Computational Genomics

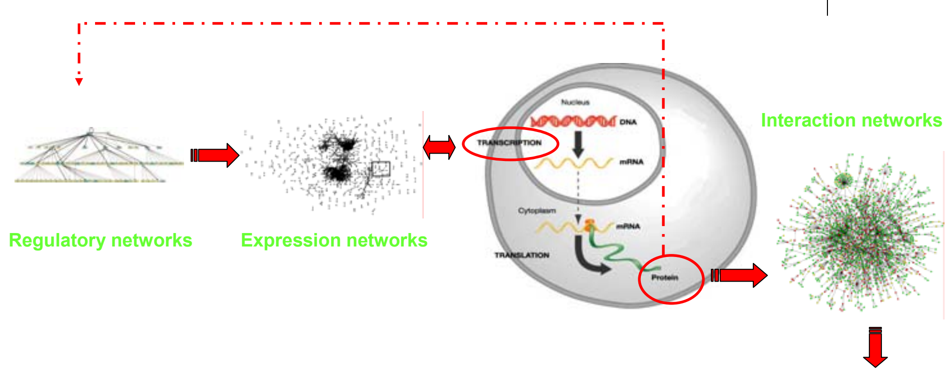
Biological Networks & Network Evolution

Eric Xing

Lecture 21, April 3, 2007

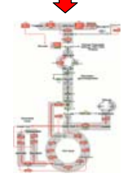
Reading:

Molecular Networks

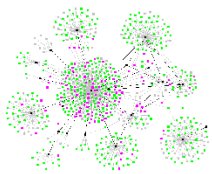


Nodes – molecules.
Links – inteactions / *relations*.

Metabolic networks

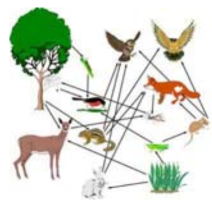


Other types of networks

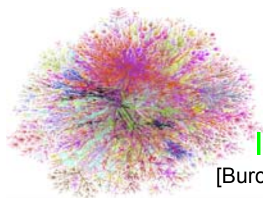


Disease Spread
[Krebs]

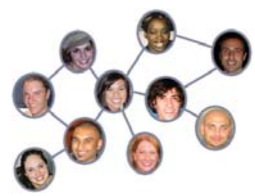
Electronic Circuit



Food Web

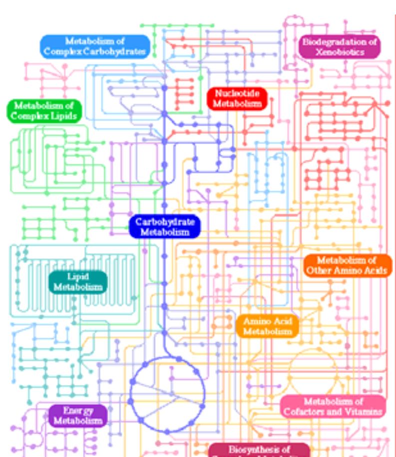


Internet
[Burch & Cheswick]

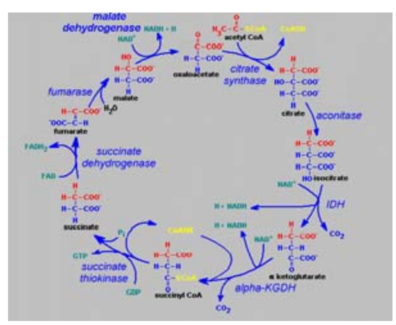


Social Network

Metabolic networks

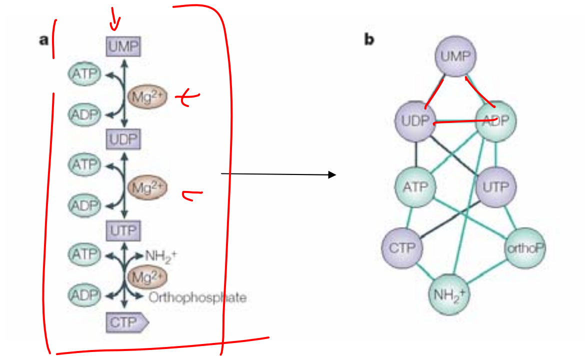


KEGG database: http://www.genome.ad.jp/kegg/kegg2.html



- *Nodes* – metabolites (0.5K).
- *Edges* – *directed* biochemichal reactions (1K).
- Reflect the cell's metabolic circuitry.

Graph theoretic description of metabolic networks



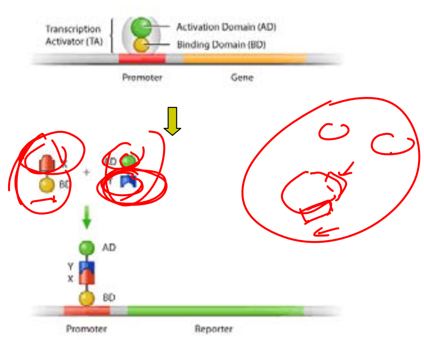
"Graph theoretic description for a simple pathway (catalyzed by Mg^{2+} -dependant enzymes) is illustrated (**a**). In the most abstract approach (**b**) all interacting metabolites are considered equally."

Barabasi & Oltvai. NRG. (2004) 5 101-113

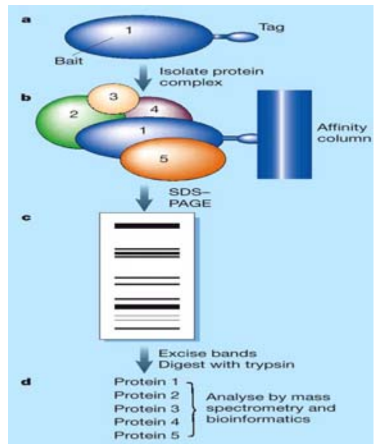
Protein Interaction Networks

- Nodes – proteins (6K).
- Edges – interactions (15K).
- Reflect the cell's machinery and signlaing pathways.

Experimental approaches



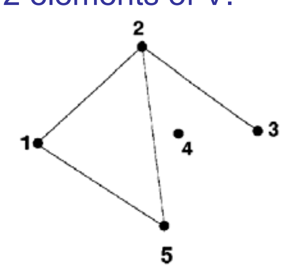
Yeast Two-Hybrid



Protein coIP

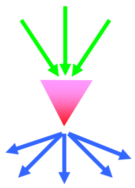
Graphs and Networks

- Graph: a pair of sets G={V,E} where V is a set of nodes, and E is a set of edges that connect 2 elements of V.
- Directed, undirected graphs
- Large, complex networks are ubiquitous in the world:
  - Genetic networks
  - Nervous system
  - Social interactions
  - World Wide Web

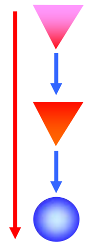


Global topological measures

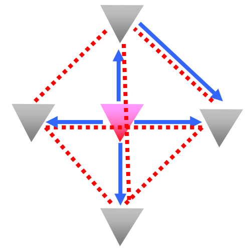
- Indicate the gross topological structure of the network



Connectivity (Degree)



Path length

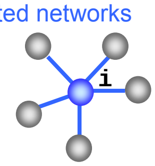


Clustering coefficient

[Barabasi]

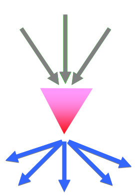
Connectivity Measures

- Node degree: the number of edges incident on the node (number of network neighbors.)
  - Undetected networks



Degree of node i = 5

    - Degree distribution *P(k)*: probability that a node has degree *k*.
  - Directed networks, i.e., transcription regulation networks (TRNs)

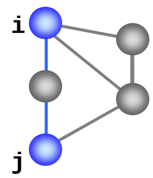


Incoming degree = 2.1
→each gene is regulated by ~2 TFs

Outgoing degree = 49.8
→each TF targets ~50 genes

Characteristic path length

- L_{ij} is the number of edges in the shortest path between vertices *i* and *j*
  - The characteristic path length of a graph is the average of the L_{ij} for every possible pair (i,j)
  - Diameter: maximal distance in the network.
    - Networks with small values of L are said to have the "small world property"

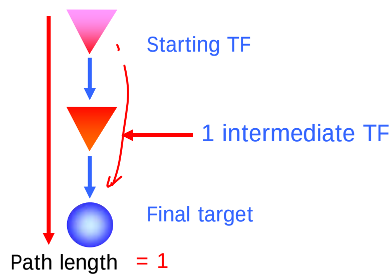


$L_{(i,j)} = 2$

- In a TRN, L'_{ij} represents the number of intermediate TFs until final target

Indicate how immediate
a regulatory response is

Average path length = 4.7



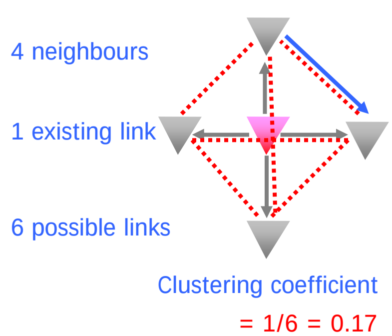
Clustering coefficient

- The clustering coefficient of node *i* is the ratio of the number E_i of edges that exist among its neighbors, over the number of edges that could exist:

$C_I=2T_I/n_I(n_I-1)$

Measure how inter-connected
the network is

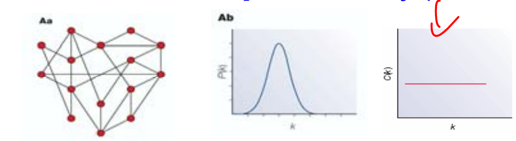
Average coefficient = 0.11



- The clustering coefficient for the entire network *C* is the average of all the C_i

A Comparison of Global Network Statistics (Barabasi & Oltvai, 2004)

A. Random Networks [Erdos and Rényi (1959, 1960)]



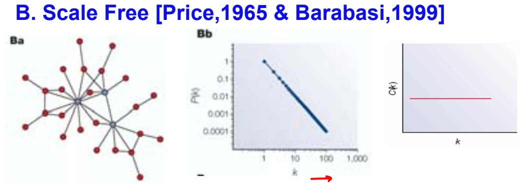
$$P(k) = \frac{e^{-\bar{k}}\bar{k}^{k}}{k!}$$

Mean path length ~ ln(k)

Phase transition:

Connected if: $p \geq \ln(k)/k$

B. Scale Free [Price,1965 & Barabasi,1999]

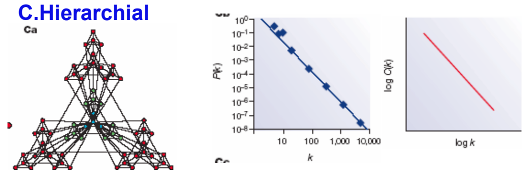


$$P(k) \sim k^{-\gamma},\ k >> 1,\ \ 2 < \gamma$$

Mean path length ~ lnln(k)

Preferential attachment. Add proportionally to connectedness

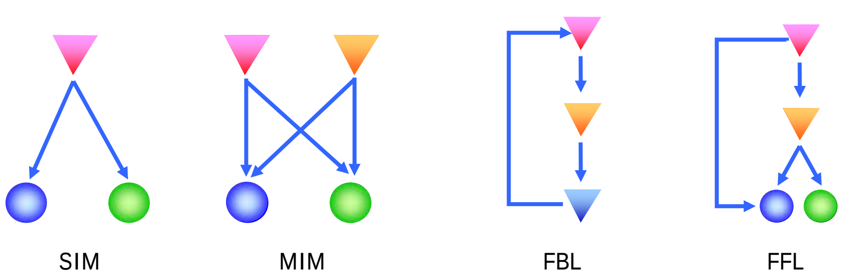
C.Hierarchial



Copy smaller graphs and let them keep their connections.

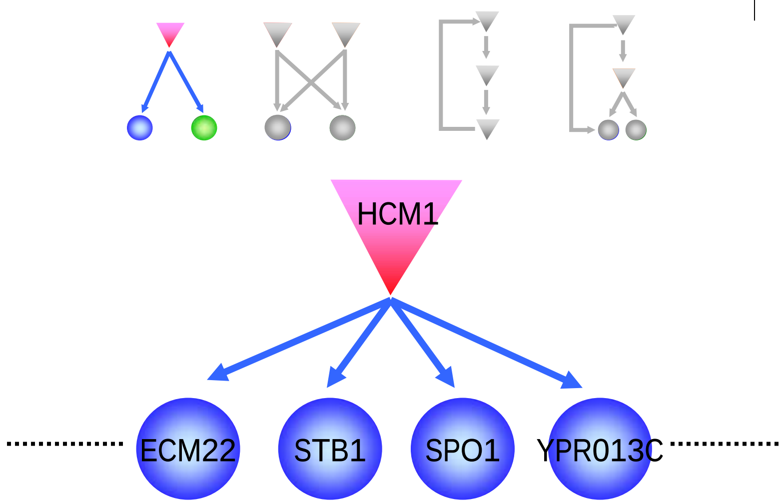
Local network motifs

- Regulatory modules within the network



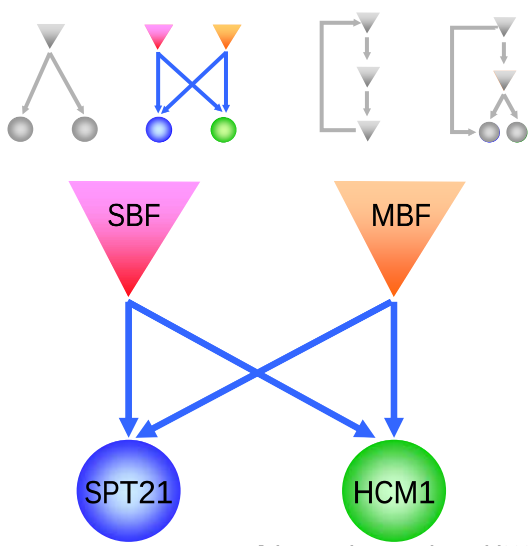
[Alon]

SIM = Single input motifs



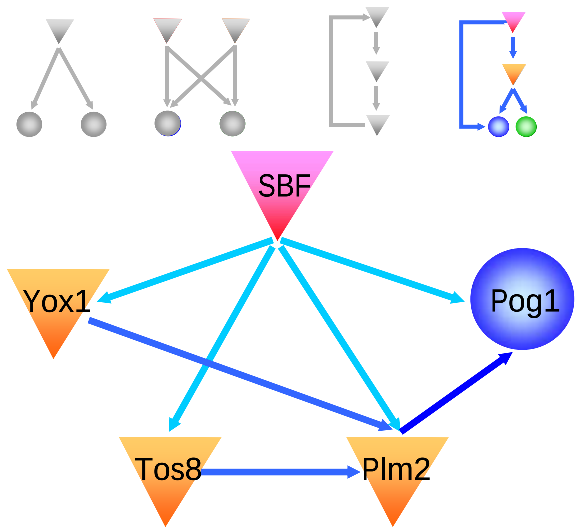
[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

MIM = Multiple input motifs



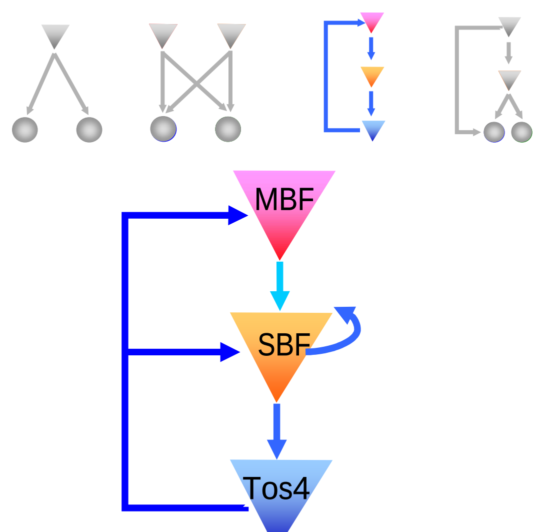
[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

FFL = Feed-forward loops



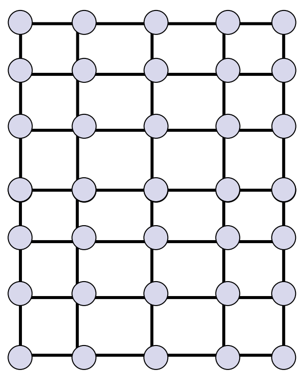
[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

FBL = Feed-back loops

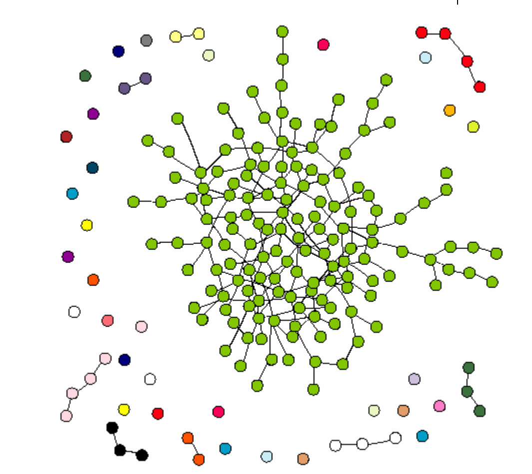


[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

What network structure should be used to model a biological network?



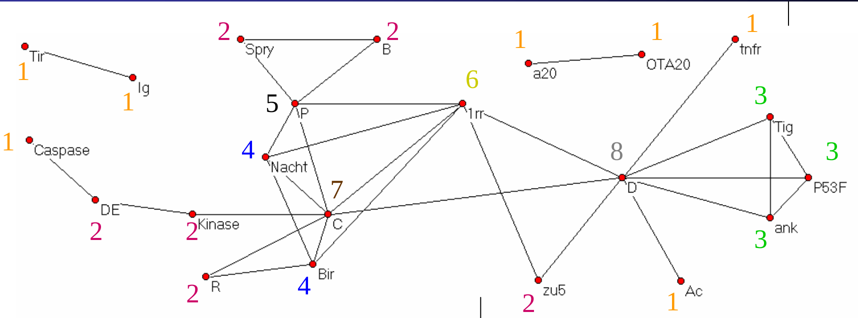
lattice



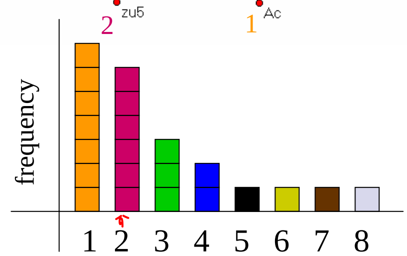
random

Strogatz S.H., *Nature* (2001) **410** 268

Calculating the degree connectivity of a network

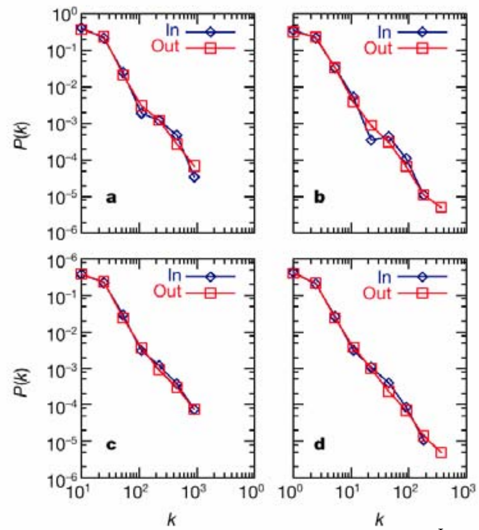


Degree connectivity distributions:



Connectivity distributions for metabolic networks

A. fulgidus
(archaea)



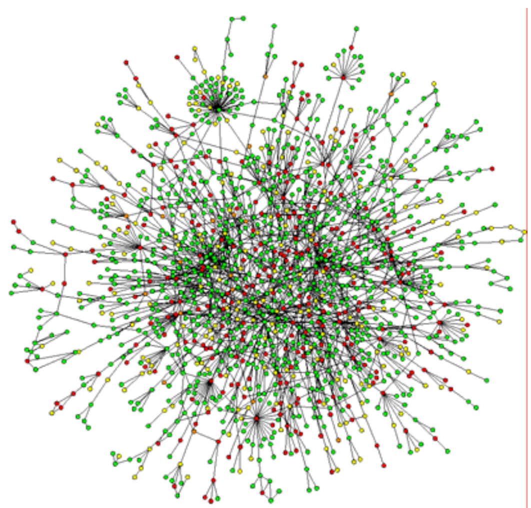
E. coli
(bacterium)

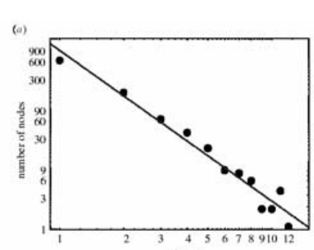
C. elegans
(eukaryote)

averaged
over 43
organisms

Jeong et al. Nature (2000) **407** 651-654

Protein-protein interaction networks



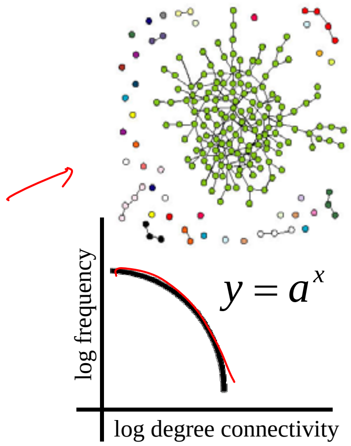


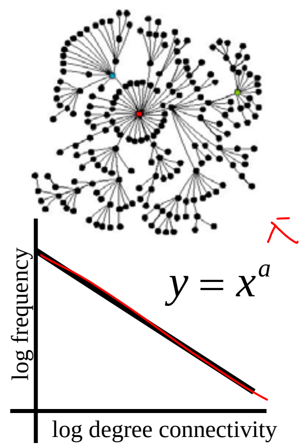
(color of nodes is explained later)

Jeong *et al. Nature* **411**, 41 - 42 (2001)
Wagner. RSL (2003) 270 457-466

Random versus scaled exponential degree distribution

- Degree connectivity distributions differs between random and observed (metabolic and protein-protein interaction) networks.

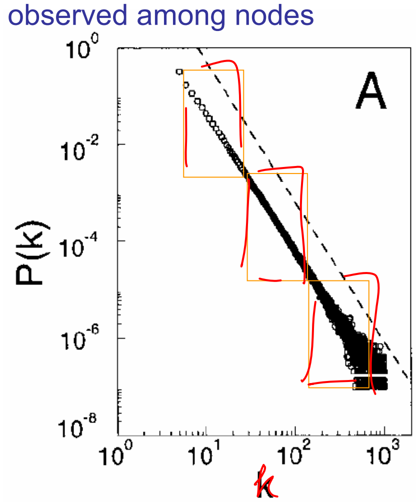




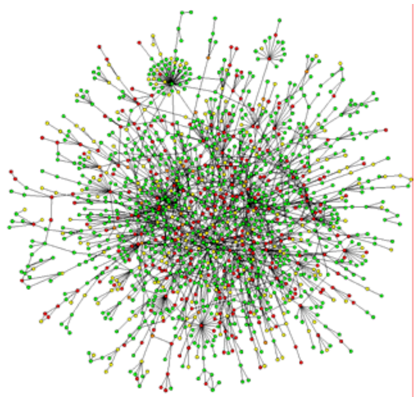
Strogatz S.H., *Nature* (2001) **410** 268

What is so "scale-free" about these networks?

- No matter which scale is chosen the same distribution of degrees is observed among nodes



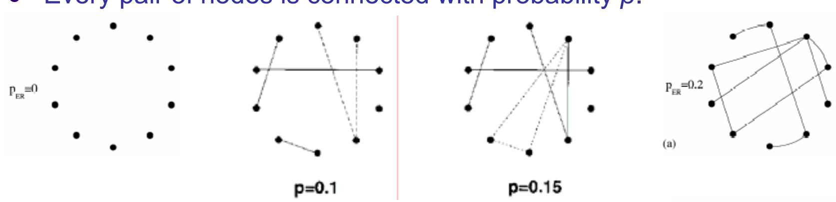
Models for networks of complex topology



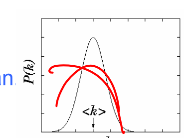
- Erdos-Renyi (1960)
- Watts-Strogatz (1998)
- Barabasi-Albert (1999)

Random Networks:
The Erdős-Rényi [ER] model (1960):

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



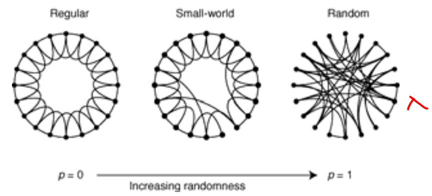
  - Mean degree: (*N*-1)*p*.
  - Degree distribution is binomial, concentrated around the mean
  - Average distance (*Np*>1): log *N*



- Important result: many properties in these graphs appear quite suddenly, at a threshold value of PER(N)
  - If PER~c/N with c<1, then almost all vertices belong to isolated trees
  - Cycles of all orders appear at PER ~ 1/N

The Watts-Strogatz [WS] model (1998)

- Start with a regular network with *N* vertices
- Rewire each edge with probability *p*



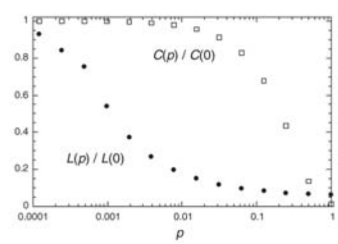
For p=0 (Regular Networks):
- high clustering coefficient
- high characteristic path length

For p=1 (Random Networks):
- low clustering coefficient
- low characteristic path length

- QUESTION: What happens for intermediate values of *p*?

WS model, cont.

- There is a broad interval of p for which L is small but C remains large



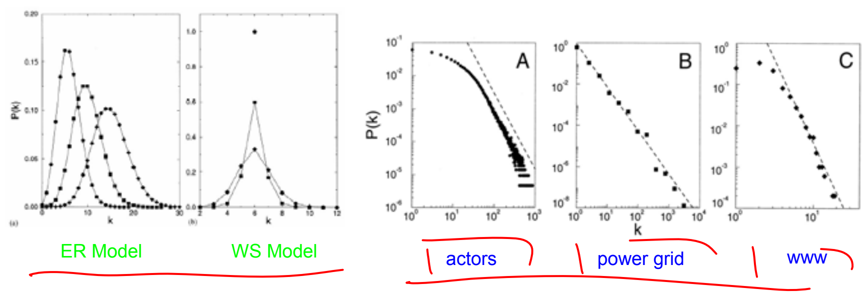
- Small world networks are common :

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Scale-free networks:
The Barabási-Albert [BA] model (1999)

- The distribution of degrees:

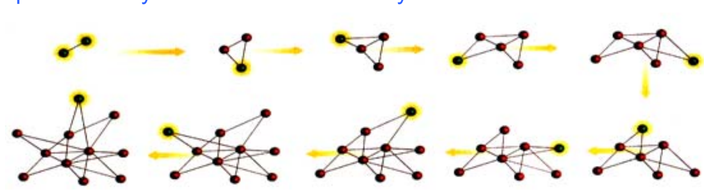


- In real network, the probability of finding a highly connected node decreases exponentially with *k*

$$P(K) \sim K^{-\gamma}$$

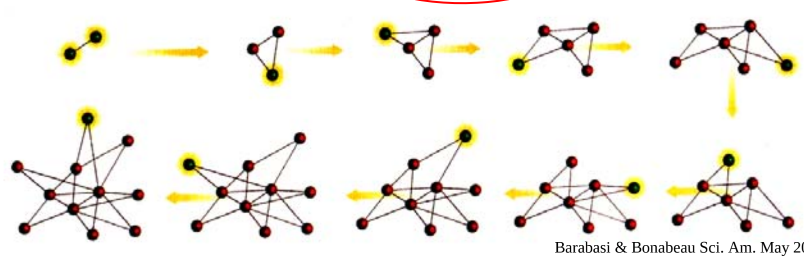
BA model, cont.

- Two problems with the previous models:
  1. N does not vary
  2. the probability that two vertices are connected is uniform
- The BA model:
  - Evolution: networks expand continuously by the addition of new vertices, and
  - Preferential-attachment (rich get richer): new vertices attach preferentially to sites that are already well connected.



Scale-free network model

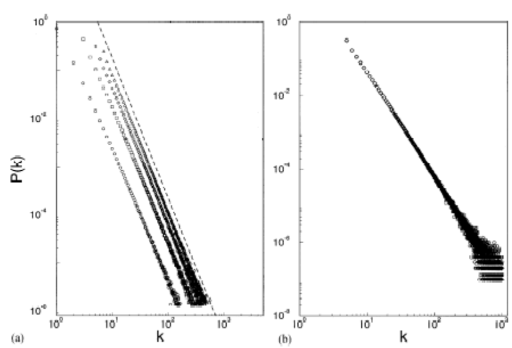
- GROWTH: starting with a small number of vertices m_0 at every timestep add a new vertex with $m \leq m_0$
- PREFERENTIAL ATTACHMENT: the probability Π that a new vertex will be connected to vertex *i* depends on the connectivity of that vertex: $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$



Barabasi & Bonabeau Sci. Am. May 2003 60-69

Barabasi and Albert. Science (1999) **286** 509-512

Scale Free Networks



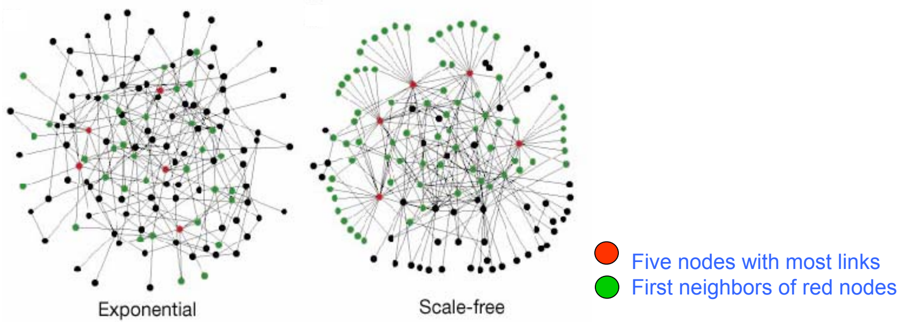
a) Connectivity distribution with N = m_0+t=300000 and m_0=m=1(circles), m_0=m=3 (squares), and m_0=m=5 (diamons) and m_0=m=7 (triangles)

b) P(k) for m_0=m=5 and system size N=100000 (circles), N=150000 (squares) and N=200000 (diamonds)

Barabasi and Albert. Science (1999) **286** 509-512

Comparing Random Vs. Scale-free Networks

- Two networks both with 130 nodes and 215 links)



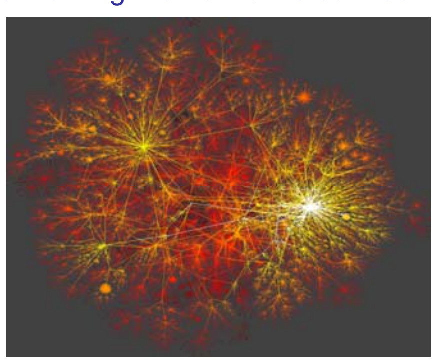
- The importance of the connected nodes in the scale-free network:
  - 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are reached.

Modified from Albert et al. Science (2000) **406** 378-382

Failure and Attack

Albert et al. Science (2000) **406** 378-382

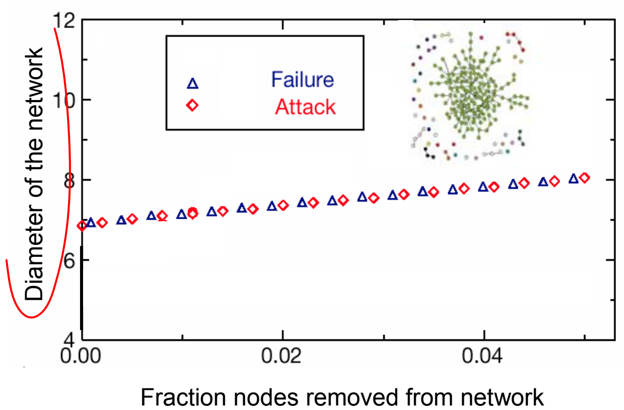
- Failure: Removal of a random node.
- Attack: The selection and removal of a few nodes that play a vital role in maintaining the network's connectivity.



a macroscopic snapshot of Internet connectivity by **K. C. Claffy**

Failure and Attack, cont.

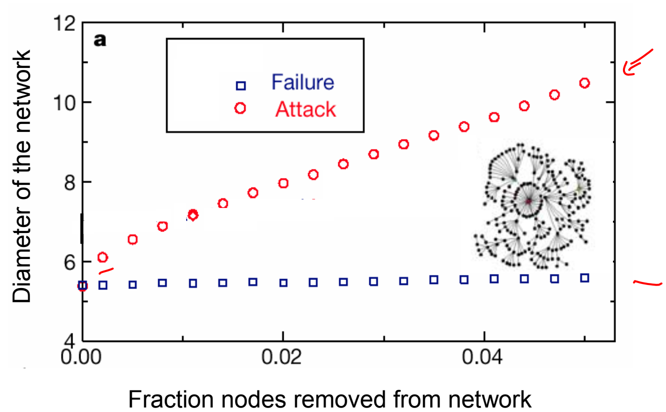
- Random networks are homogeneous so there is no difference between failure and attack



Modified from Albert et al. Science (2000) **406** 378-382

Failure and Attack, cont.

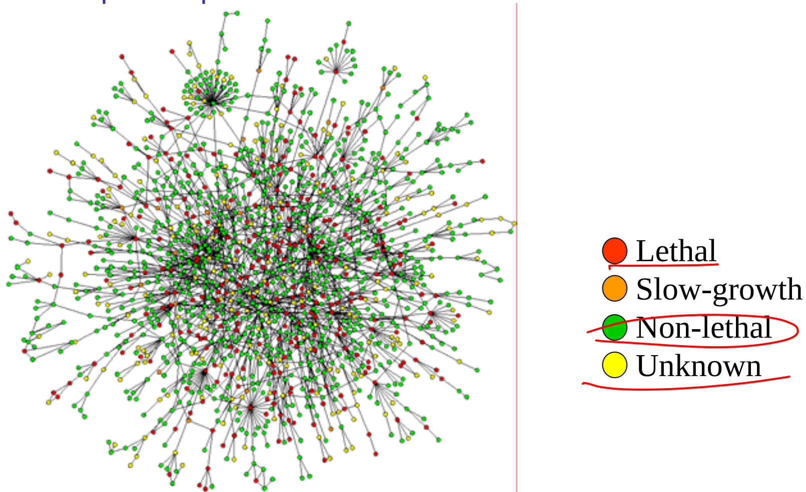
- Scale-free networks are robust to failure but susceptible to attack



Modified from Albert et al. Science (2000) **406** 378-382

The phenotypic effect of removing the corresponding protein:

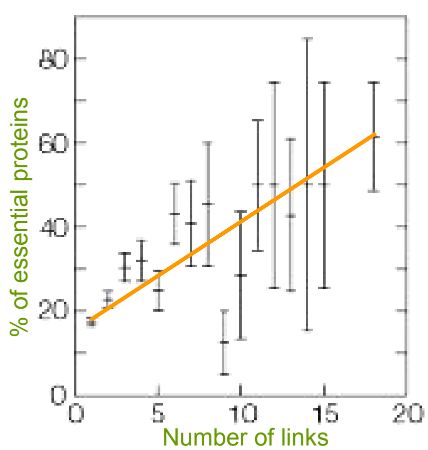
- Yeast protein-protein interaction networks



Jeong *et al. Nature* **411**, 41 - 42 (2001)

Lethality and connectivity are positively correlated

- Average and standard deviation for the various clusters.

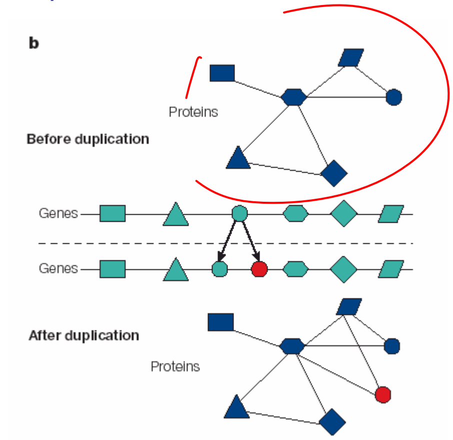


- Pearson's linear correlation coefficient = 0.75

Jeong *et al. Nature* **411**, 41 - 42 (2001)

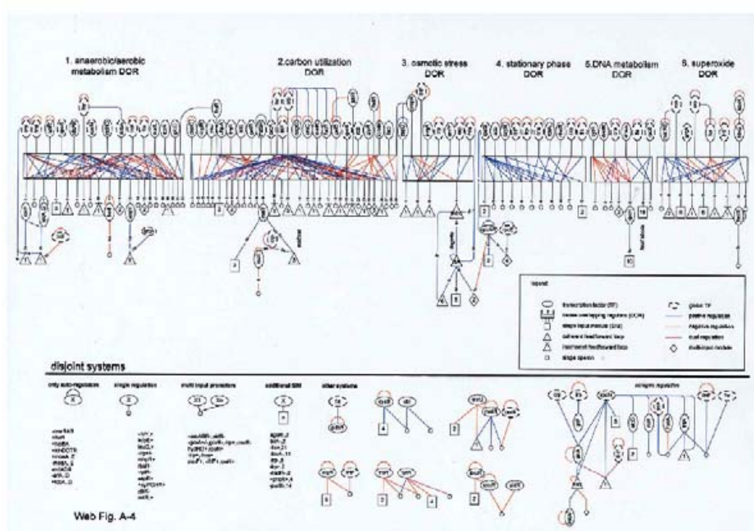
Genetic foundation of network evolution

- Network expansion by gene duplication
  - A gene duplicates
  - Inherits it connections
  - The connections can change
- Gene duplication slow ~10^{-9}/year
- Connection evolution fast ~10^{-6}/year



Barabasi & Oltvai. NRG. (2004) 5 101-113

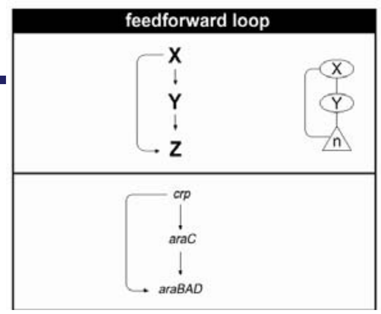
The transcriptional regulation network of Escherichia coli.

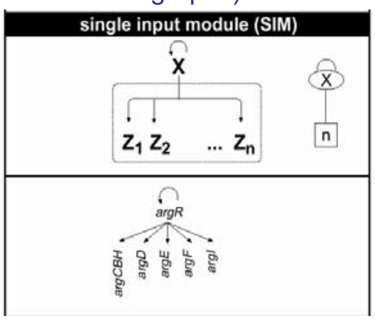


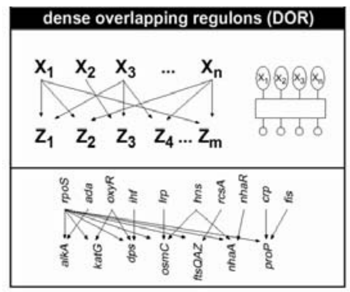
Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan & Uri Alon (2002) Nature Genetics **31** 64 - 68

Motifs in the networks

- Deployed a motif detection algorithm on the transcriptional regulation network.
- Identified three recurring motifs (significant with respect to random graphs).



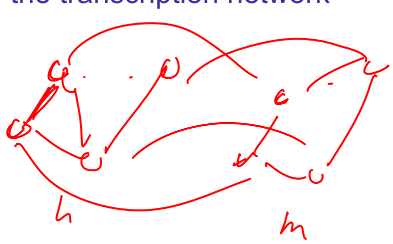


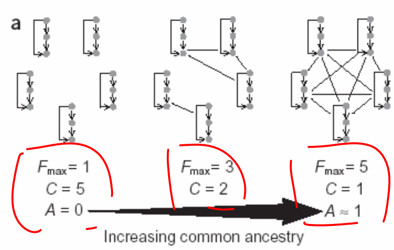


Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan & Uri Alon (2002) Nature Genetics **31** 64 - 68

Convergent evolution of gene circuits

- Are the components of the feed-forward loop for example homologous?
- Circuit duplication is rare in the transcription network





b

	Circuit type	Number of circuits	Number of families (C)	Index of common ancestry (A)	Largest circuit family (F_{max})
Yeast	Feed-forward	48	44 (46.8 ± 1.9; P = 0.08)	0.082 (0.023 ± 0.035; P = 0.08)	5 (1.9 ± 1.4; P = 0.05)
	Bi-fan	542	435 (469.0 ± 37.7; P = 0.18)	0.197 (0.135 ± 0.070; P = 0.18)	49 (41.0 ± 31.1; P = 0.33)
	MIM-2	176	168 (164.5 ± 8.8; P = 0.60)	0.045 (0.065 ± 0.050; P =0.60)	5 (7.4 ± 6.2; P = 0.59)
	Reg. chain (3)	33	33	0	1
E. coli	Feed-forward	11	11	0	1
	Bi-fan	27	27	0	1

Conant and Wagner. Nature Genetics (2003) 34 264-266

Acknowledgements

- Itai Yanai and Doron Lancet
- Mark Gerstein
- Roded Sharan
- Jotun Hein
- Serafim Batzoglou

for some of the slides modified from their lectures or tutorials

Reference

- **Barabási and Albert.** ***Emergence of scaling in random networks.*** Science **286**, 509-512 (1999).
- **Yook et al.** ***Functional and topological characterization of protein interaction networks***. Proteomics **4**, 928-942 (2004).
- **Jeong et al.** ***The large-scale organization of metabolic networks***. Nature **407**, 651-654 (2000).
- **Albert et al.** ***Error and attack tolerance in complex networks.*** Nature **406** , 378 (2000).
- **Barabási and Oltvai, Network Biology: Understanding the Cell's Functional Organization, Nature Reviews, vol 5, 2004**