A general model of web graphs

Colin Cooper[*] Alan Frieze[†]

July 4, 2001

Abstract

We describe a general model of a random graph whose degree sequence obeys a power law. Such laws have recently been observed in graphs associated with the world wide web.

1 Introduction

Essentially, there are three types of random graph or digraph used to describe *small world phenomena*; see for example [11] for an introduction to this topic. These are:

(i) *Web graphs* : Graphs evolving by the addition of new vertices and/or edges at each step t. See eg: [1], [3], [5], [6], [7], [10], [15], [16].

(ii) *Alpha-Beta graphs* : Standard random graph models with atypical degree sequences. See eg: [2], [8], [17], [18].

(iii) *Lattice graphs* : Graphs generated by perturbing regular lattices. See eg: [4], [13].

We describe the evolution of a random (multi-)graph $G(t)$ which is an example of the type of model referred to as a web graph.

Initially, at step $t = 0$, there is a single vertex v_0. At any step $t = 1, 2, ..., T, ...$ there is a birth process in which either new vertices or new edges are added. Specifically, either a procedure NEW is followed with probability $1 - \alpha$, or a procedure OLD is followed with probability α. In procedure NEW, a new vertex v is added to $G(t-1)$ with one or more edges added between v and $G(t-1)$. In procedure OLD, an existing vertex v is selected and extra edges are added at v.

The recipe for adding edges typically permits the choice of initial vertex v (in the case of OLD) and the terminal vertices (in both cases) to be made either u.a.r or according to vertex degree, or a mixture of these two based on further sampling. The number of edges added to vertex v at step t by the procedures (NEW, OLD) is given by distributions specific to the procedure. The details of these choices are given below.

A question arises about our model: Should we regard the edges as directed or undirected in relation to the sampling procedures NEW, OLD? We note that the edges have an intrinsic direction arising

[*] School of Mathematical and Computing Sciences, Goldsmiths College, University of London, London SE14 6NW, UK. E-mail: c.cooper@gold.ac.uk.. Research supported in part by the STORM Research Centre, UNL.

[†] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, U.S.A., Supported in part by NSF grant CCR-9818411. E-mail: alan@random.math.cmu.edu.

from the way they are inserted, which we can ignore or not as we please. We estimate $\bar{d}_k$, the limiting expected proportion of vertices of degree k as $t \to \infty$, and we use sampling procedures based on vertex degree. Specifically, we consider the following models:

(i) Undirected model: Sampling procedure based on vertex degree.

(ii) Directed out-model: Sampling procedure for out-edges based on out-degree.

(iii) Directed in-model: Sampling procedure for in-edges based on in-degree.

We prove for these models, that provided some degree weighted sampling occurs then $\bar{d}_k \approx Ck^{-x}$, where $x > 1$ is a function of the parameters of the model. See (3, 29, 30) for the explicit functional form of x. In contrast, if all the vertex sampling is u.a.r then $\bar{d}_k$ behaves geometrically, as in more well known models of random graphs.

We devote most of the paper to the undirected model. The other models can easily be analysed as variants of the undirected case, and are covered briefly in Section 4.

Sampling based on a mixture of in-degree and out-degree, and the estimation of $\bar{d}_{i,j}$, the expected proportion of vertices of in-degree i and out-degree j is not considered here, but is the subject of a subsequent paper [9].

The parameters of the undirected model

Our undirected model $G(t)$ has sampling parameters $\alpha, \beta, \gamma, \delta, \boldsymbol{p}, \boldsymbol{q}$ whose meaning is given below:

Choice of procedure at step t.
α: Probability that an OLD node generates edges.
$1-\alpha$: Probability that a NEW node is created.
Procedure NEW
$\boldsymbol{p} = (p_i : i \geq 1)$: Probability that new node generates i new edges.
β: Probability that choices of terminal vertices are made uniformly.
$1-\beta$: Probability that choices of terminal vertices are made according to degree.
Procedure OLD
$\boldsymbol{q} = (q_i : i \geq 1)$: Probability that old node generates i new edges.
δ: Probability that the initial node is selected uniformly.
$1-\delta$: Probability that the initial node is selected according to degree.
γ: Probability that choices of terminal vertices are made uniformly.
$1-\gamma$: Probability that choices of terminal vertices are made according to degree.

The models we study here require $\alpha < 1$, always. We assume a *finiteness* condition for the distributions $\{p_j\}$, $\{q_j\}$. This means that there exists j_0, j_1 such that $p_j = 0,\ j \geq j_0$ and $q_j = 0,\ j > j_1$.

The model creates edges in the following way: An initial vertex v is selected. If the terminal vertex w is chosen u.a.r, we say v is *assigned* to w. If the terminal vertex w is chosen according to its vertex degree, we say v is *copied* to w. In either case the edge has an intrinsic direction (v, w), which we may choose to ignore. We note that sampling according to vertex degree is equivalent to selecting an edge u.a.r and then selecting an endpoint u.a.r.

Copying

The papers [15], [16] introduce a copying model in which a new vertex v chooses an old vertex w and selects (copies) a randomly chosen set of out-neighbours of w to be its own out-neighbours. This construction leads to a larger number of small complete bipartite subgraphs than would be obtained by purely random selection of endpoints. This is an attempt to explain the occurrence of the number of small complete bipartite subgraphs found in trawls of the web.

Since our focus is on the degree sequence and since, as we show below, the construction above does not lead to a fundamentally different recurrence we will not explicitly use this method of generating edges. We will continue to use the term copy to mean that vertices are chosen with probability proportional to degree.

In terms of expected degree sequence, this definition of copying is equivalent to the case of the copying model in [16] which functions as follows: A new vertex u of out-degree i is added at each step. The choice of out-edges of u is made as follows. Firstly i provisional vertices are selected u.a.r. Now, independently for each of these i provisional vertices the following choice is made. With probability β vertex v is retained and the edge (u, v) inserted. Or, with probability $(1-\beta)$ a copied edge (u, w) is inserted instead, where w is the terminal vertex of a uniformly selected out-edge of v. This process of copying is equivalent in terms of expected degree sequence, to the version of copying we propose above for (the in-directed variant of) our model, namely selecting the terminal vertex of a random edge. For, in our model

$$\mathbf{Pr}(w \text{ is the terminal vertex of a u.a.r edge}) = \frac{d^-(w)}{|E|},$$

and in the copying model

$$\mathbf{Pr}(w \text{ is selected by copying an edge }) = \frac{d^-(w)}{|V|}\frac{1}{i},$$

where $|E| = i|V|$.

For directed copying models the expected proportion d_k^- of vertices of in-degree k is asymptotic to $Ck^{-\frac{2-\beta}{1-\beta}}$. The in-directed variant of our model gives the same result ($d_k^- \sim C'k^{-\frac{2-\beta}{1-\beta}}$) for the expected proportion of vertices of in-degree k.

In the case where the out-degree is not a constant value i,

$$\mathbf{Pr}(w \text{ is selected on copying from } u) = \frac{1}{|V|}\sum_{v \in N^-(w)} \frac{1}{d^+(v)}$$

which seems a difficult quantity to deal with, especially if there is correlation between the in-degree of w and the out-degree of v. Selecting the terminal vertex of a random edge remains an easily accessible sampling procedure of an equivalent nature.

We consider the general undirected web model to be intrinsically interesting, aside from applications to the www. Moreover, although the edges of typical www graph are directed, the idea of an undirected model has many attractions. For example, the problem of new vertices. A new entrant to the www may have its edge directed towards either a site (vertex) or an idea (terminal vertex of an edge). Existing sites (and ideas) also produce new nodes and direct edges *towards* them. Thus edges incident with new nodes have no overall preferred direction.

Notation
Let $\mu_p = \sum_{j=0}^{j_0} jp_j$, $\mu_q = \sum_{j=0}^{j_1} iq_j$ and let $\theta = 2((1-\alpha)\mu_p + \alpha\mu_q)$. To simplify subsequent notation,

we transform the parameters as follows:

$$\begin{aligned}
a &= 1+\beta\mu_p+\frac{\alpha\gamma\mu_q}{1-\alpha}+\frac{\alpha\delta}{1-\alpha},\\
b &= \frac{(1-\alpha)(1-\beta)\mu_p}{\theta}+\frac{\alpha(1-\gamma)\mu_q}{\theta}+\frac{\alpha(1-\delta)}{\theta},\\
c &= \beta\mu_p+\frac{\alpha\gamma\mu_q}{1-\alpha},\\
d &= \frac{(1-\alpha)(1-\beta)\mu_p}{\theta}+\frac{\alpha(1-\gamma)\mu_q}{\theta},\\
e &= \frac{\alpha\delta}{1-\alpha}, \qquad\qquad f=\frac{\alpha(1-\delta)}{\theta}.
\end{aligned}$$

We note that

$$c+e=a-1 \text{ and } b=d+f. \tag{1}$$

Now define the sequence $(d_0, d_1, ..., d_k, ...)$ by $d_0 = 0$ and for $k \geq 1$

$$d_k(a+bk)=(1-\alpha)p_k+(c+d(k-1))d_{k-1}+\sum_{j=1}^{k-1}(e+f(k-j))q_jd_{k-j}. \tag{2}$$

Since $a \geq 1$, this system of equations has a unique solution.

Statement of results

The main quantity we study is the random variable $D_k(t)$, the number of vertices of degree k at step t. We let $\overline{D}_k(t) = \mathbf{E}(D_k(t))$. We prove that for small k, $\overline{D}_k(t) \approx d_k t$ as $t \to \infty$.

Theorem 1. *There exists a constant $M > 0$ such that for $t, k = 1, 2, \ldots$,*

$$|\overline{D}_k(t)-td_k| \leq Mt^{1/2}\log t.$$

The number of vertices $\nu(t)$ at step t is **whp** asymptotic to $(1-\alpha)t$, see (4). It follows that

$$\bar{d}_k=\frac{d_k}{1-\alpha}.$$

The next theorem summarises what we know about the d_k:

Theorem 2.

(i) *$Ak^{-\zeta} \leq d_k \leq B\min\{k^{-1}, k^{-\zeta/j_1}\}$ where $\zeta = (1+d+f\mu_q)/(d+f)$.*

(ii) *If $j_1 = 1$ then $d_k \sim Ck^{-(1+1/(d+f))}$.*

(iii) *If $f = 0$ then $d_k \sim Ck^{-(1+1/d)}$.*

(iv) *If the* SOLUTION CONDITIONS *hold then*

$$d_k=C\left(1+O\left(\frac{1}{k}\right)\right)k^{-x},$$

where C is constant and

$$x=1+\frac{1}{d+f\mu_q}. \tag{3}$$

We say that $\{q_j : j = 1, ..., j_1\}$ is *periodic* if there exists $m > 1$ such that $q_j = 0$ unless $j \in \{m, 2m, 3m, \dots\}$.

Let

$$\phi_1(y) = y^{j_1} - \left(\frac{d + q_1 f}{b} y^{j_1 - 1} + \frac{q_2 f}{b} y^{j_1 - 2} + \cdots + \frac{q_{j_1} f}{b} \right).$$

Our SOLUTION CONDITIONS are:

S(i) $f > 0$ and either (a) $d + q_1 f > 0$ or (b) $\{q_j\}$ is not periodic.

S(ii) The polynomial $\phi_1(y)$ has no repeated roots.

We can also prove the following concentration result:

Theorem 3. *For any* $u > 0$,

$$\mathbf{Pr}(|D_k(t) - \overline{D}_k(t)| \geq u) \leq \exp\left\{ -\frac{u^2}{2tj_{\max}} \right\}$$

where $j_{\max} = \max\{j_0, j_1\}$.

2 Evolution of the degree sequence of $G(t)$

Let $\nu(t) = |V(t)|$ be the number of vertices and let $\eta(t) = |2E(t)|$ be the total degree of the graph at the end of step t. $\mathbf{E}\nu(t) = (1-\alpha)t$ and $\mathbf{E}\eta(t) = \theta t$. The random variables $\nu(t)$, $\eta(t)$ are sharply concentrated provided $t \to \infty$. Indeed $\nu(t)$ has binomial distribution $B(t, 1-\alpha)$ and so by the Chernoff bounds,

$$\mathbf{Pr}(|\nu(t) - (1-\alpha)t| \geq t^{1/2} \log t) = O(t^{-K}) \tag{4}$$

for any constant $K > 0$.

Similarly, $\eta(t)$ has expectation θt and is the sum of t independent random variables, each bounded by $\max\{j_0, j_1\}$. Hence, by Hoeffding's theorem [12],

$$\mathbf{Pr}(|\eta(t) - \theta t| \geq t^{1/2} \log t) = O(t^{-K}) \tag{5}$$

for any constant $K > 0$.

We remind the reader that $D_k(t)$ is the number of vertices of degree k at step t and that $\overline{D}_k(t)$ is its expectation. Here $\overline{D}_0(t) = 0$ for all t, $\overline{D}_1(0) = 1$, $\overline{D}_k(0) = 0$, $k \geq 2$. Then, after using (4) and (5) we see that

$$\overline{D}_k(t+1) = \overline{D}_k(t) + (1-\alpha)p_k + O(t^{-1/2}\log t) \tag{6}$$

$$+ (1-\alpha) \sum_{j=1}^{j_0} p_j \left(\frac{\beta j \overline{D}_{k-1}(t)}{(1-\alpha)t} - \frac{\beta j \overline{D}_k(t)}{(1-\alpha)t} + (1-\beta) \left(\frac{j(k-1)\overline{D}_{k-1}(t)}{\theta t} - \frac{jk\overline{D}_k(t)}{\theta t} \right) \right) \tag{7}$$

$$- \alpha \left(\frac{\delta \overline{D}_k(t)}{(1-\alpha)t} + \frac{(1-\delta)k\overline{D}_k(t)}{\theta t} \right) + \alpha \sum_{j=1}^{j_1} q_j \left(\frac{\delta \overline{D}_{k-j}(t)}{(1-\alpha)t} + \frac{(1-\delta)(k-j)\overline{D}_{k-j}(t)}{\theta t} \right) \tag{8}$$

$$+ \alpha \sum_{j=1}^{j_1} j q_j \left(\gamma \left(\frac{\overline{D}_{k-1}(t)}{(1-\alpha)t} - \frac{\overline{D}_k(t)}{(1-\alpha)t} \right) + (1-\gamma) \left(\frac{(k-1)\overline{D}_{k-1}(t)}{\theta t} - \frac{k\overline{D}_k(t)}{\theta t} \right) \right). \tag{9}$$

Here (7), (8), (9) are (respectively) the main terms of the change in the expected number of vertices of degree k due to the effect on: terminal vertices in NEW, the initial vertex in OLD and the terminal vertices in OLD. Rearranging the right hand side, we find:

$$
\begin{aligned}
&\overline{D}_k(t+1) = \overline{D}_k(t) + (1-\alpha)p_k + O(t^{-1/2}\log t) \\
&- \frac{\overline{D}_k(t)}{t}\left(\beta\mu_p + \frac{\alpha\gamma\mu_q}{1-\alpha} + \frac{\alpha\delta}{1-\alpha} + \frac{(1-\alpha)(1-\beta)\mu_p k}{\theta} + \frac{\alpha(1-\gamma)\mu_q k}{\theta} + \frac{\alpha(1-\delta)k}{\theta}\right) \\
&+ \frac{\overline{D}_{k-1(t)}}{t}\left(\beta\mu_p + \frac{\alpha\gamma\mu_q}{1-\alpha} + \frac{(1-\alpha)(1-\beta)\mu_p(k-1)}{\theta} + \frac{\alpha(1-\gamma)\mu_q(k-1)}{\theta}\right) \\
&+ \sum_{j=1}^{j_1} q_j \frac{\overline{D}_{k-j}(t)}{t}\left(\frac{\alpha\delta}{1-\alpha} + \frac{\alpha(1-\delta)(k-j)}{\theta}\right).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\overline{D}_k(t+1) = \overline{D}_k(t) + (1-\alpha)p_k + O(t^{-1/2}\log t) \\
&+ \frac{1}{t}\left((1-(a+bk))\overline{D}_k(t) + (c+d(k-1))\overline{D}_{k-1}(t) + \sum_{j=1}^{j_1} q_j(e+f(k-j))\overline{D}_{k-j}(t)\right).
\end{aligned} \tag{10}
$$

The following upper bound on d_k is claimed in Theorem 2(i).

Lemma 1. *There exists a constant $A > 0$ such that the solution of (2) satisfies $d_k \leq \frac{A}{k}$.*

Proof We proceed by induction on k and assume that k is sufficiently large. Small k can be dealt with by adjusting A. Then $p_k = 0$ and so

$$
\begin{aligned}
(a+bk)d_k &\leq (c+d(k-1))\frac{A}{k-1} + \sum_{j=1}^{j_1}(e+f(k-j))q_j\frac{A}{k-j} \\
&\leq A(d+f) + \frac{A(c+e)}{k-j_1} \\
&= Ab + \frac{A(a-1)}{k-j_1},
\end{aligned}
$$

from (1). So

$$
\begin{aligned}
d_k - \frac{A}{k} &\leq \frac{Ab}{a+bk} + \frac{A(a-1)}{(k-j_1)(a+bk)} - \frac{A}{k} \\
&= \frac{A(a-1)}{(k-j_1)(a+bk)} - \frac{Aa}{k(a+bk)} \\
&\leq 0.
\end{aligned}
$$

□

We can now prove Theorem 1, stated here for conveneience.

Theorem 4. *There exists a constant $M > 0$ such that for $t, k = 1, 2, \ldots,$*

$$
|\overline{D}_k(t) - td_k| \leq Mt^{1/2}\log t. \tag{11}
$$

Proof Let $\Delta_k(t) = \overline{D}_k(t) - td_k$. It follows from (10) and (2) that

$$\Delta_k(t) = \Delta_k(t-1)\left(1 - \frac{a+bk-1}{t}\right) + O(t^{-1/2}\log t) + \\ t^{-1}\left((c+d(k-1))\Delta_{k-1}(t-1) + \sum_{j=1}^{j_1}(e+f(k-j))q_j\Delta_{k-j}(t-1)\right). \quad (12)$$

Let L denote the hidden constant in $O(t^{-1/2}\log t)$. Assume that t, k are sufficiently large (we can adjust M to deal with small values of t, k). Let $k_0(t) = \lfloor\frac{t+1-b}{a}\rfloor$. If $k > k_0(t)$ then we observe that (i) $D_k(t) \leq \frac{tj_0}{k_0(t)} = O(1)$ and (ii) $td_k \leq t\frac{A}{k_0(t)} = O(1)$ and so (11) holds trivially.

Assume inductively that $\Delta_\kappa(\tau) \leq M\tau^{1/2}\log\tau$ for $\kappa + \tau < k + t$ and that $k \leq k_0(t)$. Then (12) and $k \leq k_0$ implies that for M large,

$$\begin{aligned} |\Delta_k(t)| &\leq L\frac{\log t}{t^{1/2}} + M(t-1)^{1/2}\log t\left(1 + \frac{1}{t}\left(c + dk + \sum_{j=1}^{j_1}(e+fk)q_j - (a-1+bk)\right)\right) \\ &= L\frac{\log t}{t^{1/2}} + M(t-1)^{1/2}\log t \\ &\leq Mt^{1/2}\log t \end{aligned}$$

provided $M \gg 2L$.

This completes the induction. □

Analysis of the difference equation (2)

Re-writing (2) we see that for $k \geq j_0$, d_k satisfies

$$d_k = d_{k-1}\frac{c+d(k-1)}{a+bk} + \sum_{j=1}^{j_1} d_{k-j}q_j\frac{e+f(k-j)}{a+bk}, \quad (13)$$

which is a linear difference equation with rational coefficients [19]. The general solution for d_k is a power law, i.e. there are constants x_L, x_U such that $Ak^{-x_L} \leq d_k \leq Bk^{-x_U}$. This is established in Lemma 2.

In the cases where $j_1 = 1$ (a new vertex generates a single edge) or $f = 0$ (old initial vertices are chosen u.a.r) a direct solution to (13) can easily be found. In general however, when $d > 0$ or $d = 0$ and $\{q_j\}$ is non-periodic, we use classical results on the solution of Laplace's difference equation, (of which (2) is an example) given in [19].

A general power law bound for d_k

The following lemma completes the proof of Theorem 2(i).

Lemma 2. *Let $p_j = 0, j \geq j_0$ and $q_j = 0, j > j_1$.*

(i) *For all $k \geq j_0$, $d_k > 0$.*

(ii) *then, for $k \geq j_0 + j_1$, there exist constants $C, D > 0$ such that*

$$Ck^{-(1+d+f\mu_q)/b} \leq d_k \leq Dk^{-(1+d+f\mu_q)/bj_1}.$$

Proof

Let I be the first index such that $p_I > 0$, so that, from (2), $d_I > 0$. As it is impossible for both c and d to be zero, the coefficient of d_{k-1} in (2) is non-zero and thus $d_k > 0$ for $k \geq I$.

For $k \geq j_0$ the recurrence (2) satisfies (13), that is

$$d_k = d_{k-1}\frac{c+d(k-1)}{a+bk} + \sum_{j=1}^{j_1} d_{k-j}q_j\frac{e+f(k-j)}{a+bk}.$$

Let $y = 1 + d + f\mu_q$, then

$$\frac{c+d(k-1)}{a+bk} + \sum_{j=1}^{j_1} q_j\frac{e+f(k-j)}{a+bk} = 1 - y/(a+bk) \geq 0$$

and thus

$$\left(1 - \frac{y}{a+bk}\right)\min\{d_{k-1}, ..., d_{k-j_1}\} \leq d_k \leq \left(1 - \frac{y}{a+bk}\right)\max\{d_{k-1}, ..., d_{k-j_1}\}. \tag{14}$$

It follows that

$$d_{j_0}\prod_{j=j_0}^{k}\left(1 - \frac{y}{a+bj}\right) \leq d_k \leq \prod_{s=0}^{\lfloor (k-(j_0+j_1))/j_1 \rfloor}\left(1 - \frac{y}{a+b(k-sj_1)}\right). \tag{15}$$

The LHS is clear. For the RHS note that $d_k \leq 1$ (as can be seen by using induction and the upper bound in (14)). When iterating d_j backwards on the RHS, we must make at least $\lfloor (k-(j_0+j_1))/l \rfloor$ iterations, and at least one value of j falls in each interval $[k-(s+1)l, k-sj_1)$. For that value of j we upper bound $(1 - y/(a+bj))$ by $(1 - y/(a+b(k-sj_1))$.

Now consider the product in the LHS of (15).

$$\begin{aligned}\log\left(\prod_{j=j_0}^{k}\left(1 - \frac{y}{a+bj}\right)\right) &= \sum_{j=j_0}^{k}\left(-\frac{y}{a+bj} - \frac{1}{2}\left(\frac{y}{a+bj}\right)^2 - \cdots\right) \\ &= O(1) - \sum_{j=j_0}^{k}\frac{y}{a+bj}.\end{aligned}$$

This justifies the lower bound of the lemma and the upper bound follows similarly for the upper bound of (15). □

The case $j_1 = 1$

We prove Theorem 2(ii). When $q_1 = 1, p_j = 0, j \geq j_0 = \Theta(1)$ the general value of d_k, $k \geq j_0$ can be found directly, by iterating the recurrence (2). Thus

$$\begin{aligned} d_k &= \frac{1}{a+bk}\left(d_{k-1}\left((a-1)+b(k-1)\right)\right) \\ &= d_{k-1}\left(1 - \frac{1+b}{a+bk}\right) \\ &= d_{j_0}\prod_{j=j_0}^{k}\left(1 - \frac{1+b}{a+jb}\right).\end{aligned}$$

Thus, for some constant C,

$$d_k \sim C(a+bk)^{-x}$$

where

$$x = 1 + \frac{1}{b} = 1 + \frac{2}{\alpha(1-\delta) + (1-\alpha)(1-\beta) + \alpha(1-\gamma)}.$$

The case $f = 0$

We prove Theorem 2(iii). The case ($f = 0$) arises in two ways. Firstly if $\alpha = 0$ so that a new vertex is added at each step. Secondly, if $\alpha \neq 0$ but $\delta = 1$ so that the initial vertex of an OLD choice is sampled u.a.r.

We first prove that for a sufficiently large absolute constant $A > 0$ and for all sufficiently large k, that

$$\frac{d_k}{d_{k-1}} = 1 - \frac{1+d}{a+dk} + \frac{\xi(k)}{k^2} \tag{16}$$

where $|\xi(k)| \leq A$.

We use induction and re-write (2) as

$$\frac{d_k}{d_{k-1}} = \frac{c + d(k-1)}{a+dk} + \sum_{j=1}^{j_1} \frac{e}{a+dk} \prod_{t=k-j}^{k-2} \frac{d_t}{d_{t-1}}. \tag{17}$$

Now use induction to write

$$\prod_{t=k-j}^{k-2} \frac{d_t}{d_{t-1}} = 1 - \frac{j-1}{a+dk} + \frac{\xi'(j,k)}{k^2} \tag{18}$$

where $|\xi'(j,k)| \leq Aj$.

Substituting (18) into (17) gives

$$\frac{d_k}{d_{k-1}} = \frac{c+d(k-1)}{a+dk} + \frac{e}{a+dk} - \frac{e\mu_q(d+1)}{(a+dk)^2} + \frac{\xi''(k)}{(a+dk)k^2}$$

where $|\xi''(k)| \leq 2A\mu_q$.

Equation (16) follows immediately and on iterating this we see that

$$d_k \sim Ck^{-\left(1+\frac{1}{d}\right)}.$$

3 Analysis of the general undirected model

Linear difference equations with rational coefficients: The method of Laplace

This section summarizes Chapter XV (pages 478-503) of *The Calculus of Finite Differences* by I. M. Milne-Thomson [19].

The equation (13) is an example of a linear difference equation with rational coefficients. It can equivalently be written as,

$$d_k(a+bk) - d_{k-1}(c+d(k-1)) - \sum_{j=1}^{k-1} d_{k-j}q_j(e+f(k-j)) = 0. \tag{19}$$

Laplace's difference equation is the name given to the equation whose coefficients are linear functions of a real variable w and an integer l. The general form of the homogeneous equation is

$$\sum_{j=0}^{l}[A_{l-j}(w+l-j)+B_{l-j}]u(w+l-j) = 0. \tag{20}$$

Thus (19) is a special case of (20) with $l = j_1$, $w = k - j_1$.

A method of solving difference equations with rational coefficients in general, and equation (20) in particular is to use the substitution

$$u(w) = \oint_C t^{w-1}v(t)dt.$$

The function $v(t)$ is obtained as the solution of the differential equation (23), given below, and C is a suitable contour of integration.

Let

$$\begin{aligned} \phi_1(t) &= A_l t^l + A_{l-1}t^{l-1} + \cdots + A_1 t + A_0 & (21) \\ \phi_0(t) &= B_l t^l + B_{l-1}t^{l-1} + \cdots + B_1 t + B_0, & (22) \end{aligned}$$

where $\phi_1(t)$ is the *characteristic equation.* The differential equation referred to, is

$$t\phi_1(t)\frac{dv(t)}{dt} - \phi_0(t)v(t) = 0. \tag{23}$$

The general method of solution requires (20) to be of the *Normal type*, namely:

N(i) Both A_l and A_0 are non-zero.

N(ii) The differential equation (23) satisfied by $v(t)$ is of the Fuchsian type.

Let the roots of the characteristic equation be $a_1, ..., a_l$ (with repetition). The condition that $v(t)$ is of the Fuchsian type, requires that $\phi_0(t)/\phi_1(t)$ can be expressed as a convergent power series of t for some $t > 0$. Thus either the roots $a_1, ..., a_l$ of the characteristic equation must be distinct, or if a is repeated ν times, then a is a root of $\phi_0(t)$ at least $\nu - 1$ times.

Assuming the roots are distinct,

$$\begin{aligned} \frac{v'(t)}{v(t)} &= \frac{\phi_0(t)}{t\phi_1(t)} \\ &= \frac{-\alpha_0}{t} + \frac{\beta_1}{t-a_1} + \cdots + \frac{\beta_l}{t-a_l}, & (24) \end{aligned}$$

and $\phi_0(t)/\phi_1(t)$ has the required series expansion. The general solution is

$$v^*(t) = t^{-\alpha_0}(t-a_1)^{\beta_1}...(t-a_l)^{\beta_l}.$$

As long as there are no repeated roots, a system of fundamental solutions $(u_j(w),\ j = 1, ..., l)$ is given by

$$u_j(w) = \frac{1}{2\pi i} \oint_{C_j} t^{w-1-\alpha_0}(t-a_1)^{\beta_1} \cdots (t-a_l)^{\beta_l} dt,$$

where C_j is a contour containing 0 and a_j but excluding the other roots. If β_j is integer the contour integral is replaced by the integral from 0 to a_j.

A specific solution for $u_j(w)$, valid for $\Re(w) > \alpha_0$, can be obtained as

$$u_j(w) \quad = \quad (a_j)^w \sum_{m=0}^{\infty} C_m B\left(\frac{w-\alpha_0+\theta-1}{\theta}, \beta_j + m + 1\right)$$

where $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$.

The variable $\theta > 1$ measures the angular separation, about the origin, of the root a_j from the other roots in the transformation $a_j z^{1/\theta} = t$ used to expand the transformed integral about $z = 1$ and obtain the above solution.

Now using the fact that $\Gamma(x) \sim \sqrt{2\pi} e^{-x} x^{x-1/2}$, as $w \to \infty$,

$$u_j(w) \sim C_j a_j^w w^{-(1+\beta_j)}(1 + O(1/w)). \tag{25}$$

Application of the technique

For convenience we let $l = j_1$ for the rest of this section.

Considering the equation (19) we see that

$$\begin{aligned} \phi_1(y) &= y^l - \left(\frac{d+q_1 f}{b} y^{l-1} + \frac{q_2 f}{b} y^{l-2} + \cdots + \frac{q_l f}{b}\right) \\ \phi_0(y) &= \frac{a}{b} y^l - \left(\frac{c+q_1 e}{b} y^{l-1} + \frac{q_2 e}{b} y^{l-2} + \cdots + \frac{q_l e}{b}\right). \end{aligned}$$

We assume that $f > 0$ so that N(i) is satisfied. Let the roots of the characteristic equation be ordered in decreasing size so that $|a_1| \geq |a_2| \geq \cdots \geq |a_l|$. Because of the SOLUTION CONDITIONS we see from Lemma 3, given below, that

$$a_1 = 1$$

and all other roots are either negative or complex and satisfy $|a| < 1$. Considering the partial fraction expansion (24) we see that

$$\phi_0(0) = -\alpha_0 \phi_1(0),$$

so that $\alpha_0 = -e/f$. Also

$$\phi_0(1) = \beta_1 \psi(1),$$

where $\psi(y) = \phi_1(y)/(y-1)$ is given by

$$\psi(z) = z^{l-1} + (1-\alpha_1)z^{l-2} + (1-\alpha_1-\alpha_2)z^{l-3} + \cdots + \\ (1-\alpha_1-\cdots-\alpha_{l-2})z + (1-\alpha_1-\cdots-\alpha_{l-1}), \tag{26}$$

and where

$$\alpha_1 = \frac{d+q_1 f}{b}, \qquad \alpha_2 = \frac{q_2 f}{b}, \dots, \alpha_l = \frac{q_l f}{b}.$$

From (1) we know $c+e = a-1$. Thus $\phi_0(1) = 1/b$, and so $\psi(1) = (d+f\mu_q)/b$. Thus $\beta_1 = 1/(d+f\mu_q)$. The other β_j require detailed knowledge of the roots of $\phi_1(t)$ and are not relevant to the asymptotic solution.

The solutions $u_j(w)$ are valid for $\Re(w) > \alpha_0 = -e/f$ which includes all $k \geq 0$.

Thus considering the root $a_1 = 1$ we see that

$$u_1(k) = Ck^{-(1+\beta_1)} \left(1 + O\left(\frac{1}{k}\right)\right)$$

where $\beta_1 = \phi_0(1)/\psi(1) = \frac{1}{d+f\mu_q}$, and $1+\beta_1$ is the parameter x of our degree sequence.

For $j \geq 2$, we use (25), giving

$$u_j(k) \to (a_j)^k k^{-(1+\beta_j)} \to 0,$$

faster than $o(1/k)$, if $|a_j| < 1$.

The specific solution for the sequence $(d_1, d_2, ..., d_k, ...)$ is

$$d_k = b_1 u_1(k) + \cdots + b_l u_l(k),$$

where $u_1(w), ..., u_l(w)$ are the fundamental solutions corresponding to the roots $a_1, ..., a_l$. We note that $b_1 \neq 0$. Indeed from Lemma 2, we know d_k obeys a power law, whereas if $b_1 = 0$, then d_k would decay exponentially as $|a_2|^k$.

Thus the error in the approximation of d_k is $O(1/k)$ from the non-asymptotic expansion of $u_1(w)$, and we conclude

$$d_k = Ck^{-\left(1+\frac{1}{d+f\mu_q}\right)} \left(1 + O\left(\frac{1}{k}\right)\right).$$

In the case where $\phi_1(t)$ has other solutions $|a_j| = 1,\ j = 2, ..., j', j' \leq l$, then the asymptotic solution d_k will be a linear combination of k-th powers of these roots.

Roots of the characteristic equation

Lemma 3. *Let $\alpha_1 = (d+q_1 f)/(d+f)$ and for $2 \leq j \leq l$, let $\alpha_j = q_j f/(d+f)$, and let*

$$\phi_1(z) = z^l - \alpha_1 z^{l-1} - \alpha_2 z^{l-2} - \cdots - \alpha_l.$$

Provided $\alpha_1 > 0$ or $\{q_j\}$ is not periodic, then the solutions of $\phi_1(z) = 0$ are

i) *An un-repeated root at $z = 1$,*

ii) *$l-1$ other (possibly repeated) roots λ satisfying $|\lambda| < 1$.*

Proof

We note the following (see Pólya & Szegő [20] p106 16,17). A polynomial $f(z)$ of the form

$$f(z) = z^n - p_1 z^{n-1} - p_2 z^{n-2} - \cdots - p_{n-1} z - p_n,$$

where $p_i \geq 0,\ i = 1, ..., n$ and $p_1 + \cdots + p_n > 0$ has just one positive zero ζ. All other zeroes z_0 of $f(z)$ satisfy $|z_0| \leq \zeta$.

Now $\alpha_i \geq 0$ and $\sum \alpha_i = 1$, and so $\phi_1(1) = 0$ and all other zeros, z_0, of $\phi_1(z)$ satisfy $|z_0| \leq 1$.

Let $\psi(z) = \phi_1(z)/(z-1)$ be as in (26). Now $\psi(1)$ is given by

$$1 + (1-\alpha_1) + (1-\alpha_1-\alpha_2) + \cdots + (1-\alpha_1-\cdots-\alpha_{l-1}) = \frac{d+f\mu_q}{d+f}, \tag{27}$$

and thus $\psi(1) \neq 0$, so that $z = 1$ is not a repeated root of ϕ_1.

Let z satisfy $\phi_1(z) = 0,\ |z| = 1,\ z \neq 1$, and let $w = 1/z$; then $\phi_1(z) = 0$ is equivalent to $h(w) = 1$, where

$$h(w) = \alpha_1 w + \alpha_2 w^2 + \cdots + \alpha_l w^l.$$

Suppose there exists $w \neq 1$, on the unit circle satisfying $h(w) = 1$. Let $T = \{w, w^2, ..., w^l\}$ then all elements of T are points on the unit circle. As $w \neq 1$, $\Re(w) < 1$ and $\Re(w^j) \leq 1,\ j = 2, ..., l$.

Now, by S(i), either $\alpha_1 > 0$ or $\alpha_1 = 0$ but $\{q_j\}$ is not periodic.

If $\alpha_1 > 0$, then

$$\sum \alpha_j \Re(w^j) \leq \alpha_1 \Re(w) + \alpha_2 + \cdots \alpha_l < 1,$$

and the conclusion, that $h(w) \neq 1$ follows.

Suppose $\alpha_1 = 0$. If $1 \notin T$, then the real part of w^j satisfies $\Re(w^j) < 1$, contradicting $h(w) = 1$. If $1 \in T$ then $w = e^{2\pi i/m}$ for some integer $m > 1$. However, as $\{q_j\}$ is not periodic, the conclusion that $h(w) < 1$ follows as before. □

The proof of Theorem 2 is now complete.

What happens if the SOLUTION CONDITIONS are not met?

If S(i) fails because $d = 0$ and $\{q_j\}$ is periodic, but S(ii) is true: then $d_k \sim \sum a^k C_a k^{-x(a)}$ where a are roots of unity satisfying $\phi_1(a) = 0$. Although do not state this case as a theorem, it is a direct consequence of the techniques in [19] and is covered by the result (25) and the discussion at the end of Section 3.

Suppose S(i) is true but S(ii) fails. One root of $\phi_1(y)$ is at $a_1 = 1$. Provided $b \neq 0$ this root is not repeated. For let $\psi(y) = \phi_1(y)/(y-1)$ then $\psi(1) = (d+f\mu_q)/(d+f)$, see (27). That the asymptotic solution is still d_k in (iv) of Theorem 2 above, can be shown (it seems) by further analysis.

If $a = 1$ is a repeated root of $\phi_1(y)$, we can expect the structure of the solution to differ.

3.1 Concentration

Here we prove Theorem 3. Fix t and condition on the following for each $\tau \leq t$: the choice of procedure NEW or OLD; the label of the initial vertex selected (OLD) or created (NEW); uniform choice of vertex or uniform choice of edge (equivalent to choice according to degree); the number T_τ of extra edges added at each step τ. Denote this conditional event by $\mathcal{A}$. In the following we will work entirely within the conditional space $\mathcal{A}$ and note that $\mathcal{A}$ is a product space over the steps $\tau \leq t$. Given $\mathcal{A}$, let $T = \sum_{\tau \leq t} T_\tau$ and let $Y_1, Y_2, \ldots, Y_T$ be the sequence of single choices of edges created. We will use the Azuma-Hoeffding martingale inequality and so we let

$$Z_i = \mathbf{E}(D_k(t) \mid Y_1, Y_2, \ldots, Y_i, \mathcal{A}) - \mathbf{E}(D_k(t) \mid Y_1, Y_2, \ldots, Y_{i-1}, \mathcal{A})$$

and prove that

$$|Z_i| \leq 2. \tag{28}$$

Noting that $T \leq tj_{\max}$, Theorem 3 follows immediately.

Fix $Y_1, Y_2, \ldots, Y_i$ and let $\hat{Y}_i = (x, \hat{v})$ be obtained from $Y_i = (x, v)$ by replacing a copying edge choice $e = (u, v)$ with an edge choice $\hat{e} = (\hat{u}, \hat{v})$. Then for each complete outcome $\mathbf{Y} = Y_1, Y_2, \ldots, Y_T$ we define a corresponding outcome $\hat{\mathbf{Y}} = Y_1, Y_2, \ldots, Y_{i-1}, \hat{Y}_i, \ldots, \hat{Y}_T$ where for $j > i$, $\hat{Y}_j$ is obtained from Y_j as follows: If Y_j creates a new edge (w, v) by randomly choosing edge (x, v) arising from Y_i, then in $\hat{Y}_j$, (w, v) is replaced by $(w, \hat{v})$.

The map $\mathbf{Y} \to \hat{\mathbf{Y}}$ is measure preserving and in going from $\mathbf{Y}$ to $\hat{\mathbf{Y}}$ only the degrees of v and $\hat{v}$ change and so the number of vertices of degree k changes by at most 2, and (28) follows.

4 Directed variants of the model

A curious phenomena of the directed models is that they are *incomplete* in the sense that the sampling procedure for terminal vertices in the out-model (resp. initial vertices in the in-model) does not need to be specified in order to estimate $\bar{d}_k^+$ (resp. $\bar{d}_k^-$). Thus these are not models, but classes of models. For the out-model, for example, terminal vertices can be picked according to *any* rule: assign, copy, direct all edges to vertex 1 etc.

Sampling based on out-degree

Let $\theta^+ = (1-\alpha)\mu_p + \alpha\mu_q$. The estimate (6-9) is replaced by

$$\overline{D}_k^+(t) = \overline{D}_k^+(t-1) + (1-\alpha)p_k + \alpha\left(\sum_j q_j\left(\tfrac{\delta}{(1-\alpha)t}\left(\overline{D}_{k-j}^+ - \overline{D}_k^+\right) + \tfrac{1-\delta}{t\theta^+}\left((k-j)\overline{D}_{k-j}^+ - k\overline{D}_k^+\right)\right)\right) + O(t^{-1/2}\log t).$$

Then for $k \geq 1$ we obtain

$$d_k^+(1 + e + fk) = (1-\alpha)p_k + \sum_{j=1}^{j_1} d_{k-j}^+ q_j (e + f(k-j)).$$

For large k this is a rational difference equation with characteristic equation

$$\phi_1(y) = y^{j_1} - \left(q_1 y^{j_1 - 1} + \cdots + q_{j_1}\right).$$

Thus provided $\phi_1(y)$ has no repeated roots, and $f > 0$,

$$x^+ = 1 + 1/(f\mu_q) = 1 + \frac{(1-\alpha)\mu_p + \alpha\mu_q}{\alpha(1-\delta)\mu_q}. \tag{29}$$

and

$$\bar{d}_k^+ \sim Ck^{-x^+}.$$

If $f = 0$ or if $\alpha = 0$ then $\bar{d}_k^+ \sim p_k$. If $f = 0$ and $\alpha > 0$ then $d_k \sim \zeta^k$ where $\zeta < 1$ is the unique positive root of

$$\phi_0(y) = y^{j_1} - \sum q_j \frac{e}{1+e} y^{j_1 - j},$$

so that the degree sequence decays exponentially.

Sampling based on in-degree

Let $\theta = \theta^+$ as given above. For $k < 0$ let $D_k^- = 0$. Then we have,

$$\begin{aligned}
D_k^-(t) &= D_k^-(t-1) + (1-\alpha)1_{k=0} \\
&+ (1-\alpha)\left(-\sum_{j=1}^{j_0} p_j \left(\tfrac{\beta j \overline{D}_k^-}{(1-\alpha)t} + \tfrac{(1-\beta)jk\overline{D}_k^-}{\theta(t-1)}\right) + \sum_{j=1}^{j_0} p_j \left(\tfrac{\beta j \overline{D}_{k-1}^-}{(1-\alpha)t} + \tfrac{(1-\beta)j(k-1)\overline{D}_{k-1}^-}{\theta t}\right)\right) \\
&+ \alpha\left(\sum_{j=1}^{j_1} q_j \left[\gamma\left(-\tfrac{j\overline{D}_k^-}{(1-\alpha)t} + \tfrac{j\overline{D}_{k-1}^-}{(1-\alpha)t}\right) + (1-\gamma)\left(-\tfrac{jk\overline{D}_k^-}{\theta t} + \tfrac{j(k-1)\overline{D}_{k-1}^-}{\theta t}\right)\right]\right).
\end{aligned}$$

For $k \geq 0$ we find,

$$\begin{aligned}
d_0^- &= \frac{(1-\alpha)^2}{(1-\alpha)(1+\beta\mu_p) + \alpha\gamma\mu_q} \\
d_k^- &= \frac{c + d(k-1)}{1+c+dk} d_{k-1} \qquad\qquad k \geq 1.
\end{aligned}$$

This is a $j_1 = 1$ case, giving an exponent

$$x^- = 1 + \frac{1}{d} = 1 + \frac{(1-\alpha)\mu_p + \alpha\mu_q}{(1-\alpha)(1-\beta)\mu_p + \alpha(1-\gamma)\mu_q}. \tag{30}$$

References

[1] M. Adler and M. Mitzenmacher, *Toward Compressing Web Graphs*, To appear in the 2001 Data Compression Conference.

[2] W. Aiello, F. Chung and L. Lu. *A random graph model for massive graphs.* Proc 32nd ACM Symposium on the Theory of Computing. (1999)

[3] R. Albert, A. Barabasi and H. Jeong. *Diameter of the world wide web.* Nature 401:103-131 (1999) see also `http://xxx.lanl.gov/abs/cond-mat/9907038`

[4] B. Bollobás and F. Chung. *The diameter of a cycle plus a random matching.* SIAM Journal of Discrete Maths 1:328-333 (1988)

[5] B. Bollobás, O. Riordan and J. Spencer, *The degree sequence of a scale free random graph process*, to appear.

[6] B. Bollobás and O. Riordan, *The diameter of a scale free random graph*, to appear.

[7] A. Broder, R. Kumar, F.Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. *Graph structure in the web.* `http://gatekeeper.dec.com/pub/DEC/SRC/publications/stata/www9.htm`

[8] C. Cooper and A. M. Frieze. *The size of the largest strongly connected component of a random digraph with a given degree sequence.*

[9] C. Cooper, A. M. Frieze. and I. N. Kovalenko. *Directed web graphs.*

[10] E. Drinea, M. Enachescu and M. Mitzenmacher, *Variations on random graph models for the web.*

[11] B. Hayes. *Graph theory in practice: Part II.* American Scientist 88:104-109. (2000). `http://www.sigmaxi.org/amsci/issues/Comsci00/compsci2000-03.html`

[12] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Jornal of the American Statistical Association.

[13] J Kleinberg. *The small-world phenomenon: An algorithmic perspective.* `http://www.cs.cornell.edu/home/kleinber/swn.ps`

[14] S.Janson, T.Łuczak and A.Ruciński, *Random Graphs*, John Wiley and Sons, 2000.

[15] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. *The web as a graph.* `www.almaden.ibm.com`

[16] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. *Stochastic models for the web graph.* `www.almaden.ibm.com`

[17] M Molloy and B. Reed. *A critical point for random graphs with a given degree sequence.* Random Structures and Algorithms 6 161-180 (1995)

[18] M Molloy and B. Reed. *The size of the giant component of a random graph with a given degree sequence.* Combinatorics, Probability and Computing.

[19] I. M. Milne-Thomson. *The Calculus of Finite Differences.* Macmillian, London (1951).

[20] G. Pólya and G. Szegő. *Problems and Theorems in Analysis I.* Springer, Berlin (1970).

This research was sponsored in part by National Science Foundation (NSF) grant no. CCR-0122581.