Advanced Topics in Data Mining
Special focus: Social Networks

Reminders

- By the end of this week/ beginning of next we need to have a tentative presentation schedule
- Each one of you should send me an email about a theme by Friday, February 22.

What did we learn in the last lecture?

What did we learn in the last lecture?

- Degree distribution
  - What are the observed degree distributions
- Clustering coefficient
  - What are the observed clustering coefficients?
- Average path length
  - What are the observed average path lengths?

What are we going to learn in this lecture?

- How to generate graphs that have the desired properties
  - Degree distribution
  - Clustering coefficient
  - Average path length

- We are going to talk about **generative models**

What is a network model?

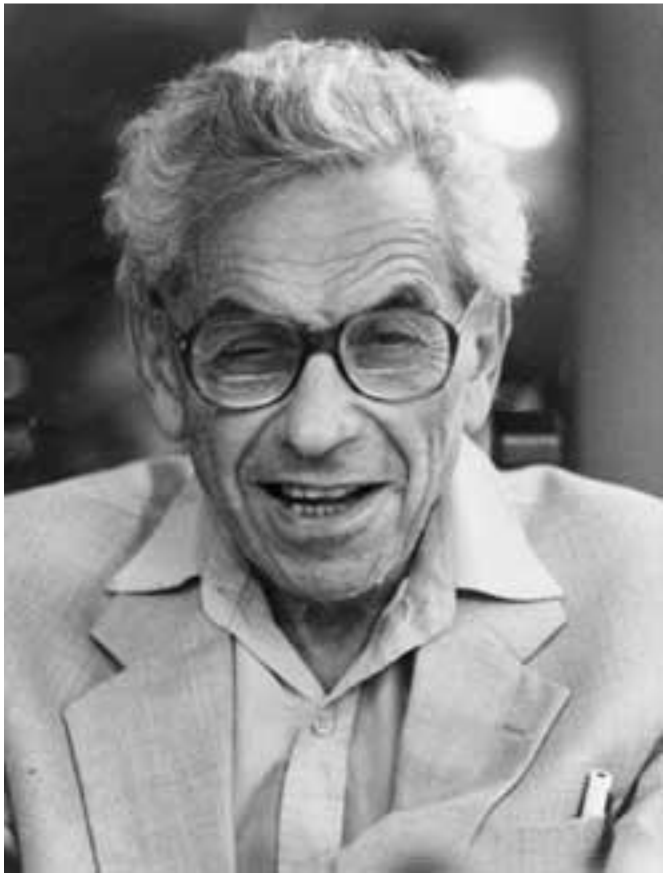
- Informally, a network model is a process (radomized or deterministic) for generating a graph
- Models of static graphs
  - input: a set of parameters Π, and the size of the graph n
  - output: a graph G(Π,n)
- Models of evolving graphs
  - input: a set of parameters Π, and an initial graph G_0
  -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

Families of random graphs

- A deterministic model D defines a single graph for each value of n (or t)

- A randomized model R defines a probability space $\langle G_n, P \rangle$ where G_n is the set of all graphs of size n, and P a probability distribution over the set G_n (similarly for t)
  - we call this a family of random graphs R, or a random graph R

Erdös-Renyi Random graphs



Paul Erdös (1913-1996)

Erdös-Renyi Random Graphs

- The $G_{n,p}$ model
  - input: the number of vertices n, and a parameter p, $0 \le p \le 1$
  - 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

Graph properties

- A property P holds almost surely (or for almost every graph), if

$$\lim_{n\to\infty} P[G \text{ has } P] = 1$$

- Evolution of the graph: which properties hold as the probability p increases?

- Threshold phenomena: Many properties appear suddenly. That is, there exist a probability p_c such that for $p<p_c$ the property does not hold a.s. and for $p>p_c$ the property holds a.s.

  - ***What do you expect to be a threshold phenomenon in random graphs?***

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

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

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

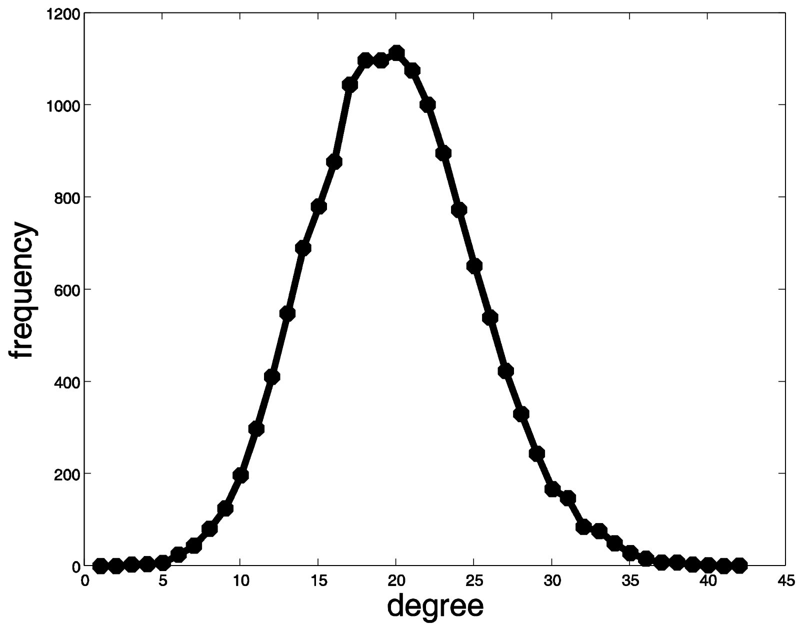
$$p(k) = P(k;z) = \frac{z^k}{k!} e^{-z}$$

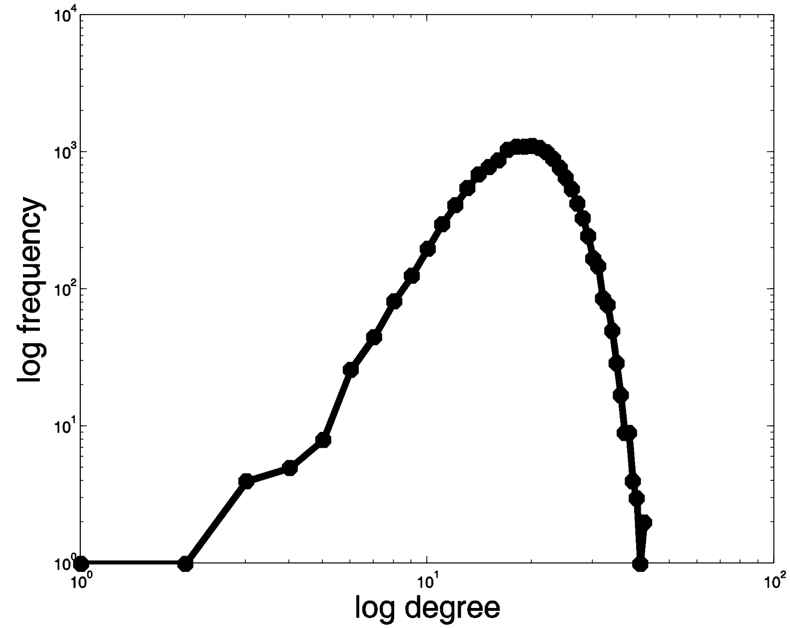
- Highly concentrated around the mean, with a tail that drops exponentially

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…

A random graph example





Departing from the Random Graph model

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

Graphs with given degree sequences

- **input:** the degree sequence **[d_1,d_2,…,d_n]**
- Can you generate a graph with nodes that have degrees **[d_1,d_2,…,d_n] ?**
- **?** ☺

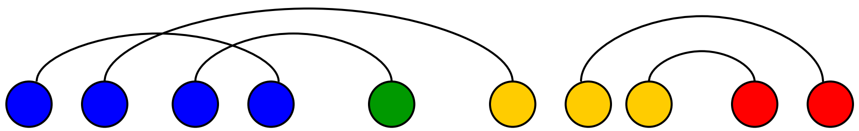
Graphs with given degree sequences

- The configuration model
  - **input:** the degree sequence $[d_1,d_2,\ldots,d_n]$
  - **process:**
    - Create d_i copies of node i
    - Take a random matching (pairing) of the copies
      - self-loops and multiple edges are allowed
- Uniform distribution over the graphs with the given degree sequence

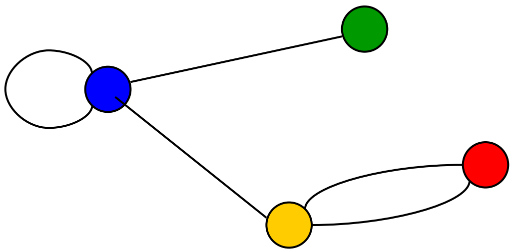
Example

- Suppose that the degree sequence is

- Create multiple copies of the nodes



- Pair the nodes uniformly at random
- Generate the resulting network



Graphs with given degree sequences

- How about **simple** graphs ?
  - No self loops
  - No multiple edges

Graphs with given degree sequences

- Realizability of degree sequences
- **Lemma:** A degree sequence **d = [d(1),…,d(n)]** with **d(1)≥d(2)≥… ≥d(n)** and **d(1)+d(2)+…+d(n)** **even** is **realizable** if and only if for every **1≤k ≤n-1** it holds that

$$\sum_{i=1}^{k} d(i) \leq k(k-1) + \sum_{i=k+1}^{n} \min\{k, d(i)\}$$

Graphs with given degree sequences -- algorithm

- Input : **d= [d(1),…,d(n)]**
- Output: No or simple graph **G=(V,E)** with degree sequence **d**
- If $\Sigma_{i=1\ldots n}$ **d(i)** is odd return "**No**"
- While 1 do
  - If there exist i with **d(i) < 0** return "**No**"
  - If **d(i)=0** for all i return the graph **G=(V,E)**
  - Pick random node **v** with **d(v)>0**
  - **S(v) =** set of nodes with the **d(v)** highest **d** values
  - **d(v) = 0**
  - For each node **w** in **S(v)**
    - **E = E\union (v,w)**
    - **d(w) = d(w)-1**

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 $[d_1,\ldots,d_t]$ 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 $\alpha = 3$

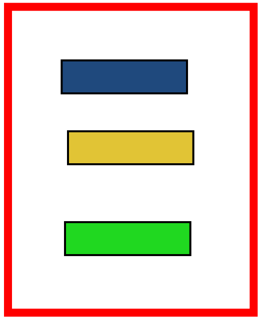
Variations of the BA model

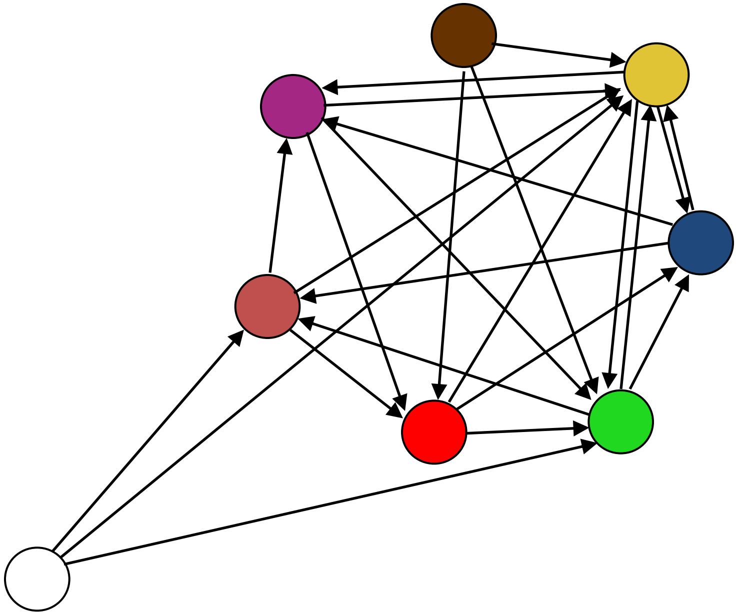
- Many variations have been considered

Copying model

- Input:
  - the out-degree d (constant) of each node
  - a parameter α
- The process:
  - Nodes arrive one at the time
  - A new node selects uniformly one of the existing nodes as a prototype
  - The new node creates d outgoing links. For the i^{th} link
    - with probability α it copies the i-th link of the prototype node
    - with probability 1- α it selects the target of the link uniformly at random

An example





Copying model properties

- Power law degree distribution with exponent $\beta = (2-\alpha)/(1-\alpha)$
- Number of bipartite cliques of size $i \times d$ is ne^{-i}

- The model has also found applications in biological networks
  - copying mechanism in gene mutations

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 ($n >> 1$)
  - Sparse connectivity (avg degree $z << n$)
  - No central node ($k_{max} << n$)
  - Large clustering coefficient (larger than in random graphs of same size)
  - Short average paths (~$\log n$, close to those of random graphs of the same size)

Mixing order with randomness

- Inspired by the work of Solmonoff and Rapoport
  - nodes that share neighbors should have higher probability to be connected
- Generate an edge between i and j with probability proportional to R_{ij}

$$R_{ij} = \begin{cases} 1 & \text{if } m_{ij} \geq z \\ \left(\frac{m_{ij}}{z}\right)^{\alpha}(1-p)+p & \text{if } 0 < m_{ij} < z \\ p & \text{if } m_{ij} = 0 \end{cases}$$

m_{ij} = number of common neighbors of i and j

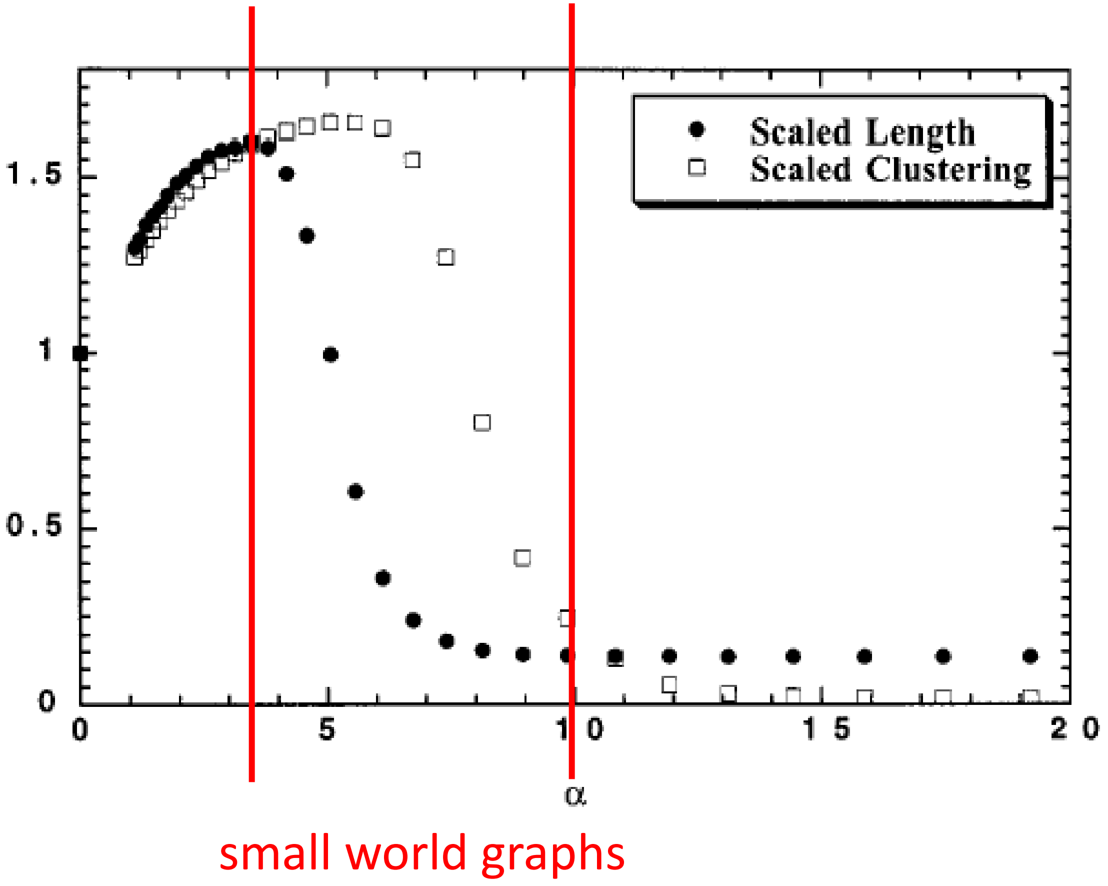
p = very small probability

- When $\alpha = 0$, edges are determined by common neighbors
- When $\alpha = \infty$ edges are independent of common neighbors
- For intermediate values we obtain a combination of order and randomness

Algorithm

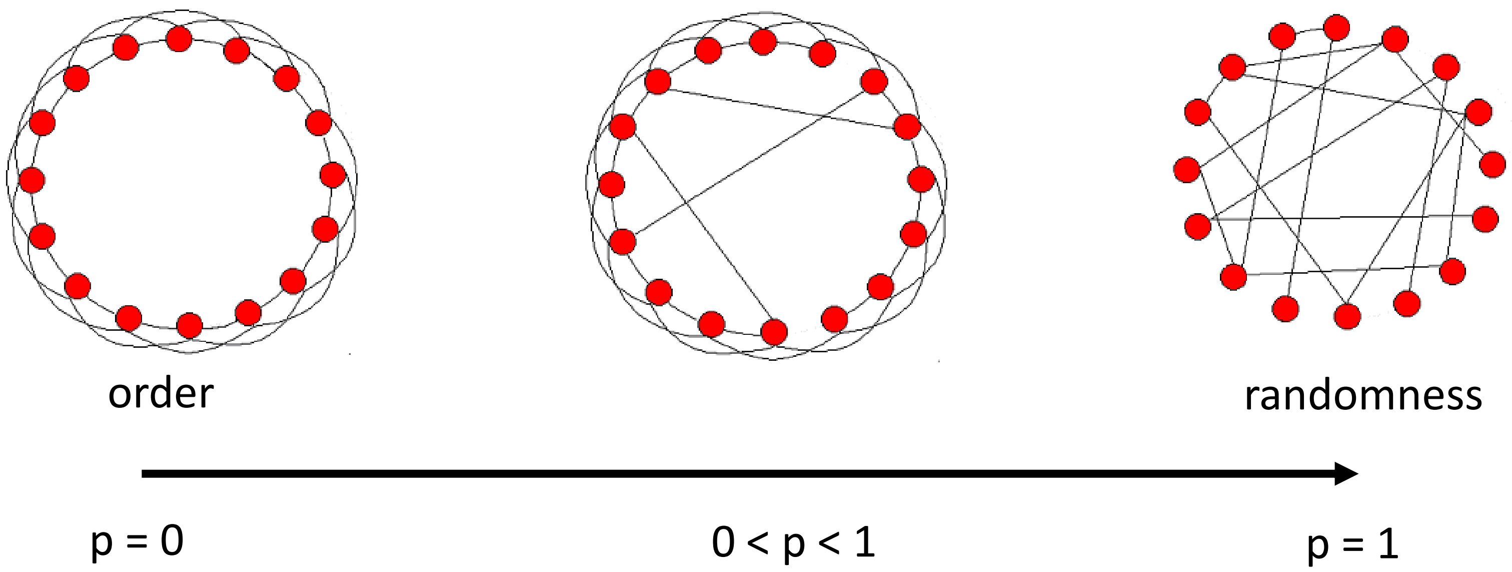
- Start with a ring
- For $i = 1 \ldots n$
  - Select a vertex j with probability proportional to R_{ij} and generate an edge (i,j)
- Repeat until z edges are added to each vertex

Clustering coefficient – Avg path length



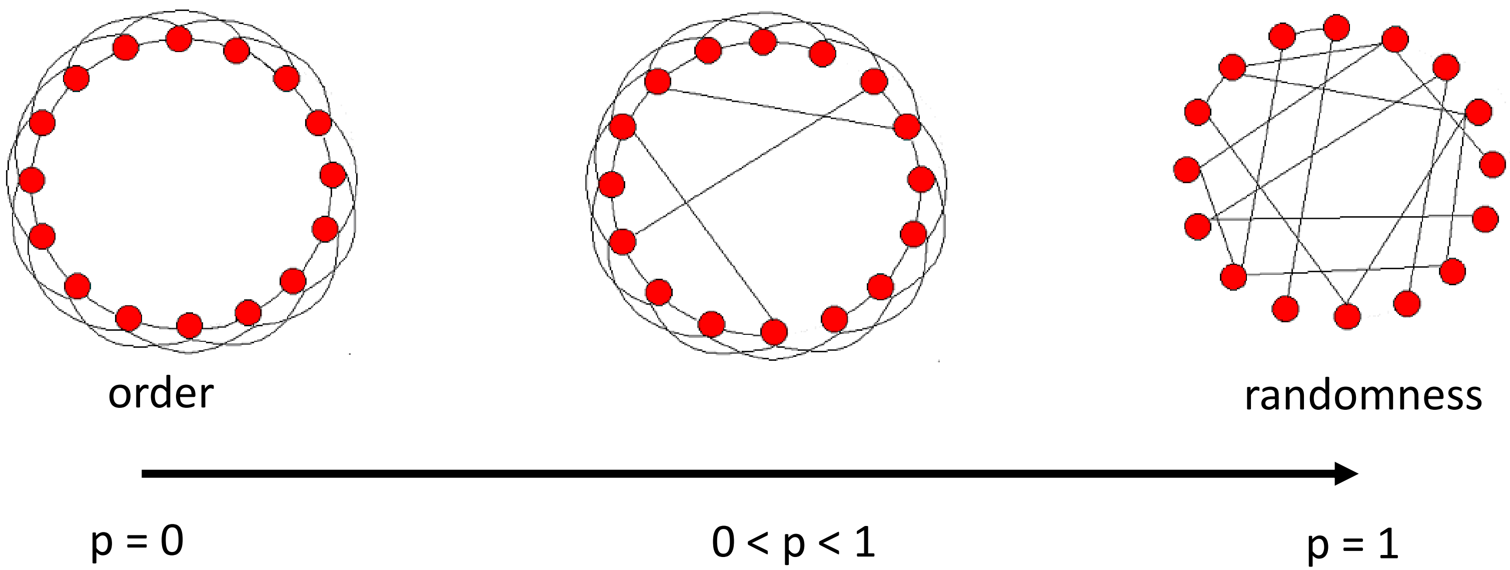
Watts and Strogatz model [WS98]

- 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"

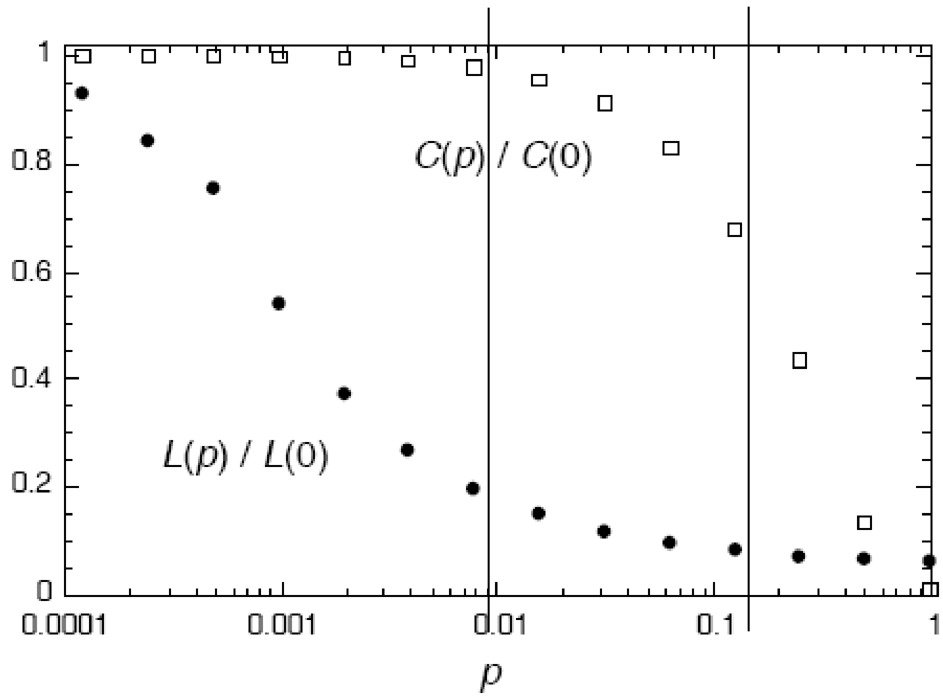


Watts and Strogatz model [WS98]

- 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”



Clustering Coefficient – Characteristic Path Length



log-scale in p

When p = 0, C = 3(k-2)/4(k-1) ~ ¾
L = n/k

For small p, C ~ ¾
L ~ logn

Next Class

- Some more generative models for social-network graphs