Internet Mathematics Vol. 6, No. 4: 437–459

A Coupled Model for the Indegree and Outdegree Analysis of the Web

P. Favati, G. Lotti, O. Menchi, and F. Romani

Abstract. We introduce a mixed model for the Web graph that simultaneously describes the inlink and outlink distributions by taking into account the interconnection of the two processes. We derive an expression for the steady-state distribution of indegrees (outdegrees) among vertices with fixed outdegree (indegree) in terms of sums of beta functions. Experimentation on subsets of the real Web shows that the proposed distributions well reproduce the behavior of the observed data.

1. Introduction

Many models have been suggested to explain the main features of the Web graph (e.g., see [Barabasi and Albert 99, Barabasi et al. 99, Kleinberg et al. 99]). Among these, particular interest has been shown in the models that deal with the graph as a directed graph, allowing one to model both indegree and outdegree distributions. The models most suitable for a description of subsets of the real Web are those based on mixed rules, including uniform attachment and preferential attachment strategies [Cooper and Frieze 01, Dorogovtsev et al. 00, Pennock et al. 02, Kumar et al. 00].

ISSN: 1542-7951 print

Most papers aim to show that both indegree and outdegree distributions satisfy a power law, but accurate observations of subsets of the real Web have shown that the link distribution corresponding to low degrees fails to fit a power law with a discrepancy larger for outlinks than for inlinks [Broder et al. 00, Caldarelli et al. 03]. A better description of either distribution can be given using the beta function [Favati et al. 08] or one of its asymptotic approximations [Pennock et al. 02]. The beta function was used earlier by [Simon 55] to describe word-frequency distributions.

All these models treat either indegrees or outdegrees, but not both simultaneously. Since the processes of inlinking and outlinking are two aspects of the same phenomenon, every pair of independent models cannot describe it completely. We expect that a single model may be devised to describe the inlink and outlink distributions together by taking into account the interconnections of the two processes. Such an integrated model would give more accurate information on the strategies that govern the evolution of links generation. We present here, as announced in [Favati et al. 08], a model that allows a full treatment of direct graphs by dealing simultaneously with both distributions. A similar approach is followed in [Bollobás et al. 03, Cooper 06], where the indegree and outdegree distributions are described in terms of the power law.

In contrast to the monodimensional case, a closed form of the solution of the bidimensional model describing the underlying stochastic process is not available, i.e., the number of pages with assigned indegree and outdegree cannot be explicitly given. However, with our approach we succeed in describing the indegree (outdegree) distributions for each fixed value of the outdegree (indegree) in terms of sums of beta functions depending on interconnected parameters derived from a unique process of link generation. Analysis of the "cross sections" of the graph is done also in [Bollobás et al. 03], but only their asymptotic expressions are given.

Experimentation carried out on real data sets shows that the proposed model is valid and that the technique used for detecting values of the parameters is effective. In Section 2 the linking model is introduced, and its steady-state solution is proposed and analyzed. In Section 3 the data and the techniques used for the experiments are presented, and the results of the experimentation are discussed. Conclusions are given in Section 4.

The models mentioned above, including the one we propose, imply a linear growth of the edges in the number of nodes, i.e., the average node degrees remain constant over time. Nonlinear models have been considered as well. It is noted in [Leskovec et al. 05] that in some networks, the average node degrees increase over time, and the authors developed a model capturing this property. However,

as noted in [Mitzenmacher 06], at this time this property has not been observed in the Web graph.

2. The Stochastic Problem

The Web can be represented as a directed graph of pages, connected by links. Two important indicators are associated with each page: the *indegree*, that is, the number of inlinks pointing to that page, and the *outdegree*, which is the number of outlinks originating from that page.

We want to examine the Web structure from both the inlink and outlink points of view. The aim is to find how many pages of the Web have a given indegree or outdegree. Hence we will examine how the number X_j of pages with indegree j and the number Y_h of pages with outdegree h depend on j and h, respectively. In order to obtain accurate information on the evolution of links generation, we perform our analysis by taking into account the interconnection between the inlink and outlink processes, and introduce the number $Z_{j,h}$ of pages with indegree j and outdegree h. We assume that any page of interest has at least one link, so $Z_{0,0} = 0$.

To find an adequate model for the function $Z_{j,h}$, a discrete-time stochastic process is considered. At any time step a new link (i.e., a new inlink and a new outlink) is created according to the following criteria:

1. With probability α_{I}, the new link points to a new page, i.e., a page having zero indegree and outdegree.
2. With probability γ_{I}, the new link points to a page chosen at random among those having zero indegree and nonzero outdegree.
3. With probability $1 - \alpha_{\mathrm{I}} - \gamma_{\mathrm{I}}$, the new link points to a page chosen among those having nonzero indegree and outdegree, according to the following rule:
   (a) With probability β_{I}, the new link points to a page chosen at random (this policy is known as *uniform attachment*).
   (b) With probability $1 - \beta_{\mathrm{I}}$, the new link points to a page chosen proportionally to its indegree (this policy is known as *preferential attachment* and expresses the concept that new links tend to attach themselves to pages already having more inlinks).

For the outlinks, analogous positions hold with the subscript O replacing the I.

2.1. The Linking Model

Let $Z_{j,h}^{(t)}$ be the number of pages having indegree j and outdegree h at time t, with $j, h \geq 0$. The time is increased by one when a link is created, and a page is counted when at least one link points toward it or away from it. At time t:

1. The number of pages is $n(t) = \sum_{j,h\geq 0} Z_{j,h}^{(t)}$.
2. The number of pages with j inlinks is $X_j^{(t)} = \sum_{h\geq 0} Z_{j,h}^{(t)}$.
3. The number of pages with h outlinks is $Y_h^{(t)} = \sum_{j\geq 0} Z_{j,h}^{(t)}$.
4. The number of links is

$$t = \sum_{j\geq 1} jX_j^{(t)} = \sum_{h\geq 1} hY_h^{(t)}. \tag{2.1}$$

We assume that initially $t = 1$, $Z_{0,0}^{(1)} = 0$, $Z_{1,0}^{(1)} = Z_{0,1}^{(1)} = 1$, and that $Z_{0,0}^{(t)} = 0$ for any t. Since only one link is created at any time step, we assume also that $Z_{j,h}^{(t)} = 0$ for $j, h > t$. When a new link is created, it points to a page having indegree j and outdegree h with probability $p_{\mathrm{I}}^{(t)}(j,h)$, and it exits from a page having indegree j and outdegree h with probability $p_{\mathrm{O}}^{(t)}(j,h)$.

These probabilities are given by two terms: the first term, according to the uniform attachment policy (a), is proportional to $Z_{j,h}^{(t)}$, while the second term, according to the preferential attachment policy (b), is proportional to the number of all existing links pointing to pages having indegree j (or exiting from pages having outdegree h). In the case of zero indegree or of zero outdegree, only the first term is present. Hence we have for $j > 0$,

$$p_{\mathrm{I}}^{(t)}(j,h) = (1-\alpha_{\mathrm{I}}-\gamma_{\mathrm{I}})\left[\frac{\beta_{\mathrm{I}}}{n_{\mathrm{I}}(t)}Z_{j,h}^{(t)} + \frac{1-\beta_{\mathrm{I}}}{t}jZ_{j,h}^{(t)}\right], \tag{2.2}$$

and for $j = 0$, $h > 0$,

$$p_{\mathrm{I}}^{(t)}(0,h) = \frac{\gamma_{\mathrm{I}}}{m_{\mathrm{I}}(t)}Z_{0,h}^{(t)}, \tag{2.3}$$

where $n_{\mathrm{I}}(t)$ and $m_{\mathrm{I}}(t)$ are the number of pages having nonzero indegree and zero indegree respectively, and

$$p_{\mathrm{O}}^{(t)}(j,h) = (1-\alpha_{\mathrm{O}}-\gamma_{\mathrm{O}})\left[\frac{\beta_{\mathrm{O}}}{n_{\mathrm{O}}(t)}Z_{j,h}^{(t)} + \frac{1-\beta_{\mathrm{O}}}{t}hZ_{j,h}^{(t)}\right], \tag{2.4}$$

and for $j > 0$, $h = 0$,

$$p_{\mathrm{O}}^{(t)}(j,0) = \frac{\gamma_{\mathrm{O}}}{m_{\mathrm{O}}(t)}Z_{j,0}^{(t)}, \tag{2.5}$$

where $n_{\mathrm{O}}(t)$ and $m_{\mathrm{O}}(t)$ are the numbers of pages having nonzero outdegree and zero outdegree respectively.

Initial positions are

$$p_{\mathrm{I}}^{(t)}(0,0) = \alpha_{\mathrm{I}}, \quad p_{\mathrm{O}}^{(t)}(0,0) = \alpha_{\mathrm{O}}.$$

It can easily be seen that

$$\sum_{h=1}^{t} p_{\mathrm{I}}^{(t)}(0,h) = \gamma_{\mathrm{I}}, \quad \sum_{j=1}^{t}\sum_{h=0}^{t} p_{\mathrm{I}}^{(t)}(j,h) = 1 - \alpha_{\mathrm{I}} - \gamma_{\mathrm{I}},$$

$$\sum_{j=1}^{t} p_{\mathrm{O}}^{(t)}(j,0) = \gamma_{\mathrm{O}}, \quad \sum_{j=0}^{t}\sum_{h=1}^{t} p_{\mathrm{O}}^{(t)}(j,h) = 1 - \alpha_{\mathrm{O}} - \gamma_{\mathrm{O}}.$$

Hence

$$\sum_{j,h=0}^{t} p_{\mathrm{I}}^{(t)}(j,h) = \sum_{j,h=0}^{t} p_{\mathrm{O}}^{(t)}(j,h) = 1.$$

The expected value of the variation of $Z_{j,h}^{(t+1)}$ with respect to $Z_{j,h}^{(t)}$ is given by

$$\mathcal{E}\big[Z_{j,h}^{(t+1)} - Z_{j,h}^{(t)}\big] = p_{\mathrm{I}}^{(t)}(j-1,h) - p_{\mathrm{I}}^{(t)}(j,h) + p_{\mathrm{O}}^{(t)}(j,h-1) - p_{\mathrm{O}}^{(t)}(j,h). \quad (2.6)$$

The equation holds also for $j = 0$ and $h = 0$, provided that

$$p_{\mathrm{I}}^{(t)}(-1,h) = p_{\mathrm{O}}^{(t)}(j,-1) = 0.$$

The model (2.6) is one of the many versions of mixed models (see, for example, [Cooper and Frieze 01, Dorogovtsev et al. 00, Pennock et al. 02]). The novelty consists in considering each link simultaneously as an inlink and an outlink and, to this purpose, in describing the behavior of $Z_{j,h}^{(t)}$, instead of that of $X_j^{(t)}$ or $Y_h^{(t)}$ separately.

The probability of generating a page with one inlink is α_{I} and with one outlink is α_{O}. The probability of transferring a page from the set of pages with zero indegree (respectively outdegree) to the set of pages with nonzero indegree (respectively outdegree) is γ_{I} (respectively γ_{O}). Hence the expected values of the variation of the different page numbers are

$$\begin{aligned}
\mathcal{E}\big[n(t+1) - n(t)\big] &= \alpha_{\mathrm{I}} + \alpha_{\mathrm{O}}, \\
\mathcal{E}\big[n_{\mathrm{I}}(t+1) - n_{\mathrm{I}}(t)\big] &= \alpha_{\mathrm{I}} + \gamma_{\mathrm{I}}, \\
\mathcal{E}\big[n_{\mathrm{O}}(t+1) - n_{\mathrm{O}}(t)\big] &= \alpha_{\mathrm{O}} + \gamma_{\mathrm{O}}, \\
\mathcal{E}\big[m_{\mathrm{I}}(t+1) - m_{\mathrm{I}}(t)\big] &= \alpha_{\mathrm{O}} - \gamma_{\mathrm{I}}, \\
\mathcal{E}\big[m_{\mathrm{O}}(t+1) - m_{\mathrm{O}}(t)\big] &= \alpha_{\mathrm{I}} - \gamma_{\mathrm{O}}.
\end{aligned} \quad (2.7)$$

2.2. The Steady-State Solution

To find the steady-state distribution of the stochastic process, we replace the expected values in (2.6) by their actual values, obtaining the difference equation

$$Z_{j,h}^{(t+1)} - Z_{j,h}^{(t)} = p_{\mathrm{I}}^{(t)}(j-1,h) - p_{\mathrm{I}}^{(t)}(j,h) + p_{\mathrm{O}}^{(t)}(j,h-1) - p_{\mathrm{O}}^{(t)}(j,h). \qquad (2.8)$$

By applying this equation recursively, we can see how $Z_{j,h}^{(t)}$ evolves as t increases. From (2.7) we obtain

$$n(t+1) = n(t) + \alpha_{\mathrm{I}} + \alpha_{\mathrm{O}}, \quad n(1) = 2,$$

and for $t \to \infty$ we have

$$n(t) \sim (\alpha_{\mathrm{I}} + \alpha_{\mathrm{O}})t,$$

meaning that the ratio $n(t)/t$, i.e., the average number of links per page, is asymptotically equal to the constant $\alpha_{\mathrm{I}} + \alpha_{\mathrm{O}}$. Analogously, we have

$$\begin{aligned} n_{\mathrm{I}}(t) &\sim (\alpha_{\mathrm{I}} + \gamma_{\mathrm{I}})t, \quad & n_{\mathrm{O}}(t) &\sim (\alpha_{\mathrm{O}} + \gamma_{\mathrm{O}})t, \\ m_{\mathrm{I}}(t) &\sim (\alpha_{\mathrm{O}} - \gamma_{\mathrm{I}})t, \quad & m_{\mathrm{O}}(t) &\sim (\alpha_{\mathrm{I}} - \gamma_{\mathrm{O}})t. \end{aligned}$$

By setting

$$\begin{aligned} \mu_{\mathrm{I}} &= \frac{(1-\alpha_{\mathrm{I}}-\gamma_{\mathrm{I}})\beta_{\mathrm{I}}}{\alpha_{\mathrm{I}}+\gamma_{\mathrm{I}}}, \quad & \mu_{\mathrm{O}} &= \frac{(1-\alpha_{\mathrm{O}}-\gamma_{\mathrm{O}})\beta_{\mathrm{O}}}{\alpha_{\mathrm{O}}+\gamma_{\mathrm{O}}}, \\ \nu_{\mathrm{I}} &= (1-\alpha_{\mathrm{I}}-\gamma_{\mathrm{I}})(1-\beta_{\mathrm{I}}), \quad & \nu_{\mathrm{O}} &= (1-\alpha_{\mathrm{O}}-\gamma_{\mathrm{O}})(1-\beta_{\mathrm{O}}), \\ \eta_{\mathrm{I}} &= \frac{\gamma_{\mathrm{I}}}{\alpha_{\mathrm{O}}-\gamma_{\mathrm{I}}}, \quad & \eta_{\mathrm{O}} &= \frac{\gamma_{\mathrm{O}}}{\alpha_{\mathrm{I}}-\gamma_{\mathrm{O}}}, \end{aligned}$$

we get from (2.2) to (2.5), for $(j,h) \neq (0,0)$,

$$p_{\mathrm{I}}^{(t)}(j,h) \sim \frac{1}{t}\pi_{\mathrm{I}}^{(j)} Z_{j,h}^{(t)} \quad \text{and} \quad p_{\mathrm{O}}^{(t)}(j,h) \sim \frac{1}{t}\pi_{\mathrm{O}}^{(h)} Z_{j,h}^{(t)},$$

where

$$\pi_{\mathrm{I}}^{(j)} = \begin{cases} \eta_{\mathrm{I}}, & \text{for } j = 0, \\ \mu_{\mathrm{I}} + \nu_{\mathrm{I}} j, & \text{for } j > 0, \end{cases}$$

and

$$\pi_{\mathrm{O}}^{(h)} = \begin{cases} \eta_{\mathrm{O}}, & \text{for } h = 0, \\ \mu_{\mathrm{O}} + \nu_{\mathrm{O}} h, & \text{for } h > 0. \end{cases}$$

To find the steady-state solution of our model, we use these quantities in (2.8) and let (2.8) hold also for $j, h > t+1$. We get

$$Z_{j,h}^{(t+1)} - Z_{j,h}^{(t)} = \frac{1}{t}\left[\pi_{\mathrm{I}}^{(j-1)} Z_{j-1,h}^{(t)} - \pi_{\mathrm{I}}^{(j)} Z_{j,h}^{(t)} + \pi_{\mathrm{O}}^{(h-1)} Z_{j,h-1}^{(t)} - \pi_{\mathrm{O}}^{(h)} Z_{j,h}^{(t)}\right].$$

This equation holds also on the boundary if we set

$$\pi_{\mathrm{I}}^{(-1)} = \pi_{\mathrm{O}}^{(-1)} = 0,$$

i.e., for $j \neq 1$ we set

$$Z_{j,0}^{(t+1)} - Z_{j,0}^{(t)} = \frac{1}{t}\Big[\pi_{\mathrm{I}}^{(j-1)} Z_{j-1,0}^{(t)} - \pi_{\mathrm{I}}^{(j)} Z_{j,0}^{(t)} - \pi_{\mathrm{O}}^{(0)} Z_{j,h}^{(t)}\Big],$$

and for $h \neq 1$ we set

$$Z_{0,h}^{(t+1)} - Z_{0,h}^{(t)} = \frac{1}{t}\Big[- \pi_{\mathrm{I}}^{(0)} Z_{0,h}^{(t)} + \pi_{\mathrm{O}}^{(h-1)} Z_{0,h-1}^{(t)} - \pi_{\mathrm{O}}^{(h)} Z_{0,h}^{(t)}\Big],$$

where for $j = 1$ the term $\pi_{\mathrm{I}}^{(j-1)} Z_{j-1,0}^{(t)}$ is replaced by α_{I}, and for $h = 1$ the term $\pi_{\mathrm{O}}^{(h-1)} Z_{0,h-1}^{(t)}$ is replaced by α_{O}.

The steady-state solution $Z_{j,h}^{(t)}$ satisfies these equations for any t, and hence we assume it to be of the form $Z_{j,h}^{(t)} = t c_{j,h}$, where $c_{j,h}$, for $(j,h) \neq (1,0)$ and $(j,h) \neq (0,1)$, satisfies

$$s_{j,h} c_{j,h} = \pi_{\mathrm{I}}^{(j-1)} c_{j-1,h} + \pi_{O}^{(h-1)} c_{j,h-1}, \tag{2.9}$$

where

$$s_{j,h} = 1 + \pi_{\mathrm{I}}^{(j)} + \pi_{\mathrm{O}}^{(h)},$$

and initial conditions are given by

$$s_{1,0} c_{1,0} = \alpha_{\mathrm{I}}, \quad s_{0,1} c_{0,1} = \alpha_{\mathrm{O}}. \tag{2.10}$$

Note that $c_{0,0}$ is not defined through $Z_{0,0}^{(t)}$.

The steady-state distributions x_j and y_h corresponding to $X_j^{(t)}$ and $Y_h^{(t)}$ are given by

$$X_j^{(t)} = t x_j, \quad \text{with } x_j = \sum_{h \geq 0} c_{j,h},$$
$$Y_h^{(t)} = t y_h, \quad \text{with } y_h = \sum_{j \geq 0} c_{j,h},$$

and still satisfy (2.1). Hence

$$\sum_{j \geq 1} j x_j = \sum_{h \geq 1} h y_h = 1.$$

2.3. Properties of the Solution

To analyze the steady-state solution, we make use of the properties of the beta complete function $B(a, b)$, which satisfies the recurrence relation

$$(a + b - 1)B(a, b) = (a - 1)B(a - 1, b). \tag{2.11}$$

In the following, the first argument of the beta function will be expressed by the addition of a constant part plus an integer variable part, as, for example,

$$B(a + j, b), \quad \text{with } j \geq 0.$$

We define the corresponding normalized function

$$B_N(a + j, b) = \frac{B(a + j, b)}{B(a, b)},$$

which starts from zero with value one.

We are particularly interested in asymptotic approximations. For $j \to \infty$ we use the notation $a(j) \approx b(j)$ to indicate that $\lim_{j\to\infty} a(j)/b(j)$ is a constant. In the case of the beta function, the following asymptotic approximation for $j \to \infty$ holds:

$$B_N(a + j, b) \approx (a + j)^{-b} \quad \text{with } \lim_{j\to\infty} \frac{B_N(a + j, b)}{(a + j)^{-b}} = \Gamma(b). \tag{2.12}$$

Lemma 2.1. *The difference equation*

$$(\rho + j)z_j = (\sigma + j - 1)z_{j-1}, \quad \textit{with } j \geq 1 + \widetilde{j},$$

where ρ and σ are independent of j, has the solution

$$z_j = z_{\widetilde{j}} \frac{B(\sigma + j, 1 + \rho - \sigma)}{B(\sigma + \widetilde{j}, 1 + \rho - \sigma)}.$$

The asymptotic approximation for $j \to \infty$ is

$$z_j \approx (\sigma + j)^{1+\rho-\sigma}.$$

Proof. Setting $a = \sigma + j$, $b = 1 + \rho - \sigma$, we see from (2.11) that $B(\sigma + j, 1 + \rho - \sigma)$ satisfies the equation. The initial condition for $j = \widetilde{j}$ sets the right normalization value. The asymptotic approximation is obtained from (2.12). □

As we will see, the steady-state solution can be expressed as a combination of linearly independent beta functions. To get information on the behavior of the function $c_{j,h}$, we examine its cross sections, i.e., we fix a value of h and let j vary, and vice versa.

Theorem 2.2. *On the boundary the steady-state solution is*

$$c_{j,0} = \frac{\alpha_{\mathrm{I}}}{\mu_{\mathrm{I}}} B_N\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j, 1 + \frac{1+\eta_{\mathrm{O}}}{\nu_{\mathrm{I}}}\right), \quad \textit{with } j \geq 1, \tag{2.13}$$

$$c_{0,h} = \frac{\alpha_{\mathrm{O}}}{\mu_{\mathrm{O}}} B_N\left(\frac{\mu_{\mathrm{O}}}{\nu_{\mathrm{O}}} + h, 1 + \frac{1+\eta_{\mathrm{I}}}{\nu_{\mathrm{O}}}\right), \quad \textit{with } h \geq 1. \tag{2.14}$$

The asymptotic approximations for j and $h \to \infty$ are

$$c_{j,0} \approx \left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j\right)^{-(1+(1+\eta_{\mathrm{O}})/\nu_{\mathrm{I}})}, \quad c_{0,h} \approx \left(\frac{\mu_{\mathrm{O}}}{\nu_{\mathrm{O}}} + h\right)^{-(1+(1+\eta_{\mathrm{I}})/\nu_{\mathrm{O}})}.$$

Proof. To get $c_{j,0}$ we apply Lemma 2.1 to the difference equation (2.9) with $h = 0$, divided by ν_{I}. For $\widetilde{j} = 1$ this equation is

$$\left(\frac{1+\mu_{\mathrm{I}}+\eta_{\mathrm{O}}}{\nu_{\mathrm{I}}} + j\right) c_{j,0} = \left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j - 1\right) c_{j-1,0}, \quad j \geq 2.$$

The solution is

$$c_{j,0} = c_{1,0} \frac{B\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j, 1 + \frac{1+\eta_{\mathrm{O}}}{\nu_{\mathrm{I}}}\right)}{B\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + 1, 1 + \frac{1+\eta_{\mathrm{O}}}{\nu_{\mathrm{I}}}\right)}, \quad j \geq 1. \tag{2.15}$$

From (2.10) we have

$$c_{1,0} = \frac{\alpha_{\mathrm{I}}}{1+\mu_{\mathrm{I}}+\nu_{\mathrm{I}}+\eta_{\mathrm{O}}}, \tag{2.16}$$

and from (2.11) we have

$$B\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + 1, 1 + \frac{1+\eta_{\mathrm{O}}}{\nu_{\mathrm{I}}}\right) = \frac{\mu_{\mathrm{I}} B\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}}, 1 + \frac{1+\eta_{\mathrm{O}}}{\nu_{\mathrm{I}}}\right)}{1+\mu_{\mathrm{I}}+\nu_{\mathrm{I}}+\eta_{\mathrm{O}}}. \tag{2.17}$$

By substituting (2.16) and (2.17) into (2.15) we get (2.13). The proof for $c_{0,h}$ is analogous. □

Theorem 2.3. *Let $\nu_{\mathrm{I}} \neq 0$, $\nu_{\mathrm{O}} \neq 0$, $\eta_{\mathrm{O}} \neq \mu_{\mathrm{O}} + \nu_{\mathrm{O}} i$, and $\eta_{\mathrm{I}} \neq \mu_{\mathrm{I}} + \nu_{\mathrm{I}} i$ for all integers i. The cross sections of $c_{j,h}$ are as follows:*

(a) *For a fixed h,*

$$c_{j,h} = \sum_{i=0}^{h} a_{\mathrm{i}}^{(h)} B_N\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j, 1 + \frac{1+\pi_{\mathrm{O}}^{(i)}}{\nu_{\mathrm{I}}}\right), \tag{2.18}$$

where

$$
\begin{aligned}
a_0^{(0)} &= \frac{\alpha_{\mathrm{I}}}{\mu_{\mathrm{I}}}, \quad a_0^{(h)} = \frac{a_0^{(h-1)}\pi_{\mathrm{O}}^{(h-1)}}{\mu_{\mathrm{O}} + \nu_{\mathrm{O}} h - \eta_{\mathrm{O}}}, \\
a_{\mathrm{i}}^{(h)} &= \frac{a_{\mathrm{i}}^{(h-1)}\pi_{\mathrm{O}}^{(h-1)}}{(h-i)\nu_{\mathrm{O}}}, \quad \textit{for } i = 1, \ldots, h-1, \\
a_h^{(h)} &= c_{0,h} - \sum_{i=0}^{h-1} a_{\mathrm{i}}^{(h)}.
\end{aligned}
$$

(b) *For a fixed j,*

$$
c_{j,h} = \sum_{i=0}^{j} b_{\mathrm{i}}^{(j)} B_N\left(\frac{\mu_{\mathrm{O}}}{\nu_{\mathrm{O}}} + h, 1 + \frac{1 + \pi_{\mathrm{I}}^{(i)}}{\nu_{\mathrm{O}}}\right), \tag{2.19}
$$

where

$$
\begin{aligned}
b_0^{(0)} &= \frac{\alpha_{\mathrm{O}}}{\mu_{\mathrm{O}}}, \quad b_0^{(j)} = \frac{b_0^{(j-1)}\pi_{\mathrm{I}}^{(j-1)}}{\mu_{\mathrm{I}} + \nu_{\mathrm{I}} j - \eta_{\mathrm{I}}}, \\
b_{\mathrm{i}}^{(j)} &= \frac{b_{\mathrm{i}}^{(j-1)}\pi_{\mathrm{I}}^{(j-1)}}{(j-i)\nu_{\mathrm{I}}}, \quad \textit{for } i = 1, \ldots, j-1, \\
b_j^{(j)} &= c_{j,0} - \sum_{i=0}^{j-1} b_{\mathrm{i}}^{(j)}.
\end{aligned}
$$

(c) *The asymptotic approximations for j and $h \to \infty$ are*

$$
\begin{aligned}
&\textit{for a fixed } h: \quad c_{j,h} \approx \left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j\right)^{-(1+(1+e_{\mathrm{O}})/\nu_{\mathrm{I}})}, \\
&\textit{for a fixed } j: \quad c_{j,h} \approx \left(\frac{\mu_{\mathrm{O}}}{\nu_{\mathrm{O}}} + h\right)^{-(1+(1+e_{\mathrm{I}})/\nu_{\mathrm{O}})},
\end{aligned} \tag{2.20}
$$

where

$$
e_{\mathrm{I}} = \min\{\eta_{\mathrm{I}}, \mu_{\mathrm{I}} + \nu_{\mathrm{I}}\}, \quad e_{\mathrm{O}} = \min\{\eta_{\mathrm{O}}, \mu_{\mathrm{O}} + \nu_{\mathrm{O}}\}.
$$

Proof. (a) Let h be fixed. For $h = 0$, expression (2.18) follows immediately from Theorem 2.2. For $h \geq 1$ we proceed by induction. For simplicity's sake in this proof we use the notation

$$
g(j, i) := B_N\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j, 1 + \frac{1 + \pi_{\mathrm{O}}^{(i)}}{\nu_{\mathrm{I}}}\right).
$$

Applying (2.11), we have

$$s_{j,i}\, g(j,i) = \pi_{\mathrm{I}}^{(j-1)}\, g(j-1,i). \tag{2.21}$$

We verify by direct substitution that (2.18) satisfies (2.9). In fact, replacing (2.18) in the left-hand side of (2.9) gives

$$L = \sum_{i=0}^{h} a_{\mathrm{i}}^{(h)}\, s_{j,h}\, g(j,i),$$

and in the right-hand side gives

$$R = \sum_{i=0}^{h} a_{\mathrm{i}}^{(h)}\, \pi_{\mathrm{I}}^{(j-1)}\, g(j-1,i) + \sum_{i=0}^{h-1} a_{\mathrm{i}}^{(h-1)}\, \pi_{\mathrm{O}}^{(h-1)}\, g(j,i).$$

Replacing $g(j-1,i)$ in R by means of (2.21), we have

$$L - R = \sum_{i=0}^{h-1} \left(a_{\mathrm{i}}^{(h)}\, s_{j,h} - a_{\mathrm{i}}^{(h)}\, s_{j,i} - a_{\mathrm{i}}^{(h-1)}\, \pi_{\mathrm{O}}^{(h-1)} \right) g(j,i).$$

Hence $L = R$ if

$$a_{\mathrm{i}}^{(h)} = \frac{a_{\mathrm{i}}^{(h-1)}\, \pi_{\mathrm{O}}^{(h-1)}}{s_{j,h} - s_{j,i}}, \quad \text{for } i = 0, \ldots, h-1,$$

where $s_{j,h} - s_{j,i} = \mu_{\mathrm{O}} + \nu_{\mathrm{O}} h - \eta_{\mathrm{O}}$ if $i = 0$ and $s_{j,h} - s_{j,i} = (h-i)\nu_{\mathrm{O}}$ if $i \geq 1$. The coefficient $a_h^{(h)}$ is found by imposing that

$$c_{j,h}\Big|_{j=0} = c_{0,h},$$

and since all the normalized beta functions of (2.18) are equal to 1, it follows that

$$a_h^{(h)} = c_{0,h} - \sum_{i=0}^{h-1} a_{\mathrm{i}}^{(h)}.$$

(b) The proof for $c_{j,h}$ with a fixed j is analogous.

(c) The dominant term of (2.18) is either the first one or the second one, according as $\eta_{\mathrm{O}} < \mu_{\mathrm{O}} + \nu_{\mathrm{O}}$ or $\eta_{\mathrm{O}} > \mu_{\mathrm{O}} + \nu_{\mathrm{O}}$. In practice, the dominant term is always

$$B_N\left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j, 1 + \frac{1 + e_{\mathrm{O}}}{\nu_{\mathrm{I}}} \right), \quad \text{with } e_{\mathrm{O}} = \min\{\eta_{\mathrm{O}}, \mu_{\mathrm{O}} + \nu_{\mathrm{O}}\}.$$

Then the asymptotic approximation appears to be independent of h, and analogously for the asymptotic approximation with a fixed j. □

Remark 2.4. Theorem 2.3 expresses $c_{j,h}$ for a fixed h as a linear combination of the $h+1$ functions

$$B_N\left(\frac{\mu_{\rm I}}{\nu_{\rm I}}+j, 1+\frac{1+\eta_{\rm O}}{\nu_{\rm I}}\right)$$

and

$$B_N\left(\frac{\mu_{\rm I}}{\nu_{\rm I}}+j, 1+\frac{1+\mu_{\rm O}+\nu_{\rm O} i}{\nu_{\rm I}}\right), \quad \text{for } i=1,\ldots,h,$$

which are linearly independent under the hypothesis that $\eta_{\rm O} \neq \mu_{\rm O}+\nu_{\rm O} i$ for all $i \leq h$. If this condition is violated, i.e., $\eta_{\rm O}=\mu_{\rm O}+\nu_{\rm O} i$ for an index $i \leq h$, then we must complete the basis by replacing $B_N(\frac{\mu_{\rm I}}{\nu_{\rm I}}+j, 1+\frac{1+\eta_{\rm O}}{\nu_{\rm I}})$ with a different function that still satisfies (2.9) and is linearly independent of all the other functions of the basis. For example, if $h=1$ and $\mu_{\rm O}+\nu_{\rm O}=\eta_{\rm O}$, it can be shown by direct substitution that such a function has the form

$$B_N\left(\frac{\mu_{\rm I}}{\nu_{\rm I}}+j, 1+\frac{1+\eta_{\rm O}}{\nu_{\rm I}}\right)\phi(j),$$

where

$$\phi(j)=\phi(j-1)+\frac{1}{s_{j,0}}, \quad \text{with } \phi(0)=0.$$

The cases $\eta_{\rm I}=\mu_{\rm I}+\nu_{\rm I} i$ and $\eta_{\rm O}=\mu_{\rm O}+\nu_{\rm O} i$ for some i are highly improbable (actually, they have never been found in experiments with real Web subsets). For this reason, they are not taken into consideration.

The inlink and outlink distributions, in contrast to the components of the steady-state solutions, can be expressed in a simple way in terms of beta functions.

Theorem 2.5. *The functions x_j and y_h satisfy the difference equations*

$$\left(1+\pi_{\rm I}^{(j)}\right)x_j=\pi_{\rm I}^{(j-1)}x_{j-1}, \quad \textit{with } \left(1+\pi_{\rm I}^{(1)}\right)x_1=\alpha_{\rm I}+\gamma_{\rm I}, \tag{2.22}$$

$$\left(1+\pi_{\rm O}^{(h)}\right)y_h=\pi_{\rm O}^{(h-1)}y_{h-1}, \quad \textit{with } \left(1+\pi_{\rm O}^{(1)}\right)y_1=\alpha_{\rm O}+\gamma_{\rm O}. \tag{2.23}$$

The solutions of these equations are

$$x_j=\frac{\alpha_{\rm I}+\gamma_{\rm I}}{\mu_{\rm I}}\ B_N\left(\frac{\mu_{\rm I}}{\nu_{\rm I}}+j, 1+\frac{1}{\nu_{\rm I}}\right), \quad \textit{with } j\geq 1, \tag{2.24}$$

$$y_h=\frac{\alpha_{\rm O}+\gamma_{\rm O}}{\mu_{\rm O}}\ B_N\left(\frac{\mu_{\rm O}}{\nu_{\rm O}}+h, 1+\frac{1}{\nu_{\rm O}}\right), \quad \textit{with } h\geq 1, \tag{2.25}$$

and their asymptotic approximations for j and $h \to \infty$ are

$$x_j \approx \left(\frac{\mu_{\mathrm{I}}}{\nu_{\mathrm{I}}} + j\right)^{-(1+1/\nu_{\mathrm{I}})}, \tag{2.26}$$

$$y_h \approx \left(\frac{\mu_{\mathrm{O}}}{\nu_{\mathrm{O}}} + h\right)^{-(1+1/\nu_{\mathrm{O}})}. \tag{2.27}$$

Proof. By summing (2.9) on h we get for any integer $n \geq 2$,

$$\begin{aligned}
&\left(1 + \pi_{\mathrm{I}}^{(j)}\right) \sum_{h=0}^{n} c_{j,h} \\
&\quad = \pi_{\mathrm{I}}^{(j-1)} \sum_{h=0}^{n} c_{j-1,h} + \sum_{h=1}^{n} \pi_{\mathrm{O}}^{(h-1)} c_{j,h-1} - \sum_{h=0}^{n} \pi_{\mathrm{O}}^{(h)} c_{j,h} \\
&\quad = \pi_{\mathrm{I}}^{(j-1)} \sum_{h=0}^{n} c_{j-1,h} - \pi_{\mathrm{O}}^{(n)} c_{j,n}.
\end{aligned}$$

Equation (2.22) follows, since $\lim_{n\to\infty} n c_{j,n} = 0$. The initial position is found in a similar way using formula (2.10) and noticing that

$$\eta_{\mathrm{I}} \sum_{h\geq 1} c_{0,h} = \eta_{\mathrm{I}}(\alpha_{\mathrm{O}} - \gamma_{\mathrm{I}}) = \gamma_{\mathrm{I}}.$$

The proof of (2.23) is analogous. To find the explicit and the asymptotic expressions of the solutions, Lemma 2.1 is exploited, as was already done in the proof of Theorem 2.2. □

Remark 2.6. Theorem 2.5 shows that while the overall linking distribution $Z_{j,h}^{(t)}$ is described by a coupled model, solutions (2.24) and (2.25) for the indegree and outdegree distributions depend on the inlink and outlink parameters in a disjointed way.

3. Real Web Experimentation

We are interested in verifying, by means of experiments, how the proposed model gives an appropriate description of the page distribution as a function of the number of inlinks and outlinks. For this purpose we fit data obtained from real Web subsets. These data were collected by a crawler that visited the Web beginning with a predetermined list of URLs and downloaded the URLs of the pages it encountered. The process was repeated recursively until a certain depth was reached. Note that the data were inevitably influenced by the different policies

implemented for crawling and by the limitation of the search. Moreover, since the pages with indegree zero could not be reached by the crawler, the corresponding data are not available in real Web subsets. Further, we must take into account that we are trying to approximate an intrinsically irregular discrete process with a continuous estimate.

To validate our model we conduct our experiments on the three data sets it-2004, sk-2005, and uk-2005, which are freely available from the WebGraph homepage [Boldi and Vigna 04].[1] These data sets were obtained from a crawl performed by UbiCrawler [Boldi et al. 04] on the `.it` domain in 2004, the `.sk` domain in 2005, and the `.uk` domain in 2005. The it-2004 graph contains 41.3 Mpages and 1.15 Glinks. The sk-2005 graph contains 50.6 Mpages and 1.95 Glinks. The uk-2005 graph contains 39.5 Mpages and 936 Mlinks.

For each subset of the real Web, we consider a sparse representation of the degree distribution

$$W = \{(j, h, W_{j,h}),\ (j,h) \in Q\},$$

where $W_{j,h}$ is the number of pages having indegree j and outdegree h and Q is the set of pairs of indices corresponding to nonzero values of $W_{j,h}$. As an example, in Figure 9 (upper) the degree distribution W of the sk-2005 graph is plotted on a log-log scale. Each point represents a triple $(j, h, W_{j,h}) \in W$. The log-log scale is usually employed in the graphic representation of Web data that span many different orders of magnitude.

From the set W we derive four subsets:

1. the set $U_1 = \{(j, W_{j,0}), j \in Q_1\}$, where Q_1 is the subset of indices j corresponding to pages with indegree j and outdegree 0;
2. the set $U_2 = \{(W_{1,h}, h), h \in Q_2\}$, where Q_2 is the subset of indices h corresponding to pages with indegree 1 and outdegree h;
3. the set $U_3 = \{(j, S_j), j \in Q_3\}$, where $S_j = \sum_h W_{j,h}$ and Q_3 is the subset of indices j such that at least one page with indegree j exists;
4. the set $U_4 = \{(h, T_h), h \in Q_4\}$, where $T_h = \sum_j W_{j,h}$ and Q_4 is the subset of indices h such that at least one page with outdegree h exists.

Note that the subset U_2 refers to pages with indegree 1 instead of indegree 0, due to the fact that the data corresponding to pages with indegree 0 are not available. The sets U_3 and U_4 represent the inlink and the outlink distributions, which are of primary interest in our investigation.

[1] See http://law.dsi.unimi.it.

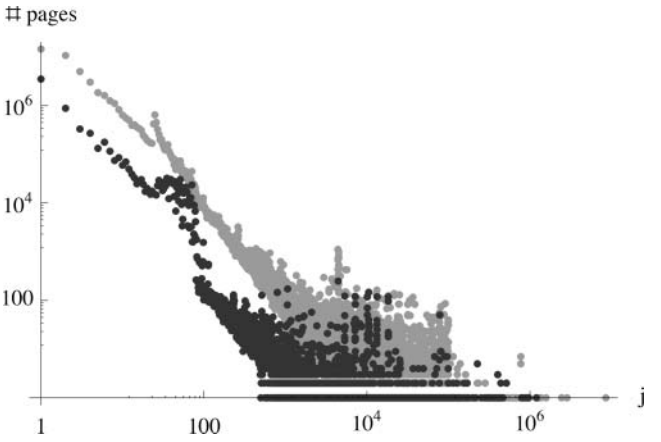

Figure 1. Sets U_1 (black dots) and U_3 (gray dots) for sk-2005.

In Figures 1 and 2, the four sets U_1–U_4 are plotted for sk-2005. These plots have a characteristic shape, with a thin head and a large tail, due to the presence of noise, more evident in the data with higher indices. It is self-evident that the shapes of inlinks and outlinks are different and that while the power law can be an acceptable approximation to the inlink distribution, this is absolutely not true for the outlink distribution.

When fitting Web data, to avoid having the different orders of magnitude affect the result, a logarithmic fit is routinely performed. Hence the following problem

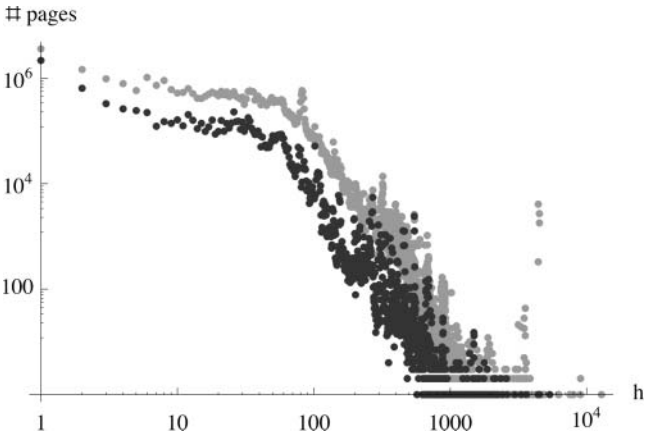

Figure 2. Sets U_2 (black dots) and U_4 (gray dots) for sk-2005.

is addressed: find the values of the parameters of the model that minimize the objective function

$$\psi = \sum_{j \in Q_1} \left(\log(tc_{j,0}) - \log W_{j,0} \right)^2 + \sum_{h \in Q_2} \left(\log(tc_{1,h}) - \log W_{1,h} \right)^2 + \sum_{j \in Q_3} \left(\log(tx_j) - \log S_j \right)^2 + \sum_{h \in Q_4} \left(\log(ty_h) - \log T_h \right)^2,$$

where the functions $c_{j,0}$, $c_{1,h}$, x_j, and y_h are of the form given in (2.13), (2.19), (2.24), and (2.25), respectively. The parameters of the model are then α_{I}, β_{I}, γ_{I}, α_{O}, β_{O}, γ_{O}, all in the interval $(0, 1)$, and the time $t > 0$.

The steady-state solution $S_j^{(t)}$ of the model is a deterministic, continuous, and monotonic function, while the real Web data consist of integer spread points. The most straightforward interpretation that gets these two facts to agree is to look at $S_j^{(t)}$ as the expected value of an integer random variable.

A cursory glance at the plots shows that the lower parts of the graphs have a more compact and regular shape, while the upper parts appear to be more spread out. This observation suggests fitting the lower envelope of the data to approximate the lower edge of the cloud and to consider an additive correction modeled by a suitable probability distribution (in [Favati et al. 08], a geometric distribution is experimentally shown to match the real Web data).

We then substitute the sets U_1 to U_4 by the corresponding lower envelopes to make the fit. In this way, the number of data in the tail is reduced, thus balancing the weight of the head and tail.

Table 1 shows the values of the parameters found by applying the fitting procedure to the above data sets.

Using the parameter values given in Table 1, we can derive the values of the parameters that characterize the asymptotic approximations (2.26) and (2.27) of the indegree and outdegree distributions given in Theorem 2.5 for the considered

Data set	α_{I}	β_{I}	γ_{I}	α_{O}	β_{O}	γ_{O}	t
it-2004	0.002	0.022	0.017	0.052	0.723	0.001	1.37×10^9
sk-2005	0.005	0.023	0.008	0.036	0.780	0.002	1.83×10^9
uk-2005	0.002	0.030	0.016	0.048	0.833	0.001	1.34×10^9

Table 1. Values of the inlink and outlink parameters and of the time.

Data set	σ_{I}	ρ_{I}	σ_{O}	ρ_{O}
it-2004	1.18	1.04	49.8	3.81
sk-2005	1.90	1.04	92.4	4.71
uk-2005	1.74	1.05	104.	6.30

Table 2. Values of the parameters of the asymptotic approximations for the indegree and outdegree distributions.

real data sets. These asymptotic approximations are rewritten as

$$x_j \approx \left(\sigma_{\mathrm{I}} + j\right)^{-(1+\rho_{\mathrm{I}})}, \tag{3.1}$$

$$y_h \approx \left(\sigma_{\mathrm{O}} + h\right)^{-(1+\rho_{\mathrm{O}})}, \tag{3.2}$$

and the values of the parameters σ_{I}, ρ_{I}, σ_{O}, ρ_{O} so obtained are listed in Table 2.

From direct inspection of the tables we note the following:

1. The values of β_{I} are much smaller than those of β_{O}, meaning that the preferential attachment is the dominant policy in the inlink distribution, while the outlink distribution is significantly ruled by the uniform attachment.

2. The values of γ_{O} are very small, meaning that a page born without outlinks acquires new outlinks with low probability. This fact has been already noted in [Bollobás et al. 03], where these pages are referred to as pages that purely provide content. As a consequence, with high probability, a new outlink is created either together with a new page or leaving a page already having outlinks. Moreover, a small γ_{O} implies that $1 + e_{\mathrm{O}}$ is nearly equal to 1, i.e., the exponent in the asymptotic approximation (2.20) is close to the exponent $-(1 + \rho_{\mathrm{I}})$ of (3.1), meaning that all the cross sections, for a fixed h, have decay behavior very similar to that of the indegree distribution.

3. The values of α_{I} are smaller than γ_{I}, meaning that the probability of pointing to a new page is lower than that of pointing to an already existing page with outdegree greater than zero and zero indegree.

4. Since the values of σ_{I} are small with respect to even low indegree values j, the asymptotic approximation (3.1) is close to a power law, confirming that the power law gives an acceptable description of the indegree distribution. In fact, in this case the log-log plot of the lower envelope of the data is nearly rectilinear (see Figure 3). The same conclusion does not hold for the asymptotic approximation (3.2) of the outdegree distribution, because the values of σ_{O} are not negligible with respect to low outdegree values of h.

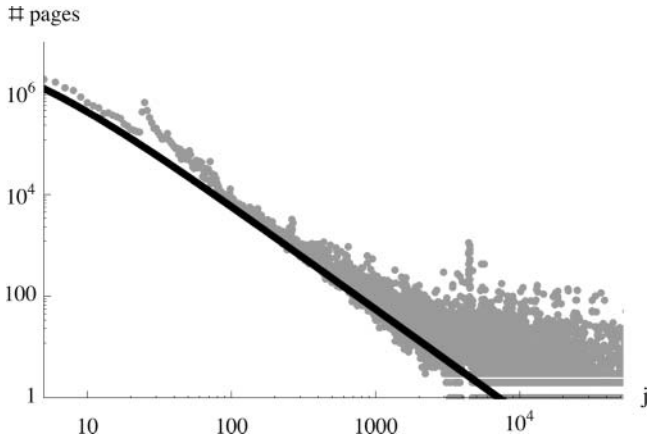

Figure 3. Fitting function x_j given in (2.24) (black solid line) for the indegree distribution (gray dots) of sk-2005.

In this case the log-log plot of the lower envelope of the data is concave downward, and the larger the σ_O value, the wider the initial region of slow decrease (see Figure 4). The values of ρ_I agree with the measures generally accepted, while the values of ρ_O are not comparable with those given in the literature, since the fit on the whole region cannot have the exponent of a power law.

Comments 1, 2, and 3 are straightforward and agree with what is generally presented in the literature, providing an implicit validation of the model.

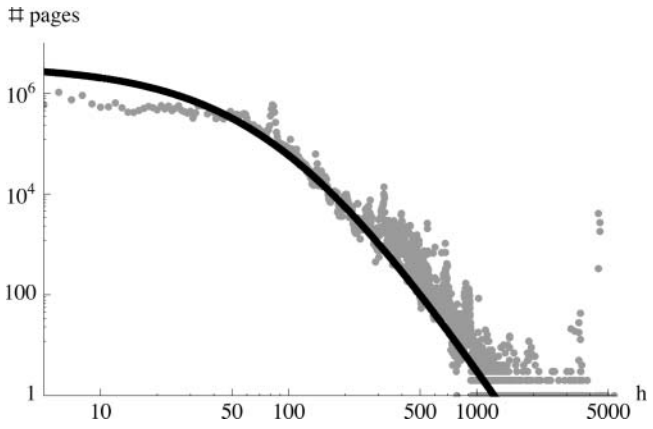

Figure 4. Fitting function y_h given in (2.25) (black solid line) for the outdegree distribution (gray dots) of sk-2005.

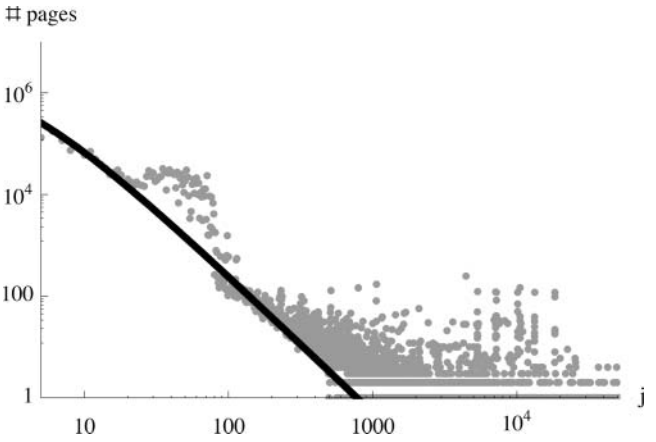

Figure 5. Fitting function $c_{j,0}$ given in (2.13) (black solid line) for the set U_1 (cross section of the degree distribution at the fixed outdegree value $h = 0$) of sk-2005 (gray dots).

Figures 3 and 4 show how well (2.24) and (2.25), with the values of the parameters given in Table 1, approximate the real indegree and outdegree distributions for sk-2005.

Figures 5–8 show how well the cross sections of the real degree distribution are fitted by the cross sections of the steady-state solution. Figures 5 and 7 refer

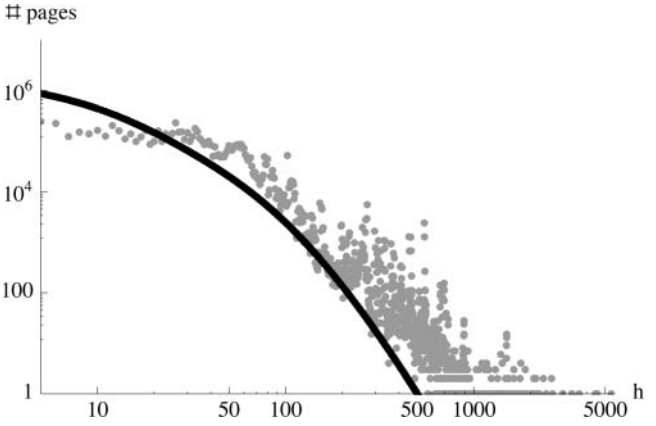

Figure 6. Fitting function $c_{1,h}$ given in (2.19) (black solid line) for the set U_2 (cross section of the degree distribution at the fixed indegree value $j = 1$) of sk-2005 (gray dots).

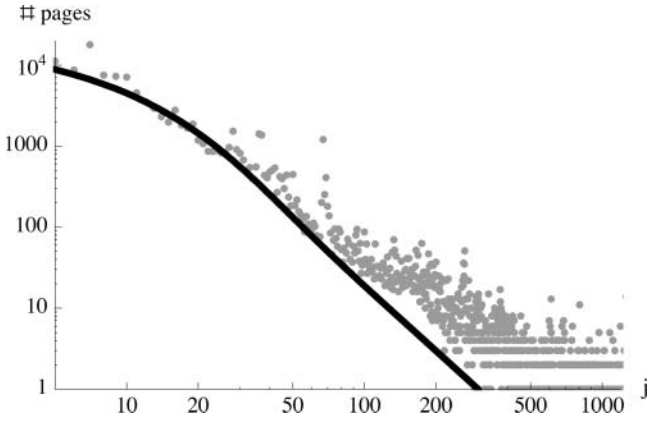

Figure 7. Fitting function $c_{j,50}$ given in (2.18) (black solid line) for the cross section of the degree distribution at the fixed outdegree value $h = 50$ of sk-2005 (gray dots).

to the indegree distribution at the values $h = 0$ (i.e., the set U_1) and $h = 50$ of the outdegree. Figures 6 and 8 refer to the outdegree distribution at the values $j = 1$ (i.e., the set U_2) and $j = 50$ of the indegree.

Finally, Figure 9 shows 3D log-log graphs of the degree distribution of sk-2005 and of the corresponding solution $Z_{j,h}^{(t)} = tc_{j,h}$ generated using the model (2.9) with the values of the parameters given in Table 1.

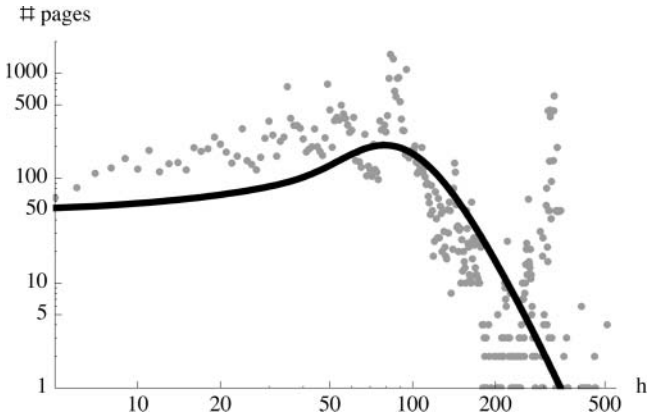

Figure 8. Fitting function $c_{50,h}$ given in (2.19) (black solid line) for the cross section of the degree distribution at the fixed indegree value $j = 50$ of sk-2005 (gray dots).

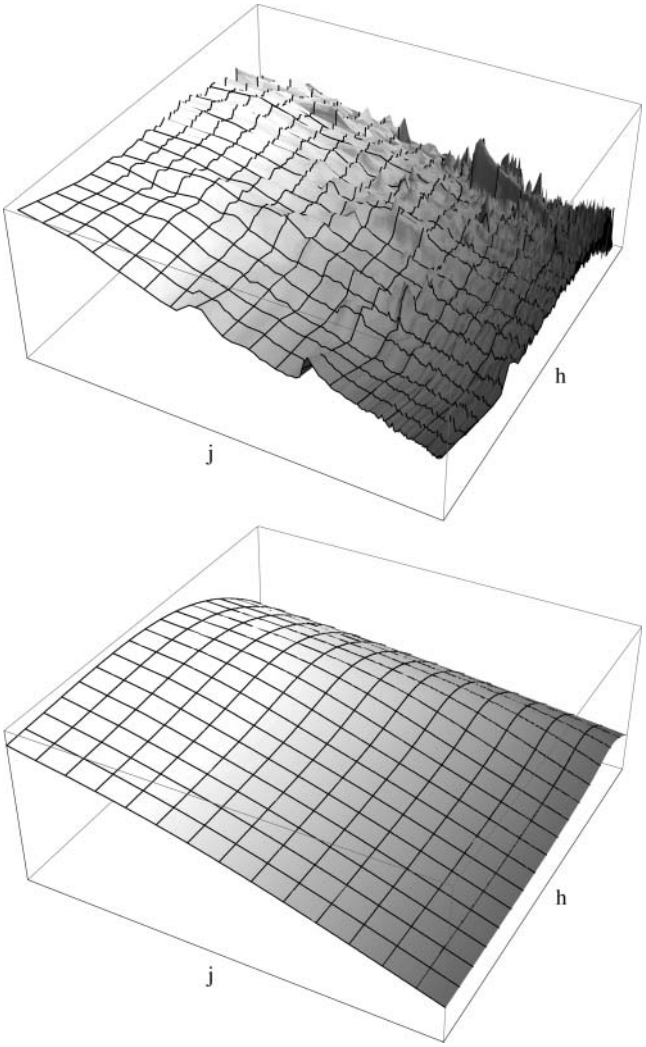

Figure 9. Degree distribution $W_{j,h}$ of sk-2005 (upper) and the corresponding solution $Z^{(t)}_{j,h} = tc_{j,h}$ generated using the model (2.9) (lower).

4. Conclusions

A coupled model for the Web graph that simultaneously describes the inlink and outlink distributions, by taking into account the interconnection of the two processes, has been proposed. This bidimensional model is a refined version of the monodimensional model presented in [Favati et al. 08].

In contrast to the monodimensional case, a closed form of the steady-state solution of the bidimensional model is not available. The asymptotic expressions of the steady-state indegree (outdegree) distributions for each fixed value of the outdegree (indegree) have already been given in the literature. With our approach, these distributions can be explicitly expressed as sums of beta functions, whose parameters depend on a unique process of link generation.

A fit procedure has been proposed to compute the parameters α_{I}, β_{I}, γ_{I}, α_{O}, β_{O}, γ_{O}, and t of the model for real Web data sets. Experimentation has shown that the proposed distributions well reproduce the behavior of the observed data.

References

[Barabasi and Albert 99] A. L. Barabasi and R. Albert. "Emergence of Scaling in Random Networks." *Science* 286 (1999), 509–512.

[Barabasi et al. 99] A. L. Barabasi, R. Albert, and H. Jeong. "Mean-Field Theory for Scale-Free Random Networks." *Physica A* 272 (1999), 173–187.

[Boldi and Vigna 04] P. Boldi and S. Vigna. "The WebGraph Framework I: Compression Techniques." In *Proc. of the Thirteenth International World Wide Web Conference (WWW 2004)*, pp. 595–601. New York: ACM, 2004.

[Boldi et al. 04] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. "Ubicrawler: A Scalable Fully Distributed Web Crawler." *Software: Practice and Experience* 34 (2004), 711–726.

[Bollobás et al. 03] B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. "Directed Scale-Free Graphs." In *Proceedings of the ACM–SIAM Symposium on Discrete Algorithms (SODA)*, pp. 132–139, 2003.

[Broder et al. 00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph Structure in the Web." In *Proceedings of the Ninth International World Wide Web Conference*, pp. 309–320, 2000.

[Caldarelli et al. 03] G. Caldarelli, P. De Los Rios, L. Laura, S. Leonardi, and S. Millozzi. "A Study of Stochastic Models for the Web Graph." Technical Report 04-03, Dip. di Informatica e Sistemistica, Università di Roma "La Sapienza," 2003.

[Cooper 06] C. Cooper. "Distribution of Vertex Degree in Web-Graphs." *Combinatorics, Probability and Computing* 15 (2006), 637–661.

[Cooper and Frieze 01] C. Cooper and A. M. Frieze. "A General Model of Undirected Web Graphs." In *Proceedings of the Ninth Annual European Symposium on Algorithms, LNCS n.2161*, pp. 500–511. Berlin: Springer, 2001.

[Dorogovtsev et al. 00] S. Dorogovtsev, J. Mendes, and A. Samukhin. "Structure of Growing Networks: Exact Solution of the Barabasi–Albert's Model." *Phys. Rev. Lett.* 85 (2000), 4633–4636.

[Favati et al. 08] P. Favati, G. Lotti, O. Menchi, and F. Romani. "A Stochastic Model for the Link Analysis of the Web." *Internet Mathematics* 3 (2008), 509–531.

[Kleinberg et al. 99] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "The Web as a graph: Measurements, Models and Methods." In *Proc. Intl. Conf. on Combinatorics and Computing*, pp. 1–18, 1999.

[Kumar et al. 00] S. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. "Stochastic Models for the Web Graph." In *Proc. 41th IEEE Symposium on Foundations of Computer Science*, pp. 57–65, 2000.

[Leskovec et al. 05] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 177–187, 2005.

[Mitzenmacher 06] M. Mitzenmacher. "The Future of Power Law Research" (editorial). *Internet Mathematics* 2 (2006), 525–534.

[Pennock et al. 02] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. "Winners Don't Take All: Characterizing the Competition for Links on the Web." *Proceedings of the National Academy of Science* 99 (2002), 5207–5211.

[Simon 55] H. A. Simon. "On a Class of Skew Distribution Functions." *Biometrika* 42 (1955), 425–440.

P. Favati, IIT–CNR, Via G. Moruzzi 1, 56124 Pisa, Italy (paola.favati@iit.cnr.it)

G. Lotti, Dipart. di Matematica, University of Parma, Viale G.P. Usberti 53/A, 43100 Parma, Italy (grazia.lotti@unipr.it)

O. Menchi, Dipart. di Informatica, University of Pisa, Largo Pontecorvo 3, 56127 Pisa, Italy (menchi@di.unipi.it)

F. Romani, Dipart. di Informatica, University of Pisa, Largo Pontecorvo 3, 56127 Pisa, Italy (romani@di.unipi.it)

Received July 28, 2008; accepted December 2, 2010.