Complex (Biological) Networks

Today: *Measuring Network Topology*

Thursday: *Analyzing Metabolic Networks*

Elhanan Borenstein

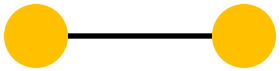
Some slides are based on slides from courses given by Roded Sharan and Tomer Shlomi

Measuring Network Topology

- Introduction to network theory
- Global Measures of Network Topology
  - Degree Distribution
  - Clustering Coefficient
  - Average Distance
- Network Motifs
- Random Network Models

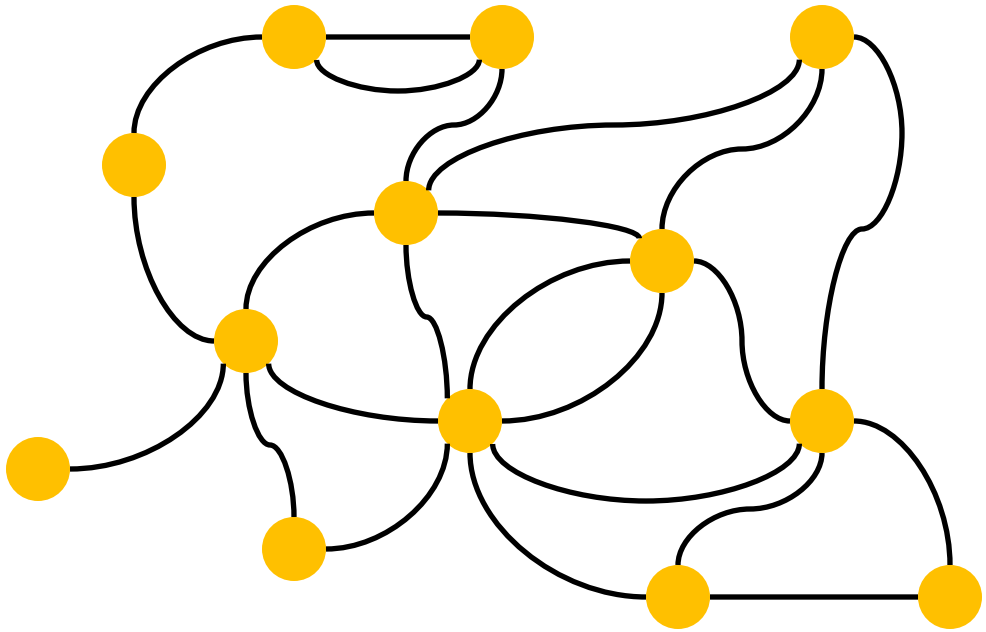
What is a Network?

- A collection of **nodes** and **links** (**edges**)
- A map of interactions or relationships



What is a Network?

- A collection of **nodes** and **links** (**edges**)
- A map of interactions or relationships



Networks vs. Graphs

- **Graph Theory**
- Definition of a graph: ***G=(V,E)***
  - V is the set of nodes/vertices (elements)
  - |V|=N
  - E is the set of edges (relations)

- One of the most well studied objects in CS
  - Subgraph finding (e.g., **clique**, **spanning tree**) and alignment
  - Graph **coloring** and graph **covering**
  - Route finding (**Hamiltonian** path, traveling salesman, etc.)
- Many problems are proven to be **NP-complete**

Networks vs. Graphs

Network theory	Graph theory
Social sciences Biological sciences	Computer science
Mostly 20th century	Since 18th century!!!
Modeling real-life systems	Modeling abstract systems
Measuring structure & topology	Solving “graph-related” questions

Why Networks?

Networks as models

Simple, visual representation of complex systems

Focus on organization (rather than on components)

Discovery (topology affects function)

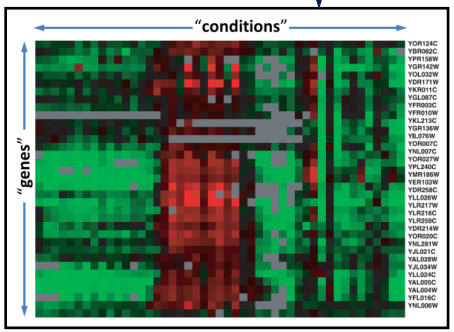
Predictive models

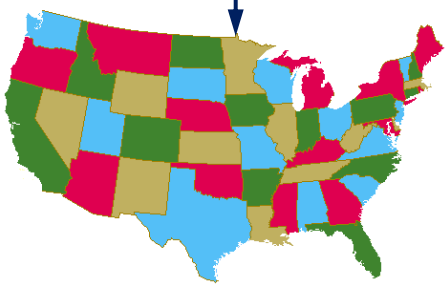
Diffusion models (dynamics)

Networks as tools

Problem representation (more common than you think)

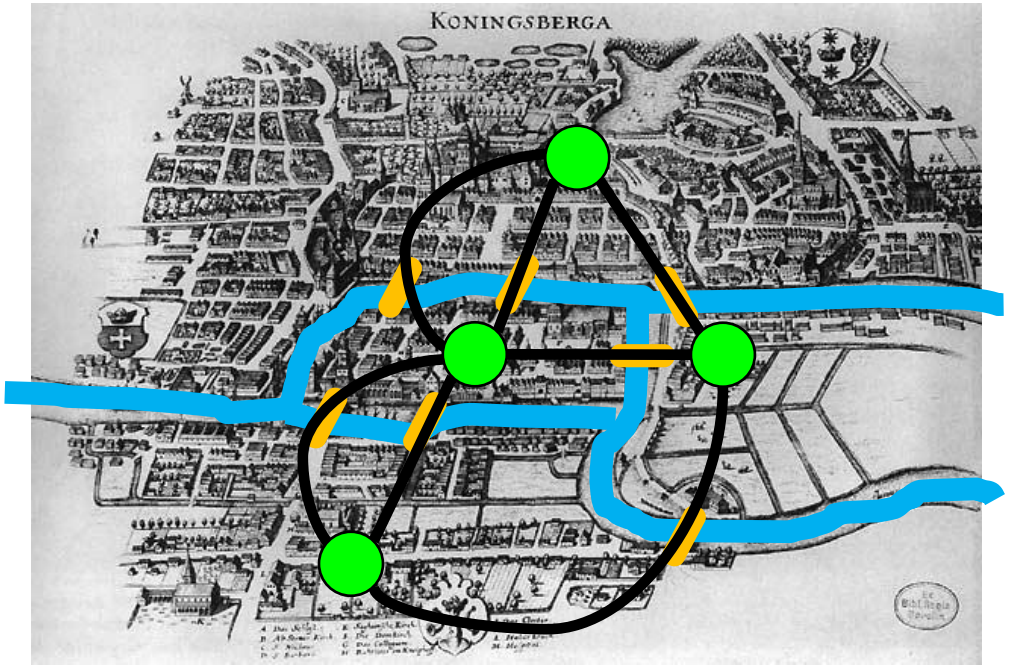
Algorithm development





The Seven Bridges of Königsberg

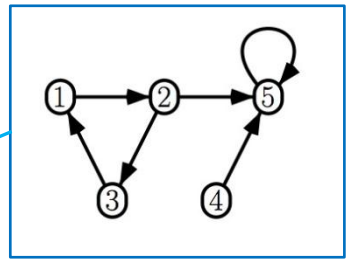
- Published by **Leonhard Euler, 1736**
- Considered the first paper in graph theory

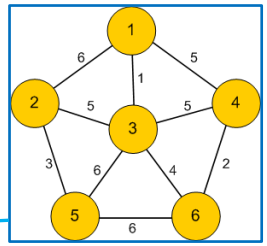


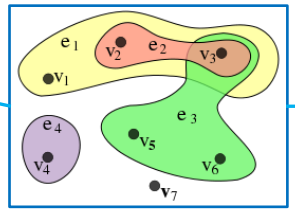
Leonhard Euler
1707 –1783

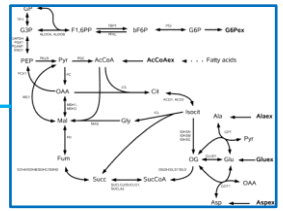
Types of Graphs/Networks

- **Edges:**
  - Directed/undirected
  - Weighted/non-weighted
  - Simple-edges/Hyperedges

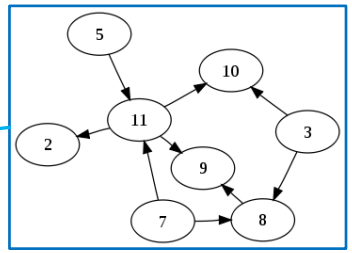


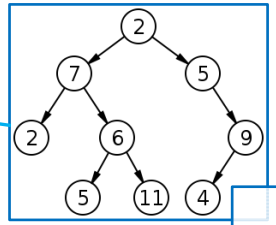


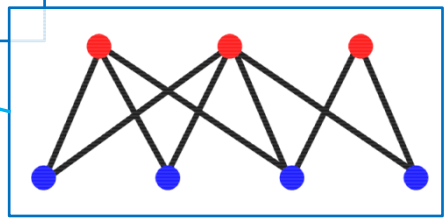




- **Special topologies:**
  - Directed Acyclic Graphs (DAG)
  - Trees
  - Bipartite networks

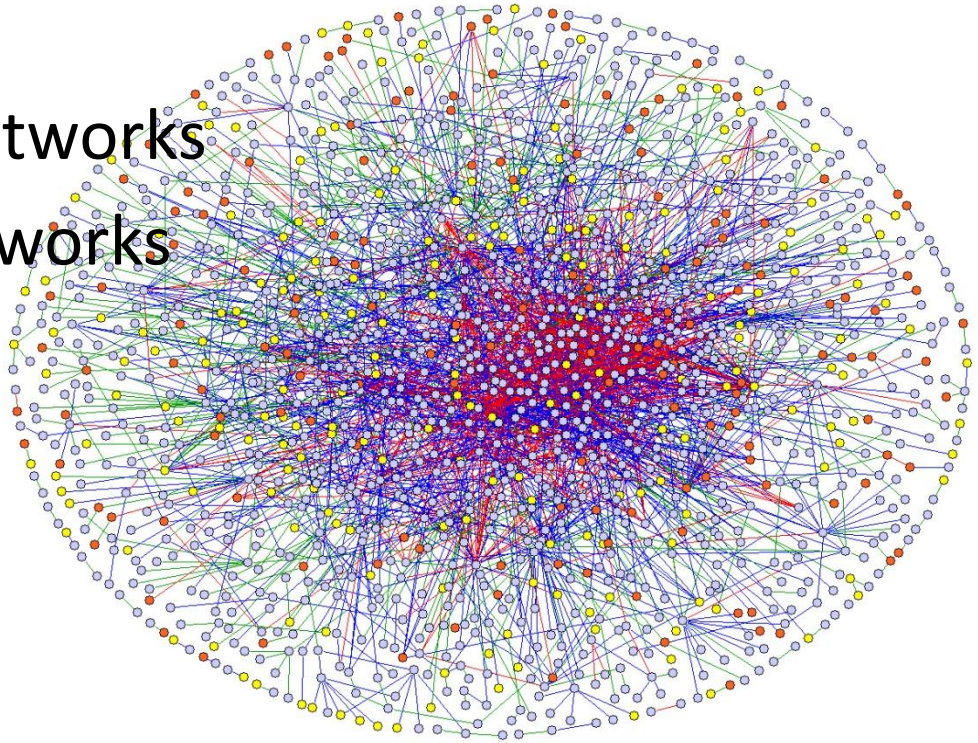






Networks in Biology

- Molecular networks:
  - Protein-Protein Interaction (PPI) networks
  - Metabolic Networks
  - Regulatory Networks
  - Synthetic lethality Networks
  - Gene Interaction Networks
  - Many more …



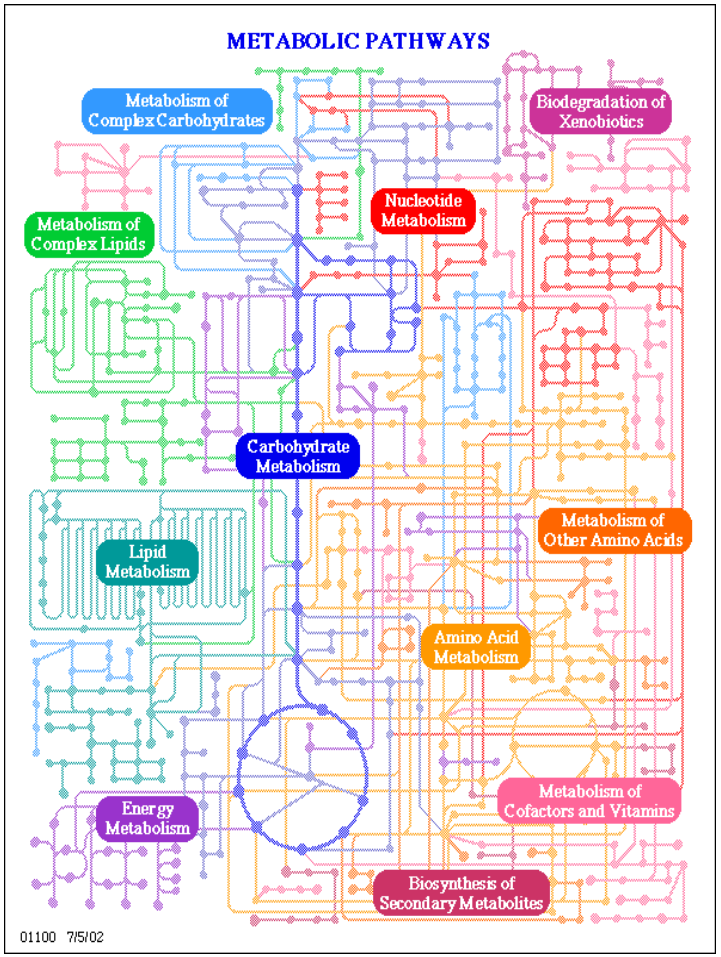
Metabolic Networks

- Reflect the set of biochemical reactions in a cell
  - **Nodes**: metbolites
  - **Edges**: biochemical reactions
  - **Additional representations!**
- Derived through:
  - Knowledge of biochemistry
  - Metabolic flux measurements

S. Cerevisiae

1062 metabolites

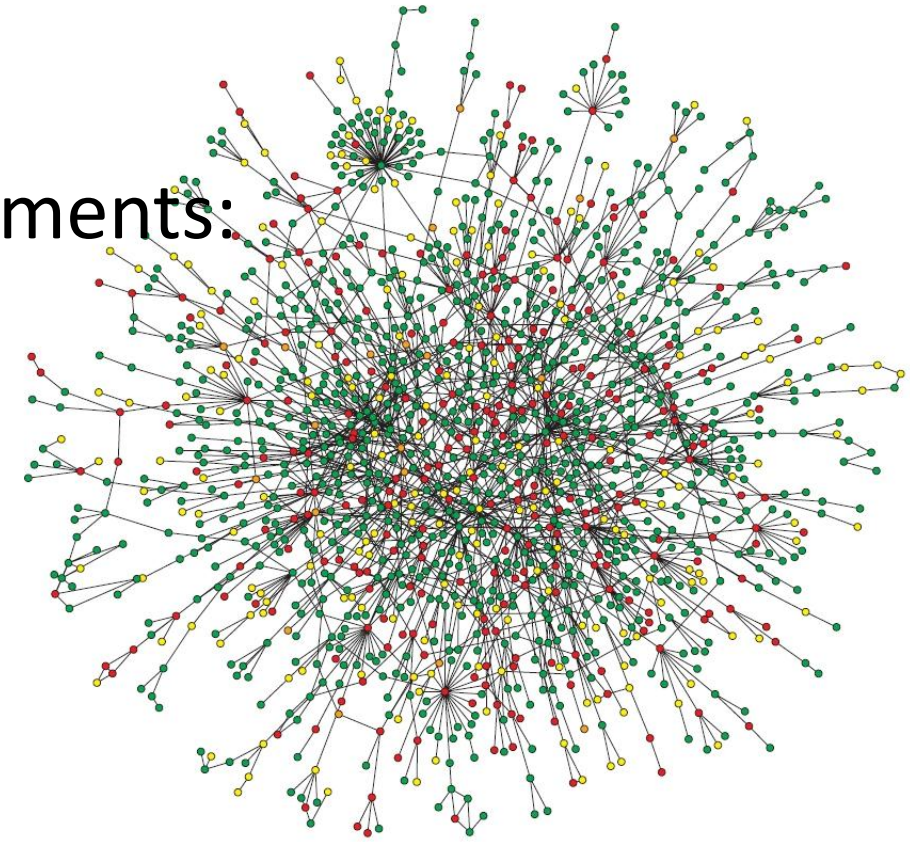
1149 reactions



Protein-Protein Interaction (PPI) Networks

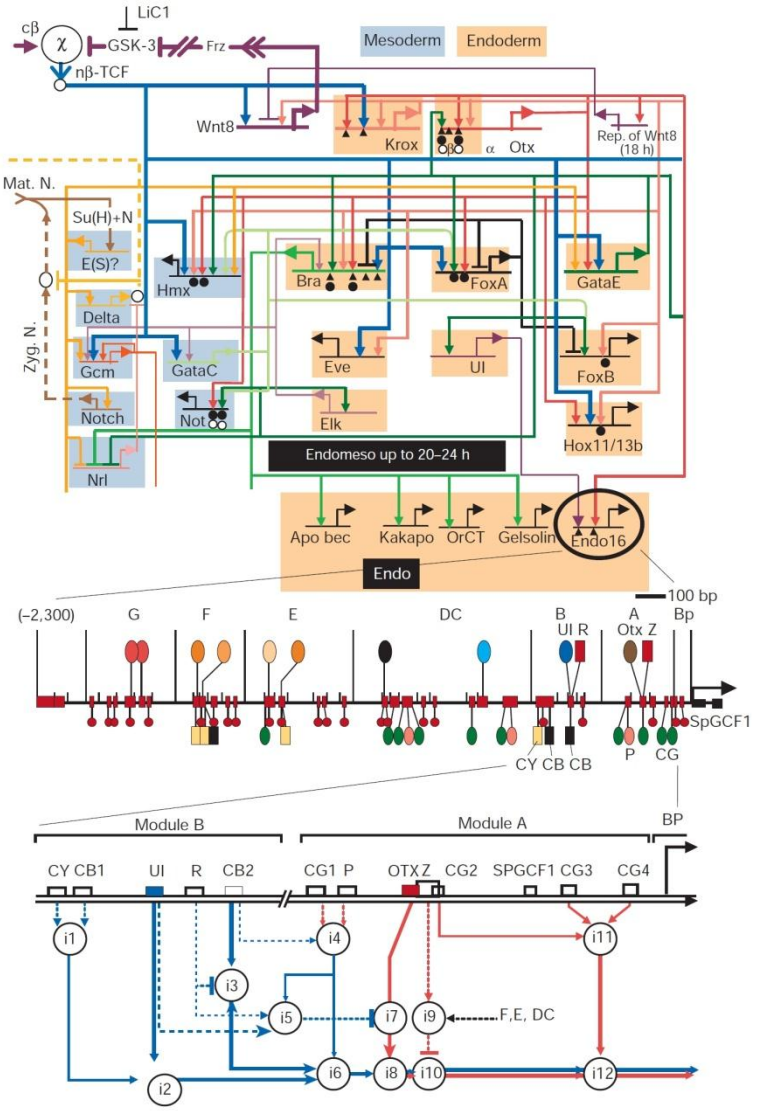
- Reflect the cell's molecular interactions and signaling pathways (**interactome**)
  - **Nodes**: proteins
  - **Edges**: interactions(?)
- High-throughput experiments:
  - Protein Complex-IP (Co-IP)
  - Yeast two-hybrid
  - Computationally

S. Cerevisiae
4389 proteins
14319 interactions

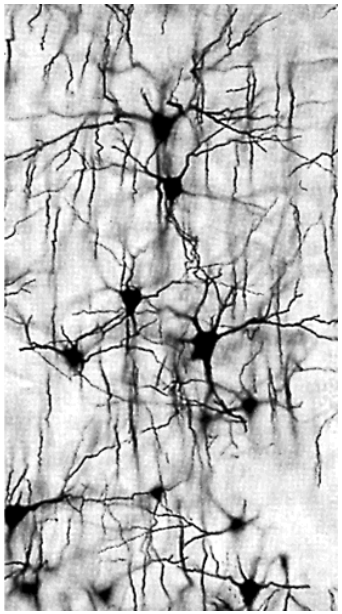


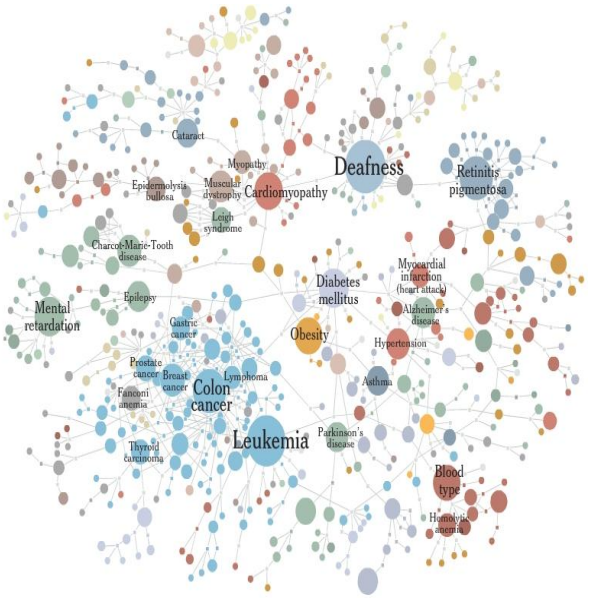
Transcriptional Regulatory Network

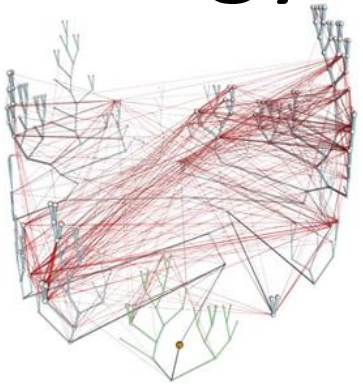
- Reflect the cell's genetic regulatory circuitry
  - **Nodes**: transcription factors (TFs) and genes;
  - **Edges**: from TF to the genes it regulates; Directed; weighted?; "almost" bipartite
- Derived through:
  - Chromatin IP
  - Microarrays
  - **Computationally**

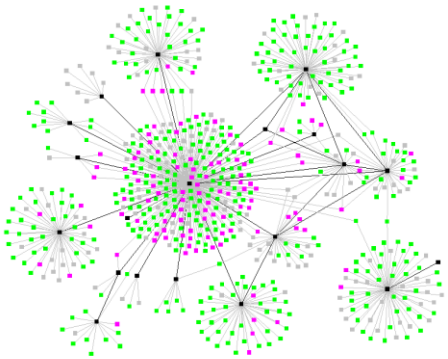


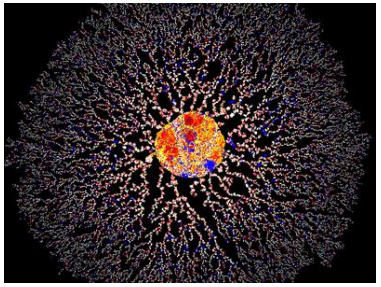
Other Networks in Biology/Medicine

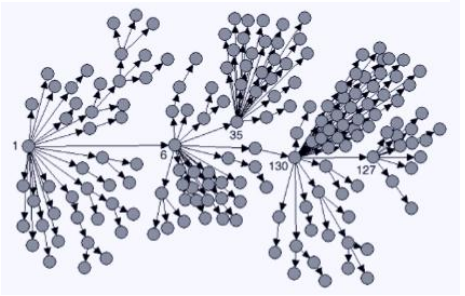


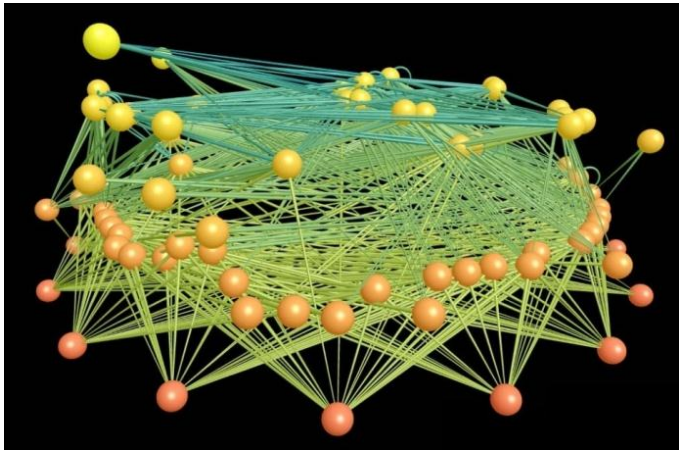


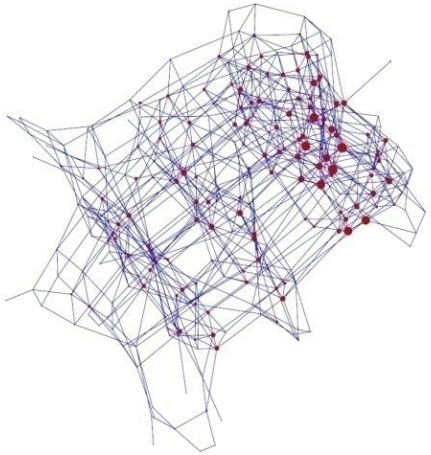






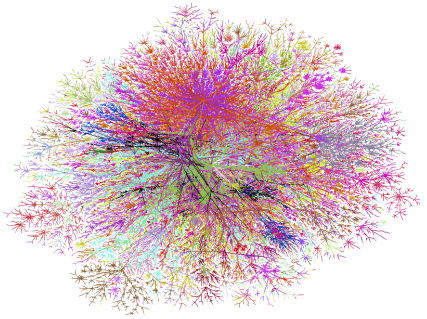


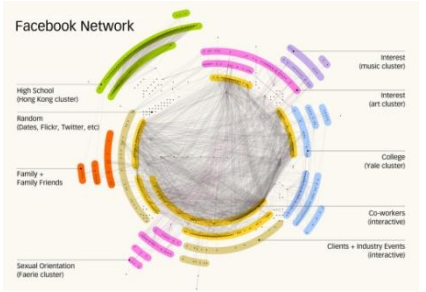




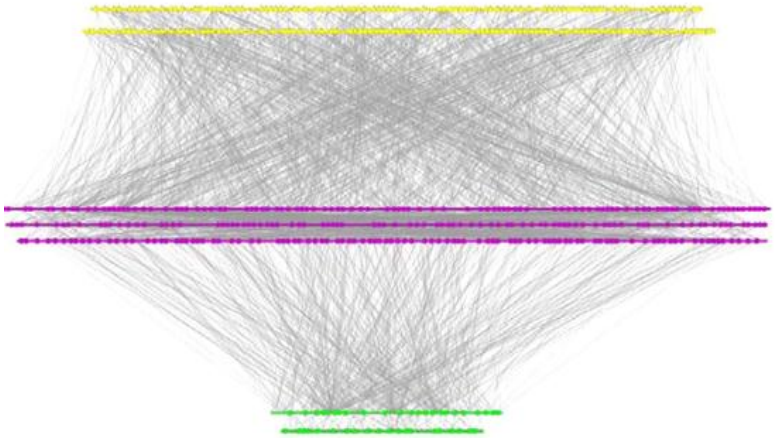
Non-Biological Networks

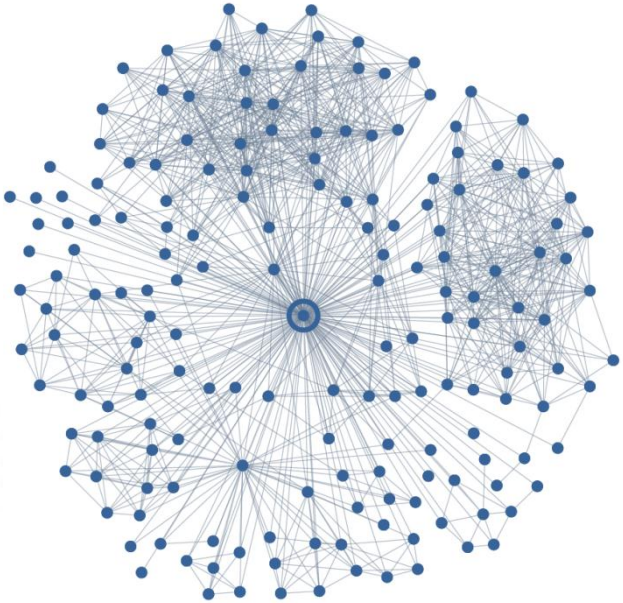
- Computer related networks:
  - WWW; Internet backbone
  - Communication and IP
- Social networks:
  - Friendship (facebook; clubs)
  - Citations / information flow
  - Co-authorships (papers); Co-occurrence (movies; Jazz)
- Transportation:
  - Highway system; Airline routes
- Electronic/Logic circuits
- Many more…

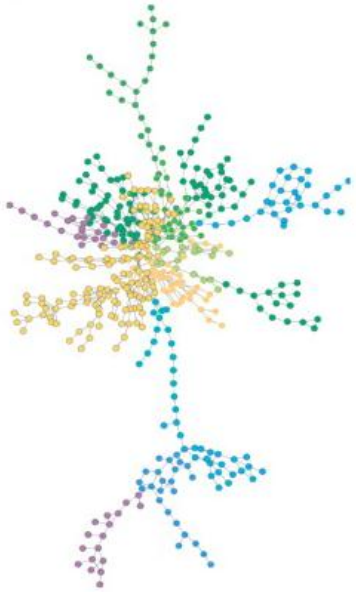


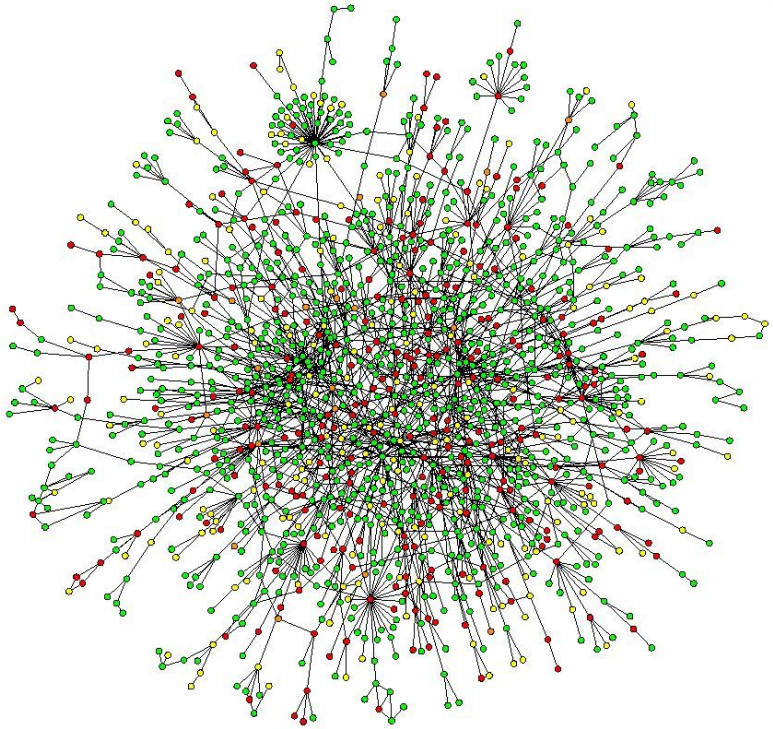


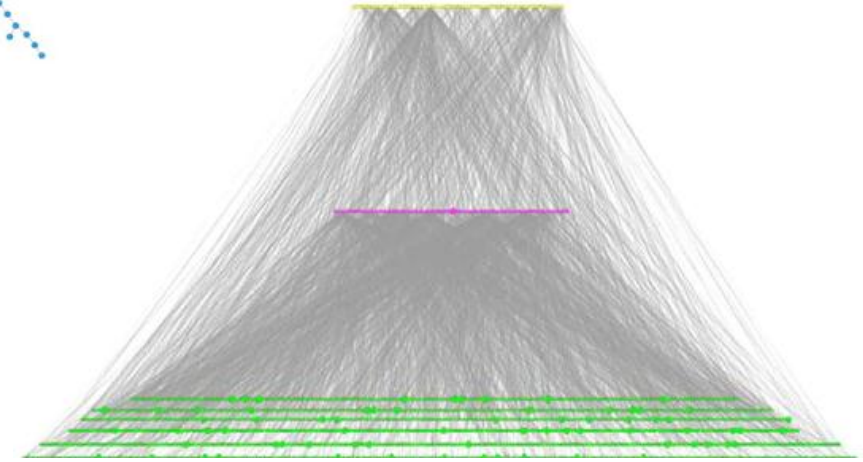
Global Measures
of
Network Topology











Comparing networks

- We want to find a way to “compare” networks.
  - “Similar” (not identical) **topology**
  - Common design principles

- We seek measures of network topology that are:
  - Simple
  - Capture **global** organization
  - Potentially “important”

Summary statistics

(equivalent to, for example, GC content for genomes)

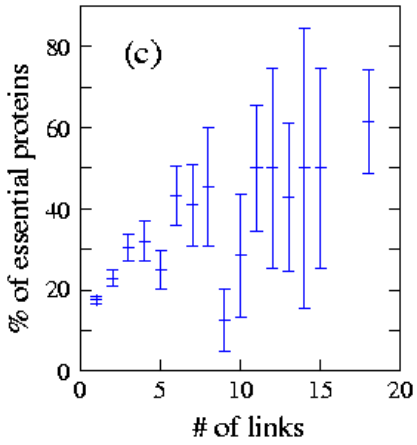
Node Degree / Rank

- Degree = Number of neighbors
- Node degree in PPI networks correlates with:
  - Gene essentiality
  - Conservation rate
  - Likelihood to cause human disease

brief communications

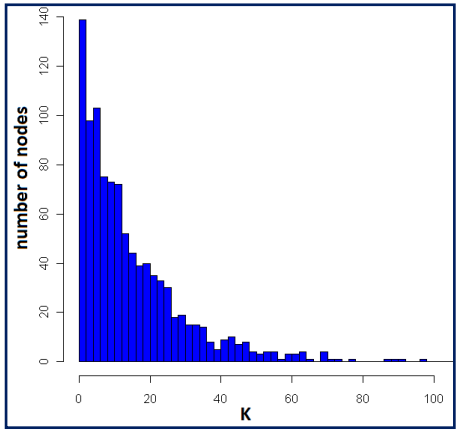
Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.



Degree Distribution

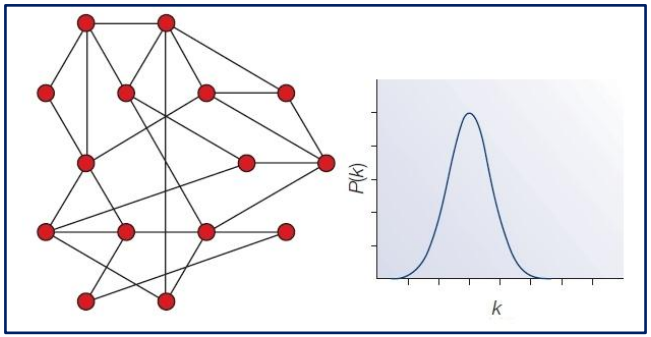
- Degree distribution P(k):
  probability that a node has a degree of exactly k



- Common distributions:

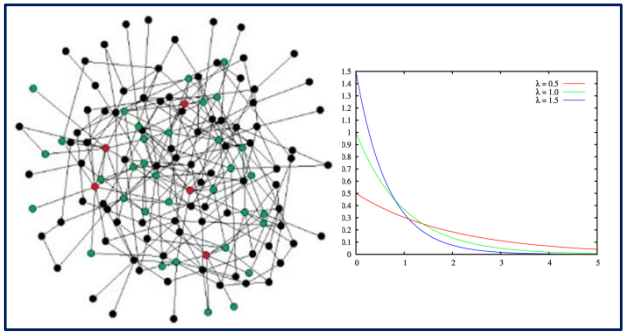
Poisson:

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



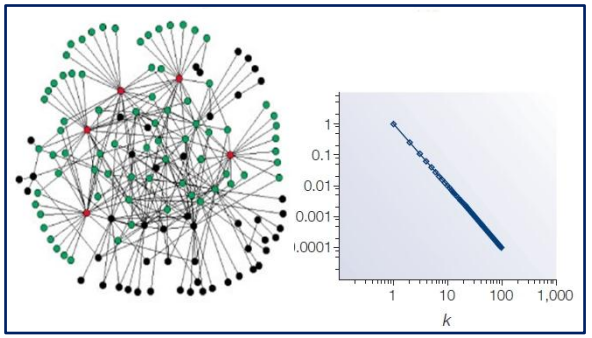
Exponential:

$$P(k) \propto e^{-k/d}$$



Power-law:

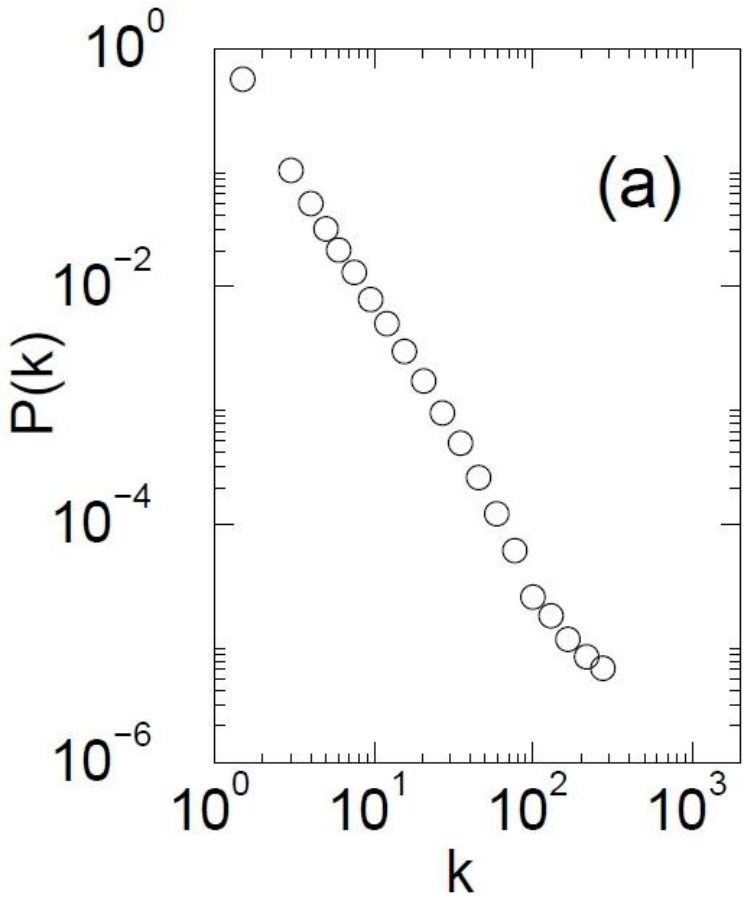
$$P(k) \propto k^{-c}, k \neq 0, c > 1$$



The Internet

- **Nodes** – 150,000 routers
- **Edges** – physical links

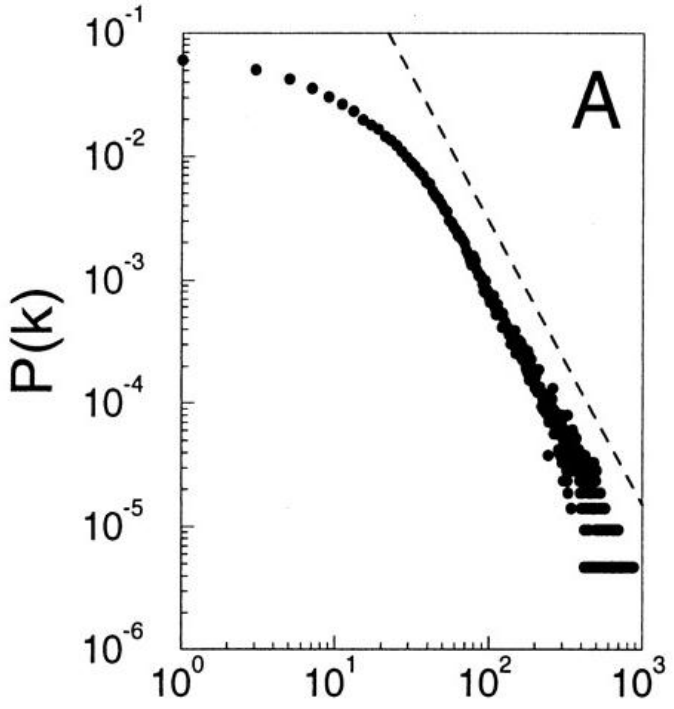
- $P(k) \sim k^{-2.3}$



Govindan and Tangmunarunkit, 2000

Movie Actor Collaboration Network

- **Nodes** – 212,250 actors
- **Edges** – co-appearance in a movie
- (<k> = 28.78)
- $P(k) \sim k^{-2.3}$

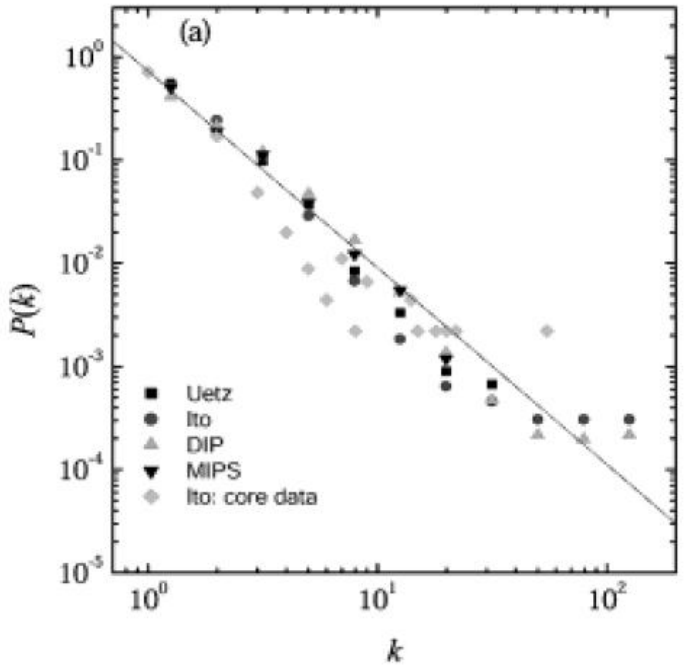


Barabasi and Albert, Science, 1999

Protein Interaction Networks

- **Nodes** – Proteins
- **Edges** – Interactions (yeast)

- $P(k) \sim k^{-2.5}$

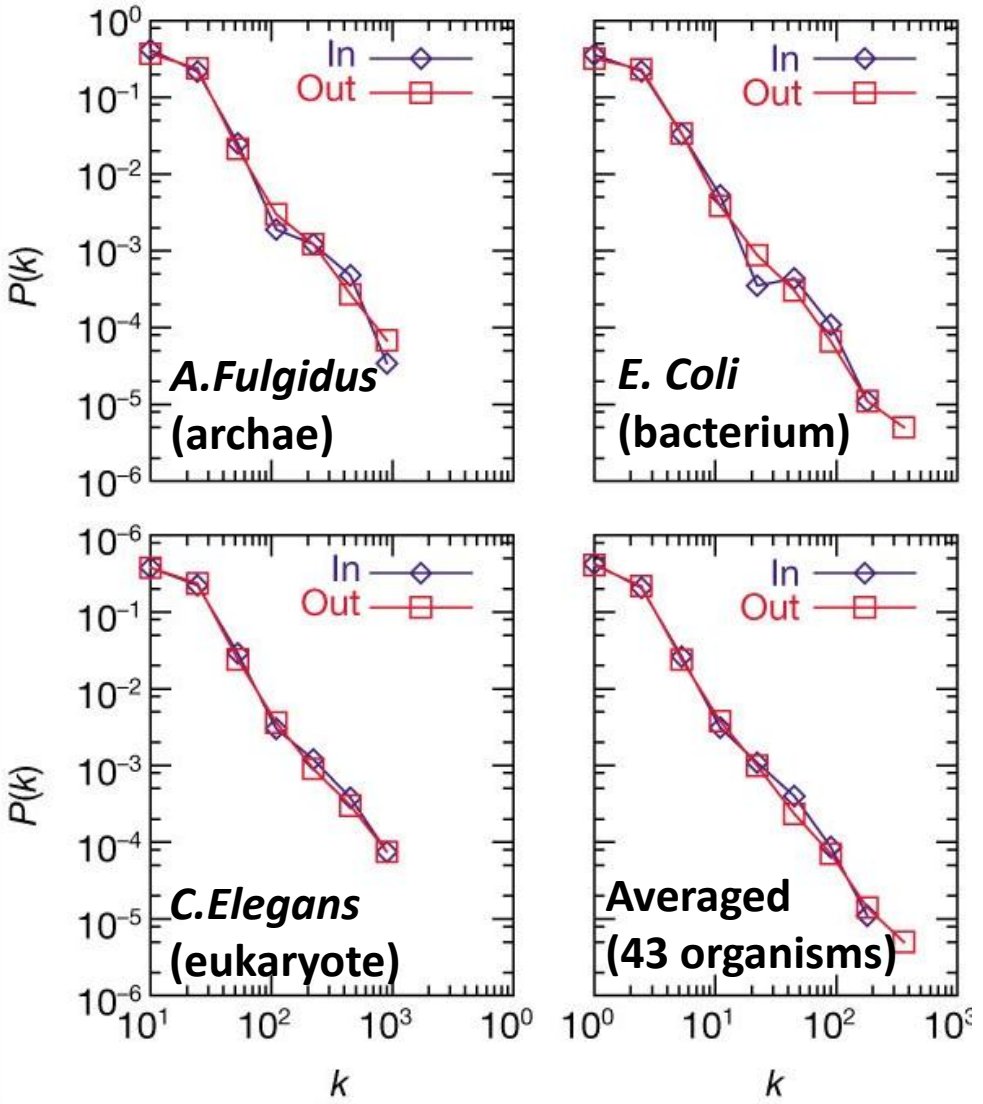


Yook et al, Proteomics, 2004

Metabolic Networks

- **Nodes** – Metabolites
- **Edges** – Reactions
- $P(k) \sim k^{-2.2\pm2}$

Metabolic networks ***across all kingdoms of life*** *are scale-free*

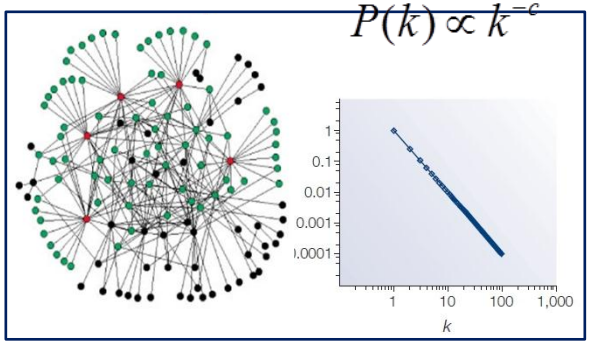


Jeong et al., Nature, 2000

$$P(k) \propto k^{-c}$$

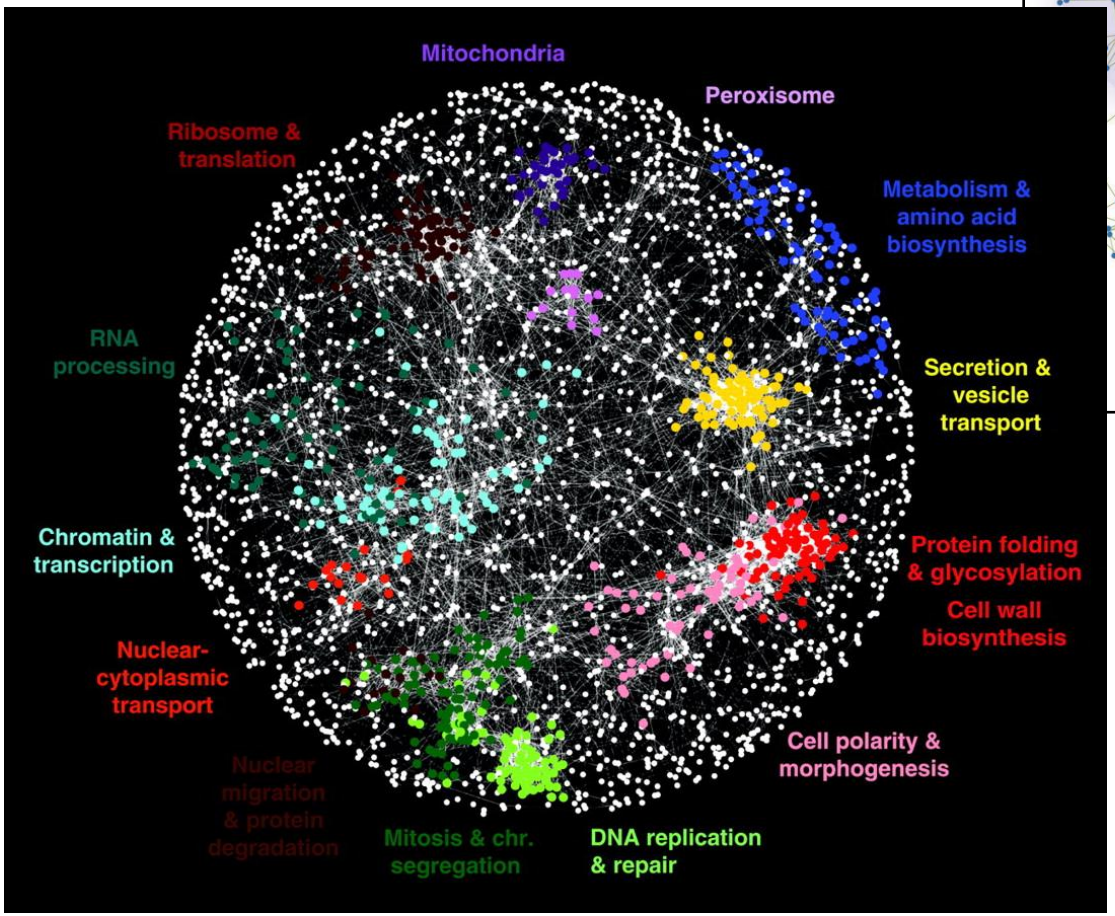
The Power-Law Distribution

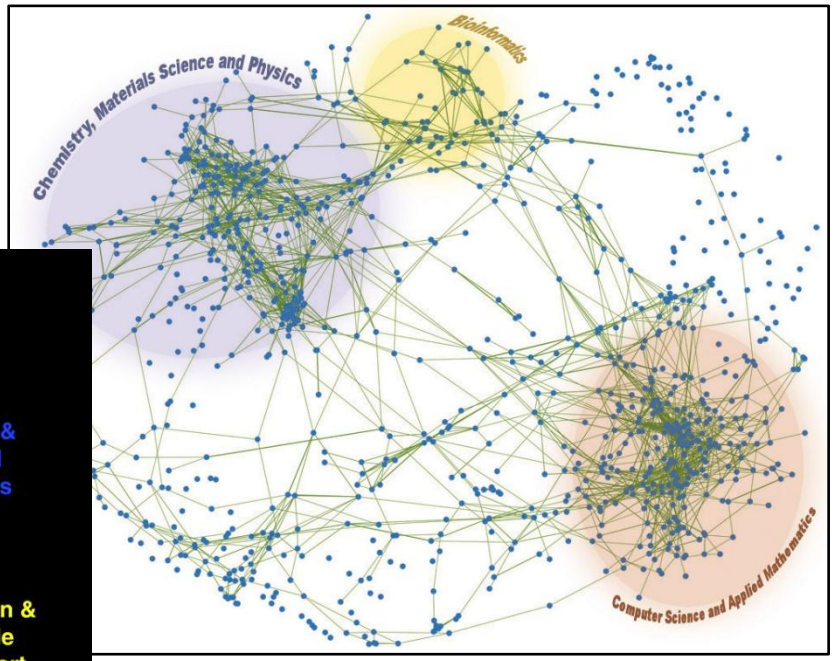
- **Power-law distribution has a "heavy" tail!**
  - Characterized by a small number of highly connected nodes, known as **hubs**
  - A.k.a. "scale-free" network



- **Hubs are crucial:**
  - Affect **error** and **attack** tolerance of complex networks (Albert et al. Nature, 2000)

Network Clustering





Costanzo et al., Nature, 2010

Clustering Coefficient (Watts & Strogatz)

- Characterizes tendency of nodes to cluster
  - "triangles density"
  - How often do my friends know each other (think "facebook")

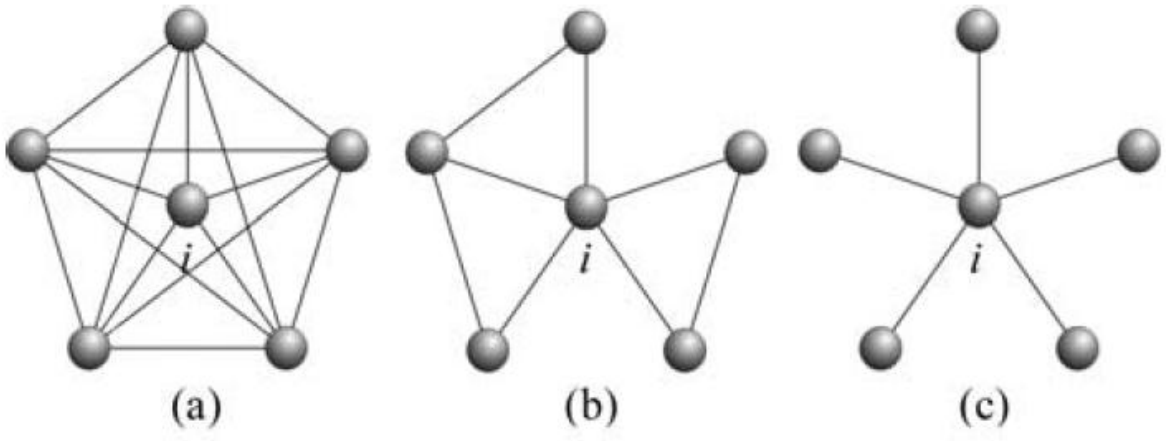
$$C_i = \frac{\#\text{ of edges among neighbors}}{\text{Max. possible }\#\text{ of edges among neighbors}} = \frac{2E_i}{d_i(d_i - 1)}$$

$$C = \frac{1}{N}\sum_v C_i$$

(if $d_i = 0$ or 1 then C_i is defined to be 0)

Clustering Coefficient: Example

- Lies in [0,1]
  - For cliques: C=1
  - For triangle-free graphs: C=0



C_i=10/10=1 C_i=3/10=0.3 C_i=0/10=0

Network Structure in Real Networks

Network	Size	$\langle k \rangle$	C	C_{rand}
WWW, site level, undir.	153, 127	35.21	0.1078	0.00023
Internet, domain level	3015 - 6209	3.52 - 4.11	0.18 - 0.3	0.001
Movie actors	225, 226	61	0.79	0.00027
LANL coauthorship	52, 909	9.7	0.43	1.8×10^{-4}
MEDLINE coauthorship	1, 520, 251	18.1	0.066	1.1×10^{-5}
SPIRES coauthorship	56, 627	173	0.726	0.003
NCSTRL coauthorship	11, 994	3.59	0.496	3×10^{-4}
Math coauthorship	70, 975	3.9	0.59	5.4×10^{-5}
Neurosci. coauthorship	209, 293	11.5	0.76	5.5×10^{-5}
E. coli, substrate graph	282	7.35	0.32	0.026
E. coli, reaction graph	315	28.3	0.59	0.09
Ythan estuary food web	134	8.7	0.22	0.06
Silwood park food web	154	4.75	0.15	0.03
Words, cooccurence	460.902	70.13	0.437	0.0001
Words, synonyms	22, 311	13.48	0.7	0.0006
Power grid	4, 941	2.67	0.08	0.005
C. Elegans	282	14	0.28	0.05

Average Distance

- Distance:

  *Length of shortest (**geodesic**) path between two nodes*

- Average distance:

  average over all connected pairs

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

Small World Networks

- **Despite their often large size, in most (real) networks there is a relatively short path between any two nodes**

Danica McKellar: 6

- "Six degrees of separation"
  (Stanley Milgram;1967)

- Collaborative distance:
  - Erdös number
  - Bacon number

Daniel Kleitman: 3

Natalie Portman: 6

Additional Measures

- Network Modularity
- Giant component
- Betweenness centrality
- Current information flow
- Bridging centrality
- Spectral density

Network Motifs

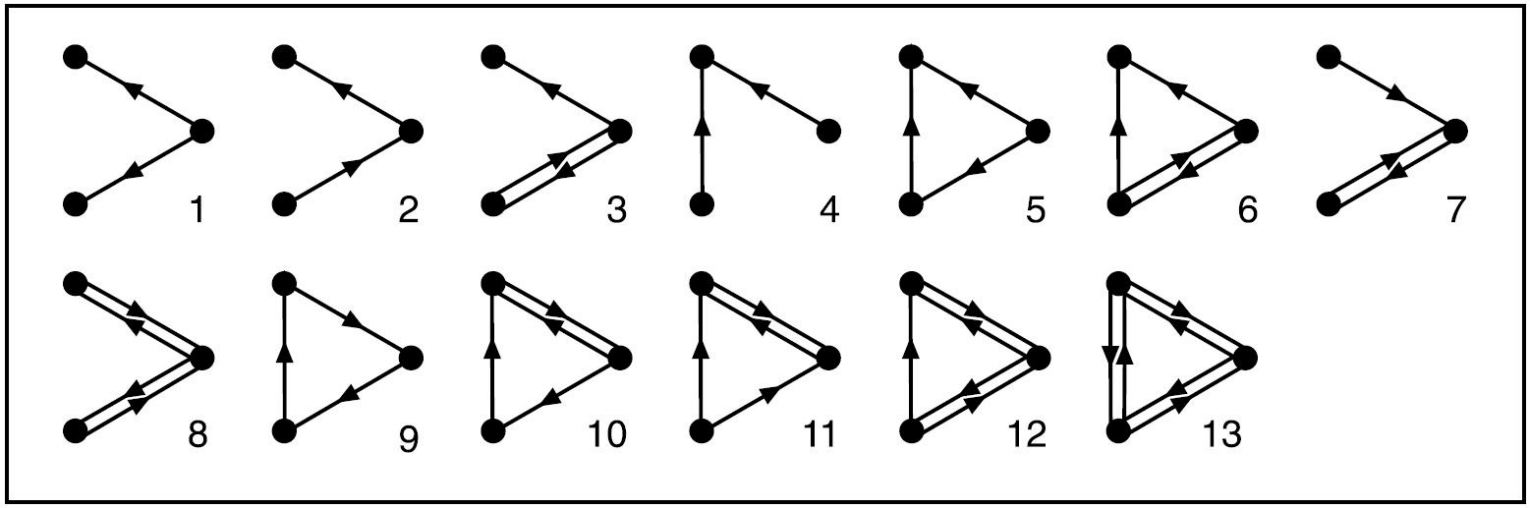
Network Motifs

- Going beyond degree distribution …
- Generalization of sequence motifs
- Basic building blocks
- Evolutionary design principles

R. Milo et al. Network motifs: simple building blocks of complex networks. Science, 2002

What are Network Motifs?

- Recurring patterns of interactions (*subgraphs*) that are significantly overrepresented (w.r.t. a background model)



13 possible 3-nodes subgraphs

(199 possible 4-node subgraphs)

R. Milo et al. Network motifs: simple building blocks of complex networks. Science, 2002

Finding motifs in the Network

1. Generate randomized networks

2a. Scan for all n-node subgraphs in the ***real*** network

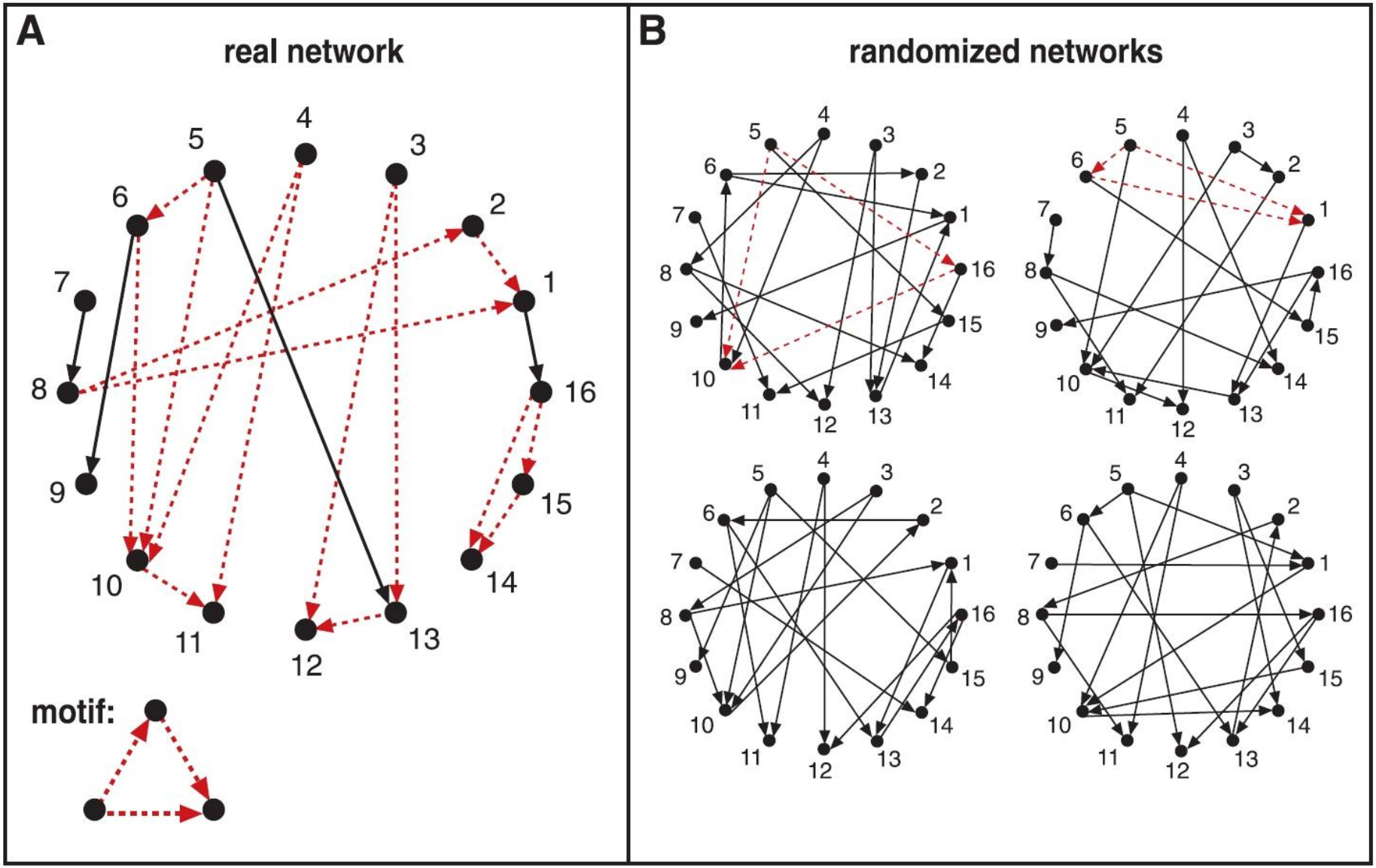
2b. Record number of appearances of each subgraph (*consider isomorphic architectures*)

3a. Scan for all n-node sub graphs in random networks

3b. Record number of appearances of each subgraph

4. Compare each subgraph's data and choose motifs

Finding motifs in the Network



Network Randomization

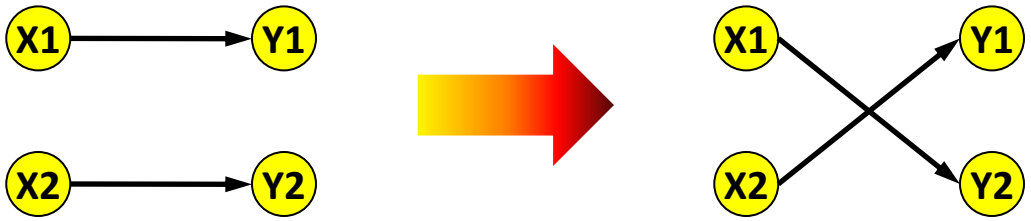
- How should the set of random networks be generated?
- Do we really want “completely random” networks?
- What constitutes a good null model?

Preserve in- and out-degree

(For motifs with n>3 also preserve distribution of smaller sub-motifs)

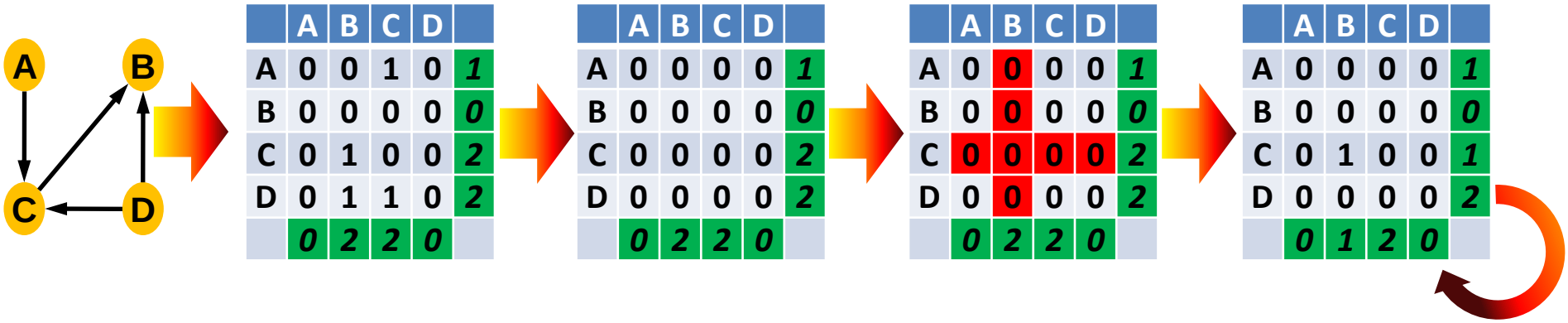
Generation of Randomized Networks

- Algorithm A (Markov-chain algorithm):
  - Start with the real network and repeatedly swap randomly chosen pairs of connections
    (X1→Y1, X2→Y2 is replaced by X1→Y2, X2→Y1)
  - Repeat until the network is well randomized
  - Switching is prohibited if the either of the connections X1→Y2 or X2→Y1 already exist



Generation of Randomized Networks

- Algorithm B (Generative):
  - Record marginal weights of original network
  - Start with an empty connectivity matrix *M*
  - Choose a row *n* & a column *m* according to marginal weights
  - If $M_{nm} = 0$, set $M_{nm} = 1$; Update marginal weights
  - Repeat until all marginal weights are 0
  - If no solution is found, start from scratch



Exact Criteria for Network Motifs

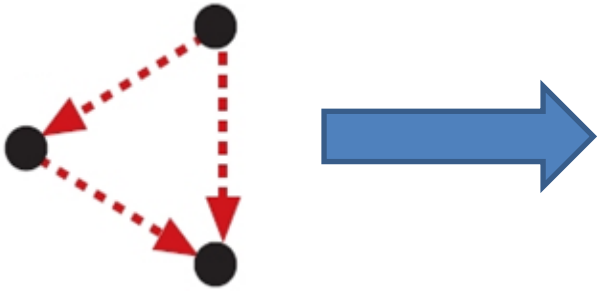
- Subgraphs that meet the following criteria:

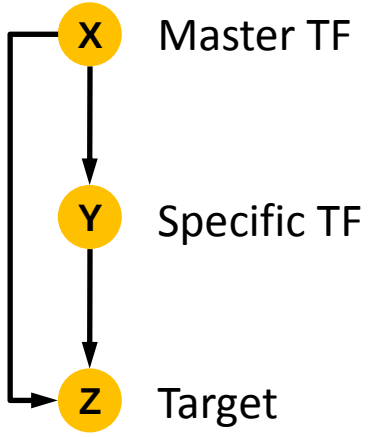
1. The probability that it appears in a randomized network an equal or greater number of times than in the real network is smaller than $P = 0.01$
2. The number of times it appears in the real network with distinct sets of nodes is at least 4
3. The number of appearances in the real network is significantly larger than in the randomized networks: $(N_{real} – N_{rand} > 0.1N_{rand})$

Feed-Forward Loops in Transcriptional Regulatory Networks

- E. Coli network
  - 424 operons (116 TFs)
  - 577 interactions
  - Significant enrichment of motif # 5

(40 instances vs. 7±3)



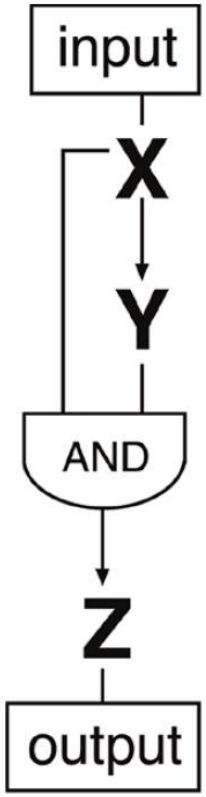


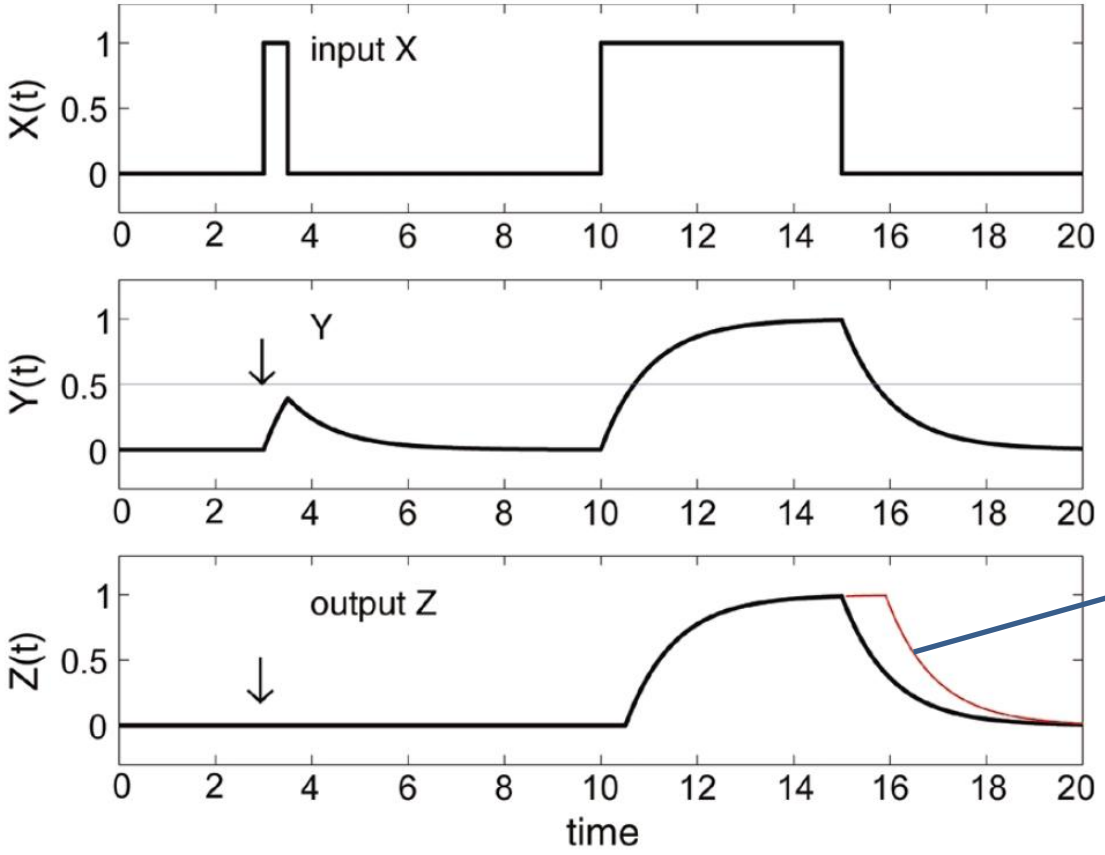
Feed-Forward Loop (FFL)

- Coherent FFLs:
  - The direct effect of x on z has the same sign as the **net** indirect effect through y
  - 85% of FFLs are coherent

S. Shen-Orr et al. Nature Genetics 2002

What's So Cool about FFLs





Boolean Kinetics

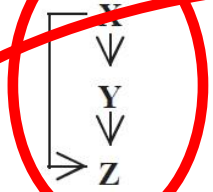
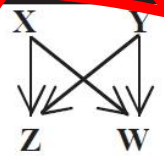
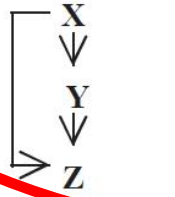
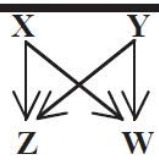
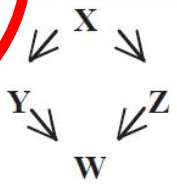
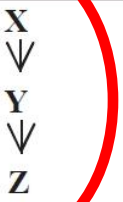
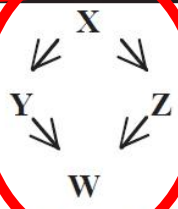
$$dY/dt = F(X, T_y) - aY$$

$$dZ/dt = F(X, T_y)F(Y, T_z) - aZ$$

A simple cascade has slower shutdown

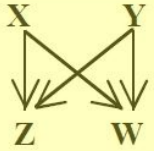
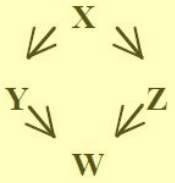
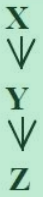
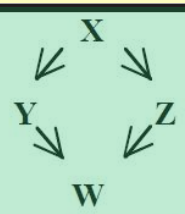
A coherent feed-forward loop can act as a circuit that rejects transient activation signals from the general transcription factor and responds only to persistent signals, while allowing a rapid system shutdown.

Network Motifs in Biological Networks

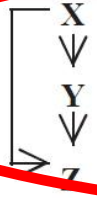
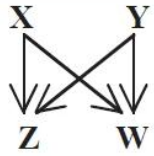
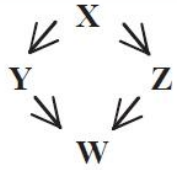
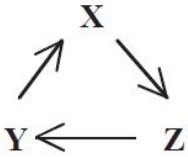
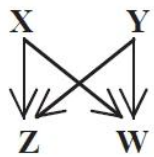
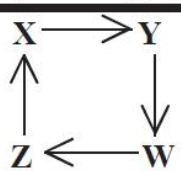
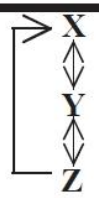
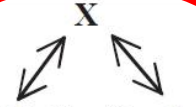
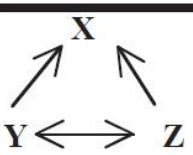
Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

FFL motif is under-represented!

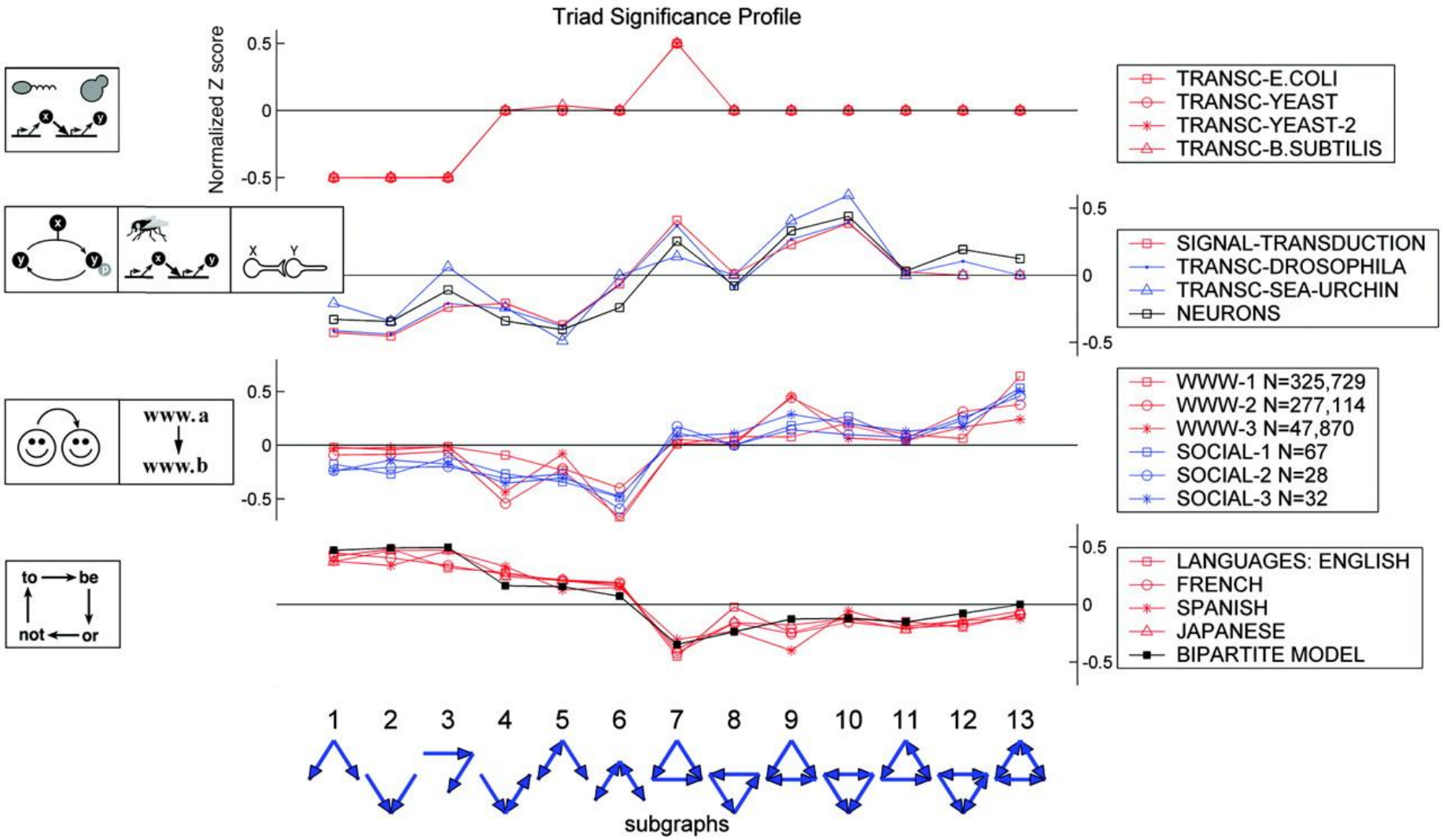
Information Flow vs. Energy Flow

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**	**FFL motif is under-represented!**		
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

Network Motifs in Technological Networks

			 Feed-forward loop			 Bi-fan			 Bi-parallel		
Electronic circuits (forward logic chips)											
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)			 Three-node feedback loop			 Bi-fan			 Four-node feedback loop		
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web			 Feedback with two mutual dyads			 Fully connected triad			 Uplinked mutual dyad		
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Network Comparison: Motif-Based Network Superfamilies



R. Milo et al. Superfamilies of evolved and designed networks. Science, 2004

Evolutionary Conservation of Motif Elements

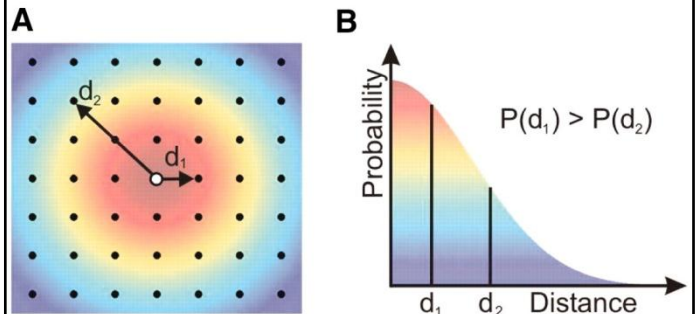
#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

Wuchty et al. Nature Genetics, 2003

Criticism of the Randomization Approach

- An incomplete null model?
- Local clustering:
  - Neighboring neurons have a greater chance of forming a connection than distant neurons
- Similar motifs are obtained in random graphs devoid of any selection rule
  - Gaussian toy network
  - Preferential-attachment rule

Gaussian “toy network"



C

Motif	Feed-forward	Bi-fan	Bi-parallel
Observed	42	347	65
Mean	27.1	18.83	43.32
Std.	5.33	8.63	6.4
Z score	2.79	38.02	3.38

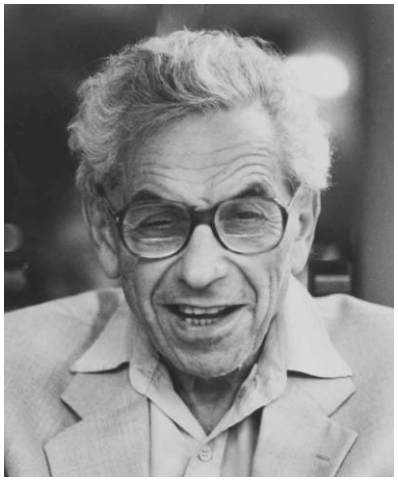
Y. Artzy-Randrup et al. Comment on “Network motifs: simple building blocks of complex networks”.

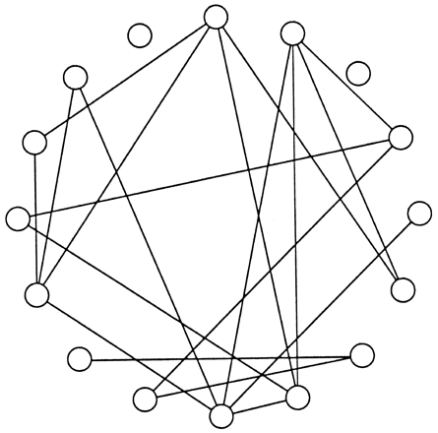
Random Network Models

1. *Random Graphs (Erdös/Rényi)*
2. *Geometric Random Graphs*
3. *The Small World Model (WS)*
4. *Preferential Attachment*

Random Graphs (Erdös/Rényi)

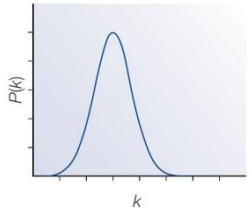
- *N* nodes
- Every pair of nodes is connected with probability *p*





Random Graphs: Properties

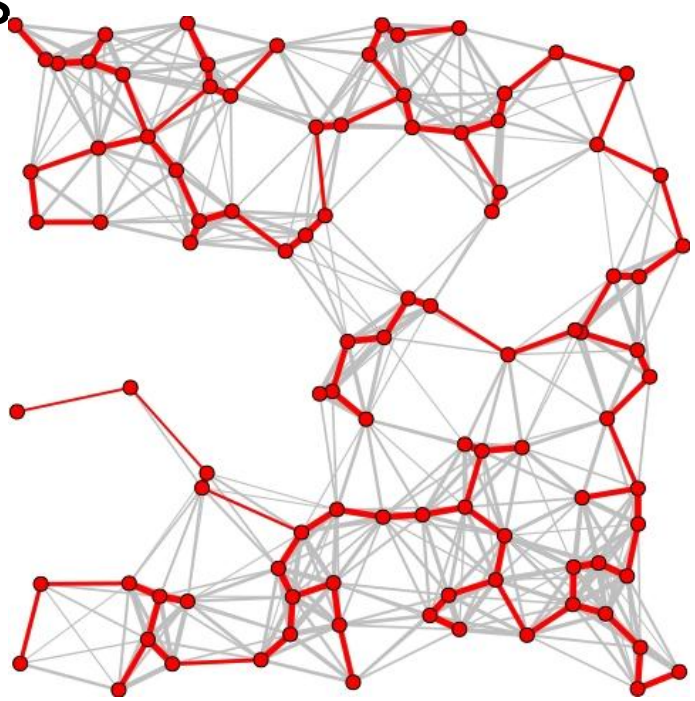
- Mean degree: *d = (N-1)p ~ Np*
- Degree distribution is binomial
  - Asymptotically Poisson: $P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \approx \frac{d^k e^{-d}}{k!}$



- Clustering Coefficient:
  - The probability of connecting two nodes at random is p
  - → Clustering coefficient is C=p
  - In many large networks p ~ 1/n → C is lower than observed
- Average distance:
  - *l~ln(N)/ln(d)* …. (think why?)
  - Small world! (and fast spread of information)

Geometric Random Graphs

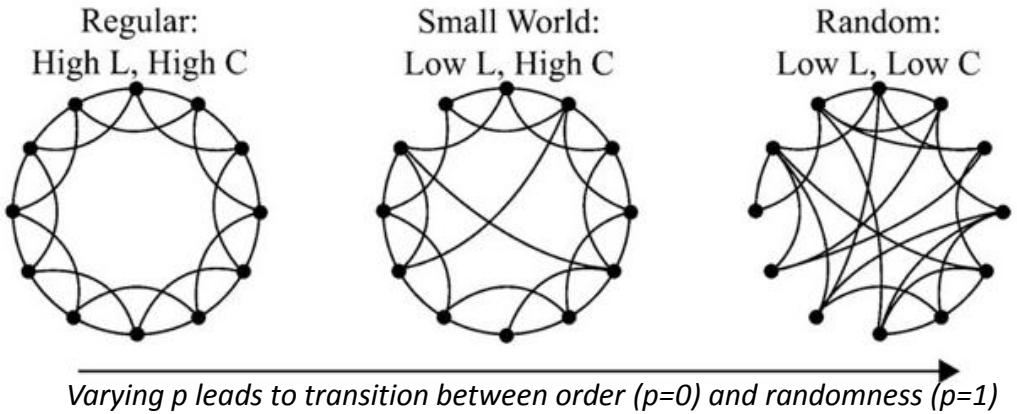
- G=(V,r)
  - V – set of points in a metric space (e.g. 2D)
  - E – all pairs of points with distance ≤ r
- Captures spatial relationships



The Small World Model (WS)

- Generate graphs with **high clustering coefficients *C*** and **small distance *l***
- Rooted in social systems

1. **Start with order** (every node is connected to its K neighbors)
2. **Randomize** (rewire each edge with probability p)



Varying p leads to transition between order (p=0) and randomness (p=1)

- **Degree distribution is similar to that of a random graph!**

Watts and Strogatz, Nature, 1998

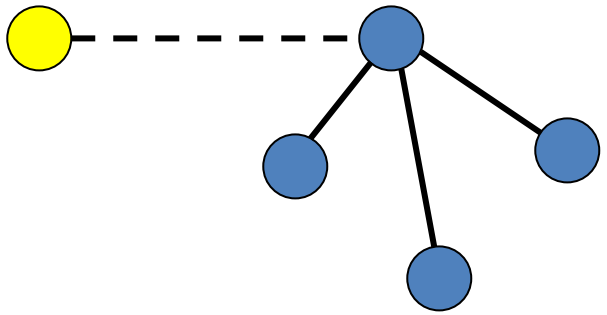
The Scale Free Model: **Preferential Attachment**

- A generative model (dynamics)
  - Growth: degree-m nodes are constantly added
  - *Preferential attachment*: the probability that a new node connects to an existing one is proportional to its degree

$$P(u \text{ connects to } v) = \frac{d(v)}{\sum d(v)}$$

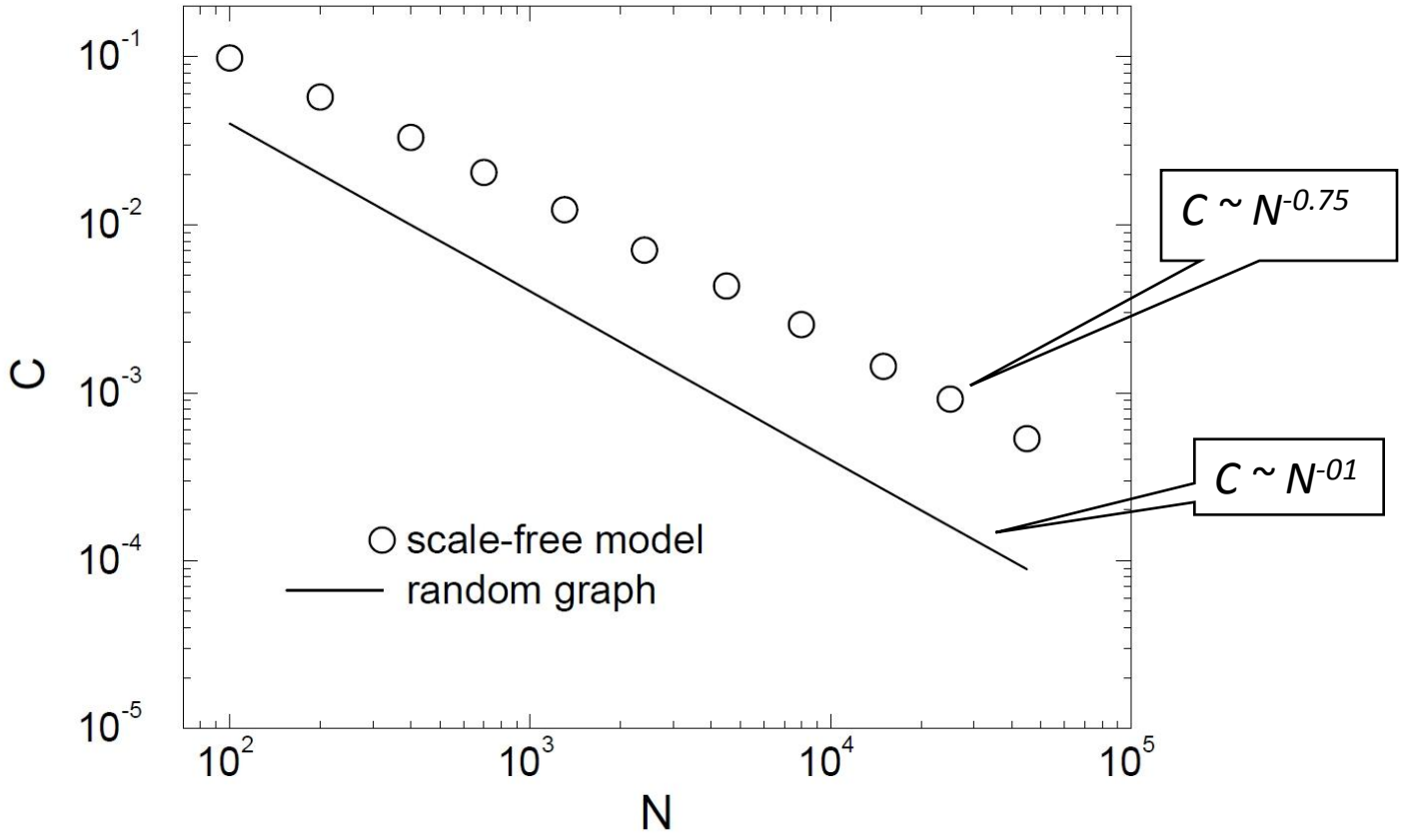
- **"The rich get richer"** principle

$$P(k) = \frac{2m(m+1)}{(k+2)(k+1)k} \sim k^{-3}$$



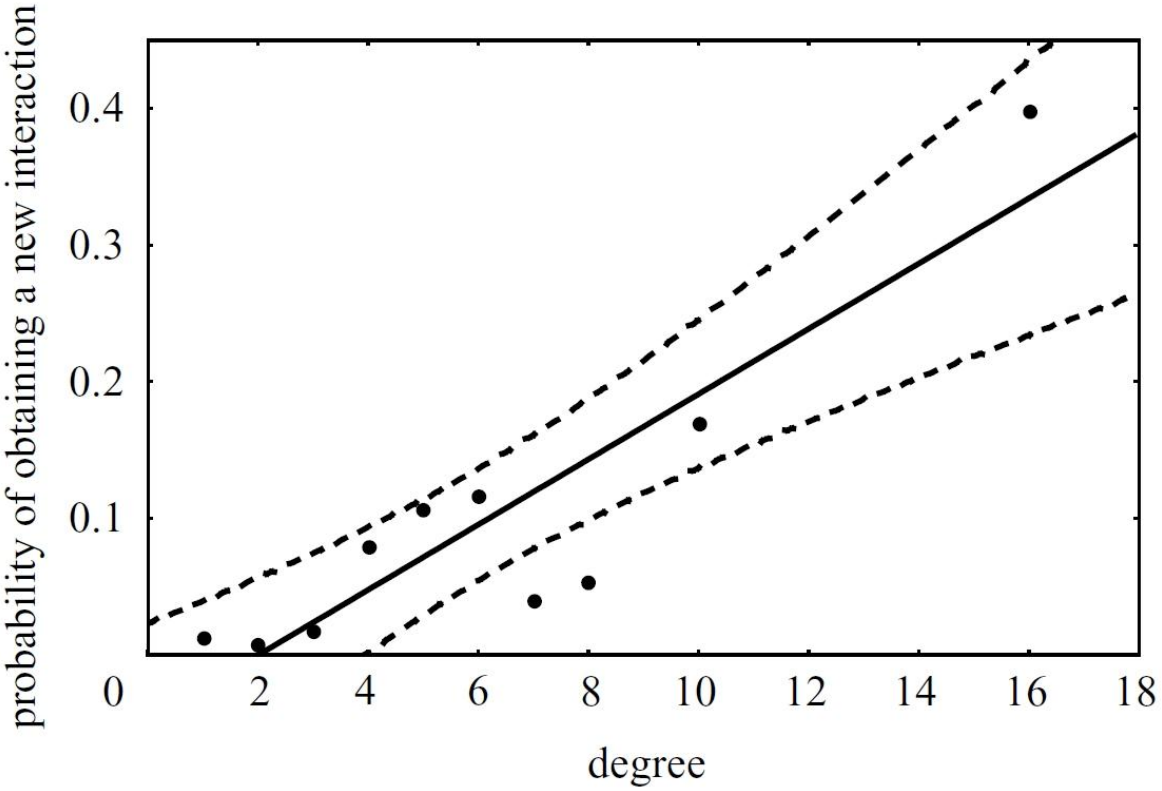
Albert and Barabasi, 2002

Preferential Attachment: Clustering Coefficient



Preferential Attachment: Empirical Evidence

- Highly connected proteins in a PPI network are more likely to evolve new interactions



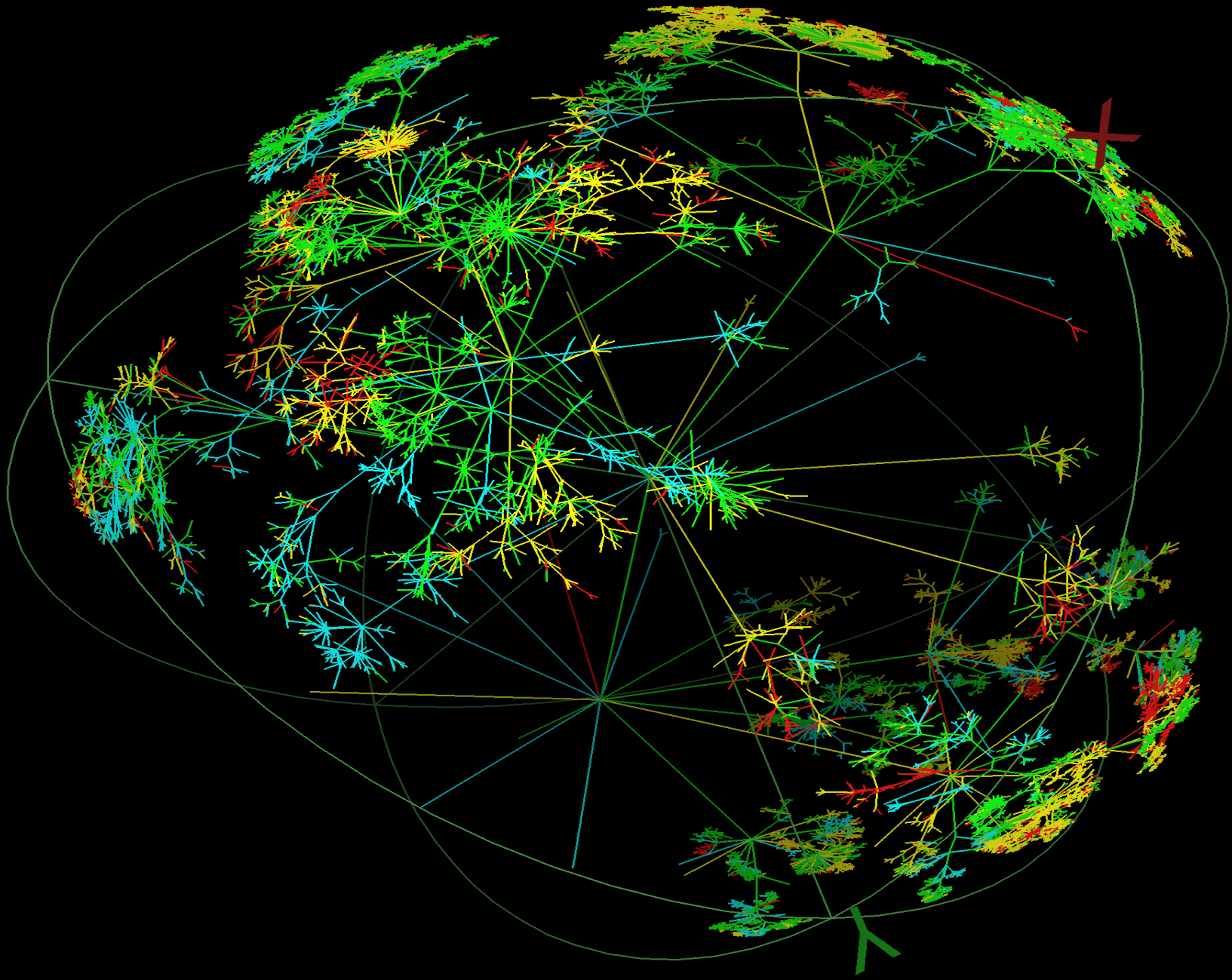
Wagner, A. Proc. R. Soc. Lond. B , 2003

Model Problems

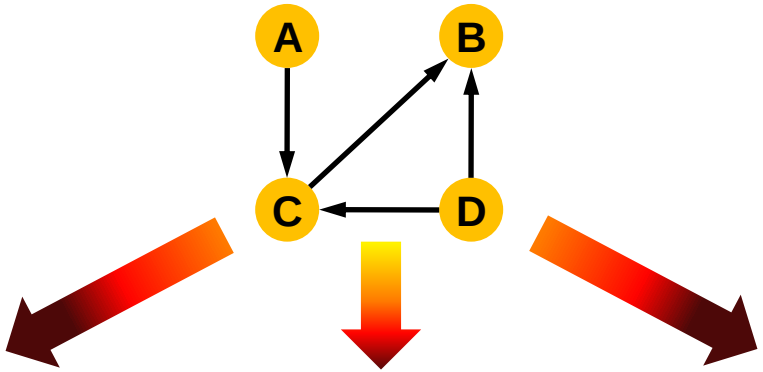
- Degree distribution is fixed
  (although there are generalizations of this method that handle various distributions)
- Clustering coefficient approaches 0 with network size, unlike real networks
- Issues involving **biological** network growth:
  - Ignores local events shaping real networks (e.g., insertions/deletions of edges)
  - Ignores growth constraints (e.g., max degree) and aging (a node is active in a limited period)

Conclusions

- No single best model!
  - Models differ in various network measures
  - Different models capture different attributes of real networks
- In literature, "*random graphs*" are most commonly used



Computational Representation of Networks



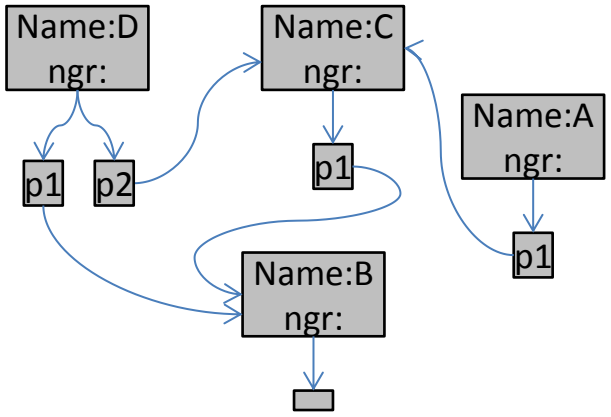
List/set of edges:
(ordered) pairs of nodes

{ (A,C) , (C,B) , (D,B) , (D,C) }

Connectivity Matrix

	A	B	C	D
A	0	0	1	0
B	0	0	0	0
C	0	1	0	0
D	0	1	1	0

Object Oriented



- **Which is the most useful representation?**