Random Models

• Model $G(n, m)$ is a probability distribution $P(G)$ over all graphs with n nodes and m edges.
Properties of model = properties of ensemble
Examples:

- graph diameter $l(G)$ means $\langle l \rangle = \sum_G P(G) l(G) = \frac{1}{\Omega} \sum_G l(G)$
- degree $\langle d(\cdot) \rangle = 2m/n$

• Model $G(n, p)$ - graphs with n nodes and independent probability p for placing an edge between two vertices (aka Erdös-Rényi model).
Properties of model = properties of ensemble where G appears with prob

$$P(G) = p^m (1-p)^{\binom{n}{2} - m}$$

and probablity of drawing a graph with m edges from the ensemble is

$$P(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2} - m} \text{ and } \langle m \rangle = \sum_{m=0}^{\binom{n}{2}} m P(m) = \binom{n}{2} p$$

- mean degree $\displaystyle\sum_{m=0}^{\binom{n}{2}} \frac{2m}{n} P(m) = \frac{2}{n}\binom{n}{2} p = (n-1)p = c$

mean degree in a graph with exactly *m* edges

- degree distribution
  - node is connected to a particular k others $q_k = p^k(1-p)^{n-1-k}$
  - node is connected to exacly k others $p_k = \binom{n-1}{k} q_k$
  - in large-scale networks $p = c/(n-1)$ can be very small, i.e.,

$$\ln((1-p)^{n-1-k}) = (n-1-k)\ln(1-c/(n-1)) \approx -(n-1-k)\frac{c}{n-1} \approx -c$$

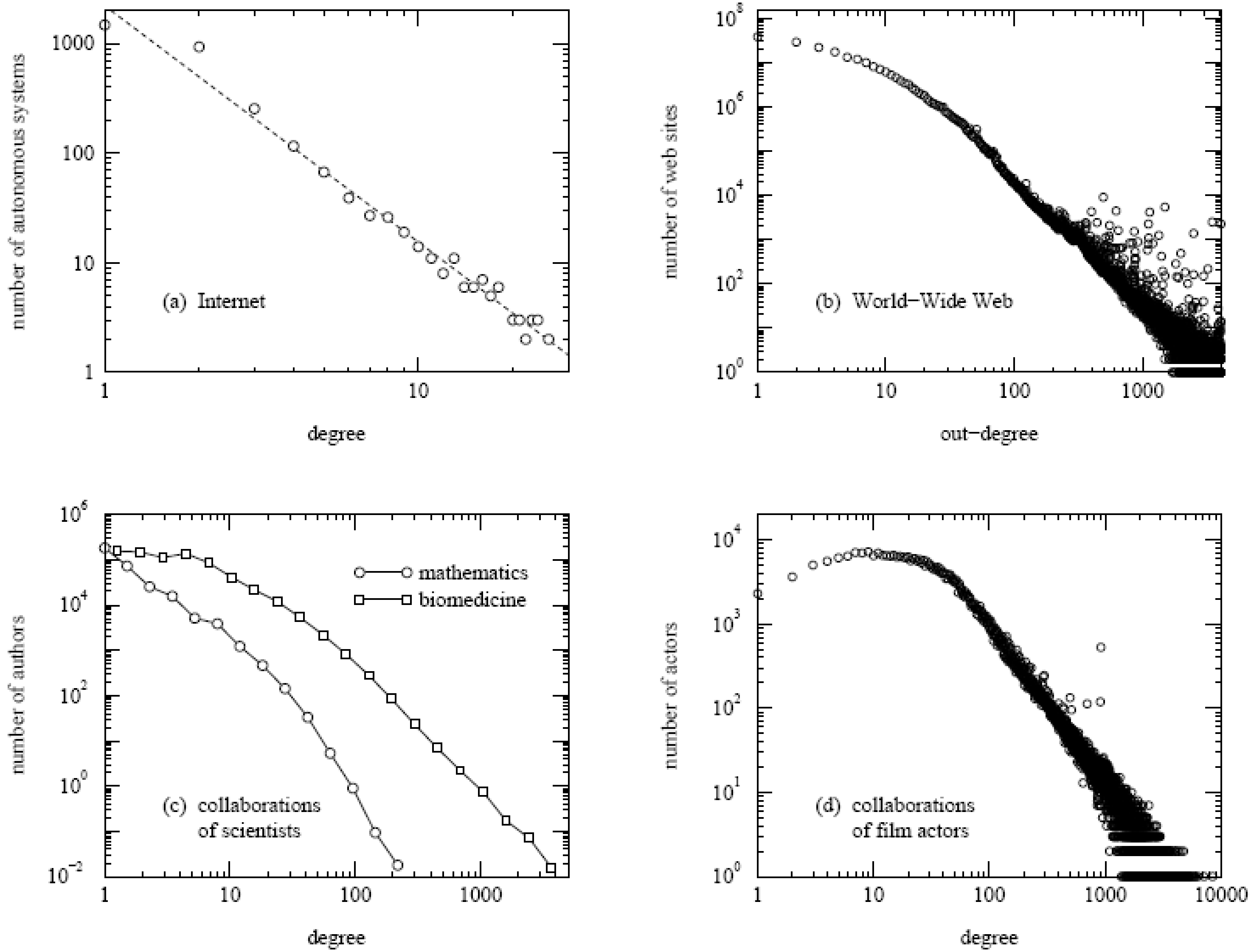
∞

Taylor series reminder: $\ln(1+\frac{1}{x}) = 2\left(A + \frac{1}{3}A^3 + \frac{1}{5}A^5 + ...\right)$, where $A = \frac{1}{2x+1}$

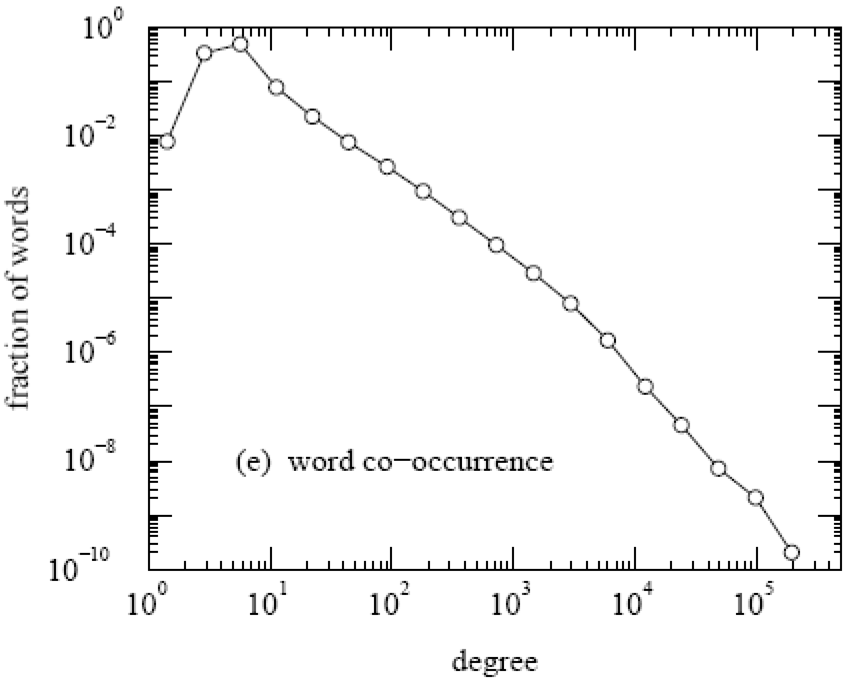
also if $\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!k!} \approx \frac{(n-1)^k}{k!}$ then

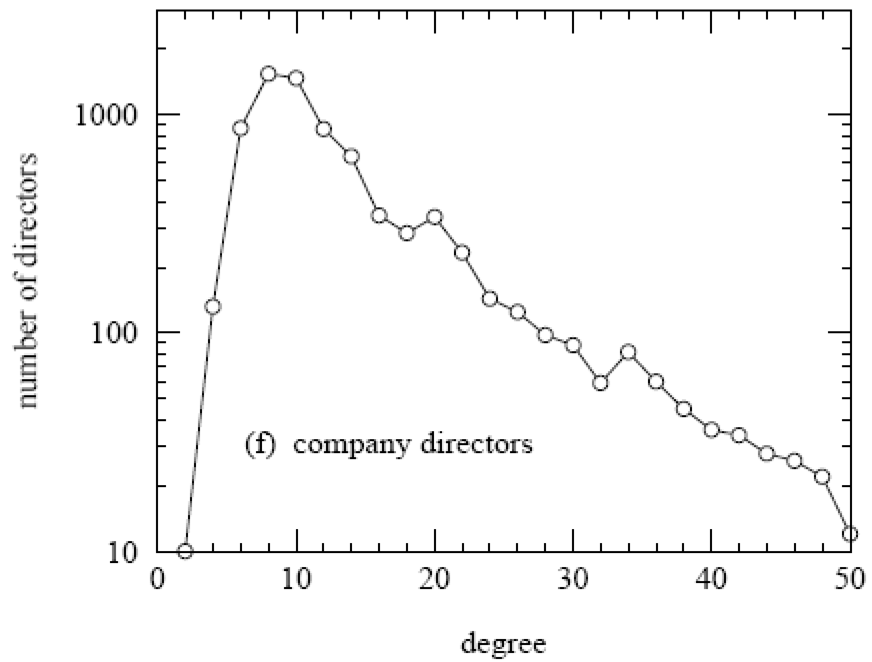
$$\boxed{p_k = \frac{(n-1)^k}{k!} p^k e^{-c} = \frac{(n-1)^k}{k!}\left(\frac{c}{n-1}\right)^k e^{-c} = e^{-c}\frac{c^k}{k!}}$$

In contrast to the degree distribution in random model …



In contrast to the degree distribution in random model …





- clustering coefficient $C = c/(n-1)$ = prob that any two nodes are neighbors

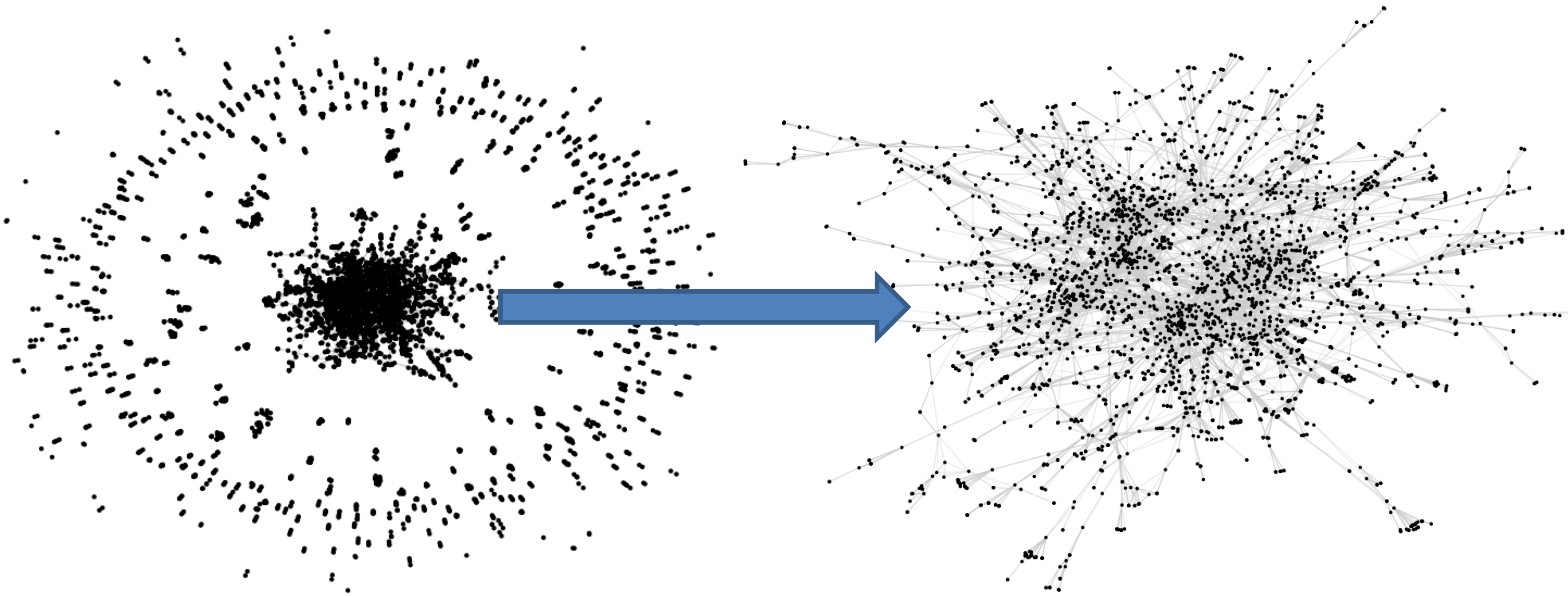
network	n	z	clustering coefficient C measured	clustering coefficient C random graph
Internet (autonomous systems)[a]	6 374	3.8	0.24	0.00060
World-Wide Web (sites)[b]	153 127	35.2	0.11	0.00023
power grid[c]	4 941	2.7	0.080	0.00054
biology collaborations[d]	1 520 251	15.5	0.081	0.000010
mathematics collaborations[e]	253 339	3.9	0.15	0.000015
film actor collaborations[f]	449 913	113.4	0.20	0.00025
company directors[f]	7 673	14.4	0.59	0.0019
word co-occurrence[g]	460 902	70.1	0.44	0.00015
neural network[c]	282	14.0	0.28	0.049
metabolic network[h]	315	28.3	0.59	0.090
food web[i]	134	8.7	0.22	0.065

Newman, “Random graphs as models of networks”

- giant component in $G(n, p)$

Giant component is a network component whose size grows in proportion to n.
u = avg fraction of vertices that do not belong to the giant component.

Q: When p=0 then |gc|=1; when p=1 then |gc|=n. What is the difference between them?



Co-authorship network

its largest connected component

- giant component in $G(n,p)$

Giant component is a network component whose size grows in proportion to n.
u = avg fraction of vertices that do not belong to the giant component.

Q: When p=0 then |gc|=1; when p=1 then |gc|=n. Is this transition smooth? Is there a point of transition?

i does not belong to gc if for every node *j* either

a) there is no edge *ij,* prob = *1-p*

b) there is edge *ij* but *j* is not in giant component, prob = *pu*

$\Pr[i \text{ does not belong to gc via } j] = 1 - p + pu$, i.e.,
total probability of not being connected to gc via any of $n-1$ other vertices is

$$u = (1-p+pu)^{n-1} = \left(1 - \frac{c}{n-1}(1-u)\right)^{n-1}$$

$$\ln u \overset{n\to\infty}{\approx} -(n-1)\frac{c}{n-1}(1-u) = -c(1-u) \;\Rightarrow\; u = e^{-c(1-u)} \;\Rightarrow\; S = 1 - e^{-cS}$$

vertices in giant component

$$S = 1 - e^{-cS}$$

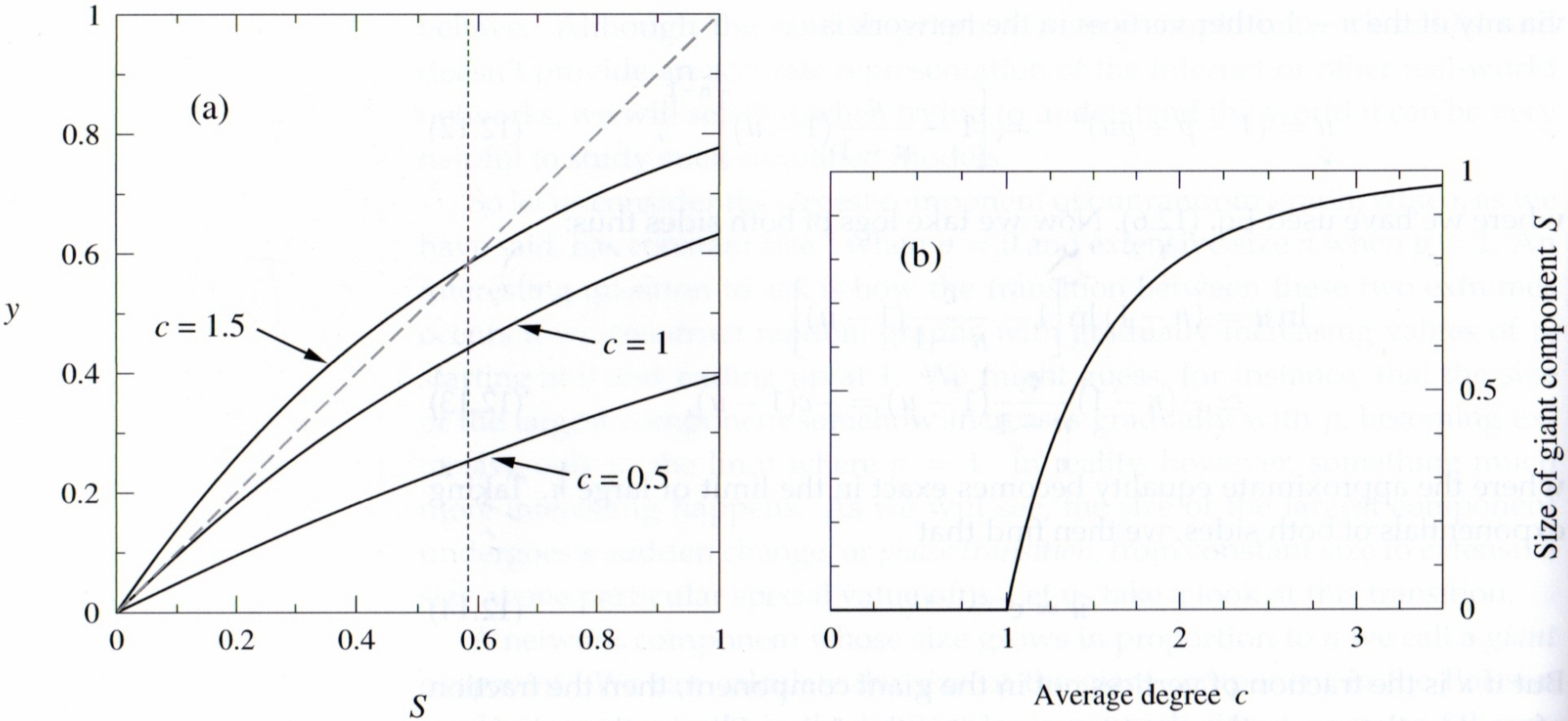


Figure 12.1: Graphical solution for the size of the giant component. (a) The three curves in the left panel show $y = 1 - e^{-cS}$ for values of c as marked, the diagonal dashed line shows $y = S$, and the intersection gives the solution to Eq. (12.15), $S = 1 - e^{-cS}$. For the bottom curve there is only one intersection, at $S = 0$, so there is no giant component, while for the top curve there is a solution at $S = 0.583\ldots$ (vertical dashed line). The middle curve is precisely at the threshold between the regime where a non-trivial solution for S exists and the regime where there is only the trivial solution $S = 0$. (b) The resulting solution for the size of the giant component as a function of c.

=> Demo in Matlab

Newman “Networks, An Introduction”

	Medline	Physics E-print Archive complete	astro-ph	cond-mat	hep-th	SPIRES	NCSTRL
total papers	2163923	98502	22029	22016	19085	66652	13169
total authors	1520251	52909	16706	16726	8361	56627	11994
first initial only	1090584	45685	14303	15451	7676	47445	10998
mean papers per author	6.4(6)	5.1(2)	4.8(2)	3.65(7)	4.8(1)	11.6(5)	2.55(5)
mean authors per paper	3.754(2)	2.530(7)	3.35(2)	2.66(1)	1.99(1)	8.96(18)	2.22(1)
collaborators per author	18.1(1.3)	9.7(2)	15.1(3)	5.86(9)	3.87(5)	173(6)	3.59(5)
size of giant component	1395693	44337	14845	13861	5835	49002	6396
first initial only	1019418	39709	12874	13324	5593	43089	6706
as a percentage	92.6(4)%	85.4(8)%	89.4(3)	84.6(8)%	71.4(8)%	88.7(1.1)%	57.2(1.9)%
2nd largest component	49	18	19	16	24	69	42
clustering coefficient C	0.066(7)	0.43(1)	0.414(6)	0.348(6)	0.327(2)	0.726(8)	0.496(6)
mean distance	4.6(2)	5.9(2)	4.66(7)	6.4(1)	6.91(6)	4.0(1)	9.7(4)
maximum distance	24	20	14	18	19	19	31

Table 1: Summary of results of the analysis of seven scientific collaboration networks. Numbers in parentheses give an estimate of the error on the least significant figures.

	Network	Type	n	m	c	S
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980
	Company directors	Undirected	7 673	55 392	14.44	0.876
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16	
	Email messages	Directed	59 812	86 300	1.44	0.952
	Email address books	Directed	16 881	57 029	3.38	0.590
	Student dating	Undirected	573	477	1.66	0.503
	Sexual contacts	Undirected	2 810			
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914
	Citation network	Directed	783 339	6 716 198	8.57	
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000
Technological	Internet	Undirected	10 697	31 992	5.98	1.000
	Power grid	Undirected	4 941	6 594	2.67	1.000
	Train routes	Undirected	587	19 603	66.79	1.000
	Software packages	Directed	1 439	1 723	1.20	0.998
	Software classes	Directed	1 376	2 213	1.61	1.000
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996
	Protein interactions	Undirected	2 115	2 240	2.12	0.689
	Marine food web	Directed	134	598	4.46	1.000
	Freshwater food web	Directed	92	997	10.84	1.000
	Neural network	Directed	307	2 359	7.68	0.967

- two giant components in $G(n, p)$?
  - Generate G in two steps: (a) with $p = c/(n-1)$ and, then (b) with $p' = c/(n-1)^{3/2}$. Now average degree is

$$c' = (n-1)(p+p') = (n-1)\left(\frac{c}{n-1} + \frac{c}{(n-1)^{3/2}}\right) = c(1+\frac{1}{\sqrt{n-1}}) \overset{n\to\infty}{\approx} c,$$

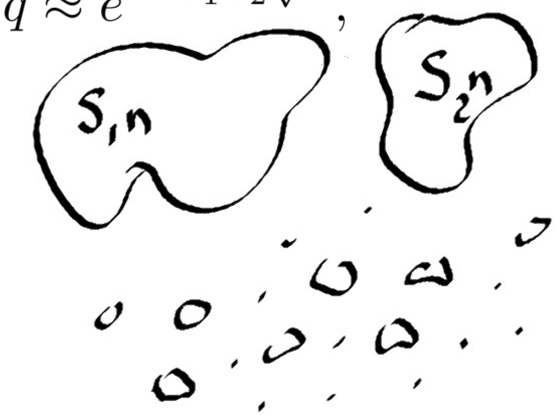
large

i.e., we generated G with the same mean degree.

  - Suppose #gc$\geq$ 2 after adding edges with prob p only $(S_1, S_2, ...)$
  - Add edges with prob p'. S_1 and S_2 remain separate with probability

$$q = (1-p')^{S_1 S_2 n^2} \;\Rightarrow\; \ln q \approx -cS_1S_2\sqrt{n} \;\Rightarrow\; q \approx e^{-cS_1S_2\sqrt{n}},$$

i.e., $q \overset{n\to\infty}{\to} 0$



Conclusion: In the limit of large *n*, the probability of existence of two separate giant components goes to zero.

- sizes of small components

  π_s is the probability that randomly chosen node belongs to a small component of size s.

  - We cannot normalize π_s to unity because some nodes may belong to the giant component, i.e.,

    fraction of nodes in gc

    $$\sum_{s=0}^{\infty} \pi_s = 1 - S.$$

  - **Observation**: small components are likely to be trees.
    Consider a small tree component of s nodes. The total number of places we can add an extra edge to is $\binom{s}{2} - (s-1)$

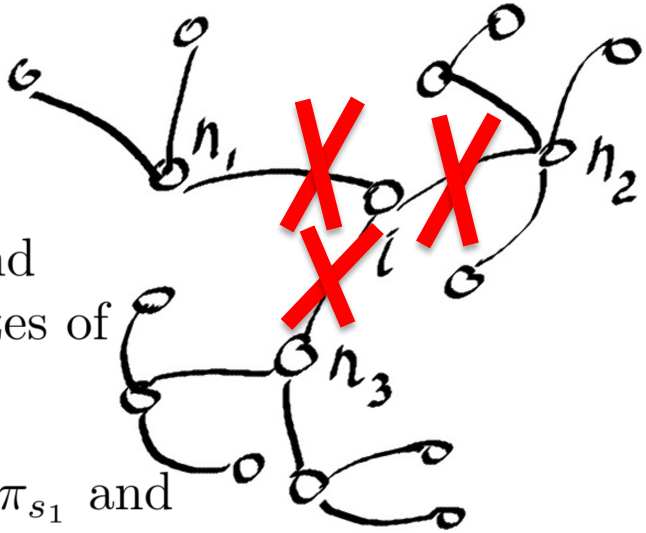
    edges in tree

    edge prob

    $$\text{Average total number of added edges } \frac{1}{2}(s-1)(s-2) \cdot \frac{c}{n-1} \xrightarrow{n \to \infty} 0$$

    the component is still tree

Calculation of π_s (the probability that randomly chosen node belongs to a small component of size s).



- Consider node i in a small (tree) component
- ... and modified network with deleted i.
  In the modified network, prob p is the same and in the limit of n the changes are negligible. Sizes of gc and sc will be indistinguishable for same p.
- Suppose $d(i) = k$ and $\Pr[n_1 \in \text{ sc of size } s_1] = \pi_{s_1}$ and

$$\Pr[\forall j \in N(i)\ n_j \in \text{ sc of size } s_j] = \Pi_{j=1}^{k} \pi_{s_j}$$

Since $\sum_{j \in N(i)} s_j = s - 1$ we have

Kronecker delta

$$p_k = e^{-c}\frac{c^k}{k!} \qquad \Pr[s|k] = \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left(\Pi_{j=1}^{k} \pi_{s_j}\right) \delta(s-1, \sum_j s_j)$$

$$\pi_s = \sum_{k=0}^{\infty} p_k \Pr[s|k] = e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left(\Pi_{j=1}^{k} \pi_{s_j}\right) \delta(s-1, \sum_j s_j)$$

$$\pi_s = \sum_{k=0}^{\infty} p_k \Pr[s|k] = e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left(\Pi_{j=1}^{k} \pi_{s_j}\right) \delta(s-1, \sum_j s_j)$$

One way to evaluate π_s is by using generating function

$$h(z) = \sum_{s=1}^{\infty} \pi_s z^s \Rightarrow \langle s \rangle = \frac{\sum_s s\pi_s}{\sum_s \pi_s} = h'(1)/(1-S) = 1/(1-c+cS).$$

see handout, pp 412-413

Average size of the small components in a random model does not grow with the number of vertices.

Average component size

$$R = \frac{2}{2-c+cS}$$

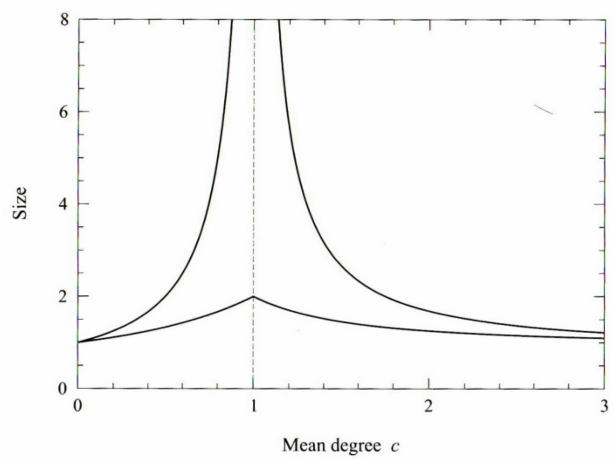


Figure 12.4: Average size of the small components in a random graph. The upper curve shows the average size $\langle s \rangle$ of the component to which a randomly chosen vertex belongs, calculated from Eq. (12.34). The lower curve shows the overall average size R of a component, calculated from Eq. (12.40). The dotted vertical line marks the point $c = 1$ at which the giant component appears. Note that, as discussed in the text, the upper curve diverges at this point but the lower one does not.

Distribution of component sizes

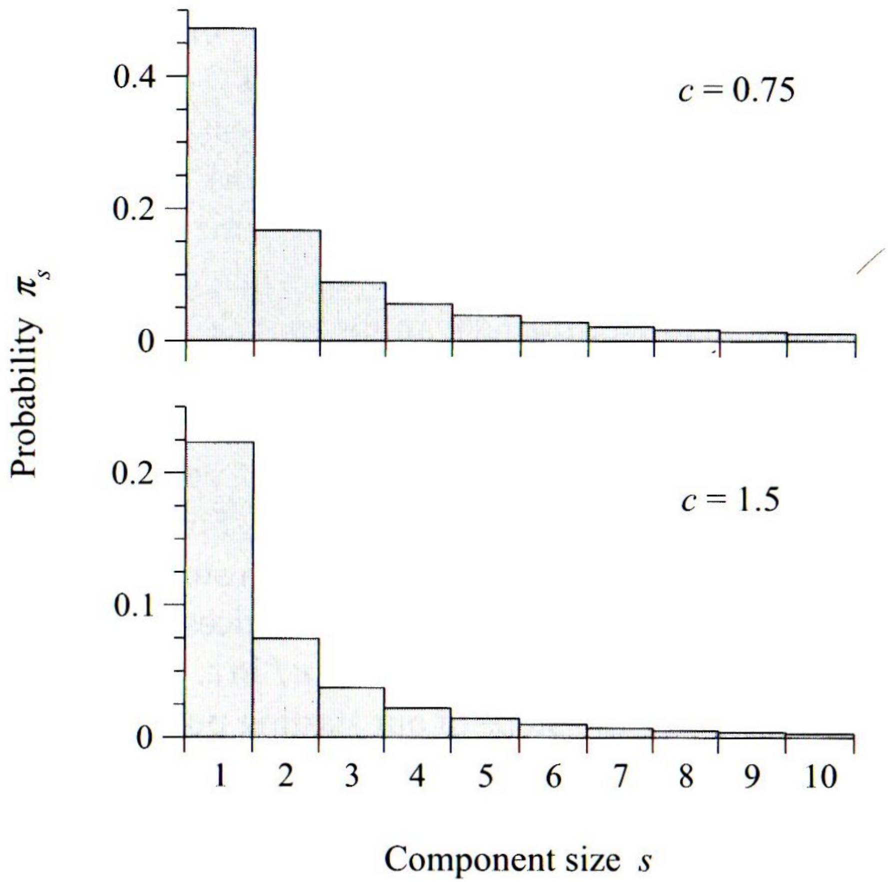


Figure 12.5: Sizes of small components in the random graph. This plot shows the probability π_s that a randomly chosen vertex belongs to a small component of size s in a Poisson random graph with $c = 0.75$ (top), which is in the regime where there is no giant component, and $c = 1.5$ (bottom), where there is a giant component.