The formation and the structure of social networks: theory and empirics

Nicolas Carayol

Université Paris Sud, ADIS

esnie – May 2007- Institut scientifique de Cargèse

Outline of the talk

1. What is a network and various applications
2. Network data and drawing issues with Pajek
3. Random networks
4. Scale free networks
5. Small worlds networks
6. Strategic network formation games
7. Efficient vs. emergent networks
8. The strategic formation of co-invention networks

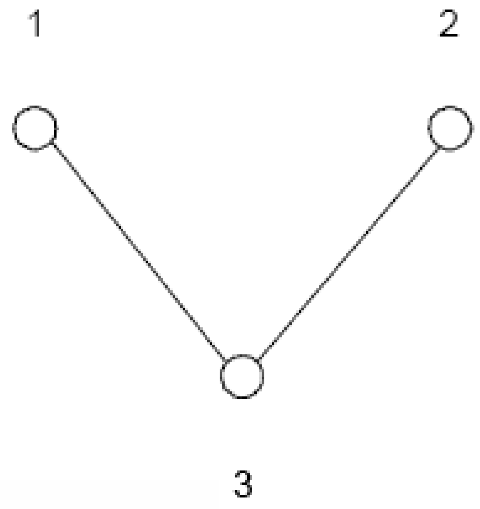
1 What is a social network?

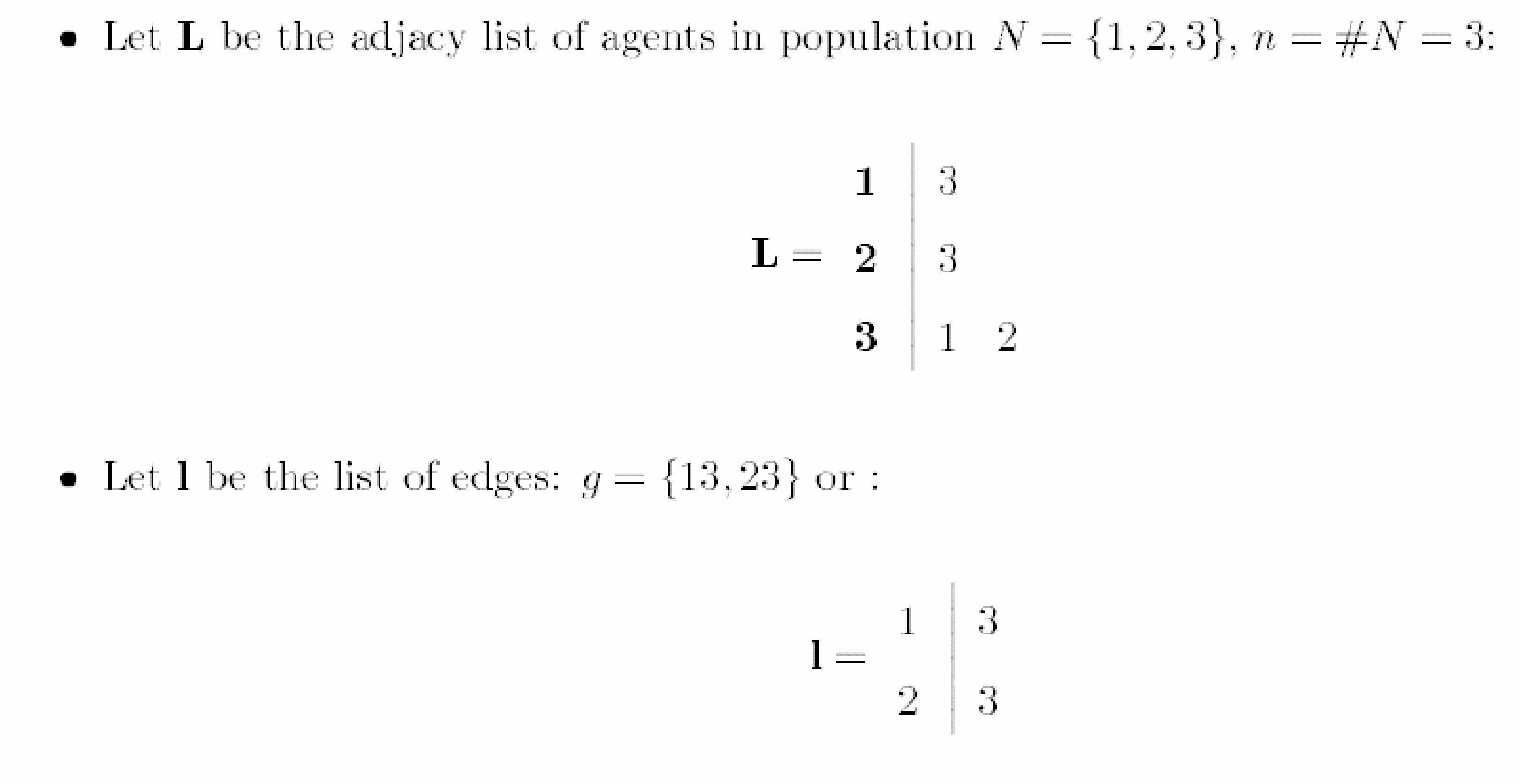
- A collection of agents
- A set of bilateral relations
- Some context of application

Many applications

- Buyer sellers networks
- R&D collaboration networks
- Collusion networks
- Public good contribution networks
- Crime networks
- Job market networks
- Opinion networks
- Company boards networks
- Stock market networks
- Marriage networks
- College dating network
- Movie actors networks
- Technology adoption networks
- …

2 Network data





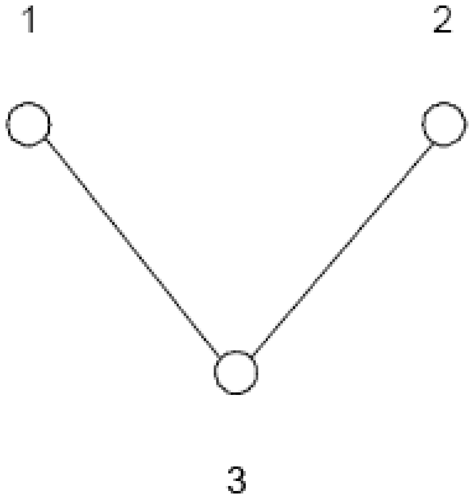
- Let $\mathbf{L}$ be the adjacy list of agents in population $N = \{1, 2, 3\}$, $n = \#N = 3$:

$$\mathbf{L} = \begin{array}{c|cc} \mathbf{1} & 3 & \\ \mathbf{2} & 3 & \\ \mathbf{3} & 1 & 2 \end{array}$$

- Let $\mathbf{l}$ be the list of edges: $g = \{13, 23\}$ or :

$$\mathbf{l} = \begin{array}{c|c} 1 & 3 \\ 2 & 3 \end{array}$$

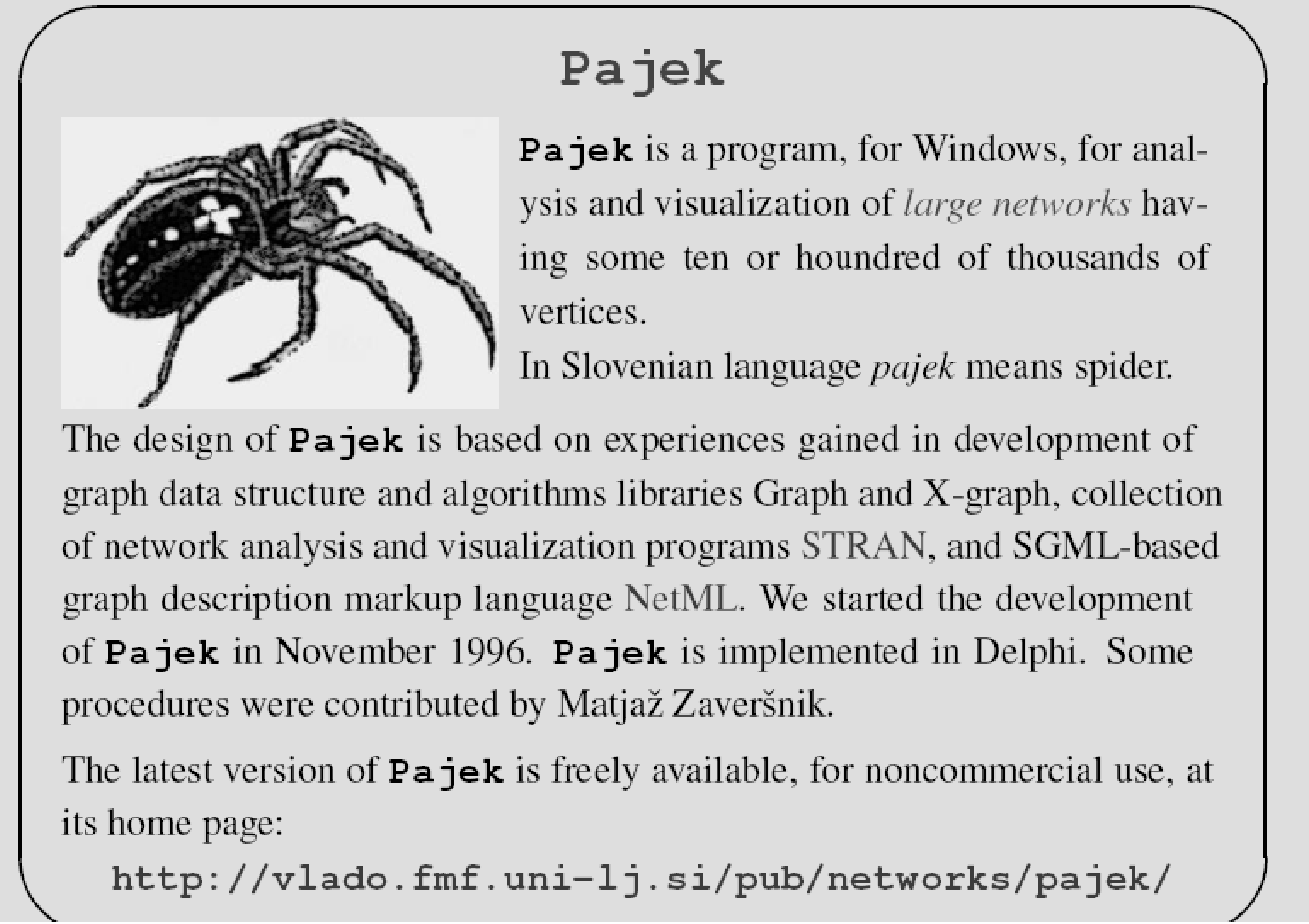
2 Network data



- Let matrix $\mathbf{G} = (g_{ij})_{n,n}$ be the $n \times n$ adjacency matrix of the network

$$\mathbf{G} = \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{array}\right] \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \end{array}$$

How to draw and perform computations on (large) networks ?



Pajek

Pajek is a program, for Windows, for analysis and visualization of *large networks* having some ten or houndred of thousands of vertices.

In Slovenian language *pajek* means spider.

The design of **Pajek** is based on experiences gained in development of graph data structure and algorithms libraries Graph and X-graph, collection of network analysis and visualization programs STRAN, and SGML-based graph description markup language NetML. We started the development of **Pajek** in November 1996. **Pajek** is implemented in Delphi. Some procedures were contributed by Matjaž Zaveršnik.

The latest version of **Pajek** is freely available, for noncommercial use, at its home page:

`http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

Draw your network with Pajek:

- Excel version of the edges list 1: network_trial.xls
- Use createpajek.exe -> network_trial.net
- Use pajek.exe
- Draw/layout/energy/Kamada-Kawai
- Other computations are available.
- You may also want to use some other softwares (e.g. ucinet)

Main questions raised in the literature

Empirical questions

- How can we measure networks ? Can we find some recurrent structural attributes ?

Theoretical questions

- How do social networks affect agents and social performance/welfare ?
- How do real social networks came to be formed ?
- Provided that agents know that they are affected by their position in networks, how can they improve their position in the networks and what are the resulting networks ?

Main questions raised **today**

Empirical questions

- **How can we measure networks ? Can we find some recurrent structural attributes ?**

Theoretical questions

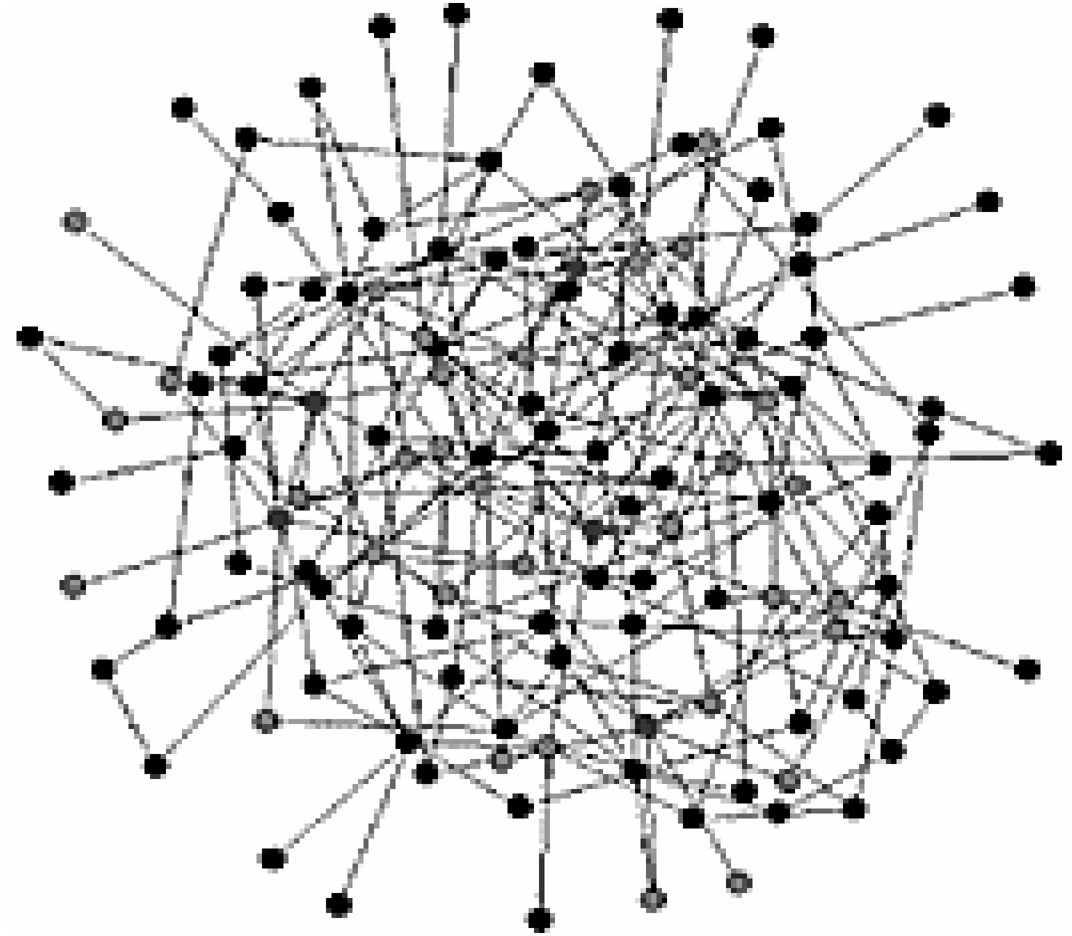
- How do social networks affect agents and social performance/welfare ?
- **How do real social networks came to be formed ?**
- **Provided that agents know that they are affected by their position in networks, how can they improve their position in the networks and what are the resulting networks ?**

3 The basic random graph model

- The measurements on real networks are usually compared against those on "random networks"

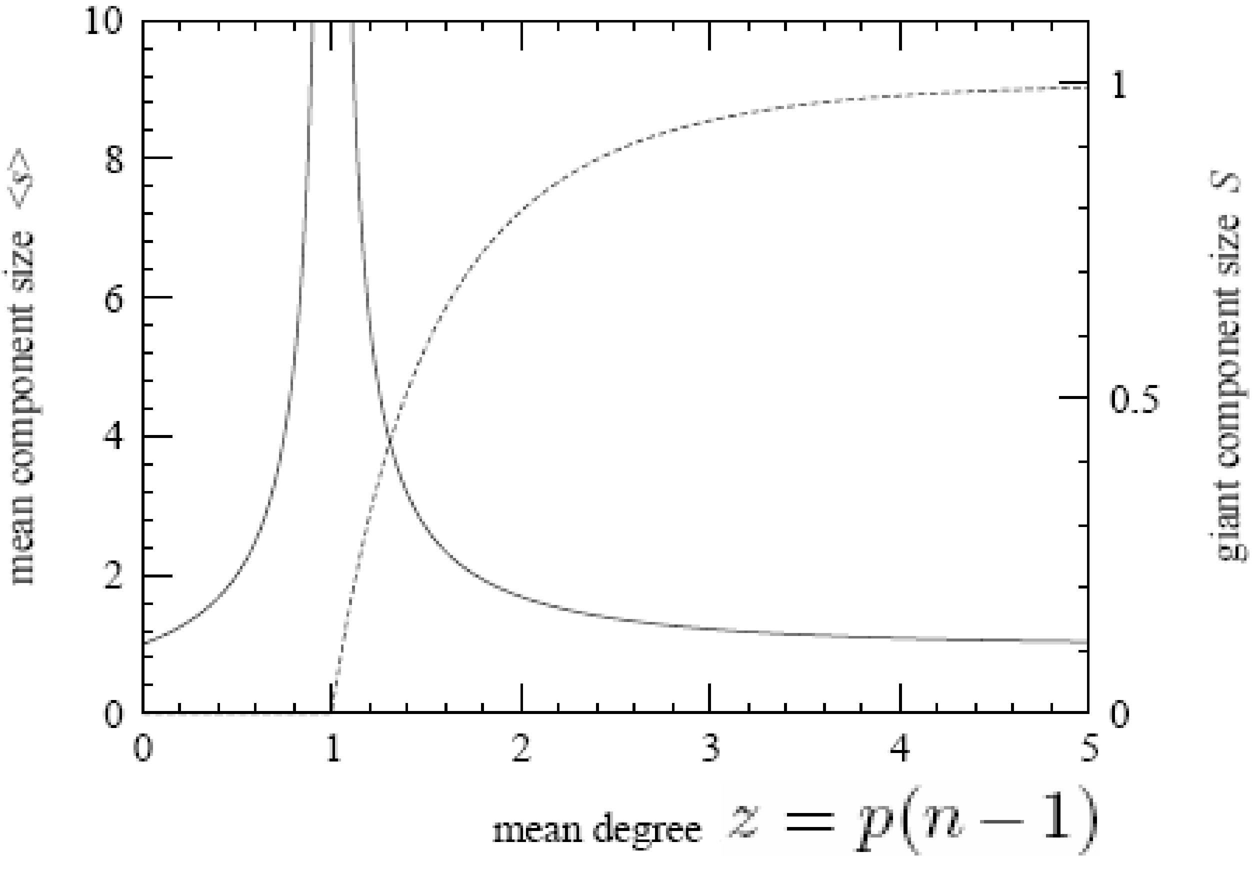
- The basic $G_{n,p}$ (Ǽrdös-Renyi) random graph model:
  - n : the number of vertices
  - $0 \leq p \leq 1$
  - for each pair of agents (i,j), generate the edge ij independently with probability p

Typical random network



The basic random graph model

- The main discovery of Erdös-Renyi, are that network properties emerge nonlinearly with p.
- Among thee properties is the size of the largest component:

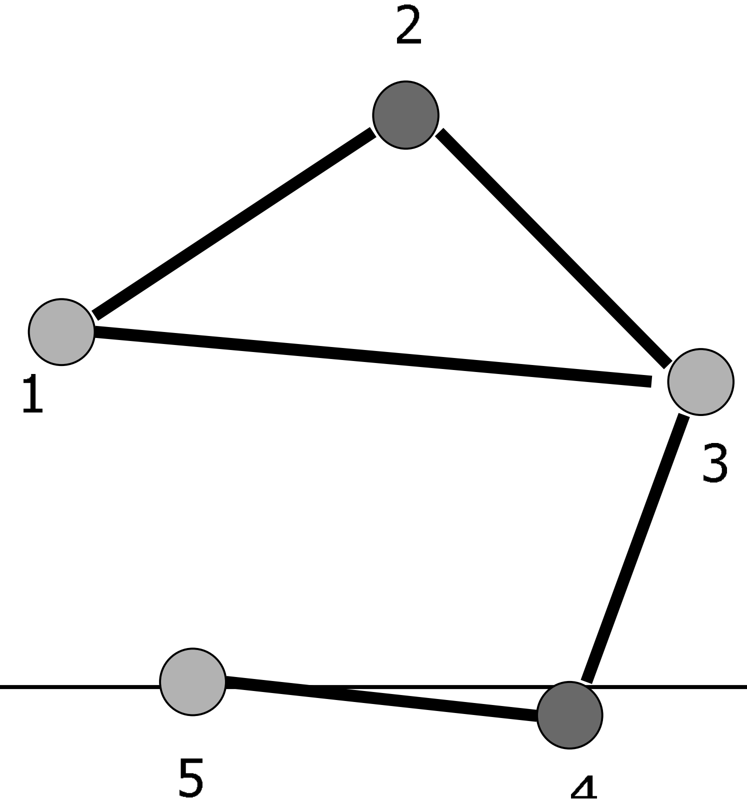


The small world phenomenon

- Milgram (69, 74) experiment :
  - Select a target in Sharon-Mass,
  - Select 296 persons (196 from Omaha-Nebraska and 100 from Boston-Mass),
  - Ask them to reach a the target, if they do not know him directly, send the letter to someone else they expect he may do, and send a report,
  - Repeat recursively.
- 64 initial reached the target – and it took in average 5.2 intermediate acquaintances to do so.
- The "six degree of separation" legend is born !
- Biased downwards but White's corrections indicate that it is probably not much more (between 6 and 8).

Definition of a path:

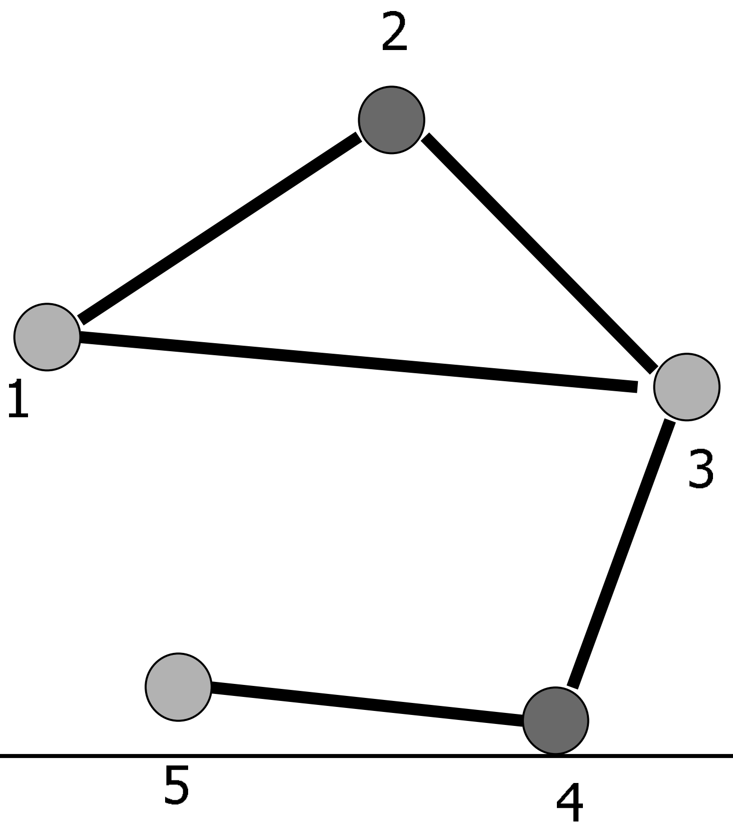
- Path from node i to node j is a sequence of edges that share common nodes from node i to node j.



• path length: number of edges on the path
• nodes i and j are connected
• cycle: a path that starts and ends at the same node

Shortest Paths

- Shortest Path from node 1 to node 4 ?



- Geodesic distance is the # of edges of the shortest path(s):
  d_{14}=2

The average path length

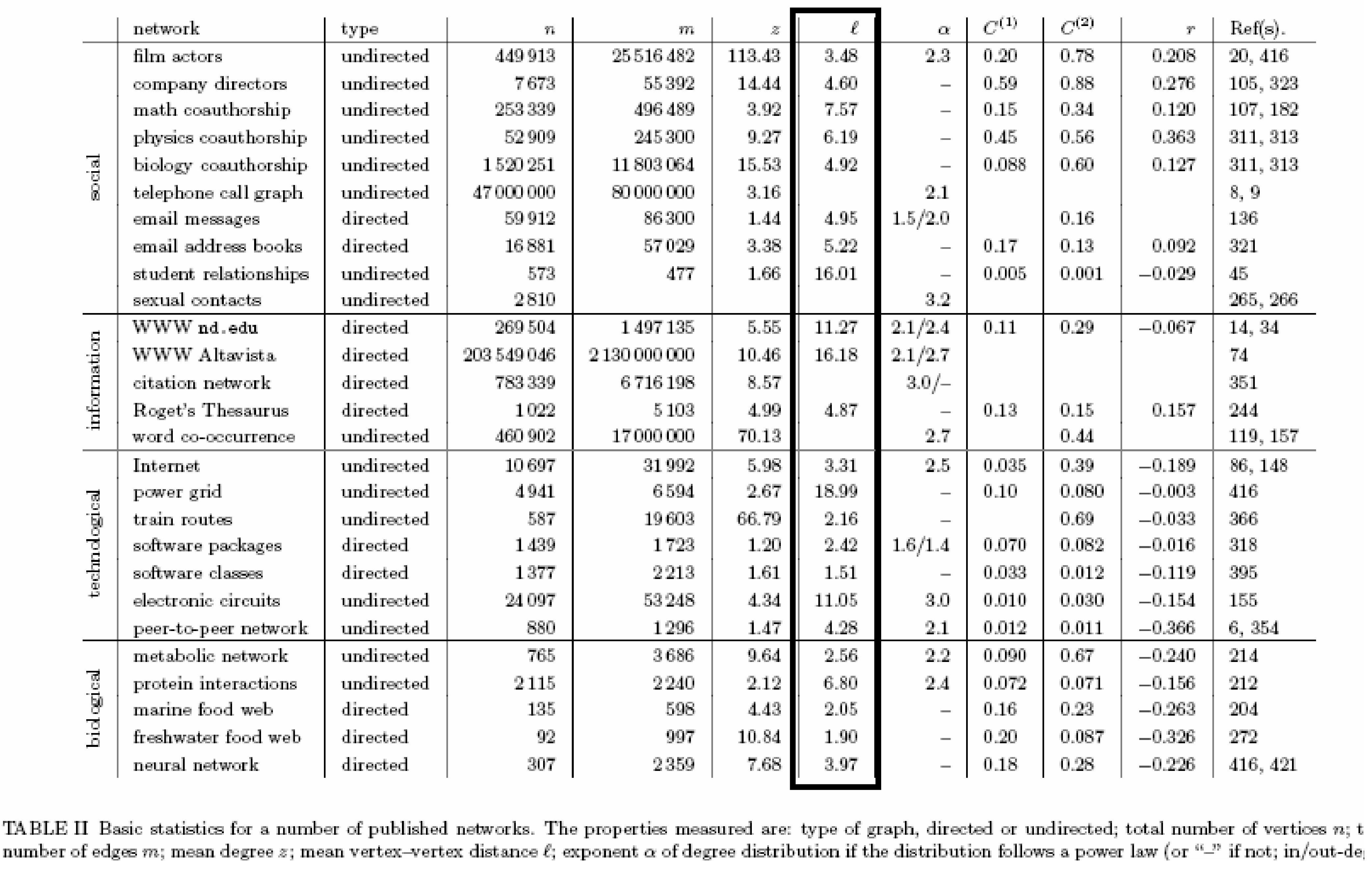
- d_{ij} = shortest path between i and j
- Characteristic average path length:

$$\ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- Harmonic mean

$$\ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d^{-1}_{ij}$$

Collective Statistics (M. Newman 2003)



	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; t
number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-de

The average path length of random networks is short !

- The average geodesic distance of a random graph (Ǽrdös-Renyi) is:

$$\ell = \log n / \log z \quad \text{with} \quad z = p(n-1)$$

which means that simple randomness is sufficient to allow (large) networks to be short.

Thus, is the random graph model a good predictor of real networks ?

NO & NO (at least)

- NO : degree distribution is incorrectly shaped

  -> leads to the "configuration model" of Molloy & Reed and the "scale free" network of Barabasi

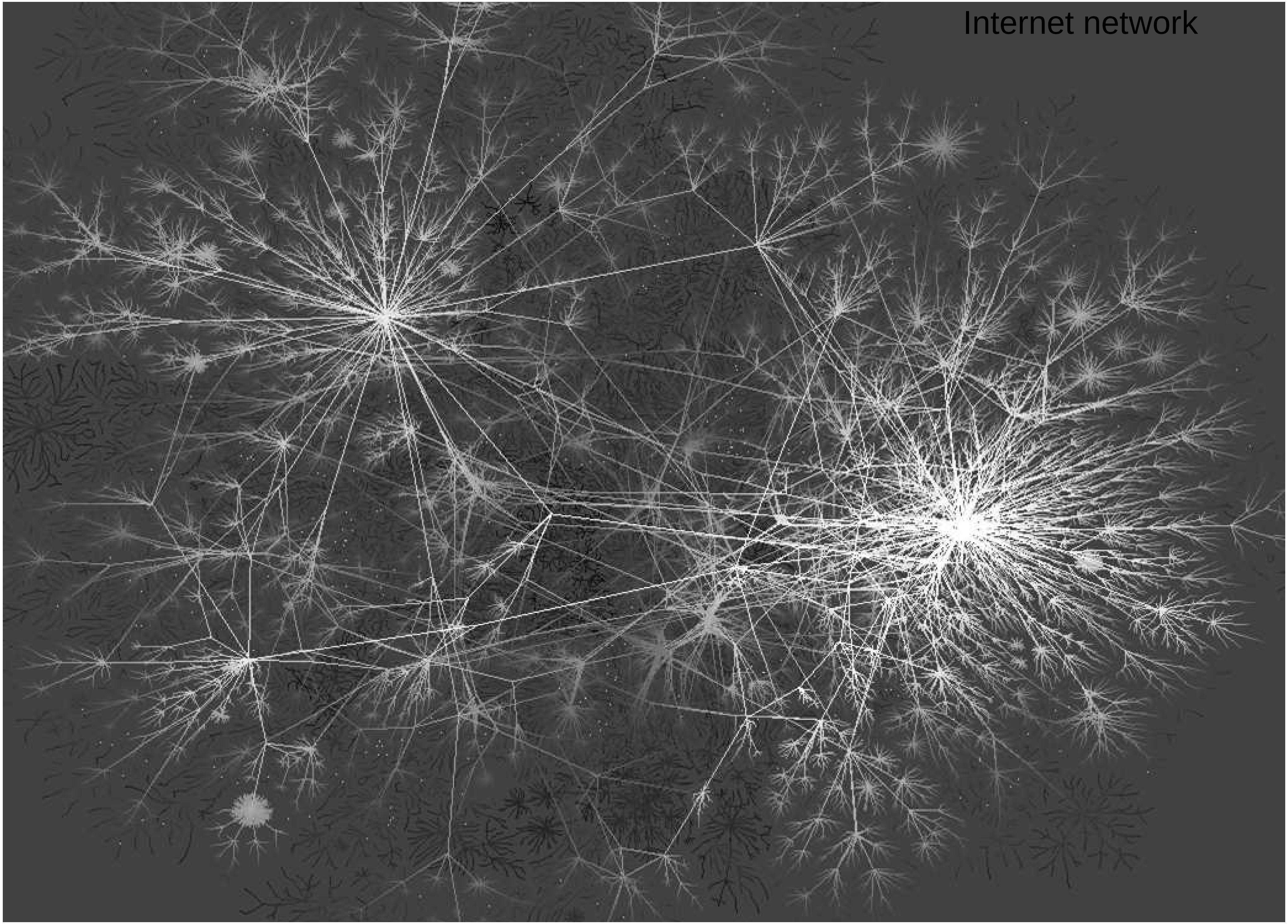
- NO : it does not generate communities as real networks do!

  -> leads to the "small world model" of Watts & Strogatz.

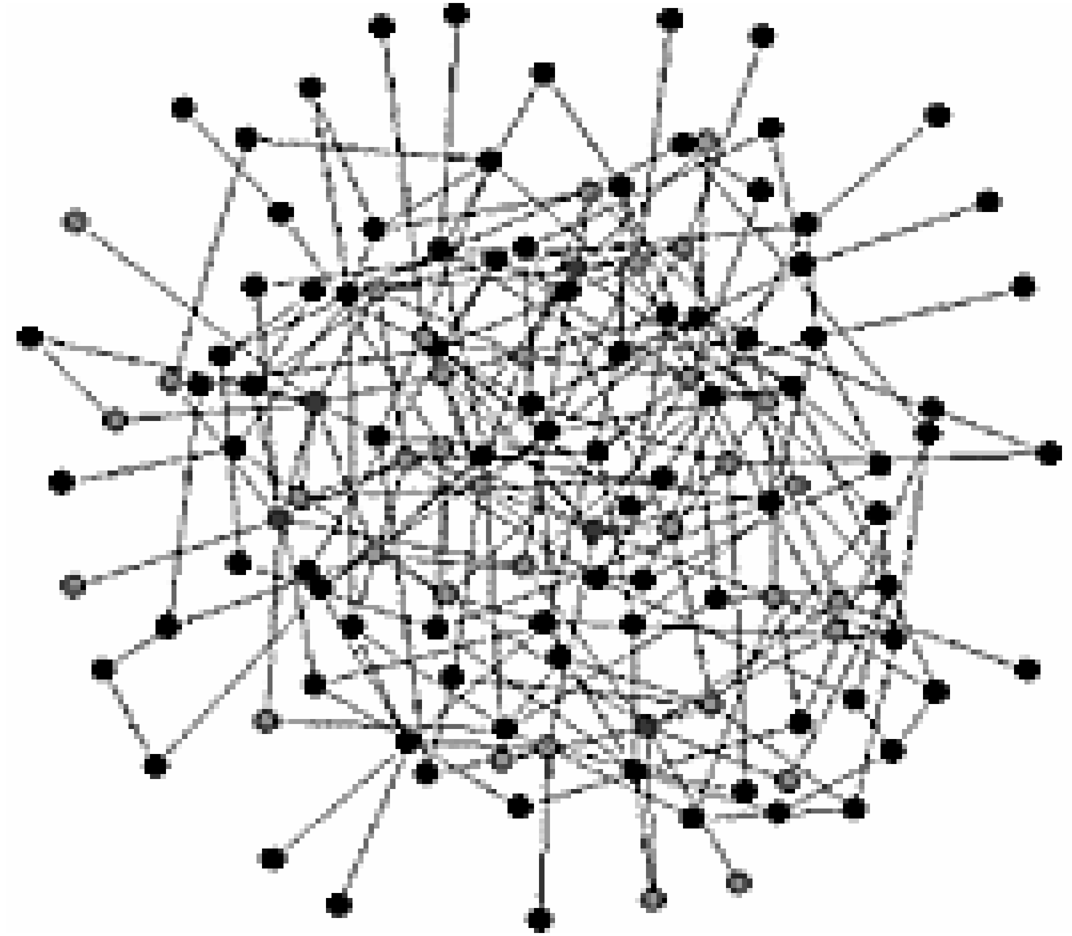
4 Scale free networks

- Let p_k denote the fraction of the agents who have exactly k neighbors, that is have degree k.

Internet network

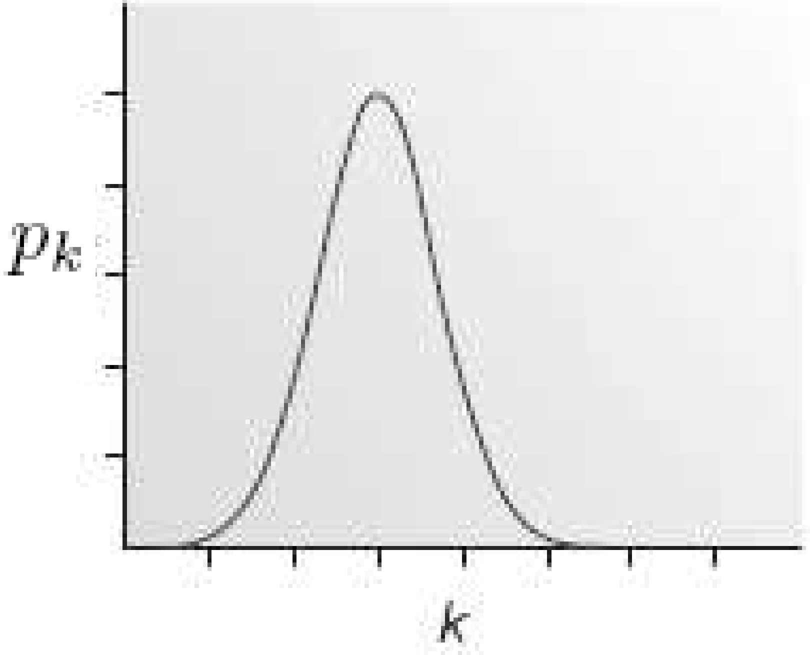


Typical random network



The basic random graph model

- The degree distribution in the random network model is Poisson.

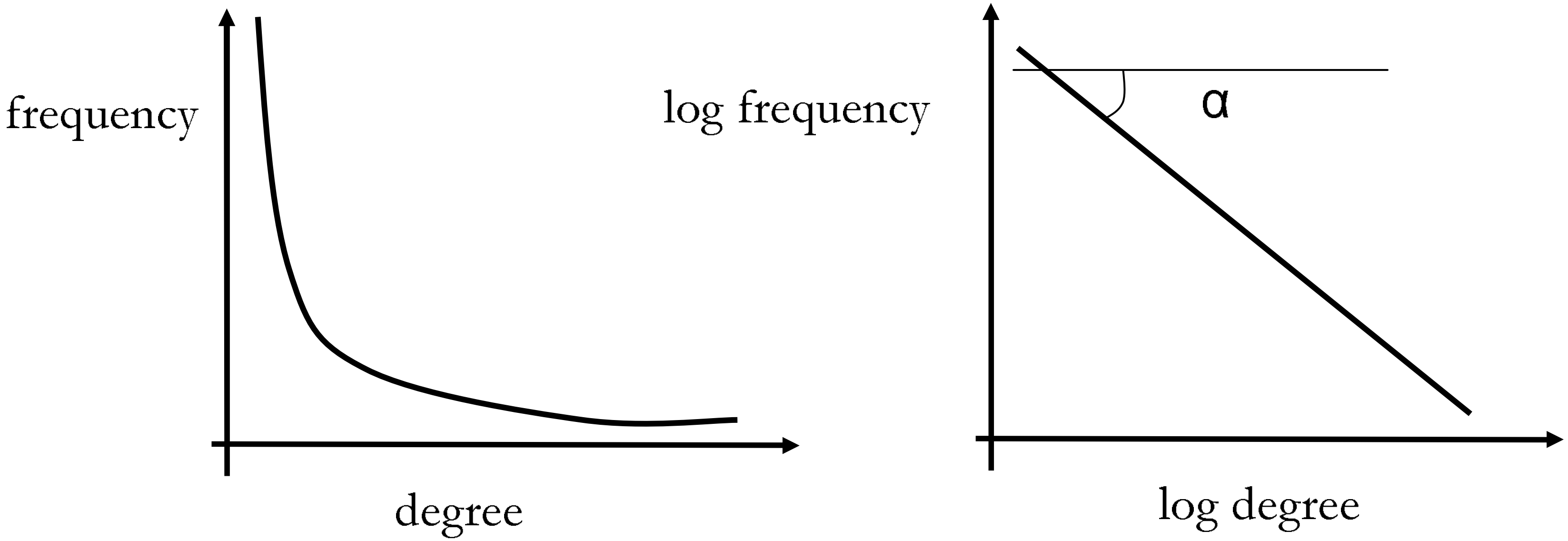


$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{z^k e^{-z}}{k!}$$

Real networks have power-law degree distribution

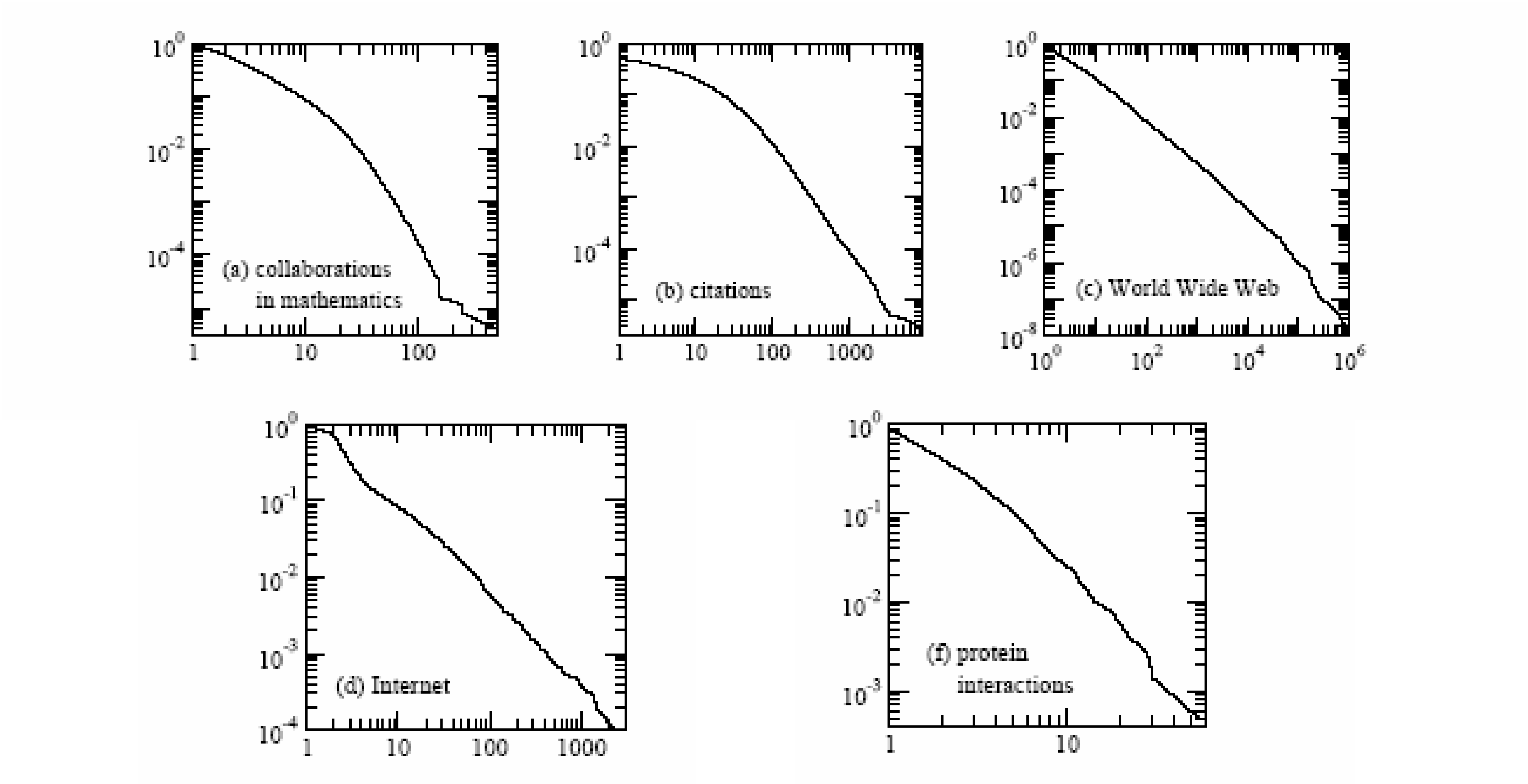
- Power-law distribution gives a line in the log-log plot

$p(k) = C k^{-\alpha}$ -> $\log p(k) = -\alpha \log k + \log C$



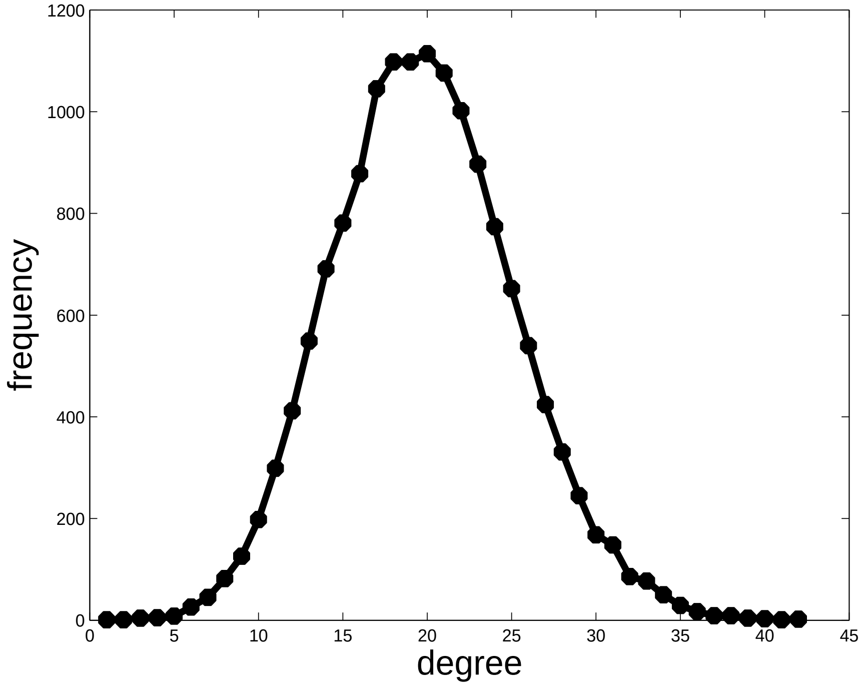
- α : power-law exponent (typically $2 \leq \alpha \leq 3$)

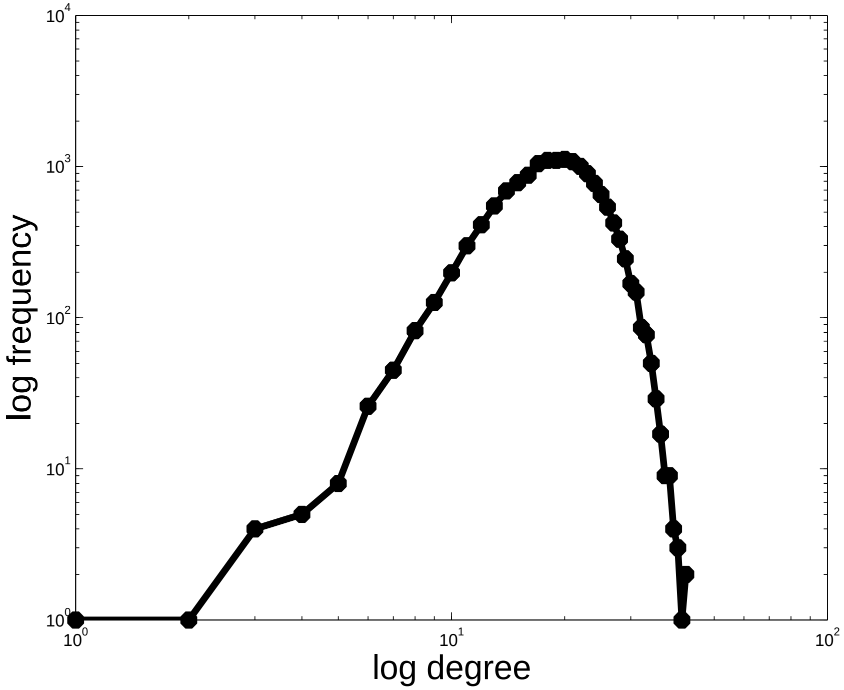
Examples



Taken from [Newman 2003]

A random graph example





The configuration model

- The configuration Model from Molloy and Reed (1988)
- A generalization of the poisson model, which allows for any ex ante specification of degree distribution.
- Let for instance:

$$p_k = \begin{cases} 0 & \text{for } k = 0 \\ k^{-\alpha}/\zeta(\alpha) & \text{for } k \geq 1 \end{cases}$$

- Results on non linear emergence of a giant component and low average distance are preserved

The Barabasi model for generating scale free networks

- Simon (1955), Price (1976), Barabasi & Albert (2001)
- Two main principles: network growth and preferential attachment.
- At each period, one node arrives.
- He connects randomly to *m* already existing nodes
- The probability it connects to a node of degree p_k is given by :

$$\frac{kp_k}{\sum_k kp_k} = \frac{kp_k}{2m}$$

- Thus at each period there are in average $m \times kp_k/2m = \frac{1}{2}kp_k$ nodes which change degree from *k* to *k+1*.

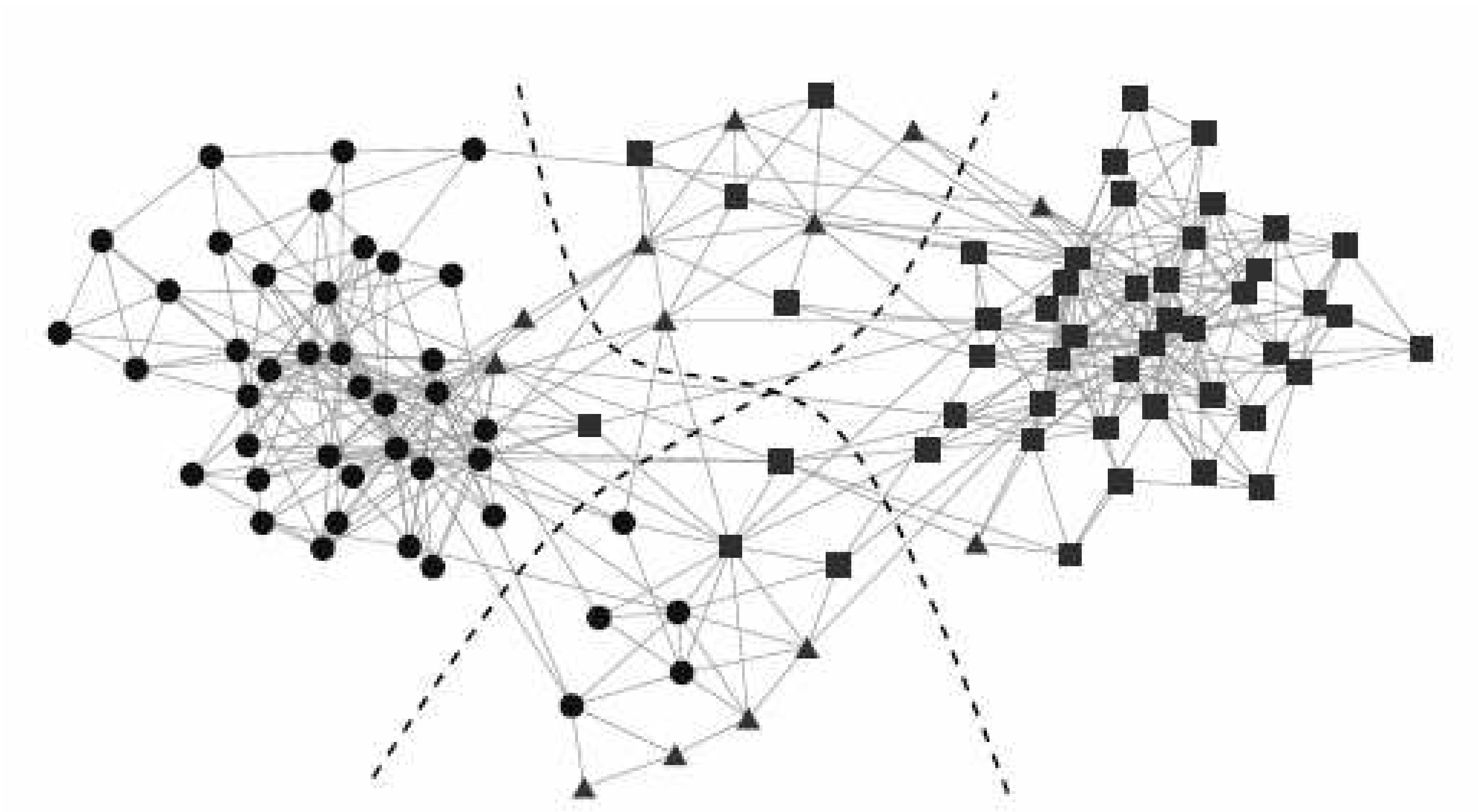
The Barabasi model for generating scale free networks ?

- Such a dynamical system leads to a network the degree distribution has been proved to be scale free, that is power distributed as follows:

$$p_k = 2m * k^{-3}$$

that's a power distribution indeed !

5 Community structures and the small world model



5 Community structures and the small world model

- In most social networks, neighborhoods tend to overlap.
- That translates in the network worlds into:

  "my neighbors have a high probability to be also neighbors together".
- In the network literature there is an index that captures this propensity: network clustering

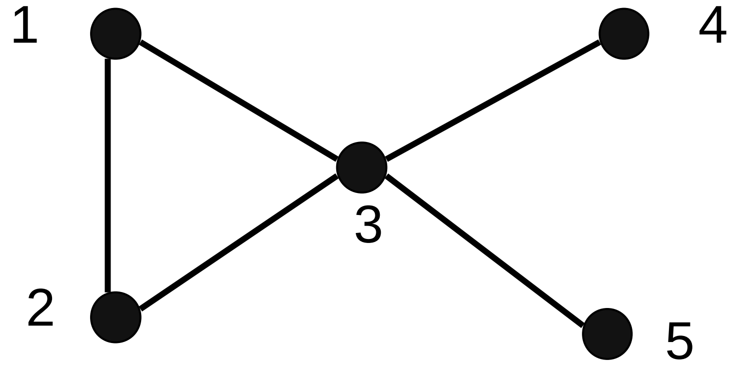
Clustering (Transitivity) coefficient

- Measures the density of triangles (local clusters) in the graph
- Two different ways to measure it:

$$C^{(1)} = \frac{\sum_{i} \text{triangles centered at node i}}{\sum_{i} \text{triples centered at node i}}$$

- The ratio of the means

Example



$$C^{(1)} = \frac{3}{1+1+6} = \frac{3}{8}$$

Clustering (Transitivity) coefficient

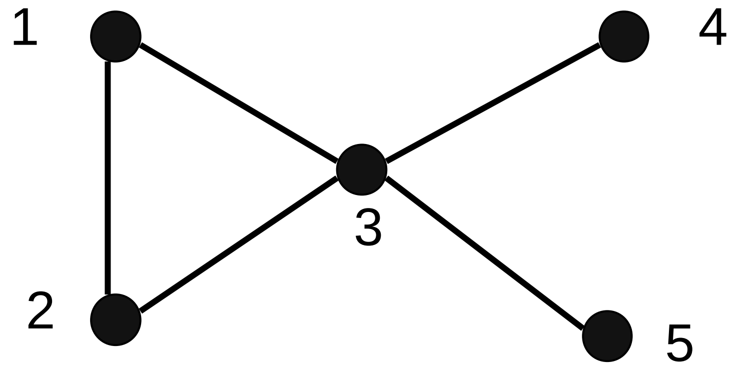
- Clustering coefficient for node i

$$C_i = \frac{\text{triangles centered at node i}}{\text{triples centered at node i}}$$

$$C^{(2)} = \frac{1}{n} C_i$$

- The mean of the ratios

Example



$$C^{(2)} = \frac{1}{5}\left(1+1+1/6\right) = \frac{13}{30}$$

$$C^{(1)} = \frac{3}{8}$$

- The two clustering coefficients give different measures
- $C^{(2)}$ increases with nodes with low degree

Clustering coefficients

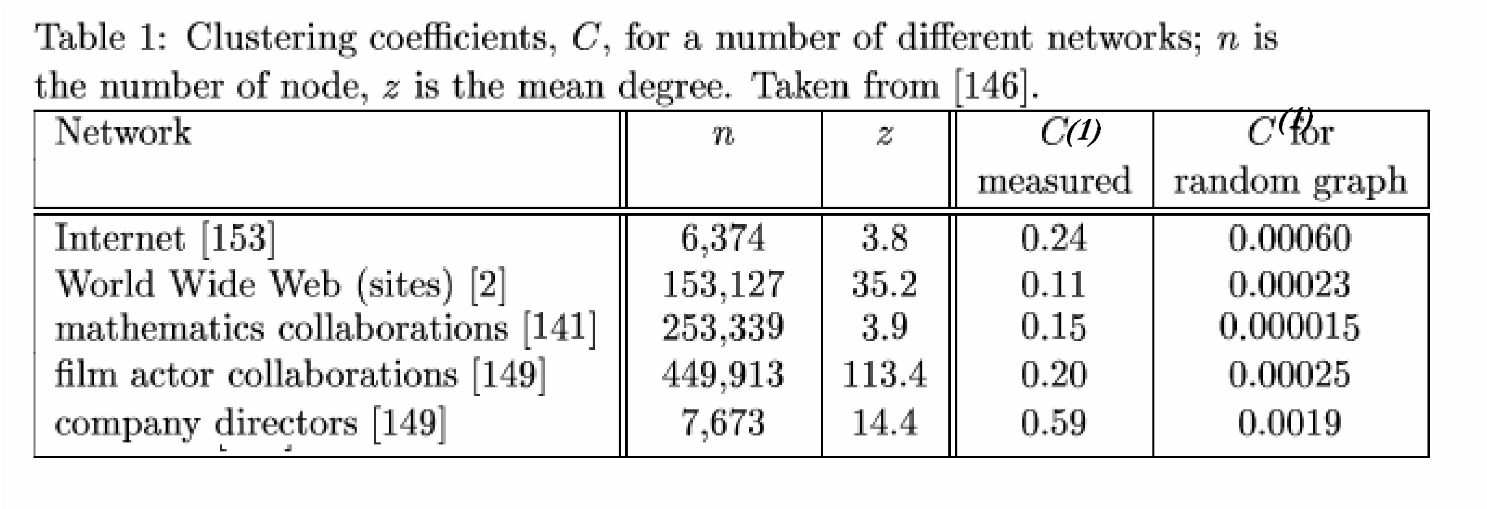


Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

Network	n	z	$C(1)$ measured	$C^{(1)}$ for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019

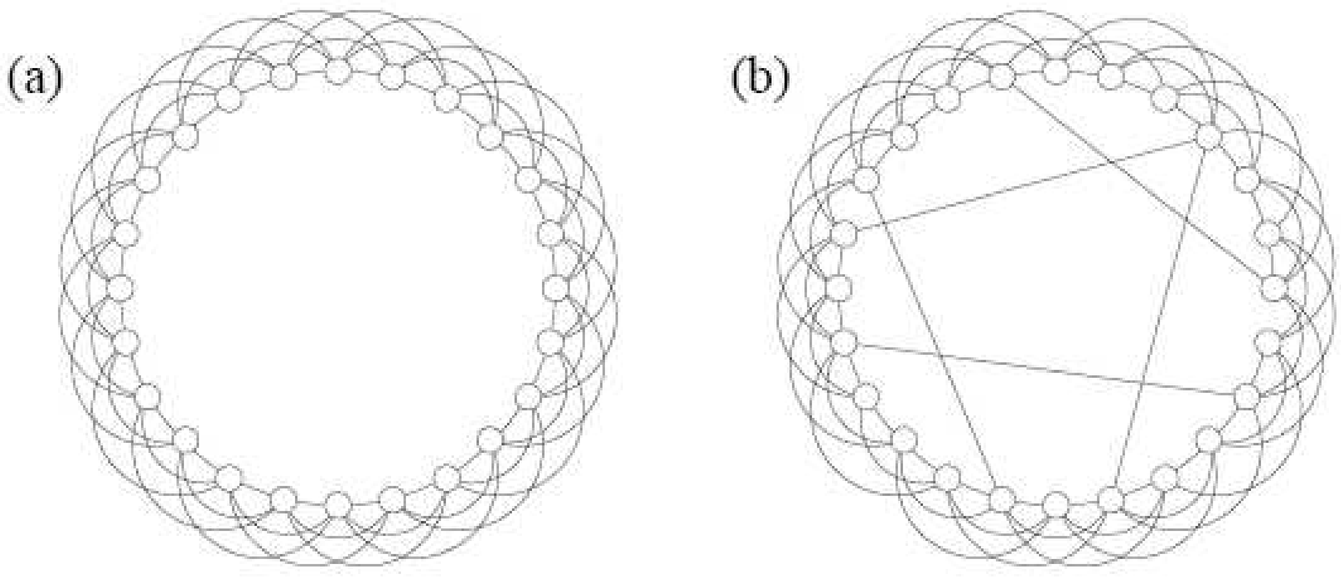
- In the standard random graphs, the probability that two of your neighbors also being neighbors is p, independently of local structure. Thus:
  - clustering coefficient C = p
  - when z is fixed C = z/n =O(1/n)y

- For instance in the configuration models, clustering is:

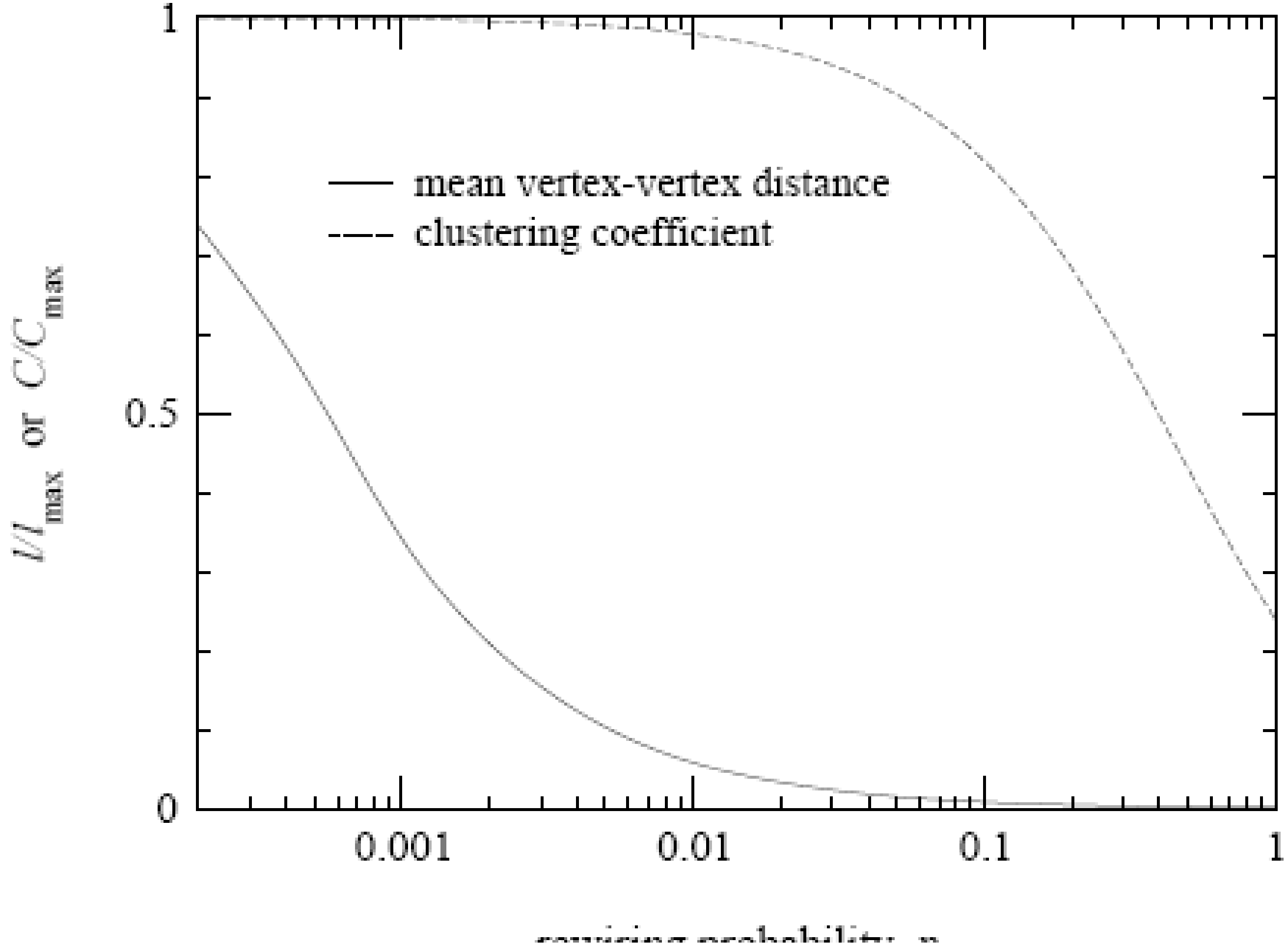
$$C^{(1)} \sim n^{-\beta}$$

The Watts & Strogatz model for generating clustered & short networks ?

- Take a 1d-lattice (a) and rewire each edge with a small proba *p*, and then reallocate one of the ends of the edge to a randomly selected node.



The Watts & Strogatz model for generating clustered & short networks ?



6 Strategic network formation games

- Networks are perceived as the equilibrium outcome of decentralized agents behaviors
- Different equilibrium notions/conceptions of network formation
  - ❑ Fully non cooperative approach - Nash networks
  - ❑ Fully Cooperative approach
  - ❑ Mixed approach - Pairwise stable networks !
- Static vs. dynamic settings
- Myopic and farsighted approaches
- Are agents allowed to form connections multilaterally on only bilaterally ?

Payoffs and efficiency

- The payoffs that i obtains from her position in the network is given by $\pi_i : \left\{g \mid g \subseteq g^N\right\} \to \Re$.

- The total value of a graph g is $\pi(g) = \sum_{i \in N} \pi_i\,(g)$

- A network $g \subseteq g^N$ is said to be efficient if $\pi(g) \geq \pi(g')$ for all $g' \subseteq g^N$.

Network formation principles and static equilibria in the mixed (coop. non coop.) approach

- Links need the consent of the two to be formed but can be deleted only if one of the two intends to.

- Pairwise stability:

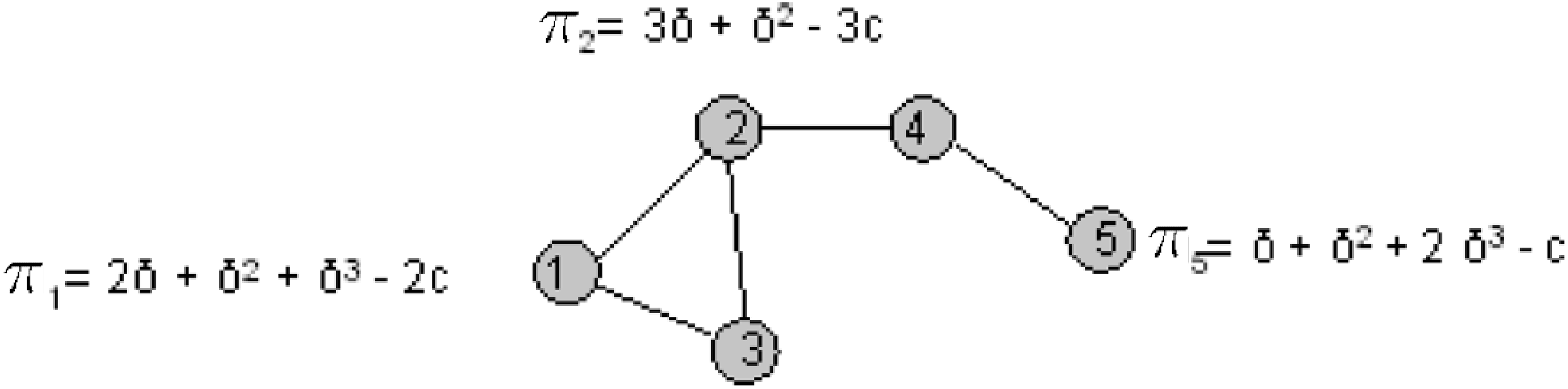
A network $g \subseteq g^N$ is pairwise stable if: i) for all $ij \in g$, $\pi_i(g) \geq \pi_i(g - ij)$ and $\pi_j(g) \geq \pi_j(g - ij)$, and ii) for all $ij \notin g$, if $\pi_i(g + ij) > \pi_i(g)$ then $\pi_j(g + ij) < \pi_j(g)$

- The payoff function of the connection model Jackson and Wolinski (1996)

$$\pi_i\left(g\right) = \sum\nolimits_j \delta^{d(i,j)} - \sum\nolimits_{j \in N_i(g)} c_{ij}$$

$$c_{ij} \equiv c$$

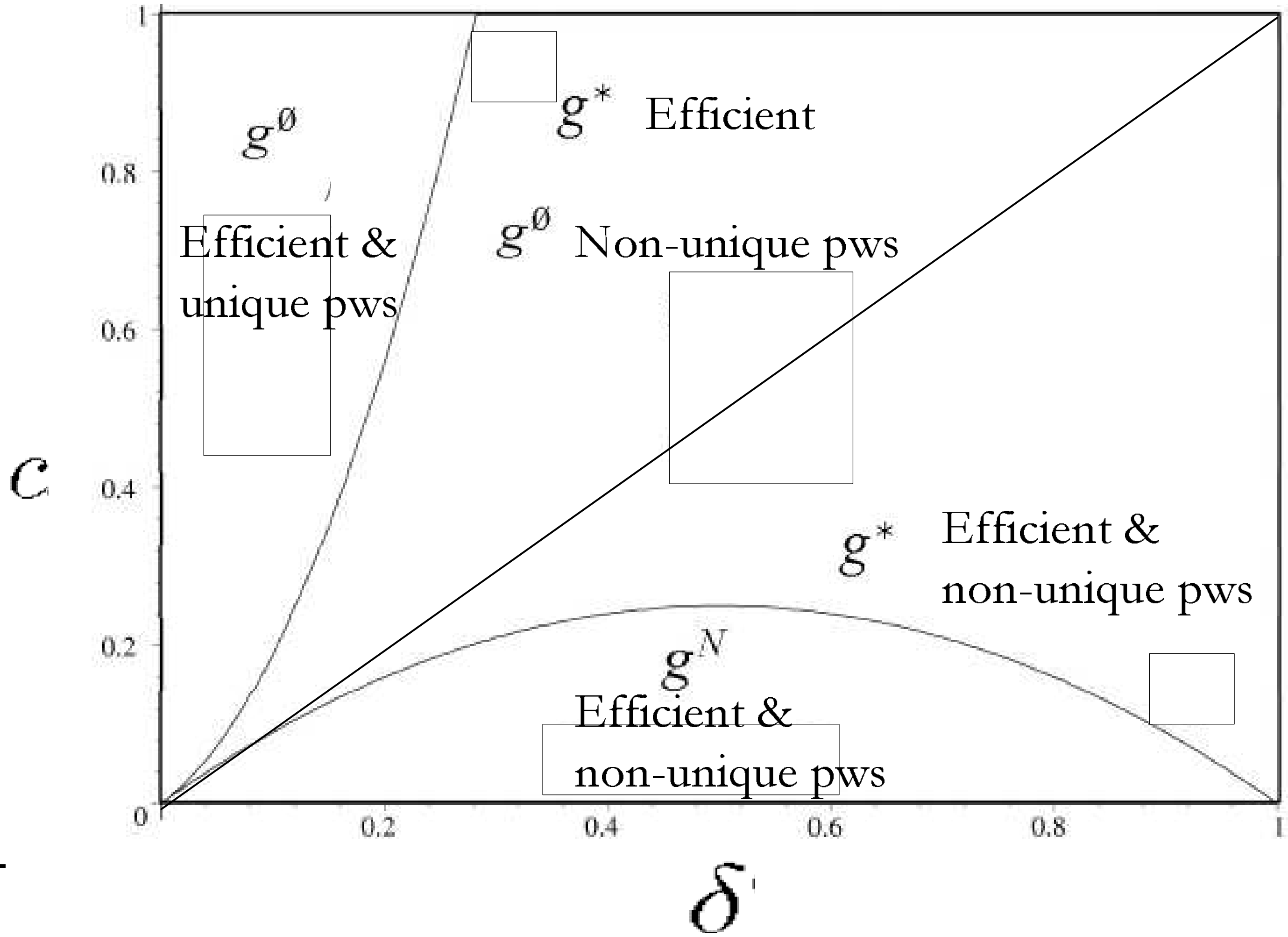
- The payoff function of the connection model Jackson and Wolinski (1996)


$\pi_2 = 3\delta + \delta^2 - 3c$
2
4
5
$\pi_5 = \delta + \delta^2 + 2\,\delta^3 - c$
$\pi_1 = 2\delta + \delta^2 + \delta^3 - 2c$
1
3

Results :

Efficiency vs. stability and transfers among players: Illustration with the connections model

Efficient and pws networks



Some remarks:

- Efficiency is fully characterized in this model (but not in all models)
- Contradictions appear between stability and efficiency → why do not allowing for transfers among players → who want to be the star ?
- Network pairwise stability is only partially characterized
- The only networks that are studied are sharp and simple ones as compared to real networks.

Other models: job contact networks

- Agents are either employed or unemployed at some moment in time
- Employees loose their jobs at some exogenous rate and information on some available positions arrive randomly
- An employed agent passes the information on jobs to his/her unemployed neighbors.
- Easy to show that agent *i* position in the network affects his welfare.

Other models: networks of firms

- Networks of cost reducing alliances (R&D)

$$\pi_i(g) = p\ q_i - q_i c_i(g)$$

$$c_i(g) = a - bn_i(g) \qquad n_i(g) = |N_i(g)|$$

$$p = \alpha - \sum_i q_i$$

- Cournot equilibrium for any given network:

$$q_i(g) = \frac{\alpha - a + nbn_i(g) - b\sum_{j\neq i} n_j(g)}{n+1}. \quad \text{and} \qquad \pi_i(g) = (q_i(g))^2$$

Other models: networks of firms

- Network efficiency:
  - the complete network is the unique efficient network,
  - because both firms and consumers surplus are increasing in the number of links formed

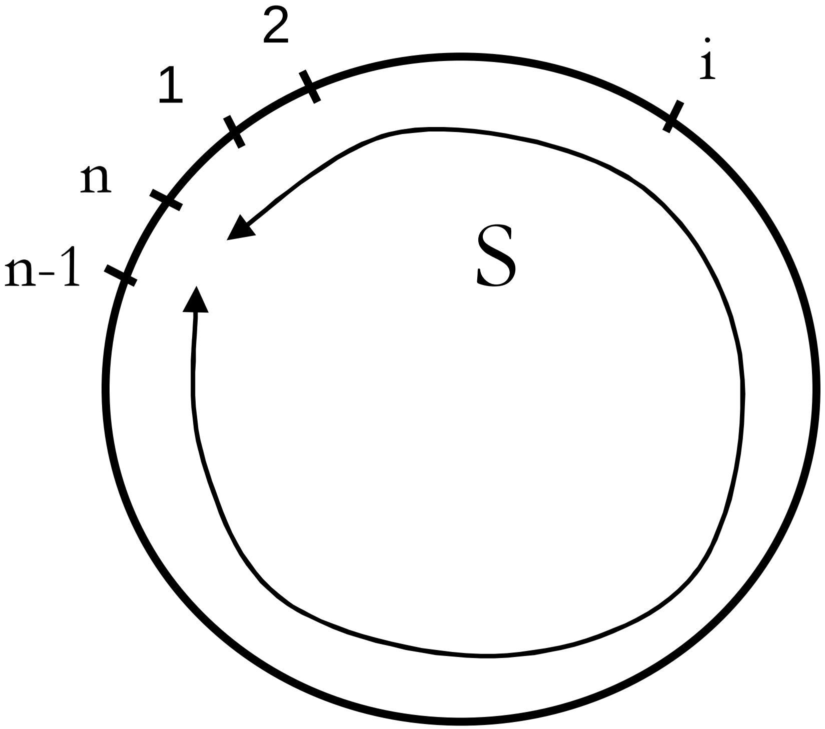
- Under Cournot competition:
  - the complete network is also the unique pairwise stable network when links formation costs are negligible
  - because profits are increasing in each companies number of connections

Network of collaboration among inventors

- Provide a slightly modified version of the connection model that nicely mimics inter-individual knowledge diffusion.
- Introduce geography (agents are localized in a ring-space).
- Study the formation of networks in a dynamic setting
- Analyze the structure of networks that emerge in this process
- Compare them with the structure of co-invention networks!
- Demonstrate that the strategic approach can explain the formation of complex real networks !

Introducing a spatial structure

$$l(i,j) = \min\{|i-j|\,; n-|i-j|\}$$



$$s_{ij} = \frac{l(i,j)}{\lceil n/2 \rceil} S$$

- Geographic distance : a ring city of dimension S

- The payoff function

$$\pi_i\left(g\right) = \sum\nolimits_{j \in N_i^2(g)} \delta^{d(i,j)} - \sum\nolimits_{j \in N_i(g)} c_{ij}$$

$$c_{ij} \equiv a_i s_{ij} \qquad\qquad a_i \sim U\left[\underline{a}, \overline{a}\right]$$

The dynamic approach

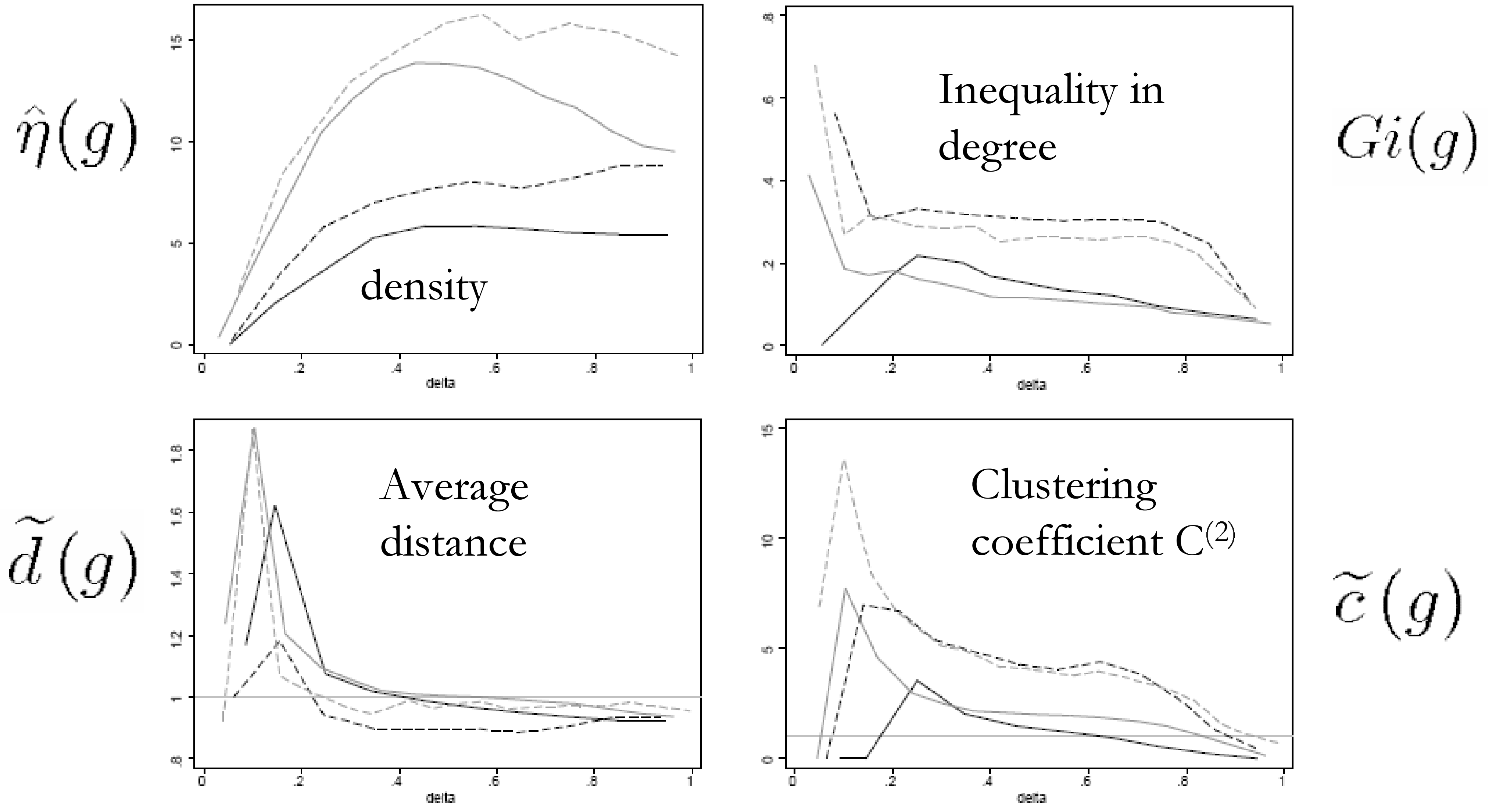
- Stochastic process:
  - Random uniform matching (Jackson and Watts, 2002) and implementation rule consistent with pairwise stability concept.
  - Agents can always make errors with a small probability but they learn with time -> a time-dependent markov chain
  - The networks that are on the support of the unique limiting distribution are said to be *emergent networks* -> Ergodicity is preserved

- Monte Carlo numerical experiments
  - Different values for n, S and σ.
  - δ varies over]0,1[

Emerging networks

(n^1, S^1, σ^1) (n^2, S^2, σ^2)

(n^2, S^1, σ^1) (n^3, S^2, σ^2)



Pictures of strategically generated complex small worlds

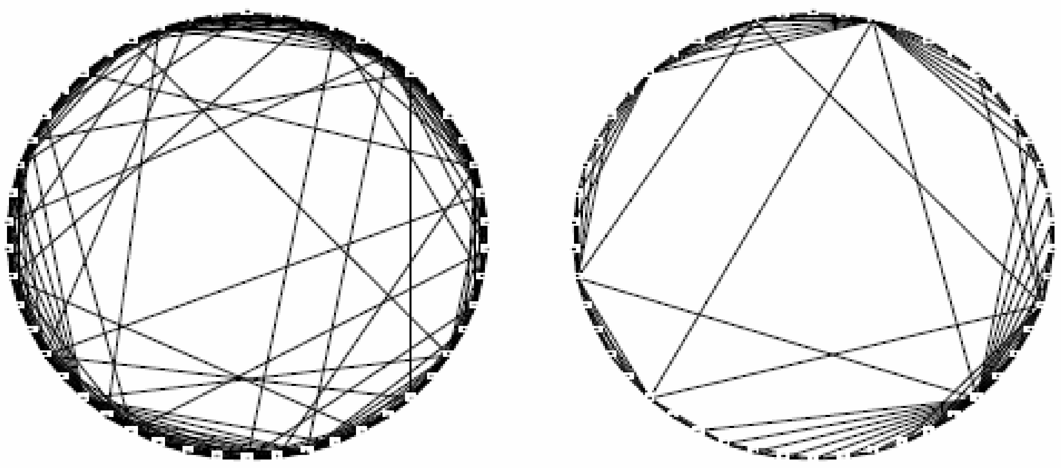
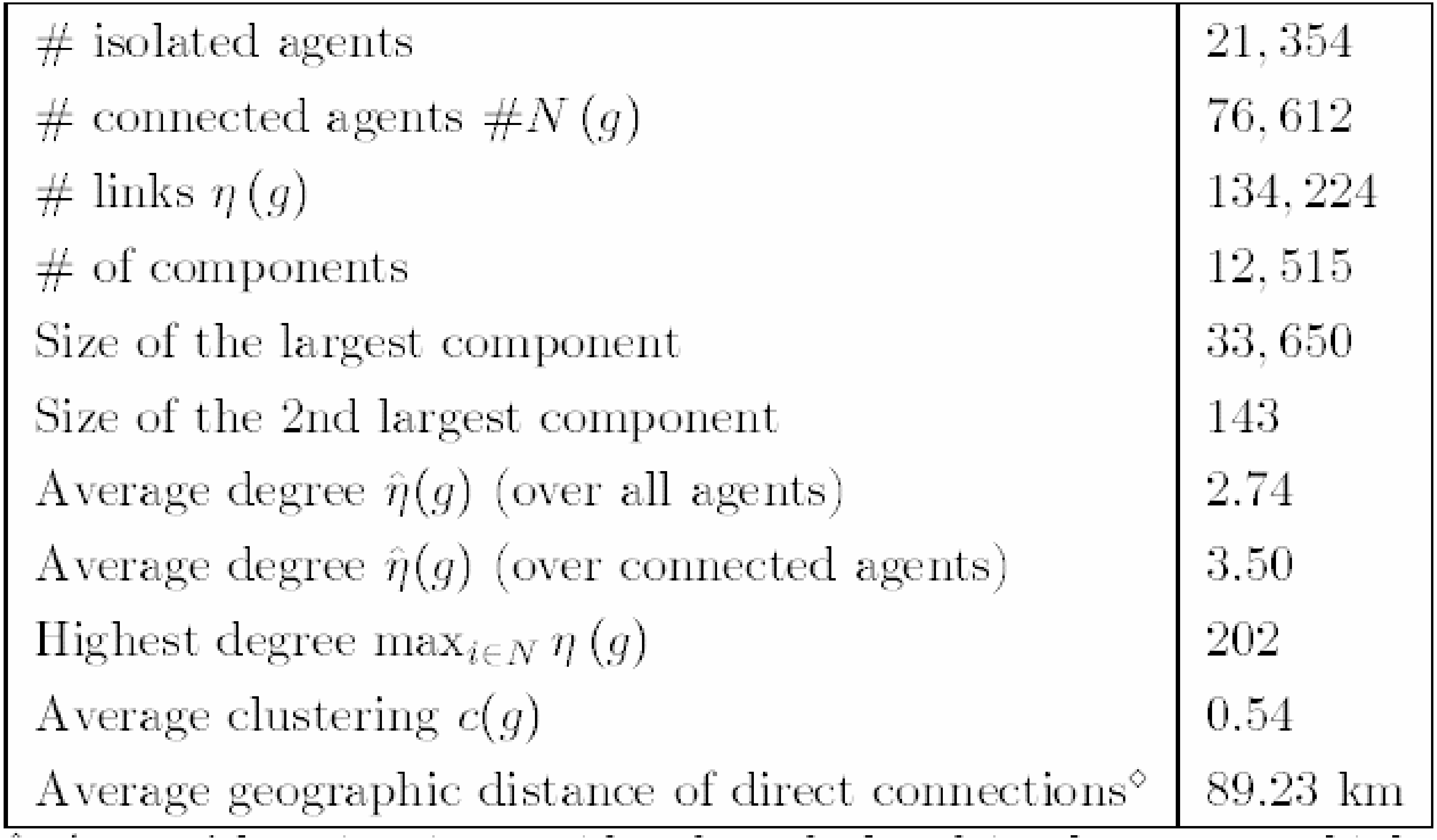


Figure 2. Typical emergent networks generated with $\delta = 0.25$, $n = n^2 = 50$ agents and two configurations of the other parameters : S^1, σ^1 (left network) and S^2, σ^2 (right network).

Co-invention network



# isolated agents	$21,354$
# connected agents $\#N(g)$	$76,612$
# links $\eta(g)$	$134,224$
# of components	$12,515$
Size of the largest component	$33,650$
Size of the 2nd largest component	143
Average degree $\hat{\eta}(g)$ (over all agents)	2.74
Average degree $\hat{\eta}(g)$ (over connected agents)	3.50
Highest degree $\max_{i \in N} \eta(g)$	202
Average clustering $c(g)$	0.54
Average geographic distance of direct connections$^{\diamond}$	89.23 km

Comparability problems :

- In the theory: a given set of agents who can form links with each others & long run equilibrium
- In the empirics: not every agent can connect to all others: unobserved (cognitive or institutional) boundaries

A component-based methodology:

- Rely on components as populations so as to approximate isolated population
- Consider only components which exhibit no evolution in the recent past ("dead networks").

	Empirics			**Theory**		
	No recent link formation			$\delta = 0.25$		
	$15 \leq n \leq 25$	$30 \leq n \leq 70$		$n^1 = 20$	$n^2 = 50$	
	all S	$S < \overline{S}$	$S \geq \overline{S}$	σ^1, S^1	σ^1, S^1	σ^2, S^2
Average degree $\hat{\eta}(g)$	3.74	8.11	4.18	3.78	10.71	5.67
Average distance $d(g)$	2.51	3.22	3.14	2.43	2.02	2.25
Average clustering $c(g)$	0.73	0.71	0.75	0.63	0.63	0.70
Gini coefficient $Gi(g)$	0.31	0.33	0.37	0.22	0.16	0.33
Total # of agents	$(1,671)$	(504)	(496)	$(10,000)$	$(10,000)$	$(10,000)$

Table 3. Various structural measures computed on empirical and theoretical emergent networks. $\overline{S}$ is the median geographical size of components of population $35 \leq n \leq 65$.

Degree distribution

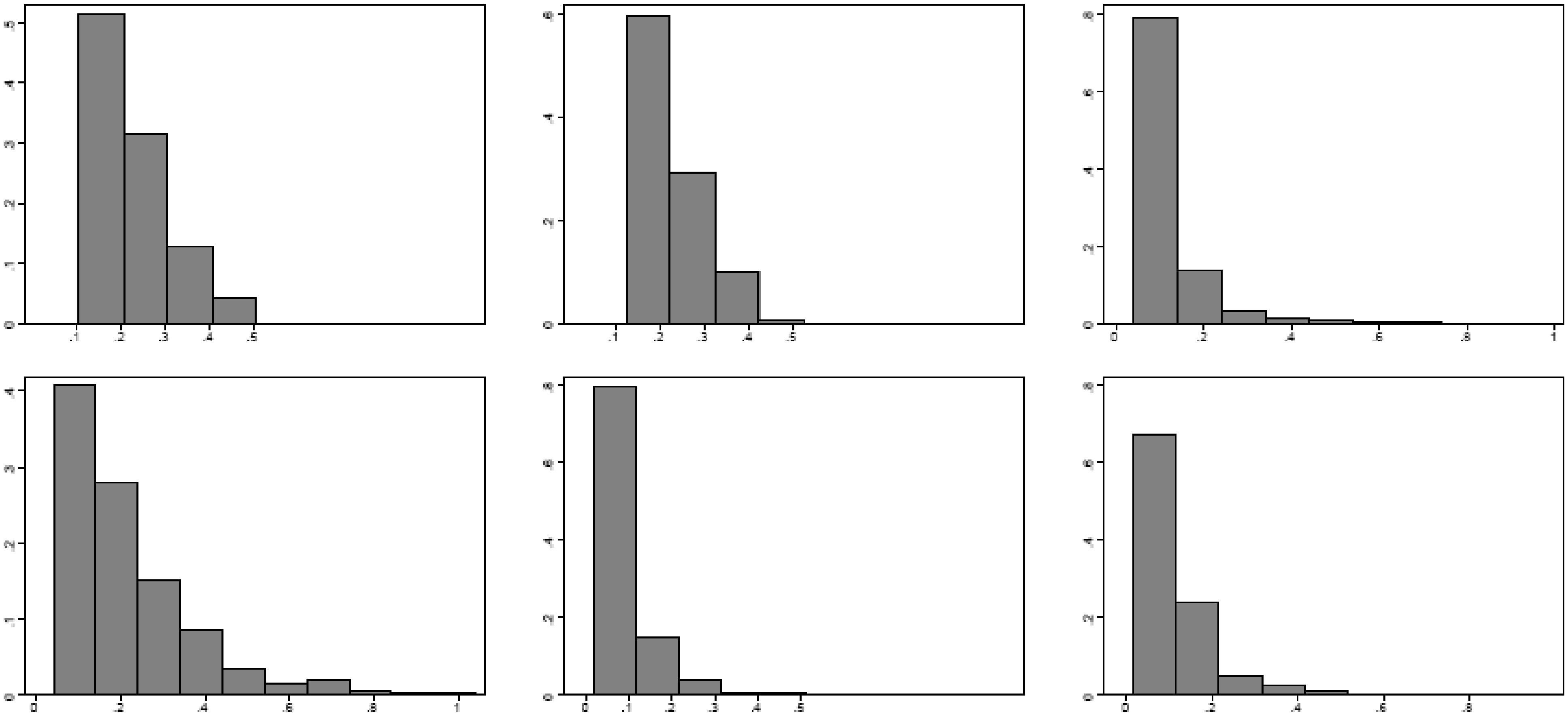
$$\rho_C(k/(\#C-1)) = \frac{1}{\#C}\sum_{i\in C} 1\{\eta_i(g)=k\}$$

$n^1 = 20$

σ^1, S^1

$n^2 = 50$

σ^1, S^1 | σ^2, S^2



$15 \leq n \leq 25$

all S

$35 \leq n \leq 65$

$S < \overline{S}$ | $S \geq \overline{S}$

Distribution of links according to geographical distance

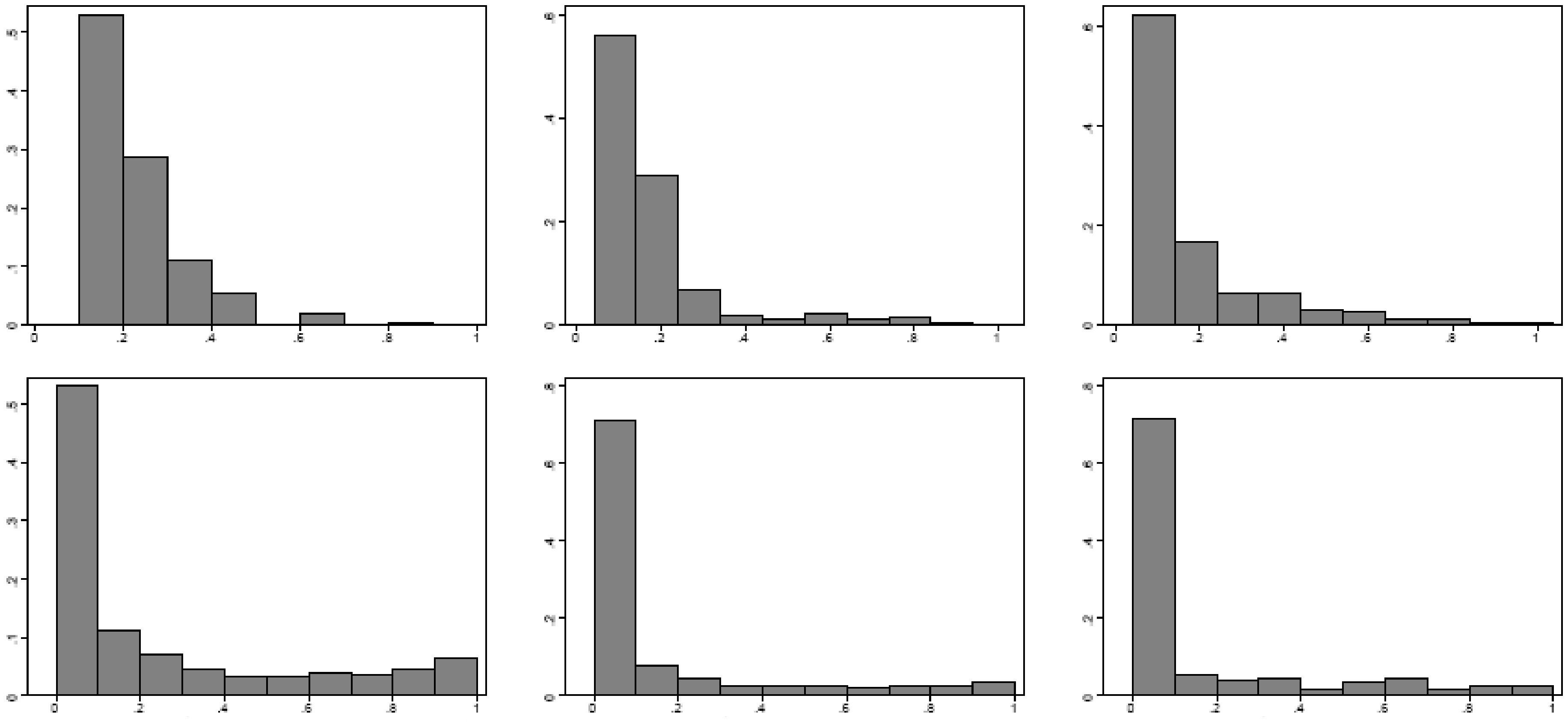
$$\phi_C(h/S) = \frac{1}{\#\{ij \,|ij \in g \text{ st } i,j \in C\}} \sum_{ij \in g \text{ st } i,j \in C} 1\{s_{ij} = h\}$$

$n^1 = 20$

σ^1, S^1

$n^2 = 50$

σ^1, S^1 | σ^2, S^2



$15 \leq n \leq 25$

all S

$35 \leq n \leq 65$

$S < \overline{S}$ | $S > \overline{S}$

Concluding remarks

- Networks is a theoretically rich tool
- Sill full of applications still unexplored
- Rapidly increasing topic in economics

- People in Paris: June Networks - Program
  - PhD Ccourse from Matt Jackson at Paris-sud, June 13th,
  - Workshop at Insead, Fontainebleau, June 18th,
  - Seminar at Paris Sud, by Matt Jackson, June 19th,
  - International conference at Carré des Sciences, Paris, June 28-29th, *www.adislab.org* .