Internet Mathematics Vol. 3, No. 4: 509-531

A Stochastic Model for the Link Analysis of the Web

Paola Favati, Grazia Lotti, Ornella Menchi, and Francesco Romani

Abstract. The behavior of inlink and outlink distributions appears to be one of the most studied properties of the web structure. The literature agrees that the inlink distribution follows a power law, but no such agreement exists for the outlink distribution. Accurate observations show that in the low-degree region the link distribution fails to fit a power law with a discrepancy larger for outlinks than for inlinks. Moreover, a power law, as well as any continuous function, does not fit the scattered behavior shared by both the link distributions for large-degree values. The linking model we consider here is a mixed one, based on both the preferential attachment strategy and the uniform attachment strategy. A new approximation technique is devised to detect the parameters of the steady state solution that describe a real data set. A stochastic technique is suggested to describe the scattering of the data. With these techniques the model appears to be well suited for describing both inlink and outlink distributions. The experimentation on subsets of the World Wide Web and of Wikipedia shows that our approach produces an approximation more adequate than the power law. This approximation suggests that the two attachment strategies play a different role in the inlink and the outlink cases.

1. Introduction

Developing a realistic model of the World Wide Web is a relevant task for many reasons: for example, designing and testing web applications, studying the time evolution of the web, and detecting statistical peculiarities of subsets of the web. Inlink and outlink distributions represent important elements of such a model.

In the literature it is widely accepted [Albert et al. 99, Barabasi and Albert 99, Barabasi et al. 99, Kumar et al. 99] that the inlinks follow a power law [Adamic 02], and some experimental observations of real web data (see, for

1542-7951/06 $0.50 per page

example, [Broder et al. 00]) validate this distribution, with an exponent around 2.1. The power law behavior is shared by many man-made and naturally occurring phenomena, which have been studied even before the web era (see, for example, [Simon 55] and [Mitzenmacher 03] with its extensive bibliography).

More accurate observations showed that in the low-degree region the link distribution fails to fit a power law with a discrepancy larger for outlinks than for inlinks [Caldarelli et al. 03]. It has been conjectured that the pages with a low number of outlinks might follow a Poisson distribution or a combination of Poisson and power law distributions [Broder et al. 00]. Since the processes of inlinking and outlinking are two aspects of the same linking phenomenon, we expect that just one model may be devised to describe the inlink and outlink distributions. Hence, we aim at studying a model that suits the two distributions for different values of its parameters.

Several models have been proposed to explain the power law distribution, most of them based on preferential attachment. According to it, a new link points to a page with a probability proportional to the current degree of the page. A frequently-cited preferential attachment model is the Barabasi-Albert model [Barabasi and Albert 99, Barabasi et al. 99]. To give newer pages a larger chance to compete for links, mixed models have been proposed, which combine both preferential attachment and a uniform baseline probability of attachment [Cooper and Frieze 01, Dorogovtsev et al. 00, Pennock et al. 02]. Actually, mixed models are sufficient to explain the observed deviation from the power law behavior (see, for example, [Pennock et al. 02]).

In this paper we adopt a mixed model that depends on three parameters: two parameters α and β of a probabilistic nature and the time t at which the phenomenon is analyzed. The steady state solution of the model is expressed in terms of a beta function. Detecting the values of the parameters that correspond to the link distribution of a given real data set is not an easy task. A least-squares fit of the sought-after solution to the data can be difficult, due to the huge dimension of the data. This difficulty can be dealt with by reducing the number of the parameters [Pennock et al. 02]. Moreover, the fitting procedure applied to an inlink distribution can lead to values of the parameters not consistent with the model. To overcome this drawback, we suggest applying the least-squares fit to suitably preprocessed data.

The fit of the beta function to the preprocessed data leads to a nonlinear least-squares procedure. Thus, good approximations of the beta function that ease this task would be welcome. Actually, we find a good approximation to the inverse of the solution, which turns out to be a Yule function. The multiplicative form of such a function allows a linear least-squares fit in the reversed log-log space. This result would be interesting by itself if coupled with a model describing the reversed data. However, in the absence of such a model, the Yule function obtained by the fit can be used to approximate the parameters of the beta function at a low cost.

The log-log plots of web data subsets show characteristic noisy tails. Since the beta function is continuous and monotonic, a simple rounding cannot reproduce such a scattered behavior. This observation suggests the application of a stochastic correction, according to an integer probability function $P(n)$. It seems reasonable to suppose that $P(n)$ is a decreasing function related to the value of the beta function.

The experimentation, made for both the inlink and outlink distributions of real data sets, shows that the proposed model is valid and that the outlined procedures for detecting correct values of the parameters are effective. The preferential attachment appears to be the dominant policy in the inlink distribution, while the outlink distribution appears to be significantly ruled by the uniform attachment.

In Section 2 the linking model is introduced and its steady state solution is analyzed; in Section 3 the difficulties inherent to the fit procedures are examined; in Section 4 approximations of the beta function are studied and an approximation of the inverse of the beta function is proposed in terms of the Yule function. In Section 5 the data used for the experiments are presented together with the proposed technique for the stochastic correction, and the results of the experimentation are presented and discussed. Conclusions and future developments are given in Section 6.

2. The Stochastic Problem

At an abstract level the web can be represented as a direct graph of nodes, called *pages*, connected by arcs, called *links*. The terms *inlink* and *outlink* are used to indicate a link pointing to a page and a link originating from a page, respectively. Two important indicators are associated with each page: the *in-degree*, i.e., the number of inlinks pointing to that page, and the *out-degree*, i.e., the number of outlinks originating from that page.

We want to examine the web structure from both the inlink and the outlink points of view: the aim is to find how many pages of the web have a given in-degree (or out-degree) j. In the following the term *degree* will be used for either in-degree or out-degree. Hence, we will examine how the number X_j of pages with degree j depends on j. The analysis of large subsets of the web shows that there are many pages with small degrees and few pages with large degrees. This behavior can be approximately described by a power law, i.e., $X_j = a\, j^{-k}$, with constant parameters a and k.

To find an adequate model for the function X_j, a discrete-time stochastic process is considered. At any time step a new link is created: with probability α it is connected to a page having zero degree, and with probability $1 - \alpha$ it is connected to a page having nonzero degree. For the inlinks this means that the new link points with probability α to a page without inlinks (possibly because

it was created recently) and with probability $1-\alpha$ to a page already having inlinks. In the latter case, the most common inlink model is based on the two following assumptions:

1. With probability β the new inlink points to a page chosen at random. This policy is known as *uniform attachment*. If it was the only policy applied, then all the pages would acquire approximately the same number of inlinks.

2. With probability $1-\beta$ the new inlink points to a page chosen proportionally to the in-degree of that page. This policy is known as *preferential attachment* and expresses the concept that new links tend to attach themselves to pages that already have many inlinks.

For the outlink model the same stochastic process can be considered. The probability α refers to the creation of a link pointing out from a page without outlinks. In this case assumption 2 can be given the following interpretation: the new outlink tends to exit from a page already having many outlinks.

As we will see from the experiments, the values of the parameters α and β detected to describe the inlink distribution and the outlink distribution are very different and agree with the expectations. In fact, it seems reasonable to expect that the outlink process is independent from the importance of the target page and depends on the characteristics of the source page. Hence, the outlink process has mainly a random feature. On the contrary, we expect that the inlink process relies more on the importance of the target page.

2.1. The Linking Model

Let $X_j^{(t)}$ be the number of pages having degree j at time t. The time is increased by one when a link is created, and a page is counted only when at least one link points to it (for the inlink distribution) or out from it (for the outlink distribution). For the inlinks this reflects what is made in practice, since the web data are collected through a crawler that cannot reach a page without inlinks.

At time t, the number of pages is $n(t) = \sum_j X_j^{(t)}$ and the number of links is $t = \sum_j j\, X_j^{(t)}$. The expected value of the variation of $n(t+1)$ with respect to $n(t)$ is

$$\mathcal{E}\big[n(t+1) - n(t)\big] = \alpha. \tag{2.1}$$

We assume that initially $t = 1$ and $X_1^{(1)} = 1$.

When a new link is created, the probability $p(j,t)$ that it is connected with a page having degree j is given by two terms: the first term, according to assumption 1, is proportional to $X_j^{(t)}$, and the second term, according to assumption 2, is proportional to the number of all existing links connected with pages having

degree j, i.e., $j\,X_j^{(t)}$. Hence,

$$p(j,t) = (1-\alpha)\,\left[\frac{\beta}{n(t)}\,X_j^{(t)} + \frac{1-\beta}{t}\,j\,X_j^{(t)}\right]. \tag{2.2}$$

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

$$\mathcal{E}\big[X_j^{(t+1)} - X_j^{(t)}\big] = p(j-1,t) - p(j,t), \quad j = 2,\ldots,t. \tag{2.3}$$

The equation holds also for $j = 1$ and for $j = t+1$ provided that we set

$$p(0,t) = \alpha \quad \text{and} \quad X_{t+1}^{(t)} = 0.$$

Model (2.3) is one of the many versions of mixed models (see, for example, [Cooper and Frieze 01, Dorogovtsev et al. 00, Pennock et al. 02]).

2.2. The Steady State Solution

To find the steady state distribution of the stochastic process, following [Simon 55] we replace the expected values in equations (2.1) and (2.3) by their actual values, obtaining the difference equations

$$n(t+1) = n(t) + \alpha, \quad n(1) = 1, \tag{2.4}$$

and

$$X_j^{(t+1)} - X_j^{(t)} = p(j-1,t) - p(j,t), \quad j = 1,\ldots,t+1. \tag{2.5}$$

By applying recursively this equation we can see how $X_j^{(t)}$ evolves when t increases. For this reason we will call $X_j^{(t)}$ the *transient solution*.

For large t from Equation (2.4) we get

$$n(t) \sim \alpha t, \tag{2.6}$$

meaning that the ratio $1/\alpha$, which represents the average number of links per page, is asymptotically constant (compare with [Pennock et al. 02], where this ratio is assumed to be constant at any time and taken as a parameter of the model). Approximation (2.6) is substituted into equation (2.5), giving for $t \geq 1$ the recursion

$$\begin{cases} X_j^{(t+1)} - X_j^{(t)} = \alpha - \dfrac{q_1}{t}\,X_1^{(t)}, & \text{for } j = 1, \\ X_j^{(t+1)} - X_j^{(t)} = \dfrac{1}{t}\big(q_{j-1}X_{j-1}^{(t)} - q_j X_j^{(t)}\big), & \text{for } j = 2,\ldots,t+1, \\ X_j^{(t+1)} = 0, & \text{for } j = t+2, \end{cases} \tag{2.7}$$

where $q_j = (1-\alpha)\,\Big(\dfrac{\beta}{\alpha} + (1-\beta)j\Big)$.

The steady state solution $S_j^{(t)}$ is found by dropping the third equation from system (2.7) and letting the second equation hold for any $j \geq 2$, i.e.,

$$S_j^{(t+1)} - S_j^{(t)} = \begin{cases} \alpha - \dfrac{q_1}{t}\, S_1^{(t)}, & \text{for } j = 1, \\ \dfrac{1}{t}\big(q_{j-1} S_{j-1}^{(t)} - q_j S_j^{(t)}\big), & \text{for } j \geq 2. \end{cases} \tag{2.8}$$

The solution $S_j^{(t)}$ can be expressed in terms of the complete beta function, as can be seen by direct substitution:

$$S_j^{(t)} = \zeta\, B\big(\sigma + j, \rho + 1\big), \quad \text{for} \quad j \geq 1, \tag{2.9}$$

where

$$\zeta = \frac{\alpha\, \rho\, t}{(\sigma + \rho + 1)\, B\big(\sigma + 1, \rho + 1\big)}, \quad \sigma = \frac{\beta}{\alpha(1-\beta)}, \quad \rho = \frac{1}{(1-\alpha)(1-\beta)}.$$

Since $S_j^{(t)}$ satisfies

$$S_j^{(t)} = \frac{q_{j-1}}{1 + q_j}\, S_{j-1}^{(t)},$$

it can be shown that for any $\rho \geq 1$

$$\sum_{j=1}^{\infty} S_j^{(t)} = \alpha\, t \quad \text{and} \quad \sum_{j=1}^{\infty} j\, S_j^{(t)} = t. \tag{2.10}$$

Log-log plots are usually employed for the graphic representation of the data in the web context. Figure 1 shows the steady state solution $S_j^{(t)}$ compared with

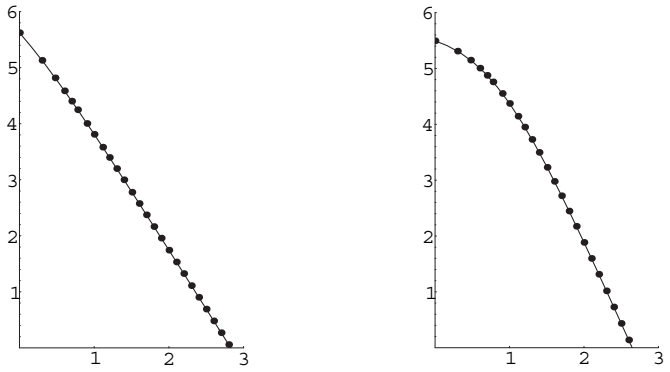

Figure 1. The dots show the transient solution $X_j^{(t)}$ at time $t = 8 \times 10^6$ in the cases $\alpha = 0.1$, $\beta = 0.1$ (left) and $\alpha = 0.15$, $\beta = 0.5$ (right). The solid lines represent the corresponding steady state solution $S_j^{(t)}$.

the transient solution $X_j^{(t)}$ at time $t = 8 \times 10^6$, corresponding to two different choices of the parameters α and β. In both cases $S_j^{(t)}$ appears to be a very good approximation of $X_j^{(t)}$. Actually this happens for any α and β.

3. The Fitting Problem

A technique to verify if the proposed model suits the World Wide Web consists of finding the parameters α, β, and t that best fit the real data. Hence, the following problem should be dealt with: given a subset of the web

$$W = \big\{(j, W_j), \quad j \in J\big\}, \tag{3.1}$$

where W_j is the number of pages having degree j and J is the set of the indices of the components W_j effectively present, find the parameters $(\alpha, \beta) \in \mathcal{A} = (0,1)\times(0,1)$ and $t > 0$ such that the function $S_j^{(t)}$ of the form (2.9) best describes W_j. A fitting procedure can be applied to answer this problem. When j varies in J, both the data and the beta function span many orders of magnitude. To avoid the data with smaller indices having too much weight in the fit, it is preferable to take the logarithm of the data, i.e., to consider the minimum problem

$$\min_{\alpha,\,\beta,\,t} \sum_{j\in J} (\log W_j - \log S_j^{(t)})^2, \quad \text{with} \quad (\alpha,\, \beta) \in \mathcal{A}. \tag{3.2}$$

Solving problem (3.2) is not as easy as it appears, due to both the huge dimension of the data and the nonlinear dependence of the logarithm of the beta function on α and β. These two facts contribute to the unstable behavior of the minimization, which requires a high-cost nonlinear procedure. Approximations of the beta function allowing a linear fit procedure would be fruitful. By addressing this problem, we found a useful approximation of the inverse of the beta function, which will be described in the next section.

4. Approximations of the Beta Function

4.1. Asymptotic Approximations of the Beta Function

A simple approximation can be obtained by using the asymptotic series expansion of the beta function

$$B(a,b) \propto \Gamma(b) a^{-b} \frac{1 - (1 + O(1/a)b(b-1)}{(2a)}. \tag{4.1}$$

Setting

$$\nu = \zeta\, \Gamma(\rho+1) = \frac{\alpha\, \rho\, t\, \Gamma(\rho+1)}{(\sigma+\rho+1)B\big(\sigma+1, \rho+1\big)},$$

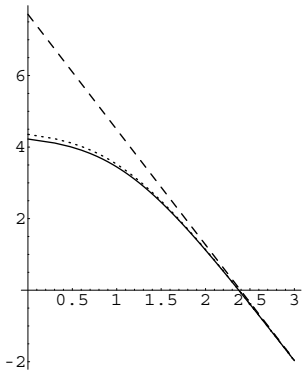

Figure 2. Steady state solution $S_j^{(t)}$ for $\alpha = 0.1$, $\beta = 0.5$, and $t = 10^6$ (solid line), approximation $V_j^{(t)}$ (dotted line), and Zipf's approximation $Z_j^{(t)}$ (dashed line).

we have

$$S_j^{(t)} \sim V_j^{(t)}, \quad \text{where} \quad V_j^{(t)} = \frac{\nu}{(\sigma + j)^{\rho+1}}, \tag{4.2}$$

and for j sufficiently large

$$S_j^{(t)} \sim Z_j^{(t)}, \quad \text{where} \quad Z_j^{(t)} = \frac{\nu}{j^{\rho+1}}. \tag{4.3}$$

The approximation $Z_j^{(t)}$ is frequently used to state that for any t the steady state solution satisfies a Zipf's law [Zipf 49], i.e., a decreasing power law. The approximation $V_j^{(t)}$ is similar to the one obtained in [Pennock et al. 02].

In the log-log scale the graph of $Z_j^{(t)}$ is a straight line. Hence, it is evident that approximating $S_j^{(t)}$ by a function of form (4.3) may be inadequate for moderately large values of j, as in the case of Figure 1 (right). Figure 2 shows $S_j^{(t)}$ for $\alpha = 0.1$, $\beta = 0.5$, and $t = 10^6$ (solid line), the approximation $V_j^{(t)}$ (dotted line), and the Zipf's approximation $Z_j^{(t)}$ (dashed line). Clearly, in this case the Zipf's approximation $Z_j^{(t)}$ is adequate only for asymptotic values of j.

The approximation error of $Z_j^{(t)}$ with respect to $S_j^{(t)}$ can be very large for small values of α and large values of β, while the approximation error of $V_j^{(t)}$ is much smaller. Unless the log-log plot of the real data shows a linear behavior, $Z_j^{(t)}$ cannot be exploited for best fitting the data. Approximation $V_j^{(t)}$ would be more adequate, but unfortunately also its logarithm does not depend linearly on the parameters.

4.2. Approximating the Inverse of the Beta Function

An interesting approximation is found by considering the inverse of the beta function. For simplicity sake, we denote in this section the steady state solution $S_j^{(t)}$ by $f(j)$, i.e.,

$$y = f(j) = \zeta\, B\big(\sigma + j, \rho + 1\big).$$

The image of f is $(0, y_{\max}]$ with $y_{\max} = S_1^{(t)}$. For any α, β, and t, function f is monotone decreasing. Let $j = g(y)$ denote its inverse. An explicit form of g is not known, so we consider the inverse of the approximating function $v(j) = V_j^{(t)}$, which is

$$j = v^{-1}(y) = \Big(\frac{\nu}{y}\Big)^{1/(\rho+1)} - \sigma.$$

Function $v^{-1}(y)$ appears to be a good approximation of $g(y)$. Unfortunately its logarithm also cannot be expressed linearly on the parameters. But we note that

$$v^{-1}(y)/\Big(\frac{\nu}{y}\Big)^{1/(\rho+1)} = 1 - \phi\, y, \quad \text{where} \quad \phi = \frac{\sigma}{y}\,\Big(\frac{y}{\nu}\Big)^{1/(\rho+1)}.$$

Since $\phi\ y$ is small, $1 - \phi\ y$ can be approximated by a function of the form b^y, where b is a constant less than 1 but close to 1. The smaller is ϕ, the better is the approximation. This fact suggests the consideration of, as possible approximation to $g(y)$, a function of the form

$$h(y) = c\, b^y\, y^{-r}, \tag{4.4}$$

where c, b, and r are suitable parameters. We expect r to be close to $1/(\rho + 1)$. Functions of the form (4.4) are known as *Yule functions*.

The form of function (4.4) has been chosen mainly because its logarithm can be expressed as a linear combination of the basis functions 1, y, $\log y$, that is,

$$\log h(y) = \log c\ + y\, \log b - r\, \log y. \tag{4.5}$$

In this way the parameters c, b, and r of a function $h(y)$ that suit given real data can be determined by applying a linear least-squares fit.

It remains to show that the inverse of $f(j)$ is indeed well approximated by a function of the form (4.4), for any $t > 0$ and $(\alpha, \beta) \in \mathcal{A}$. Actually, this analysis will be restricted to a closed subset $\mathcal{S}$ of $\mathcal{A}$:

$$\mathcal{S} = \{(\alpha, \beta)\ :\quad \alpha_1 \le \alpha \le \alpha_2, \quad \beta_1 \le \beta \le \beta_2\},$$

where $\alpha_1 > 0$, $\alpha_2 < 1$, and $\beta_1 > 0$ and $\beta_2 < 1$ are chosen in such a way to suit the World Wide Web (in the experiments $0.02 \le \alpha \le 0.12$ and $0.05 \le \beta \le 0.8$). We construct a procedure that maps the triple $(\alpha,\ \beta,\ t)$ defining $f(j)$ into the triple $(c,\ b,\ r)$ defining the best approximation $h(y)$ of the inverse of $f(j)$. Another procedure, which maps $(c,\ b,\ r)$ into $(\alpha,\ \beta,\ t)$, is also constructed because it will be needed later on.

4.3. Constructing $h(y)$ from $f(j)$ and Vice Versa

Procedure `betatoYule` maps (α, β, t) into (c, b, r):

1. Consider a finite set of points $x_i \in [1, t]$. Let
$$F = \{(x_i, f(x_i))\}$$
be the corresponding set of sampled pairs of f, and let
$$F_{\text{rev}} = \{(f(x_i), x_i)\}$$
be the set of the reversed pairs, i.e., the set of the sampled pairs of g.

2. Find an approximation of the form (4.4) of the data F_{rev} by applying to the pairs $\{(\log f(x_i), \log x_i)\}$ a least-squares fit with respect to the basis $\{1, y, \log y\}$. Let
$$q(y) = d_1 + d_2\, y - d_3 \log y$$
be the resulting fit. Then $h(y) = \exp(q(y))$, i.e., $h(y)$ has the form (4.4) with
$$c = \exp(d_1), \quad b = \exp(d_2), \quad r = d_3.$$

Then, procedure `betatoYule` provides us with the functions $c = c(\alpha, \beta, t)$, $b = b(\alpha, \beta, t)$, and $r = r(\alpha, \beta, t)$.

Theorem 4.1. *The functions $c = c(\alpha, \beta, t)$, $b = b(\alpha, \beta, t)$, and $r = r(\alpha, \beta, t)$ are continuously differentiable with respect to $t > 0$ and $(\alpha, \beta) \in \mathcal{S}$.*

Proof. Let $\varphi_1(y) = 1$, $\varphi_2(y) = y$, and $\varphi_3(y) = \log y$ be the bases used in the fit, and let
$$\{(y_i, z_i) = (\log f(x_i), \log x_i)\}$$
be the set of points of the fit. Then, $\boldsymbol{d} = [d_1, d_2, d_3]^T$ is the solution of the system
$$A^T A \boldsymbol{d} = A^T \boldsymbol{z},$$
where $a_{i,k} = \varphi_k(y_i)$, $k = 1, 2, 3$. Since the functions $\varphi_k(y)$ are linearly independent, the matrix A has full rank, and the vector $\boldsymbol{d}$ is a continuously differentiable function of the elements of A. Moreover, $f(j)$ is continuously differentiable with respect to $(\alpha, \beta) \in \mathcal{S}$, hence also the element of A are continuously differentiable functions of α and β. It follows that $c = \exp(d_1)$, $b = \exp(d_2)$, and $r = d_3$ are continuously differentiable functions of α, β, and t. □

By direct inspection, it turns out that c is a monotone decreasing function of α and increasing with β and t, that b is a monotone increasing function of α and t and decreasing with β, and that r is a monotone decreasing function of α, β,

and t, with a weak dependence on α and t. Moreover, when α, β, and t increase, $r(\alpha, \beta, t)$ approaches

$$1/(\rho + 1) = 1 - \frac{1}{1 + (1 - \alpha)(1 - \beta)}. \tag{4.6}$$

We can now consider the procedure `Yuletobeta` that associates to $h(y)$ a function $f(j)$ such that $h(y)$ is the best approximation of the inverse of $f(j)$. More precisely, the procedure `Yuletobeta` receives as input the triple $(\widetilde{c}, \widetilde{b}, \widetilde{r})$, actual parameters of a Yule function $\widetilde{h}(y)$, and returns the triple $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{t})$, which solves the system

$$\begin{cases} c(\alpha, \beta, t) = \widetilde{c} \\ b(\alpha, \beta, t) = \widetilde{b} \\ r(\alpha, \beta, t) = \widetilde{r}. \end{cases} \tag{4.7}$$

The monotonicity of the functions $c(\alpha, \beta, t)$, $b(\alpha, \beta, t)$, and $r(\alpha, \beta, t)$ guarantees the uniqueness of the solution of (4.7), while the existence of a solution of (4.7) satisfying $(\widetilde{\alpha}, \widetilde{\beta}) \in \mathcal{S}$ is not guaranteed.

Problem (4.7) is solved iteratively. The initial approximation $(\alpha^{(0)}, \beta^{(0)}, t^{(0)})$ can be found by taking into account expressions (2.10) and (4.6), i.e.,

$$t^{(0)} = \sum_{j \in J} j\, W_j, \qquad \alpha^{(0)} = \frac{1}{t^{(0)}} \sum_{j \in J} W_j, \qquad \beta^{(0)} = 1 - \frac{\widetilde{r}}{(1 - \alpha^{(0)})(\widetilde{r} - 1)}.$$

We suggest the discrete Newton-Raphson method to solve (4.7). A sufficient condition for its local convergence is the continuously differentiability of the functions of (4.7), which has been proved in Theorem 4.1, provided that the initial approximation $(\alpha^{(0)}, \beta^{(0)}, t^{(0)})$ is sufficiently close to the solution $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{t})$.

Using procedure `betatoYule`, we are able to measure the effectiveness of the approximation of the inverse of $f(j)$ by $h(y)$ through the relative error

$$E = \max_{(\alpha,\beta) \in \mathcal{S}} \epsilon(\alpha, \beta, t), \quad \text{where} \quad \epsilon(\alpha, \beta, t) = \max_i \frac{|\log h\big(f(x_i)\big) - \log x_i|}{|\log x_i|}.$$

The computation, repeated for different values of t and (α, β) chosen uniformly in $\mathcal{S}$, has shown that E increases with t and with β. For example, when $t = 10^6$, we have found that $E \sim 7\%$ for $\beta = 0.1$ and $E \sim 18\%$ for $\beta = 0.8$. Then, we think that, due to the low level of precision required in the determination of the parameters, the approximation is sufficiently good.

By the way, the Yule function has already been used as a direct approximation of the steady state solution [Martindale 95, Martindale and Konopka 96], with a better result than by using the Zipf's function, but still unsatisfactory. This depends on the fact that the Yule function is a good approximation of the inverse of the beta function but not of the beta function itself.

4.4. Modeling the Reversed Data

The fact that the inverse of the beta function is well approximated by a Yule function of the form (4.4), suggests a possible alternative to the modeling of the data, i.e., the modeling of the reversed data. Let W_{rev} be the reverse of W, i.e., the set whose elements are the pairs (W_j, j), with $j \in J$. This approach requires solving the minimum problem

$$\min_{c,\,b,\,r} \sum_{j \in J} (\log h(W_j) - \log j)^2, \tag{4.8}$$

where the function $\log h$ depends linearly on the parameters, as shown by (4.5). Let $\widetilde{c}$, $\widetilde{b}$, and $\widetilde{r}$ be the values of the parameters corresponding to the solution

$$\widetilde{h}(y) = \widetilde{c}\,\widetilde{b}^{y}\, y^{-\widetilde{r}} \tag{4.9}$$

of (4.8). Thus, we propose the Yule function (4.9) for the description of the (reversed) degree distribution of the web. This would motivate the search of a model describing the reversed data and leading to a steady state solution of the form (4.4).

Besides being an accurate description of the reversed W, function (4.9) can be usefully employed for approximating the parameters of the model by means of the procedure `Yuletobeta`, instead of solving directly problem (3.2).

5. Experiments

5.1. Real Web Data

Real web data are collected by a crawler that visits the web starting from a predetermined list of URLs and downloads the URLs of the pages it encounters. The process is repeated recursively until a certain depth is reached. Due to the different policies implemented for the crawling and the limitation of the search, the data is inevitably contaminated by random noise. This results in an imprecise count of the degrees, more incorrect for the out-degrees, since the pages with no inlink cannot be reached and their outlinks cannot be counted.

The experiments were conducted on the three data sets WB, UK, and IT, freely available from the WebGraph homepage http://webgraph-data.dsi.unimi.it/. WB was obtained from the 2001 crawl performed by the WebBase crawler. The resulting graph has 118M pages and 1G links. The data provided by WebBase [Hirai et al. 00] was filtered to eliminate invalid links and to normalize URLs. UK and IT were obtained from a crawl performed by UbiCrawler on the `.uk` domain in 2002 and `.it` domain in 2004, respectively [Boldi et al. 02a, Boldi et al. 02b]. The UK graph contains 18.5M pages and 300M links. The IT graph

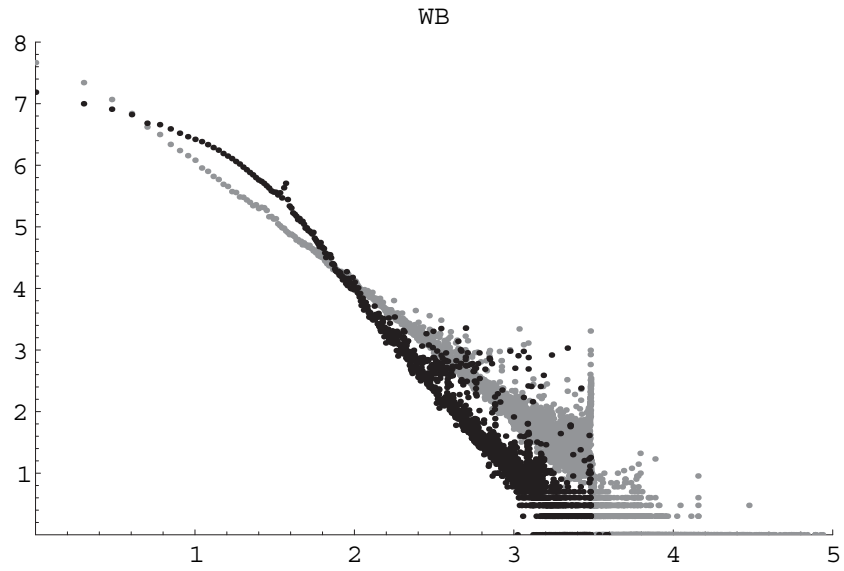

Figure 3. Link distributions for the WB graph (grey dots for the inlink distribution, black dots for the outlink distribution).

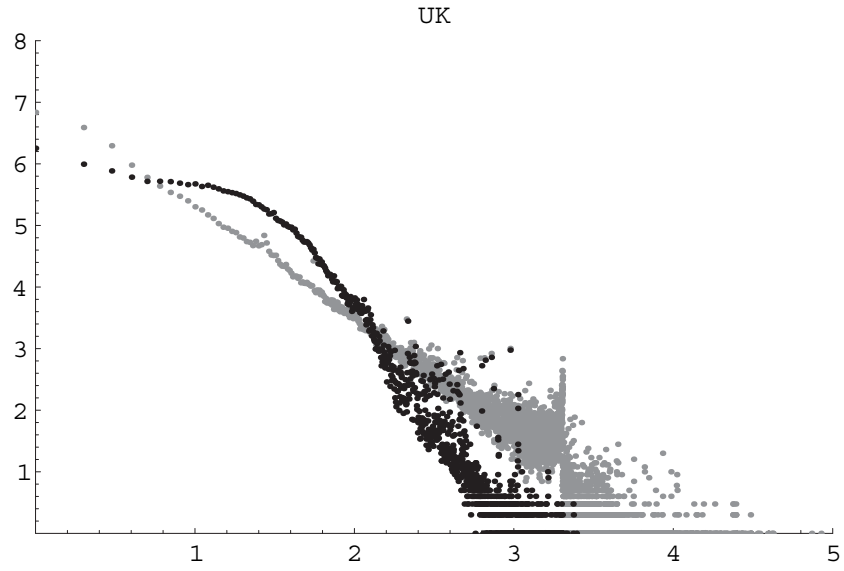

Figure 4. Link distributions for the UK graph (grey dots for the inlink distribution, black dots for the outlink distribution).

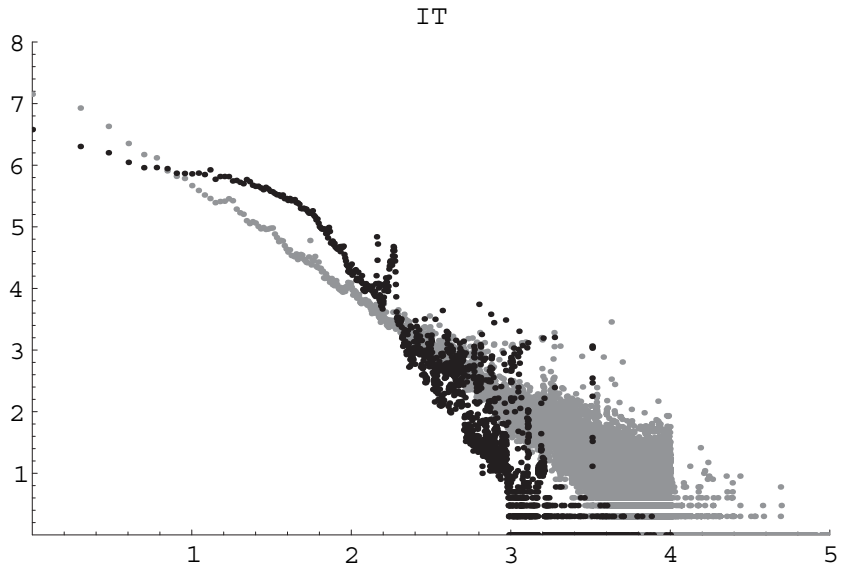

Figure 5. Link distributions for the IT graph (grey dots for the inlink distribution, black dots for the outlink distribution).

contains 41.3M pages and 1.15G links. This crawl was limited to 10,000 pages per host and maximum in-host depth 16.

In Figures 3, 4, and 5, the link distribution sets W are plotted for the WB, UK, and IT graphs, respectively, in log-log scale. Each point represents a pair $(j, W_j) \in W$.

It is evident that the shapes of inlinks and outlinks are different and that, while the power law can be a good approximation for the inlink distribution, this is not true for the outlink distribution. We are interested in verifying, by means of the experiments, that the proposed model holds for both inlinks and outlinks. More precisely, we are interested in estimating the parameters α, β, and t, characteristic of the model, that best account for the link distribution of the previous data sets, comparing our results with those given in the literature.

5.2. Integer Approximation

We cannot help noticing the presence in the plots of the characteristic noisy tails. Actually, the corresponding plots in the natural scale (not log-log scale) would show that the noise occurs everywhere, not only for large degrees (see, for example, Figure 6, where an initial section of the IT inlink distribution is shown in natural scale). However, as already mentioned in Section 3, we are interested in dealing with data in log-log scale.

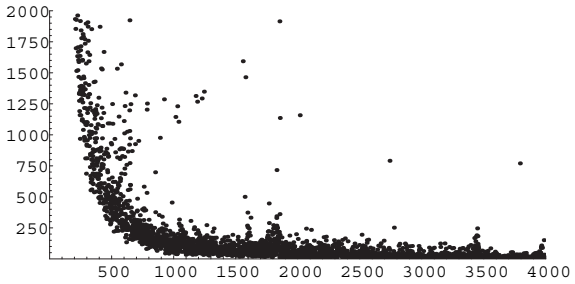

Figure 6. Initial section of the IT inlink distribution in natural scale.

The steady state solution $S_j^{(t)}$ of the model is a deterministic, continuous, and monotonic function, while the real web data consists 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. With this approach the parameters of the model could be determined by the least-squares fit of the experimental data with a beta function or of the reversed data with a Yule function. Unfortunately, in this way negative values of α and β, not compatible with the model, are found for the UK and IT inlink distributions. A cursory glance of the plots shows that the lower part of the graphs has a more compact and regular shape, while the upper part appears to be more spread (also the natural scale plots show a significant thickening towards the lower edge of the cloud). This observation suggests fitting the lower envelope of the data to approximate the lower edge of the cloud.

For a set of data W of the form (3.1), the lower envelope is defined as the subset $L(W)$ of all the pairs (j, W_j) such that no other pair (k, W_k) exists in W with both $k < j$ and $W_k \leq W_j$. Let L be the set of the first components of $L(W)$. In the low-degree region the lower envelope almost coincides with the original data, while in the high-degree region the lower envelope provides a lower bound to the tail of the original data and many points are discarded. In this way the head and the tail are equally weighted in the log-log scale.

The parameters of the model are obtained by either solving problem (3.2) or problem (4.8) and applying procedure `Yuletobeta` with j varying in the set L instead of the set J. Let us denote by s_j the corresponding steady state solution.

The experiments give compatible values of the parameters in all the cases. The use of the lower envelope leads us to conjecture that the integer values of the data are generated as the rounded sum of two terms:

- the values of the continuous approximation s_j,
- the realizations of an integer nonnegative random variable ξ_j with probability function $P_j(n)$ decreasing with n.

A suitable distribution with this property can be the geometric distribution whose probability function is $P_j(n) = p_j(1-p_j)^{n-1}$. The parameter p_j, which represents the probability that the success occurs at the first trial, will be linked to s_j. In the present context a success at the nth trial means an additive correction of $n-1$ to s_j. Since the expected value of the geometric random variable is $1/p_j$, the expected value of the additive correction is $1/p_j - 1$, and, on the basis of experimental observations, we hypothesize that $1/p_j - 1 = s_j^{\eta}$, where η is a suitable parameter.

5.3. The Results of the Experiments on Real Web Data

The experiments have been conducted on the six data sets W described in Section 5.1 according to the procedure described above, with the aim to

1. verify that the two techniques, (a) solving directly problem (3.2) or (b) solving problem (4.8) and applying procedure `Yuletobeta`, both give effective approximations of the data, producing comparable results;

2. validate the "additive conjecture" by tuning the parameter η that makes the tails of the data well reproduced.

The following tables and figures summarize the results of the experimentation for the six data sets. Table 1 lists the values of α^*, β^*, and t^* computed by solving directly (3.2) and the corresponding values $\rho^* + 1$ and σ^*. Table 2 lists the parameters $\widetilde{c}$, $\widetilde{r}$, and $\widetilde{b}$ computed by solving (4.8). Table 3 lists the values of $\widetilde{\alpha}$, $\widetilde{\beta}$, and $\widetilde{t}$ computed by solving (4.7) and the corresponding values $\widetilde{\rho}+1$ and $\widetilde{\sigma}$.

By comparing Tables 1 and 3, we note that α^* and β^* are slightly larger than the corresponding $\widetilde{\alpha}$ and $\widetilde{\beta}$. But, when we compare the resulting plots, we don't appreciate any difference. This observation is a further demonstration that the inverse of a beta function is a function of the Yule form and, hence, that fitting the reversed data by means of a Yule function is an adequate strategy.

It is interesting to compare the values of ρ^*+1 and $\widetilde{\rho}+1$ given in the tables with the exponents of the Zipf's functions mentioned in the literature, namely 2.1 for the inlinks and 2.7 for the outlinks. In the case of the inlinks, the experiments led to reconstructed beta functions sufficiently close to the Zipf's function commonly reported in the literature. For the outlinks the difference between our values and the exponent given in the literature is much greater. This difference is mainly due to the choice of weighting in the same way the head and the tail in the log-log scale. In order to test what would happen with a different choice of the weights, we have performed also the following experiment: we have applied our model to the data grouped into buckets of exponentially increasing sizes, as suggested in [Pennock et al. 02], and we have obtained smaller values of ρ, closer to those given in the literature. The corresponding fit curves lie inside the fat tail but do not bend enough in the low-degree region, not giving a good approximation

	t^*	α^*	β^*	ρ^*+1	σ^*
WB inlink	9.9×10^8	0.12	0.097	2.3	0.91
UK inlink	2.9×10^8	0.066	0.060	2.1	0.97
IT inlink	6.8×10^8	0.065	0.089	2.2	1.5
WB outlink	9.1×10^8	0.11	0.58	3.7	13
UK outlink	2.8×10^8	0.071	0.77	5.7	47
IT outlink	9.7×10^8	0.054	0.71	4.7	46

Table 1. Values t^*, α^*, and β^* of the parameters t, α, and β, respectively, for the data sets WB, UK, and IT and the corresponding values ρ^*+1 and σ^*.

	$\widetilde{c}$	$1-\widetilde{b}$	$\widetilde{r}$
WB inlink	6496	2.0×10^{-8}	0.45
UK inlink	4108	1.2×10^{-7}	0.48
IT inlink	6258	7.7×10^{-8}	0.47
WB outlink	1560	1.7×10^{-7}	0.30
UK outlink	654	2.2×10^{-6}	0.23
IT outlink	1414	1.2×10^{-6}	0.28

Table 2. Values $\widetilde{c}$, $\widetilde{r}$, and $\widetilde{b}$ of the parameters c, r, and b, respectively, for the data sets WB, UK, and IT.

	$\widetilde{t}$	$\widetilde{\alpha}$	$\widetilde{\beta}$	$\widetilde{\rho}+1$	$\widetilde{\sigma}$
WB inlink	$10.\times10^8$	0.11	0.096	2.2	0.93
UK inlink	3.0×10^8	0.061	0.057	2.1	0.98
IT inlink	7.1×10^8	0.060	0.082	2.2	1.5
WB outlink	8.2×10^8	0.10	0.57	3.6	13
UK outlink	2.2×10^8	0.066	0.76	5.5	48
IT outlink	6.8×10^8	0.049	0.68	4.3	44

Table 3. Values $\widetilde{t}$, $\widetilde{\alpha}$, and $\widetilde{\beta}$ of the parameters t, α, and β, respectively, for the data sets WB, UK, and IT and the corresponding values $\widetilde{\rho}+1$ and $\widetilde{\sigma}$.

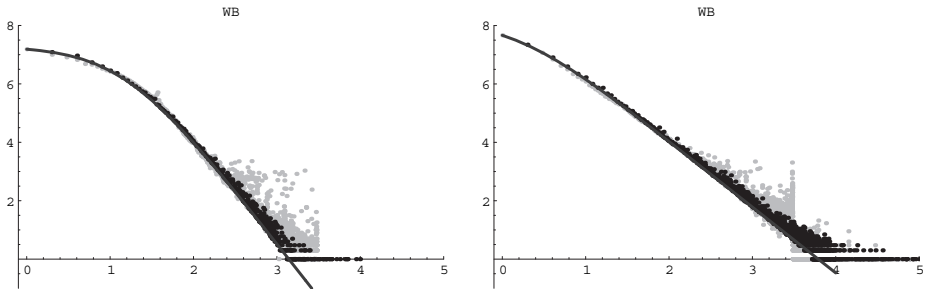

Figure 7. The WB data (grey points), the continuous approximation s_j (black line) fitted on the lower envelope, and the integer approximation (black points). Outlink distribution on the left; inlink distribution on the right.

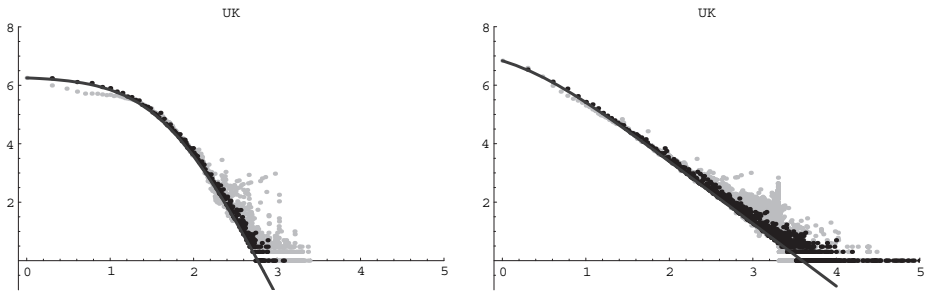

Figure 8. The UK data (grey points), the continuous approximation s_j (black line) fitted on the lower envelope, and the integer approximation (black points). Outlink distribution on the left; inlink distribution on the right.

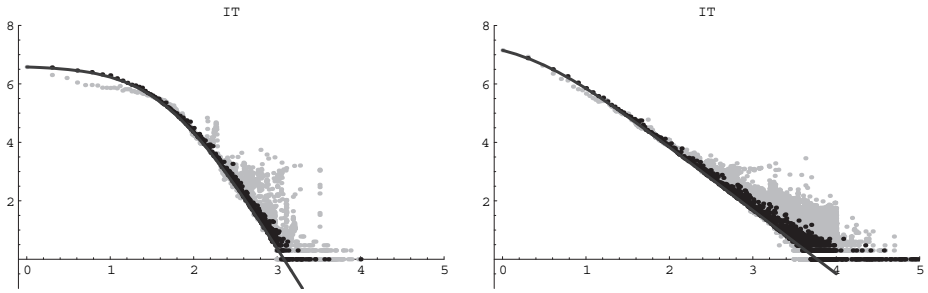

Figure 9. The IT data (grey points), the continuous approximation s_j (black line) fitted on the lower envelope, and the integer approximation (black points). Outlink distribution on the left; inlink distribution on the right.

of the head (in fact, for the beta function, the closer to 2 the value of $\rho + 1$, the more rectilinear the curve).

It is worth noting that the values of β are much smaller for the inlinks than for the outlinks. This means that the preferential attachment is the dominant policy in the inlink distribution, while the outlink distribution appears to be significantly ruled by the uniform attachment.

We give now a set of figures (Figures 7–9) that validates the additive conjecture. The parameter η has been experimentally tuned to the value 0.8.

5.4. Wikipedia Data

Wikipedia is an online and free content encyclopedia, structured as a graph. Since in many aspects the web graph and the Wikigraph are very similar [Buriol et al. 06], it could be interesting to analyze if such a similarity extends also to the connectivity distributions of the two graphs. For this reason, we have tested our model on data obtained by downloading the complete collection [Wikimedia 07] (2.2M pages) of English Wikipedia articles, updated to November 2006. The graph structure has been extracted by using the programs of [Senellart 07].

The parameters obtained for this experiment are listed in Table 4, and the resulting approximation is shown in Figure 10.

	t^*	α^*	β^*	$\rho^* + 1$	σ^*
Wiki inlink	2.5×10^7	0.063	0.32	2.6	7.6
Wiki outlink	2.9×10^7	0.037	0.72	4.7	69

Table 4. Values t^*, α^*, and β^* of the parameters t, α, and β, respectively, for the Wikipedia data sets and the corresponding values $\rho^* + 1$ and σ^*.

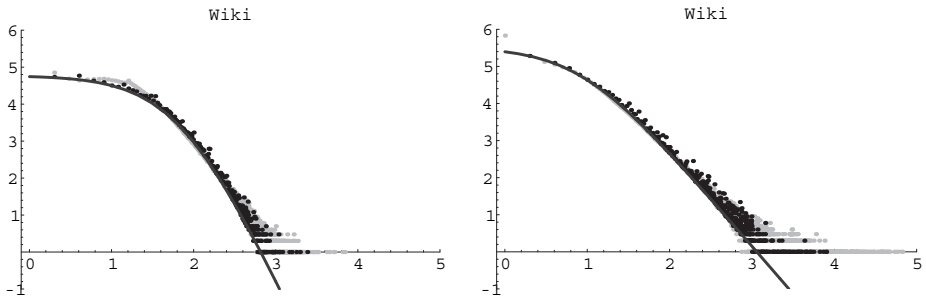

Figure 10. The Wikipedia data (grey points), the continuous approximation s_j (black line) fitted on the lower envelope, and the integer approximation (black points). Outlink distribution on the left; inlink distribution on the right.

For the inlink distribution, the value of β obtained for the Wikipedia data set is much larger than the corresponding values obtained for the web data sets (listed in the first three rows of Table 1), showing that the preferential attachment policy is less dominant for Wikipedia than for the web subsets. This means that the choice of pointing to a selected article depends not only on the importance of the article, but also on its contents, which are of a typical random nature.

Instead, the outlink distributions of the considered Wikipedia and web subsets appear to be very similar, the more evident difference being the value of α^*. The lower value found for Wikipedia indicates a denser graph, that is, a greater average number of outlinks per article.

6. Conclusions

In this paper a unique model for both the inlink and the outlink distributions has been considered. This model, thanks to an accurate approximation of its steady state solution, seems to be well suited to describe the two connectivity distributions, as it appears evident from the figures. In particular, for the outlink distribution, a unitary model that applies to the whole connectivity region appears to be preferable to models based on the superposition of two different probability distributions, acting on two different connectivity regions [Broder et al. 00].

Two procedures have been proposed to compute the parameters α, β, and t of the model for real subsets of data. They turn out to be equivalent from the point of view of the accuracy of the results, and the second one, which acts on the reverse data, is faster. The values of the parameters computed for the considered data sets are reasonable.

In particular, for the web subsets, the parameters found for the inlink distributions are coherent with the measures given in the literature, confirming that the power law gives an acceptable description of the distribution (in fact, in this case the log-log plot of the lower envelope of the data is nearly rectilinear). The parameters found in the outlink cases give values of the exponent $\rho + 1$ different from the one generally accepted when the power law is considered for the fit. This is not surprising; in fact, an "additional degree of freedom is also sufficient to explain the often large deviation from power law behavior observed in the low connectivity region" [Pennock et al. 02], and a reasonable fit on the whole connectivity region cannot have the same exponent of such a power law. An analysis of the computed values of β shows that the preferential attachment is the dominant policy in the inlink distribution, while the outlink distribution appears to be significantly ruled by the uniform attachment.

On the contrary, the analysis of the Wikipedia set shows that the power law does not give an acceptable description even in the inlink case.

Possible developments of this research could be the following:

1. Find a simple model describing the reversed data and leading to a steady state solution of Yule form. In this way the computation of the model parameters would require only a linear fit.
2. Examine discrete probability distributions for a better description of the fluctuations of the experimental data, different from the geometric one considered here.
3. Study a unitary model describing the inlink and outlink distributions by taking into account the interconnections of the two processes. Such an integrated model would give more accurate information on the strategies that rule the evolution of the links generation.

Acknowledgments. The authors wish to thank Dr. Marco Righi for his help in downloading and computing the Wikipedia data and the referees for their useful comments.

References

[Adamic 02] L. A. Adamic. "Zipf, Power-Laws, and Pareto—A Ranking Tutorial." Available at http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html, 2002.

[Albert et al. 99] R. Albert, H. Jeong, and A. L. Barabasi. "Diameter of the World-Wide Web." *Nature* 401 (1999), 130.

[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 et al. 02a] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. "UbiCrawler: A Scalable Fully Distributed Web Crawler." In *Proceedings of the Eighth Australian World Wide Web Conference*. Available at http://ausweb.scu.edu.au/aw02/papers/refereed/vigna/paper.html, 2002.

[Boldi et al. 02b] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. "UbiCrawler: Scalability and Fault-Tolerance Issues." In *Poster Proceedings of Eleventh International World Wide Web Conference*. Available at http://www2002.org/CDROM/poster/162/index.html, 2002.

[Brin and Page 98] S. Brin and L. Page. "The anatomy of a large-scale hypertextual web search engine." *Proceedings of WWW7, Computer Networks* 30:1–7 (1998), 107–117.

[Broder et al. 00] A. Broder, R. Kumar, F. Maghoul, P. Prabhakar, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. "Graph Structure in the Web." In *Proceedings of the 9th International World Wide Web Conference*. Available at http://www9.org/w9cdrom/160/160.html, 2000.

[Buriol et al. 06] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. "Temporal Analysis of the Wikigraph." In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 45–51. Los Alamitos, CA: IEEE Press, 2006.

[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, Universita' di Roma "La Sapienza," 2003.

[Cooper and Frieze 01] C. Cooper and A. M. Frieze. "A General Model of Undirected Web Graphs." In *Algorithms—ESA 2001: 9th Annual European Symposium, Aarhus, Denmark, August 28-31, 2001, Proceedings*, Lecture Notes in Computer Science 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.

[Hirai et al. 00] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. "WebBase: A Repository of Web Pages." In *Proceedings of the Nineth International World Wide Web Conference*. Available at http://www9.org/w9cdrom/296/296.html, 2000.

[Kumar et al. 99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "Trawling the Web for Emerging Cyber Communities." *Proceedings of WWW8, Computer Networks* 31:11–16 (1999), 1481–1493.

[Martindale 95] C. Martindale. "Fame More Fickle than Fortune: On the Distribution of Literary Eminence." *Poetics* 23 (1995), 219–234.

[Martindale and Konopka 96] C. Martindale and A. K. Konopka. "Oligonucleotide Frequencies in DNA Follow a Yule Distribution." *Computers Chem.* 20 (1996), 35–38.

[Mitzenmacher 03] M. Mitzenmacher. "A Brief History of Generative Models for Power Law and Lognormal Distributions." *Internet Mathematics* 1:2 (2003), 226–251.

[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.

[Zipf 49] G. K. Zipf. *Human Behavior and the Principle of Least Effort.* New York: Hafner, 1949.

[Senellart 07] Pierre Senellart. "Wikipedia Related Stuff." Available at http://pierre.senellart.com/software/wikipedia/index.en, 2007.

[Wikimedia 07] "Wikimedia Downloads." Available at http://download.wikimedia.org/, 2007.

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

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

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

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

Received March 1, 2007; accepted July 5, 2007.