Financial
Industrial Goods
Utilities
Healthcare


Capital Goods
Transportation
Technology
Consumer Cyclicals
Financials
Utilities
Consumer Staples
Energy
Basic Materials
EARLY EXPANSION
LATE EXPANSION
EARLY CONTRACTION
LATE CONTRACTION

Classical K-means



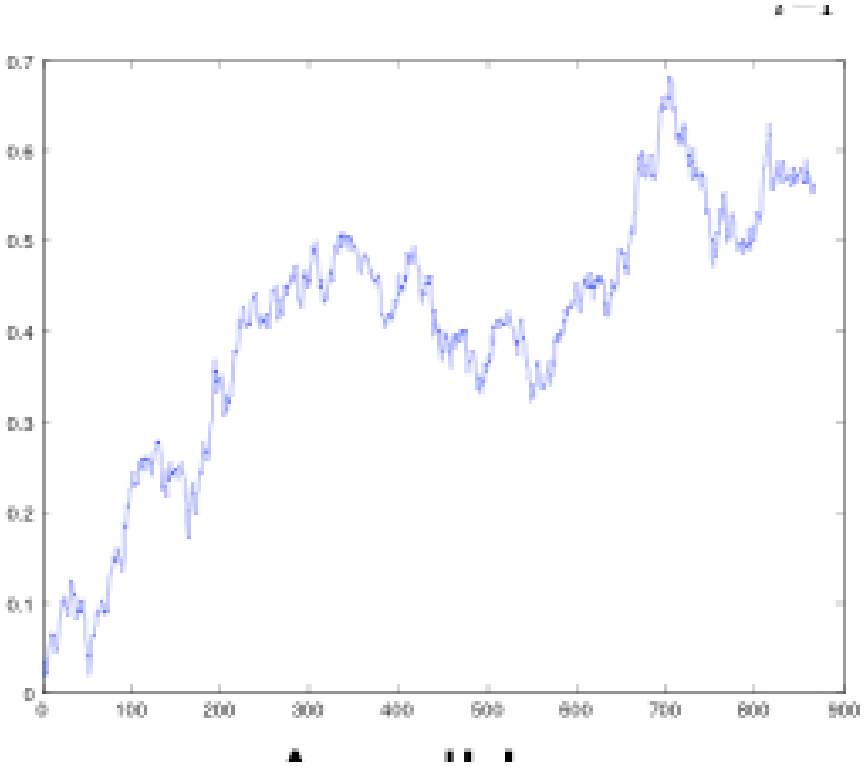
Key Questions

How much information has been captured?

Requires developing Models and statisitics to compare them....

Random model: pretend our stocks are all independent and random

(better models described in Lecture 2, Part B!)



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

Compare to actual model via certain statisitics. In this case UNIVERSAL properties related to Eigenvlaues....

(Leture 1,Part B we will see why this is a good coice!)

Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series

Vasiliki Plerou,[1,2] Parameswaran Gopikrishnan,[1] Bernd Rosenow,[3] Luís A. Nunes Amaral,[1] and H. Eugene Stanley[1]

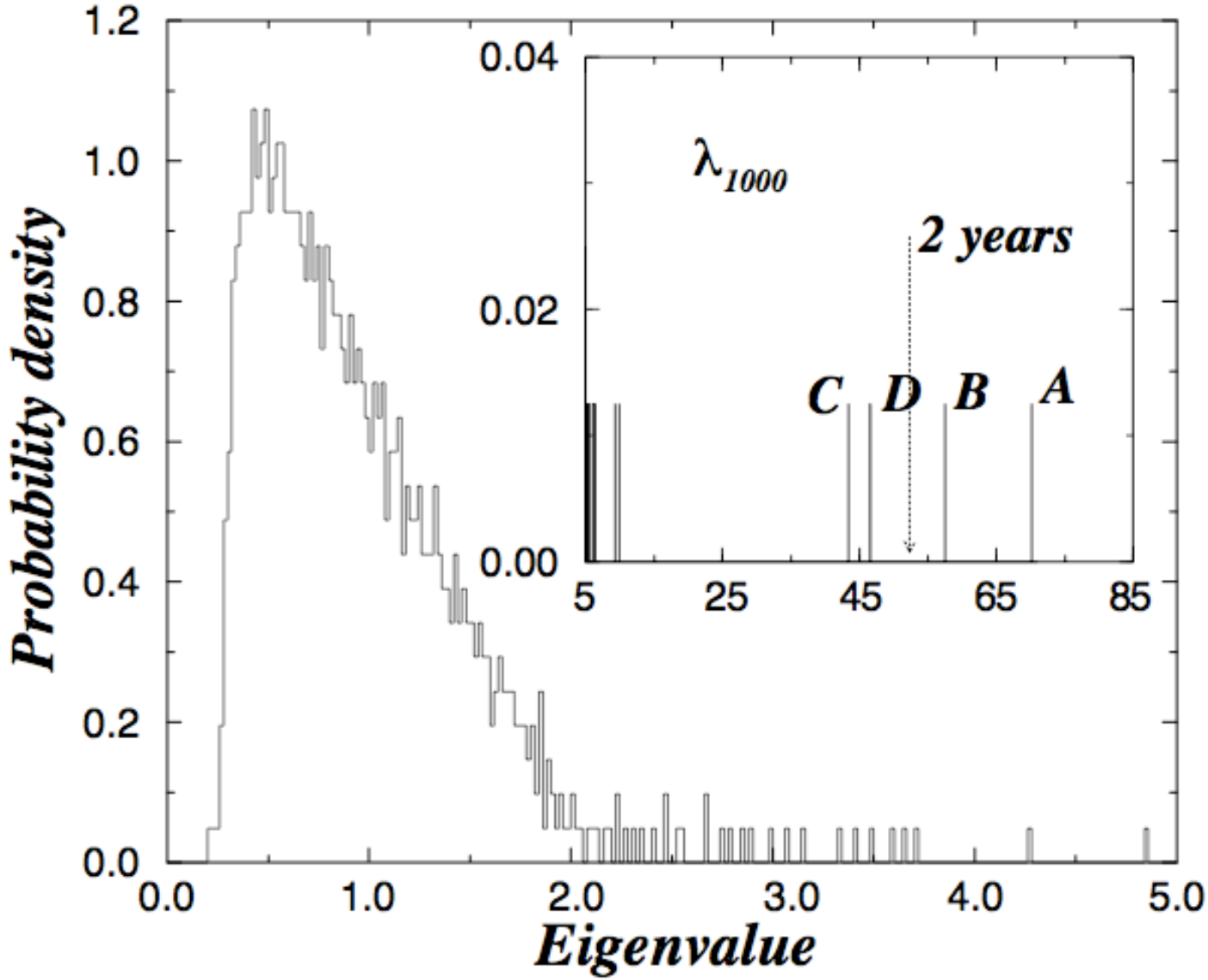
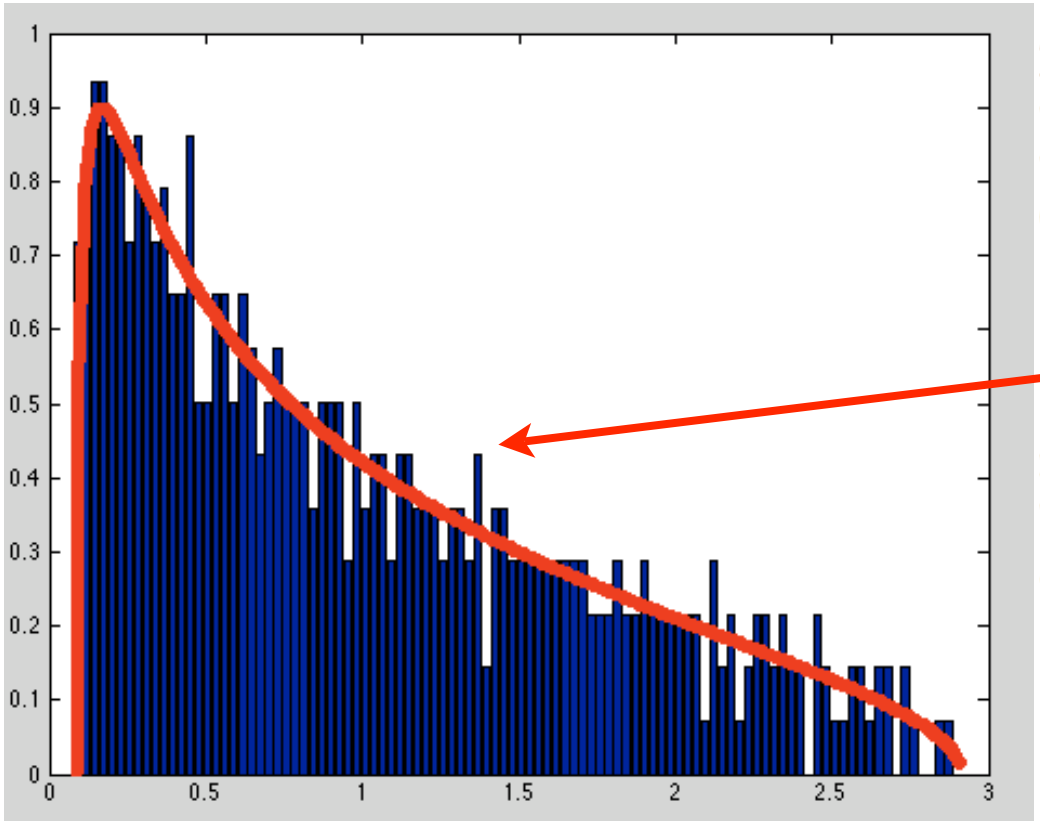


FIG. 1. The probability density function of the eigenvalues c
the normalized cross-correlation matrix **C** for the 1000 larges
U.S. companies in the TAQ database for the 2-year perio



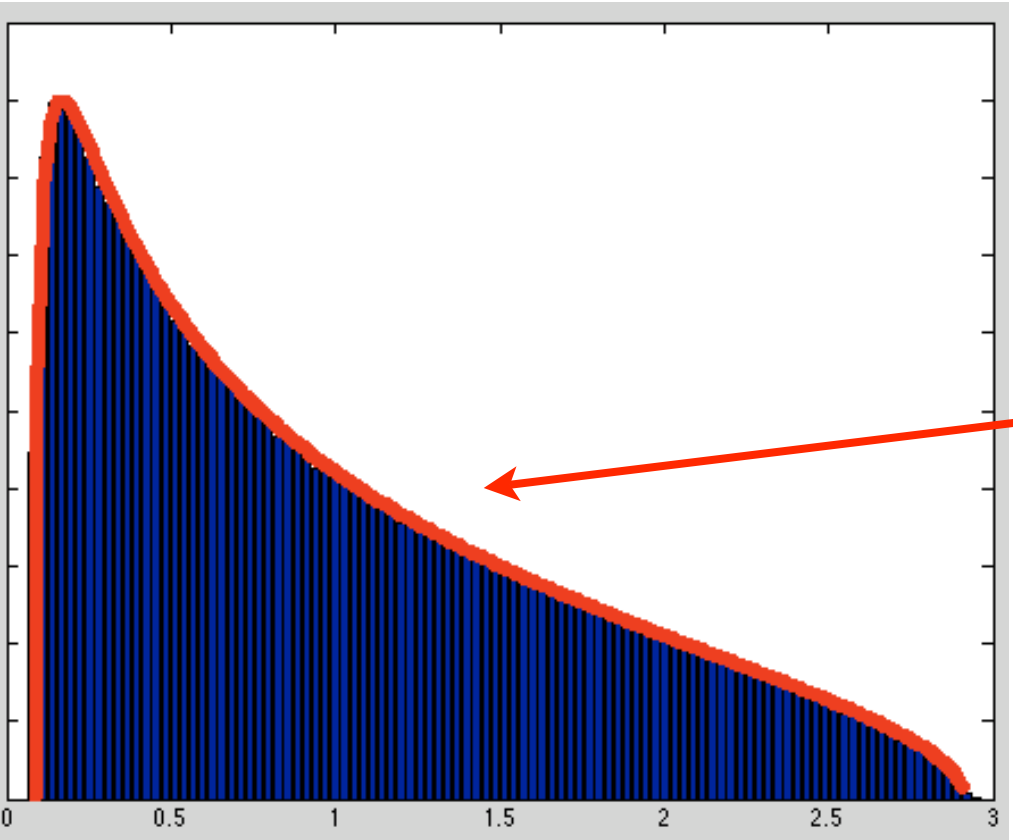
Statistical properties of random matrices such as R are known [26,27]. Particularly, in the limit $N \to \infty$, $L \to \infty$, such that $Q \equiv L/N$ (>1) is fixed, it was shown analytically [27] that the probability density function $P_{\mathrm{rm}}(\lambda)$ of eigenvalues λ of the random correlation matrix R is given by

$$P_{\mathrm{rm}}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \qquad (6)$$

for λ within the bounds $\lambda_- \leqslant \lambda_i \leqslant \lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalues of R, respectively, given by

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}. \qquad (7)$$

Histogram,
1 runs,
N=500,
L=1000



Statistical properties of random matrices such as R are known [26,27]. Particularly, in the limit $N \to \infty$, $L \to \infty$, such that $Q \equiv L/N$ (>1) is fixed, it was shown analytically [27] that the probability density function $P_{\mathrm{rm}}(\lambda)$ of eigenvalues λ of the random correlation matrix R is given by

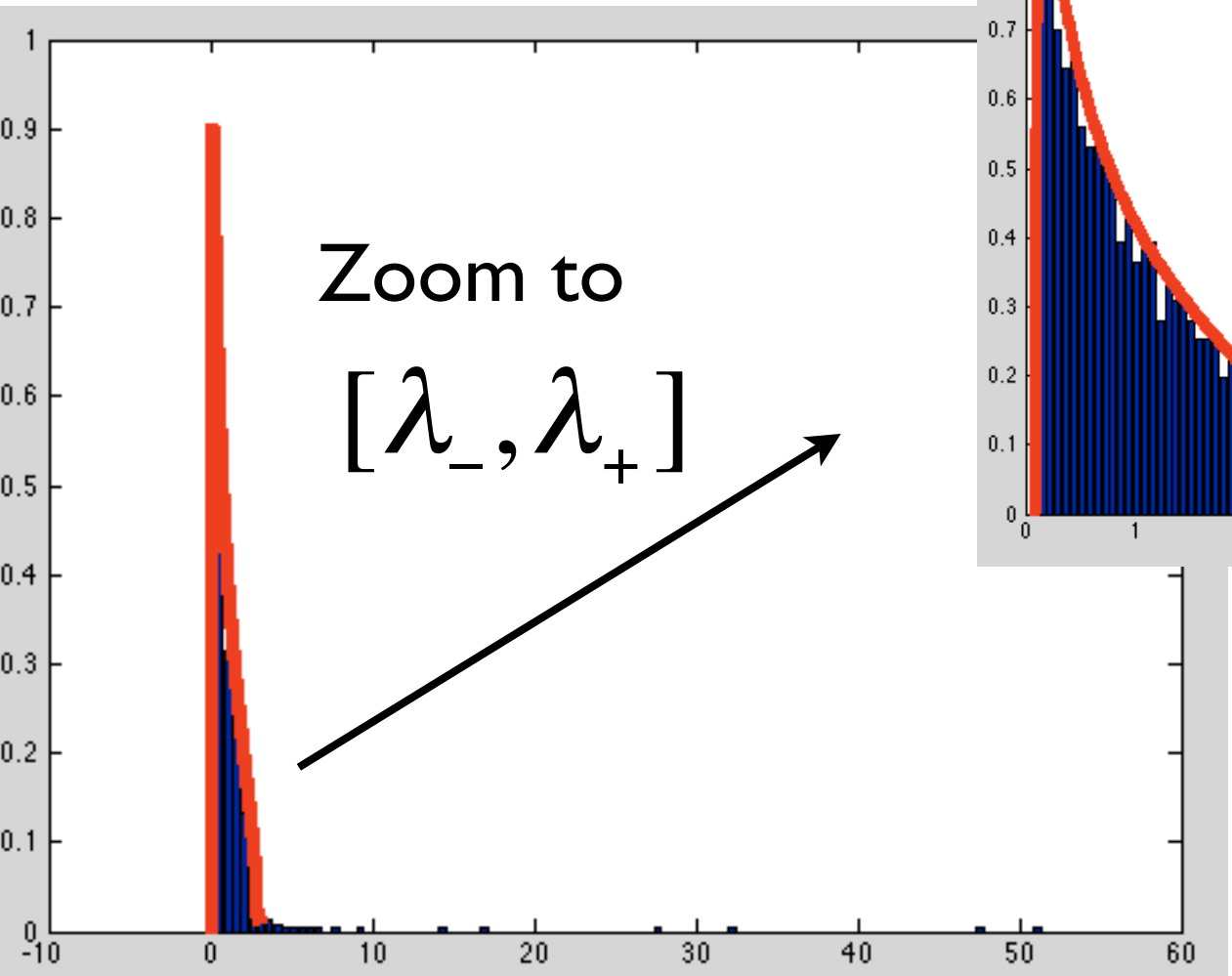
$$P_{\mathrm{rm}}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \qquad (6)$$

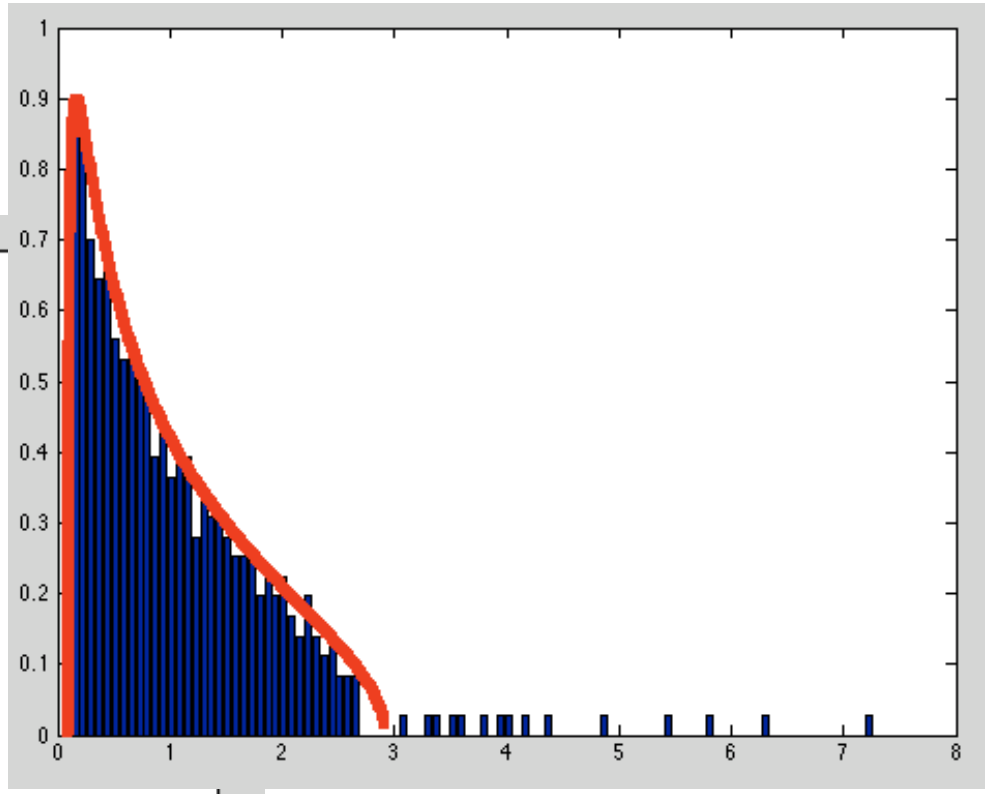
for λ within the bounds $\lambda_- \leqslant \lambda_i \leqslant \lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalues of R, respectively, given by

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}. \qquad (7)$$

Histogram,
200 runs,
LN=500,
L=1000

Simulate a Market
(Will discuss CAREFULLY in Lecture 2)



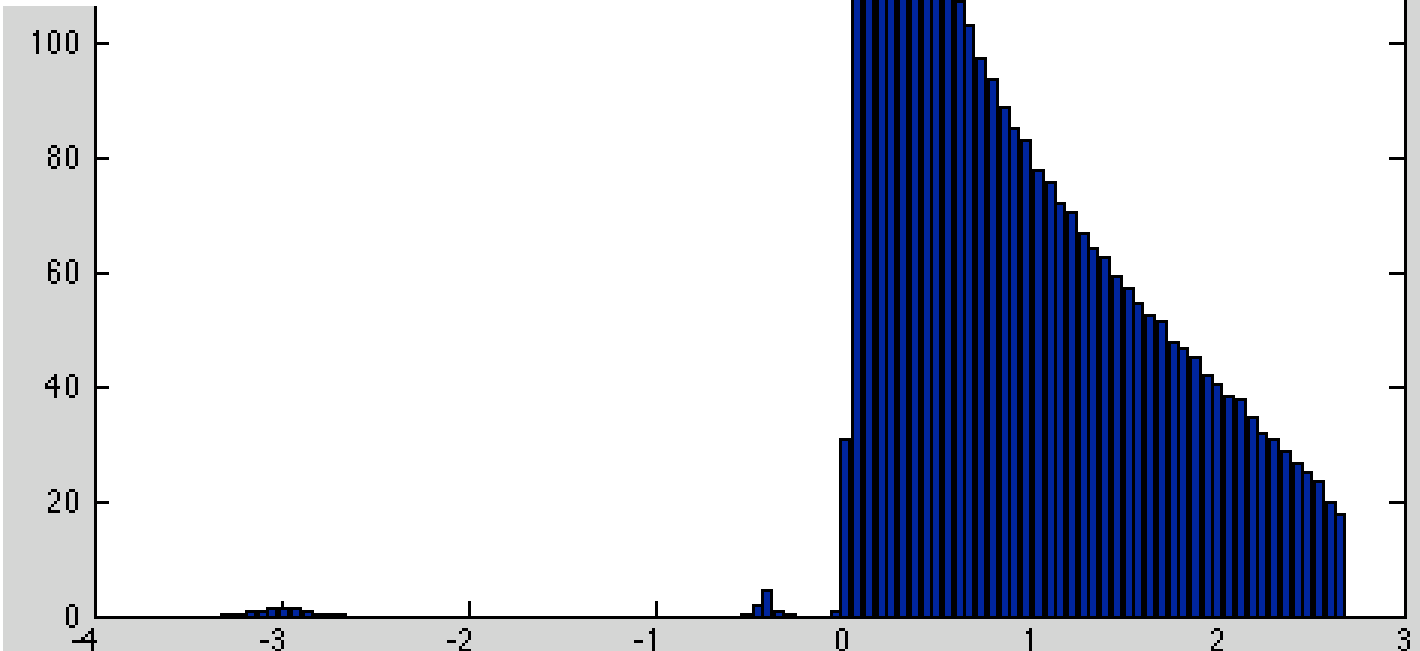


Scale so that the median agrees with Null

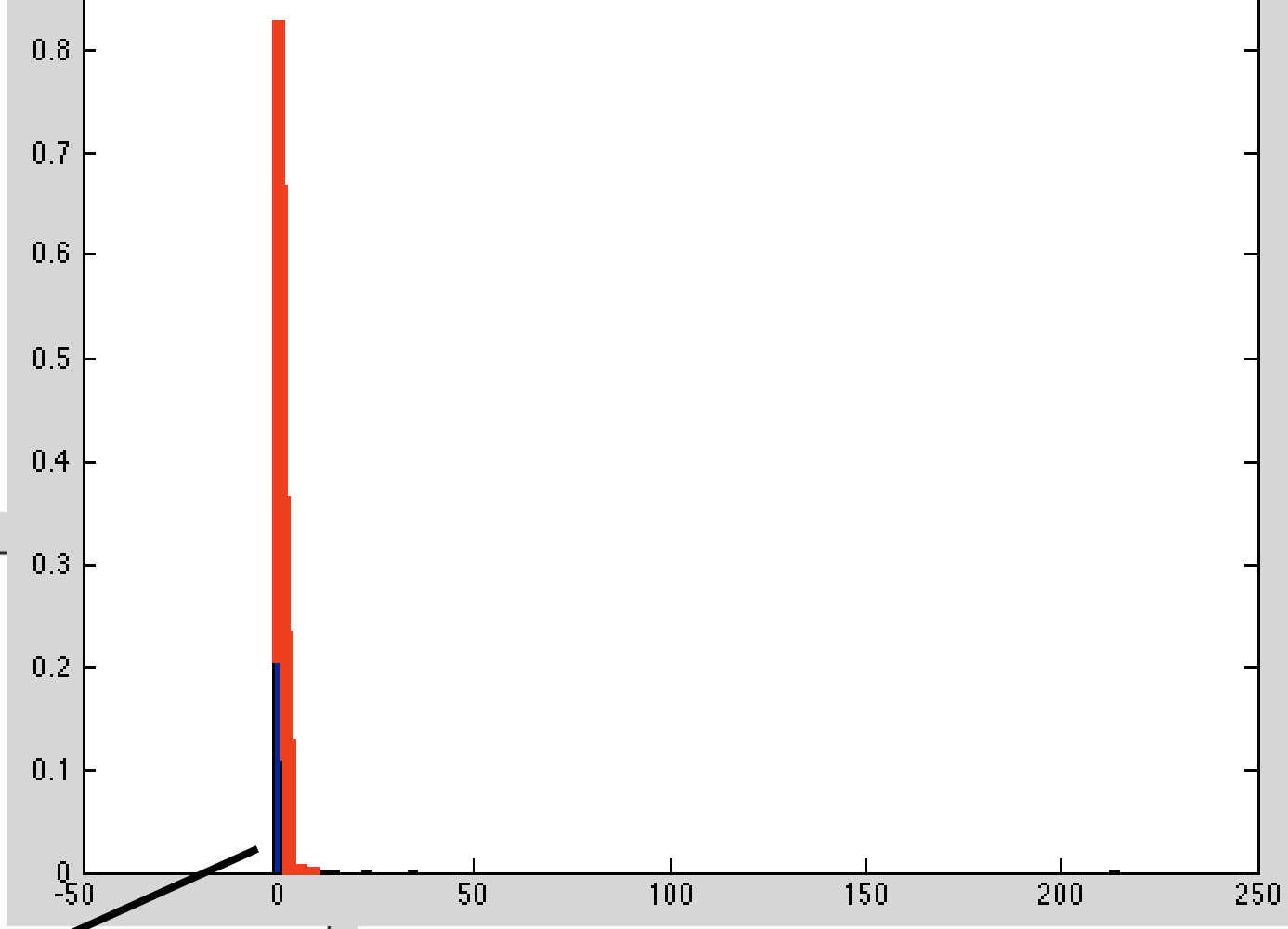
The real market has NaN structure... Different Null.

0.0002	0.0020	-0.0239	-0.0048	-0.0080	-0.0089	0.0051
0.0059	0.0194	0.0478	0.0066	0.0023	0.0139	0.0074
0.0015	0.0245	-0.0191	0.0004	0.0011	-0.0119	0.0037
0.0028	0.0012	0.0068	-0.0043	0.0042	0.0131	-0.0224
-0.0102	-0.0151	0.0232	0.0004	-0.0049	-0.0200	-0.0079
0.0010	-0.0010	0.0189	-0.0044	0.0018	NaN	-0.0028
-0.0006	-0.0064	-0.0194	0.0079	0.0040	-0.0081	-0.0019
0.0065	-0.0077	-0.0513	0.0065	0.0001	-0.0028	0.0118
-0.0015	0.0037	0.0194	0.0065	-0.0017	-0.0138	0.0146
-0.0044	NaN	-0.0139	0.0073	0.0042	0.0227	0.0177
-0.0095	0.0351	0.0190	-0.0064	0.0104	0.0032	-0.0182
-0.0037	-0.0182	-0.0314	-0.0077	-0.0011	-0.0125	-0.0232
-0.0122	-0.0017	0.0106	0.0073	-0.0040	-0.0232	0.0196
0.0053	-0.0112	0.0004	-0.0004	-0.0061	-0.0021	-0.0118
-0.0002	0.0077	-0.0169	0.0026	-0.0048	-0.0105	-0.0028
0.0016	0.0001	-0.0093	-0.0060	0.0044	-0.0080	-0.0061
NaN	NaN	NaN	NaN	NaN	NaN	NaN
-0.0016	0.0057	0.0101	-0.0052	0.0081	0.0034	0.0085
0.0070	0.0076	0.0221	0.0056	0.0065	0.0144	0.0120
0.0010	-0.0056	0	-0.0034	0.0021	0.0046	0.0048
-0.0025	-0.0021	-0.0102	-0.0039	-0.0058	-0.0083	-0.0091

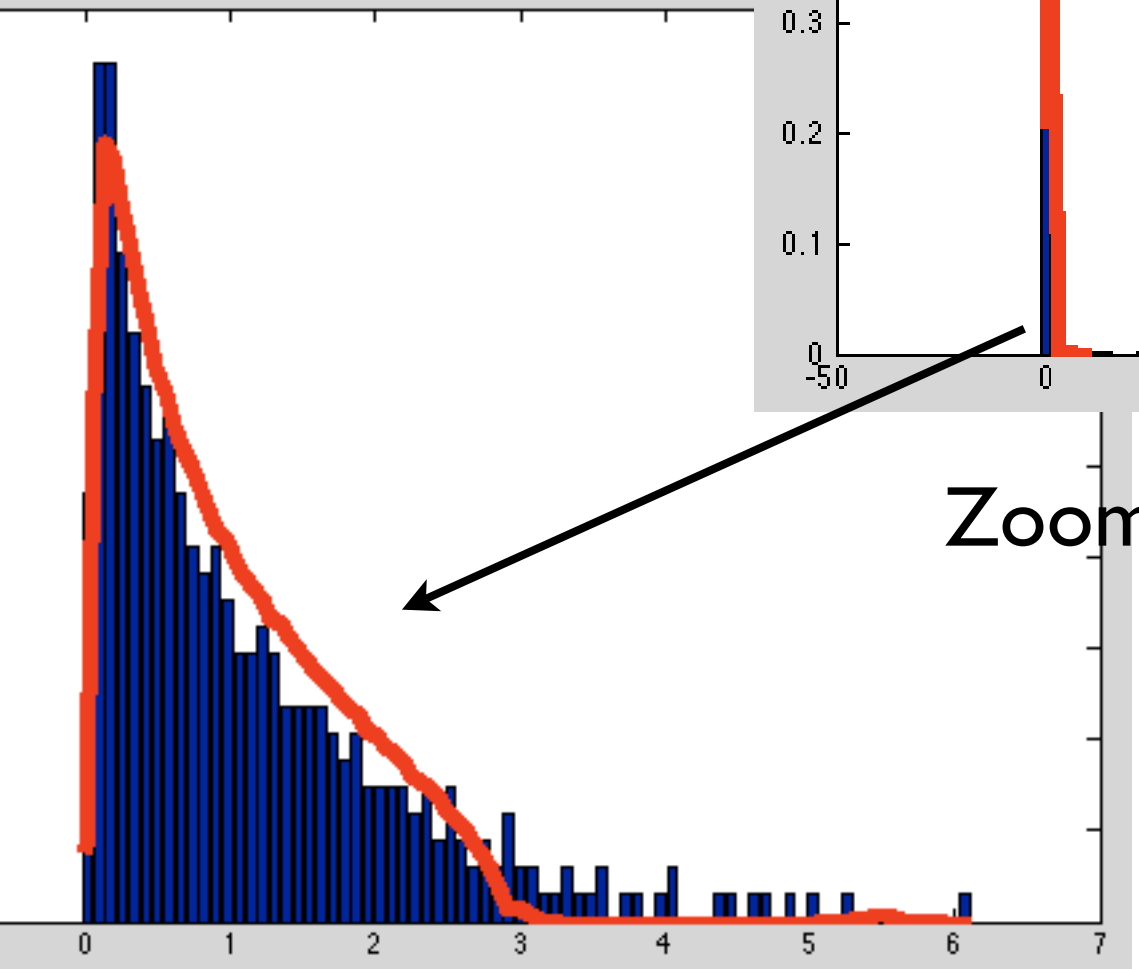
Null formed 200 trials with same NaNs



Actual market compared with Null



Zoom In



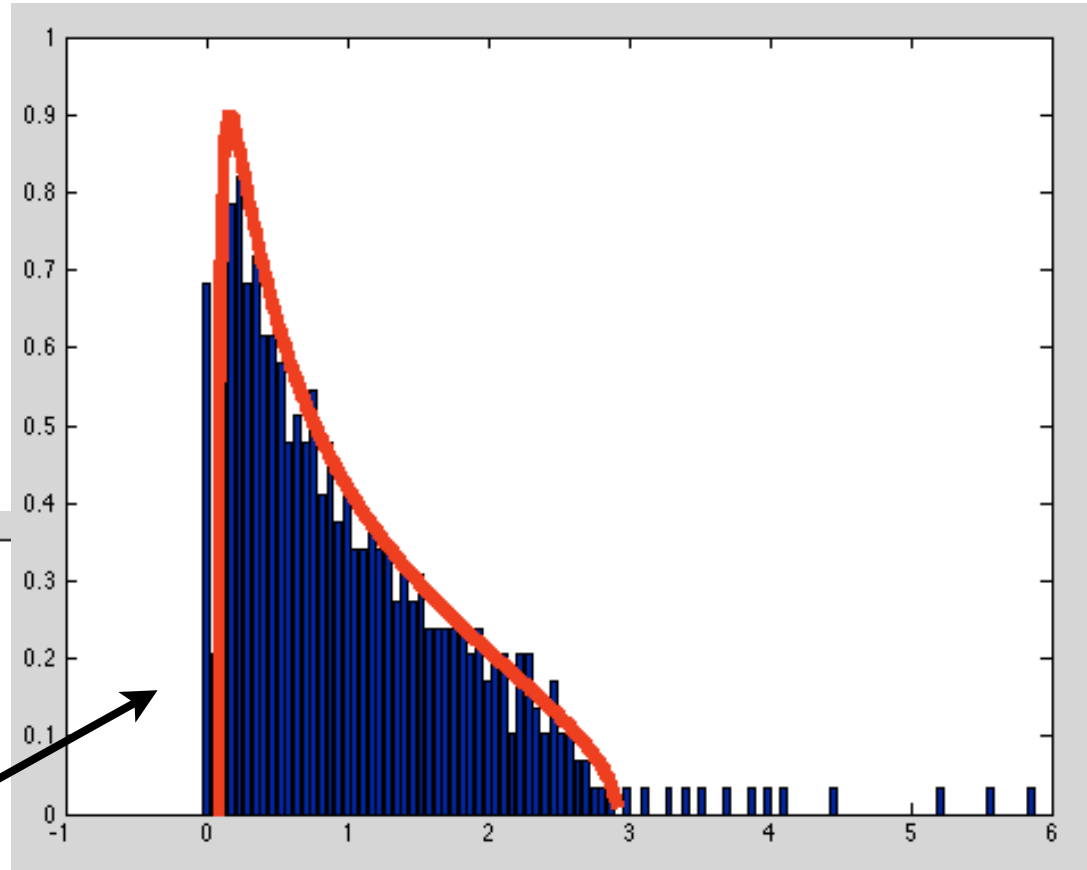
Partition Scrubbing

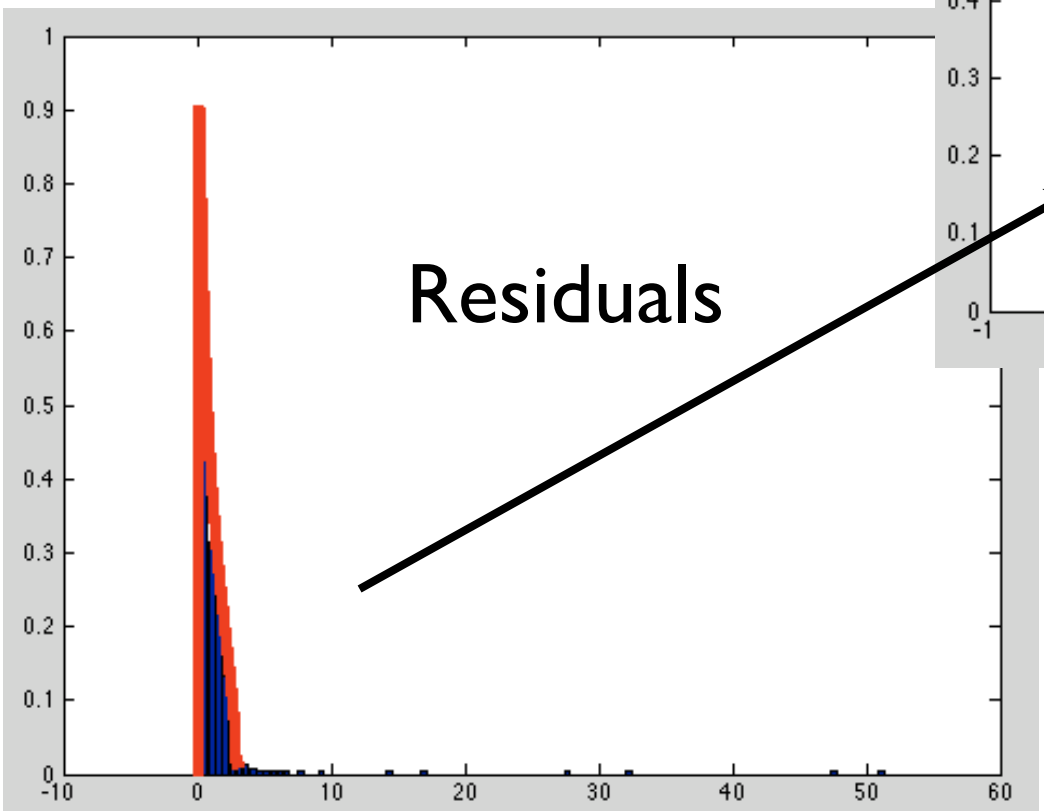
$$X_i = \epsilon_i + \sum_{j=1}^{K} c_{ij} V_j$$

where we identify the V_i as the mean series of the stocks in the k^{th} cluster.

Now we can measure how much our model explains with the residuals....

Constructing a model to explain what you see, here we remove 20 clusters....





Residuals

Model Comparison: How do we quantify how close we are?

The eigenvalue spacing are powerful statistics...

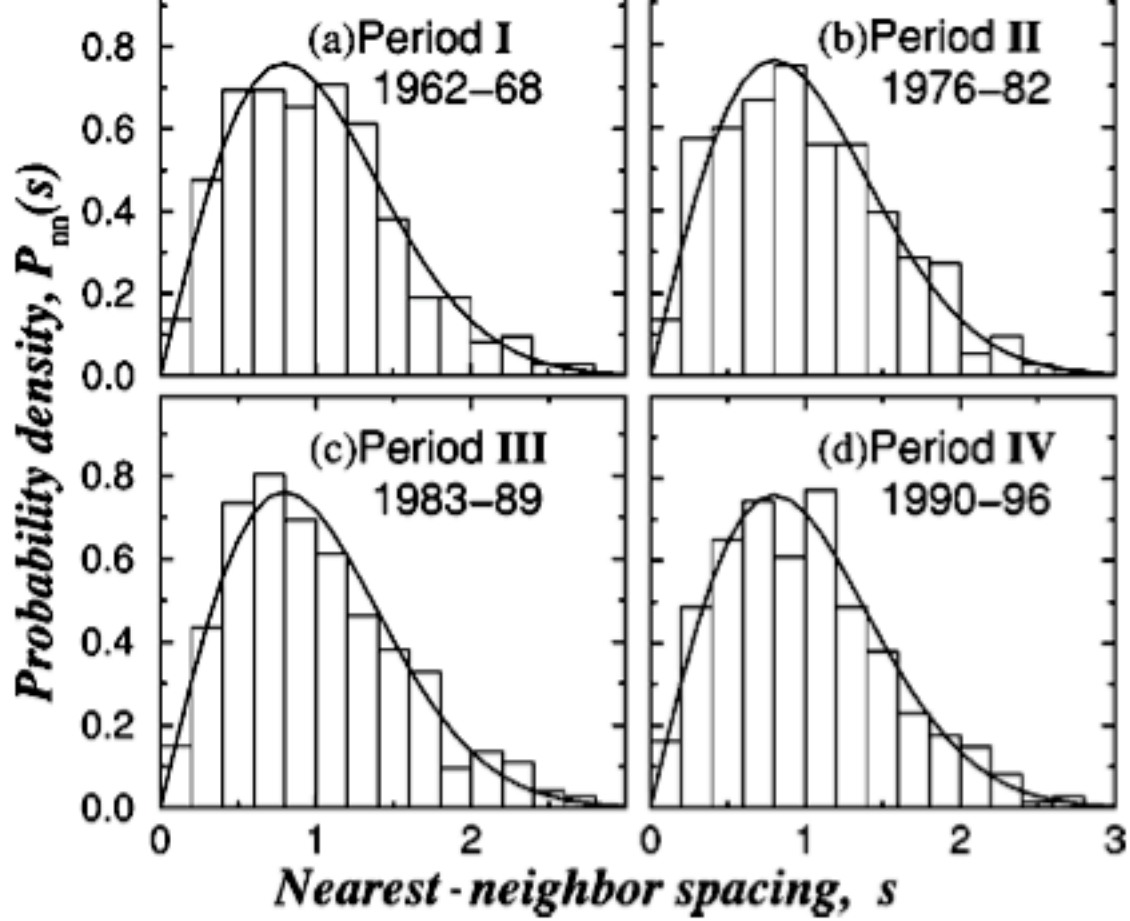


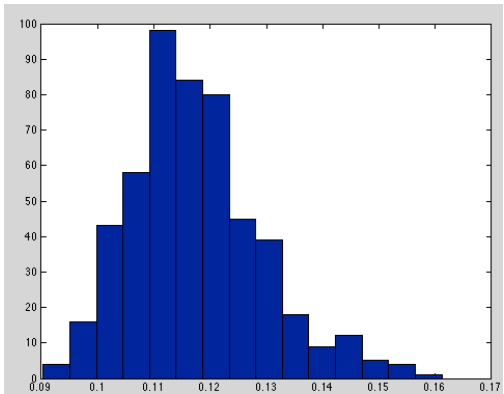
FIG. 6. Nearest-neighbor spacing distribution $P(s)$ of the unfolded eigenvalues ξ_i of C computed from the daily returns of 422 stocks for the 7-yr periods (a) 1962–1968, (b) 1976–1982, (c) 1983–1989, and (d) 1990–1996. We find good agreement with the GOE result (solid curve). The unfolding was performed by using

PHYSICAL REVIEW E, VOLUME 65, 066126

Random matrix approach to cross correlations in financial data

Vasiliki Plerou,[1,2,*] Parameswaran Gopikrishnan,[1] Bernd Rosenow,[1,3] Luís A. Nunes Amaral,[1] Thomas Guhr,[4,5] and H. Eugene Stanley[1]

$$Deviance = \sqrt{\sum_{k=2}^{N-W} \left((\lambda_k - \lambda_{k+W}) - (\hat{\lambda}_k - \hat{\lambda}_{k+W})\right)^2}$$

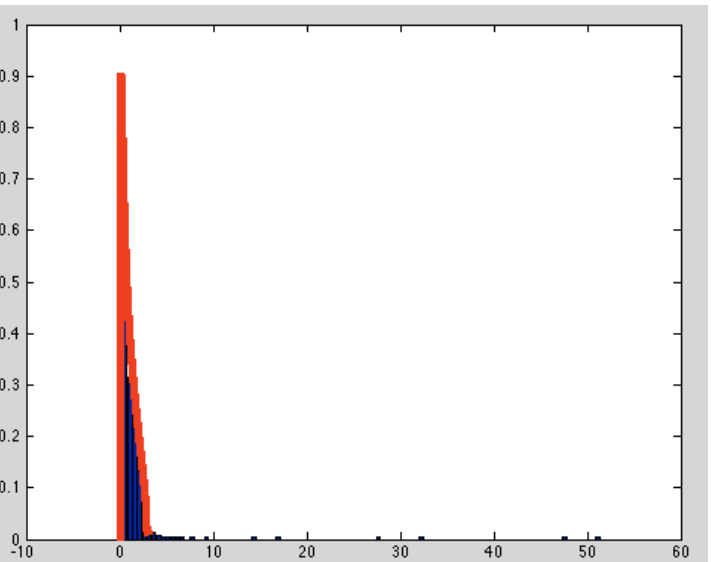


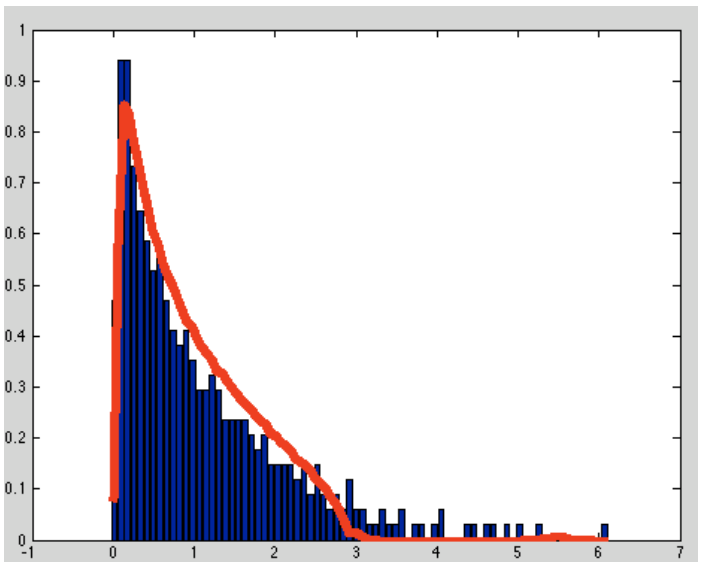
W=5

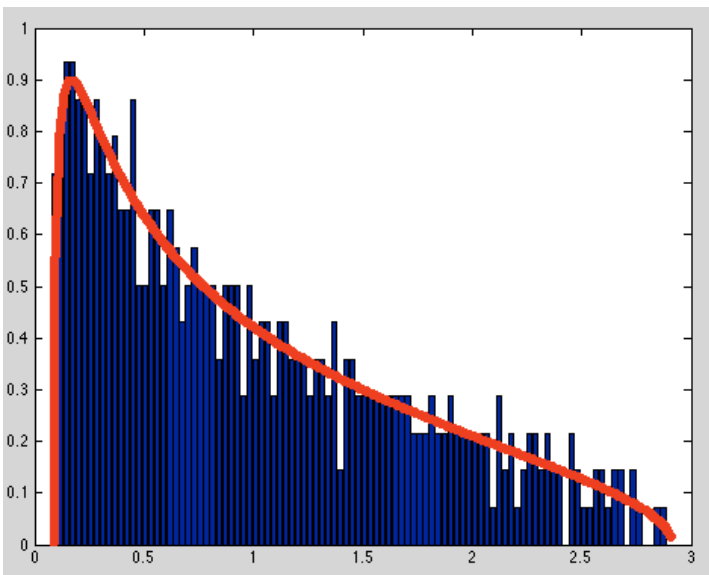
Deviance = 188.3

Deviance = 3.22

Mean Deviance = 0.1175







Complex Model

20 Partition Scrubbed

Null

Big Issues:

What are the eigenvalues trying to tell us?

Can they help us get a glipmse into sytems' the underlying geometry?...

.....the dimension, the number of clusters, the topology, the geometry...