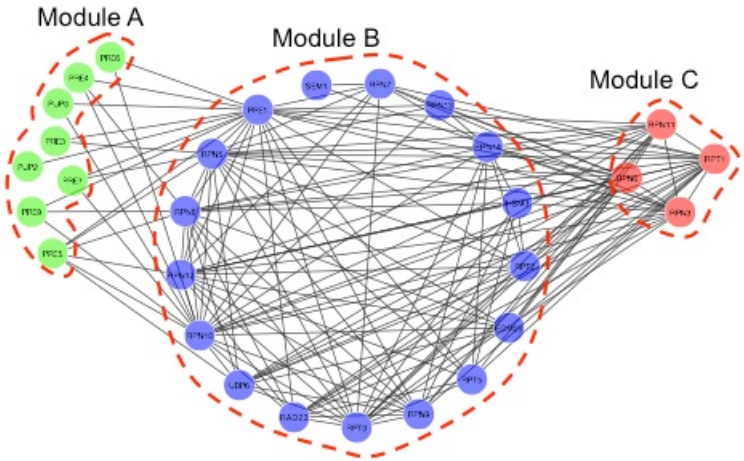
ENGINEERING
TEXAS A&M UNIVERSITY

Functional Module identification in Biological Networks



Xiaoning Qian

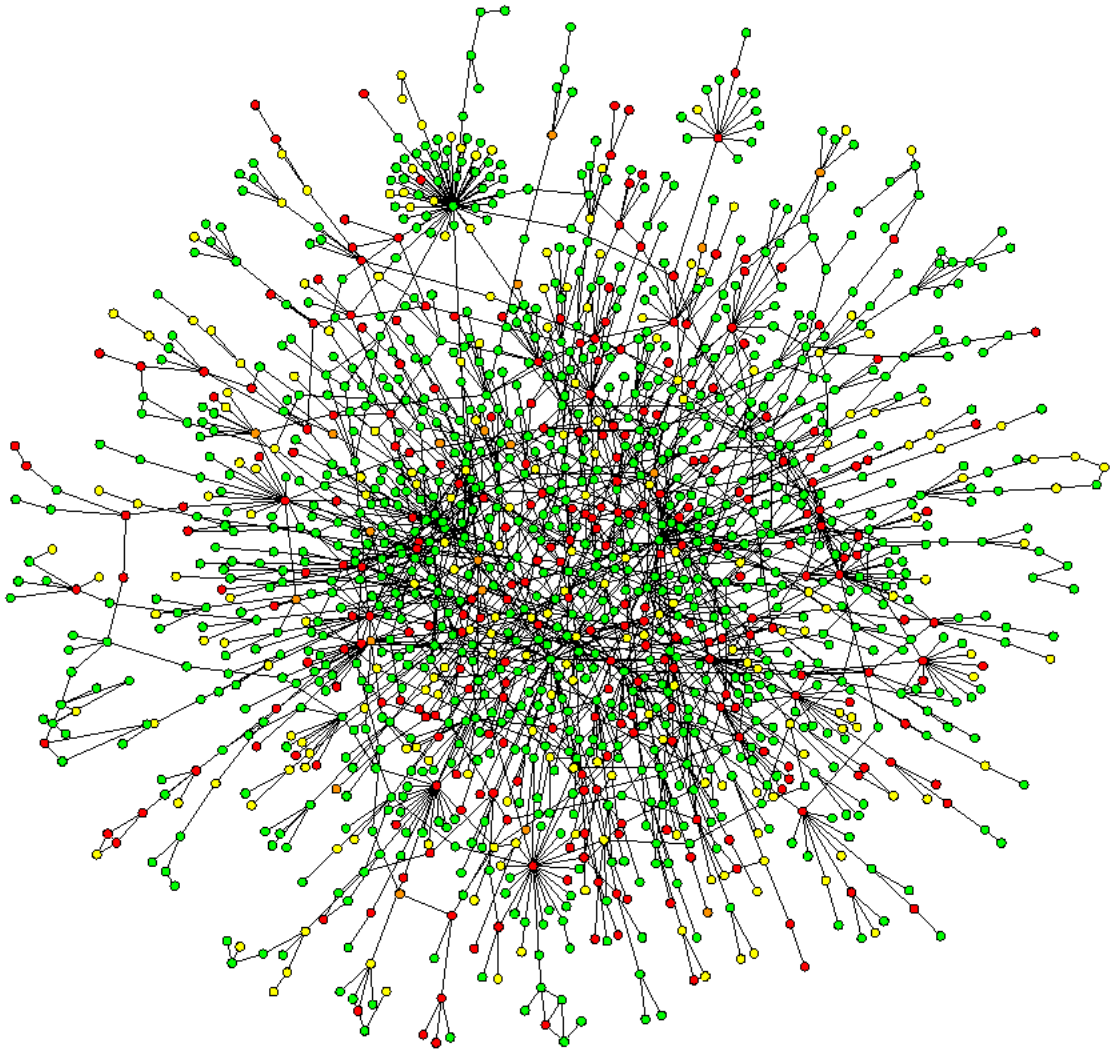
Department of Electrical & Computer Engineering;
Center for Bioinformatics & Genomic Systems Engineering
Texas A&M University

Acknowledgements

- Dr. Yijie Wang
- Siamak Zamani Dadaneh

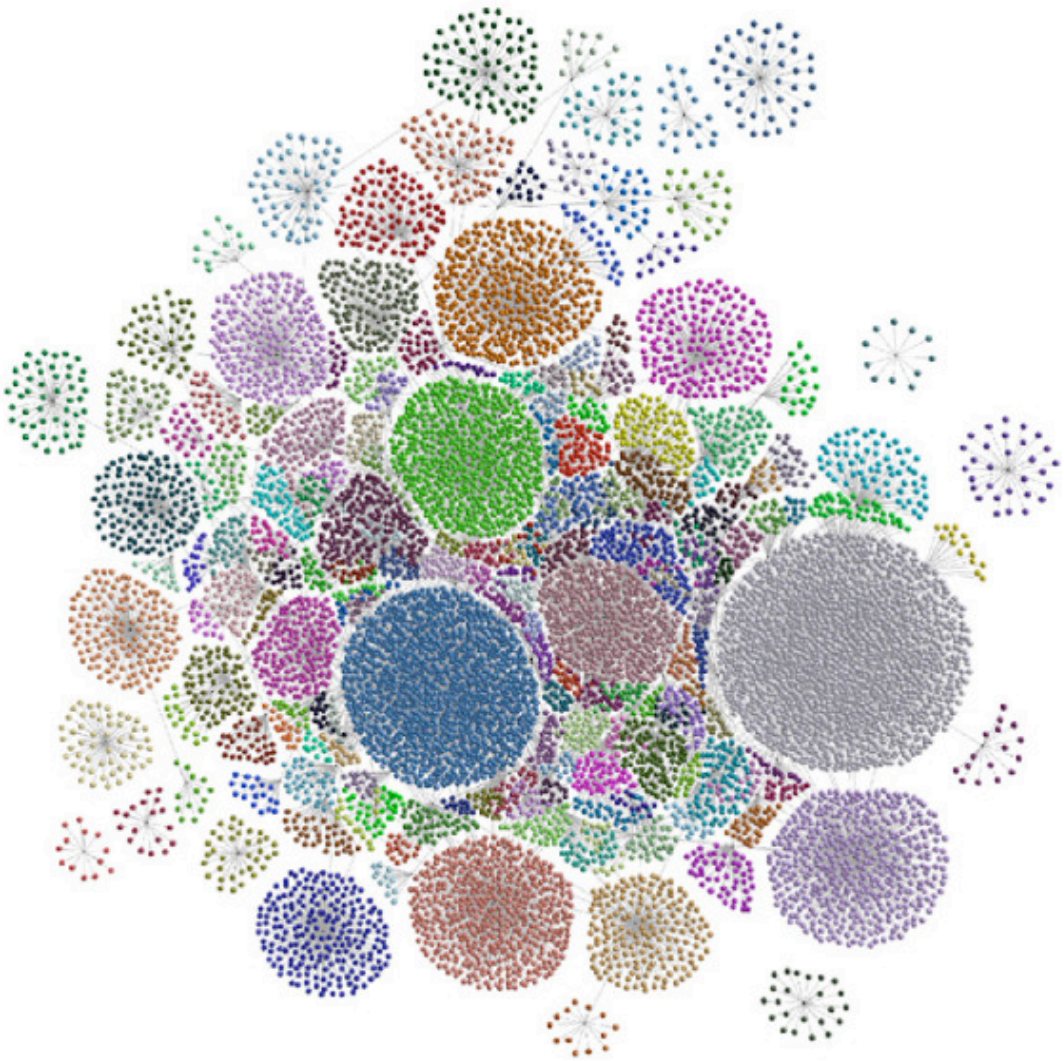
- Profs. Byung-Jun Yoon and Mingyuan Zhou

Protein-Protein Interaction Networks



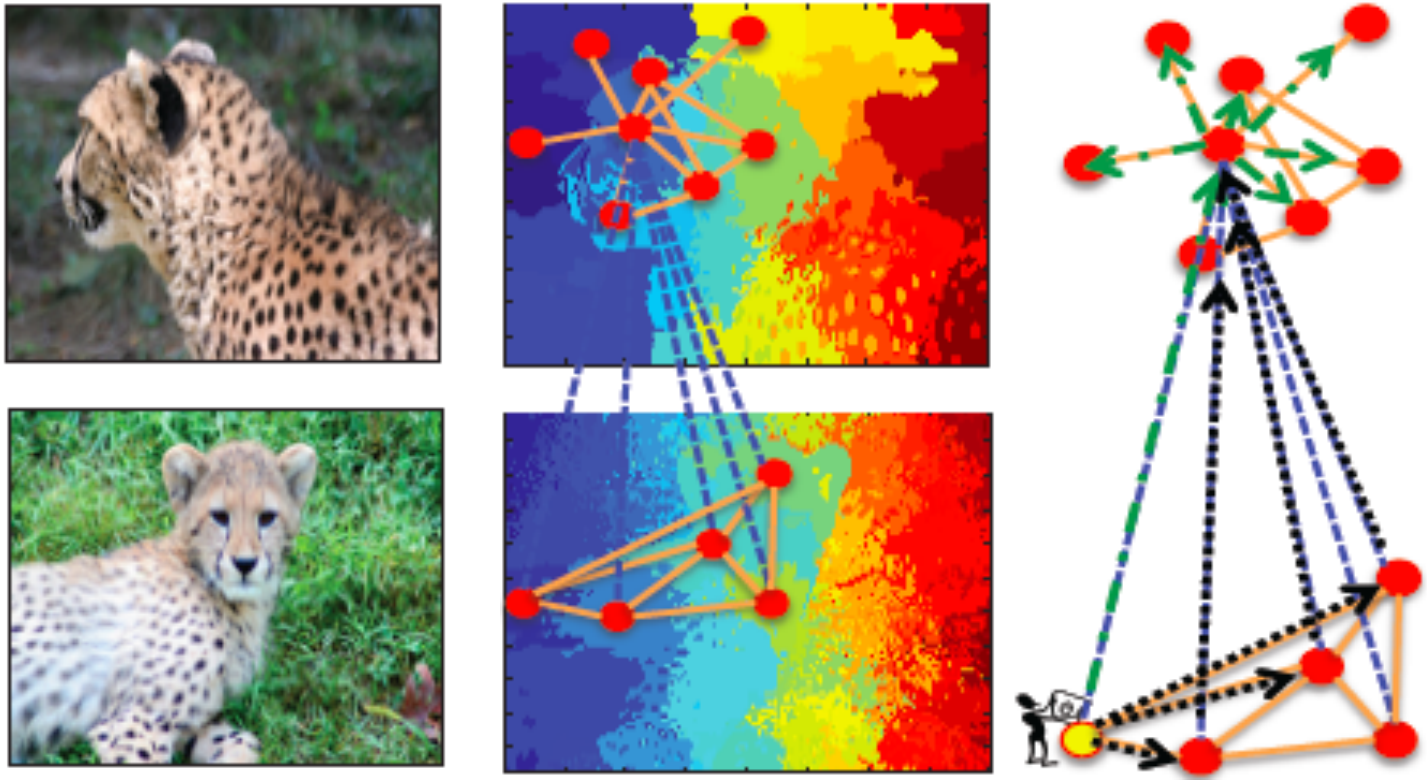
Jeong H, Mason S, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks, **Nature** 441: 41-42.

Gene Co-Expression Networks

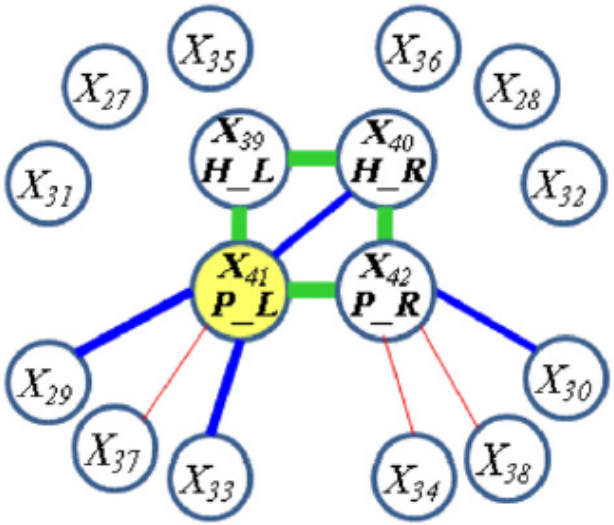


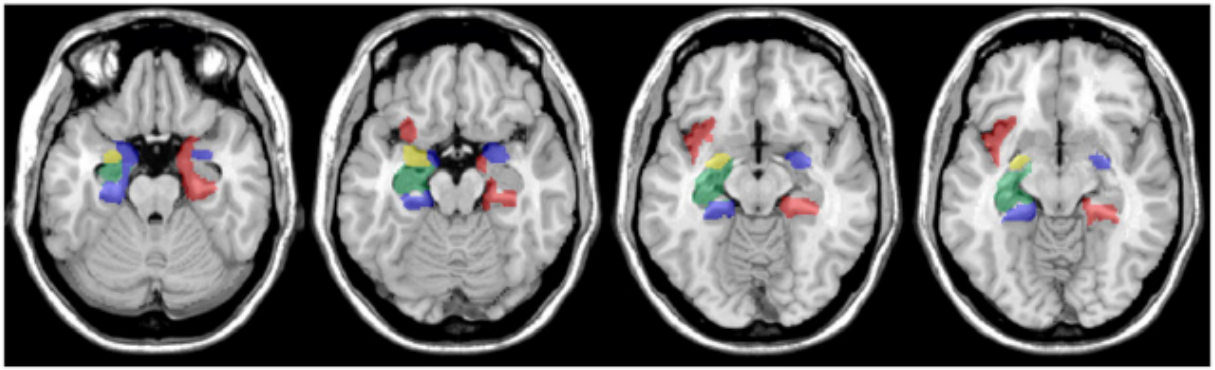
Tamasin N Doig, et al. Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment. BMC Genomics, 14:469, 2013.

Images → Graph Representations

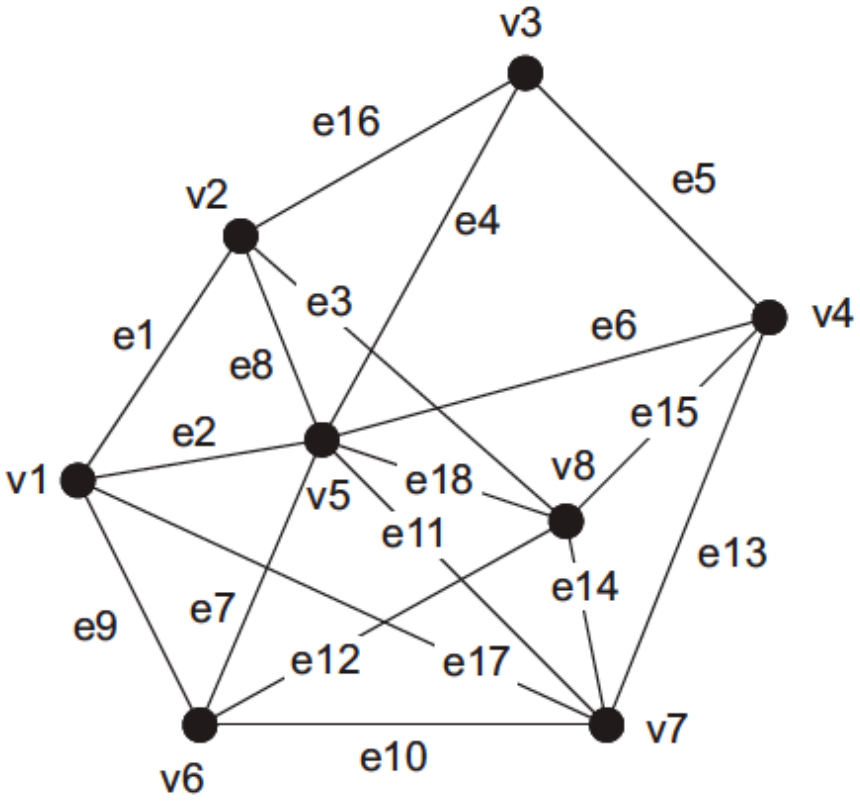


Images → Graph Representations





Networks (Graphs)



Networks (Graphs)

- Graph representation: $G = \{V, E\}$
- Adjacency matrix: $A: A_{ij} = 1 \text{ if } (i,j) \in E$
- Graph Laplacian: $L = D - A$
  - It is symmetric for undirected graphs: $L = BB^T$
  - It is non-negative definite.
  - It is directly related to graph bi-partitioning (cut size).

Network Clustering

- Clustering: assign vertices into groups such that there are many edges within groups but very few across groups.
- Graph partitioning:
  - The algorithm has to divide the given graph.
  - Number and size of partitions are typically fixed.
- Community detection:
  - Some vertices may not belong to any group.
  - Number and size of partitions are not fixed.
- Clustering is NP hard.

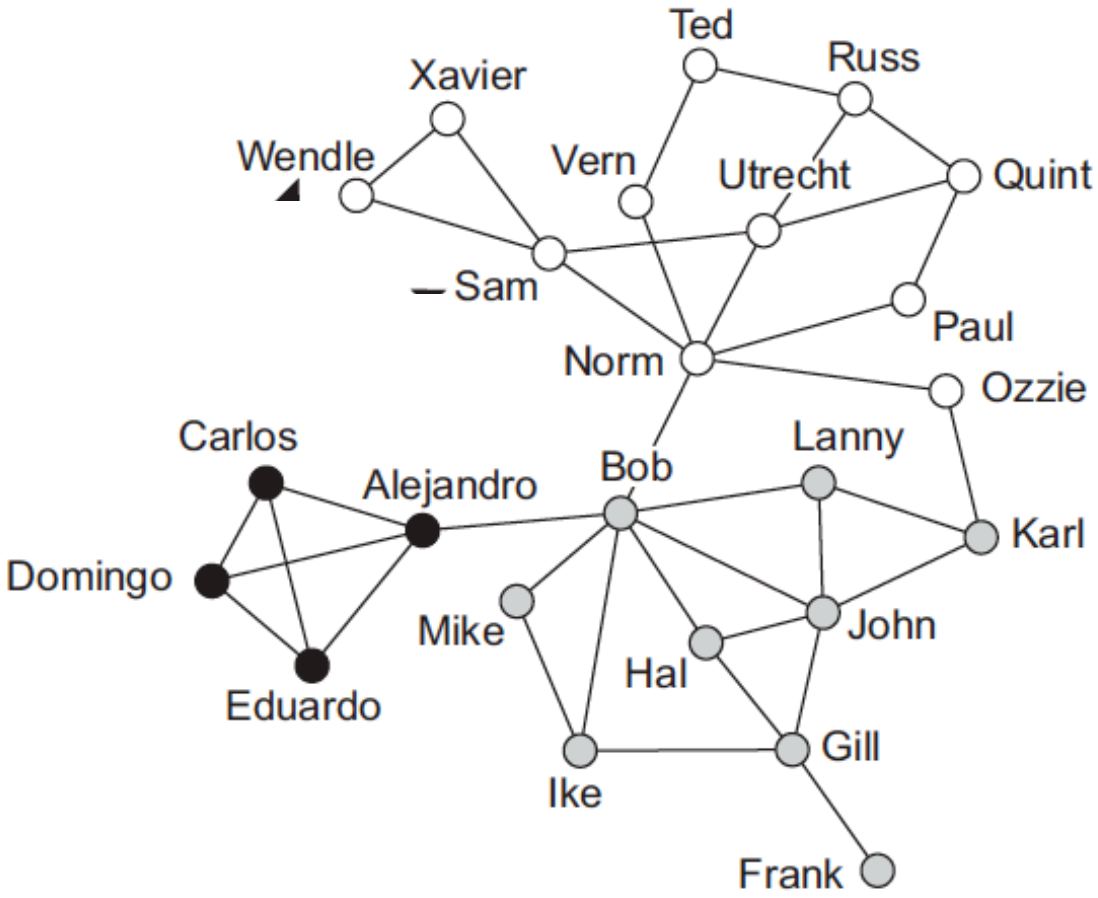
Module Identification

Question

How do we define and identify biologically meaningful modules in biological networks?

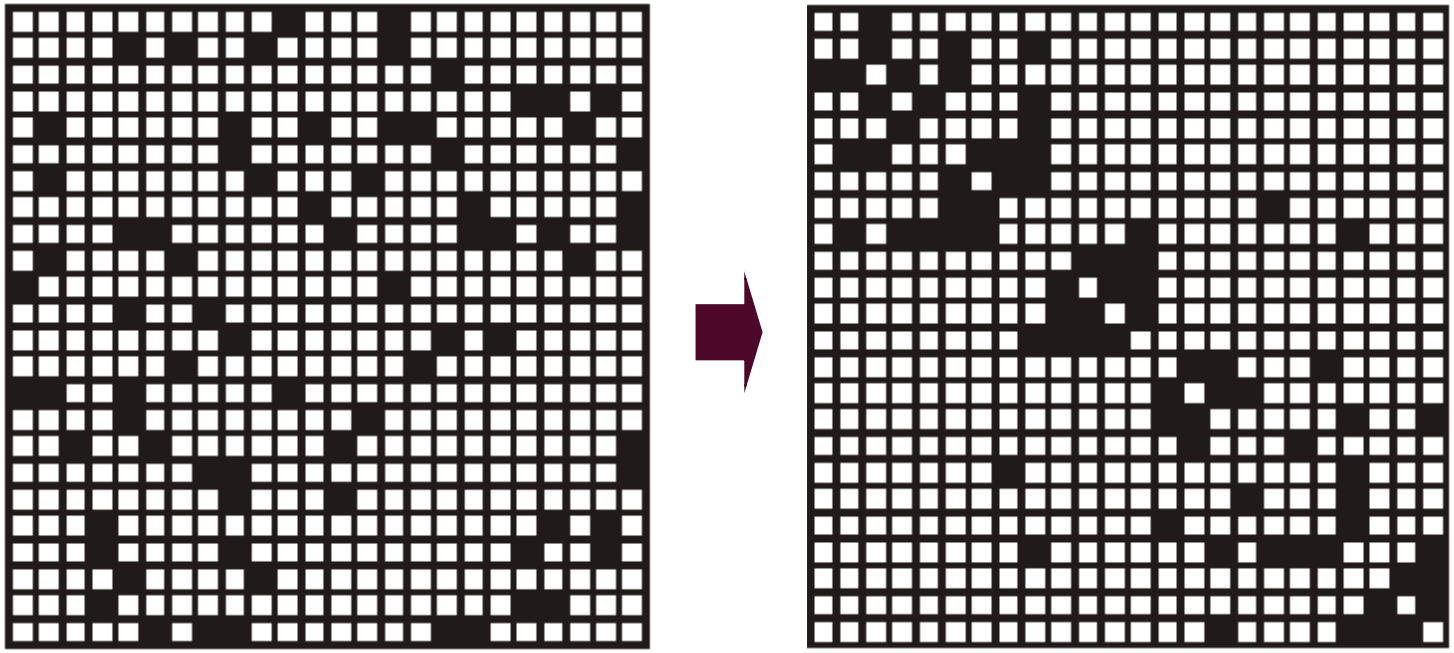
Community (Module) Identification

- Group "similar" nodes.



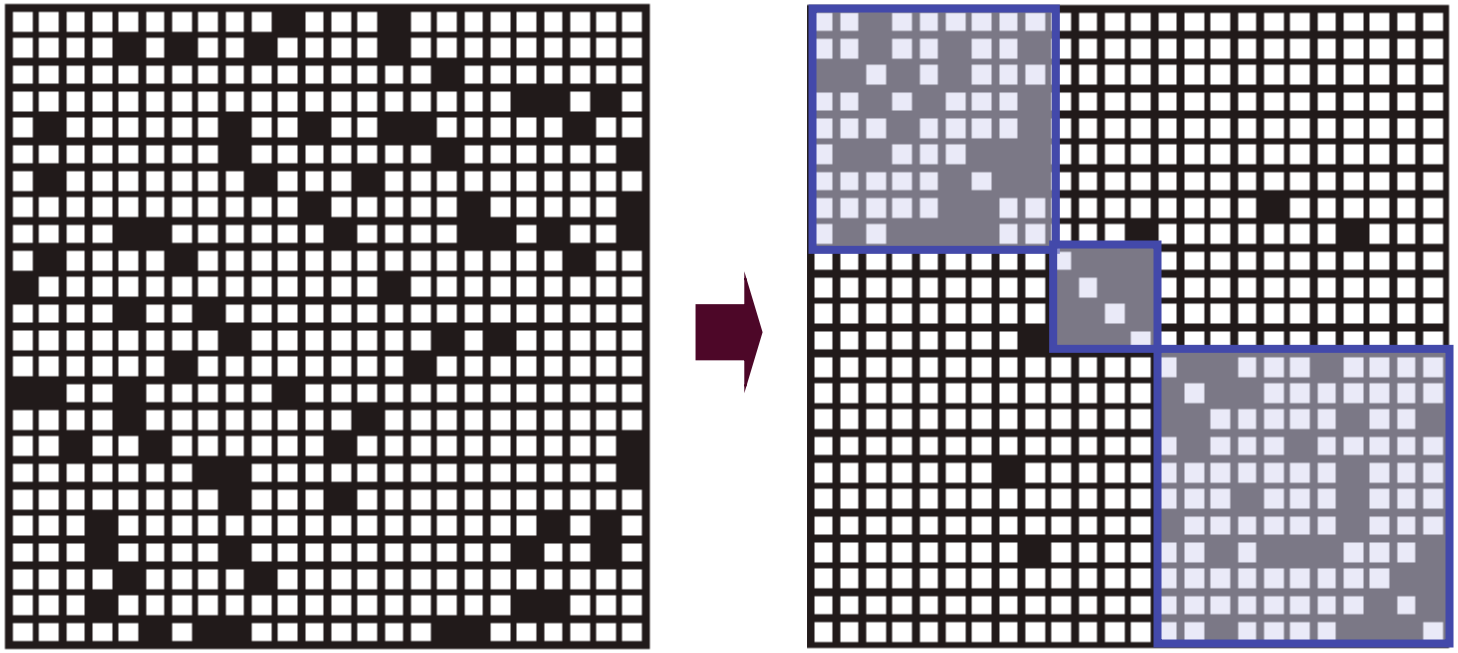
Community (Module) Identification

- Group "similar" nodes.



Community (Module) Identification

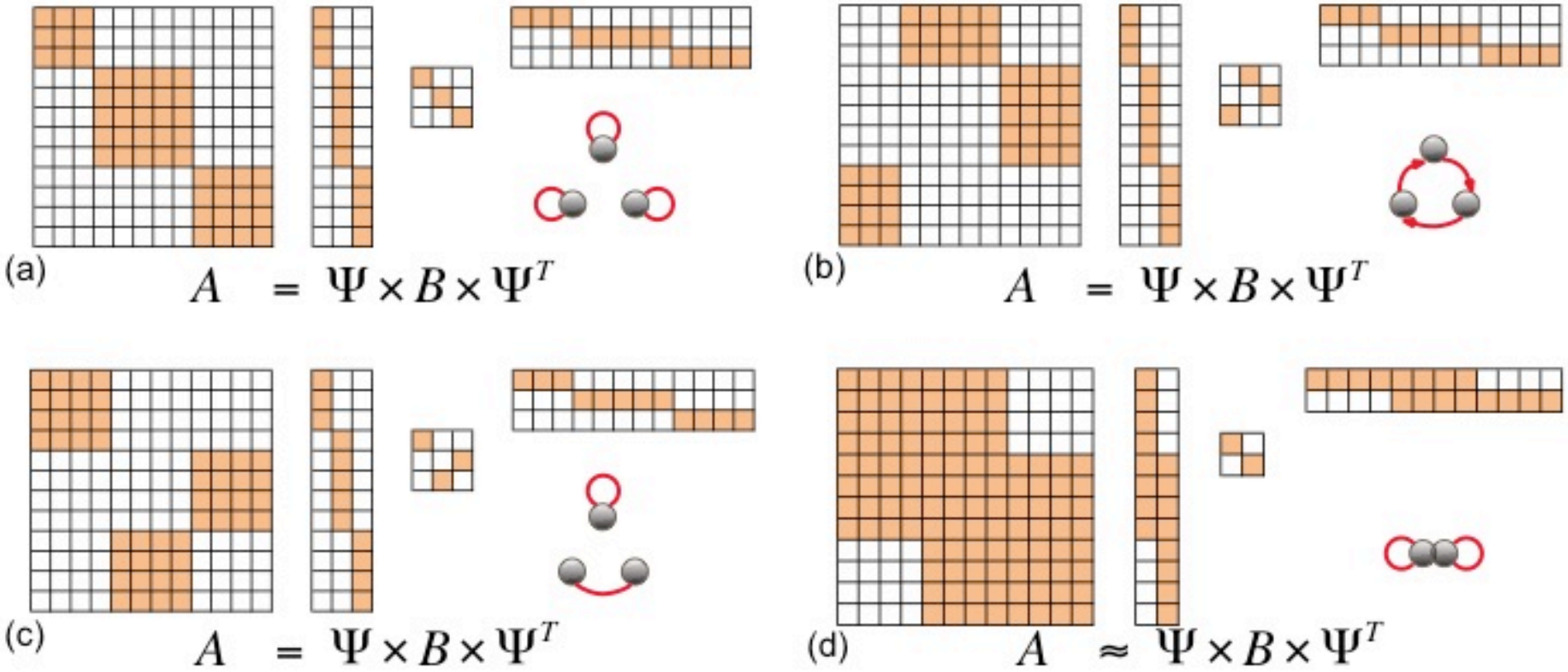
- Group "similar" nodes.



Formulations & Solution Algorithms

- Group "similar" nodes.
  - "Modularity"-based formulations
    - Clique, k-plex, other network motif based algorithms
    - Heuristic algorithms (greedy)
    - Integer programming
    - Spectral algorithms – Random Walk on Graph
    - Hybrid algorithms
  - Interaction-pattern-based formulations
    - Non-negative matrix factorization (NMF)
    - Edge partition models (Stochastic Block Models)

Non-negative Matrix Factorization



Wang Y and Qian X (2016) Clustering of Noisy Graphs via Non-negative Matrix Factorization With Sparsity Regularization. Texas A&M ECE Tech Report (http://www.ece.tamu.edu/~xqian/FMItutorial.html)

Modularity

- Topologically, nodes within the same modules have higher than "expected" connectivity.
  - This is a "diagonal" block model.
  - Many existing algorithms search for these densely connected modules.

$$\max_{\mathbf{c}} Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

Modularity

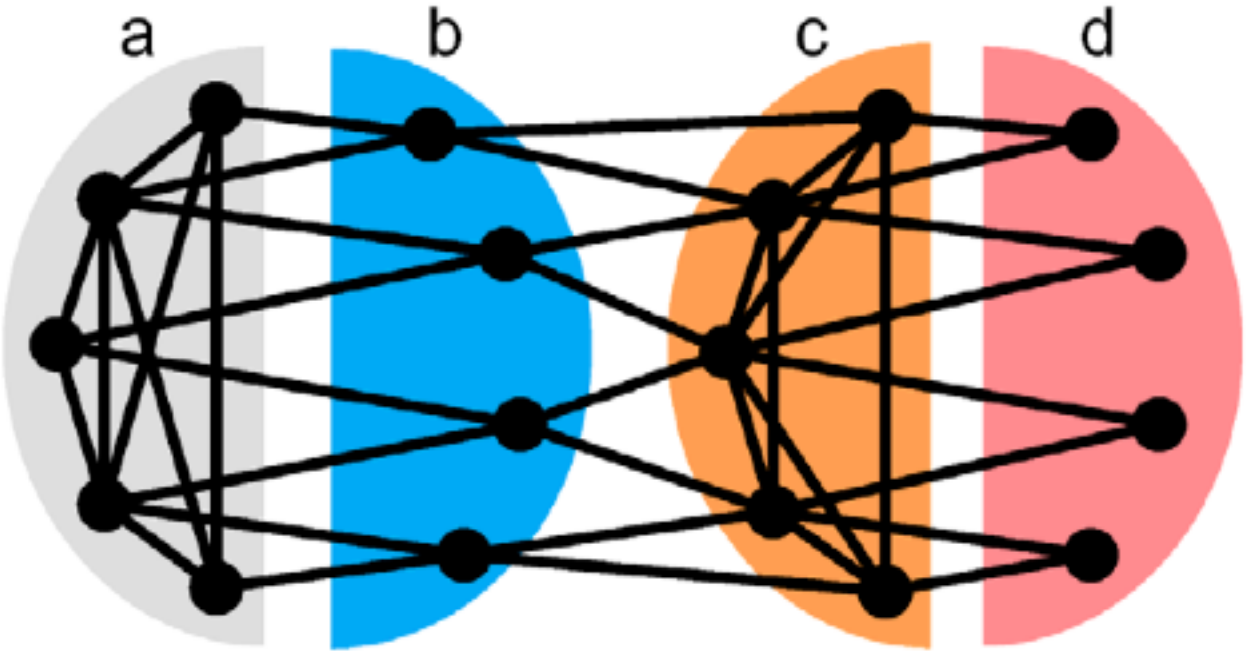
- Topologically, nodes within the same modules have higher than "expected" connectivity.
  - This is a "diagonal" block model.
  - Many existing algorithms search for these densely connected modules.

Question

Do molecules always work together by dense connections?

Blockmodel Module Identification

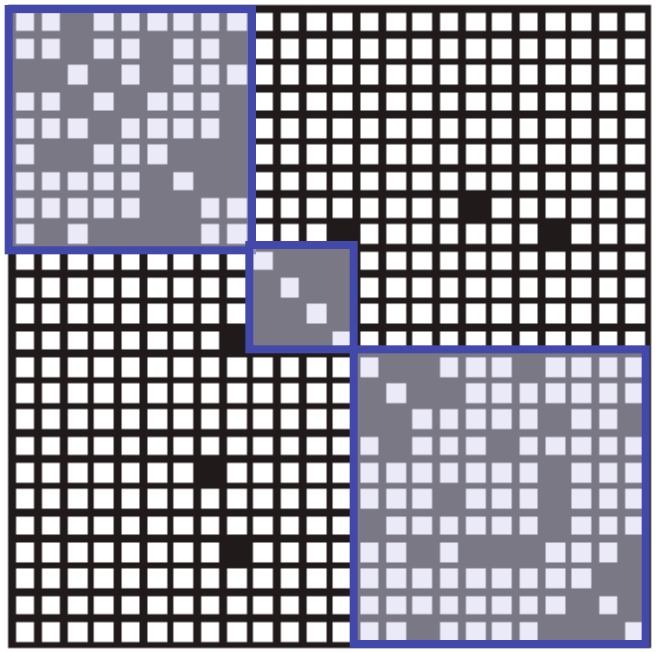
- Signal transduction (e.g. transmembrane) proteins do not have high connectivity among themselves but interact with similar other types of proteins.



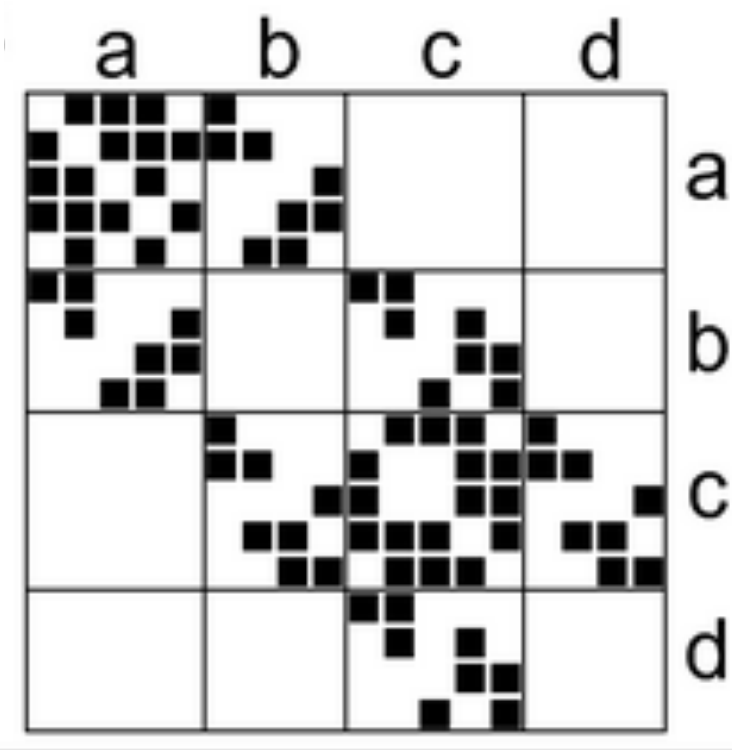
Pinkert S, *et al.*. (2010) Protein interaction networks---more than mere modules, **PLoS Comput Biol 6**(1).

Blockmodel Module Identification

- Signal transduction (e.g. transmembrane) proteins do not have high connectivity among themselves but interact with similar other types of proteins.



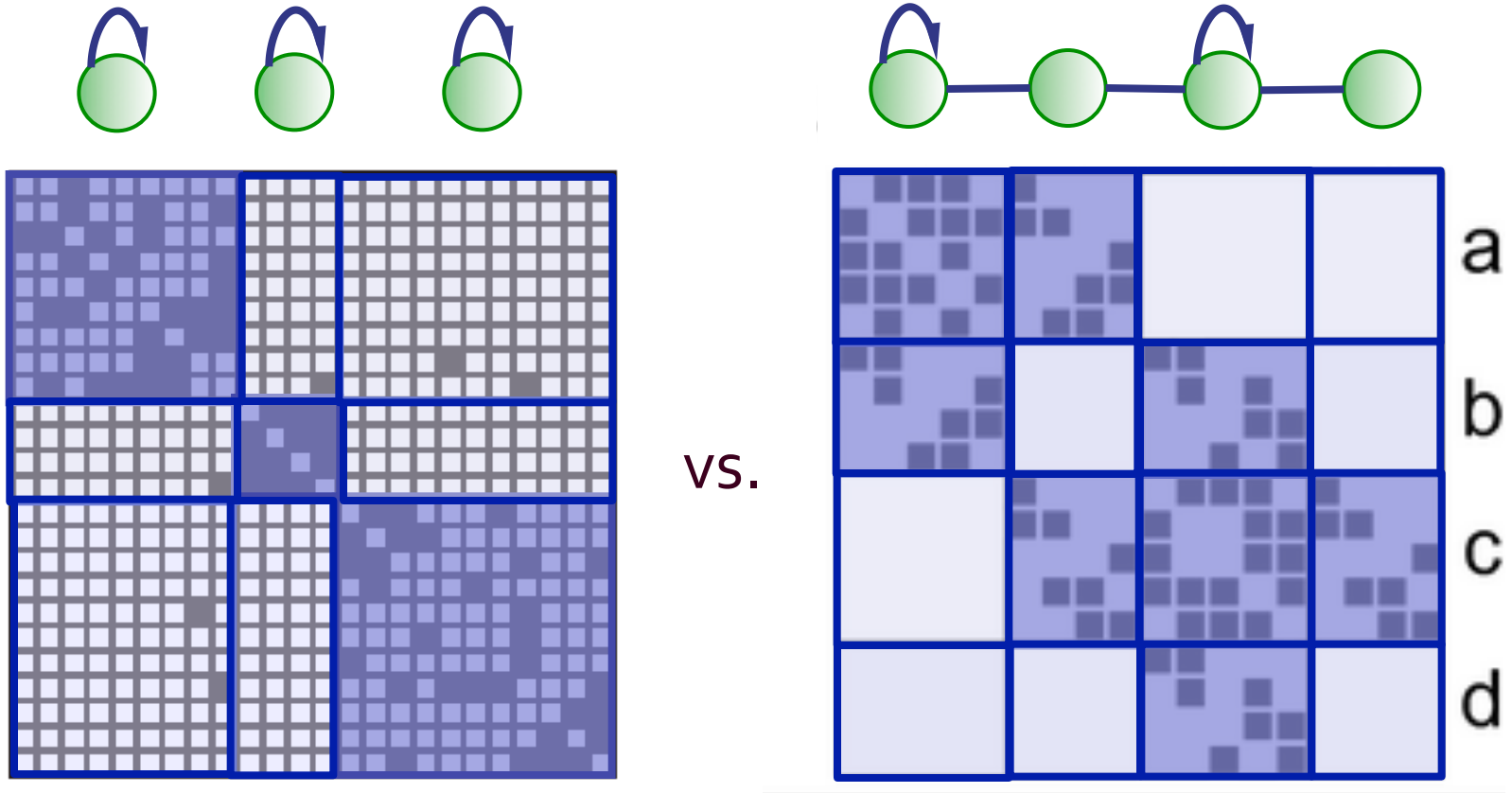
vs.



Pinkert S, *et al.*. (2010) Protein interaction networks---more than mere modules, **PLoS Comput Biol 6**(1).

Blockmodel "Modularity"

- More general blockmodel "modularity" by introducing a virtual image graph



Blockmodel "Modularity"

- More general blockmodel "modularity" by introducing a virtual image graph

$$\min_{\mathbf{c}} \frac{1}{2m} \sum_{ij} (A_{ij} - B_{c_i c_j})(w_{ij} - p_{ij})$$

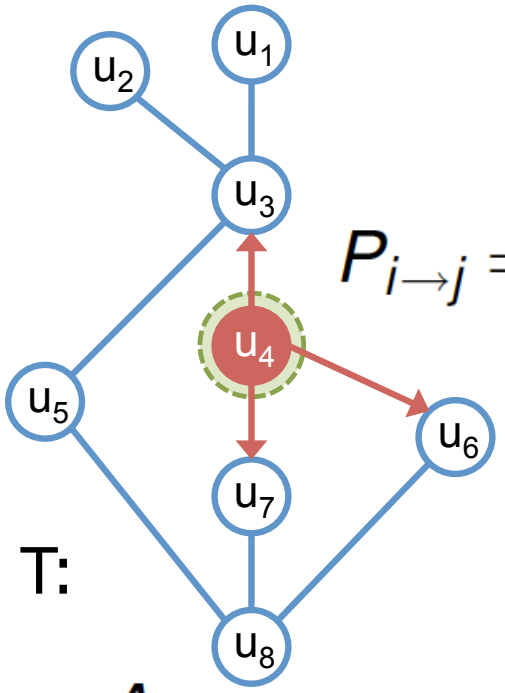
$$\max_{\mathbf{c}} Q^* = \frac{1}{2m} \sum_{kl}^{q} \left| \sum_{ij} (w_{ij} - p_{ij}) \delta(c_i, k) \delta(c_j, l) \right|$$

  - However, it is computationally hard to get the global optimum due to the inherent combinatorial complexity of the resulting optimization problem.

Random Walk on Graphs

- Random walk on graphs is a Markov chain
  - Transition matrix: $P = D^{-1}A$
  - For a connected graph, there is a steady-state distribution:

$$\pi_i = \frac{k_i}{2m}$$



$$P_{i \to j} = \frac{A_{ij}}{k_i}$$

  - Transition probability from set S to T:

$$P_{S \to T} = \frac{\sum_{i \in S; j \in T} \pi_i P_{i \to j}}{\pi(S)} = \frac{\sum_{i \in S; j \in T} A_{ij}}{\sum_{i \in S} k_i}$$

Random Walk and Graph Partitioning

- Solving graph cut:
  - Graph Laplacian: $L = D - A$
  - The second smallest eigenvalue quantifies connectivity.
  - Normalized cut is for graph partition by solving:

$$Lx = \lambda Dx$$

  - Essentially, it uses the second smallest eigenvector of

$$D^{-1}(D - A)$$

Maila M and Shi J (2001) A random walks view of spectral segmentation , **AI and STATISTICS** 2001.

Random Walk and Graph Partitioning

- Solving graph cut:
  - Graph Laplacian: $L = D - A$
  - The second smallest eigenvalue quantifies connectivity.
  - Normalized cut is for graph partition by solving:

$$Lx = \lambda Dx$$

  - Essentially, it uses the second smallest eigenvector of

$$D^{-1}(D - A) \qquad \text{vs.} \qquad P = D^{-1}A$$

Maila M and Shi J (2001) A random walks view of spectral segmentation , **AI and STATISTICS** 2001.

Random Walk and Graph Partitioning

- Normalized cut:
  - The objective function of normalized cut is for balanced graph partition:

$$NCut(S,\bar{S}) = \left(\frac{1}{\sum_{i\in S} k_i} + \frac{1}{\sum_{i\in \bar{S}} k_i}\right) \sum_{i\in S, j\in \bar{S}} A_{ij}$$

$$NCut(S,\bar{S}) = P_{S\to\bar{S}} + P_{\bar{S}\to S}$$

  - We can define this as the conductance of S on graph:

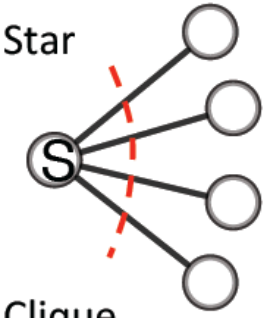
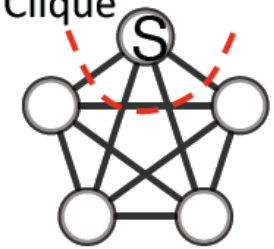
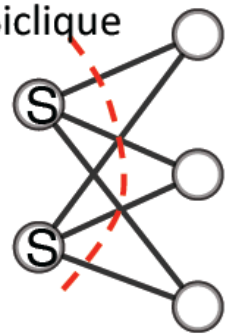
$$\Phi_P(S) = P_{S\to\bar{S}}$$

  - Hypothesis: Functional modules have low conductance to the rest of the graph.

Maila M and Shi J (2001) A random walks view of spectral segmentation , **AI and STATISTICS** 2001.

Blockmodel Module Identification

- Do we capture general blockmodel modules with conductance?

Motifs	Transition Matrix (P)	$\Phi(S)$
Star 	$\begin{matrix} 0 & 1 & 1 & 1 & 1 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \end{matrix}$	1
Clique 	$\begin{matrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{matrix}$	1
Biclique 	$\begin{matrix} 0 & 0 & 1/2 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 1/2 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \end{matrix}$	1

Wang Y and Qian X (2014) Blockmodel module identification in PPI through Markov random walk, **Bioinformatics** doi:btt569.

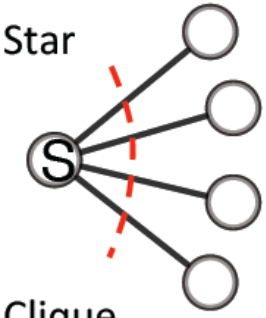
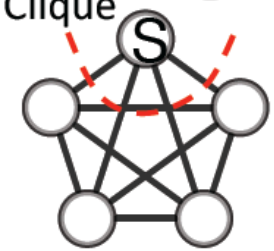
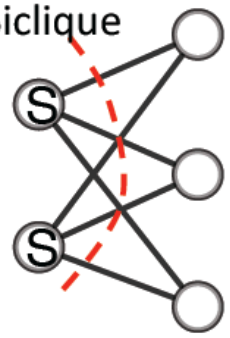
Blockmodel Module Identification

- Can we define a new conductance which captures interaction patterns instead of "modularity"?
  - Intuition: If two nodes interact with similar nodes, it is more probable that they can reach each other in two steps.
  - Two-hop transition matrix: $P^2 = P \times P$
  - As the steady-state distribution does not change, we can define a new two-hop conductance:

$$\Phi_{P^2}(S) = P^2_{S \to \bar{S}}$$

Blockmodel Module Identification

- Do we capture general blockmodel modules with conductance?

Motifs	Transition Matrix (P)	$\Phi(S)$	Transition Matrix ($P \times P$)	$\Phi(S)$
Star 	$\begin{matrix} 0 & 1 & 1 & 1 & 1 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \end{matrix}$	1	$\begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \end{matrix}$	0
Clique 	$\begin{matrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{matrix}$	1	$\begin{matrix} 1/4 & 3/16 & 3/16 & 3/16 & 3/16 \\ 3/16 & 1/4 & 3/16 & 3/16 & 3/16 \\ 3/16 & 3/16 & 1/4 & 3/16 & 3/16 \\ 3/16 & 3/16 & 3/16 & 1/4 & 3/16 \\ 3/16 & 3/16 & 3/16 & 3/16 & 1/4 \end{matrix}$	3/4
Biclique 	$\begin{matrix} 0 & 0 & 1/2 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 1/2 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \end{matrix}$	1	$\begin{matrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \end{matrix}$	0

Wang Y and Qian X (2014) Blockmodel module identification in PPI through Markov random walk, **Bioinformatics** doi:btt569.

Blockmodel Module Identification

- We search for modules as low conductance sets:

$$\min_{S_1,\ldots,S_q} \sum_i \Phi_{P2}(S_i)$$

$$s.t. \qquad \cup_i S_i = V;\ S_i \cap S_j = \oslash, \forall i \neq j$$

  - Module assignment matrix (binary): X n-by-q
  - After algebraic manipulations, we can prove that

$$\min_{S_1,\ldots,S_q} \sum_i \Phi_{P2}(S_i) \Leftrightarrow \max_X tr\left(\frac{X^T A D^{-1} A X}{X^T D X}\right)$$

Blockmodel Module Identification

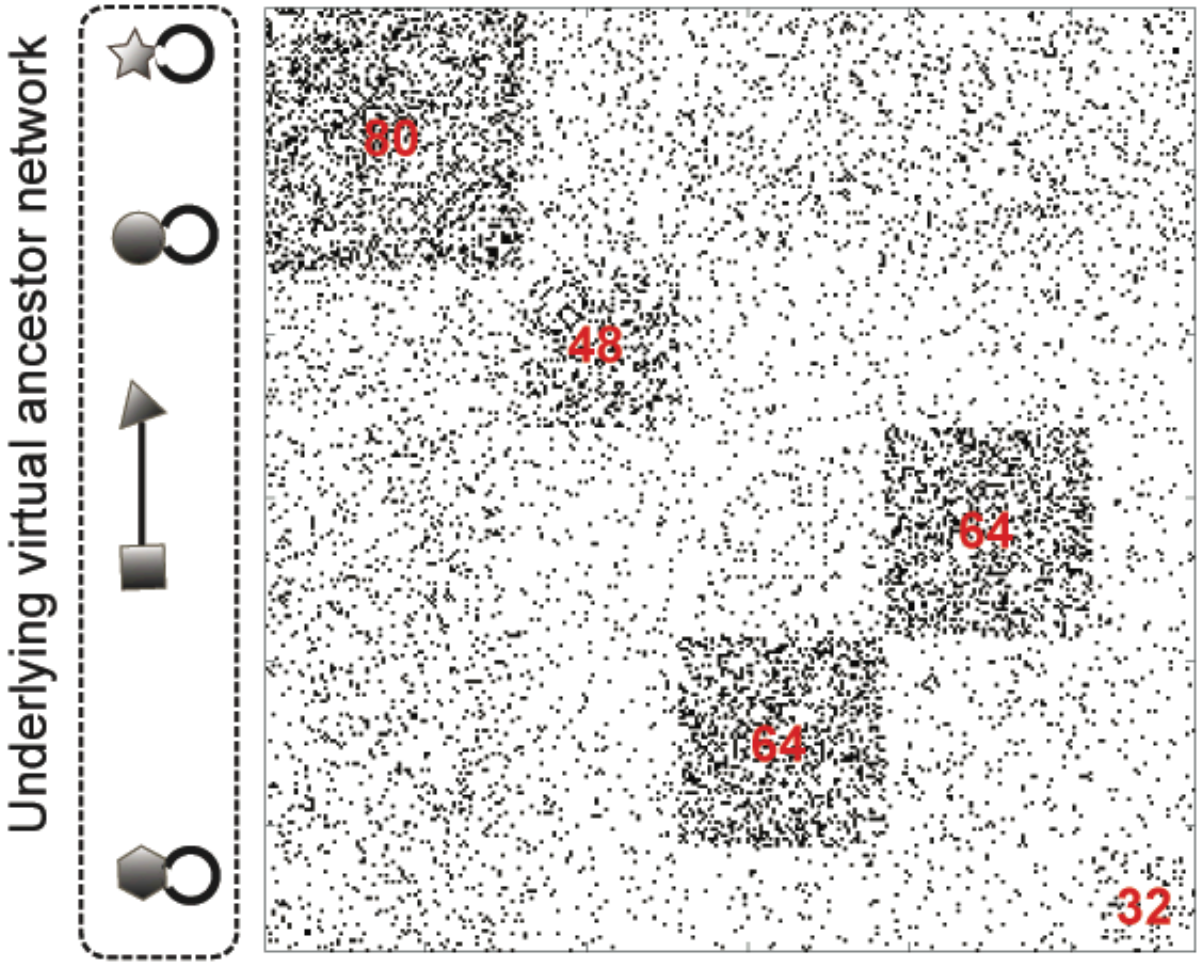
- We can solve :

$$\max_X tr\left(\frac{X^T A D^{-1} A X}{X^T D X}\right)$$
$$s.t. \quad X1_q = 1_n; X \in \{0,1\}^{n \times q}$$

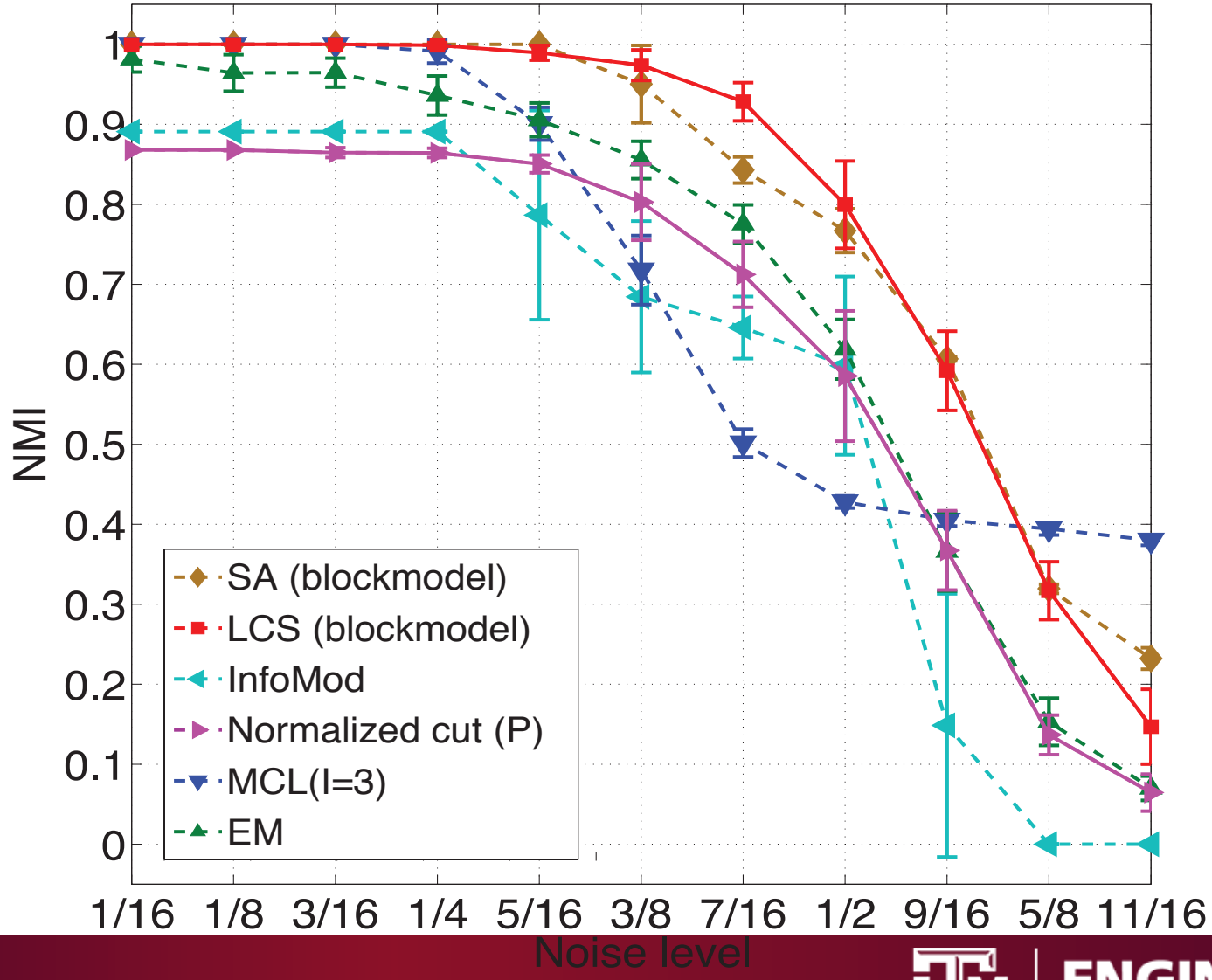
  - This final optimization problem can be solved by semi-definite programming or spectral approximate method.

Experimental Results

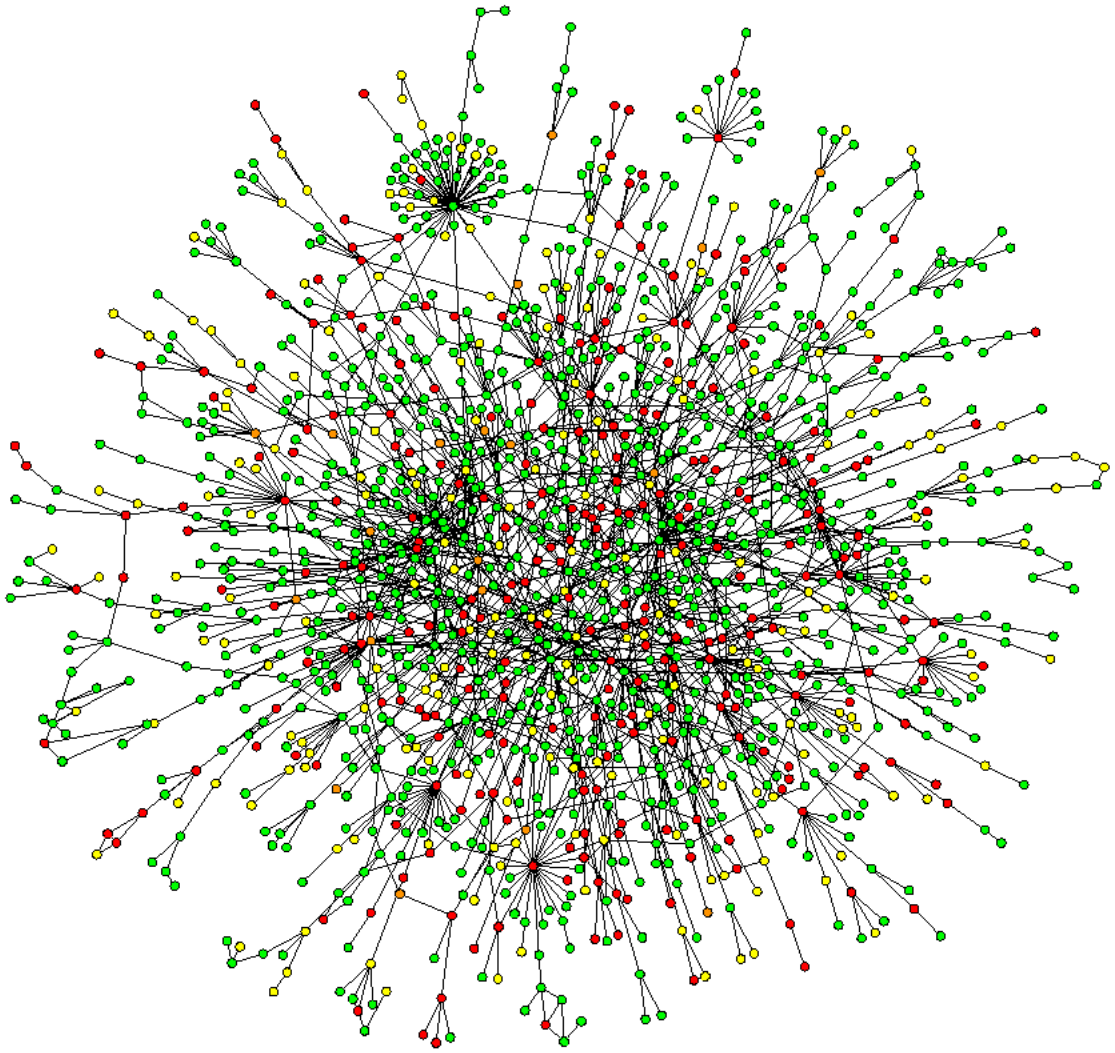
Synthetic Networks



Synthetic Networks



Biological Networks



Biological Networks

We collect yeast (*Sce*) PPI network from DIP and human (*Hsa*) PPI network from HPRD.

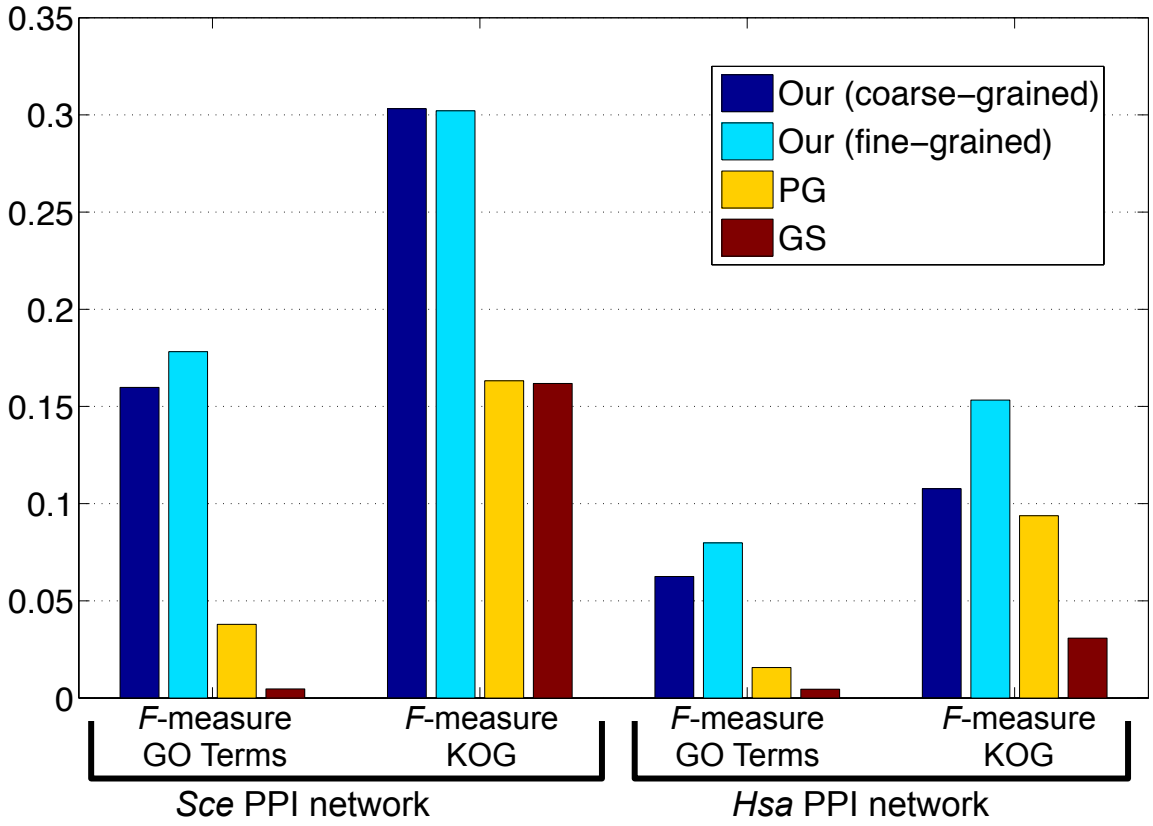
There is no ground truth regarding functional modules in these real-world PPI networks.

The performance is measured by F-measures based on Gene Ontology (GO) terms and KOG categories.

We check whether an identified module can be associated with specific GO terms or KOG categories.

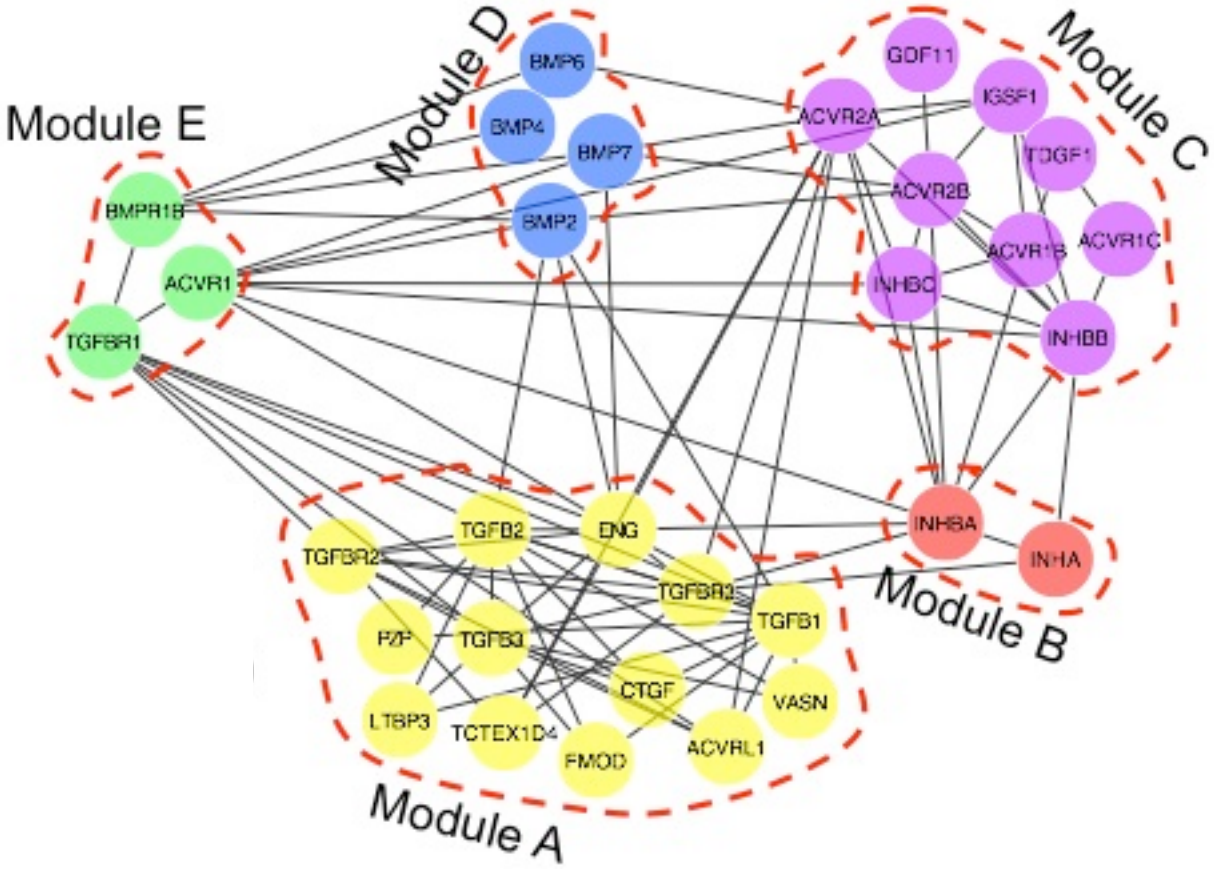
$$F = 2 \times precision \times recall / (precision + recall)$$

Biological Networks

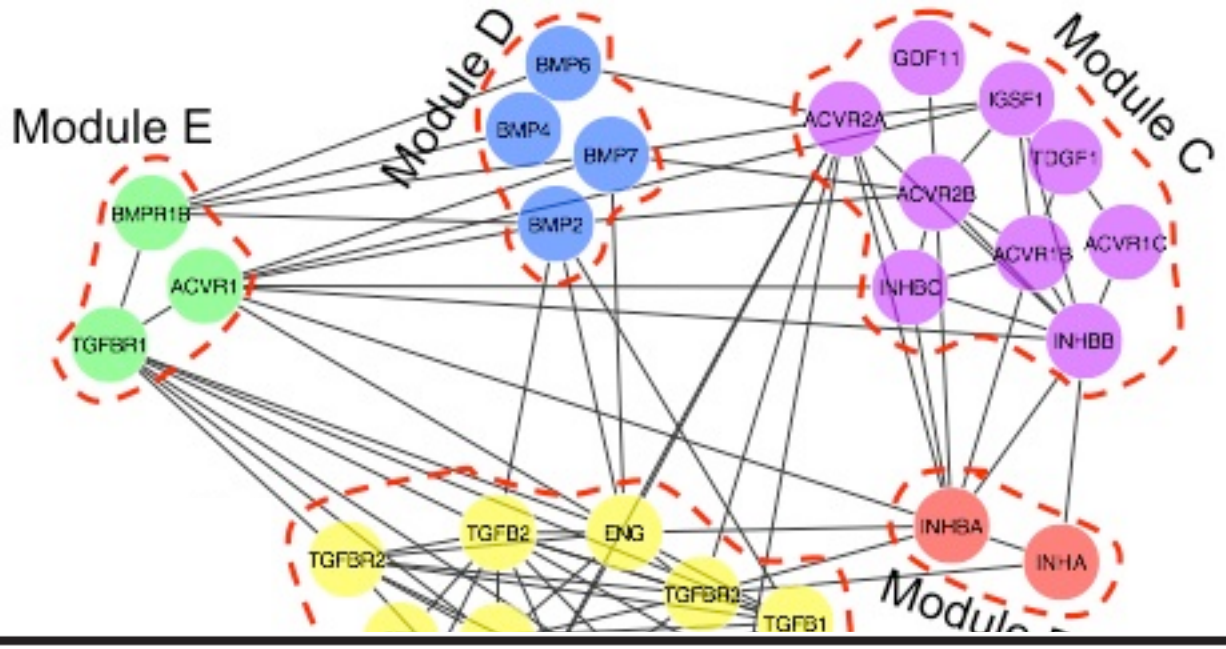


Coarse-grained: Average module size is around 10.
Fine-grained: Average module size is around 5.

Biological Networks

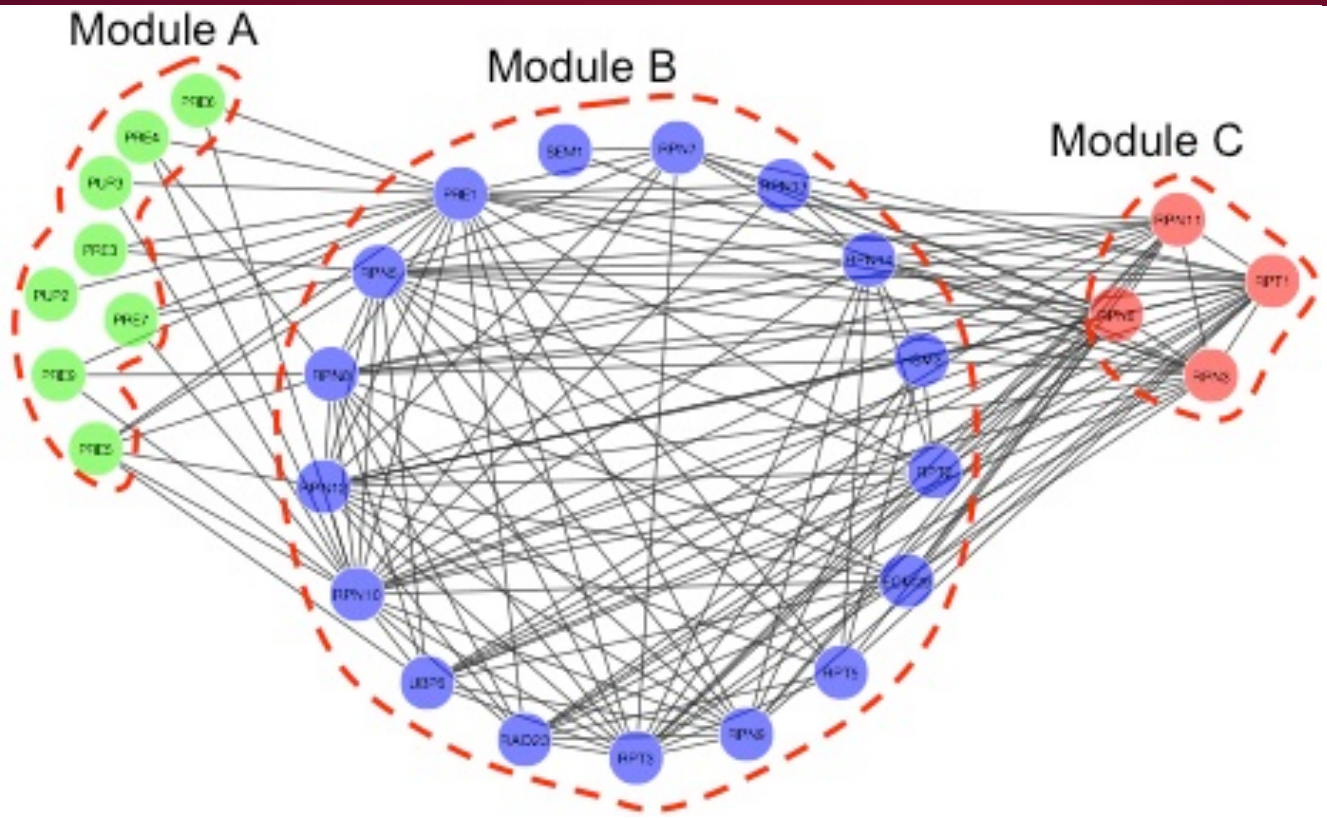


Biological Networks

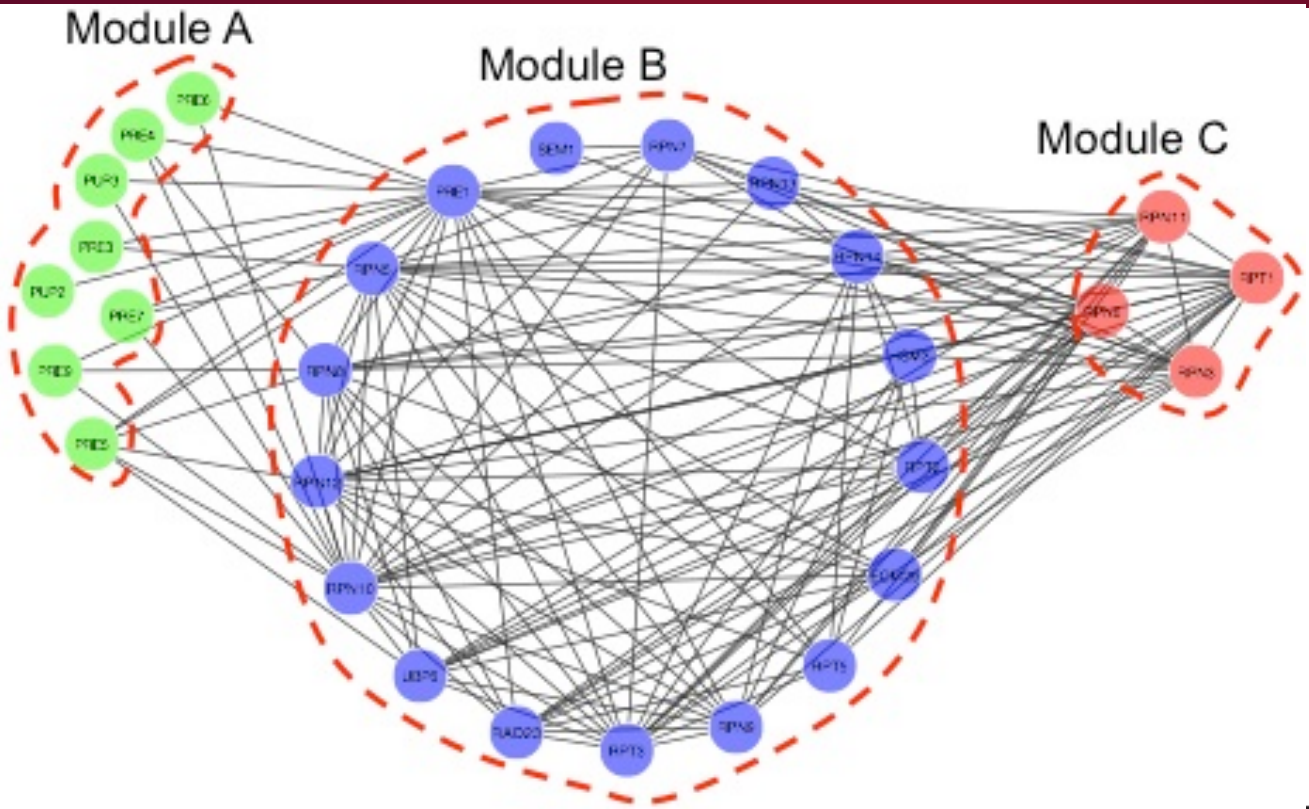


Modules	Enriched GO terms	p-value
A	transforming growth factor beta	4.30e-7
B	hemoglobin biosynthetic process	6.51e-5
C	transmembrane receptor protein serine	9.03e-9
D	fibroblast growth factor	1.15e-7
E	transforming growth factor	7.06e-8

Biological Networks



Biological Networks



Modules	Enriched GO terms	*p*-value
A	proteasome core complex	6.42e-21
B	proteasome complex	4.30e-32
C	proteasome regulatory particle	3.81e-9

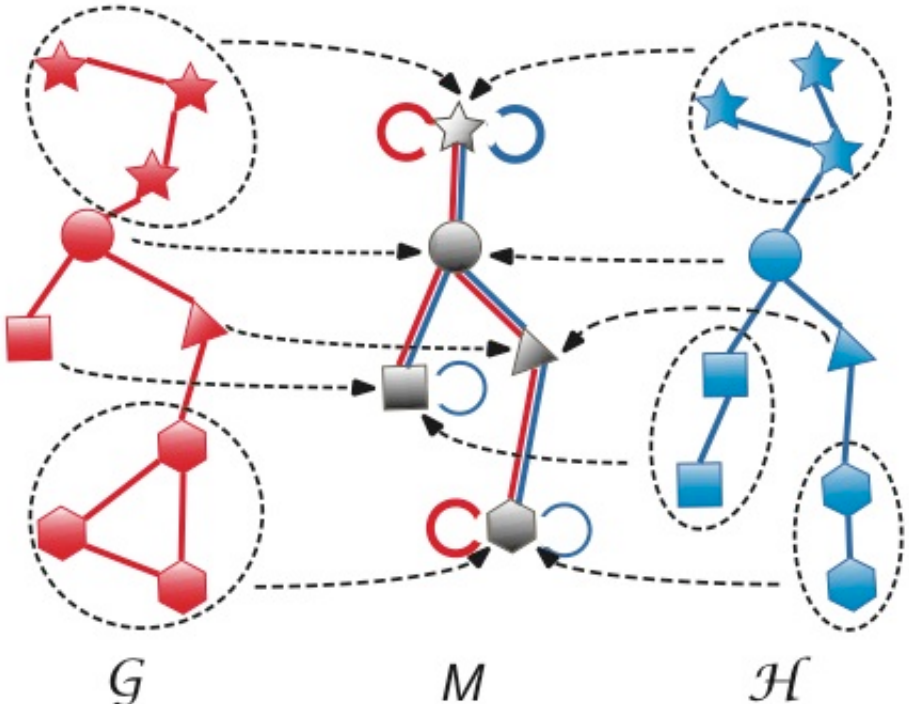
Ongoing Research

- The current publicly accessible data may not be complete or accurate. → Will integration of network data help improve clustering?
  1. Random Walk across networks
  2. Generative models: Edge Partition Models with Bayesian Computation

- What about vertex properties?
  1. Deep models: Graph Convolutional Networks

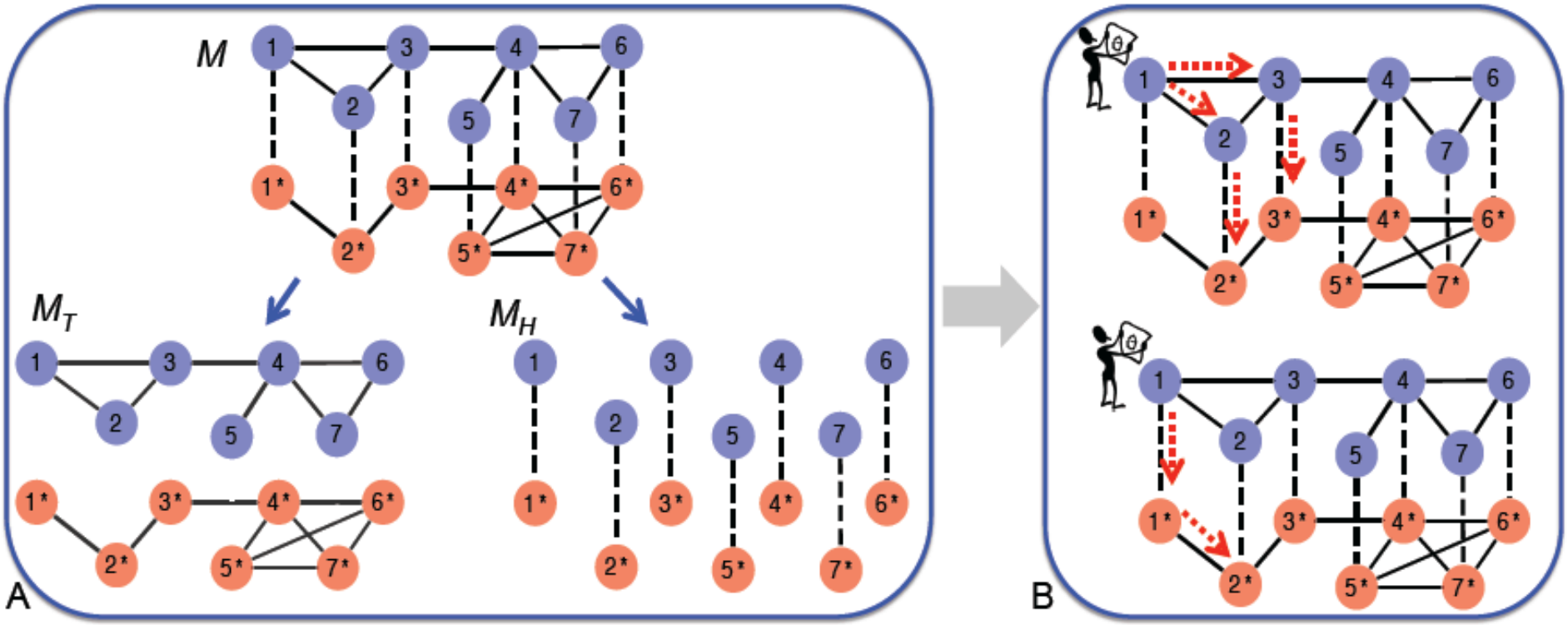
Interactomic data are noisy.

- The current publicly accessible data may not be complete or accurate.
  - Can we borrow strengths from data across different data sources or even different organisms?



Random Walk across Networks

- Random walk across networks to integrate both topology and constituent similarity between nodes.



Wang Y and Qian X (2014) Joint clustering of protein interaction networks through Markov random walk, **APBC 2014**.

Random Walk across Networks

- Random walk across networks can either first take a step within a network or take a step across networks.
- We focus on the two-hop random walk again:

$$P = \frac{1}{2}P_{A\bar{S}} + \frac{1}{2}P_{S\bar{A}}$$

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}_{N\times N} \qquad S = \begin{bmatrix} 0 & S_{12} \\ S_{12}^T & 0 \end{bmatrix}_{N\times N}$$

Blockmodel Module Identification

- We can similarly solve :

$$\begin{aligned} \max \quad & trace\left(\frac{X^T \bar{P} X}{X^T D_{\bar{P}} X}\right) \\ s.t. \quad & X 1_k = 1_N, x_{i\ell} \in \{0,1\} \end{aligned}$$

- We can again solve this by semi-definite programming or spectral approximate method.
- Note that the time complexity is linear with respect to the number of networks.

Experimental Results

Biological Networks

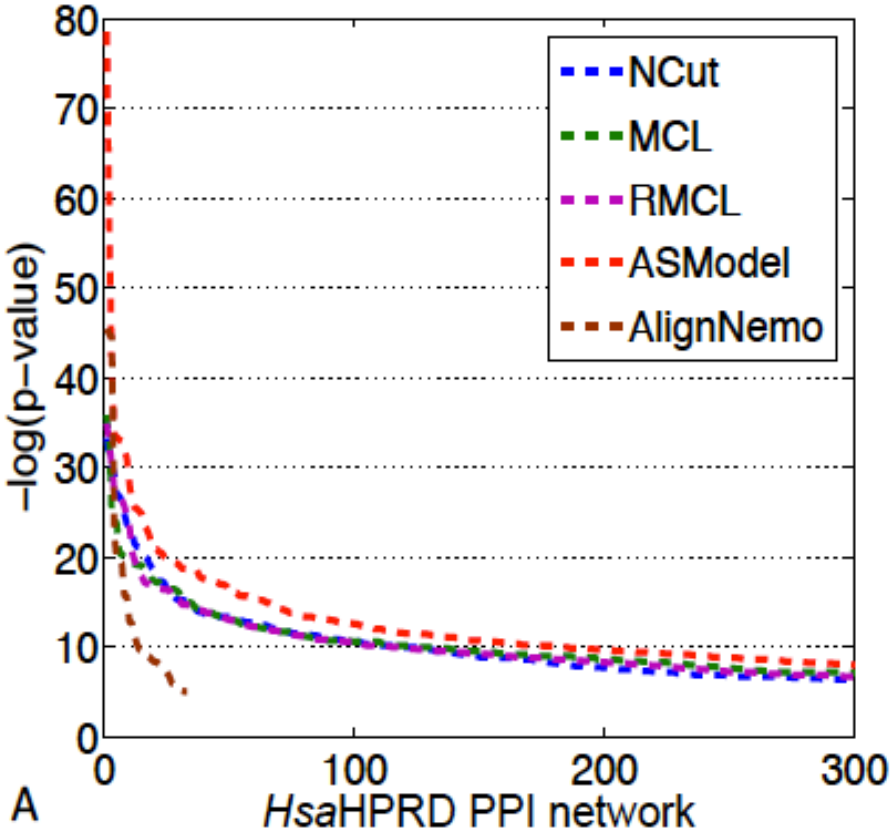
We construct two human (*Hsa*) PPI networks from HPRD as well as from PIPs.

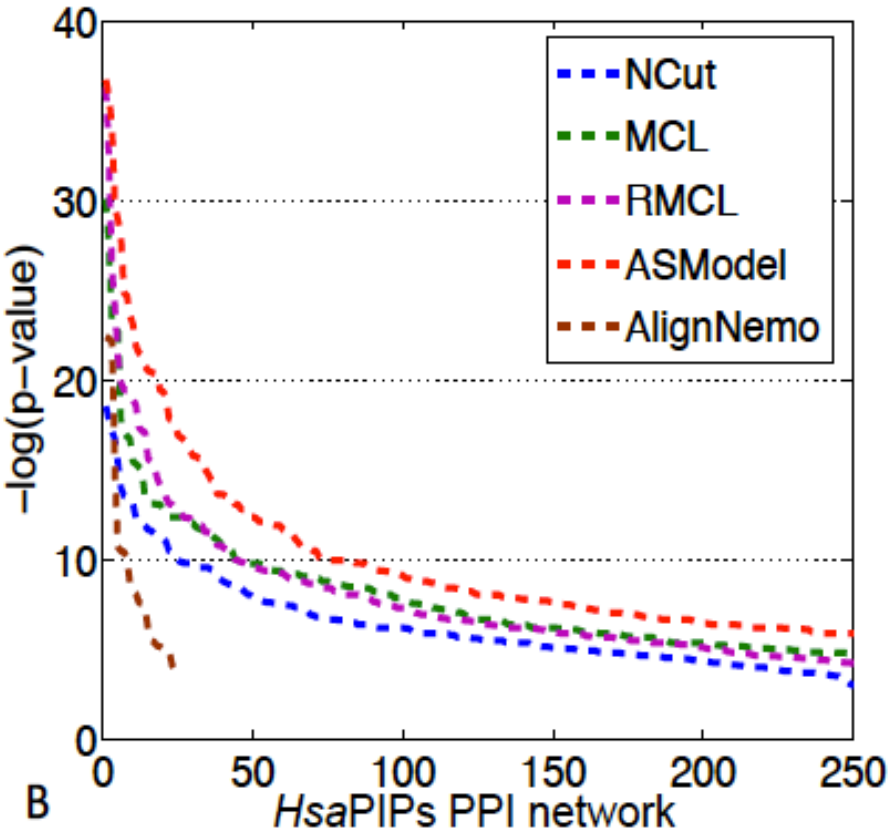
We also simultaneously cluster human PPI network from HPRD and yeast (*Sce*) network from DIP.

There is no ground truth regarding functional modules in these real-world PPI networks.

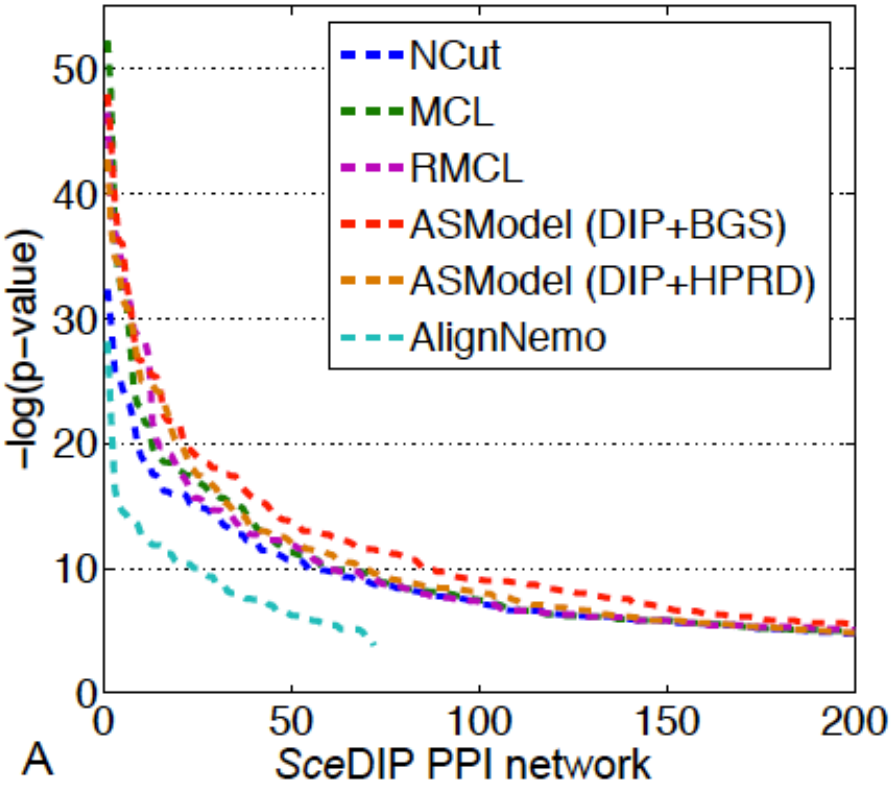
We compare the top GO enriched clusters based on the average –log(p-value).

Two is more than one.





Two is more than one.



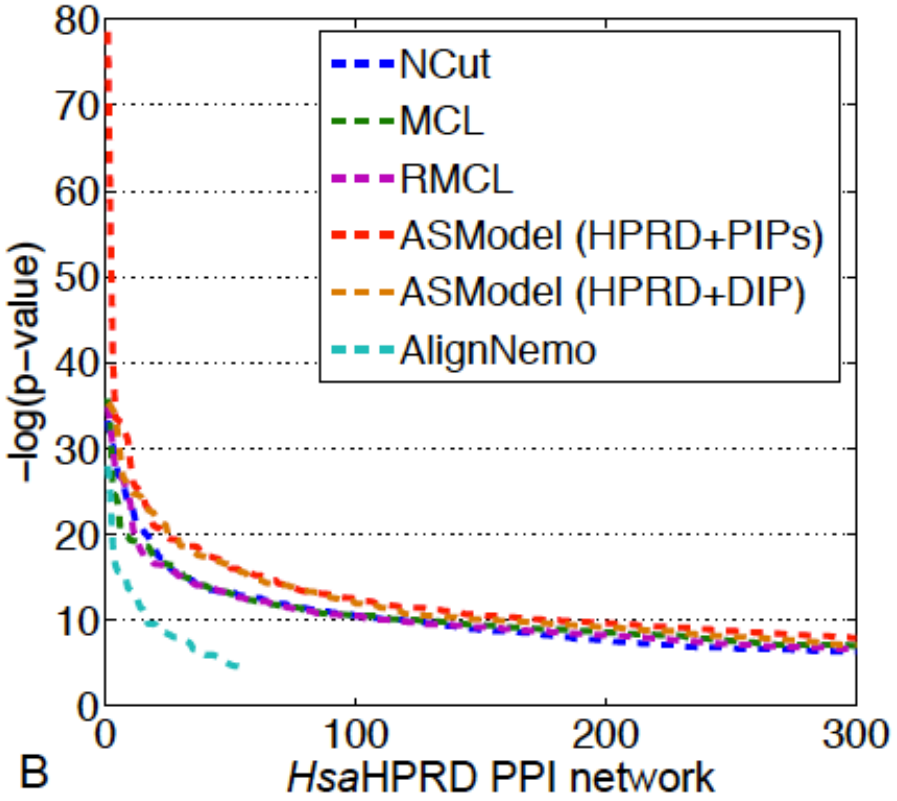
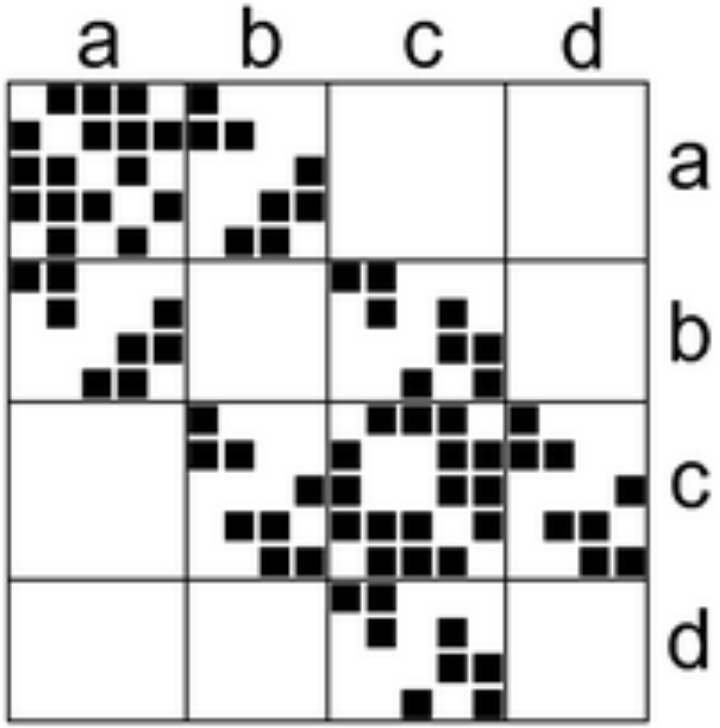


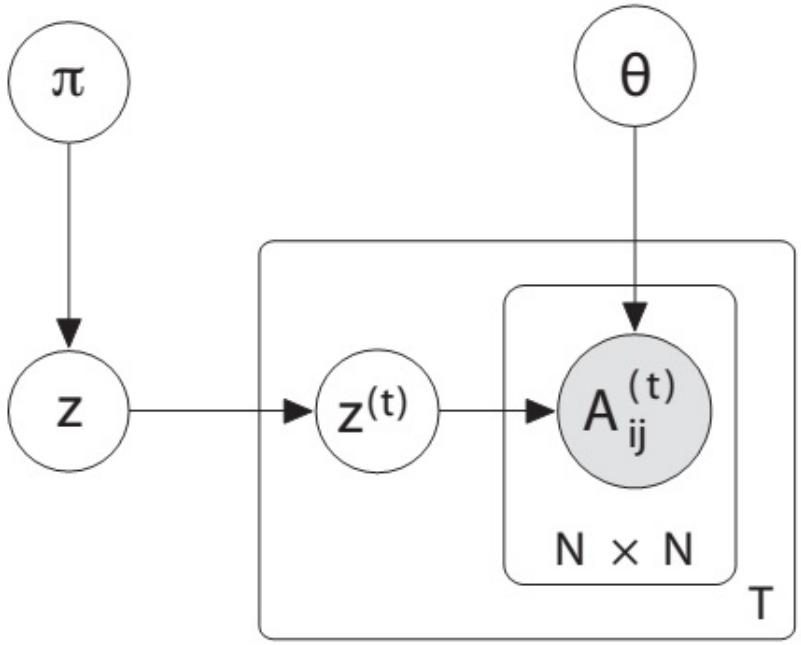
Image Co-Segmentation

- Joint network clustering helps identify better modules.
- Random walk for blockmodel module identification, which can be extended to other applications.

Ongoing Research

- Bayesian multiple network clustering
  - Generative models: Edge Partition Models (Stochastic Block Model) with Bayesian Computation





Zamani Dadaneh S and Qian X (2016) Bayesian module identification from multiple noisy networks. EURASIP JBSB.

Stochastic Block Model (SBM)

- A network is represented by a binary adjacency matrix $\boldsymbol{A}$

- Module membership can be captured by a vector $\boldsymbol{z}$, which has a multinomial distribution with the prior $\boldsymbol{\pi}$

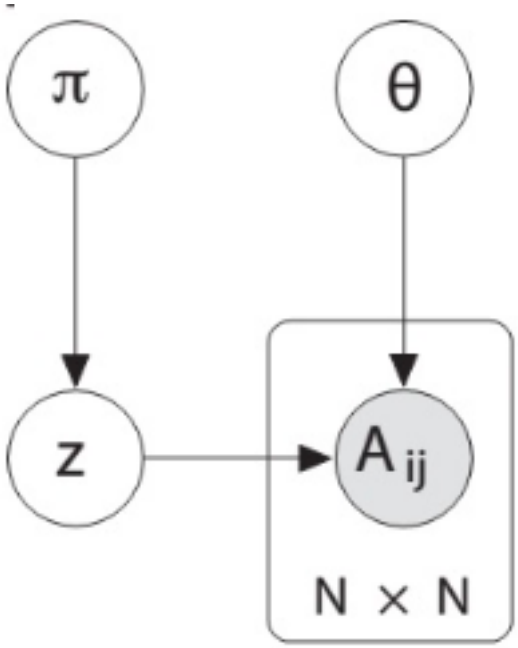
- Probability of the existence of an edge between two nodes ($A_{ij} = 0$ or 1) is governed by θ_{ij} , depending on whether $z_i = z_j$

Zamani Dadaneh S and Qian X (2016) Bayesian module identification from multiple noisy networks. EURASIP JBSB.

Stochastic Block Model (SBM)

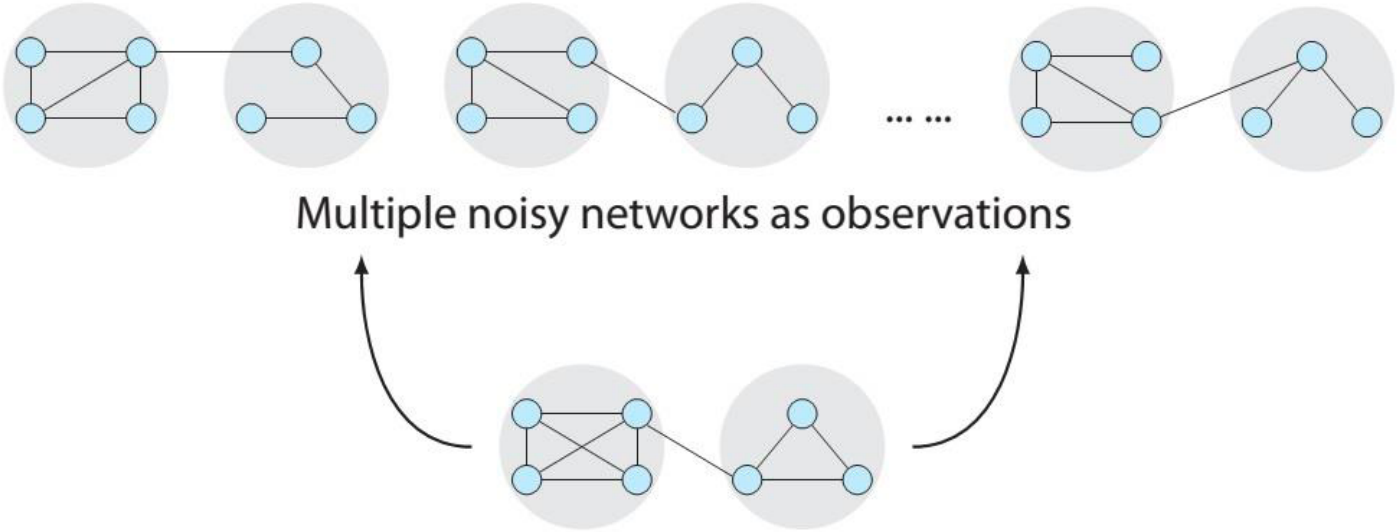
- SBM Formulation

$$
\begin{aligned}
p(\vec{z}|\vec{\pi}) &\equiv \prod_{\mu=1}^{K} \pi_{\mu}^{n_{\mu}} \\
p(\mathbf{A}|\vec{z},\vec{\pi},\vec{\theta}) &\equiv \theta_c^{c_+}(1-\theta_c)^{c_-}\theta_d^{d_+}(1-\theta_d)^{d_-} \\
p(\vec{\theta}) &\equiv \mathcal{B}(\theta_c;\tilde{c}_{+_0},\tilde{c}_{-_0})\mathcal{B}(\theta_d;\tilde{d}_{+_0},\tilde{d}_{-_0}) \\
p(\vec{\pi}) &\equiv \mathcal{D}(\vec{\pi};\tilde{\vec{n}})
\end{aligned}
$$

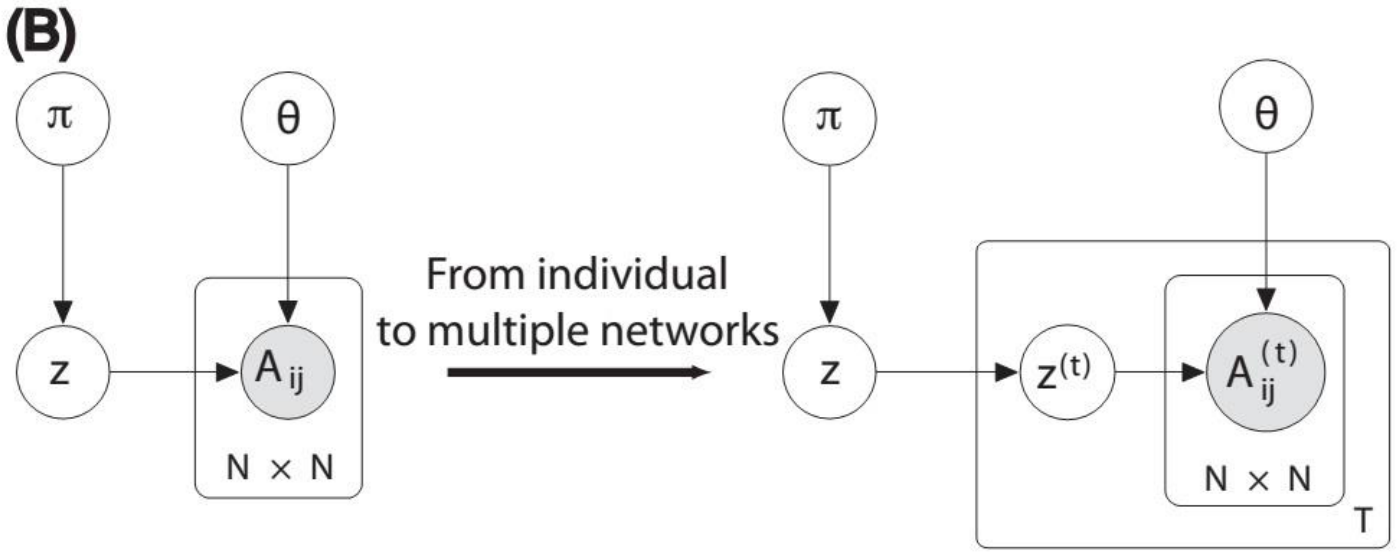


Hofman JK and Wiggins CH (2008) Bayesian approach to network modularity. Phys. Rev. Lett. 100, 258701 .

Hierarchical SBM for Multiple Nets



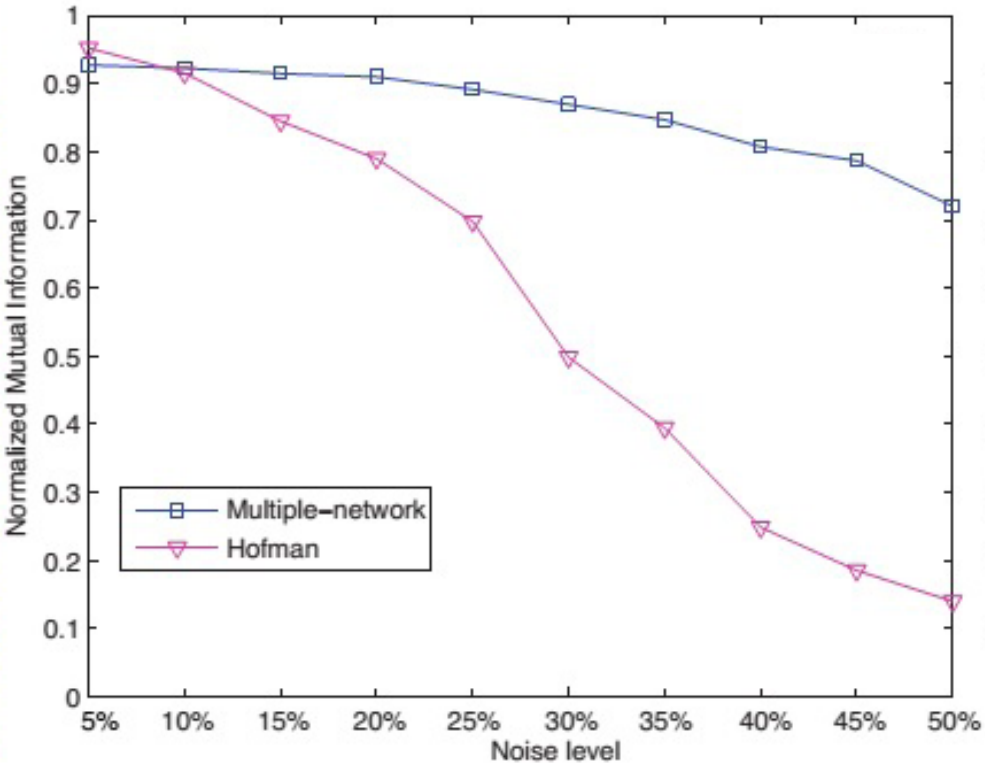
(A) Latent root graph captures the underlying modular structure.

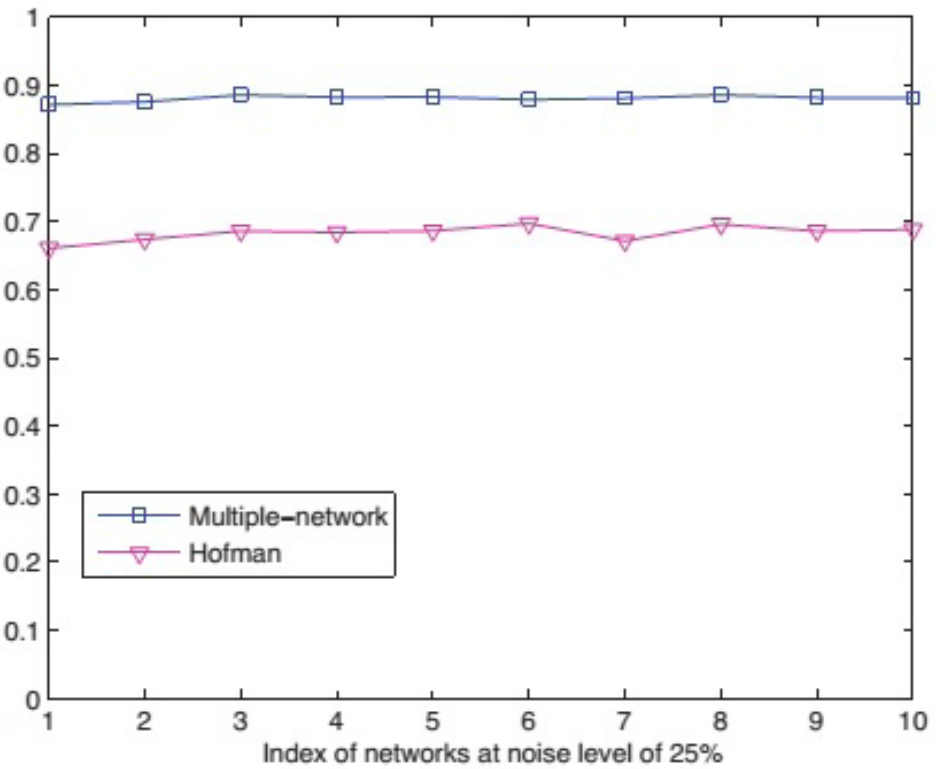


Zamani Dadaneh S and Qian X (2016) Bayesian module identification from multiple noisy networks. EURASIP JBSB.

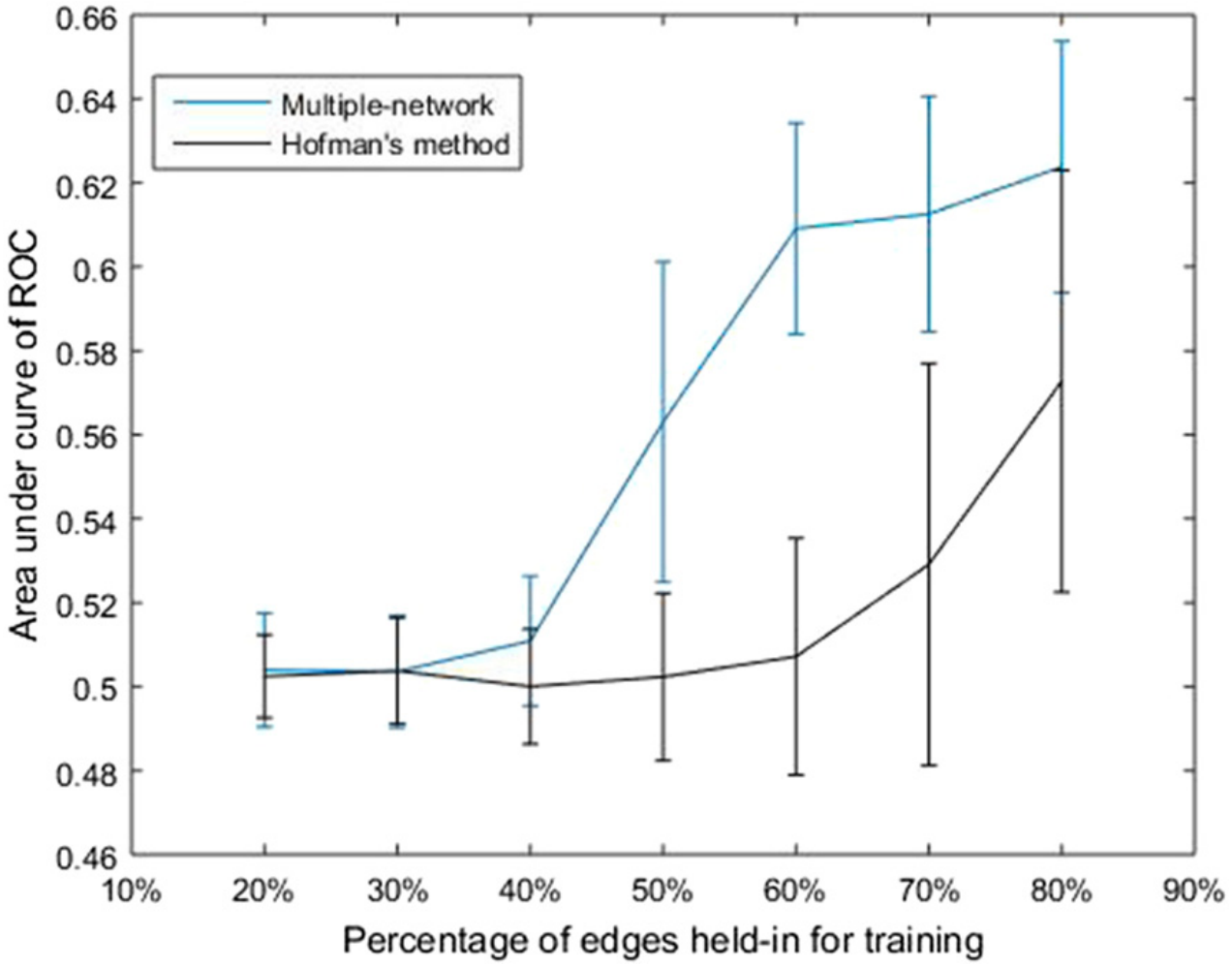
Experimental Results

Integrating 10 noisy networks



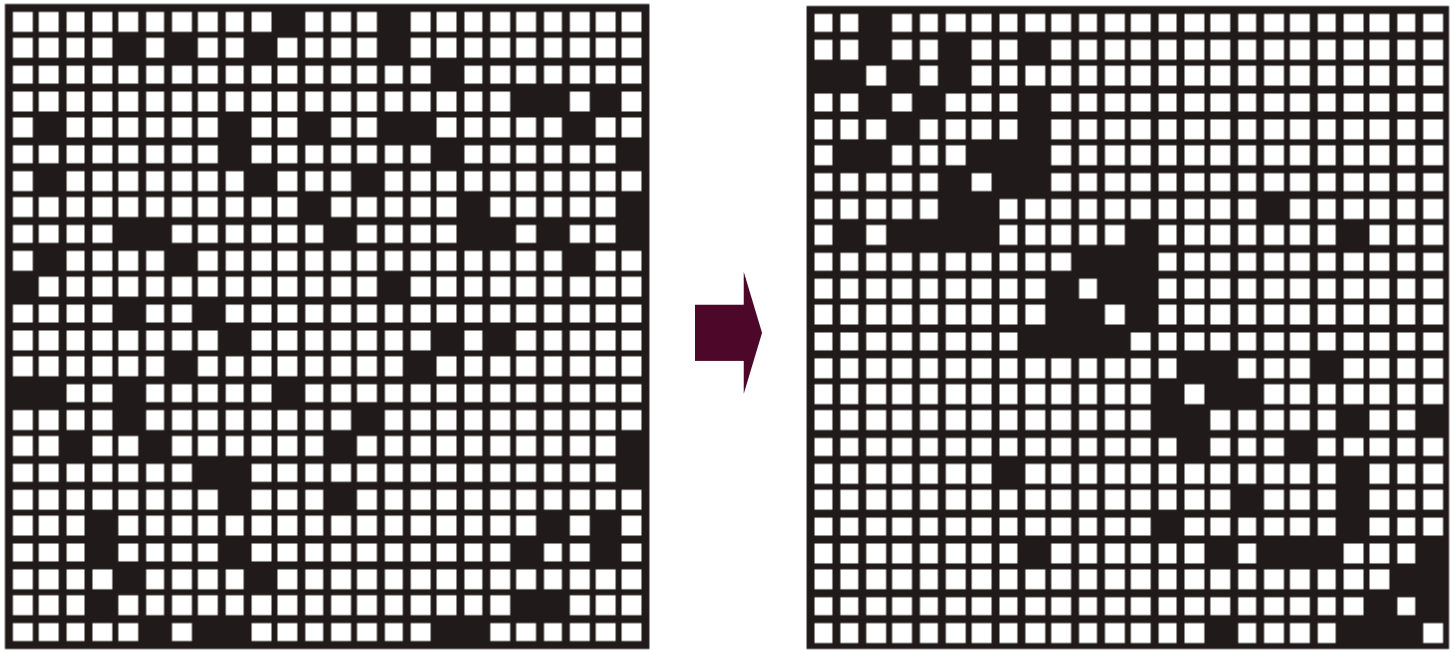


Edge Prediction



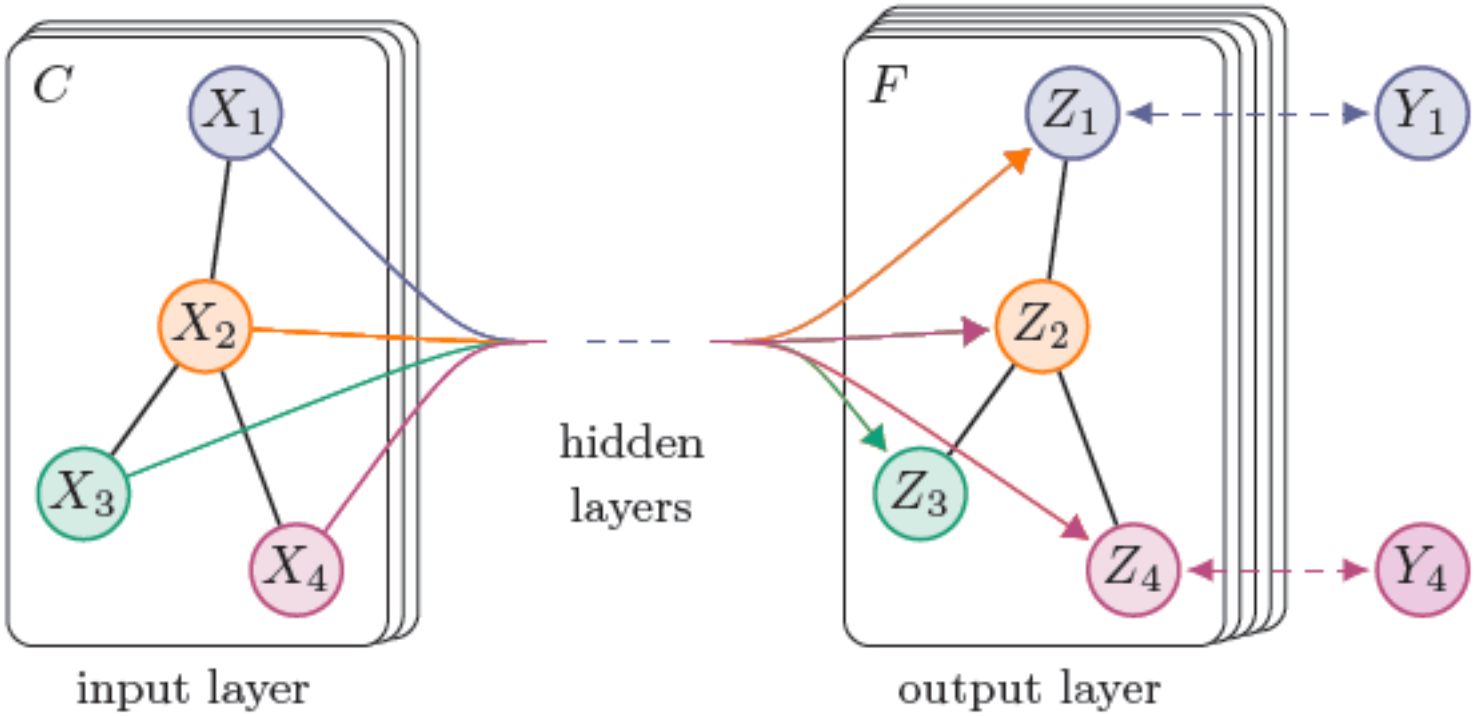
Ongoing Research

- What about vertex properties?
  1. Deep models: Graph Convolutional Networks



Graph Convolution Networks (GCN)

- What is GCN?

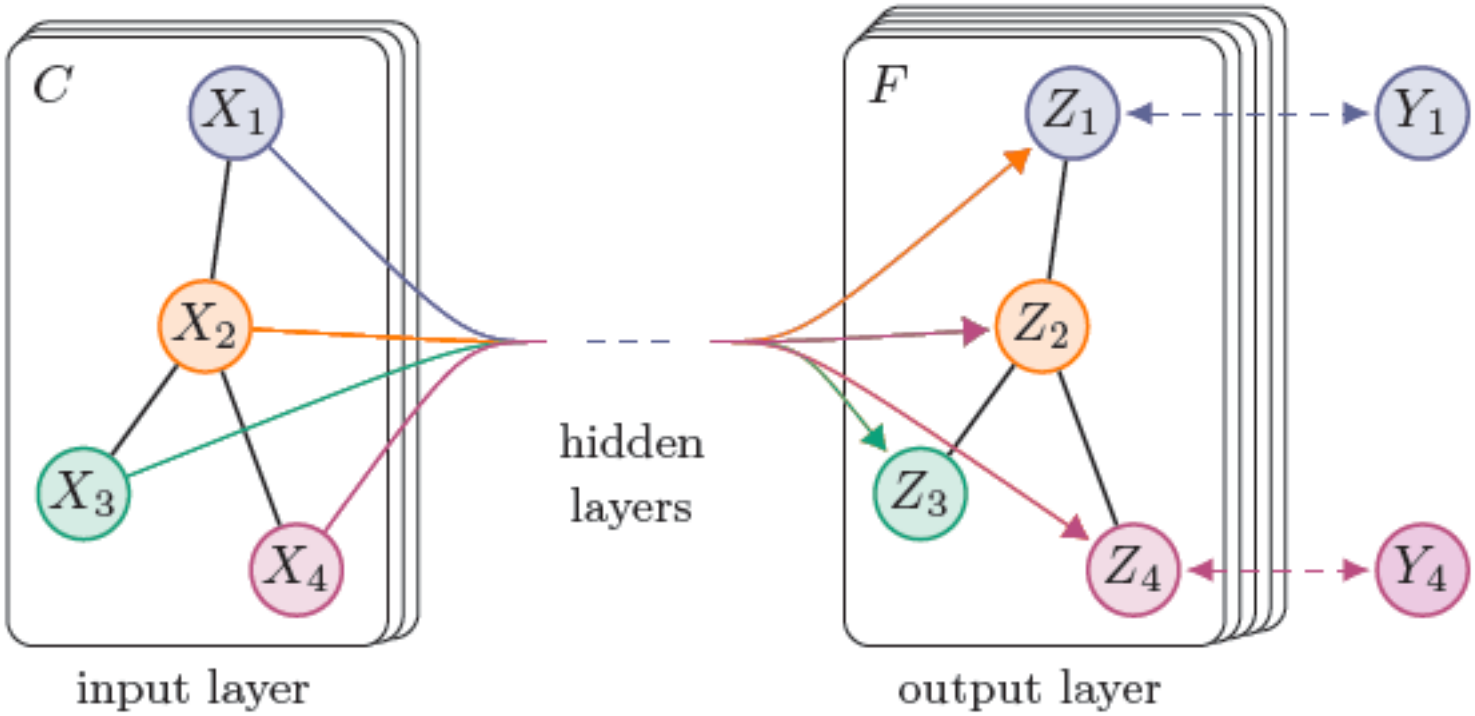


$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

Wipf T and Wellling M (2017) Semi-Supervised Classification with Graph Convolutional Networks. ICLR.

Graph Convolution Networks (GCN)

- What is GCN?



$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

Wipf T and Wellling M (2017) Semi-Supervised Classification with Graph Convolutional Networks. ICLR.

Conclusions

- It is more appropriate to investigate interaction patterns for biological modules.
- Markov models is one class of appropriate models that enables effective clustering of biological networks.
- There are more challenges in module identification: definitions, optimization, and evaluation due to empirical properties of "measured" biological networks.

Acknowledgements

- Dr. Yijie Wang, Siamak Zamani Dadaneh, and other students who have been working on the problem.
- Drs. Byung-Jun Yoon, Mingyuan Zhou, and other collaborators for insightful discussion.
- NSF Awards #1244068, #1447235, #1547557, #1553281, and NIH R21DK092845 for their kind funding support.

Thank you!

Any Questions?

Q&A