

Structure of biological networks

Outline

- Review of Lecture 1
- Representations
- Basic concepts
- Real vs random networks
- Biological questions

Representation

- Graphs, digraphs, weighted graphs
- Bipartite graphs
- Hyper-graphs
- Stochiometry matrices

Graphs

$$G=(N,E)$$

$$N=\{n_1, n_2, \dots, n_N\}$$

$$E=\{e_1, e_2, \dots, e_M\}$$

$$e_k=\{n_i, n_j\}$$

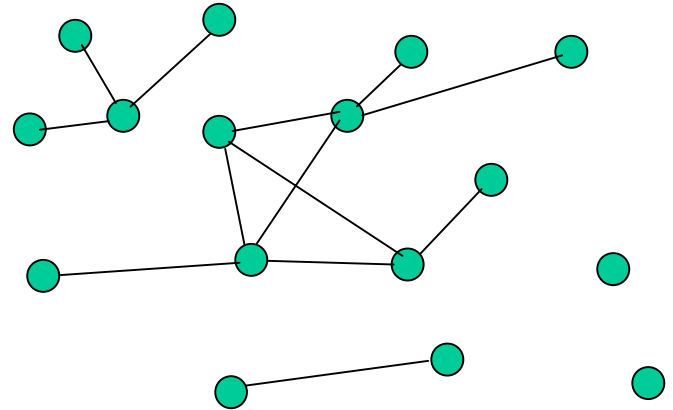
Constraints:

- no duplicate edges
- no loops etc

E.g.

Nodes: proteins

Edges: interactions



Mutligraph

$$G=(N,E)$$

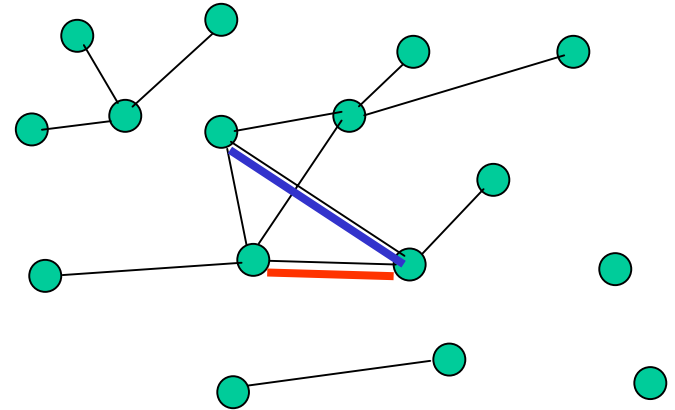
$$N=\{n_1 n_2, \dots n_N\}$$

$$E=\{e_1 e_2, \dots e_M\}$$

$$e_k=\{n_i n_j\}$$

+ duplicate edges

i.e. $e_m=\{n_i n_j\}$



E.g.

Nodes: proteins

Edges: interactions of different sort:

binding and similarity

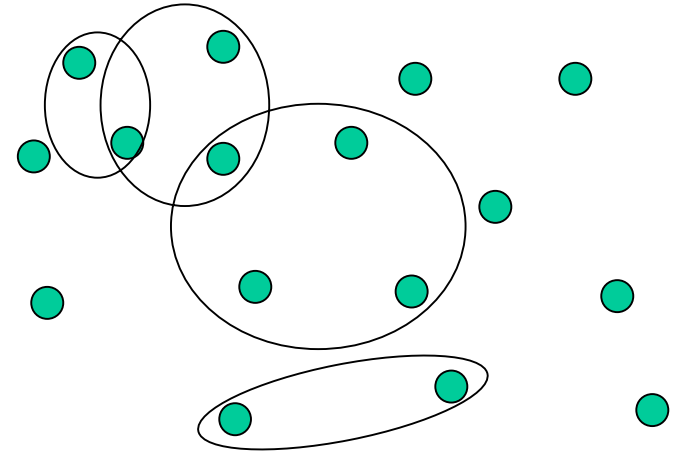
Hypergraphs

$$G=(N,E)$$

$$N=\{n_1, n_2, \dots, n_N\}$$

$$E=\{e_1, e_2, \dots, e_M\}$$

$$e_x=\{n_i, n_j, n_k, \dots\}$$



Constraints:

- no duplicate edges
- no loops etc

E.g.

Nodes: proteins

Edges: protein complexes

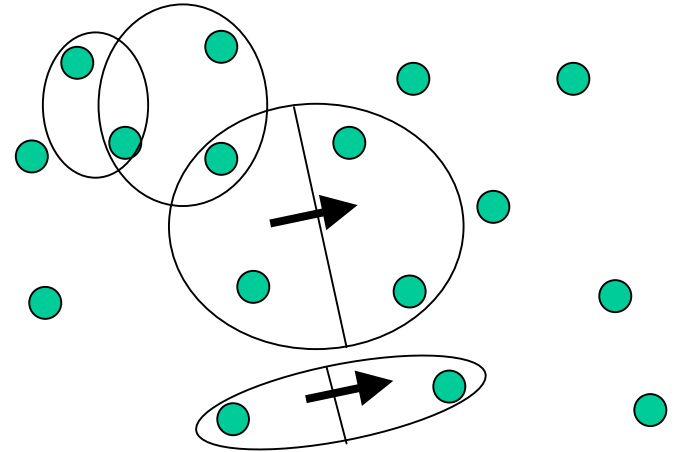
Directed Hypergraphs

$$G=(N,E)$$

$$N=\{n_1, n_2, \dots, n_N\}$$

$$E=\{e_1, e_2, \dots, e_M\}$$

$$e_x = \{n_i, n_j \dots \mid n_k, n_l \dots\}$$



E.g.

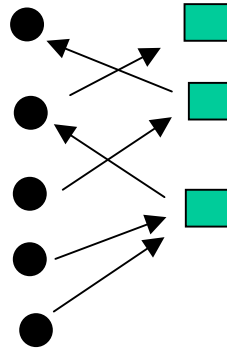
Nodes: substances

Edges: chemical reactions $A + B \rightarrow C + D$

$$X = \{A, B \dots \mid C, D \dots\}$$

Bipartite graphs

substances reactions



Bipartite graphs

substances reactions

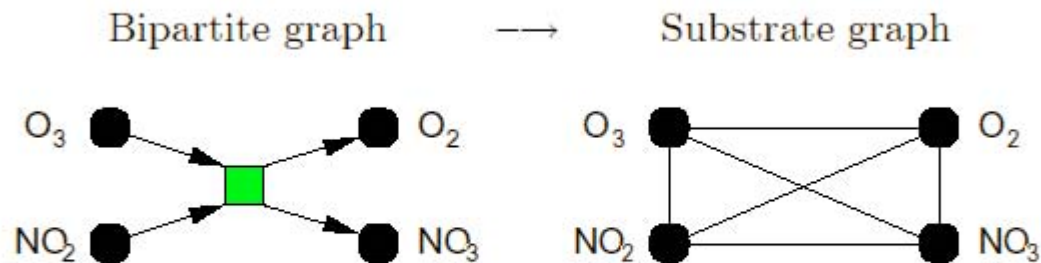
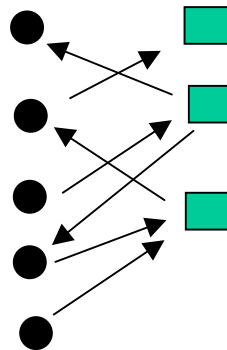


Fig. 2. One-mode projection from a bipartite graph to a substrate graph.

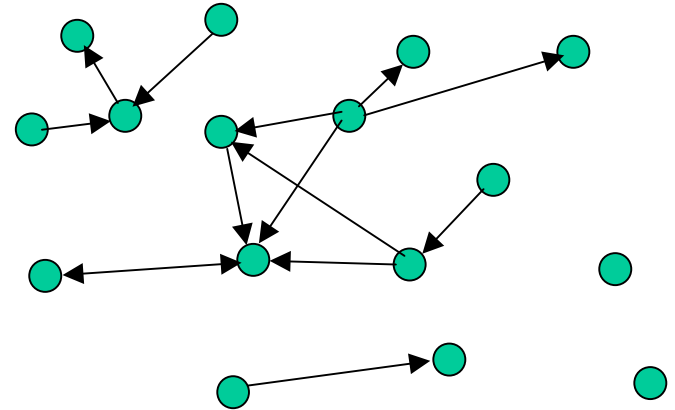
Directed Graphs

$$G=(N,E)$$

$$N=\{n_1, n_2, \dots, n_N\}$$

$$E=\{e_1, e_2, \dots, e_M\}$$

$$e_k=\{n_i, n_j\}$$



Constraints:

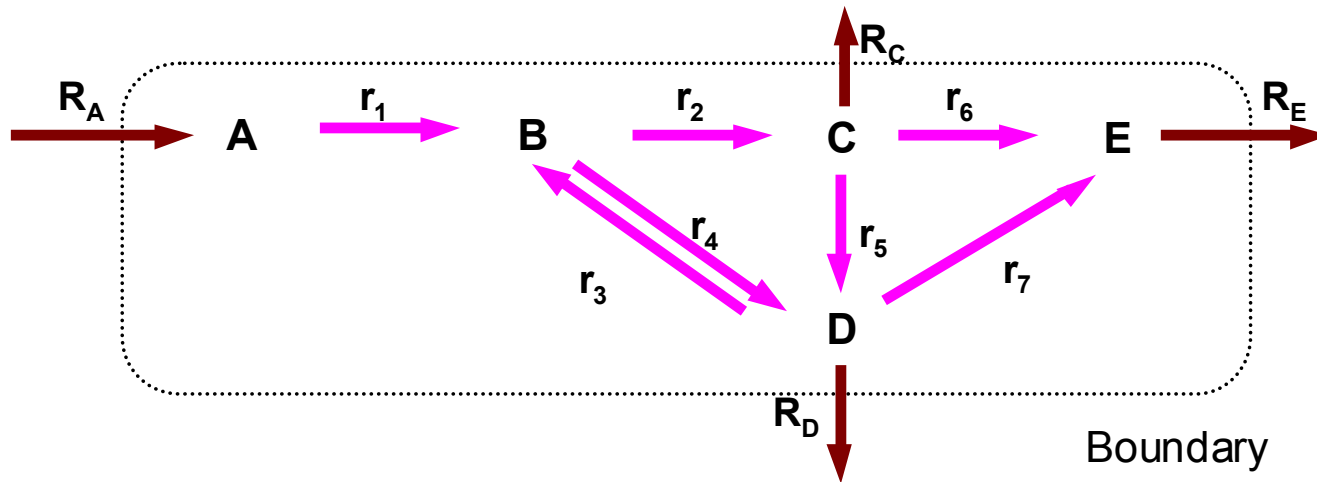
- ~~no duplicate edges~~
- no loops etc

E.g.

Nodes: genes and their products

Edges from A to B: gene A regulates expression of gene B

Flux Balance Analysis



Steady state Mass Balance

$$\text{A: } -r_1 = -R_A$$

$$\text{B: } -r_1 + r_4 - r_2 - r_3 = 0$$

$$\text{C: } +r_2 - r_5 - r_6 = +R_C$$

$$\text{D: } +r_3 + r_5 - r_4 - r_7 = +R_D$$

$$\text{E: } +r_6 + r_7 = +R_E$$

$$S \cdot v = b$$



$$\begin{bmatrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Internal fluxes

$$= \begin{bmatrix} -R_A \\ 0 \\ R_C \\ R_D \\ R_E \end{bmatrix}$$

Transportation fluxes

Summary

- Different problems -> different graphs

Basic concepts

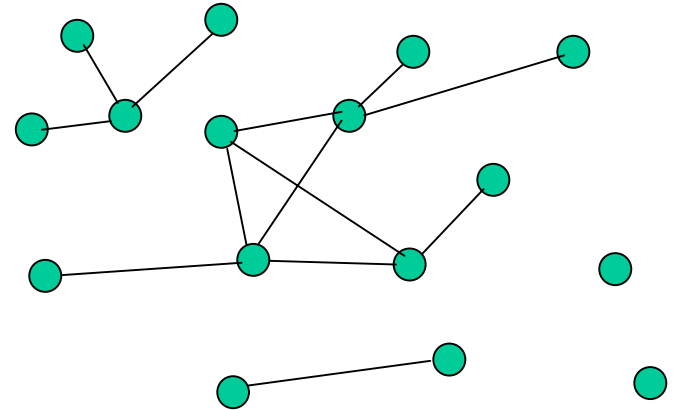
- Degree, components, path length, diameter
- Clustering coefficient
- Centrality and betweenness
- Random graphs: giant component, component size distribution

Basic concepts

Nodes: proteins

Edges: interactions

- Node degree
- Components
- Shortest path
- Diameter=<Shortest Path>

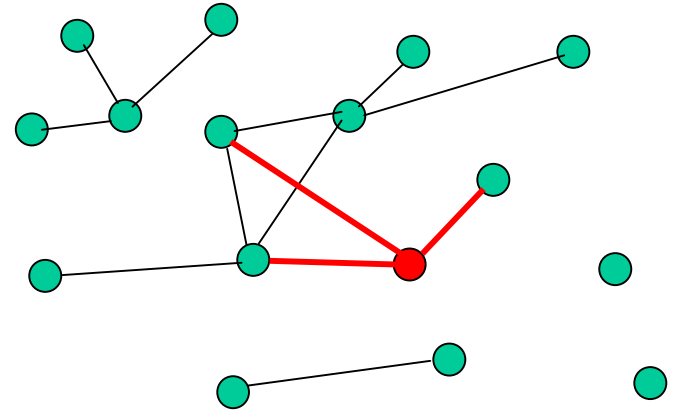


Basic concepts

Nodes: proteins

Edges: interactions

- Node degree $d(k)$
- Components
- Shortest path
- Diameter = <Shortest Path>

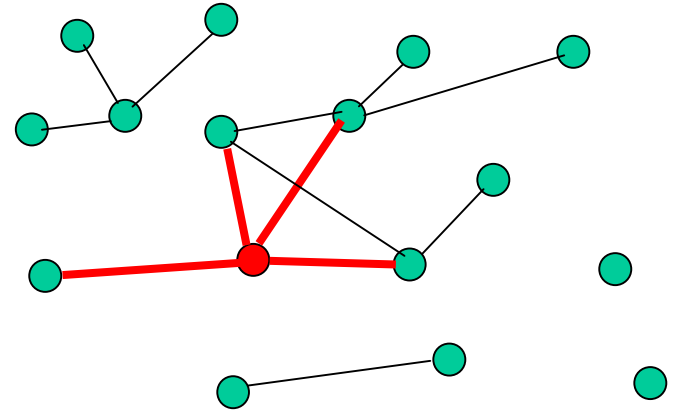


Basic concepts

Nodes: proteins

Edges: interactions

- Node degree $d(k)$
- Components
- Shortest path
- Diameter = <Shortest Path>

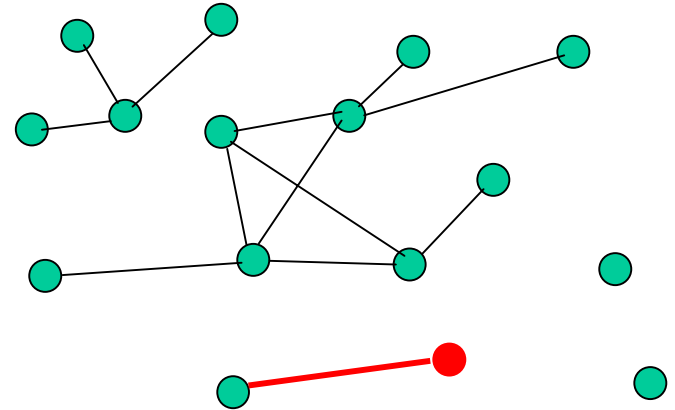


Basic concepts

Nodes: proteins

Edges: interactions

- Node degree $d(k)$
- Components
- Shortest path
- Diameter = <Shortest Path>



Basic concepts

Nodes: proteins

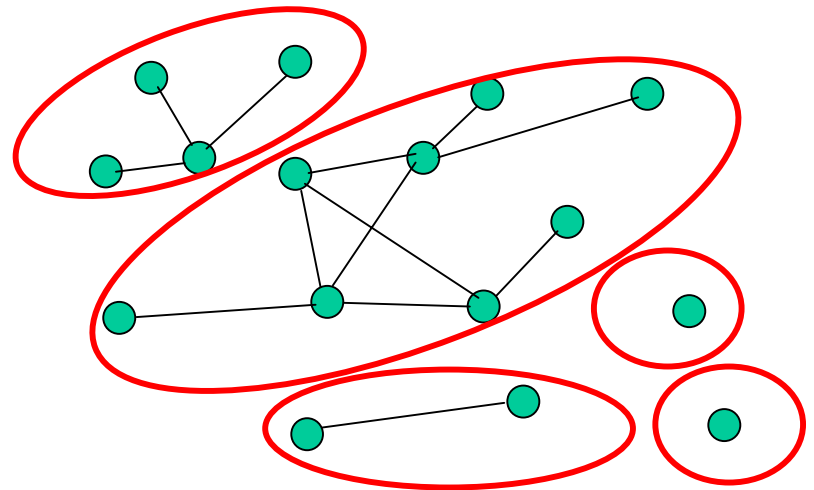
Edges: interactions

- Node degree

- **Components**

- Shortest path

- Diameter=<Shortest Path>



Basic concepts

Nodes: proteins

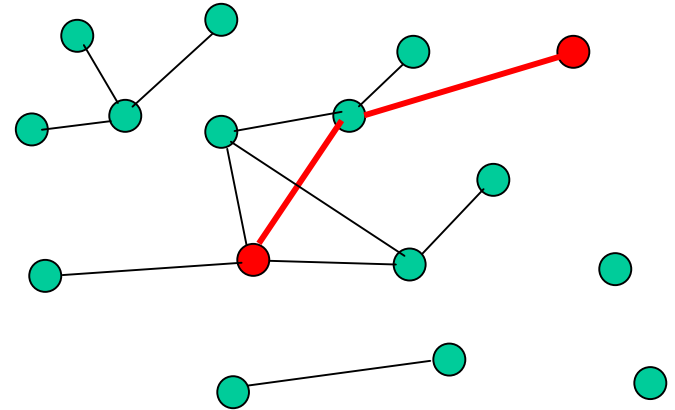
Edges: interactions

- Node degree

- Components

- Shortest path

- Diameter=<Shortest Path>



Basic concepts

Nodes: proteins

Edges: interactions

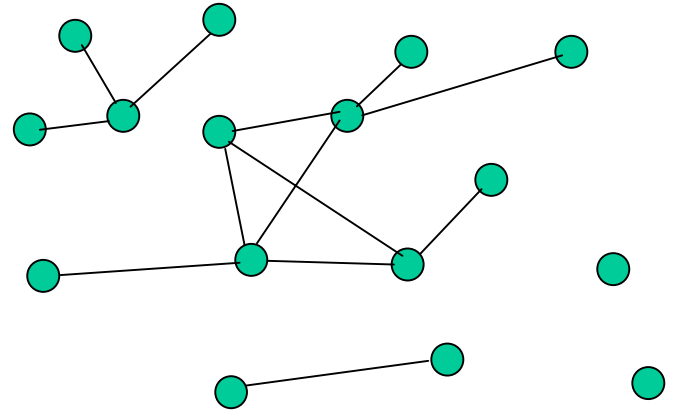
- Node degree

- Components

- Shortest path l_{ij}

- Diameter=<Shortest Path>

$$l = \frac{2}{N(N-1)} \sum_{i,j=1..N} l_{ij}$$

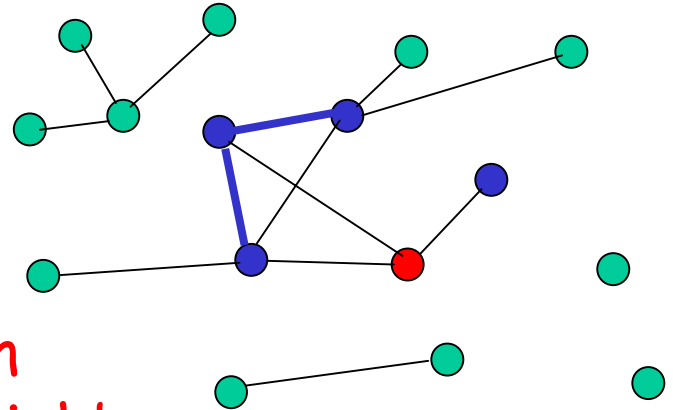


Basic concepts

- Clustering coefficient of node i :
 $C(i) = 2E(i) / (k(i)(k(i)-1))$

= fraction of realized interaction
between the neighbors
= # of triangles that include i / $(k(i)(k(i)-1))$

- Clustering coefficient of graph G :
 $C = \frac{1}{N} \sum_{i=1..N} C(i)$



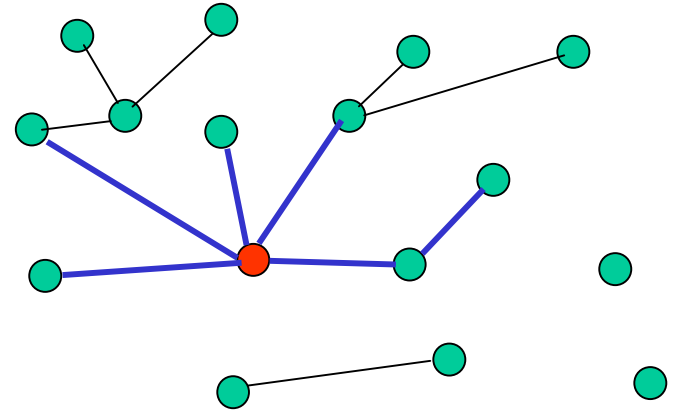
Basic concepts

"Centrality"

- **Betweenness**

- of node k

$$g(k) = \sum_{i,j=1..N} \frac{s_{ij}(k)}{s_{ij}}$$



where s_{ij} is the number of SHORTEST paths from i to j

$s_{ij}(k)$ the number of such paths that go through k

- edge m

$$g(m) = \sum_{i,j=1..N} \frac{s_{ij}(m)}{s_{ij}}$$

COMPARE REAL NETWORKS WITH RANDOM ONES

in all these parameters

WHAT IS A RANDOM GRAPH ?

- Wagner et al 2000
- Barabasi et al 2000
- AB model
- Robustness

ON THE EVOLUTION OF RANDOM GRAPHS

by

P. ERDÖS and A. RÉNYI

*Institute of Mathematics
Hungarian Academy of Sciences, Hungary*

1. Definition of a random graph

Let $E_{n,N}$ denote the set of all graphs having n given labelled vertices $V_1, V_2, \dots, V_n$ and N edges. The graphs considered are supposed to be not oriented, without parallel edges and without slings (such graphs are sometimes called linear graphs). Thus a graph belonging to the set $E_{n,N}$ is obtained by choosing N out of the possible $\binom{n}{2}$ edges between the points $V_1, V_2, \dots, V_n$, and therefore the number of elements of $E_{n,N}$ is equal to $\binom{\binom{n}{2}}{N}$. A random graph $\Gamma_{n,N}$ can be defined as an element of $E_{n,N}$ chosen at random, so that each of the elements of $E_{n,N}$ have the same probability to be chosen, namely $1/\binom{\binom{n}{2}}{N}$. There is however an other slightly different point of view, which has some advantages. We may consider the formation of a random graph as a stochastic process defined as follows: At time $t=1$ we choose one out of the $\binom{n}{2}$ possible edges connecting the points $V_1, V_2, \dots, V_n$, each of these edges having the same probability to be chosen; let this edge be denoted by e_1 . At time $t=2$ we choose one of the possible $\binom{n}{2} - 1$ edges, different from e_1 , all these being equiprobable. Continuing this process at time $t=k+1$ we choose one of the $\binom{n}{2} - k$ possible edges different from the edges $e_1, e_2, \dots, e_k$ already



Paul Erdős and A. Rényi,

On the evolution of random graphs

Magyar Tud. Akad. Mat. Kut. Int. Kozl. 5 (1960) 17-61.

Random graphs

- Erdos-Reyni
 - start with N nodes
 - pick two at random and connect them
(exclude loops and double edges)
 - continue till M edges are formed

Degree: $\langle k \rangle = 2M/N$

k distribution: Poisson

diameter l : $l \sim \log N$

clustering coef: $C \sim 0$

Random graphs

- Erdos-Reyni

$$p=2M/(N(N-1)) \quad c=2M/N$$

18.1 Theorem 5.4 of [1]

Theorem 1 Consider an Erdős-Renyi random graph $\mathcal{G}(n, p)$. Set $p = c/n$, fixed $c > 0$.

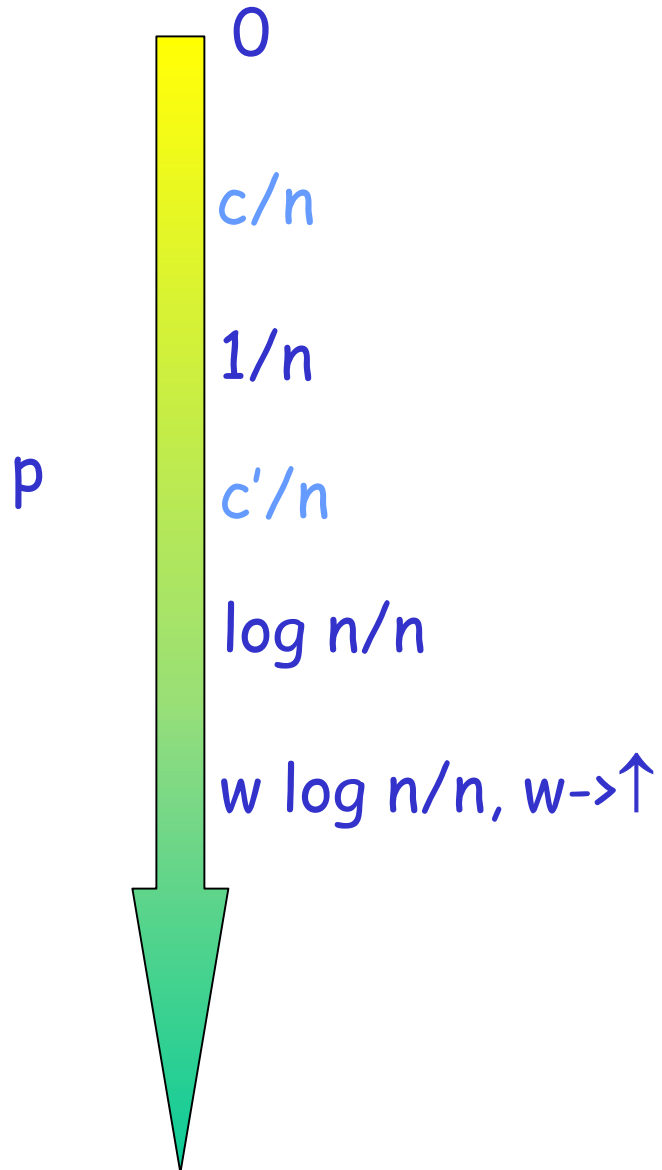
Study $n \rightarrow \infty$ limits.

Assertions hold with prob $\rightarrow 1$ as $n \rightarrow \infty$.

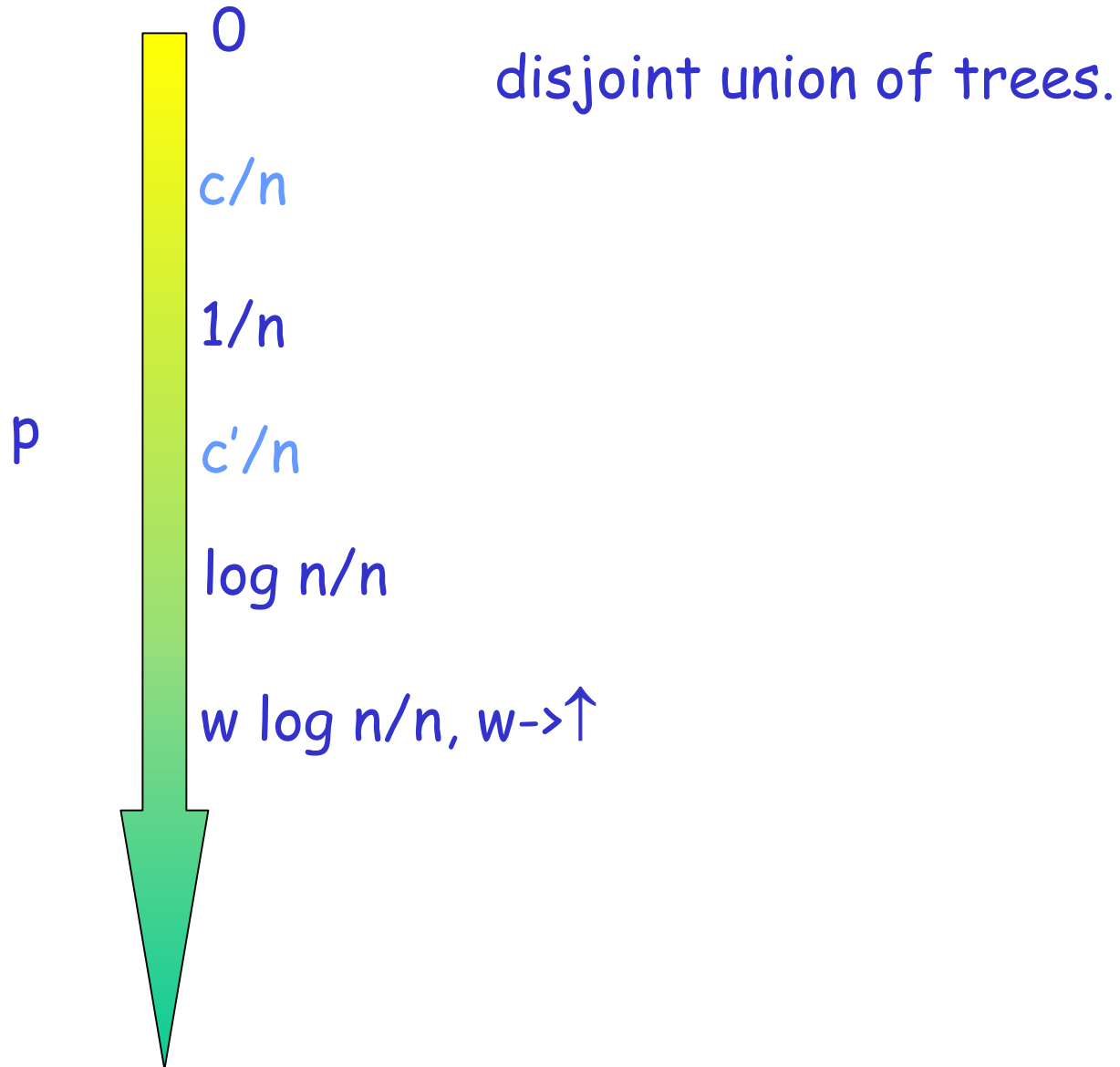
(a) If $c < 1$ the largest component size is at most $\frac{3}{(1-c)^2} \log n$.

(b) If $c > 1$ there is a unique giant component with $(1 + o(1))\beta n$ vertices, where $\beta = \beta(c) > 0$ solves $\beta + e^{-\beta c} = 1$. The second largest component size is at most $\frac{16c}{(c-1)^2} \log n$.

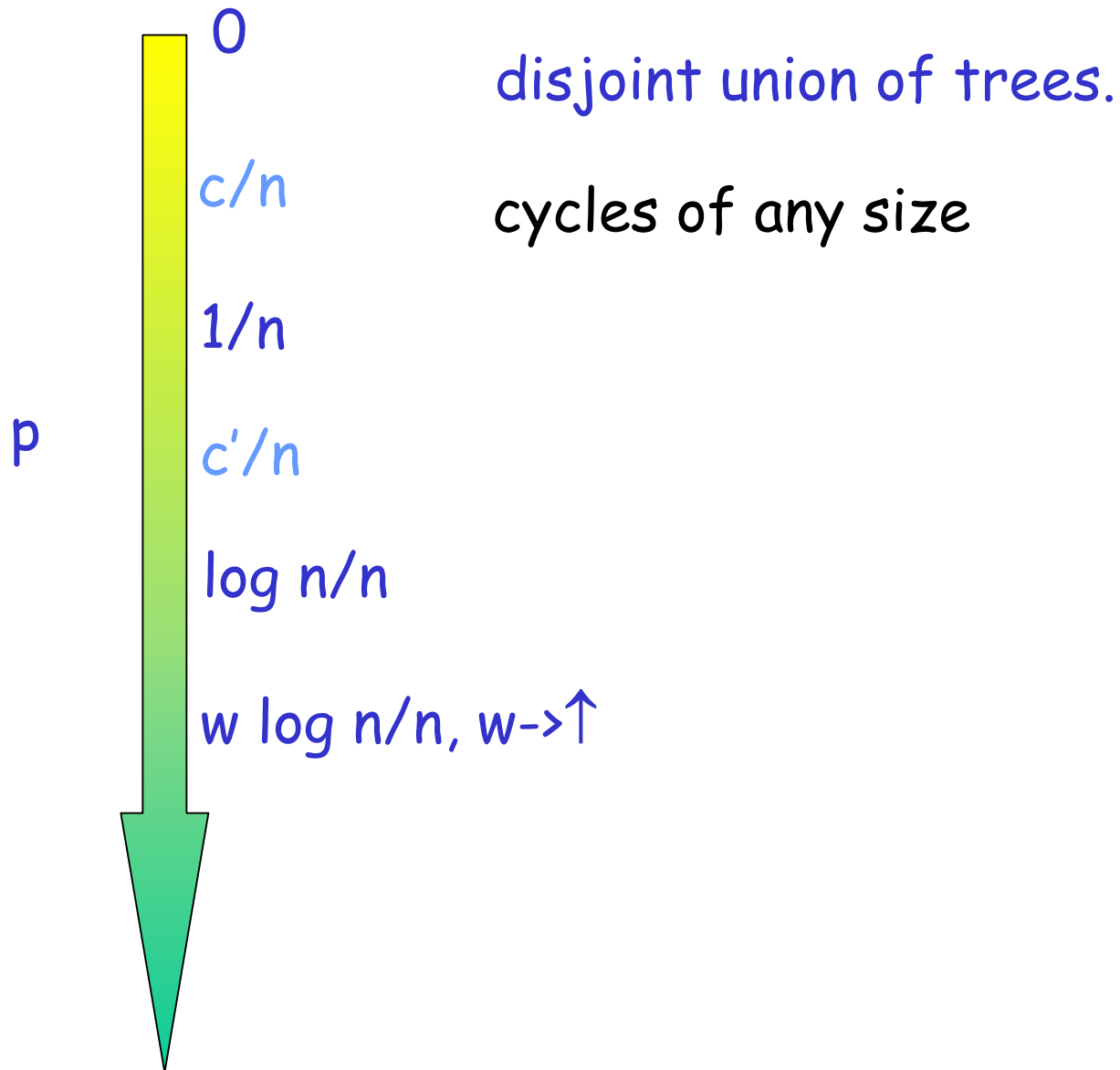
The evolution of random graphs $G(n,p)$



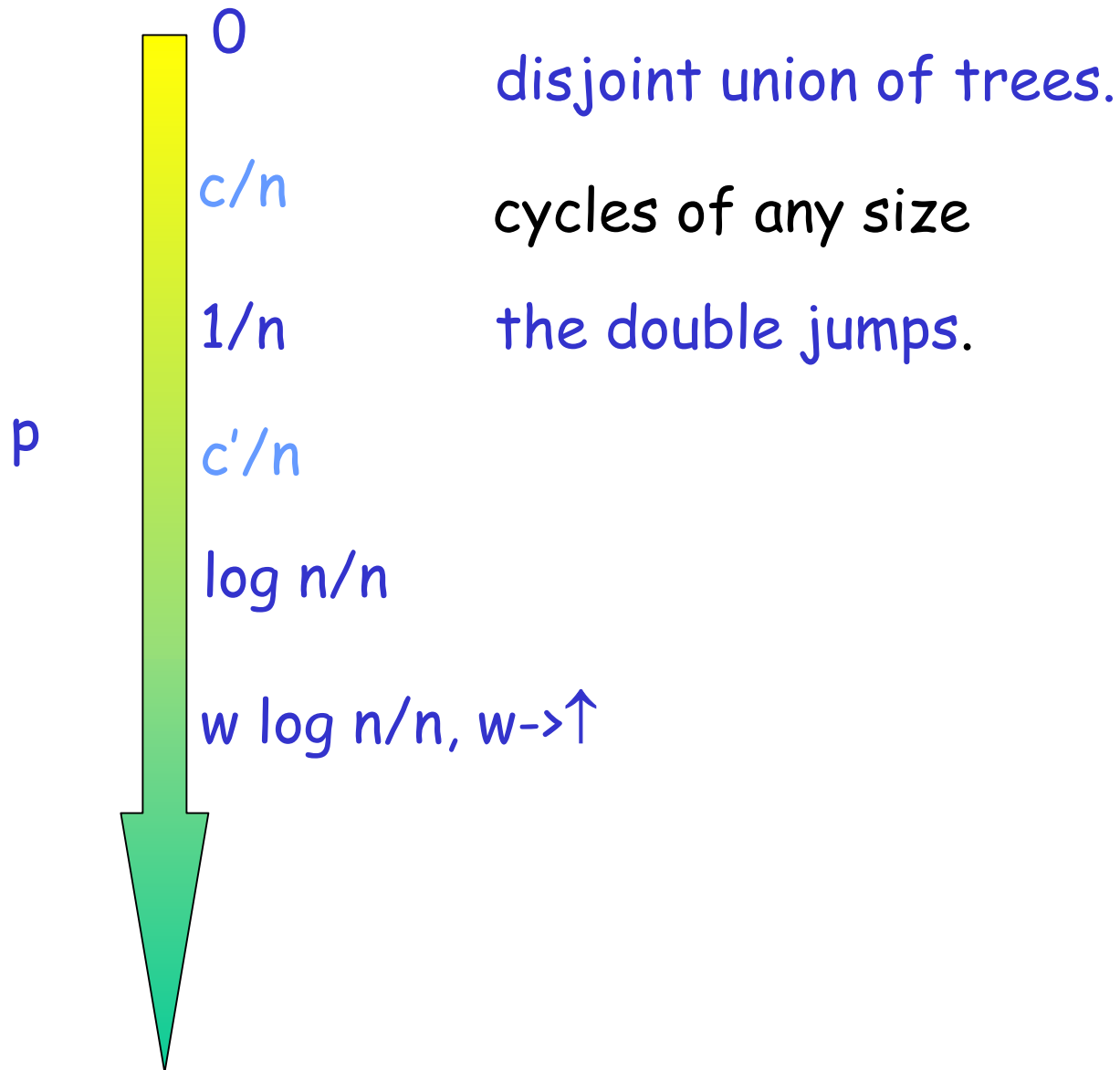
The evolution of random graphs $G(n,p)$



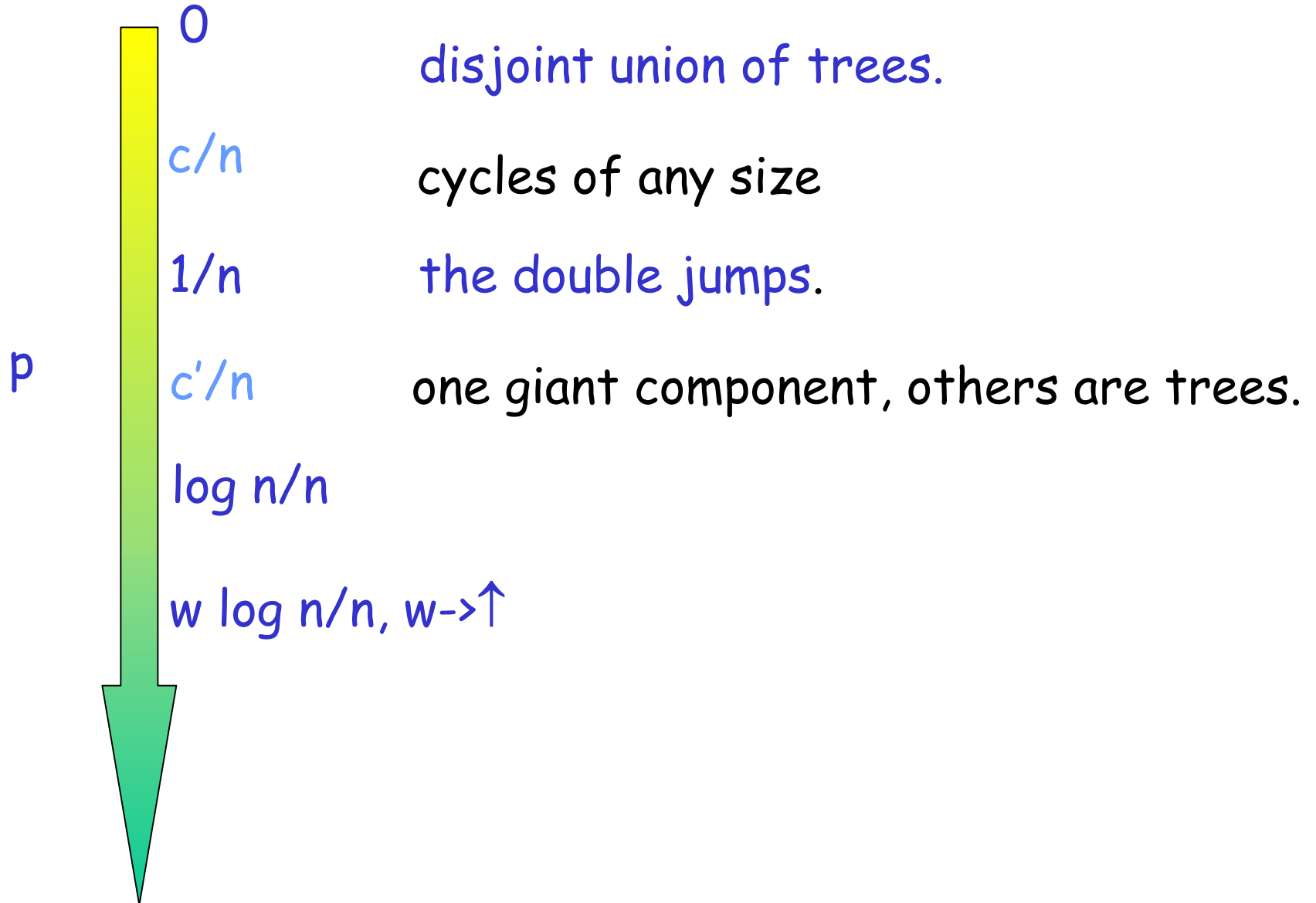
The evolution of random graphs $G(n,p)$



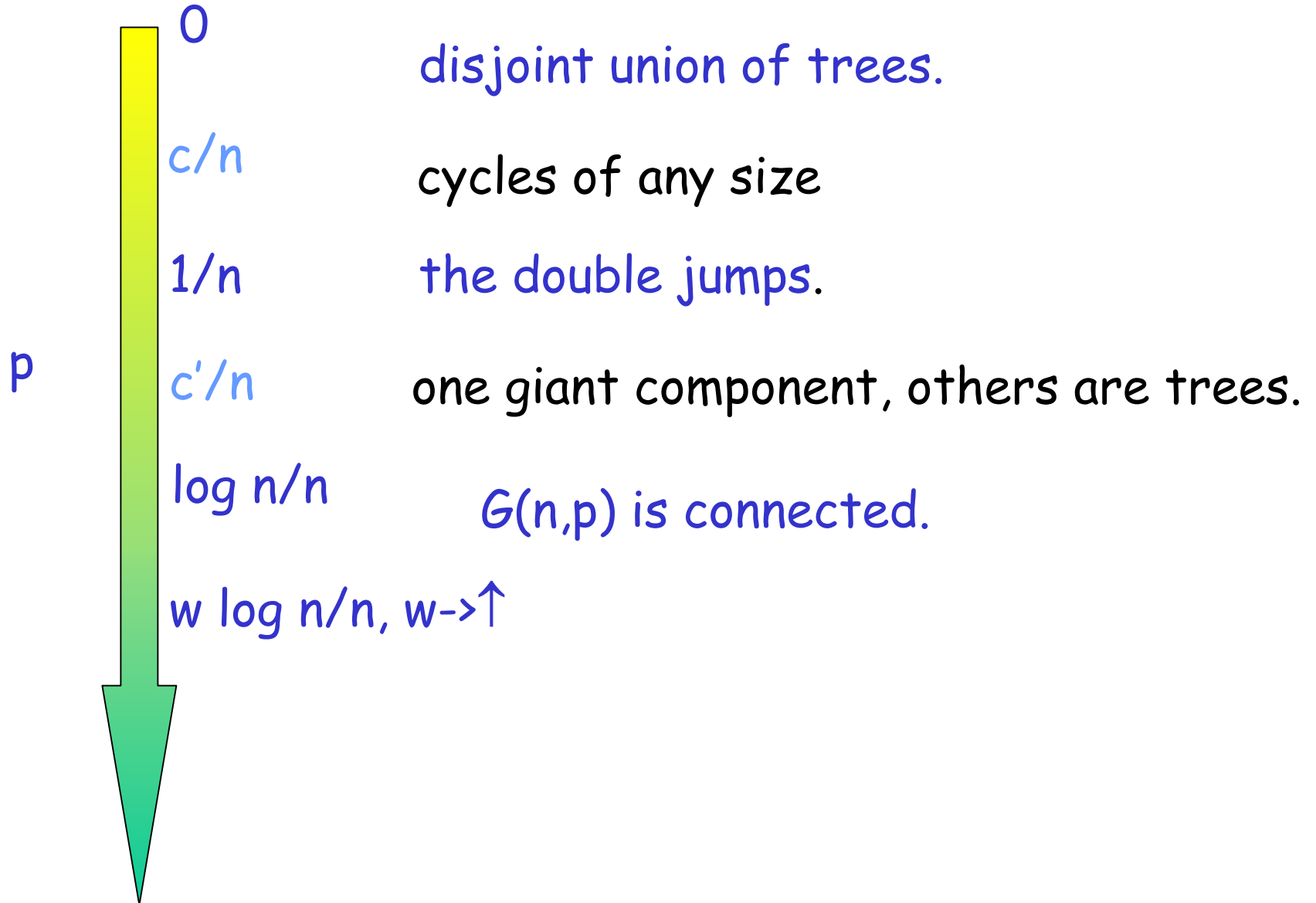
The evolution of random graphs $G(n,p)$



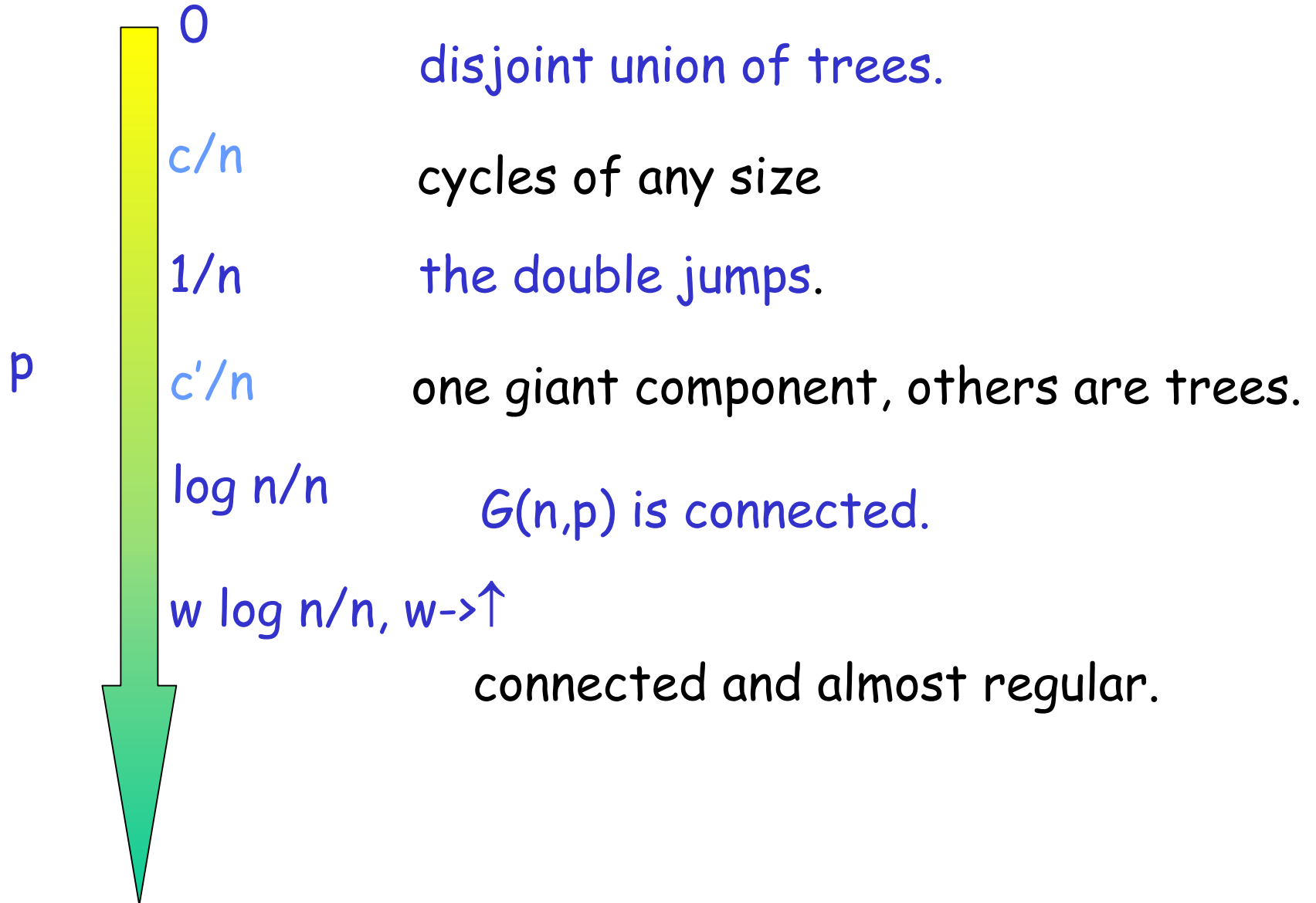
The evolution of random graphs $G(n,p)$



The evolution of random graphs $G(n,p)$



The evolution of random graphs $G(n,p)$



FINALLY..

COMPARE REAL NETWORKS
WITH RANDOM GRAPHS

it's not a random graph!

Table 1
Comparison of Statistical Features Between Random Graphs and the Yeast Protein Interaction Network

		RANDOM GRAPHS	
	YEAST	ER	PL ($\tau = 2.5$)
Whole graph			
Nodes	985	984.02 (10.39)	970.7 (81.57)
Degree	1.83	1.85 (0.98)	1.64 (1.76)
No. of components	163	108 (8)*	266.3 (30.6)*
Giant component			
Nodes	466	624.0 (38.7)*	336.9 (86)
Degree	2.3	2.07 (1.05)	2.50 (2.6)
Clustering coefficient ($\times 10^{-3}$)	22	0.59 (0.9)*	4.02 (2.3)*
Characteristic path length	7.14	15.88 (1.76)*	6.01 (1.14)

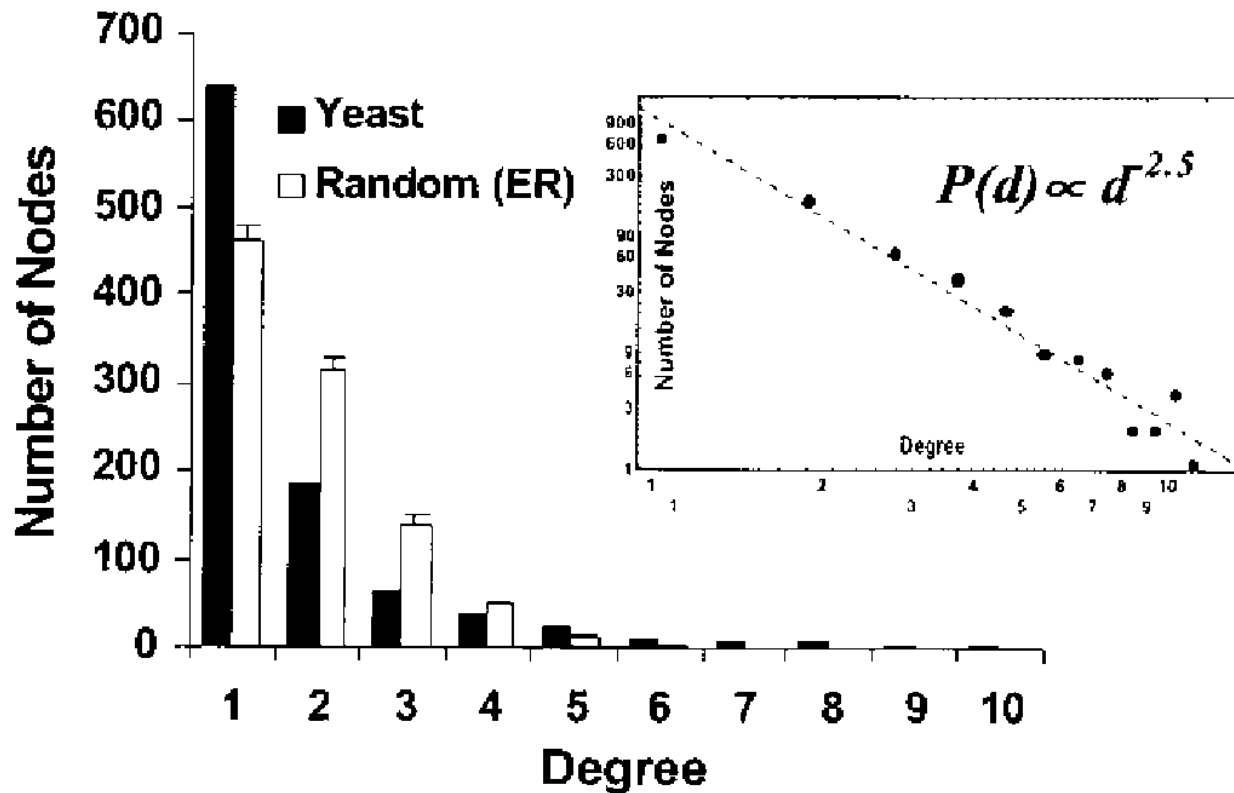
it's not a random graph!

Andreas Wagner

Mol. Biol. Evol. 18(7):1283–1292. 2001

a)

Degree distribution



IT'S ALMOST **SCALE-FREE (=POWER-LAW) GRAPH**

it's not a random graph!

The large-scale organization of metabolic networks

H. Jeong*, B. Tombor†, R. Albert*, Z. N. Oltvai† & A.-L. Barabási*

* Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA

† Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA

In a cell or microorganism, the processes that generate mass, energy, information transfer and cell-fate specification are seamlessly integrated through a complex network of cellular constituents and reactions¹. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of

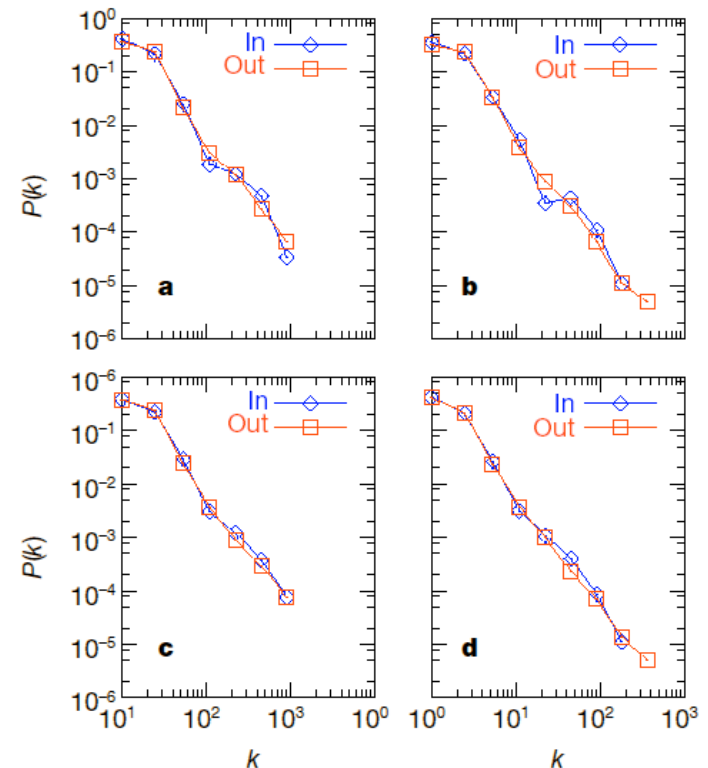
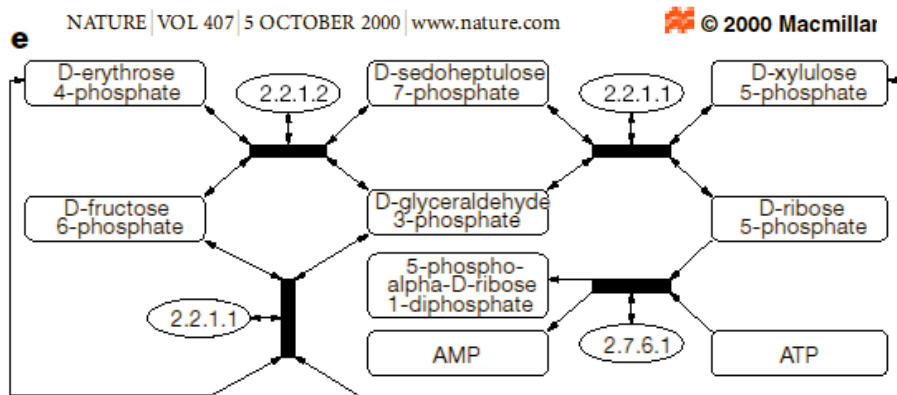


Figure 2 Connectivity distributions $P(k)$ for substrates. **a**, *Archaeoglobus fulgidus* (archae); **b**, *E. coli* (bacterium); **c**, *Caenorhabditis elegans* (eukaryote), shown on a log-log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. k_{in} (k_{out}) corresponds to the number of reactions in which a substrate participates as a product (educt). The characteristics of the three organisms shown in **a–c** and the exponents γ_{in} and γ_{out} for all organisms are given in Table 1 of the Supplementary Information. **d**, The connectivity distribution averaged over all 43 organisms.

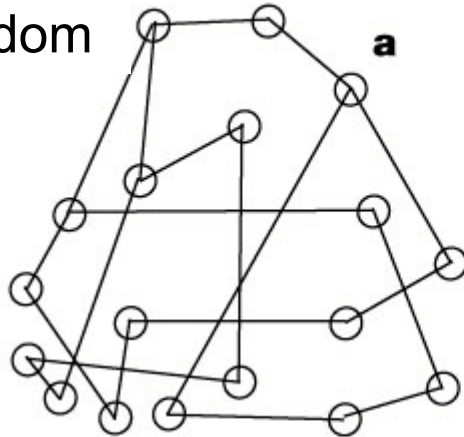
IT'S ALMOST **SCALE-FREE (=POWER-LAW) GRAPH**

Random vs power-law

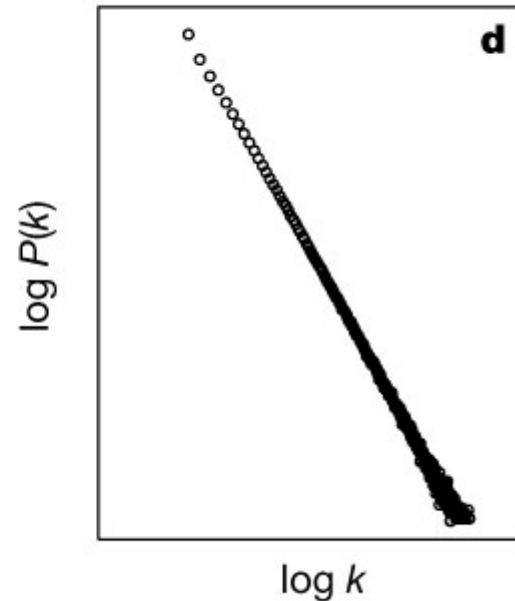
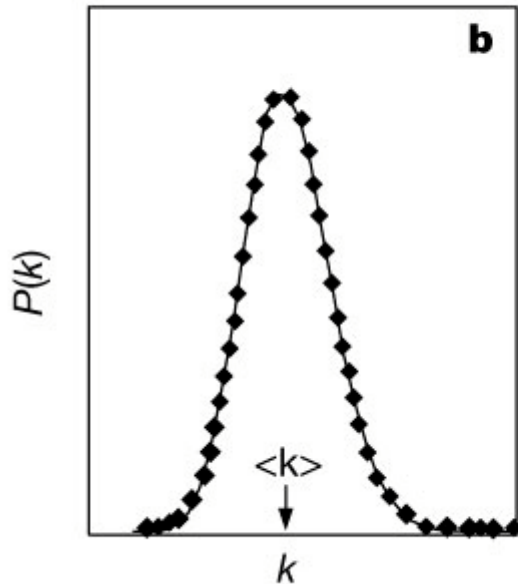
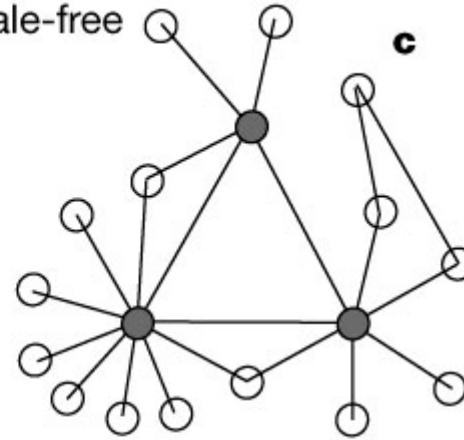
Barabasi A et.al. Nature:411(2001)

Wagner A Mol Biol Evol:18(2001)

Random



Scale-free



Other scale-free networks

Barabasi A et.al. Nature:411(2001)

- Metabolic network
- Network of social interactions:
scientific collaborations, actors in films
- The Internet:
links, physical connections

Distributions of number of sexual partnerships have power law decaying tails and finite variance

Fredrik Liljeros^{*,†}, Christofer R. Edling[†], H. Eugene Stanley[‡], Y. Åberg[†],
and Luís A. Nunes Amaral^{**}

**Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden <liljeros@sociology.su.se>*

† Department of Sociology, Stockholm University, S-106 91 Stockholm, Sweden.

‡ Centre for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215, USA.

*** Department of Chemical Engineering, Northwestern University, Evanston, IL 60208, USA.*

The web of human sexual contacts

Promiscuous individuals are the vulnerable nodes to target in safe-sex campaigns.

Unlike clearly defined 'real-world' networks¹, social networks tend to be subjective to some extent^{2,3} because the perception of what constitutes a social link may differ between individuals. One unambiguous type of contact is sexual contact, and human sexual behaviour of a representative sample of individuals⁴ to reveal the features of a sexual-contact network. The cumulative distribution of the number of different sexual partners that an individual has over a year decays as a scale-free power law. This has a similar exponential decay as the distribution of the number of different sexual partners that an individual has over a year decays as a scale-free power law. The scale-free nature of human sexual contacts is the most efficient way to prevent sexually transmitted diseases. Many real-world networks

are. The response rate was 59%, which corresponds to 2,810 respondents. Two independent analyses of non-response error revealed that elderly people, particularly women, are under-represented in the sam-

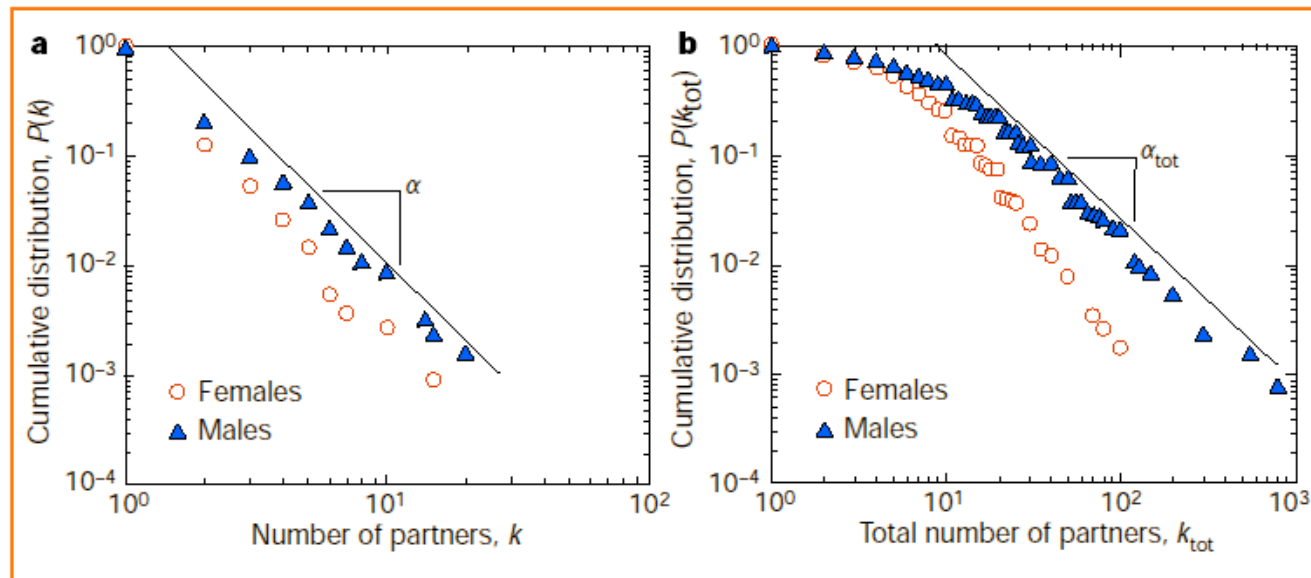


Figure 2 Scale-free distribution of the number of sexual partners for females and males. **a**, Distribution of number of partners, k , in the previous 12 months. Note the larger average number of partners for male respondents: this difference may be due to 'measurement bias' — social expectations may lead males to inflate their reported number of sexual partners. Note that the distributions are both linear, indicating scale-free power-law behaviour. Moreover, the two curves are roughly parallel, indicating similar scaling exponents. For females, $\alpha = 2.54 \pm 0.2$ in the range $k > 4$, and for males, $\alpha = 2.31 \pm 0.2$ in the range $k > 5$. **b**, Distribution of the total number of partners k_{tot} over respondents' entire lifetimes. For females, $\alpha_{\text{tot}} = 2.1 \pm 0.3$ in the range $k_{\text{tot}} > 20$, and for males, $\alpha_{\text{tot}} = 1.6 \pm 0.3$ in the range $20 < k_{\text{tot}} < 400$. Estimates for females and males agree within statistical uncertainty.

Random vs power-law

Barabasi A et.al. Nature:411(2001)

Wagner A Mol Biol Evol:18(2001)

The network of protein-protein interactions
(and other molecular biological networks)
are scale-free networks!

WHY?

- Scale-free networks are "better"...

OR/AND

- Biological networks became scale-free due to evolution.

Random

- Removal of a randomly picked node significantly increases **the average path**.
- All nodes are of equal “importance”.

Scale-free

- Removal of a random node slightly increases **the average path**.
- Removal of a highly-connected node leads to drastic increase of **the average path!**

POWER-LAW NETWORKS

- Tolerant to random “attacks”,
- But more sensitive to targeted attacks!

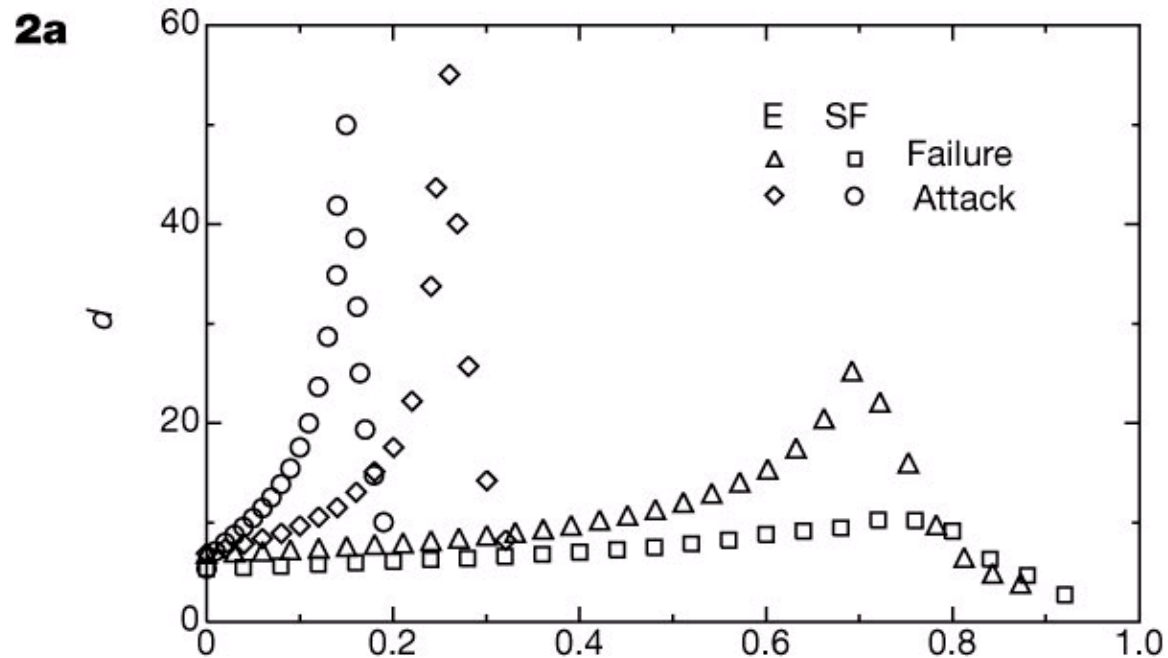


Figure 2 Changes in the diameter d of the network as a function of the fraction f of the removed nodes. **a**, Comparison between the exponential (E) and scale-free (SF) network models, each containing $N = 10,000$ nodes and 20,000 links (that is, $k = 4$).

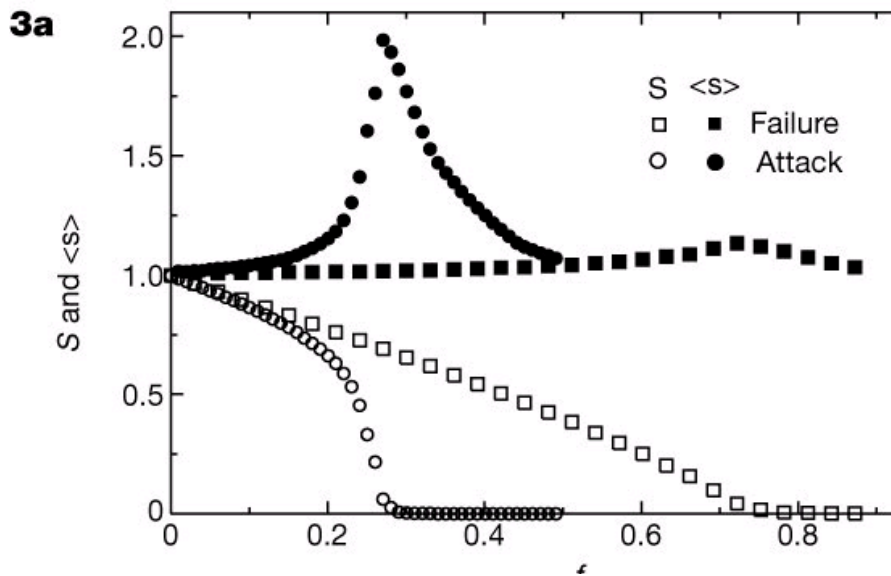
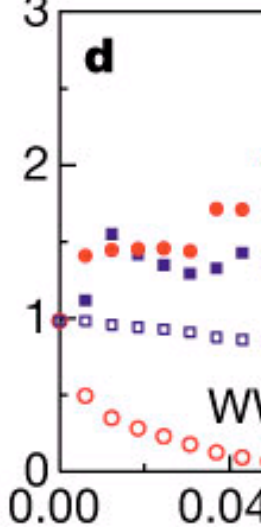
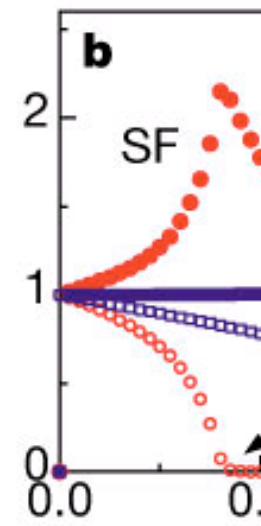
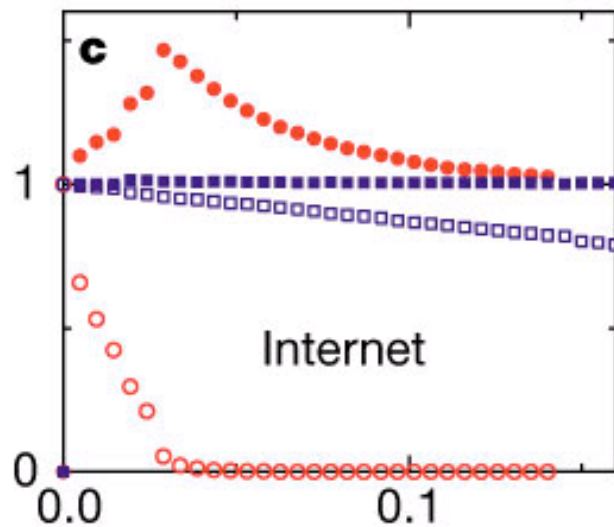
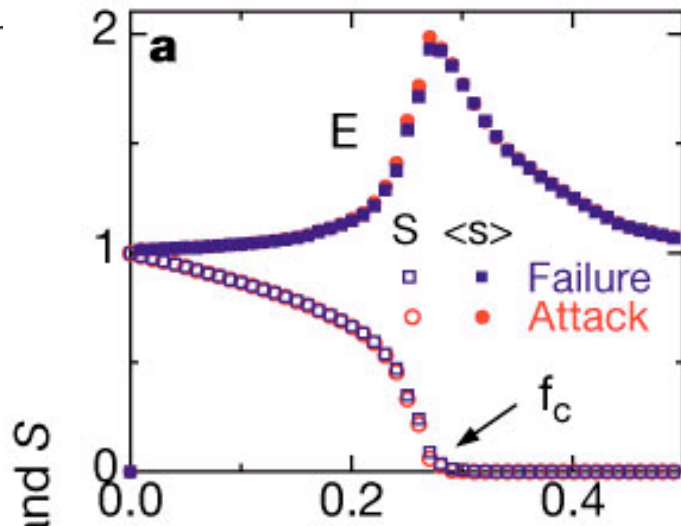


Figure 3 Network fragmentation under random removal of the largest cluster S (open symbols) and the average size $\langle S \rangle$ (filled symbols) as a function of the fraction of removed nodes f . The size S is defined as the fraction of nodes in the largest cluster for $f=0$. **a**, Fragmentation of the exponential network under random removal and attacks (circles). **b**, Fragmentation of the scale-free network under random removal (blue squares) and attacks (red circles). The critical fraction f_c is indicated by the arrow, indicating that the main cluster is completely deflated.



f

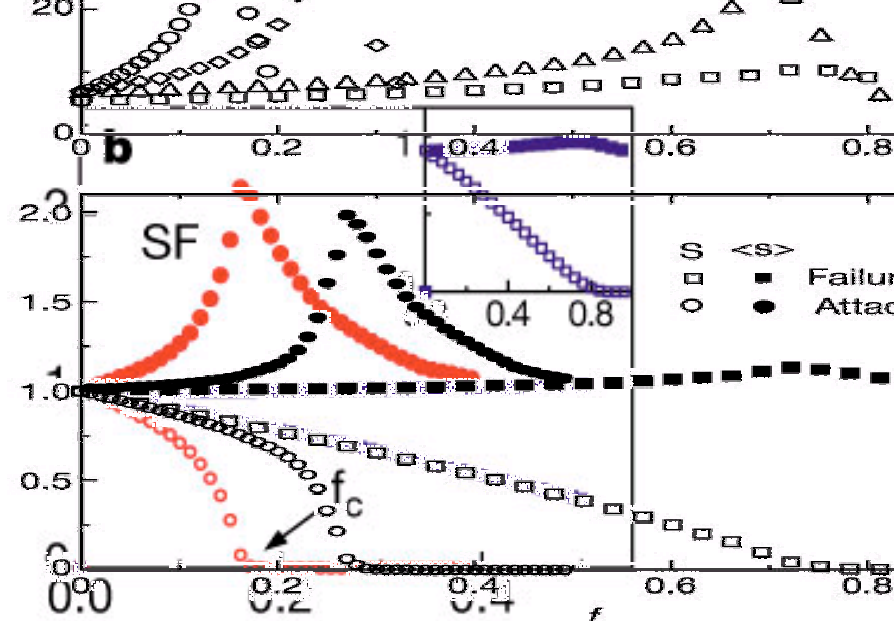


Figure 3 Network fragmentation under random failures and attacks. The relative size of the largest cluster S (open symbols) and the average size of the isolated clusters s (filled symbols) as a function of the fraction of removed nodes f for the same systems as in Fig. 2. The size S is defined as the fraction of nodes contained in the largest cluster (that is, $S = 1$ for $f = 0$). **a**, Fragmentation of the exponential network under random failures (squares) and attacks (circles). **b**, Fragmentation of the scale-free network under random failures (blue squares) and attacks (red circles). The inset shows the error tolerance curves for the whole range of f , indicating that the main cluster falls apart only after it has been completely deflated

Evolution of scale graphs (AB model)

1. Growth

2. Preferential attachment

Alberts and Barabasi 2000

Herbert A. Simon 1967

Evolution of graphs

- Growth

1. start with m_0 nodes
 2. add a node with m edges
 3. connect these edges to existing nodes
- at timestep t : $t+m_0$ nodes, tm edges

Evolution of graphs

- Preferential attachment

Probability Π of connection to node i depends on the degree k_i of this node.

E.g.
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

“Rich gets richer”

Evolution of power-law graphs

1. Growth

2. Preferential attachment

Rich gets richer:



Better evolution of graphs

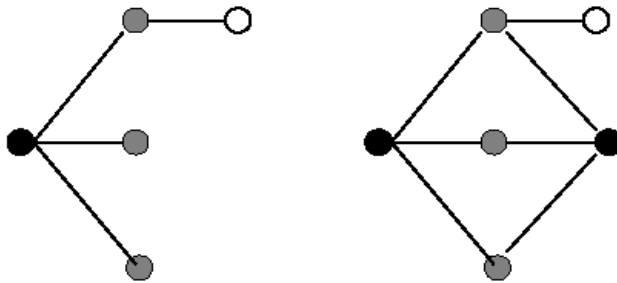
A. Wagner, M.Lassig, A.Maritan etc

- Gene duplication
- Mutations
- Preferential attachment

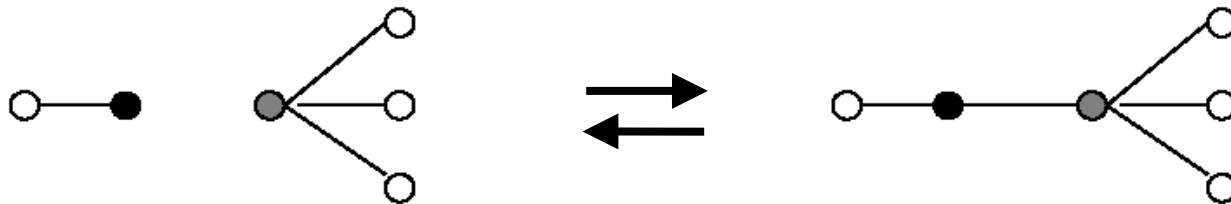
More biological **neutral** evolution of graphs

A.Wagner, M.Lassig, A.Maritan, S.Redner etc.

- Gene duplication



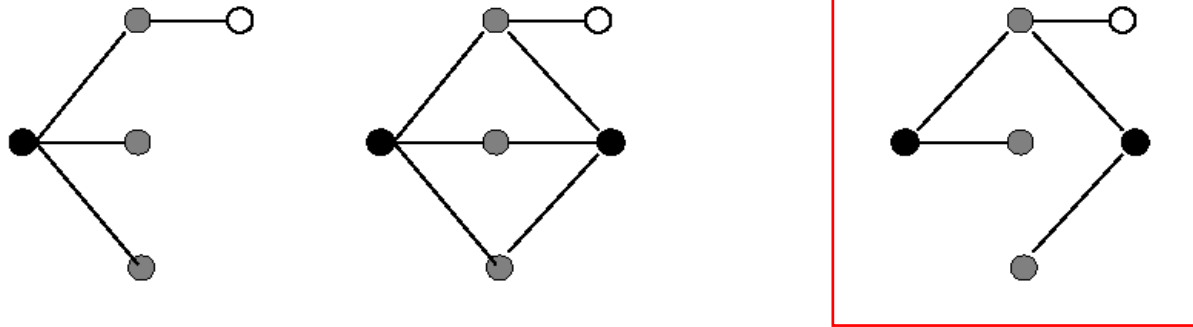
- Mutations



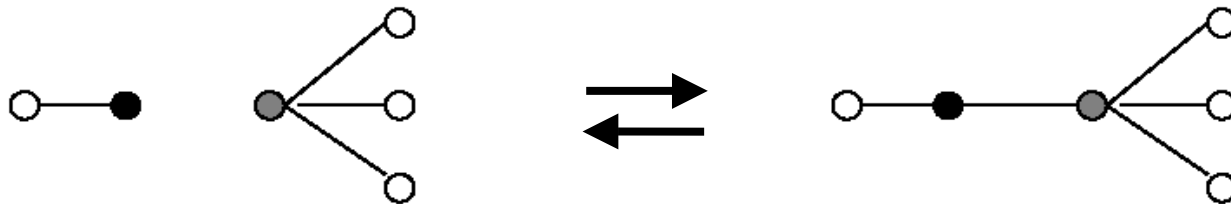
More biological neutral evolution of graphs

A.Wagner, M.Lassig, A.Maritan, S.Redner etc.

- Gene duplication



- Mutations (rich gets richer)



=> Broad (not power-law) distribution!

Summary

- Structures of some biological network are known (partially and with lots of mistakes!)
- They do not look like random graphs, more like power-law graphs.
- Such structures can be
 - results of neutral evolution
 - results of selection(for robustness or other properties)

Biological questions

- Lethality and centrality
- Degree anti-correlations
- Regulatory motifs
- Communities

Biological questions

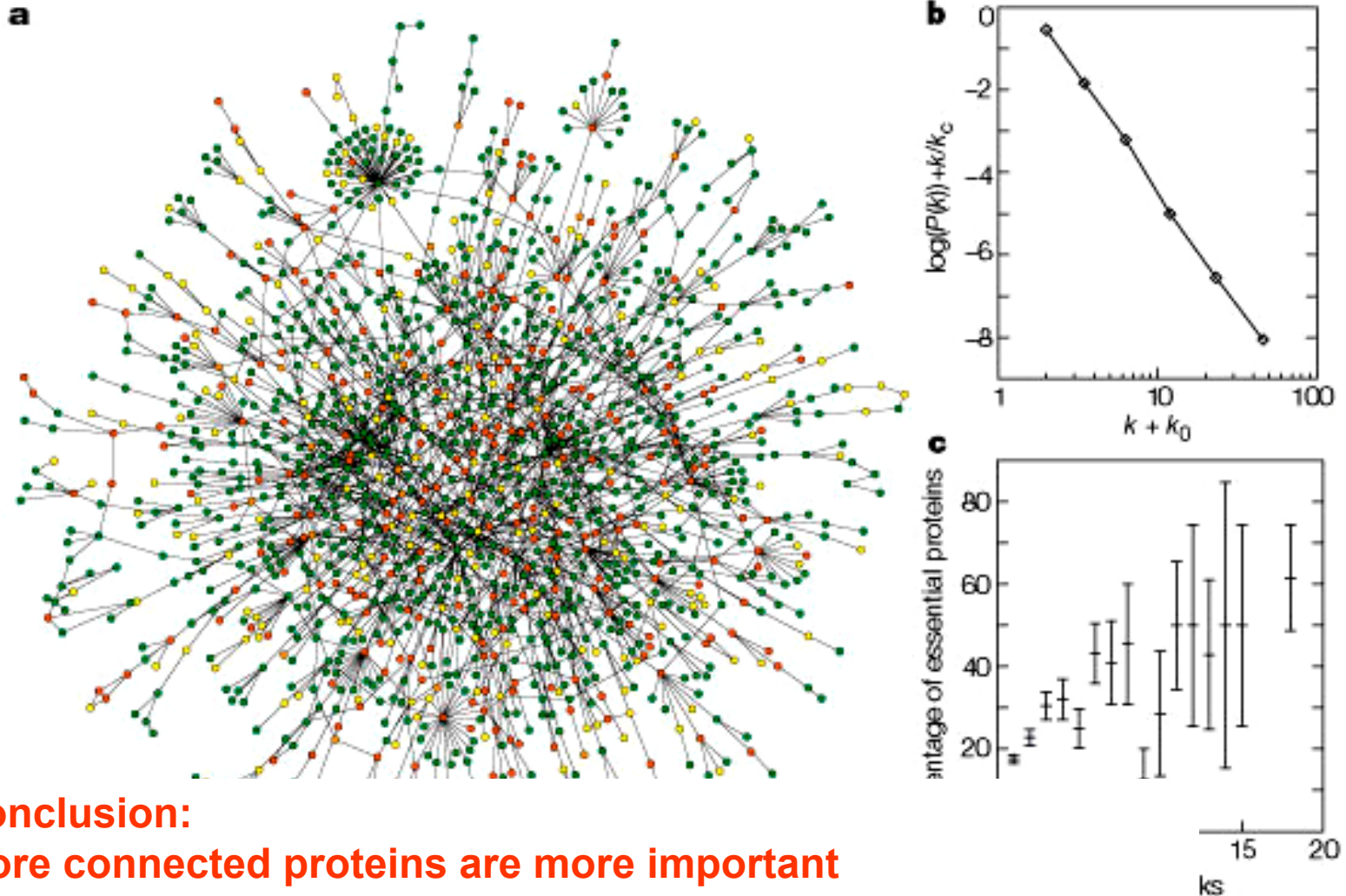
- Lethality and centrality

Jeong H, Mason SP, Barabasi AL, Oltvai ZN.

[Lethality and centrality in protein networks.](#)

Nature. 2001 May 3;411(6833):41-2.

Biological questions



Biological questions

- Degree anti-correlations

Specificity and Stability in Topology of Protein Networks

Sergei Maslov¹ and Kim Sneppen²

3 MAY 2002 VOL 296 SCIENCE www.sciencemag.org

conclusion:
MODULARITY

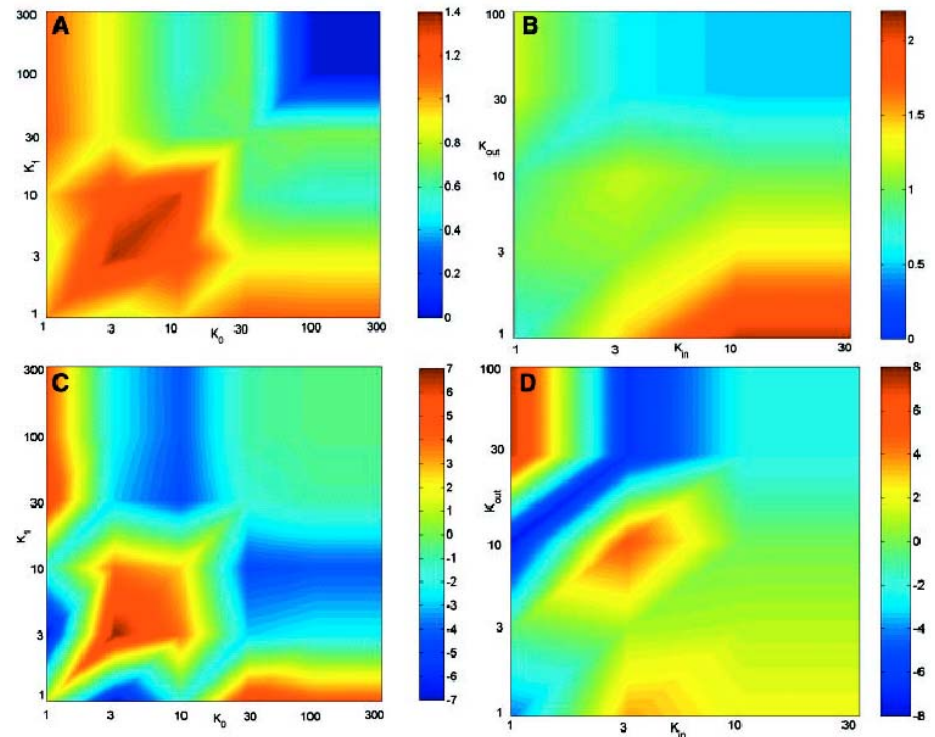


Fig. 2. Correlation profiles of protein interaction and regulatory networks in yeast. **(A)** The ratio $P(K_0, K_1)/P_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with total numbers of interaction partners given by K_0, K_1 correspondingly, directly interact with each other in the full set of (2), while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network. **(B)** The same as in **(A)** but for a protein with the in-degree K_{in} to be regulated by that with the out-degree K_{out} in the transcription regulatory network (3). **(C)** Z-scores for connectivity correlations from **(A)**: $Z(K_0, K_1) = [P(K_0, K_1) - P_r(K_0, K_1)]/\sigma_r(K_0, K_1)$, where $\sigma_r(K_0, K_1)$ is the standard deviation of $P_r(K_0, K_1)$ in 1000 realizations of a randomized network. **(D)** As in **(C)**, but for incoming and outgoing links in the transcription regulatory network. To improve statistics, the connectivities in all four panels of this figure were logarithmically binned into 2 bins per decade.

Biological questions

- Degree anti-correlations

Specificity and Stability in Topology of Protein Networks

Sergei Maslov¹ and Kim Sneppen²

3 MAY 2002 VOL 296 SCIENCE WWW.

conclusion:
MODULARITY

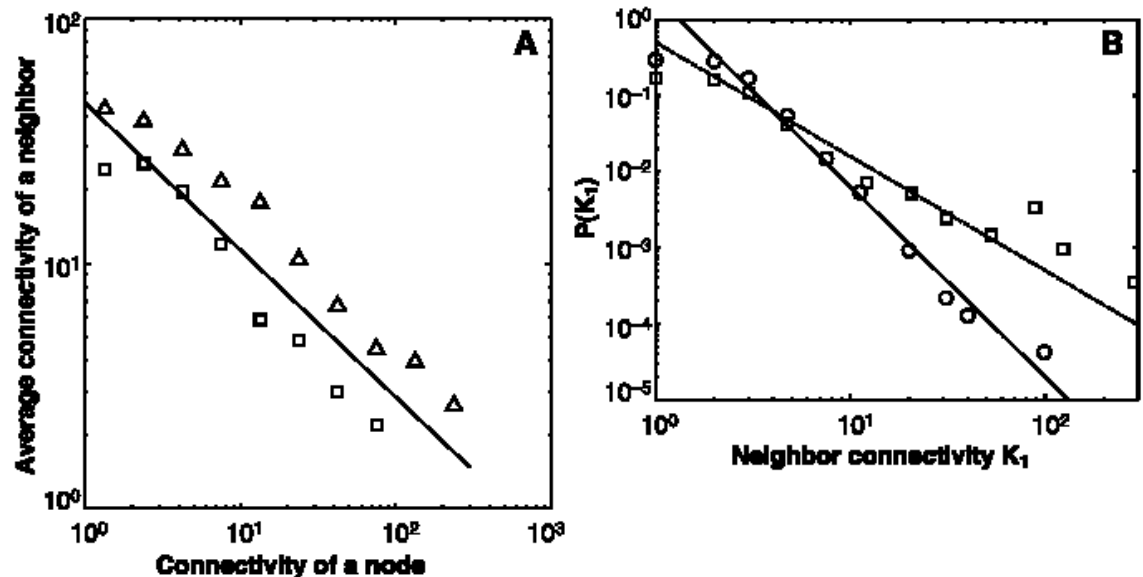


Fig. 3. Correlations in connectivities of neighbors. (A) The average connectivity $\langle K_1 \rangle$ of nearest neighbors of proteins with the connectivity K_0 in the physical interaction network (triangles) and the regulatory network (squares). The solid line is a power law fit, $\propto 1/K_0^{0.6}$. (B) The probability distribution of connectivities K_1 in the physical interaction network, calculated separately for neighbors of proteins with small connectivity $K_0 \leq 3$ (squares), and with large connectivity $K_0 \geq 100$ (circles). Lines are power laws $\propto 1/K_1^{1.5}$ (dashed) and $\propto 1/K_1^{2.5}$ (solid).

Biological questions

- Degree anti-correlations

Specificity and Stability in Topology of Protein Networks

Sergei Maslov¹ and Kim Sneppen²

3 MAY 2002 VOL 296 SCIENCE www.sciencemag.org

conclusion:

MODULARITY

SEE RELATED PAPER BY *T. Ideker*
IN THE READING LIST

Biological questions

- Regulatory motifs

(see papers by *U.Alon*
and *T.Ideker*)

Biological questions

- Communities

(see papers by *M.Newman*
and *L.Mirny*)

Scheme of basic concepts

– Characteristics

- diameter
- clustering
- degree distribution
- betweenness
- communities/modules/motifs

– Models

- random
- scale-free
- WS

– Properties

- attack vulnerability