Self Similar (Scale Free, Power Law) Networks (I)

E6083: lecture 4

Prof. Predrag R. Jelenković

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
{predrag}@ee.columbia.edu

February 7, 2007

Outline

Outline

Introduction

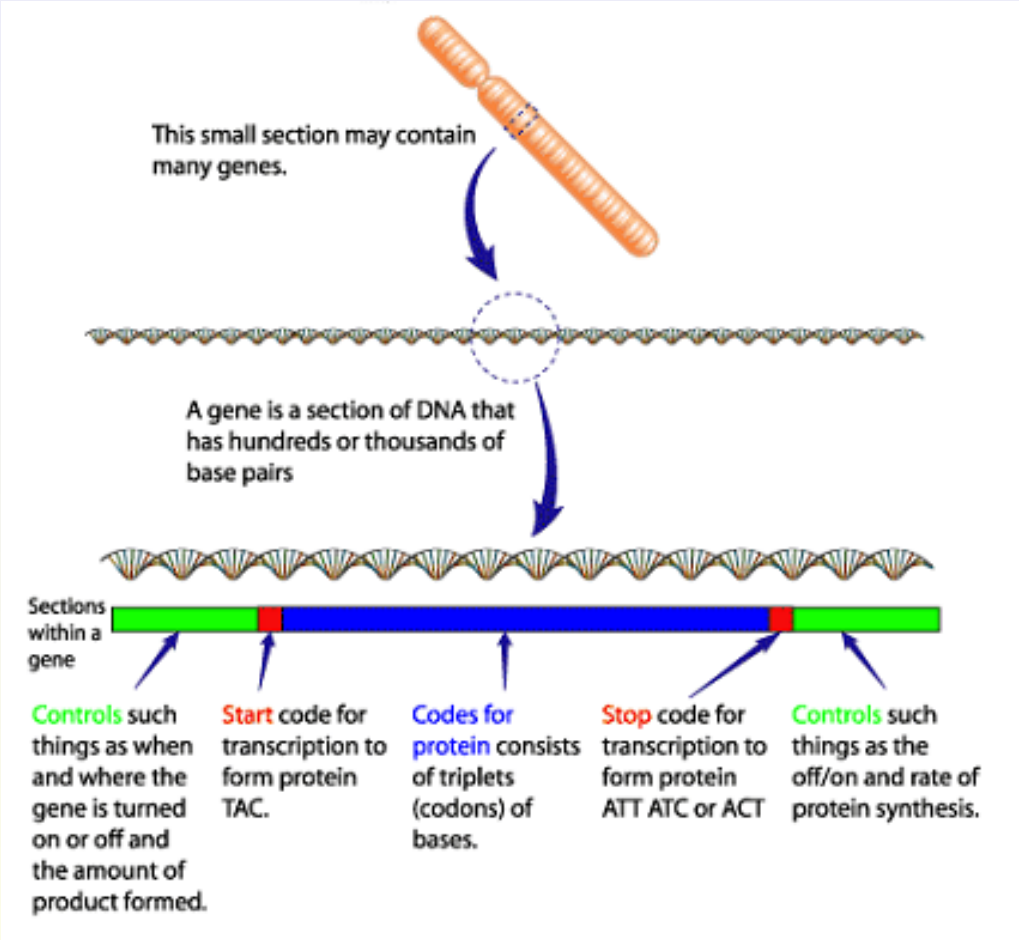
Network symbols

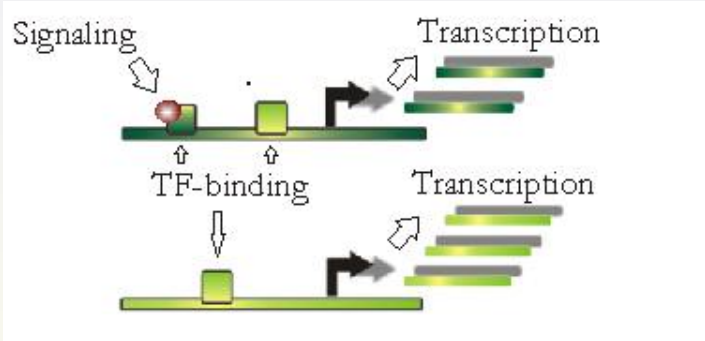
Nodes: biological objects (proteins, genes)
Edges: interaction between nodes

Examples

Network	node	Edges
Metabolic networks	metabolites	interaction
Transcriptional interactions	genes	regulation
Protein folding networks	residue	folding neighbors

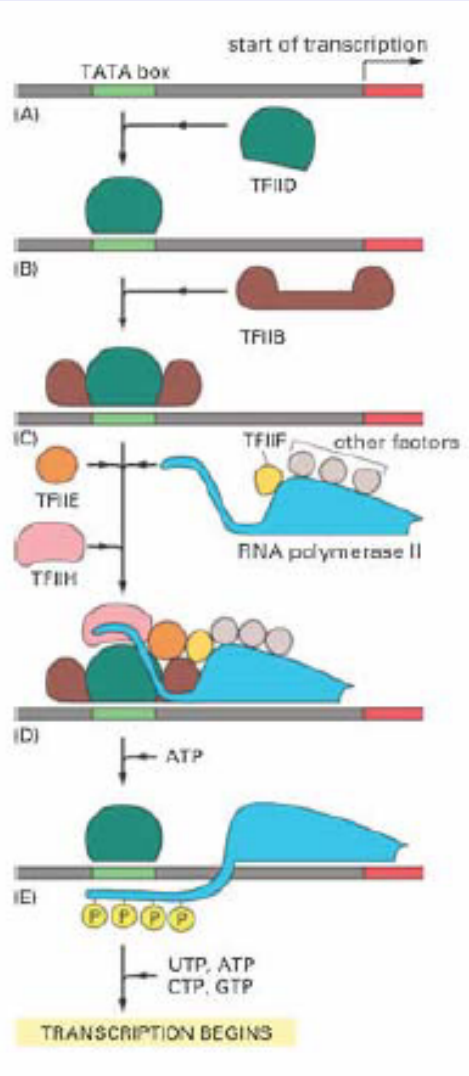
DNA & Genes


This small section may contain many genes.
A gene is a section of DNA that has hundreds or thousands of base pairs
Sections within a gene
Controls such things as when and where the gene is turned on or off and the amount of product formed.
Start code for transcription to form protein TAC.
Codes for protein consists of triplets (codons) of bases.
Stop code for transcription to form protein ATT ATC or ACT
Controls such things as the off/on and rate of protein synthesis.


Signaling
Transcription
TF-binding
Transcription

Gene Transcription Regulation

- The transcription rate is controlled by the promoter.
- Transcription Factors (TF), including activators and repressors, binds the sites in promoter.
- TFs are regulated by other TFs, and form a network.
- TFs are encoded in genes.



Gene Regulatory Network

Cell's gene regulatory network refers to the coordinated on and off switching of genes by regulatory proteins that bind to non-coding DNA.

How to discover edges?

Most work in this area has focused on reconstructing the network from data/experiments, for example, find the correlation function ρ of the number of proteins, the hypothesis is that if two genes are positively/negatively regulated, then ρ is close to ± 1, meaning, A appears with high probability if B is present, then... Also, some researchers use mutual information as a measure of gene "closeness".

Outline

http://www.biochemj.org/bj/381/0001/bj3810001.htm

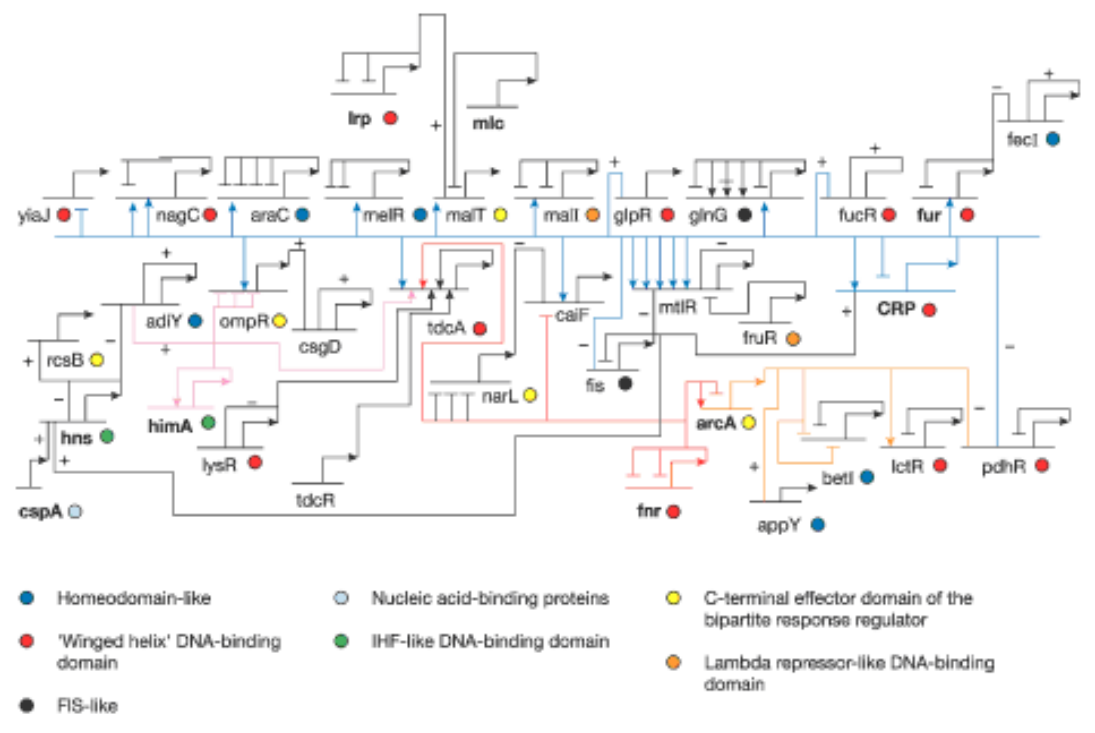


Figure: Regulatory network of transcription factors (TFs) in E. coli.

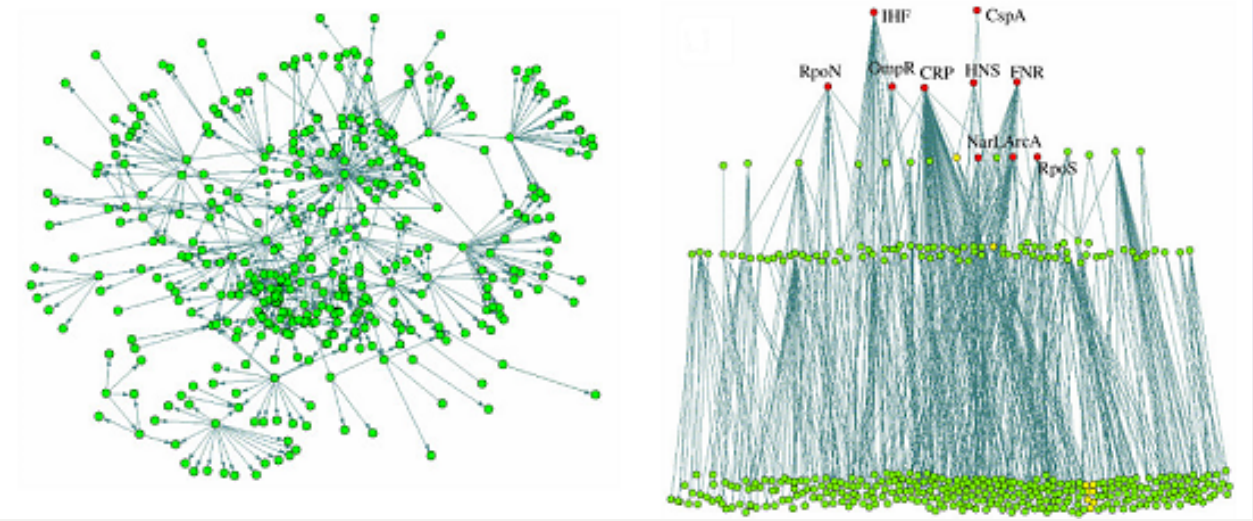


Figure: Hierarchical structure and modules in the E. coli transcriptional regulatory network

The original unorganized network vs. the hierarchical regulation structure. Nodes in the graph are operons. Links show the transcriptional regulatory relationships. The global regulators found in this work are shown in red.

Operons in different modules are shown in different colors. The ten global regulators form the core part of the network. The periphery modules are connected mainly through the global regulators. Depending on the connectivity between the modules and their connectivity to the global regulators, these modules can be further grouped to larger modules at a higher level.

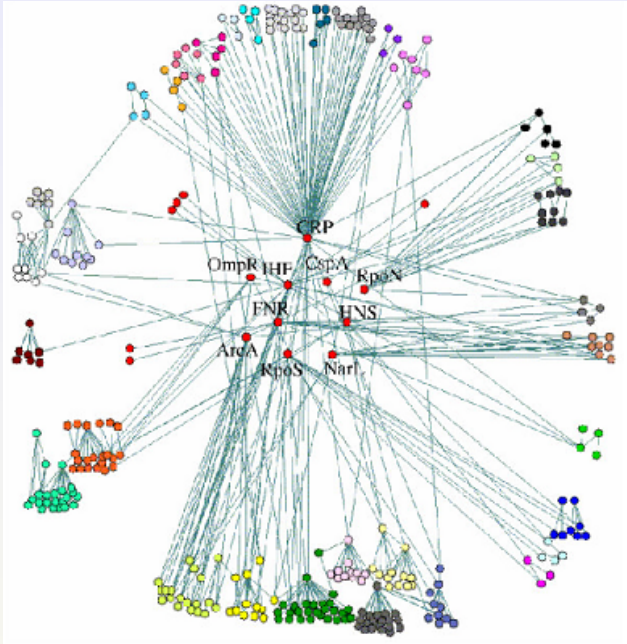
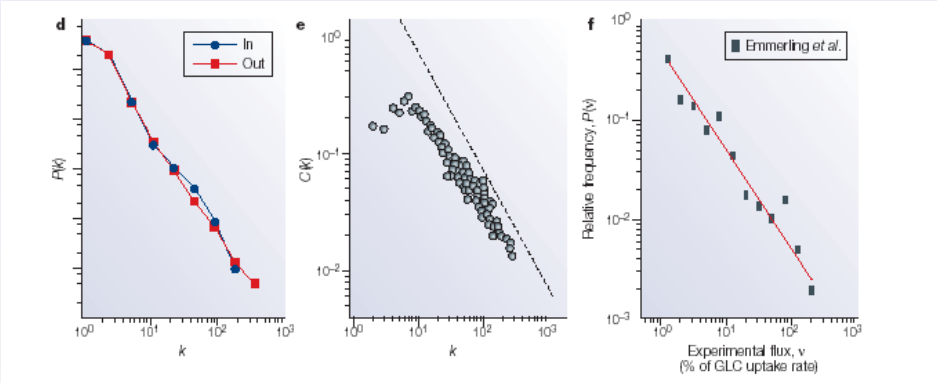


Figure: Functional modules in the transcriptional regulatory network of E. coli

Characterizing metabolic networks of E. Coli

Network biology (Barabasi & Oltvai, Nature, 2004)
(d) The degree distribution, $P(k)$ of the metabolic network illustrates its scale-free topology.
(e) The scaling of the clustering coefficient $C(k)$ (defined later) with the degree k illustrates the hierarchical architecture of metabolism.



(f) The flux distribution in the central metabolism of E. Coli follows a power law, which indicates that most reactions have small metabolic flux, whereas a few reactions, with high fluxes, carry most of the metabolic activity.

Questions

- What is the topology of this network?
- Are there basic structures (subgraphs/subnetworks, motifs)?
- How do we model the operations of regulatory networks? (analogy circuits: gates, logic?)
- How does evolution change regulatory networks? Impact of natural selection (fitness), motifs..
- Resilience to attacks (targeted or random), disease, etc.
- We could have a whole course on gene regulatory networks (Spring 2008).

Power Law Random Graph– Scale Free Network

The observations of power-law distributions in the connectivity of complex networks came as a surprise to researchers deeply rooted in the tradition of random networks.

Traditional random graph - Erdos Renyi model VS Scale Free Network - Barabási model

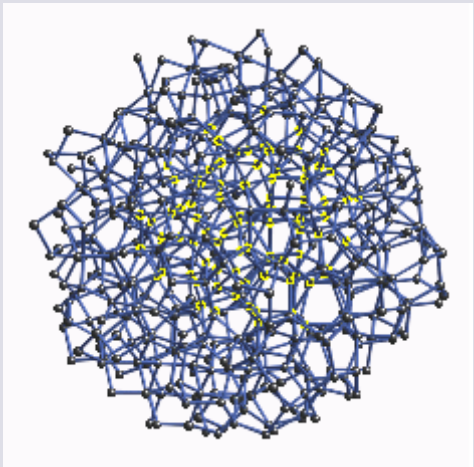


Figure: Concentrated Degree distribution: $\approx$ Poisson

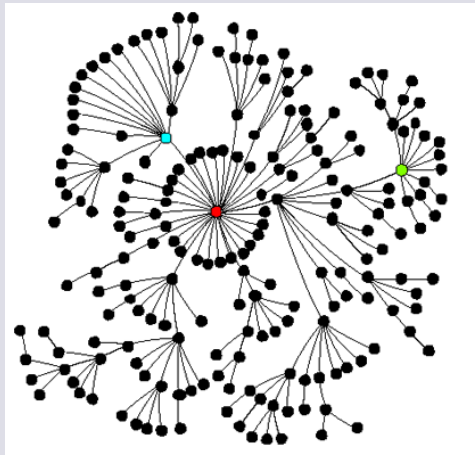


Figure: Power Law Degree distribution

Outline

Introduction to Erdös Rényi Model

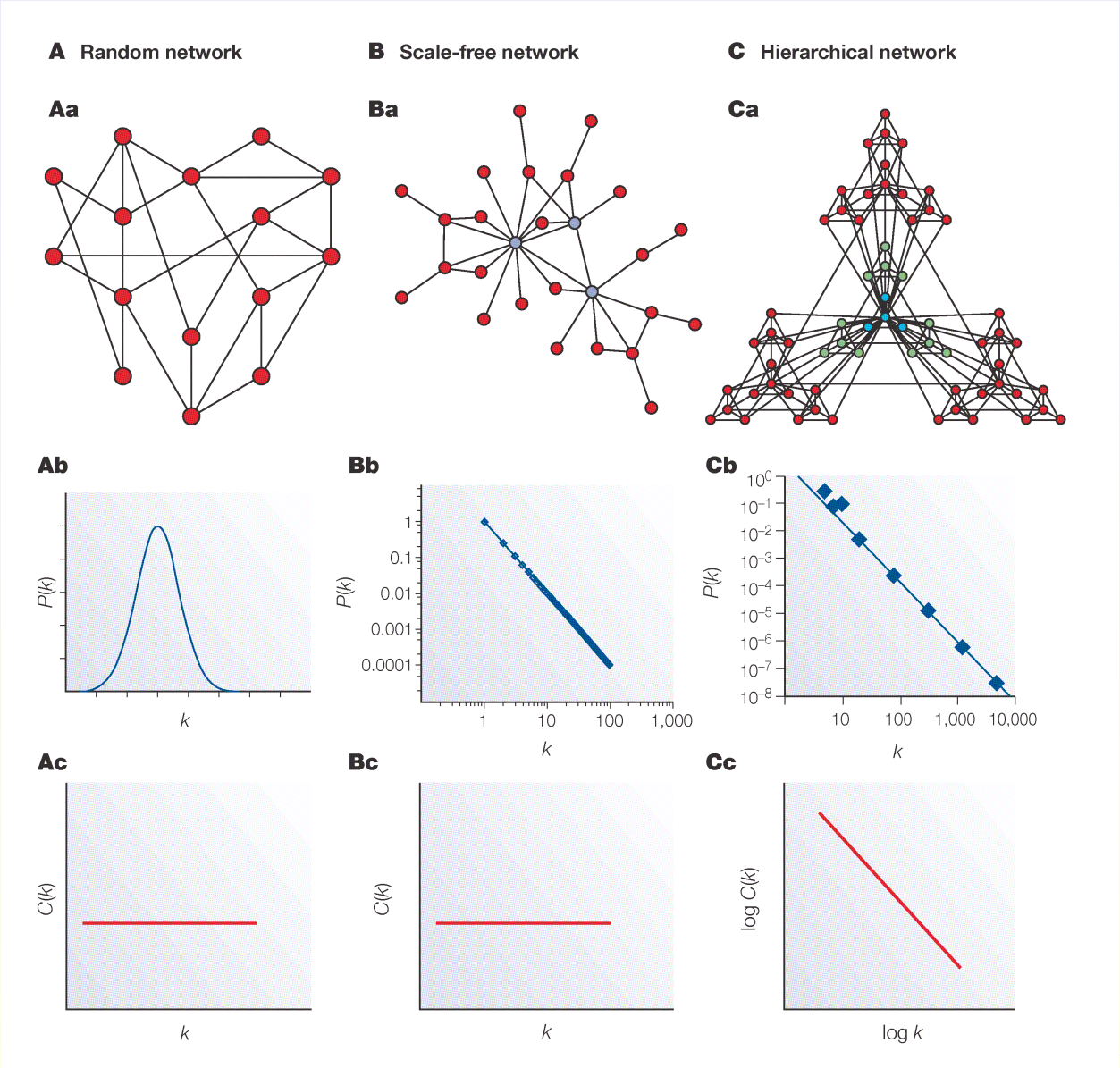
- $G(n,p)$ is a graph with n nodes where an edge has probability p to be selected.
- Average degree $d = \mathbb{E}D = p(n-1) \approx pn$;
  $\mathbb{P}[D = k] = \binom{m}{k} p^k (1-p)^{m-k} \approx (d^k/k!)exp(-d)$.
  Sharply concentrated around its mean, i.e., Poisson-like.
- Percolation transition, threshold behavior at d = 1.
- If $d < 1$, then with high probability the network is forming mostly trees and no component is larger than $\log n$.
- If $d > 1$, there is a unique "giant component".

The Clustering Coefficient of a Network

- Let $N(u)$ denote the set of neighbors of u in a graph: $N(u) = \{v : (u, v) \in G\}$.
- The clustering coefficient of u:
  let $k =\mid N(u) \mid$ (i.e., the number of neighbors of u);
  $\binom{k}{2}$ = max possible # of edges between vertices in $N(u)$;
  $c(u)$ = (actual # of edges between vertices in $N(u)$)/$\binom{k}{2}$.
- $0 \leq c(u) \leq 1$; measure of cliquishness of u's neighborhood.
- Clustering coefficient of a graph: average of $c(u)$ over all vertices.

Real networks often have high clustering $C_{real} \gg C_{rnd}$.

Main Network Types



Internet Topology (Michalis Faloutsos, Petros Faloutsos & Christos Faloutsos 1999)

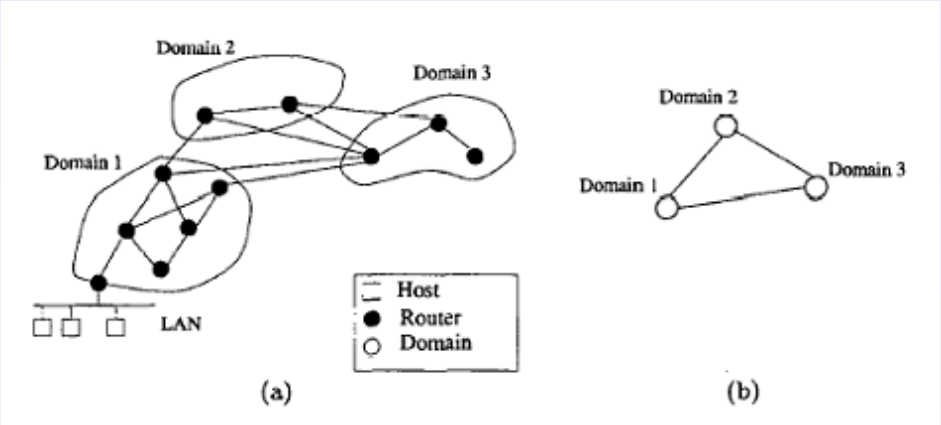


Figure: The structure of Internet at a) the router level and b) the inter-domain level. The hosts connect to routers in LANs.

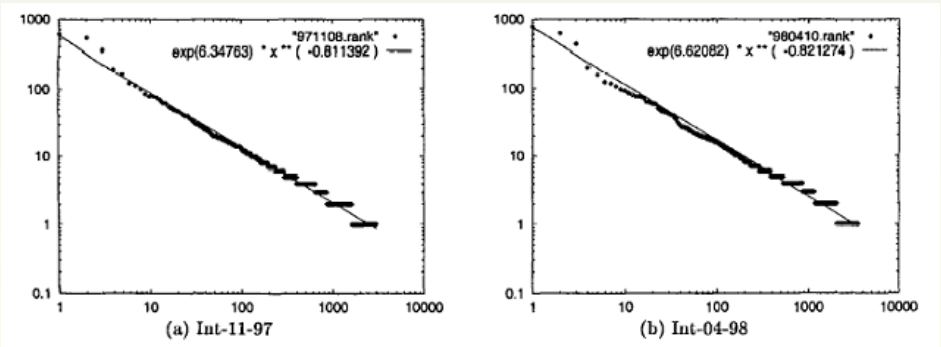


Figure: Log-log plot of the outdegree d, versus the rank in the sequence of decreasing outdegree. Data in Nov 97 and April 98.

Internet Topology

- The outdegree, indegree distribution follow power laws.
- The total number of pairs of nodes within h hops follow power laws.
- The eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ follow power laws. (the eigenvalues of a graph are closely related to many basic topological properties such as the diameter, the number of edges, the number of spanning trees, the number of connected components...)

Outline

Many Natural Networks

- A heavy-tailed degree distribution: a small but distinctive number of high-degree vertices serve as hubs.
- Few connected components: often only 1 or a small number independent of network size
- Small diameter: often growing only logarithmically with network size
- A high degree of clustering

Mechanisms resulting power law random graph

Preferential Attachment: the rich get richer

As new connections form, they attach to a node with a probability proportional to the existing number of connections (growth and preferential attachment).

Copying Models

The linear growth copying model was introduced by Kleinberg et al. in 1999.

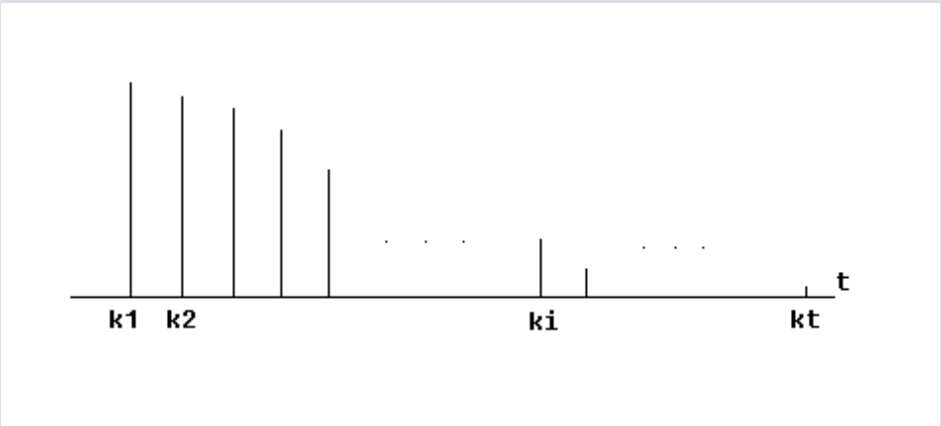
New mechanism: generalized random walk (GRW)

The evolvement of large scale systems (e.g., self-assemble DNA network, Internet, social network) is attributed to rules lying into two categories: global information and local information.
(The preceding copying model can be viewed as a special case of random walk attachments.)

Preferential Attachment

Informal derivation

At time 0, one node is present, and at each step $t++$, a new vertex is added, with one undirected edge preferentially ($\sim k_i$) attached to one existing node. Assume that vertex i was added to the system at time t_i.



Then, at time t,

$$\frac{dk_i}{dt} = \frac{k_i}{2t} \Rightarrow k_i = \left(\frac{t}{t_i}\right)^{1/2}.$$

Preferential Attachment: Continued...

Thus, the degree D distribution is obtained by

$$\mathbb{P}[D > x] = \mathbb{P}[t_i < t/x^2] = \frac{t}{x^2(t+1)},$$

which implies $\mathbb{P}[D = x] \sim \frac{1}{x^3}$.

A rigorous analysis of preferential attachment was first given by Bollobás et al. Let the number of vertices of G_n with indegree equal to d be $X_n(d)$, and consider G_n as one graph from the process $\{G_t : 0 \leq t \leq n\}$. The martingale

$$X_t = \mathbb{E}[X_n(d) \mid G_t]$$

satisfies that $\mid X_{t+1} - X_t \mid$ is bounded by two.
Applying Azuma-Hoeffding inequality, we obtain that $X_n(d)$ is very concentrated around its mean, and thus only need to compute $\mathbb{E}[X_n(d)]$.

Preferential Attachment: Continued...

What if attaching proportional to k_i^α?

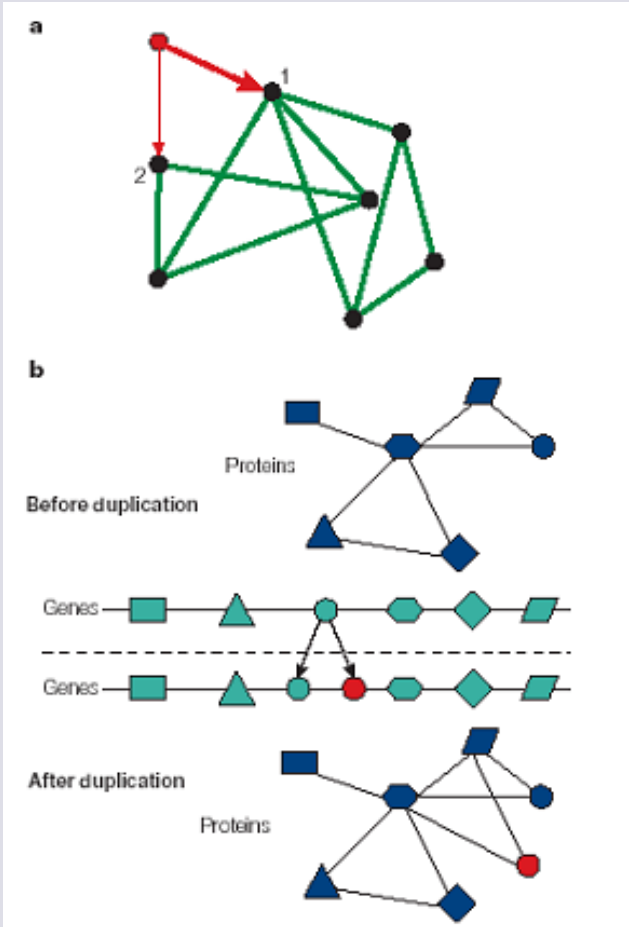
- If $\alpha > 1$, eventually one person gets all the links.
  There is a finite time after which no one else gets anything!
- If $\alpha < 1$, the degree distribution follows a stretched exponential.

Limitation of preferential attachment

- Global information.
- Number of nodes increases linearly.

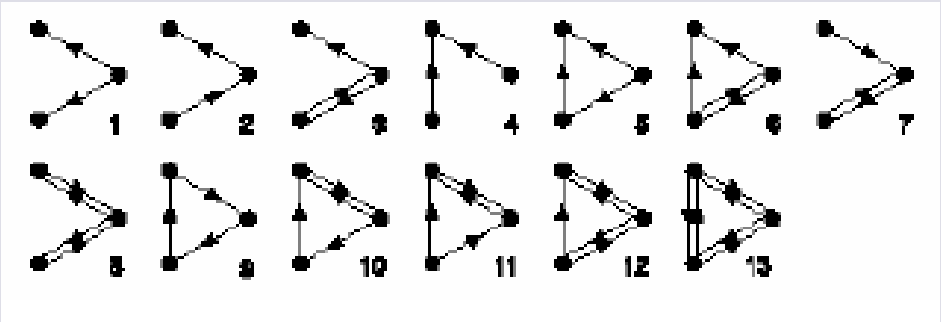
The origin of the scale-free topology in biological networks

The new protein has the same structure as the old one, so they both interact with the same proteins. Therefore proteins with a large number of interactions tend to gain links more often, as it is more likely that they interact with the protein that has been duplicated.



Discovering motifs

Motifs are those patterns which occur significantly more frequently in real than in equivalent randomized networks.



Look for all possible two- or three-node configurations.

Yeast Regulatory Network Motifs

Lee et al, Science 2002

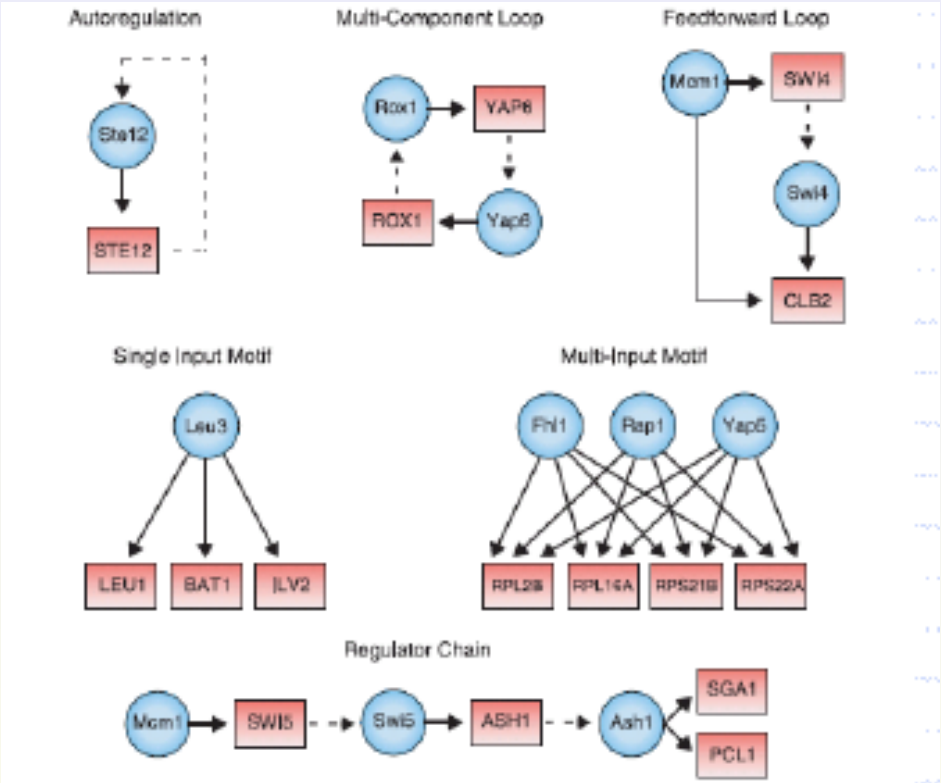


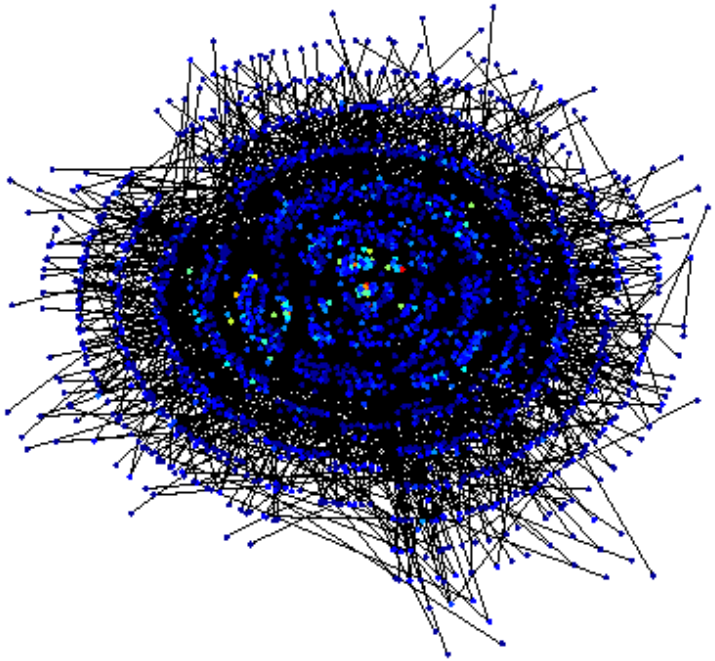
Fig. 3. Examples of network motifs in the yeast regulatory network. Regulators are represented by blue circles; gene promoters are represented by red rectangles. Binding of a regulator to a promoter is indicated by a solid arrow. Genes encoding regulators are linked to their respective regulators by dashed arrows. For example, in the autoregulation motif, the Ste12 protein binds to the *STE12* gene, which is transcribed and translated into Ste12 protein. These network motifs were uncovered by searching binding data with various algorithms. For details on the algorithms used and a full list of motifs found, see [78].

Motifs Of The Yeast Protein Network

S. Wuchty, Z. Oltvai & A.-L. Barabasi, Nature Genetics, 2003

#	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

Scale free network caused by random walk



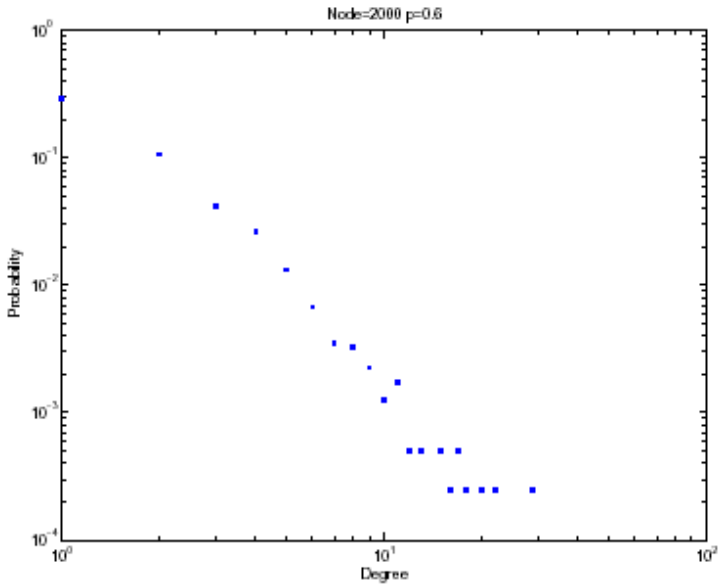


Figure: Node=2000, Random Walk $p = 0.6$.