A Stochastic Model for the Evolution of the Web

Mark Levene, Trevor Fenner, George Loizou and Richard Wheeldon
Birkbeck College, University of London
London WC1E 7HX, U.K.
{mark,trevor,george,richard}@dcs.bbk.ac.uk

Abstract

Recently several authors have proposed stochastic models of the growth of the Web graph that give rise to power-law distributions. These models are based on the notion of preferential attachment leading to the "rich get richer" phenomenon. However, these models fail to explain several distributions arising from empirical results, due to the fact that the exponent they predict is not consistent with the data. To address this problem we extend the evolutionary model of the Web graph by including a non-preferential component and viewing the stochastic process in terms of an urn transfer model. By making this extension we can now explain a wider variety of empirically discovered power-law distributions provided the exponent is greater than two. These include: the distribution of incoming links, the distribution of outgoing links, the distribution of pages in a Web site and the distribution of visitors to a Web site.

1 Introduction

A power-law distribution is a function of the form

$$f(i) = C\ i^{-\tau},$$

where C and τ are positive constants. Power-law distributions are *scale-free* in the sense that if i is rescaled by multiplying it by a constant, then $f(i)$ would still be proportional to $i^{-\tau}$.

Power-law distributions are abundant, for example Zipf's law [Rap82], which states that relative frequency of words in a text is inversely proportional to their rank, and Lotka's law [Nic89], which is an inverse square law stating that the number of authors making n contributions is proportional to n^{-2}. (We refer the reader to [Sch91] for more examples of power-law distributions.)

Recently several researchers have detected power-law distributions in the Internet [FFF99] and World-Wide-Web [BKM$^+$00, DKM$^+$01] topologies. In order to understand how these power-law distributions emerge and how the Web has evolved and is evolving, several researchers have recently been studying stochastic models of graphs which give rise to such distributions. One particular power-law phenomenon that has attracted attention is the distribution of incoming links to a Web page. This distribution is important, since a link from Web page P to Web page Q can be viewed as a recommendation of page Q; thus Web pages having more incoming links are more highly recommended and therefore potentially of higher quality. This observation is the basis of Google's PageRank algorithm [Hen01].

Albert et al. [ABJ00] studied a stochastic model of growth and *preferential attachment*, where new links to existing Web pages are added in proportion to the number of incoming links these Web pages already have. Their theoretical model predicts an exponent $\tau = 3$, which is not in agreement with the value of approximately 2.1 obtained from the study reported in [BKM$^+$00]. Dorogovtsev et al. [DMS00a] generalise Albert et al.'s model and predict an exponent greater than two. More precisely, they obtain the value $2 + A/m$ for the exponent, where A is the initial attractiveness of a newly created Web page and m is the number of new links added to the Web graph at each step of the stochastic process. This exponent value is consistent with the empirical value of the exponent of the distribution of incoming links provided A/m is sufficiently small. Bornholdt and Ebel [BE00] pointed out that the stochastic process proposed by Simon [Sim55] in 1955 can also offer an explanation of the power-law distribution. (We note that during the period of 1959-1961 there was a fierce debate between Mandelbrot and Simon in *Information and Control* on the validity of Simon's model [Man59].) In reply to Bornholdt and Ebel, Dorogovtsev et al. [DMS00b] note that the model they describe in [DMS00a] essentially coincides with Simon's model.

The models discussed above are based on the process of preferential attachment and do not take into account the fact that links may also be added or removed randomly through a *non-preferential process*. By this we mean that the probability of adding or removing a link to a particular Web page may be influenced by factors other than the popularity of that Web page, where popularity is measured by the number of incoming links. Our main contribution in this paper is to extend Simon's model [Sim55] with a non-preferential component and view the stochastic process in terms of an urn transfer model [JK77]. (We note that at the end of Section 3 of his seminal paper Simon suggested adopting a mixture of preferential and non-preferential components but did not develop the idea.) By making this extension we can explain a wider variety of empirically discovered power-law distributions than can be explained with Simon's original model. These include: the distribution of incoming links, the distribution of outgoing links, the distribution of pages in a Web site and the distribution of visitors to a Web site.

The rest of the paper is organised as follows. In Section 2 we present an urn transfer model that generalises Simon's original model. In Section 3 we demonstrate how this can provide a stochastic model for the evolution of the Web that is consistent with a wide range of empirical data. Finally, in Section 4 we give our concluding remarks. The proofs of some of the mathematical results are given in the Appendix. As far as we are aware our convergence proof given in the Appendix is the first formal proof validating Simon's model − it does not rely on the mean-field theory approach, as for example in [BAJ99].

2 An Urn Transfer Model

We now present an *urn transfer model* [JK77] for a stochastic process that we will use in Section 3 to analyse the evolution of the Web graph. Our model is an extension of Simon's stochastic process [Sim55], which was originally described in terms of the underlying process leading to the distribution of words in a piece of text. Simon's stochastic process is essentially a birth process, where there is a constant probability p that the next word is a new word and, given that the next word has already appeared, its probability of occurrence is proportional to the previous number of occurrences of that word. We extend Simon's model by setting the probability of occurrence of a word, given that it has already appeared, to be a weighted average of the preferential probability, as described above, and the uniform probability if all words were equiprobable. As we noted in the introduction, this extension was already proposed by Simon at the end of Section 3 of his paper. Simon set out this extension in equation (3.7) and presented a tentative solution in equation (3.8). As we will see in Section 3, this extension of Simon's model makes sense in the context of the Web, since, for example, links to a Web site are often added or removed in a random fashion without taking into consideration the "importance" of the site in terms of how many links it already has. We now describe our urn model in detail.

We assume a countable number of urns, urn_i, $i = 1, 2, 3, \ldots$, where each ball in urn_i has i pins attached to it. Initially, at stage $k = 1$ of the stochastic process, all the urns are empty except urn_1 which has one ball in it. Let $F_i(k)$ be the number of balls in urn_i after k steps of the stochastic process, so $F_1(1) = 1$, and let p and α be parameters, with $0 < p < 1$ and $\alpha > -1$. Then, at stage $k+1$ of the stochastic process for $k \geq 1$, one of two things may occur:

(i) with probability p_{k+1}, where

$$p_{k+1} = 1 - \frac{(1-p)\sum_{i=1}^{k}(i+\alpha)F_i(k)}{k(1+\alpha p) + \alpha(1-p)}, \tag{1}$$

a new ball is added to urn_1 (provided that $0 \leq p_{k+1} \leq 1$) or,

(ii) with probability $1 - p_{k+1}$, an urn is selected $-$ urn_i being chosen with probability

$$\frac{(1-p)(i+\alpha)F_i(k)}{k(1+\alpha p) + \alpha(1-p)}, \tag{2}$$

for $1 \leq i \leq k$; then one ball from urn_i is transferred to urn_{i+1}. This is equivalent to attaching an additional pin to the ball chosen from urn_i and moving it to its "correct" urn. (We note that the denominator appearing in (1) and (2) has been chosen so that the expected value of p_{k+1} is p; see (8) below.)

At each stage we either add a new ball with one pin or add a pin to an existing ball and move the ball to the next urn up, so at stage k the total number of pins is k, i.e.

$$\sum_{i=1}^{k} iF_i(k) = k.$$

It is obvious that $F_i(k) = 0$ for any $i > k$. We call the above model the p_k*-model*.

Let $B(k) = \sum_i F_i(k)$, the total number of balls in all the urns. We can now simplify (1) to

$$p_{k+1} = 1 - \frac{(1-p)(k + \alpha B(k))}{k(1+\alpha p) + \alpha(1-p)}. \tag{3}$$

Since it is clear from (3) that $p_{k+1} < 1$, in order for p_{k+1} to be well defined, we must have $p_{k+1} \geq 0$ for $k \geq 1$, i.e.

$$(1-p)(k + \alpha B(k)) \leq k(1+\alpha p) + \alpha(1-p). \tag{4}$$

In the Appendix we show that p_{k+1} is always well defined (i.e. non-negative) for all k when $p \geq 1/2$, but only if

$$\alpha \leq \frac{p}{1-2p} \tag{5}$$

when $p < 1/2$. In the discussion at the end of Section 3 we suggest that, in practice, starting from a typical initial configuration of balls in the urns, it is likely that p_{k+1} will be well defined for all k, even if (5) does not hold.

We next make a small digression to explain our use of (2) rather than the more natural definition of the probability as

$$\frac{(1-p)(i+\alpha)F_i(k)}{\sum_{i=1}^{k}(i+\alpha)F_i(k)} = \frac{(1-p)(i+\alpha)F_i(k)}{k+\alpha B(k)}. \tag{6}$$

In order to find a solution for the expected value of $F_i(k)$, we would need to take the expected value of (6); this is problematic since the random variables $B(k)$ and $F_i(k)$ are *not* independent and it is therefore not clear how to calculate the expectation of the right-hand expression in (6). We observe that this problem does not arise in Simon's original model [Sim55], since in this case we have $\alpha = 0$ and the denominator reduces to the constant k in both (2) and (6). In our case, when α is not necessarily zero, by using (2) instead of the more natural (6), there is no problem in computing the expectation of $F_i(k)$, since the parameter p is a constant, allowing us to find the expected value of $F_i(k)$ by using the linearity of expectations.

In the Appendix we prove the following results for the expectations of $B(k)$ and p_k for $k > 1$, namely

$$E(B(k)) = E\Big(\sum_{i=1}^{k} F_i(k)\Big) = 1 + (k-1)p \tag{7}$$

and

$$E(p_k) = p. \tag{8}$$

(We note that $E(B(1)) = B(1) = 1$.)

Thus, in terms of expectations (i.e. using a mean-field theory approach), it is possible to describe the urn transfer model as a "more natural" stochastic process, where at each stage k, for $k > 1$, either

(i) a new ball is inserted into urn_1 with probability p, or

(ii) with probability $1-p$ an urn is chosen, the probability of choosing urn_i being proportional to $(i+\alpha)F_i(k)$, and then one ball from urn_i is transferred to urn_{i+1}.

We stress that, since this model uses the expectations of the random variables p_{k+1} rather than the random variables themselves, it is only an approximation of our urn transfer model. This model, which we call the *p-model*, is in fact the "more natural" model discussed above that uses (6) instead of (2).

We note that we could modify the initial condition of our stochastic process so that, for example, urn_1 would initially contain $\delta > 1$ balls instead of one, or more generally that a finite number of urns would initially be non-empty with some prescribed number of balls in each. As can be seen from the development of the model below, as k tends to infinity, such a change in the initial conditions will not have an effect on the asymptotic distribution of the balls in the urns.

We call the transfer of a ball as a result of (ii) above a mixture of *preferential* and *non-preferential* transfer. When $\alpha = 0$, then the transfer is purely preferential otherwise non-preferential transfer takes a part in the process.

Following Simon [Sim55], we now state the equations for the p_k-model. For $i > 1$ (including $i > k$), we have:

$$E_k(F_i(k+1)) = F_i(k) + \beta_k \Big((i-1+\alpha) F_{i-1}(k) - (i+\alpha) F_i(k) \Big), \tag{9}$$

where $E_k(F_i(k+1))$ is the expected value of $F_i(k+1)$ given the state of the model at stage k, and

$$\beta_k = \frac{1-p}{k(1+\alpha p) + \alpha(1-p)},$$

the normalising constant used in (2).

Equation (9) gives the expected number of balls in urn_i as the previous number of balls in that urn plus the difference between the probability of increasing the number of balls in urn_i, which is equal to the probability of choosing urn_{i-1} in step (ii) of our urn transfer model, and the probability of decreasing the number of balls in urn_i, which is equal to the probability of choosing urn_i.

In the boundary case, $i = 1$, we have

$$E_k(F_1(k+1)) = F_1(k) + p_{k+1} - \beta_k (1+\alpha) F_1(k), \tag{10}$$

which describes the expected number of balls in urn_1, which is the previous number of balls in the first urn plus the difference between the probability of inserting a new ball in urn_1 and the probability of transferring a ball from urn_1 to urn_2.

Now letting

$$\beta = \frac{1-p}{1+\alpha p},$$

we see that $k\beta_k \approx \beta$ for large k. In fact, for $k \geq 1$,

$$\beta - k\beta_k = \alpha \beta \beta_k. \tag{11}$$

Using the facts that $0 < p < 1$ and $\alpha > -1$, it is also easy to see that

$$0 < \beta < 1, \tag{12}$$

and, for $k \geq 1$,

$$0 < \beta_k < \frac{1}{k+\alpha}. \tag{13}$$

Since the right-hand sides of (9) and (10) are linear in the random variables, using (8), we may take expectations to obtain

$$E(F_i(k+1)) = E(F_i(k)) + \beta_k\Big((i-1+\alpha)E(F_{i-1}(k)) - (i+\alpha)E(F_i(k))\Big) \tag{14}$$

for $i > 1$, and

$$E(F_1(k+1)) = E(F_1(k)) + p - \beta_k(1+\alpha)E(F_1(k)). \tag{15}$$

In order to solve equations (14) and (15) we show that $E(F_i(k))/k$ tends to a limit f_i, as k tends to infinity. Assume for the moment that this is the case, then, letting k tend to infinity, $E(F_i(k+1)) - E(F_i(k))$ tends to f_i and $\beta_k E(F_i(k))$ tends to βf_i; so (14) and (15) yield

$$f_i = \beta\Big((i-1+\alpha)f_{i-1} - (i+\alpha)f_i\Big) \tag{16}$$

for $i > 1$, and

$$f_1 = p - \beta(1+\alpha)f_1. \tag{17}$$

Now let us *define* f_i, $i \geq 1$, by the recurrence relation (16) with boundary condition (17). We may then let

$$E(F_i(k)) = k(f_i + \epsilon_{i,k}), \tag{18}$$

and in the Appendix we prove that $\epsilon_{i,k}$ tends to zero as k tends to infinity. This justifies our assumption that $E(F_i(k))/k$ tends to f_i as k tends to infinity. We therefore see that f_i is the asymptotic expected rate of increase of the number of balls in urn_i.

From (16) and (17) we obtain

$$f_i = \frac{\beta(i-1+\alpha)}{1+\beta(i+\alpha)}\; f_{i-1} \tag{19}$$

and

$$f_1 = \frac{p}{1+\beta(1+\alpha)}, \tag{20}$$

respectively.

Now, on repeatedly using (19), we get

$$\begin{aligned} f_i &= \frac{\rho\; p\; (1+\alpha)\; (2+\alpha)\; \cdots\; (i-1+\alpha)}{(1+\rho+\alpha)\; (2+\rho+\alpha)\; \cdots\; (i+\rho+\alpha)} \\ &= \frac{\rho\; p\; \Gamma(i+\alpha)\; \Gamma(1+\rho+\alpha)}{\Gamma(1+\alpha)\; \Gamma(i+1+\rho+\alpha)}, \end{aligned} \tag{21}$$

where $\rho = 1/\beta$ and Γ is the gamma function [AS72, 6.1].

It follows that for large i, on using Stirling's approximation [AS72, 6.1.39], we have

$$f_i \sim C\; i^{-(1+\rho)}, \tag{22}$$

where C is independent of i and $\sim$ means *is asymptotic to*. Thus we have derived in (22) a general power-law distribution for f_i, with exponent $1 + \rho$. An obvious consequence of (19) is that $f_i > f_{i+1}$, i.e. asymptotically there are more balls in urn_i than in urn_{i+1}.

It follows from (6), (7) and (8) that (16) and (17) will also hold for the asymptotic distribution for the p-model obtained using the mean-field theory approach. Hence, on the assumption that this approach is valid, the asymptotic distribution will be the same as for the p_k-model, as given by (21) and (22).

When $\alpha = 0$ then the extended model reduces to Simon's original model and by increasing α the exponent will increase accordingly. In any case the exponent is always greater than 2, so the expected number of pins per ball is finite. The constraint that $\rho > 1$ is equivalent to the condition that $\alpha > -1$. Another way to understand this constraint is that if $\alpha \leq -1$ then the first urn will never be chosen in case (ii) of the stochastic process, and thus no ball will ever be transferred out of urn_1. When ρ is close to 1 we obtain Lotka's law [Nic89], which is an inverse square power-law; see also Price's cumulative advantage distribution leading to Lotka's law [Pri76, KH95].

In many real situations, such as the Web, p is generally small. For example, if we interpret balls as Web pages and the number of pins attached to a ball as the number of links incoming to that Web page, then we expect the ratio of pages to links to be quite small, say 0.1, and thus the exponent of the power-law to be just over two. If the value of p and the power-law exponent are obtained from empirical evidence, we may find a discrepancy from Simon's original model, i.e. when $\alpha = 0$. Our current extension can explain this discrepancy through the non-preferential component as long as the exponent is greater than two.

3 A Model for the Evolution of the Web

We now describe a discrete stochastic process by which the Web graph could evolve. At each time step the state of the Web graph is a directed graph $G = (N, E)$, where N is its node set and E is its link set. In this case $F_i(k)$, $i \geq 1$, is the number of nodes in the Web graph having i incoming links; $F_i(k)$ induces an equivalence class of nodes in N all having i incoming links. We note that although we have chosen i to denote the number of incoming links, i could alternatively denote the number of outgoing links, the number of pages in a Web site or any other reasonable parameter.

Consider the evolution process of the Web graph with respect to the number of nodes having i incoming links at the kth step of the process. Initially G contains just a single node. At each step one of two things can happen. With probability p a new node is added to G having one incoming link. In the p-model, this is equivalent to placing a new ball in urn_1. With probability $1 - p$ a node is chosen with probability proportional to $(i + \alpha)$, and then an additional incoming link is added to this node. In the p-model this is equivalent to a mixture of preferential and non-preferential transfer of a ball from urn_i to urn_{i+1}; the mixture level depends on the value of the parameter α.

We now look at some of the measurements of the Web graph which were reported recently. Broder et al. [BKM$^+$00] reported a power-law distribution with exponent 2.09 for the number of incoming links (referred to as *inlinks*) to a node. This value was derived from a 203 million node crawl of the Web graph. The average number of inlinks per Web page was measured at about 8 [KRRT99], which gives us a value of 0.125 for p. We can compute α by

$$\alpha = \frac{\rho(1-p) - 1}{p}.$$

Thus a more accurate model of the stochastic process generating the distribution of incoming links would assume $\alpha \approx -0.37$ rather than $\alpha = 0$. (It would not be unreasonable in this case to assume Simon's model, i.e. $\alpha = 0$, which would give an exponent of 2.14, since the small difference in the exponents may be due to statistical error.)

Looking at the outgoing links (referred to as *outlinks*) from a node, Broder et al. [BKM$^+$00] reported a power-law distribution with exponent 2.72. Moreover, the average number of outlinks per Web page was measured at about 7.2 [KRR$^+$00], which gives a value of 0.14 for p. Thus to get an exponent of 2.72 we would have to assume $\alpha \approx 3.42$. However, Simon's original model would predict an exponent of about 2.16 for outlinks, similar to that for inlinks.

Another interpretation of i is the number of pages within Web sites (referred to as *webpages*). In this case, Huberman and Adamic [HA99] reported a power-law distribution with exponent 1.85, derived from a 250,000 Web site crawl. Our model cannot explain this observation as the exponent is less than two. A more recent result from a private communication with Adamic reported an exponent of 2.2, derived from a 1.6 million Web site crawl; the difference is possibly due to a different crawling strategy. To calculate p we can estimate the size of the Web to be 2.1 billion pages [MM00] distributed over approximately 113.5 million Web sites (this number, which was reported on www.netsizer.com during the first quarter of 2001, refers to the number of Internet hosts, so it is an over-estimate of the number of Web sites). Thus we can derive a value 0.054 for p; in reality p will be even closer to zero. To get an exponent of 2.2 we would have to assume $\alpha \approx 2.50$. This gives a more accurate description than we would obtain from Simon's original model, which would predict an exponent of 2.06.

As a final interpretation, let i be the number of users visiting a Web site during the course of a day (referred to as *visitors*). In this case, Adamic and Huberman [AH00] reported a power-law distribution with exponent 2.07, derived from access logs of 60,000 AOL users accessing 120,000 Web sites. Now, from www.netsizer.com we obtain the statistic that in the USA there were 72.7 million Internet hosts and 166.6 million users at the beginning of 2001. Moreover, from www.netvalue.com we obtain the statistic that, on average, users visit about 1.93 different Web sites per day. So, we derive the value $72.7/(166.6 * 1.93) \approx 0.226$ for p, on the assumption that each Web site gets visited at least once per day. Thus to get an exponent of 2.07 we would have to assume $\alpha \approx -0.76$. However, Simon's original model would predict an exponent of about 2.29.

In order to validate our model, we programmed a simulation of the stochastic model using the parameter values we have derived for p and α and compared the exponent values obtained with the reported empirical values. (Our simulation is in the spirit of Simon and Van Wormer's Monte Carlo simulation, whose intention was to test how good the estimates of the original model are [SV63].) We repeated the simulation five times using the p_k-model, and five times using the p-model. Each simulation was carried out for 200,000 iterations, and for the purpose of regression we considered only the first 25 urns. The results of our simulations are presented in Table 1; in all cases the correlation coefficient of the regression analysis was close to one. The discrepancy between the simulated values and the empirical values can be attributed in part to the fact that (22) is only an asymptotic approximation to (21). It is also possible that running the simulations for a much larger number of iterations would give more accurate results.

Interpretation	Empirical	p_k-model	p-model
inlinks	2.09	2.096	2.094
outlinks	2.72	2.714	2.675
webpages	2.2	2.122	2.208
visitors	2.07	2.131	2.179

Table 1: Power law exponents of simulation results

For outlinks and webpages we restarted the p_k-model simulation whenever the computed value of p_{k+1} was ill-defined, i.e. negative; only a moderate number of restarts were necessary. From (3) it can be shown that, for $k > 1$, $|\,p_{k+1} - p_k\,|$ is of the order of $1/k$. This indicates that for large k it is very unlikely that p_{k+1} will be ill-defined, given that p_j is well defined for $j \leq k$. In practice, if instead of starting with just one ball in urn_1, we start from a typical initial configuration with a modest number of balls in the urns, it is likely that p_{k+1} will be well defined for all k.

Batch	Overall	$k \leq 10$	Average k	Max k
1	66	89%	3.78	21
2	63	90%	3.86	26
3	63	90%	3.34	13
4	60	88%	3.68	30
5	63	90%	3.73	22
6	65	92%	3.49	22
7	61	92%	3.50	17
8	64	94%	3.54	22
9	64	86%	4.19	34
10	59	92%	3.49	21
Average	63	90%	3.66	23

Table 2: Statisitcs for restarts, with $p = 0.15$ and $\alpha = 3.5$

To illustrate this point, let us now examine more closely the situation regarding restarts for outlinks, rounding off p to be 0.15 and α to be 3.5. It can be verified that the probability that p_3 be ill-defined is 0.15, that p_4 be ill-defined is about 0.1905, that p_5 be ill-defined is about 0.1769 and that p_6 be ill-defined is 0. Thus the total probability of p_k being ill-defined for $k \leq 6$ is about 0.5174. Table 2 shows the values of a simulation where the p_k-model was run 1000 times in batches of 100 runs each. Whenever p_k was ill-defined for a given run, this run was considered to be a restart and the simulation moved on to the next run. The second column shows the numbers of restarts within the batch, the third column shows the percentage of the restarts observed with p_k ill-defined for $k \leq 10$, the fourth column shows the average stage at which the restarts became ill-defined and the fifth column shows the maximum stage at which any restart became ill-defined. Thus, if the process is well defined for, say, 50 or more stages, it is very unlikely to become ill-defined at a later stage.

4 Concluding Remarks

We have extended Simon's classical stochastic model by adding to it a non-preferential component which is combined with preferential attachment. When viewing this stochastic process in terms of an urn transfer model, this amounts to choosing balls proportional to the number of times they have previously been selected (i.e. the number of pins) plus a constant $\alpha > -1$. From the equations of this process we derived an asymptotic formula for the exponent of the resulting power-law distribution. As far as we are aware our proof given in the Appendix is the first formal proof of the convergence of Simon's model; unlike in previous work, we do not rely on the mean-field theory approach.

Utilising our result we are able to explain several power-law distributions in the Web graph, which we now summarise. For the distribution of incoming links we derived $\alpha \approx -0.37$, for the distribution of outgoing links we derived $\alpha \approx 3.42$, for the distribution of pages in a Web site we derived $\alpha \approx 2.50$ and, finally, for the distribution of visitors to a Web site we derived $\alpha \approx -0.76$. In all cases our extension of Simon's original model can better explain the exponent of the power-law distribution, implying that there is some mixture of preferential and non-preferential attachment in the selection process.

The power law distribution that we have established can be stated as a hypothesis: *in order to explain the evolution of the Web graph both preferential and non-preferential processes are at work*. This hypothesis is more consistent with empirical data than the one which utilises only preferential attachment. Our model is still limited to the cases where the exponent of the power-law distribution is greater than two. We are currently investigating a possible model which could yield an exponent less than two.

A Appendix : Proofs

We first prove (7) and (8). Since at stage $k+1$ we add a new ball with probability p_{k+1},

$$E_k(B(k+1)) = B(k) + p_{k+1},$$

so, taking expectations,

$$E(B(k+1)) = E(B(k)) + E(p_{k+1}). \tag{23}$$

Lemma A.1 For $k > 1$,

$$E(B(k)) = E\left(\sum_{i=1}^{k} F_i(k)\right) = 1 + (k-1)p \tag{24}$$

and

$$E(p_k) = p. \tag{25}$$

Proof. We prove the result by induction on k. For $k = 2$, remembering that $B(1) = 1$ and using (3), it is easy to see that $p_2 = p$, and thus, by using (23), that

$$E(B(2)) = 1 + E(p_2) = 1 + p.$$

Now assume that (24) and (25) hold for some k, $k > 1$. Then,

$$E(p_{k+1}) = 1 - \frac{(1-p)(k + \alpha E(B(k)))}{k(1+\alpha p) + \alpha(1-p)} = 1 - \frac{(1-p)(k + \alpha(1 + (k-1)p))}{k(1+\alpha p) + \alpha(1-p)} = p$$

and thus, using (23),

$$E(B(k+1)) = 1 + (k-1)p + p = 1 + kp. \quad \Box$$

We now consider condition (4) needed for p_{k+1} to be well defined.

Lemma A.2 p_{k+1} is always well defined (i.e. non-negative) for all k when $p \geq 1/2$, but only if

$$\alpha \leq \frac{p}{1-2p}$$

when $p < 1/2$.

Proof. In order that $p_{k+1} \geq 0$, condition (4) must hold. This is equivalent to

$$\alpha(B(k) - 1) \leq \frac{kp\ (1+\alpha)}{1-p}. \tag{26}$$

There are three cases to consider:

(I) When $\alpha = 0$, there are no restrictions on p.

(II) When $-1 < \alpha < 0$, it is straightforward to see that again there are no restrictions on p since, in this case, the maximum value of the left-hand side of (26) is zero, when $B(k) = 1$.

(III) When $\alpha > 0$, we see from (26) that we must have

$$p \geq \frac{\alpha(B(k)-1)}{\alpha(B(k)-1) + k(1+\alpha)}.$$

Setting $B(k)$ to its maximum value k, this requires that

$$p \geq \frac{\alpha(k-1)}{\alpha(2k-1) + k}, \tag{27}$$

which will hold for all k provided

$$p \geq \frac{\alpha}{2\alpha+1},$$

in particular this holds for all α when $p \geq 1/2$. However, for $p < 1/2$, for (27) to hold for all k we need

$$\alpha \leq \frac{p}{1-2p}. \quad \square$$

We conclude by proving that as k tends to infinity $E(F_i(k))/k$ tends to f_i, justifying our derivation of the asymptotic distribution of the balls in the urns. We first state some useful properties of f_i, which may be verified directly using (16) and (17).

Lemma A.3

(I) For all $i \geq 1$, $0 < f_i < 1$ and $f_i > f_{i+1}$.

(II) $\sum_{i=1}^{\infty} f_i = p$ and $\sum_{i=1}^{\infty} i f_i = 1$. $\square$

Theorem A.4 For all $i \geq 1$,

$$\lim_{k\to\infty} \frac{E(F_i(k))}{k} = f_i.$$

Proof. Using (18) to rewrite (14) and (15), we obtain, for $i > 1$,

$$(k+1)(f_i + \epsilon_{i,k+1}) = k(f_i + \epsilon_{i,k}) + k\beta_k(i-1+\alpha)(f_{i-1} + \epsilon_{i-1,k}) - k\beta_k(i+\alpha)(f_i + \epsilon_{i,k}) \quad (28)$$

and, for $i = 1$,

$$(k+1)(f_1 + \epsilon_{1,k+1}) = k(f_1 + \epsilon_{1,k}) - k\beta_k(1+\alpha)(f_1 + \epsilon_{1,k}) + p. \quad (29)$$

Equations (16) and (17) may be written in a similar form as

$$(k+1)f_i = kf_i + \beta(i-1+\alpha)f_{i-1} - \beta(i+\alpha)f_i \quad (30)$$

and

$$(k+1)f_1 = kf_1 - \beta(1+\alpha)f_1 + p. \quad (31)$$

For $i > 1$, subtracting (30) from (28) yields:

$$(k+1)\epsilon_{i,k+1} = k\epsilon_{i,k} + k\beta_k(i-1+\alpha)\epsilon_{i-1,k} - k\beta_k(i+\alpha)\epsilon_{i,k} + (k\beta_k - \beta)\Big((i-1+\alpha)f_{i-1} - (i+\alpha)f_i\Big).$$

Using (16) and (11) this simplifies to

$$(k+1)\epsilon_{i,k+1} = \Big(1 - \beta_k(i+\alpha)\Big)k\epsilon_{i,k} + \beta_k(i-1+\alpha)k\epsilon_{i-1,k} - \alpha\beta_k f_i. \quad (32)$$

Similarly, for $i = 1$, subtracting (31) from (29) and using (11), we obtain:

$$(k+1)\epsilon_{1,k+1} = \Big(1 - \beta_k(1+\alpha)\Big)k\epsilon_{1,k} + \alpha\beta\beta_k(1+\alpha)f_1. \tag{33}$$

From (32), by virtue of Lemma A.3(I) and (13), for $i > 1$, we have

$$(k+1)\mid\epsilon_{i,k+1}\mid \leq \Big(1 - \beta_k(i+\alpha)\Big)k\mid\epsilon_{i,k}\mid + \beta_k(i-1+\alpha)k\mid\epsilon_{i-1,k}\mid + \mid\alpha\mid\beta_k. \tag{34}$$

We now define

$$\delta_k = \max_{i\geq 1}\mid\epsilon_{i,k}\mid = \max_{1\leq i\leq k+1}\mid\epsilon_{i,k}\mid. \tag{35}$$

(The two maxima are equal since, by (18), $\epsilon_{i,k} = -f_i$ for $i > k$, and f_i is monotonic decreasing.)

On using (35), inequality (34) yields

$$(k+1)\mid\epsilon_{i,k+1}\mid \leq (1-\beta_k)k\delta_k + \mid\alpha\mid\beta_k \tag{36}$$

for $i > 1$.

Similarly, from (33), on using (35) together with (12), (13) and Lemma A.3(I), it follows that

$$(k+1)\mid\epsilon_{1,k+1}\mid \leq \Big(1 - \beta_k(1+\alpha)\Big)k\delta_k + \mid\alpha\mid\beta_k(1+\alpha). \tag{37}$$

We show by induction on k that

$$k\delta_k \leq \max\{1, \mid\alpha\mid\}. \tag{38}$$

From (18) and (35) we see that $\delta_1 = \max\{1 - f_1, f_2\}$. So, by Lemma A.3(I), (38) holds for $k = 1$.

Now assume that (38) holds for some $k \geq 1$. So, for $i > 1$, (36) gives

$$(k+1)\mid\epsilon_{i,k+1}\mid \leq (1-\beta_k)\max\{1, \mid\alpha\mid\} + \mid\alpha\mid\beta_k \leq \max\{1, \mid\alpha\mid\}. \tag{39}$$

Similarly, for $i = 1$, (37) gives

$$(k+1)\mid\epsilon_{1,k+1}\mid \leq \Big(1 - \beta_k(1+\alpha)\Big)\max\{1, \mid\alpha\mid\} + \mid\alpha\mid\beta_k(1+\alpha) \leq \max\{1, \mid\alpha\mid\}. \tag{40}$$

Therefore, from (39) and (40), $(k+1)\delta_{k+1} \leq \max\{1, \mid\alpha\mid\}$.

So, by induction, (38) holds for all $k \geq 1$. Thus, to conclude the proof, we note that, as k tends to infinity, δ_k tends to 0, so $\epsilon_{i,k}$ tends to 0 for all i. □

References

[ABJ00] R. Albert, A.-L. Barbási, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, 281:69–77, 2000.

[AH00] L.A. Adamic and B.A. Huberman. The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce*, 1:5–12, 2000.

[AS72] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Dover, New York, NY, 1972.

[BAJ99] A.-L. Barbási, R. Albert, and H. Jeong. Mean-field theory for scale free random networks. *Physica A*, 272:173–189, 1999.

[BE00] S. Bornholdt and H. Ebel. World-Wide Web scaling exponent from Simon's 1955 model. *Condensed Matter Archive*, cond-mat/0008465, 2000.

[BKM+00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, A. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph strucutre in the web. *Computer Networks and ISDN Systems*, 30:309–320, 2000.

[DKM+01] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the Web. In *Proceedings of International Conference on Very Large Data Bases*, pages 69–78, Rome, 2001.

[DMS00a] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85:4633–4636, 2000.

[DMS00b] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. WWW and internet models from 1955 till our day and the "popularity is attractive" principle. *Condensed Matter Archive*, cond-mat/0009090, 2000.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 251–262, Cambridge, Ma., 1999.

[HA99] B.A. Huberman and L.A. Adamic. Evolutionary dynamics of the World Wide Web. *Nature*, 399:131, 1999.

[Hen01] M.R. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 1:45–50, 2001.

[JK77] N.L. Johnson and S. Kotz. *Urn Models and their Application: An Approach to Modern Discrete Probability.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, NY, 1977.

[KH95] M. Koenig and T. Harrell. Lotka's law, Price's urn, and electronic publishing. *Journal of the American Society for Information Science*, 46:386–388, 1995.

[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The web as a graph. In *Proceedings of ACM Symposium on Principles of Database Systems*, pages 1–10, Dallas, Tx., 2000.

[KRRT99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *Proceedings of International Conference on Very Large Data Bases*, pages 639–650, Edinburgh, 1999.

[Man59] B. Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by H.A. Simon. *Information and Control*, 2:90–99, 1959.

[MM00] B.H. Murray and A. Moore. Sizing the internet. White paper, Cyveillance, July 2000.

[Nic89] P.T. Nicholls. Bibliometric modeling processes and the empirical validity of Lotka's law. *Journal of the American Society of Information Science*, 40:379–385, 1989.

[Pri76] D. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society of Information Science*, 27:292–306, 1976.

[Rap82] A. Rapoport. Zipf's law revisited. In H. Guiter and M.V. Arapov, editors, *Studies on Zipf's Law*, pages 1–28. Studienverlag Brockmeyer, Bochum, Germany, 1982.

[Sch91] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* W.H. Freeman, New York, NY, 1991.

[Sim55] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[SV63] H.A. Simon and T. Van Wormer. Some Monte Carlo estimates of the Yule distribution. *Behavrioral Science*, 8:203–210, 1963.