Graphs and Networks
Oct 31: Lecture 16
Daniel A. Spielman

Some Graphs

The World-Wide Web,
(or a managable portion of it)

Internet:
Connections among routers, or
BGP relations among Autonomous Systems

Actors, with edges if have appeared in the same movie

Co-authorship
vertices = authors
edges between pairs who've published together

Some Graphs

Electrical Grid:

vertices = generators, transformers, substations

edges = high-power transmisison lines

Neural Network of worm – Caenorhabditis elegans

Conformation space of lattice polymer chain

vertex = configuration

edge = can reach one from another

Protein interaction network (yeast S. verevisiae)

vertex = protein

edge between proteins that interact

Some Graphs

From data in R^n

data point → vertex

edge between u and v

of wt $\frac{1}{dist(u,v)}$

or wt $e^{-\frac{1}{\sigma^2}(dist(u,v)^2)}$

Some Graphs

Netflix Challenge:

100 million ratings of 1-5, from
480,000 customers
18,000 movies

$1,000,000 to whoever can best predict scores.

Average score has RMS error 1.0540
Cinematch RMS error 0.9525
To win grand prize, need **0.8572**

Current leader: 0.9052

Small-World/Low Diameter

Bollobas-Chung ('88):

A cycle plus random matching probably has diameter O(log n)

Strogatz-Watts (Nature vol 393, 4 June 1998):

Experiments on k-regular graphs on n-nodes

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}
Film actors	3.65	2.99
Power grid	18.7	12.4
C. elegans	2.65	2.25

Actors: n = 225226, k = 61

Power grid: n = 4,941, k = 2.67

C. elegans, n = 282, k = 14

Diameter in scientific collaborations

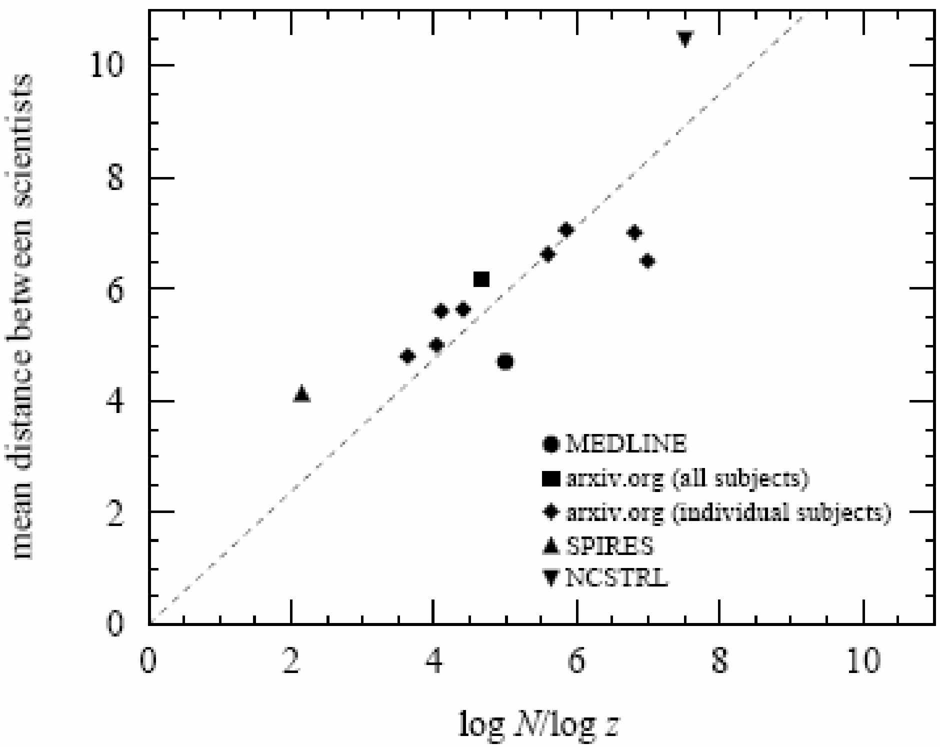


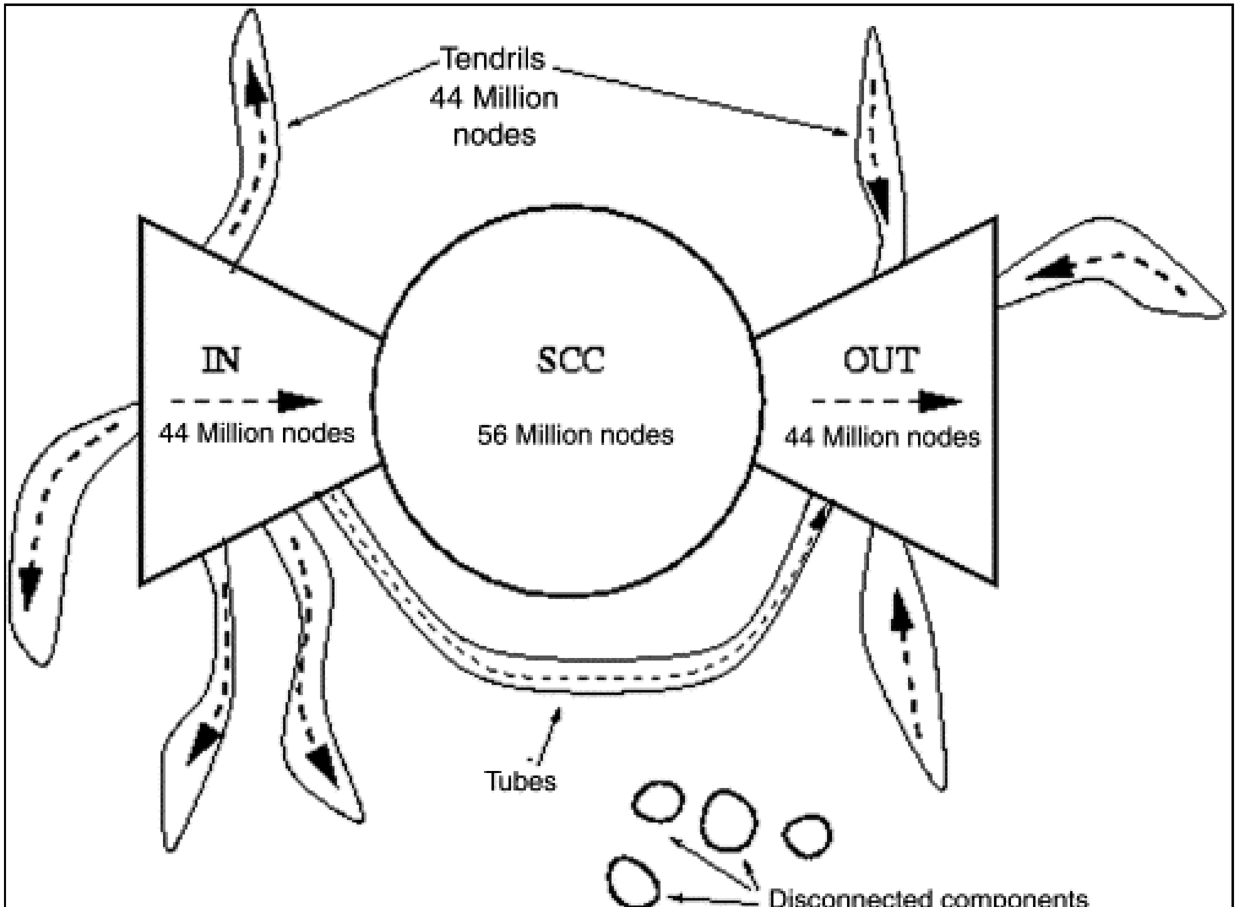
FIG. 3. Average distance between pairs of scientists in the various communities, plotted against the average distance on a random graph of the same size and average coordination number. The dotted line is the best fit to the data which also passes through the origin.

The Structure of Scientific Collaboration Networks, M.E.J. Newman
http://arxiv.org/abs/cond-mat/0007214/

Graph Structure in the Web

Broder et. al., Computer Networks 33, (2000) pp.309-320

Altavista Crawl of 200M pages, 1.5B links



Graph Structure in the Web

Broder et. al., Computer Networks 33, (2000) pp.309-320

For two random nodes, directed path exists with prob 25%

Edge type	In-links (directed)	Out-links (directed)	Undirected
Average connected distance	16.12	16.18	6.83

Breadth-first search from random nodes in SCC:

Measure	Minimum depth	Average depth	Maximum depth
In-links	475	482	503
Out-links	430	434	444

As the table shows, from some nodes in the SCC it is possible to complete the search at distance 475, while from other nodes distance 503 is required. This allows us to conclude that the directed diameter of SCC is at least 28.

Clustering Coefficient:

If node i has k nbrs, could be k(k-1)/2 triangles at node k.
C_i = number of triangles at node i / (k(k-1)/2)

$C = \text{ave}_i \; C_i$

alternatively, can use

C = 6 x number triangles / number length-two paths

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Scientific Collaboration Networks

	MEDLINE	Los Alamos e-Print Archive				SPIRES	NCSTRL
		complete	astro-ph	cond-mat	hep-th		
total papers	2156769	98502	22029	22016	19085	66652	13169
total authors	1388989	52909	16706	16726	8361	56627	11994
first initial only	1006412	45685	14303	15451	7676	47445	10998
mean papers per author	5.5(4)	5.1(2)	4.8(2)	3.65(7)	4.8(1)	11.6(5)	2.55(5)
mean authors per paper	2.966(2)	2.530(7)	3.35(2)	2.66(1)	1.99(1)	8.96(18)	2.22(1)
collaborators per author	14.8(1.1)	9.7(2)	15.1(3)	5.86(9)	3.87(5)	173(6)	3.59(5)
cutoff z_c	7300(2700)	52.9(4.7)	49.0(4.3)	15.7(2.4)	9.4(1.3)	1200(300)	10.7(1.6)
exponent τ	2.5(1)	1.3(1)	0.91(10)	1.1(2)	1.1(2)	1.03(7)	1.3(2)
size of giant component	1193488	44337	14845	13861	5835	49002	6396
first initial only	892193	39709	12874	13324	5593	43089	6706
as a percentage	87.3(7)%	85.4(8)%	89.4(3)	84.6(8)%	71.4(8)%	88.7(1.1)%	57.2(1.9)%
2nd largest component	56	18	19	16	24	69	42
mean distance	4.4(2)	5.9(2)	4.66(7)	6.4(1)	6.91(6)	4.0(1)	9.7(4)
maximum distance	21	20	14	18	19	19	31
clustering coefficient C	0.072(8)	0.43(1)	0.414(6)	0.348(6)	0.327(2)	0.726(8)	0.496(6)

Emergence of Scaling in Random Networks, Barabasi and Albert, Science, vol 286, 15 Oct 1999

Degree distributions

Power-law when

x_k = # nodes deg k

$\approx c \cdot k^{-\alpha}$

for some constants c, α

$$p_k = \frac{x_k}{\sum_i x_i}$$

get $\log(p_k) = \log(c) - \alpha \log(k)$

Power-law degree distributions

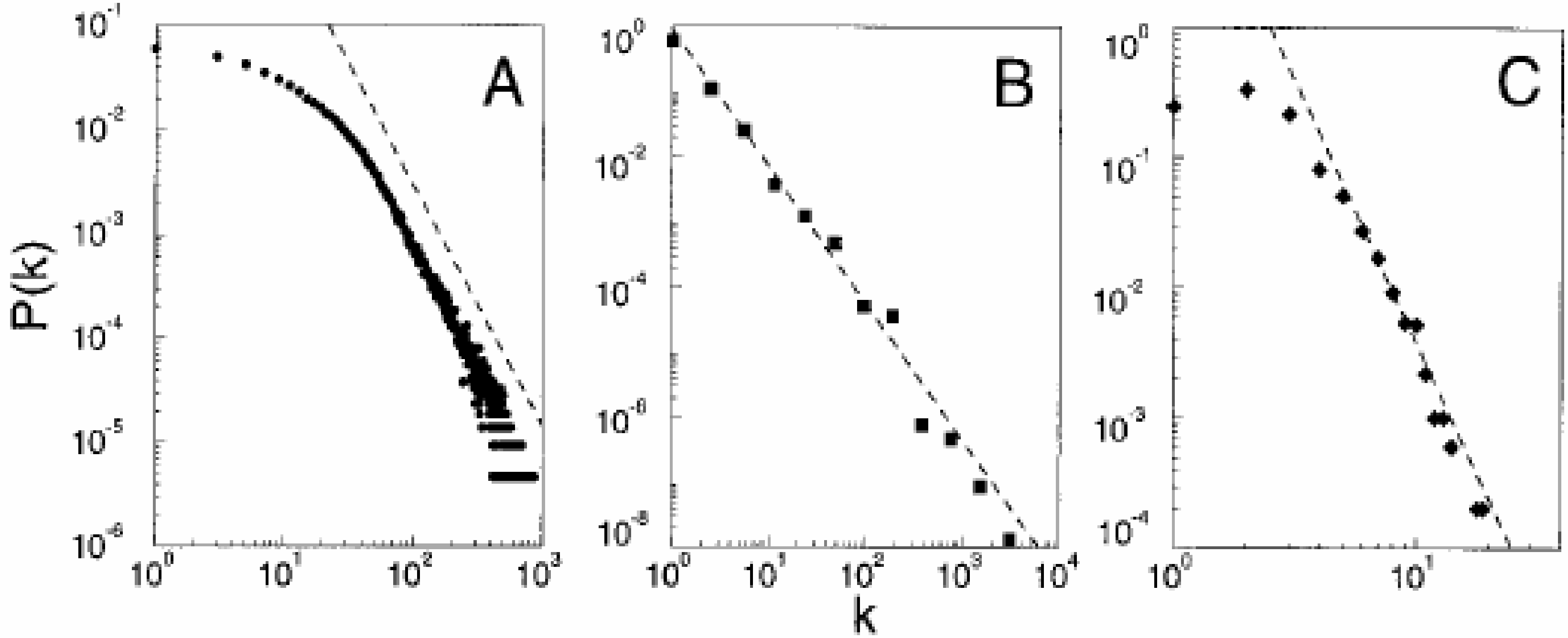


Fig. 1. The distribution function of connectivities for various large networks. (**A**) Actor collaboration graph with $N = 212{,}250$ vertices and average connectivity $\langle k \rangle = 28.78$. (**B**) WWW, $N = 325{,}729$, $\langle k \rangle = 5.46$ (6). (**C**) Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$.

Emergence of Scaling in Random Networks, Barabasi and Albert, Science, vol 286, 15 Oct 1999

Power-Law Degree Distributions?

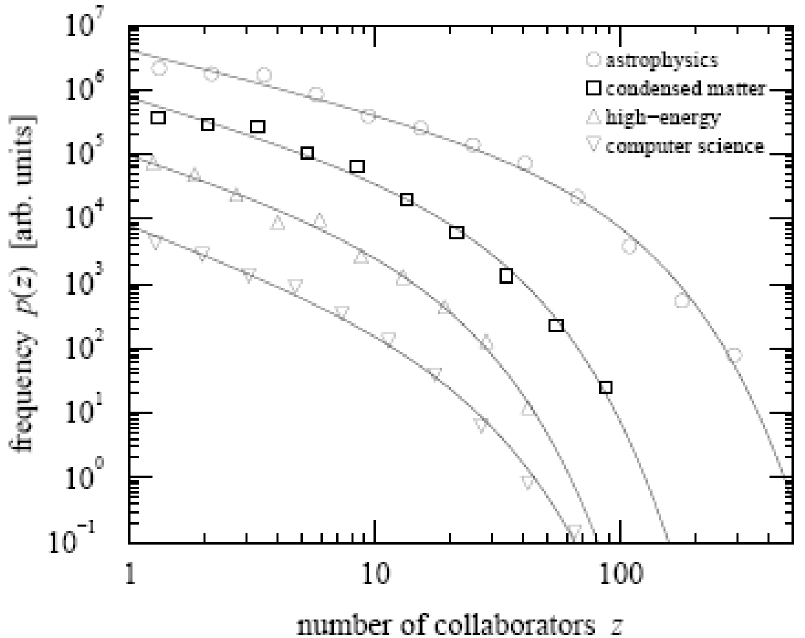


FIG. 1. Histograms of the number of collaborators of scientists in four of the databases studied here. The solid lines are least-squares fits to Eq. (1).

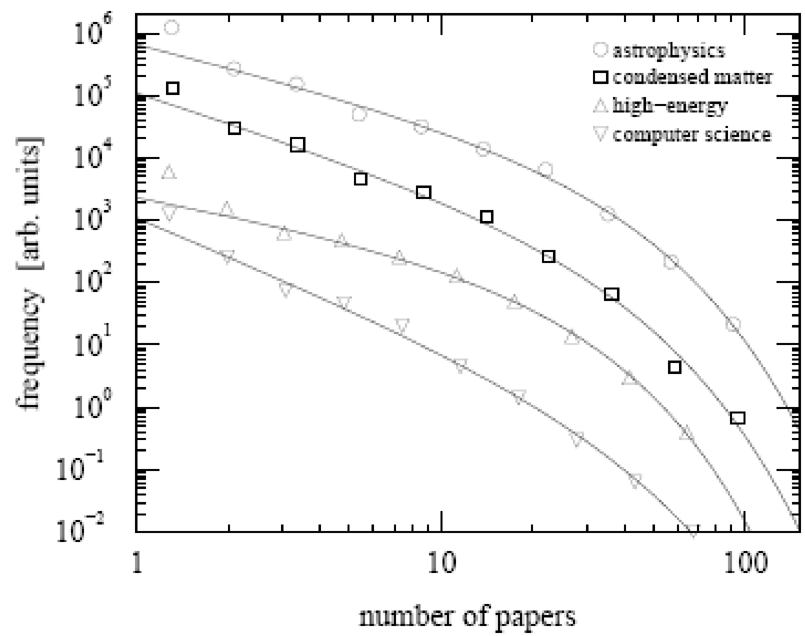


FIG. 2. Histograms of the number of papers written by scientists in four of the databases. As with Fig. 1, the solid lines are least-squares fits to Eq. (1).

used. However, our data are well fitted by a power-law form with an exponential cutoff:

$$P(z) \sim p^{-\tau} e^{-z/z_c}, \qquad (1)$$

The Structure of Scientific Collaboration Networks, M.E.J. Newman
http://arxiv.org/abs/cond-mat/0007214/

Diameter in scientific collaborations

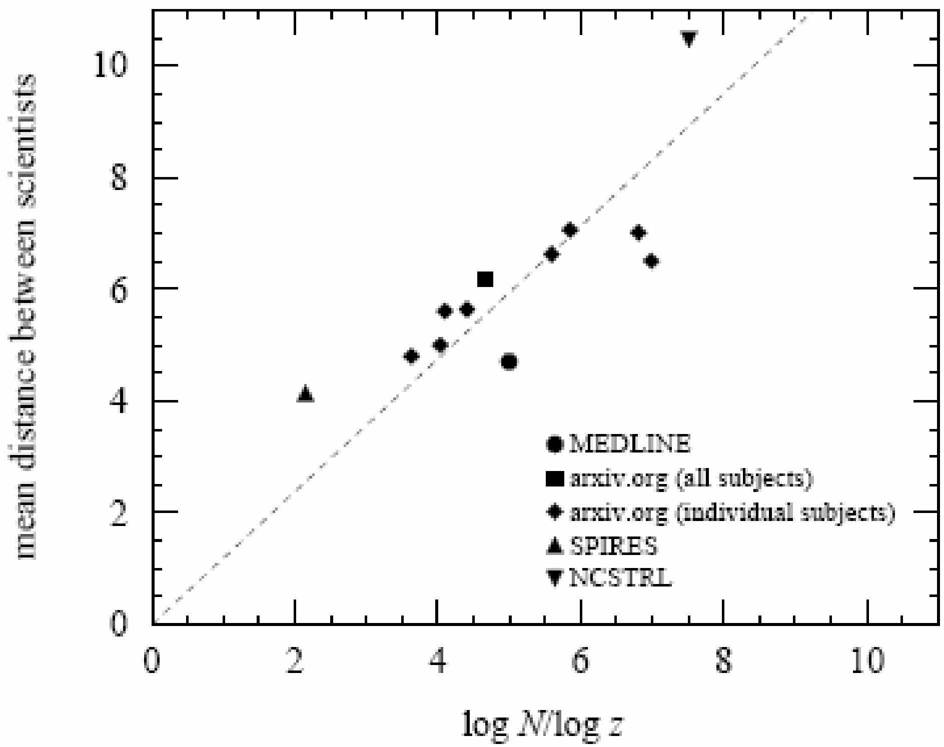


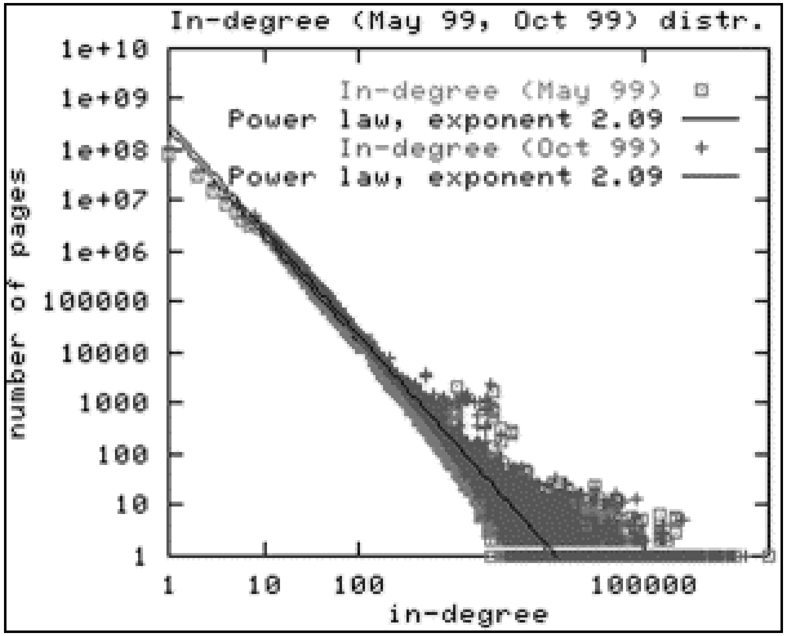
FIG. 3. Average distance between pairs of scientists in the various communities, plotted against the average distance on a random graph of the same size and average coordination number. The dotted line is the best fit to the data which also passes through the origin.

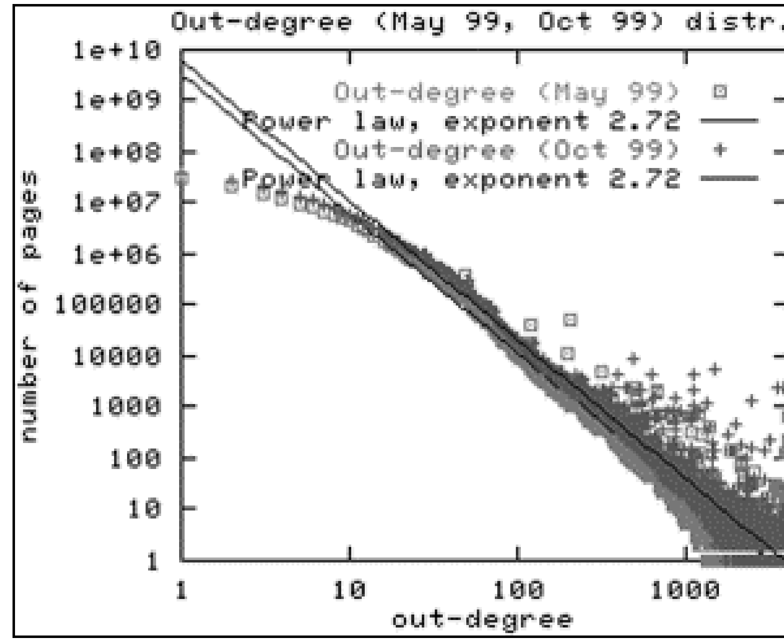
Explain below-line points by power-law degree distribution:

Bollobas-Riordan.

The Structure of Scientific Collaboration Networks, M.E.J. Newman
http://arxiv.org/abs/cond-mat/0007214/

Power-law degree distributions





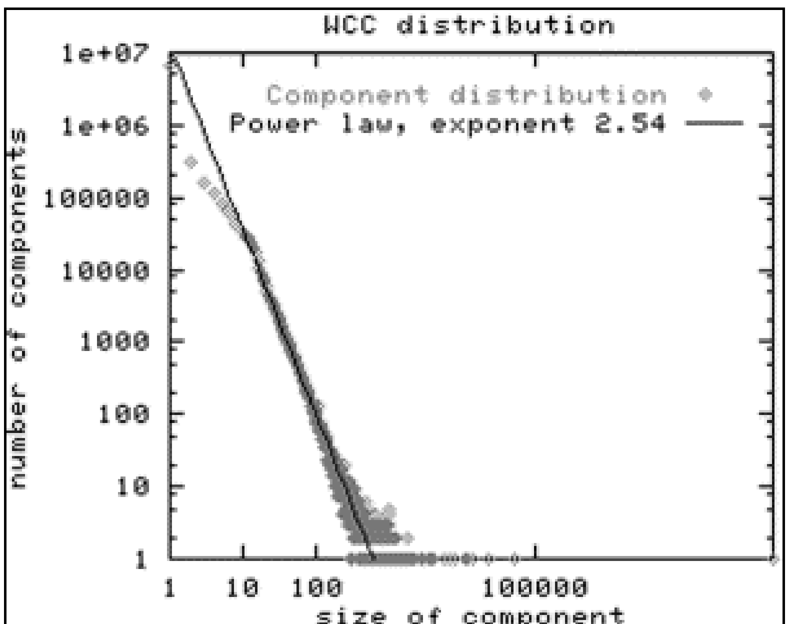
Graph structure in the Web, Broder et. al.,

Computer Networks Vol 33, No 1-6 , June 2000, pp. 309-320

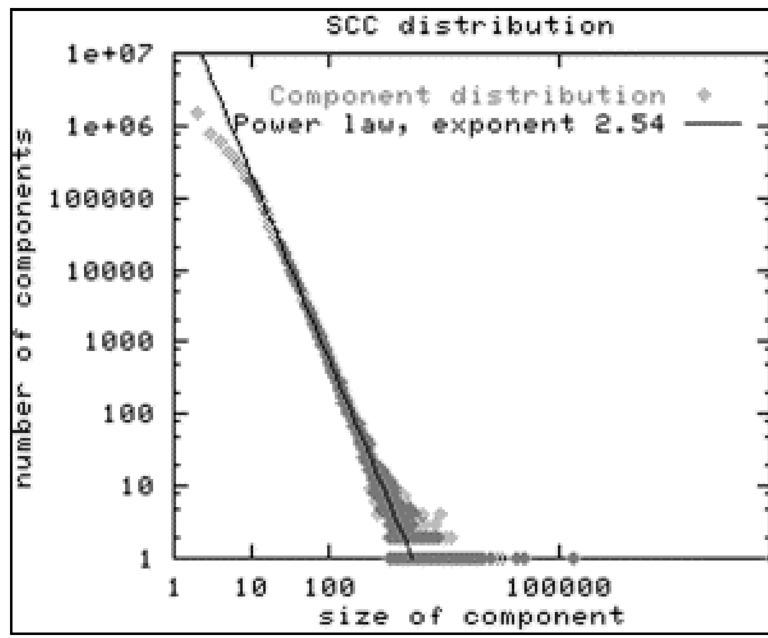
Analysis of Web Graph

Graph structure in the Web, Broder et. al.,
Computer Networks Vol 33, No 1-6 , June 2000, pp. 309-320

Weakly-connected components (traverse edge either way)
largest had 186m pages = 91%



Strongly-connected components (only following links)
largest had 56m pages = 28%



“Scale-Free”

If $\Pr[X > x] \sim c x^{-\alpha}$

$$\Pr[X > x \mid X > w] = \frac{\Pr[X > x]}{\Pr[X > w]}$$

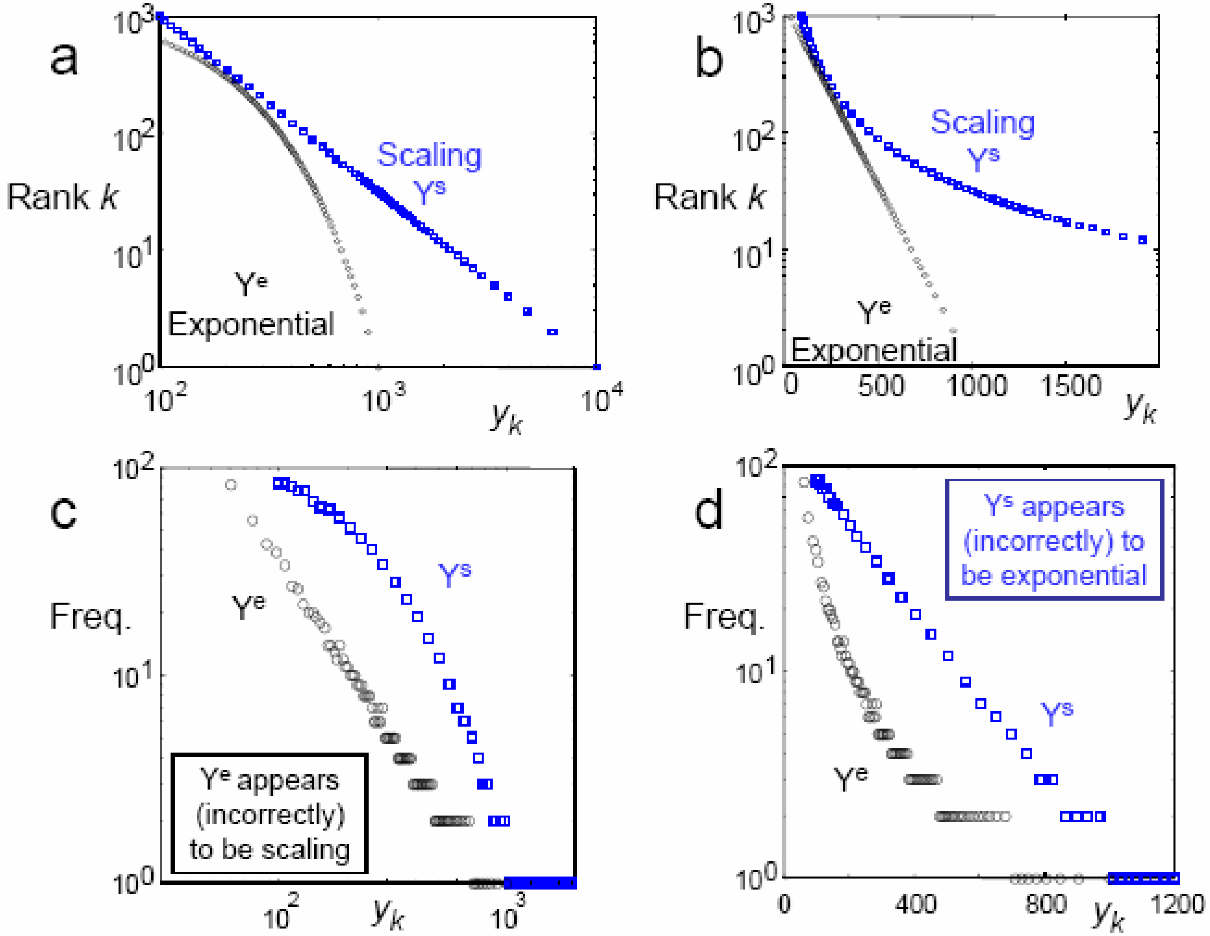
$$\sim \left(\frac{x}{w}\right)^{-\alpha}$$

In contrast,

if $\Pr[X > x] \sim e^{-\lambda x}$,

$$\Pr[X > x \mid X > w] \sim e^{-\lambda(x-w)}$$

Problems with these plots:



Towards a theory of scale-free graphs:
Li, Alderson, Ranaka, Doyle, Willinger,
arXiv:cond-mat/0501169 v2 18 Oct 2005

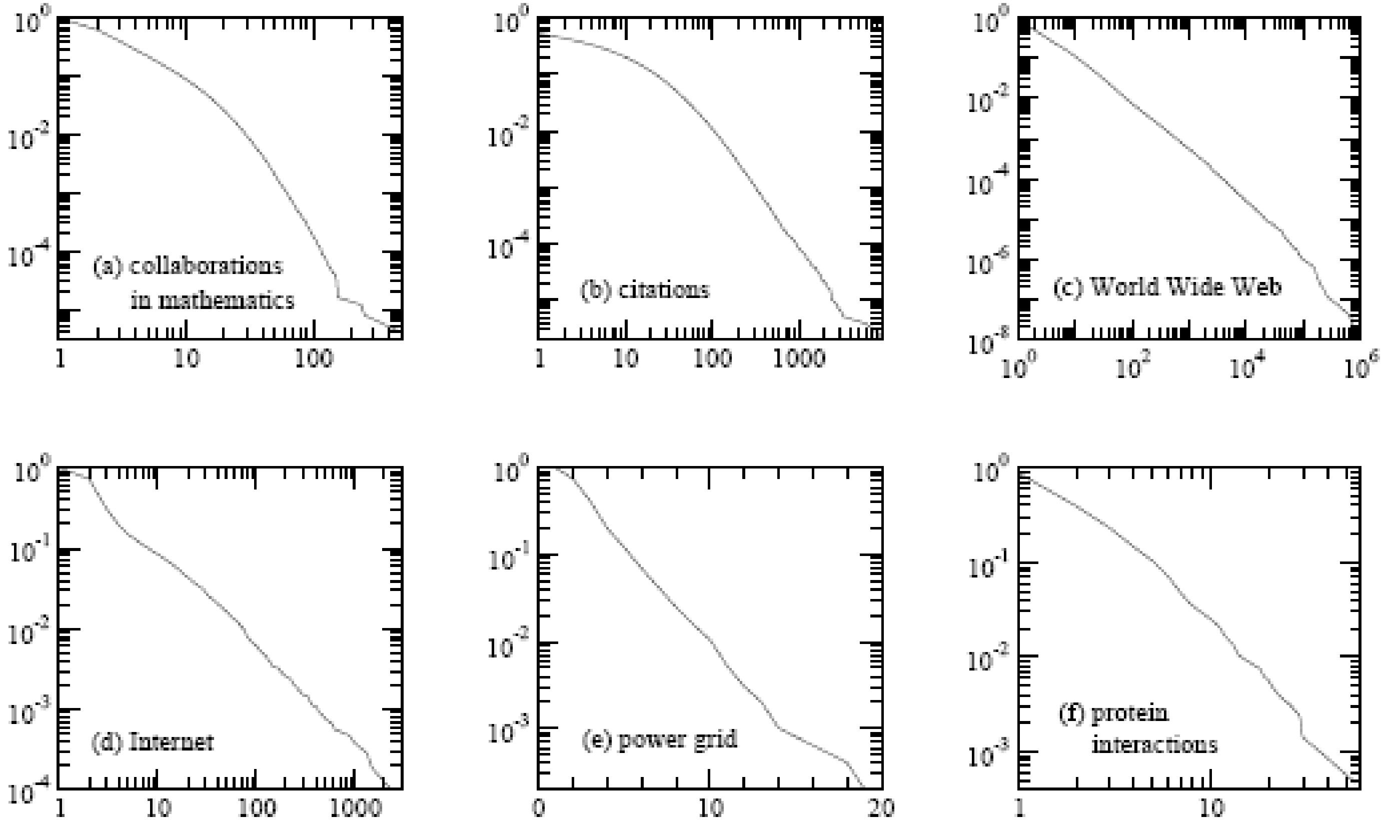
Better to take complementary CDF:

$$P_k = \sum_{k'=k}^{\infty} p_{k'} = \Pr[X \geq k]$$

if $p_k \sim k^{-\alpha}$

$$P_k \sim k^{-(\alpha-1)}$$

Power-Law Degree Distributions?



The Structure and Function of Complex Networks
M.E.J. Newman, cond-mat/0303516 v1 15 Mar 2003

Assortative

If high-degree nodes likely connected

If p_k = prob node has deg k

prob end of rand edge deg $k \sim k p_k$

prob end of rand edge has k more edges $\sim (k+1) p_{k+1}$

$$q_k = \frac{(k+1) p_{k+1}}{\sum_j j p_j}$$

Assortative

	network	n	r
real-world networks	physics coauthorship[a]	52 909	0.363
	biology coauthorship[a]	1 520 251	0.127
	mathematics coauthorship[b]	253 339	0.120
	film actor collaborations[c]	449 913	0.208
	company directors[d]	7 673	0.276
	Internet[e]	10 697	−0.189
	World-Wide Web[f]	269 504	−0.065
	protein interactions[g]	2 115	−0.156
	neural network[h]	307	−0.163
	food web[i]	92	−0.276
models	random graph[u]		0
	Callaway *et al.*[v]		$\delta/(1+2\delta)$
	Barabási and Albert[w]		0

Assortative Mixing in Networks, M.E.J. Newman
cond-mat/0205405 v1 20 May 2002

	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

The Structure and Function of Complex Networks

M.E.J. Newman, cond-mat/0303516 v1 15 Mar 2003