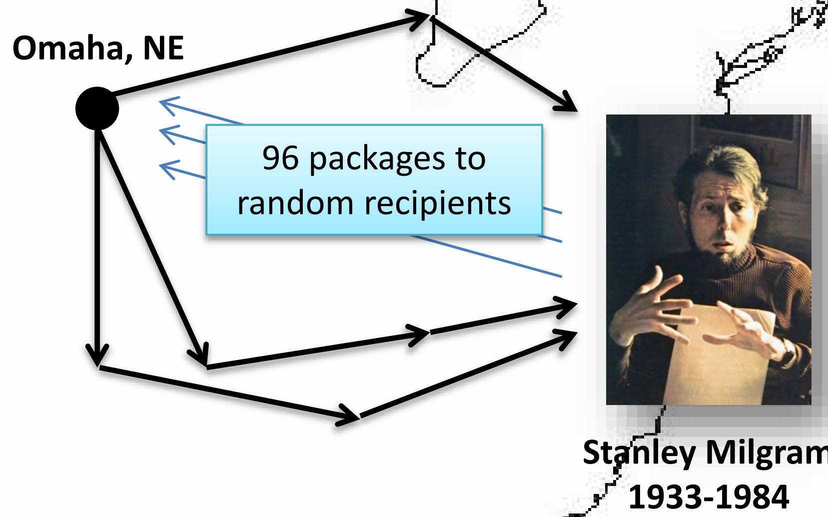
Small-world Phenomena

- *Name, address, occupation* of the target were known; no sending was allowed
- 18 packages returned back to Boston
- mean path result was just 5.9 steps
- small-world effect was confirmed in many other experiments

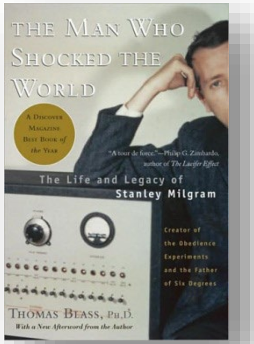


Bonus observations in the experiment

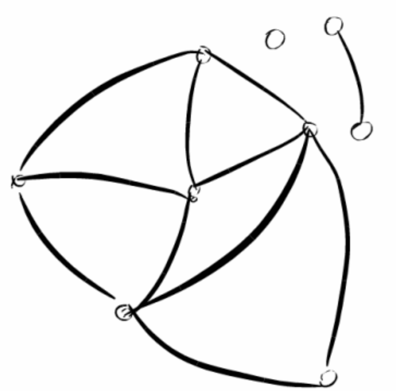
- most of the packages were received through 3 target's friends
- people are good in finding short paths (later was shown that it is hard to find shortest path without knowing full information)

Similar experiments

- emails: only 384 out of 24K were received/ results confirmed, 4 steps
- Microsoft .NET Messenger Service: 6.6 people



Degree Distributions



$$p_0 = \frac{1}{9},\ p_1 = \frac{2}{9},\ p_2 = \frac{1}{9},\ p_3 = \frac{2}{9},\ p_4 = \frac{3}{9}$$

The probability that randomly chosen node has degree *k*

Each node is connected independently with probability *p* to *n-1* nodes

Classical undirected random graph models $G_{n,p}$

choose *k* neigh among *n-1*

probability of being connected to exactly k neighbors

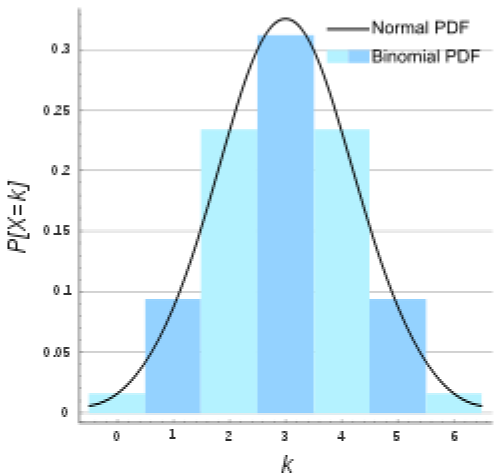
$$\binom{n-1}{k} p^k (1-p)^{n-1-k} \quad \text{(Binomial distribution)}$$

The probability of being connected to exactly *k* nodes

when graphs are small

when graphs are large (*n* is assumed to be large, mean degree is approximately constant as the network grows)

$$\frac{(np)^k}{k!} e^{-np} \quad \text{(Poisson distribution)}$$



What if the network is large?

Each node is connected independently with probability *p* to *n-1* nodes

Classical undirected random graph models $G_{n,p}$

choose *k* neigh among *n-1*

probability of being connected to exactly k neighbors

$$\binom{n-1}{k} p^k (1-p)^{n-1-k} \quad \text{(Binomial distribution)}$$

p_k is the probability of being connected to exactly *k* nodes

when graphs are small

When **graphs are large** then *n* is assumed to be large, and **mean degree is approximately constant *c*** as the network grows. For example, the number of your friends does not grow with the population in the world.
Let *p = c/(n-1)* then we can write

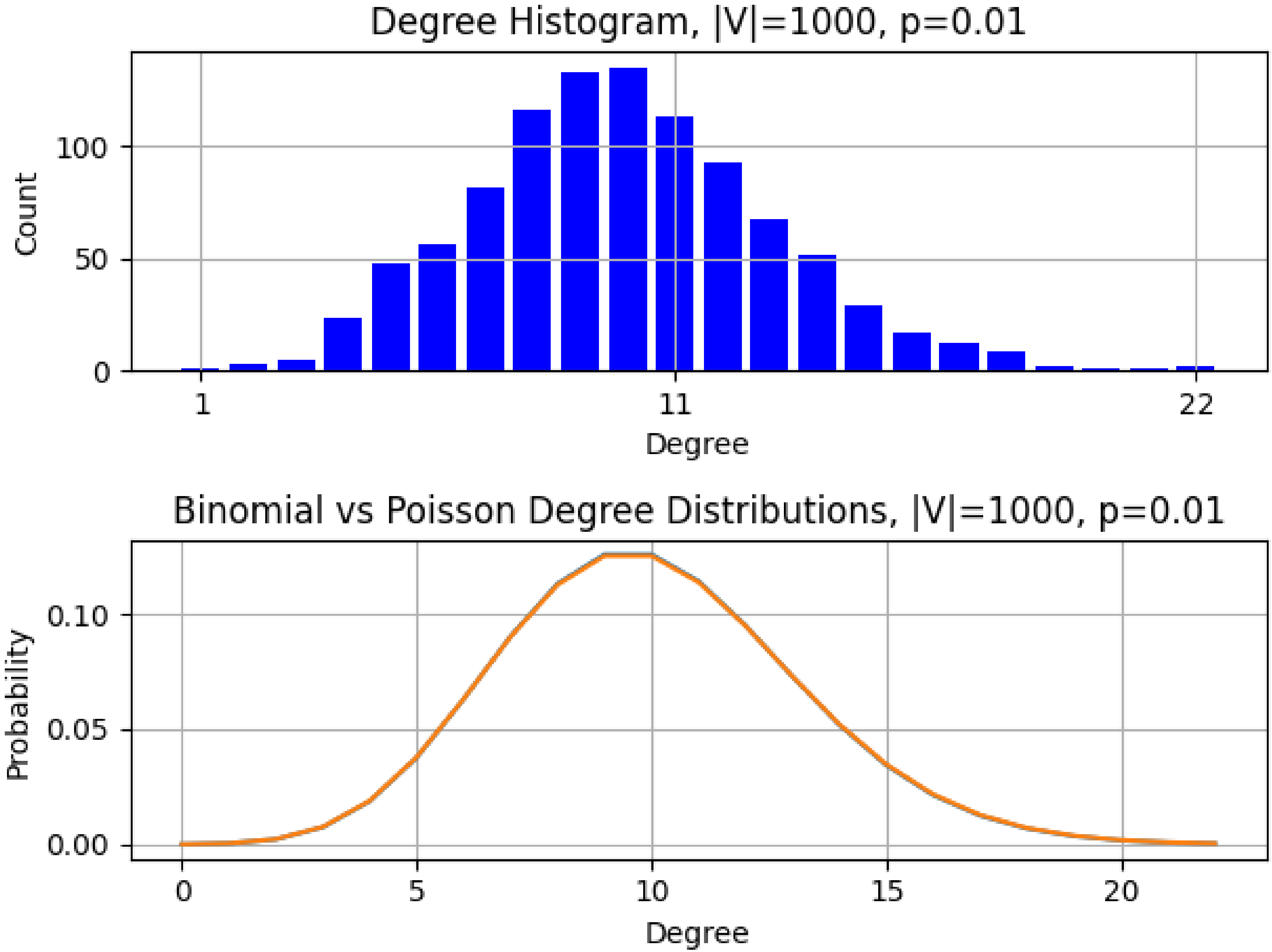
$$\ln(1-p)^{n-1-k} = (n-1-k)\ln(1-p) = (n-1-k)\ln(1-\tfrac{c}{n-1}) \approx -(n-1-k)\tfrac{c}{n-1} \approx -c \;\Rightarrow\; \text{Taking exponents of both sides } (1-p)^{n-1-k} = e^{-c}$$

Also

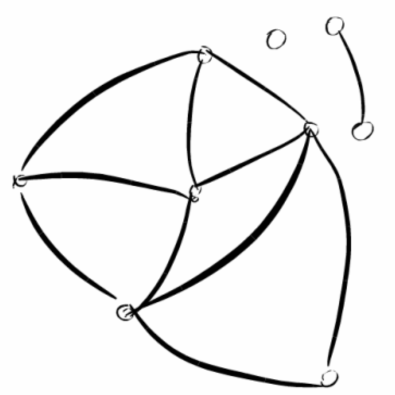
$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!k!} \approx \frac{(n-1)^k}{k!}, \text{ so}$$

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

Poisson distribution


Degree Histogram, |V|=1000, p=0.01
Count
100
50
0
1
11
22
Degree
Binomial vs Poisson Degree Distributions, |V|=1000, p=0.01
Probability
0.10
0.05
0.00
0
5
10
15
20
Degree

Degree Distributions

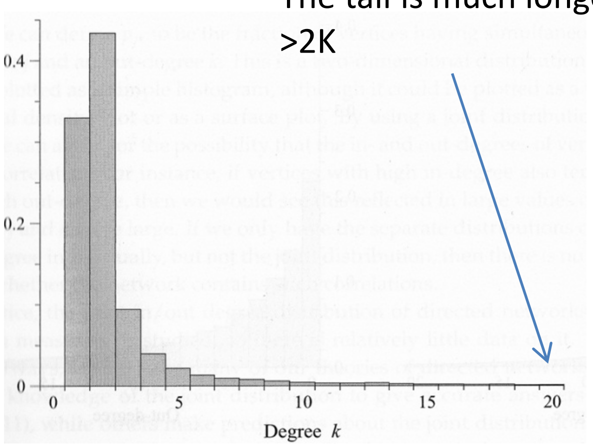


$$p_0 = \frac{1}{9},\ p_1 = \frac{2}{9},\ p_2 = \frac{1}{9},\ p_3 = \frac{2}{9},\ p_4 = \frac{3}{9}$$

The probability that a randomly chosen node has degree *k*

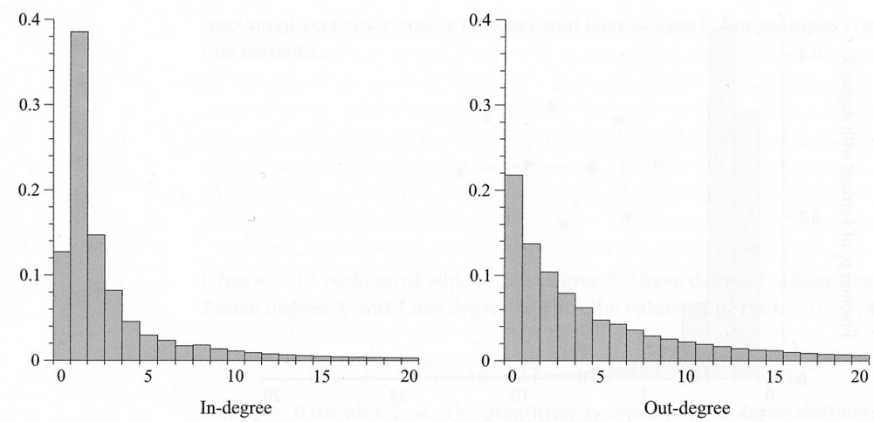
The tail is much longer >2K

Fraction p_k of vertices with degree k



Internet at the level of autonomous systems

Fraction of vertices

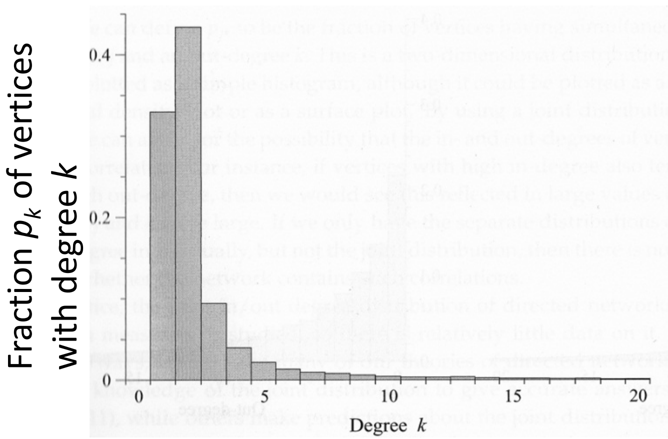


World Wide Web

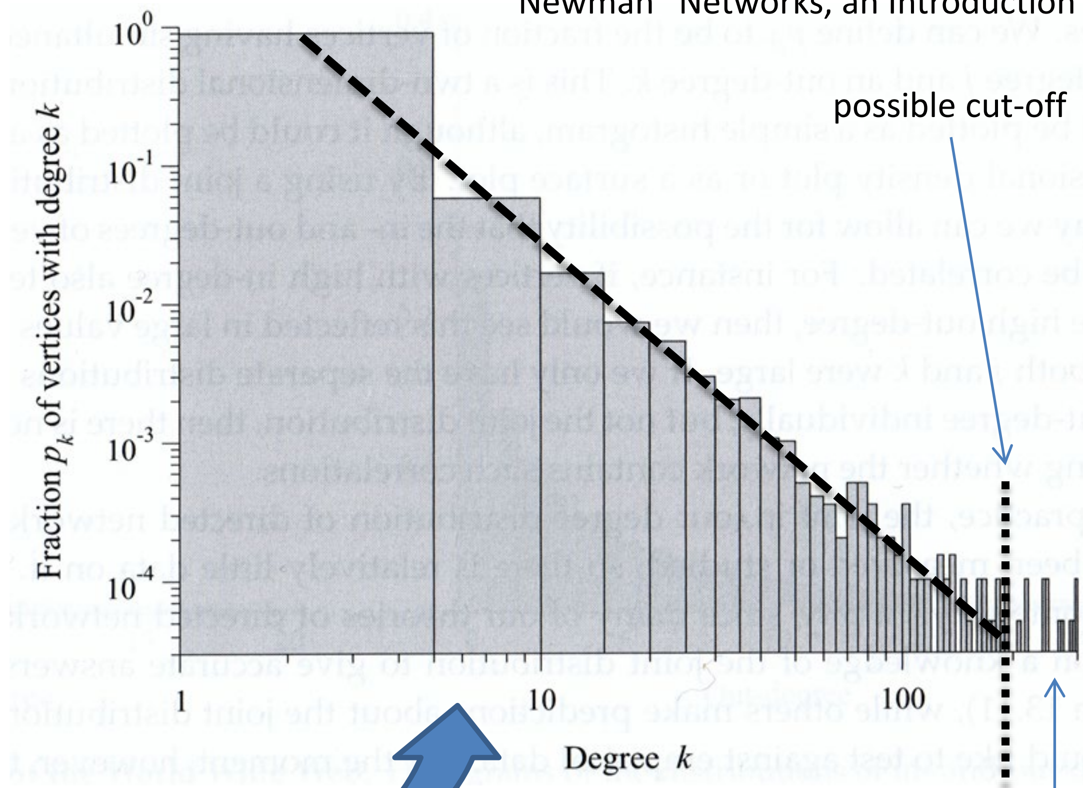
Newman "Networks, an Introduction"

Power Laws (aka scale-free)

Newman "Networks, an Introduction"



Internet at the level of autonomous systems



logarithmic scales; bigger range of bins

$$\ln p_k = -\alpha \ln k + c \text{ or } p_k = Ck^{-\alpha}, \text{ where } C = e^c$$

$$\text{typical } \alpha \in [2,3] \text{ (see handout Table 8.1)}$$

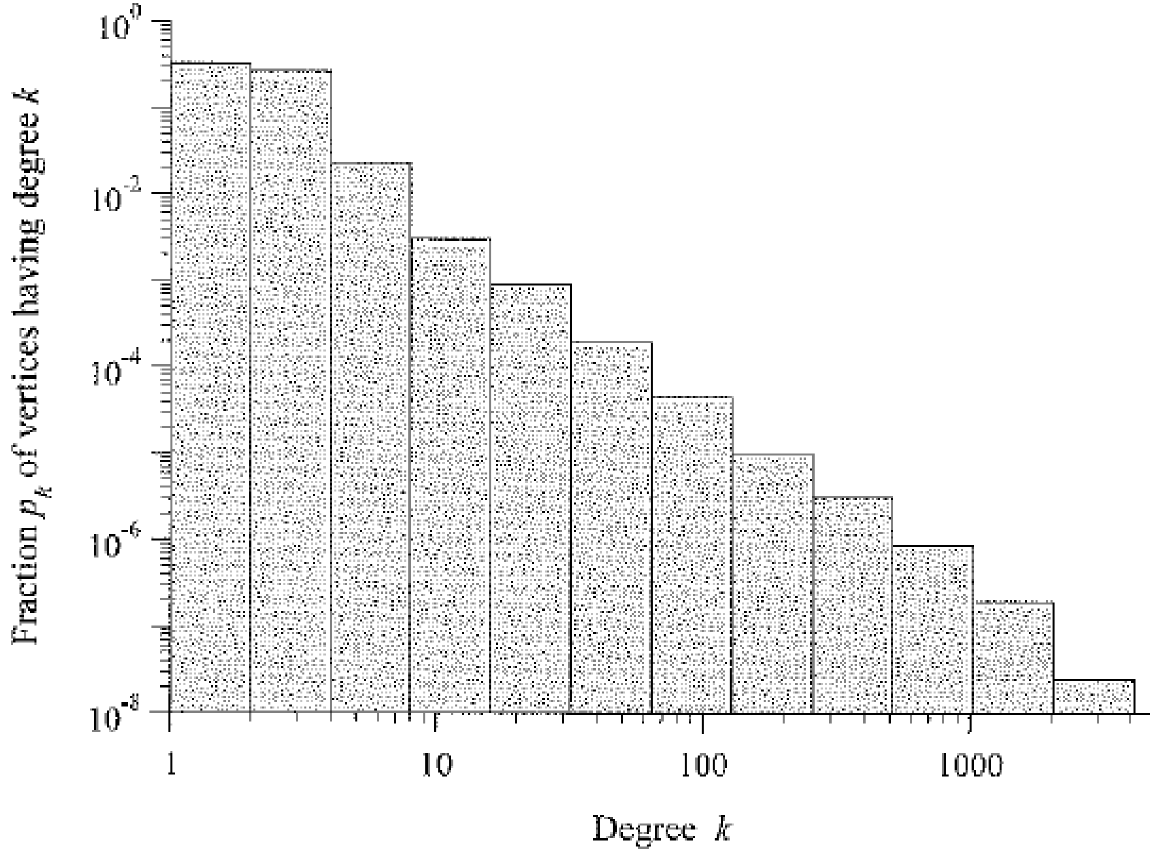
Problem of histograms: statistics is poor at the tail of the distribution

Solution I: different sizes of bins

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1			
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16	
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029
	Sexual contacts	Undirected	2 810					3.2			
Information	WWW `nd.edu`	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7			
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–			
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44	
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.003
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226

Table 8.1: Basic statistics for a number of networks. The properties measured are: type of network, directed or undirected; total number of vertices n; total number of edges m; mean degree c; fraction of vertices in the largest component S (or the largest weakly connected component in the case of a directed network); mean geodesic distance between connected vertex pairs ℓ; exponent α of the degree distribution if the distribution follows a power law (or "-" if not; in/out-degree exponents are given for directed graphs); clustering coefficient C from Eq. (7.41); clustering coefficient C_{WS} from the alternative definition of Eq. (7.44); and the degree correlation coefficient r from Eq. (7.82). The last column gives the citation(s) for each network in the bibliography. Blank entries indicate unavailable data.

Power Laws: Logarithmic Binning



- Bin 1 covers degrees in [1,2)
- Bin 2 covers degrees in [2, 4)
- Bin 3 covers degrees in [4, 8)
- …

Width of bins can vary

Figure 8.6: Histogram of the degree distribution if the Internet, created using logarithmic binning. In this histogram the widths of the bins are constant on a logarithmic scale, meaning that on a linear scale each bin is wider by a constant factor than the one to its left. The counts in the bins are normalized by dividing by bin width to make counts in different bins comparable.

Cumulative Distribution

Probability at a random vertex has degree k or greater

$P_k = \sum_{k'=k}^{\infty} p_{k'}$

Let p_k follows a power law in its tail, i.e., $p_k = Ck^{-\alpha}$ for $k \geq k_{\min}$. Then

$$P_k = C \sum_{k'=k}^{\infty} k'^{-\alpha}$$
$$\approx C \int_k^{\infty} k'^{-\alpha}\, \mathrm{d}k' = \frac{C}{\alpha - 1} k^{-(\alpha-1)}$$

$$\alpha = 1 + N \left(\sum_i \ln \frac{d(i)}{k_{\min} - 1/2} \right)^{-1}$$

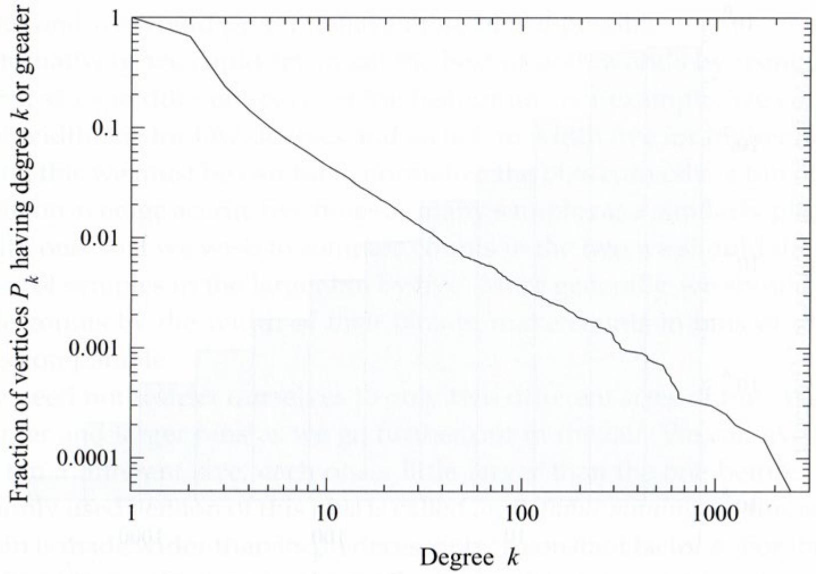


Figure 8.7: Cumulative distribution function for the degrees of vertices on the Internet. For a distribution with a power-law tail, as is approximately the case for the degree distribution of the Internet, the cumulative distribution function, Eq. (8.4), also follows a power law, but with a slope 1 less than that of the original distribution.

Newman "Networks, an Introduction"

Advantages:
- no bins
- easy calculation
- can be plotted as normal function at log-log scale
- binning loses the information; cumulative distribution preserves everything

Disadvantages
- less easy to interpret than histograms
- successive points are correlated

Cumulative Distribution

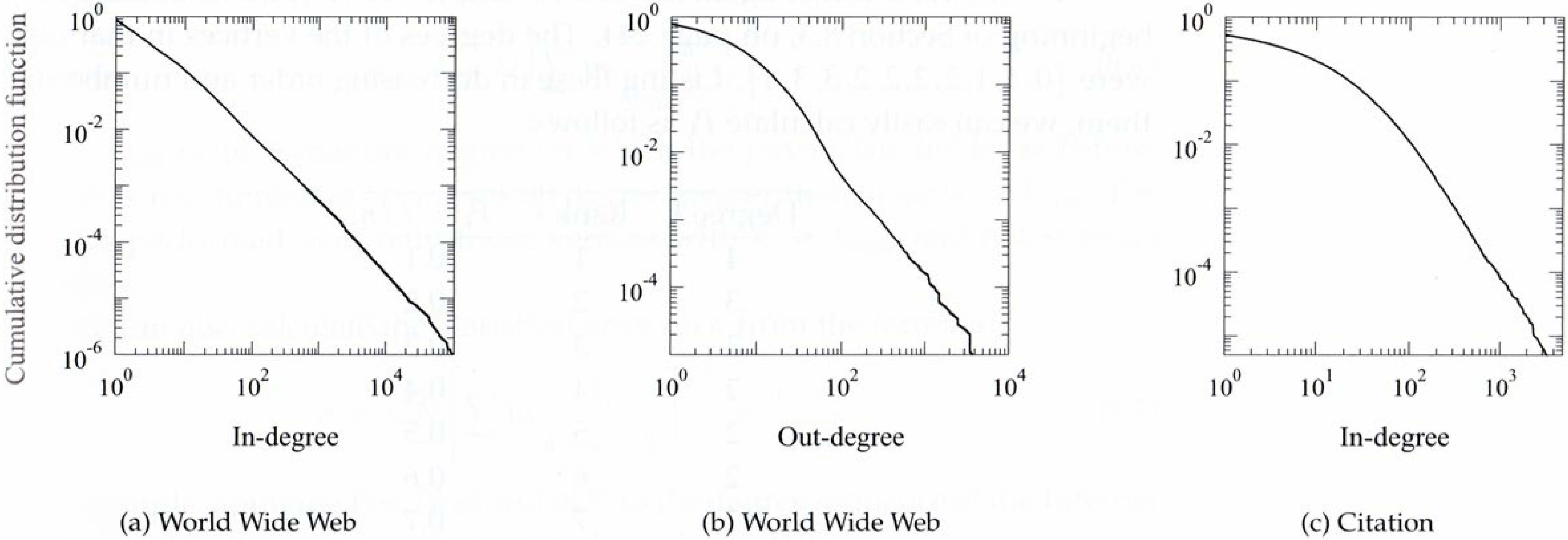


Figure 8.8: Cumulative distribution functions for in- and out-degrees in directed networks. (a) The in-degree distribution of the World Wide Web, from the data of Broder *et al.* [56]. (b) The out-degree distribution for the same Web data set. (c) The in-degree distribution of a citation network, from the data of Redner [280]. The distributions follow approximate power-law forms in each case.

From Newman "Networks, an Introduction"

Power Laws

More examples: city populations, moon craters, solar flares, computer files, words frequencies in human languages, hits on web pages, publications per scientist, book sales, …

Normalization: we have to find C such that $\sum_{k=0}^{\infty} p_k = 1$
After eliminating $k = 0$

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha)}, \text{ i.e., } p_k = \frac{k^{-\alpha}}{\zeta(\alpha)}, \text{ where } p_0 = 0$$

Riemann zeta function

$$\frac{1}{\zeta(s)} = \prod_{p \text{ is prime}} \left(1 - \frac{1}{p^{-s}}\right)$$

However, pure power-law behavior is not perfect for real-world networks
Normalization over the tail:

incomplete Riemann zeta function

$$p_k = \frac{k^{\alpha}}{\sum_{k=k_{\min}}^{\infty} k^{-\alpha}} = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})}$$

or if we approximate it then $C \approx 1/\left(\int_{k_{\min}}^{\infty} k^{-\alpha} dk\right) = (\alpha - 1) k_{\min}^{\alpha - 1}$

Moments: The mth moment of the distribution is defined as

$$\langle k^m \rangle = \sum_{k=0}^{\infty} k^m p_k = \sum_{k=0}^{k_{\min}-1} k^m p_k + C \sum_{k=k_{\min}}^{\infty} k^m k^{-\alpha}$$

if power law begins with some k_{min}

mth moment exists (finite) when $\alpha > m+1$ (integrate the second term)

Remark: This estimate works for arbitrarily large network with the same power law distribution. For finite network $\langle k^m \rangle = \frac{1}{n} \sum_{i \in V} d(i)^m$

Another interesting question is where the majority of the distribution of x lies. For any power law with exponent $\alpha > 1$, the median is well defined. That is, there is a point $x_{\frac{1}{2}}$ that divides the distribution in half so that half the measured values of x lie above $x_{\frac{1}{2}}$ and half lie below.

$$\int_{x_{1/2}}^{\infty} p(x)\,\mathrm{d}x = \tfrac{1}{2}\int_{x_{\min}}^{\infty} p(x)\,\mathrm{d}x,$$

Point that divides distribution in two halves

$$x_{1/2} = 2^{1/(\alpha-1)} x_{\min}.$$

Further reading: Newman “Power laws, Pareto distributions and Zipf’s law”

Top-heavy distributions or 80/20 rule: how many edges are connected to the highest degree vertices?

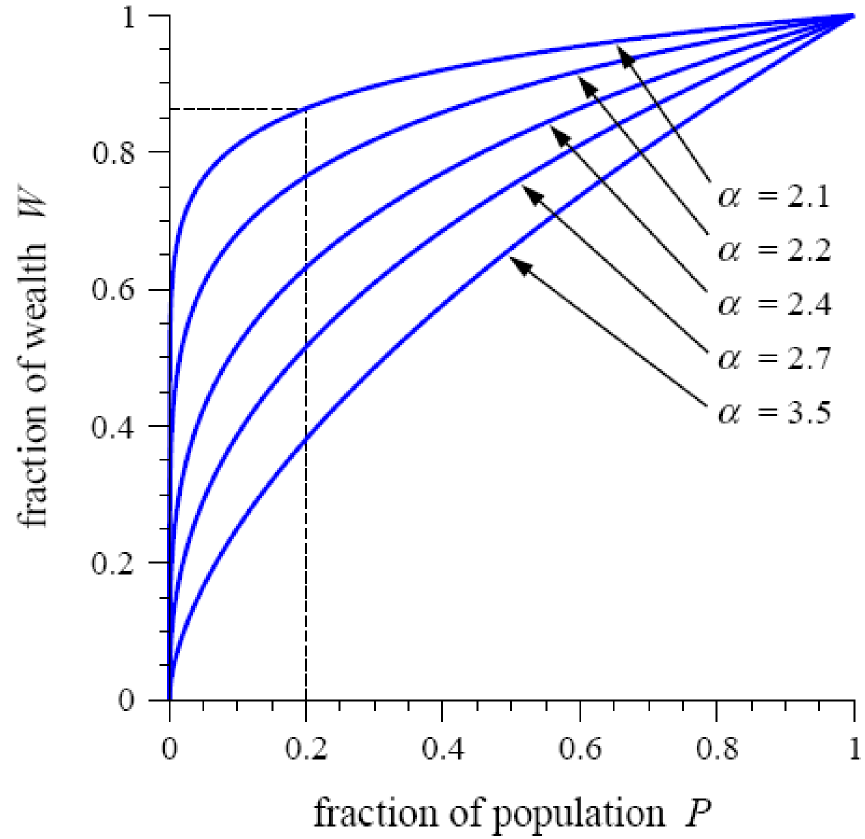
A fraction of edges attached to the highest degree vertices

↓

$$W = P^{(\alpha-2)/(\alpha-1)}$$

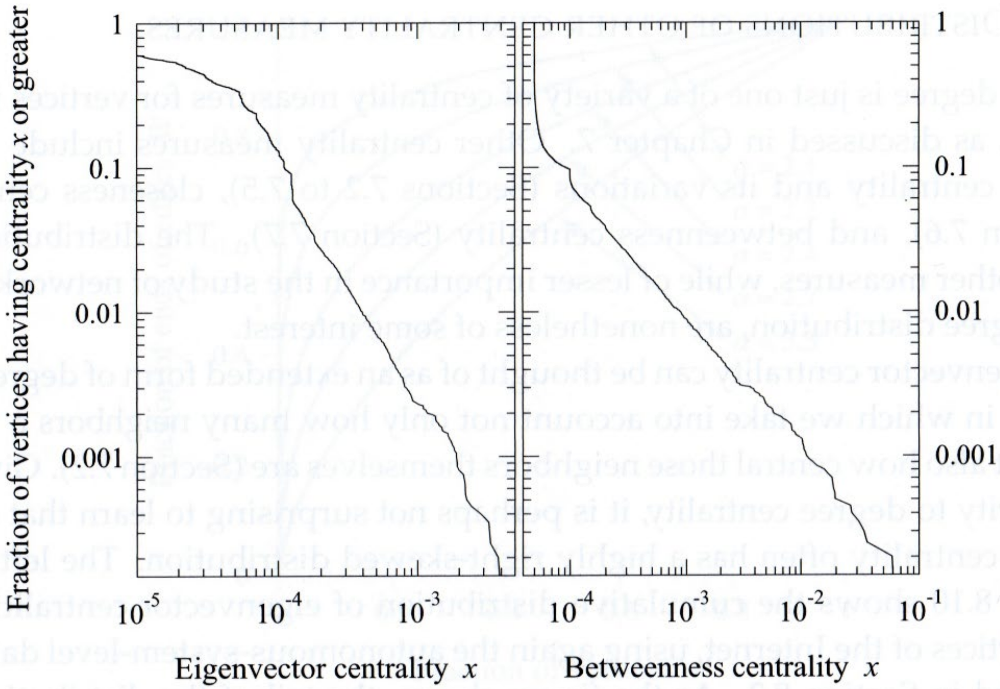
↑

A fraction of highest degree vertices

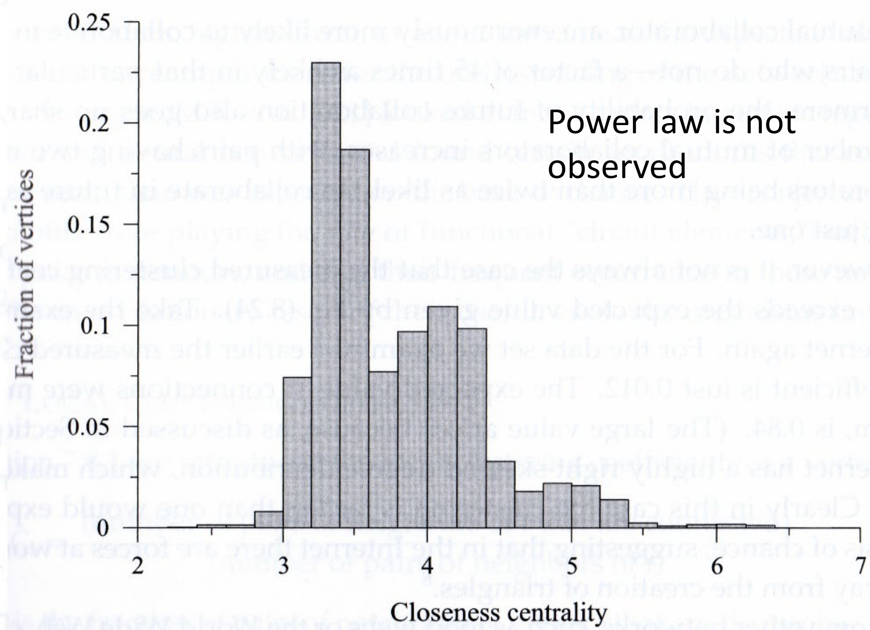


Example 1: According to various estimations, 50-60% of the incoming links point to 1% of the "reach" nodes.
Example 2: In scientific citation networks, about 8% of papers are cited by more than 50% of all papers.

Further reading: Newman "Power laws, Pareto distributions and Zipf's law"



Cumulative distributions for Internet nodes



Noncumulative histogram for Internet nodes

An exception to this pattern is the closeness centrality, which is the mean of distances from a vertex to all other reachable vertices. The values of the closeness centrality are typically limited to a rather small range from a lower bound of 1 to an upper bound of order log *n*, and this means that their distribution cannot have a long tail.

Homework: paper review + computational part
Submit by 10/8/2020

1. (20%) Newman "Power laws, Pareto distributions and Zipf's law"
2. (80%) Computational part
   - Download network "as-22july06" from the Sparse matrix collection
   - Plot the degree distribution histogram
   - Plot the cumulative degree distribution function
   - Compute power law parameters *C*, and α

Clustering Coefficient and Transitivity

A triangle is a complete subgraph of G with 3 vertices.

$\lambda(G)$ = number of triangles in G; $\lambda(v)$ is defined accordingly; $\lambda(G) = \frac{1}{3}\sum_v \lambda(v)$

A triple is a subgraph of G with 3 nodes and 2 edges.

A triple is a *triple at v* if v incident with both edges.

$$\tau(v) = \binom{d(v)}{2} = \frac{d^2(v) - d(v)}{2}, \;\; \tau(G) = \sum_v \tau(v)$$

We define *clustering coefficient* as $c(v) = \lambda(v)/\tau(v)$.

Given $V' = \{v \in V | d(v) \geq 2\}$ we define the clustering coefficient of G as

$$C(G) = \frac{1}{V'} \sum_{v \in V'} c(v)$$

Transitivity of G is defined as

$$T(G) = \frac{3\lambda(G)}{\tau(G)}$$

Clustering Coefficient and Transitivity

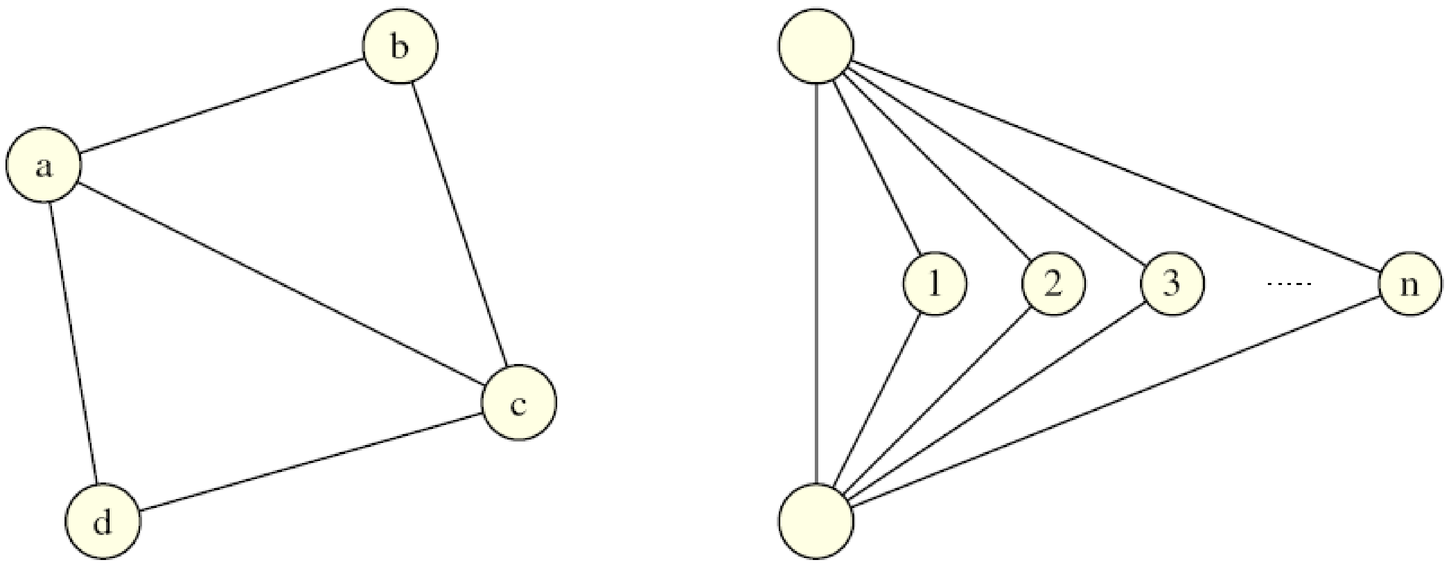


Fig. 11.2. On the left: Graph with clustering coefficients: $c(a) = c(c) = 2/3$, $c(b) = c(d) = 1$, $C(G) = \frac{1}{4}(2 + 4/3) \approx 0.83$ and transitivity $T(G) = 3 \cdot 2/8 = 0.75$. On the right: family of graphs where $T(G) \to 0$, $C(G) \to 1$ for $n \to \infty$.

Clustering Coefficient and Transitivity

Transitivity by Bollobas and Riordan

$$T(G) = \frac{\sum_{v \in V'} \tau(v) c(v)}{\sum_{v \in V'} \tau(v)}$$

- If all nodes have the same degree then $C(G) = T(G)$
- If all clustering coefficients are equal then $C(G) = T(G)$

Computing Clustering Coefficient

Computing cc = computing triples (trivial, how?) + computing triangles
Computing triangles = $O(nd_{max}^2)$ – trivial, $O(n^{2.376})$ – mat-mat multiplication

Approximation for very large networks

$X_i \in [0, M]$ is a random independent and identically distributed variable; k is number of samples; ϵ is error bound

Hoeffding inequality

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} X_i - \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} X_i\right]\right| \geq \epsilon\right) \leq e^{\frac{-2k\epsilon^2}{M^2}}$$

Lemma: If we consider the constant error bound then there exist algorithms that approximate the clustering coefficients for each node $c(v)$ and the transitivity $T(G)$ in time $O(n)$. The clustering coefficient $C(G)$ can be approximated in time in O(1).

Homework: "Approximating clustering-coefficient and transitivity" (submit review by 10/13/2019)