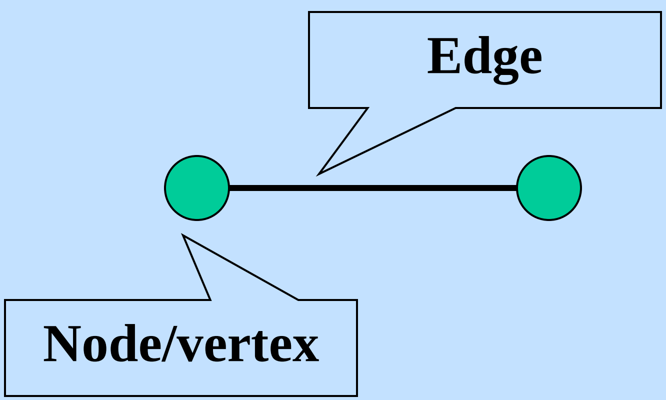
Networks & Modules

1. Networks
2. Protein complex identification
3. Pathway identification

Networks

- Represent relations between elements.
- *Nodes* – elements (towns).
- *Edges* – relations (roads).

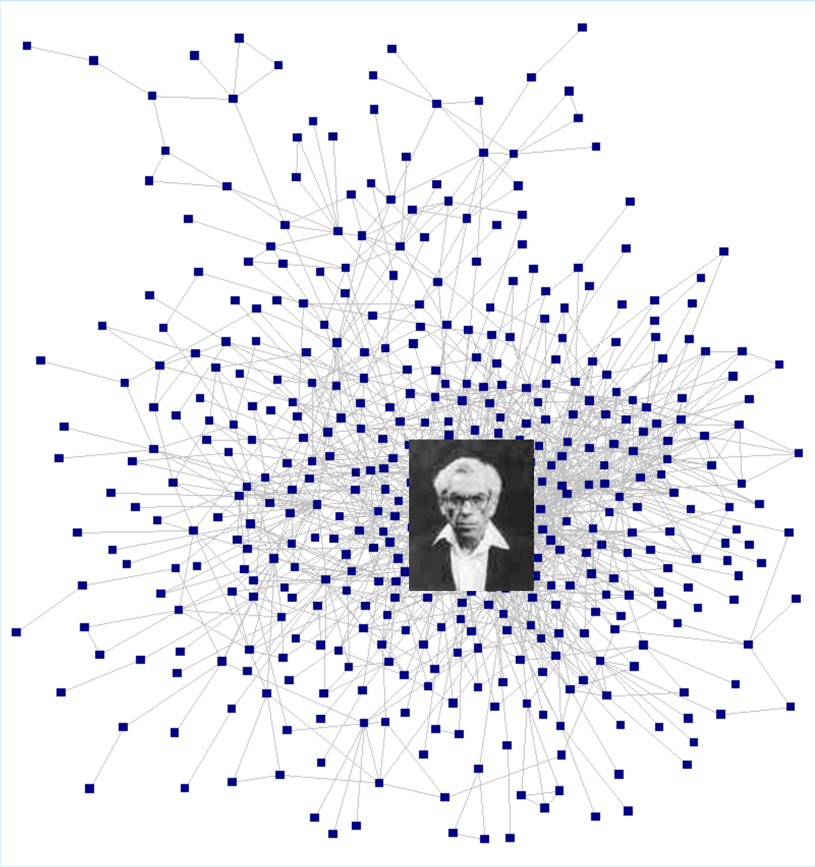


It's a Small World

Milgram'67: six degrees of separation.

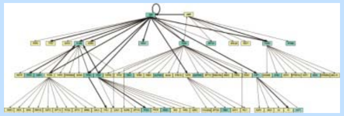
Collaboration Networks

Nodes – collaborators (scientists)
Edges – acts of collaboration (joint articles)

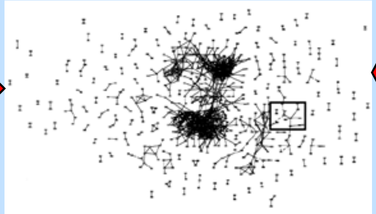


http://www.orgnet.com/Erdos.html

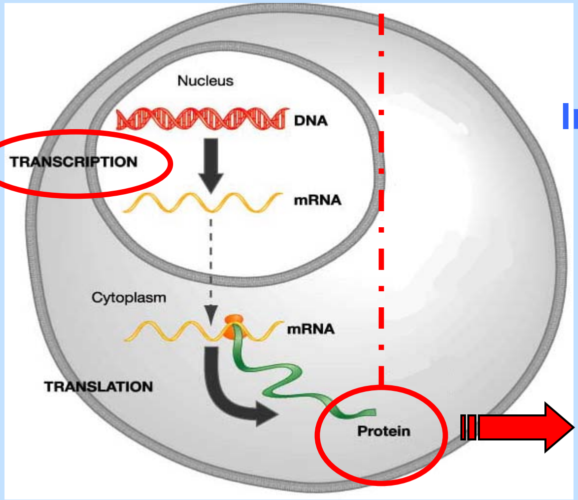
Molecular Networks



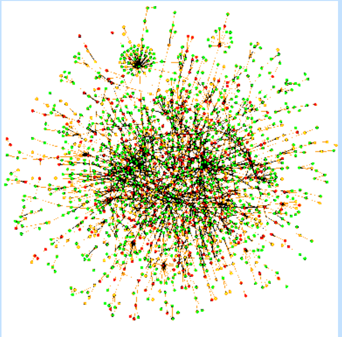
Regulatory networks

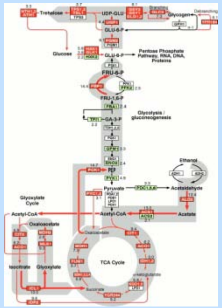


Expression networks



Interaction networks





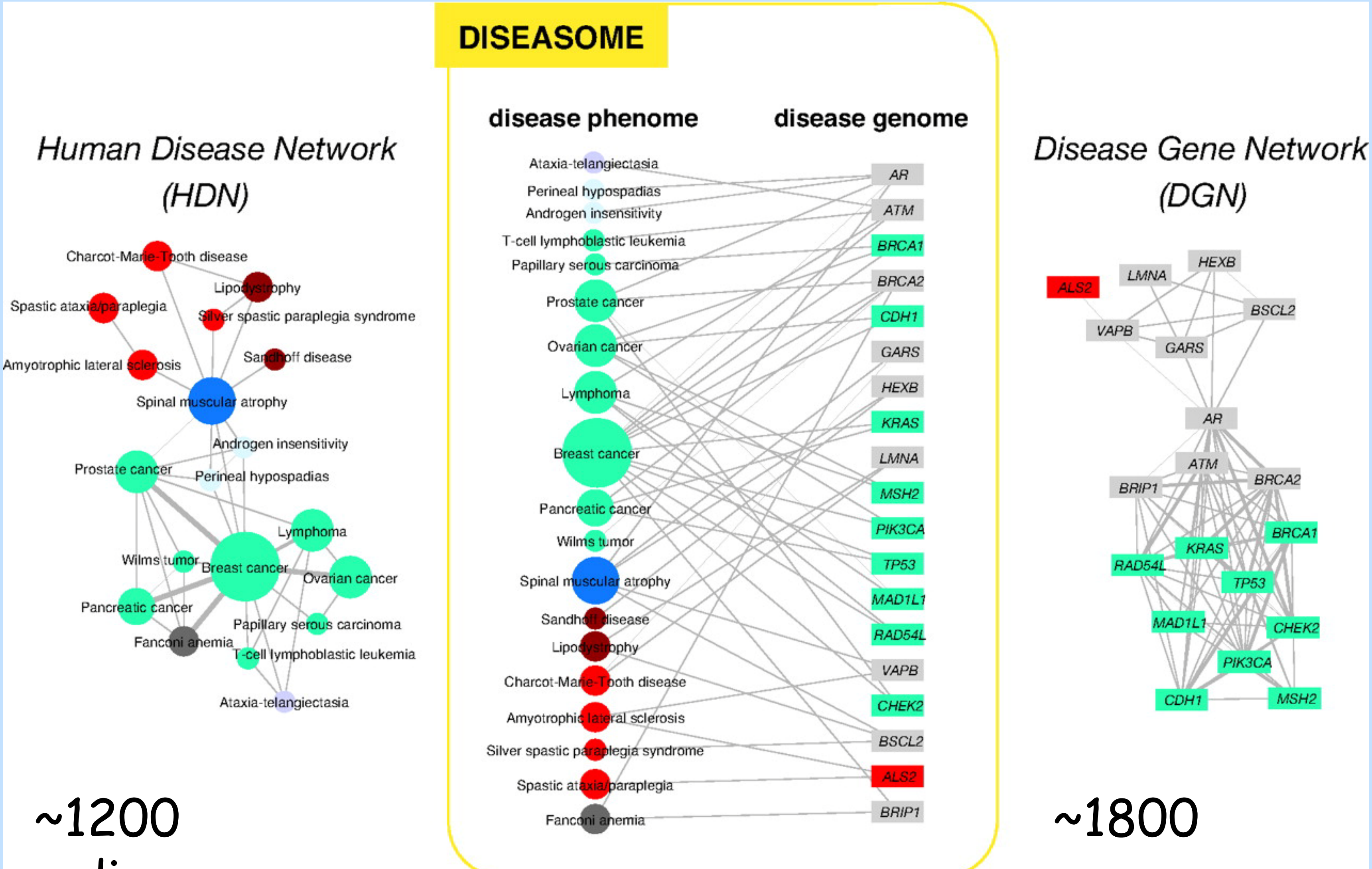
Metabolic networks

Nodes – molecules

Edges – interactions / similarity

The human disease network

Kwang-Il Goh*†‡§, Michael E. Cusick†‡¶, David Valle‖, Barton Childs‖, Marc Vidal†‡¶**, and Albert-László Barabási*†‡**

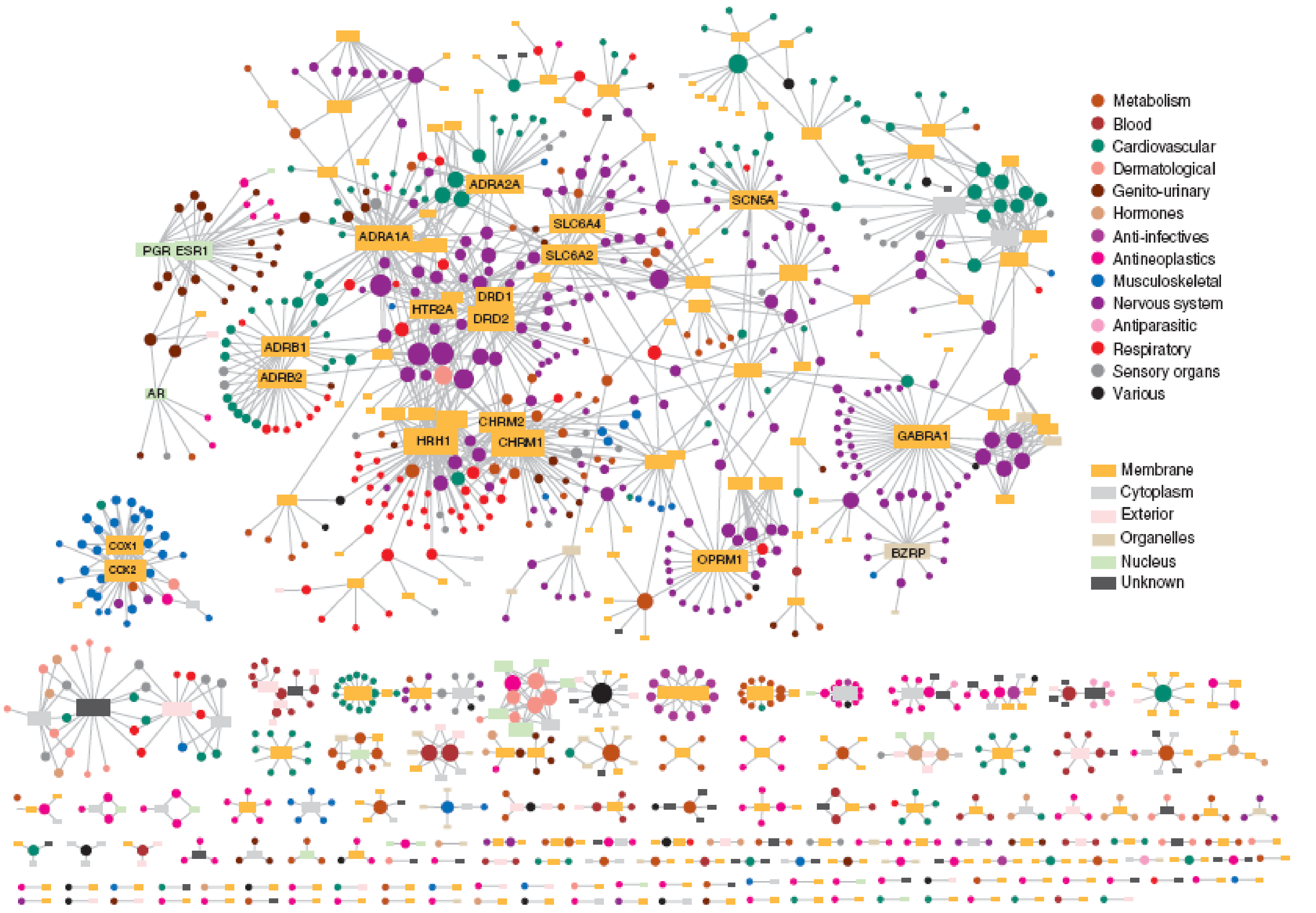


~1200 diseases

~1800 genes

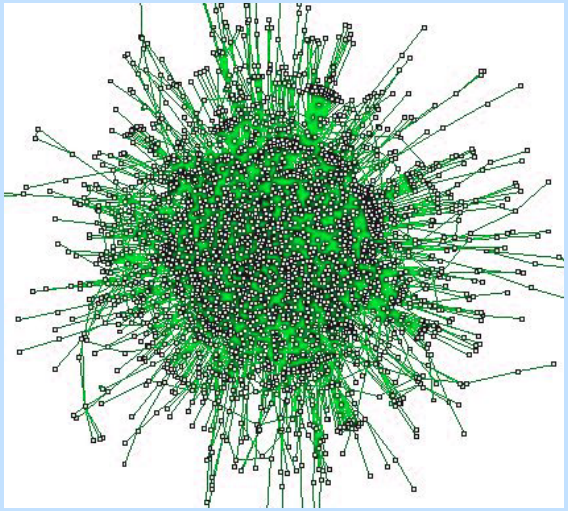
Drug–target network

Muhammed A Yıldırım[1,2,3], Kwang-Il Goh[1,4,5], Michael E Cusick[1,2], Albert-László Barabási[1,4,6] & Marc Vidal[1,2]

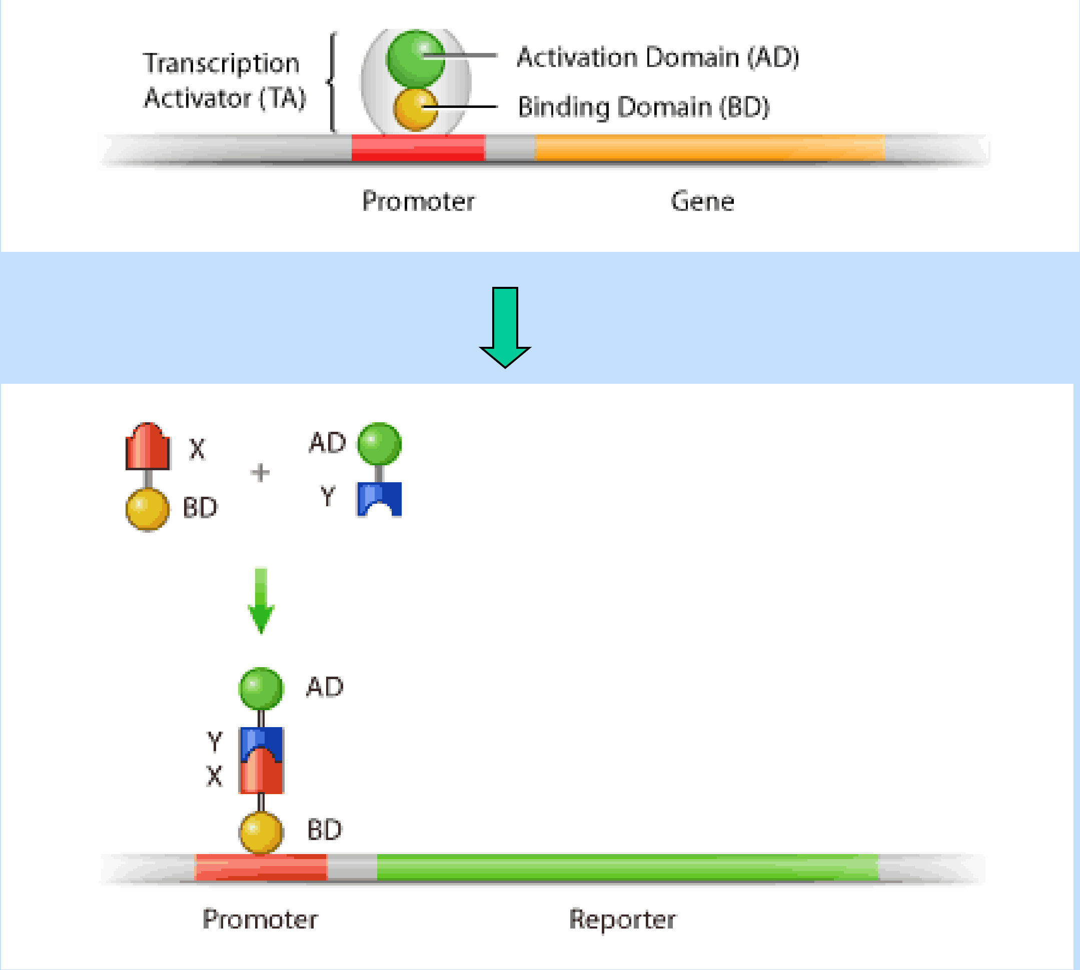


Protein-protein Interaction Networks

- *Nodes* – proteins (6K).
- *Edges* – interactions (40K).
- Reflect the cell’s machinery and signlaing pathways.
- Measured by high-throughput technologies:
  - yeast two-hybrid
  - co-immunoprecipitation



Yeast Two-Hybrid

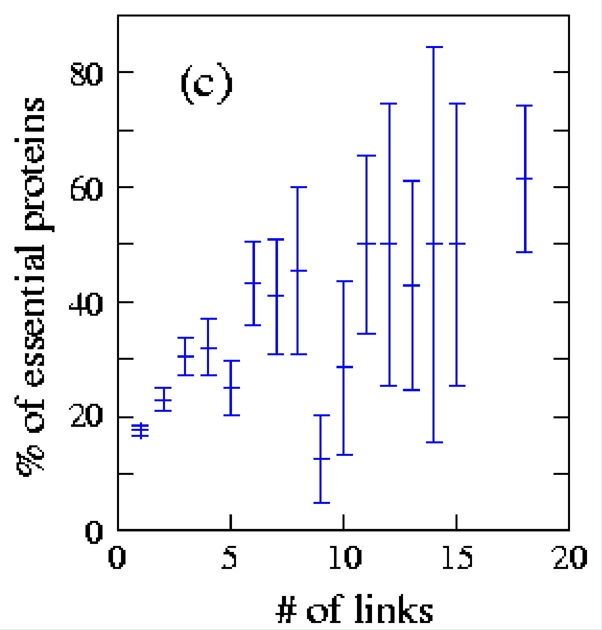


http://www.bioteach.ubc.ca/MolecularBiology/AYeastTwoHybridAssay/

Network properties: Degree

Why is degree important?

- *Degree*: #neighbors.
- Local characterization of a node.
- Indicates its centrality in the network.



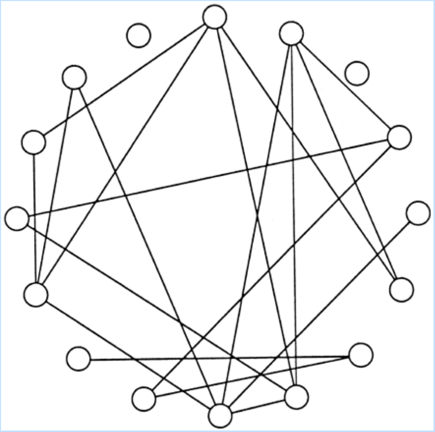
Jeong et al, Nature 411, 41-2, 2001

Degree Distribution

- *Degree distribution P(k)*: probability that a node has degree *k*.
- For directed graphs, two distributions: in-degree and out-degree.
- Average degree: $$d \equiv \sum_{k \geq 0} kP(k)$$
- Number of edges: *Nd/2*.

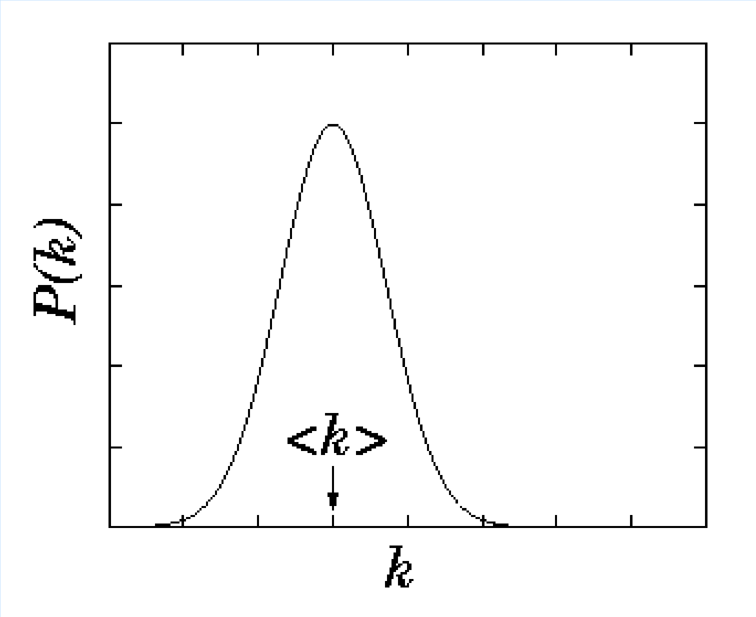
Random Networks (Erdös/Rényi)

- N nodes.
- Every pair of nodes is connected with probability p.

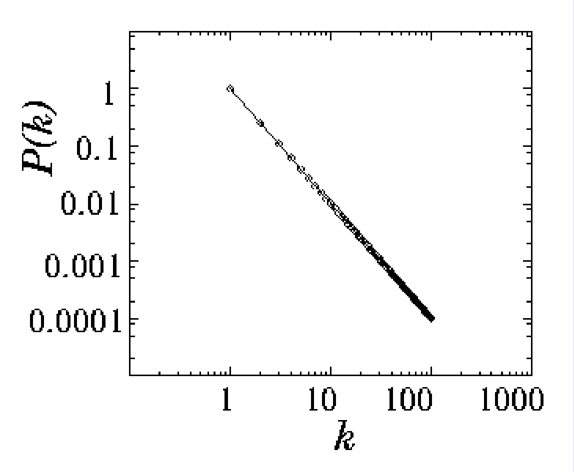


- Mean degree: $d=(N-1)p \sim Np$.
- Degree distribution is binomial, asymptotically Poisson:

$$P(k) = \frac{e^{-d} d^k}{k!}$$

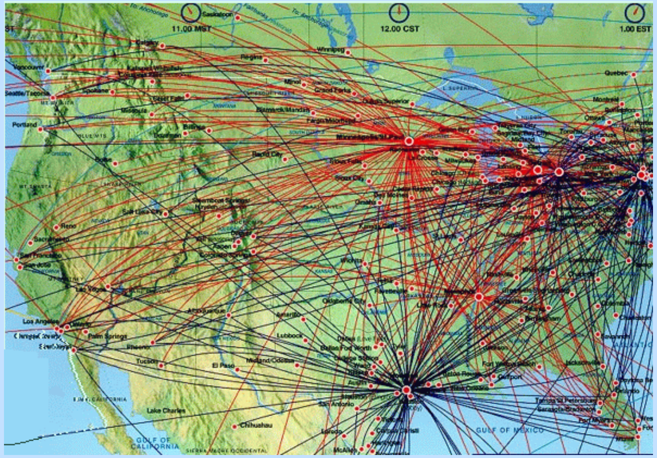


Scale-Free Networks (Barabasi&Albert'99)



- Power-law degree distribution

$$P(k) \propto k^{-c}, k \neq 0, c > 1$$



- Characterized by a small number of highly connected nodes, known as *hubs*.

Scale-Free Distribution

- A distribution *p(x)* that is scale-invariant, i.e.:

$p(ax)=g(a)p(x)$

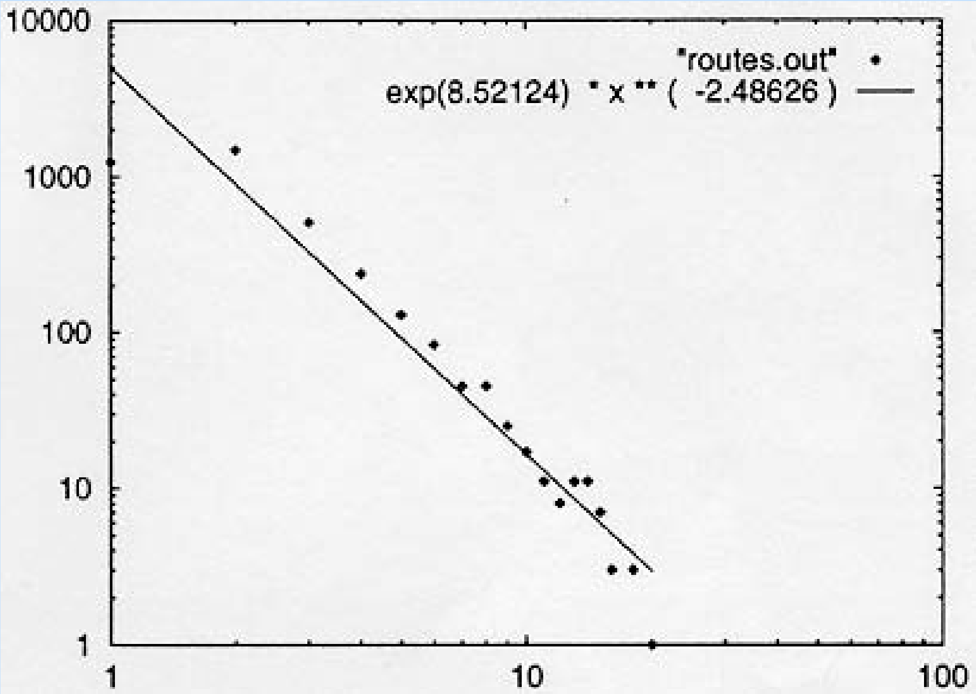
- It can be shown that the only scale free distributions are power-law distributions!!!

Are Real Networks Random or Scale-Free?

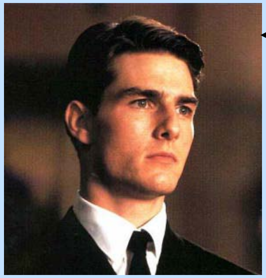
The Internet

Nodes – routers.
Edges – physical links.

***P(k) ~ k* $^{-2.5}$**
(Faloutsos et al.'99)



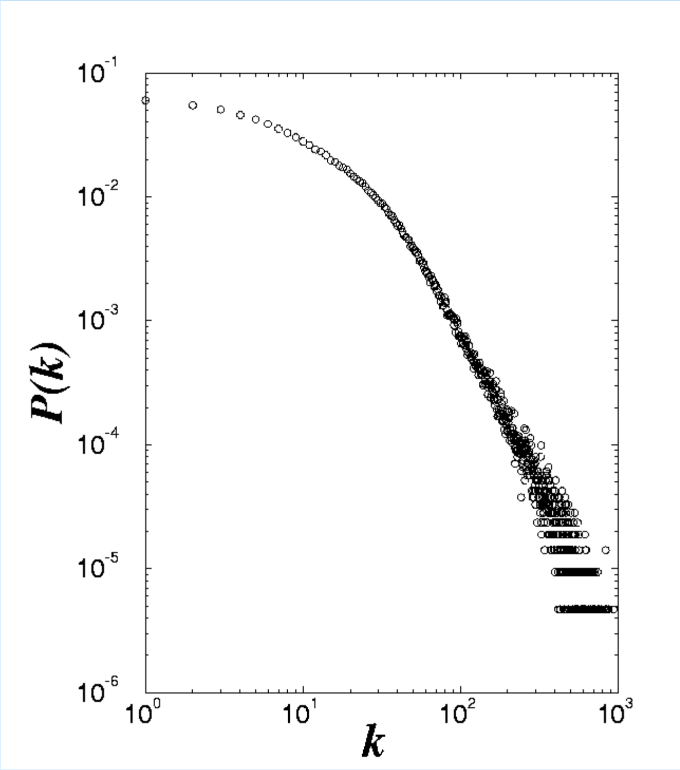
Film Actors



Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

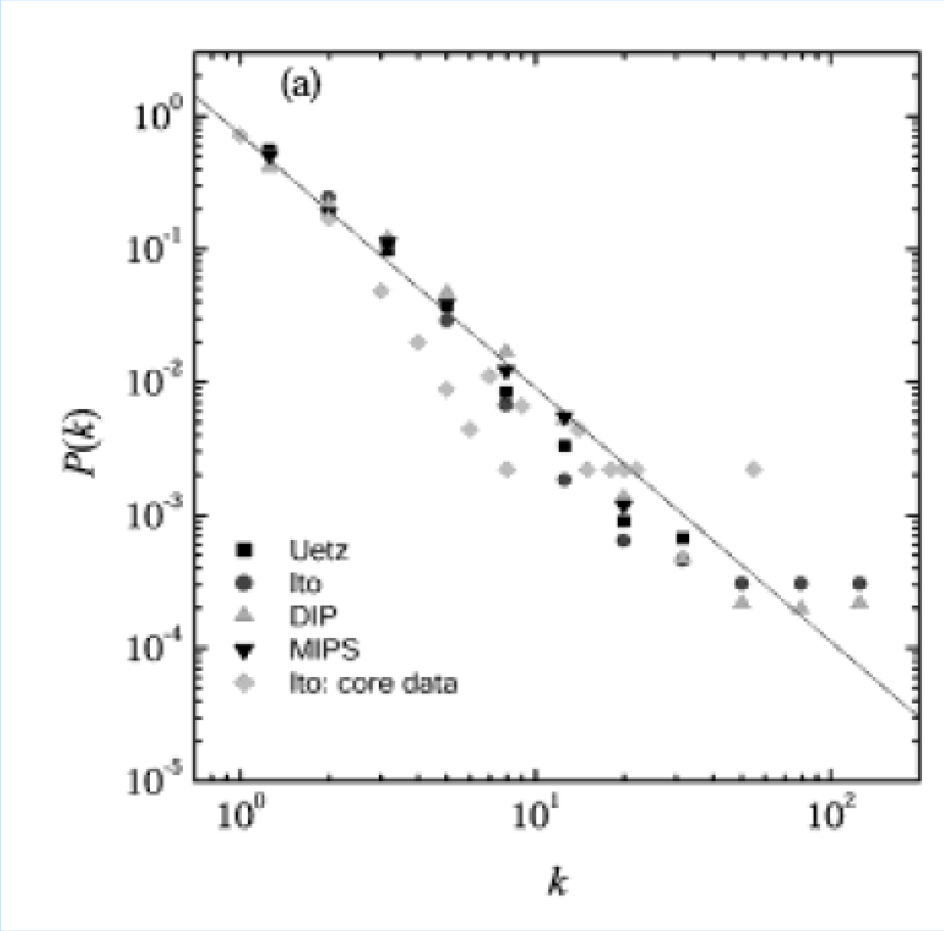
Nodes – actors.
Edges – joint movies.

$\boldsymbol{P(k) \sim k^{-2.3}}$
(Barabasi&Albert'99)



Protein Interaction Networks

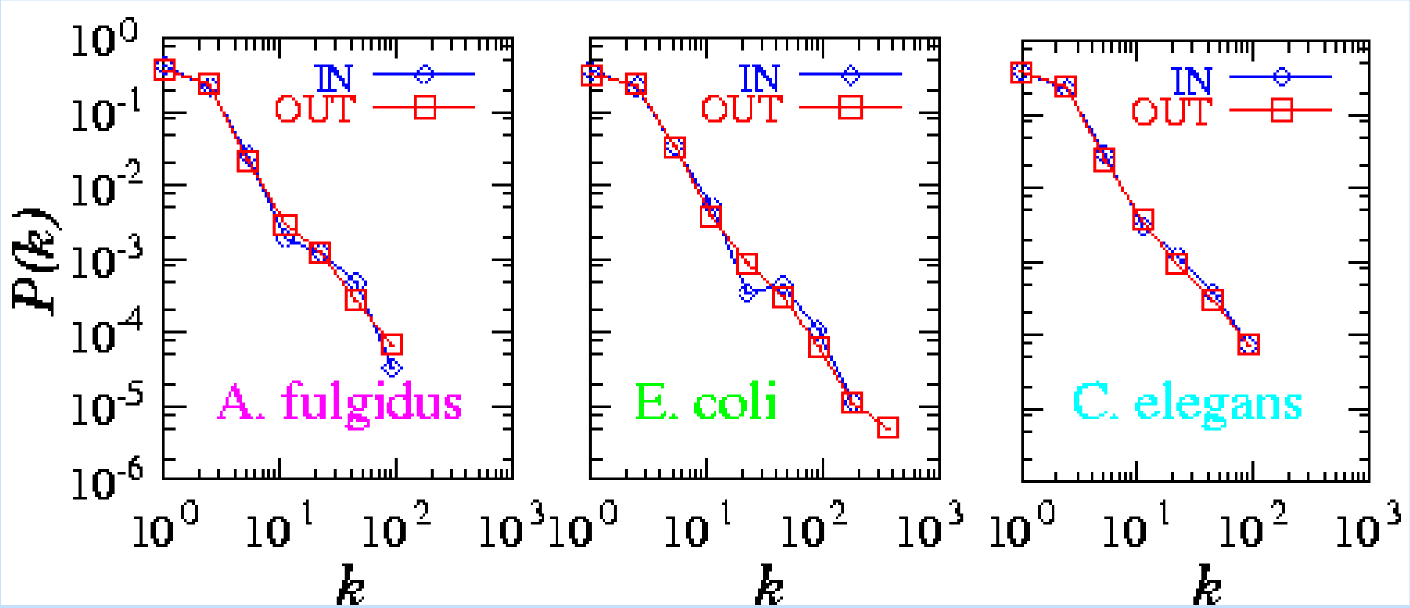
- *Nodes* – proteins.
- *Edges* – interactions.



$P(k) \sim k^{-2.5}$
(Yook et al.'04)

Metabolic Networks

- *Nodes* – metabolites.
- *Edges* – biochemichal reactions.



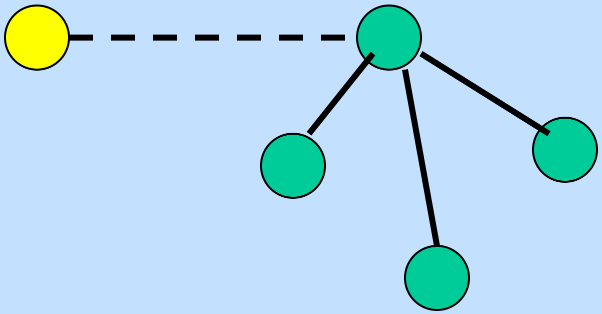
Metabolic networks from all kingdoms of life are scale-free

c=2.2±0.2 (Jeong et al.’00)

Why Are Real Networks Scale-Free?

Scale-Free Model (Barábasi & Albert)

- **Growth:** nodes are constantly added.
- **Preferential attachment:** the probability that a new node connects to existing ones is proportional to their degree.



In resulting network:

$$P(k) \approx k^{-3}$$

Relevance to biology?

Clustering

Clustering Coefficient
(Watts & Strogatz)

- Characterizes tendency of nodes to cluster

$$C(v) = \frac{\#\{\text{pairs of connected neighbors of v}\}}{d(v)(d(v)-1)/2}$$

$$C = \frac{1}{N}\sum_{v} C(v)$$

(if $d(v)=0,1$ then $C(v)$ is defined to be 0)

- Lies in [0,1].

- What is C for random graphs?

Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

(Taken from Pruzlj'05)

Shortest Paths

Small World

- What is the avg. distance in a random network?
- Fact: a random network is locally tree-like (exponential growth of neighbors with distance)
- d^i vertices on avg. are at distance *i* or closer from a vertex.
- Since $N \sim d^l$ we have $l \sim \ln N / \ln d$ – *small world effect*.
- Implies fast spread of information.

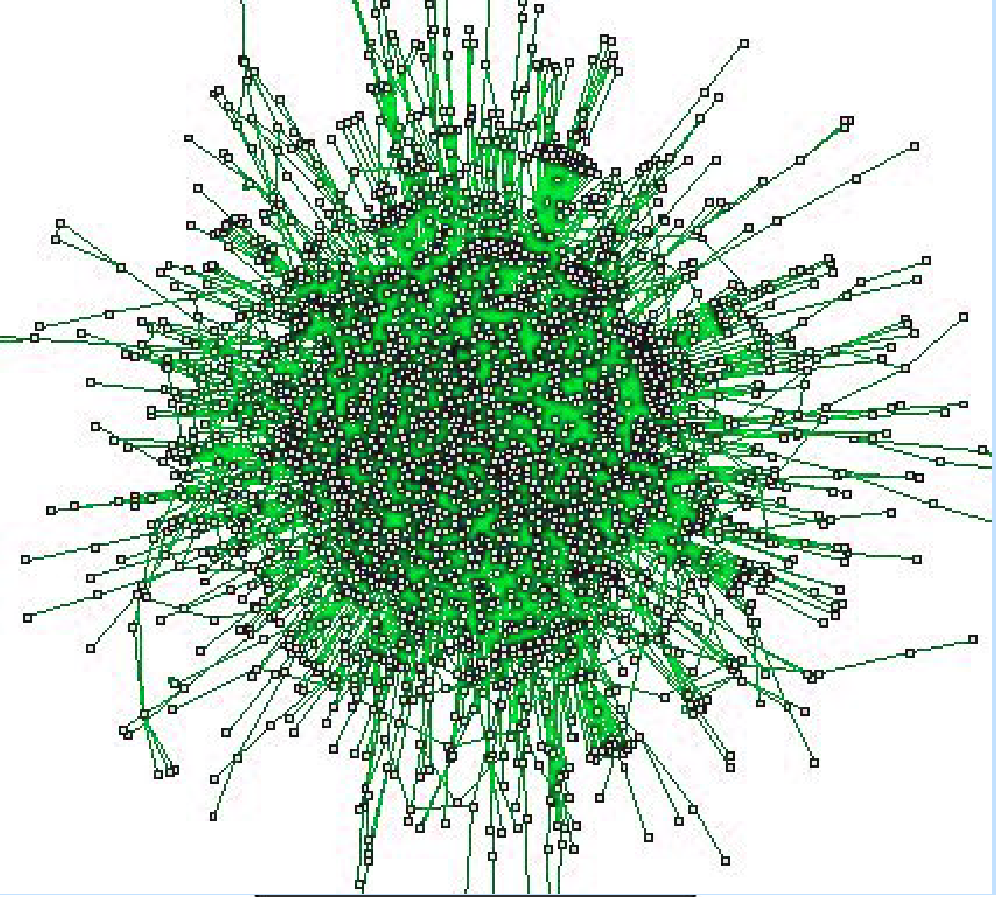
	network	type	n	m	z	ℓ
social	film actors	undirected	449 913	25 516 482	113.43	3.48
	company directors	undirected	7 673	55 392	14.44	4.60
	math coauthorship	undirected	253 339	496 489	3.92	7.57
	physics coauthorship	undirected	52 909	245 300	9.27	6.19
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92
	telephone call graph	undirected	47 000 000	80 000 000	3.16	
	email messages	directed	59 912	86 300	1.44	4.95
	email address books	directed	16 881	57 029	3.38	5.22
	student relationships	undirected	573	477	1.66	16.01
	sexual contacts	undirected	2 810			
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18
	citation network	directed	783 339	6 716 198	8.57	
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87
	word co-occurrence	undirected	460 902	17 000 000	70.13	

(Taken from Newman'03)

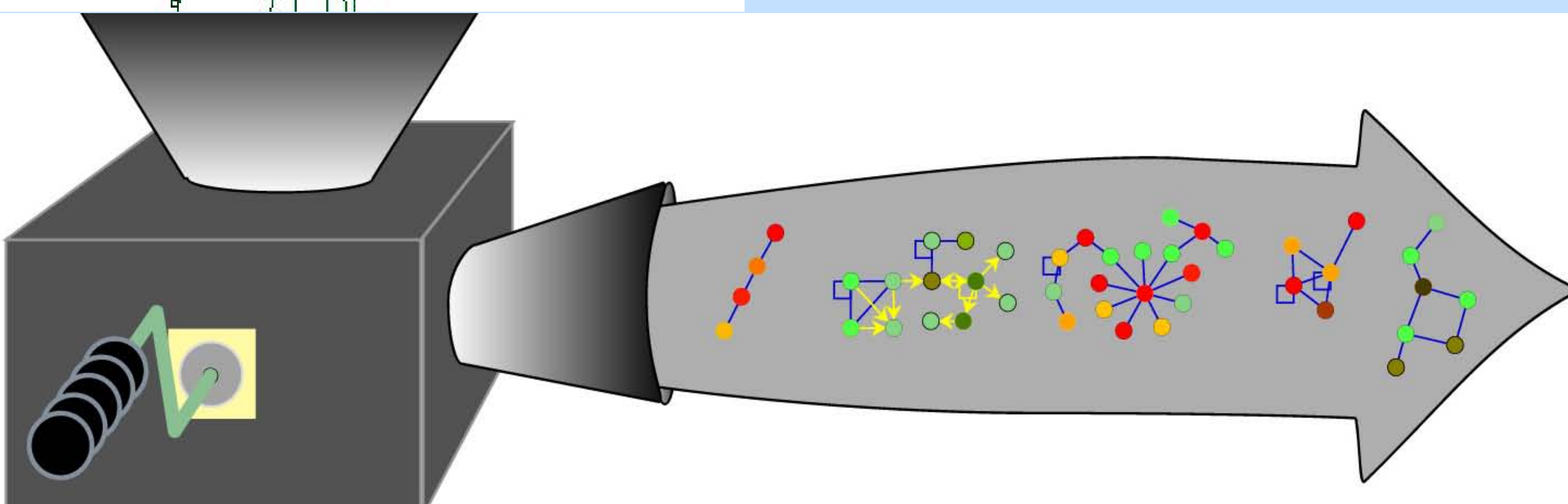
Module Identification

Gene/Protein Modules

- A *module* is a set of genes/proteins performing a distinct biological function.
- Characterized by a coherent behavior of its genes w.r.t. a certain biological property.
- Examples:
  - *transcriptional module*: a set of co-expressed genes sharing a common function.
  - *protein complex*: assembly of proteins that build up some cellular machinery.
  - *signaling pathway*: a chain of interacting proteins propagating a signal in the cell.

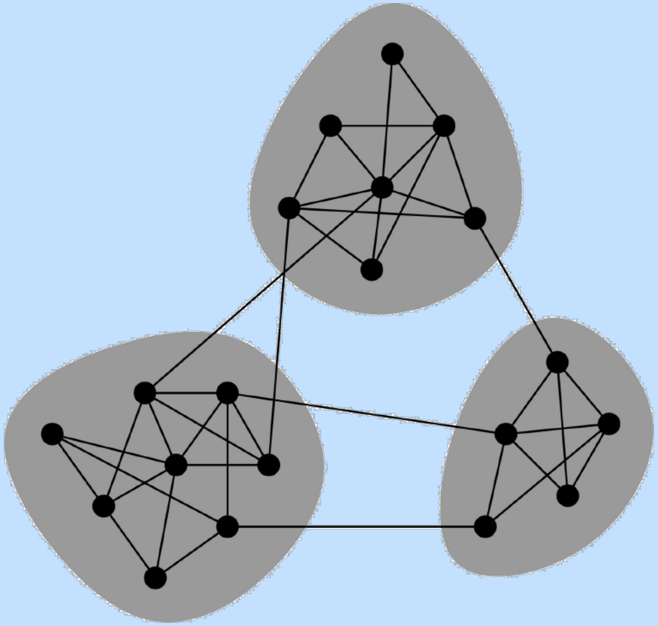


Distilling Modules from Networks



Modularity and Community Structure in Networks

M.E.J Newman, PNAS 2006



Modularity of a division (Q)

Q = #(edges within groups) - E(#(edges within groups in a RANDOM graph with same node degrees))

Trivial division: all vertices in one group

==> Q(trivial division) = 0

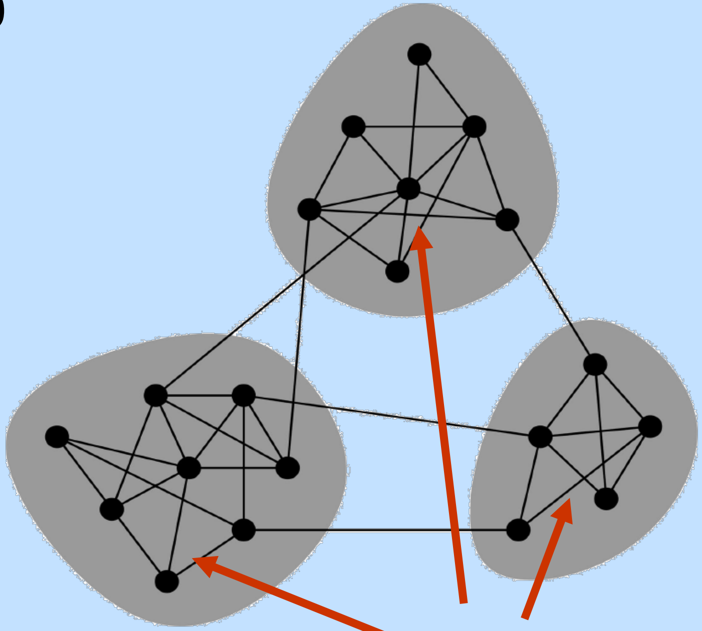
k_i = degree of node i

M = $\sum k_i$ = 2|E|

Aij = 1 if (i,j)$\in$E, 0 otherwise

Eij = expected number of edges between i and j in a random graph with same node degrees.

Lemma: Eij $\approx k_i * k_j$ / M



Q = $\sum$(Aij - ki*kj/M | i,j in the same group)

Division into two groups

Q = ∑(Aij - ki*kj/M | i,j in the same group)

- Suppose we have n vertices {1,...,n}
- **s** - {±1} vector of size n.
  Represent a 2-division:
  - si == sj iff i and j are in the same group
  - ½ (si*sj+1) = 1 if si==sj, 0 otherwise
- ==>

$$Q = \frac{1}{2}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{M})(s_i s_j + 1)$$

Division into two groups (2)

$$Q = \frac{1}{2}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{M})(s_i s_j + 1)$$

Since $\sum_{i,j} A_{ij} = \sum_i k_i = M$

$$Q = \frac{1}{2}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{M}) s_i s_j$$

B = the modularity matrix
- symmetric

$$Q = \frac{1}{2}\boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s}$$

where $$B_{ij} = A_{ij} - \frac{k_i k_j}{M}$$

Division into two groups (3)

B is symmetric $\Rightarrow$ **B** is diagonalizable (real eigenvalues)

B's eigenvalues

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$$

B's orthonormal eigenvectors

$$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_n$$

$$\boldsymbol{B}\boldsymbol{u}_i = \beta_i \boldsymbol{u}_i$$

$$Q = \frac{1}{2}\boldsymbol{s}^T \boldsymbol{B}\boldsymbol{s} \quad \left[\boldsymbol{s} = \sum_i a_i \boldsymbol{u}_i\right] \quad \Rightarrow \quad Q = \frac{1}{2}\sum_i \beta_i a_i^2$$

- Which vector **s** maximizes Q?
  - clearly **s** ~ **u1** maximizes Q, but **u1** may not be {±1} vector
  - Heuristic: maximize the projection of s on u1 (a_1): choose si= +1 if $u1_i$>0, si=-1 otherwise

Identifying protein pathways

Finding Simple Paths

Problem: Given a graph $G=(V,E)$ and a parameter k, find a simple path of length k in G.

- NPC by reduction from Hamiltonian path.
- Trivial algorithm runs in $O(n^k)$.
- We will be interested in a *fixed parameter* algorithm (Downey & Fellows ’92) – i.e., time is exponential in k but polynomial in n.

Color Coding [AYZ'95]

<u>Problem:</u> Given a graph *G=(V,E)* and a parameter *k*, find a simple path with *k* vertices (length *k-1*) in *G*.

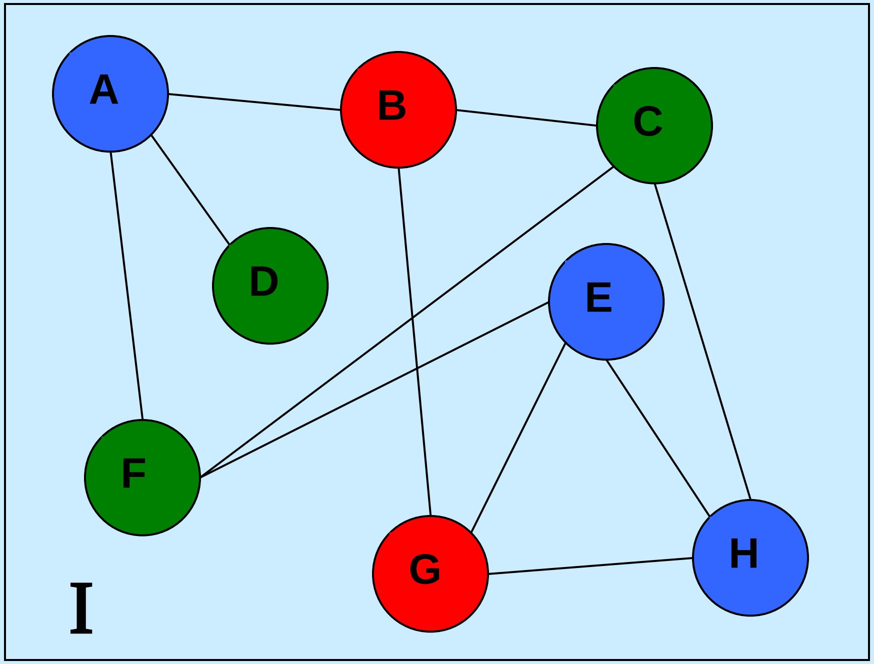
<u>Algorithm:</u> Randomly color vertices with *k* colors, and find a *colorful* path (distinct colors).

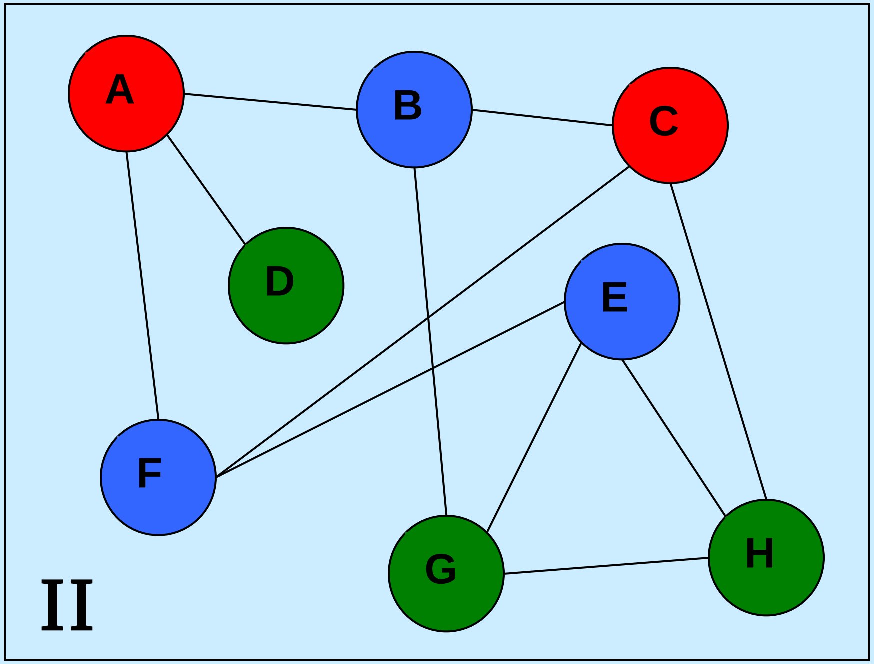
$$c : V \to [1,k]; S \in 2^{[1,k]}$$

$$P(v,S) = \max_{u:(u,v)\in E,\, c(u)\in S-\{c(v)\}} P(u, S-\{c(v)\})$$

<u>Main idea:</u> only 2^k color subsets vs. n^k node subsets.

Coloring Example





- Two different colorings on toy graph, *k*=3
- In coloring **I**, P(A,RGB) is built C->BC->ABC
- In coloring **II**, P(A,RGB) is built G->BG->ABG
- ABC is not colorful in coloring **II**

Randomization Analysis

- A colorful path is simple, but a simple path may not be colorful *under a given coloring*
- Solution: run multiple independent trials.
- After one trial:

$$\Pr(Success) = k!/k^k \geq 1/e^k$$

Color Coding [AYZ'95]

<u>Complexity:</u>

– Space complexity is $O(2^k n)$.

– Colorful path found by DP in $O(km2^k)$.

– $O(e^k)$ iterations are sufficient.

– Overall time is $2^{O(k)}m$.

– Note that the exponential part involves the parameter only, that is, the problem is *fixed parameter tractable*.

Comparison of Running Times

Path length	Color coding	Exhaustive
8	435	866
9	2,149	15,120
10	11,650	--

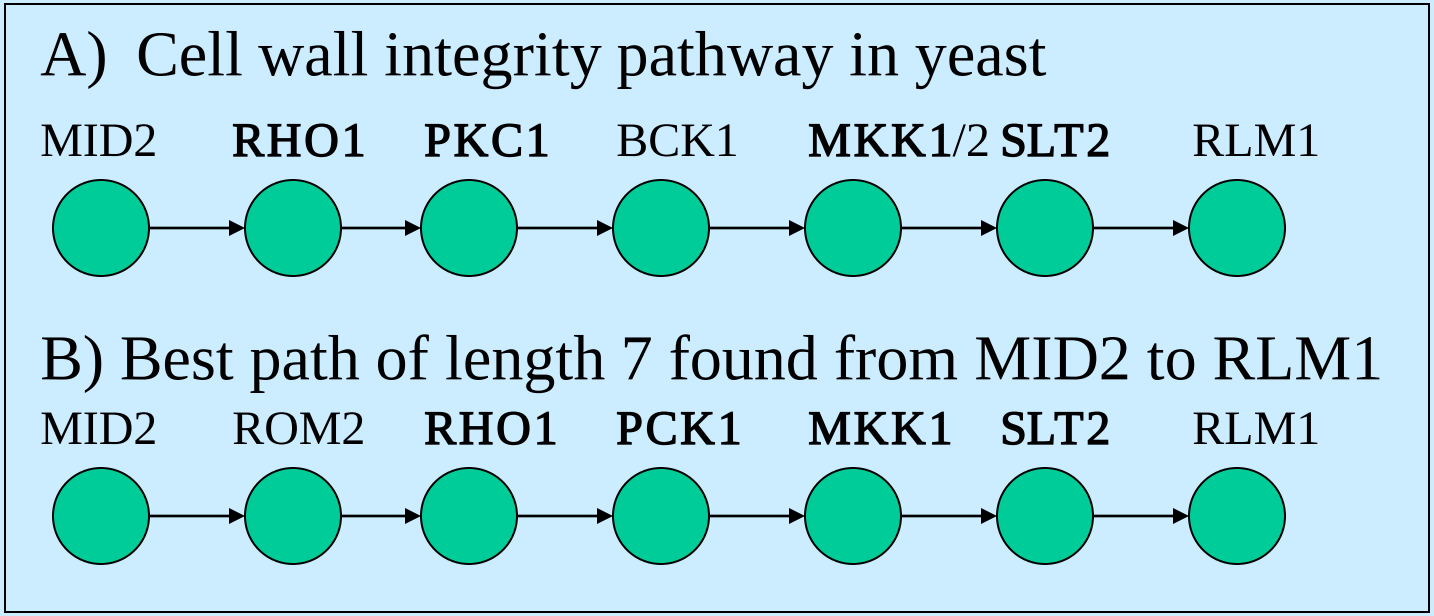
- ~4500 vertices, ~14500 edges.

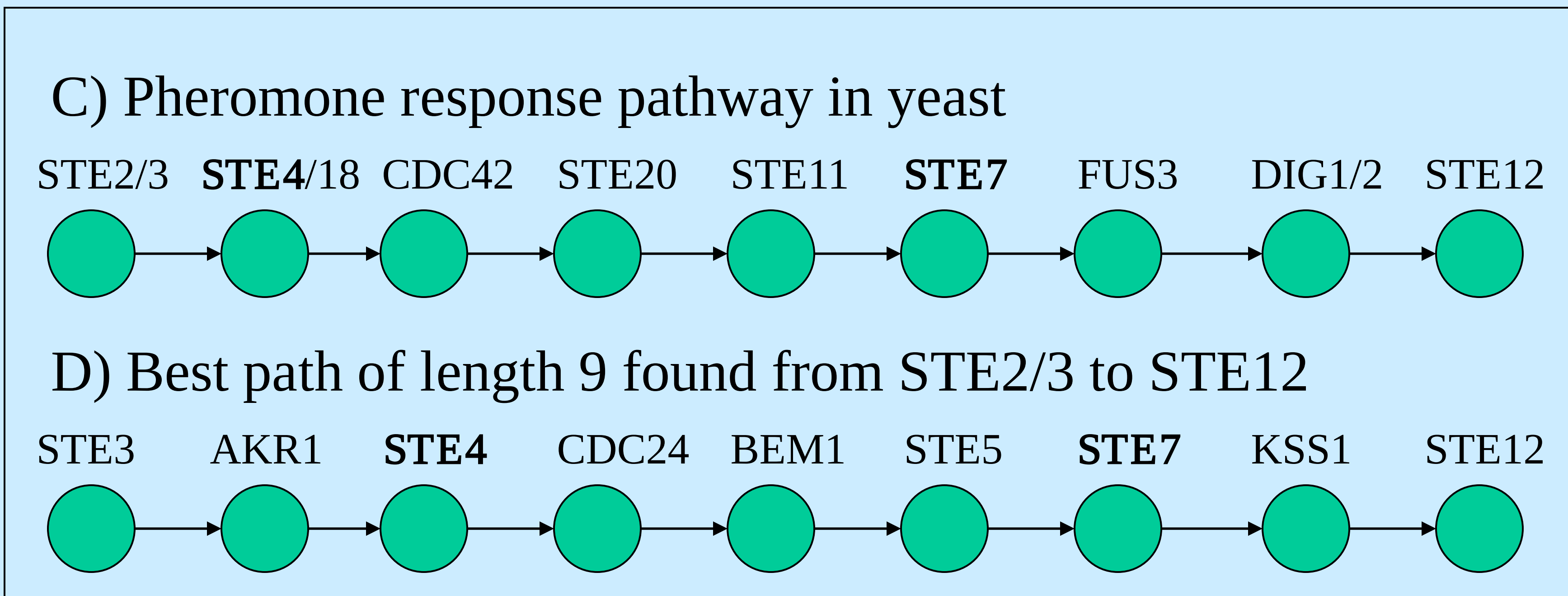
Scott et al. JCB 2005

Biologically-Motivated Constraints

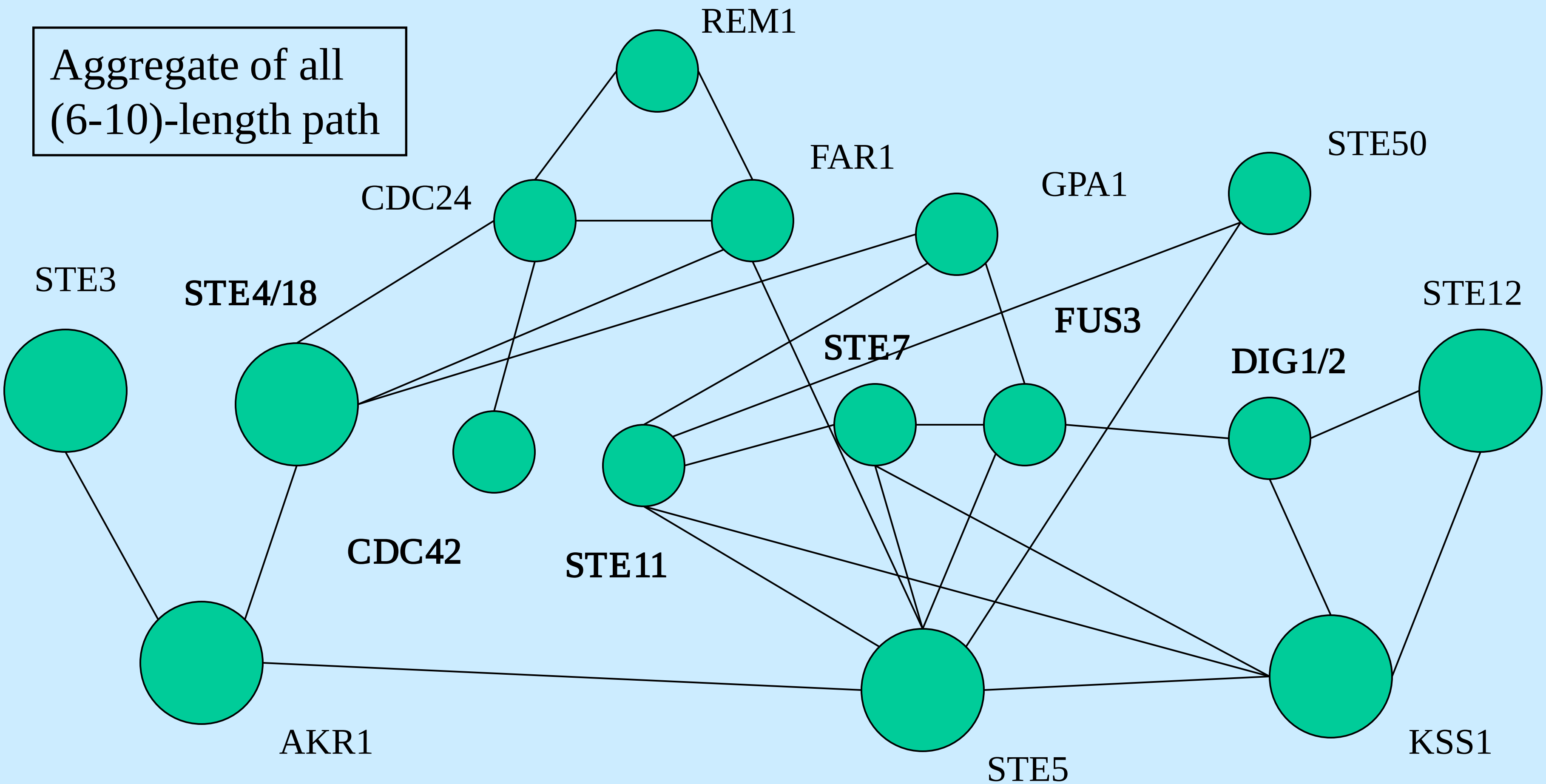
- Color-Coding gives an algorithmic basis, now introduce biologically motivated extensions.
- Can introduce edge weights (confidence).
- Can constrain the start or end of a path by type.
  - Steffen et al. ’02: pathways from membrane to TF.
- Can force the inclusion of a specific protein on the path by …

Appl. to yeast





A Closer Look at Pheromone Response



The real pathway (main chain):