CS6220: DATA MINING TECHNIQUES

Chapter 11: Advanced Clustering Analysis

Instructor: Yizhou Sun

yzsun@ccs.neu.edu

April 10, 2013

Chapter 10. Cluster Analysis: Basic Concepts and Methods

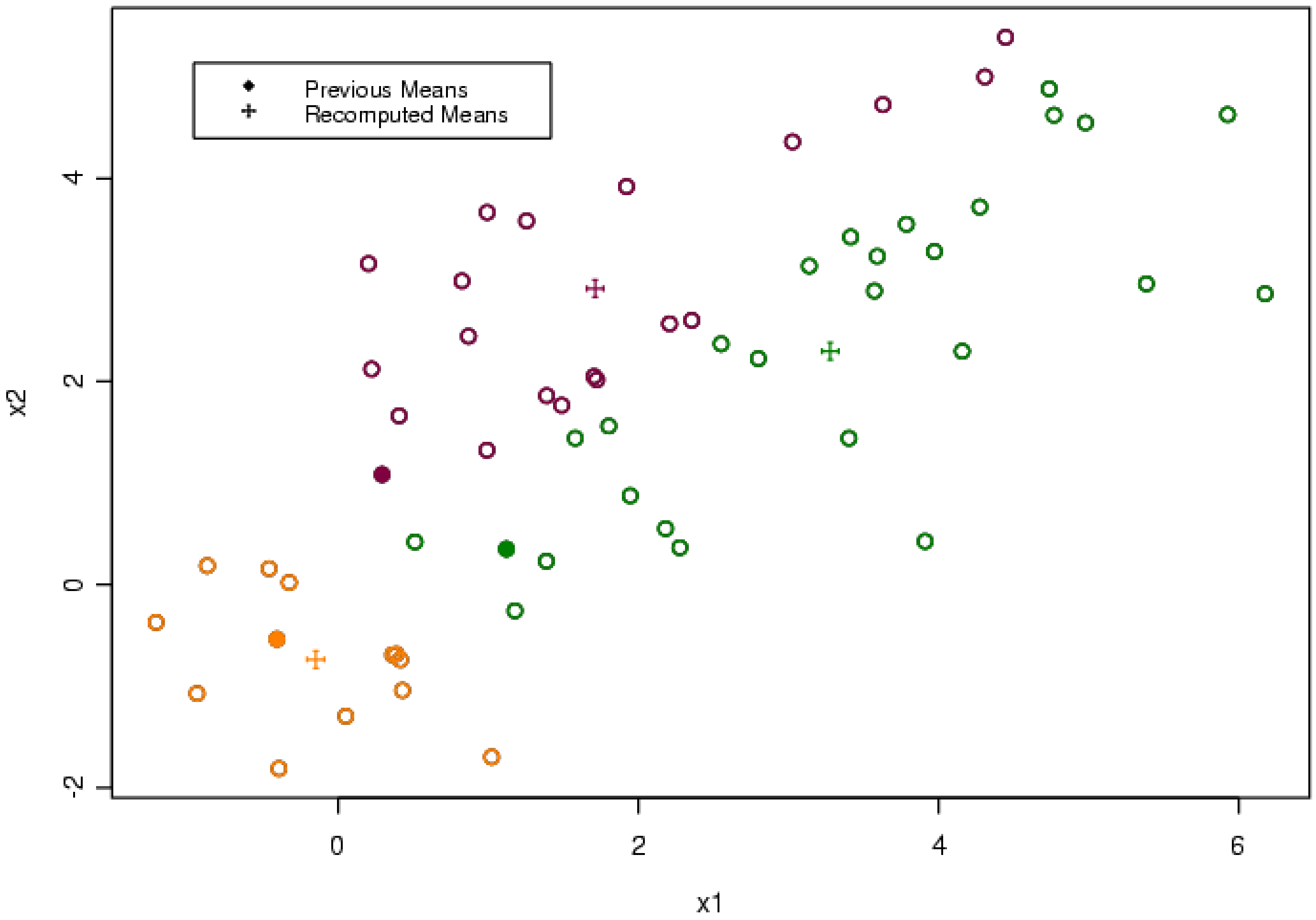
- Beyond K-Means
  - K-means
  - EM-algorithm
  - Kernel K-means
- Clustering Graphs and Network Data
- Summary

Recall K-Means

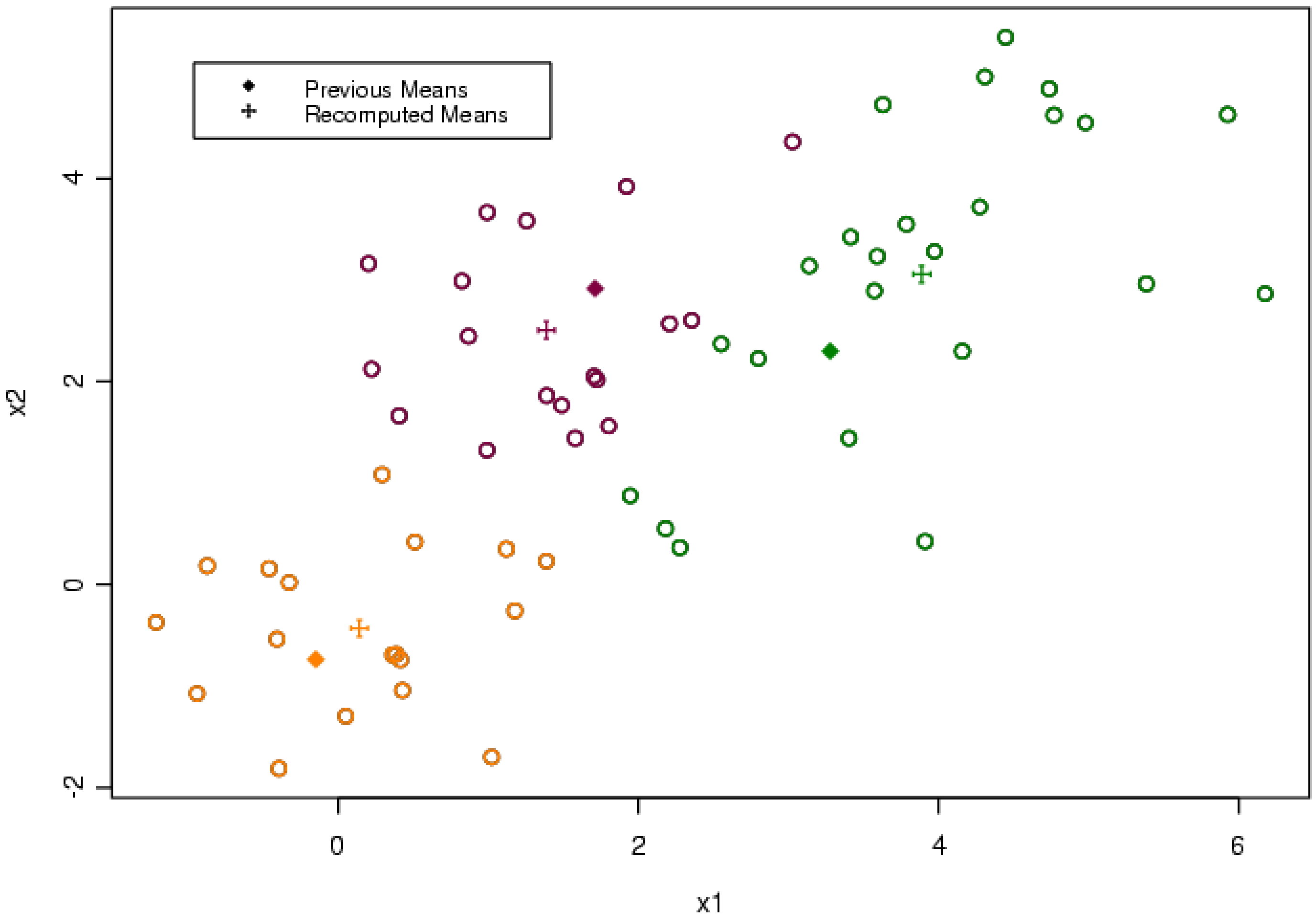
- Objective function
  - $J = \sum_{j=1}^{k} \sum_{C(i)=j} ||x_i - c_j||^2$
  - Total within-cluster variance
- Re-arrange the objective function
  - $J = \sum_{j=1}^{k} \sum_{i} w_{ij} ||x_i - c_j||^2$
    - Where $w_{ij} = 1, if\ x_i\ belongs\ to\ cluster\ j; w_{ij} = 0, otherwise$
  - Looking for:
    - The best assignment w_{ij}
    - The best center c_j

Solution of K-Means

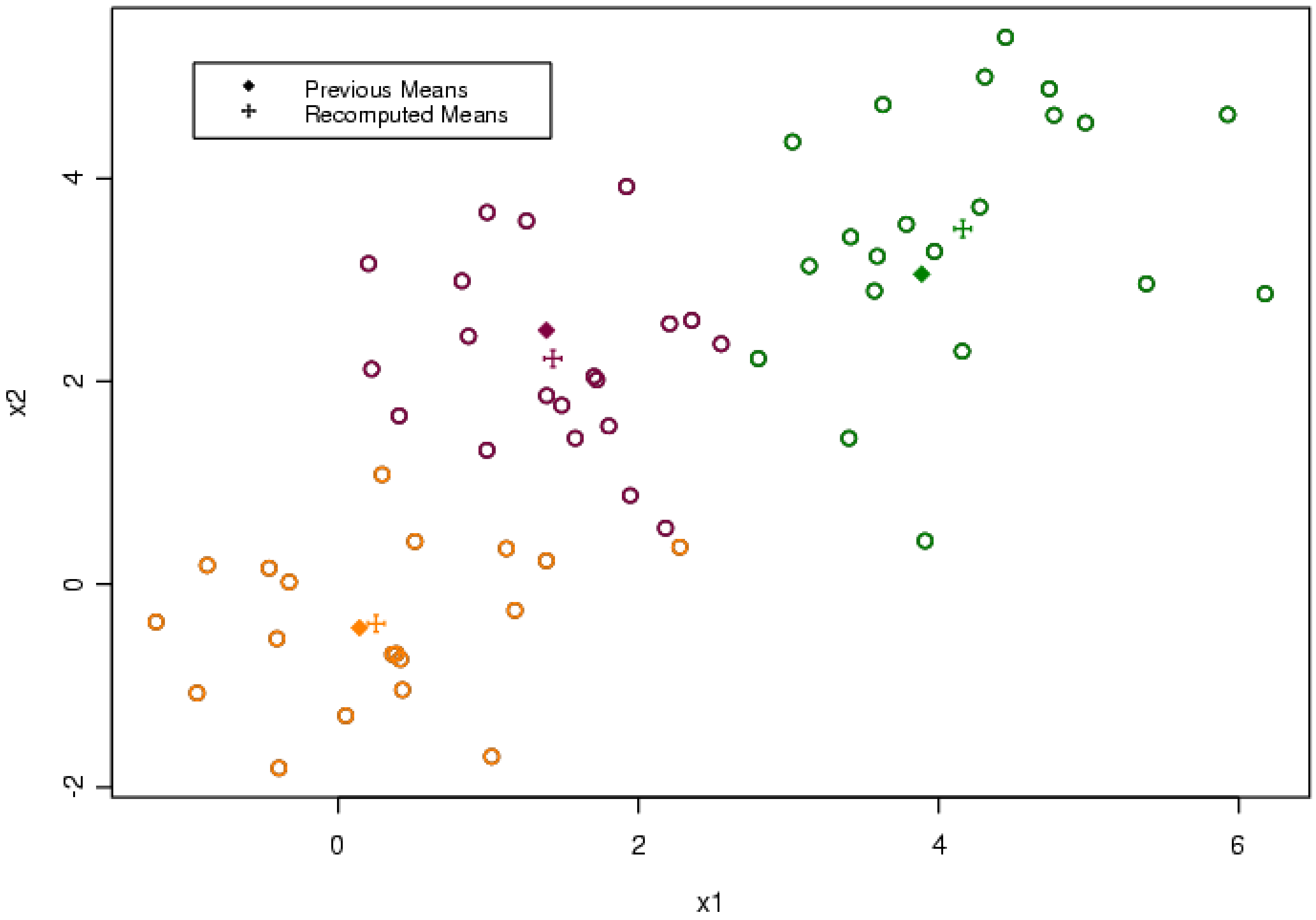
- Iterations
  - Step 1: Fix centers c_j, find assignment w_{ij} that minimizes J
    - => $w_{ij} = 1, if \ ||x_i - c_j||^2$ is the smallest
  - Step 2: Fix assignment w_{ij}, find centers that minimize J
    - => first derivative of J = 0
    - => $\frac{\partial J}{\partial c_j} = -2 \sum_{j=1}^{k} \sum_i w_{ij}(x_i - c_j) = 0$
    - =>$c_j = \frac{\sum_i w_{ij} x_i}{\sum_i w_{ij}}$
      - Note $\sum_i w_{ij}$ is the total number of objects in cluster j

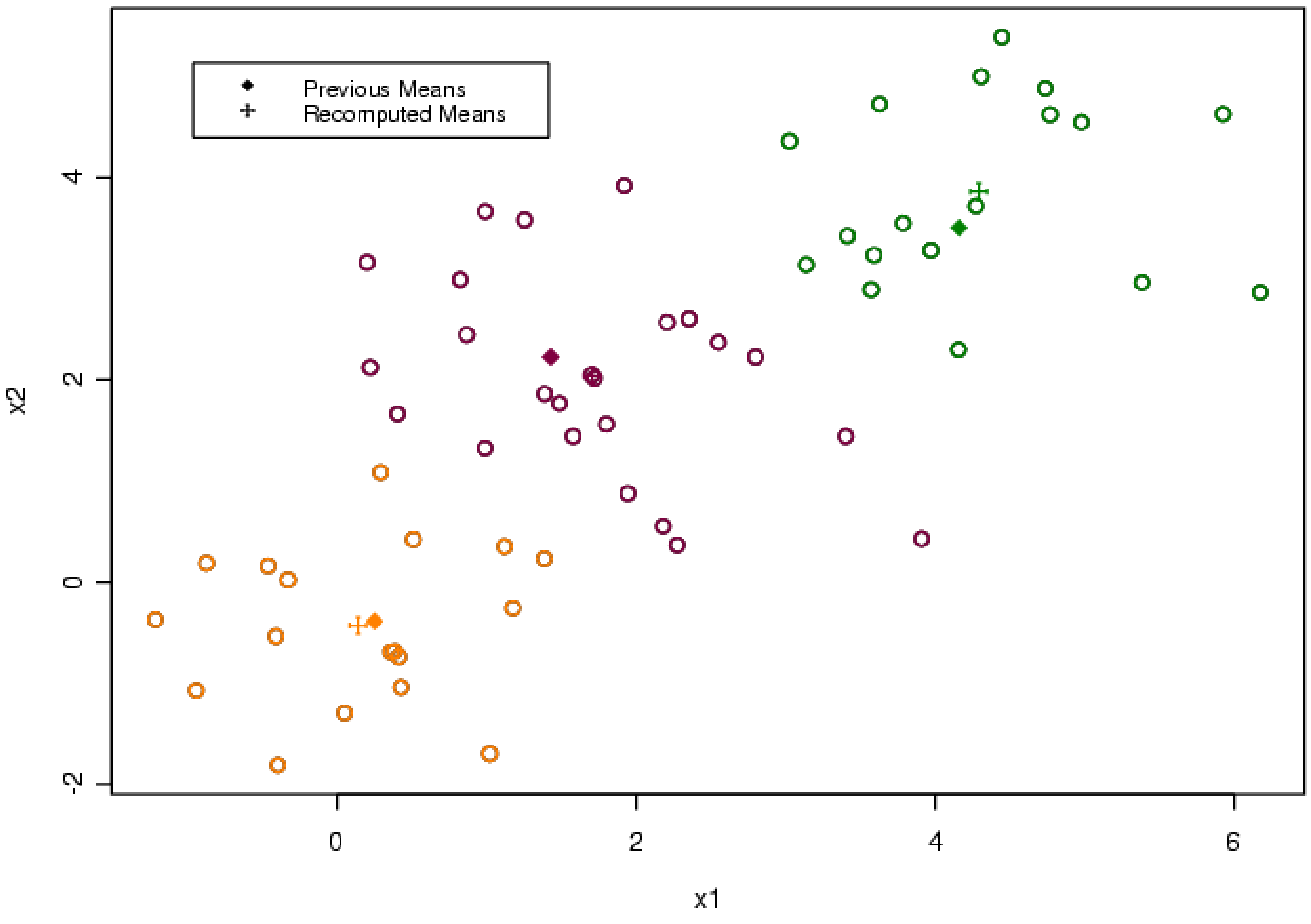


Previous Means
Recomputed Means
x2
x1
-2
0
2
4
6

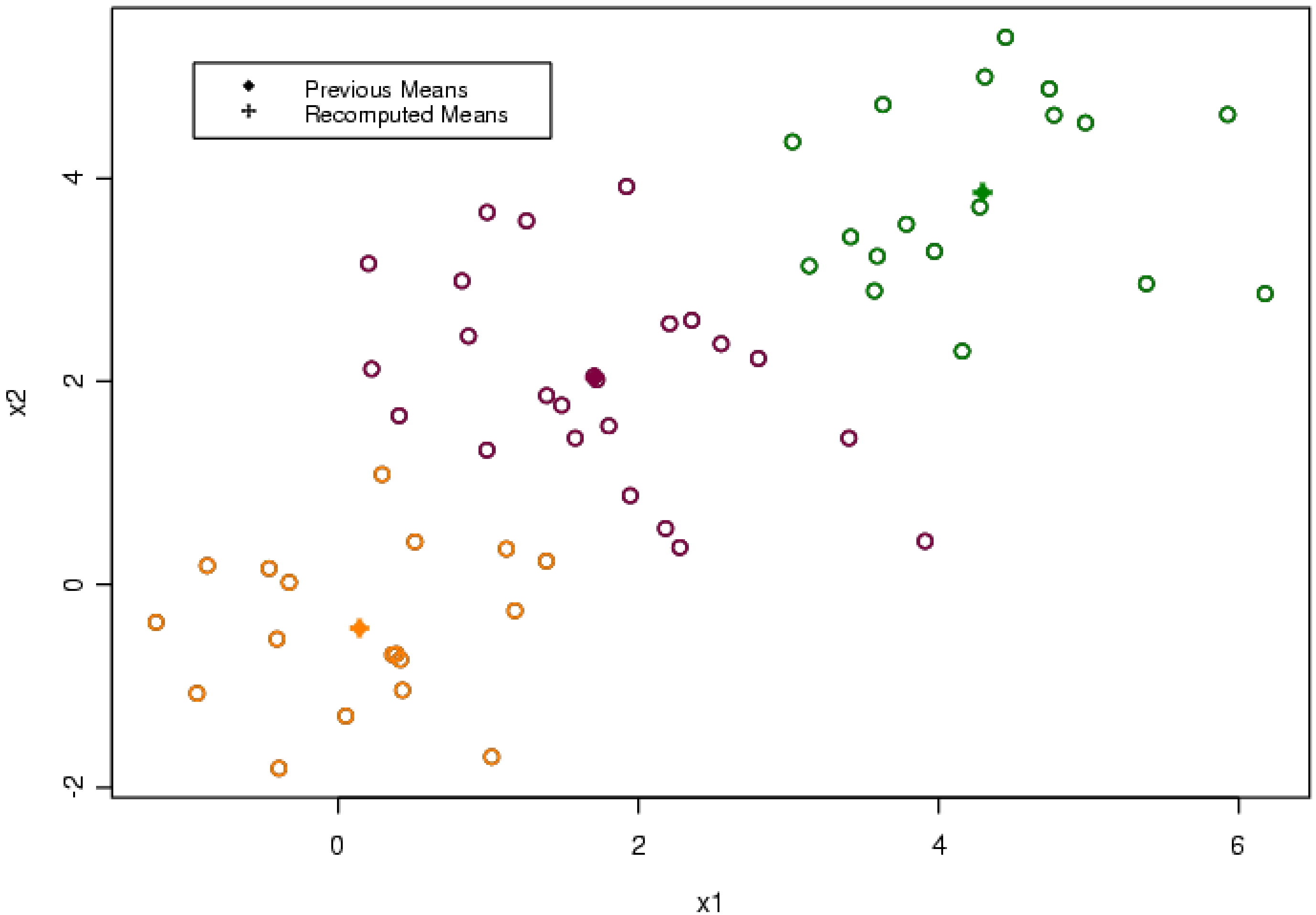


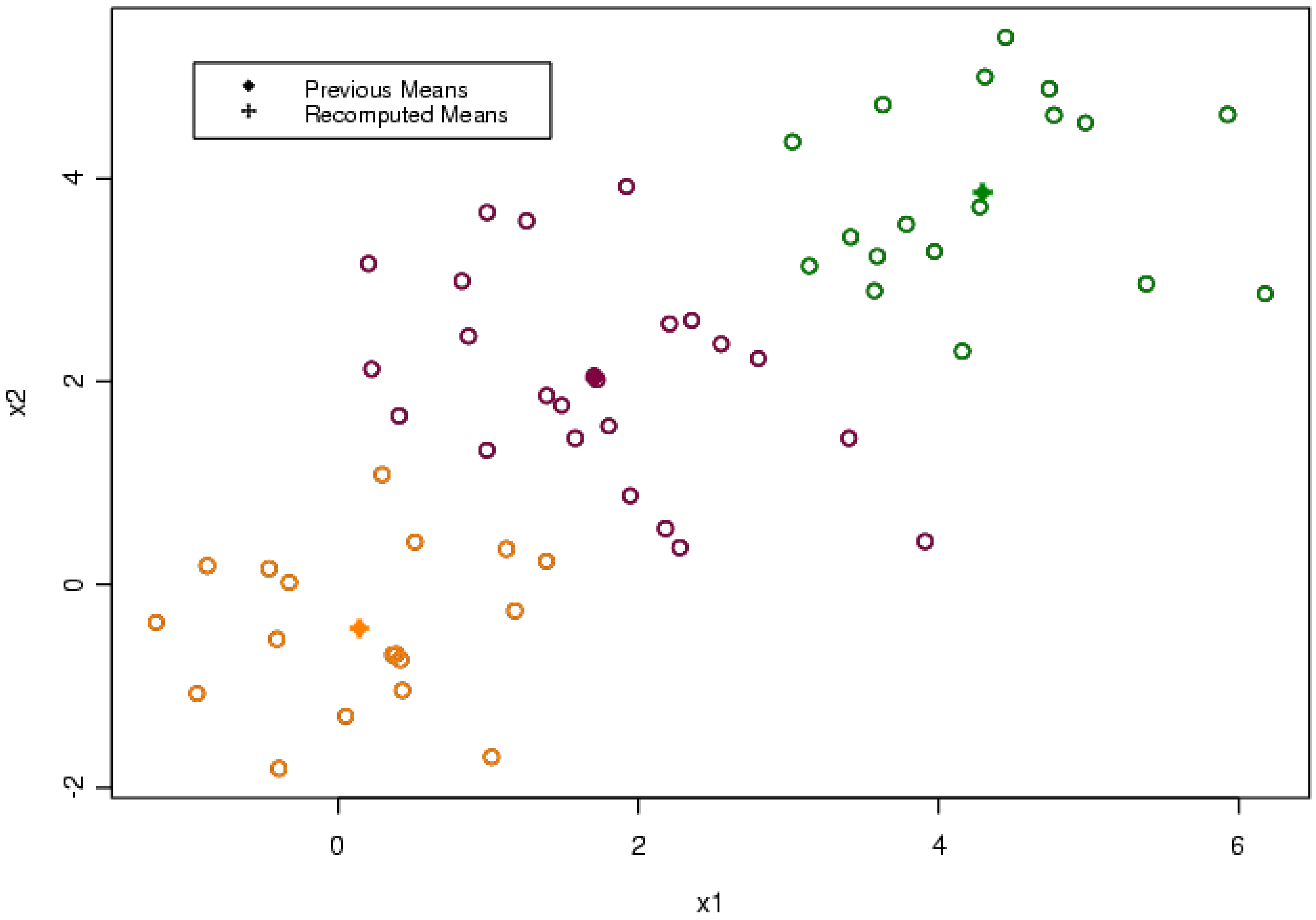
Previous Means
Recomputed Means
x2
x1
-2
0
2
4
6


Previous Means
Recomputed Means
x2
x1
-2
0
2
4
6



Previous Means
Recomputed Means
x1
x2
0
2
4
6
-2
0
2
4

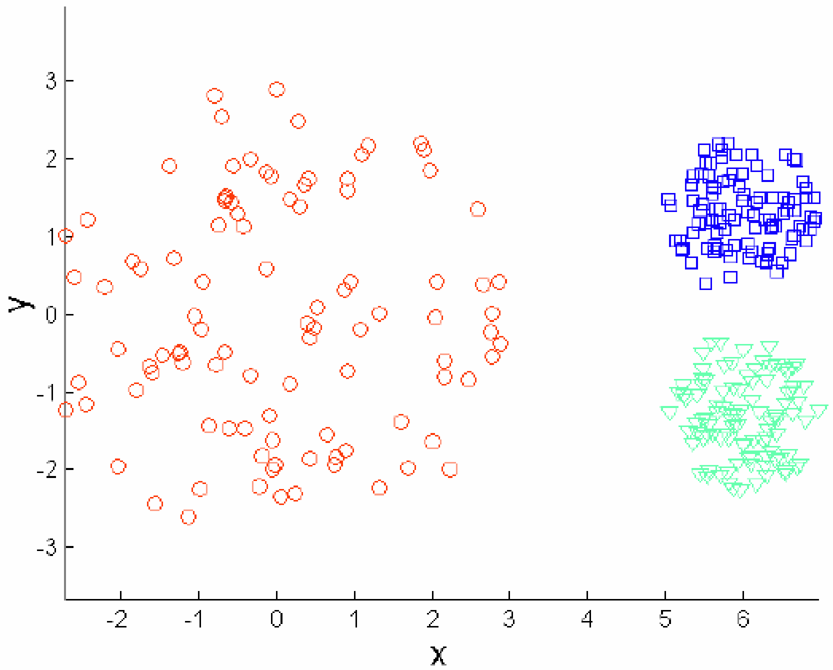

Previous Means
Recomputed Means
x1
x2
0
2
4
6
-2
0
2
4


Previous Means
Recomputed Means
x2
x1
-2
0
2
4
6

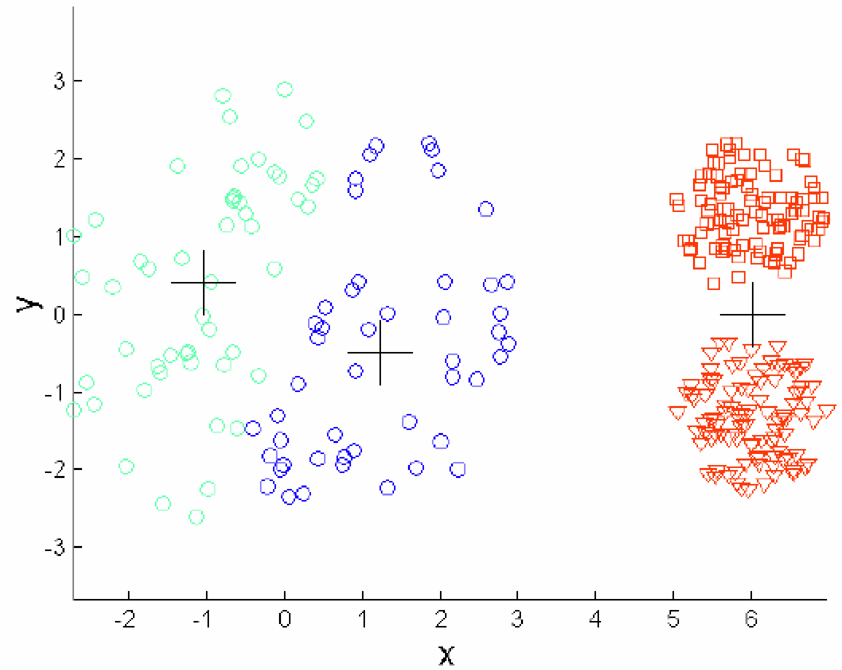
Limitations of K-Means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-Spherical Shapes

Limitations of K-Means: Different Density and Size

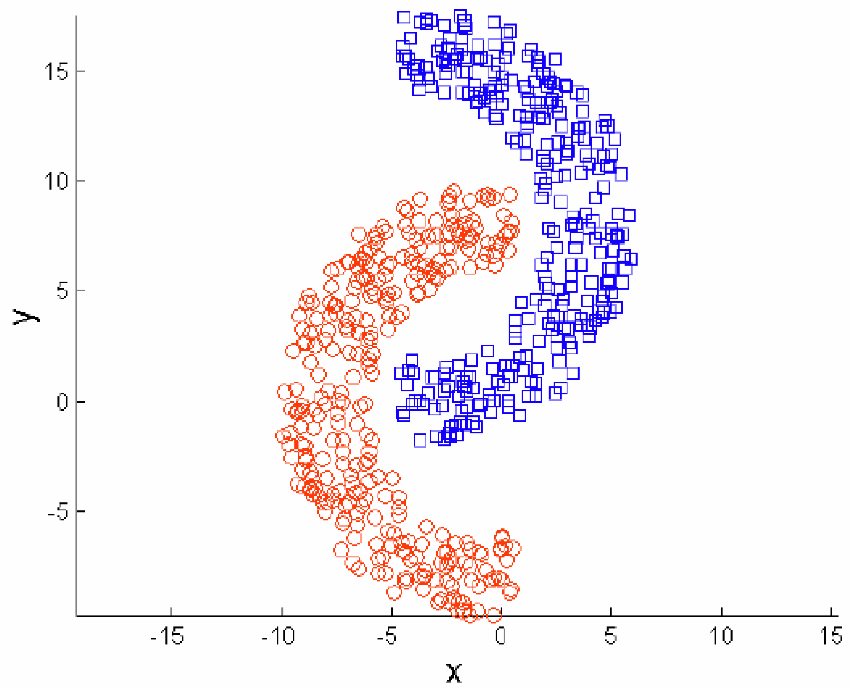


Original Points

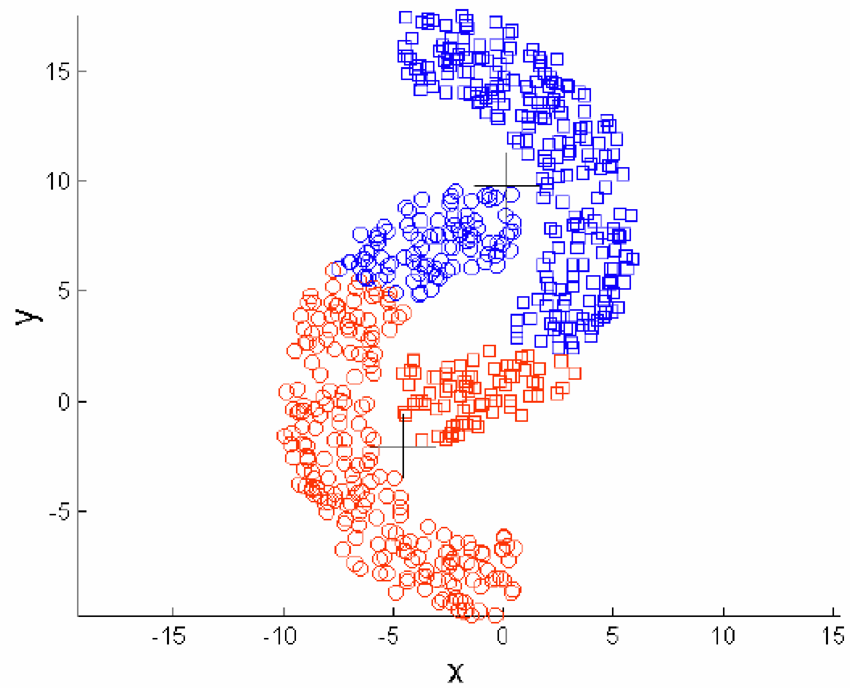


K-means (3 Clusters)

Limitations of K-Means: Non-Spherical Shapes



Original Points



K-means (2 Clusters)

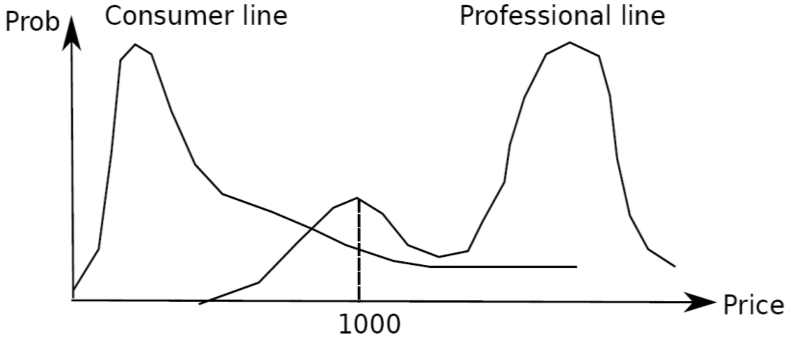
Fuzzy Set and Fuzzy Cluster

- Clustering methods discussed so far
  - Every data object is assigned to exactly one cluster
- Some applications may need for fuzzy or soft cluster assignment
  - Ex. An e-game could belong to both entertainment and software
- Methods: fuzzy clusters and probabilistic model-based clusters
- Fuzzy cluster: A fuzzy set S: $F_S : X \rightarrow [0, 1]$ (value between 0 and 1)

Probabilistic Model-Based Clustering

- Cluster analysis is to find hidden categories.
- A hidden category (i.e., *probabilistic cluster)* is a distribution over the data space, which can be mathematically represented using a probability density function (or distribution function).

- Ex. categories for digital cameras sold
  - consumer line vs. professional line
  - density functions f_1, f_2 for C_1, C_2
  - obtained by probabilistic clustering



- A **mixture model** assumes that a set of observed objects is a mixture of instances from multiple probabilistic clusters, and conceptually each observed object is generated independently
- **Our task**: infer a set of *k* probabilistic clusters that is mostly likely to generate *D* using the above data generation process

Mixture Model-Based Clustering

- A set C of k probabilistic clusters $C_1, \ldots, C_k$ with probability density functions $f_1, \ldots, f_k$, respectively, and their probabilities $\omega_1, \ldots, \omega_k$.
- Probability of an object o generated by cluster C_j is $$P(o|C_j) = \omega_j f_j(o)$$
- Probability of o generated by the set of cluster $\boldsymbol{C}$ is $$P(o|\boldsymbol{C}) = \sum_{j=1}^{k} \omega_j f_j(o)$$
- Since objects are assumed to be generated independently, for a data set D = $\{o_1, \ldots, o_n\}$, we have,

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

- Task: Find a set C of k probabilistic clusters s.t. $P(D|\boldsymbol{C})$ is maximized

The EM (Expectation Maximization) Algorithm

- **The (EM) algorithm:** A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models.
  - **E-step** assigns objects to clusters according to the current fuzzy clustering or parameters of probabilistic clusters
    - $w_{ij}^t = p(z_i = j | \theta_j^t, x_i) \propto p(x_i | C_j^t, \theta_j^t) p(C_j^t)$
  - **M-step** finds the new clustering or parameters that minimize the sum of squared error (SSE) or the expected likelihood
    - Under uni-variant normal distribution assumptions:
    - $\mu_j^{t+1} = \frac{\sum_i w_{ij}^t x_i}{\sum_i w_{ij}^t}; \sigma_j^2 = \frac{\sum_i w_{ij}^t \left\| x_i - c_j^t \right\|^2}{\sum_i w_{ij}^t}; p(C_j^t) \propto \sum_i w_{ij}^t$
- More about mixture model and EM algorithms: http://www.stat.cmu.edu/~cshalizi/350/lectures/29/lecture-29.pdf

K-Means: Special Case of Gaussian Mixture Model

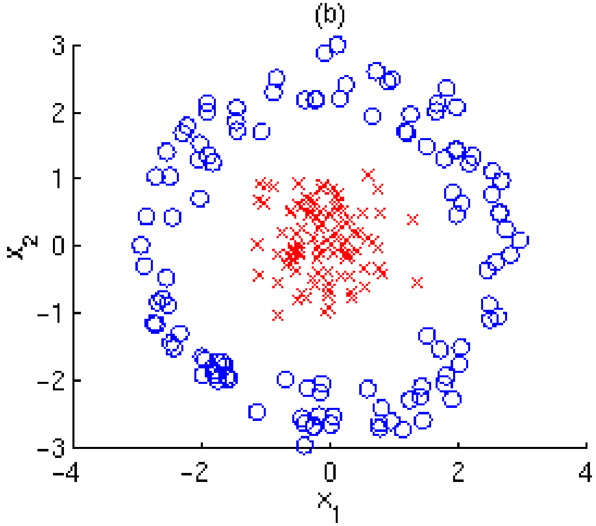
- When each Gaussian component with covariance matrix $\sigma^2 I$
  - Soft K-means
- When $\sigma^2 \rightarrow 0$
  - Soft assignment becomes hard assignment

Advantages and Disadvantages of Mixture Models

- Strength
  - Mixture models are more general than partitioning
  - Clusters can be characterized by a small number of parameters
  - The results may satisfy the statistical assumptions of the generative models
- Weakness
  - Converge to local optimal (overcome: run multi-times w. random initialization)
  - Computationally expensive if the number of distributions is large, or the data set contains very few observed data points
  - Need large data sets
  - Hard to estimate the number of clusters

Kernel K-Means

- How to cluster the following data?



- A non-linear map: $\phi: R^n \rightarrow F$
  - Map a data point into a higher/infinite dimensional space
  - $x \rightarrow \phi(x)$
- Dot product matrix K_{ij}
  - $K_{ij} = <\phi(x_i), \phi(x_j)>$

Solution of Kernel K-Means

- Objective function under new feature space:
  - $J = \sum_{j=1}^{k} \sum_i w_{ij} ||\phi(x_i) - c_j||^2$
- Algorithm
  - By fixing assignment w_{ij}
    - $c_j = \sum_i w_{ij}\, \phi(x_i) / \sum_i w_{ij}$
  - In the assignment step, assign the data points to the closest center
    - $d(x_i, c_j) = \left\| \phi(x_i) - \frac{\sum_{i'} w_{i'j} \phi(x_{i'})}{\sum_{i'} w_{i'j}} \right\|^2 =$

$$\phi(x_i) \cdot \phi(x_i) - 2 \frac{\sum_{i'} w_{i'j} \phi(x_i) \cdot \phi(x_{i'})}{\sum_{i'} w_{i'j}} + \frac{\sum_{i'} \sum_l w_{i'j} w_{lj} \phi(x_{i'}) \cdot \phi(x_l)}{(\sum_{i'} w_{i'j})\hat{}2}$$

Do not really need to know $\phi(x)$, *but only* K_{ij}

Advatanges and Disadvantages of Kernel K-Means

- **Advantages**
  - Algorithm is able to identify the non-linear structures.
- **Disadvantages**
  - Number of cluster centers need to be predefined.
  - Algorithm is complex in nature and time complexity is large.
- **References**

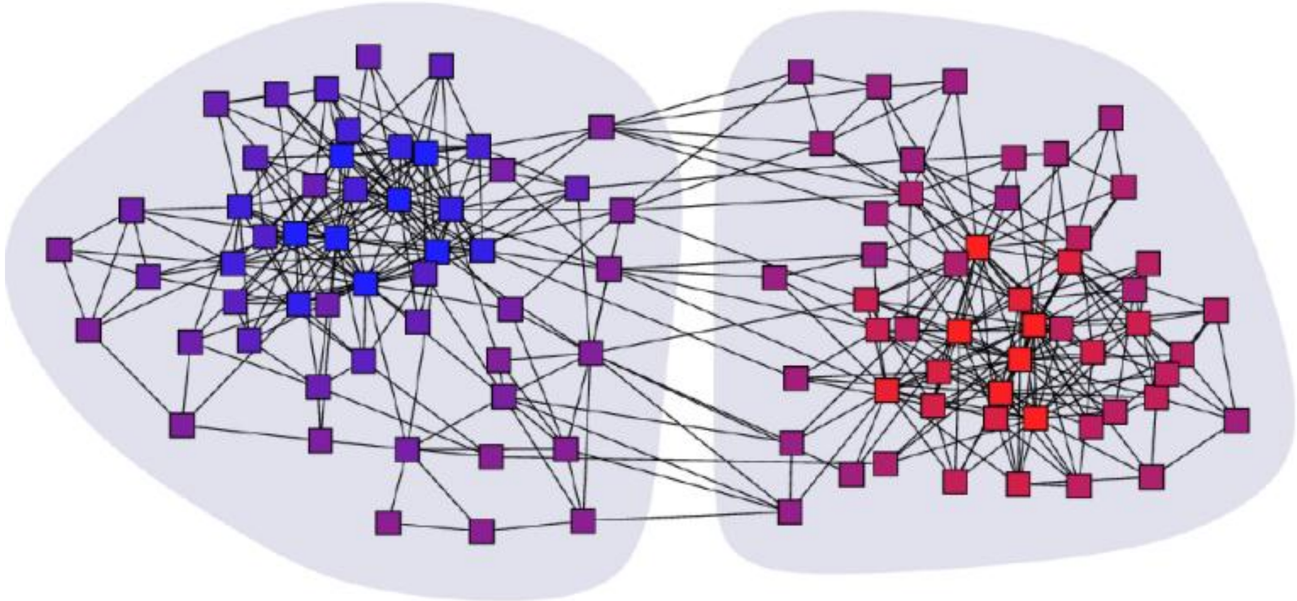
  - Kernel k-means and Spectral Clustering by Max Welling.
  - Kernel k-means, Spectral Clustering and Normalized Cut by Inderjit S. Dhillon, Yuqiang Guan and Brian Kulis.
  - An Introduction to kernel methods by Colin Campbell.

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Beyond K-Means
  - K-means
  - EM-algorithm for Mixture Models
  - Kernel K-means
- Clustering Graphs and Network Data
- Summary

Clustering Graphs and Network Data

- Applications
  - Bi-partite graphs, e.g., customers and products, authors and conferences
  - Web search engines, e.g., click through graphs and Web graphs
  - Social networks, friendship/coauthor graphs



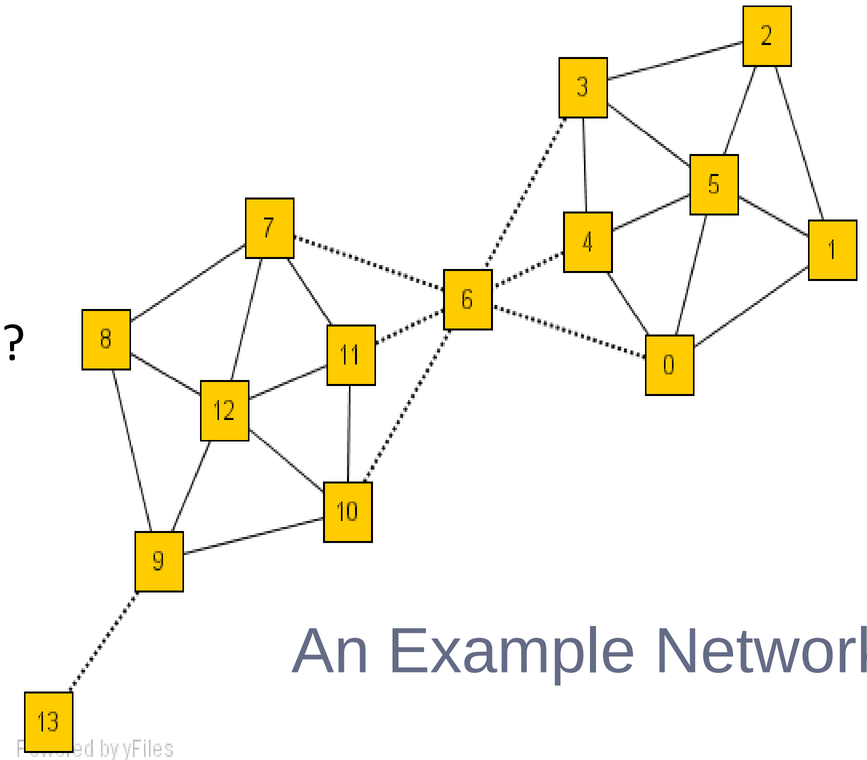
Clustering books about politics [Newman, 2006]

Algorithms

- Graph clustering methods
  - Density-based clustering: SCAN (Xu et al., KDD'2007)
  - Spectral clustering
  - Modularity-based approach
  - Probabilistic approach
  - Nonnegative matrix factorization
  - …

SCAN: Density-Based Clustering of Networks

- How many clusters?
- What size should they be?
- What is the best partitioning?
- Should some points be segregated?

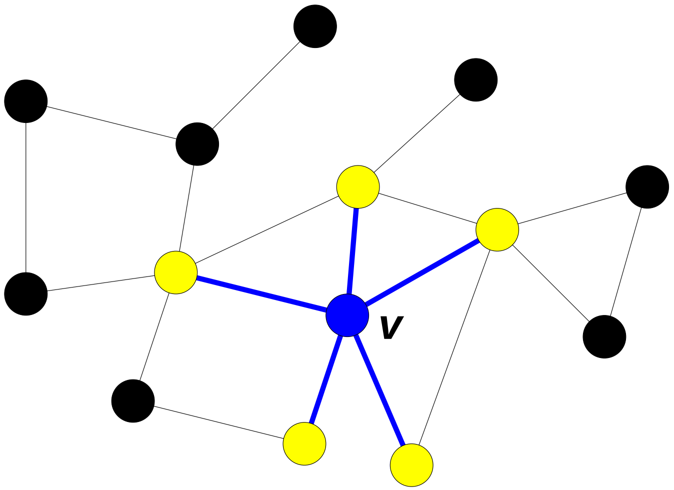


An Example Network

- Application: Given simply information of who associates with whom, could one identify clusters of individuals with common interests or special relationships (families, cliques, terrorist cells)?

A Social Network Model

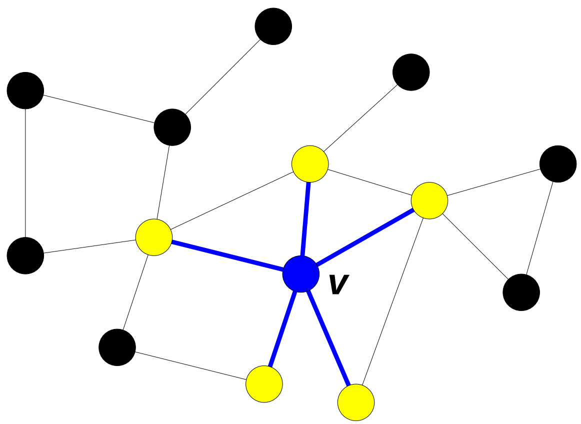
- Cliques, hubs and outliers
  - Individuals in a tight social group, or clique, know many of the same people, regardless of the size of the group
  - Individuals who are hubs know many people in different groups but belong to no single group. Politicians, for example bridge multiple groups
  - Individuals who are outliers reside at the margins of society. Hermits, for example, know few people and belong to no group
- The Neighborhood of a Vertex
  - Define $\Gamma(\nu)$ as the immediate neighborhood of a vertex (i.e. the set of people that an individual knows)



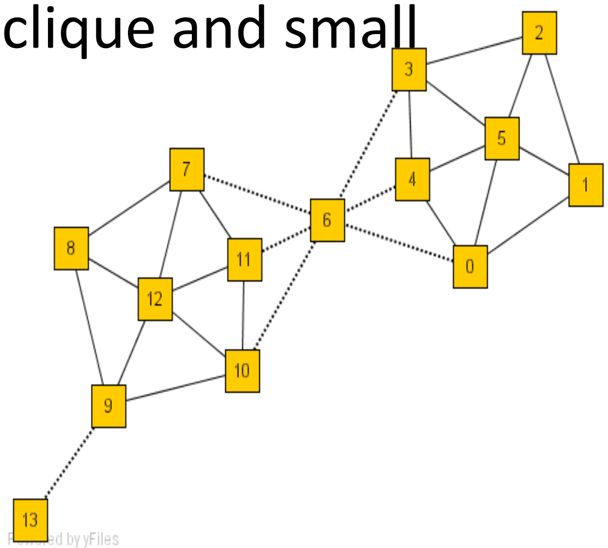
Structure Similarity

- The desired features tend to be captured by a measure we call Structural Similarity

$$\sigma(v,w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)||\Gamma(w)|}}$$



- Structural similarity is large for members of a clique and small for hubs and outliers



Structural Connectivity [1]

- ε-Neighborhood: $N_{\varepsilon}(v) = \{w \in \Gamma(v) \mid \sigma(v,w) \geq \varepsilon\}$
- Core: $CORE_{\varepsilon,\mu}(v) \Leftrightarrow |N_{\varepsilon}(v)| \geq \mu$
- Direct structure reachable:

$$DirRECH_{\varepsilon,\mu}(v,w) \Leftrightarrow CORE_{\varepsilon,\mu}(v) \wedge w \in N_{\varepsilon}(v)$$

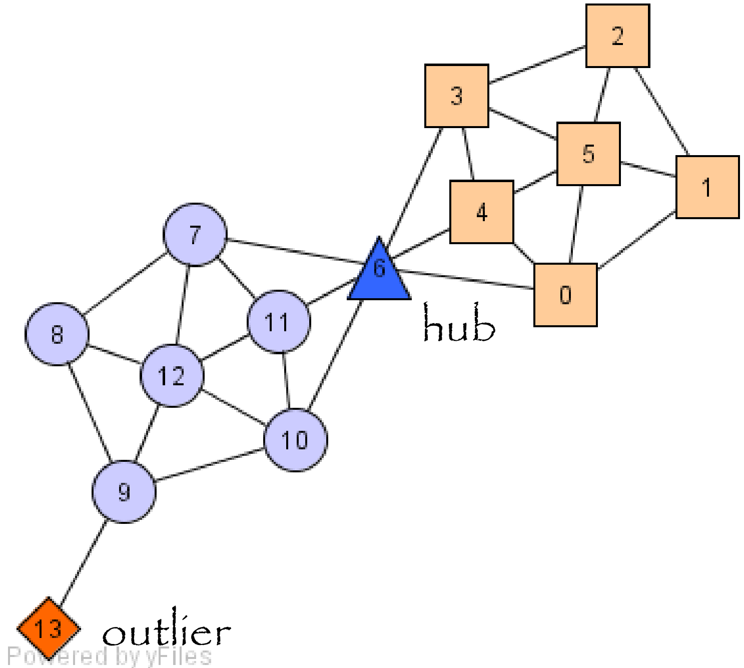
- Structure reachable: transitive closure of direct structure reachability
- Structure connected:

$$CONNECT_{\varepsilon,\mu}(v,w) \Leftrightarrow \exists u \in V : RECH_{\varepsilon,\mu}(u,v) \wedge RECH_{\varepsilon,\mu}(u,w)$$

[1] M. Ester, H. P. Kriegel, J. Sander, & X. Xu (KDD'96) “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases

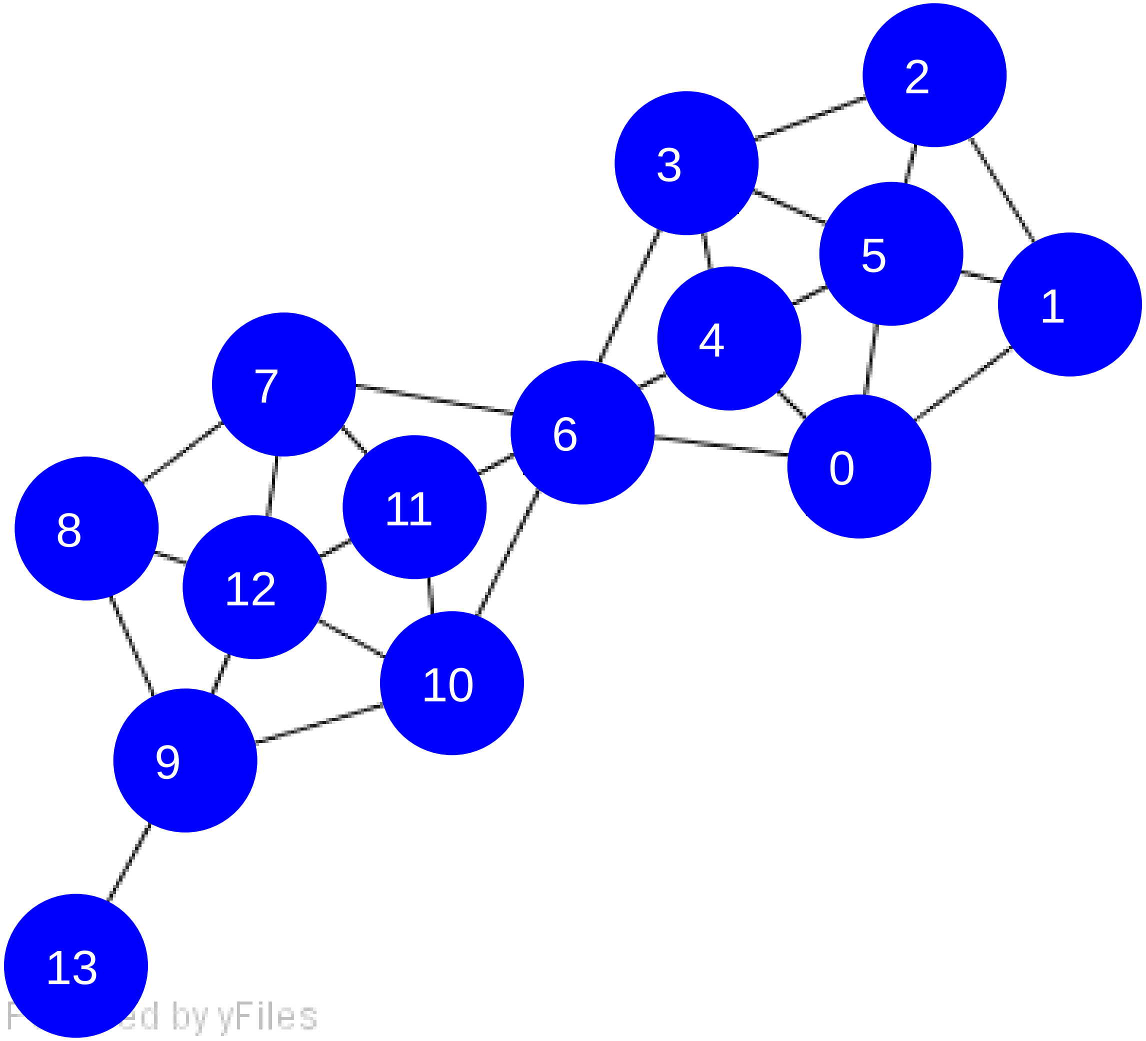
Structure-Connected Clusters

- Structure-connected cluster C
  - Connectivity: $\forall v, w \in C : CONNECT_{\varepsilon,\mu}(v,w)$
  - Maximality: $\forall v, w \in V : v \in C \wedge REACH_{\varepsilon,\mu}(v,w) \Rightarrow w \in C$
- Hubs:
  - Not belong to any cluster
  - Bridge to many clusters
- Outliers:
  - Not belong to any cluster
  - Connect to less clusters



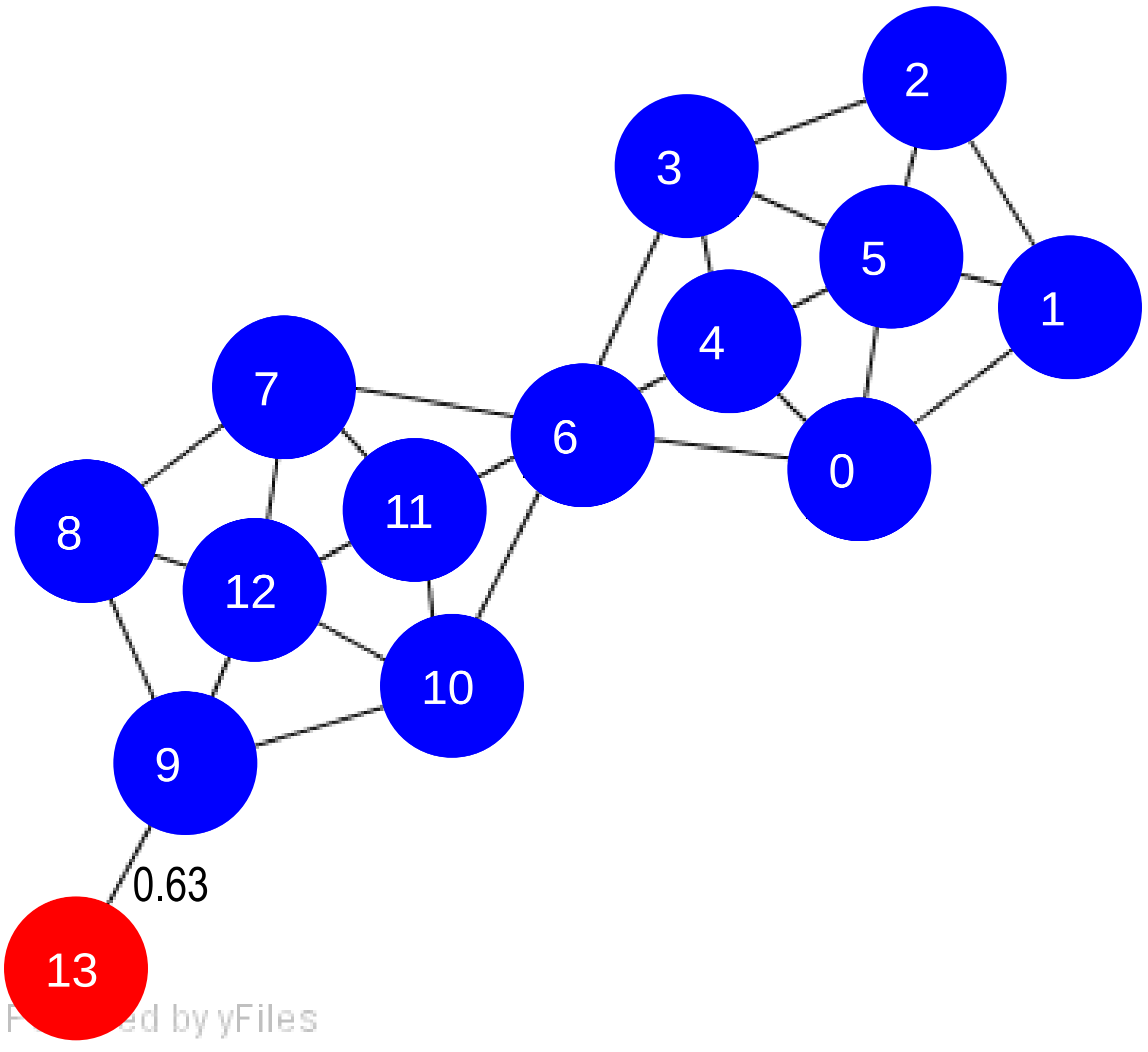
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



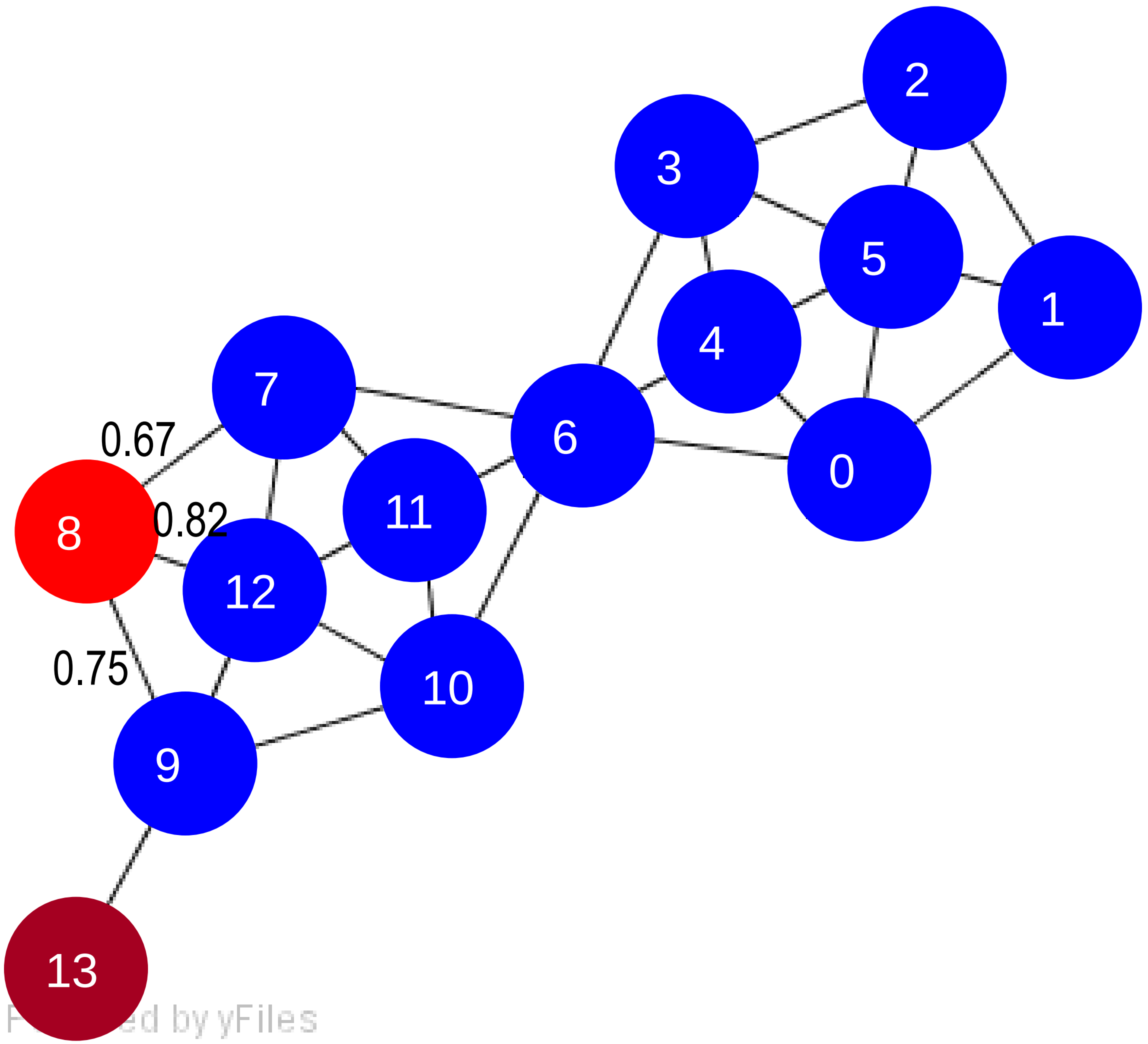
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



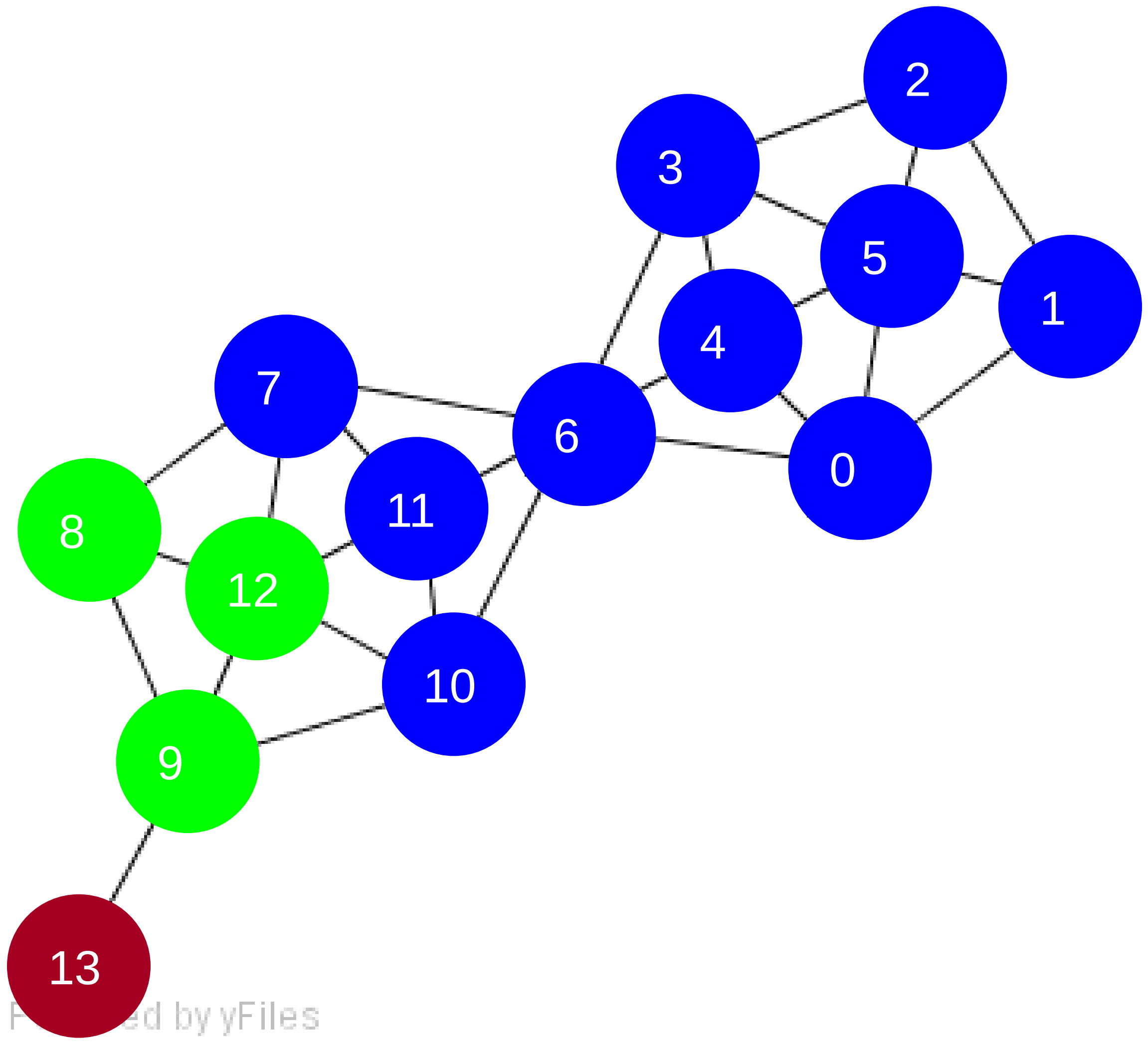
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



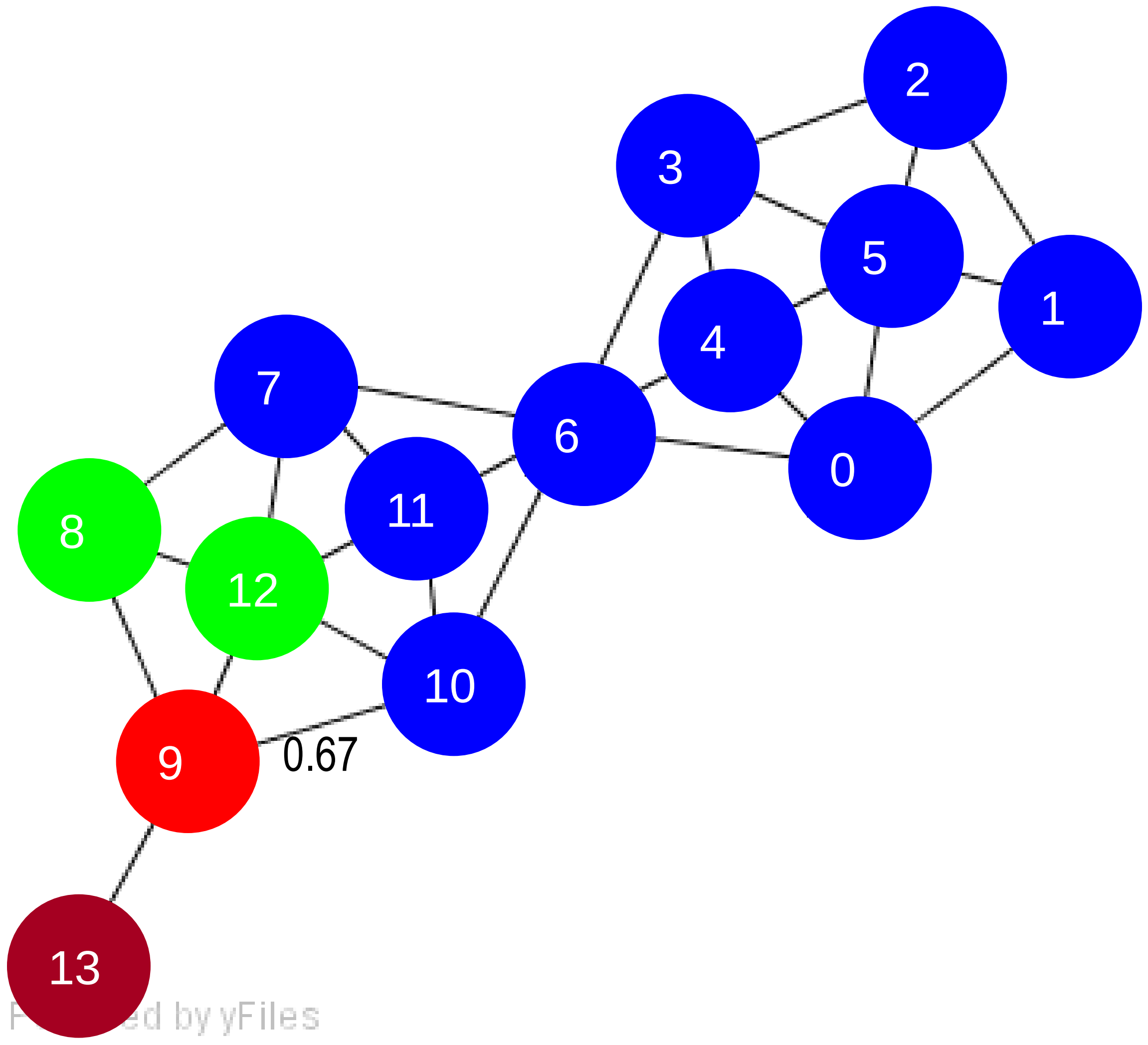
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



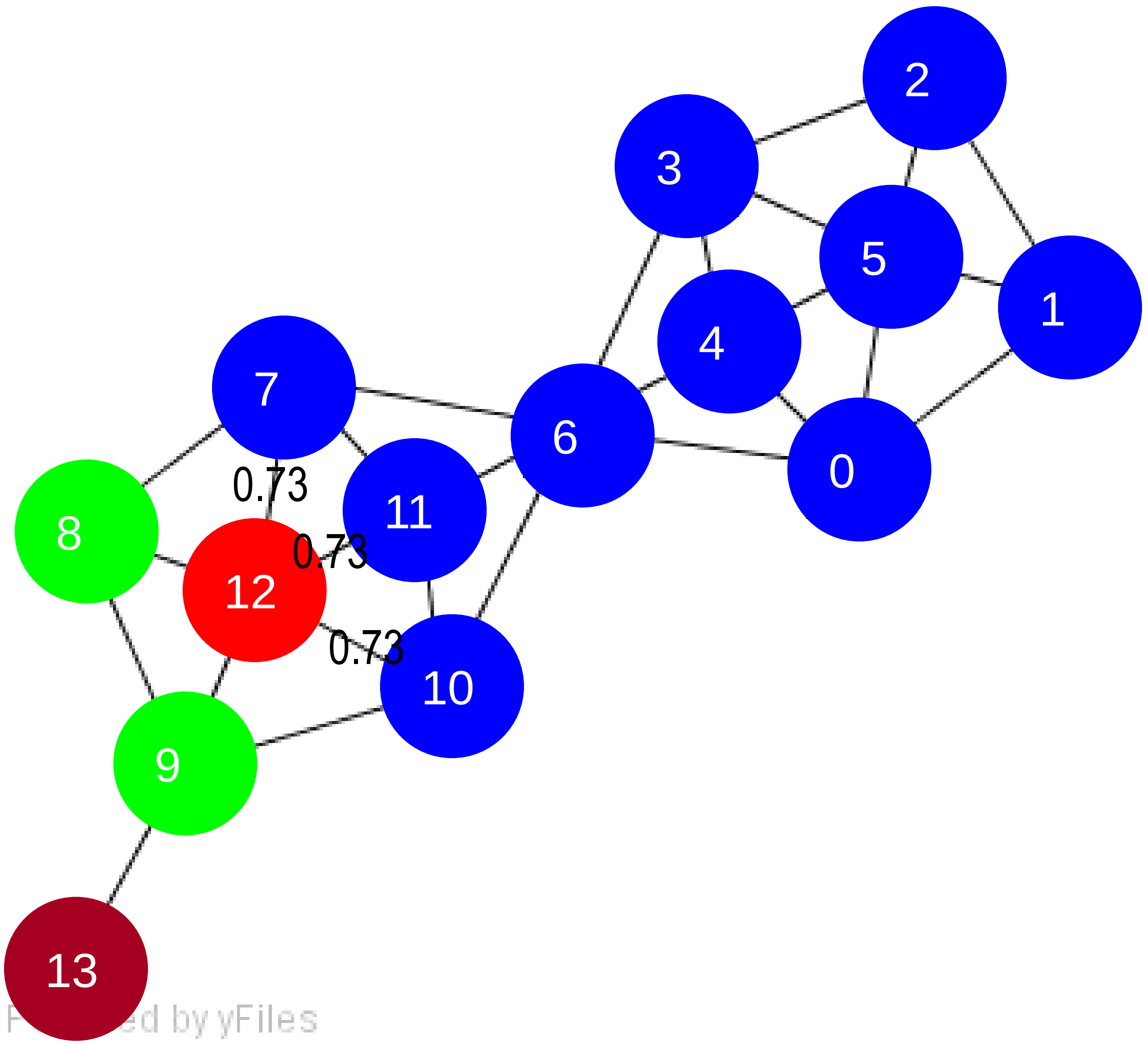
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



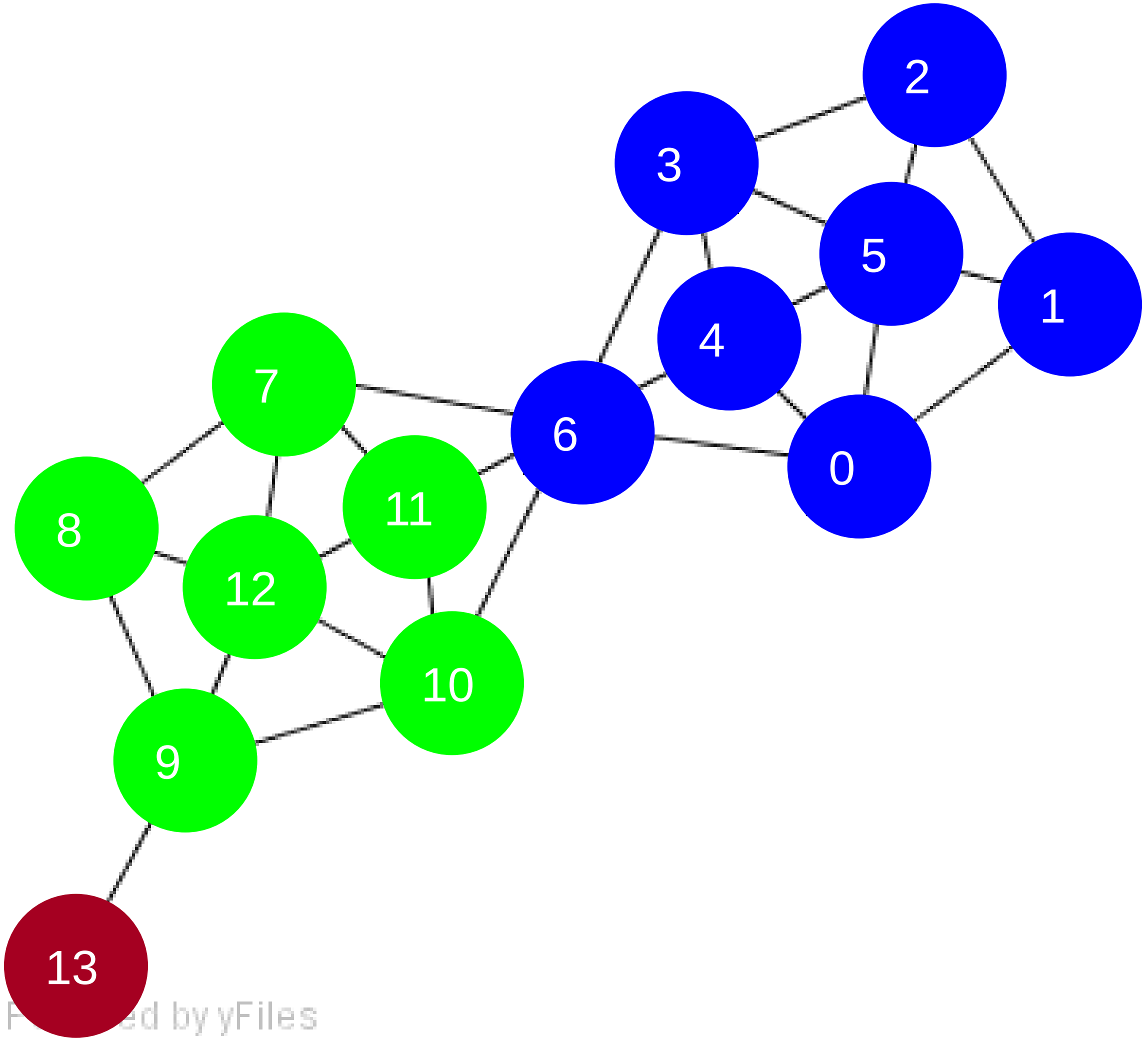
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



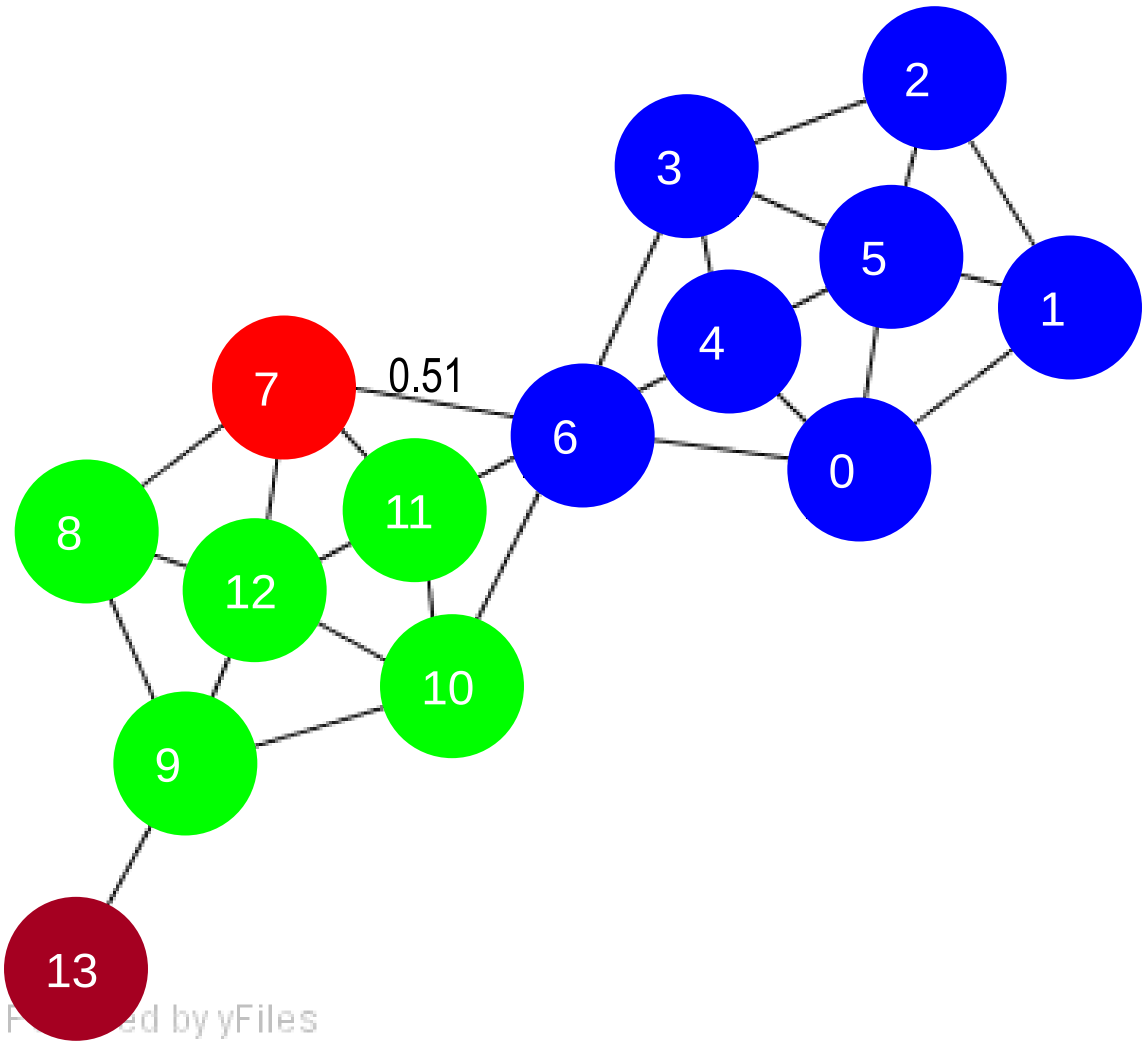
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



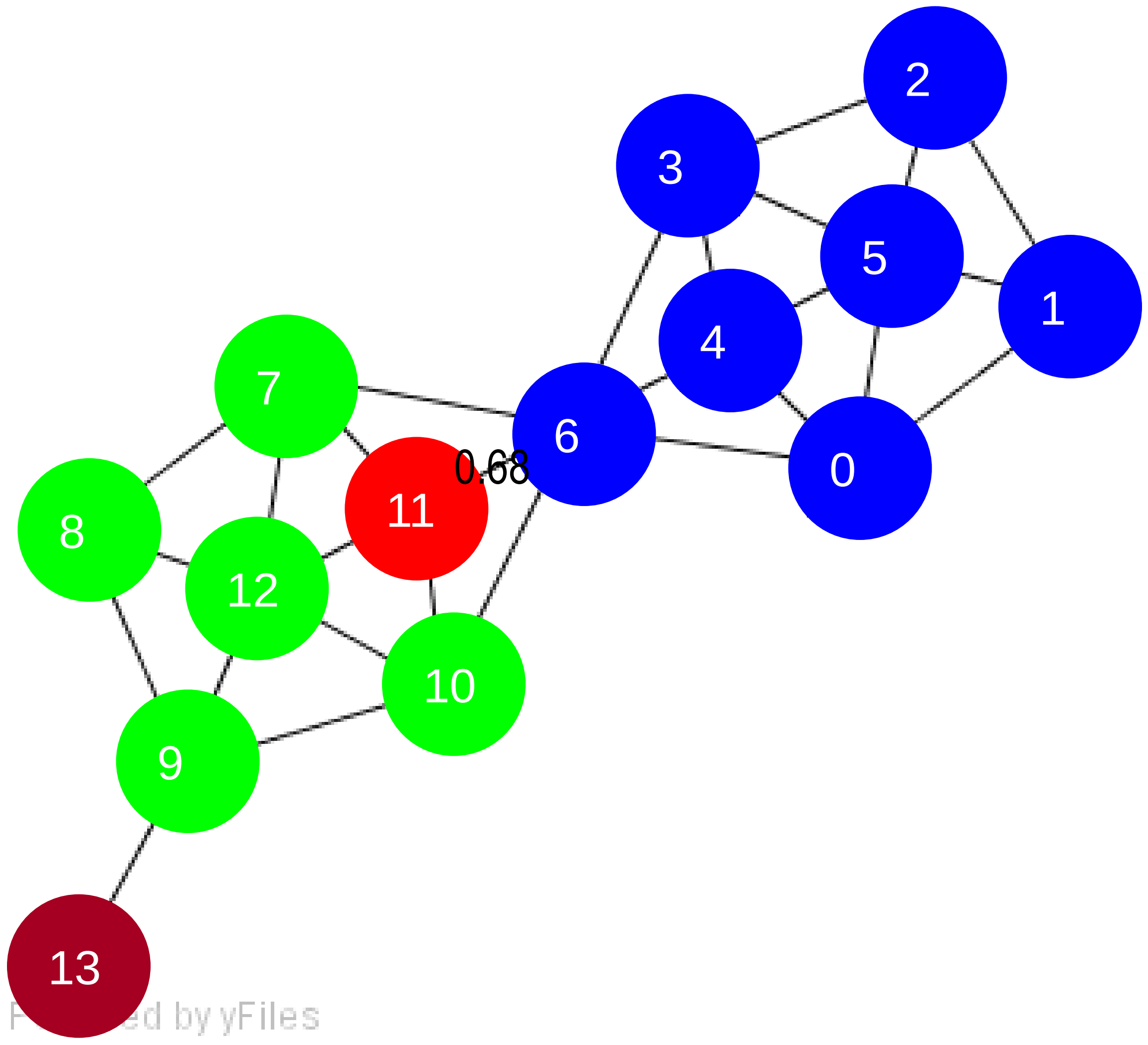
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



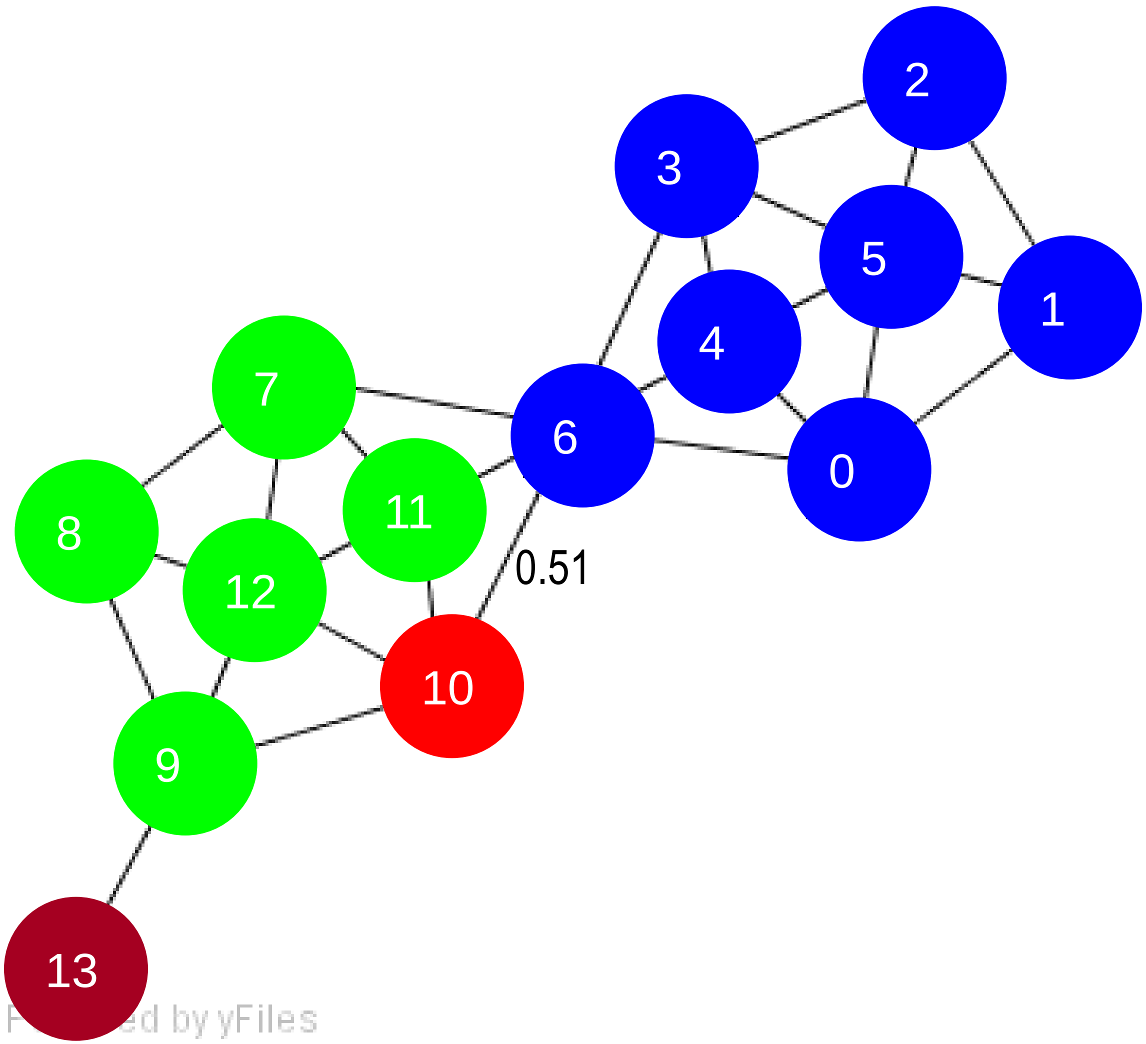
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



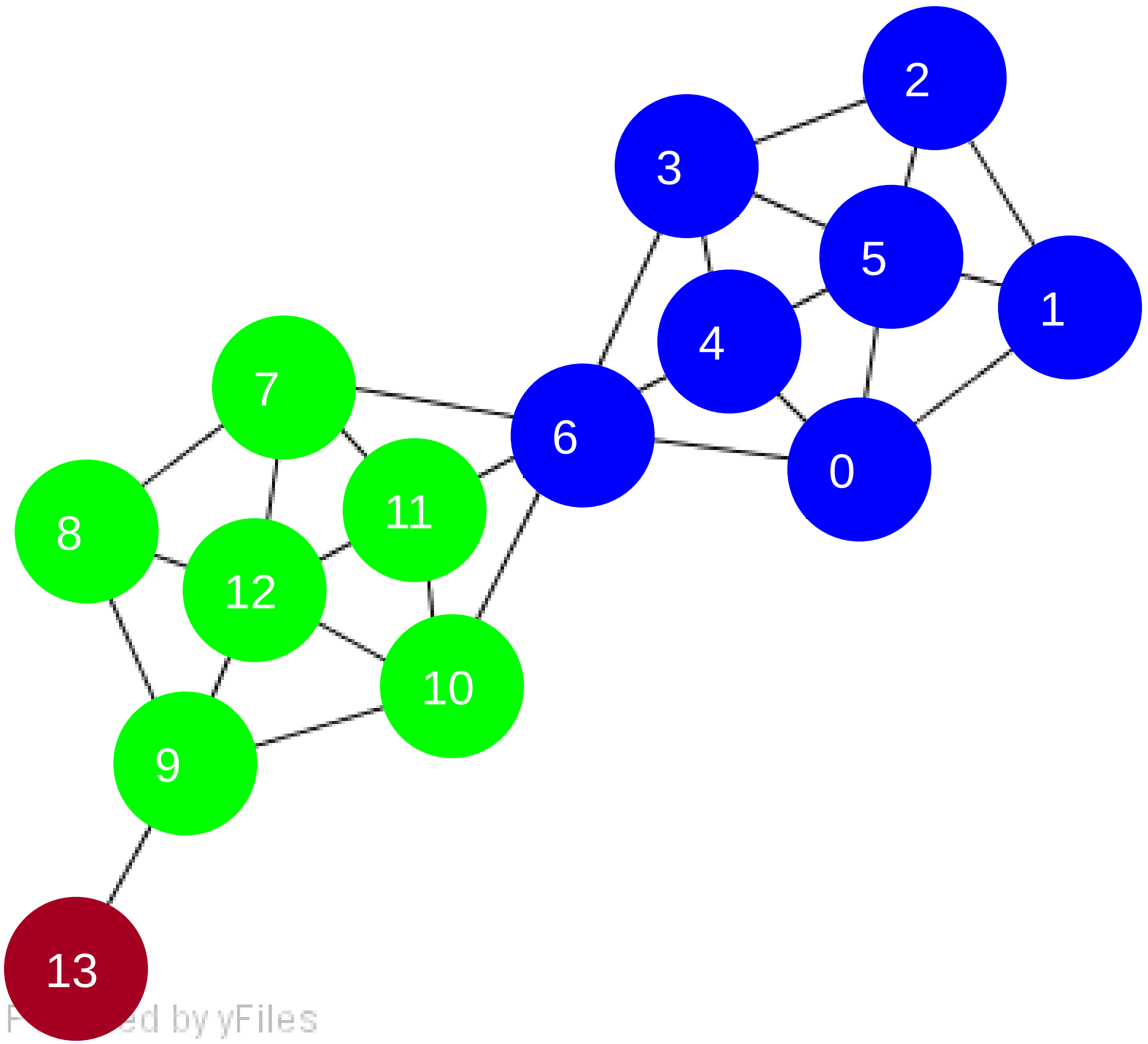
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



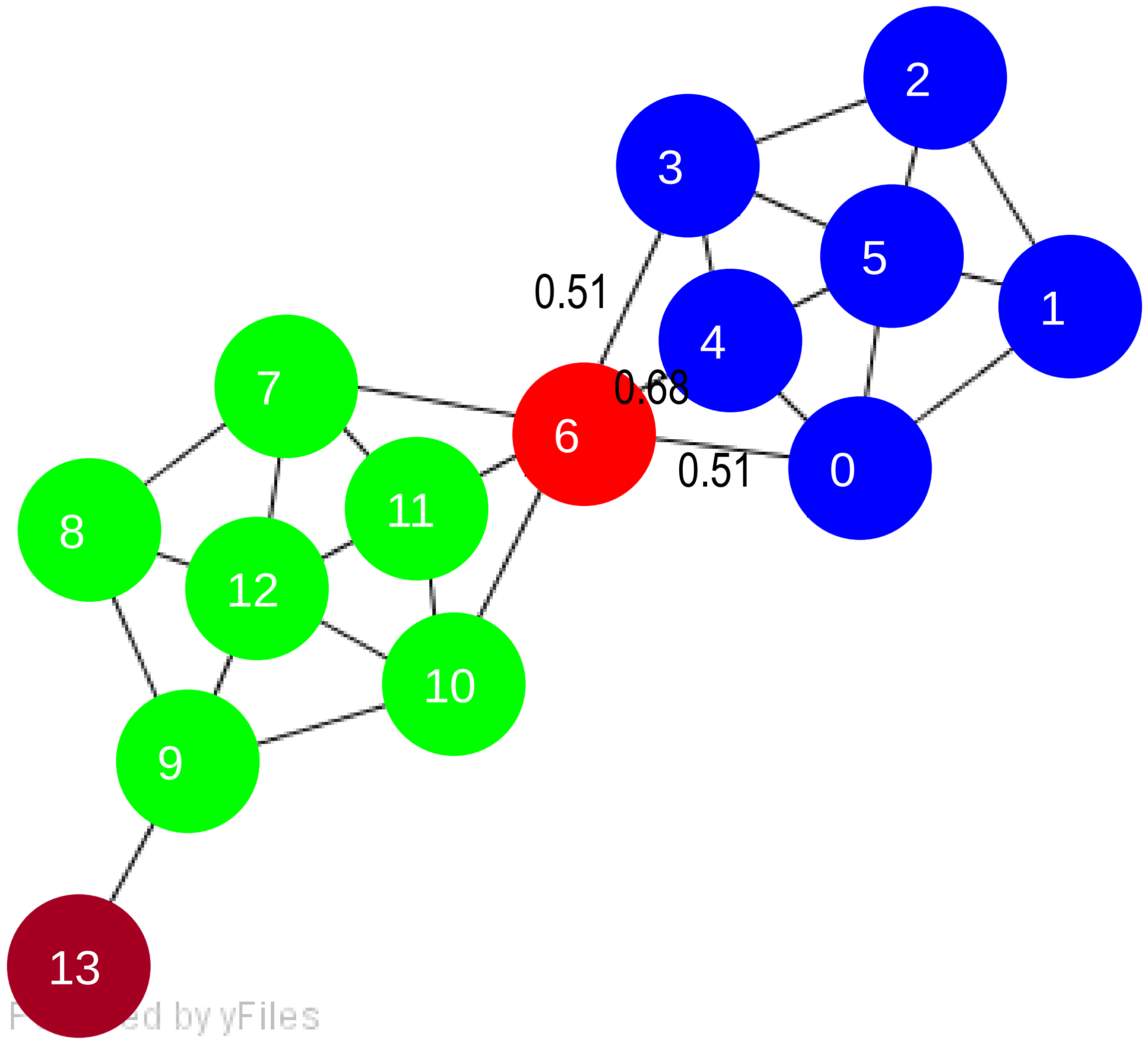
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



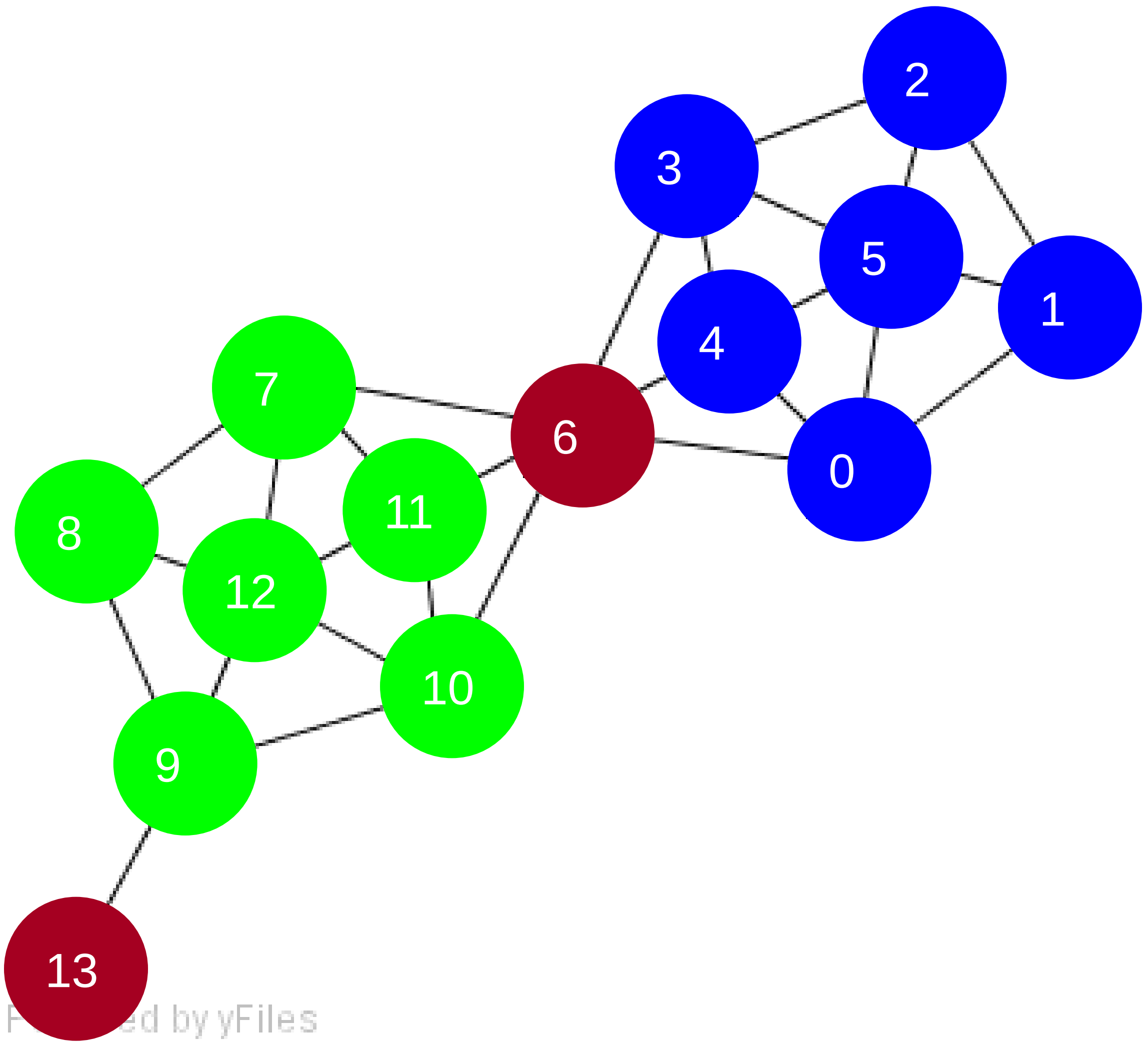
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



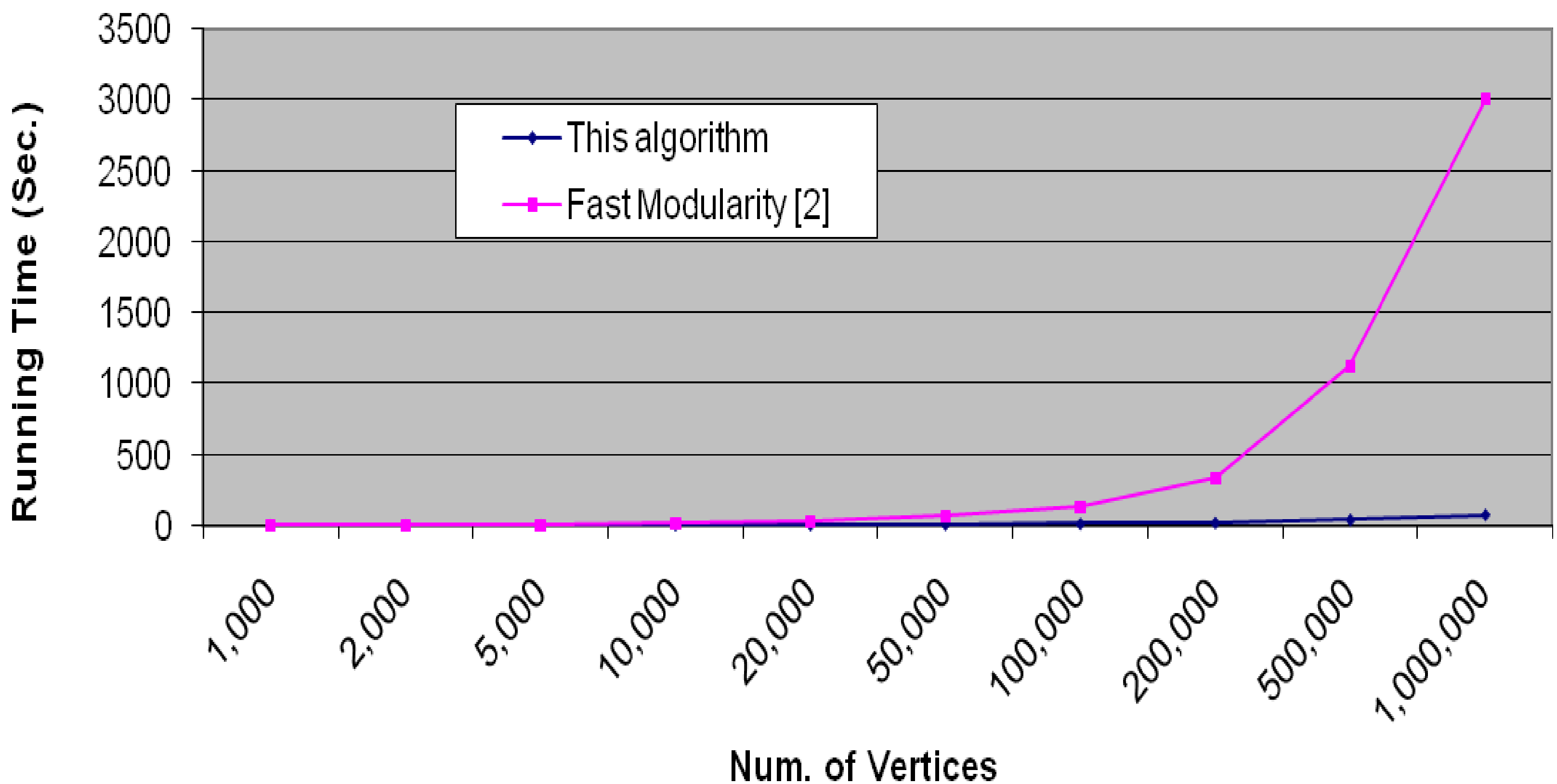
Algorithm

$\mu = 2$
$\varepsilon = 0.7$



Running Time

- Running time = $O(|E|)$
- For sparse networks = $O(|V|)$



[2] A. Clauset, M. E. J. Newman, & C. Moore, *Phys. Rev. E* **70**, 066111 (2004).

Spectral Clustering

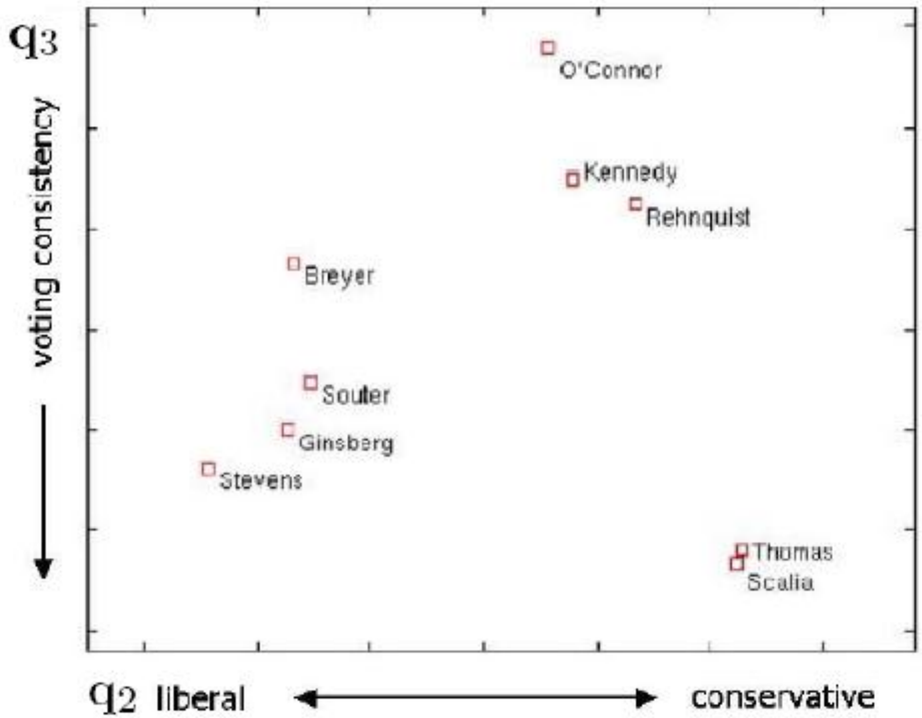
- Reference: ICDM'09 Tutorial by Chris Ding
- Example:
  - Clustering supreme court justices according to their voting behavior

Number of times (%) two Justices voted in agreement

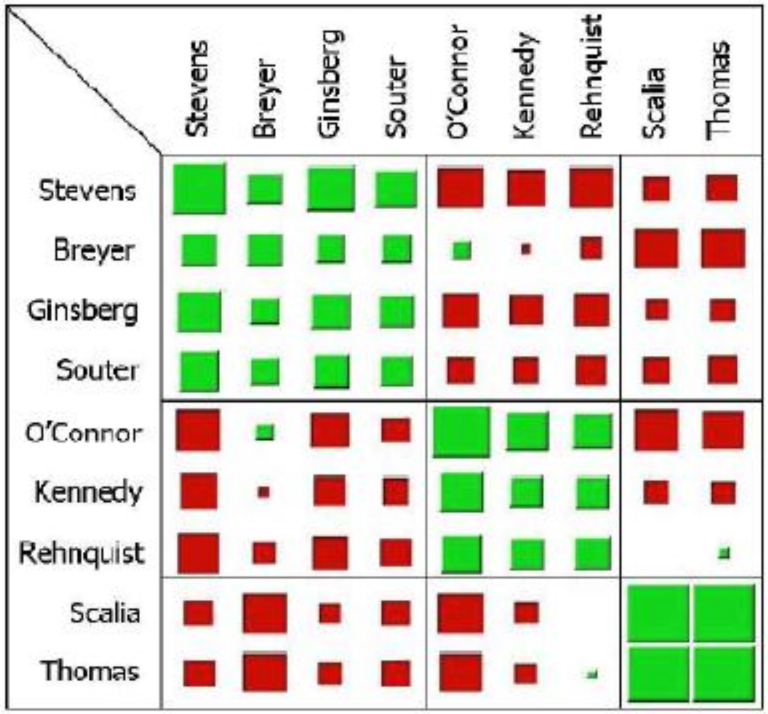
	Ste	Bre	Gin	Sou	O'Co	Ken	Reh	Sca	Tho
Stevens	–	62	66	63	33	36	25	14	15
Breyer	62	–	72	71	55	47	43	25	24
Ginsberg	66	72	–	78	47	49	43	28	26
Souter	63	71	78	–	55	50	44	31	29
O'Connor	33	55	47	55	–	67	71	54	54
Kennedy	36	47	49	50	67	–	77	58	59
Rehnquist	25	43	43	44	71	77	–	66	68
Scalia	14	25	28	31	54	58	66	–	79
Thomas	15	24	26	29	54	59	68	79	–

Table 1: From the voting record of Justices 1995 Term – 2004 Term, the number of times two justices voted in agreement (in percentage). (Data source: from July 2, 2005 *New York Times*. Originally from *Legal Affairs*; *Harvard Law Review*)

Example: Continue



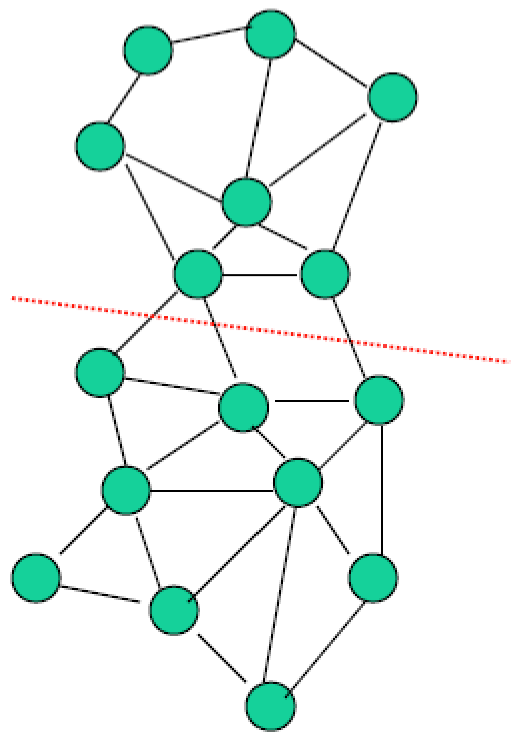
$$C = \mathbf{q}_2\mathbf{q}_2^T + \mathbf{q}_3\mathbf{q}_3^T$$



- Three groups in the Supreme Court:
  - Left leaning group, center-right group, right leaning group.

Spectral Graph Partition

- Min-Cut
  - Minimize the # of cut of edges



Objective Function

2-way Spectral Graph Partitioning

Partition membership indicator: $q_i = \begin{cases} 1 & \textbf{if } i \in A \\ -1 & \textbf{if } i \in B \end{cases}$

$$J = CutSize = \frac{1}{4}\sum_{i,j} w_{ij}[q_i - q_j]^2$$

$$= \frac{1}{4}\sum_{i,j} w_{ij}[q_i^2 + q_j^2 - 2q_i q_j] = \frac{1}{2}\sum_{i,j} q_i[d_i\delta_{ij} - w_{ij}]q_j$$

$$= \frac{1}{2} q^T (D - W) q$$

Relax indicators q_i from discrete values to continuous values, the solution for min $J(q)$ is given by the eigenvectors of

$$(D - W)q = \lambda q$$

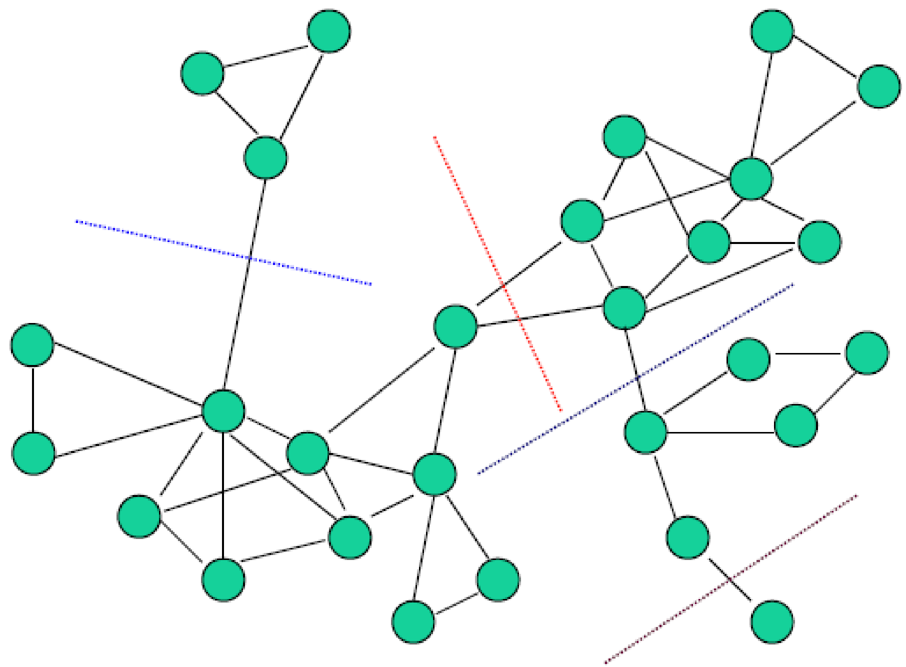
(Fiedler, 1973, 1975)

(Pothen, Simon, Liou, 1990)

Minimum Cut with Constraints

minimize cutsize without explicit size constraints

But where to cut ?



Need to balance sizes

New Objective Functions

$$s(A,B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$$

- Ratio Cut (Hangen & Kahng, 1992)

$$J_{Rcut}(A,B) = \frac{s(A,B)}{|A|} + \frac{s(A,B)}{|B|}$$

- Normalized Cut (Shi & Malik, 2000)

$$d_A = \sum_{i \in A} d_i$$

$$J_{Ncut}(A,B) = \frac{s(A,B)}{d_A} + \frac{s(A,B)}{d_B}$$

$$= \frac{s(A,B)}{s(A,A) + s(A,B)} + \frac{s(A,B)}{s(B,B) + s(A,B)}$$

- Min-Max-Cut (Ding et al, 2001)

$$J_{MMC}(A,B) = \frac{s(A,B)}{s(A,A)} + \frac{s(A,B)}{s(B,B)}$$

Other References

- A Tutorial on Spectral Clustering by U. Luxburg http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/Luxburg07_tutorial_4488%5B0%5D.pdf

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Beyond K-Means
  - K-means
  - EM-algorithm
  - Kernel K-means
- Clustering Graphs and Network Data
- Summary

Summary

- Generalizing K-Means
  - Mixture Model; EM-Algorithm; Kernel K-Means
- Clustering Graph and Networked Data
  - SCAN: density-based algorithm
  - Spectral clustering

Announcement

- HW #3 due tomorrow
- Course project due next week
  - Submit final report, data, code (with readme), evaluation forms
  - Make appointment with me to explain your project
    - I will ask questions according to your report
- Final Exam
  - 4/22, 3 hours in class, cover the whole semester with different weights
  - You can bring two A4 cheating sheets, one for content before midterm, and the other for content after midterm
- Interested in research?
  - My research area: Information/social network mining