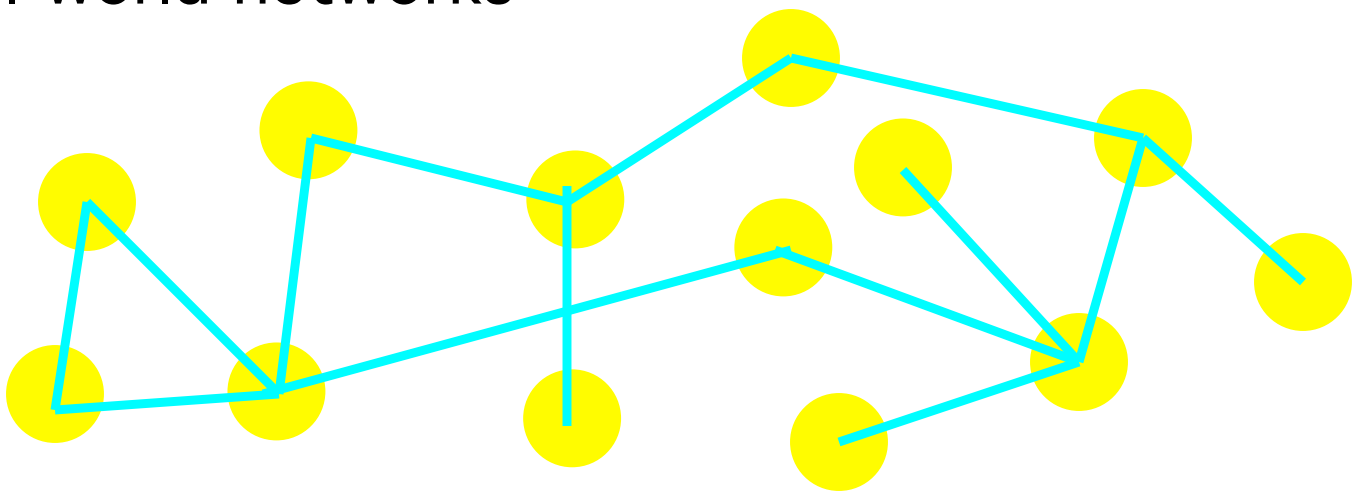
Social Networks

- Characterize the connections between identifiable parts of a system
- Graph theory
  - Vertices = nodes of a graph
  - Edges = connections between vertices
  - Degree of a vertex = # of connections that it has
- Scale-free networks
- Small-world networks



Random Networks (Erdos & Renyi, 1960)

- N vertices, connected by an edge with probability p
- Degrees follow a Poisson distribution
  - Poisson distribution approximates Binomial if P is small and N is large (e.g. accidents, prairie dogs, customers). The probability of obtaining x occurrences of A when the average number of occurrences is □ is:

$$F(x) = e^{-\lambda}\frac{\lambda^{x}}{x!}$$

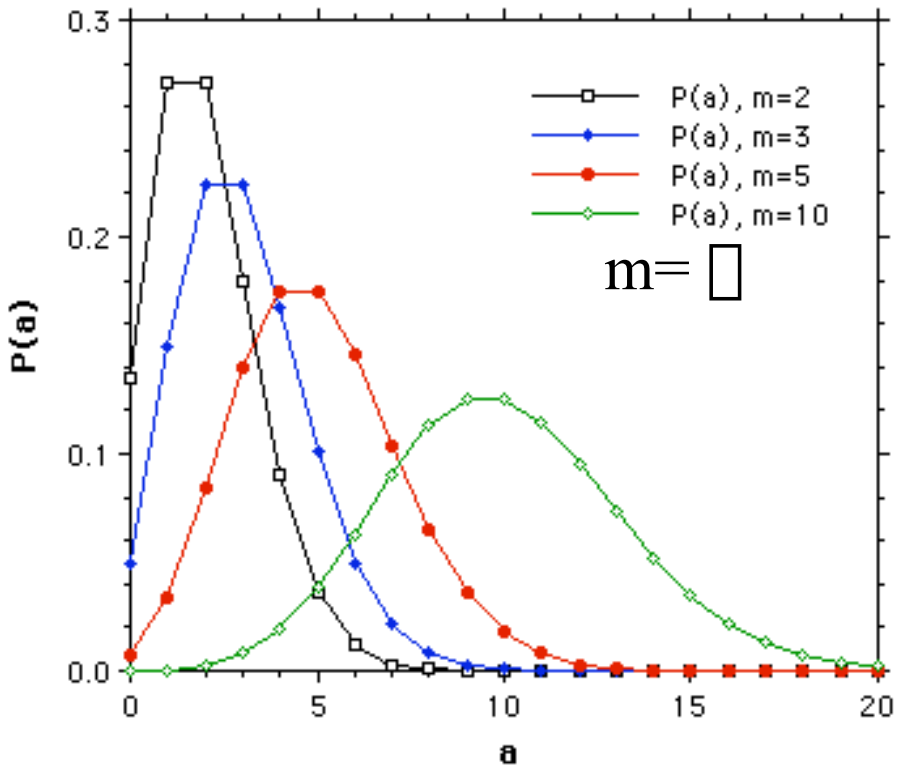
For networks, x = degree of a vertex,

F(x) = number of vertices with x connections

  - Binomial function describes the probability of obtaining x occurrences of event A when each of N events is independent of the others, and the probability of event A on any trial is P:

$$F(x) = \frac{N!}{x!(N-x)!}P^{x}(1-P)^{N-x}$$

Poisson distribution of degrees in a random network



$$\square = N\left[\frac{N \square 1}{k}\right] p^k (1 \square p)^{N \square 1 \square k}$$

□=Where N = Number of vertices, p = probability of each pair of vertices being connected, k= number of edges

Probability of finding a highly connected vertex decreases exponentially for k >> average k

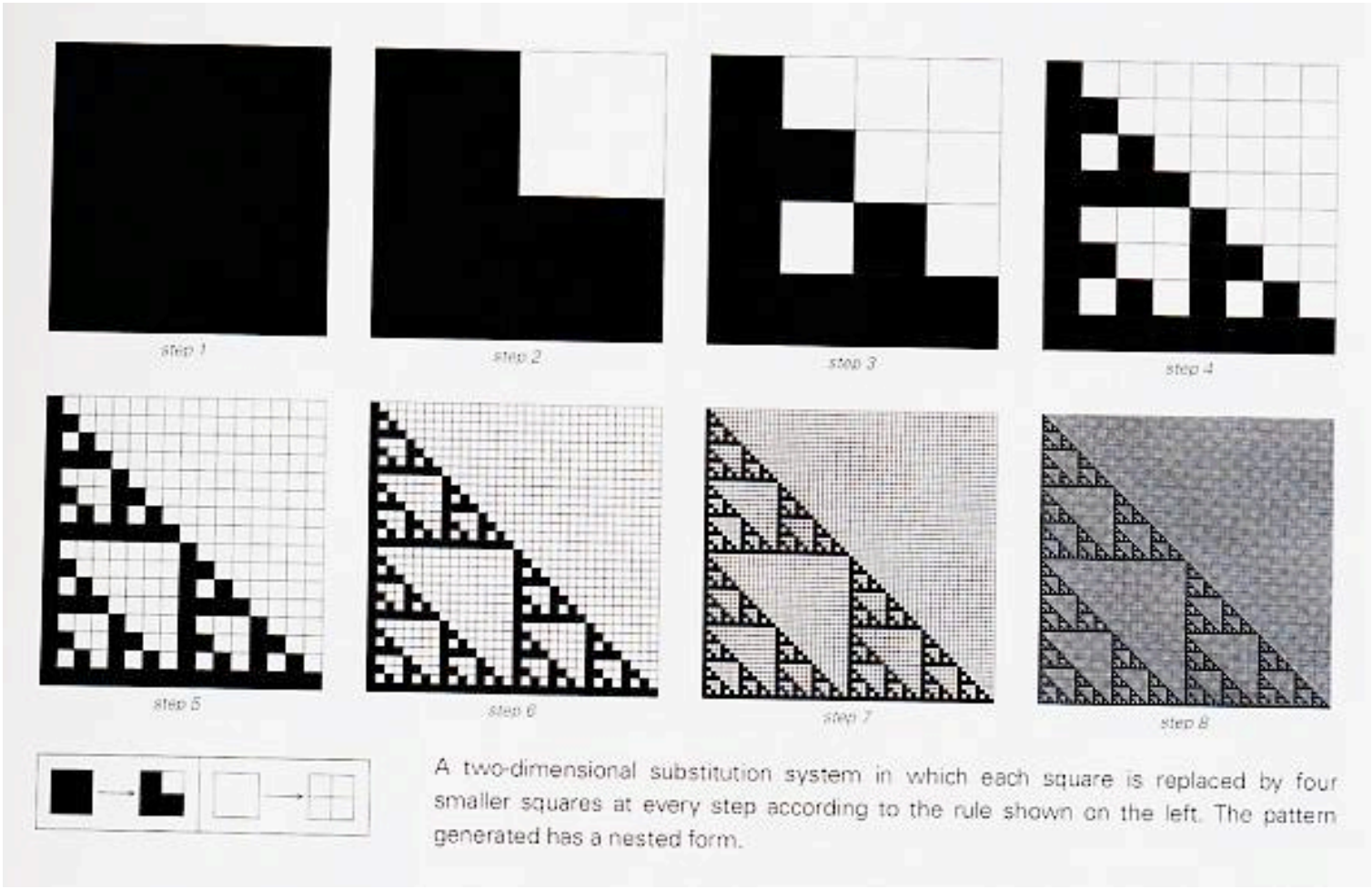
Scale-free Networks

- Very uneven distribution of connections. Some nodes have very high degrees of connectivity (hubs), while most have small degrees
- Scale-free means that the description of a system does not change as a function of the magnification (scale) used to view the system
  - fractals = self similar patterns with fractional dimensionality
  - Power law distribution of degrees: high connectivity is unlikely but occurs more often than predicted by random network

$$P(x) \quad x^{-a}$$

  - Power laws show up as straight lines when plotted on log-log coordinates, with the slope of the line = -a
  - Power laws are scale free because if x is rescaled (multiplied by a constant), then P(x) is still proportional to x^{-a}
    - if $P(x) = x^{-2}$, then P(10*x) is still proportional to x^{-2}. $P(x)= 10^{-2} * x^{-2}$

2-D substitution systems



A two-dimensional substitution system in which each square is replaced by four smaller squares at every step according to the rule shown on the left. The pattern generated has a nested form.

Cantor's Set

A = 1/3, N= 2, so D=log(2)/log(3)

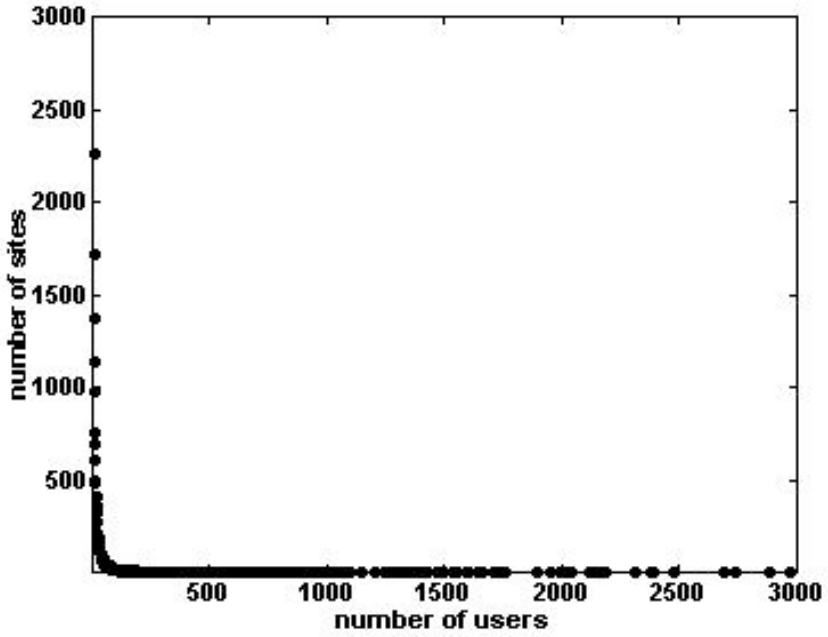
A = 1/9, N= 4, so D=log(4)/log(9)

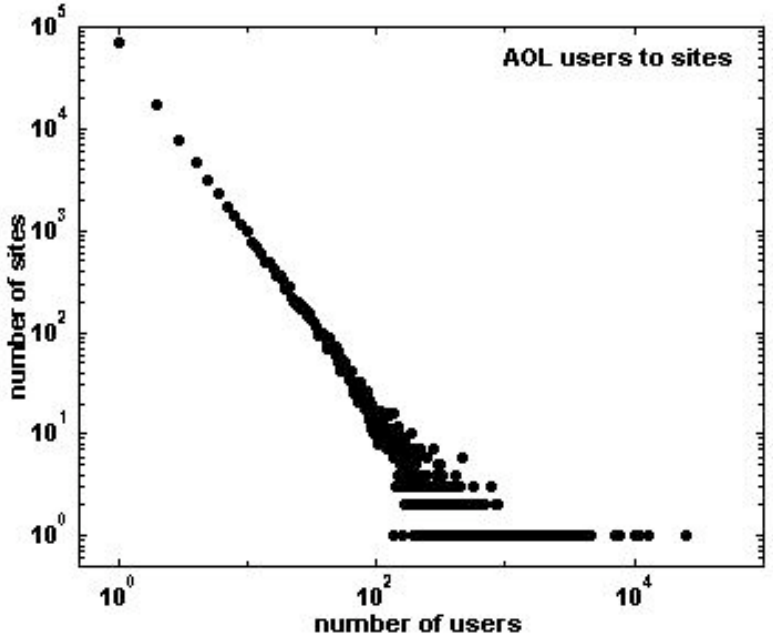
$$A = \left(\frac{1}{3^T}\right),\ N = 2^T, D = \frac{\log\left(2^T\right)}{\log\left(3^T\right)} = \frac{T\log(2)}{T\log(3)} \approx 0.6309$$

Dimensionality is between 0 and 1

Power laws

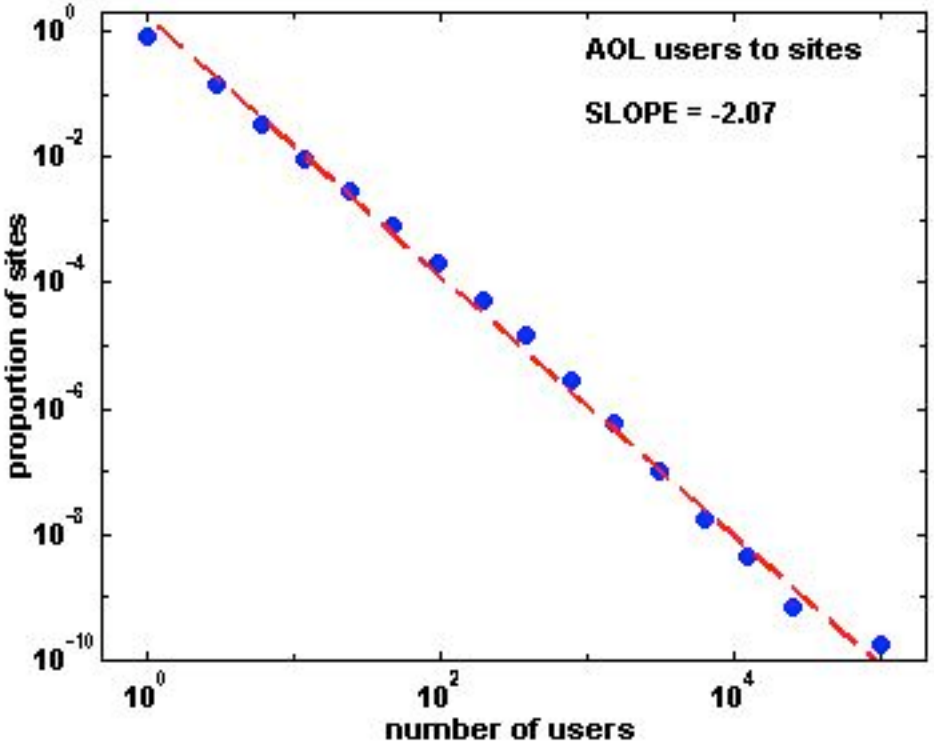
A mathematical relation that forms linear plots when data is transformed into log-log coordinates





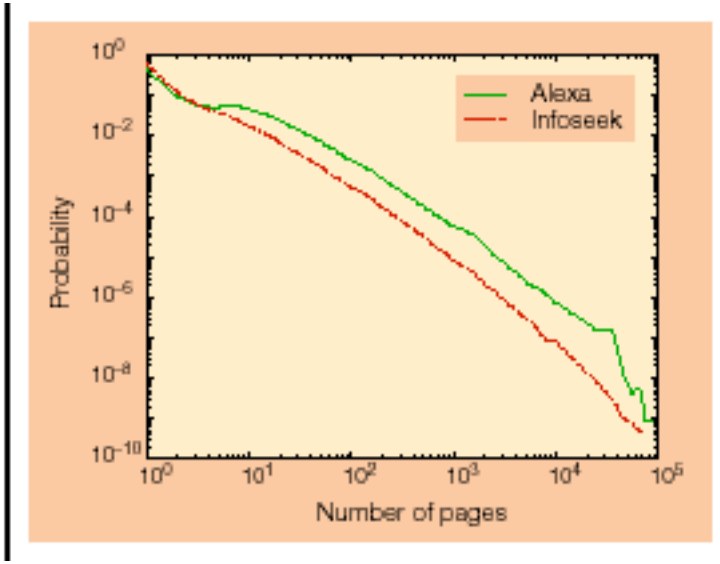
Very few sites have many users

Power laws



Most sites are visited by few users

P(site has X visitors) = $CX^{-2.07}$

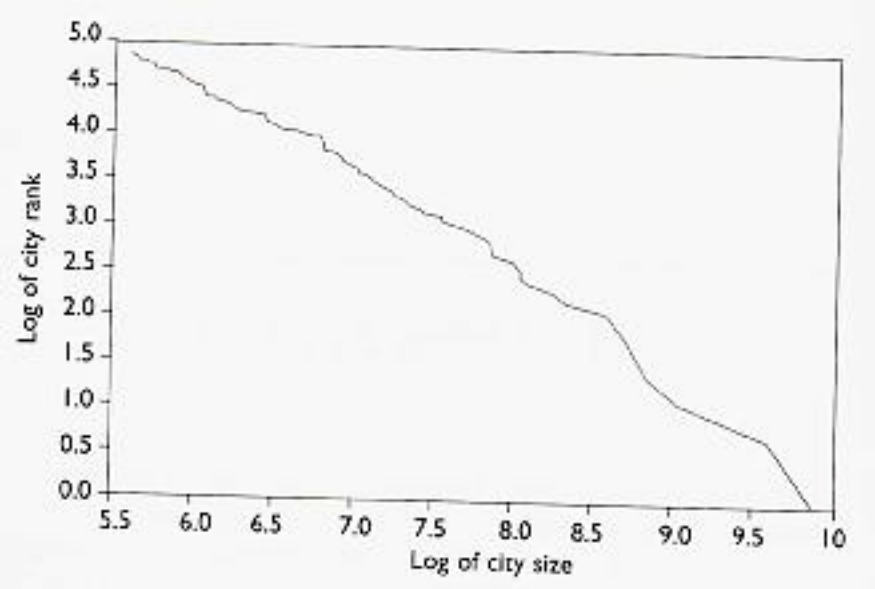


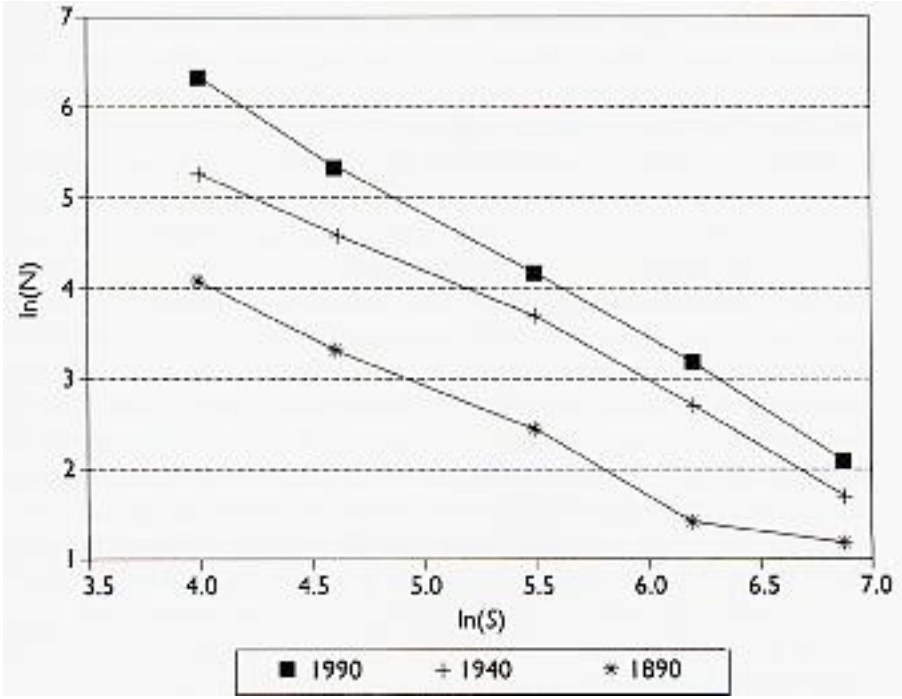
Huberman & Adamic (1999)

Most sites have few pages

P(site has X pages) = $CX^{-1.8}$

Large events are rare, small events common



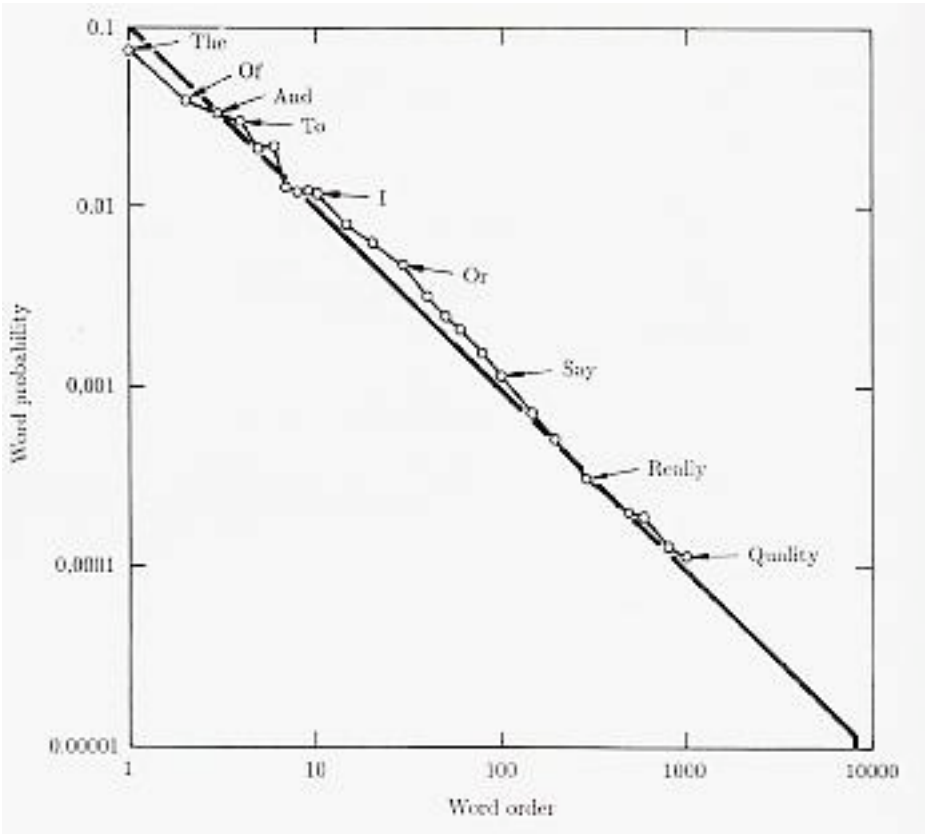


The population of a city is inversely proportional to its rank

$$\text{Log (rank of city)} = \text{Log (population)}^{-P}$$

$$P=1$$

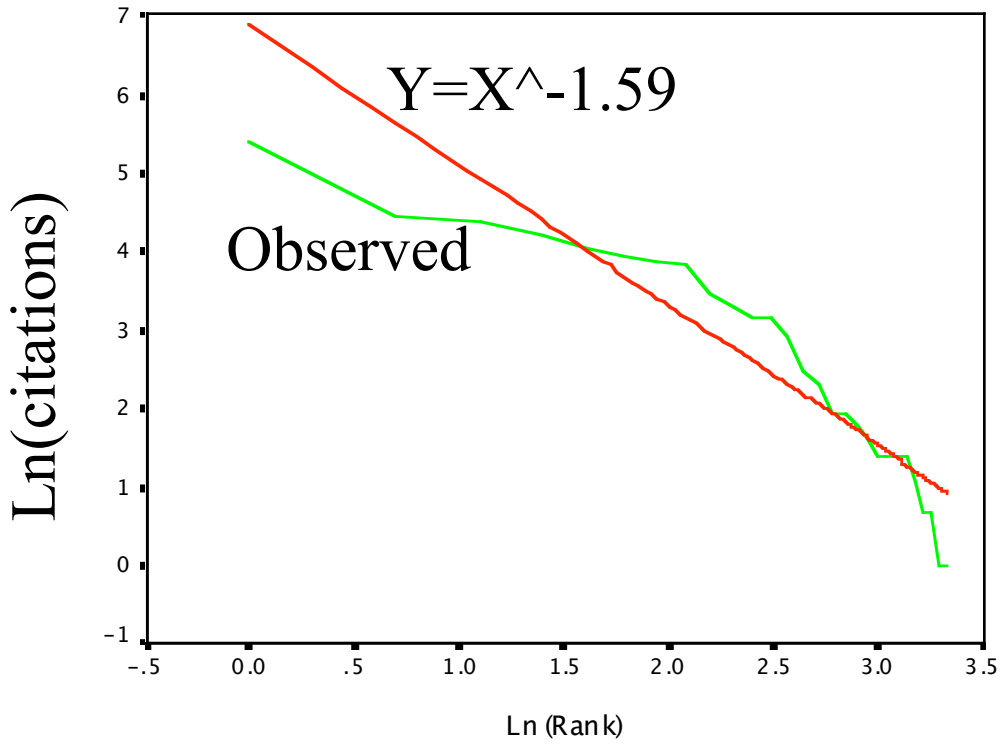
Zipf's Law



Log (probability of word) = Log (rank of word)$^{-P}$

P=1

Power Law in Bibliometrics

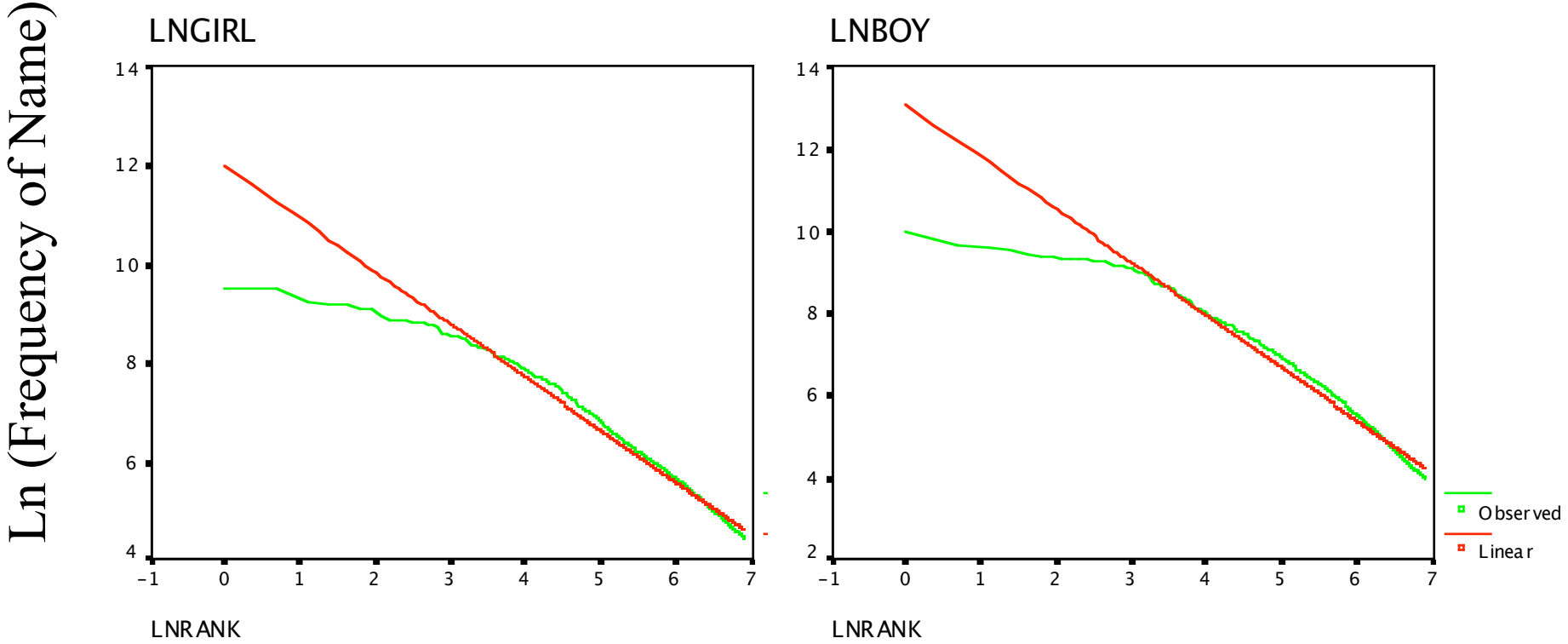


Ln (citations to Goldstone) = Ln (rank of citation)$^{-1.59}$

Power Law in Baby names
Social Security data, 1990s

Michael	21243	Ashley	14108
Christopher	16421	Jessica	14090
Matthew	15851	Emily	10345
Joshua	14973	Sarah	10109
Jacob	13086	Samantha	10096
Andrew	12281	Brittany	9016
Daniel	12178	Amanda	8982
Nicholas	12072	Elizabeth	7745
Tyler	11739	Taylor	7329
Joseph	11646	Megan	7266

Social Security data, 1990s

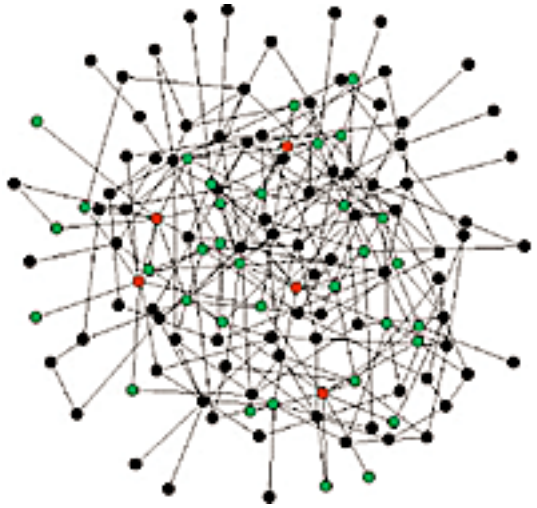


Apparent departure from power law, but only for a very small number of most common names. Ln(10) = 2.3, so only 10 data points out of 100 make up deviation

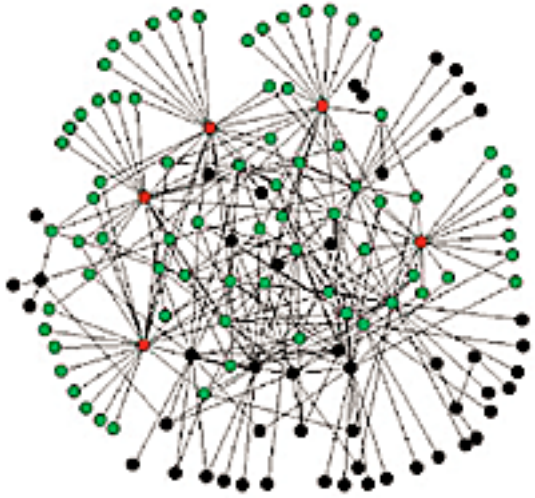
Boy exponent = -1.34, Girl exponent = -1.11

Naming for boys is more “elitist” than girls (faster drop-off of frequency with rank)

Scale-free Networks



Random Graph

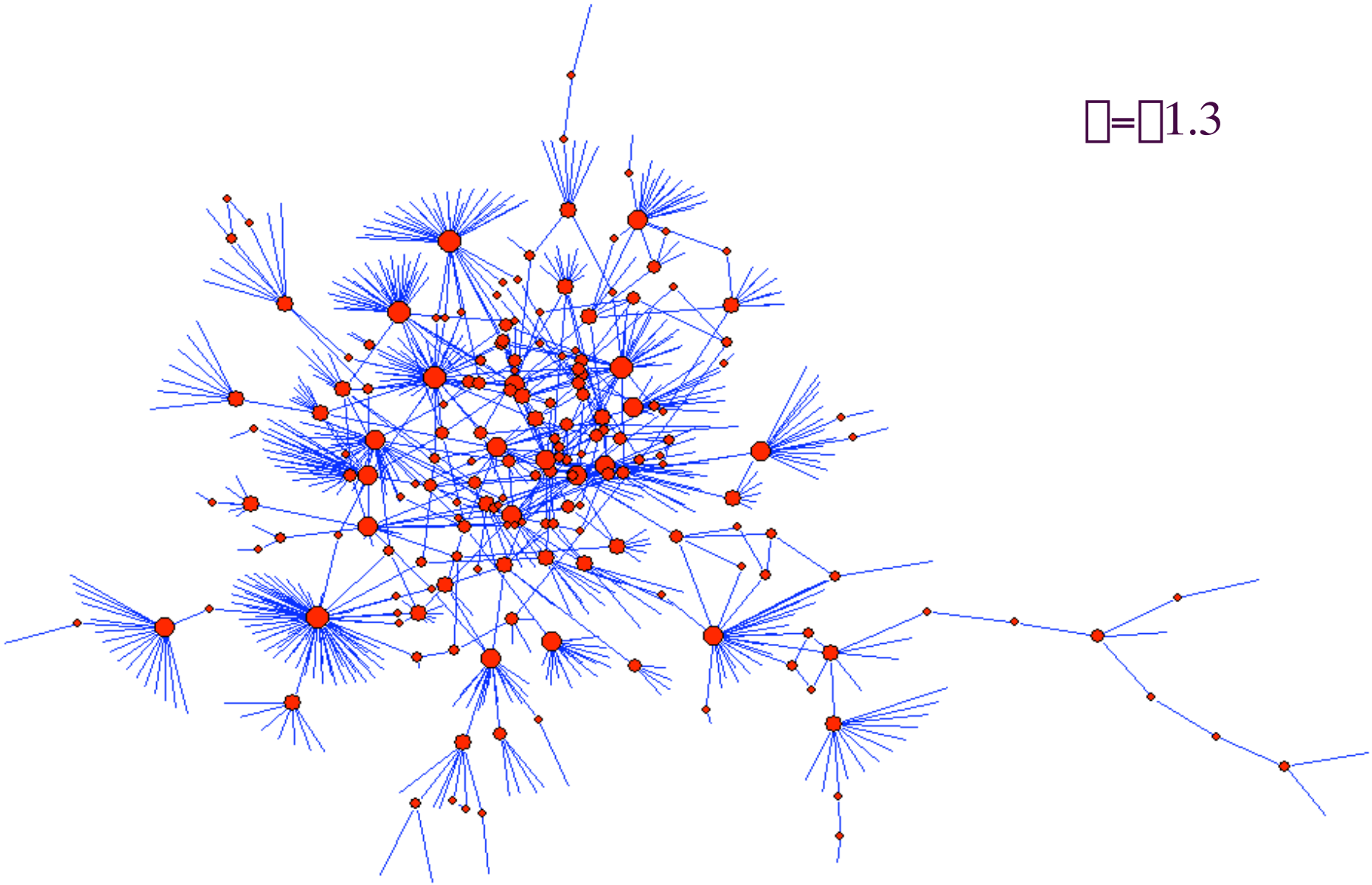


Scale-free Graph

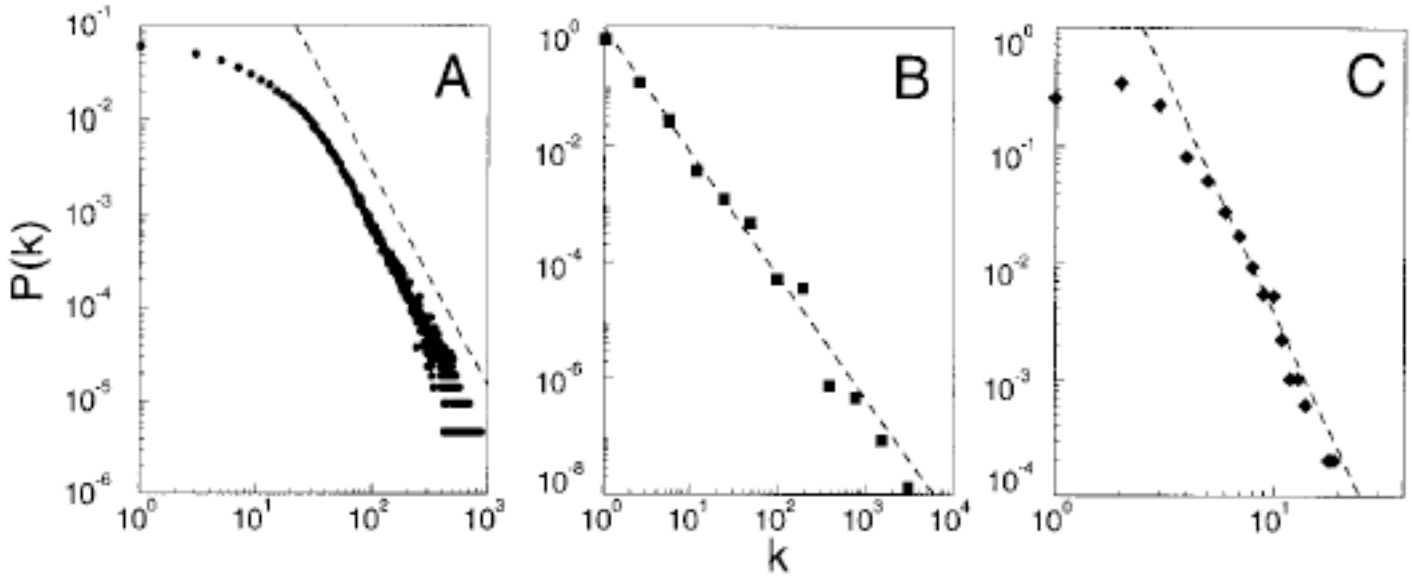
In the random graph, the 5 most connected nodes are connected to 27% of all nodes.

In the scale-free graph, the 5 most connected nodes are connected to 60% of all nodes.

Colorado Springs High-risk Sex Contacts



Barabasi & Albert, 1999



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

A = actor collaborations, N=212,250 edges, average connectivity <k> = 28.78, exponent $\gamma = 2.3$

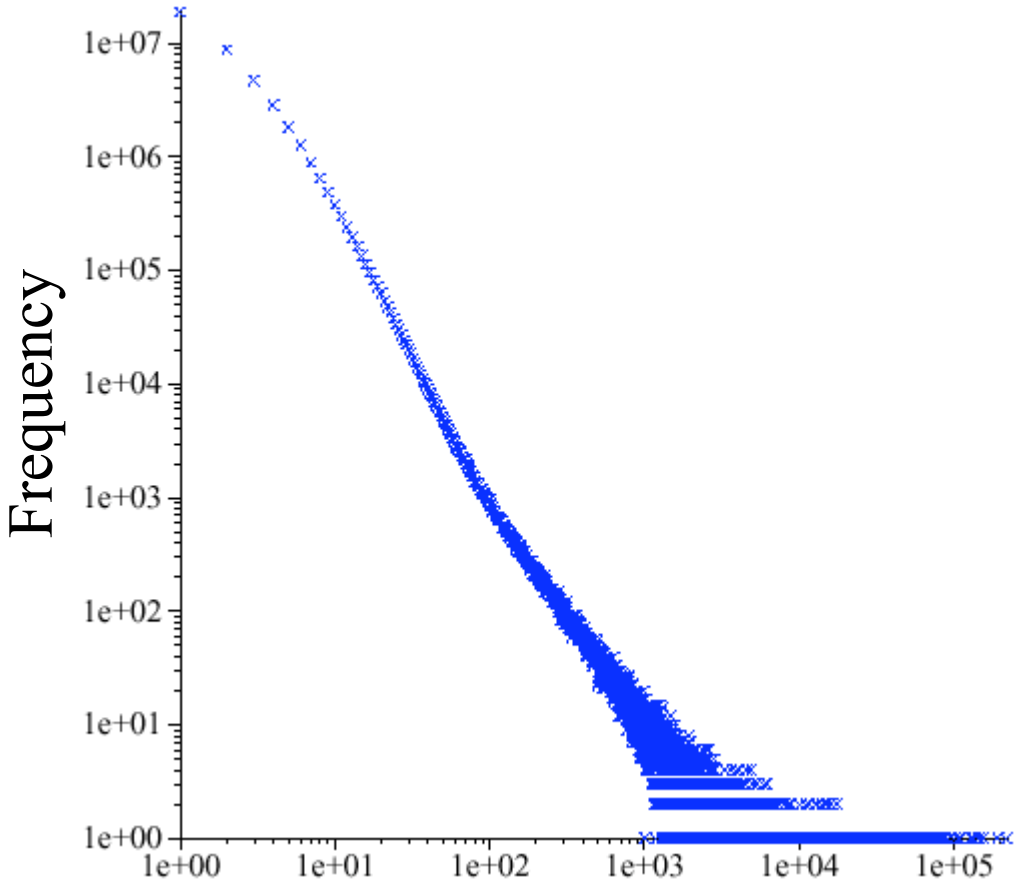
B = WWW, N = 325,729, <k> = 5.46, $\gamma = 2.3$

C = Power grid, N = 4941, <k> = 2.67, $\gamma = 4$

Redner (1998): probability that a paper is cited k times $\sim k^{-3}$

- Number of nodes having *k* links to other nodes (*degree k)* scales as a power law: $k^{-\gamma}$
- Exponent γ is in the interval 2-3 for most real networks.
- Example: The Bell Labs call graph: Calls made between 53 million phone numbers in a single day

•Aiello, W., F. Chung and L. Lu, 2000 Proc. 32nd ACM Symp. Theor. Comp.



Most numbers are called by only a few people

Where do scale free networks come from?

- Growth plus preferential attachment
  - Growth - networks do not start with all vertices established. Rather, networks accumulate vertices with time
  - Preferential attachment - a vertex that already has a large number of edges connecting it to other vertices will tend to attract still more edges. Rich get richer.
    - well known actors get more parts
    - well cited papers get more citations
- Formal model
  - Growth: start with m_0 vertices, and add new vertices one by one, each with m edges
  - Preferential attachment: probability that new vertex will connect to Vertex i is based on k_i the degree of I:

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

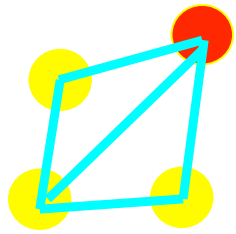
  - Predicts $\gamma = 3$
    - to generalize, use directed graphs, or edge deletion

Properties of scale free networks

- Robust to network failures (Albert, Jeong, & Barabasi, 2000)
  - Networks tend to stay connected, and average path length continues to be small, if random vertices are deleted
  - The probability of deleting a hub (vertex with high k) is small
- Vulnerable to targeted attacks
  - Targeted attacks specifically remove hubs
  - This is a positive property for negative networks
    - Decrease spread of AIDS by changing behavior of a small number of highly connected individuals
- Basic model does not predict high degrees of clustering
  - vertices connected to a vertex are often directly connected themselves. Clustering coeficient: $C_i = \frac{2E_i}{k_i(k_i-1)}$.

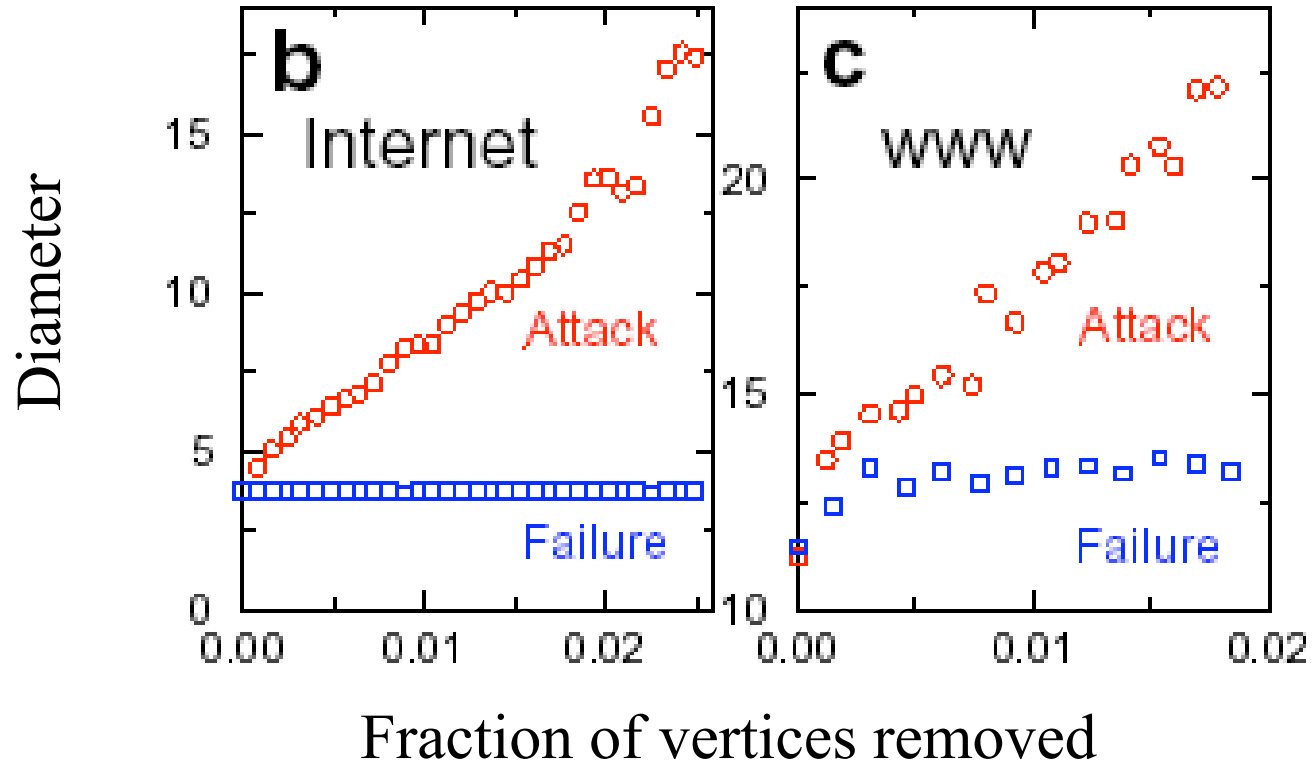
E_i = number of edges between i's neighbors

k_i = degree of vertex i



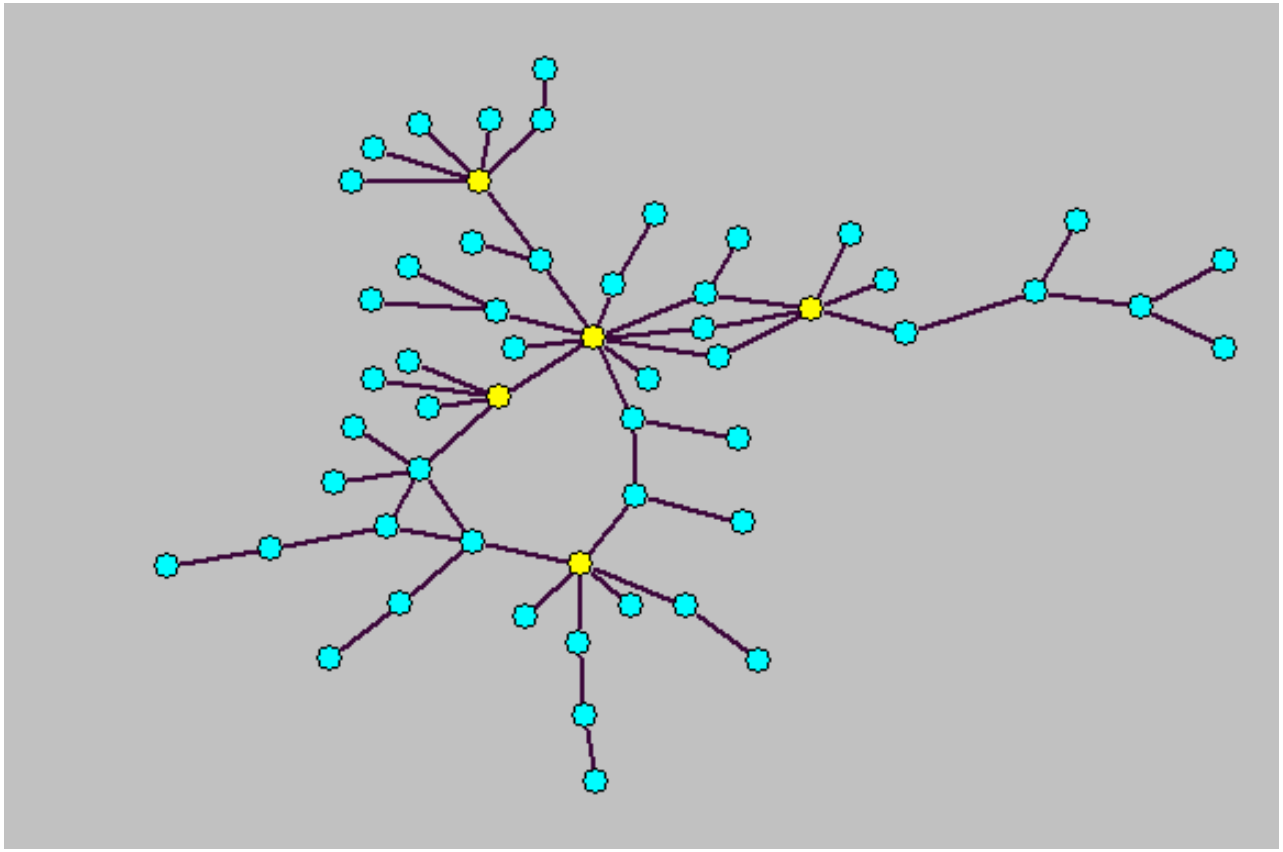
$C_{red}=2*2/3*2=.667$

Scale free networks are robust to failures but vulnerable to smart attacks

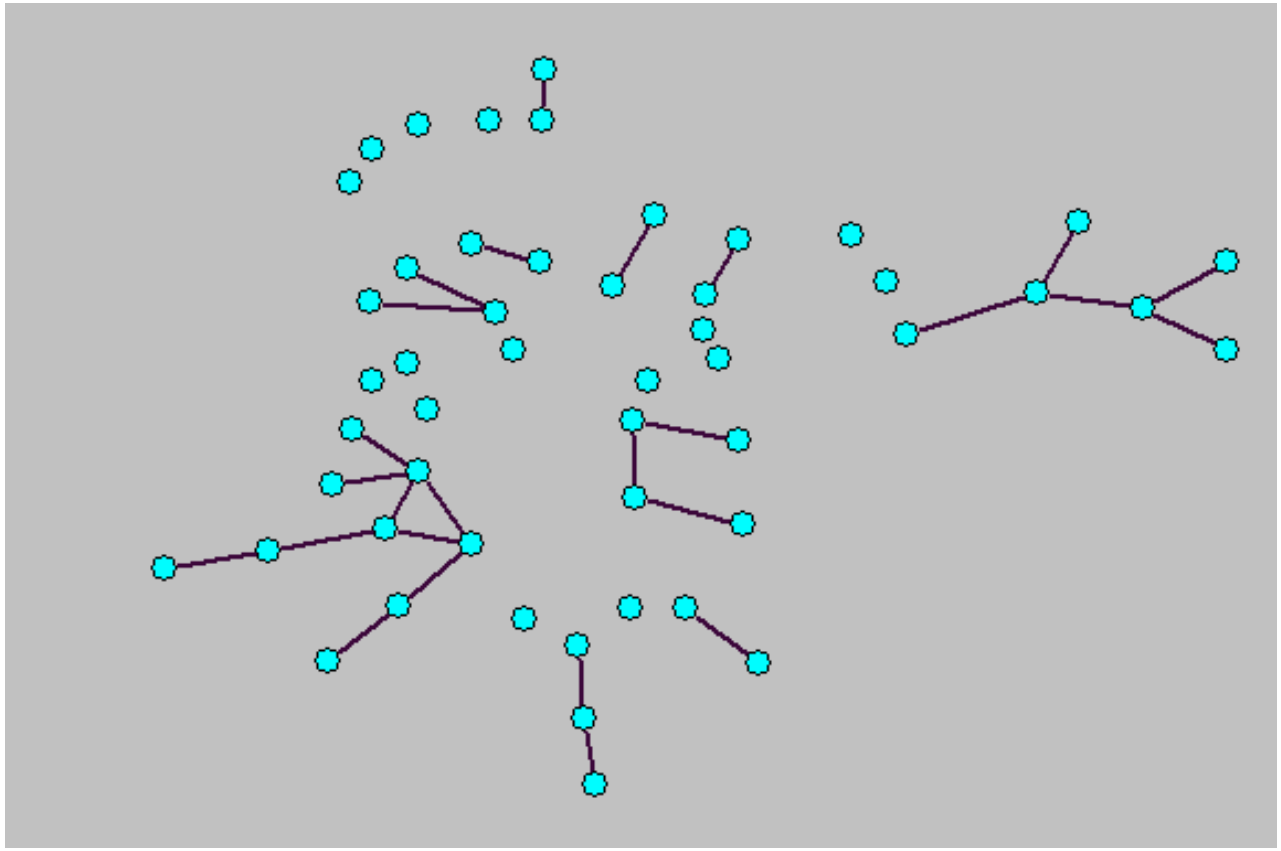


Diameter = Average path length connecting each of the N(N-1) directed connections between vertices

Hubs make the network fragile to node disruption



Hubs make the network fragile to node disruption



Degree, Clustering, & Path Length

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Albert & Barabasi (2002)

Power law exponents, degree, path length

TABLE II. The scaling exponents characterizing the degree distribution of several scale-free networks, for which $P(k)$ follows a power law (2). We indicate the size of the network, its average degree $\langle k \rangle$, and the cutoff κ for the power-law scaling. For directed networks we list separately the indegree (γ_{in}) and outdegree (γ_{out}) exponents, while for the undirected networks, marked with an asterisk (*), these values are identical. The columns l_{real}, l_{rand}, and l_{pow} compare the average path lengths of real networks with power-law degree distribution and the predictions of random-graph theory (17) and of Newman, Strogatz, and Watts (2001) [also see Eq. (63) above], as discussed in Sec. V. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

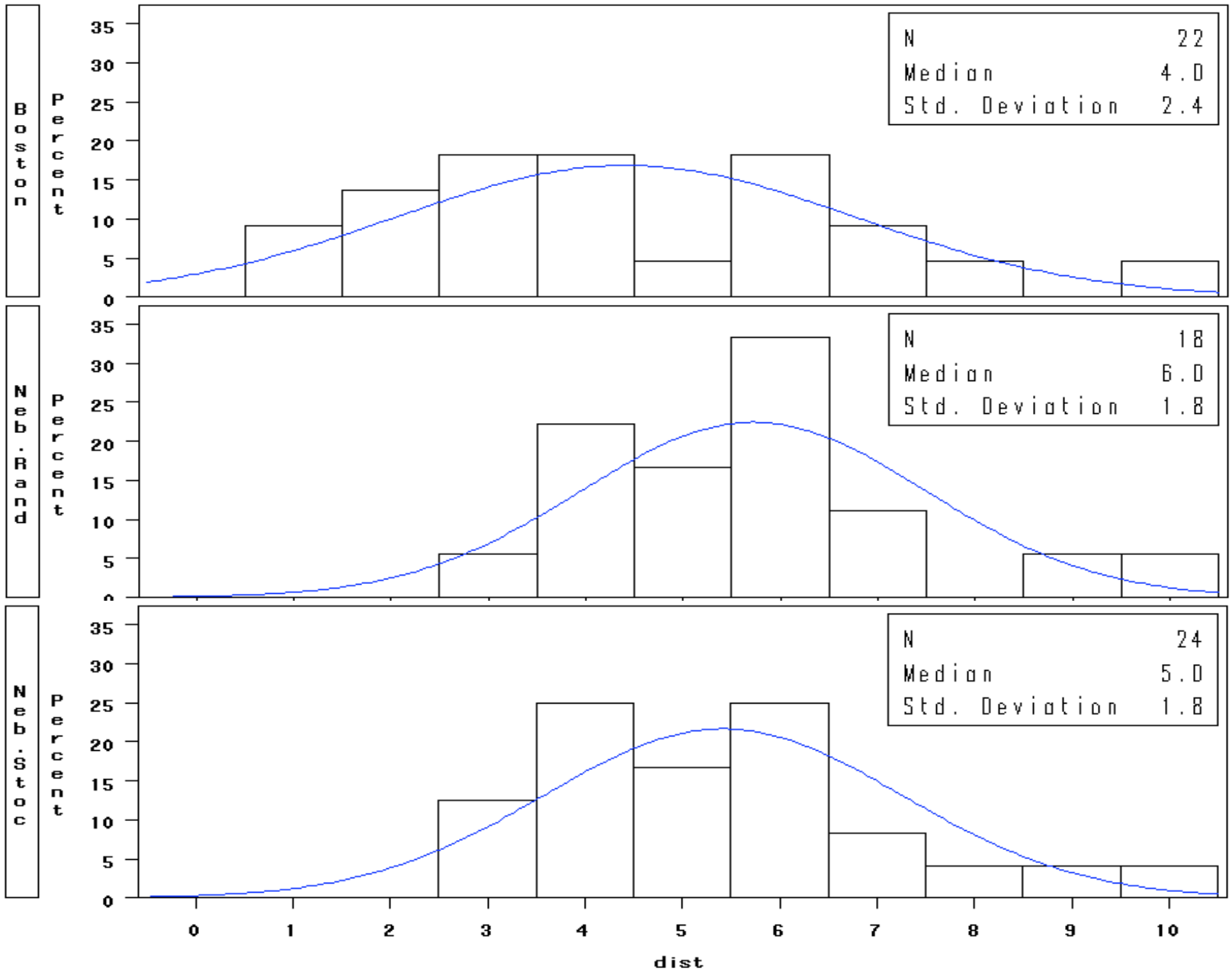
Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}	ℓ_{real}	ℓ_{rand}	ℓ_{pow}	Reference	Nr.
WWW	325 729	4.51	900	2.45	2.1	11.2	8.32	4.77	Albert, Jeong, and Barabási 1999	1
WWW	4×10^7	7		2.38	2.1				Kumar *et al.*, 1999	2
WWW	2×10^8	7.5	4000	2.72	2.1	16	8.85	7.61	Broder *et al.*, 2000	3
WWW, site	260 000				1.94				Huberman and Adamic, 2000	4
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2	4	6.3	5.2	Faloutsos, 1999	5
Internet, router*	3888	2.57	30	2.48	2.48	12.15	8.75	7.67	Faloutsos, 1999	6
Internet, router*	150 000	2.66	60	2.4	2.4	11	12.8	7.47	Govindan, 2000	7
Movie actors*	212 250	28.78	900	2.3	2.3	4.54	3.65	4.01	Barabási and Albert, 1999	8
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2	4	2.12	1.95	Newman, 2001b	9
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1	6	5.01	3.86	Barabási *et al.*, 2001	10
Co-authors, math.*	70 975	3.9	120	2.5	2.5	9.5	8.2	6.53	Barabási *et al.*, 2001	11
Sexual contacts*	2810			3.4	3.4				Liljeros *et al.*, 2001	12
Metabolic, *E. coli*	778	7.4	110	2.2	2.2	3.2	3.32	2.89	Jeong *et al.*, 2000	13
Protein, *S. cerev.**	1870	2.39		2.4	2.4				Jeong, Mason, *et al.*, 2001	14
Ythan estuary*	134	8.7	35	1.05	1.05	2.43	2.26	1.71	Montoya and Solé, 2000	14
Silwood Park*	154	4.75	27	1.13	1.13	3.4	3.23	2	Montoya and Solé, 2000	16
Citation	783 339	8.57			3				Redner, 1998	17
Phone call	53×10^6	3.16		2.1	2.1				Aiello *et al.*, 2000	18
Words, co-occurrence*	460 902	70.13		2.7	2.7				Ferrer i Cancho and Solé, 2001	19
Words, synonyms*	22 311	13.48		2.8	2.8				Yook *et al.*, 2001b	20

(γ=2 is equivalent to Zipf's law with exponent of 1)

Small World Networks

- Elements of a system are frequently connected to each other via a short path
  - Milgram's lost letter experiments (from Omaha to Boston stock broker)
  - Six-degrees of separation to Kevin Bacon and Paul Erdos
- Regular Lattices
  - Vertices only connected to their neighbors (ring-worlds)
  - "Large world" networks because one must go through many intermediaries to form some paths. Path length is proportional to n/2k (n=# vertices, k = degree)
- Random graphs
  - Vertices connected randomly
  - Average path length is short, proportional to ln(n)/ln(k)
  - Unfortunately network has no clusters/cliques
- Small world networks (Watts & Strogatz, 1998)
  - Intermediates between regular and random graphs
  - degree has a poisson distribution (like random graphs)

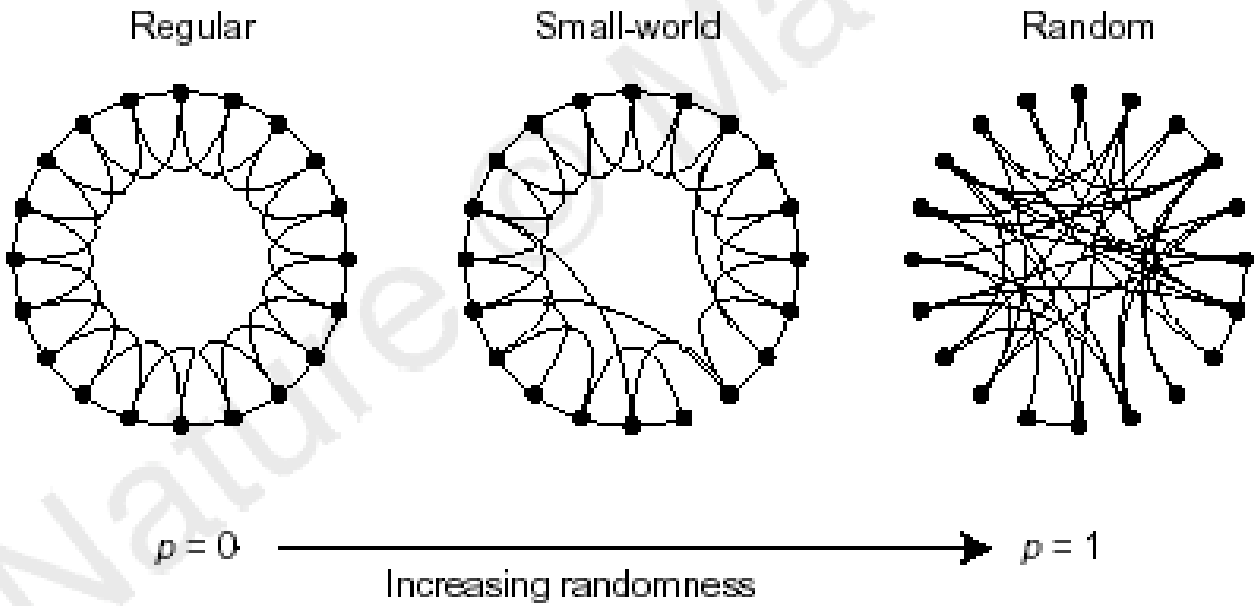
Milgram, 1967



The Oracle of Bacon
http://www.cs.virginia.edu/oracle/

Bacon Number	# of People
0	1
1	1667
2	129780
3	344712
4	82578
5	6635
6	782
7	121
8	23
9	1
10	1

Small World Networks



Constructing a small world network (Watts, 1999)

Start with regular graph

Rewire each edge with probability p

Small World Networks

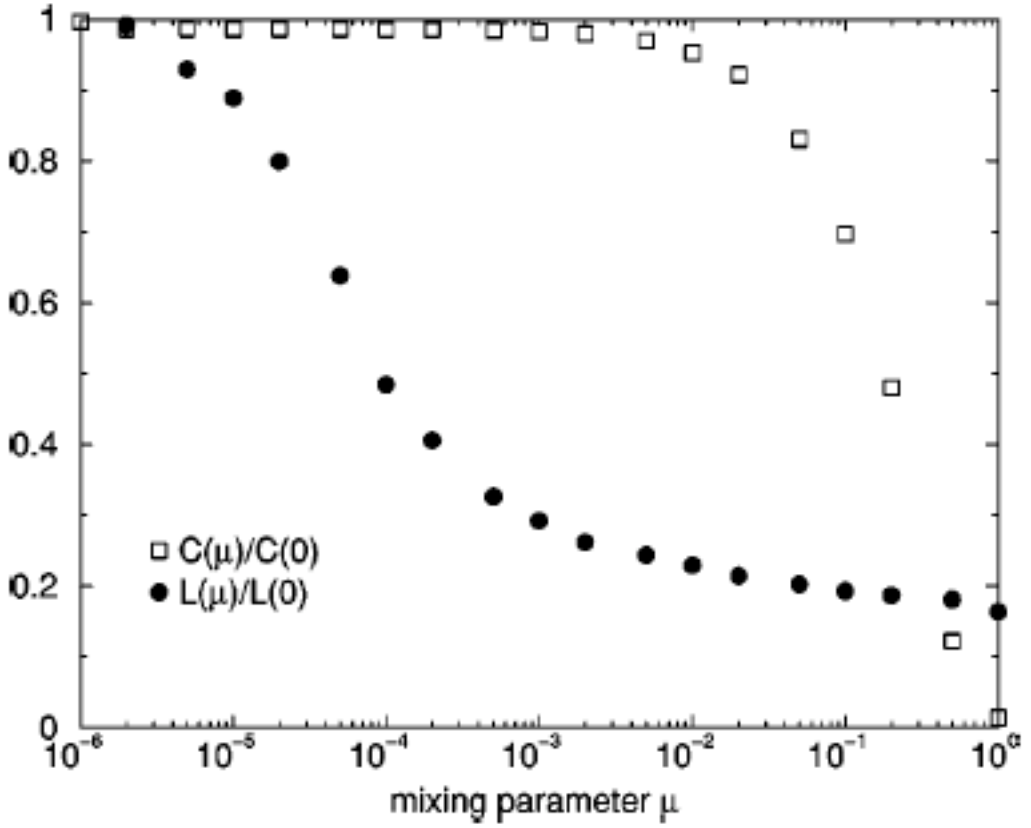


FIG. 1. Small-world effect in scale-free networks. Introducing the ratio $\mu \ll 1$ of random links into the highly clustered scale-free networks drastically reduces the typical distance L between nodes. However, the strongly interconnected neighborhoods of the original model ($\mu = 0$) are preserved, as the clustering coefficient remains at its large value. Only when μ reaches the order of 1 does the clustering coefficient drop significantly. All plotted values are averages over 100 independent realizations. The networks have $N = 10^4$ nodes with average degree $\langle k \rangle = 20$.

As random rewirings increase, clustering coefficient (C) and average path length (L) both decrease

But, for a large range of rewiring probabilities, it is possible to have short path lengths but still clusters

Divide by C(O) and L(O) because normalize according to regular graph

Small world networks frequently occur (Watts & Strogatz, 1998)

Table 1 Empirical examples of small-world networks

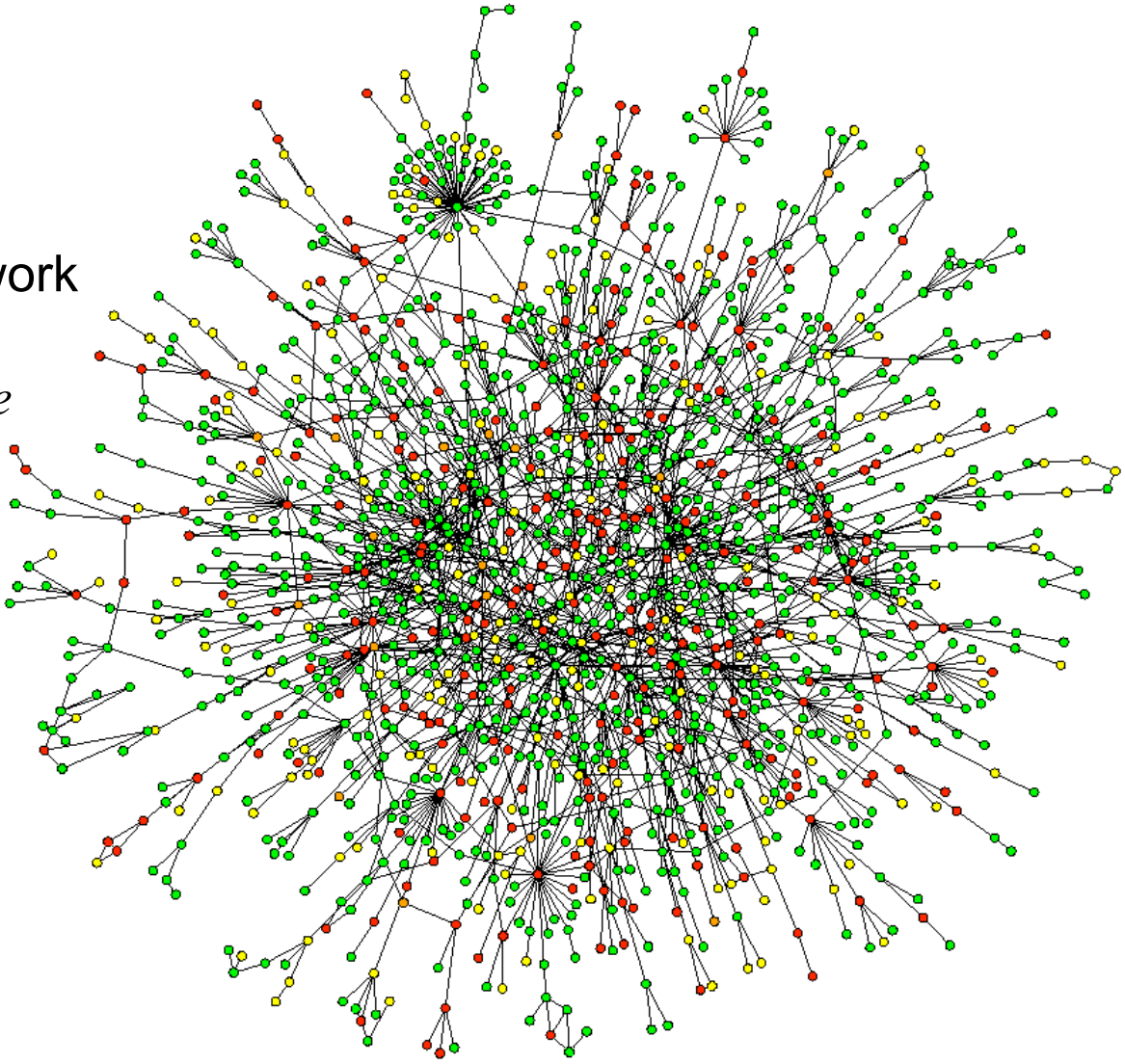
	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Characteristic path length L and clustering coefficient C for three real networks, compared to random graphs with the same number of vertices (n) and average number of edges per vertex (k). (Actors: $n = 225{,}226$, $k = 61$. Power grid: $n = 4{,}941$, $k = 2.67$. *C. elegans*: $n = 282$, $k = 14$.) The graphs are defined as follows. Two actors are joined by an edge if they have acted in a film together. We restrict attention to the giant connected component[16] of this graph, which includes ~90% of all actors listed in the Internet Movie Database (available at http://us.imdb.com), as of April 1997. For the power grid, vertices represent generators, transformers and substations, and edges represent high-voltage transmission lines between them. For *C. elegans*, an edge joins two neurons if they are connected by either a synapse or a gap junction. We treat all edges as undirected and unweighted, and all vertices as identical, recognizing that these are crude approximations. All three networks show the small-world phenomenon: $L \gtrsim L_{random}$ but $C \gg C_{random}$.

Disease infection: If only a few long-range connections, diseases spread very quickly (path length is short)

Biochemical Networks

Protein Interaction Network of common Yeast cell *Saccharomyces Cerevisiae* Jeong, *et al.*, 2001, Nature 411, 41.



Al-Quaeda

- Links between 9/11 hijackers and known associates.
- (Courtesy of Valdis Krebs, *Uncloaking Terrorist Networks*, First Monday **7**, no 4, April 2002)

Open issues in social networks

- Integrating properties of scale free and small world networks
  - Clusters
  - Hubs
- Ways of characterizing elements in a network