A Spectral Clustering Approach to Optimally Combining Numerical Vectors with a Modular Network

Motoki Shiga, Ichigaku Takigawa, Hiroshi Mamitsuka

Bioinformatics Center, ICR, Kyoto University, Japan

KDD 2007, San Jose, California, USA, August 12-15 2007

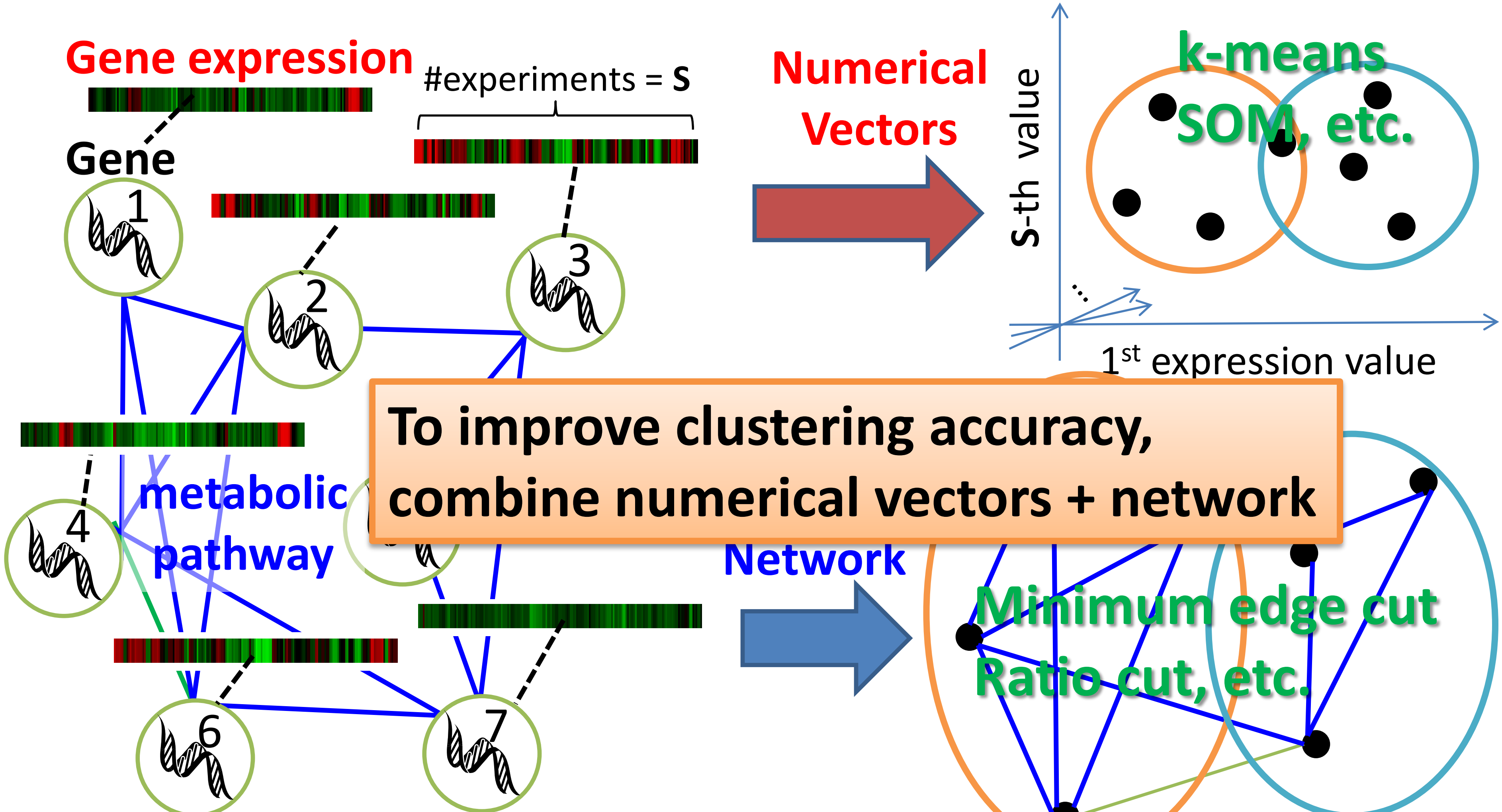
Table of Contents

Heterogeneous Data Clustering

Heterogeneous data : various information related to an interest

Ex. Gene analysis : gene expression, metabolic pathway, …, etc.

Web page analysis : word frequency, hyperlink, …, etc.



To improve clustering accuracy, combine numerical vectors + network

M. Shiga, I. Takigawa and H. Mamitsuka, ISMB/ECCB 2007.

Related work : semi-supervised clustering

- **Local property**

  Neighborhood relation

  -must-link edge, cannot-link edge

- **Hard constraint** (K. Wagstaff and C. Cardie, 2000.)
- **Soft constraint** (S. Basu etc., 2004.)

  - Probabilistic model (Hidden Markov random field)

Proposed method

- **Global property** (network modularity)
- **Soft constraint**

  -Spectral clustering

Table of Contents

1. Motivation
   Clustering for heterogeneous data
   (numerical + network)

2. Proposed method
   Spectral clustering (numerical vectors + a network)

3. Experiments
   Synthetic data and real data

4. Summary

Spectral Clustering

L. Hagen, etc., IEEE TCAD, 1992., J. Shi and J. Malik, IEEE PAMI, 2000.

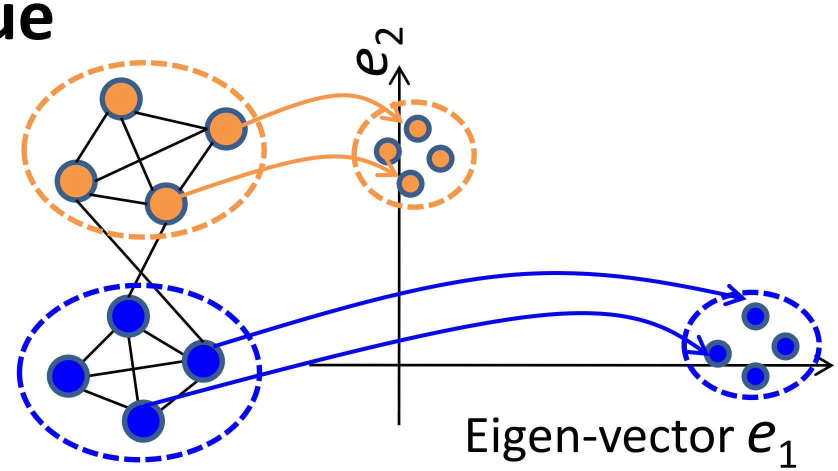
1. Compute affinity(dissimilarity) matrix **M** from data
2. To optimize cost

   $J(\mathbf{Z}) = \mathrm{tr}\{\mathbf{Z}^T\ \mathbf{M}\ \mathbf{Z}\}$ subject to $\mathbf{Z}^T\mathbf{Z}=\mathbf{I}$ ***Trace optimization***

   where $\mathbf{Z}(i,k)$ is 1 when node *i* belong to cluster *k*, otherwise 0,

   compute **eigen-values and -vectors of matrix M**

   by relaxing Z(*i*,k) to a real value

   Each node is by one or more computed **eigenvectors**

3. Assign a cluster label to each node (by k-means)

Cost combining numerical vectors with a network

$$J = \mathrm{tr}\{Z^T M Z\}$$
$$= (1-\omega) J_{num}(Z) + \omega J_{net}(Z)$$

Cost of **numerical vector** **network**

cosine dissimilarity

$$J_{num}(Z) = \frac{1}{2} - \mathrm{tr}\left(\frac{Z^T (2N)^{-1} Y Z}{Z^T Z}\right)$$

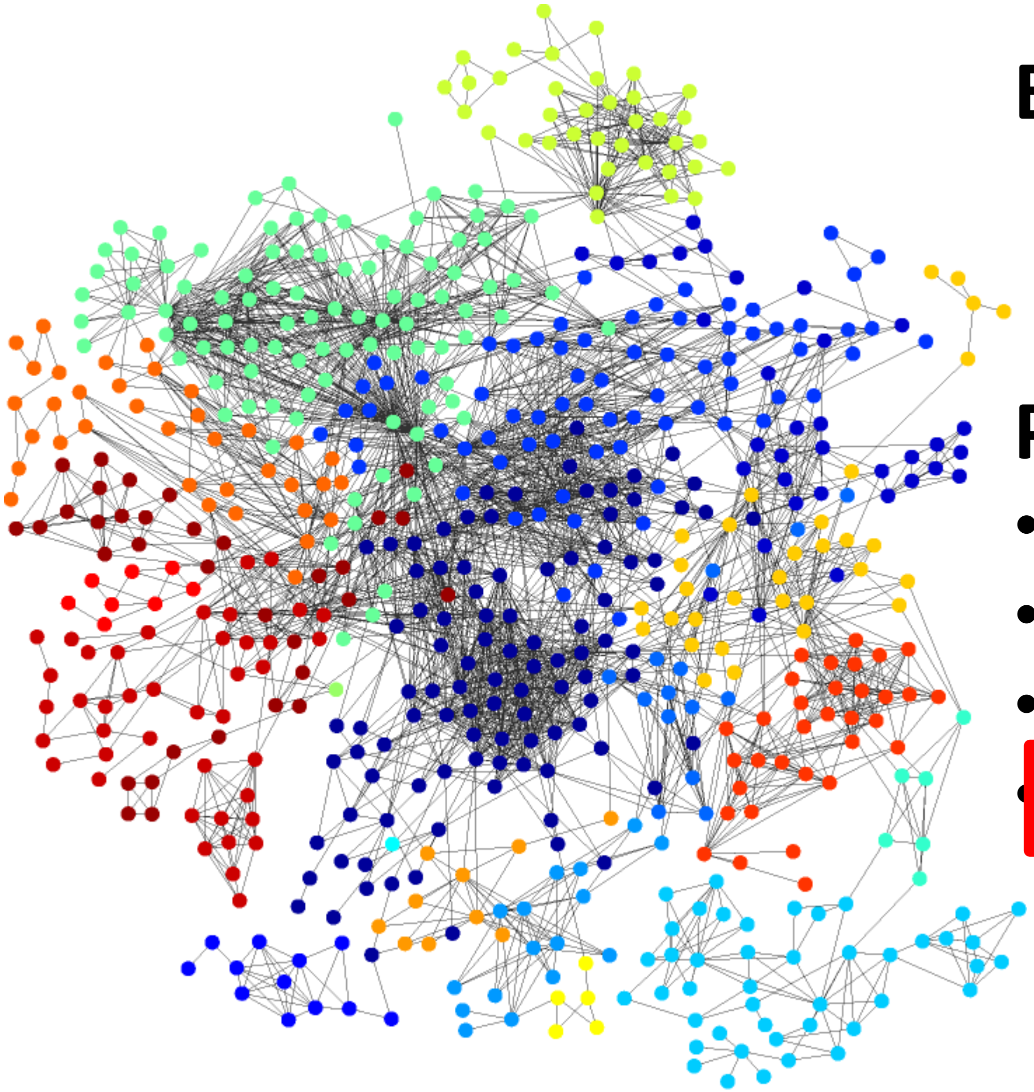
What cost?

N : #nodes,

Y : inner product of normalized numerical vectors

*To define a **cost of a network**,*
*use **a property of complex networks***

Complex Networks



Ex. Gene networks,
WWW,
Social networks, …, etc.

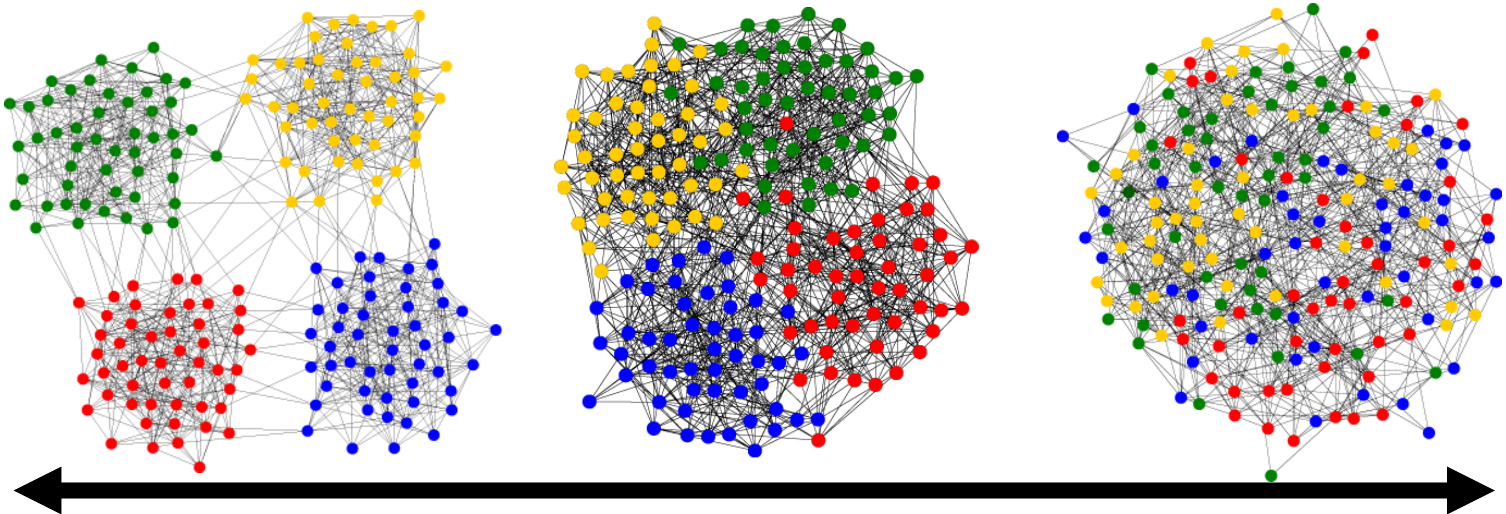
Property

- Small world phenomena
- Power law
- Hierarchical structure
- Network modularity

Ravasz, et al., Science, 2002.
Guimera, et al., Nature, 2005.

Normalized Network Modularity

= density of intra-cluster edges



High ⟷ Low

$$\tilde{Q}(\mathcal{Z}) = \sum_{k=1}^{K} \frac{N}{|\mathcal{Z}_k|} \left\{ \frac{L(\mathcal{Z}_k, \mathcal{Z}_k)}{L} - \left(\frac{L(\mathcal{Z}_k, \mathcal{Z})}{L} \right)^2 \right\}$$

normalize by cluster size

intra-edges

total edges

Z : set of whole nodes

Z_k : set of nodes in cluster k

L(A,B) : #edges between A and B

Guimera, et al., Nature, 2005., Newman, et al., Phy. Rev. E, 2004.

Cost Combining Numerical Vectors with a Network

$$J = \mathrm{tr}\{Z^T M Z\}$$
$$= (1-\omega) J_{num}(Z) + \omega J_{net}(Z)$$

Cost of **numerical vector** **network**

cosine dissimilarity

$$J_{num}(Z) = \frac{1}{2} - \mathrm{tr}\left(\frac{Z^T (2N)^{-1} Y Z}{Z^T Z}\right)$$

Normalized modularity

(Negative)

$$J_{net}(Z) = -\mathrm{tr}\left(\frac{Z^T N \left(\frac{1}{L^2} D - \frac{1}{L} W\right) Z}{Z^T Z}\right)$$

$$\tilde{Z} = \frac{Z}{\sqrt{Z^T Z}}$$

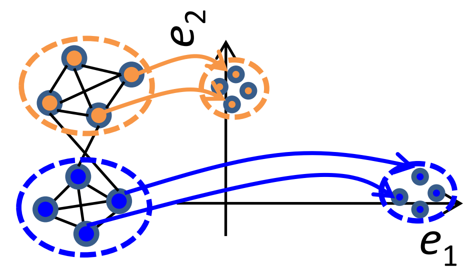
M_ω

$$= \mathrm{tr}\{\tilde{Z}^T \left(\frac{\omega N}{L^2} D - \frac{\omega N}{L} W - \frac{1-\omega}{2N} Y\right) \tilde{Z}\}$$

Our Proposed Spectral Clustering

for ω = 0…1

1. Compute matrix $\mathbf{M}_\omega = \frac{\omega N}{L^2}\boldsymbol{D} - \frac{\omega N}{L}\boldsymbol{W} - \frac{1-\omega}{2N}\boldsymbol{Y}$
2. To optimize cost $J(\mathbf{Z}) = \mathrm{tr}\{\mathbf{Z}^T \mathbf{M}_\omega \mathbf{Z}\}$ subjet to $\mathbf{Z}^T\mathbf{Z}=\mathbf{I}$, compute eigen-values and -vectors of matrix $\mathbf{M}_\omega$ by relaxing elements of **Z** to a real value



   Each node is represented by K-1 eigen-vectors

3. Assign a cluster label to each node by k-means. (k-means outputs $\mathrm{Cost}_{\mathrm{spectral}}$ in spectral space.)

end

- **Optimize weight ω**

$$\omega^* \leftarrow \mathrm{argmin}_{0\leq\omega\leq1}\mathrm{Cost}_{\mathrm{spectral}}$$

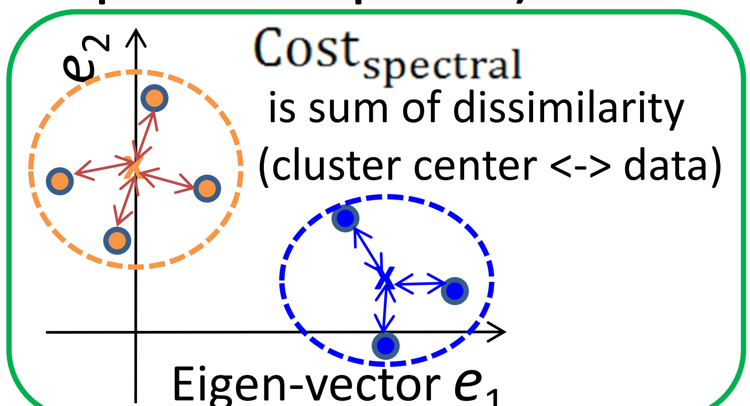
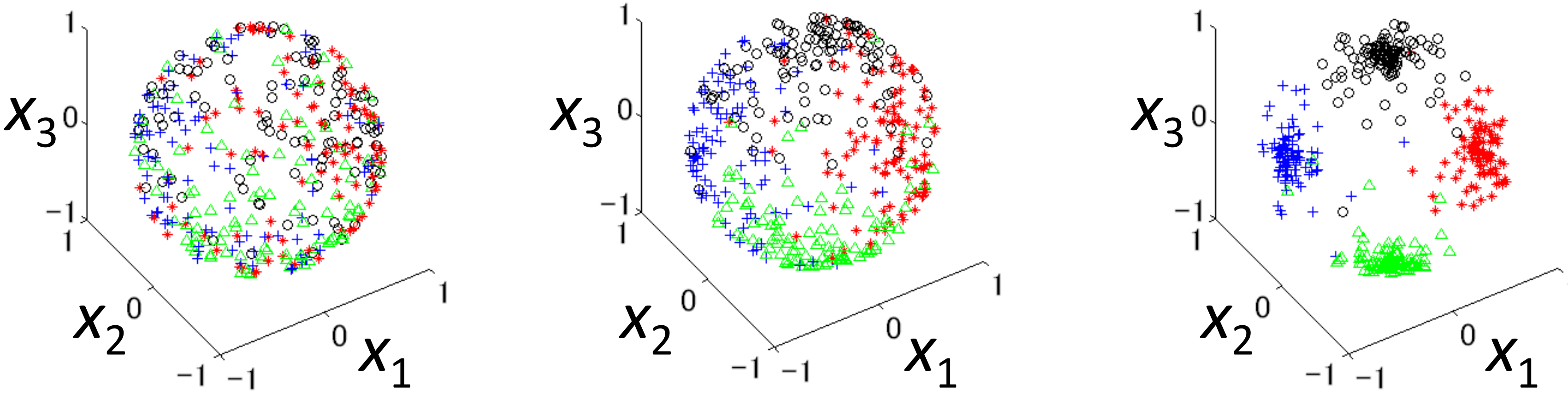


Table of Contents

Synthetic Data

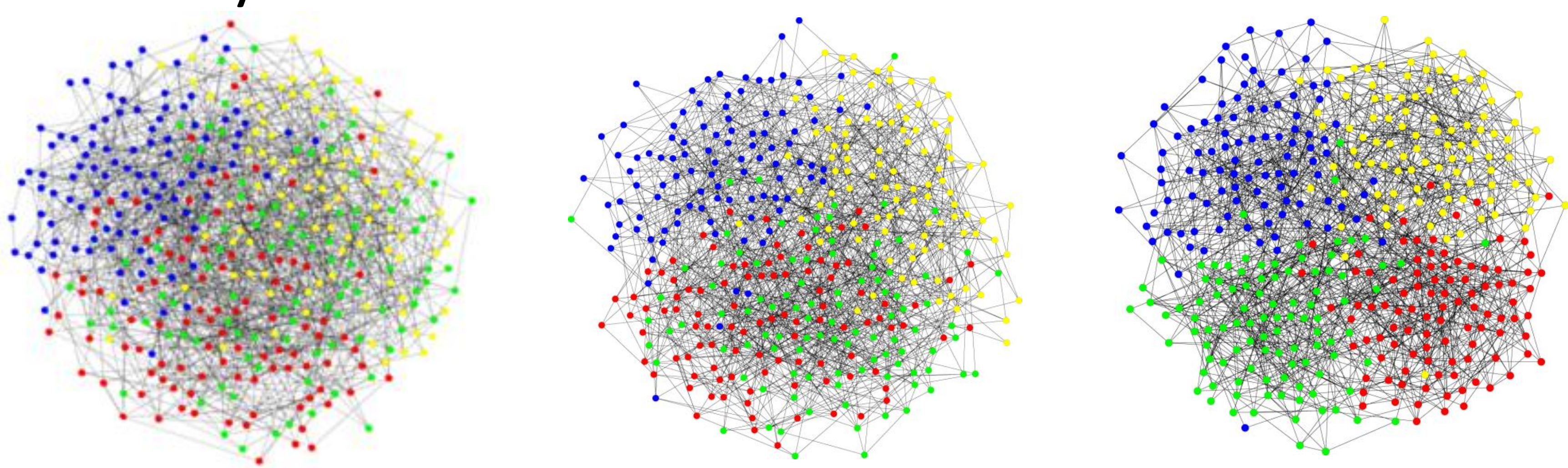
Numerical vectors (von Mises-Fisher distribution)

$\vartheta = 1$ 5 50



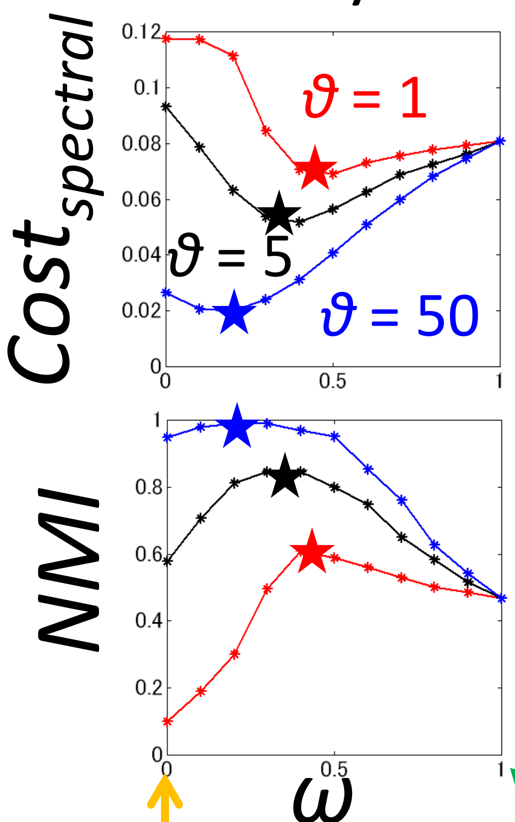
Network (Random graph) #nodes = 400, #edges = 1600

Modularity = 0.375 0.450 0.525

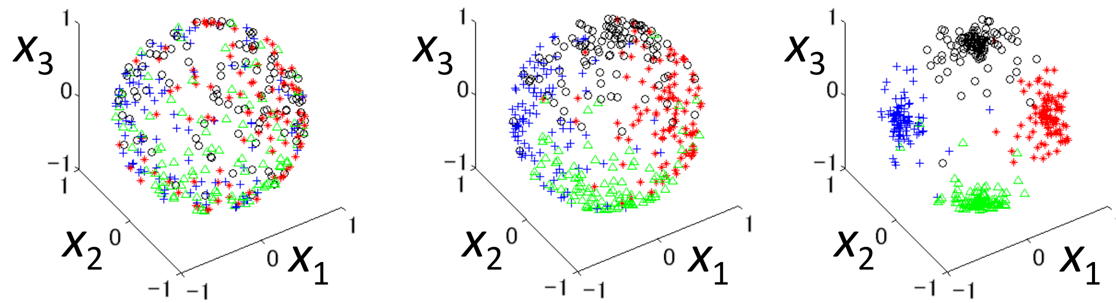


Results for Synthetic Data

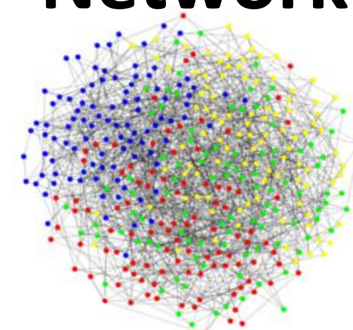
Modularity = 0.375



Numerical vectors



Network



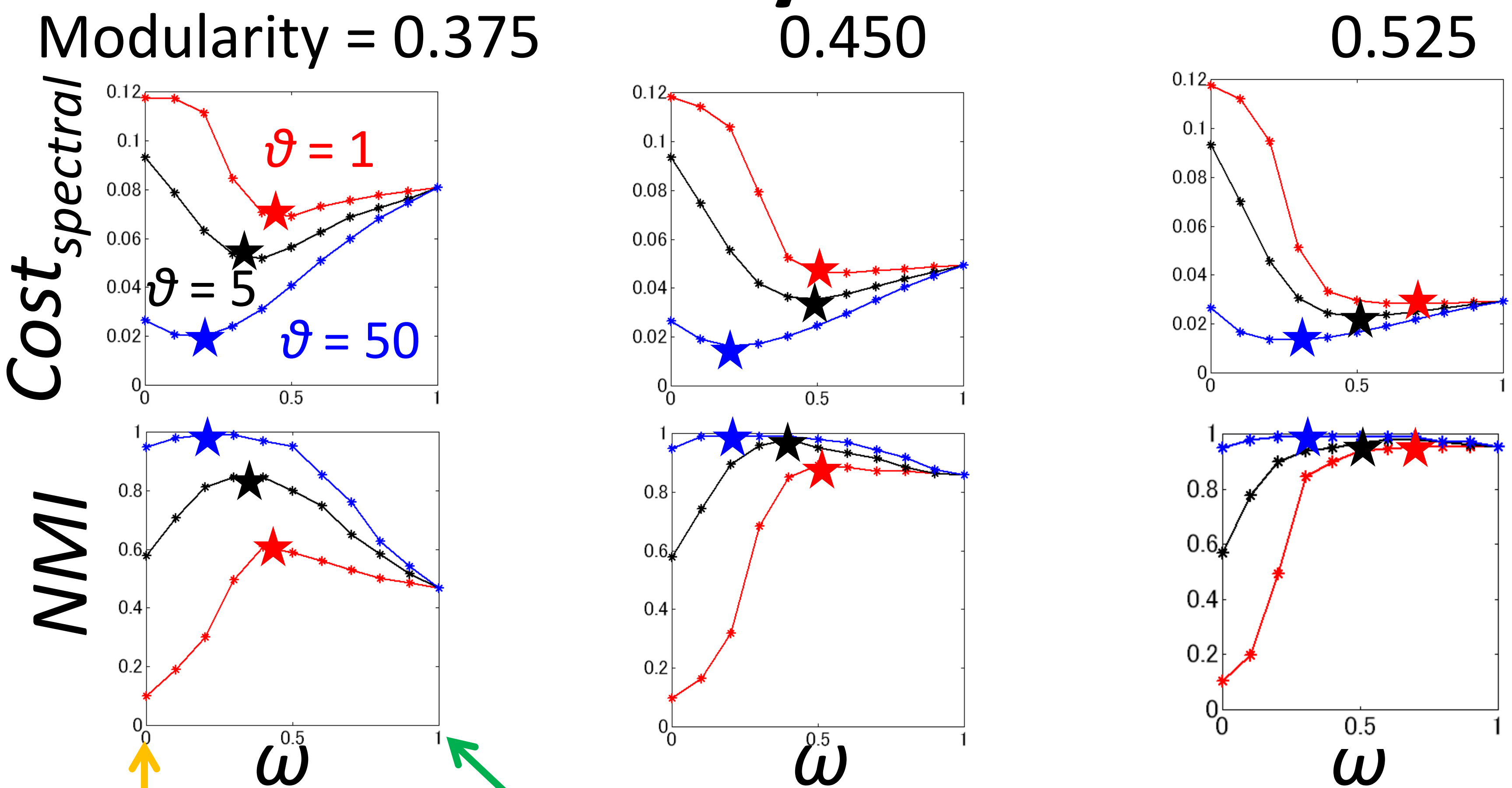
#nodes = 400, #edges = 1600
Modularity = 0.375

Numerical vectors only
(**k-means**)

Network only
(**maximum modularity**)

- Best NMI (Normalized Mutual Information) is in $0 < \omega < 1$
- Can be optimized using $\text{Cost}_{\text{spectral}}$

Results for Synthetic Data



Numerical vectors only
(**k-means**)

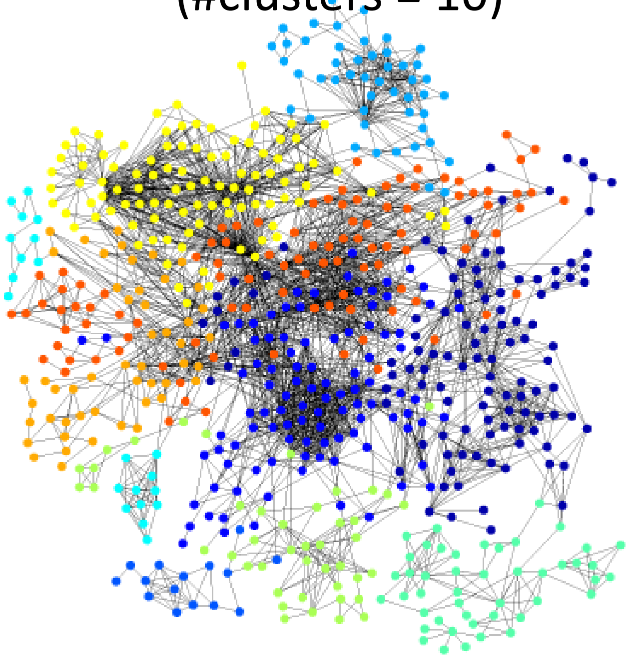
Network only
(**maximum modularity**)

- Best NMI (Normalized Mutual Information) is in $0 < \omega < 1$
- Can be optimized using $\text{Cost}_{\text{spectral}}$

Synthetic Data (Numerical Vector) + Real Data (Gene Network)

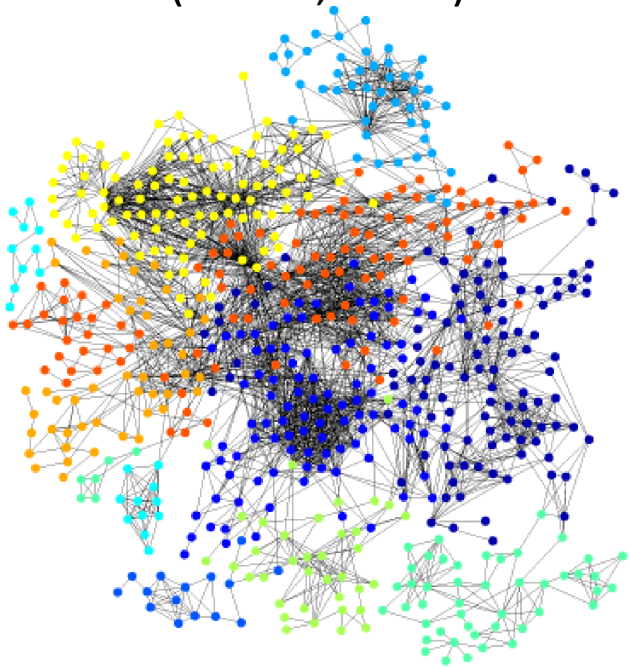
True cluster

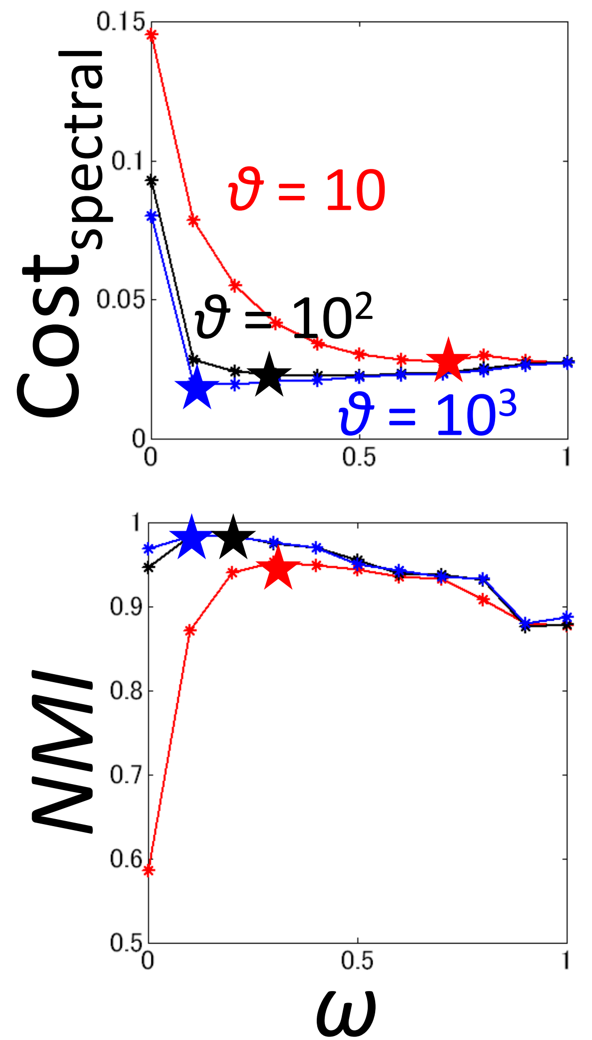
(#clusters = 10)



Resultant cluster

(ω=0.5, ϑ=10)





Gene network

by KEGG metabolic pathway

- Best NMI is in $0 < \omega < 1$
- Can be optimized using Cost$_{spectral}$

Summary

- **New spectral clustering method proposed**
  combining numerical vectors with a network
  - **Global network property** (normalized network modularity)
  - Clustering can be optimized by the weight
- **Performance confirmed experimentally**
  - Better than numerical vectors only and a network only
  - **Optimizing the weight** with synthetic dataset and semi-real dataset

Thank you for your attention!