Advanced Algorithms and Models for Computational Biology
-- a machine learning approach

Network Algorithms

Eric Xing

Lecture 23, April 12 & 17, 2006

Reading

Mining and analyzing networks

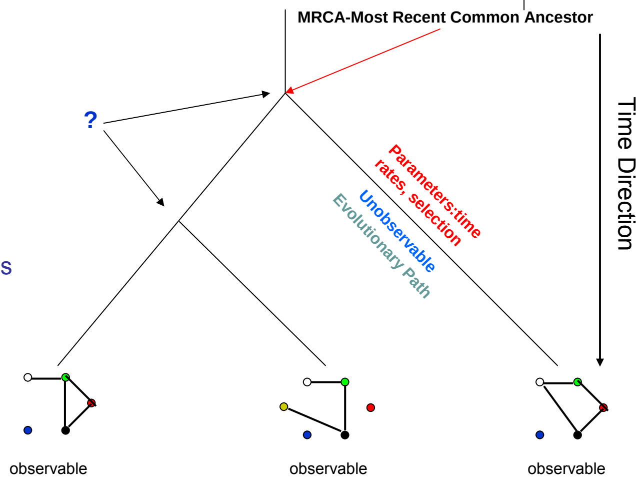
- Identifying Signaling Pathways
  - color-coding technique (Alon, Yuster and Zwick. 1995) and generalizations (Scott et al. RECOMB 2005)
- Identifying Interaction Complexes (clique-like structures)
  - Statistical subgraph scoring (Sharan et al. RECOMB 2004)
- Network alignment
  - PathBLAST: identify conserved *pathways* (Kelley et al 2003)
  - MaWISh: identify conserved *multi-protein complexes* (Koyuturk et al 2004)
  - Nuke: Scalable and General Pairwise and Multiple Network Alignment (Flannick, Novak, Srinivasan, McAdams, Batzoglou 2005)
- Network Dynamics
  - Sandy: backtracking to find active sub-network (Luscombe et al, Nature 2005)
- Node function inference
  - Stochastic block models (Aroldi et al, 2006)
  - Latent space models (Hoff, 2004)
- Link prediction
  - Naïve Bayes classifier, Bayesian network
  - MRF

Network evolution

3 Problems:

1. Test all possible relationships.
2. Examine unknown internal states.
3. Explore unknown paths between states at nodes.

→ Network alignment



MRCA-Most Recent Common Ancestor

?

Time Direction

Parameters:time rates, selection

Unobservable

Evolutionary Path

observable

observable

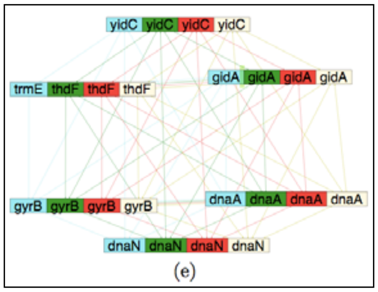
observable

Motivation

- Sequence alignment seeks to identify conserved DNA or protein sequence
  - Intuition: conservation implies functionality

```
EFTPPVQAAYQKVVAGV   (human)
DFNPNVQAAFQKVVAGV   (pig)
EFTPPVQAAYQKVVAGV   (rabbit)
```

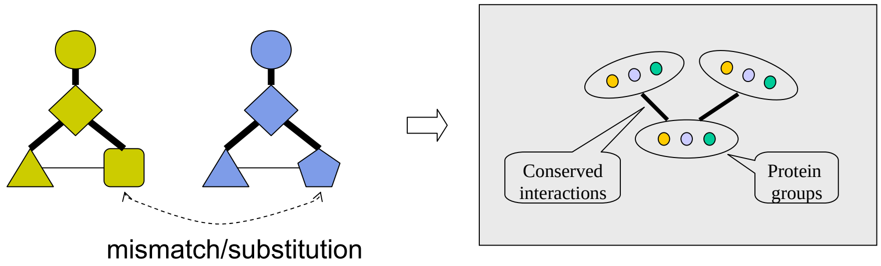
- By similar intuition, subnetworks conserved across species are likely functional modules



(e)

Network Alignment

- “Conserved” means two subgraphs contain proteins serving **similar** functions, having **similar** interaction profiles
  - Key word is similar, not identical



- Product graph:
  - Nodes: groups of sequence-similar proteins, one per species.
  - Edges: conserved interactions.

Scoring Scheme

- Given two protein subsets, one in each species, with a many-to-many correspondence between them, we wish:
  - Each subset induces a dense subgraph.
  - Matched protein pairs are sequence-similar.
- Two hypothesis:
  - **Conserved complex model**: matched pairs are similar.
  - **Random model**: matched pairs are randomly chosen.

$$L(C,C') = L(C)\cdot L(C')\cdot \prod_{u,v\ \text{matched}} \frac{\Pr(S_{u,v}\mid \text{similar})}{\Pr(S_{u,v}\mid \text{random})}$$

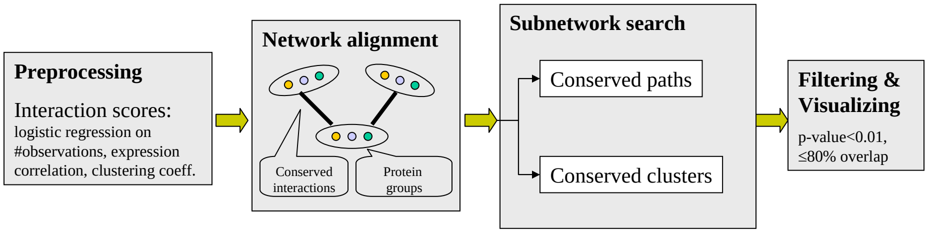
Similarity (BLAST E-value)

Scoring Scheme cont.

- For multiple networks: run into problem of scoring a multiple sequence alignment.
- Need to balance edge and vertex terms.

- Practical solution:
  - Sensible threshold for sequence similarity.
  - Nodes in alignment graph are filtered accordingly.
  - Node terms are removed from score.

Multiple Network Alignment



Preprocessing

Interaction scores: logistic regression on #observations, expression correlation, clustering coeff.

Network alignment

Subnetwork search

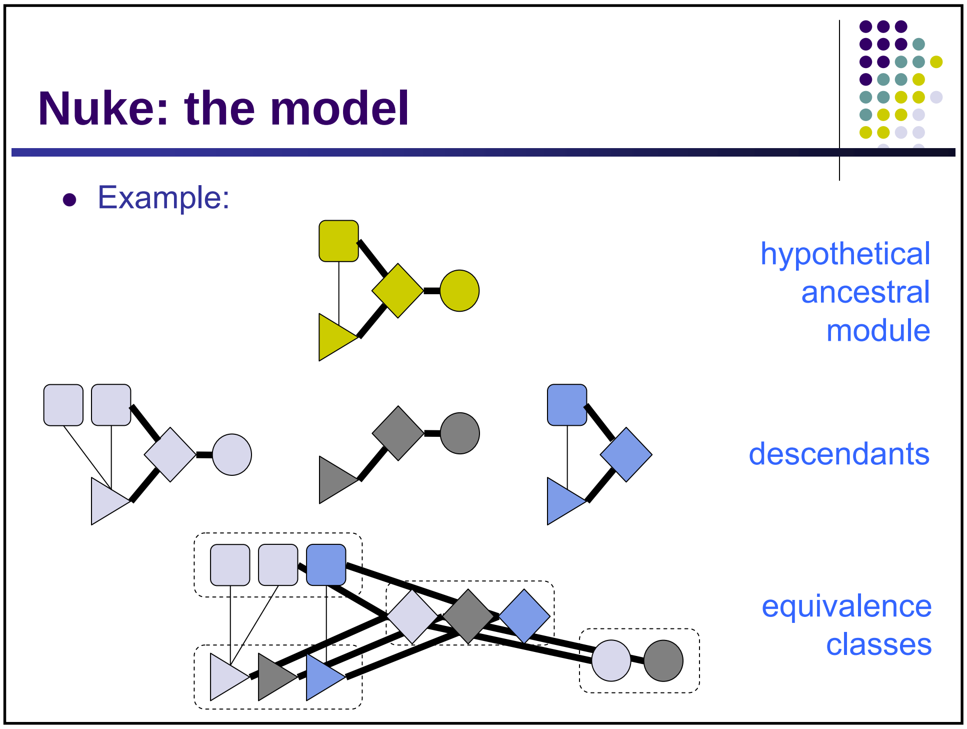
Conserved paths

Conserved clusters

Filtering & Visualizing

p-value<0.01, ≤80% overlap

- Two recent algorithms:
  - ???, Sharan et al. PNAS 2005
  - Nuke: Flannick, Novak, Srinivasan, McAdams, Batzoglou 2005



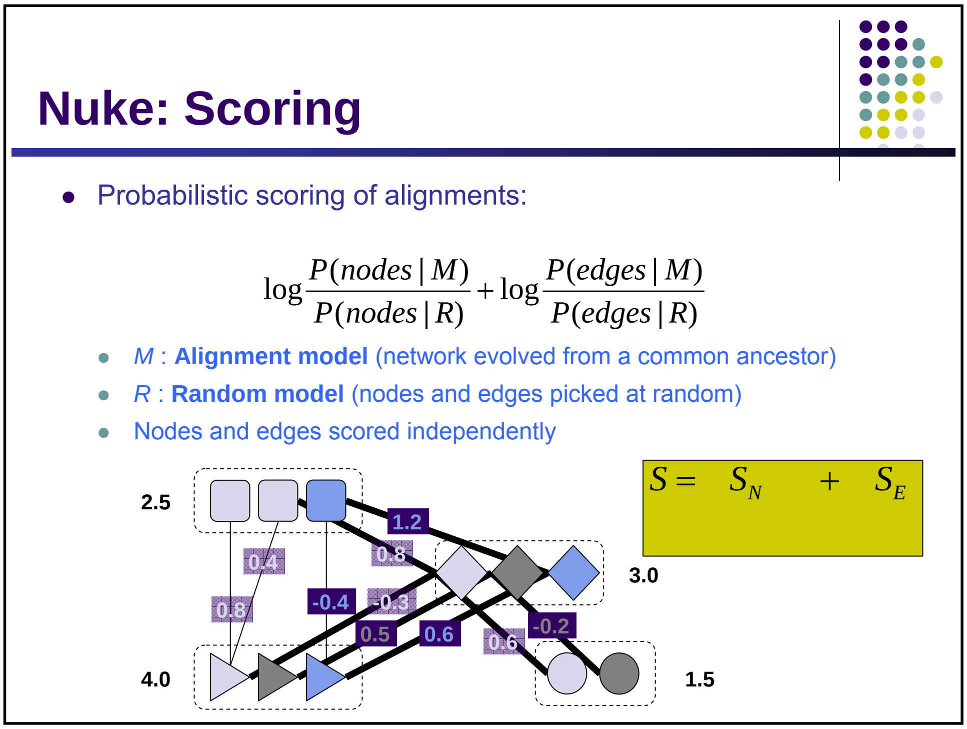
Nuke: the model

- Example:

hypothetical ancestral module

descendants

equivalence classes



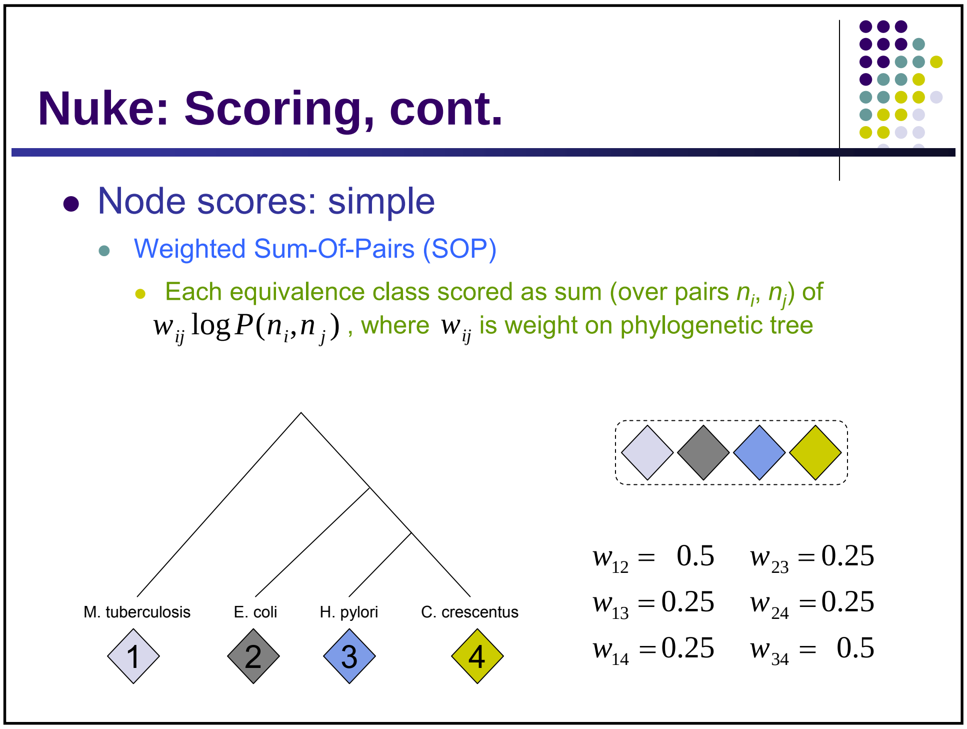
Nuke: Scoring

- Probabilistic scoring of alignments:

$$\log\frac{P(nodes \mid M)}{P(nodes \mid R)} + \log\frac{P(edges \mid M)}{P(edges \mid R)}$$

  - *M* : **Alignment model** (network evolved from a common ancestor)
  - *R* : **Random model** (nodes and edges picked at random)
  - Nodes and edges scored independently

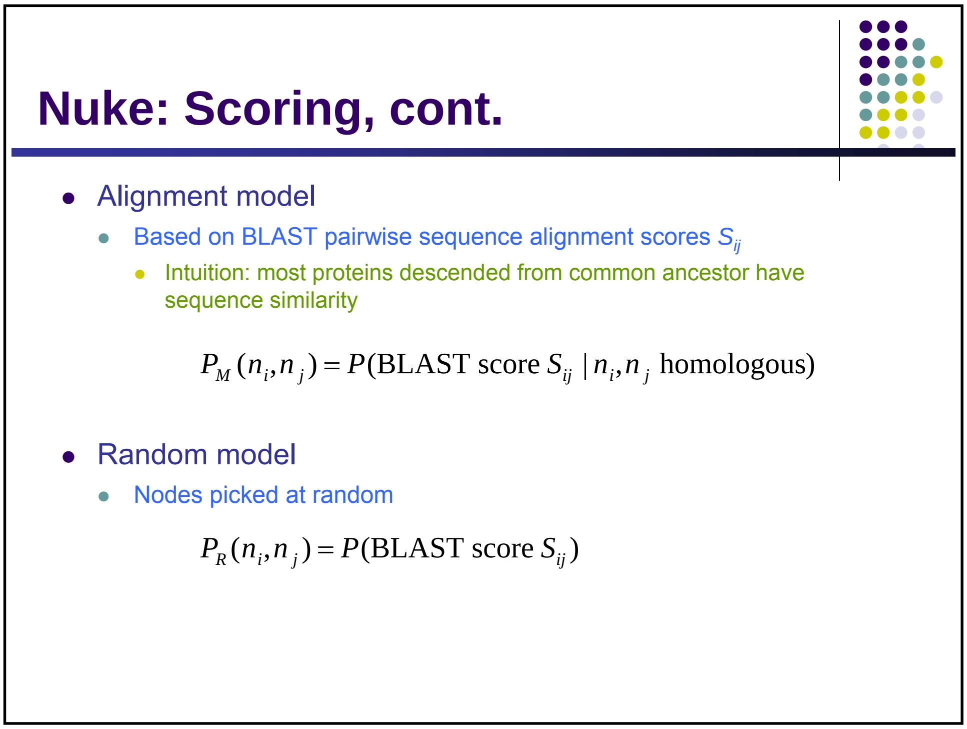
$$S = S_N + S_E$$



Nuke: Scoring, cont.

- Node scores: simple
  - Weighted Sum-Of-Pairs (SOP)
    - Each equivalence class scored as sum (over pairs n_i, n_j) of $w_{ij} \log P(n_i, n_j)$, where w_{ij} is weight on phylogenetic tree

$$w_{12} = 0.5 \quad w_{23} = 0.25$$
$$w_{13} = 0.25 \quad w_{24} = 0.25$$
$$w_{14} = 0.25 \quad w_{34} = 0.5$$



Nuke: Scoring, cont.

- Alignment model
  - Based on BLAST pairwise sequence alignment scores S_{ij}
    - Intuition: most proteins descended from common ancestor have sequence similarity

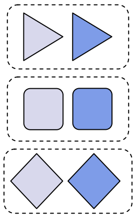
$$P_M(n_i, n_j) = P(\text{BLAST score } S_{ij} \mid n_i, n_j \text{ homologous})$$

- Random model
  - Nodes picked at random

$$P_R(n_i, n_j) = P(\text{BLAST score } S_{ij})$$

Nuke: Scoring, cont.

- Edge scores: more complicated
  - Edge scores in earlier aligners rewarded high edge weights
    - But this biases towards clique-like topology!
  - Don’t want solely conservation either
    - This alignment has highly conserved (zero-weight) edges:



Non-trivial tradeoff in pairwise alignment of full networks

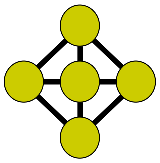
ESMs: A New Edge-Scoring Paradigm

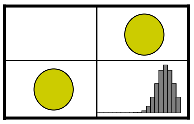
- Idea: assign each node a *label* from a finite alphabet ∑, and define edge likelihood in terms of labels it connects
  - During alignment, assign labels which maximize score

- *E*: Symmetric matrix of probability distributions, *E*(*x*, *y*) is distribution of edge weights between nodes labeled *x* and *y*

ESMs: A New Edge-Scoring Paradigm

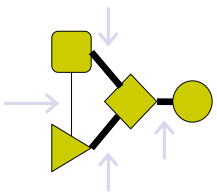
- Idea: assign each node a *label* from a finite alphabet ∑, and define edge likelihood in terms of labels it connects
  - During alignment, assign labels which maximize score
- *E*: Symmetric matrix of probability distributions, *E*(*x*, *y*) is distribution of edge weights between nodes labeled *x* and *y*
- Simplest case is *clique ESM*
  - 1x1 matrix: ∑ contains a single label
  - Duplicates edge-scoring of aligners which search for cliques

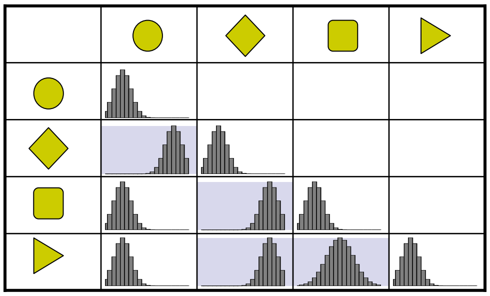




ESMs: A New Edge-Scoring Paradigm

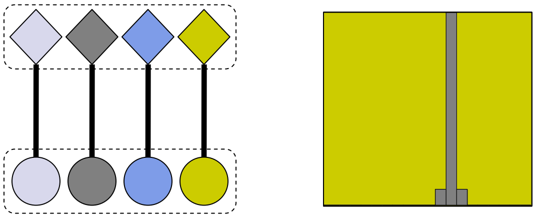
- For query-to-database alignment, use a *module ESM*
  - One label for each node in query module
    - Tractable because queries are usually small (~10-40 nodes)
  - For each pair of nodes (n_i, n_j) in query, let E(i, j) be a Gaussian centered at c_{ij} = weight of (n_i, n_j) edge





ESMs: A New Edge-Scoring Paradigm

- Multiple alignment gives us more information about conservation
  - Can iteratively improve ESM to adjust mean and deviation based on weights of edges between aligned pairs of query nodes
    - Easily implemented using kernel density estimation (KDE)

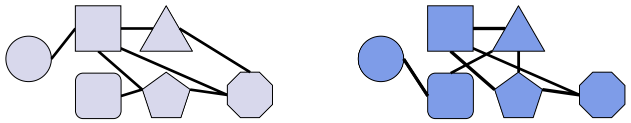


A General Network Aligner: Algorithm

- Given this model of network alignment and scoring framework, how to efficiently find alignments between a pair of networks (N_1, N_2)?
- Constructing every possible set of equivalence classes clearly prohibitive

A General Network Aligner: Algorithm

- Idea: seeded alignment
  - Inspired by seeded sequence alignment (BLAST)
  - Identify regions of network in which “good” alignments likely to be found
    - MaWISh does this, using high-degree nodes for seeds
    - Can we avoid such strong topological constraints?



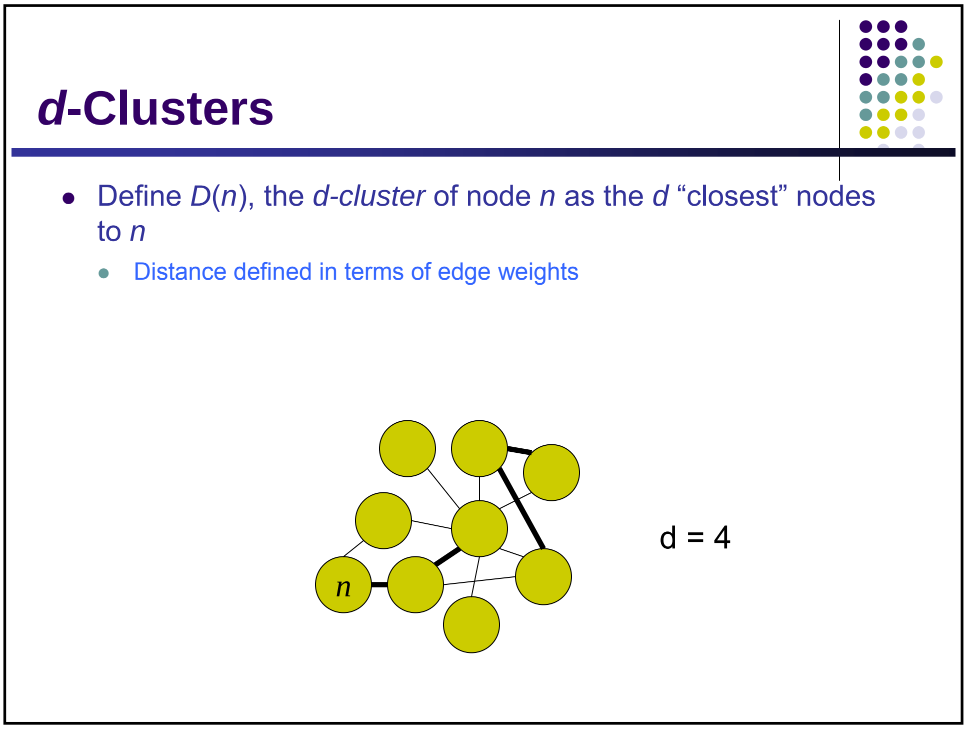
Seed

↓

Extend

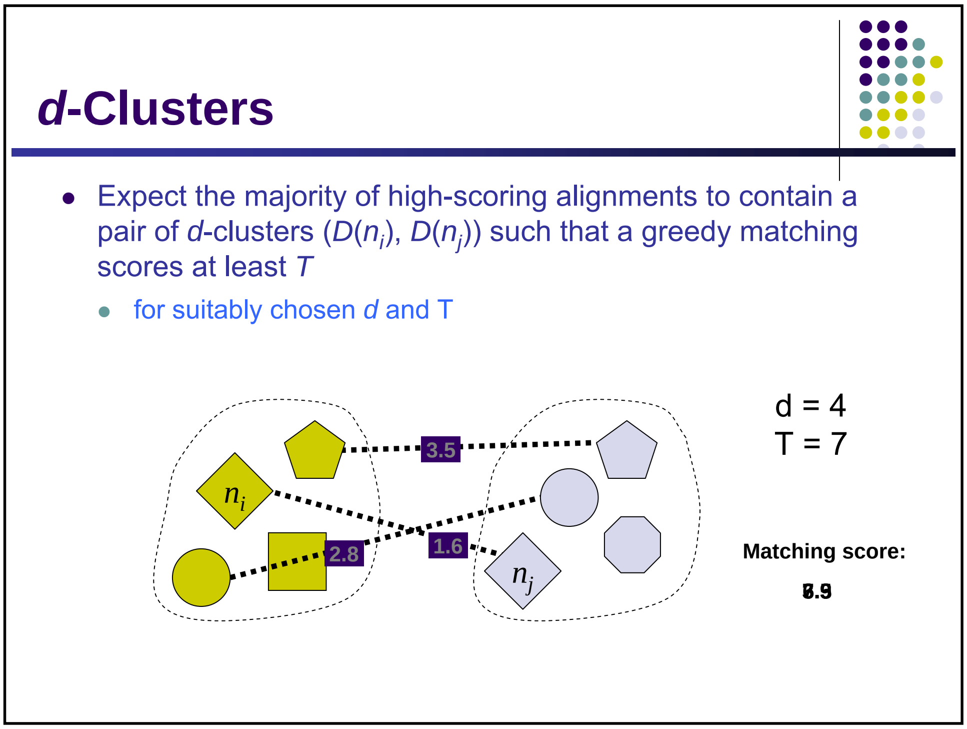
d-Clusters: Intuition

- “Good” alignments typically have:
  - a significant number of nodes with high sequence similarity
    - Implied by the node scoring function, which prefers aligning nodes with high BLAST scores
  - with mostly conserved connected components
    - Implied by the edge scoring function which prefers conserved edge weights



d-Clusters

- Define *D*(*n*), the *d-cluster* of node *n* as the *d* “closest” nodes to *n*
  - Distance defined in terms of edge weights

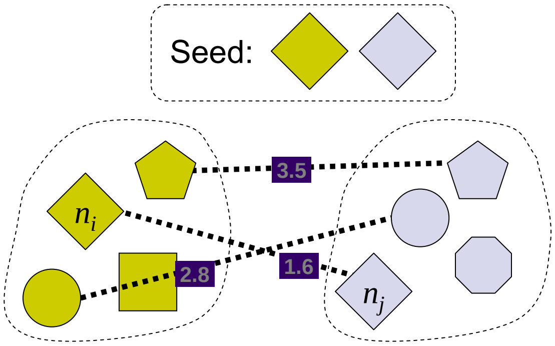


d-Clusters

- Expect the majority of high-scoring alignments to contain a pair of *d*-clusters ($D(n_i)$, $D(n_j)$) such that a greedy matching scores at least *T*
  - for suitably chosen *d* and T

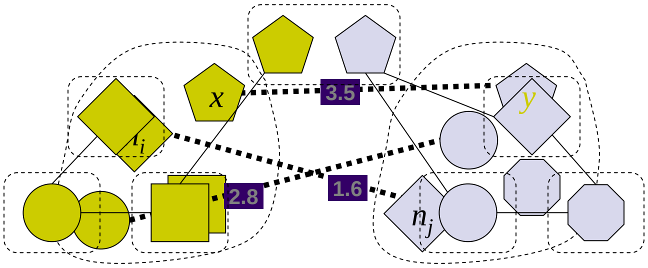
d-Clusters

- Seeding algorithm: for each $n_i \in N_1$ and $n_j \in N_2$, emit (n_i, n_j) as a seed if matching score exceeds T



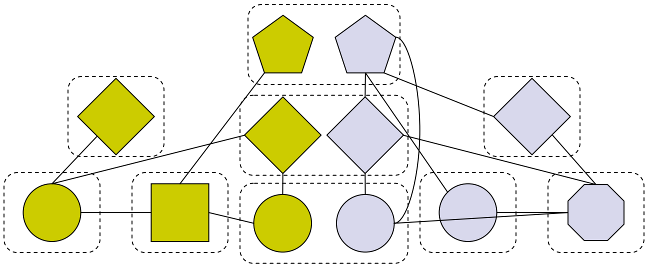
Extending seeds

- Given a pair of *d*-cluster seeds $(D(n_i), D(n_j))$, want to find highest-scoring alignment containing this seed
- Start by forming an equivalence class consisting of $x \in D(n_i)$ and $y \in D(n_j)$ maximizing $S_N(x, y)$
  - All other $m \in N_1 \cup N_2$ are singleton equivalence classes



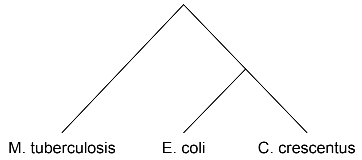
Extending seeds

- Extend greedily:
  - Define the *frontier* (*F*) as the set of all already-aligned nodes **and** their neighbors in each network
  - Picking nodes *s*, *t* ∈ *F*, ┘ and label *L* ∈ ∑, which maximally increase alignment score:
    - Merge equivalence classes [*s*] and [*t*]
    - Relabel the resulting equivalence class to *L*



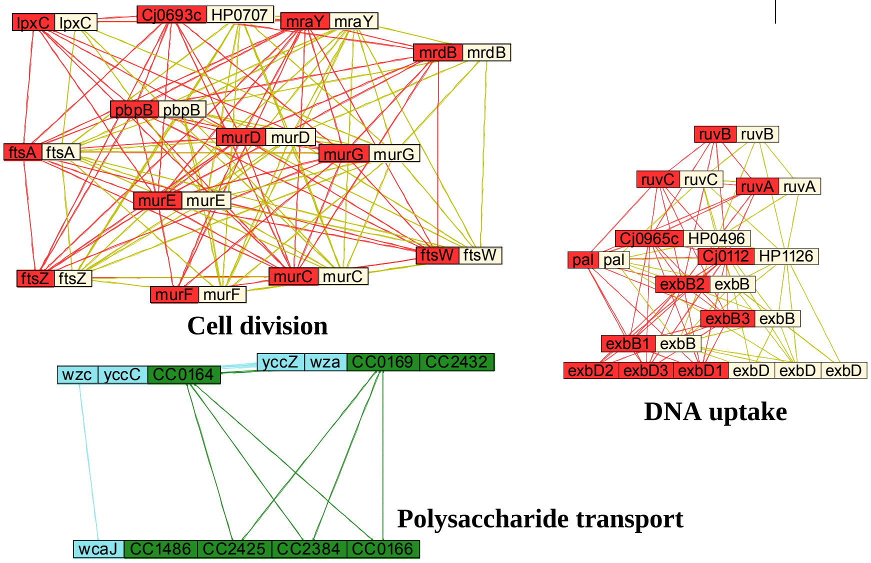
Multiple Alignment

- Progressive alignment technique
  - Used by most multiple sequence aligners



- Simple modification of implementation to align *alignments* rather than *networks*
  - Node scoring already uses weighted SOP
  - Edge scoring remains unchanged

Pairwise alignments

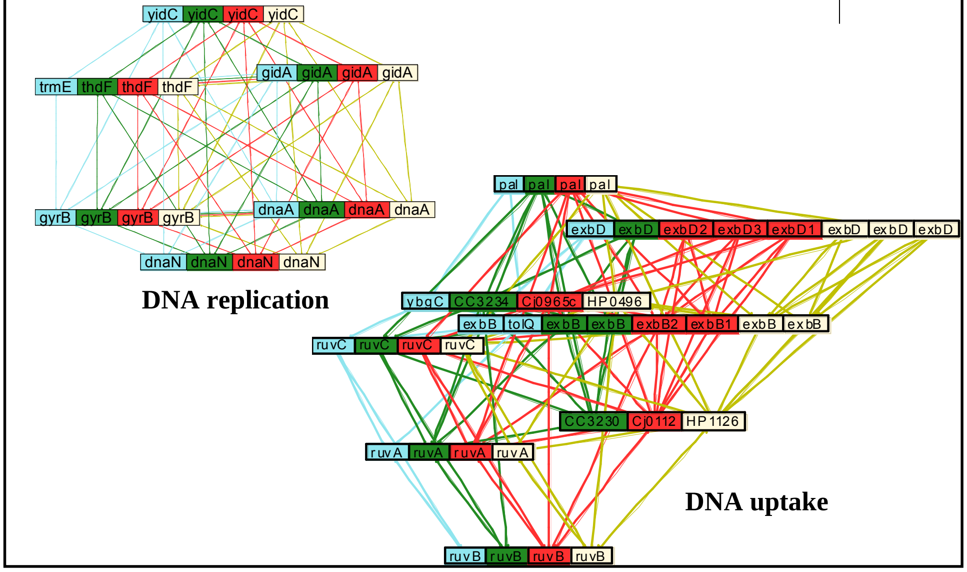


Cell division

DNA uptake

Polysaccharide transport

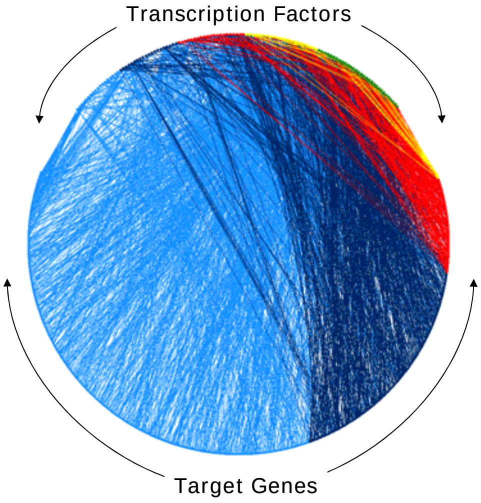
Multiple alignments



DNA replication

DNA uptake

Dynamic Yeast TF network



- Analyzed network as a static entity
- But network is *dynamic*
  - Different sections of the network are active under different cellular conditions
- Integrate gene expression data

[Luscombe et al, *Nature*]

Gene expression data

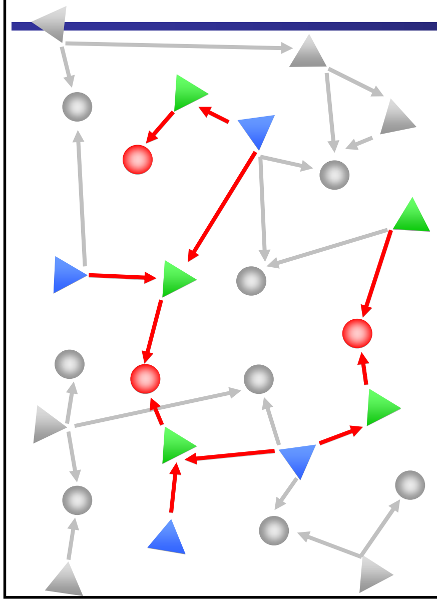
- Genes that are differentially expressed under five cellular conditions

Cellular condition	No. genes
Cell cycle	437
Sporulation	876
Diauxic shift	1,876
DNA damage	1,715
Stress response	1,385

- Assume these genes undergo transcription regulation

[Luscombe et al, *Nature*]

Backtracking to find active sub-network



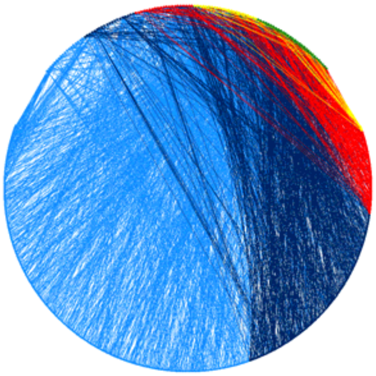
- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

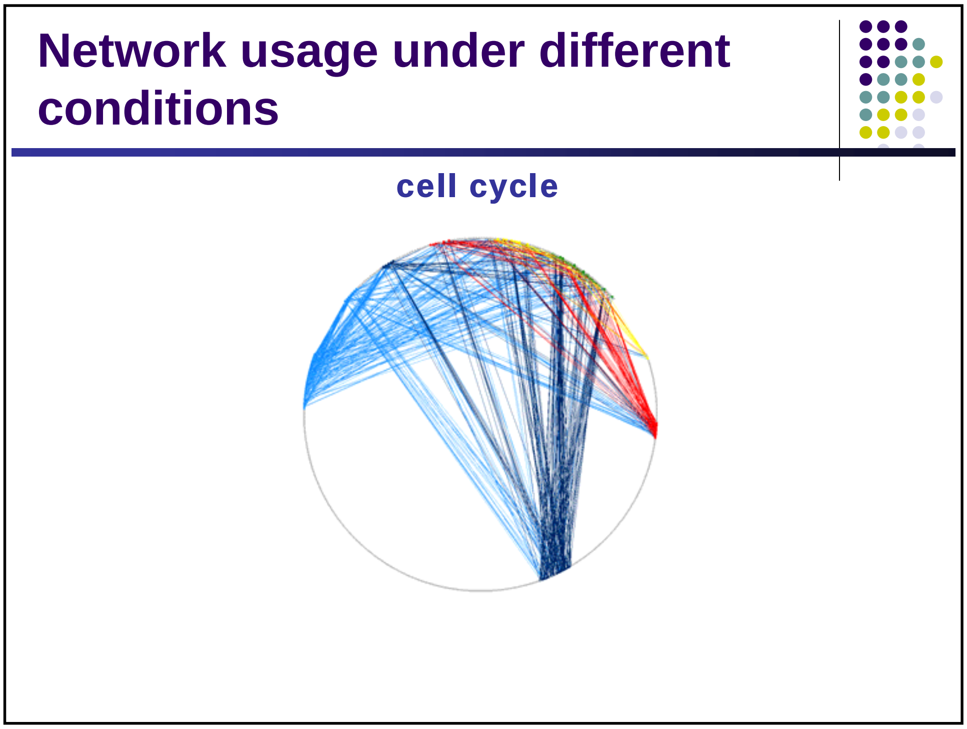
Active regulatory sub-network

[Luscombe et al, *Nature*]

Network usage under different conditions

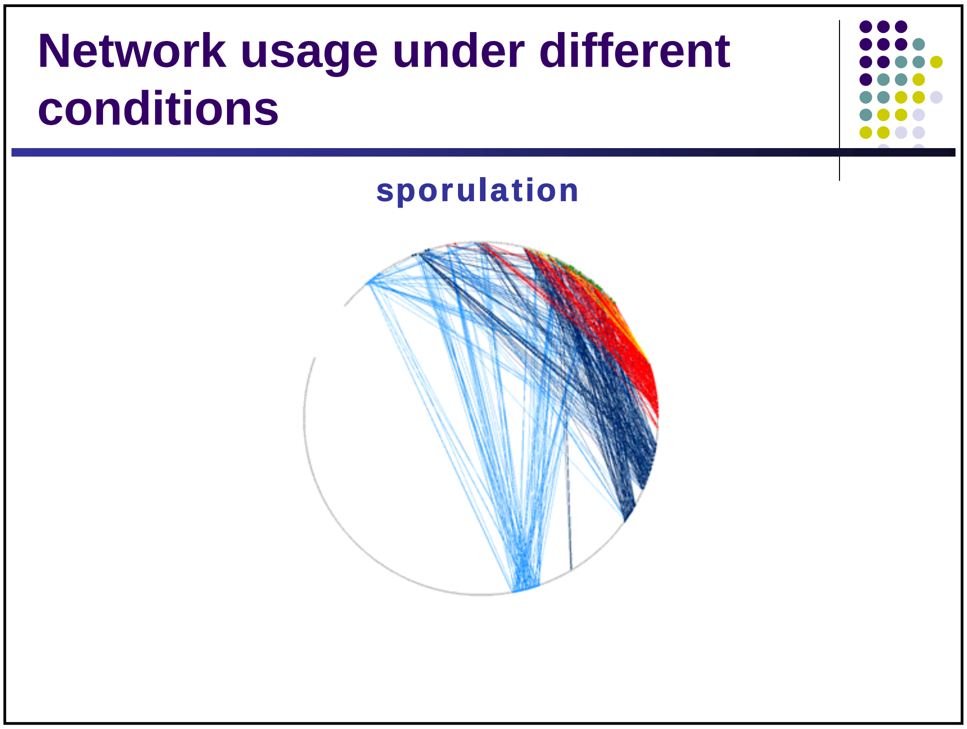
static





Network usage under different conditions

cell cycle

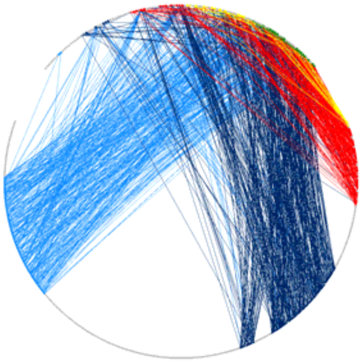


Network usage under different conditions

sporulation

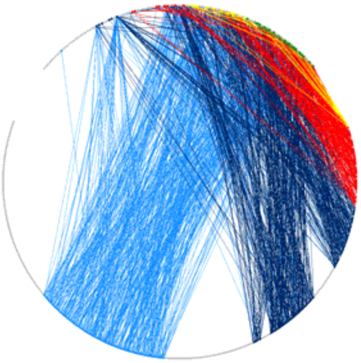
Network usage under different conditions

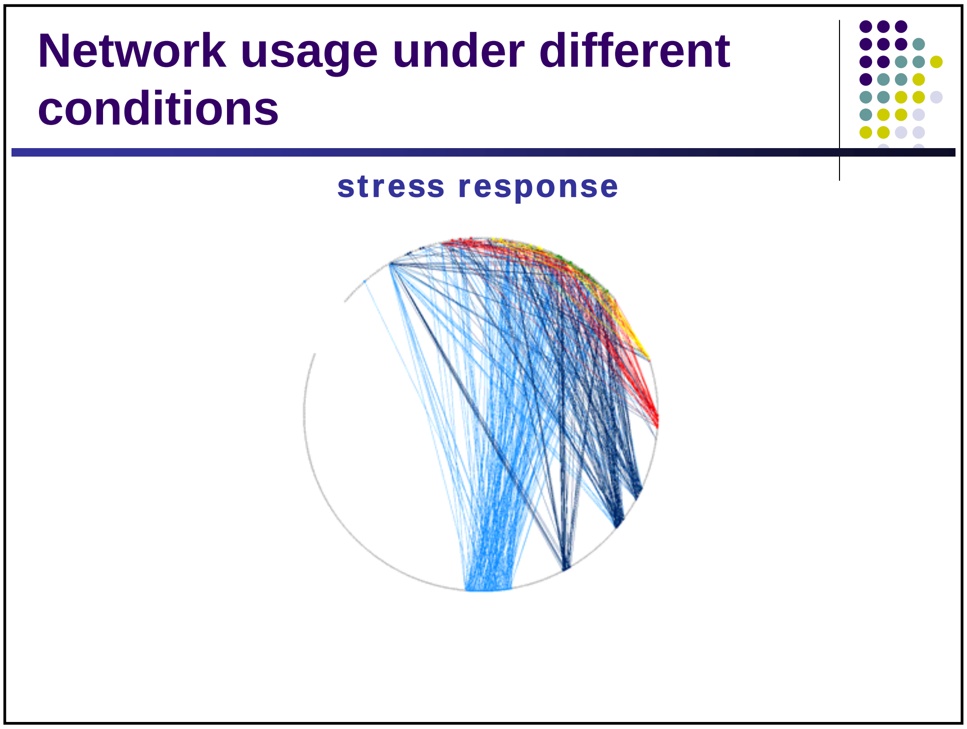
diauxic shift



Network usage under different conditions

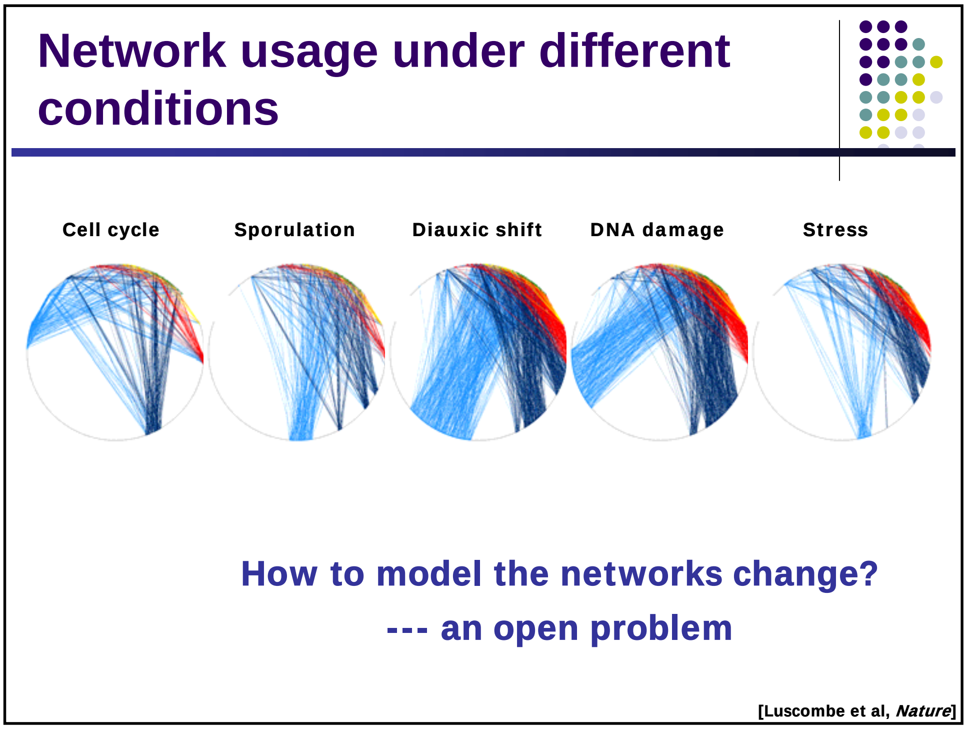
DNA damage





Network usage under different conditions

stress response



Network usage under different conditions

Cell cycle

Sporulation

Diauxic shift

DNA damage

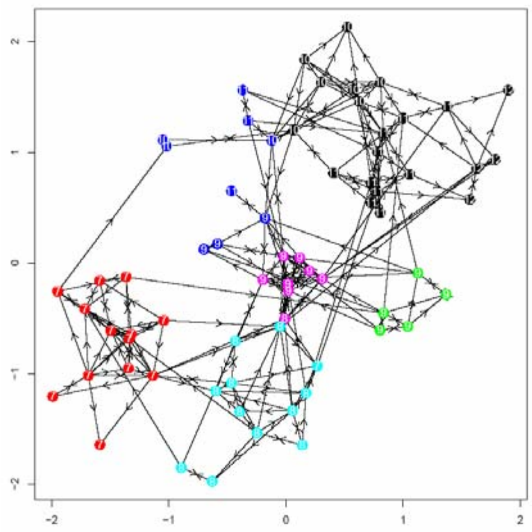
Stress

How to model the networks change?

--- an open problem

[Luscombe et al, *Nature*]

Node Clustering



Dissecting Social Networks

White et al: From logical role systems to empirical social structures

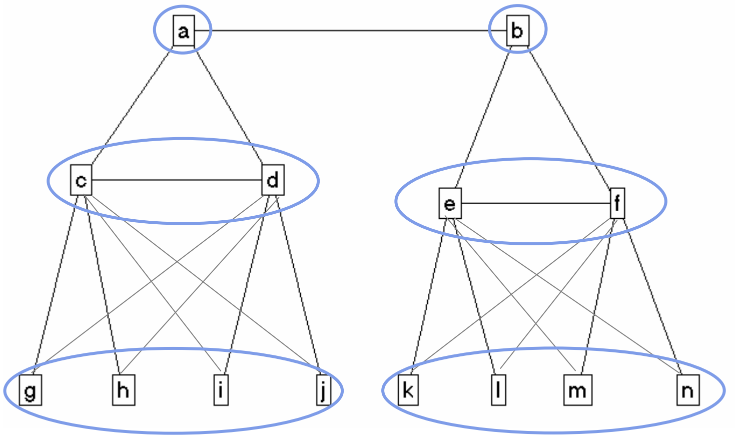
"We can express a ***role*** through a **relation** (or set of relations) and thus a social system by the inventory of roles. If roles equate to ***positions*** in an exchange system, then we need only identify particular aspects of a position. But what aspect?"

Structural Equivalence:

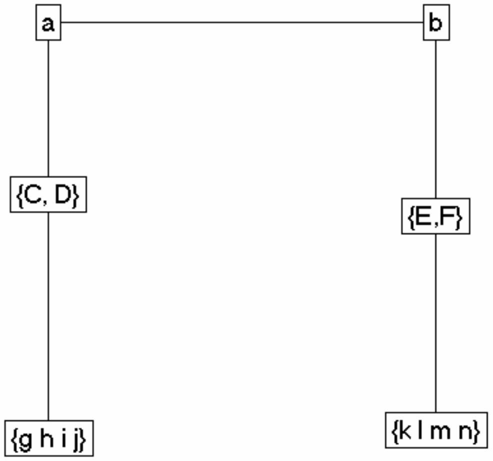
Two actors are *structurally equivalent* if they have the same types of ties to the same people.

Structural Equivalence

- Two actors are structurally equivalent if they have the same types of ties to the same people.

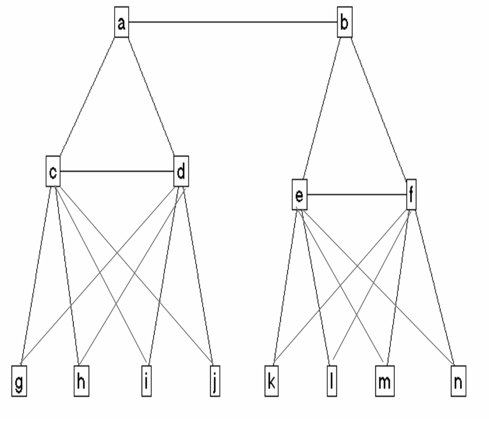


Structural Equivalence



Graph reduced to positions

Classical Blockmodeling



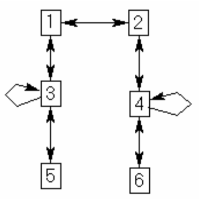
0	1	1	1	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	1	0	0	0	0	0	0	0	0
1	0	0	1	0	0	1	1	1	1	0	0	0	0
1	0	1	0	0	0	1	1	1	1	0	0	0	0
0	1	0	0	0	1	0	0	0	0	1	1	1	1
0	1	0	0	1	0	0	0	0	0	1	1	1	1
0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0

Blockmodeling is the process of identifying these types of positions. A ***block*** is a section of the adjacency matrix - a “group” of structurally equivalent people.

Cohesive Subgroups

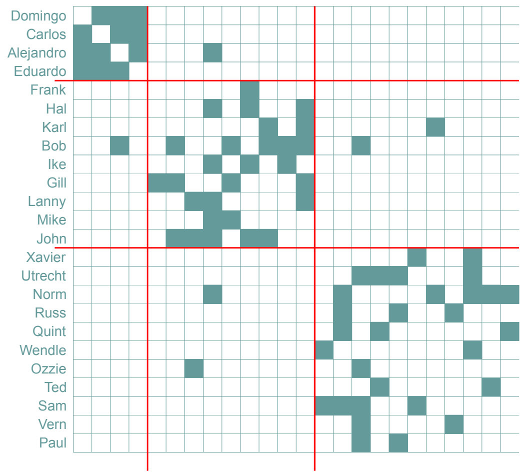
	1	2	3		4		5				6			
1	.	1	1	1	0	0	0	0	0	0	0	0	0	0
2	1	.	0	0	1	1	0	0	0	0	0	0	0	0
3	1	0	.	1	0	0	1	1	1	1	0	0	0	0
	1	0	1	.	0	0	1	1	1	1	0	0	0	0
4	0	1	0	0	.	1	0	0	0	0	1	1	1	1
	0	1	0	0	1	.	0	0	0	0	1	1	1	1
5	0	0	1	1	0	0	.	0	0	0	0	0	0	0
	0	0	1	1	0	0	0	.	0	0	0	0	0	0
	0	0	1	1	0	0	0	0	.	0	0	0	0	0
	0	0	1	1	0	0	0	0	0	.	0	0	0	0
6	0	0	0	0	1	1	0	0	0	0	.	0	0	0
	0	0	0	0	1	1	0	0	0	0	0	.	0	0
	0	0	0	0	1	1	0	0	0	0	0	0	.	0
	0	0	0	0	1	1	0	0	0	0	0	0	0	.

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	1	0	1	0
4	0	1	0	1	0	1
5	0	0	1	0	0	0
6	0	0	0	1	0	0

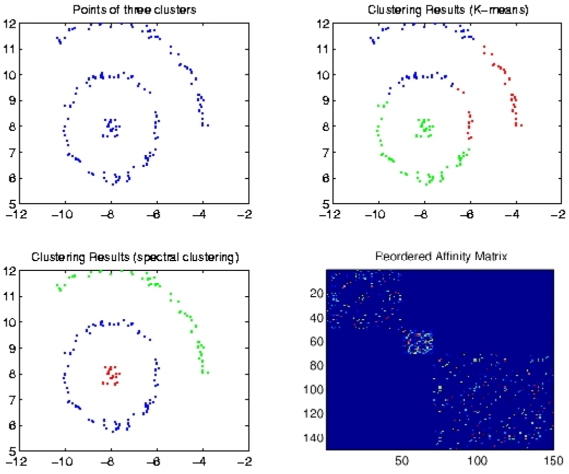


Structural equivalence thus generates 6 positions in the network

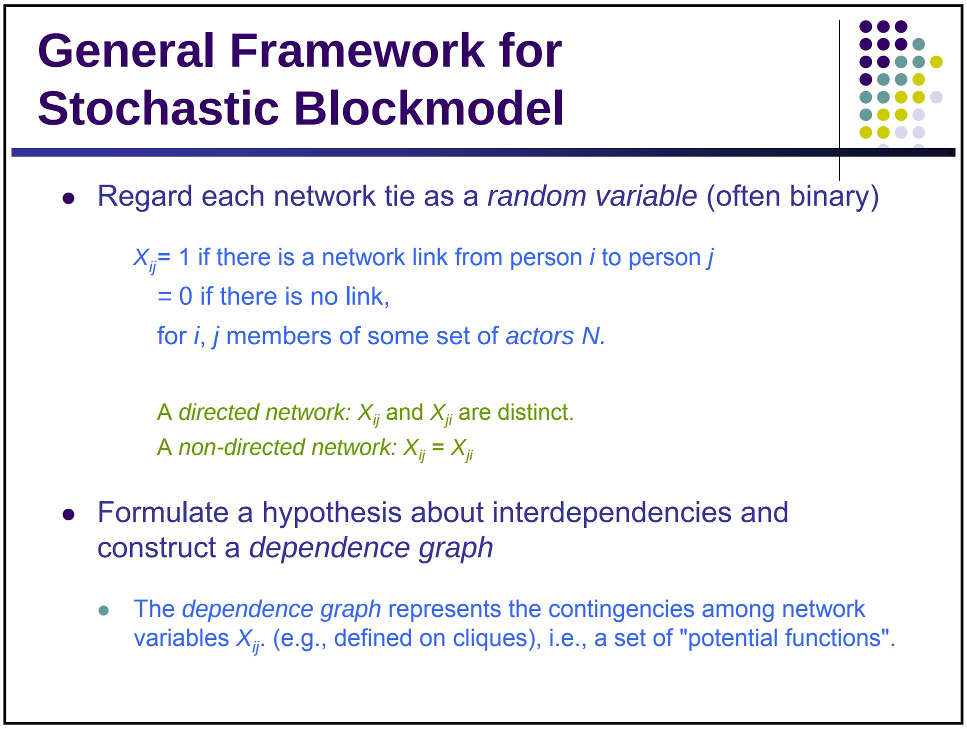
Stochastic Cohesive Subgroups



Spectral Clustering



- ❑ Minimize total transition probability of single-step between cluster random walk
- ❑ Each object has a unique cluster membership



General Framework for Stochastic Blockmodel

- Regard each network tie as a *random variable* (often binary)

  X_{ij}= 1 if there is a network link from person *i* to person *j*

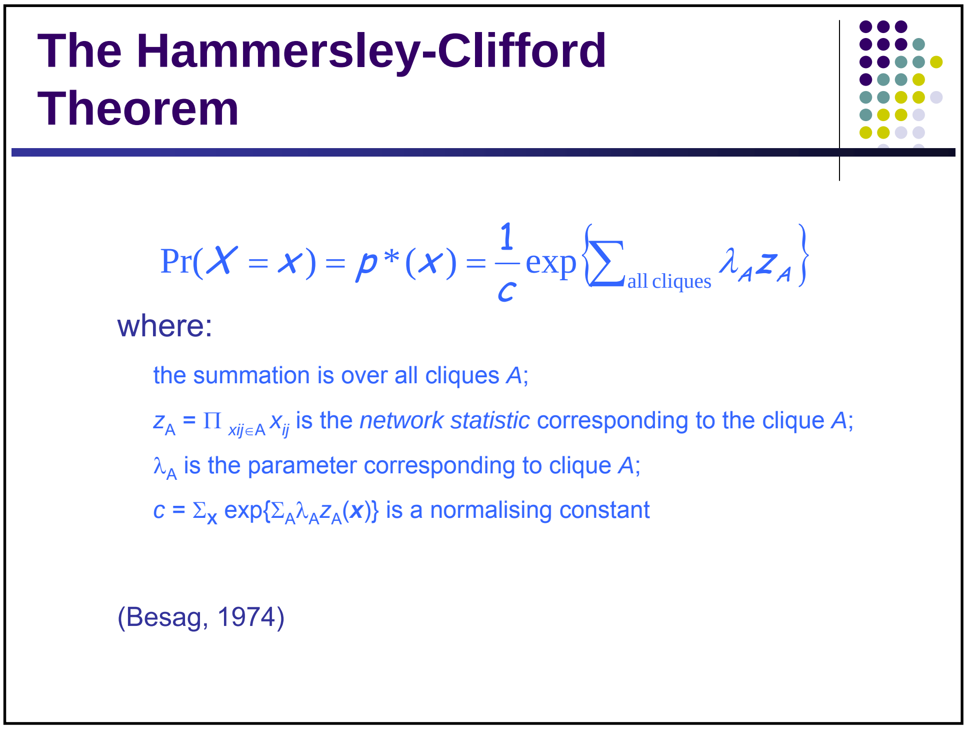
  = 0 if there is no link,

  for *i*, *j* members of some set of *actors N.*

  A *directed network:* X_{ij} and X_{ji} are distinct.

  A *non-directed network:* $X_{ij} = X_{ji}$

- Formulate a hypothesis about interdependencies and construct a *dependence graph*
  - The *dependence graph* represents the contingencies among network variables X_{ij}. (e.g., defined on cliques), i.e., a set of "potential functions".



The Hammersley-Clifford Theorem

$$\Pr(X = x) = p^*(x) = \frac{1}{c}\exp\left\{\sum_{\text{all cliques}} \lambda_A z_A\right\}$$

where:

the summation is over all cliques *A*;

$z_A = \Pi_{x_{ij} \in A} x_{ij}$ is the *network statistic* corresponding to the clique *A*;

λ_A is the parameter corresponding to clique *A*;

$c = \Sigma_x \exp\{\Sigma_A \lambda_A z_A(\mathbf{x})\}$ is a normalising constant

(Besag, 1974)

Bernoulli blockmodels

- Suppose actors are either in block 1 or 2, and pairwise potentials
- Hammersley-Clifford:

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/c)\exp\{\Sigma_{i,j}\, \lambda_{ij}\, x_{ij}\}$$

- Block homogeneity:

  $\lambda_{ij} = \theta_{11}$ if *i* and *j* both in block 1

  $\lambda_{ij} = \theta_{12}$ if *i* in block 1 and *j* in block 2, etc.

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/c)\exp\{\theta_{11} L_{11} + \theta_{12} L_{12} + \theta_{21} L_{21} + \theta_{22} L_{22}\}$$

where L_{rs} is the number of edges from block r to block s.

- Extendable to multiple blocks

A Latent Mixture Membership Blockmodel

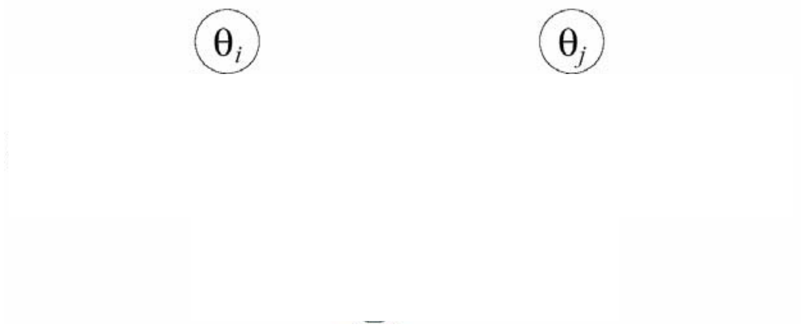
Motivation

- In many networks (e.g., biological network, citation networks), each node may be "multiple-class", i.e., has multiple functional/topical aspects.
- The interaction of a node (e.g., a protein) with different nodes (partners) may be under different function context.
- Prior knowledge of group interaction may be available.

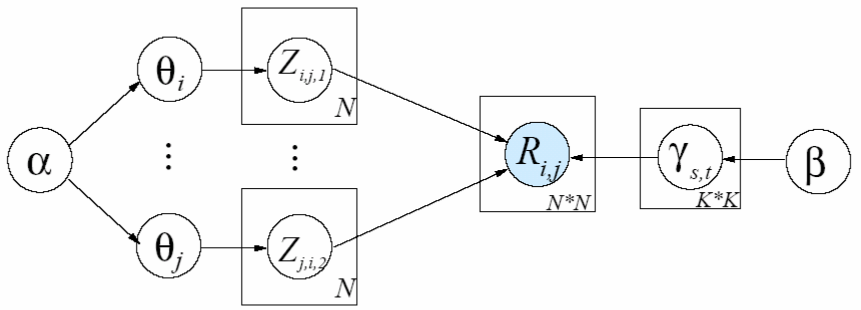
A Latent Mixture Membership Blockmodel

Topic vector of node *i*

Topic vector of node *j*



A Hierarchical Bayesian LMMB



For each object i=1,…,N:

$\theta_i \sim \mathrm{Dirichlet}(\alpha)$

For each topic-pair (*s,t*):

$\gamma_{s,t} \sim \mathrm{Beta}(\beta)$

For each pair of object (*i,j*)

$Z_{i,j,1} \sim \mathrm{Multi}(\theta_i)$

$Z_{i,j,2} \sim \mathrm{Multi}(\theta_j)$

$R_{i,j} \sim \mathrm{Bernoulli}\left(\rho\gamma_{z_{i,j,1},z_{i,j,2}} + (1-\rho)\delta_0\right)$

Variational Inference

- The Joint likelihood:

$$p(r,z,\theta,\gamma)=\prod_i \theta_i^{\sum_j z_{i,j,1}+z_{i,j,2}+\alpha-1}\times\gamma_{m,n}^{\sum_{i,j} r_{i,j} z^m_{i,j,1} z^n_{i,j,2}+\beta_1-1}(1-\gamma_{m,n})^{\sum_{i,j}(1-r_{i,j})z^m_{i,j,1}z^n_{i,j,2}+\beta_2-1}$$

- GMF approximation:

$$\mathrm{q}(\mathbf{r},\mathbf{z},\theta,\gamma\mid\alpha,\beta)=\left(\prod_{i=1}^{N}\mathrm{q}(\theta_i\mid\mu_i)\right)\times\left(\prod_{s=1,t=1}^{K}\mathrm{q}(\gamma_{s,t}\mid\nu_{s,t})\right)\times\left(\prod_{i=1,j=1}^{N}\mathrm{q}(z_{i,j,1},z_{i,j,2},r_{i,j}\mid\varphi_{i,j})\right)$$

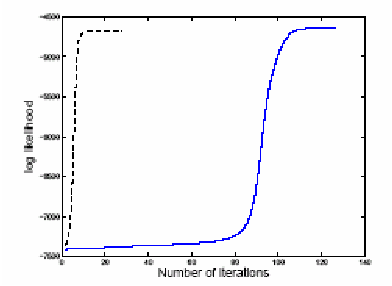
$$\mu_i=\alpha+\sum_j\left\langle z_{i,j,1}\right\rangle+\sum_j\left\langle z_{i,j,2}\right\rangle$$

$$\nu_{s,t}=\beta+\sum_{i,j} r_{s,t}\left\langle z_{i,j,1}z_{i,j,2}\right\rangle$$

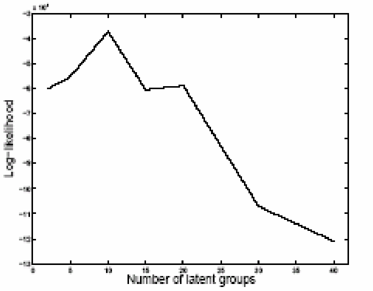
- MF approximation: …

$$\mathrm{q}(\mathbf{r},\mathbf{z},\theta,\gamma\mid\alpha,\beta)=\left(\prod_{i=1}^{N}\mathrm{q}(\theta_i\mid\mu_i)\right)\times\left(\prod_{s=1,t=1}^{K}\mathrm{q}(\gamma_{s,t}\mid\nu_{s,t})\right)\times\left(\prod_{i=1,j=1}^{N}\mathrm{q}(z_{i,j,1}\mid\phi_{i,j,1})\mathrm{q}(z_{i,j,2}\mid\phi_{i,j,1})\mathrm{q}(r_{i,j}\mid\varphi_{i,j})\right)$$

Experiments



Convergence

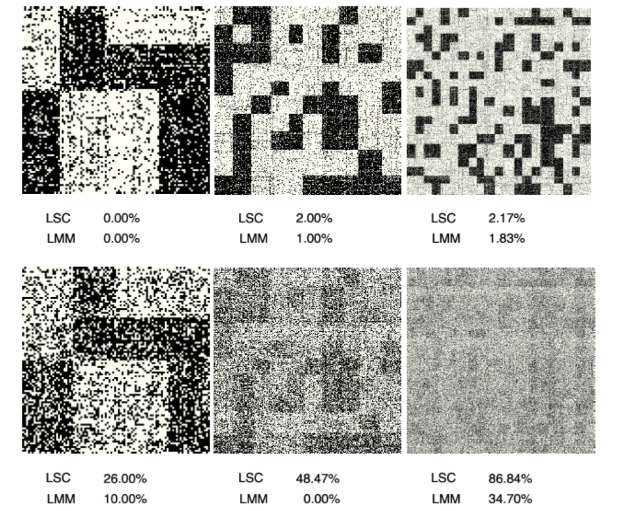


Model Selection

LMMB and SC on Simulated Data

stringent

diffused



100 300 600

Protein-Protein Interaction Data

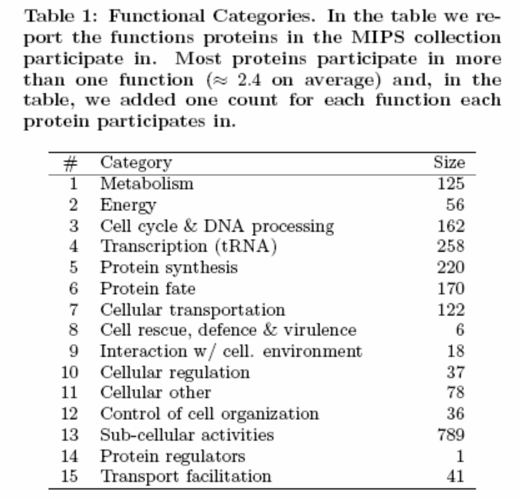
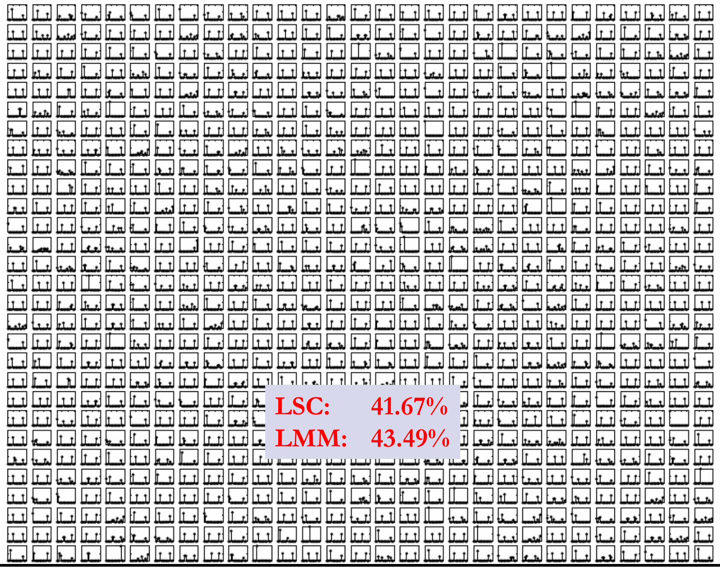


Table 1: Functional Categories. In the table we report the functions proteins in the MIPS collection participate in. Most proteins participate in more than one function (≈ 2.4 on average) and, in the table, we added one count for each function each protein participates in.

#	Category	Size
1	Metabolism	125
2	Energy	56
3	Cell cycle & DNA processing	162
4	Transcription (tRNA)	258
5	Protein synthesis	220
6	Protein fate	170
7	Cellular transportation	122
8	Cell rescue, defence & virulence	6
9	Interaction w/ cell. environment	18
10	Cellular regulation	37
11	Cellular other	78
12	Control of cell organization	36
13	Sub-cellular activities	789
14	Protein regulators	1
15	Transport facilitation	41

Inferred Membership



LSC: 41.67%
LMM: 43.49%

Supervised Prediction of Membership

- Learning q and g from training data and predict r :

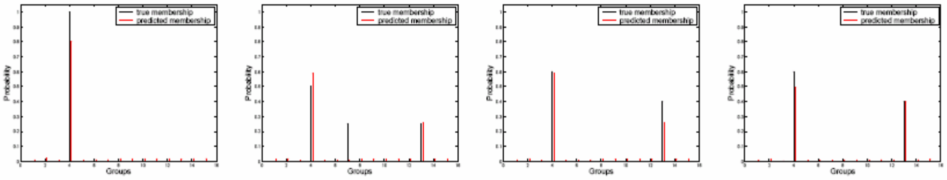


Figure 5: Predicted (red) versus true (black) mixed-membership probabilities for four example proteins.

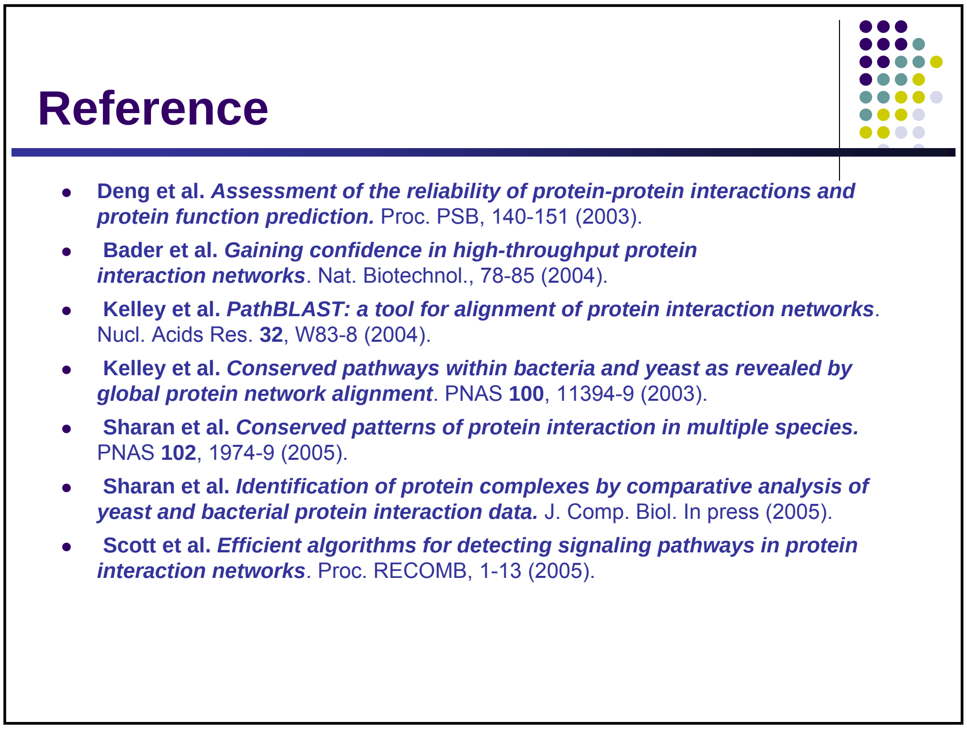
Accuracy: 45.12%

Summary of LMMB

- A stochastic block model
- Each node can play "multiple roles", and its ties with other nodes can be explained by different roles
- Hierarchical Bayesian formalism
- Efficient variational inference

Acknowledgements

- Mark Gerstein
- Roded Sharan
- Jotun Hein
- Batzoglou



Reference

- **Deng et al.** ***Assessment of the reliability of protein-protein interactions and protein function prediction.*** Proc. PSB, 140-151 (2003).
- **Bader et al.** ***Gaining confidence in high-throughput protein interaction networks***. Nat. Biotechnol., 78-85 (2004).
- **Kelley et al.** ***PathBLAST: a tool for alignment of protein interaction networks***. Nucl. Acids Res. **32**, W83-8 (2004).
- **Kelley et al.** ***Conserved pathways within bacteria and yeast as revealed by global protein network alignment***. PNAS **100**, 11394-9 (2003).
- **Sharan et al.** ***Conserved patterns of protein interaction in multiple species.*** PNAS **102**, 1974-9 (2005).
- **Sharan et al.** ***Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data.*** J. Comp. Biol. In press (2005).
- **Scott et al.** ***Efficient algorithms for detecting signaling pathways in protein interaction networks***. Proc. RECOMB, 1-13 (2005).