



A Tutorial on Spectral Clustering

Chris Ding

Computational Research Division
Lawrence Berkeley National Laboratory
University of California

Supported by Office of Science, U.S. Dept. of Energy



Some historical notes

- Fiedler, 1973, 1975, graph Laplacian matrix
- Donath & Hoffman, 1973, bounds
- Pothen, Simon, Liou, 1990, Spectral graph partitioning (many related papers there after)
- Hagen & Kahng, 1992, Ratio-cut
- Chan, Schlag & Zien, multi-way Ratio-cut
- Chung, 1997, Spectral graph theory book
- Shi & Malik, 2000, Normalized Cut



Spectral Gold-Rush of 2001

9 papers on spectral clustering

- Meila & Shi, *AI-Stat 2001*. Random Walk interpretation of Normalized Cut
- Ding, He & Zha, *KDD 2001*. Perturbation analysis of Laplacian matrix on sparsely connected graphs
- Ng, Jordan & Weiss, *NIPS 2001*, K-means algorithm on the embedded eigen-space
- Belkin & Niyogi, *NIPS 2001*. Spectral Embedding
- Dhillon, *KDD 2001*, Bipartite graph clustering
- Zha et al, *CIKM 2001*, Bipartite graph clustering
- Zha et al, *NIPS 2001*. Spectral Relaxation of K-means
- Ding et al, *ICDM 2001*. MinMaxCut, Uniqueness of relaxation.
- Gu et al, K-way Relaxation of NormCut and MinMaxCut



Part I: Basic Theory, 1973 – 2001

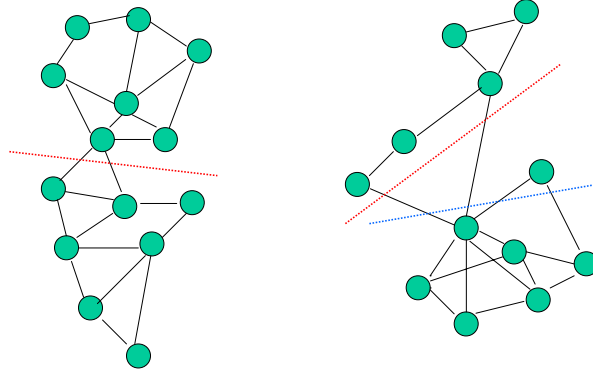


Spectral Graph Partitioning

MinCut: **min** cutsize

cutsize = # of cut edges

Constraint on sizes: $|A| = |B|$



Tutorial on Spectral Clustering, ICML 2004, Chris Ding © University of California

5



2-way Spectral Graph Partitioning

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

$$\begin{aligned} J = \text{CutSize} &= \frac{1}{4} \sum_{i,j} w_{ij} [q_i - q_j]^2 \\ &= \frac{1}{4} \sum_{i,j} w_{ij} [q_i^2 + q_j^2 - 2q_i q_j] = \frac{1}{2} \sum_{i,j} q_i [d_i \delta_{ij} - w_{ij}] q_j \\ &= \frac{1}{2} q^T (D - W) q \end{aligned}$$

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)

Tutorial on Spectral Clustering, ICML 2004, Chris Ding © University of California

6



Properties of Graph Laplacian

Laplacian matrix of the Graph: $L = D - W$

- L is semi-positive definite $x^T L x \geq 0$ for any x .
- First eigenvector is $q_1 = (1, \dots, 1)^T = e^T$ with $\lambda_1 = 0$.
- Second eigenvector q_2 is the desired solution.
 - The smaller λ_2 , the better quality of the partitioning. Perturbation analysis gives

$$\lambda_2 = \frac{\text{cutsizesize}}{|A|} + \frac{\text{cutsizesize}}{|B|}$$

- Higher eigenvectors are also useful

Tutorial on Spectral Clustering, ICML 2004, Chris Ding © University of California

7



Recovering Partitions

From the definition of cluster indicators:
Partitions A, B are determined by:

$$A = \{i \mid q_2(i) < 0\}, B = \{i \mid q_2(i) \geq 0\}$$

However, the objective function $J(q)$ is
insensitive to additive constant c :

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

Thus, we sort q_2 to increasing order, and cut in the
middle point.

Tutorial on Spectral Clustering, ICML 2004, Chris Ding © University of California

8



Multi-way Graph Partitioning

- Recursively applying the 2-way partitioning
 - Recursive 2-way partitioning
 - Using Kernigan-Lin to do local refinements
- Using higher eigenvectors
 - Using q_3 to further partitioning those obtained via q_2 .
- Popular graph partitioning packages
 - Metis, Univ of Minnesota
 - Chaco, Sandia Nat'l Lab



2-way Spectral Clustering

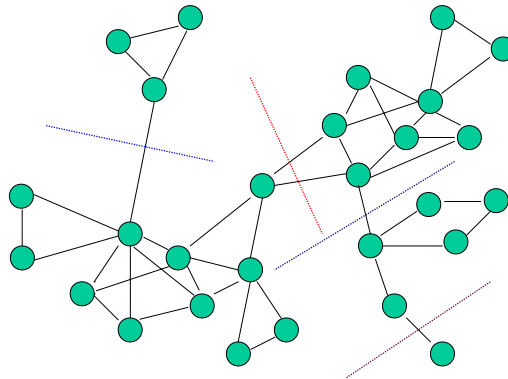
- Undirected graphs (pairwise similarities)
- Bipartite graphs (contingency tables)
- Directed graphs (web graphs)



Spectral Clustering

min cutsize , without explicit size constraints

But where to cut ?



Need to balance sizes



Clustering Objective Functions

- Ratio Cut

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

- Normalized Cut

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

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

- Min-Max-Cut

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

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

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



Ratio Cut (Hagen & Kahng, 1992)

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

Size Balance $J_{Rcut}(A,B) = \frac{s(A,B)}{|A|} + \frac{s(A,B)}{|B|}$ (Wei & Cheng, 1989)

Cluster membership indicator: $q(i) = \begin{cases} \sqrt{n_2 / n_1 n} & \text{if } i \in A \\ -\sqrt{n_1 / n_2 n} & \text{if } i \in B \end{cases}$

Normalization: $q^T q = 1, q^T e = 0$

Substitute q leads to $J_{Rcut}(q) = q^T (D - W) q$

Now relax q , the solution is 2nd eigenvector of L



Normalized Cut (Shi & Malik, 1997)

Min similarity between A & B: $s(A,B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$

Balance weights $J_{Ncut}(A,B) = \frac{s(A,B)}{d_A} + \frac{s(A,B)}{d_B}$ $d_A = \sum_{i \in A} d_i$

Cluster indicator: $q(i) = \begin{cases} \sqrt{d_B / d_A d} & \text{if } i \in A \\ -\sqrt{d_A / d_B d} & \text{if } i \in B \end{cases}$ $d = \sum_{i \in G} d_i$

Normalization: $q^T D q = 1, q^T D e = 0$

Substitute q leads to $J_{Ncut}(q) = q^T (D - W) q$

$\min_q q^T (D - W) q + \lambda (q^T D q - 1)$

Solution is eigenvector of $(D - W) q = \lambda D q$



MinMaxCut (Ding et al 2001)

Min similarity between A & B: $s(A,B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$

Max similarity within A & B: $s(A,A) = \sum_{i \in A} \sum_{j \in A} w_{ij}$

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

Cluster indicator: $q(i) = \begin{cases} \sqrt{d_B/d_A} & \text{if } i \in A \\ -\sqrt{d_A/d_B} & \text{if } i \in B \end{cases}$

Substituting,

$$J_{MMC}(q) = \frac{1 + \sqrt{d_B/d_A}}{J_m + \sqrt{d_B/d_A}} + \frac{1 + \sqrt{d_A/d_B}}{J_m + \sqrt{d_A/d_B}} - 2 \quad J_m = \frac{q^T W q}{q^T D q}$$

Because $\frac{dJ_{MMC}(J_m)}{dJ_m} < 0 \quad \min J_{mmc} \Rightarrow \max J_m(q)$

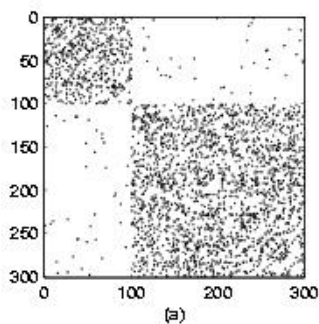
$$\Rightarrow Wq = \xi Dq \Rightarrow (D - W)q = \lambda Dq$$



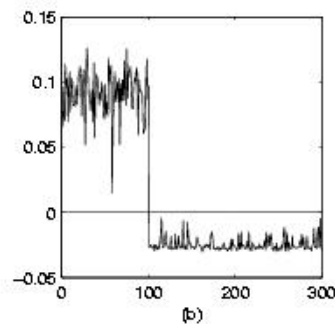
A simple example

2 dense clusters, with sparse connections between them.

Adjacency matrix



Eigenvector q_2





Comparison of Clustering Objectives

- If clusters are **well separated**, all three give **very similar** and **accurate** results.
- When clusters are **marginally separated**, NormCut and MinMaxCut give better results
- When clusters **overlap significantly**, MinMaxCut tend to give more **compact** and **balanced** clusters.

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

Cluster Compactness $\Rightarrow \max s(A, A)$



2-way Clustering of Newsgroups

Newsgroups	RatioCut	NormCut	MinMaxCut
Atheism	63.2 ± 16.2	97.2 ± 0.8	97.2 ± 1.1
Comp.graphics			
Baseball	54.9 ± 2.5	74.4 ± 20.4	79.5 ± 11.0
Hockey			
Politics.mideast	53.6 ± 3.1	57.5 ± 0.9	83.6 ± 2.5
Politics.misc			



Cluster Balance Analysis I: Random Graph Model

- Random graph: edges are randomly assigned with probability p : $0 \leq p \leq 1$.
- RatioCut & NormCut show no size dependence

$$J_{\text{Rcut}}(A, B) = \frac{p |A| |B|}{|A|} + \frac{p |A| |B|}{|B|} = np = \text{constant}$$

$$J_{\text{Ncut}}(A, B) = \frac{p |A| |B|}{p |A| (n-1)} + \frac{p |A| |B|}{p |B| (n-1)} = \frac{n}{n-1} = \text{constant}$$

- MinMaxCut favors balanced clusters: $|A|=|B|$

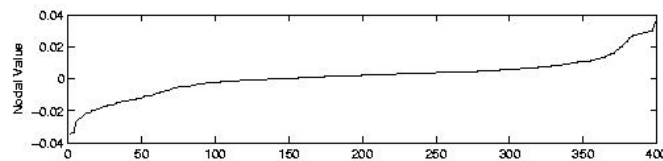
$$J_{\text{MMC}}(A, B) = \frac{p |A| |B|}{p |A| (|A|-1)} + \frac{p |A| |B|}{p |B| (|B|-1)} = \frac{|B|}{|A|-1} + \frac{|A|}{|B|-1}$$



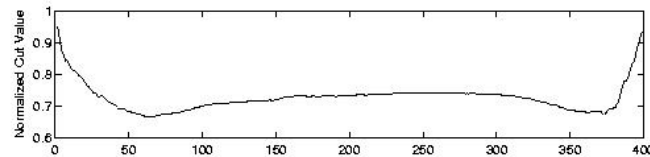
2-way Clustering of Newsgroups

Cluster Balance

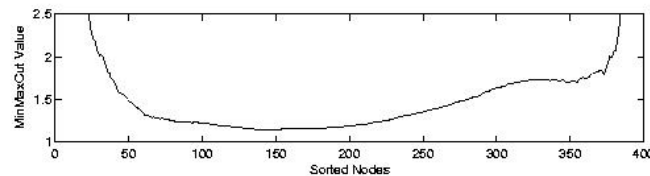
Eigenvector



$J_{\text{Ncut}}(i)$



$J_{\text{MMC}}(i)$





Cluster Balance Analysis II: Large Overlap Case

$$f = \frac{s(A, B)}{(1/2)[s(A, A) + s(B, B)]} > 0.5$$

Conditions for skewed cuts:

$$\text{NormCut} : s(A, A) \geq \left(\frac{1}{2f} - \frac{1}{2}\right)s(A, B) = s(A, B) / 2$$

$$\text{MinMaxCut} : s(A, A) \geq \frac{1}{2f} s(A, B) = s(A, B)$$

Thus MinMaxCut is much less prone to skewed cuts



Spectral Clustering of Bipartite Graphs

Simultaneous clustering of rows and columns
of a contingency table (adjacency matrix B)

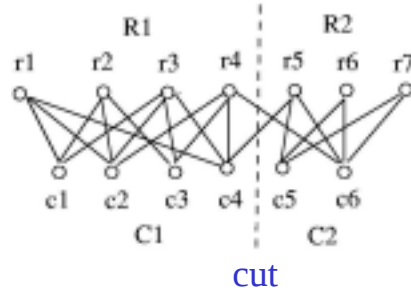
Examples of bipartite graphs

- Information Retrieval: word-by-document matrix
- Market basket data: transaction-by-item matrix
- DNA Gene expression profiles
- Protein vs protein-complex



Spectral Clustering of Bipartite Graphs

Simultaneous clustering of **rows** and **columns**
(adjacency matrix B)



$$s(B_{R_1, C_2}) = \sum_{r_i \in R_1} \sum_{c_j \in C_2} b_{ij}$$

min **between-cluster** sum of weights: $s(R_1, C_2), s(R_2, C_1)$

max **within-cluster** sum of weights: $s(R_1, C_1), s(R_2, C_2)$

$$J_{MMC}(C_1, C_2; R_1, R_2) = \frac{s(B_{R_1, C_2}) + s(B_{R_2, C_1})}{2s(B_{R_1, C_1})} + \frac{s(B_{R_1, C_2}) + s(B_{R_2, C_1})}{2s(B_{R_2, C_2})}$$

(Ding, AI-STAT 2003)

Tutorial on Spectral Clustering, ICML 2004, Chris Ding © University of California

23



Bipartite Graph Clustering

Clustering indicators for rows and columns:

$$f(i) = \begin{cases} 1 & \text{if } r_i \in R_1 \\ -1 & \text{if } r_i \in R_2 \end{cases} \quad g(i) = \begin{cases} 1 & \text{if } c_i \in C_1 \\ -1 & \text{if } c_i \in C_2 \end{cases}$$

$$B = \begin{pmatrix} B_{R_1, C_1} & B_{R_1, C_2} \\ B_{R_2, C_1} & B_{R_2, C_2} \end{pmatrix} \quad W = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \quad \mathbf{q} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}$$

Substitute and obtain

$$J_{MMC}(C_1, C_2; R_1, R_2) = \frac{s(W_{12})}{s(W_{11})} + \frac{s(W_{12})}{s(W_{22})}$$

f, g are determined by

$$\left[\begin{pmatrix} D_r & \\ & D_c \end{pmatrix} - \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \right] \begin{pmatrix} f \\ g \end{pmatrix} = \lambda \begin{pmatrix} D_r & \\ & D_c \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix}$$

Tutorial on Spectral Clustering, ICML 2004, Chris Ding © University of California

24



Clustering of Bipartite Graphs

Let

$$\tilde{B} = D_r^{-1/2} B D_c^{-1/2}, z = \begin{pmatrix} u \\ v \end{pmatrix} = Dq = \begin{pmatrix} D_r^{1/2} f \\ D_c^{1/2} g \end{pmatrix}$$

We obtain

$$\begin{pmatrix} 0 & \tilde{B} \\ \tilde{B}^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \lambda \begin{pmatrix} u \\ v \end{pmatrix}$$

Solution is SVD:
$$\tilde{B} = \sum_{k=1}^m u_k \lambda_k v_k^T$$

(Zha et al, 2001, Dhillon, 2001)



Clustering of Bipartite Graphs

Recovering row clusters:

$$R_1 = \{r_i, | f_2(i) < z_r\}, R_2 = \{r_i, | f_2(i) \geq z_r\},$$

Recovering column clusters:

$$C_1 = \{c_i, | g_2(i) < z_c\}, C_2 = \{c_i, | g_2(i) \geq z_c\},$$

$z_r = z_c = 0$ are dividing points. Relaxation is invariant up to a constant shift.

Algorithm: search for optimal points $i_{\text{cut}}, j_{\text{cut}}$, let $z_r = f_2(i_{\text{cut}}), z_c = g_2(j_{\text{cut}})$, such that $J_{\text{MMC}}(C_1, C_2; R_1, R_2)$

is minimized.

(Zha et al, 2001)



Clustering of Directed Graphs

Min directed edge weights between A & B:

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

Max directed edges within A & B:

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

- Equivalent to deal with $\tilde{W} = W + W^T$
- All spectral methods apply to $\tilde{W}$
- For example, web graphs clustered in such way

(He, Ding, Zha, Simon, ICDM 2001)



K -way Spectral Clustering $K \geq 2$



K-way Clustering Objectives

- Ratio Cut

$$J_{\text{Rcut}}(C_1, \dots, C_K) = \sum_{\langle k, l \rangle} \left(\frac{s(C_k, C_l)}{|C_k|} + \frac{s(C_k, C_l)}{|C_l|} \right) = \sum_k \frac{s(C_k, G - C_k)}{|C_k|}$$

- Normalized Cut

$$J_{\text{Ncut}}(C_1, \dots, C_K) = \sum_{\langle k, l \rangle} \left(\frac{s(C_k, C_l)}{d_k} + \frac{s(C_k, C_l)}{d_l} \right) = \sum_k \frac{s(C_k, G - C_k)}{d_k}$$

- Min-Max-Cut

$$J_{\text{MMC}}(C_1, \dots, C_K) = \sum_{\langle k, l \rangle} \left(\frac{s(C_k, C_l)}{s(C_k, C_k)} + \frac{s(C_k, C_l)}{s(C_l, C_l)} \right) = \sum_k \frac{s(C_k, G - C_k)}{s(C_k, C_k)}$$



K-way Spectral Relaxation

- Prove that the solution lie in the subspace spanned by the first k eigenvectors
- Ratio Cut
- Normalized Cut
- Min-Max-Cut



K-way Spectral Relaxation

Unsigned cluster indicators: $h_1 = (1 \cdots 1, 0 \cdots 0, 0 \cdots 0)^T$
 $h_2 = (0 \cdots 0, 1 \cdots 1, 0 \cdots 0)^T$
 $\dots\dots\dots$

Re-write: $h_k = (0 \cdots 0, 0 \cdots 0, 1 \cdots 1)^T$

$$J_{\text{Rcut}}(h_1, \dots, h_k) = \frac{h_1^T (D - W) h_1}{h_1^T h_1} + \dots + \frac{h_k^T (D - W) h_k}{h_k^T h_k}$$

$$J_{\text{Ncut}}(h_1, \dots, h_k) = \frac{h_1^T (D - W) h_1}{h_1^T D h_1} + \dots + \frac{h_k^T (D - W) h_k}{h_k^T D h_k}$$

$$J_{\text{MMC}}(h_1, \dots, h_k) = \frac{h_1^T (D - W) h_1}{h_1^T W h_1} + \dots + \frac{h_k^T (D - W) h_k}{h_k^T W h_k}$$



K-way Ratio Cut Spectral Relaxation

Unsigned cluster indicators: $x_k = (0 \cdots 0, \overbrace{1 \cdots 1}^{n_k}, 0 \cdots 0)^T / n_k^{1/2}$

Re-write: $J_{\text{Rcut}}(x_1, \dots, x_k) = x_1^T (D - W) x_1 + \dots + x_k^T (D - W) x_k$
 $= \text{Tr}(X^T (D - W) X) \quad X = (x_1, \dots, x_k)$

Optimize : $\min_X \text{Tr}(X^T (D - W) X), \text{subject to } X^T X = I$

By K. Fan's theorem, optimal solution is
eigenvectors: $X = (v_1, v_2, \dots, v_k)$, $(D - W)v_k = \lambda_k v_k$

and lower-bound

$$\lambda_1 + \dots + \lambda_k \leq \min J_{\text{Rcut}}(x_1, \dots, x_k)$$

(Chan, Schlag, Zien, 1994)



K-way Normalized Cut Spectral Relaxation

Unsigned cluster indicators:

$$y_k = D^{1/2} (0 \cdots 0, \overbrace{1 \cdots 1}^{n_k}, 0 \cdots 0)^T / \| D^{1/2} h_k \|$$

Re-write:

$$\begin{aligned} J_{\text{Ncut}}(y_1, \dots, y_k) &= y_1^T (I - \tilde{W}) y_1 + \dots + y_k^T (I - \tilde{W}) y_k \\ &= \text{Tr}(Y^T (I - \tilde{W}) Y) \end{aligned} \quad \tilde{W} = D^{-1/2} W D^{-1/2}$$

Optimize : $\min_Y \text{Tr}(Y^T (I - \tilde{W}) Y)$, **subject to** $Y^T Y = I$

By K. Fan's theorem, optimal solution is

eigenvectors: $Y = (v_1, v_2, \dots, v_k)$, $(I - \tilde{W}) v_k = \lambda_k v_k$

$$(D - W) u_k = \lambda_k D u_k, \quad u_k = D^{-1/2} v_k$$

$$\lambda_1 + \dots + \lambda_k \leq \min J_{\text{Ncut}}(y_1, \dots, y_k) \quad (\text{Gu, et al, 2001})$$



K-way Min-Max Cut Spectral Relaxation

Unsigned cluster indicators:

$$y_k = D^{1/2} h_k / \| D^{1/2} h_k \| \quad \tilde{W} = D^{-1/2} W D^{-1/2}$$

Re-write:

$$J_{\text{MMC}}(y_1, \dots, y_k) = \frac{1}{y_1^T \tilde{W} y_1} + \dots + \frac{1}{y_k^T \tilde{W} y_k} - k$$

Optimize : $\min_Y J_{\text{MMC}}(Y)$, **subject to** $Y^T Y = I$, $y_k^T \tilde{W} y_k > 0$.

Theorem. Optimal solution is by eigenvectors:

$$Y = (v_1, v_2, \dots, v_k), \quad \tilde{W} v_k = \lambda_k v_k$$

$$\frac{k^2}{\lambda_1 + \dots + \lambda_k} - k \leq \min J_{\text{MMC}}(y_1, \dots, y_k) \quad (\text{Gu, et al, 2001})$$



K-way Spectral Clustering

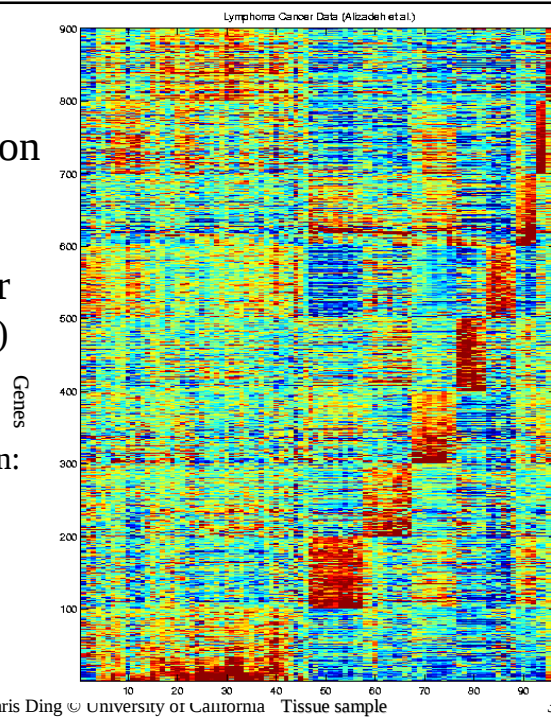
- Embedding (similar to PCA subspace approach)
 - Embed data points in the subspace of the K eigenvectors
 - Clustering embedded points using another algorithm, such as K -means (Shi & Malik, Ng et al, Zha, et al)
- Recursive 2-way clustering (standard graph partitioning)
 - If desired K is not power of 2, how optimally to choose the next sub-cluster to split? (Ding, et al 2002)
- Both above approach **do not use** K -way clustering objective functions.
- Refine the obtained clusters using the K -way clustering objective function typically improve the results (Ding et al 2002).

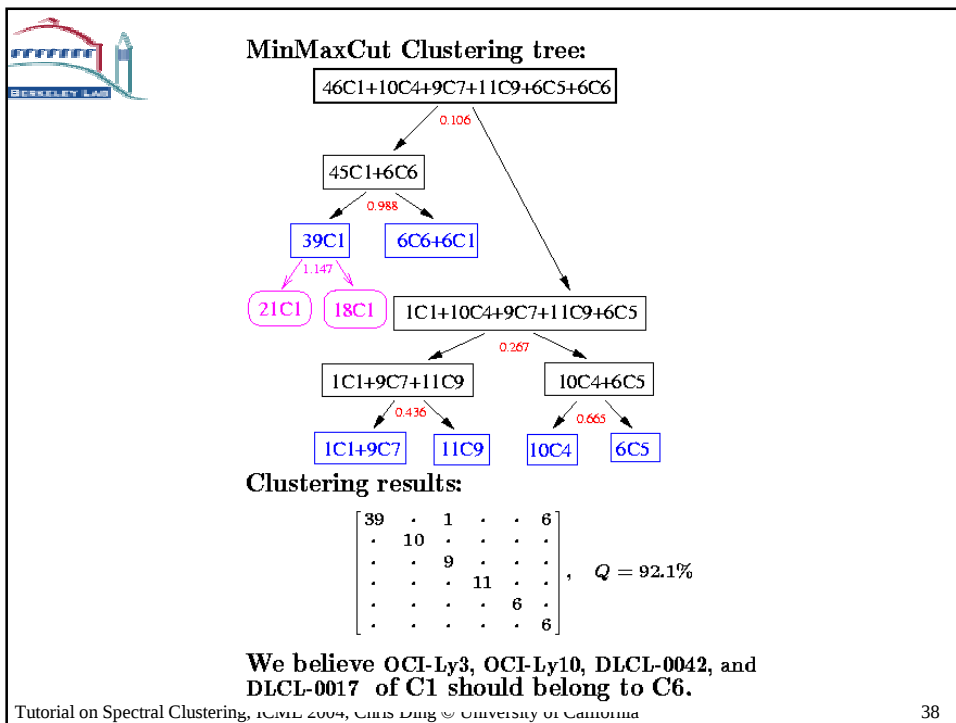
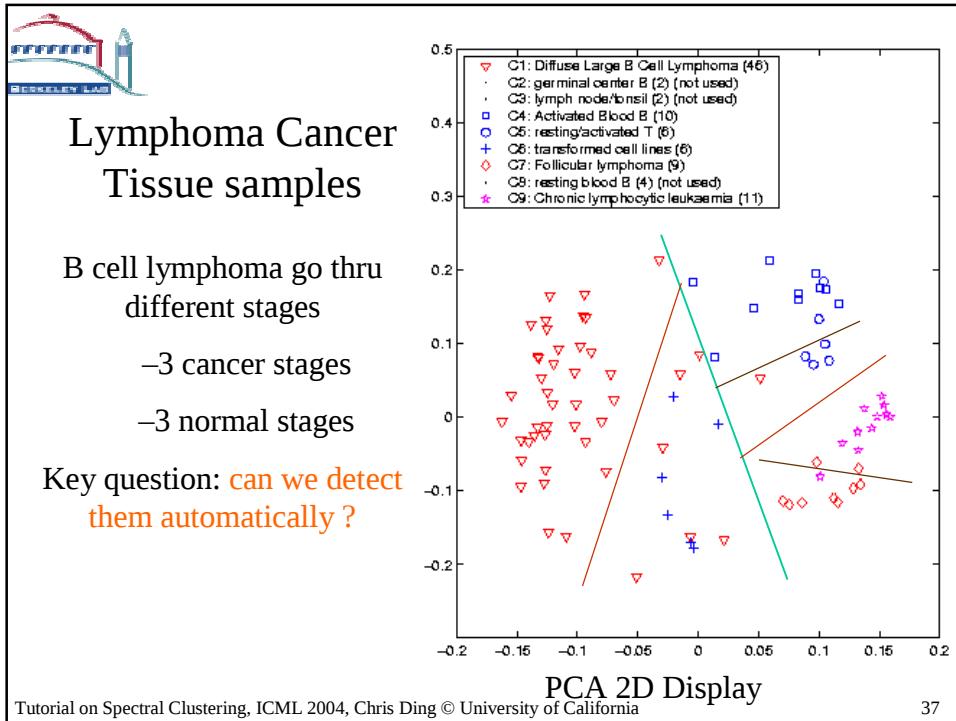


DNA Gene expression

Lymphoma Cancer
(Alizadeh et al, 2000)

Effects of feature selection:
Select 900 genes out of
4025 genes







Brief summary of Part I

- Spectral graph partitioning as origin
- Clustering objective functions and solutions
- Extensions to bipartite and directed graphs
- Characteristics
 - Principled approach
 - Well-motivated objective functions
 - Clear, un-ambiguous
 - A framework of rich structures and contents
 - **Everything is proved rigorously** (within the relaxation framework, i.e., using continuous approximation of the discrete variables)
- Above results mostly done by 2001.
- More to come in Part II