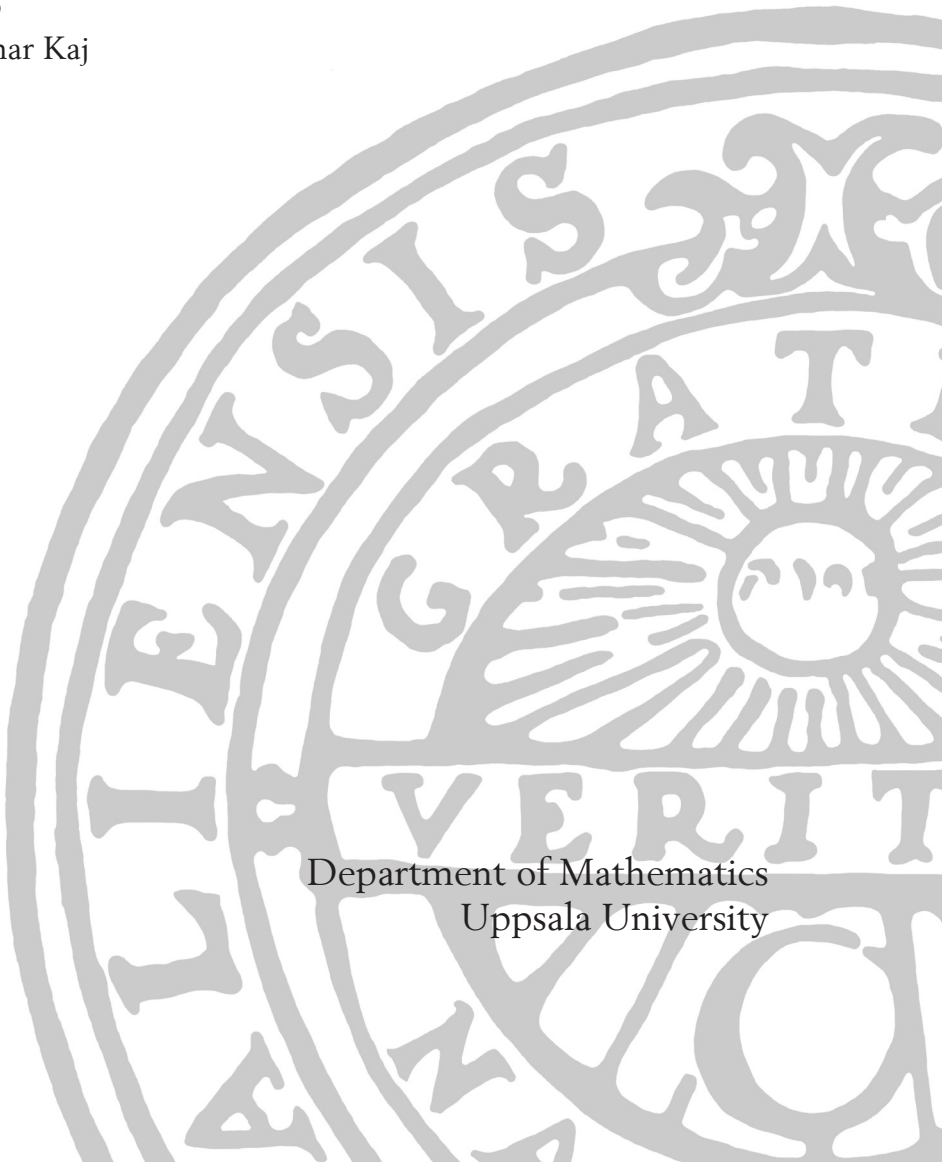
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2012:12

Random graphs as models of Peer-to-Peer social networks

Janne Räisänen

Examensarbete i matematik, 15 hp
Handledare och examinator: Ingemar Kaj
Juni 2012



Department of Mathematics
Uppsala University

Abstract

This thesis will present a study of using different types of random graphs to model real-world networks. Three random graph models are examined with various properties, such as growing graphs and preferential attachment. The resulting degree distribution for each model is investigated, and some numerical simulations are done on Matlab to study the average values of degrees, shortest path lengths and clustering coefficients. These results will be then compared between the models and corresponding data collected from observations done on social connections between peers in a Peer-to-Peer network.

Sammanfattning

Rapporten kommer att presentera en studie av olika typer av slumpgrafer som modeller för riktiga nätverk. Tre olika slumpgraf modeller kommer att studeras med egenskaper som växande grafer och preferentiell kantsannolikhet. Den resulterande gradfördelningen undersöks hos varje modell, och några numeriska simuleringar utförs på Matlab för att studera medelvärden av grader, distanser mellan noder och klustringskoefficienter. Slutligen kommer dessa resultat att jämföras mellan modeller och motsvarande data från observationer som gjorts på sociala kontakter mellan peers i ett Peer-to-Peer nätverk.

Acknowledgements

I would like to thank my supervisor Ingemar Kaj for all the help and guidance with this thesis work.

Contents

1 Introduction

Understanding the structure of a real-world network and studying its evolution has been an object of great interest for scientists in various fields. A widely studied network is the world-wide web, where web pages that consists of text and pictures are linked together via hyperlinks that allow users to navigate from one page to another. Sociologists are often investigating networks that represent various social interactions between people or groups of people. Neurological networks that represent patterns of connections between brain cells, and power grids consisting of generating stations connected by high-voltage lines are other examples of networks.

Often one is interested in certain topological properties of networks, such as the number of social connections that an individual has in a given society, the smallest possible number of hyperlinks that exists between two web pages, or the number of communities that a society forms. Sometimes the evolution of a network is studied, to see whether the network during its generation converges to some particular state for a large number of nodes. However, many real networks are quite large, and in most cases it is very difficult to describe them analytically. To avoid this problem, complex topology, uncertainty and the lack of regularity in real networks have been described with simple random graph models. With recent improvements in computational power and with the aid of numerical analysis, such topological information is increasingly available in more complex random graph models as well. Therefore, large data sets can now easily be stored and studied, and this has had a significant impact in the empirical studies on large networks.

Studies of the random graph models have shown results similar to the characteristics of many real networks. Some models have shown that the probability of a node in a network interacts with k other nodes decays as a power law. This result indicates that large networks self-organize into a scale-free state. Measurements have also shown that many real networks show a fundamental property called the small-world effect, which means that typical distances between nodes are small. Another common property of real networks is that clustering form, representing circles of individuals in which every member knows every other member.

This thesis will introduce a few random graph models with different properties. Each model will be implemented in Matlab and a number of numerical simulations will be done in order to estimate the average values of these properties. Results from the simulations will be then compared with corresponding data collected from a Peer-to-Peer social network.

2 Definitions

The purpose of this chapter is to introduce some probability distributions and basic concepts in graph theory. The knowledge of these probability distributions will be useful in the next chapter where degree distributions of random graph models will be studied. The main objective of introducing graph theory is to explain how to evaluate degrees of vertices, shortest path lengths between a pair of vertices and clustering coefficients. These three properties will be then used in the last chapter to analyse and compare with data collected from a real network.

2.1 Probability distributions

A discrete random variable $X \sim Bin(n, p)$ have a binomial distribution with parameters n and p if the probability function is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{1}$$

for some $k = 0, \ldots, n$. The binomial distribution can be used to determine the probability of having a given number of successes k in a sequence of n independent trials with two possible outcomes. The outcome is successful with probability p and failure occurs with probability $1 - p$. The expected value is $E(X) = np$.

A discrete random variable $Y \sim Po(\lambda)$ have a Poisson distribution with a parameter $\lambda > 0$ if the probability function is given by

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{2}$$

for some $k = 0, 1, 2, \ldots$. The Poisson distribution can be used when the events occur randomly in time and the number of events are viewed in a time interval. When a binomially distributed random variable have a small enough p value, this distribution can be used as an approximation. The expected value is $E(Y) = \lambda$.

A discrete random variable $Z \sim Geo(p)$ have a geometric distribution with a parameter p if the probability function is given by

$$P(Z = k) = p(1-p)^k, \tag{3}$$

for some $k = 0, 1, 2, \ldots$. The Geometric distribution can be used to determine the probability of having a number of failures k before getting a single success. Each trial has a success probability p. The expected value is $E(Z) = (1-p)/p$.

A random variable X have a power law distribution if for some constant $c > 0$ and exponent $\gamma > 0$ the probability function is given by

$$P(X \geq x) \sim cx^{-\gamma}. \tag{4}$$

In a power law distribution the tails fall asymptotically according to the power of γ. Such a distribution leads to heavier tails than other common distributions, such as the Poisson and Geometric distributions.

The figures below display Matlab simulations for the Poisson, Geometric and Power law probability functions for different choices of parameters. Bottom figures show simulations of the Power law function where the log-log plot on the right shows the heavier tail more clearly. These plots will be compared with the corresponding figures of degree distributions in the next chapter.

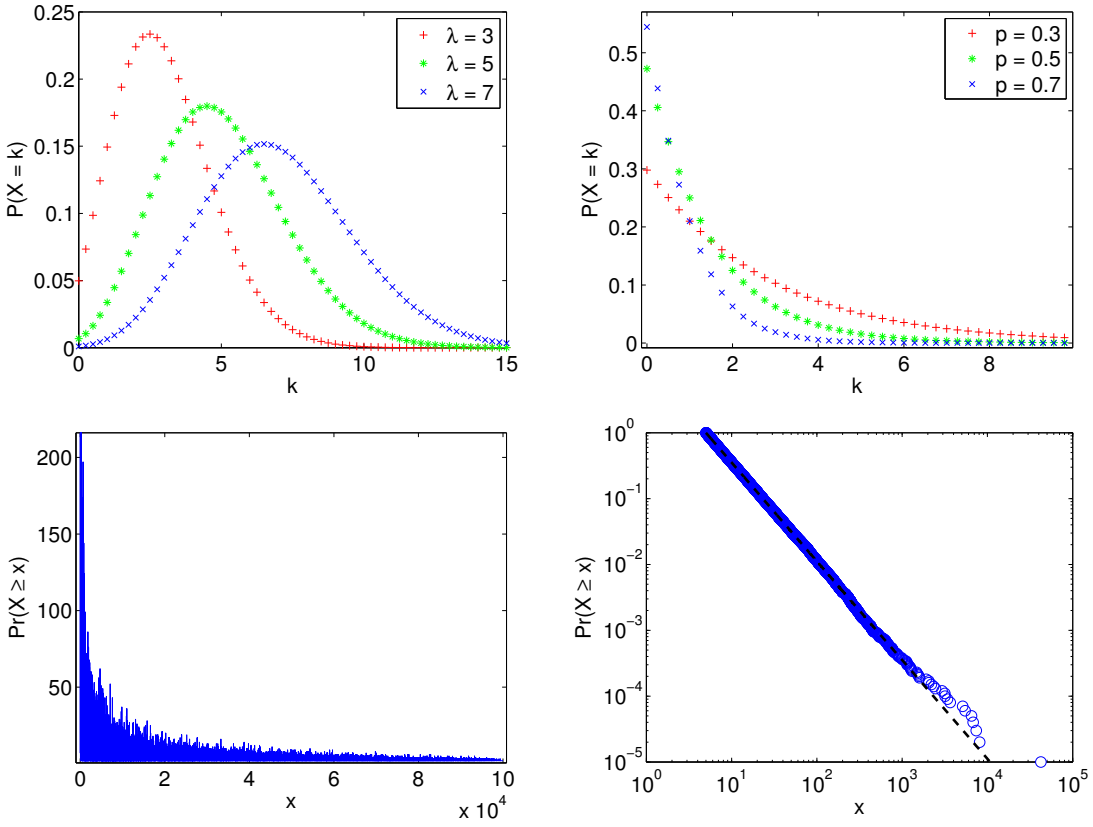


Figure 1: *Simulation results for the Poisson (top-left), Geometric (top-right) and Power law (bottom) probability functions.*

2.2 Graph theory

A graph is a pair $G = (V, E)$ of sets, where V is the set of vertices and $E \subseteq V \times V$ is the set of edges in G. An edge connecting two vertices $x \in V$ and $y \in V$ is denoted by an unordered pair $\{x, y\} \in E$ if the edge is undirected, or by an ordered pair $(x, y) \in E$ if the edge is directed from x to y. If E is a set of directed edges, then $G = (V, E)$ is called a directed graph, otherwise the graph is undirected. G is called a multi-graph if there are at least two directed (same direction) or undirected edges in E connecting the same two vertices x and y.

A way of storing the information of a graph is to represent it with an adjacency matrix A. The adjacency matrix of an undirected graph is the matrix with elements A_{ij} such that $A_{ij} = 1$ if there is an edge between i and j. If all edges are undirected, it follows that $A_{ij} = A_{ji}$. When the graph is directed, $A_{ij} = 1$ if there is an edge from i to j. Also, in a multi-graph $A_{ij} = m$ if there are m edges between i and j. If $A_{ij} = 0$ then there are no edges between i and j.

An x-y walk in G is a finite alternating sequence

$$x = x_0, e_1, x_1, e_2, x_2, e_3, \ldots, e_{n-1}, x_{n-1}, e_n, x_n = y \tag{5}$$

of vertices and edges from G, starting at vertex x and ending at vertex y and involving the n edges $e_i = \{x_{i-1}, x_i\}$, where $1 \leq i \leq n$. The length l of a walk equals to the number of edges in the walk. If no vertex of the walk occurs more than once, then the walk is called a path, and if $x = y$, then it is a closed path or a cycle. The total number of paths of length r in a graph G is given by

$$N_{ij}^{(r)} = \sum_{k,l,\ldots,m=1}^{n} A_{ik} A_{kl} \cdots A_{mj} = [A^r]_{ij} \,. \tag{6}$$

Thus, the shortest (geodesic) path between vertices i and j is the smallest value of r such that $[A^r]_{ij} > 0$. The diameter of a graph is the length of the longest geodesic path between any pair of vertices in the graph for which a path actually exists.

If G is an undirected graph, then the degree k_i of vertex i is the number of edges in G that are connected to vertex i. For an undirected graph of n vertices the degree can be written as

$$k_i = \sum_{j=1}^{n} A_{ij}. \tag{7}$$

For a directed graph, the incoming degree k_i^{in} of vertex i is the number of edges in G that are incident into i and the outgoing degree k_i^{out} is the number of edges incident from i. The in- and out-degrees in the terms of the adjacency matrix are

$$k_i^{in} = \sum_{j=1}^{n} A_{ij}, \quad k_i^{out} = \sum_{j=1}^{n} A_{ji}. \tag{8}$$

Transitivity or clustering occurs in a graph when a vertex u is connected to vertex v, and v is connected to w, then u is also connected to w. The relation between these three vertices forms a cycle of length three. Let the clustering coefficient C of a graph G to be the fraction of three times the number of triangles $3N_\Delta$ divided by the number of connected triples N_3

$$C = \frac{3N_\Delta}{N_3}, \tag{9}$$

where a triangle is a cycle of length three, and a connected triple means three vertices uvw with edges (u, v) and (v, w). We have the factor of three in the numerator because each cycle gets counted three times when we count the connected triples in the graph.

3 Random graphs

Traditionally complex networks have been studied with graph theoretical methods. However, it has been observed that most networks in the real world have commonly occurring topological properties that are quite difficult to detect with deterministic methods. This section will propose the use of random graph models to study the structure of these networks. Random graphs include some random elements when they are created, while they still show some commonly recurring properties similar to those of the networks in the real world.

A random graph is constructed by giving fixed values to a set of parameters, where the graph is random in other aspects. A random graph model is defined as an ensemble of graphs, which is a probability distribution over possible graphs. For example, let us fix the number of vertices n and edges m for a random graph model $G(n, m)$. Then this model is defined as a probability distribution $P(G)$ over all graphs G in which $P(G) = 1/\omega$ for simple graphs with n vertices and m edges and $P(G) = 0$ otherwise, where ω is the total number of such simple graphs.

This chapter is going to introduce three random graph models with equilibrium and non-equilibrium characteristics. Equilibrium network models are used to model a network of a given size and describe their topology at a given time point. Non-equilibrium network models are used to model the growth of the network, where they explain how the networks came to be in the first place. Other properties that these models will include are uniform and preferential attachment. Each model will be implemented in Matlab and the graphs are generated in simulations. The goal in this chapter is to introduce the models and evaluate the degree distribution with analytical and numerical methods for each model.

3.1 The Poisson random graph model

Let $X \sim Bin(n-1, p)$ be a binomially distributed random variable with parameters n and p. A simple example of a random graph model is where we fix the number of vertices n and the independent probability p of connecting an edge from a given vertex to each of the $n-1$ other vertices. [1] The probability of being connected to k other vertices and not to any of the others is $p^k(1-p)^{n-1-k}$. Since there are $\binom{n-1}{k}$ ways to choose those k other vertices, the total probability of being connected

to exactly k others is

$$P\left(X=k\right)=\binom{n-1}{k}p^{k}\left(1-p\right)^{n-1-k}. \tag{10}$$

The number of graphs in a model with exactly n vertices and m edges is equal to the number of ways of picking the positions of the edges from the $\binom{n}{2}$ distinct pairs of vertices. Each of these pairs appears with the same probability $P(G)$ and hence the total probability of generating a graph with m edges from the ensemble is

$$P\left(m\right)=\binom{\binom{n}{2}}{m}p^{m}\left(1-p\right)^{\binom{n}{2}-m}. \tag{11}$$

Then the average number of edges $\langle m\rangle$ is

$$\langle m\rangle=\sum_{m=0}^{\binom{n}{2}}mP\left(m\right)=\binom{n}{2}p. \tag{12}$$

The average degree in a graph with exactly m edges is $\langle k\rangle=2m/n$, and hence the average degree in this model is

$$\langle k\rangle=\sum_{m=0}^{\binom{n}{2}}\frac{2m}{n}P\left(m\right)=\frac{2}{n}\binom{n}{2}p=\left(n-1\right)p. \tag{13}$$

Therefore, the probability p that any two vertices are neighbours is given by

$$p=\frac{\langle k\rangle}{n-1}. \tag{14}$$

This implies that p will become vanishingly small as $n\to\infty$, which gives

$$\begin{aligned}\ln\left(\left(1-p\right)^{n-1-k}\right)&=\left(n-1-k\right)\ln\left(1-\frac{\langle k\rangle}{n-1}\right)\\&\approx-\left(n-1-k\right)\frac{\langle k\rangle}{n-1}\approx-\langle k\rangle,\end{aligned} \tag{15}$$

where the logarithm is expanded as a Taylor series, and the equalities become exact as $n\to\infty$. Taking exponentials of both sizes gives

$$\left(1-p\right)^{n-1-k}=e^{-\langle k\rangle}, \tag{16}$$

for a large number of vertices. Further, for large n we have

$$\binom{n-1}{k}=\frac{(n-1)!}{(n-1-k)!k!}\approx\frac{(n-1)^{k}}{k!} \tag{17}$$

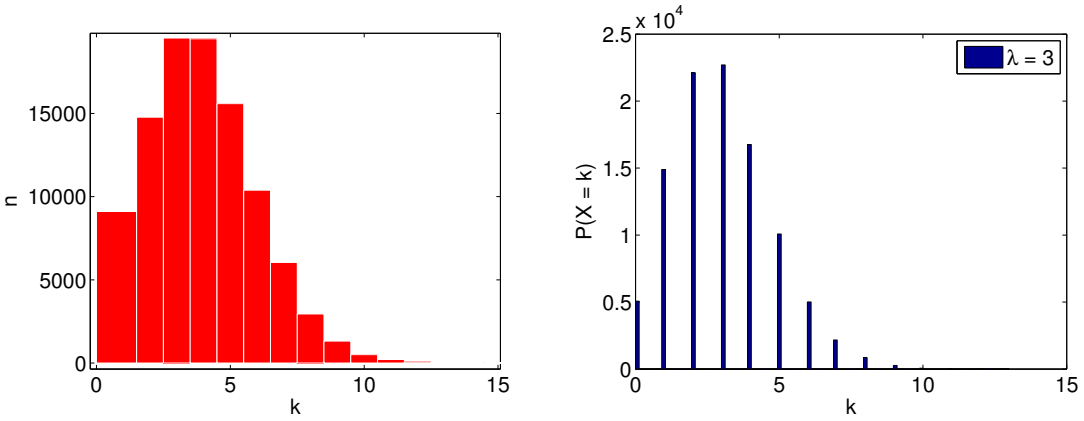


Figure 2: *Degree distribution of the Poisson random graph model (left). Poisson probability density function (right).*

and thus equation (10) becomes

$$\begin{aligned} P(X=k) &= \frac{(n-1)^k}{k!} p^k e^{-\langle k\rangle} \\ &= \frac{(n-1)^k}{k!} \left(\frac{\langle k\rangle}{n-1}\right)^k e^{-\langle k\rangle} \\ &= \frac{\langle k\rangle^k e^{-\langle k\rangle}}{k!}, \end{aligned} \tag{18}$$

which turns out to have a Poisson degree distribution when $n \to \infty$.

The average clustering coefficient $\langle c\rangle$ for this model is the same as the probability p of connecting an edge between a pair of vertices, since the probability that any two vertices are neighbours is the same. Therefore, the average clustering coefficient is given by

$$\langle c\rangle = \frac{\langle k\rangle}{n-1}. \tag{19}$$

This implies that clustering depends on the number of vertices in the graph. A potential problem with this model is when $n \to \infty$, the average clustering coefficient goes to zero.

The average number of vertices with a shortest path length l away from a randomly chosen vertex is $\langle k\rangle^l$. Since this number grows exponentially with l it doesn't take long path lengths before the number of vertices reached is equal to the total number of vertices in the graph. Thus, the average shortest path length between two vertices is given by

$$\langle k\rangle^l \approx n \quad \Leftrightarrow \quad l \approx \frac{\ln n}{\ln\langle k\rangle}. \tag{20}$$

At this point, every vertex is approximately within a length l of our starting point, implying that the diameter of the network is approximately $\ln n / \ln\langle k\rangle$.

3.2 Network growth and preferential attachment

Recent studies suggest that many large networks have a degree distribution that follows a power law distribution. We have seen in the previous section that the Poisson random graph model does not produce this feature. While the goal of the Poisson model is to construct a graph with correct topological features, modelling scale-free networks will put the emphasis on capturing the network dynamics of non-equilibrium network models. The idea behind these models is to study the structures of networks and their properties that these non-equilibrium processes produce. They offer the explanation of why the network should have a particular degree distribution in the first place.

The origin of the power law degree distribution in networks was first addressed by Barabasi and Albert [9], who argued that the scale-free nature of real networks is rooted in two mechanisms common in many real networks, growth and preferential attachment.

The Poisson random graph model starts with a fixed number of vertices that are then connected uniformly at random. However, in a non-equilibrium network model the initial graph state consists a small number of vertices n_0 and at every time point $t > n_0$ we add a new vertex with m edges that connects the new vertex to other pre-existing vertices in the graph. After t time-steps this algorithm results in a network with $n = t + n_0$ nodes and mn edges. This process is also seen in most real networks that start from a small number of vertices and generate systems which grow by the continuous addition of new vertices. For instance, the world-wide web grows exponentially in time by the addition of new web pages.

Preferential attachment means that when choosing the vertices to which the new vertex connects, the probability Π that a new vertex will be connected to vertex i depends on the degree k_i of vertex i, such that

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \tag{21}$$

The uniform attachment property assume that the probability that two nodes are connected is independent of the vertices' degree. However, most real networks exhibit preferential attachment. A web page will more likely include hyperlinks to popular documents with already high degree, because such highly connected documents are easy to find and thus well known.

The remaining part of this section will introduce two non-equilibrium network models with uniform and preferential attachment. These models start with $n_0 = 2$ vertices and $2m$ edges in the graph, for some initial degree $m = 1, 2, \ldots$. After a new

vertex is inserted into the graph, m edges will be connected from the new vertex to some pre-existing vertices picked either uniformly or preferentially at random. It is possible to form a multi-graph by connecting multiple edges to the same vertex. The insertion process will be repeated until all of the n vertices are in the graph.

3.2.1 A non-equilibrium model with uniform attachment

Consider a graph sequence

$$\{G_t(m)\}_{t=2}^{n} = G_2(m), G_3(m), \ldots, G_n(m) \tag{22}$$

which for every time-step t represents a number of t vertices and mt edges in the graph. Let

$$V(G_t) = \{v_1, \ldots, v_t\} \tag{23}$$

be the set of vertices in $G_t(m)$, and $k_i(t)$ the degree of a vertex v_i. For $m \geq 1$ we start with a multi-graph $G_2(m)$ which consists of two vertices connected to each other by $2m$ edges. Then, conditionally on $G_t(m)$, in order to obtain $G_{t+1}(m)$, we add a new vertex v_{t+1}. This vertex will be connected to itself or a pre-existing vertex v_i of $G_t(m)$ with probability

$$\Pr(\{v_{t+1}, v_i\} \mid G_t(m)) = \frac{1}{t+1}, \tag{24}$$

for $i = 1, \ldots, t+1$, and to prevent self-loops the probability is given by

$$\Pr(\{v_{t+1}, v_i\} \mid G_t(m, \delta)) = \begin{cases} \frac{1}{t}, & i = 1, \ldots, t, \\ 0, & i = t+1. \end{cases} \tag{25}$$

Let $p_i(k, t)$ be the probability that vertex i has degree k at time t. The equation which describes the evolution of this probability is given by

$$p_i(k, t+1) = \frac{1}{t} p_i(k-1, t) + \left(1 - \frac{1}{t}\right) p_i(k, t), \tag{26}$$

where $1/t$ is the probability that a vertex gets connected, and $1 - 1/t$ is the probability that the vertex remains in the same state. [2] The total degree distribution of the graph is

$$P(k, t) = \frac{1}{t} \sum_{i=1}^{t} p_i(k, t), \tag{27}$$

and the initial δ_0 and boundary δ_b conditions are

$$\delta_0 = P_2(k, t = 2), \quad \delta_b = P_t(k, t > 2). \tag{28}$$

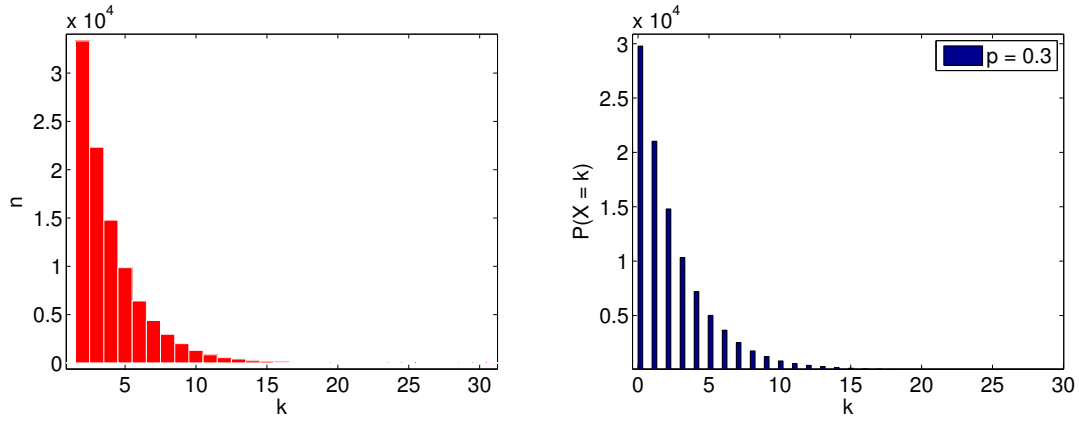


Figure 3: *Degree distribution of the non-equilibrium network model with uniform attachment (left). Geometric probability density function (right).*

Applying the sum on both sides of equation (26) gives us the equation in terms of the total degree distribution

$$(t+1)\,P\,(k,t+1) = P\,(k-1,t) + tP\,(k,t) - P\,(k,t) + \delta_b, \tag{29}$$

The continuum limit of this equation is

$$\begin{aligned}(t+1)\,P\,(k,t+1) - tP\,(k,t) &= \frac{\partial\,(tP\,(k,t))}{\partial t} \\ &= P\,(k-1,t) - P\,(k,t) + \delta_b. \end{aligned} \tag{30}$$

Now let

$$P\,(k) = P\,(k, t \to \infty)\,, \tag{31}$$

where it follows that the degree distribution of this model for large values of t is

$$\begin{aligned} & P\,(k) = P\,(k-1) - P\,(k) + \delta_b \\ \iff \quad & 2P\,(k) - P\,(k-1) = \delta_b \\ \iff \quad & P\,(k) = 2^{-k}. \end{aligned} \tag{32}$$

The Matlab simulations above shows that the degree distribution of this model closely resembles the geometric distribution.

3.2.2 A non-equilibrium model with preferential attachment

In this model [10] we have a graph sequence,

$$\{G_t\,(m,\delta)\}_{t=2}^{n} = G_2\,(m,\delta)\,, G_3\,(m,\delta)\,, \ldots, G_n\,(m,\delta) \tag{33}$$

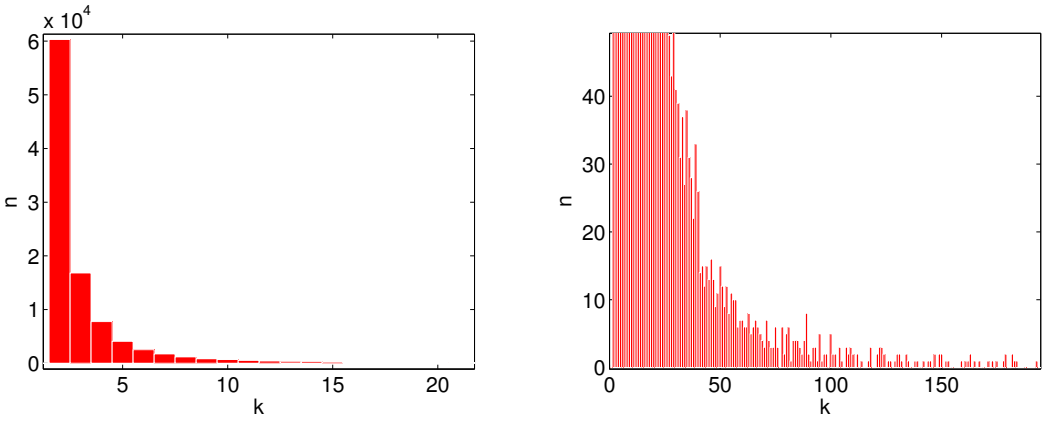


Figure 4: *Degree distribution of the non-equilibrium network model with preferential attachment. Both figures represent the same data, the left figure shows the highest point on the curve and the second figure is zoomed in to show how far right on the x-axis the tail goes.*

which for every time point t represents a graph of t vertices and mt edges for some $m = 1, 2, \ldots$. The extra parameter δ makes the model more general. Let

$$V\left(G_t\right) = \{v_1, \ldots, v_t\} \tag{34}$$

be the set of vertices in $G_t\left(m, \delta\right)$, and $k_i(t)$ the degree of a vertex v_i. For $m \geq 1$ and $\delta > -m$ we start with a multi-graph $G_2\left(m, \delta\right)$ which consists of two vertices connected to each other by $2m$ edges. Then, conditionally on $G_t\left(m, \delta\right)$, in order to obtain $G_{t+1}\left(m, \delta\right)$, we add a new vertex v_{t+1}. This vertex will be connected to itself or a pre-existing vertex v_i of $G_t\left(m, \delta\right)$ with probability,

$$\Pr\left(\{v_{t+1}, v_i\} \mid G_t\left(m, \delta\right)\right) = \begin{cases} \frac{k_i(t)+\delta}{t(2+\delta)+(1+\delta)}, & i = 1, \ldots, t, \\ \frac{1+\delta}{t(2+\delta)+(1+\delta)}, & i = t+1, \end{cases} \tag{35}$$

and if self-loops are excluded, we get

$$\Pr\left(\{v_{t+1}, v_i\} \mid G_t\left(m, \delta\right)\right) = \begin{cases} \frac{k_i(t)+\delta}{t(2+\delta)}, & i = 1, \ldots, t, \\ 0, & i = t+1. \end{cases} \tag{36}$$

The difference between the previous model and this one is that the probability to connect a vertex is directly proportional to the degree of the vertex. This suggests the influence of the preferential attachment property in this model. Since the degrees of vertices are updated after each edge is attached, the attachment probabilities for each vertex will be updated as well.

In order to find out the degree distribution of this model, let us first define the Gamma function

$$\Gamma\left(\alpha\right) = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{37}$$

for $\alpha > 0$. With partial integration it can be shown that

$$\Gamma(n+1) = n\Gamma(n), \quad \Gamma(n) = (n-1)!. \tag{38}$$

Let $p_i(k,t)$ be the probability that vertex i has degree k at time t. The equation which describes the evolution of this probability is given by

$$p_i(k,t+1) = \frac{(k-1)+\delta}{t(2+\delta)} p_i(k-1,t) + \left(1 - \frac{k+\delta}{t(2+\delta)}\right) p_i(k,t), \tag{39}$$

where $(k-1+\delta)/t(2+\delta)$ is the probability that a vertex gets connected, and $1-(k+\delta)/t(2+\delta)$ is the probability that the vertex remains in the same state. The total degree distribution of the graph is

$$P(k,t) = \frac{1}{t}\sum_{i=1}^{t} p_i(k,t). \tag{40}$$

Letting $P(k) = P(k, t \to \infty)$, the total degree distribution turns out to be

$$P(k) = \left(2 + \frac{\delta}{m}\right) \frac{\Gamma(k+\delta)\,\Gamma\left(m+2+\delta+\frac{\delta}{m}\right)}{\Gamma(m+\delta)\,\Gamma\left(k+3+\delta+\frac{\delta}{m}\right)}, \tag{41}$$

for $k \geq m$, and $p_k = 0$ for $k = 0, \ldots, m-1$. [10] This can be further simplified to get

$$P(k) = C_{m,\delta} k^{-\gamma}, \tag{42}$$

for an exponent $\gamma = 3 + \delta/2$ and some constant $C_{m,\delta}$. This shows that the non-equilibrium network model with preferential attachment has a power law degree distribution. Comparing these three random graph models suggests that both growth and preferential attachment are required to produce graphs with a power law degree distribution.

4 Application: Peer-to-Peer networks

Peer-to-Peer is a communications model of a network in which each peer can be acting as a client or server requesting and providing resources. These networks do not depend on dedicated servers, instead the communication occurs directly between peers. Many of today's applications are based on the Peer-to-Peer model. Some examples of these applications are file-sharing and IP-telephony software such as BitTorrent and Skype.

The objective of this chapter is to study the network structure and data of a Peer-to-Peer social network. Some properties of the random graph models will be compared with the corresponding properties of the Peer-to-Peer network. The data used for this thesis consists of a statistical analysis that captures social associations of distributed peers in resource sharing. The analysis was done by an experimental machine acting as a monitor to record the traffic passing through the network. It recorded information such as which peer answered a query of which other peer. The experimental machine ran on the network from 5 hours to 3 days. It usually connected 300 normal peers and 30 other super nodes. The traffic data it recorded involved 1000 to 200000 peers. [6]

4.1 Network structure and data

In the article, basic properties of the networks such as the degree distributions have been particularly studied. The results have shown that these networks demonstrate the small-world effect, as the shortest path length between peers is small. Moreover, most of the peers are pure resource providers, contributing to the high resource availability of Peer-to-Peer networks in resource sharing. Comparing with the peers that do not contribute any resources are only a small fraction of the whole network. For peers that have more than one connection, their undirected degree distributions follow a power-law distribution.

Investigations on betweenness and correlations suggest that dynamics of peer social networks are not dominated by a few highly connected peers. In fact, the peer degrees suggest that active providers are connected between each other and by active requesters. The collected social networks studied in the article are only some small snapshots of the large-scale and continuously changing Peer-to-Peer networks.

The possible social links between peers were discovered from the collected data to form corresponding networks. A directed connection was created from peer A to peer

B if B was a query answerer of A. The strength of this connection indicated how many queries B answered A. The stronger a connection strength is, the more important the end peer is to the other peer of the connection. A connection strength with value 1 suggests a single communication, and hence a weak association. Strength with a constantly high value suggests the end peer is a frequent resource provider of the start peer, and hence a long-term social relation. However, the connection strength may decay over time in the absence of any contribution from the end peer. Among tens or hundreds of thousands of peers, only a few of them acted as both requesters and providers. These peers play a major role in Peer-to-Peer social networks as they contribute essential links to the networks. These peers are hence called major peers.

The average values of Table 1 will be studied and compared with results from the random graph simulations. Here the average degree values vary between 3.84 and 4.56, which indicates that there is an unequal amount of in and out degrees. Whereas the random graph models studied in this thesis have only undirected edges. This means that the simulations will not show any variance in this aspect. As long as there is a fixed number of undirected edges, the average degree of a vertex is fixed as well.

	SN1	SN5	SN6
n	42186	112921	191679
m	81083	230500	415037
$\langle k \rangle$	3.84	4.56	4.5
$\langle l \rangle$	5.45	6.77	8.5
$\langle c \rangle$	0.019	0.021	0.015

Table 1: *Selected data from the article. SN1, SN5 and SN6 are three snapshots of the Peer-to-Peer network recorded at three distinct time points. First two rows include number of vertices n and edges m. Last three rows display values of average degrees $\langle k \rangle$, shortest path lengths $\langle l \rangle$ and clustering coefficients $\langle c \rangle$.*

4.2 Analysis

In this section three models will be simulated, where the average values of degrees, shortest path lengths and clustering coefficients are evaluated. Each snapshot of the network includes approximately between 40,000 and 200,000 vertices and between 80,000 and 400,000 edges. Because of the difficulty of storing information about

graphs of these sizes in an adjacency matrix, smaller samples of graphs will be generated in the simulations.

4.2.1 The Poisson random graph model

As shown before, the expected values of the properties for this model can be found analytically. To evaluate the average degree of a vertex $\langle k \rangle$ in the graph, we simply take twice the number of edges m and divide by the number of vertices n. Since m is fixed to be twice the number of vertices n, the average degree value of a vertex turns out to be $\langle k \rangle = 4$ for all simulations. The average values of shortest path length $\langle l \rangle$ between a pair of vertices and the clustering coefficient $\langle c \rangle$ are given by

$$\langle l \rangle = \frac{\ln n}{\ln \langle k \rangle}, \quad \langle c \rangle = \frac{\langle k \rangle}{n-1}. \tag{43}$$

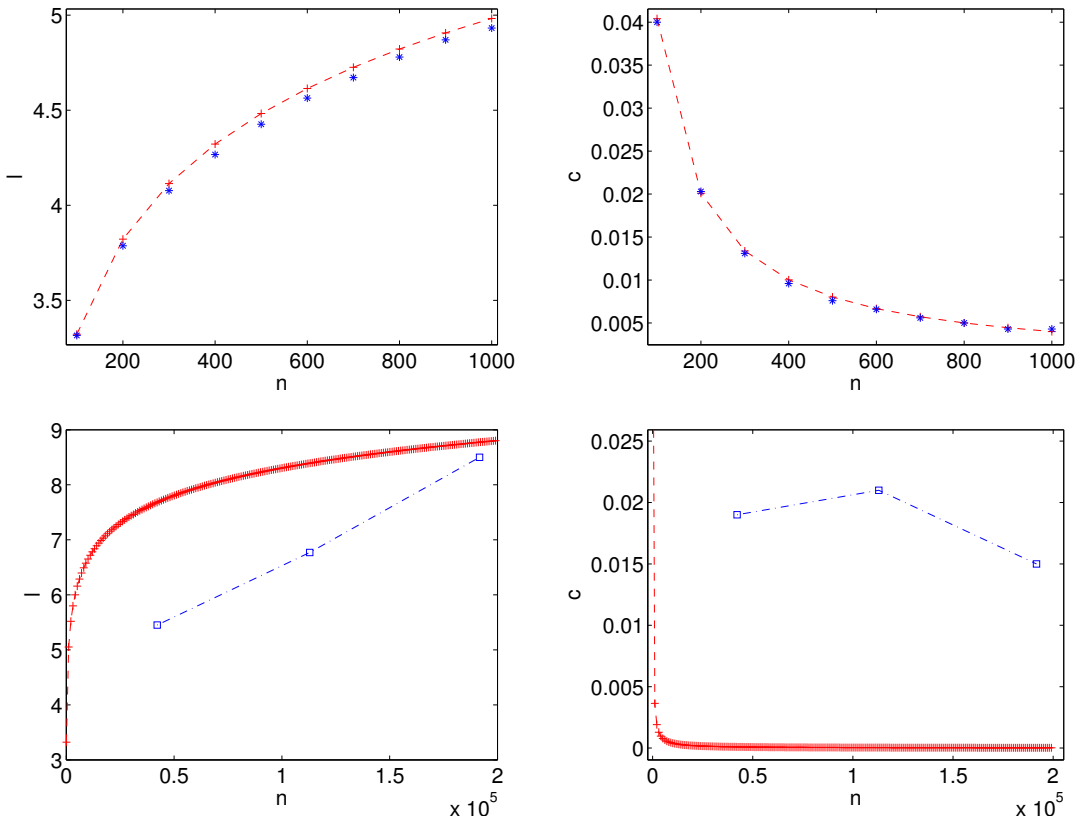


Figure 5: *Results for the Poisson model. It displays the average shortest path lengths $\langle l \rangle$ (left) and clustering coefficients $\langle c \rangle$ (right) for different sized graphs. Top figures are comparisons between data from simulations (*) and theoretical values (+) of the Poisson model. Bottom figures are comparisons between the model (*) and data points from the Peer-to-Peer network (squares).*

The first two sub-figures show results from simulations where the number of vertices are between 100 and 1000, and edges have been set to be twice the number of vertices. For both average shortest path lengths and clustering coefficients we can draw the conclusion that the values from the algorithms correspond closely with the theoretical values.

For the simulations on the bottom two sub-figures, the number of vertices have been set between 1000 and 200000, and edges are twice the number of vertices. It shows that the average shortest path lengths of each Peer-to-Peer network have smaller values than the values of the model. However, network SN6 is quite close to the value given by the model. The average clustering coefficient in this model depends on the value of number of nodes, and it can be seen that the value drops very low quickly and it looks like the clustering in this model does not correspond well with the Peer-to-Peer network data.

4.2.2 The non-equilibrium network models

For the non-equilibrium network models there are no available analytic methods to evaluate average values of the model properties. The simulations will be done for both models and the figures below illustrate the difference between these models for both the average shortest path lengths and clustering coefficients.

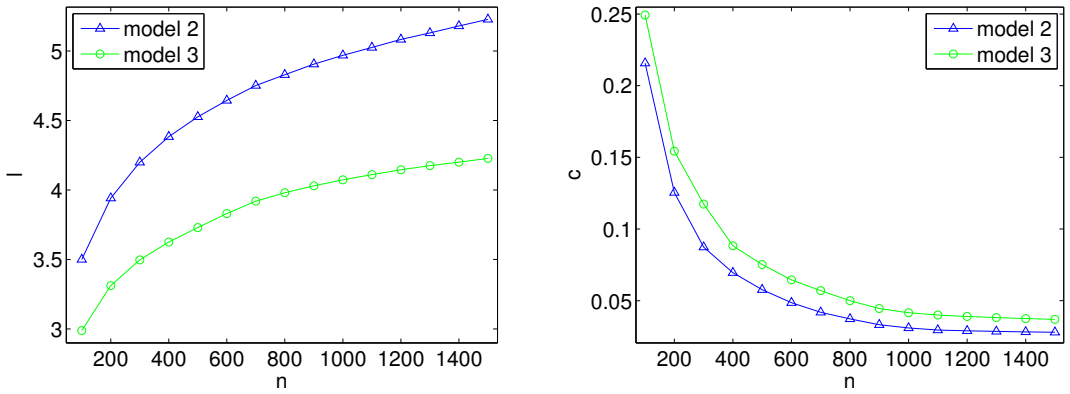


Figure 6: *Results for the non-equilibrium network models. It displays the average shortest path lengths $\langle l \rangle$ (left) and clustering coefficients $\langle c \rangle$ (right) for different sized graphs. Here model 2 is the non-equilibrium model with uniform attachment and model 3 is the non-equilibrium model with preferential attachment.*

Both sub-figures show the results from simulations where the number of vertices and minimum degree have been set to $n = 1500$ respective $m = 2$. It can be seen on the left figure that the model with preferential attachment has significantly shorter

path lengths compared to the model with uniform attachment. Also, on the right figure the model with preferential attachment shows higher values of clustering. However, there is a smaller difference between the mean values of the clustering coefficients.

4.3 Results

The random graph model with a Poisson degree distribution can rarely model real world networks accurately. This was clear when comparing the simulations results against data from the Peer-to-Peer network. The biggest difference is shown with the average clustering coefficient which drops quickly for large number of nodes in the graph. Another significant aspect in which the properties of this model diverge from those of real networks is the shape of their degree distribution. Real networks typically have right-skewed degree distributions, with most vertices having low degree but with a small number of high-degree hubs in the tail of the distribution.

The non-equilibrium model with uniform attachment has a geometric degree distribution, which like the Poisson distributed model does not show any number of high-degree hubs in the tail of the distribution. The average shortest path lengths do not differ much from the values observed in the Poisson model. This model starts with the highest values, but when $n > 900$ the Poisson model shows highest mean values of the shortest path lengths. However, the average clustering coefficients are much higher for the non-equilibrium model for all sizes of graphs.

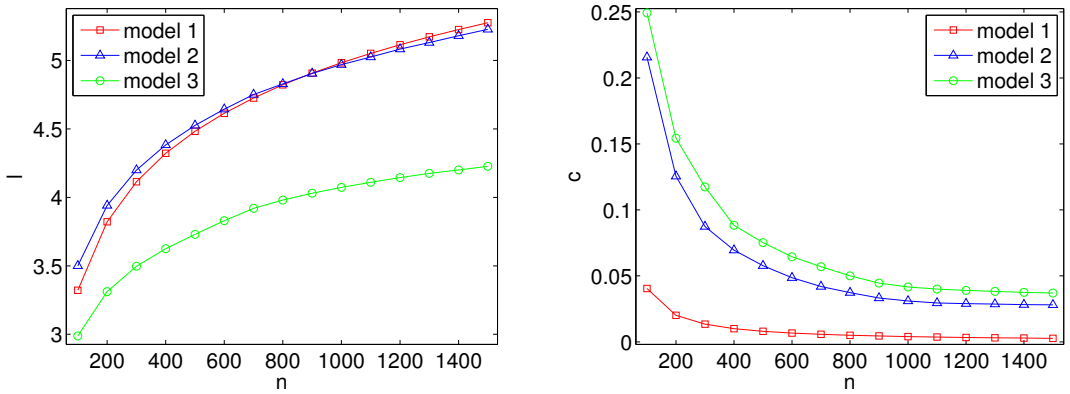


Figure 7: *Results and comparison between each random graph model. It displays the average shortest path lengths $\langle l \rangle$ (left) and clustering coefficients $\langle c \rangle$ (right) for different sized graphs. Here model 1 is the Poisson random graph model, model 2 is the non-equilibrium model with uniform attachment and model 3 is the non-equilibrium model with preferential attachment.*

When uniform attachment is exchanged with preferential attachment, it turns out that the second non-equilibrium model shows a degree distribution proportional to a power law distribution for graphs with large number of vertices. This makes a big difference for the average shortest path lengths which give much lower values than the other two models. However, the average clustering coefficients tend to be quite close to the non-equilibrium model with uniform attachment.

5 Conclusions

This report presented a study of random graphs as models of Peer-to-Peer social networks. Three random graph models and their resulting degree distributions were studied and it turns out that both growth and preferential attachment are needed to generate a graph with a scale-free degree distribution.

The application chosen for this thesis was a Peer-to-Peer social network that represents social associations of distributed peers in resource sharing. This particular network has a degree distribution that follows a scale-free distribution. The application was presented in the previous chapter, and some numerical simulations were done for each random graph model. The results from these simulations were compared between the models and the corresponding data collected from the network. From the simulations it can be seen that the model with growth and preferential attachment shows much lower values for the average shortest path lengths than the other two models, and the values for the average clustering coefficient are higher compared with the other models. Estimation of the linear regression for larger number of vertices suggests that the non-equilibrium network model with preferential attachment shows the closest correspondence with the empirical data observed from the Peer-to-Peer network.

As the Peer-to-Peer network grows by starting with a given small number of nodes, new nodes join the network within a given time period and make connections to nodes that already exists in the network. Therefore, it is clear that non-equilibrium network models are more suitable to model this network than equilibrium network models, such as the Poisson random graph. Also, because the nodes that are already highly connected are more likely to get more connections than nodes with lower degrees, it makes sense to have a model with some variant of the preferential attachment property.

The three models used for the simulations are still quite simple, for instance directed and weighted edges have not been taken into account. As most real networks have edges that are directed and weighted, it is expected that by constructing models and specifying these two attributes will provide a better and more accurate scenario for their study and modelling. Also, there are many more models that could be studied as well. For example, a combination of an equilibrium network that was left out is the equilibrium network model with preferential attachment, and a generalized scale-free model which is based on non-linear preferential attachment.

A Matlab implementations

This appendix includes Matlab code for the random graph models. The pathlength and clustering algorithms used here are found in the contest (controllable test matrices) random network toolbox for Matlab at `http://www.mathstat.strath.ac.uk/research/groups/numerical_analysis/contest/toolbox`.

A.1 The Poisson random graph model

```
function G = erdosm(n,m,N)
% The Poisson random graph model with a fixed number of
% edges attached uniformly at random
%    INPUT     n — number of vertices
%              m — number of edges
%              N — number of graphs to simulate
%    OUTPUT    G — resulting graph

S = 0; C = 0;

% for each graph g
for g = 1:N
    % initial state (n vertices and 0 edges)
    G = zeros(n);

    % connect all m edges
    for e = 1:m
        % pick two vertices i and j uniformly at random
        r = randperm(n);
        i = r(1); j = r(2);
        % assign the values into the matrix
        G(i,j) = G(i,j) + 1;
        G(j,i) = G(j,i) + 1;
    end

    % find the shortest path lengths between a pair of vertices
    L = pathlength(G);
    % calculate the average shortest path length S
    S = S + sum(L(:));
    % calculate the average clustering coefficient C
    C = C + clustering(G,'ave');
end
```

A.2 A non-equilibrium model with uniform attachment

```
function G = pref2(n,d,N)
% A non-equilibrium network model with uniform attachment
%    INPUT     n - number of vertices
%              d - initial degree of a new vertex
%              N - number of graphs to simulate
%    OUTPUT    G - resulting graph

S = 0; C = 0;

% for each graph g
for g = 1:N
    % pre-allocate the adjacency matrix of size n
    G = zeros(n);
    % initial state (2 vertices and 2d edges)
    G(1,2) = d*2;
    G(2,1) = d*2;

    % for each new vertex v
    for v = 3:n
        % connect all d edges one at a time
        for e = 1:d
            % choose the second vertex uniformly at random
            ind = [v, ceil(rand*v)];
            % connect vertices v and ddistr(r)
            G(ind(1),ind(2)) = G(ind(1),ind(2)) + 1;
            % undirected graphs have a symmetric adjacency matrix
            G(ind(2),ind(1)) = G(ind(2),ind(1)) + 1;
        end
    end

    % find the shortest path lengths between a pair of vertices
    L = pathlength(G);
    % calculate the average shortest path length S
    S = S + sum(L(:));
    % calculate the average clustering coefficient C
    C = C + clustering(G,'ave');
end
```

A.3 A non-equilibrium model with preferential attachment

```
function G = pref(n,d,N)
% A non-equilibrium network model with preferential attachment
%    INPUT     n - number of vertices
%              d - initial degree of a new vertex
%              N - number of graphs to simulate
%    OUTPUT    G - resulting graph

S = 0; C = 0;

% for each graph g
for g = 1:N
    % pre-allocate the adjacency matrix of size n
    G = zeros(n);
    % initial state (2 vertices and 2d edges)
    G(1,2) = d*2;
    G(2,1) = d*2;

    % for each new vertex v
    for v = 3:n
        % connect all d edges one at a time
        for e = 1:d
            % ddistr stores information of the degree distribution
            % e.g. if [1 1 1 1 2 2 2 3], then vertex 2 has ...
               degree 3
            % and a probability 3/8 of being connected
            ddistr = [];
            % count the total degree of each vertex
            td = sum(G');
            % exclude vertices with zero degree
            nztd = td(td~=0);
            % for each non-zero row (vertex)
            for j = 1:nnz(td)
                for i = 1:nztd(j)
                    ddistr = [ddistr j];
                end
            end
            % choose an index uniformly at random
            r = ceil(rand*nnz(ddistr));
            % choose the second vertex with preferential attachment
            ind = [v, ddistr(r)];
            % connect vertices v and ddistr(r)
```

```
41              G(ind(1),ind(2)) = G(ind(1),ind(2)) + 1;
42              % undirected graphs have a symmetric adjacency matrix
43              G(ind(2),ind(1)) = G(ind(2),ind(1)) + 1;
44          end
45      end
46
47      % find the shortest path lengths between a pair of vertices
48      L = pathlength(G);
49      % calculate the average shortest path length S
50      S = S + sum(L(:));
51      % calculate the average clustering coefficient C
52      C = C + clustering(G,'ave');
53  end
```

References

[1] M.E.J. Newman. Networks : an introduction. *Oxford University Press*, 2010.

[2] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks : From biological nets to the internet and www. *Oxford University Press*, 2003.

[3] R.P. Grimaldi. Discrete and combinatorial mathematics : An applied introduction. *Pearson Education*, 2003.

[4] S.E. Alm and T. Britton. Stokastik: Sannolikhetsteori och statistikteori med tillämpningar. *Liber*, 2008.

[5] R. Steinmetz and K. Wehrle. Peer-to-peer systems and applications. *Springer-Verlag Berlin and Heidelberg GmbH and Co. K*, 2005.

[6] Y. Sun F. Wang, Y. Moreno. Structure of peer-to-peer social networks. *Physical Review E 73, 036123*, 2006.

[7] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature, Volume 393, p.440-442*, 1998.

[8] M.E.J. Newman. Random graphs with clustering. *Phys. Rev. Lett. 103, 058701*, 2009.

[9] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics 74, 47*, 2002.

[10] R. van der Hofstad. Random graphs and complex networks. *Lecture notes, http://www.win.tue.nl/ rhofstad/NotesRGCN.pdf*, 2012.