Descriptive network analysis

Sorokin Semen

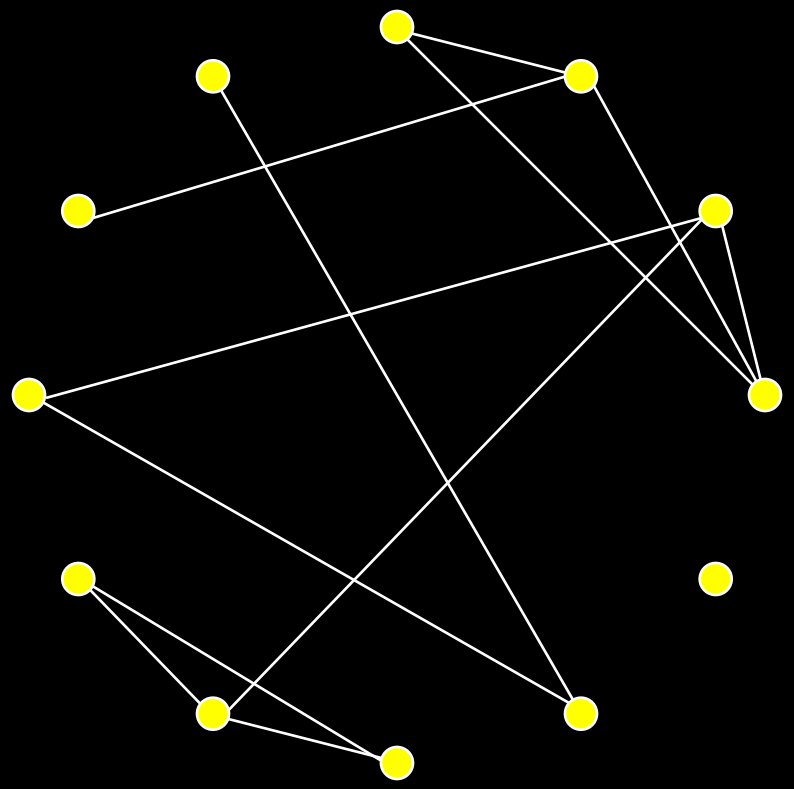
Society as a Graph

People are represented as *nodes.*

Relationships are represented as *edges.*

(Relationships may be acquaintanceship, friendship, co-authorship, etc.)

Allows analysis using tools of mathematical graph theory



Six Degrees of Separation

Milgram (1967)

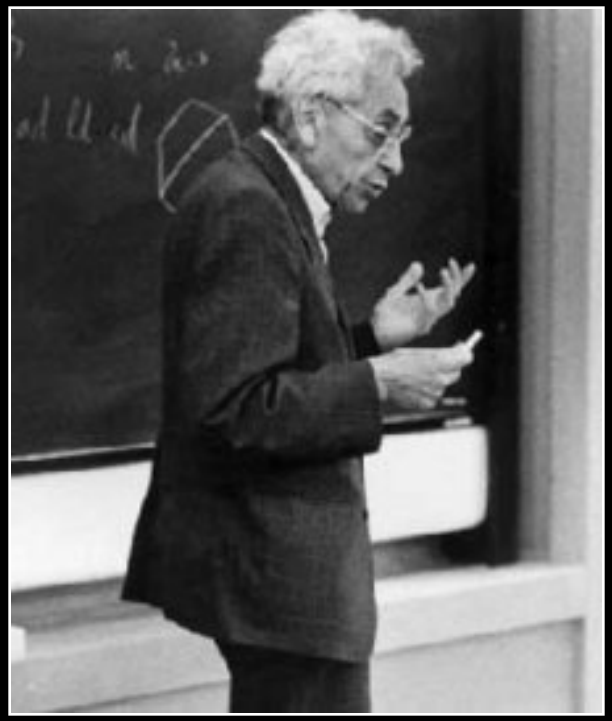
The experiment:

- Random people from Nebraska were to send a letter (via intermediaries) to a stock broker in Boston.
- Could only send to someone with whom they were on a first-name basis.

Among the letters that found the target, the average number of links was six.

Stanley Milgram (1933-1984)

Erdős Number



Paul Erdős (1913-1996)

Number of links required to connect scholars to Erdős, via co-authorship of papers

Erdős wrote 1500+ papers with 507 co-authors.

Jerry Grossman's (Oakland Univ.) website allows mathematicians to compute their Erdos numbers:

http://www.oakland.edu/enp/

Connecting path lengths, among mathematicians only:

- average is *4.65*
- maximum is *13*

Unlike Bacon, Erdos has better centrality in his network

Random Graphs

Erdős and Renyi (1959)

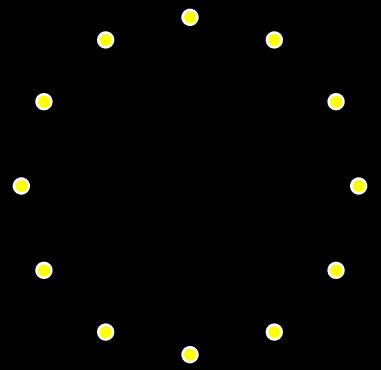
N nodes

A pair of nodes has probability *p* of being connected.

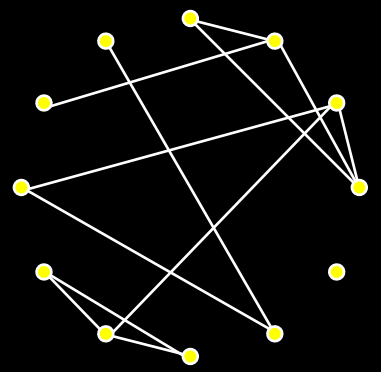
Average degree, $k \approx pN$

$N = 12$

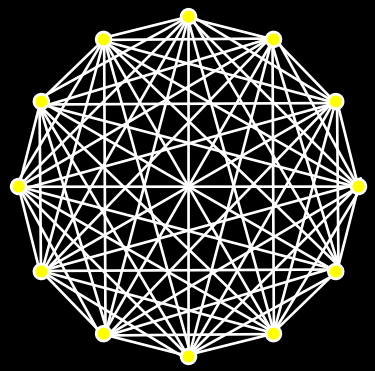
$p = 0.0$; $k = 0$



$p = 0.09$; $k = 1$



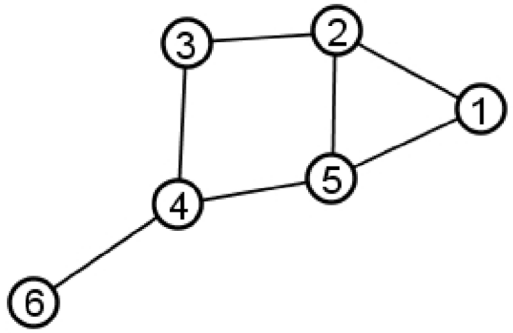
$p = 1.0$; $k \approx$ N



Graph connectivity

- The *distance* $d_G(v_i, v_j)$ between two vertices is the number of edges in the shortest path from v_i to v_j
- Graph *diameter* is the largest shortest path: $D = \max_{i,j} d_G(v_i, v_j)$
- Average path length:

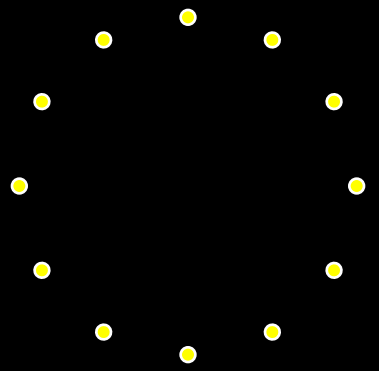
$$\langle L \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_G(v_i, v_j)$$



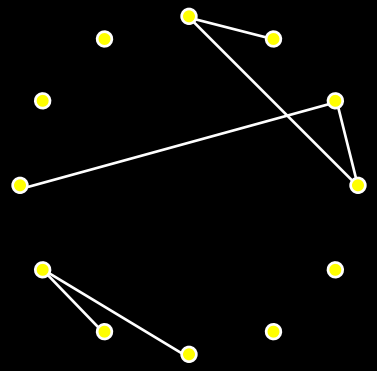
Random Graphs

Erdős and Renyi (1959)

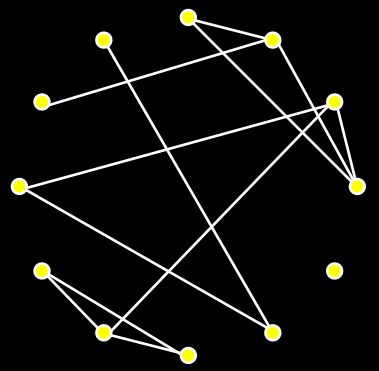
Let's look at…



$p = 0.0$; $k = 0$



$p = 0.045$; $k = 0.5$

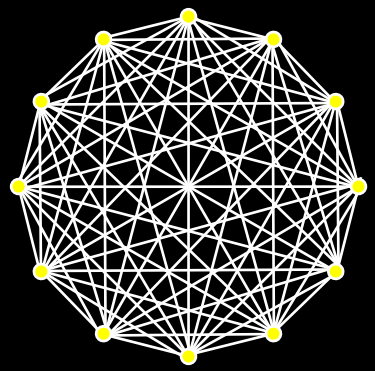


$p = 0.09$; $k = 1$

Size of the largest connected cluster

Diameter (maximum path length between nodes) of the largest cluster

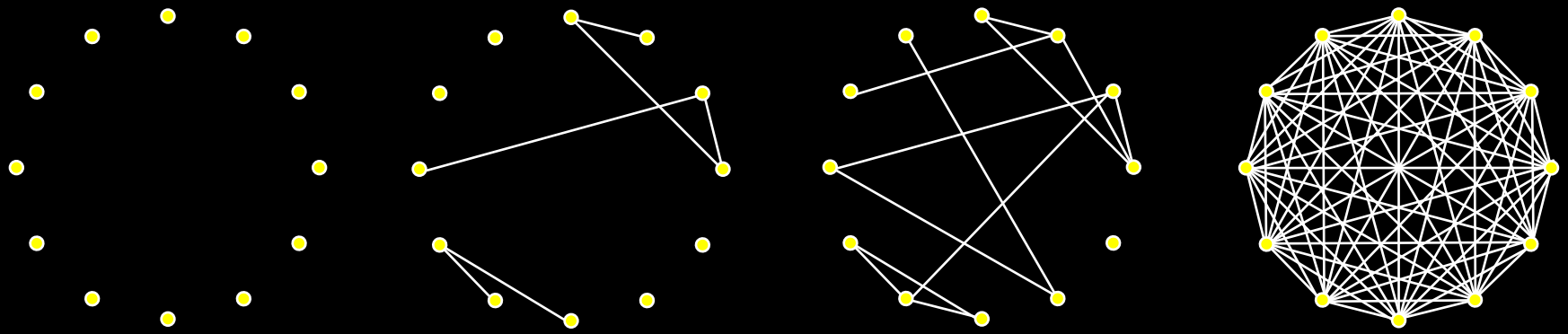
Average path length between nodes (if a path exists)



$p = 1.0$; $k \approx N$

Random Graphs

Erdős and Renyi (1959)



$p = 0.0$; $k = 0$	$p = 0.045$; $k = 0.5$	$p = 0.09$; $k = 1$	$p = 1.0$; $k \approx N$
Size of largest component			
1	5	11	12
Diameter of largest component			
0	4	7	1
Average path length between (connected) nodes			
0.0	2.0	4.2	1.0

Random Graphs

Erdős and Renyi (1959)

If $k < 1$:

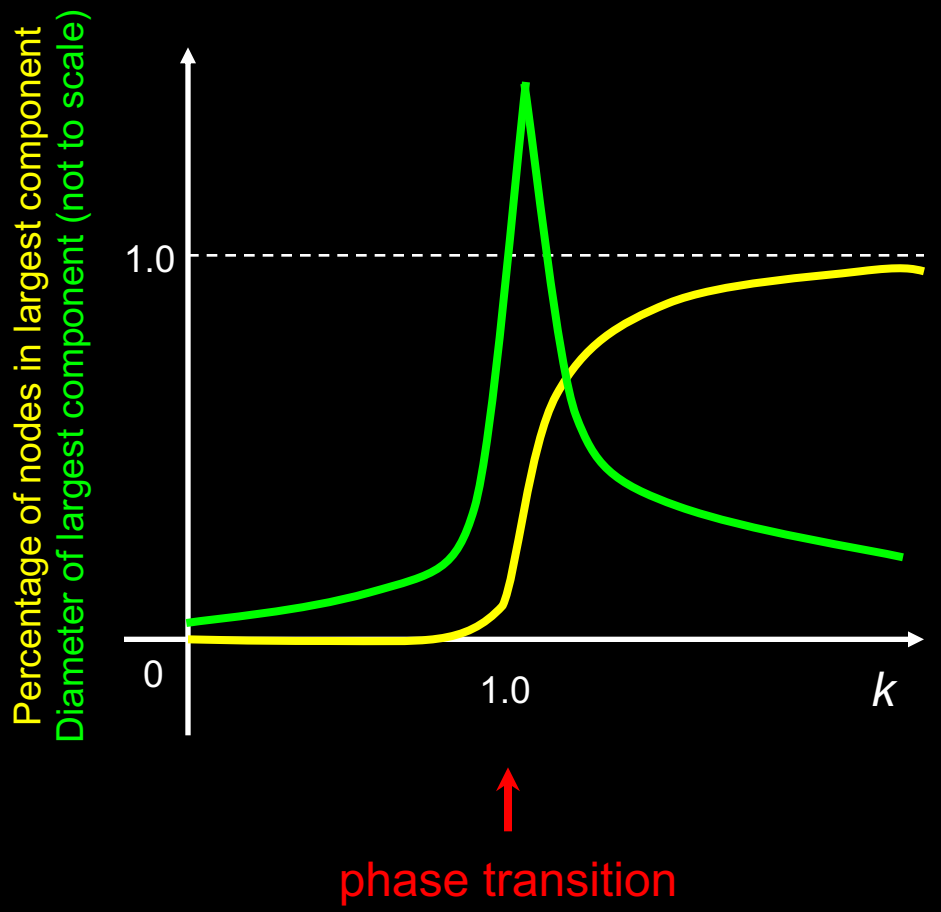
- small, isolated clusters
- small diameters
- short path lengths

At k = 1:

- a *giant component* appears
- diameter peaks
- path lengths are high

For k > 1:

- almost all nodes connected
- diameter shrinks
- path lengths shorten



Random Graphs

Erdős and Renyi (1959)

What does this mean?

Random Graphs

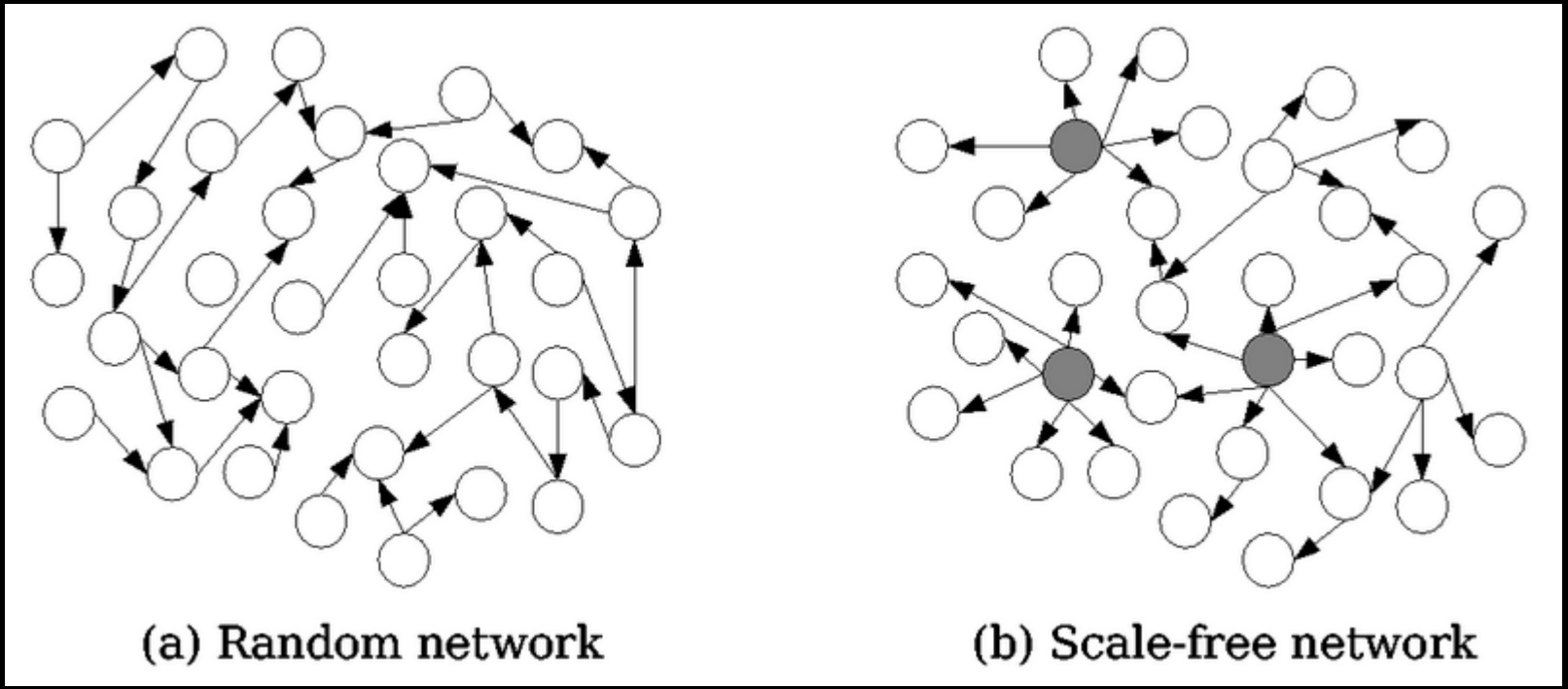
Erdős and Renyi (1959)

What does this mean?

- If connections between people can be modeled as a random graph, then…
  - Because the average person easily knows more than one person (k >> 1),
  - We live in a “small world” where within a few links, we are connected to anyone in the world.

Random vs. Real Social networks

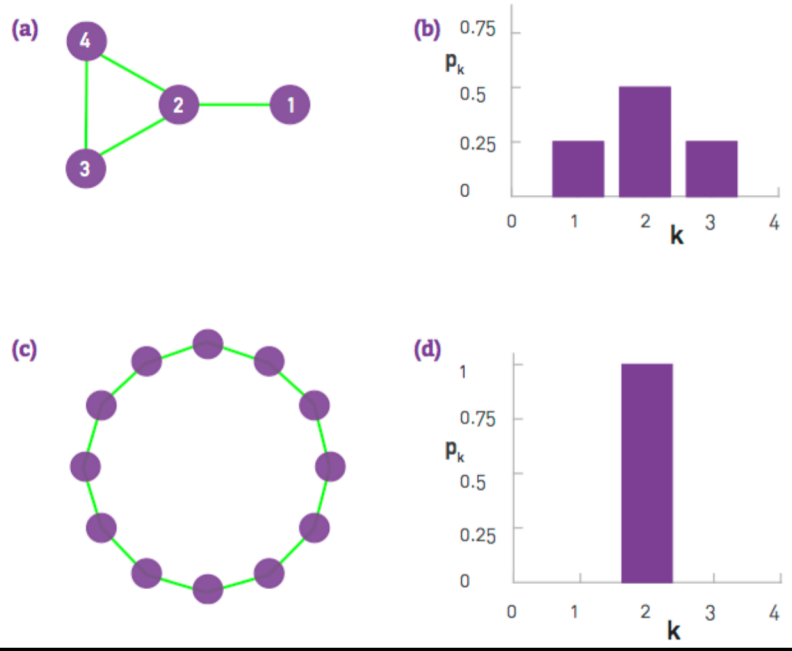
- Random network models introduce an edge between any pair of vertices with a probability *p*
  - The problem here is NOT randomness, but rather the distribution used (which, in this case, is *uniform*)
- Real networks are not exactly like these
  - Tend to have a relatively few nodes of high connectivity
  - These networks are called "Scale-free" networks
    - Macro properties scale-invariant



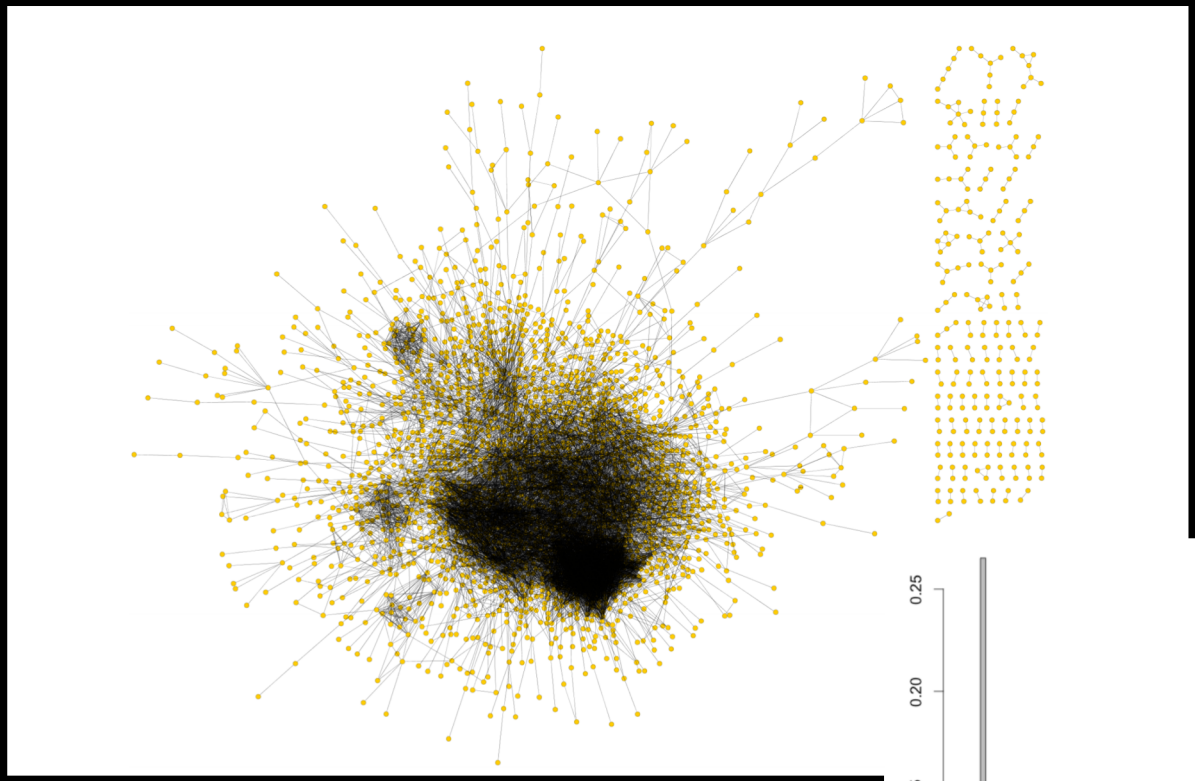
Degree Distribution

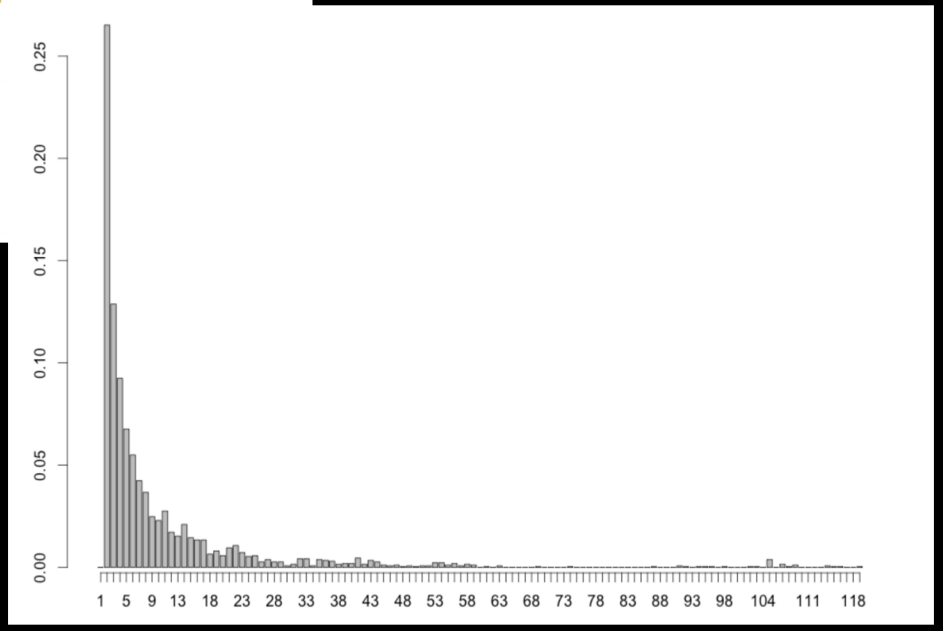
- k_i - node degree, $k_i = 1, 2, \ldots k_{\max}$
- n_k - number of nodes with degree k, total nodes $n = \sum_k n_k$
- Degree distribution is a fraction of the nodes with degree k

$$P(k_i = k) = P(k) = P_k = \frac{n_k}{\sum_k n_k} = \frac{n_k}{n}$$

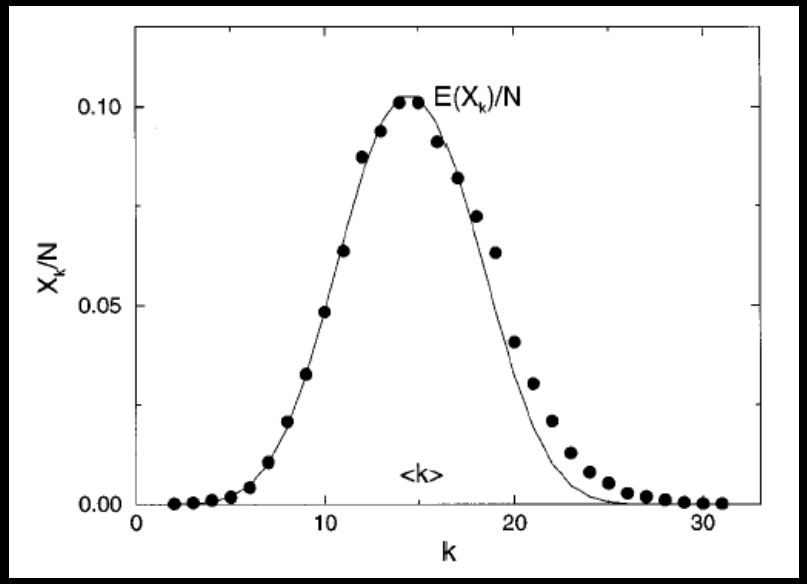


Graph example

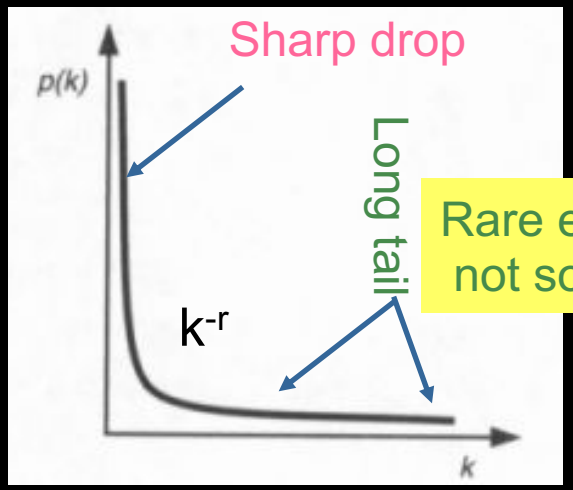




Degree Distribution & Power Laws



Degree distribution of a random graph,
N = 10,000 p = 0.0015 k = 15.
(Curve is a Poisson curve, for comparison.)



Rare events are not so rare!

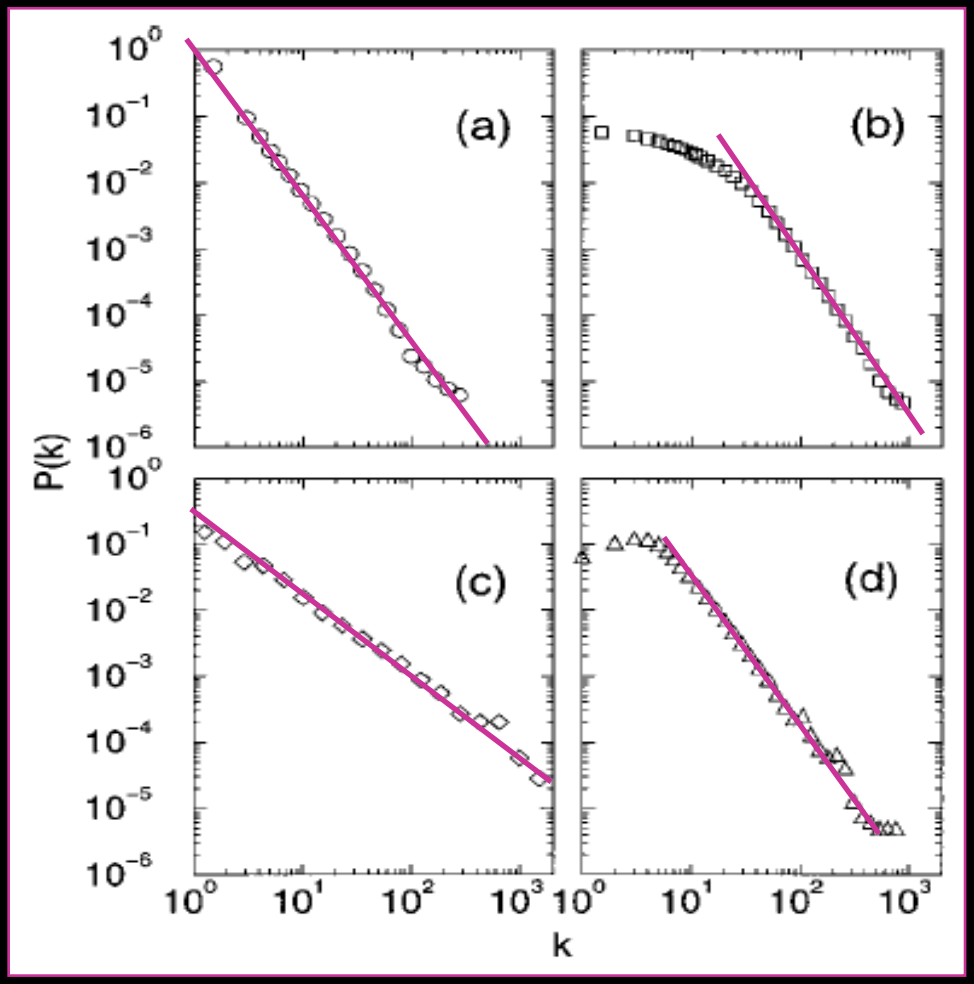
But, many real-world networks exhibit a *power-law* distribution.

→also called "Heavy tailed" distribution

Typically 2<r<3. For web graph
r ~ 2.1 for in degree distribution
2.7 for out degree distribution

Power Laws

Albert and Barabasi (1999)

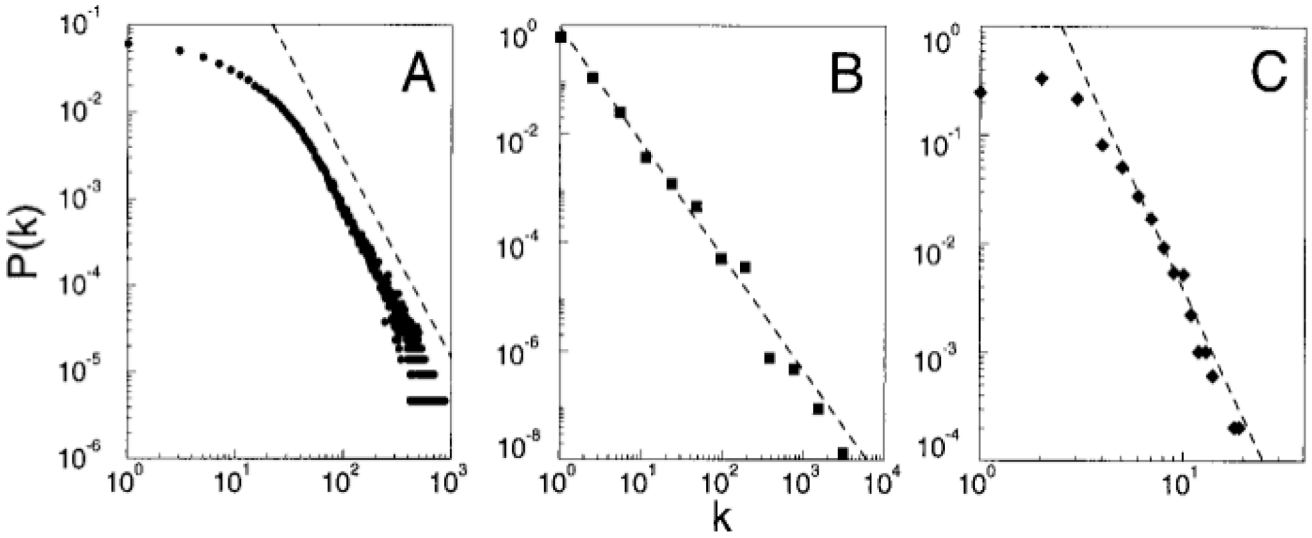


Power-law distributions are straight lines in log-log space.

$y=k^{-r} \rightarrow \log y = -r \log k \rightarrow ly = -r\, lk$

Power laws in real networks:
(a) WWW hyperlinks
(b) co-starring in movies
(c) co-authorship of physicists
(d) co-authorship of neuroscientists

Log-log scale



Actor collaboration graph, N=212,250 nodes, $\langle k \rangle = 28.8, \gamma = 2.3$

WWW, N = 325,729 nodes, $\langle k \rangle = 5.6, \gamma = 2.1$

Power grid data, N = 4941 nodes, $\langle k \rangle = 5.5, \gamma = 4$

Barabasi et.al, 1999

Zipf's Law: Power law distriubtion between rank and frequency

Digression

- In a given language corpus, what is the approximate relation between the frequency of a k^{th} most frequent word and $(k+1)^{th}$ most frequent word?

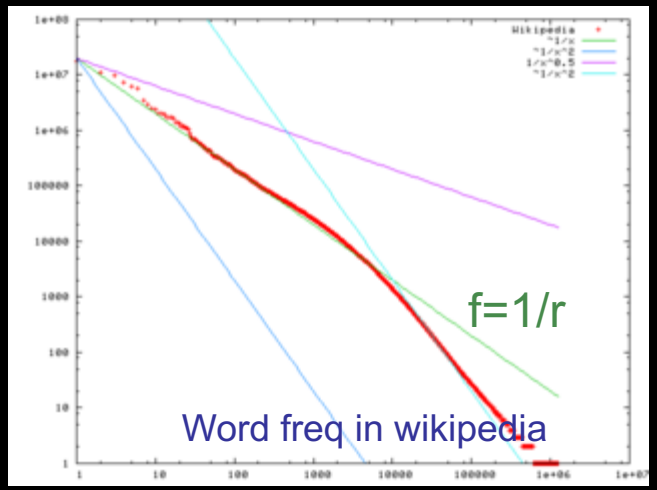
Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$$

For s>1 $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} < \infty.$

Most popular word is twice as frequent as the second most popular word!



Law of categories in Marketing…

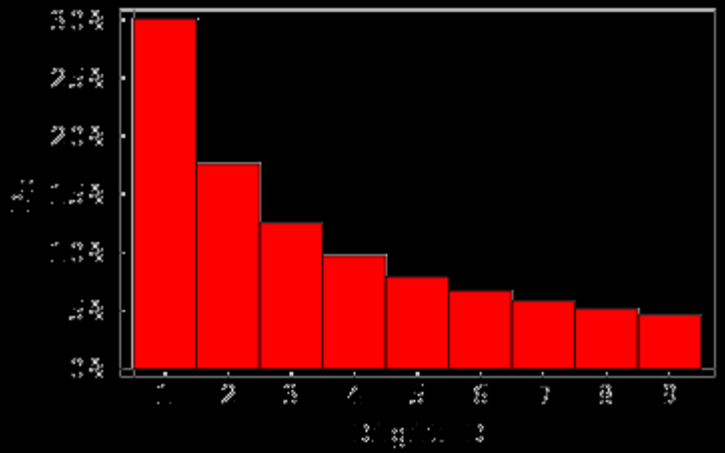
What is the explanation for Zipf's law?

- Zipf's law is an empirical law in that it is observed rather than "proved"
- Many explanations have been advanced as to why this holds.
- Zipf's own explanation was "principle of least effort"
  - Balance between speaker's desire for a small vocabulary and hearer's desire for a large one (so meaning can be easily disambiguated)
- Alternate explanation— "rich get richer" –popular words get used more often
- Li (1992) shows that just random typing of letters with space will lead to a "language" with zipfian distribution..

Benford's law (aka first digit phenomenon)

How often does the digit 1 appear in numerical data describing natural phenomenon?

- You would expect 1/9 or 11%



This law holds so well in practice that it is used to catch forged data!!

WHY?

If there exists a universal distribution, it must be scale invariant (i.e., should work in any units)

→ starting from there we can show that the distribution must satisfy the differential eqn

$x\,P'(x) = -P(x)$

For which, the solution is $P(x)=1/x$!

D	P_D	D	P_D
1	0.30103	6	0.0669468
2	0.176091	7	0.0579919
3	0.124939	8	0.0511525
4	0.09691	9	0.0457575
5	0.0791812		

http://mathworld.wolfram.com/BenfordsLaw.html

Digression begets its own digression

Power Laws & Scale-Free Networks

"The rich get richer!"

Examples of Scale-free networks (i.e., those that exhibit power law distribution of in degree)

- Social networks, including collaboration networks. An example that have been studied extensively is the collaboration of movie actors in films.
- Protein-interaction networks.
- Sexual partners in humans, which affects the dispersal of sexually transmitted diseases.
- Many kinds of computer networks, including the World Wide Web.

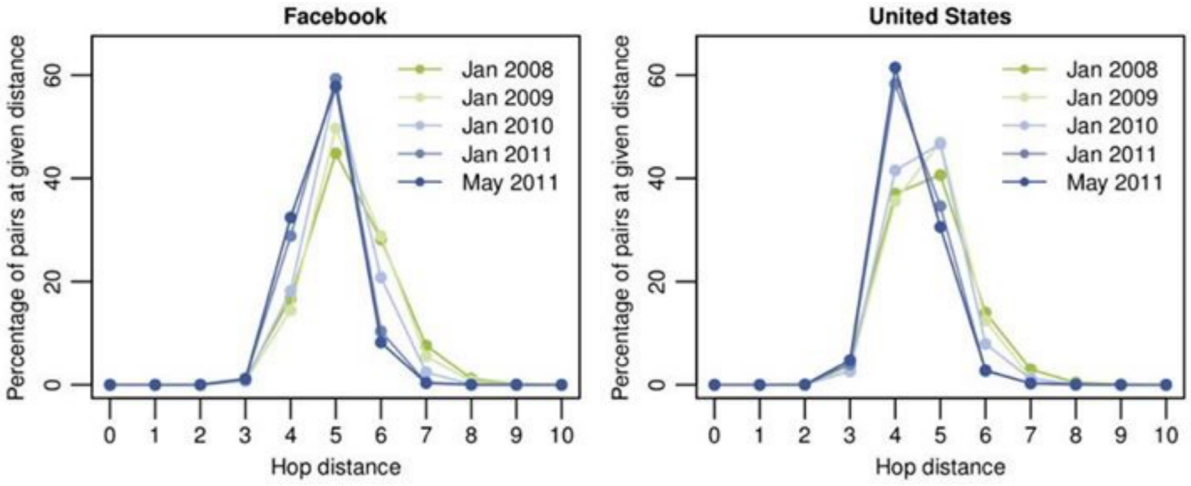
Power-law distribution of node-degree arises if

(but *not* "only if")

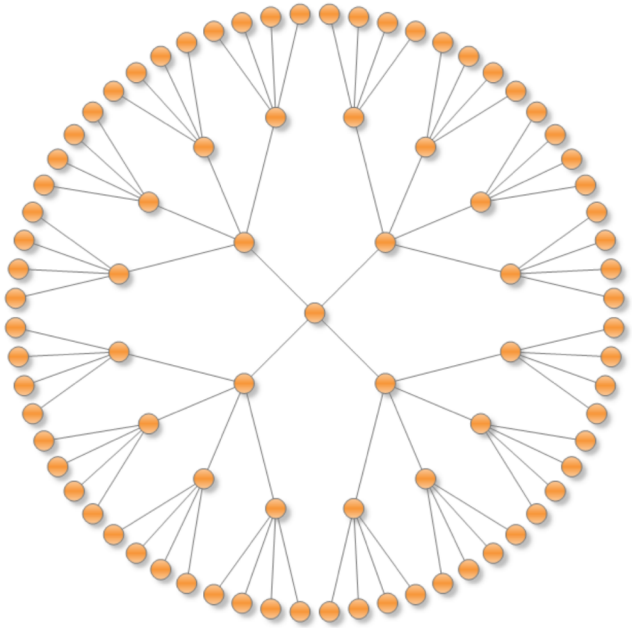
- As Number of nodes grow edges are added in proportion to the number of edges a node already has.
  - Alternative: Copy model—where the new node copies a random subset of the links of an existing node
    - Sort of close to the WEB reality

Graph average path length in Small world

- Email graph:
  D. Watts (2001), 48,000 senders, $\langle L \rangle \approx 6$
- MSN Messenger graph:
  J. Lescovec et al (2007), 240mln users, $\langle L \rangle \approx 6.6$
- Facebook graph:
  L. Backstrom et al (2012), 721 mln users, $\langle L \rangle \approx 4.74$



Simple model



An estimate: $z^d = N$, $d = \log N / \log z$

$N \approx 6.7$ bln, $z = 50$ friends, $d \approx 5.8$.

Summary

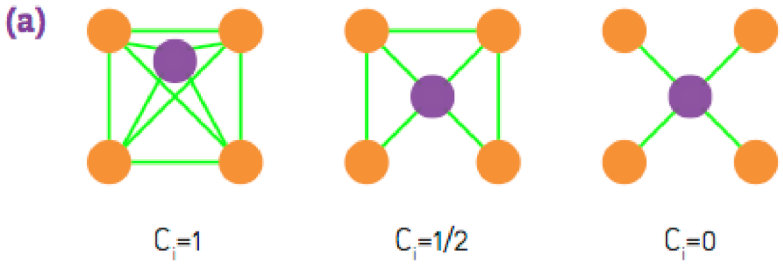
- A network is considered to exhibit small world phenomenon, if its diameter is approximately logarithm of its size (in terms of number of nodes)
- Most uniform random networks exhibit small world phenomena
- Most real world networks are not uniform random
  - Their in degree distribution exhibits power law behavior
  - However, most power law random networks also exhibit small world phenomena
- The fact that a network exhibits small world phenomenon doesn't mean that an agent with strictly local knowledge can efficiently navigate it (i.e, find paths that are O(log(n)) length
  - It is always possible to find the short paths if we have global knowledge
    - This is the case in the FOAF (friend of a friend) networks on the web

Clustering coefficient (local)

How neighbors of a given node connected to each other

- *Local clustering coefficient* (per vertex):

$$C_i = \frac{\text{number of links in } \mathcal{N}_i}{k_i(k_i-1)/2}$$



- Average clustering coefficient:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$