Complex scale-free networks with tunable power-law exponent and clustering

E.R. Colman *, G.J. Rodgers

Department of Mathematical Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

HIGHLIGHTS

- We introduce a novel network evolution process that models a network of citations.
- Links are made from new nodes to m random nodes and l neighbours of each one.
- The degree distribution is scale-free and the power-law exponent is $(2l+1)/l$.
- By selecting l and m the clustering can be tuned between 0 and 1.
- When l and m are random variables the degree distribution is also scale free.

ARTICLE INFO

ABSTRACT

We introduce a network evolution process motivated by the network of citations in the scientific literature. In each iteration of the process a node is born and directed links are created from the new node to a set of target nodes already in the network. This set includes m "ambassador" nodes and l of each ambassador's descendants where m and l are random variables selected from any choice of distributions p_l and q_m. The process mimics the tendency of authors to cite varying numbers of papers included in the bibliographies of the other papers they cite. We show that the degree distributions of the networks generated after a large number of iterations are scale-free and derive an expression for the power-law exponent. In a particular case of the model where the number of ambassadors is always the constant m and the number of selected descendants from each ambassador is the constant l, the power-law exponent is $(2l+1)/l$. For this example we derive expressions for the degree distribution and clustering coefficient in terms of l and m. We conclude that the proposed model can be tuned to have the same power law exponent and clustering coefficient of a broad range of the scale-free distributions that have been studied empirically.
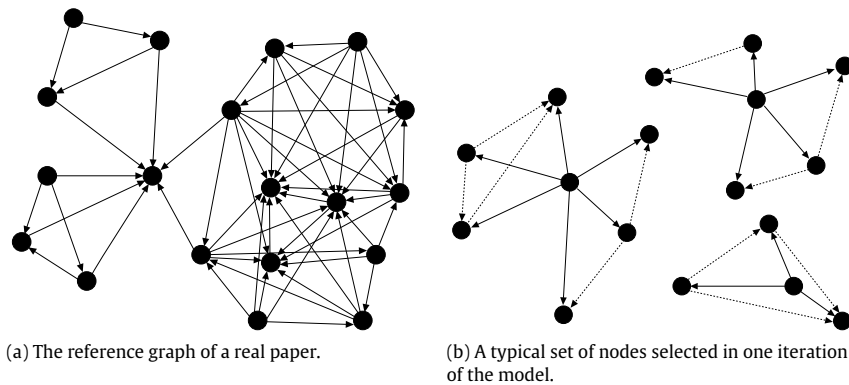
© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Networks, in recent years, have become ubiquitous in the modelling of complex systems. Many fields of study, including for example biology [1], economics [2], and epidemiology [3], employ network based models to mimic the large numbers of agents that interact in the systems they study. It is often found that with the aid of an appropriate network model, the macroscopic behaviour of a complex system can be reproduced with very few assumptions being made about the constituent agents themselves. Typically the system being modelled will be reduced to a set of vertices and a set of vertex pairs called edges. Vertices and edges may represent things like web pages and the hyper-links between them [4], people and their friendships [5], or transport hubs and the transport links between them [6]. In many cases the structure of the network exhibits non-trivial statistical properties such as a high level of clustering, short average path lengths and small numbers

---

* Corresponding author. Tel.: +44 7821353915.
 E-mail address: Ewan.Colman@brunel.ac.uk (E.R. Colman).

(a) The reference graph of a real paper. (b) A typical set of nodes selected in one iteration of the model.

Fig. 1. (a) Shows the sub-network know as the reference graph of a real paper. The nodes represent the papers it cites and the edges represent the citations between them, it is highly clustered and many of the nodes are descendants of others. *Source:* Ref. [23]. (b) Shows a typical sub-network structure of nodes that the proposed model links to, in this case three nodes are selected initially and 3, 5, and 6 of their descendants are also selected, the dotted lines represent the other links between descendants.

of highly connected vertices. An example that is frequently used to illustrate these properties is the network of citations in the scientific literature. Here we present a stochastic model that closely resembles this network but also has the potential to have many other applications.

A crude but arguably effective measure of the worth of a scientific paper is the number citations made to it from other existing scientific articles. Empirical studies have shown that the number of articles with k citations (i.e. cited by k other articles) is proportional to k^{-3} [7,8]. This distribution has certain properties we might expect, namely that the vast majority of papers written have few citations, creating little or no impact on future research, whereas a very small number of papers are extremely significant and have a very large number of citations. By modelling each paper as a node (vertex) and drawing directed edges from each paper to the papers it cites, it has been shown that the correct degree distribution is reproduced using *preferential attachment*; the process of creating nodes sequentially and linking them to nodes selected randomly with probability proportional to their degree (originally discussed in Ref. [9] although the term was coined later in Ref. [10]). The implication of this result is that authors of scientific articles are more likely to choose to cite articles that are already well cited rather than ones that have few or no citations. The attractiveness of highly connected nodes can be explained by a number of processes for example *redirection* [11], where nodes are selected randomly and a link is formed between one of its neighbours and a new node, and random walk models [12–14] where the new node is linked to the nodes occupied by random walkers on the network.

There is a growing literature offering more accurate representations of the way in which the citation network develops, much of this work can be found in the fields of Scientometrics, Bibliometrics, Informetrics and Webometrics [15–17]. A significant amount has been written concerning models that not only agree with the empirical data regarding degree distributions but also agree with other properties, for example in Ref. [17] the evolution of the citation network model is motivated by a coupling with the network of co-authors, other models account for the effect of time on the probability of receiving a citation [18,19]. The model in Ref. [20] introduces tunable clustering (quantified by the clustering coefficient [21]) by extending the preferential attachment model with an additional Triad Formation (TF) step. For each node that is introduced to the network, a node is selected by preferential attachment and linked to, then each neighbouring node is selected with probability p and also linked to from the new node, resulting in a triangle (triad) of edges. The forest fire model described in Ref. [22] extends the Triad Formation model by selecting multiple neighbours of the initially selected node, the process continues by then linking to a number of the neighbours of those neighbours and so on, at each stage a random variable from the binomial distribution determines the number of neighbours selected. In Ref. [23] the forest fire model, along with other models that attempt to mimic the network of citations in even greater detail, is tested against empirical data. The authors also examine the way the articles cited by any one paper, call it i, relate to one another forming a sub-network called a reference graph of i (see Fig. 1(a)). They observed that a clique structure is prevalent, i.e small groups of nodes that all link to each other, and incorporated this finding into their own model.

Much of the literature suggests that the high levels of clustering found in citation networks is a consequence of each author's choice to cite papers that are found in the bibliographies the other papers they cite. This has been observed empirically [24], and modelled using a TF process where the initial nodes are selected randomly (rather than preferentially) [25]. A power-law degree distribution was found with an exponent that varies depending on the Triad Formation probability p, however, this model does not exhibit the exponential out-degree distribution observed in the data [26].

The models mentioned above and those considered in this paper belong to the class of evolving directed clustered scale-free networks that have applications beyond citation networks, the world-wide web being another well studied example. In these models the distributions of in-degree and out-degree are treated separately, often driven by a preferential linking mechanism where the probability of adding an edge from a node i is proportional to the out-degree of i, similarly the probability that the link will end at node j is proportional to the in-degree of j [27]. Correlations between the in-degree and
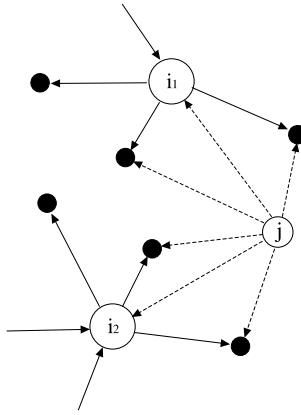
Fig. 2. The new node j attaches to 2 randomly selected nodes i_1 and i_2 as well as 2 randomly descendants shown here in black, the dashed lines represent the new edges that are added in this iteration whereas the solid ones were added previously. This example illustrates one iteration in the growth process when $m = 2$ and $l = 2$.

out-degree of nodes in such networks have been shown to emerge [28]. A detail of citation networks that makes analysis substantially easier is that the out-degree of a node i is fixed from the moment it is created. Consequently the evolution of the out-degree distribution can be disregarded, moreover we can control the out-degree distribution through an appropriate parametrization and ultimately answer the question of how the distribution of bibliography sizes affects the topology of the network.

In this paper we introduce a variant of the TF model that uses a different parameter set to those previously studied. Using the distributions for the number of initial citations and the number of copied citations (which together give the out-degree distribution) as parameters, we show that the networks created by this process may have power-law in-degree distribution with any exponent greater than 2, and a clustering coefficient that ranges between 0 and 1. In Section 2 we describe the stochastic process that iteratively grows the network. We describe a simple case of the model in Section 3 and solve for the power law exponent of the in-degree distribution in terms of two input parameters. In Section 4 we formulate an expression for the in-degree distribution in terms of two input probability distributions then in Section 5 we find an expression for the clustering coefficient for the model in Section 3. In Section 6 we present numerical results that confirm the results of Sections 3 and 4. In Section 7 we discuss the strengths of our models and suggest how this work might be continued.

2. The model

Starting from a finite random network, at each iteration a node j is introduced and directed links are formed between j and a set of nodes that already exist in the network. Letting p_l and q_m be the probability distributions of the discrete random variables l and m, links are formed by the following process:

1. The value m is selected with probability q_m and steps 2 and 3 are repeated a m times.
2. The value l is selected with probability p_l, a node i in the network is randomly selected from those which have out-degree l or greater, the edge $j \rightarrow i$ is added. Borrowing the terminology used in Ref. [22] we will refer to i as an "ambassador".
3. l of i's descendants are randomly selected and directed edges are added from j to each of these.

We are primarily interested in expressing the degree distributions for both incoming and outgoing edges and the clustering coefficient of the network as the number of iterations grows very large in terms of p_l and q_m ($l, m \in \mathbb{N}$). In the next section we solve for a simplified model where l and m are fixed (i.e. $p_r = \delta_{rl}$ and $q_r = \delta_{rm}$), we present the general solution in the section that follows.

3. Attachment to m random nodes and l of each of their descendants

We examine the network generated by the process described in Section 2 when $p_r = \delta_{rl}$ and $q_r = \delta_{rm}$, in this section we derive the degree distribution of this network. In this simplified model the growth of the network depends only on the fixed values l and m, thus the process can be described concisely as follows; in each iteration, m ambassador nodes are randomly selected, l descendants of each ambassador are also selected, then a new node j is attached to each of the selected nodes (see Fig. 2). We are interested in calculating the probability $P(k)$ of finding a node with in-degree k, in the citation model this represents the proportion of articles that are cited by k other papers. Let N be the total number of nodes, N increases by 1 with each iteration and every node has an out-degree of $m(l+1)$, the number of edges as N grows large is $E = m(l+1)N$. Consider a typical node i with in-degree k. There are two possible events which may cause the degree of i to increase to $k+1$: i can either be selected as one of the m ambassador nodes, or it can be selected as a descendant of another node j. In any given iteration, i will be selected as an ambassador with probability m/N. Alternatively j will be selected as an ambassador

with probability m/N and then i will be one of the l selected descendants of j with probability $l/m(l+1)$. Let $P(i; k)$ denote the probability that in one iteration a node i with degree k is selected. Since there are k potential ancestors to i,

$$P(i; k) = \frac{m}{N}\left(1 + \frac{lk}{m(l+1)}\right).$$ (1)

It is possible for the same node to be selected two or more times in one iteration, for example if two of the selected ambassador nodes are a distance of one or two edges from each other. Since this possibility becomes less likely as N increases we do not account for it in our calculations. Let I_k be the number of nodes with in-degree k. For $k \geq 1$, I_k changes over time according to the rate equation

$$\frac{\partial I_k}{\partial N} = \frac{m}{N}\left[\left(1 + \frac{l(k-1)}{m(l+1)}\right)I_{k-1} - \left(1 + \frac{lk}{m(l+1)}\right)I_k\right].$$ (2)

The first term on the right hand side accounts for the creation of a node of in-degree k that occurs when one of the new edges attaches to a node of in-degree $k-1$, the second term accounts for the destruction of a node of in-degree k when it is attached to by one of the new edges. For $k = 0$ the rate equation is

$$\frac{\partial I_0}{\partial N} = 1 - \frac{m}{N}I_0.$$ (3)

We are interested in finding $P_{in}(k)$ the probability of a node having in-degree k when N is very large. By assuming $P(k)$ grows linearly with N when N is large, we substitute $I_k = NP_{in}(k)$ into Eq. (2) to find

$$\left(\frac{l+1+m(l+1)}{l} + k\right)P_{in}(k) = \left(\frac{m(l+1)-l}{l} + k\right)P_{in}(k-1)$$ (4)

for $l \neq 0$. From Eq. (3) we also find

$$P_{in}(0) = \frac{1}{1+m}$$ (5)

and thus the in-degree distribution is expressed

$$P_{in}(k) = \frac{1}{m+1}\frac{\Gamma[(2l+1+m(l+1))/l]\Gamma[k+m(l+1)/l]}{\Gamma[m(l+1)/l]\Gamma[k+(2l+1+m(l+1))/l]}.$$ (6)

For large enough values of k, $P_{in}(k)$ has power-law form

$$P_{in}(k) \sim k^{-\gamma} \quad \text{where } \gamma = \frac{2l+1}{l}.$$ (7)

4. Solution to the general model

Let $P_{out}(s)$ denote the proportion of nodes in the network that have out-degree s. Note that at the time of its creation, the out-degree of a node is fixed and, unlike its in-degree, does not change over time. Therefore, for sufficiently large networks, $P_{out}(s)$ is equal to the probability of creating a node with out-degree s within a single iteration. This can be written

$$P_{out}(s) = \sum_{n=1}^{\infty} q_n P\left(\sum_{i=1}^{n} x_i = s\right)$$ (8)

where the x_i are integer random variables that equal $l+1$ with probability p_l.

To calculate the in-degree distribution we again construct a rate equation from the probability that the degree of a typical node i will increase in one iteration. Let T_l be the number of nodes have out-degree greater or equal to $l-1$, for the node i to be randomly selected as the ambassador in step 2 it must be one of these nodes. The probability that this is the case, multiplied by the probability that i is the one node randomly selected from the T_l nodes available, forms the probability that i is the ambassador given that l is the number of descendants chosen in step 2. Summing over all values of l returns $P_a(i)$ the probability that any node i is selected as an ambassador, thus

$$P_a(i) = \sum_{l=1}^{\infty} p_l \frac{T_l}{N} \times \frac{1}{T_l}$$

$$= \frac{1}{N}.$$ (9)

Suppose i has in-degree k, and that as per step 2 only nodes with out-degree l or greater can be selected as an ambassador. Suppose also that the ambassador is an ancestor of i and has out-degree s where $s \geq l$ (see Fig. 3). The expectation of the
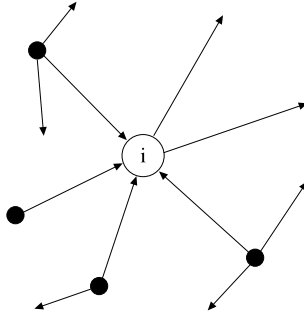
Fig. 3. At each time-step i can be selected as the initial node, alternatively one of the black nodes may be selected and then i is selected as one of its descendants.

number of nodes that satisfy these conditions is $kP_{out}(s)$. The probability that each one is selected is $1/N$ given by Eq. (9). Once selected, the probability that of the s descendants i is one of those selected in Step 3, is l/s. Taking the product and summing over all l and all possible values of s returns $P_d(i; k)$ the probability that any node i with degree k is selected as a descendant, therefore

$$P_d(i; k) = \frac{k}{N}\Phi \tag{10}$$

where

$$\Phi(p, q) = \sum_{l=1}^{\infty} \sum_{s=l}^{\infty} \frac{lp_l P_{out}(s)}{s}. \tag{11}$$

The probability of a node i with degree k being linked to during step 2 or 3 of the process is $P_a(i) + P_d(i; k)$. Summing again over all possible values of m, the probability that the degree of i will increase by 1 during any iteration is

$$P(i; k) = \frac{\langle m \rangle}{N}(1 + k\Phi) \tag{12}$$

where

$$\langle m \rangle = \sum_{m=1}^{\infty} mq_m. \tag{13}$$

The associated rate equation is constructed in exactly the same way as Eq. (2), thus

$$\frac{\partial I_k}{\partial N} = \frac{\langle m \rangle}{N}[(1 + (k-1)\Phi)I_{k-1} - (1 + k\Phi)I_k]. \tag{14}$$

Letting $P_{in}(k) = I_k/N$ be the proportion of nodes that have in-degree at large N, Eq. (14) becomes

$$\left(\frac{1 + \langle m \rangle}{\langle m \rangle \Phi} + k\right) P_{in}(k) = \left(\frac{1 - \Phi}{\Phi} + k\right) P_{in}(k - 1). \tag{15}$$

The rate equation for I_0 solves to find $P(0) = 1/(1 + \langle m \rangle)$ and thus

$$P_{in}(k) = \frac{1}{\langle m \rangle + 1} \frac{\Gamma[(1 + \langle m \rangle)/\langle m \rangle \Phi]\Gamma[k + (1 - \Phi)/\Phi]}{\Gamma[(1 - \Phi)/\Phi]\Gamma[k + (1 + \langle m \rangle)/\langle m \rangle \Phi]}. \tag{16}$$

For large values of k, $P_{in}(k)$ has a power-law form

$$P_{in}(k) \sim k^{-\gamma} \quad \text{where } \gamma = 1 + \frac{1}{\langle m \rangle \Phi}. \tag{17}$$

5. Clustering

The clustering coefficient of a node i is defined as the number of edges between the neighbours of i divided by the number of pairs of nodes from the neighbours of i. If node i has d neighbours (ancestors and descendants) then this is

$$C_i = \frac{2E_i}{d(d - 1)} \tag{18}$$

where E_i is the number of edges between the neighbours of i. Let $E(k)$ be the expectation of E_i when i has in-degree k, also let $\Theta(k)$ be the expectation of the number of times i has been selected as the ambassador node during step 2 of any previous iteration. From Eqs. (9) and (10) we see that the kth edge is $(k-1)\Phi$ times more likely to be added as a result of one of i's neighbours being an ambassador rather than i being selected as an ambassador itself, so

$$\Theta(k) = \sum_{i=0}^{k-1} \frac{1}{1+i\Phi}. \tag{19}$$

We find $\bar{C}$, the mean of C_i over all nodes i in the network we studied in Section 3 where $p_r = \delta_{rl}$ and $q_r = \delta_{rm}$. This is the sum over all k of the product of $P_{\text{in}}(k)$ given by Eq. (6), and $C(k)$ the expectation of the clustering of a node of degree k. The contribution to E_i made by each neighbour j of i depends on the way in which the link was originally created, there are four cases to be considered. The first is where the link $i \to j$ was added when i was introduced to the network, in this case the expected contribution to E_i is the number of edges in the reference graph of i (see Fig. 1). The second case is where i was selected as an ambassador and l edges are added to E_i. In the third case D_a edges are counted for those that were added when i was selected as a descendant of the ambassador j where the link from j to i was originally formed when i was selected as an ambassador in a previous iteration. Lastly, D_d edges are counted for those that were added when i was selected as a descendant of the ambassador j where the link from j to i was formed when i was originally selected as a descendant. Then

$$E(k) = E(0) + \Theta l + (k - \Theta)\left[\frac{\Theta}{k} D_a + \frac{1-\Theta}{k} D_d \right]. \tag{20}$$

When a node i is added to the network, E_i includes the edges between each ambassador node and its descendants as well as the edges between those descendants. The probability that an edge exists between two descendants of the same node is $C(0)$ so the expectation of E_i is

$$E(0) = m\left(l + \binom{l}{2} C(0) \right). \tag{21}$$

Combining this with Eq. (18) and solving gives

$$C(0) = \frac{2l}{(m-1)(l+1)^2 + 2l}. \tag{22}$$

In the instance where an ambassador node j is selected and i is linked to as one of j's descendants, the new node will link to a further $l-1$ neighbours of $m(l+1) - 1$ possible descendants of j, those that are also neighbours of i will be counted in E_i. If j originally formed a link with i by selecting i as an ambassador, then l of i's descendants are also descendants of j, hence the expectation of the number of neighbours of i that are linked to is

$$D_a = \frac{l(l-1)}{m(l+1) - 1}. \tag{23}$$

If j originally formed a link with i by selecting i as the descendant of some other node, the expected number of links between j and any of i's neighbours is $1 + (l-1)C(0)$ so the expectation of the number of i's neighbours linked to is

$$D_d = \frac{[1 + (l-1)C(0)](l-1)}{m(l+1) - 1}. \tag{24}$$

Combining Eqs. (18)–(20), (22)–(24) gives an expression for the clustering of a node of in-degree k in terms of m and l ($l, m \geq 1$), multiplying by $P_{\text{in}}(k)$ given by Eq. (6) and summing over all k gives the mean clustering for the entire network. The clustering coefficient tends to 0 as m grows large. As l grows large the clustering also tends to 0 except when m is equal to one, in which case it tends to 1 (see Fig. 4).

6. Numerical results

It should be emphasized that the results found in previous sections are mean field approximations as N tends to infinity, it therefore cannot immediately be assumed that the derived results will be a fair description of any individual network grown following the proposed process. We consider the following:

1. Correlations between out-degree of a node and the in-degree of its descendants. Specifically in Eq. (10) where it is assumed that the out-degree of the neighbours of node i (i.e the black nodes in Fig. 3) follow the distribution $P_{\text{out}}(s)$ regardless of the in-degree of i. In reality this might not be the case; imagine, for example, a node j with relatively large out-degree and i as one of its descendants, selecting j as an ambassador in future iterations is relatively unlikely to result in selecting i again unless the new node also has large out-degree (more specifically a large value of l in step 2 of the iteration), so the expectation is for i to have few ancestors each with large out-degree (the opposite is true if the out-degree of j is small). The effect of this has not been considered analytically, instead we show numerically that in practice there is no significant deviation from the mean field result.
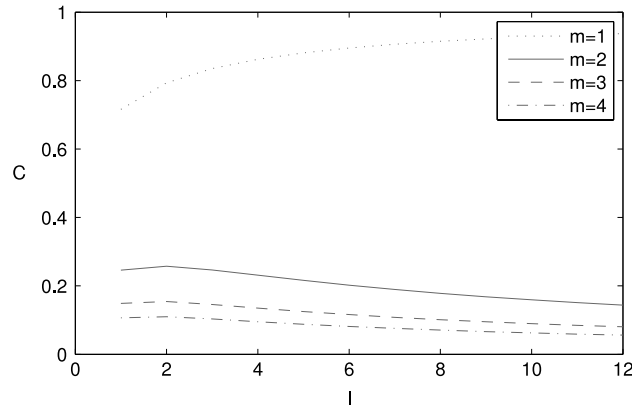
Fig. 4. Using the formulae in Section 5 the average clustering coefficient is plotted on the vertical axis for the first 4 values of m, against l on the horizontal axis. The clustering tends to zero as l grows large for all values of m with the exception of $m = 1$.
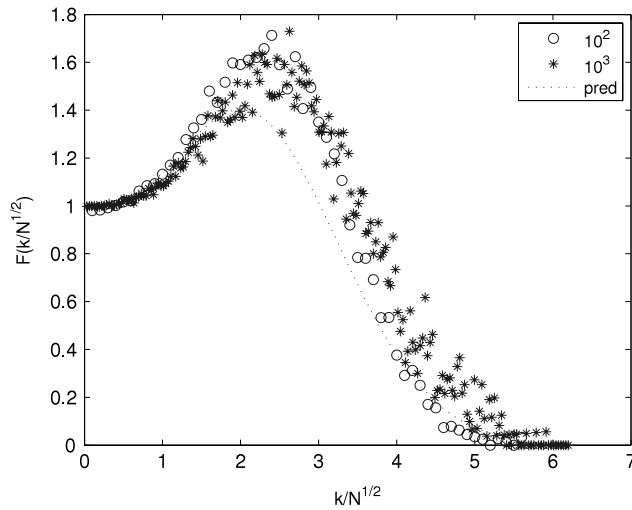

Fig. 5. The correction function $F(\xi)$ in Eq. (25) for the simplest case of the model. Here $m = 1, l = 1$ and an initial condition of three nodes, one with out-degree 2 connected to the other two. The dashed line represents the analytical prediction from Ref. [29].

2. Finite size corrections. For finite networks of size N, the existence of a largest degree $k_{\max}$ means the power-law degree distribution must fail around the largest values of k. These effects have been investigated for particular classes of preferential attachment based network [14,29]. Once the asymptotic mean field solution $P(k)$ is known, the solution to average degree distribution on a network of size N is

$$N_k(N) \simeq NP(k)F(\xi) \quad \text{where } \xi = k/k_{\max}. \tag{25}$$

From the generated data (discussed below) we observe in Fig. 5 a similar form of scaling function $F(\xi)$ as observed in Refs. [14,29]. The function $F(\xi)$ can be derived by considering the average of all possible values of $N_k(N)$ for every N starting from an initial value for $N_1(1)$, under the specific circumstances however, the initial conditions must be chosen carefully for each possible choice of our parameters. It is impractical to derive $F(\xi)$ for every possibility here, instead we show that the model passes a suitable goodness-of-fit test even when finite size effects are neglected.

In the numerical tests we grew a network in three phases, initially a small number of nodes with large out degree are created (the degree must be large enough to allow T_l to be non-zero for all l), then a phase of creating new nodes with a random number of out links to randomly selected nodes already in the network, finally the process described in Section 2 is applied for a large number of iterations. To assess the goodness-of-fit of the results in Eqs. (6) and (16) we compare the degree distribution of a simulated network of size N to the distribution given by drawing N values from a pseudo-random number generator adapted to output the value k with probability given by Eq. (16). The degree distribution of the simulated network is then compared against the mean field prediction Eq. (16) using a suitable measure of similarity, in this case we choose the Kolmogorov–Smirnov statistic. Lastly, over a large number of trials (we chose 10^3) the pseudo-random distribution is measured against the model, the p-value for this test is the proportion of trials in which the simulated data is closer to the model (i.e. a lower KS statistic) than the random data. Here we have followed the methodology of Ref. [30], developed for
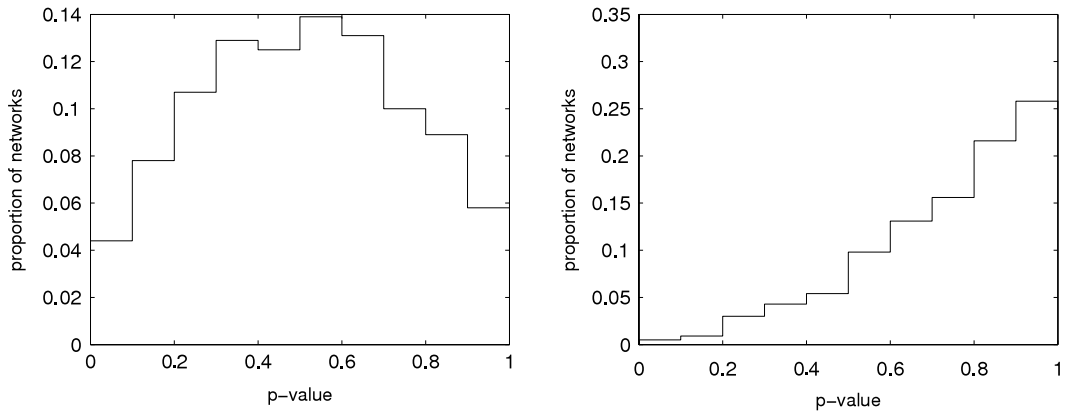
Fig. 6. Histograms showing the numerically derived frequencies that given p-values were achieved. The networks in the figure to the left use the parameters $p_r = (1/3)(\delta_{r1} + \delta_{r2} + \delta_{r3})$ and $q_r = (1/3)(\delta_{r2} + \delta_{r3} + \delta_{r4})$ and the networks in the figure to the right use the parameters $p_r = \delta_{r1}$ and $q_r = \delta_{r1}$, each network has $N = 10^3$.
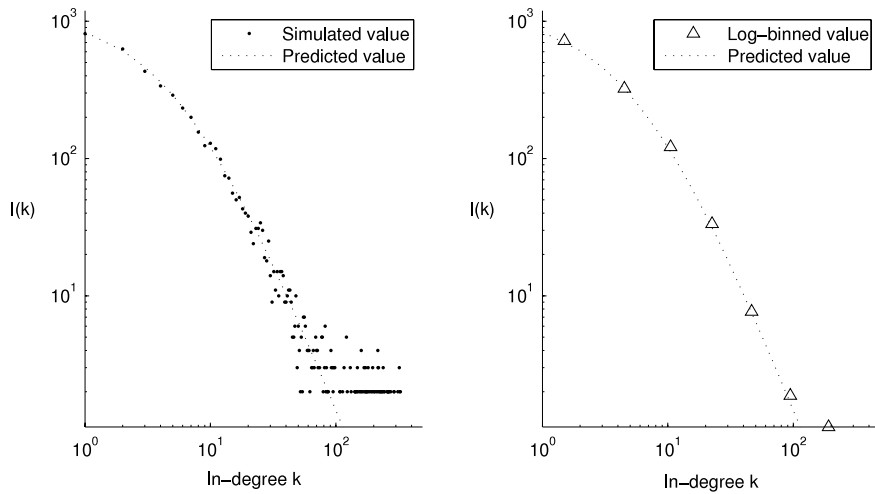

Fig. 7. The number of nodes $I(k)$ of in-degree k for each value of k, results here are taken from the simulation of the model when $p_r = \delta_{r3}$ and $q_r = \delta_{r4}$, the dotted line shows the predicted result derived from Eq. (6), the right hand figure shows that the log-binned values agree very well with the prediction with the exception of the largest values of k. This network contained 6×10^3 nodes, the first 300 were added randomly.

use in empirical studies where the data are not likely to be as clean as those generated in a computer simulation. The authors suggest that a p-value greater than 0.1 is evidence enough for the model to be accepted. We ran this test for 10^3 networks first using the pair of distributions $p_r = (1/3)(\delta_{r1} + \delta_{r2} + \delta_{r3})$ and $q_r = (1/3)(\delta_{r2} + \delta_{r3} + \delta_{r4})$ then another 10^3 networks using $p_r = \delta_{r3}$ and $q_r = \delta_{r4}$. Fig. 6 shows the proportion of these trials that achieved particular p-values, while the p-value varies greatly, only a very small proportion are less than 0.1.

We ran the simulation for a large number of different distributions p_l and q_m and found that the numerical results agreed with the analytically derived formulae, Figs. 7 and 8 show two typical examples. The log-binned values are the means of I_k over a ranges of k that increases logarithmically with k. In these examples the first bin is just the first value of I_1, the second is the mean of I_2 and I_3, the third is the mean of the next 4 values and so on. We were able to compute the clustering coefficient only for networks no more than approximately 10^3 nodes, we found that for networks where the out-degree of the nodes is large the simulated result tended to be higher than the analytical result, this exposes the assumption in the analytical calculations that ambassador nodes will not be close to each other in the network. This discrepancy gets smaller as the network grows larger as one would expect.

7. Remarks

There are two particular strengths of this model that are worth highlighting. The first is tunability; the feature that a wide range of results for the clustering and power-law exponent can be achieved by inputting the appropriate parameter values. In the simplified model l can be tuned to achieve any exponent between 2 and 3, by adjusting m the clustering is tunable to a restricted range of values (see Fig. 4). It is not difficult to find distributions in the full model that allow the
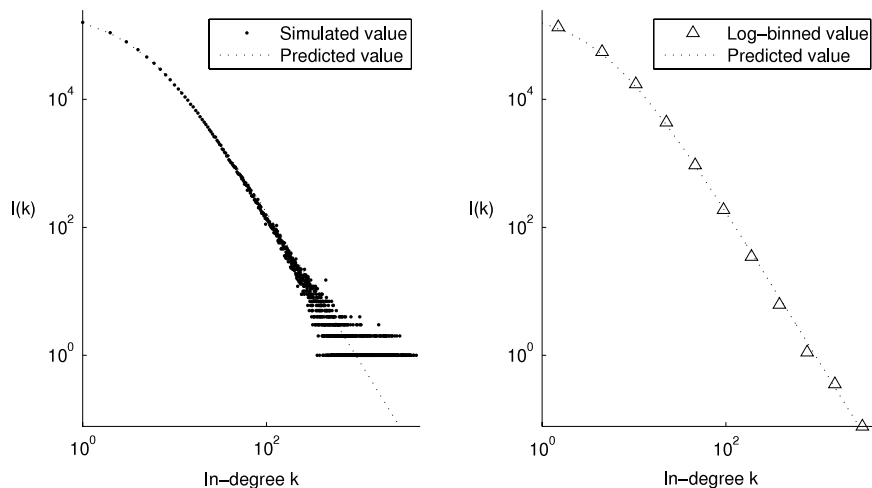
Fig. 8. The in-degree from the simulation of the model when $p_r = (1/3)(\delta_{r1} + \delta_{r2} + \delta_{r3})$ and $q_r = (1/3)(\delta_{r2} + \delta_{r3} + \delta_{r4})$, the dotted line shows the predicted result given by Eq. (16). This network contained 10^6 nodes, the first 10^3 were added randomly.

clustering to be tuned to any value between 0 and 1, however as we showed in Section 4 the exponent in the distribution depends on both m and l. Tunable networks are particularly useful to study processes on networks such as epidemic spread; since the results they obtain depend largely on the topologies of the underlying networks, adjustability allows the extent of the effects of clustering and degree distribution to be analysed in greater detail [31]. The second strength of this model is its generality; the property that there are a wide range of parameter values that can be used as input to the model. As there are no restrictions on the probability distributions involved it is possible to choose those that most closely match the empirical data. A possible analysis would involve approximating the distributions p and q by measuring the distribution of the number of citations in the bibliography of each paper in a dataset and the distribution of citations that are also included in the bibliographies of other cited papers.

Acknowledgment

References

[1] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, A. Barabási, The large-scale organization of metabolic networks, Nature 407 (2000) 651–654.
[2] T. Ohnishi, H. Takayasu, M. Takayasu, Hubs and authorities on Japanese inter-firm network: characterization of nodes in very large directed networks, Progress of Theoretical Physics Supplement 179 (2009) 157–166.
[3] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical Review Letters 86 (2001) 3200–3203.
[4] A.-L. Barabsi, R. Albert, H. Jeong, Scale-free characteristics of random networks: the topology of the world-wide web, Physica A: Statistical Mechanics and its Applications 281 (2000) 69–77.
[5] G. Kossinets, D. Watts, Empirical analysis of an evolving social network, Science 311 (2006) 88–90.
[6] B. Jiang, A topological pattern of urban street networks: universality and peculiarity, Physica A: Statistical Mechanics and its Applications 384 (2007) 647–655.
[7] D.J.d.S. Price, Networks of scientific papers, Science 149 (1965) 510–515.
[8] S. Redner, How popular is your paper? an empirical study of the citation distribution, The European Physical Journal B. Condensed Matter and Complex Systems 4 (1998) 131–134. http://dx.doi.org/10.1007/s100510050359.
[9] D.D.S. Price, A general theory of bibliometric and other cumulative advantage processes, Journal of the American Society for Information Science 27 (1976) 292–306.
[10] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47–97.
[11] P.L. Krapivsky, S. Redner, Organization of growing random networks, Physical Review E 63 (2001) 066123.
[12] A. Vázquez, Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations, Physical Review E 67 (2003) 056104.
[13] J. Saramäki, K. Kaski, Scale-free networks generated by random walkers, Physica A: Statistical Mechanics and its Applications 341 (2004) 80–86.
[14] T. Evans, J. Saramäki, Scale-free networks from self-organization, Physical Review E 72 (2005) 026138.
[15] L. Bornmann, H.-D. Daniel, What do citation counts measure? a review of studies on citing behavior, Journal of Documentation 64 (2008) 45–80.
[16] L. Leydesdorff, S. Milojević, Scientometrics, Elsevier, 2012, pp. 1–20.
[17] K. Börner, J.T. Maru, R.L. Goldstone, The simultaneous evolution of author and paper networks, Proceedings of the National Academy of Sciences of the United States of America 101 (2004) 5266–5273.
[18] M. Wang, G. Yu, D. Yu, Measuring the preferential attachment mechanism in citation networks, Physica A: Statistical Mechanics and its Applications 387 (2008) 4692–4698.
[19] Z.-X. Wu, P. Holme, Modeling scientific-citation patterns and other triangle-rich acyclic networks, Physical Review E 80 (2009) 037101.
[20] P. Holme, B.J. Kim, Growing scale-free networks with tunable clustering, Physical Review E 65 (2002) 026107.
[21] R. Albert, A. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47.
[22] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, ACM, New York, NY, USA, 2005, pp. 177–187.

[23] F.-X. Ren, H.-W. Shen, X.-Q. Cheng, Modeling the clustering in citation networks, Physica A: Statistical Mechanics and its Applications 391 (2012) 3533–3539.
[24] M.V. Simkin, V.P. Roychowdhury, Read before you cite! 2002. ArXiv preprint cond-mat/0212043.
[25] M.V. Simkin, V.P. Roychowdhury, Copied citations create renowned papers? 2003. ArXiv preprint cond-mat/0305150.
[26] A. Vazquez, Statistics of citation networks, 2001. ArXiv preprint cond-mat/0105031.
[27] B. Tadic, Dynamics of directed graphs: the world-wide web, Physica A: Statistical Mechanics and its Applications 293 (2001) 273–284.
[28] P.L. Krapivsky, G.J. Rodgers, S. Redner, Degree distributions of growing networks, Physical Review Letters 86 (2001) 5401–5404.
[29] P. Krapivsky, S. Redner, Finiteness and fluctuations in growing networks, Journal of Physics A: Mathematical and General 35 (2002) 9517.
[30] A. Clauset, C.R. Shalizi, M.E. Newman, Power-law distributions in empirical data, SIAM Review 51 (2009) 661–703.
[31] X. Wu, Z. Liu, How community structure influences epidemic spread in social networks, Physica A: Statistical Mechanics and its Applications 387 (2008) 623–630.