




Complex networks II

Albert Díaz-Guilera

<http://diaz-guilera.net/>

@anduviera

PHYSCOMP² Universitat de Barcelona



Characterizing complex networks

Levels of characterization

- ▶ Microscale: role of nodes in the network (centrality, degree, betweenness, ...)
- ▶ Macroscale: distributions, statistical properties
- ▶ Mesoscale: motifs, modules, communities, ...

Microscale



- ▶ Centrality
- ▶ Degree (local perspective)
- ▶ Other measures (global perspective)

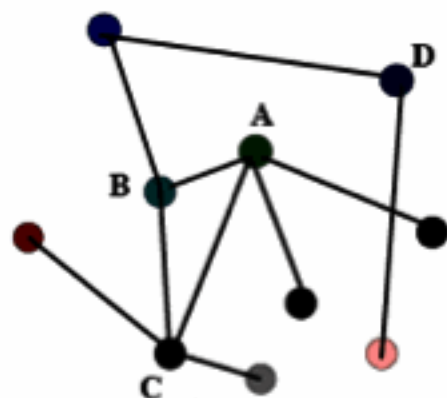
(a)

Degree (microscopic scale)

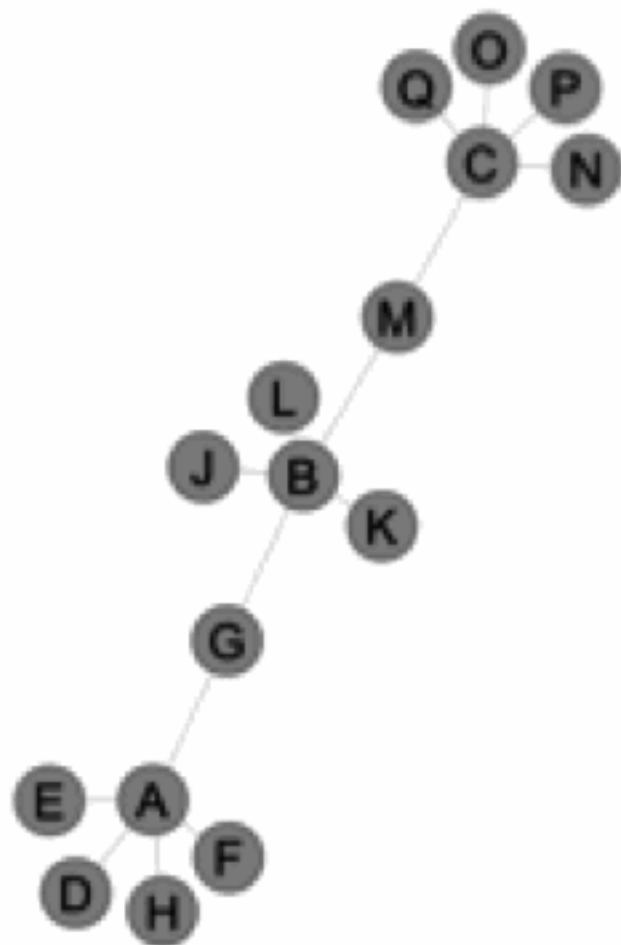
- ▶ Number of links that a node has
- ▶ It corresponds to the local centrality in social network analysis
- ▶ It measures how important is a node with respect to its nearest neighbors

Distance between two nodes

- ▶ Number of links that make up the shortest-path between two nodes



- ▶ Centrality: nodes that are “close” to many other nodes in the network.
- ▶ **Closeness centrality:** average distance from a given node to the other nodes



	A,C	B	G,M	J,K,L	La resta
Local (grau)	5	5	2	1	1
Global (distància)	43	33	37	47	57

Betweenness

- ▶ Measures the “intermediary” role in the network
- ▶ It is a set of matrices, one for each node

B_{ij}^k **Ratio of shortest paths between i and j that go through k**

$0 \leq B_{ij}^k \leq 1$ **There can be more than one geodesic between i and j**

$$B_k = \sum_{ij} B_{ij}^k$$

- ▶ B_k is a measure of the centrality, in terms of flow, of node k

Eigenvector centrality

- ▶ Generalization of degree
- ▶ Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

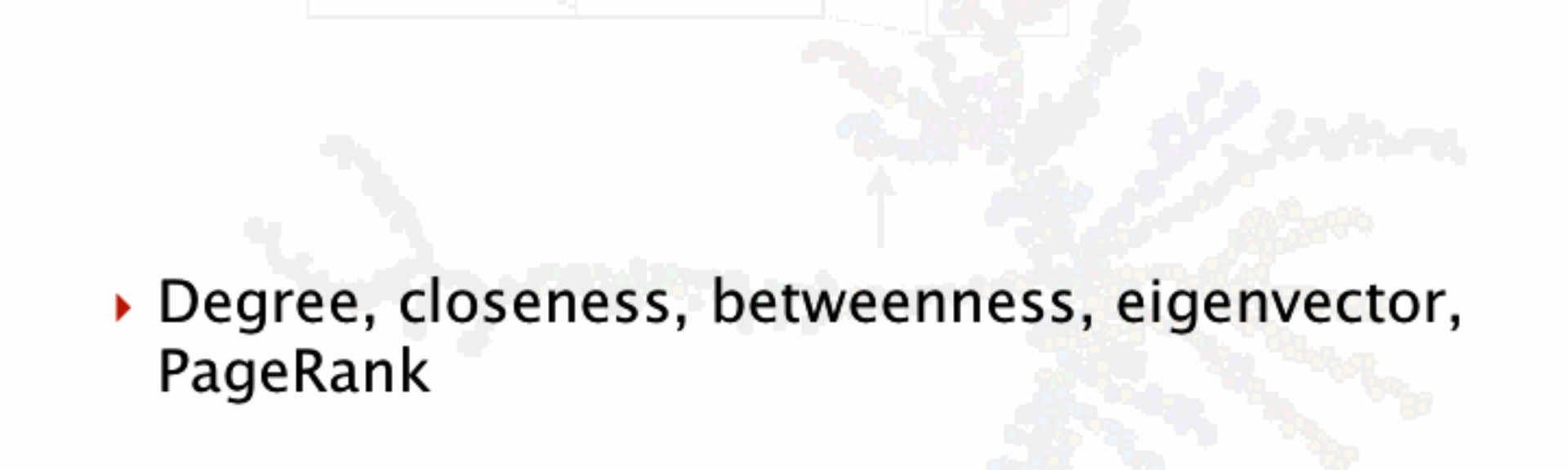
$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j=1}^N a_{i,j} x_j$$

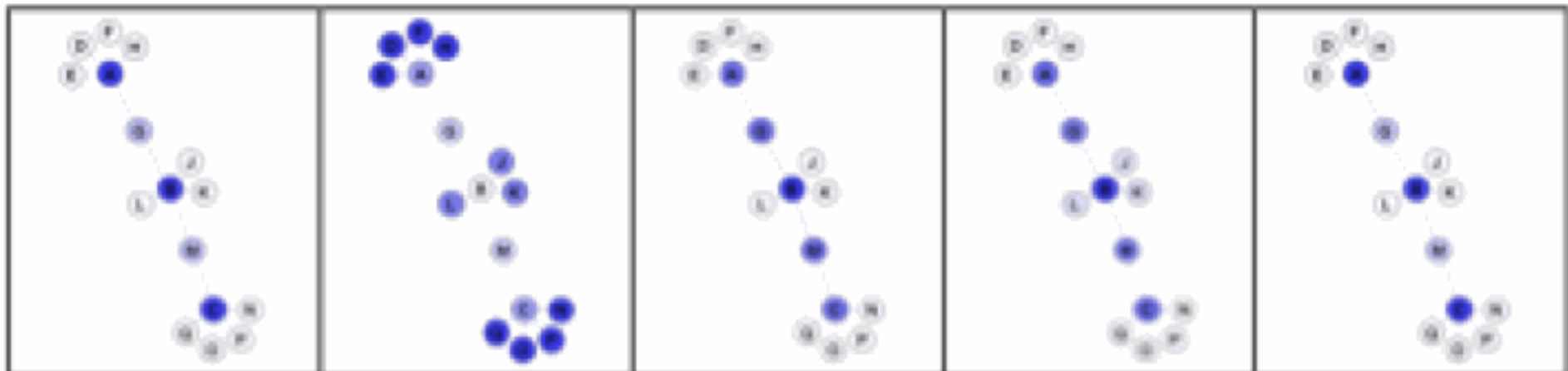
$$\mathbf{Ax} = \lambda \mathbf{x}$$

Page-rank

- ▶ Originally developed by the founders of Google (Page and Brin)
- ▶ Related to the probability that a random walker arrives to a given node
- ▶ Recursive relation (very fast convergence)

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

- 
- ▶ Degree, closeness, betweenness, eigenvector, PageRank

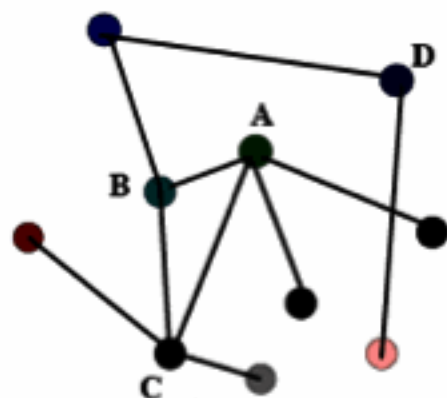


Macroscale

- ▶ Shortest-paths
- ▶ Clustering (of the network)
- ▶ Distributions (degree, betweenness, ...)
- ▶ Statistical properties: clustering
- ▶ Correlations

Distance between two nodes

- ▶ Number of links that make up the shortest-path between two nodes



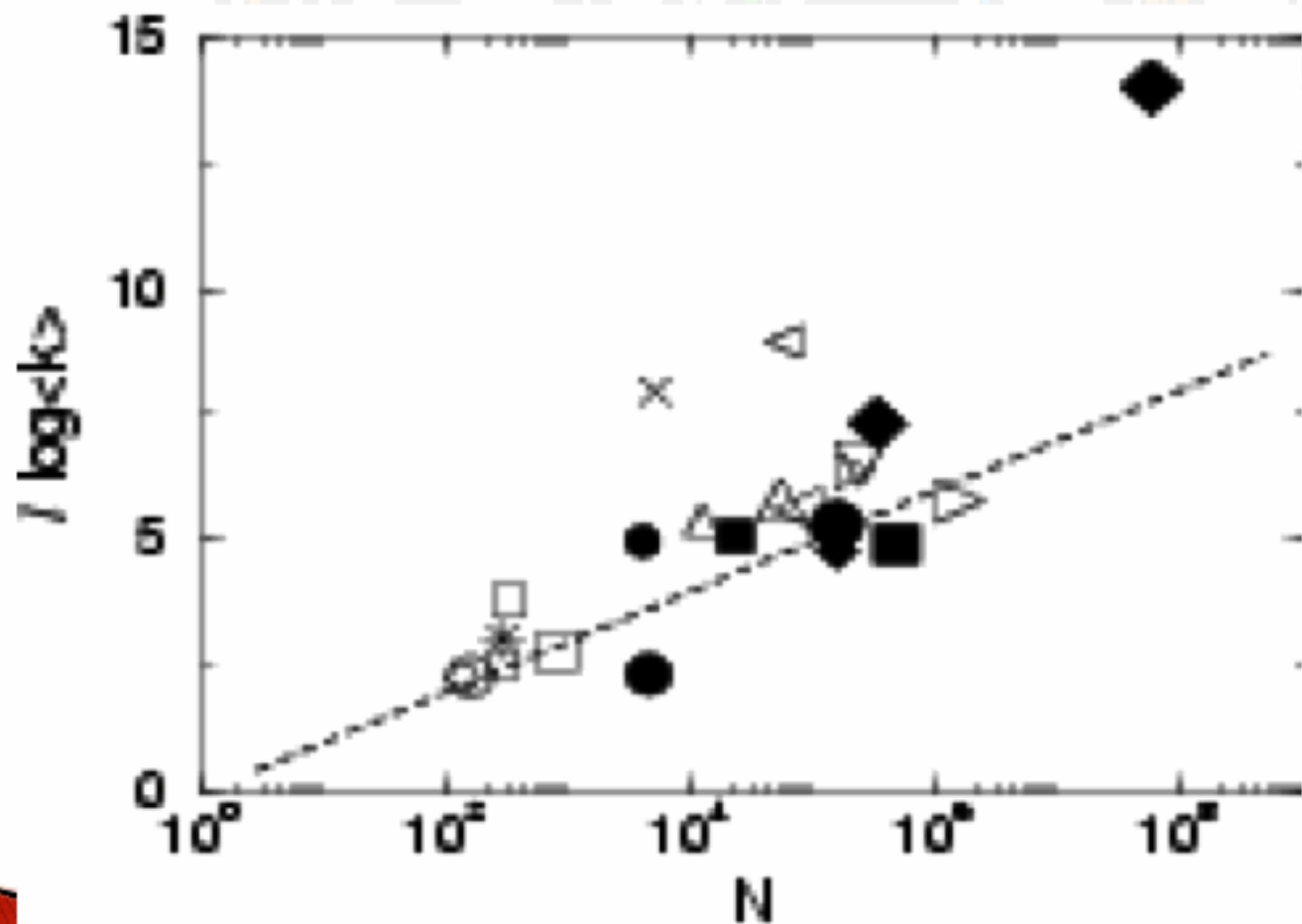
- ▶ Centrality: nodes that are “close” to many other nodes in the network.
- ▶ Global centrality: defined as the sum of minimum distances to any other nodes in the networks

Global centrality of the whole network?

Mean shortest path = average over all pairs of nodes in the network

Diameter: largest distance between a pair of nodes in the network

What do we find?



Clustering

- ▶ Cycles in social network analysis language
- ▶ Circles of friends in which every member knows each other

Clustering coefficient

- ▶ Clustering coefficient of a node

$$C_i = \frac{E_i}{k_i(k_i-1)/2}$$

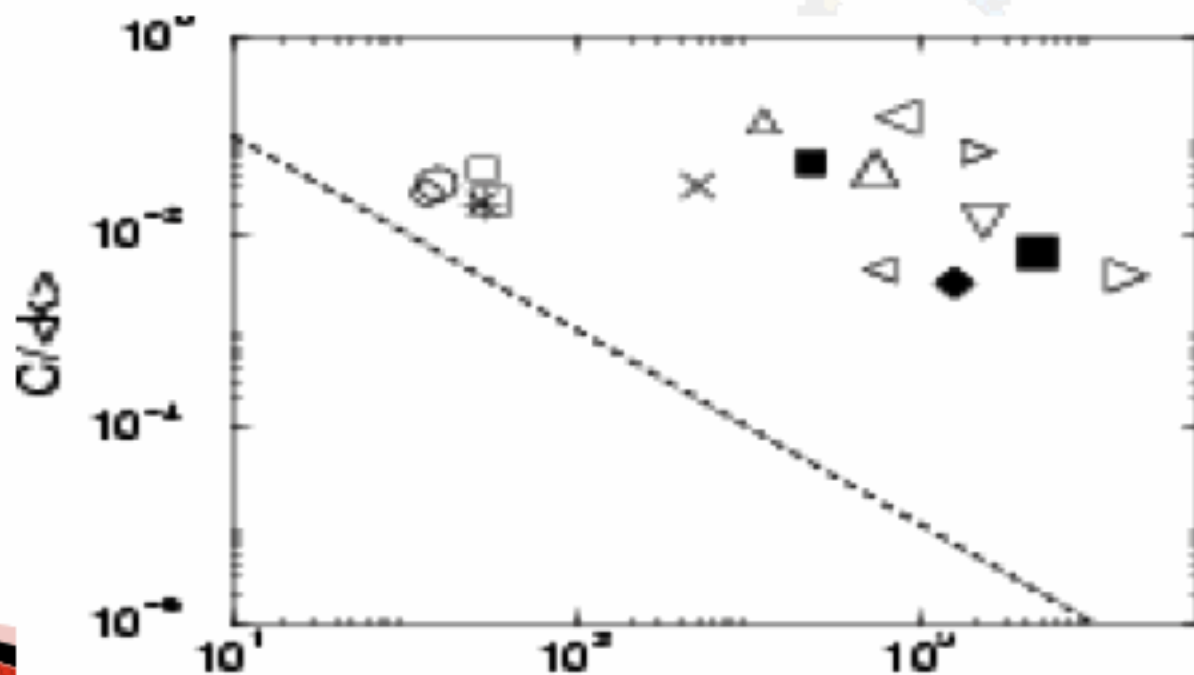
- ▶ Clustering coefficient of the network

$$C = \frac{1}{N} \sum_{i=1}^N C_i$$

- ▶ Alternative definition: ratio between total number of triangles and possible

What happens in real networks?

- ▶ The clustering coefficient is much larger than it is in an equivalent random network

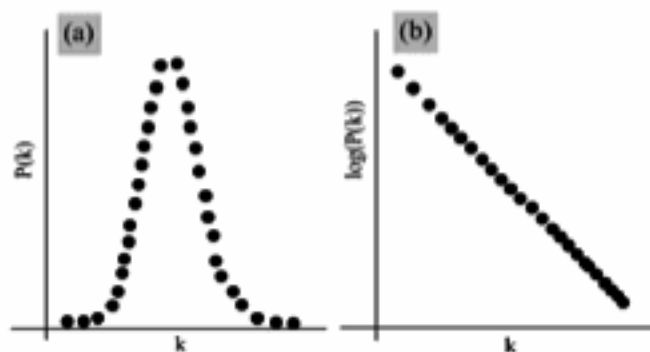


Degree distribution (macroscopic scale)

- ▶ Gives an idea of the spread in the number of links the nodes have
- ▶ $P(k)$ is the probability that a randomly selected node has k links

What should we expect?

- ▶ In regular lattices all nodes are identical
- ▶ In random networks the majority of nodes have approximately the same degree

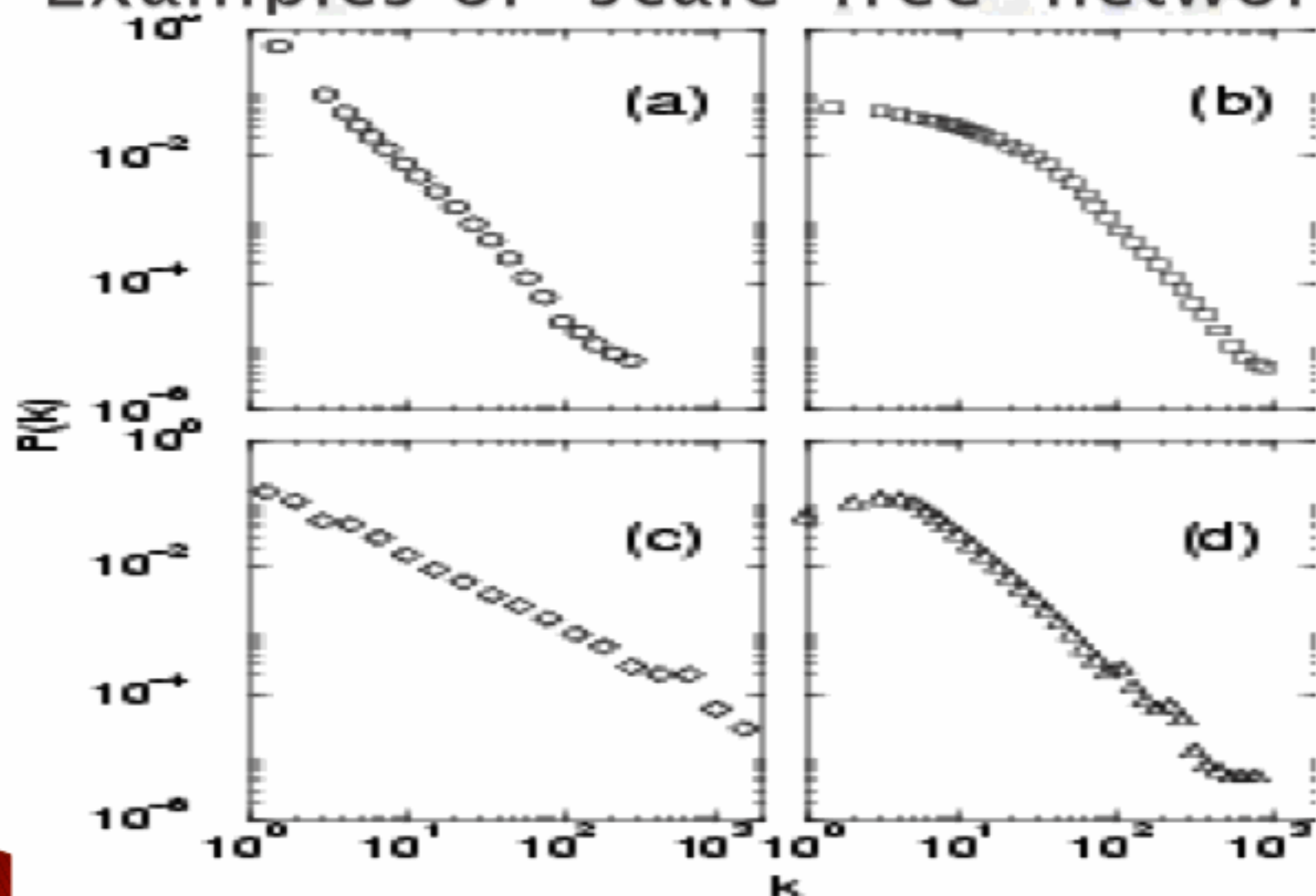


- ▶ Real-world networks: this distribution has a power tail

$$P(k) \approx k^{-\gamma}$$

"scale-free" networks

Examples of “scale-free” networks




Universality

- ▶ We find the same macroscopic behavior in systems which have completely different origins
- ▶ Engineered or self-organized
- ▶ Biology, transportation, social, ...
- ▶ Which are the basic underlying principles?
- ▶ Models (simple models) that can explain the basic trends

Correlations

- ▶ Degree correlations: expected degree of the neighbors of a node as a function of its degree

$$k_{\text{nn}}(k) = \sum_{k'} k' P(k'|k)$$



The correct mathematical way to quantify such a measure is the *conditioned probability* $p(k_1|k_2)$ to have a vertex with degree k_1 at one side of the edge when at the other site of the edge the degree is k_2 .

We have two constraints on the conditioned probability. The first one is given by normalization condition

$$\sum_{k_1} p(k_1|k_2) = 1. \quad (1.10)$$

For non oriented graphs the same quantity obeys the detailed balance distribution (Boguñá and Pastor-Satorras, 2002)

$$k_2 p(k_1|k_2) P(k_2) = k_1 p(k_2|k_1) P(k_1) \quad (1.11)$$

This balance equation simply states that the number of edges going from vertex k_1 to vertex k_2 must be equal to the number of edges going from vertex k_2 to vertex k_1 .

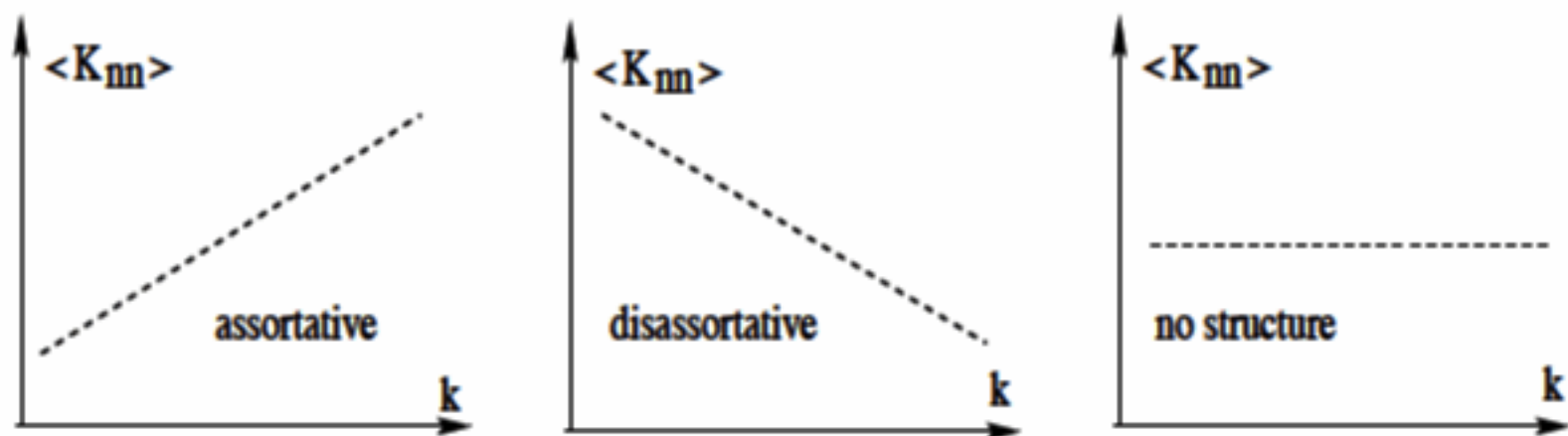


FIG. 1.10. The three possible behaviour of the average degree of the neighbours $\langle K_{nn} \rangle$ versus the degree k of the vertex origin.

Assortativity

Following (Callaway, Hopcroft, Kleinberg, Newman and Strogatz, 2001; Newman, 2002a) we define

$$r = \frac{1}{\sigma^2} \sum_{k_1, k_2} (p(k_1|k_2) - q_{k_1}q_{k_2}) \quad (1.16)$$


where

- q_k is the normalized distribution for the “remaining degree” of vertices. Remaining degree is the degree of a vertex without the edge considered in the link. In formulas, this means that the remaining degree of vertex i is given by $k_i - 1$. The normalized distribution for such a quantity is then given by:

$$q_k = (k + 1)P(k + 1) / \sum_{i=1, N} iP(i).$$

- The σ^2 is the variance of the above quantity:

$$\sigma^2 = \sum_{i=1, N} k^2 q_k - (\sum_{i=1, N} k q_k)^2$$



Network	n	r
Physics Co-authorship	52909	0.363
Biology Co-authorship	1520251	0.127
Mathematics Co-authorship	253339	0.120
Film Actors Collaboration	253339	0.208
Company Directors	7673	0.276
Internet	10697	-0.189
Protein Interactions	2115	-0.156
Marine food web	134	-0.247
Little Rock Lake	92	-0.276

Table 1.1 *Order and assortative coefficient for various networks*

Mesoscale

► Motifs

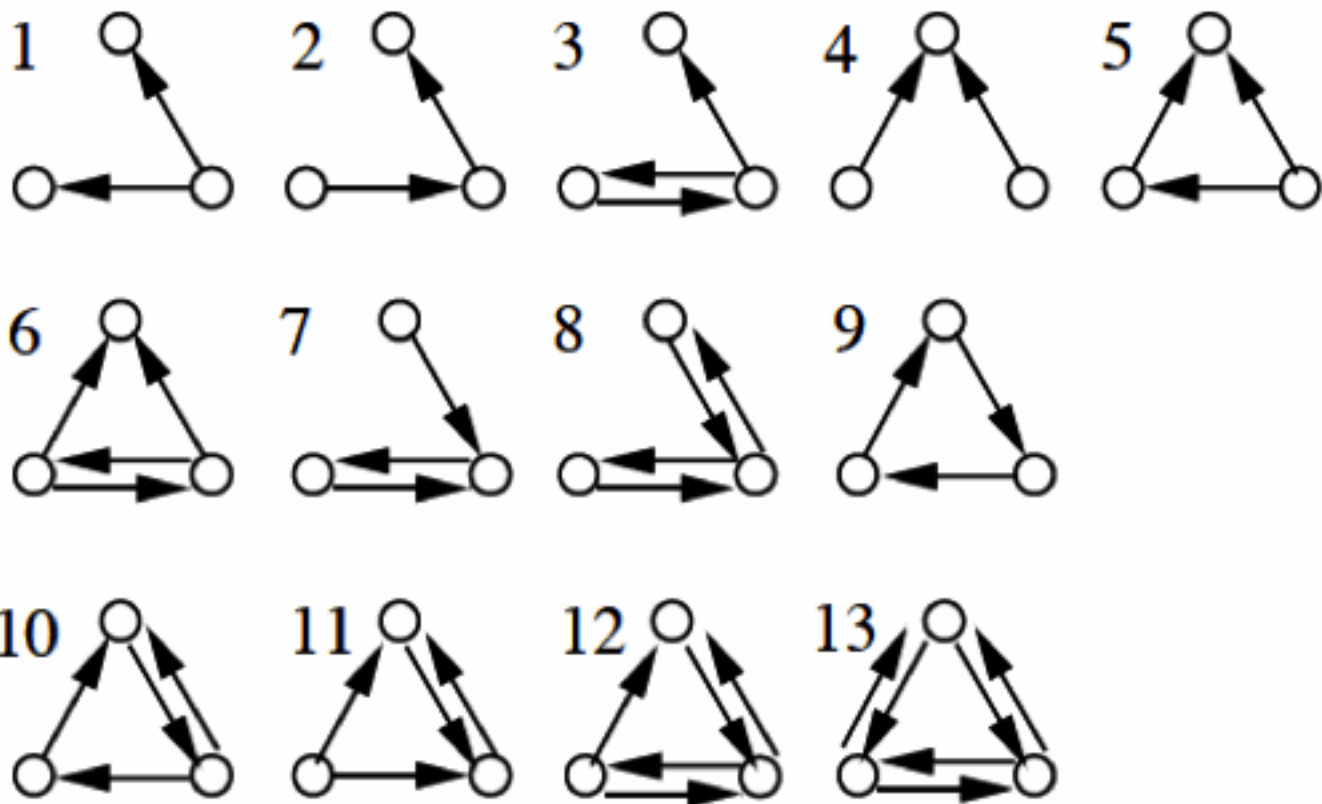
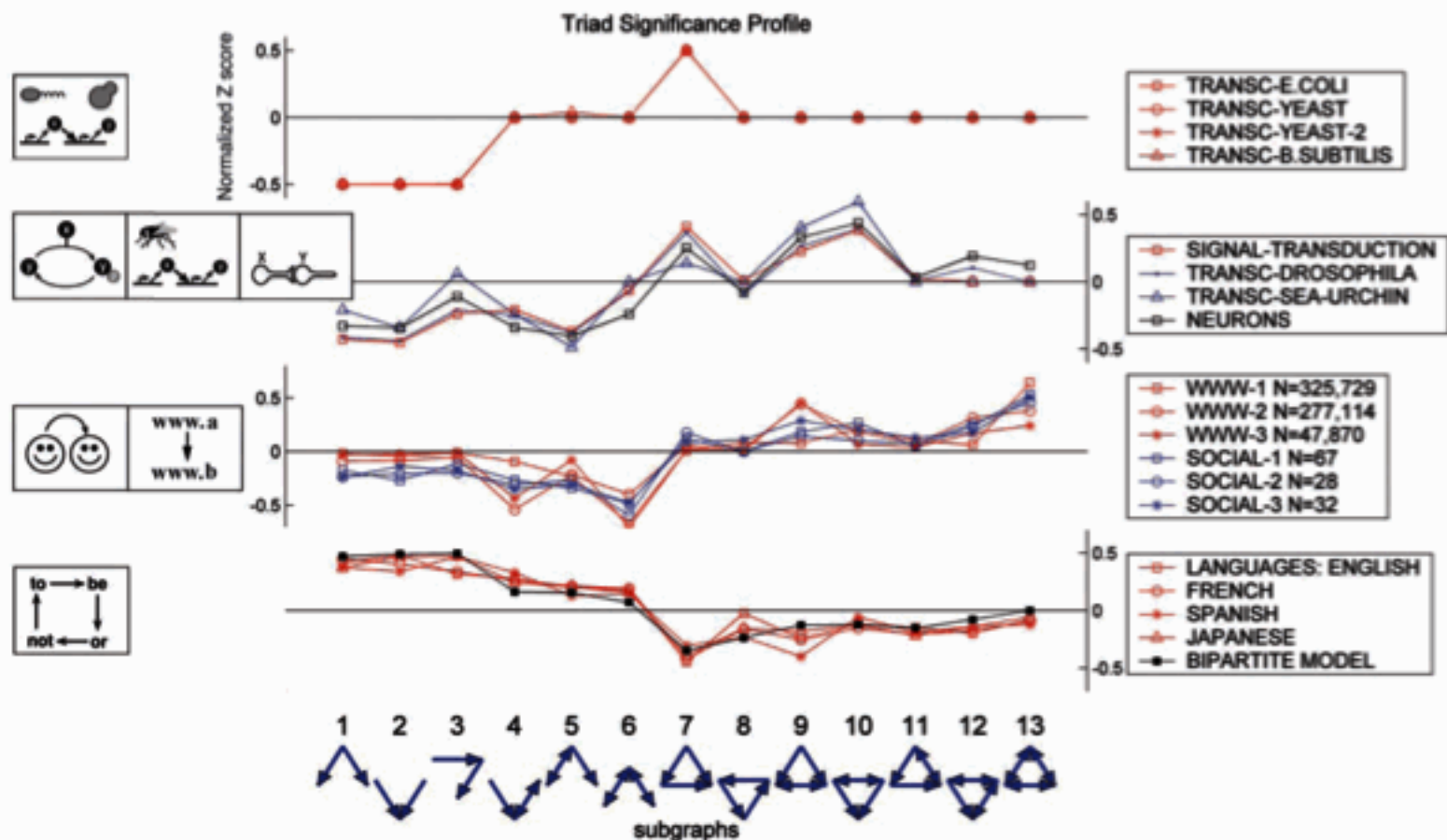


FIG. 2.2. The basic 13 elementary motifs that can be drawn in an oriented graph of three vertices. The table made for four vertices motifs has 199 entries.

Motifs

- ▶ How often some motifs appear compared with a random network



Null models

- ▶ What should we take as reference?
- ▶ Random? How random(ness)?
- ▶ Reshuffling

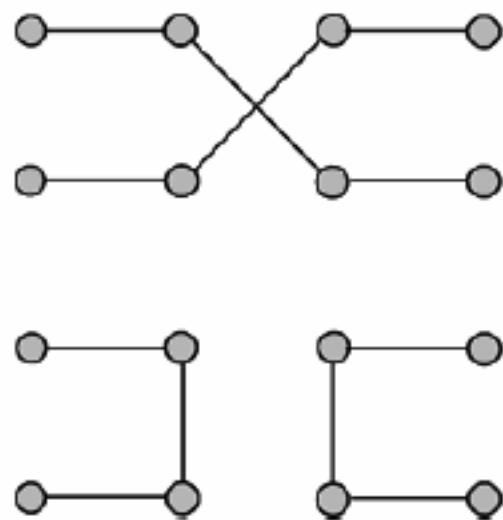


FIG. 2.3. A possible way to rearrange edges keeping the same size, order and degree sequence.

Null models with geographical constraints

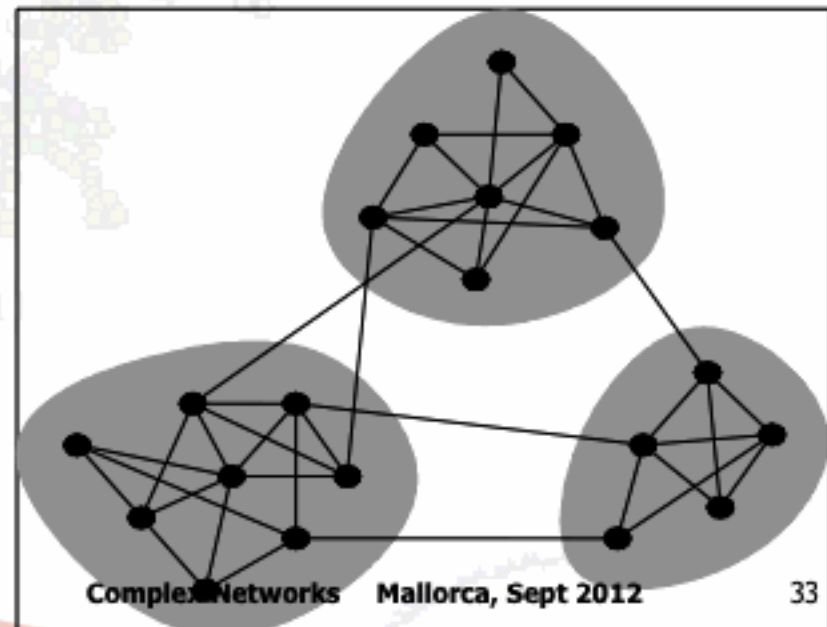
- ▶ Geographic Constraints on Social Network Groups. Onnela J-P, Arbesman S, González MC, Barabási A-L, Christakis NA (2011). PLoS ONE 6(4): e16939.
- ▶ Uncovering space-independent communities in spatial networks. P. Expert, T.S. Evans, V.D. Blondel and R. Lambiotte. PNAS, 108 7663–7668 (2011)



Communities: describing the mesoscale

A technical problem
A management problem

- ▶ Existence of communities or modules in networks
- ▶ Technical issue: finding the best partition
- ▶ Management issue: finding meaningful partitions



Communities

- ▶ Social networks are formed by communities
- ▶ According to:
 - Political reasons
 - Religion
 - Education
 - Scientific disciplines
 -

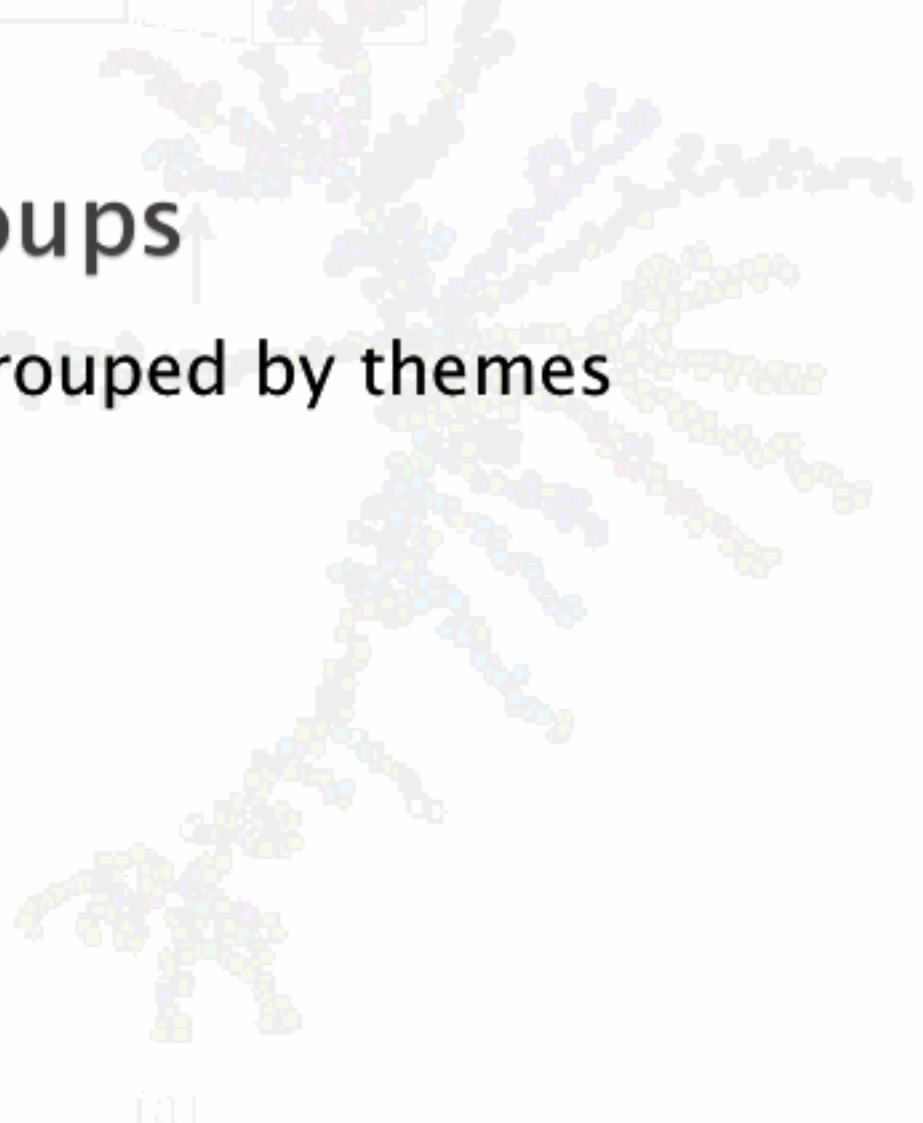


Clusters

- ▶ Technological networks
 - Internet: connections according to geographical proximity
 - Power grids

Thematic groups

- ▶ World-Wide Web: grouped by themes



Modules

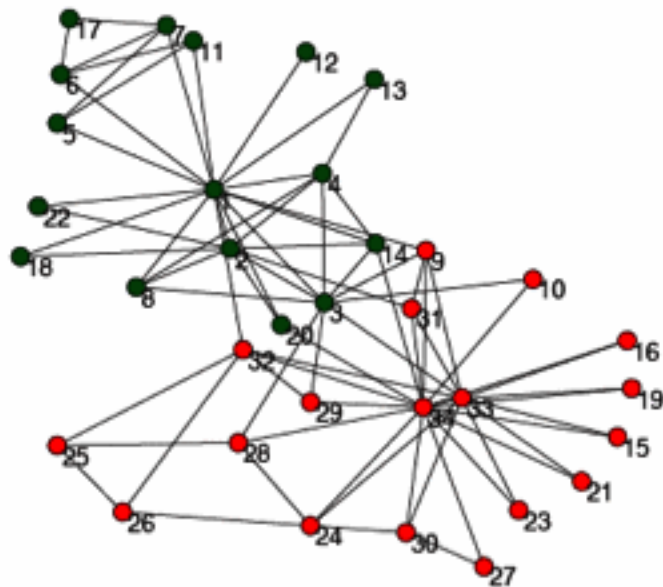
- ▶ Biological networks
- ▶ Gene regulatory networks: groups are functional modules

Technical issue

- ▶ We have to identify the communities
- ▶ How many possible partitions into communities?
- ▶ NP problem to find the best one

Communities: intuitive picture

- ▶ Definition: subsets of nodes that are more densely linked, when compared with the rest of the network



Zachary's Karate club

Partition



- ▶ A partition is a division of the network into groups, communities or clusters
- ▶ The question is: Which of all possible partitions is the best?
- ▶ NP problem
- ▶ Community detection:
 - From computer scientists
 - To statistical physicists (Girvan–Newman, PNAS 99, 7821, 2002)

Quantifying a partition

- **Modularity:**

$$Q = \sum_i (e_{ii} - a_i^2)$$

- e_{ij} : fraction of total links starting at a node in partition i and ending at a node in partition j
- a_i : fraction of links connected to i
- a_i^2 : number of intracommunity links

Methods of community identification

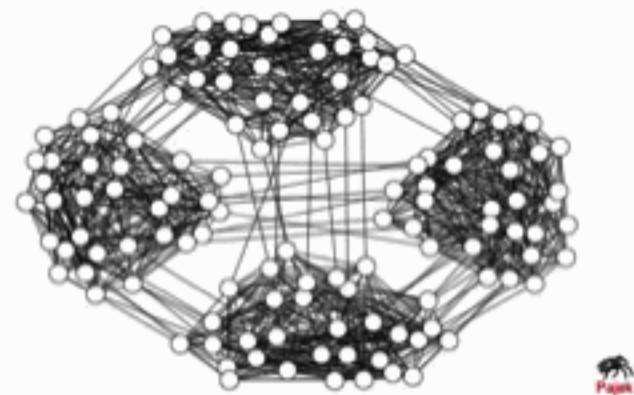
- ▶ L. Danon, J. Duch, A.D-G, A. Arenas J. Stat. Mech. (2005) P09008
 - Link removal methods
 - Agglomerative methods
 - Maximizing modularity
 - Spectral analysis methods
 - Based on physics: resistor networks, q-Potts model
- ▶ **More recent reviews:**
 - [S. Fortunato, *Community detection in graphs*](#) (Phys. Rep. 486, 75–174, 2010)

Computational costs

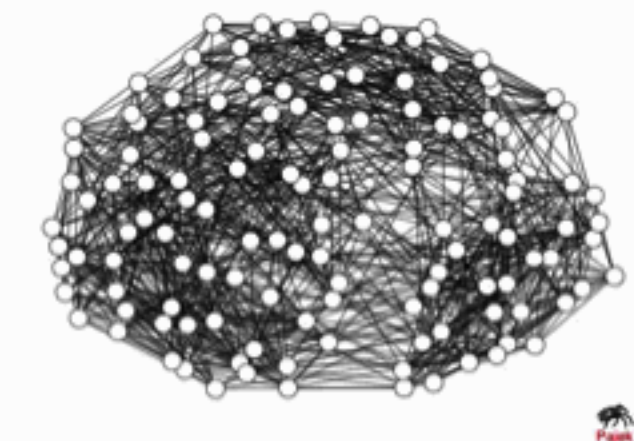
Reference	Alias	Order
(Newman and Girvan, 2004)	NG	$O(m^2n)$
(Girvan and Newman, 2002)	GN	$O(n^2m)$
(Fortunato et al., 2004)	FLM	$O(n^4)$
(Radicchi et al., 2004)	RCCLP	$O(n^2)$
(Newman, 2004b)	NF	$O(n \log^2 n)$
(Donetti and Muñoz, 2004),	DMSA	$O(n^3)$
(Donetti and Muñoz, 2004),	DMCA	$O(n^3)$
(Eckmann and Moses, 2002)	EM	$O(m \langle k^2 \rangle)$
(Zhou and Lipowsky, 2005)	ZL	$O(n^3)$
(Reichardt and Bornholdt, 2004)	RB	unknown
(Bagrow and Boltt, 2004)	BB	$O(n^3)$
(Duch and Arenas, 2005)	DA	$O(n^2 \log n)$
(Capocci et al., 2004)	CSCC	$O(n^2)$
(Wu and Huberman, 2004)	WH	$O(n + m)$

Comparing algorithms

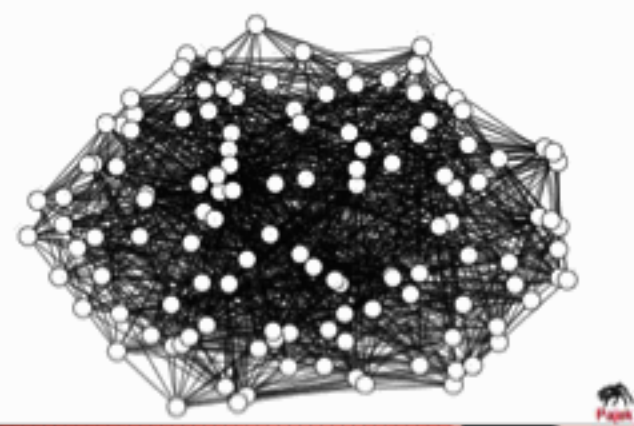
- ▶ *ad-hoc* networks (Newman–Girvan, PRE 69, 026113, 2004)
 - 128 nodes
 - 4 communities of 32 nodes each
 - Each node has 16 links:
 - z_{in} internal nodes within the community
 - z_{out} nodes out of its community



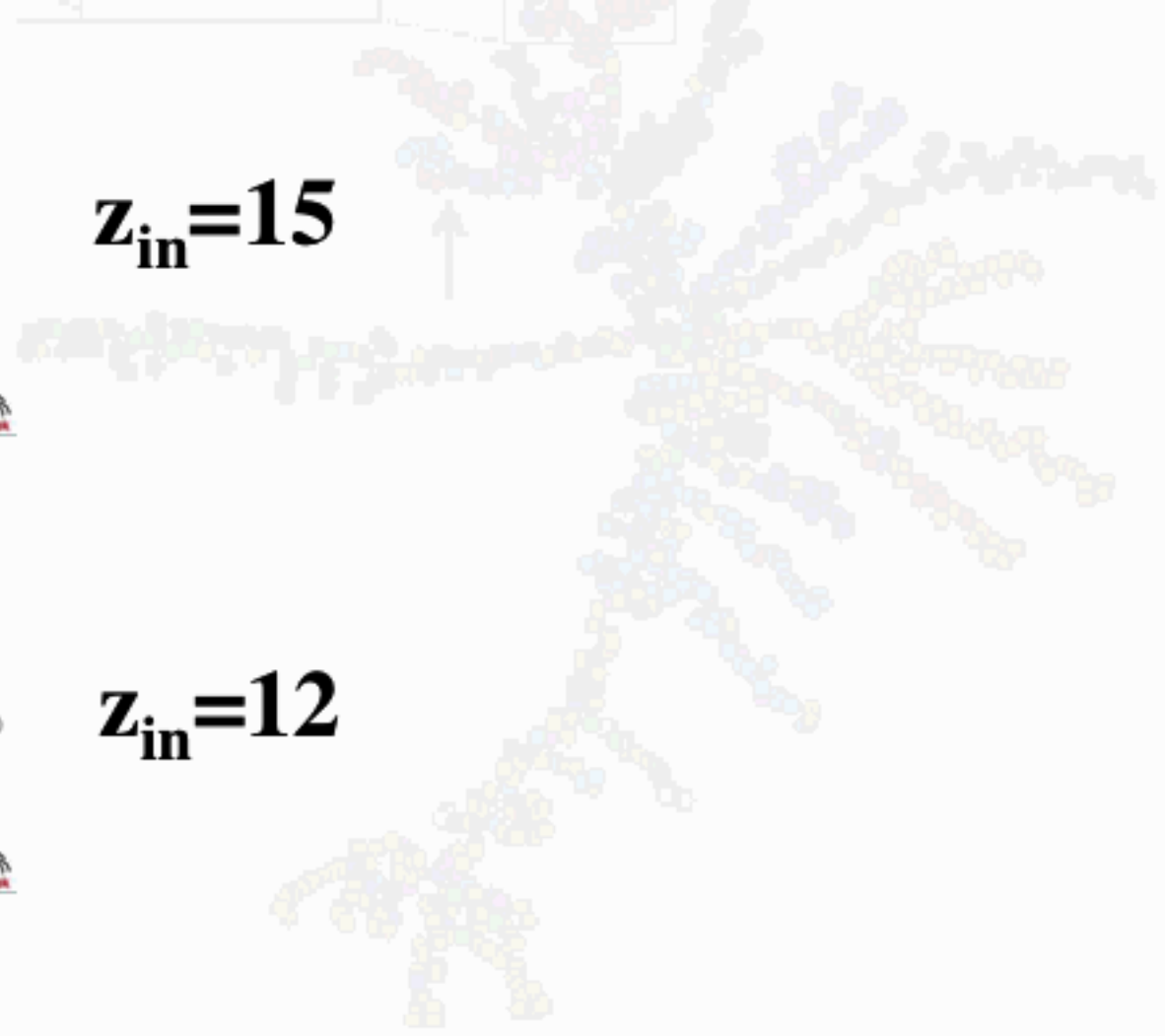
$z_{in}=15$



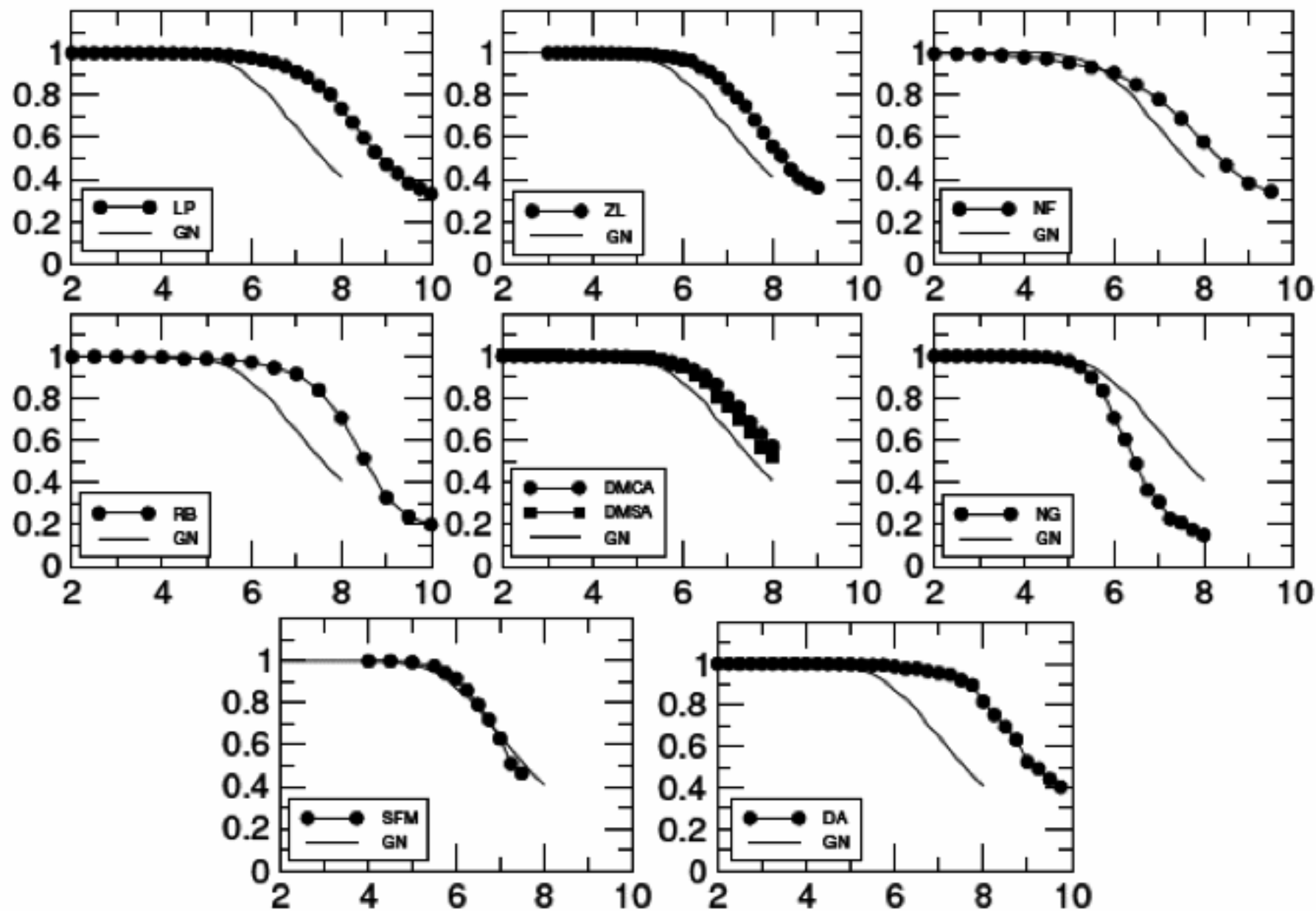
$z_{in}=12$



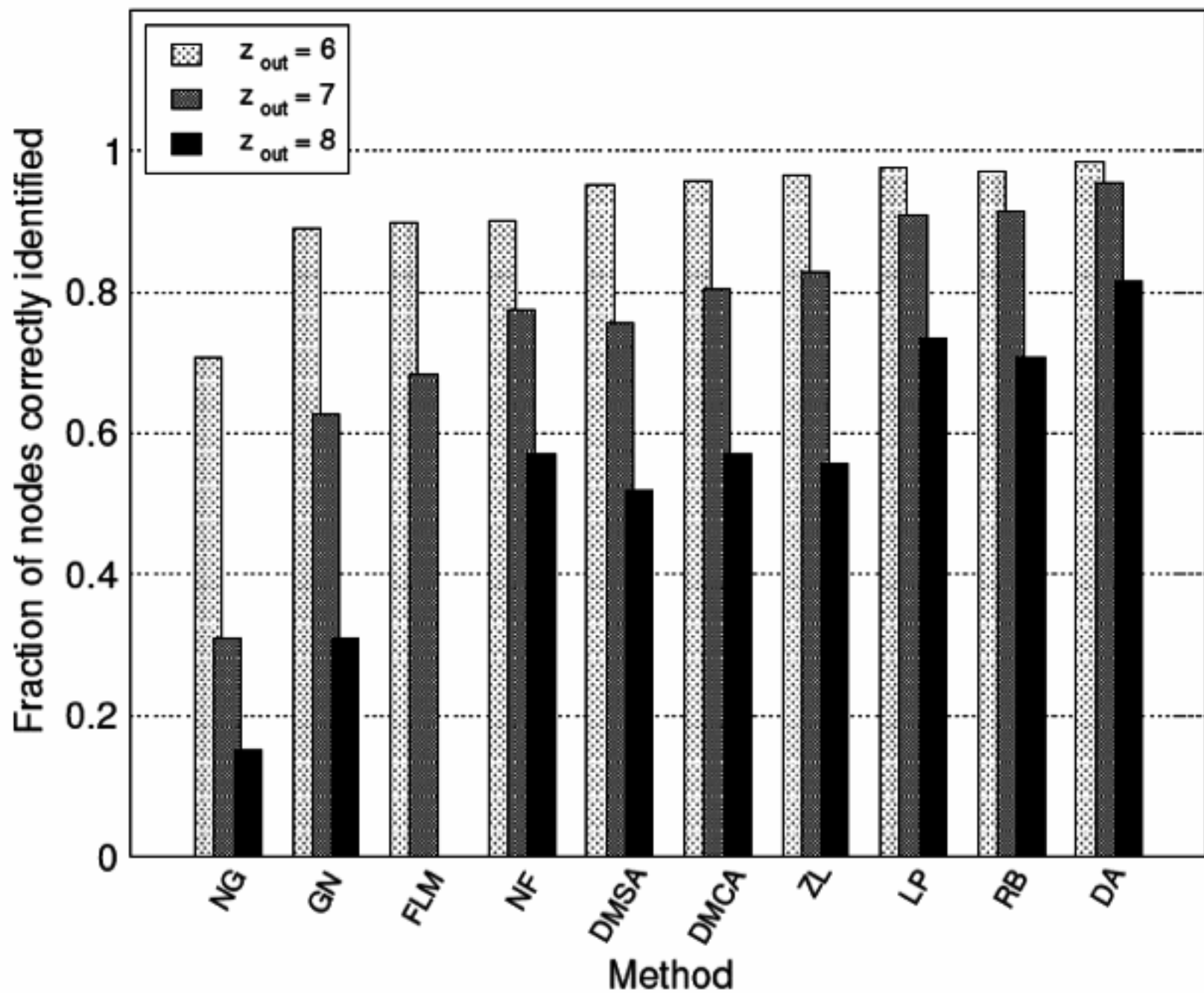
$z_{in}=8$



Fraction of correctly identified nodes



Average number of out links z_{out}



Identifying communities

- ▶ Identifying what communities are
- ▶ Managerial point of view:
 - How a company is organized
 - How powerful is the formed informal chart

Two networks



- ▶ E-mail network at Universitat Rovira i Virgili
- ▶ FisEs

URV



Importance from management

Unravel the real (informal) organization behind the formal chart

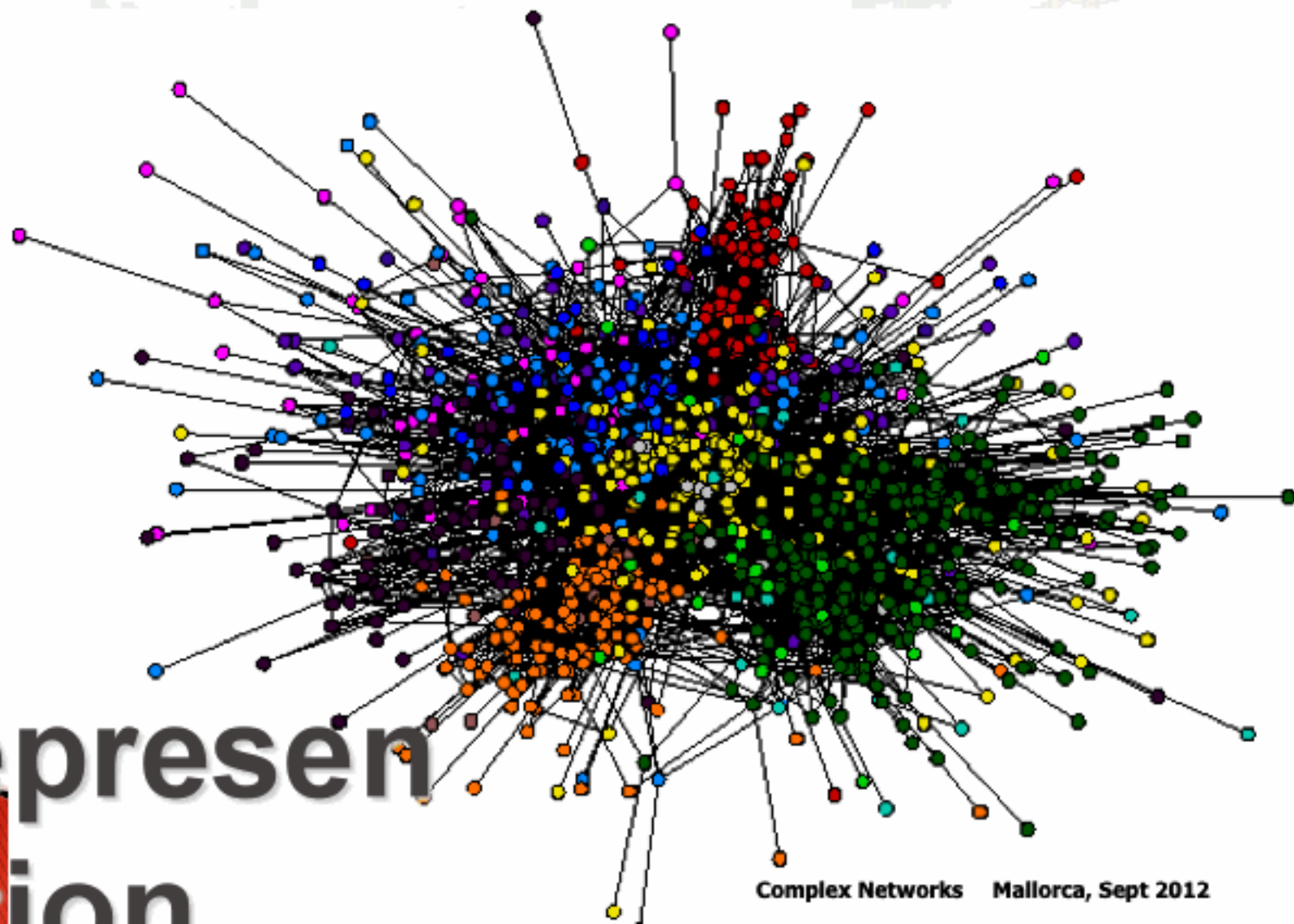
"If the formal organization is the skeleton of a company, the informal is the central nervous system... Complex webs of social ties form every time colleagues communicate and solidify over time into surprisingly stable networks."

D. Krackhardt and J. R. Hanson, Harvard Business Review, 71, 104-113 (1993)

Data acquisition to construct the e-mail network of the URV

- Node => e-mail address
- Link => bidirectional e-mails between nodes (undirected graph)
- Number of users approx. 1700 (professors, technicians, administrators, graduate students)
- We consider only e-mails sent within the University during the first 3 months of 2002 (stable network)
- Non “spam” mail: (neglect >50 recipients)

Email at URV



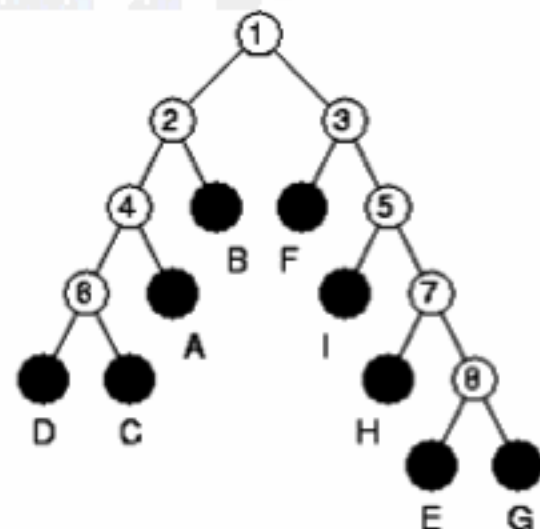
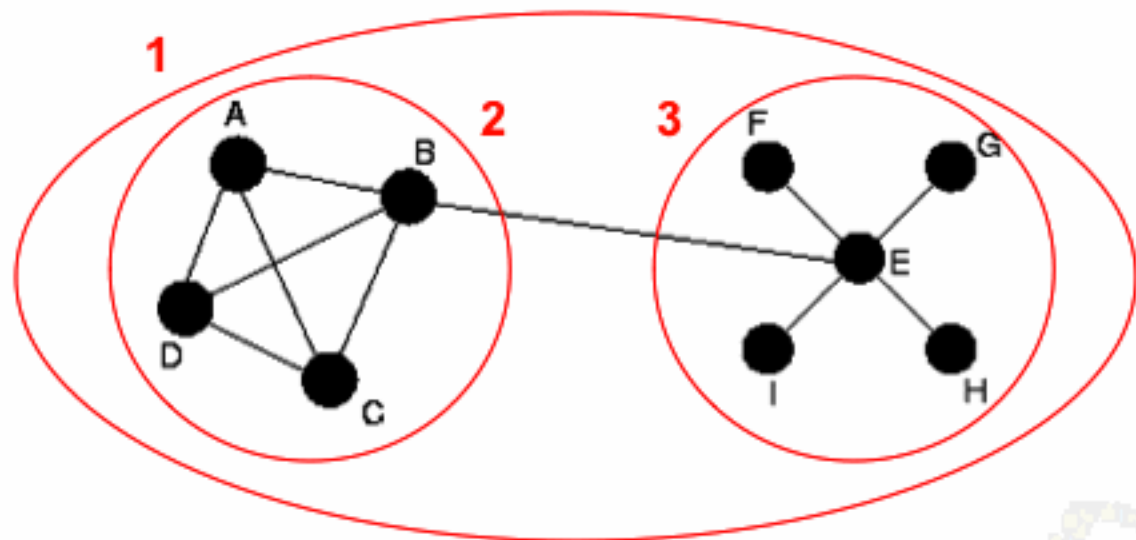
Representen
tation

Community identification: the Girvan & Newman (GN) algorithm*

- **Definition:** Betweenness of a link = # minimum paths connecting pairs of nodes that go through that link
- **Idea in GN algorithm:** The links which connect highly clustered communities have a higher link betweenness. Then cut these links to separate communities.

*Girvan and Newman, PNAS USA 99, 7821–7826

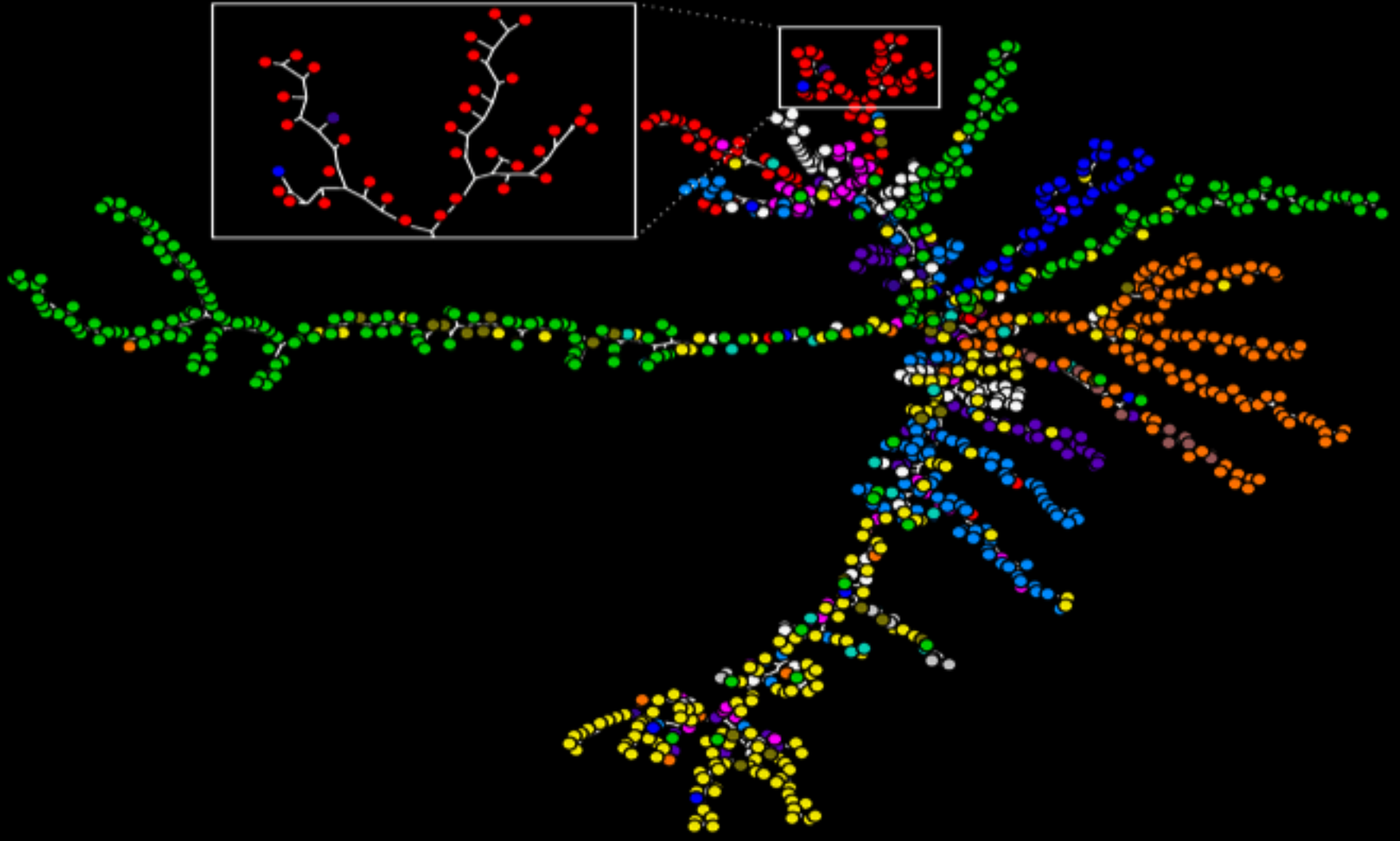
Communities



A network containing two clear communities linked by BE. Since there is no more community structure, the rest of the nodes will be separated one by one generating a binary tree with two branches corresponding to the two communities.

Leaders are at the tips of the branches

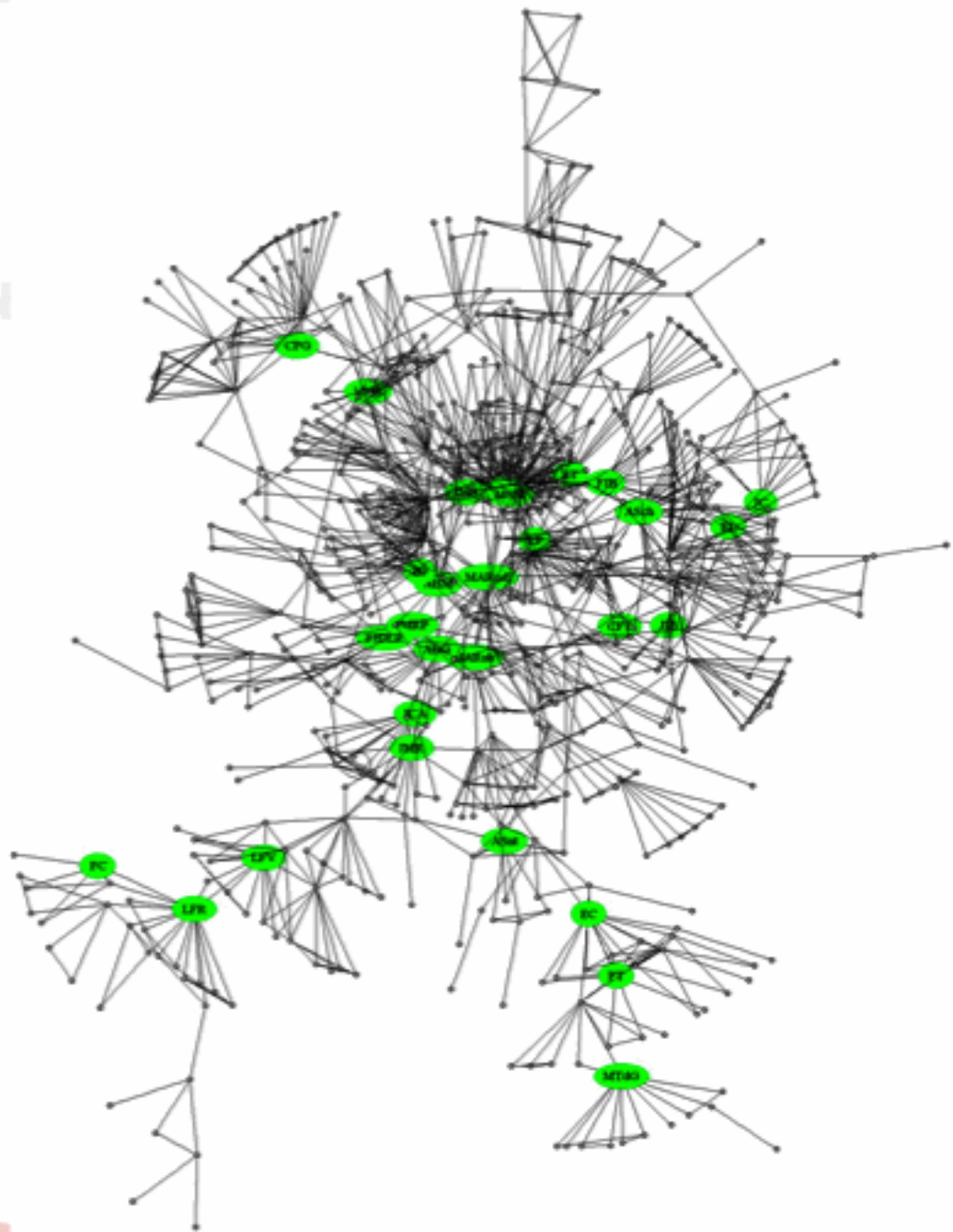
Communities in URV



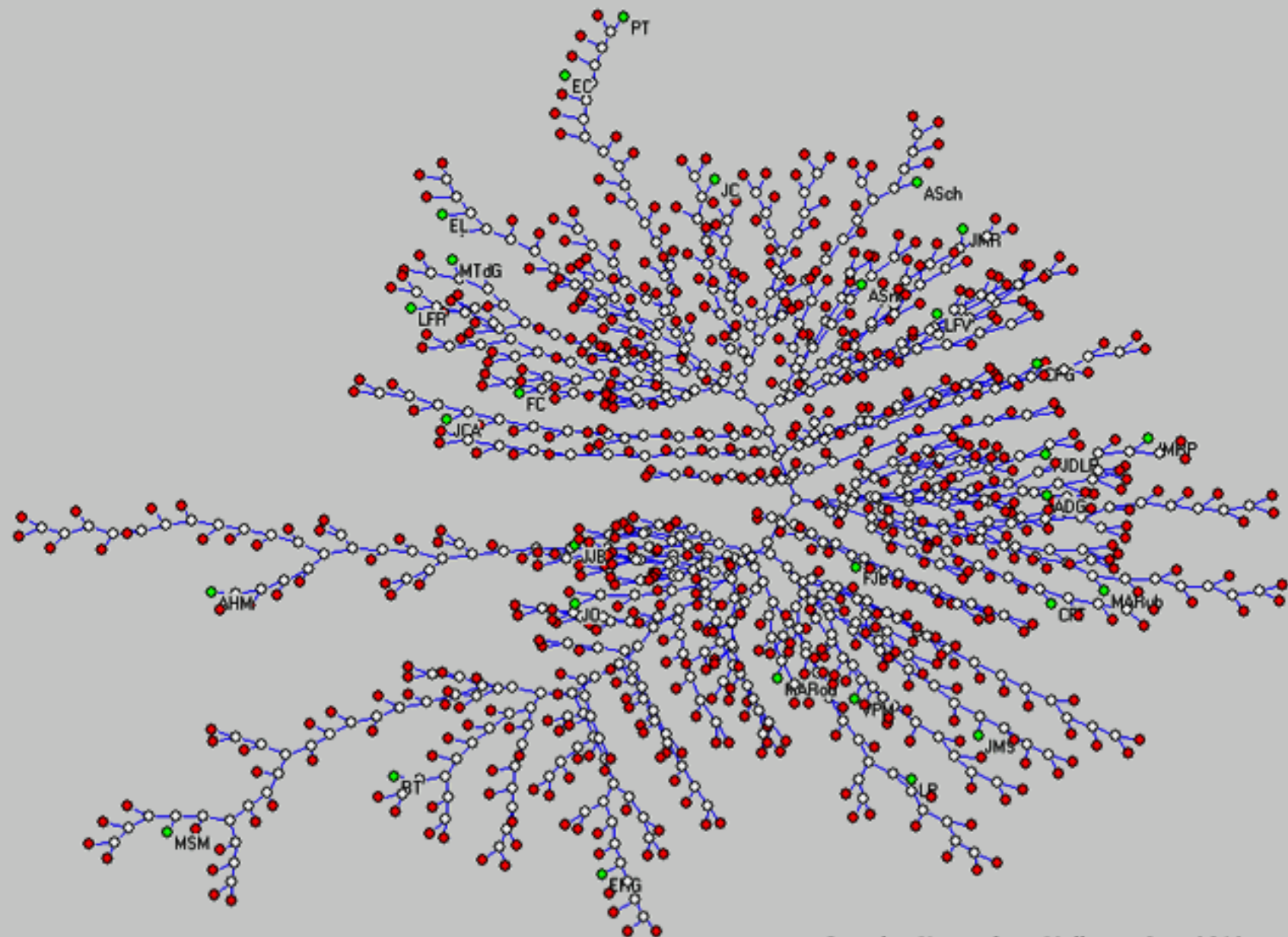
FisEs network

(principal
component)

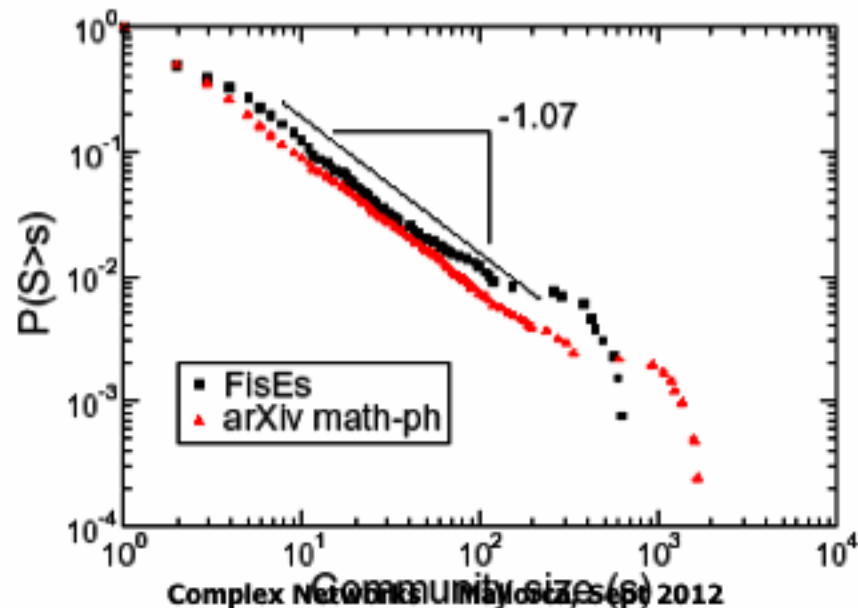
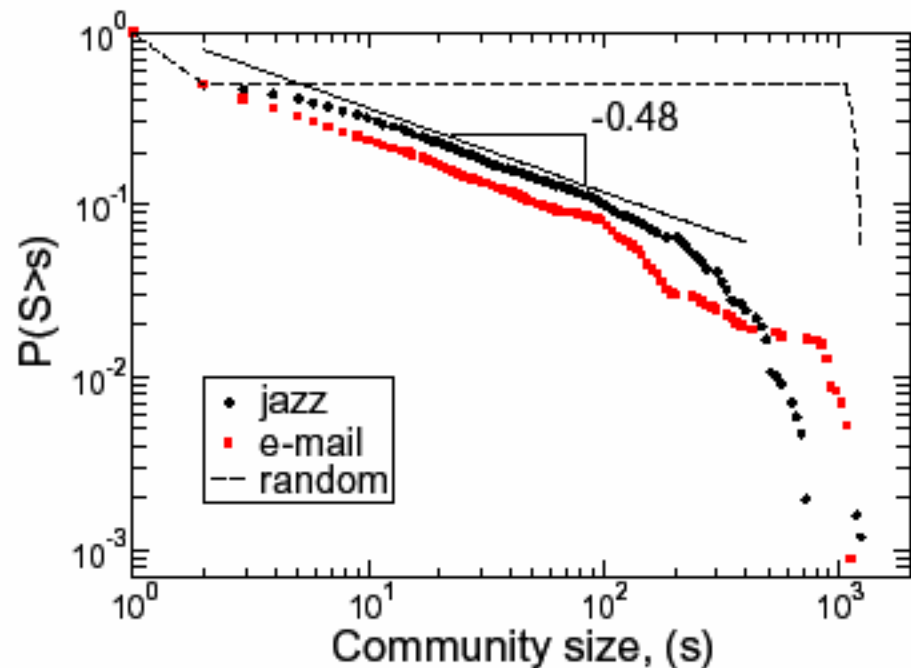
785 total
655 pc (84%)



Community structure



Self-similarity





Models describing simple
properties of complex
networks

1 Erdos–Renyi: random graph model

- ▶ Definition: N labeled nodes connected by n links which are chosen randomly from the $N(N-1)/2$ possible links
- ▶ There are $\binom{N(N-1)/2}{n}$ graphs with N nodes and n links

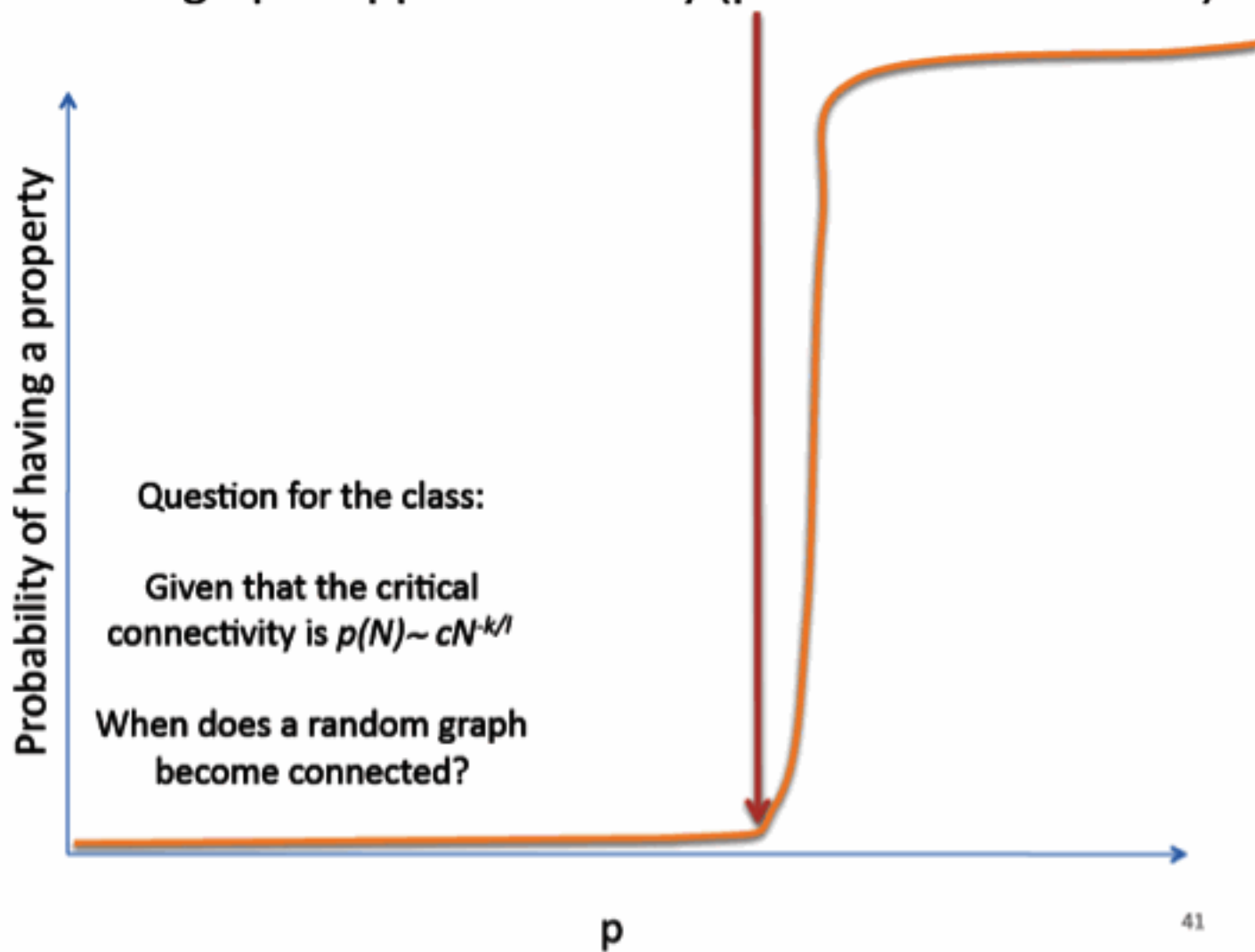
Alternative definition

- ▶ Binomial model: start with N nodes, every pair of nodes being connected with probability p
- ▶ The expected total number of links, n , is a random variable
 - $E(n) = pN(N-1)/2$

Mean connectivity

- ▶ $\langle k \rangle = pN$
- ▶ If $p \propto N^{-1}$ then $\langle k \rangle$ is a constant
- ▶ If $0 < \langle k \rangle < 1$ almost surely all clusters are either trees or clusters containing exactly one cycle
- ▶ At $\langle k \rangle = 1$ the structure changes abruptly. Cycles appear and a giant cluster develops

Subgraphs appear suddenly (percolation threshold)



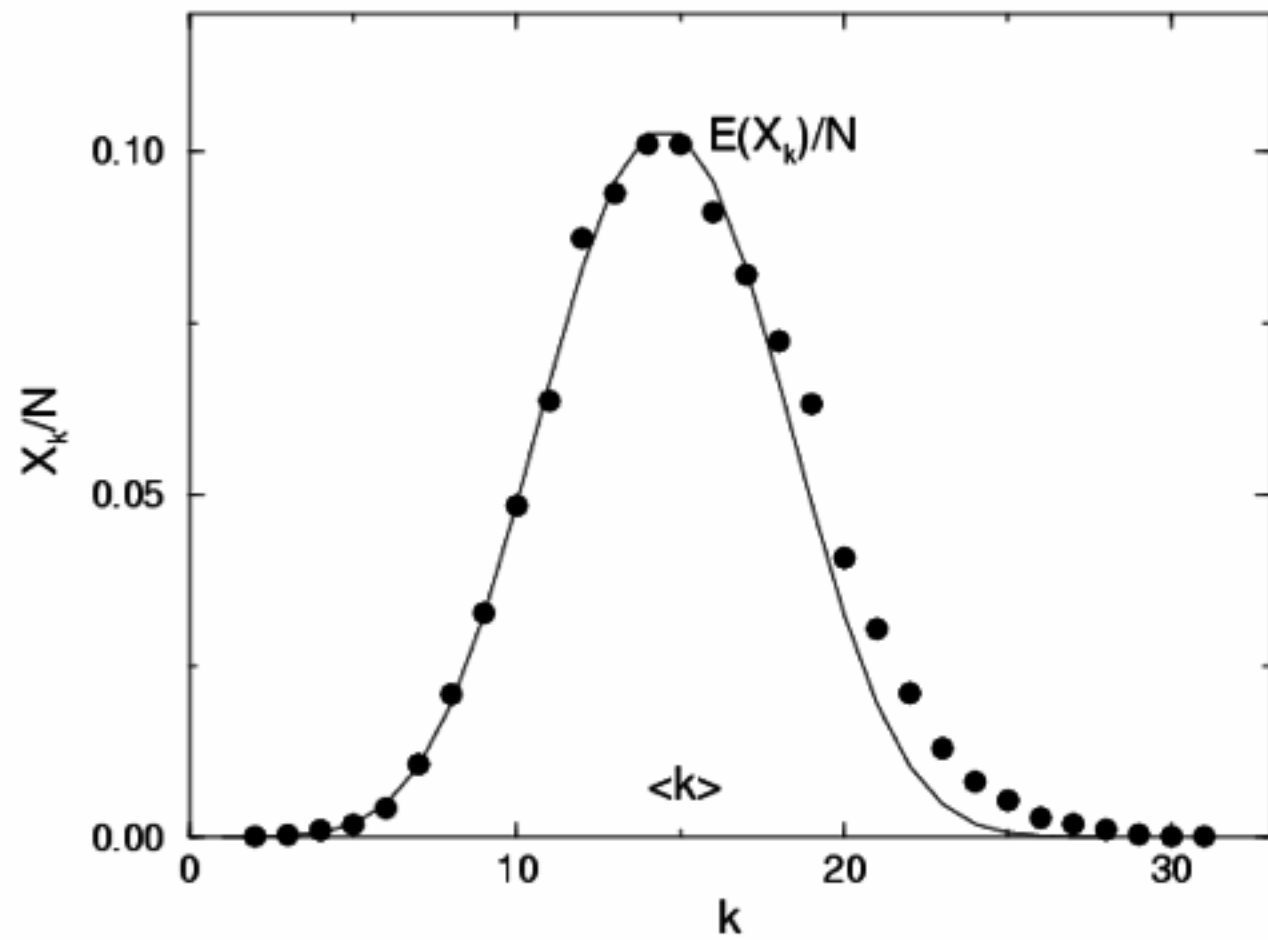
Degree distribution

- ▶ The degree of a node follows a binomial distribution (in a random graph with p)

$$P(k_i = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

- ▶ Probability that a given node has a connectivity k
- ▶ For large N , Poisson distribution

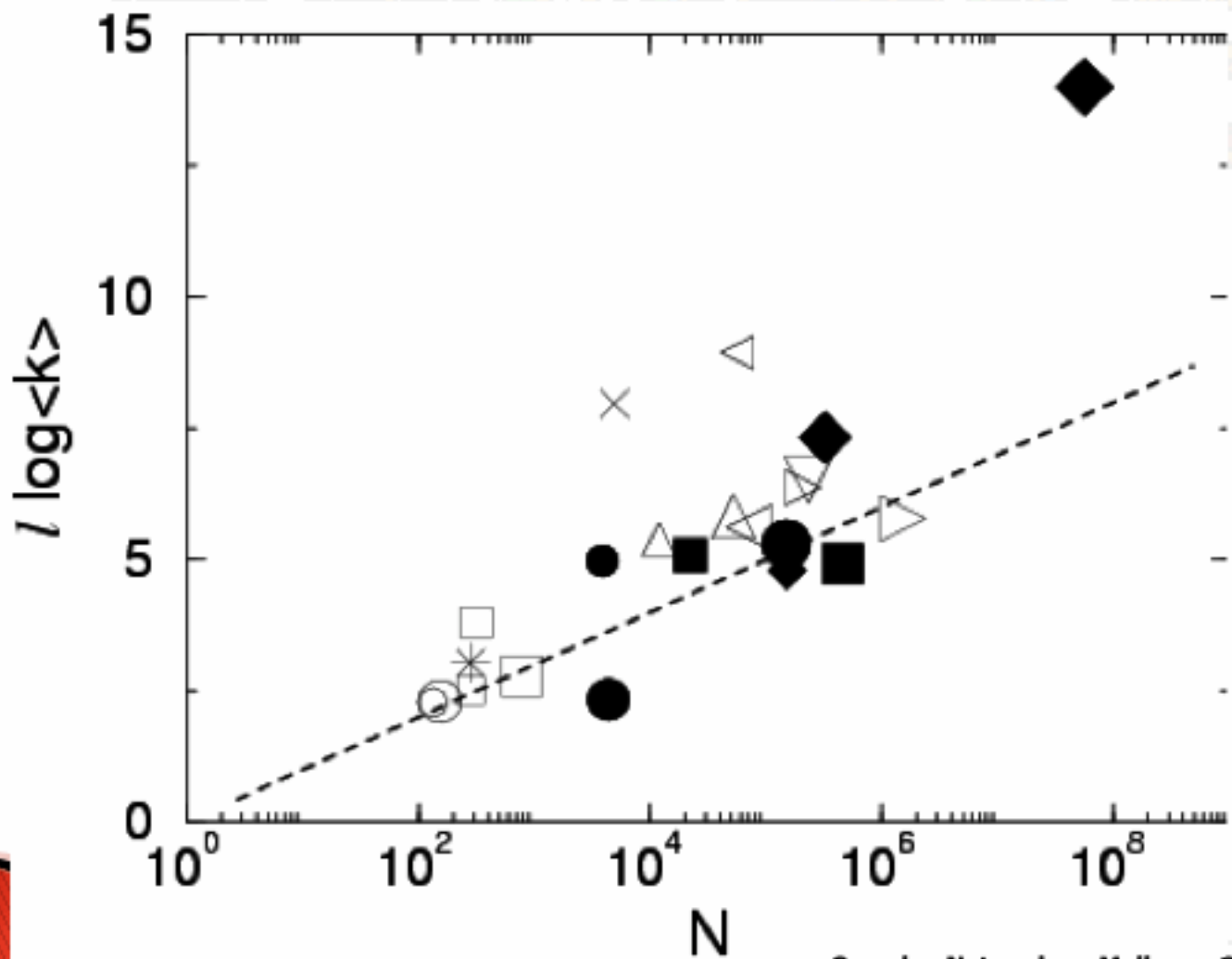
$$P(k) \approx e^{-pN} \frac{(pN)^k}{k!} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



Mean short path

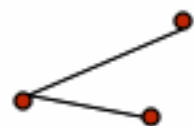
- ▶ Assume that the graph is homogeneous
- ▶ The number of nodes at distance l are $\langle k \rangle^l$
- ▶ How to reach the rest of the nodes?
- ▶ l_{rand} to reach all nodes $\Rightarrow k^l = N$

$$l_{\text{rand}} \approx \frac{\ln N}{\ln \langle k \rangle} \approx \frac{\ln N}{\ln pN}$$



Clustering coefficient

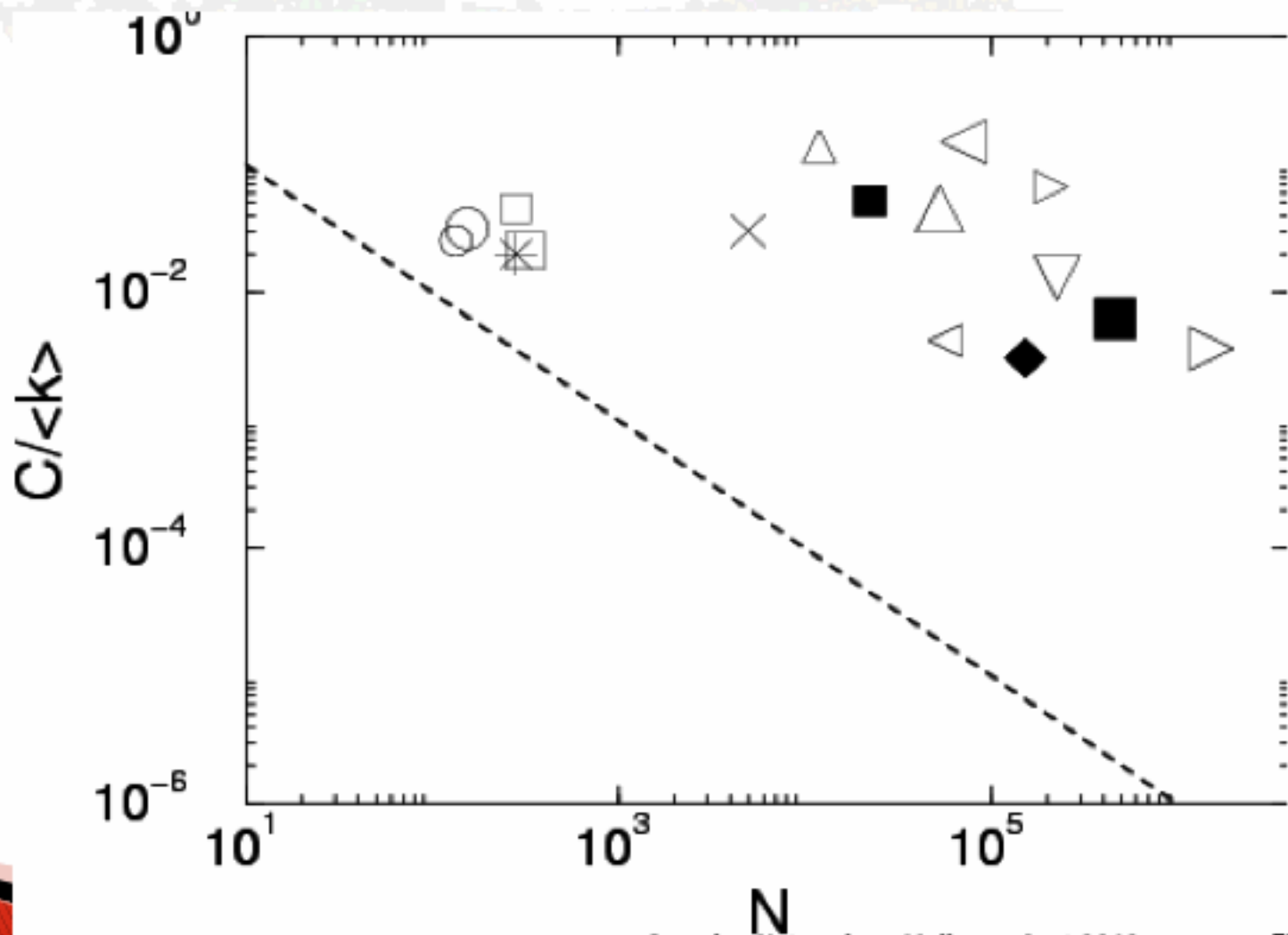
- ▶ Probability that two nodes are connected (given that they are connected to a third)?



$$C_{rand} = p = \frac{\langle k \rangle}{N}$$

$$\frac{C_{rand}}{\langle k \rangle} \approx \frac{1}{N}$$

while it is constant for real networks



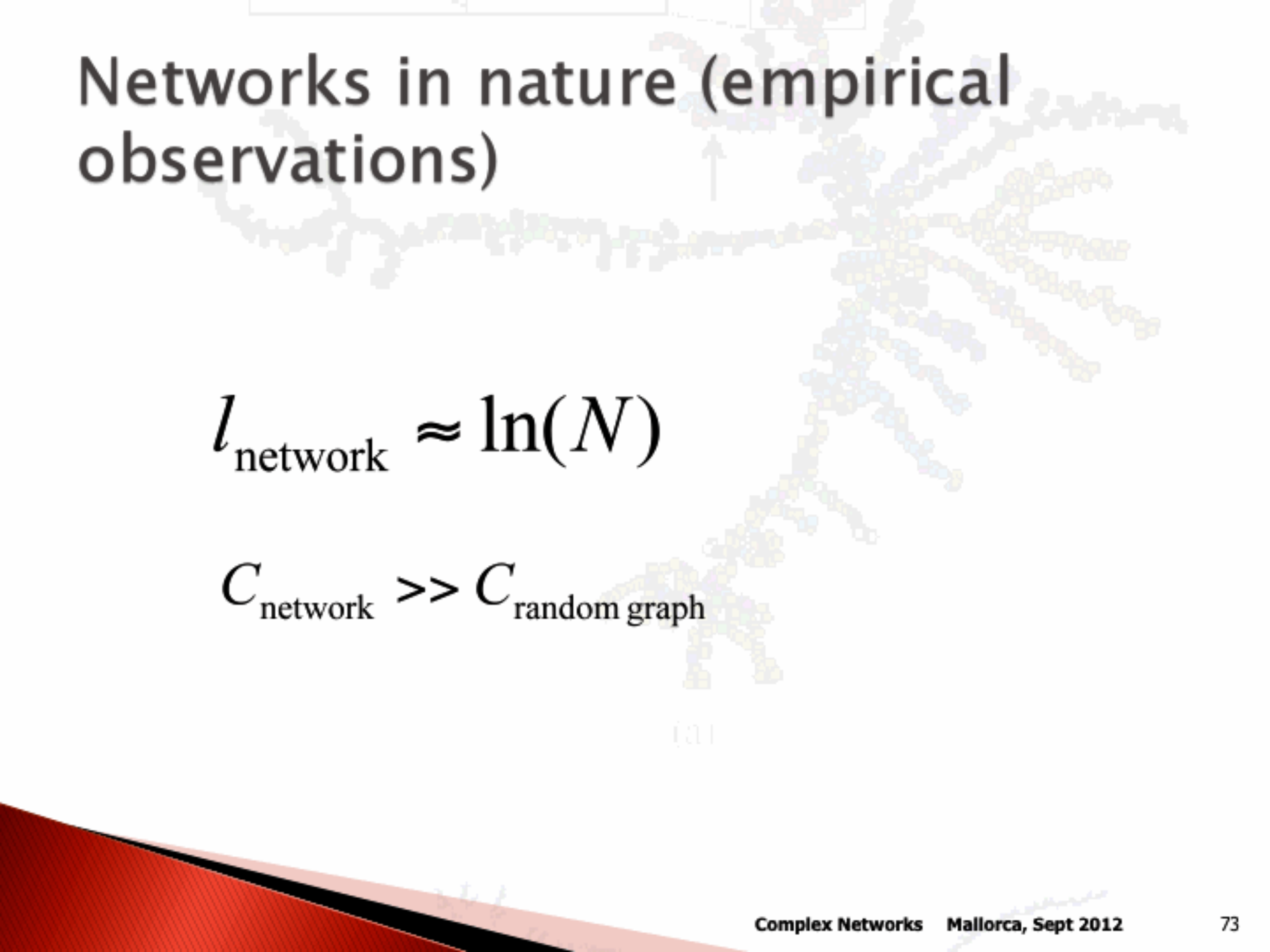
Generalized random graphs

- ▶ Any degree distribution, clustering, ...

2 Watts–Strogatz: small-world model

- ▶ Small world: the average shortest path length in a real network is small
- ▶ Six degrees of separation (Milgram, 1967)
- ▶ Local neighborhood + long-range friends
- ▶ A random graph is a small world

Networks in nature (empirical observations)



$$l_{\text{network}} \approx \ln(N)$$

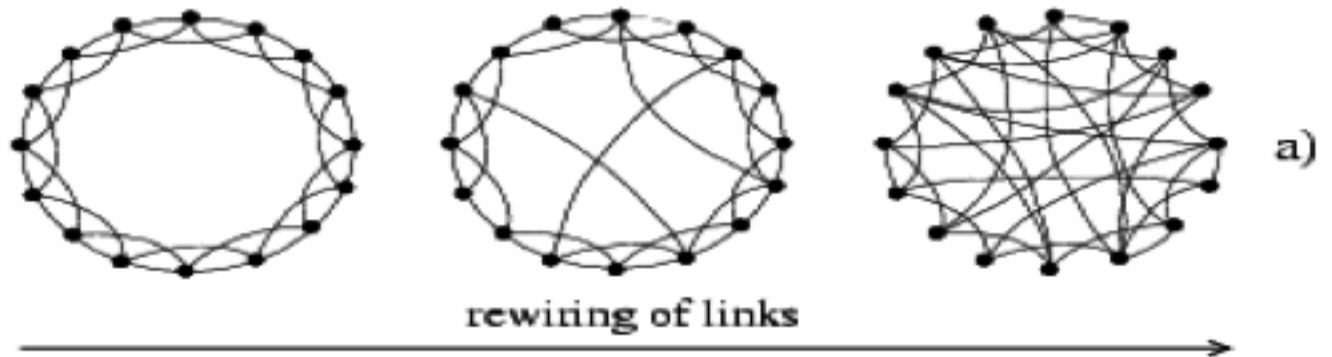
$$C_{\text{network}} \gg C_{\text{random graph}}$$

(a)

Model proposed

- ▶ Crossover from regular lattices to random graphs
- ▶ Tunable
- ▶ Small world network with (simultaneously):
 - Small average shortest path
 - Large clustering coefficient (not obeyed by RG)

Two ways of constructing



Original model

- ▶ Each node has $K \geq 4$ nearest neighbors (local)
- ▶ Probability p of rewiring to randomly chosen nodes
- ▶ p small: regular lattice
- ▶ p large: classical random graph

$p=0$ Ordered lattice

$$l \approx \frac{N}{2K} \gg 1$$

$$C = \frac{3(K-2)}{4(K-1)}$$

(a)

$p=1$ Random graph

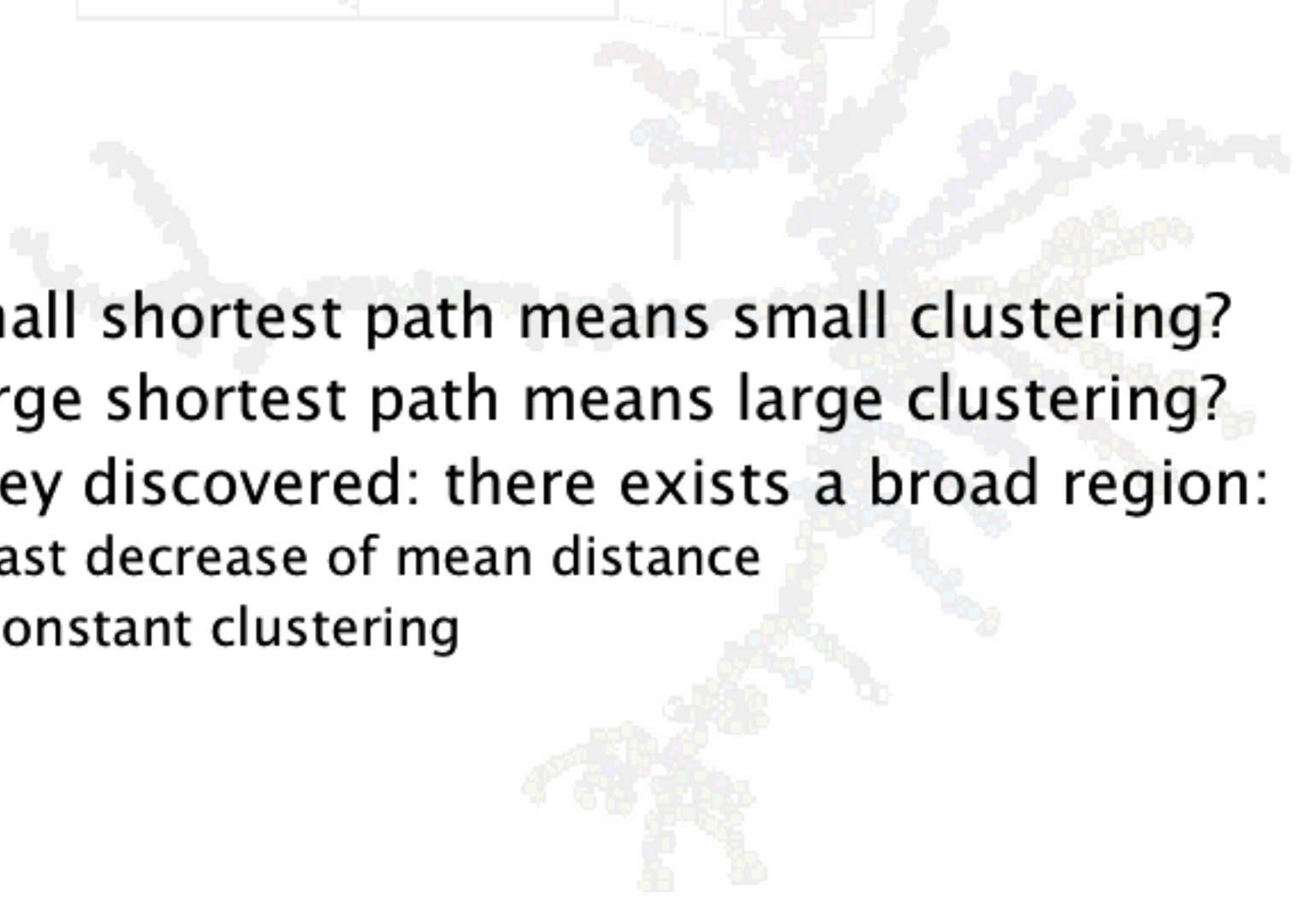
$$l \approx \frac{\ln N}{\ln K}$$

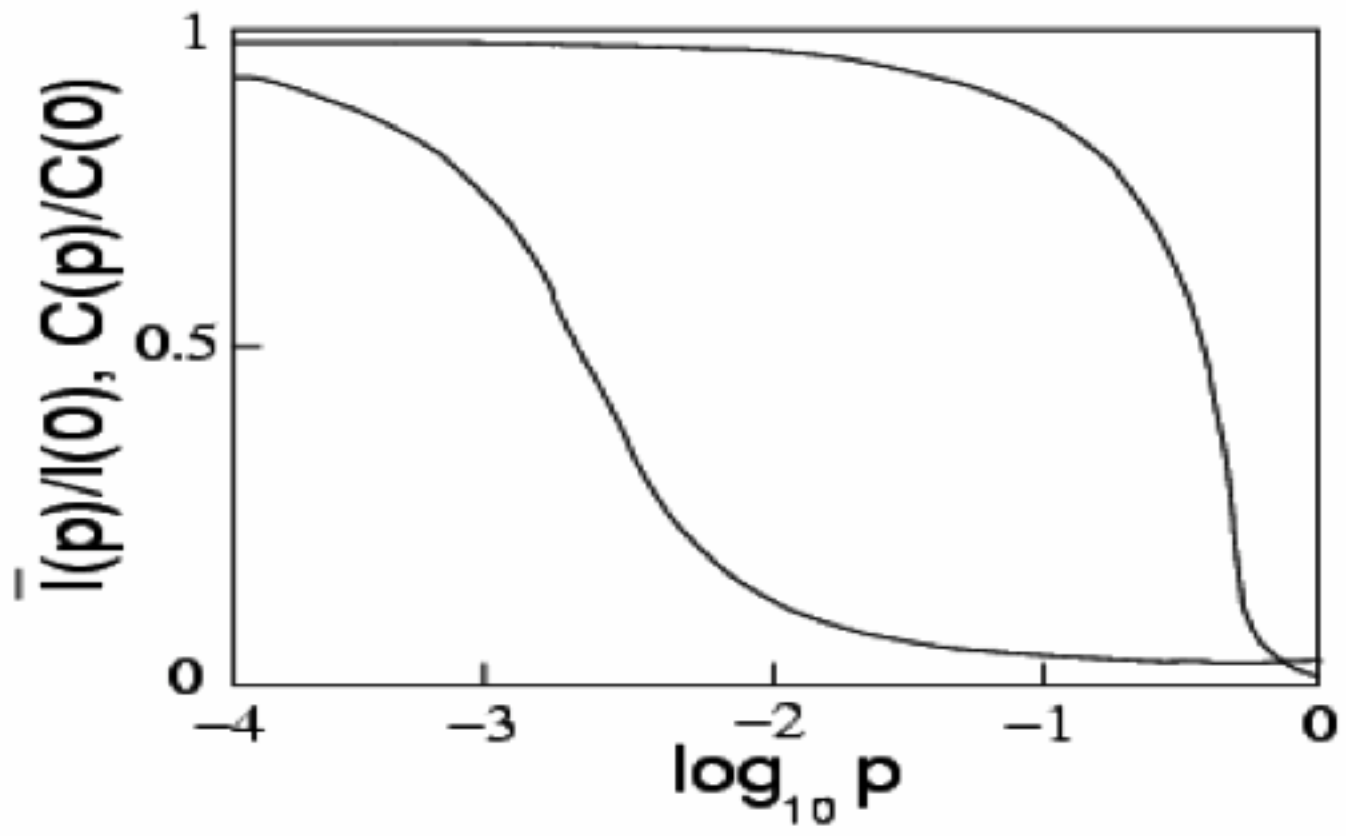
small

$$C \approx \frac{K}{N}$$

small

(a)

- 
- ▶ Small shortest path means small clustering?
 - ▶ Large shortest path means large clustering?
 - ▶ They discovered: there exists a broad region:
 - Fast decrease of mean distance
 - Constant clustering

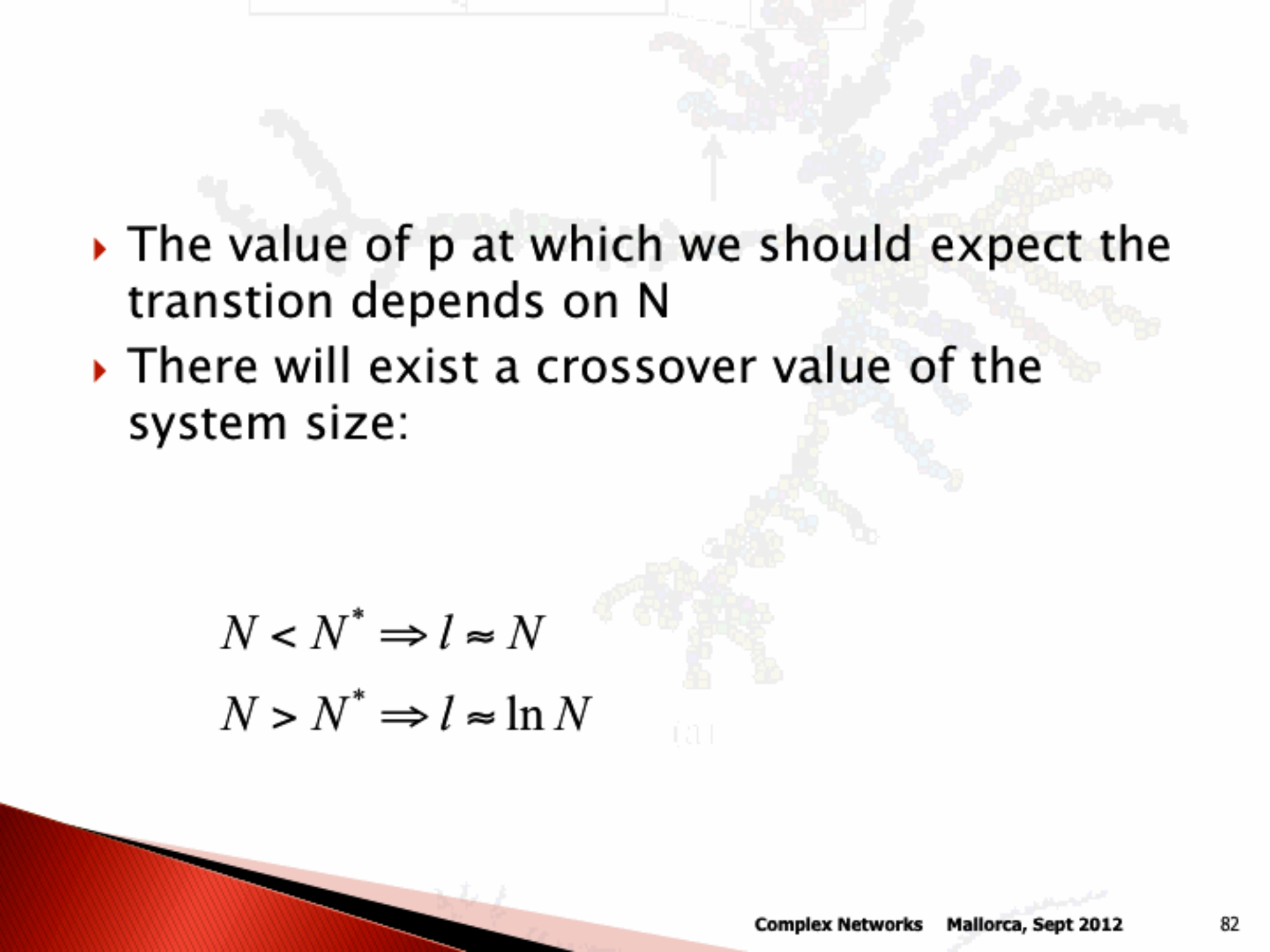


Average shortest path

$$l(p \rightarrow 0) \approx N$$

$$l(p \rightarrow 1) \approx \ln N$$

- ▶ Rapid drop of l , due to the appearance of short-cuts between nodes
- ▶ It starts to decrease when $p \geq 2/NK$ (existence of one short cut)

- 
- ▶ The value of p at which we should expect the transition depends on N
 - ▶ There will exist a crossover value of the system size:

$$N < N^* \Rightarrow l \approx N$$

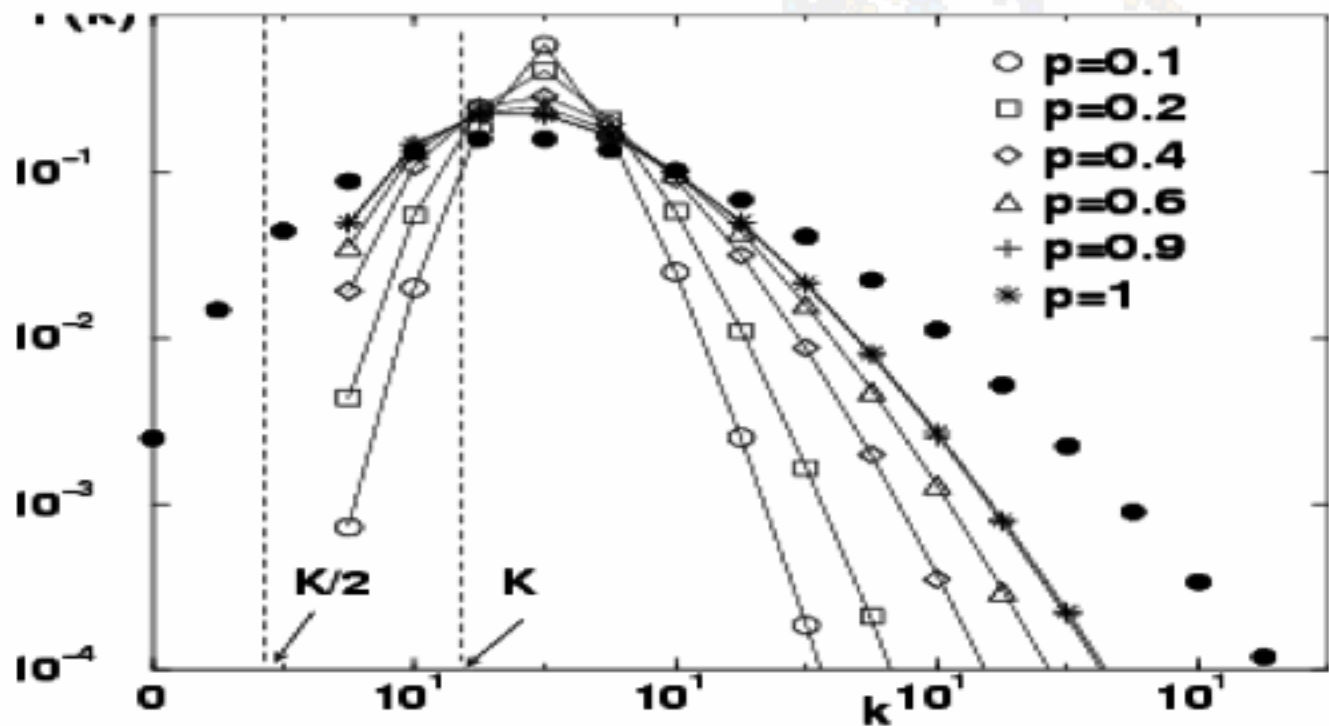
$$N > N^* \Rightarrow l \approx \ln N$$

Degree distribution

- ▶ $p=0$ delta-function
- ▶ $p>0$ broadens the distribution
- ▶ Edges left in place with probability $(1-p)$
- ▶ Edges rewired towards i with probability $1/N$

$$P(k) = \sum_{n=0}^{f(k,K)} C_{K/2}^n (1-p)^n p^{K/2-n} \frac{(pK/2)^{k-K/2-n}}{(k-K/2-n)!} e^{-pK/2}$$

for $k \geq K/2$, where $f(k,K) = \min(k-K/2, K/2)$

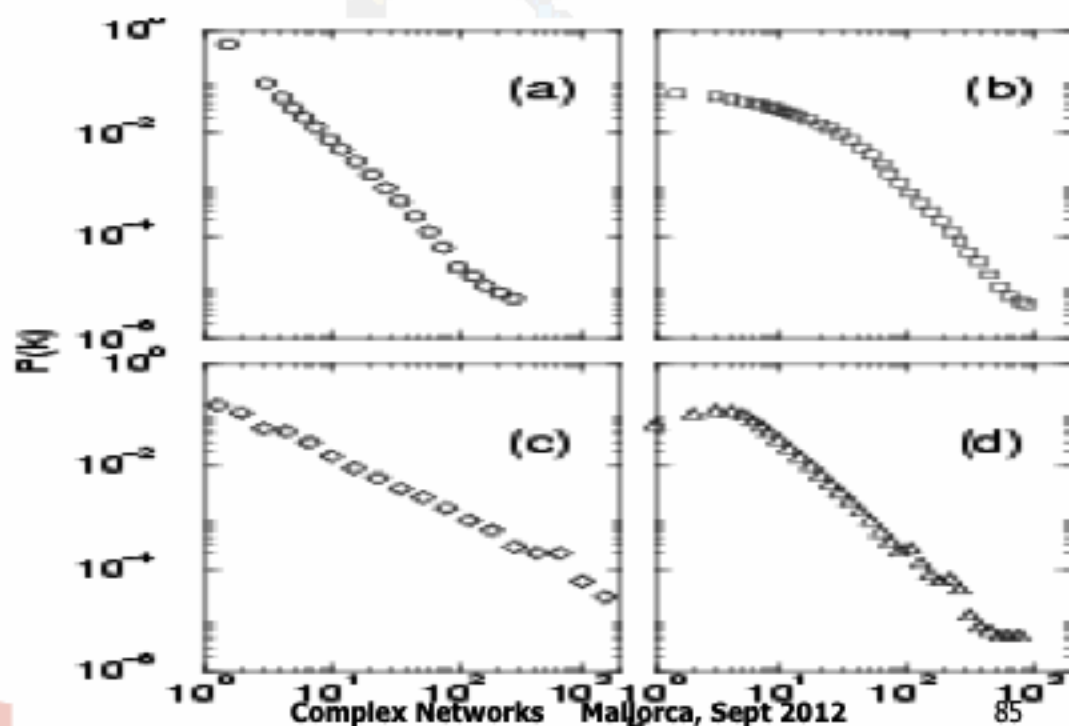


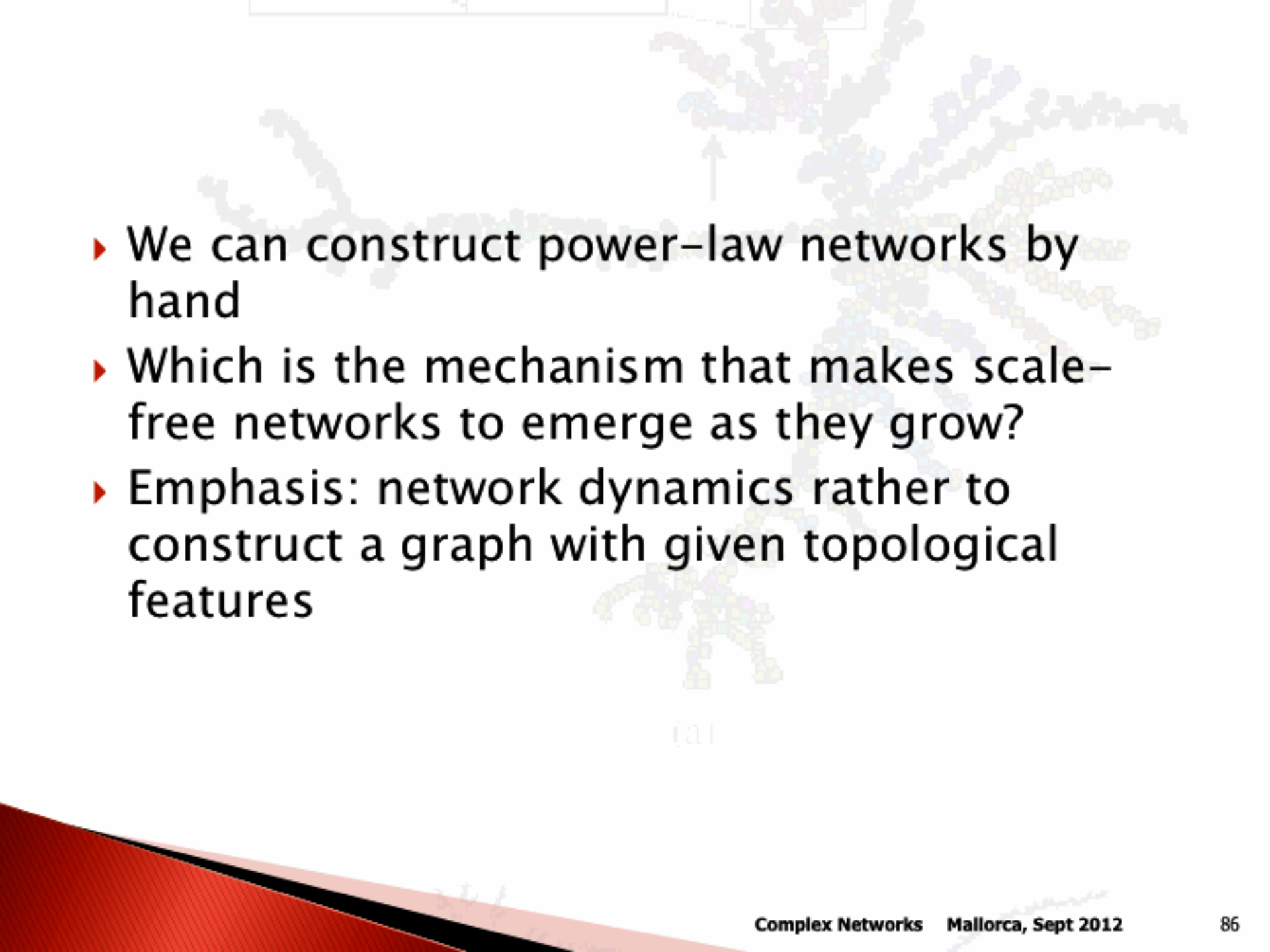
only one edge is rewired

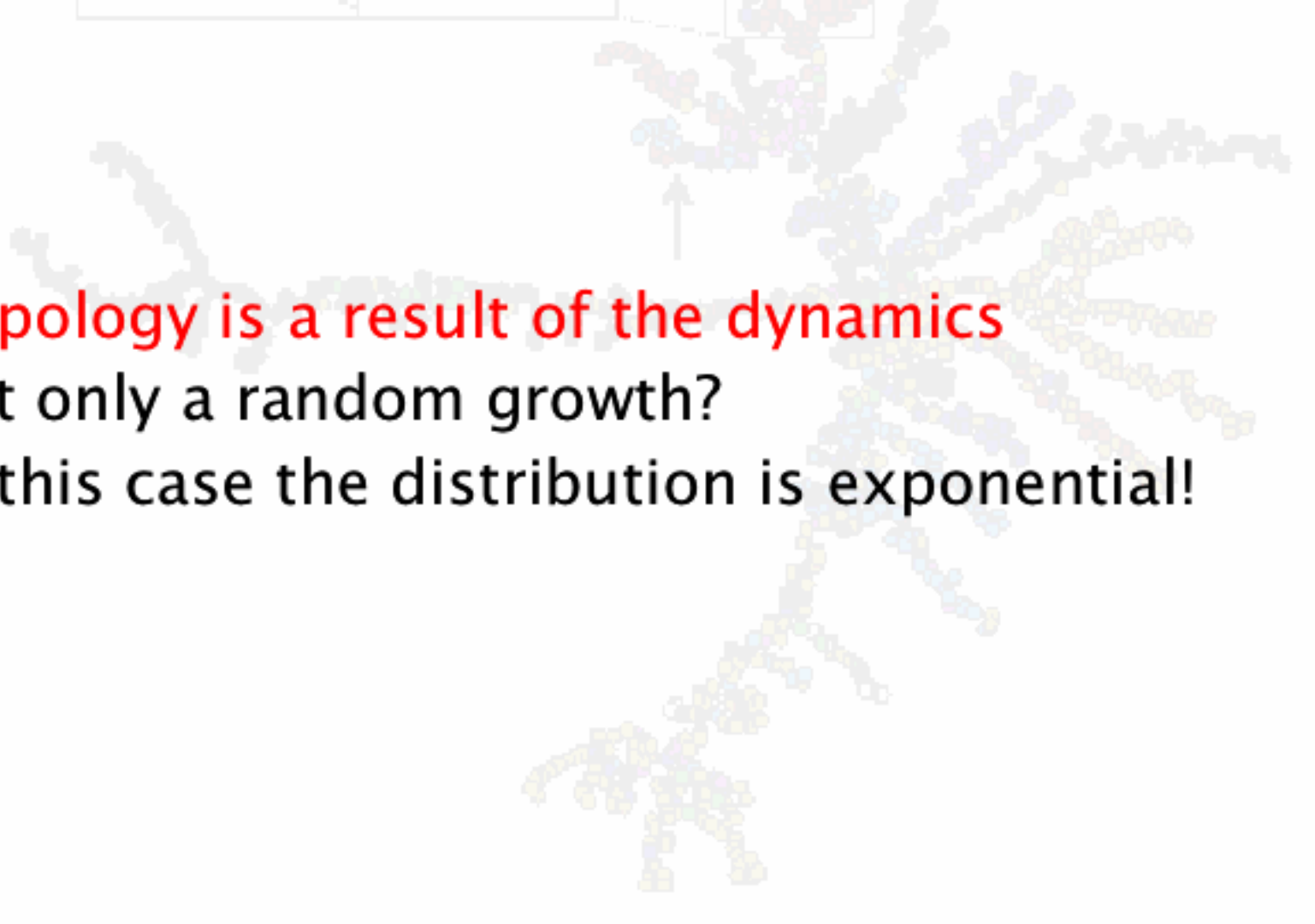
exponential decay, all nodes have similar number of links

3 Barabasi–Albert: scale-free network

- ▶ Many large networks are scale free
- ▶ The degree distribution has a power-law behavior for large k (far from a Poisson distribution)
- ▶ Random graph theory and the Watts–Strogatz model cannot reproduce this feature



- 
- ▶ We can construct power-law networks by hand
 - ▶ Which is the mechanism that makes scale-free networks to emerge as they grow?
 - ▶ Emphasis: network dynamics rather to construct a graph with given topological features

- 
- ▶ **Topology is a result of the dynamics**
 - ▶ But only a random growth?
 - ▶ In this case the distribution is exponential!

Barabasi–Albert model (1999)

- ▶ Two generic mechanisms common in many real networks
 - Growth (www, research literature, ...)
 - Preferential attachment (idem): attractiveness of popularity
- ▶ The **two** are necessary

Growth

- ▶ $t=0$, m_0 nodes
- ▶ Each time step we add a new node with m ($\leq m_0$) edges that link the new node to m different nodes already present in the system

Preferential attachment: rich gets richer

- ▶ When choosing the nodes to which the new connects, the probability Π that a new node will be connected to node i depends on the degree k_i of node i

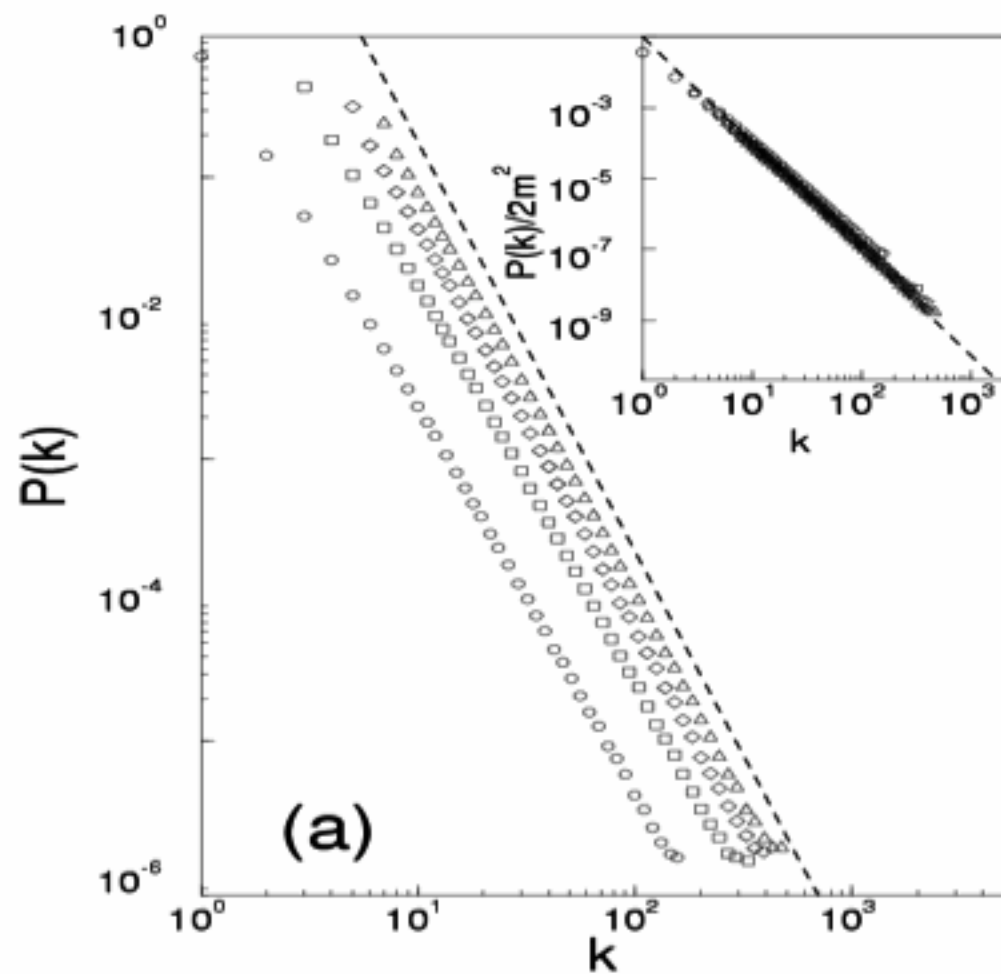
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

Linear attachment (more general models)
Sum over all existing nodes

Numerical simulations

- ▶ Power-law $P(k) \approx k^{-\gamma}$ $\gamma_{SF} = 3$
- ▶ The exponent does not depend on m (the only parameter of the model)

(a)



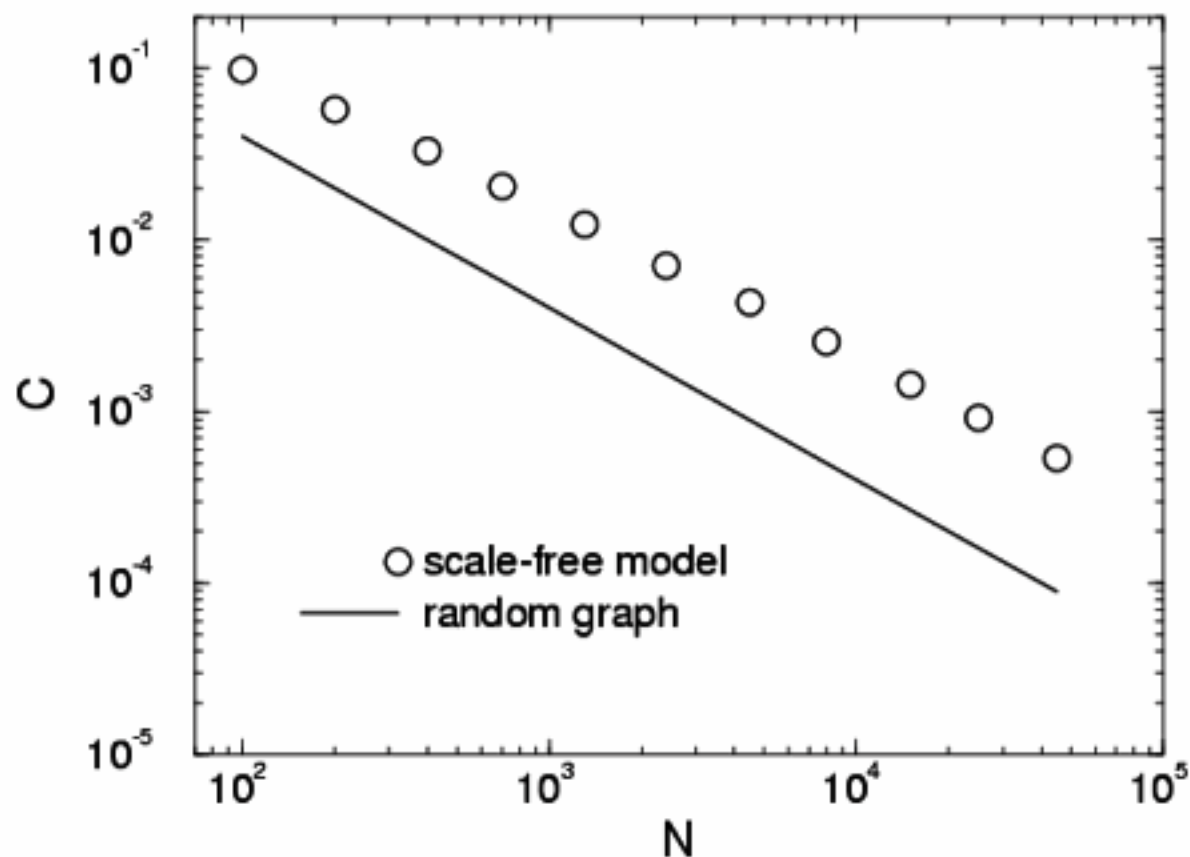
$\gamma=3$. different m 's. $P(k)$ changes. γ not

Degree distribution

- ▶ Analytically

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}$$

Clustering coefficient



5 times larger

$$C_{\text{SF}} \sim N^{-0.75}$$

$$C_{\text{RG}} = \langle k \rangle N^{-1}$$

SW: C is independent of N

Hubs

