# Multi-view Spectral Clustering with Applications in Gene Coexpression Networks

Shuqin Zhang

School of Mathematical Sciences
Fudan University

Joint with M. Ng (HKBU), H. Zhao (Yale)

June 8, 2015

# Outline


1. Introduction
2. Module Identification in Multiple Networks
3. Numerical Experiments
4. Conclusions

# Introduction

- Networks are widely applied to model different types of complex systems.
- Many networks have the community/module structure property.
- Intuitively, a module is a cohesive group of nodes that connected “more densely" to each other than to the nodes in other modules.
- Modules may correspond to some functional units or play similar roles.

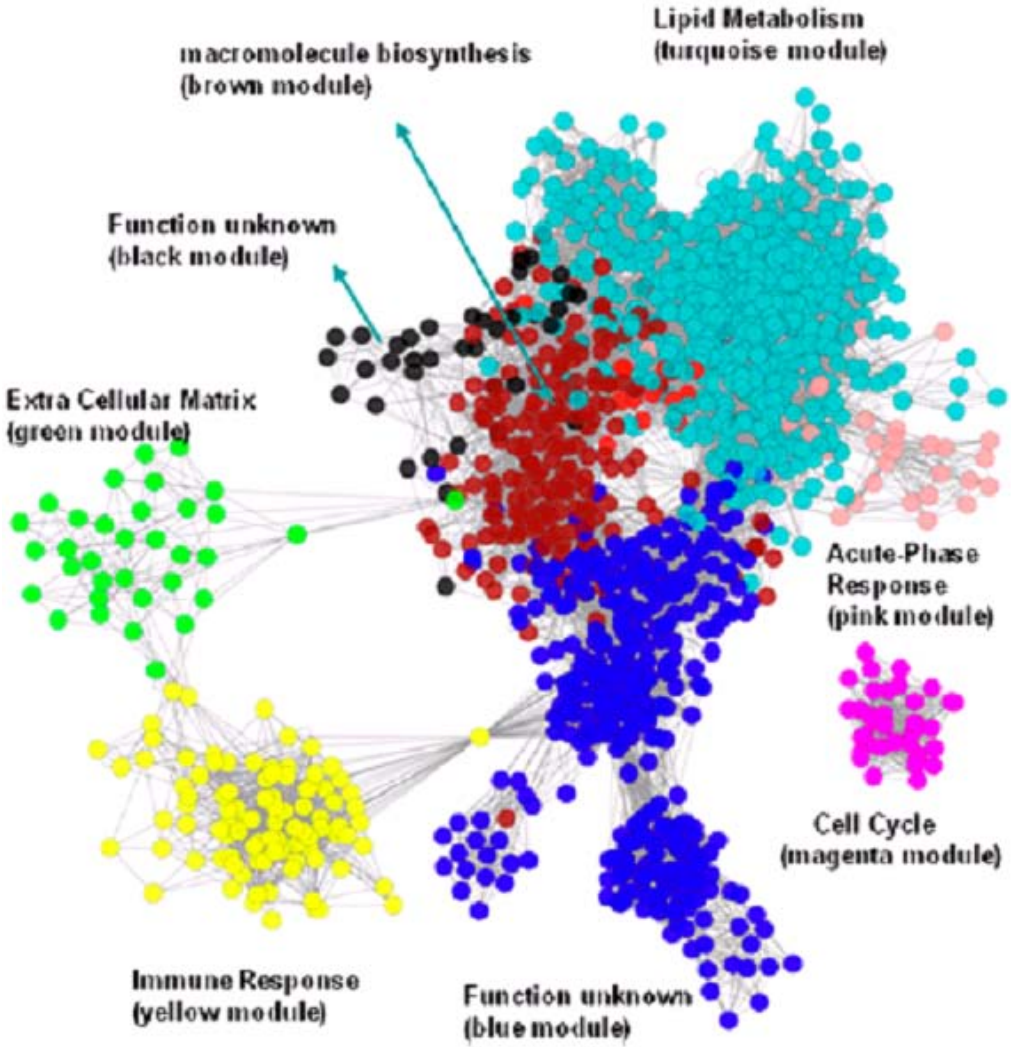



Figure : Human liver cohort(HLC) gene co-expression network

X.Yang,B.Zhang, et al. Genome Research, 20,1020-1036,2010

# Mathematical and computational methods

- Clustering techniques;
- Modularity optimization;
- Spectral clustering;
- Information-theoretic framework;
- Markov time sweeping method;
- Minimum-cut, K-clique percolation, etc..

# Introduction

- A large amount of data on different levels for the same objects are generated in recent years.
  Example:
  Social network data from Facebook, google+, weibo,wechat, etc..
  Citation network data from different journals for the same authors.
  Gene expression data from different cancers.
- Integration of the data on different levels can provide a better way of understanding, classifying and grouping objects for analysis and applications.

# Functional module identification from multiple biological networks

- The modules in a single network may not be stable due to the noise in the data or the tuning of parameters when building the networks.
- Identification of modules from multiple gene co-expression networks for the patients' different tissues can discover the subtle signals that may not be clear in one specific tissue.
- By identifying the common modules in the networks constructed from patients having different diseases, we can obtain the common factors of them.
- By integrating networks for different species, we can study the conservation and evolvement relations of them.

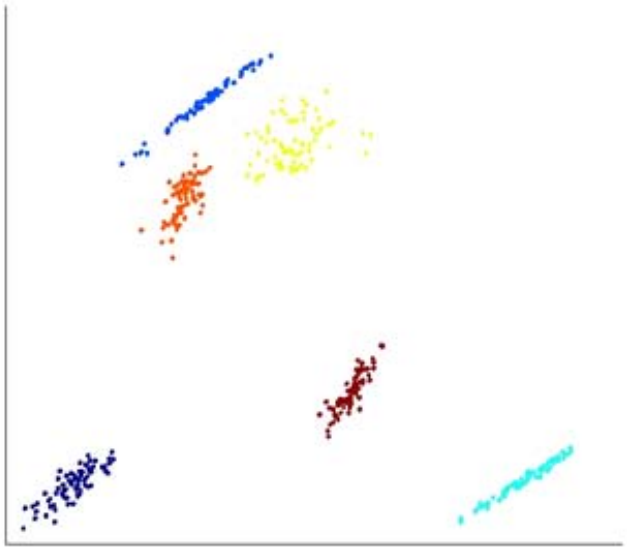
# Weakness of the existing methods

- Most existing methods were developed under the assumption that the underlying modules/clusters for all considered networks/data sets are the same.
- The heuristic algorithms by subgraph searching tend to find the small subgraphs compared to the general concept of modules.
- The computation speed may be slow.

# Introduction: Spectral clustering

Clustering is a basic step of analyzing a large data set.

- Maximize inter(between)-cluster distance
- Minimize intra(within)-cluster distance



left: Kmeans works!
right: Kmeans does not work, spectral clustering works!

# Spectral clustering

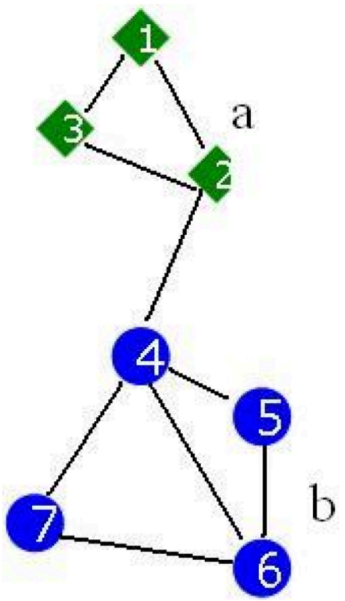
Spectral clustering has been one of the most popular modern clustering techniques in recent years.

- Preprocessing
  Construct the similarity matrix and the graph representing the data set.
- Spectral representation
  - Form the associated Laplacian matrix
  - Compute eigenvalues and eigenvectors of the Laplacian matrix.
  - Map each point to a lower-dimensional representation based on one or more eigenvectors.
- Clustering
  Assign points to two or more classes, based on the new representation.

# Basic Concepts

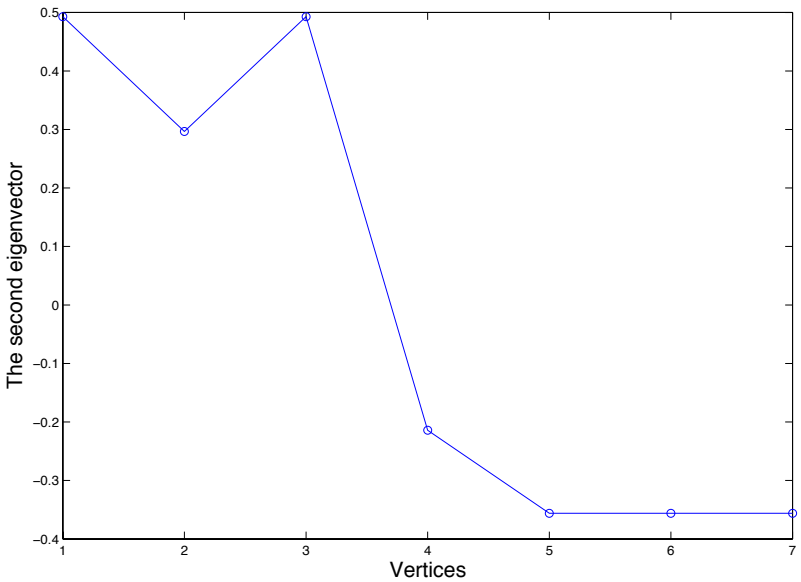
- Suppose the adjacency matrix of a network $G(V, E)$ with $n$ vertices is $A$. If there is an edge between nodes $i$ and $j$, $A_{ij} = 1$, otherwise, $A_{ij} = 0$.
- Degree of the vertex $i$: $d_i = \sum_{j=1}^{n} A_{ij}$. $D = diag(d_i)$.

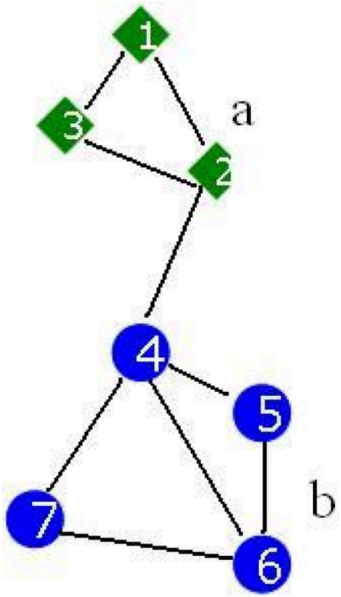
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$




- $D = diag(2, 3, 2, 4, 2, 3, 2)$.

- Laplacian matrix $L = D - A$.



**Algorithm:**

Input: Data set $(X_1, X_2, \cdots, X_N)$, and $K$, which is the number of clusters.

1. Construct the similarity graph $A$, where $A_{ij}$ describes the similarity between the point $X_i$ and $X_j$;
2. Compute the matrix $L = D - A$;
3. Compute the $K$ eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K$ corresponding to the $K$ smallest eigenvalues of matrix $L$;
4. Construct a new matrix $T \in R^{N \times K}$, with columns $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K$;
5. Cluster the points constructed from each row of matrix $T$ with $k$-means clustering into clusters $C_1, C_2, \cdots, C_K$.

Output: Index of vertices in each cluster.

# Graph cut point of view

Given a partition of $V$ into $K$ sets $C_1, \cdots, C_K$, we define $K$ indicator vectors $f_1, \cdots, f_k$ where $f_j$ is the indicator vector for $C_j$. Define:

$$\text{cut}(C_k, \overline{C}_k) = \frac{1}{2} \sum_{i \in C_k, j \in \overline{C}_k} A_{ij}$$

| graph cut | indicator vectors | objective | constraints |
|---|---|---|---|
| Mcut=$\sum_{k=1}^{K} \text{cut}(C_k, \overline{C_k})$ | $f_{ik} = \begin{cases} 1, & \text{if } i \in C_k, \\ 0, & \text{otherwise,} \end{cases}$ | $\text{Tr}(F^T L F)$ | |
| Rcut=$\sum_{k=1}^{K} \frac{\text{cut}(C_k, \overline{C_k})}{\lvert C_k \rvert}$ | $f_{ik} = \begin{cases} 1/\sqrt{\lvert C_k \rvert}, & \text{if } i \in C_k, \\ 0, & \text{otherwise,} \end{cases}$ | $\text{Tr}(F^T L F)$ | $F^T F = I$ |
| Ncut=$\sum_{k=1}^{K} \frac{\text{cut}(C_k, \overline{C_k})}{vol(C_k)}$ | $f_{ik} = \begin{cases} 1/\sqrt{vol(C_k)}, & \text{if } i \in C_k, \\ 0, & \text{otherwise,} \end{cases}$ | $\text{Tr}(F^T L F)$ | $F^T D F = I$ |

# Module identification in multiple networks

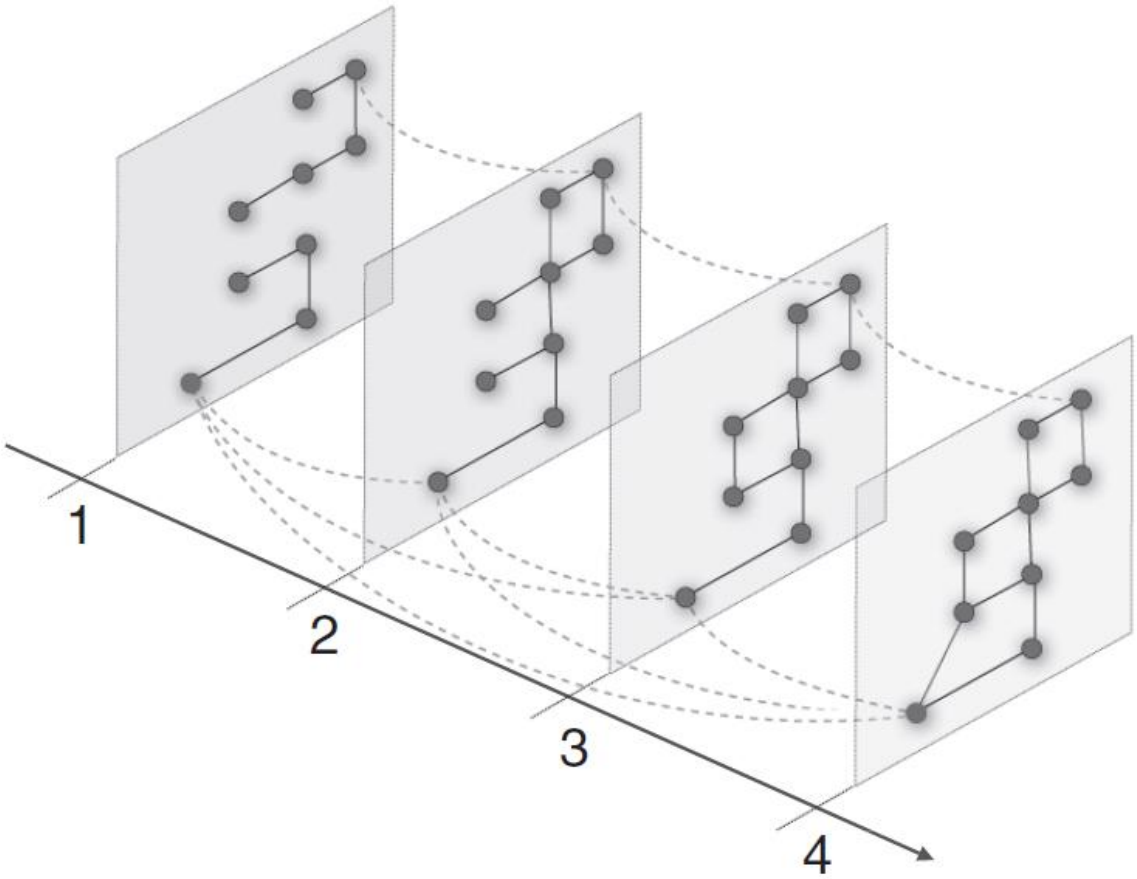



Figure : Schematic of a multislice network

P. J. Mucha, et al., Community structure in time-dependent, multiscale, and multiplex networks, Science, 328, 876-878, 2010.

# Multi-view spectral clustering

Clustering Goal

1. The intra-graph nodes between different modules have very small similarities and those within the same clusters have very high similarities;
2. Partitions in different graphs should be highly consistent.

# Notations

- Suppose we have $M$ different graphs $G_1, G_2, \cdots, G_M$, which are constructed from $N$ objects under $M$ different relations;
- Each graph consists of $N$ nodes, which compose $K$ clusters.
- The adjacency matrix for graph $G_m$ is $A_m$, where $A_m(i,j) = 1$, if there is a relation between object $i$ and object $j$, otherwise $A_m(i,j) = 0$.
- We use $D_m$ to denote the diagonal matrix with the diagonal entries being the degree of the corresponding node.
- We construct $N \times N$ Laplacian matrix $L_m$ for $G_m$.

# Two-view spectral clustering

- Let $\mathbf{U}^m$ be the assignment of the $N$ objects into $K$ clusters in $G_m$, where

$$\mathbf{U}^m_{ik} = \begin{cases} 1, & \text{if } i \in C_k \text{ for the } m-th \text{ graph}, \\ 0, & \text{otherwise}, \end{cases}$$

  $i = 1, 2, \cdots, N, k = 1, 2, \cdots, K, m = 1, 2 \cdots, M.$

- To do clustering in each view, we use the standard spectral clustering. That is:

$$\min \sum_{k=1}^{K} \frac{(\mathbf{U}^m_{.,k})^T L_m \mathbf{U}^m_{.,k}}{(\mathbf{U}^m_{.,k})^T \mathbf{U}^m_{.,k}}, \text{ for } m = 1, 2$$

- Key step: define the similarity between the clusters in the two views.
- Define the similarity function between the clusters in $G_1$ and $G_2$:

$$S(\mathbf{U}^1, \mathbf{U}^2) = \sum_{k=1}^{K} \frac{\mathbf{U}^1_{.,k} \cdot \mathbf{U}^2_{.,k}}{\|\mathbf{U}^1_{.,k}\|_2 \|\mathbf{U}^2_{.,k}\|_2}.$$

- We need to maximize this consistency.
- Combine these three terms together, we have:

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} \left( \frac{(\mathbf{U}^1_{.,k})^T L_1 \mathbf{U}^1_{.,k}}{(\mathbf{U}^1_{.,k})^T \mathbf{U}^1_{.,k}} + \frac{(\mathbf{U}^2_{.,k})^T L_2 \mathbf{U}^2_{.,k}}{(\mathbf{U}^2_{.,k})^T \mathbf{U}^2_{.,k}} \right) - \beta \sum_{k=1}^{K} \frac{\mathbf{U}^1_{.,k} \cdot \mathbf{U}^2_{.,k}}{\|\mathbf{U}^1_{.,k}\|_2 \|\mathbf{U}^2_{.,k}\|_2}, \\ s.t. \quad & \sum_{k=1}^{K} \mathbf{U}^m_{.,k} = 1, \ \text{for} \ \ m = 1, 2. \end{aligned}$$

$\beta$ is the parameter to control the contributions from intra- and inter-graph connections.

- Let $\mathbf{U}_{.k} = (\frac{\mathbf{U}^1_{,k}{}^T}{\|\mathbf{U}^1_{,k}\|_2}, \frac{\mathbf{U}^2_{,k}{}^T}{\|\mathbf{U}^2_{,k}\|_2})^T, \mathbf{U}^T\mathbf{U} = 2I_K$. $I_K$ is the $K$-by-$K$ identity matrix.
- Define:

$$\mathbf{B} = \mathbf{B}_{within} + \beta\mathbf{B}_{across}$$

$$\mathbf{B}_{within} = \begin{pmatrix} L_1 & \mathbf{0} \\ \mathbf{0} & L_2 \end{pmatrix},$$

$$\mathbf{B}_{across} = \begin{pmatrix} \mathbf{0} & -I_N \\ -I_N & \mathbf{0} \end{pmatrix}.$$

- The optimization problem is relaxed to:

$$\min \mathrm{Tr}(\mathbf{U}^T\mathbf{B}\mathbf{U}), s.t. \mathbf{U}^T\mathbf{U} = 2I_K,$$

# Extension to multiple networks

- Define

$$\Psi(\mathbf{U}^1, \cdots, \mathbf{U}^M) = \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{{\mathbf{U}^m_{\cdot,k}}^T L_m \mathbf{U}^m_{\cdot,k}}{\|\mathbf{U}^m_{\cdot,k}\|_2^2} - \beta \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{l=l, l\neq m}^{M} \frac{{\mathbf{U}^m_{\cdot,k}}^T \mathbf{U}^l_{\cdot,k}}{\|\mathbf{U}^m_{\cdot,k}\|_2 \|\mathbf{U}^l_{\cdot,k}\|_2}$$

- The optimization problem is formulated as:

$$\begin{aligned} \min \quad & \Psi(\mathbf{U}^1, \cdots, \mathbf{U}^M) \\ s.t. \quad & \mathbf{U}^m_{i,k} \in \{0, 1\}, i = 1, 2, \cdots, M, k = 1, 2, \cdots, K, \\ & \sum_{k=1}^{K} \mathbf{U}^m_{\cdot,k} = \mathbf{1}, \text{for } m = 1, 2, \cdots, M. \end{aligned} \tag{1}$$

- Similarly, we define:

$$\mathbf{B} = \mathbf{B}_{within} + \beta \mathbf{B}_{across}$$

$$\mathbf{B}_{within} = \begin{bmatrix} L_1 & 0 & \cdots & 0 \\ 0 & L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_M \end{bmatrix},$$

$$\mathbf{B}_{across} = \begin{bmatrix} \mathbf{0} & -I_N & \cdots & -I_N \\ -I_N & \mathbf{0} & \cdots & -I_N \\ \vdots & \vdots & \ddots & \vdots \\ -I_N & -I_N & \cdots & \mathbf{0} \end{bmatrix}.$$

- The optimization problem is relaxed to:

$$\begin{aligned} \min \quad & \mathrm{Tr}(\mathbf{U}^T \mathbf{B} \mathbf{U}) \\ s.t. \quad & \mathbf{U}^T \mathbf{U} = \mathbf{I}_K. \end{aligned}$$

**Algorithm:**

Input: Adjacency matrix $A_m, m = 1, 2 \cdots, M$, and $K$, which is the number of clusters.

1. Compute the matrices $L_m = D_m - A_m$, $m = 1, 2, \cdots, M$;
2. Construct the matrix $\mathbf{B}$;
3. Compute the $K$ eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K$ corresponding to the $K$ smallest eigenvalues of matrix $B$;
4. Construct a new matrix $T \in R^{MN \times K}$, with columns $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K$;
5. Cluster the points constructed from each row of matrix $T$ with $k$-means clustering into clusters $C_1, C_2, \cdots, C_K$;
6. For each cluster, divide the points into $M$ sets according to their original graph label.

Output: Index of nodes in each cluster.

# Analysis of multi-view spectral clustering

## Theorem

- *(i)* $\mathbf{B} + (M-1)I_{NM}$ *is a positive semidefinite matrix and 0 is the smallest eigenvalue.*
- *(ii) The multiplicity of zero eigenvalue of* $\mathbf{B} + (M-1)I_{NM}$ *is equal to the number of connected components of the graph composed of* $G_1, \cdots, G_M$ *and the connections linking the same node.*

# Graph Cut Point of View

Let $\{C_m^{(1)}, C_m^{(2)}, \cdots, C_m^{(K)}\}$ be a partition of $G_m$ $(1 \leq m \leq M)$.

- Define the intra-graph cut between the cluster $C_m^{(k)}$ and its complement $\overline{C_m^{(k)}}$ as

$$\Phi_m(C_m^{(k)}, \overline{C_m^{(k)}}) = \frac{1}{2} \sum_{i \in C_m^{(k)}, j \in \overline{C_m^{(k)}}} A_m(i,j),$$

- The consistency weight between $C_m^{(k)}$ in $G_m$ and $C_{m'}^{(k)}$ in $G_{m'}$ is defind as

$$\Psi(C_m^{(k)}, C_{m'}^{(k)}) = \sum_{s \in C_m^{(k)}, t \in C_{m'}^{(k)}} s \triangle t,$$

where $s \triangle t$ is equal to 1 (or 0) if $s = t$ (or $s \neq t$).

# Ratio cut for multiple graphs

Define:

$$
\begin{aligned}
& J_1(\{C_m^{(1)}, C_m^{(2)}, \cdots, C_m^{(K)}\}_{m=1}^M) \\
= & \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{1}{|C_m^{(k)}|}\Phi_m(C_m^{(k)}, \overline{C_m^{(k)}}) - \frac{\beta}{M}\sum_{m,m'=1,m\neq m'}^{M}\sum_{k=1}^{K}\frac{\Psi(C_m^{(k)}, C_{m'}^{(k)})}{\sqrt{|C_m^{(k)}||C_{m'}^{(k)}|}}
\end{aligned}
$$

**Theorem**

*Multi-view spectral clustering is a relaxation version of minimization of $J_1$.*

# Numerical Experiments

- Demonstration with synthetic data
- Comparison with other methods
- Applications in biological data sets

# Demonstration with synthetic data

- Data: Downloaded from http://www.biostat.pitt.edu/bioinfo/publication.htm
- Network construction: we first calculated the Pearson correlation coefficient between any two genes. Then if its absolute value is greater than some given value, we assign an edge between them; otherwise, there is no edge. We tried different thresholds such that these networks have approximately scale free property. The average degree of these three networks are all between 3 and 4.

Table : Module identification results for the simulated data.

| data | SD | Cluster | $\alpha$ | $No_{isolated}$ | $Accu_{sep}$ | $Accu_{int}$ |
|---|---|---|---|---|---|---|
| data_0 | 0.2 | 0,6,8 | 0.85 | 19 | 0.88 | 0.97 |
| _Noise0. | 0.4 | | 0.70 | 6 | 0.95 | |
| | 0.8 | | 0.50 | 7 | 0.91 | |
| data_1 | 0.2 | 1,4,6 | 0.85 | 0 | 0.99 | 1.00 |
| _Noise0. | 0.4 | | 0.75 | 8 | 0.91 | |
| | 0.8 | | 0.55 | 16 | 0.96 | |
| data_2 | 0.2 | 8,12,13 | 0.85 | 0 | 0.99 | 1.00 |
| _Noise0. | 0.4 | | 0.70 | 4 | 0.91 | |
| | 0.8 | | 0.50 | 15 | 0.78 | |
| data_3 | 0.2 | 4,11,12 | 0.90 | 0 | 1.00 | 1.00 |
| _Noise0. | 0.4 | | 0.80 | 5 | 0.91 | |
| | 0.8 | | 0.60 | 8 | 0.93 | |
| data_4 | 0.2 | 1,2,10 | 0.80 | 1 | 0.97 | 1.00 |
| _Noise0. | 0.4 | | 0.65 | 8 | 0.79 | |
| | 0.8 | | 0.50 | 29 | 0.55 | |

'SD' is the standard deviation of the noise, '$\alpha$' is the cutoff for building the gene coexpression networks, '$No_{isolated}$' is the number of isolated nodes in each network,

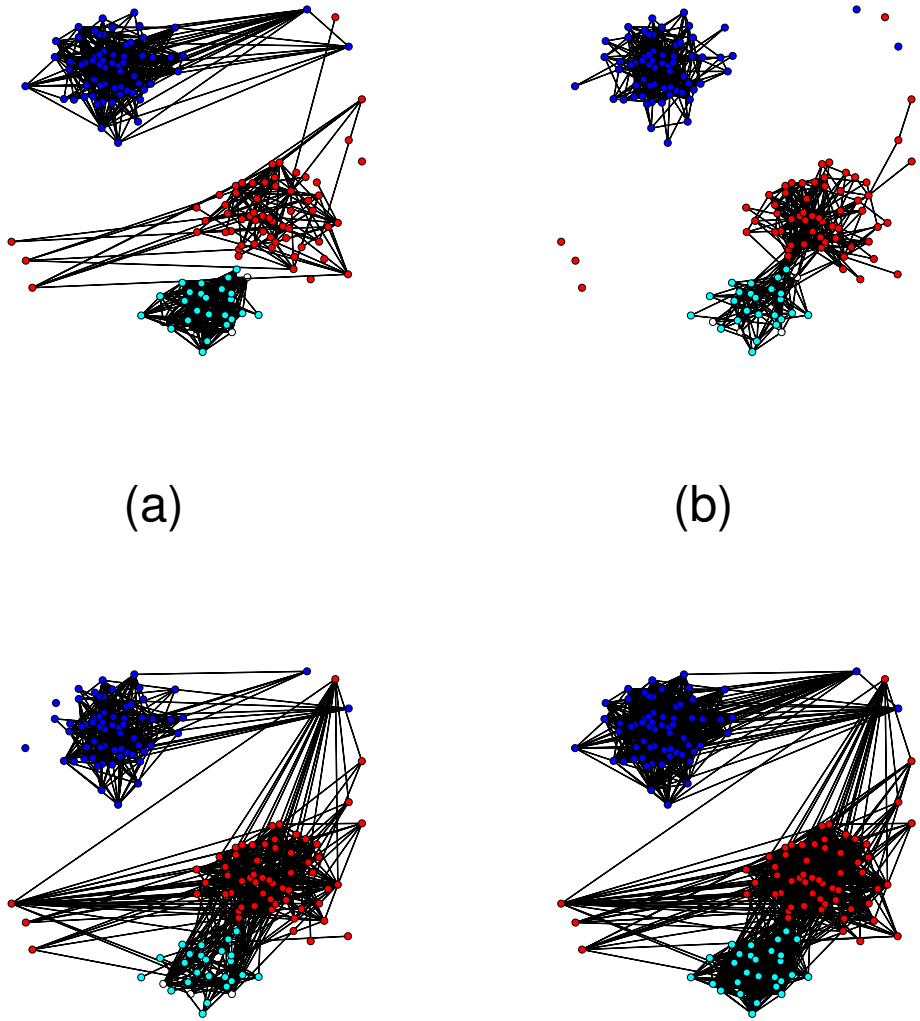



Figure : The three identified consistent modules.

# Comparison with other methods

- Simulation setting: We consider the following four connection probabilities:

$$P_1 = \frac{1}{n}\begin{pmatrix} 16 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 17 \end{pmatrix}, P_2 = \frac{1}{n}\begin{pmatrix} 16 & 0.4 & 0.6 \\ 0.4 & 18 & 0.55 \\ 0.6 & 0.55 & 17 \end{pmatrix},$$

$$P_3 = \frac{1}{n}\begin{pmatrix} 16 & 0.8 & 1.2 \\ 0.8 & 18 & 1.1 \\ 1.2 & 1.1 & 17 \end{pmatrix}, P_4 = \frac{1}{n}\begin{pmatrix} 16 & 1.2 & 1.8 \\ 1.2 & 18 & 1.65 \\ 1.8 & 1.65 & 17 \end{pmatrix}.$$

  We generated 50 networks for each setting.

- $$\text{Identification accuracy} = \frac{TP + TN}{TP + TF + FP + FN},$$

  where 'TP', 'TN', 'FP', and 'FN' represent the number of the true positive, true negative, false positive, and false negative.

# Comparison partners

- Affinity Aggregation for Spectral Clustering (AASC)

$$\begin{aligned} \min \quad & \sum_k v_k^2 \mathrm{Tr}(\mathbf{f}^T(\mathbf{D}_k - \mathbf{W}_k)\mathbf{f}), \\ s.t. \quad & \mathbf{f}^T\mathbf{f} = I. \end{aligned}$$

- Co-Regularized multi-view Spectral Clustering (CRSC)

$$\begin{aligned} \min \quad & \sum_v \mathrm{Tr}({U^{(v)}}^T L_{(v)} U^{(v)}) - \lambda \sum_{v \neq w} \mathrm{Tr}(U^{(v)} {U^{(v)}}^T U^{(w)} {U^{(w)}}^T), \\ s.t. \quad & {U^{(v)}}^T U^{(v)} = I. \end{aligned}$$

- Nonnegative Matrix Factorization Clustering (NMFC)

$$\begin{aligned} \min \quad & \sum_{g=1}^{2} \|A_g - H_g H_g^T\|_F^2 + \lambda_1 \sum_{g=1}^{2} \|H_g - H\|_1 + \lambda_2 \|H\|_1 \\ s.t. \quad & \sum_{k=1}^{K} (H_g)_{ik} = 1, (H_g)_{ik}, H_{ik} \geq 0. \end{aligned}$$

- Optimized data fusion for K-means Laplacian Clustering (OKLC)

$$\begin{aligned} \min \quad & \sum_{m} \mathrm{Tr}(\lambda \mathbf{f}^T L_m \mathbf{f} + (1-\lambda) \mathbf{f}^T G_m \mathbf{f}), \\ s.t. \quad & \mathbf{f}^T \mathbf{f} = I. \end{aligned}$$

- Case 1: The number of nodes for each cluster is given by (50,50,50), and (50,50,50). Then the total size the three common modules is 150.

Table : Comparison of the identification accuracy for the networks with the same module sizes across different networks.

| Setting | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|
| AASC | **1.00(0.00)** | **1.00(0.00)** | **1.00(0.00)** | **1.00(0.01)** |
| CRSC | 1.00(0.01) | 0.99(0.01) | 0.99(0.01) | 0.98(0.14) |
| NMFC | 0.96(0.08) | 0.97(0.07) | 0.97(0.06) | 0.98(0.05) |
| OKLC | 0.99(0.01) | 0.99(0.01) | 0.99(0.01) | 0.98(0.03) |
| Our method | **1.00(0.00)** | **1.00(0.00)** | **1.00(0.00)** | **1.00(0.01)** |

- Case 2: The number of nodes for each cluster is given by (50,50,50), and (30,90,30). Then the total size the three common modules is 110.

Table : Comparison of the identification accuracy for the networks with different module sizes across different networks.

| Setting | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|
| AASC | 0.72(0.01) | 0.71(0.01) | 0.72(0.01) | 0.72(0.01) |
| CRSC | 1.00(0.01) | 0.95(0.12) | 0.87(0.24) | 0.75(0.31) |
| NMFC | 0.69(0.08) | 0.68(0.03) | 0.66(0.04) | 0.65(0.04) |
| OKLC | 0.67(0.10) | 0.68(0.08) | 0.66(0.09) | 0.65(0.10) |
| Our method | **1.00(0.00)** | **0.98(0.01)** | **0.93(0.08)** | **0.80(0.13)** |

# The parameter $\beta$

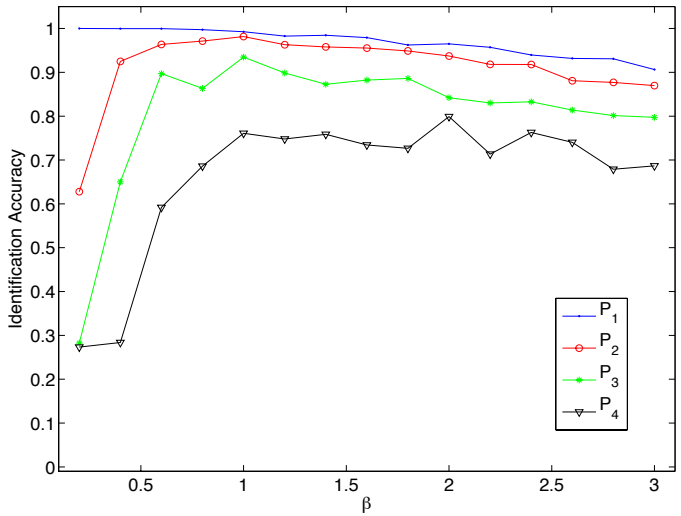



Figure : The average identification accuracy for different values of $\beta$ for Case 2.

# Applications in biological data

- We downloaded the TCGA gene expression data for three cancers: ovarian cancer (OV), glioblastoma multiforme (GBM), and lung squamous cell carcinoma (LUSC) from TCGA website.
- There are 588 OV samples, 594 GBM samples, and 134 LUSC samples. For each cancer, we computed the variance of all the genes across the samples, and selected the first 1500 genes with largest variance. Then we took the union of the genes for further study. The total number of genes considered is 2756.

- We finally identified 13 common clusters. We did enrichment analysis for Gene Ontology (GO, biological process) and KEGG pathways for these modules with DAVID.

Table : Module information for multiple cancers with our proposed method

| Module | Size | Density | $N_{GO}$ | $N_{KEGG}$ |
|---|---|---|---|---|
| Module 1 | 12 | 0.6566 | 15 | 1 |
| Module 2 | 6 | 1.0000 | NA | NA |
| Module 3 | 41 | 0.6858 | 104 | 4 |
| Module 4 | 14 | 0.4359 | 7 | 0 |
| Module 5 | 11 | 0.8788 | 2 | 0 |
| Module 6 | 7 | 0.6190 | 15 | 2 |
| Module 7 | 8 | 0.6310 | 0 | 0 |
| Module 8 | 217 | 0.2033 | 262 | 20 |
| Module 9 | 7 | 1.0000 | 0 | 0 |
| Module 10 | 11 | 0.7697 | 2 | 0 |
| Module 11 | 6 | 1.0000 | NA | NA |
| Module 12 | 6 | 0.7111 | 9 | 0 |
| Module 13 | 77 | 0.7412 | 16 | 3 |

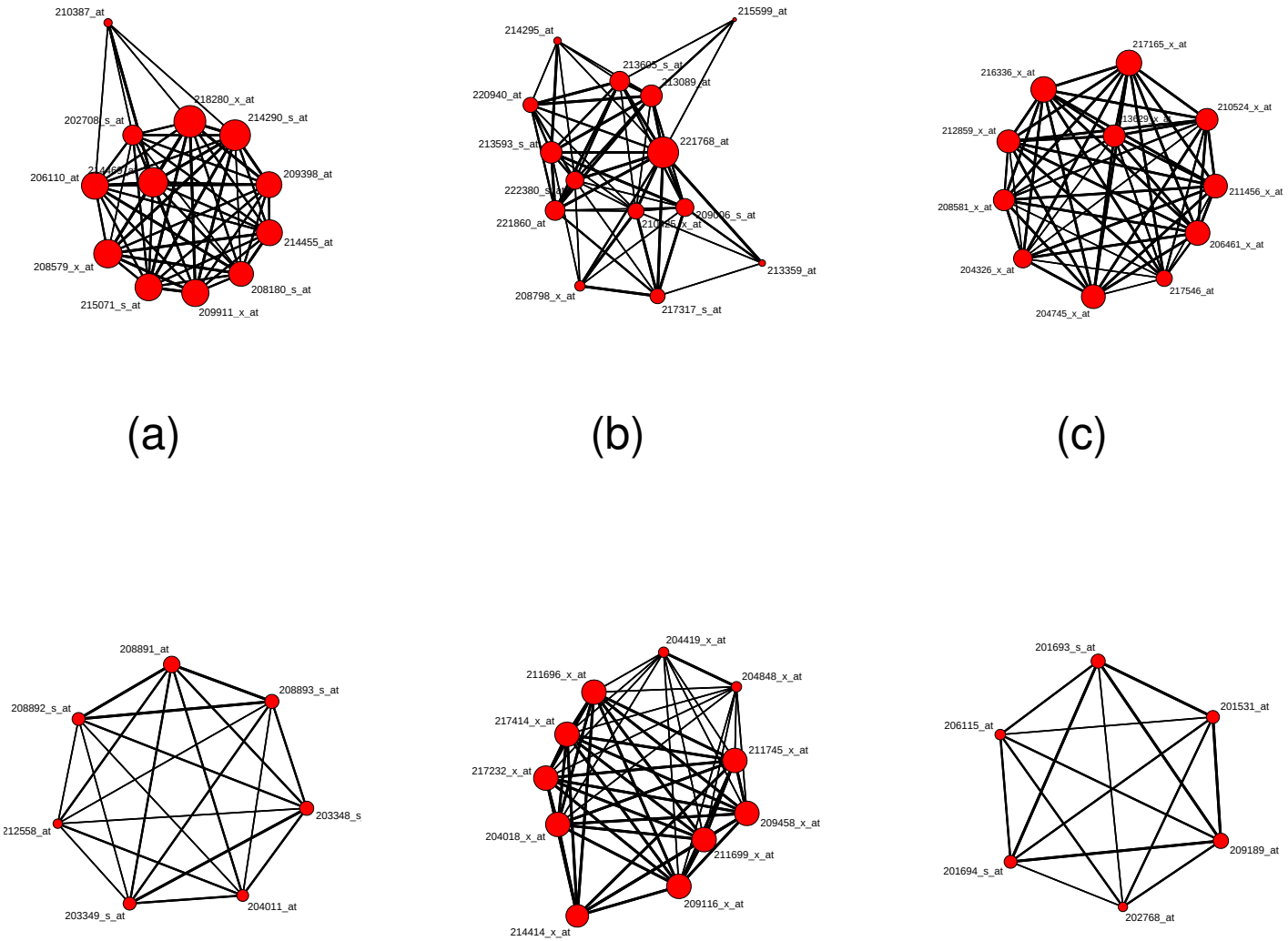



Figure : The identified modules 1, 4, 5, 6, 10, 12 for the networks constructed

- Three complete graphs: Module 2, Module 9, and Module 11. Module 2 and Module 11 correspond to the gene family GAGE, and CD24(CD24L4), respectively.
- The genes in Module 2 belong to GAGE family, which is completely silent in normal adult tissues, except testis., but expressed in a variety of tumor tissues, such as stomach cancer, ovarian carcinoma, and uterine cervical carcinoma.
- The genes in Module 11 belong to Cd24 and CD24L4 family. These genes appear to be highly expressed in a large variety of human cancers, such as ovarian cancer, nonsmall cell lung cancer, colorectal cancer, and they have a high correlation with invasiveness.
- Module 9 corresponds to the control probes.

# Comparison with the known cancer-associated genes

- We checked the associated genes with these three cancers in KEGG. One common gene is *tp*53 (translated to the protein *p*53). Although this gene is not in our final networks with our construction strategy, we still found its related pathway: hsa04115: p53 signaling pathway. Five genes: Cdk1, CCNB2, rrm2, CCNB1, and CCNE2 in Module 3 are included in this pathway.
- These genes enrich 166 GO terms, of which 65 are consistent with those enriched by our identified modules. Among these 65 terms, 2, 25, 6, and 32 terms are from Module 1, Module 3, Module 6, and Module 8, respectively. The most significantly enriched 5 terms of the cancer-associated genes are related to cell cycle.

- These cancer-associated genes enrich a total of 18 KEGG pathways, with 5 of them may be related to all cancers. Two of them are the same as that enriched by Module 3, which are hsa04115: p53 signaling pathway, and hsa04110: Cell cycle. These suggest that Module 3 plays an important role in all cancers.
- Another notable module is Module 6, which is composed of 7 genes. 6 of 15 enriched GO terms are the same as those enriched by the cancer-associated genes. Such information suggests that these biological processes may have close relations to cancers.

# Comparison with AASC

- AASC identified 14 modules.
- The genes in Module 2 and 13 identified by AASC are distributed in 4 modules with our method.

Table : Common modules obtained with our method and AASC

| Our method | AASC | $N_{intersect}$ | Similarity |
|---|---|---|---|
| Module 1 | Module 12 | 12 | 1.0000 |
| Module 2 | Module 10 | 6 | 1.0000 |
| Module 3 | Module 5 | 41 | 1.0000 |
| Module 5 | Module 4 | 11 | 1.0000 |
| Module 8 | Module 14 | 151 | 0.5625 |
| Module 9 | Module 8 | 7 | 1.0000 |
| Module 10 | Module 9 | 9 | 0.9045 |
| Module 11 | Module 11 | 6 | 1.0000 |
| Module 12 | Module 7 | 4 | 0.6667 |
| Module 8 | Module 1 | 26 | 0.3461 |
| Module 8 | Module 6 | 6 | 0.1663 |
| Module 13 | Module 3 | 61 | 0.8430 |

# Gene co-expression networks for different tissues of morbidly obese patients

- GEO Accession number: GSE24294. We focused on the 459 subjects with data available for liver, omental and subcutaneous adipose tissues. The original data were measured on 40,638 probes. After the preprocessing, we got 17,282 common genes of these three tissues. We selected the first 1,800 most differentially expressed genes of each tissue. The total number of the union of these genes is 2637.
- The average degree of gene co-expression networks for liver, omental and subcutaneous adipose tissues is 13.2, 18.7, and 13.0, respectively. After we removed all the common genes with no connections, each network has a total number of 1873 genes.

- We identified 11 modules.

Table : Module information for the morbidly obese patients

| Module | Size | Density | $N_{GO}$ | $N_{KEGG}$ |
|---|---|---|---|---|
| Module 1 | 4 | 1.0000 | 65 | 0 |
| Module 2 | 67 | 0.1728 | 12 | 1 |
| Module 3 | 11 | 0.4545 | 2 | 0 |
| Module 4 | 9 | 0.6111 | 1 | 0 |
| Module 5 | 7 | 1.0000 | NA | NA |
| Module 6 | 6 | 0.5778 | 8 | 0 |
| Module 7 | 13 | 0.2265 | 16 | 1 |
| Module 8 | 12 | 0.2525 | 16 | 6 |
| Module 9 | 73 | 0.1752 | 32 | 5 |
| Module 10 | 385 | 0.1226 | 447 | 27 |
| Module 11 | 6 | 0.8444 | 1 | 1 |

- Two complete graph modules: Module 1 and Module 5.
- Module 1 is composed of the genes: SAA1, SAA2, SAA3P, and SAA4 in all the three tissues. It enriches the GO terms: acute-phase response, acute inflammatory response, inflammatory response, response to wounding, and defense response.
- Module 5 is composed of genes: GAGE3, GAGE4, GAGE5, GAGE6, GAGE7, GAGE7B, and GAGE8, which are from the same gene family. These genes are expressed in a variety of tumor tissues as shown in the previous example.

# Comparison with the known obesity-associated genes

- We checked the obesity related genes in http://omim.org/entry/601665, and did GO terms enrichment for this gene list. They enrich 127 GO terms, among which 7, 1, 3 and 35 terms are the same as those in Modules 1, 8, 7 and Module 10, respectively.
- The 7 terms in Module 1 are mainly related to negative regulation of several responses, such as defense response, and inflammatory response.
- The 3 consistent terms in Module 7 are regulation of transcription from RNA polymerase II promoter, DNA-dependent positive regulation of transcription, and positive regulation of RNA metabolic process. These biological processes show that obesity should be related to the start of transcription.

- Module 3 enriched the GO terms: oxygen transport, and gas transport. Half of the genes carrying out the function of oxygen transport and gas transport are in this module. In the obese situation, oxygen consumption is increased in the obese as a result of the metabolic activity of the excess fat and the increased
- The consistent term in Module 8 is: response to organic substance. This shows that the obese people and the normal people may respond differently to this biological process, which implies that organic food or not may not be the cause for obesity.
- The 35 consistent terms in Module 10 are mainly related to the regulation of some processes.

# Integration of networks helps informative module identification

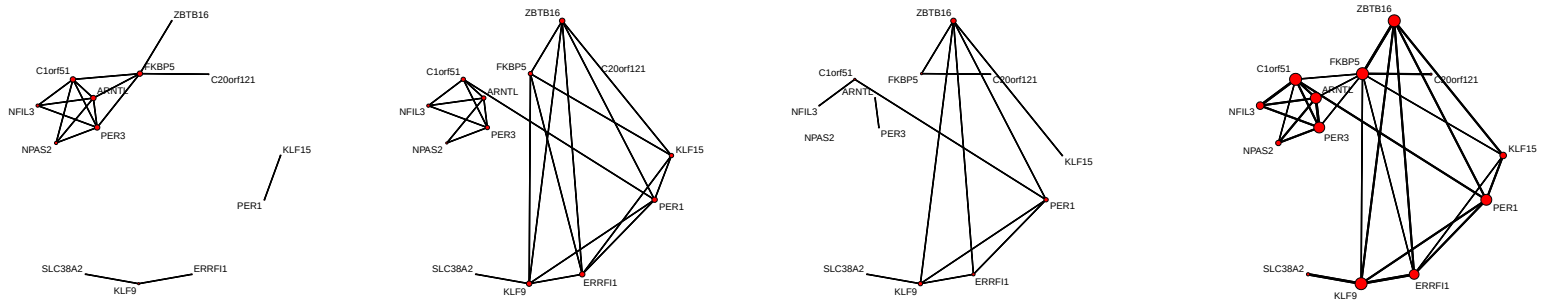



Figure : The structure of Module 7 in the three different tissues of morbidly obese patients. (a) liver, (b) omental, (c) subcutaneous adipose tissue, (d) the combined structure of the module.

# Conclusions

- We extended spectral clustering to multi-view data sets.
- Numerical experiments show the good performance of the method.
- Application in gene coexpression networks found several meaningful modules.

Thank you!