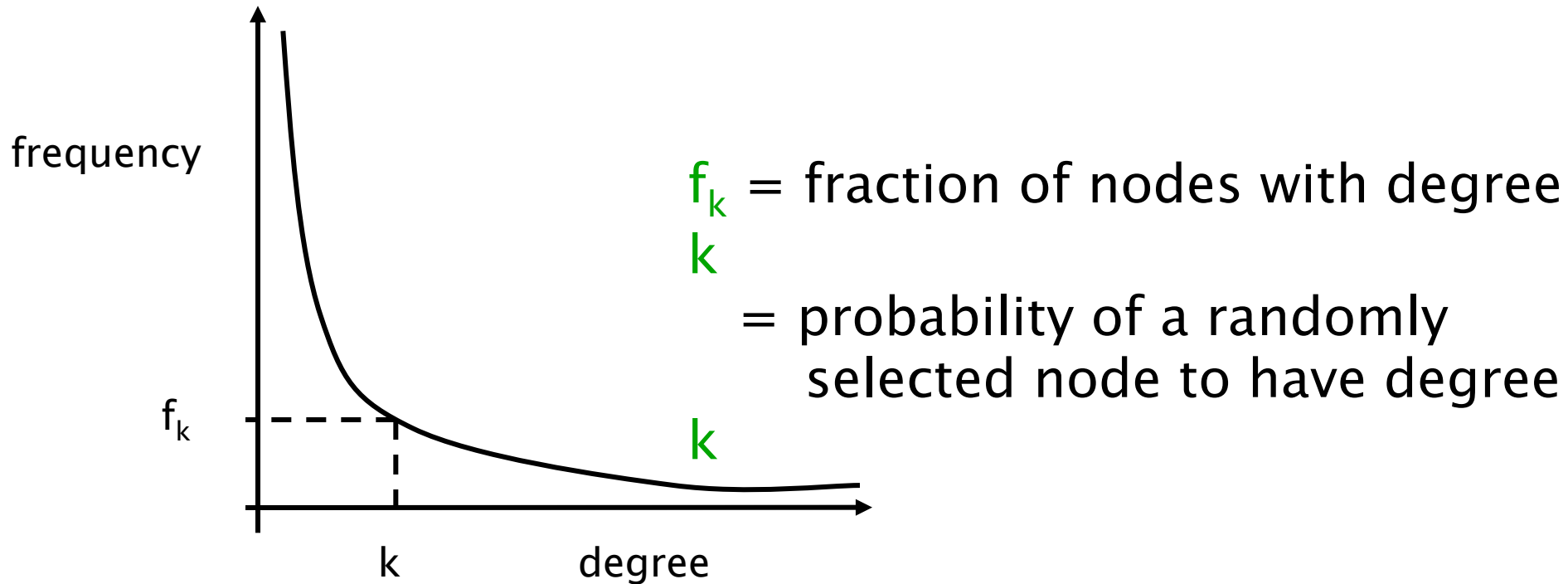


Basics of network analysis and network models

Measuring Networks

- Degree distributions
- Small world phenomena
- Clustering Coefficient
- Mixing patterns
- Degree correlations
- Communities and clusters

Degree distributions



- Problem: find the probability distribution that best fits the observed data

Power-law distributions

- The degree distributions of most real-life networks follow a power law

$$p(k) = Ck^{-\alpha}$$

- Right-skewed/Heavy-tail distribution
 - there is a non-negligible fraction of nodes that has very high degree (hubs)
 - **scale-free**: no characteristic scale, average is not informative

- In stark contrast with the random graph model!
 - Poisson degree distribution, $z=np$

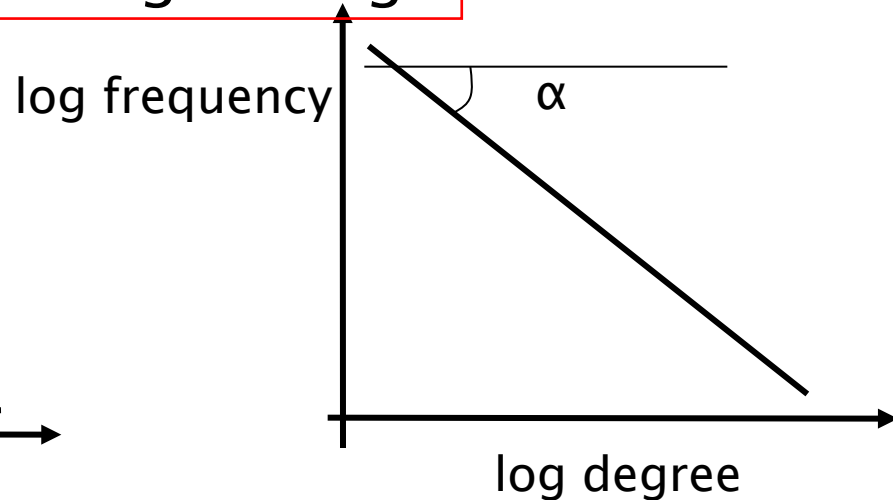
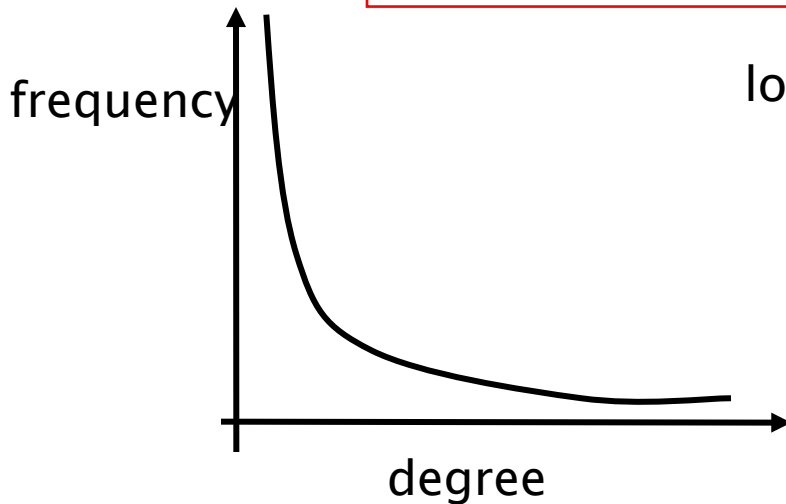
$$p(k) = P(k; z) = \frac{z^k}{k!} e^{-z}$$

- highly concentrated around the mean
 - the probability of very high degree nodes is exponentially small

Power-law signature

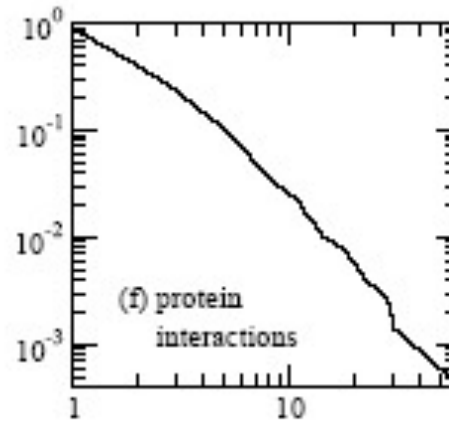
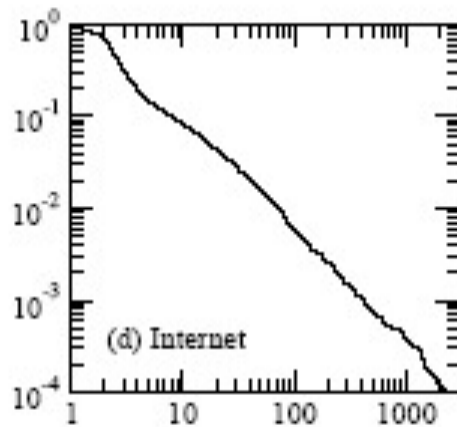
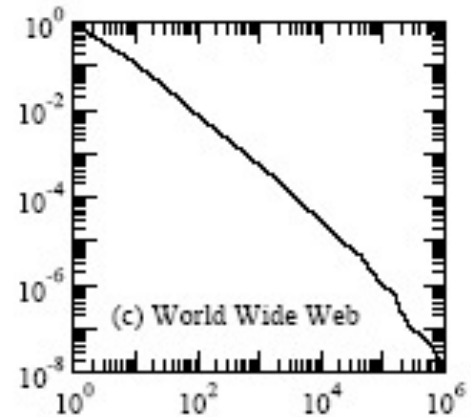
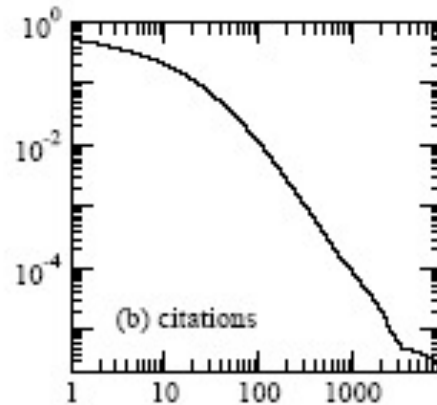
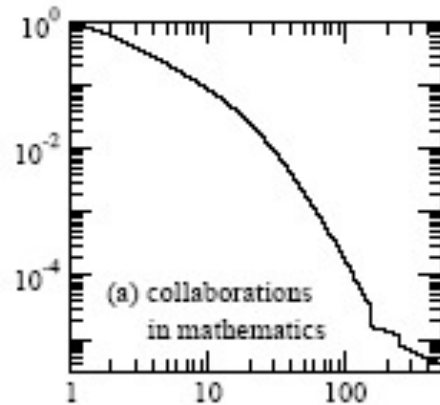
- Power-law distribution gives a line in the log-log plot

$$\log p(k) = -\alpha \log k + \log C$$



- α : power-law exponent (typically $2 \leq \alpha \leq 3$)

Examples



Taken from [Newman 2003]

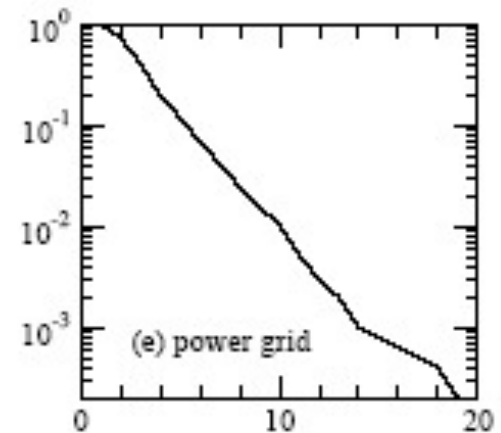
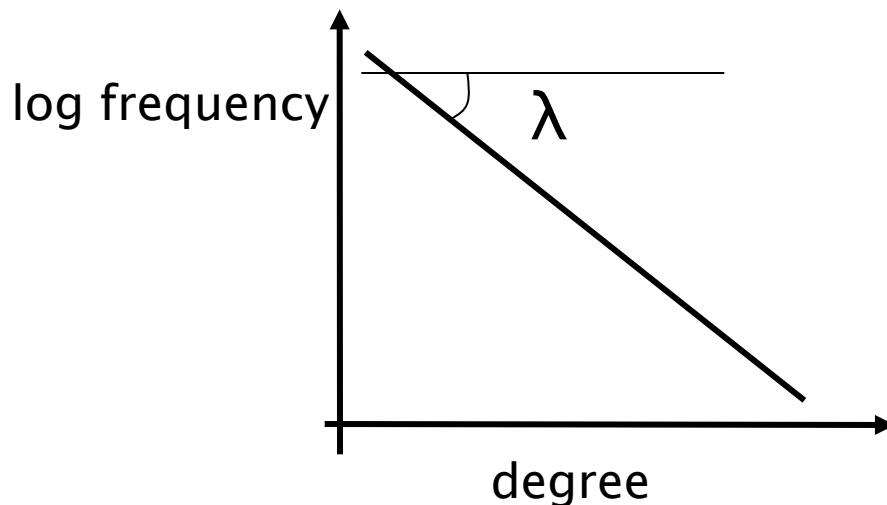
Exponential distribution

- Observed in some technological or collaboration networks

$$p(k) = \lambda e^{-\lambda k}$$

- Identified by a line in the log-linear plot

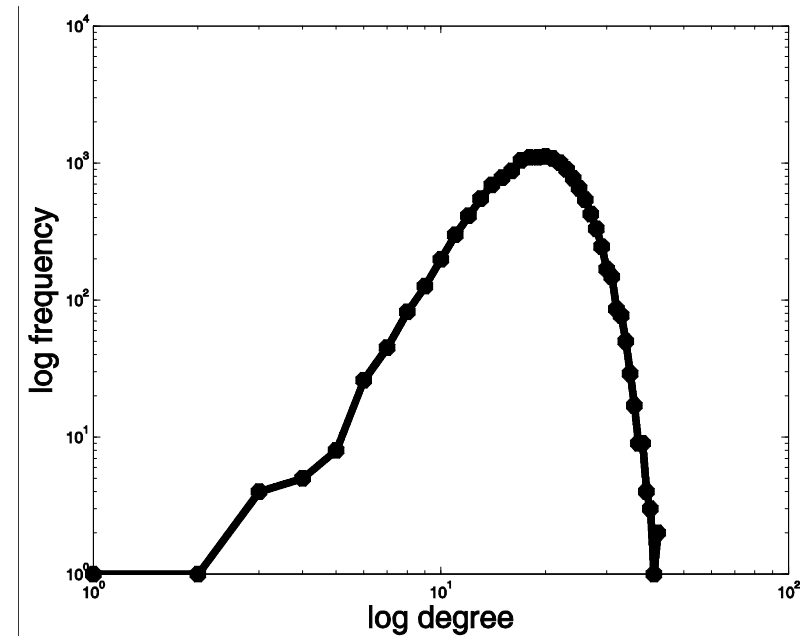
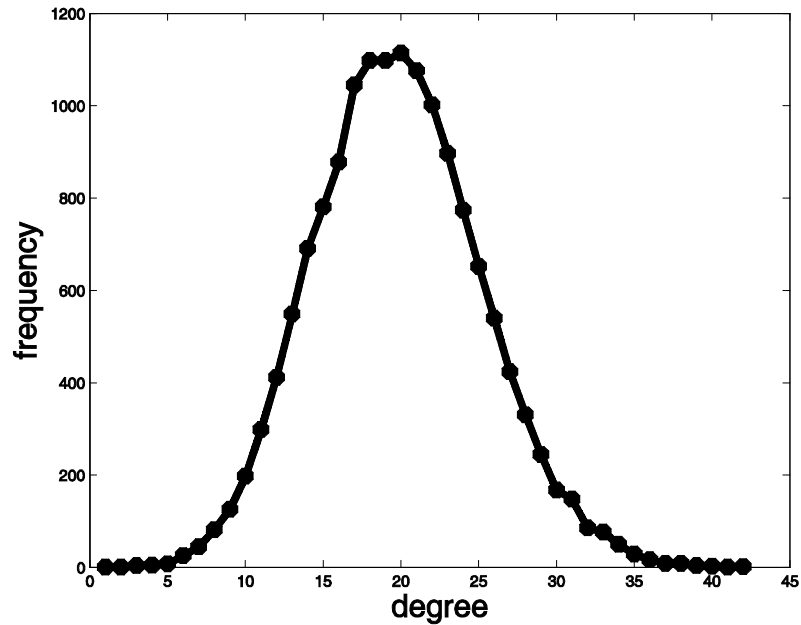
$$\log p(k) = -\lambda k + \log \lambda$$



The basic random graph model

- The measurements on real networks are usually compared against those on “random networks”
- The basic $G_{n,p}$ (Erdős–Renyi) random graph model:
 - n : the number of vertices
 - $0 \leq p \leq 1$
 - for each pair (i,j) , generate the edge (i,j) independently with probability p

A random graph example



Average/Expected degree

- For random graphs $z = np$
- For power-law distributed degree
 - if $\alpha \geq 2$, it is a constant
 - if $\alpha < 2$, it diverges

Maximum degree

- For random graphs, the maximum degree is highly concentrated around the average degree z
- For power law graphs

$$k_{\max} \approx n^{1/(\alpha-1)}$$

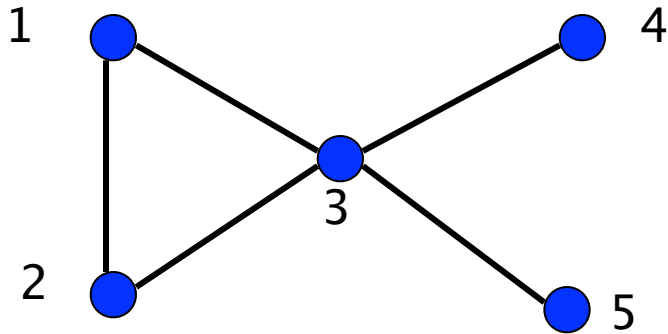
Clustering (Transitivity) coefficient

- Measures the density of triangles (local clusters) in the graph
- Two different ways to measure it:

$$C^{(1)} = \frac{\sum_i \text{triangles centered at node } i}{\sum_i \text{triples centered at node } i}$$

- The ratio of the means

Example



$$C^{(1)} = \frac{3}{1+1+6} = \frac{3}{8}$$

Clustering (Transitivity) coefficient

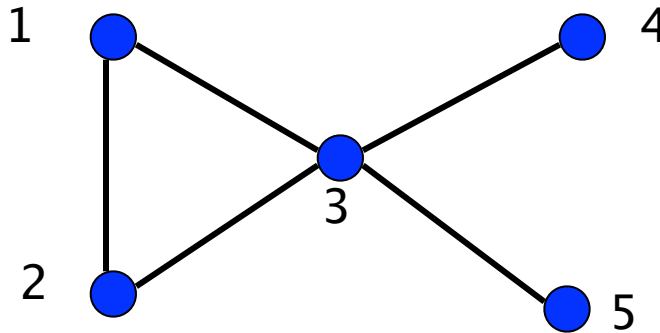
- Clustering coefficient for node i

$$C_i = \frac{\text{triangles centered at node } i}{\text{triples centered at node } i}$$

$$C^{(2)} = \frac{1}{n} C_i$$

- The mean of the ratios

Example



$$C^{(2)} = \frac{1}{5} (1 + 1 + 1/6) = \frac{13}{30}$$

$$C^{(1)} = \frac{3}{8}$$

- The two clustering coefficients give different measures
- $C^{(2)}$ increases with nodes with low degree

Clustering coefficient for random graphs

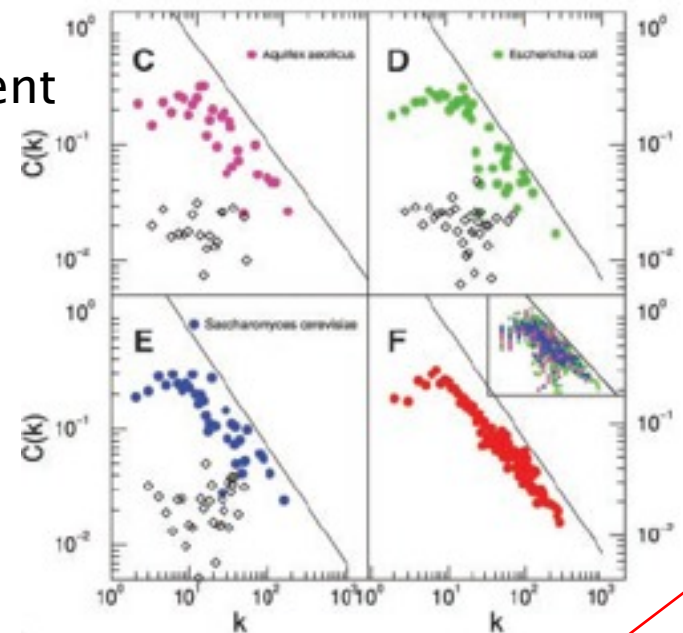
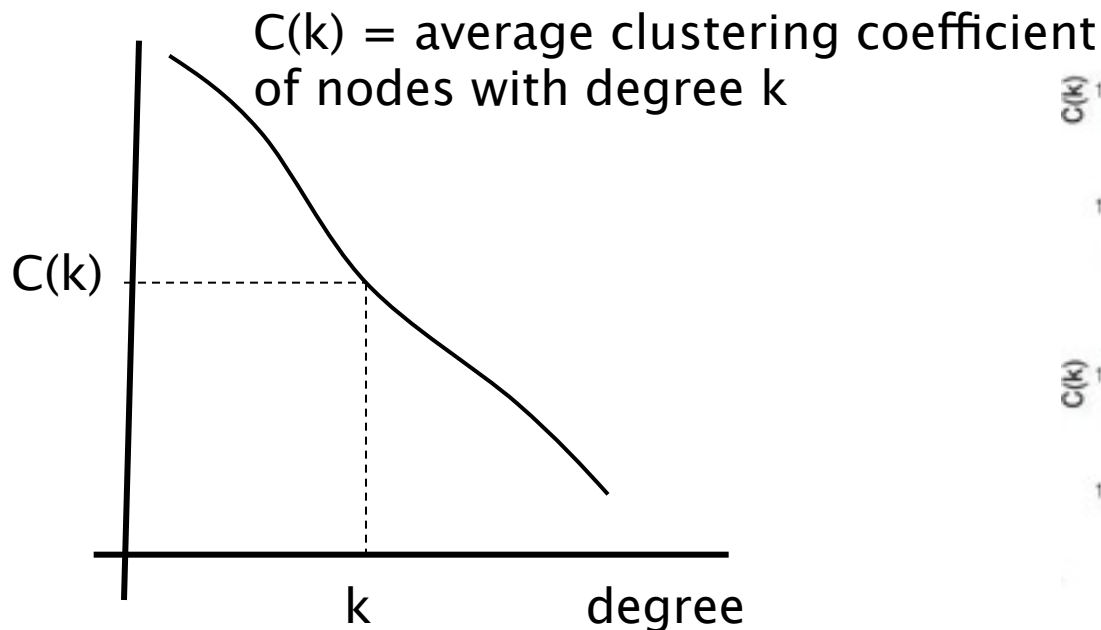
- The probability of two of your neighbors also being neighbors is p , independent of local structure
 - clustering coefficient $C = p$
 - when z is fixed $C = z/n = O(1/n)$

Table 1: Clustering coefficients, C , for a number of different networks; n is the number of nodes, z is the mean degree. Taken from [146].

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

The $C(k)$ distribution

- The $C(k)$ distribution is supposed to capture the hierarchical nature of the network
 - when constant: no hierarchy
 - when power-law: hierarchy



The small-world experiment

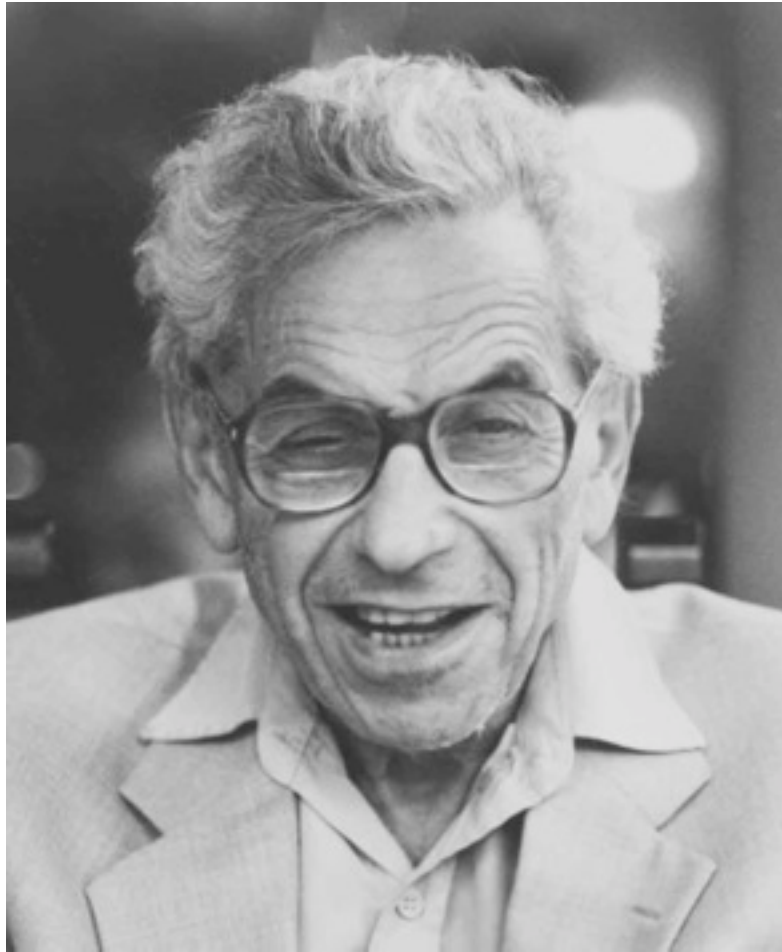
- Milgram 1967
- Picked 300 people at random from Nebraska
- Asked them to get the letter to a stockbroker in Boston – they could bypass the letter through friends they knew on a first-name basis
- How many steps does it take?
 - Six degrees of separation: (play of John Guare)

Six Degrees of Kevin Bacon



- Bacon number:
 - Create a network of Hollywood actors
 - Connect two actors if they co-appeared in some movie
 - Bacon number: number of steps to Kevin Bacon
- As of Dec 2007, the highest (finite) Bacon number reported is 8
- Only approx 12% of all actors cannot be linked to Bacon
- What is the Bacon number of Elvis Presley?

Erdos numbers?



The small-world experiment

- 64 chains completed
 - 6.2 average chain length (thus “six degrees of separation”)
- Further observations
 - People that owned the stock had shortest paths to the stockbroker than random people
 - People from Boston area have even closer paths

Measuring the small world phenomenon

- d_{ij} = shortest path between i and j

- Diameter:

$$d = \max_{i,j} d_{ij}$$

- Characteristic path length:

$$l = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- Harmonic mean

$$l^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1}$$

- Also, distribution of all shortest paths

Is the path length enough?

- Random graphs have diameter

$$d = \frac{\log n}{\log z}$$

- $d = \log n / \log \log n$ when $z = \omega(\log n)$
- Short paths should be combined with other properties
 - ease of navigation
 - high clustering coefficient

Degree correlations

- Do high degree nodes tend to link to high degree nodes?
- Pastor Satorras et al.
 - plot the mean degree of the neighbors as a function of the degree

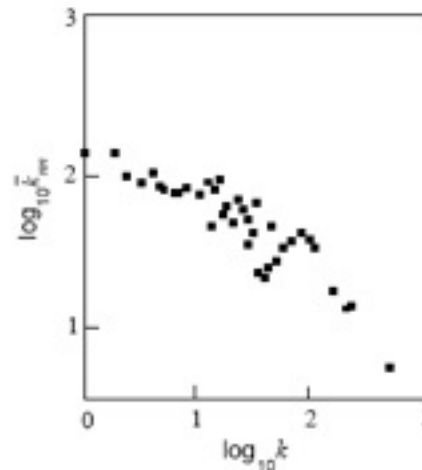


FIG. 3.13. Correlations of the degrees of nearest-neighbour vertices (autonomous systems) in the Internet at the interdomain level (after Pastor-Satorras, Vázquez, and Vespignani 2001). The empirical dependence of the average degree of the nearest neighbours of a vertex on the degree of this vertex is shown in a log-log scale. This empirical dependence was fitted by a power law with exponent approximately 0.5.

Connected components

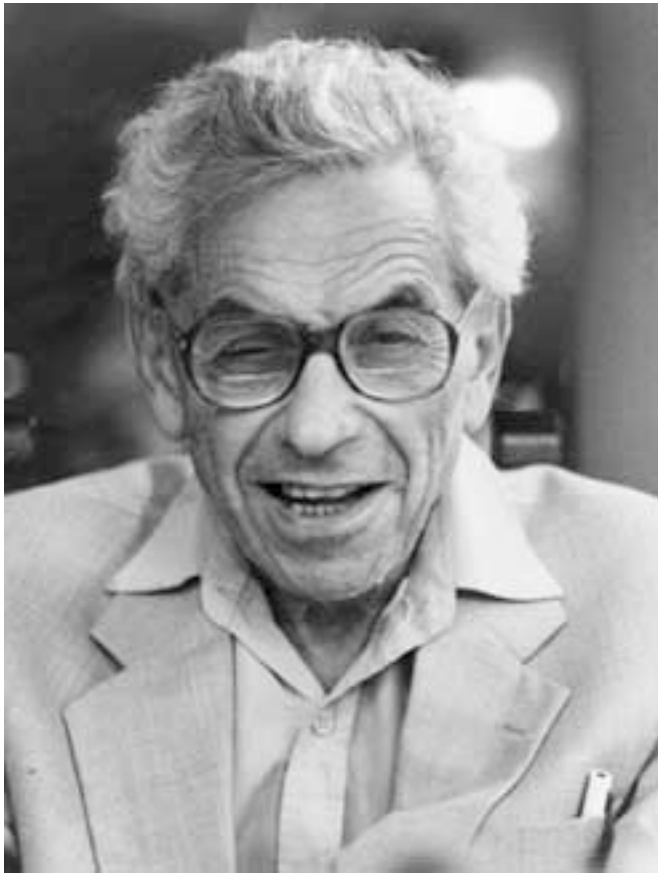
- For undirected graphs, the size and distribution of the connected components
 - is there a **giant component**?
- For directed graphs, the size and distribution of strongly and weakly connected components

Generative models of graphs

What is a network model?

- Informally, a network model is a **process** (randomized or deterministic) for generating a graph
- Models of **static** graphs
 - **input**: a set of parameters Π , and the size of the graph n
 - **output**: a graph $G(\Pi, n)$
- Models of **evolving** graphs
 - **input**: a set of parameters Π , and an initial graph G_0
 - **output**: a graph G_t for each time t

Erdős–Renyi Random graphs



Paul Erdős (1913–1996)

Erdős–Renyi Random Graphs

- The $G_{n,p}$ model
 - **input**: the number of vertices n , and a parameter p , $0 \leq p \leq 1$
 - **process**: for each pair (i,j) , generate the edge (i,j) independently with probability p
- Related, but not identical: The $G_{n,m}$ model
 - **process**: select m edges uniformly at random

The giant component

- Let $z=np$ be the average degree
- If $z < 1$, then almost surely, the largest component has size at most $O(\ln n)$
- if $z > 1$, then almost surely, the largest component has size $\Theta(n)$. The second largest component has size $O(\ln n)$
- if $z = \omega(\ln n)$, then the graph is almost surely connected.

The phase transition

- When $z=1$, there is a phase transition
 - The largest component is $O(n^{2/3})$
 - The sizes of the components follow a power-law distribution.

Random graphs degree distributions

- The degree distribution follows a **binomial**

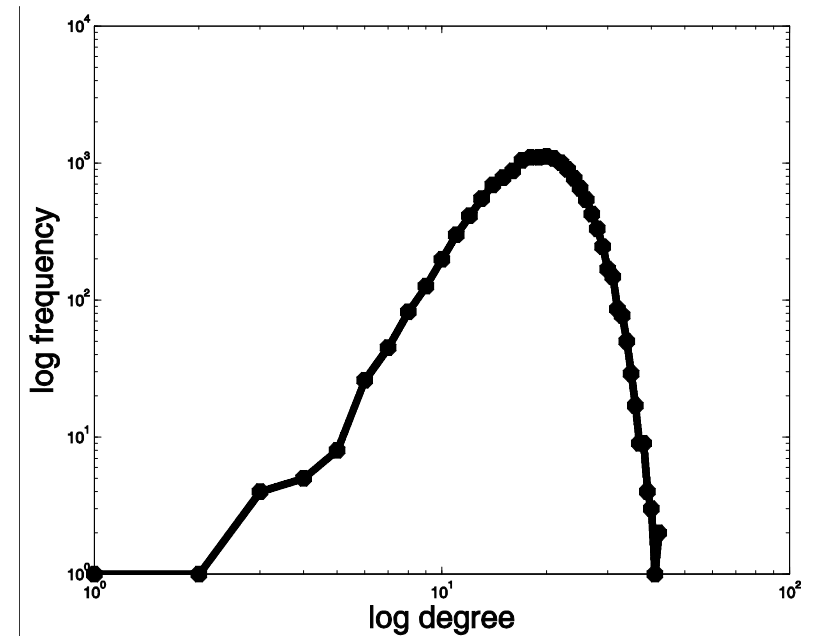
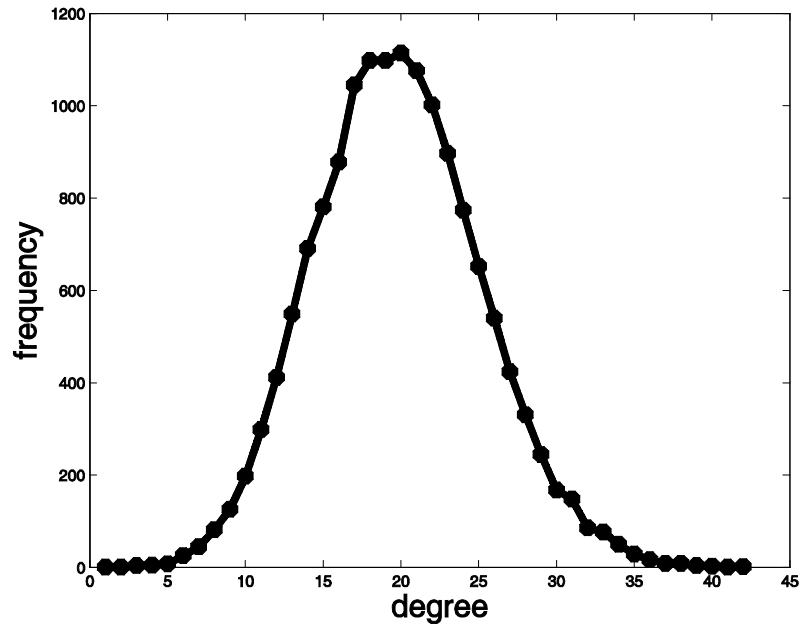
$$p(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- Assuming **$z=np$** is fixed, as **$n \rightarrow \infty$** , **$p(k)$** is approximated by a **Poisson** distribution

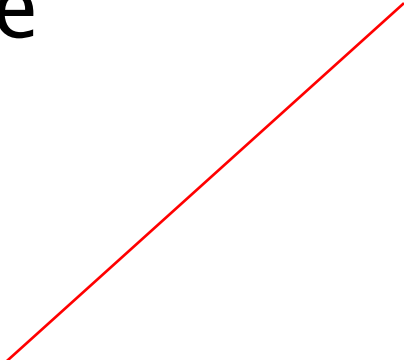
$$p(k) = P(k; z) = \frac{z^k}{k!} e^{-z}$$

- Highly concentrated around the mean, with a tail that drops exponentially

A random graph degree distribution



Random graphs and real life

- A beautiful and elegant theory studied exhaustively
 - Random graphs had been used as idealized network models
 - Unfortunately, they don't capture reality...
- 

Departing from the Random Graph model

- We need models that better capture the characteristics of real graphs
 - degree sequences
 - clustering coefficient
 - short paths

How can we generate data with power-law degree distributions?

Preferential Attachment in Networks

- First considered by [Price 65] as a model for citation networks
 - each new paper is generated with m citations (mean)
 - new papers cite previous papers with probability proportional to their indegree (citations)
 - what about papers without any citations?
 - each paper is considered to have a “default” citation
 - probability of citing a paper with degree k , proportional to $k+1$
- Power law with exponent $\alpha = 2 + 1/m$

Barabasi–Albert model

- The BA model (undirected graph)
 - **input**: some initial subgraph G_0 , and m the number of edges per new node
 - **the process**:
 - nodes arrive one at the time
 - each node connects to m other nodes selecting them with probability proportional to their degree
 - if $[d_1, \dots, d_t]$ is the degree sequence at time t , the node $t+1$ links to node i with probability

$$\frac{d_i}{\sum_i d_i} = \frac{d_i}{2mt}$$

- Results in power-law with exponent $\alpha = 3$

Small world Phenomena

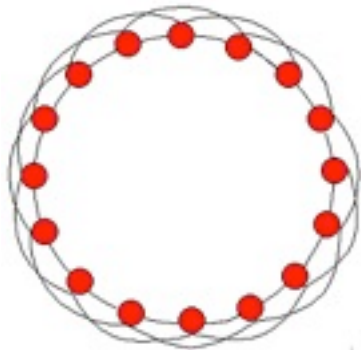
- So far we focused on obtaining graphs with power-law distributions on the degrees. What about other properties?
 - **Clustering coefficient**: real-life networks tend to have high clustering coefficient
 - **Short paths**: real-life networks are “**small worlds**”
 - this property is easy to generate
 - Can we combine these two properties?

Small-world Graphs

- According to Watts [W99]
 - Large networks ($n \gg 1$)
 - Sparse connectivity (avg degree $z \ll n$)
 - No central node ($k_{\max} \ll n$)
 - Large clustering coefficient (larger than in random graphs of same size)
 - Short average paths ($\sim \log n$, close to those of random graphs of the same size)

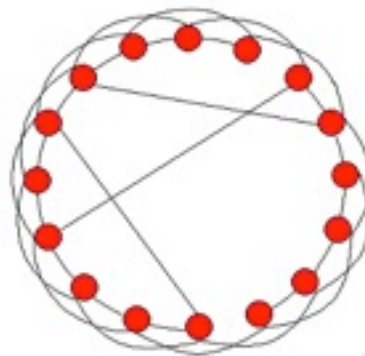
Watts and Strogatz model [WS98]

- Start with a ring, where every node is connected to the next z nodes
- With probability p , **rewire** every edge (or, add a **shortcut**) to a uniformly chosen destination.
 - Granovetter, “The strength of weak ties”

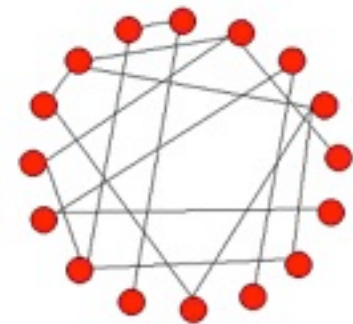


order

$p = 0$



$0 < p < 1$

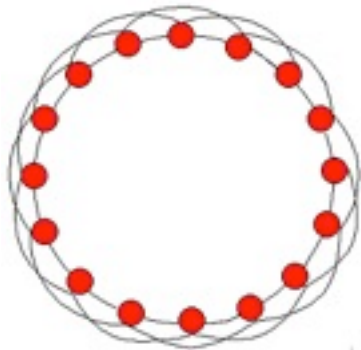


randomness

$p = 1$

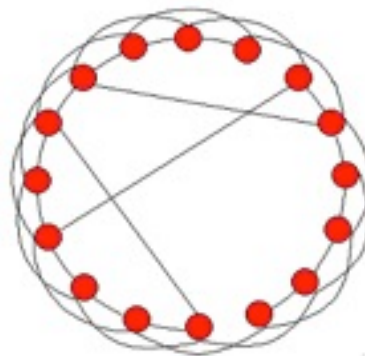
Watts and Strogatz model [WS98]

- Start with a ring, where every node is connected to the next z nodes
- With probability p , **rewire** every edge (or, add a **shortcut**) to a uniformly chosen destination.
 - Granovetter, “The strength of weak ties”

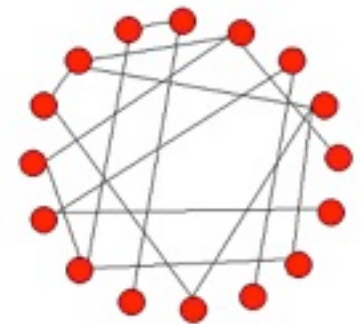


order

$p = 0$



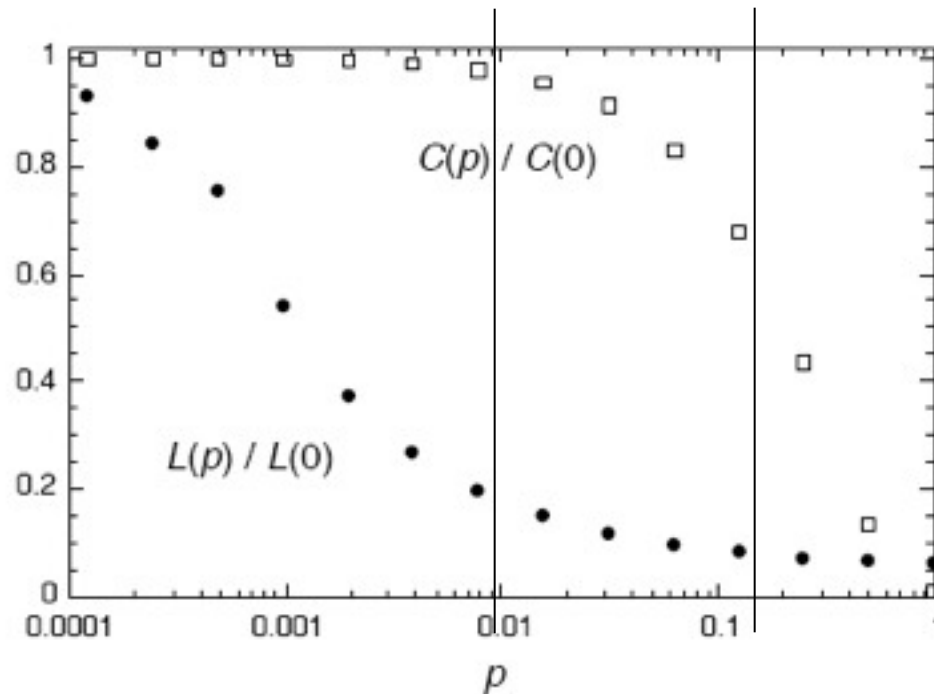
$0 < p < 1$



randomness

$p = 1$

Clustering Coefficient – Characteristic Path Length



log-scale in p

When $p = 0$, $C = 3(k-2)/4(k-1) \sim 3/4$
 $L = n/k$

For small p , $C \sim 3/4$
 $L \sim \log n$