

Biological Networks Analysis

Degree Distribution and Network Motifs

Genome 559: Introduction to Statistical and
Computational Genomics

Elhanan Borenstein

A quick review

- Ab initio gene prediction

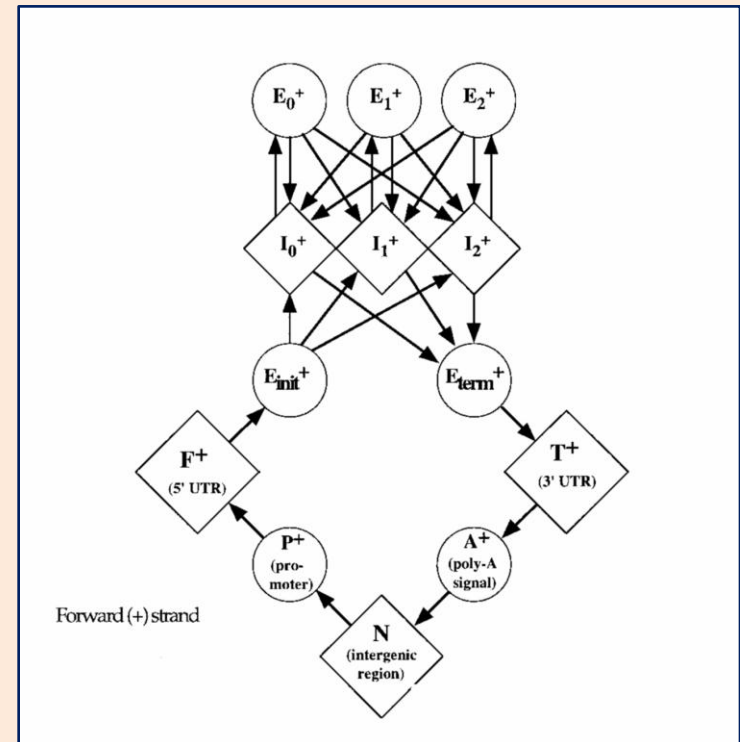
- Parameters:

- Splice donor sequence model
- Splice acceptor sequence model
- Intron and exon length distribution
- Open reading frame
- More ...

- Markov chain

- States
- Transition probabilities

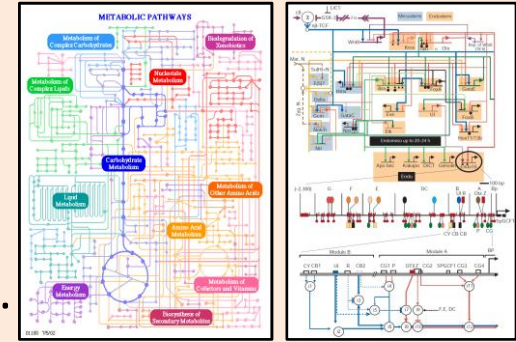
- Hidden Markov Model (HMM)



A quick review

- **Networks:**

- Networks vs. graphs
- A collection of **nodes** and **links**
- Directed/undirected; weighted/non-weighted, ...
- Networks as models vs. networks as tools

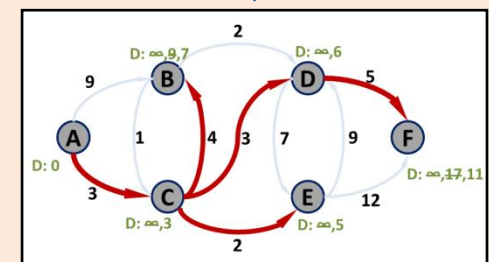
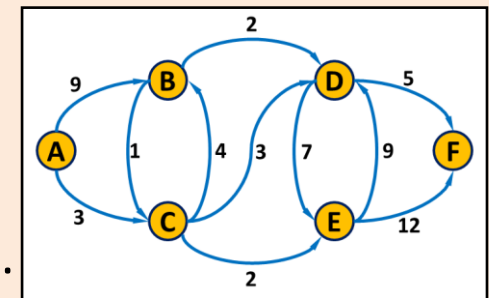


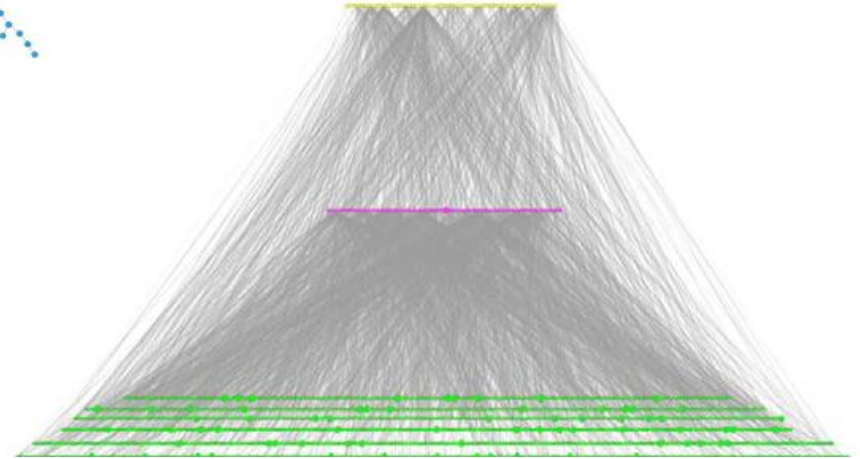
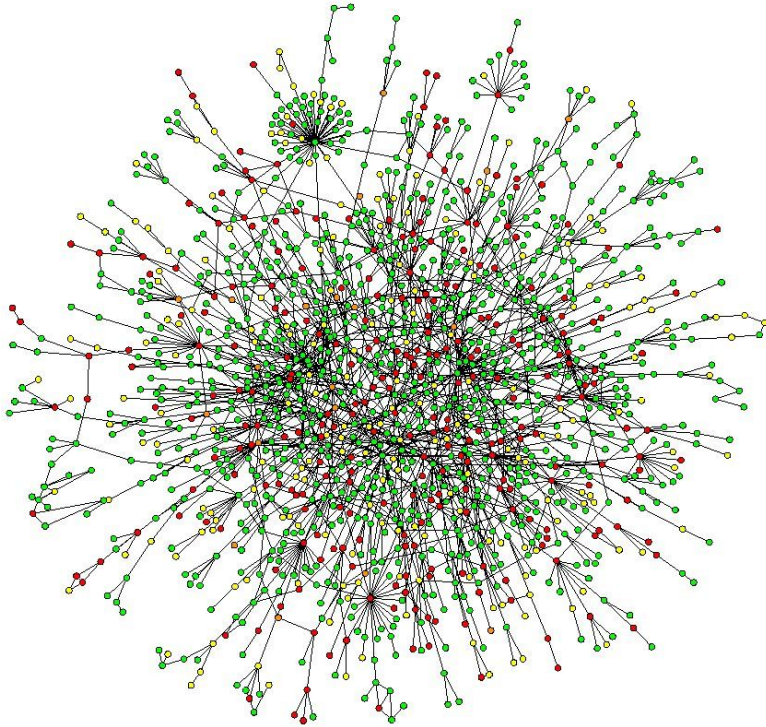
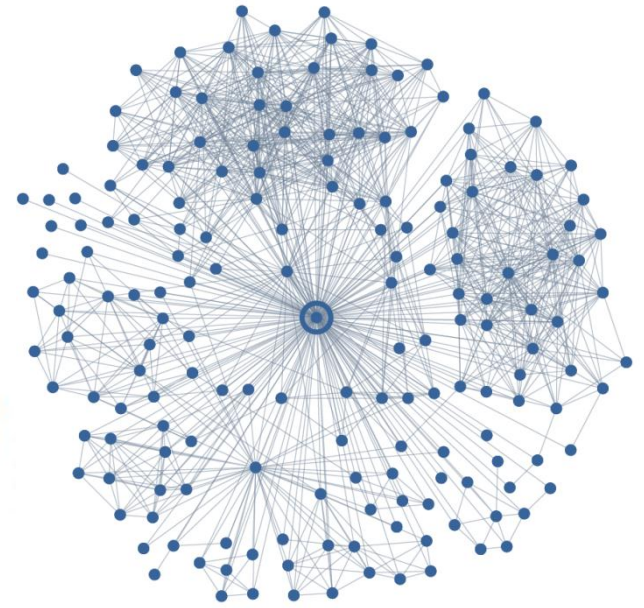
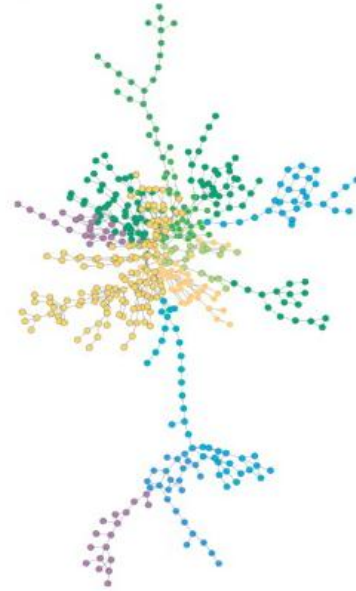
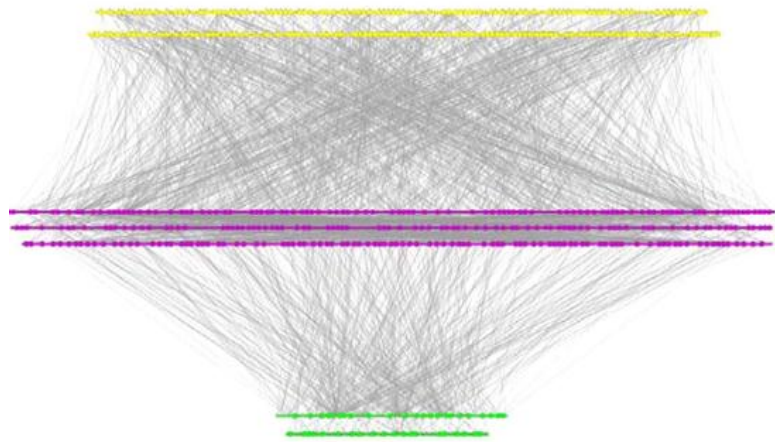
- Many types of biological networks

- The shortest path problem

- Dijkstra's algorithm

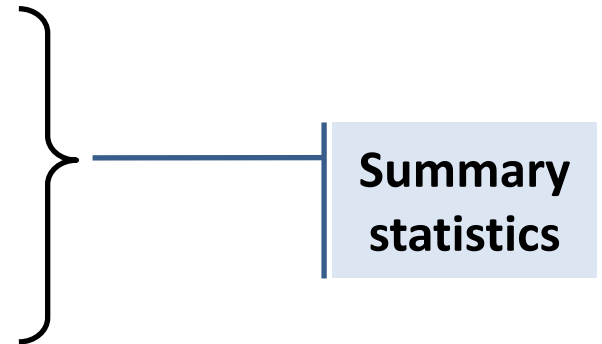
1. **Initialize:** Assign a distance value, D , to each node. Set $D=0$ for *start* node and to infinity for all others.
2. **For each unvisited neighbor of the current node:** Calculate tentative distance, D^t , through current node and if $D^t < D$: $D \leftarrow D^t$. Mark node as visited.
3. **Continue with the unvisited node with the smallest distance**





Comparing networks

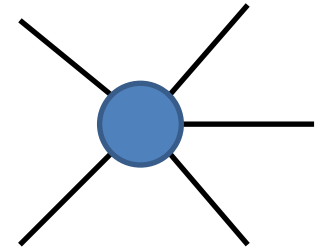
- We want to find a way to “compare” networks.
 - “Similar” (not identical) **topology**
 - “Common” **design principles**
- We seek measures of network topology that are:
 - Simple
 - Capture **global** organization
 - Potentially “important”



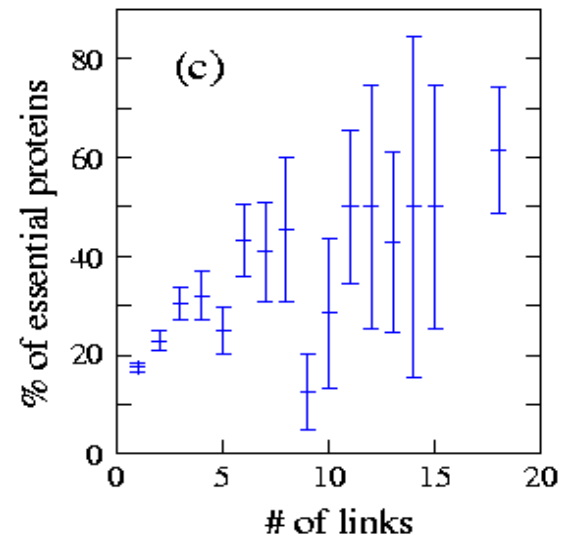
(equivalent to, for example, GC content for genomes)

Node degree / rank

- Degree = Number of neighbors



- Node degree in PPI networks correlates with:
 - Gene essentiality
 - Conservation rate
 - Likelihood to cause human disease



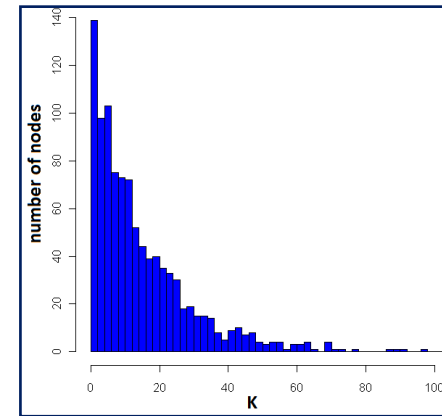
brief communications

Lethality and centrality in protein networks

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

Degree distribution

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



- Common distributions:

Poisson:

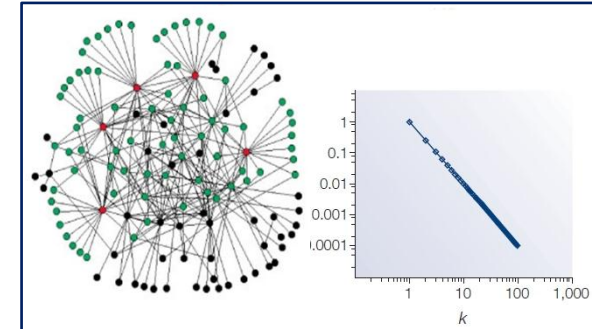
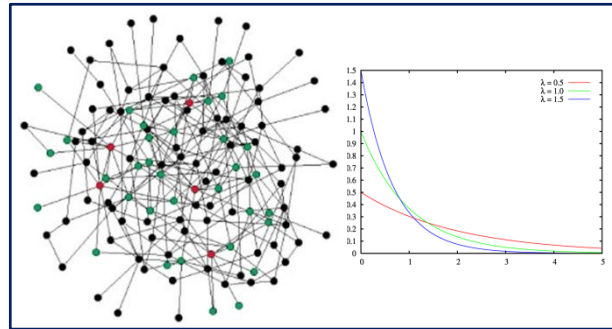
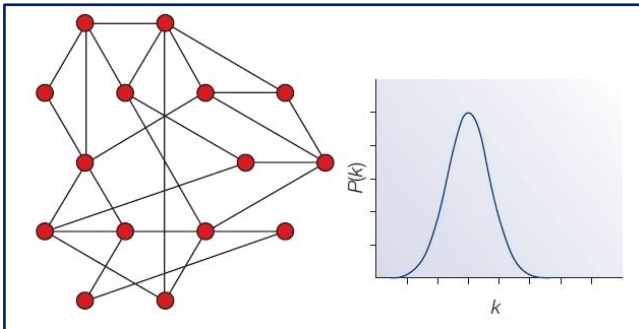
$$P(k) = \frac{e^{-d} d^k}{k!}$$

Exponential:

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

Power-law:

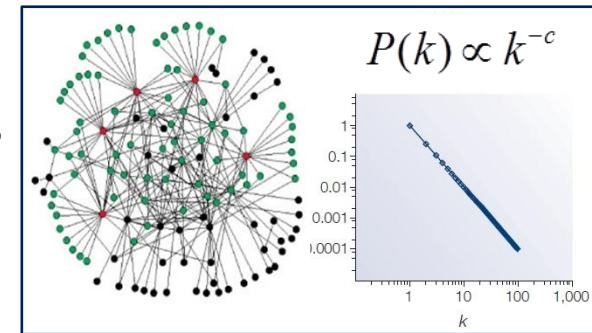
$$P(k) \propto k^{-c}, k \neq 0, c > 1$$



The 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



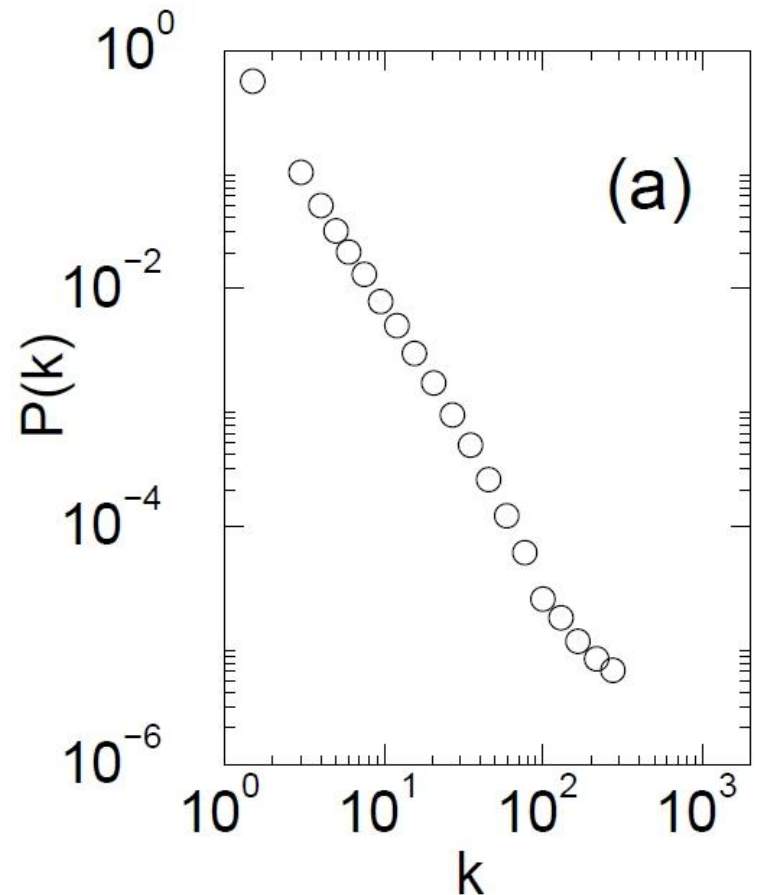
- **Hubs are crucial:**

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



The Internet

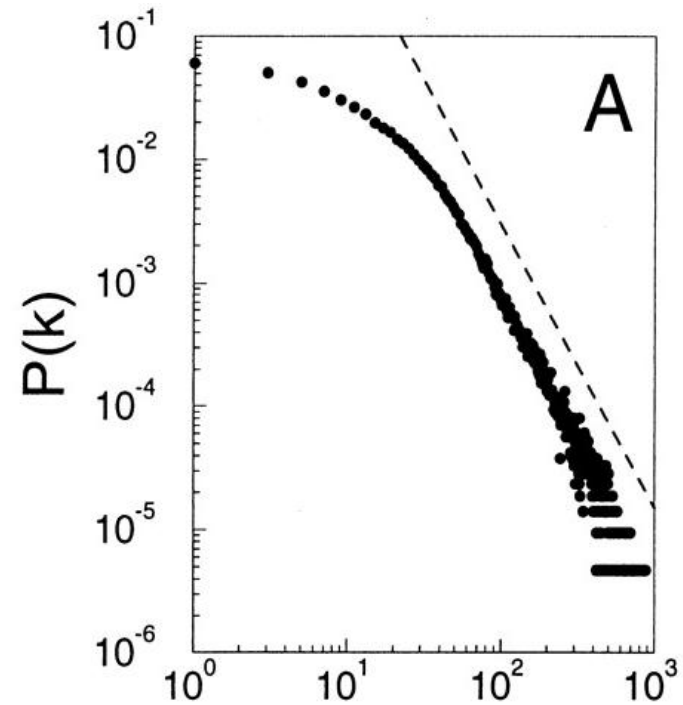
- **Nodes** – 150,000 routers
- **Edges** – physical links
- $P(k) \sim k^{-2.3}$



Movie actor collaboration network

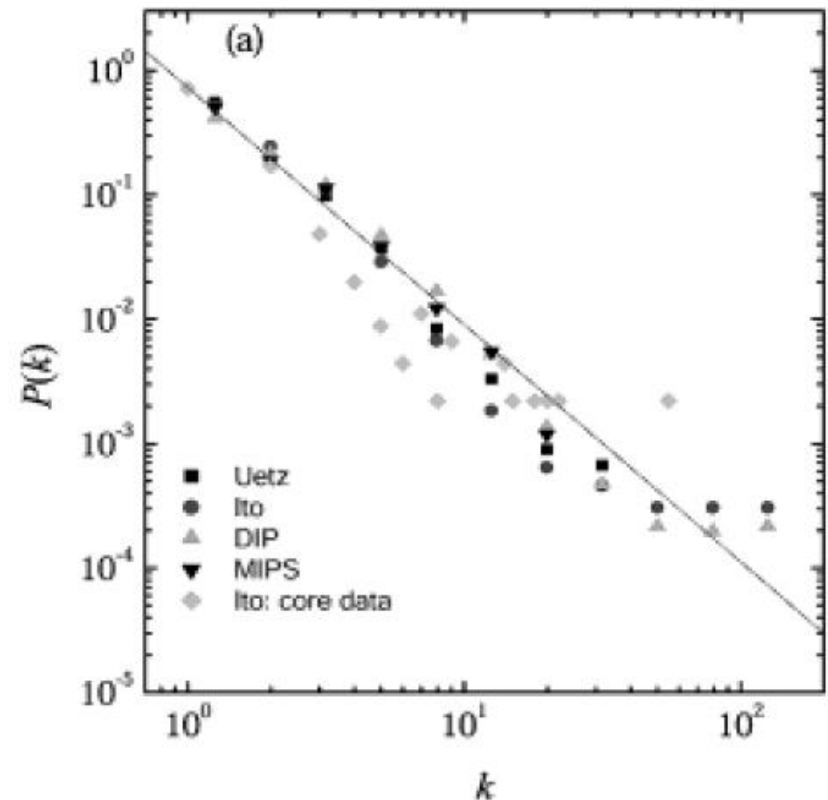


- **Nodes** – 212,250 actors
- **Edges** – co-appearance in a movie
- $P(k) \sim k^{-2.3}$



Protein protein interaction networks

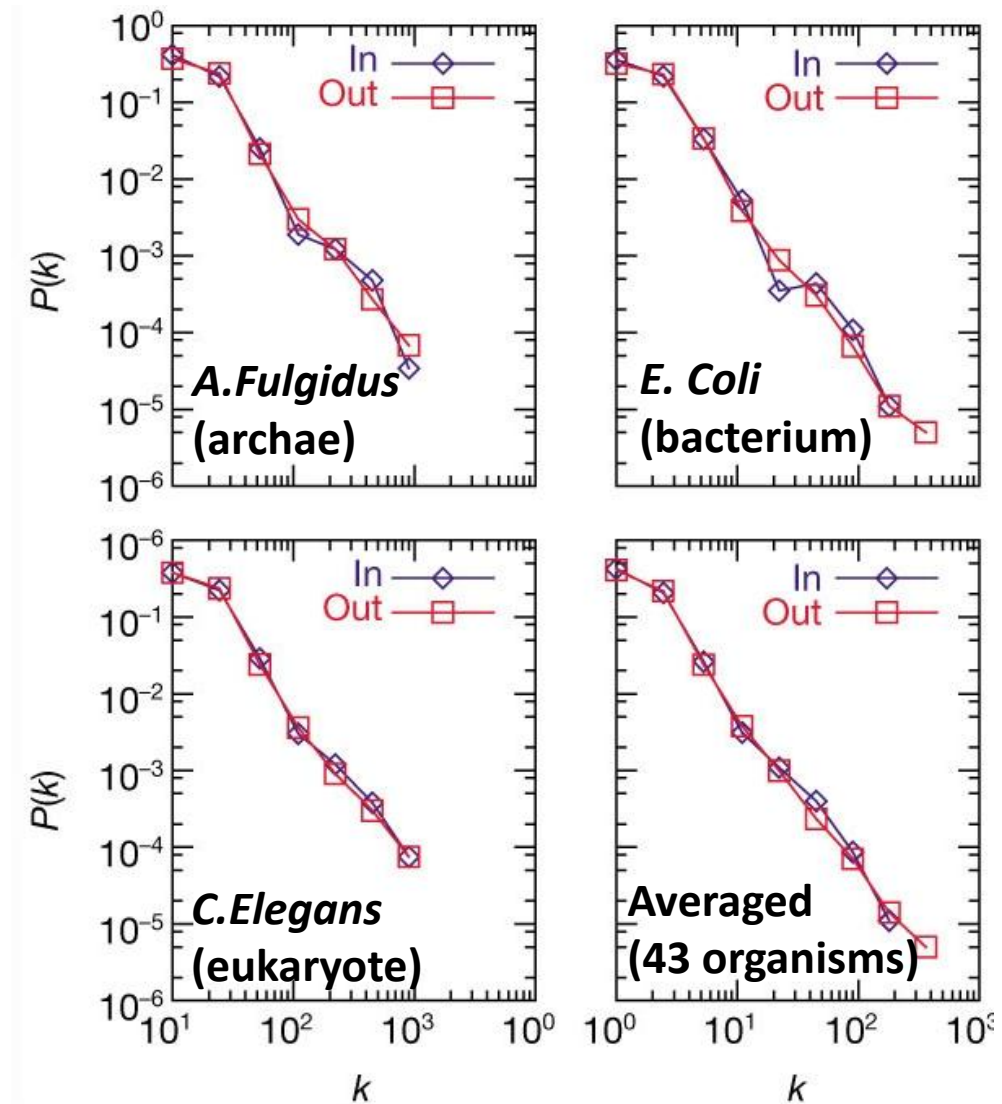
- **Nodes** – Proteins
- **Edges** – Interactions (yeast)
- $P(k) \sim k^{-2.5}$



Metabolic networks

- **Nodes** – Metabolites
- **Edges** – Reactions
- $P(k) \sim k^{-2.2 \pm 2}$

Metabolic networks across all kingdoms of life are scale-free



Why do so many real-life networks exhibit a power-law degree distribution?

- Is it “selected for”?
- Is it expected by chance?
- Does it have anything to do with the way networks evolve?
- Does it have functional implications?

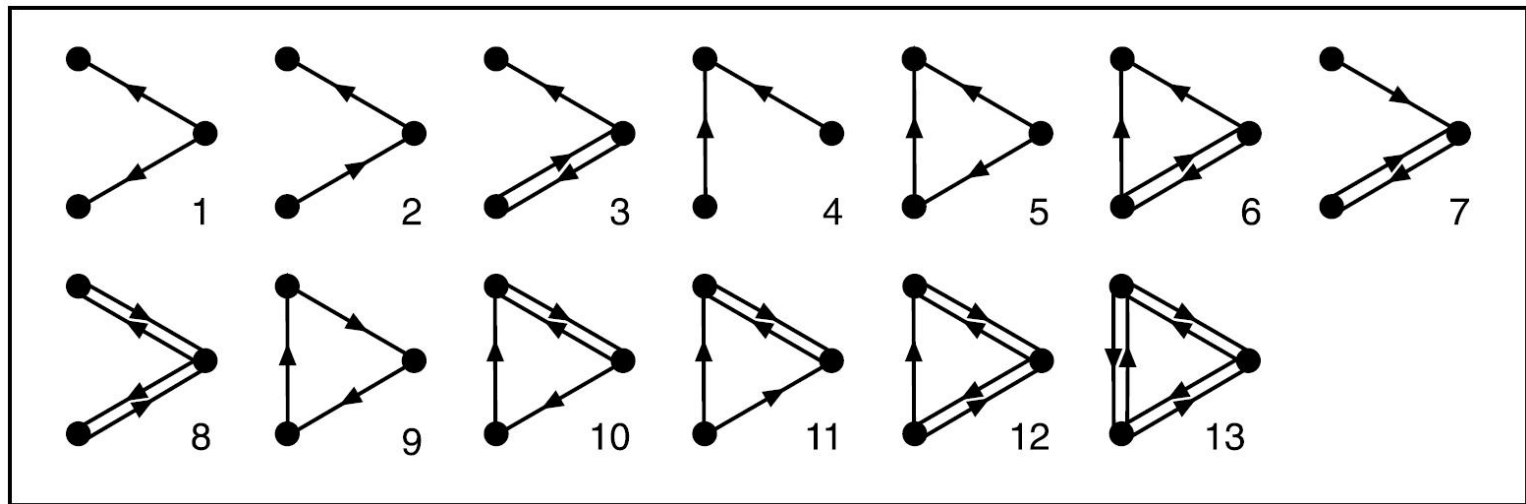


Network motifs

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

What are network motifs?

- Recurring patterns of interaction (*sub-graphs*) that are significantly **overrepresented** (w.r.t. a background model)

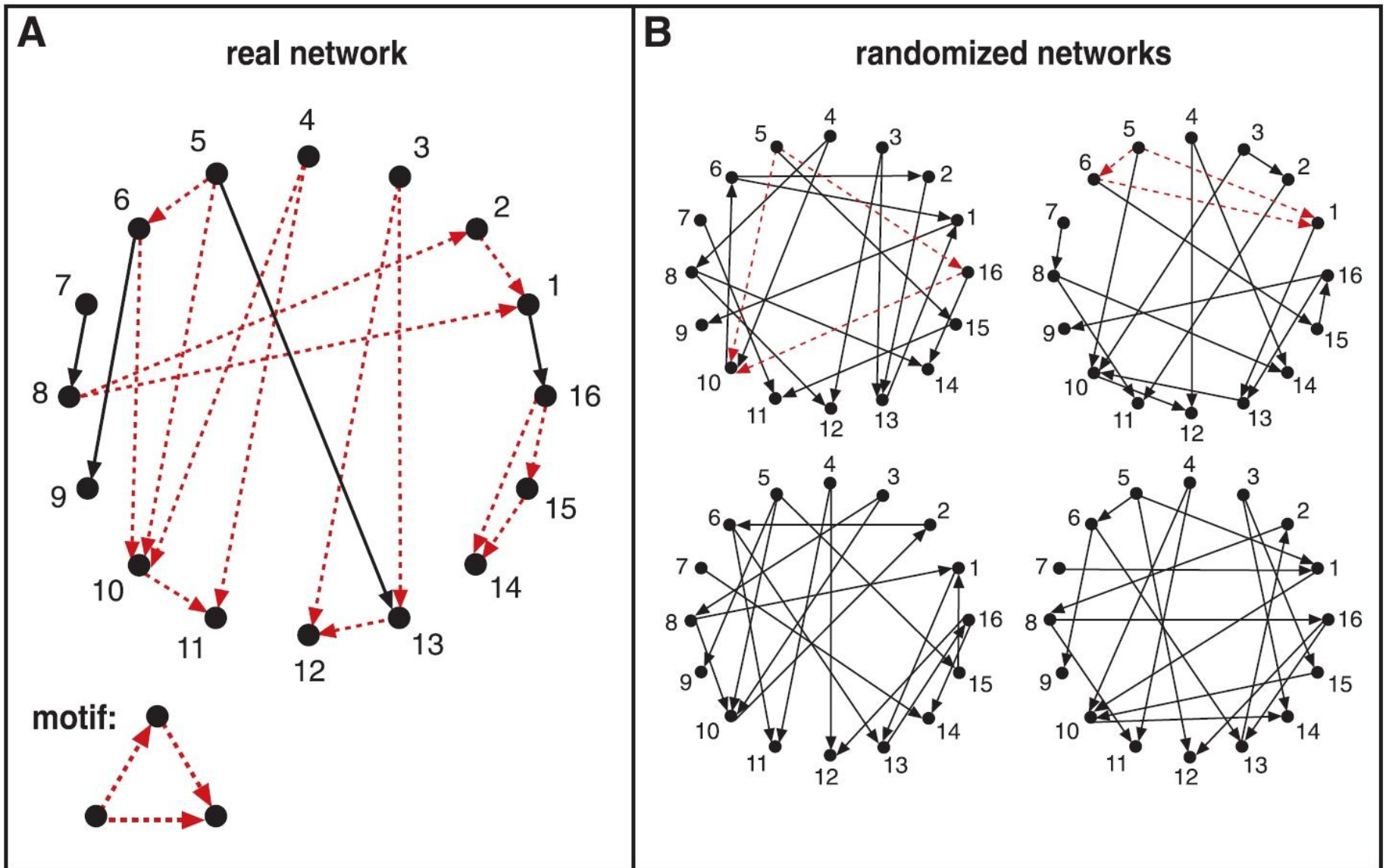


13 possible 3-nodes sub-graphs
(199 possible 4-node sub-graphs)

Finding motifs in the network

- 1a. Scan all n-node sub-graphs in the *real* network
- 1b. Record number of appearances of each sub-graph (*consider isomorphic architectures*)
2. Generate a large set of random networks
- 3a. Scan for all n-node sub-graphs in **random** networks
- 3b. Record number of appearances of each sub-graph
4. Compare each sub-graph's data and identify 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?

Network randomization

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

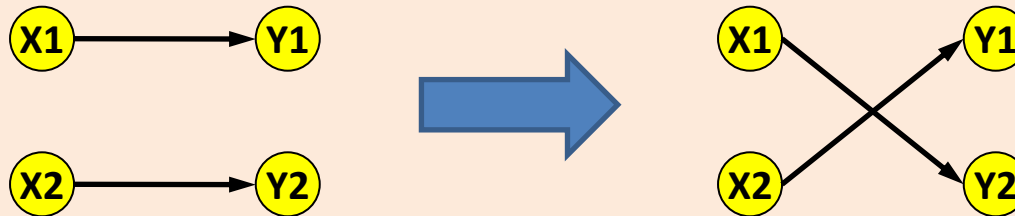


Preserve in- and out-degree

Generation of randomized networks

Network randomization algorithm :

- Start with the real network and repeatedly swap randomly chosen pairs of connections
($X1 \rightarrow Y1, X2 \rightarrow Y2$ is replaced by $X1 \rightarrow Y2, X2 \rightarrow Y1$)



(Switching is prohibited if either of the $X1 \rightarrow Y2$ or $X2 \rightarrow Y1$ already exist)

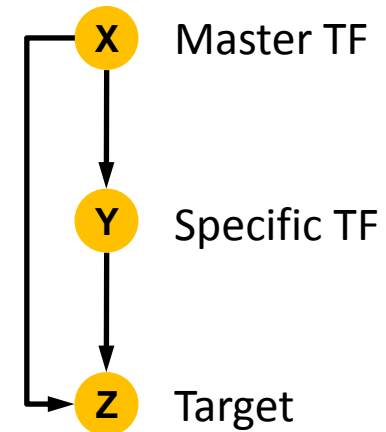
- Repeat until the network is “well randomized”

Motifs in transcriptional regulatory networks

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



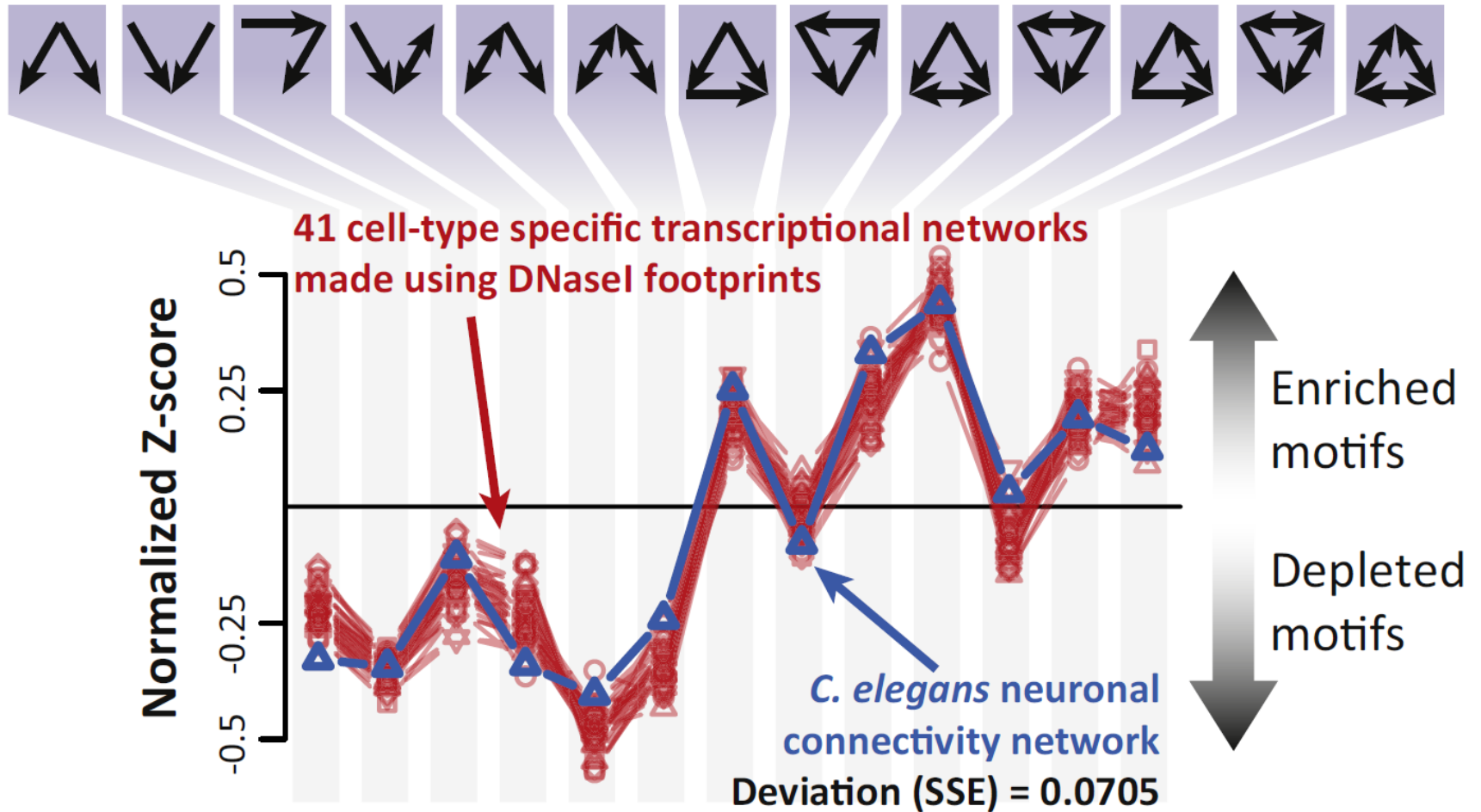
(40 instances vs. 7 ± 3)



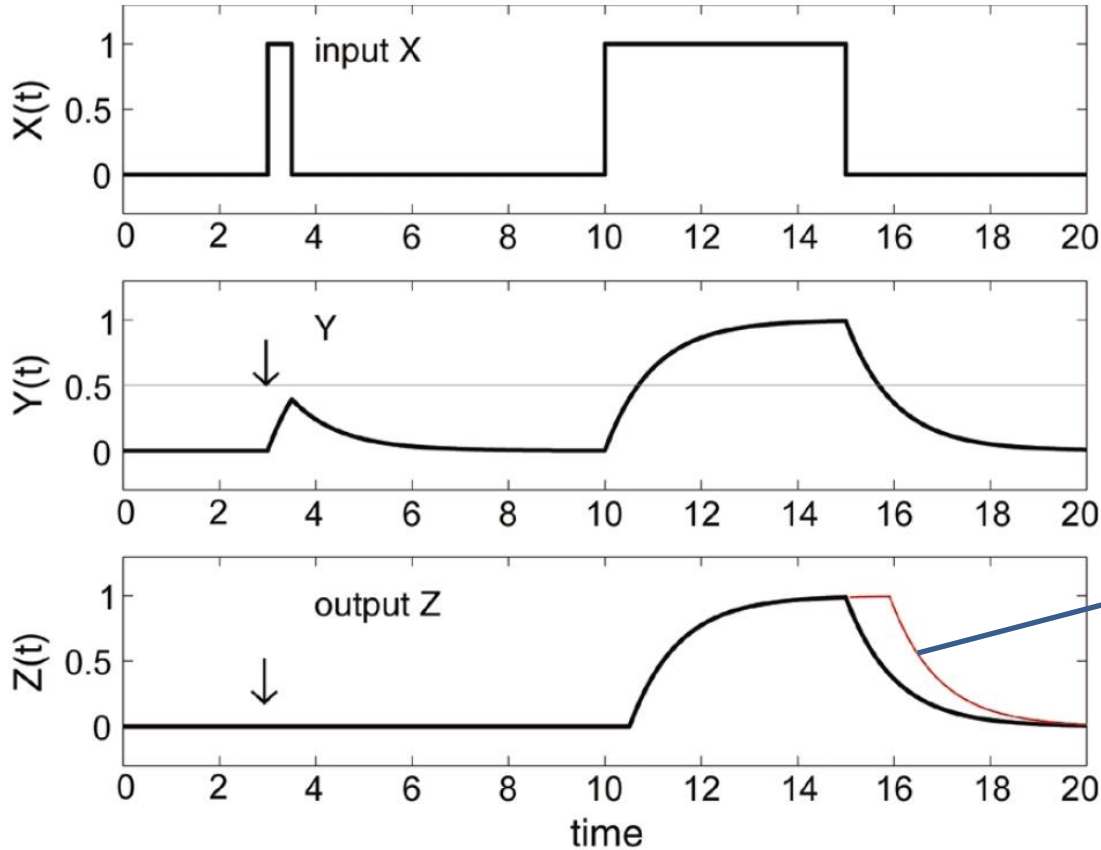
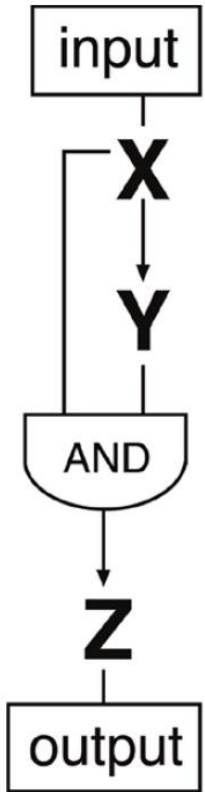
**Feed-Forward Loop
(FFL)**

Motifs in transcriptional regulatory networks

- Human cell-specific networks



What's so interesting about FFLs



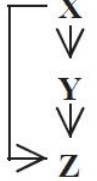
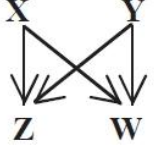
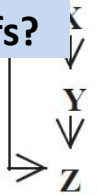
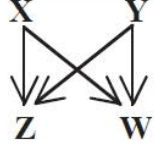
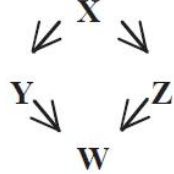
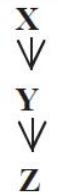
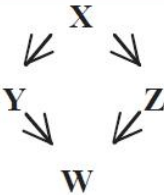
Boolean Kinetics

$$\begin{aligned} dY / dt &= F(X, T_y) - aY \\ dZ / dt &= F(X, T_y)F(Y, T_z) - aZ \end{aligned}$$

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 for a rapid system shutdown.

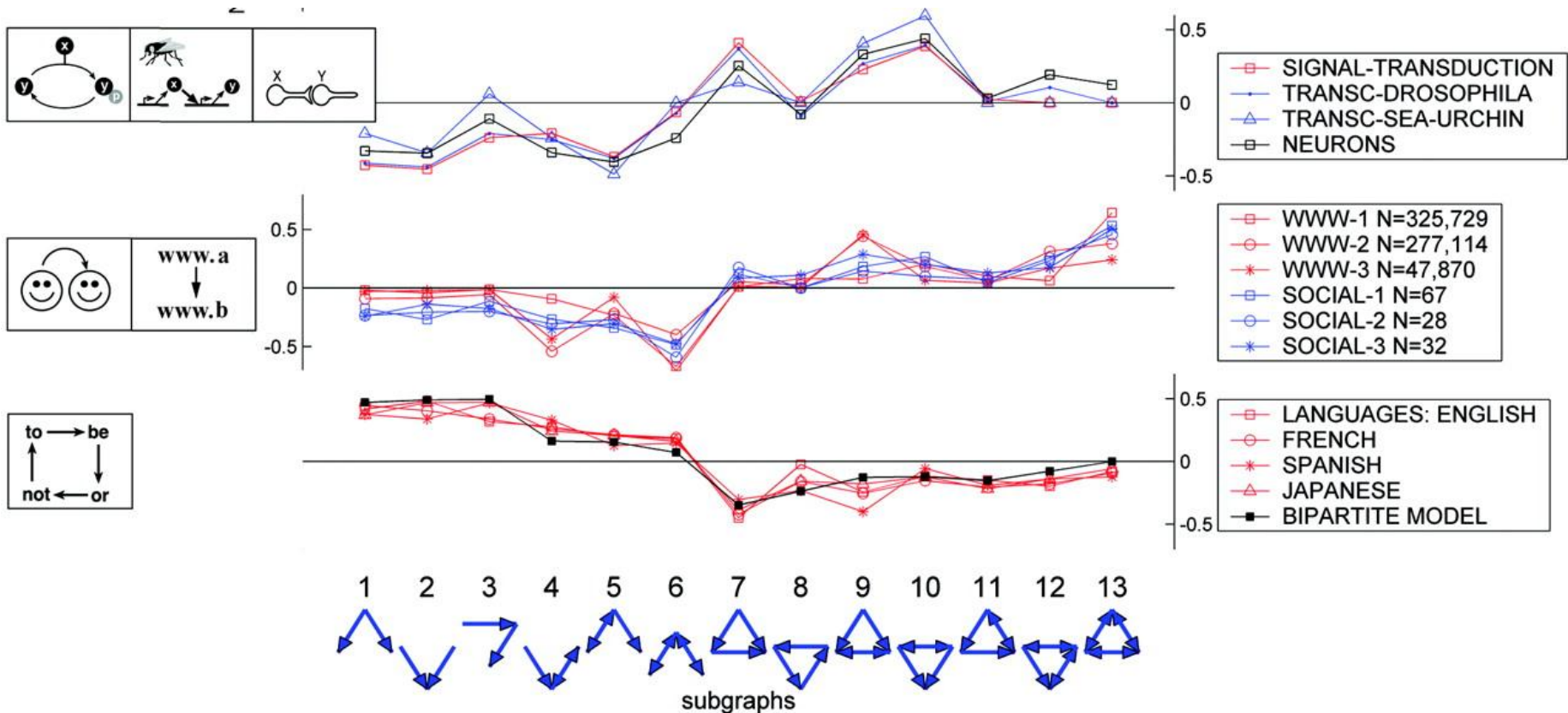
Network motifs in biological networks

Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score
Gene regulation (transcription)				Feed-forward loop			Bi-fan				
<i>E. coli</i>	424			7 ± 3	10	203	47 ± 12	13			
<i>S. cerevisiae</i> *	685			11 ± 4	14	1812	300 ± 40	41			
Neurons				Feed-forward loop			Bi-fan			Bi-parallel	
<i>C. elegans</i> †	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			

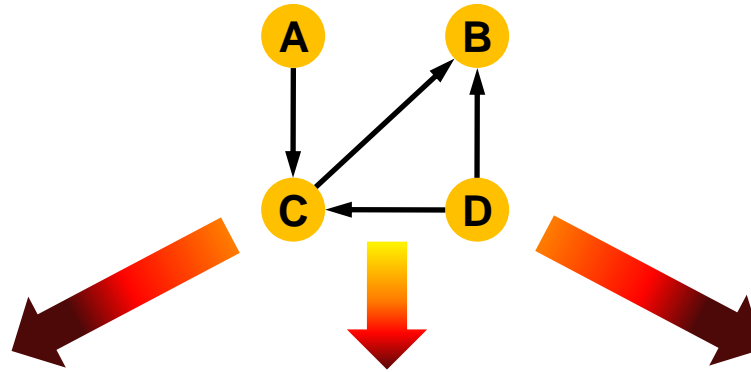
Why do these networks have similar motifs?

Why is this network so different?

Motif-based network super-families



Computational representation of networks



List of edges:

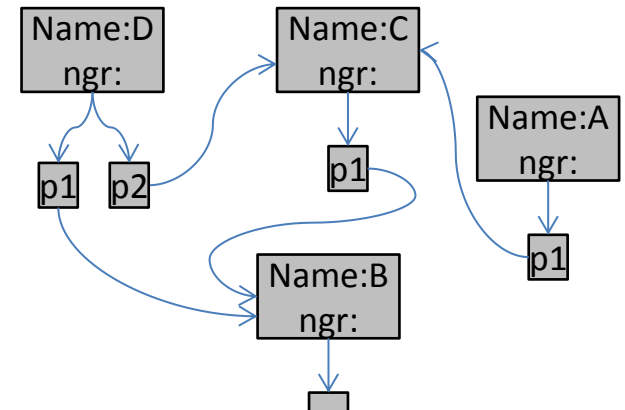
(ordered) pairs of nodes

[(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?

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

