4.10 Modeling (Why?)

Data models are grouped into two major types: descriptive models and process models. *Descriptive models* aim to illustrate the major features of a (typically static) data set, such as statistical patterns of article citation counts, networks of citations, individual differences in citation practice, the composition of knowledge domains, or the identification of research fronts as indicated by new yet highly cited papers. *Process models* or predictive models aim to simulate, statistically describe, or formally reproduce statistical and dynamic characteristics of interest.

4.10.1 Random Graph Model

The random graph model generates a graph that has a fixed number of nodes which are connected randomly by undirected edges, see Figure 4.17 (left). The number of edges depends on a specified probability. The edge probability is chosen based on the number of nodes in the graph. The model most commonly used for this purpose was introduced by Gilbert (Gilbert, 1959). This is known as the *G(n,p)* model with *n* being the number of vertices and *p* the linking probability. The number of edges created according to this model is not known in advance. Erd?s-Rényi introduced a similar model where all the graphs with *m* edges are equally probable and *m* varies between *0* and *n(n-1)/2* (Erd?s & Rényi, 1959). This is known as the *G(n,m)* model. The degree distribution for this network is Poissonian, see Figure 4.17 (right).

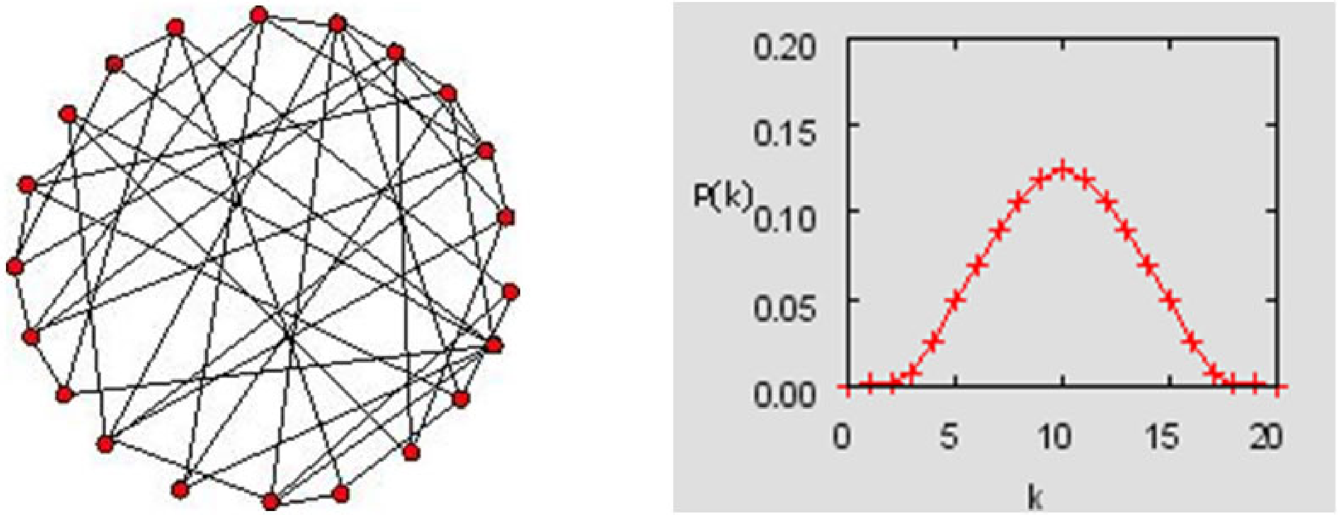
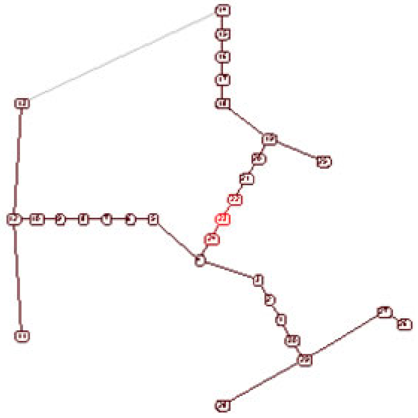


Figure 4.17: Random graph and its Poissonian node degree distribution

Very few real world networks are random. However, random networks are a theoretical construct that is well understood and their properties can be exactly solved. They are commonly used as a reference, e.g., in tests of network robustness and epidemic spreading (Batagelj & Brandes).
In the Sci2 Tool, the random graph generator implements the *G(n,p)* model by Gilbert. Run *'Modeling > Random Graph'* and input the total number of nodes in the network and their wiring probability. The output is a network in which each pair of nodes is connected by an undirected edge with the probability specified in the input.
A wiring probability of 0 would generate a network without any edges and a wiring probability of 1 with *n* nodes will generate a network with *(n-1)* edges. The wiring probability should be chosen dependent on the number of vertices. For a large number of vertices the wiring probability should be smaller.

4.10.2 Watts-Strogatz Small World

A small-world network is one whose majority of nodes are not directly connected to one another, but still can reach any other node via very few edges. It can be used to generate networks of any size. The degree distribution is almost Poissonian for any value of the rewiring probability (except in the extreme case of rewiring probability zero, for which all nodes have equal degree). The clustering coefficient is high until beta is close to 1, and as beta approaches one, the distribution becomes Poissonian. This is because the graph becomes increasingly similar to an Erd?s-Rényi Random Graph, see Figure 4.18. (Watts & Strogatz, 1998; Inc. Wikimedia Foundation, 2009).



$$P(k) = \sum_{n=0}^{f(k,K)} C^{n}_{K/2} \left(1-\beta\right)^{n} \beta^{K/2-n} \frac{(\beta K/2)^{k-K/2-n}}{(k-K/2-n)!} e^{-\beta K/2}$$

Figure 4.18: Small world graph (left) and its node degree distribution equation (right)

Small world properties are usually studied to explore networks with tunable values for the average shortest path between pairs of nodes and a high clustering coefficient. Networks with small values for the average shortest path and large values for the clustering coefficient can be used to simulate social networks, unlike ER random graphs, which have small average shortest path lengths, but low clustering coefficients.

The algorithm requires four inputs: the number *n* of nodes of the network, the number *k* of initial neighbors of each node (the initial configuration is a ring of nodes), the probability of rewiring the edges (which is a real number between 0 and 1), and the seed of a random number generator. The network is built following the original prescription of Watts and Strogatz, i.e., by starting from a ring of nodes each connected to the k nodes and by rewiring each edge with the specified probability. The algorithm run time is *O(kn).*

Run with *'Modeling > Watts-Strogatz Small World'* and input 1000 nodes, 10 initial neighbors, and a rewiring probability of 0.01 then compute the average shortest path and the clustering coefficient and verify that the former is small and the latter is relatively large.

4.10.3 Barabási-Albert Scale Free Model

The Barabási-Albert (BA) model is an algorithm which generates a scale-free network by incorporating growth and preferential attachment. Starting with an initial network of a few nodes, a new node is added at each time step. Older nodes with a higher degree have a higher probability of attracting edges from new nodes. The probability of attachment is given by

$$P(k_i)=\frac{k_i}{\Sigma_j k_j}$$

.

The initial number of nodes in the network must be greater than two and each of these nodes must have at least one connection. The final structure of the network does not depend on the initial number of nodes in the network. The degree distribution of the generated network is a power law with a scaling coefficient of *-3* (Barabási & Albert, 1999; Barabási & Albert, 2002). Figure 4.19 shows the network on the left and the probability distribution on a log-log scale on the right.

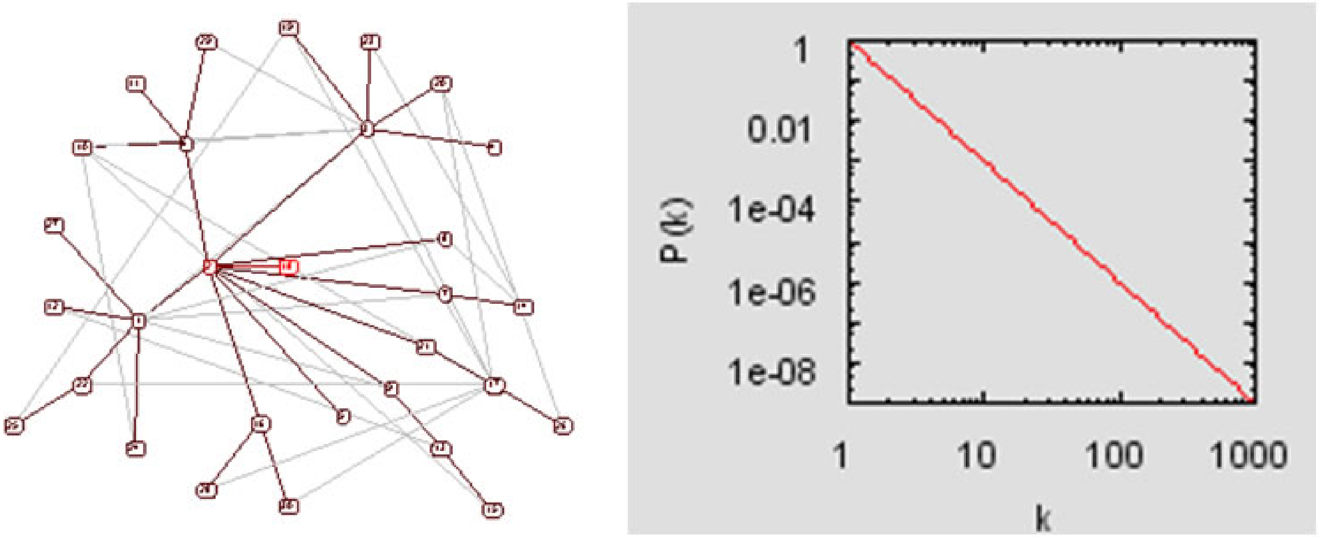


Figure 4.19: Scale free graph (left) and its node degree distribution (right)

This is the simplest known algorithm to generate a scale-free network. It can be applied to model undirected networks such as the collaboration network among scientists, the movie actor network, and other social networks where the connections between the nodes are undirected. However, it cannot be used to generate a directed network.

The inputs for the algorithm are the number of initial nodes, the number of initial edges for a new node, and the seed of a random number generator. The algorithm starts with the initial number of nodes that are fully connected. At each time step, a new node is generated with the initial number of edges. The probability of attaching to an existing node is calculated by dividing the degree of an existing node by the total number of edges. If this probability is greater than zero and greater than the random number obtained from a random number generator then an edge is attached between the two nodes. This is repeated in each time step.

Run with *'Modeling > Barabási-Albert Scale Free Model'* and a time step of around *1000*, initial number of nodes *2*, and number of edges *1* in the input. Lay out and determine the number and degree of highly connected nodes via *'Analysis > Unweighted and Undirected > Degree Distribution'* using the default value. Plot node degree distribution using Gnuplot.