Advanced Topics in Data Mining
Special focus: Social Networks

Goal of the class

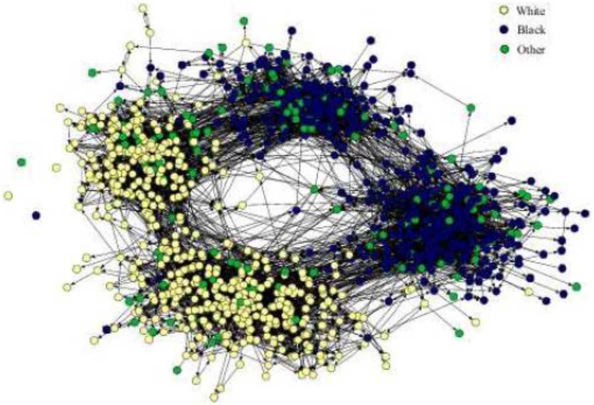
- Address major trends in the analysis of social-network data
- Get you involved and interested
- Do something fun and cool

What is a social network?

- Facebook
- LinkedIn
- ….
- The network of your friends and acquaintances
- Social network is a graph $G=(V,E)$
  - V: set of users
  - E: connections/friendships among users

Social Networks

- Links denote a social interaction
  - Networks of acquaintances
  - collaboration networks
    - actor networks
    - co-authorship networks
    - director networks
  - phone-call networks
  - e-mail networks
  - IM networks
  - Bluetooth networks
  - sexual networks
  - home page/blog networks



Themes in data analysis for social networks

- Measure characteristics of social networks (Measurements)
  - How many hops apart are two random Facebook users
- Design models that capture the generation process of network data (Generative Models)
  - Generate graphs with the same properties as real social network graphs
- Algorithmic problems related to (Algorithmic SN analysis)
  - Information propagation
  - Advertising
  - Expertise finding
  - Privacy

Structure and function of the class

- **Material:** Mostly based on recent papers related to social-network analysis.
  - Some papers and links are already posted on the website of the class
  - Other interesting papers can be found in the proceedings of : KDD, WWW, WSDM, ICDM… conferences

- **Goal:** Understand the material in these papers and (hopefully) extend it

Structure and function of the class

- Introductory lectures
- Paper presentations (20%)
- Projects and Project Presentation (50%)
- Project Report (otherwise called reaction paper) (20%)
- Class Participation (10%)

Introductory Lectures

- Measurements in networks

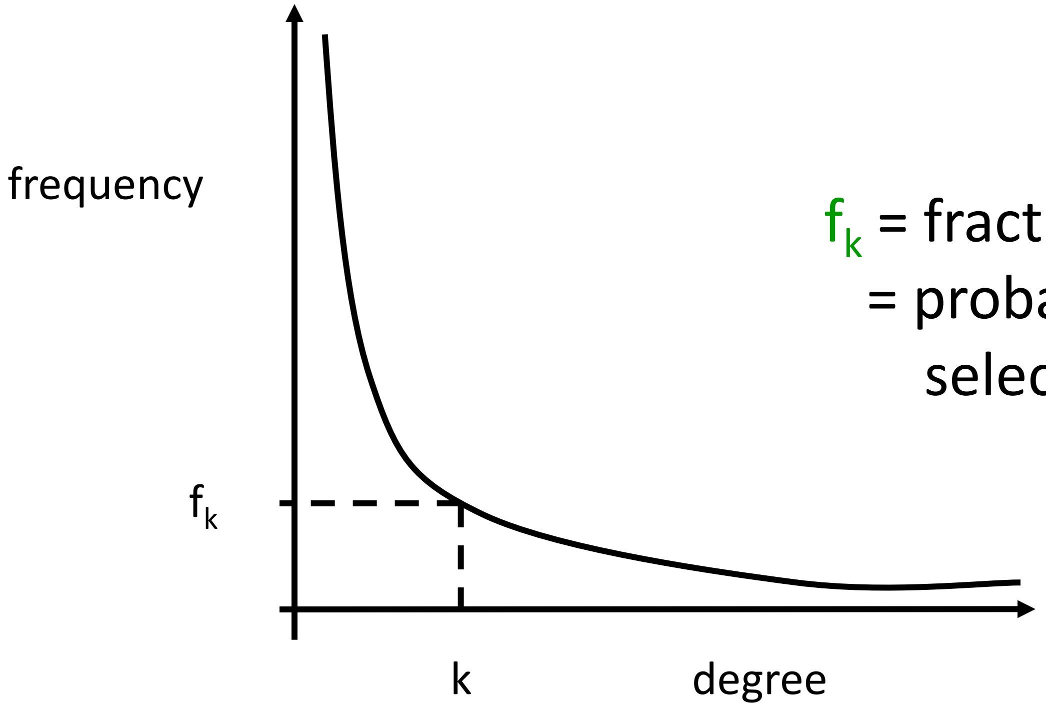
- Generative models

- Algorithmic topics
  - Introduction to information propagation
  - Expertise location
  - Privacy

Measuring Networks

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

Degree distributions



f_k = fraction of nodes with degree k
= probability of a randomly selected node to have degree k

- 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

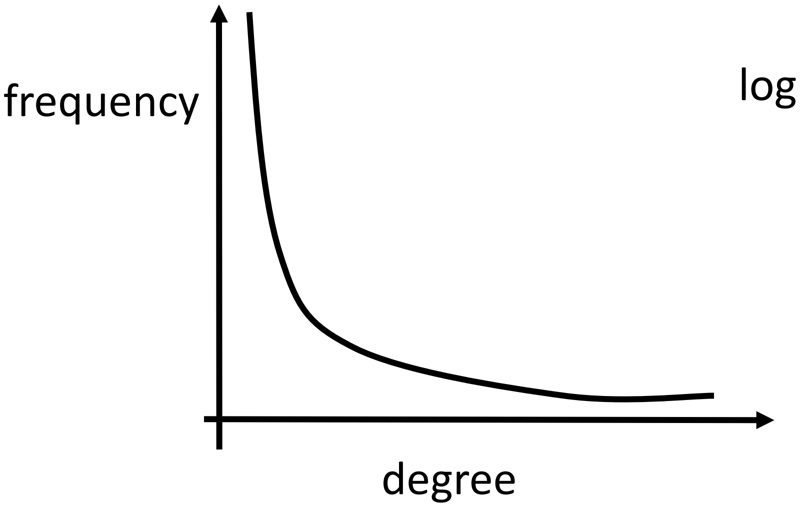
$$p(k) = P(k; z) = \frac{z^k}{k!} e^{-z}$$

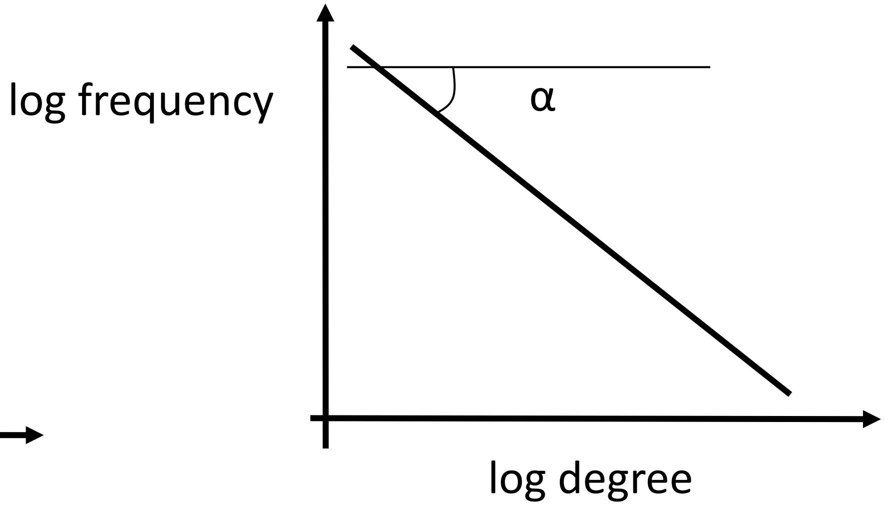
  - highly concentrated around the mean
  - the probability of very high degree nodes is exponentially small

Power-law signature

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

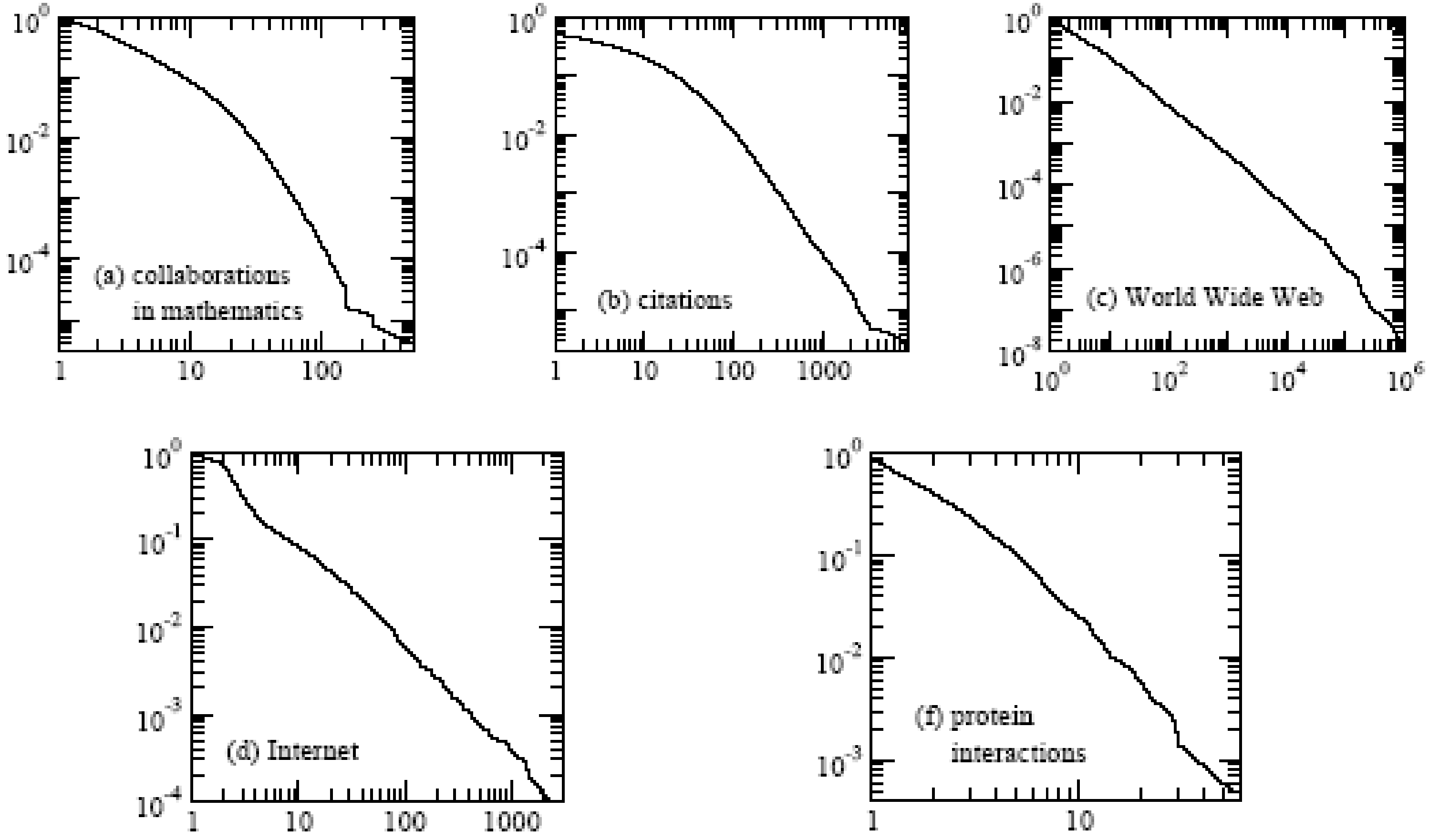
$$\log p(k) = -\alpha \log k + \log C$$





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

Examples



Taken from [Newman 2003]

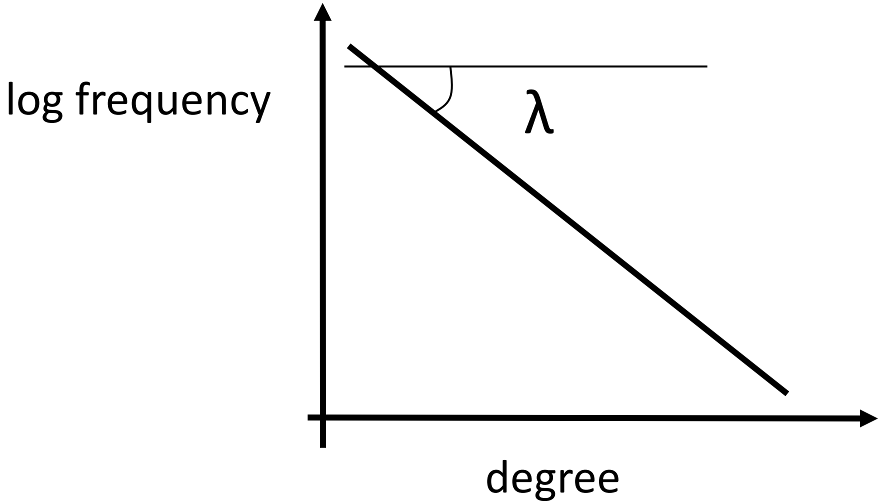
Exponential distribution

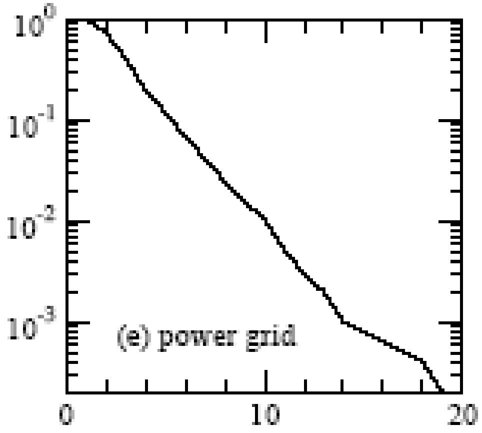
- Observed in some technological or collaboration networks

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

- Identified by a line in the log-linear plot

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



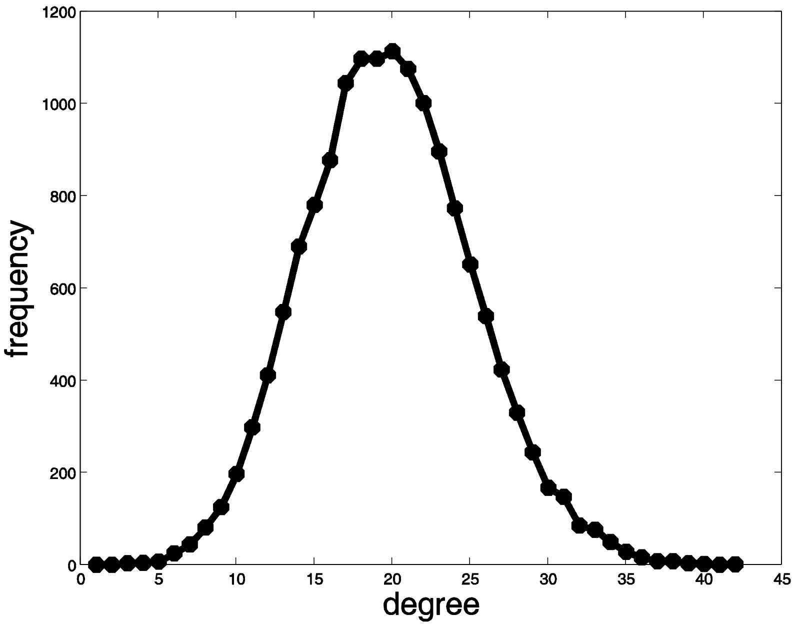


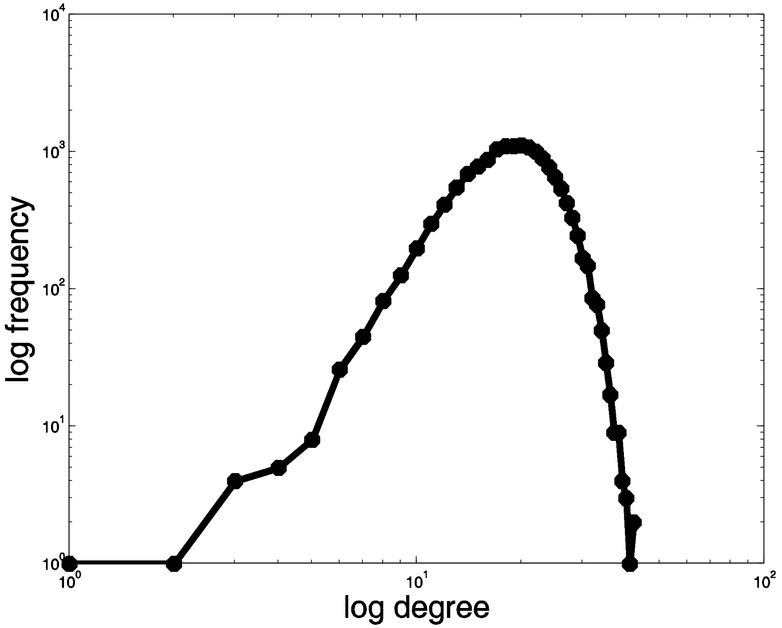
The basic random graph model

- The measurements on real networks are usually compared against those on “random networks”

- The basic $G_{n,p}$ (Erdös-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

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

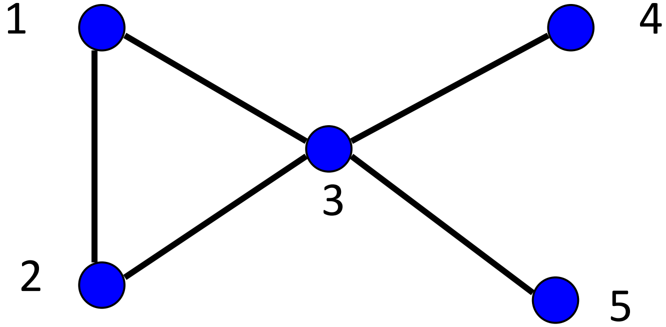
Clustering (Transitivity) coefficient

- 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

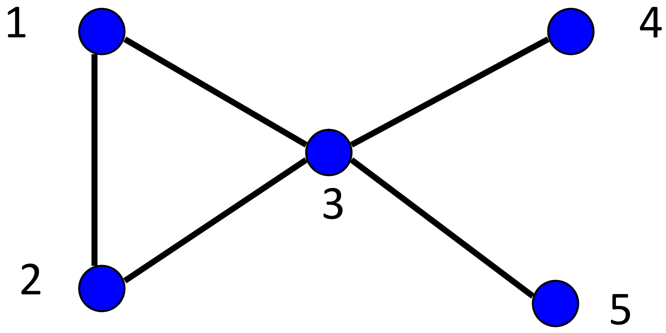
- Clustering coefficient for node i

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

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

- The mean of the ratios

Example



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

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

- 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)

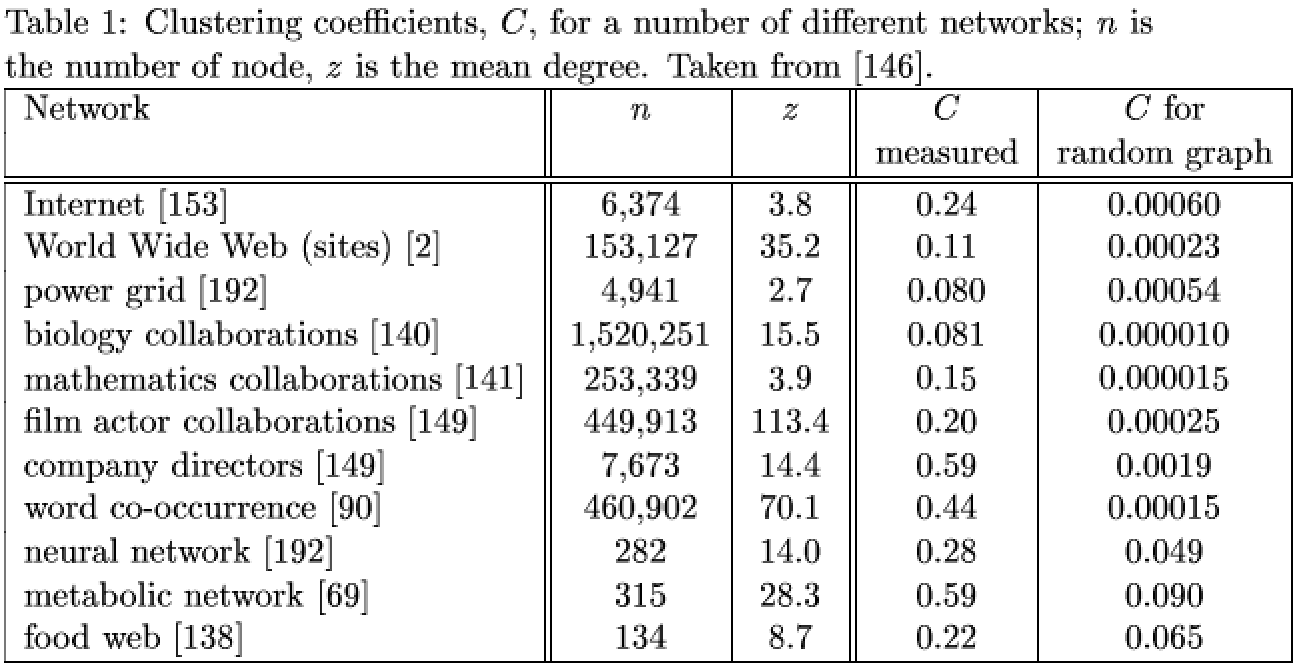


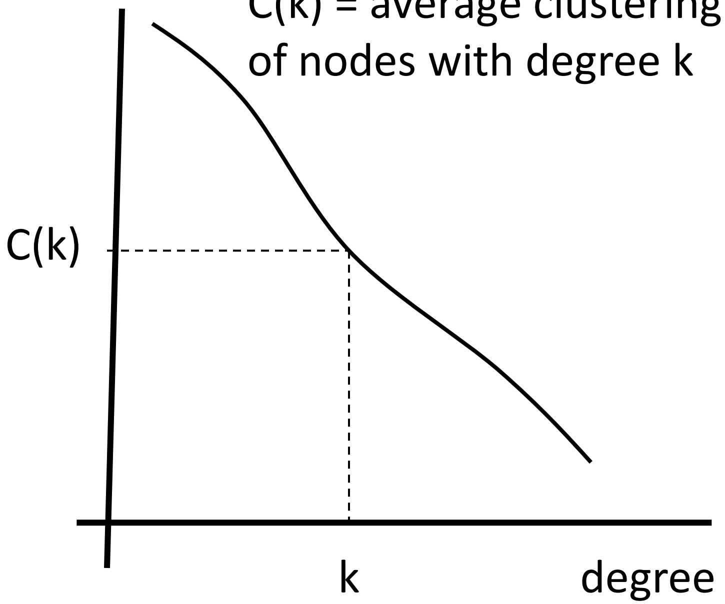
Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

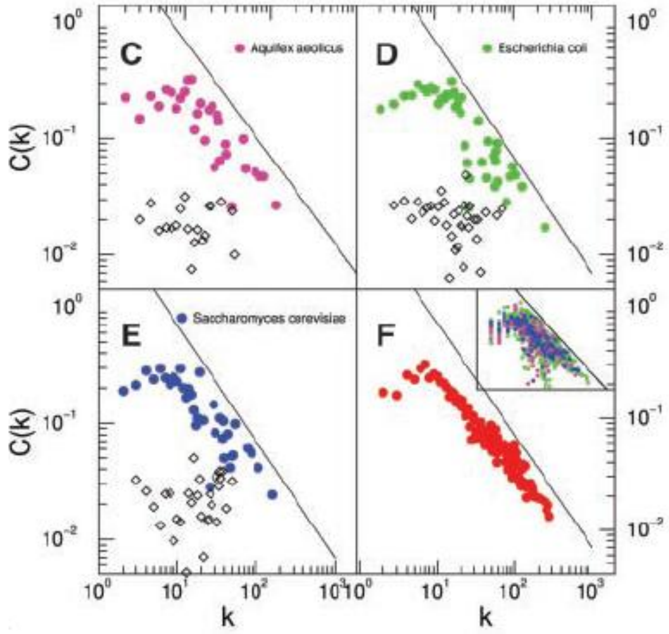
Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor 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

C(k) = average clustering coefficient of nodes with degree k





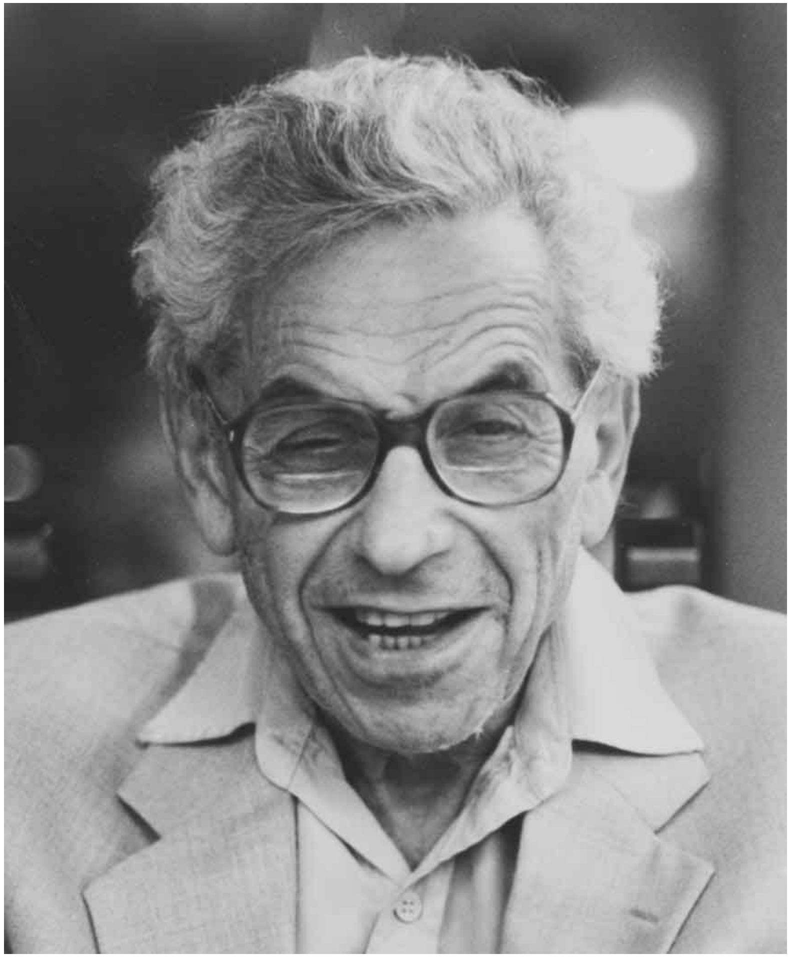
The small-world experiment

- 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 Prisley?

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:

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

- Harmonic mean

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

- 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 Satoras et al.
  - plot the mean degree of the neighbors as a function of the degree

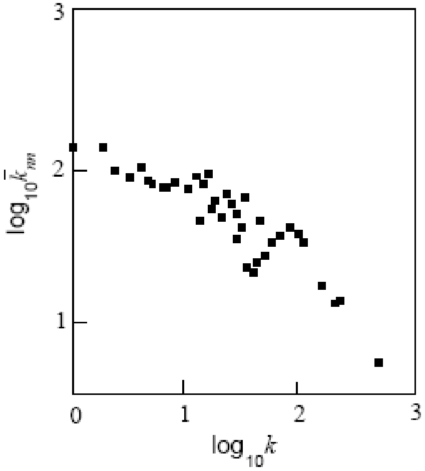


FIG. 3.13. Correlations of the degrees of nearest-neighbour vertices (autonomous systems) in the Internet at the interdomain level (after Pastor-Satorras, Vázquez, and Vespignani 2001). The empirical dependence of the average degree of the nearest neighbours of a vertex on the degree of this vertex is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

Degree correlations

- Newman
  - compute the correlation coefficient of the degrees of the two endpoints of an edge
  - assortative/disassortative

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2},$$

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

Graph eigenvalues

- For random graphs
  - semi-circle law

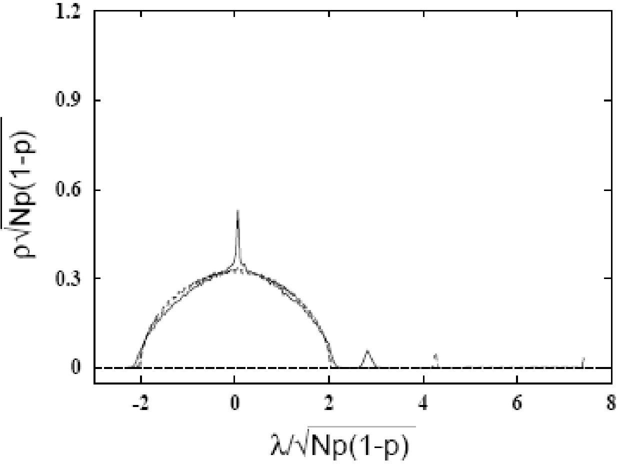
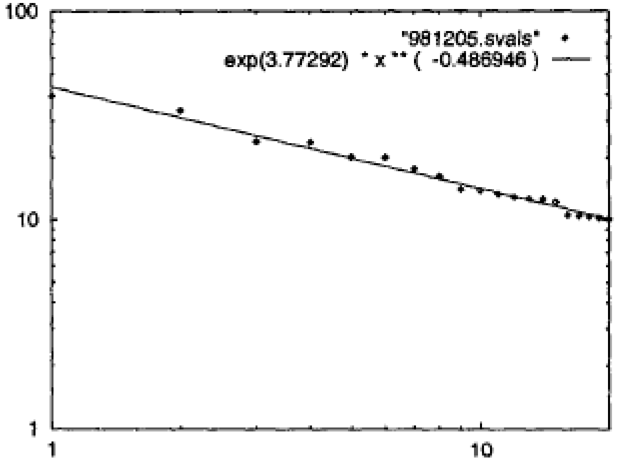


FIG. 10. Rescaled spectral density of three random graphs having $p = 0.05$ and size $N = 100$ (continuous line), $N = 300$ (dashed line) and $N = 1000$ (short-dashed line). The isolated peak corresponds to the principal eigenvalue. After Farkas *et al.* 2001.

- For the Internet (Faloutsos[3])



(a) Int-12-98

Next class

- What is a good model that generates graphs in which power law degree distribution appears?

- What is a good model that generates graphs in which small-world phenomena appear?