PageRank and The Random Surfer Model

Prasad Chebolu* Páll Melsted*

Abstract

In recent years there has been considerable interest in analyzing random graph models for the Web. We consider two such models - the Random Surfer model, introduced by Blum et al. [7], and the PageRank-based selection model, proposed by Pandurangan et al. [18]. It has been observed that search engines influence the growth of the Web. The PageRank-based selection model tries to capture the effect that these search engines have on the growth of the Web by adding new links according to Pagerank. The PageRank algorithm is used in the Google search engine [1] for ranking search results.

We show the equivalence of the two random graph models and carry out the analysis in the Random Surfer model, since it is easier to work with. We analyze the expected in-degree of vertices and show that it follows a powerlaw. We also analyze the expected PageRank of vertices and show that it follows the same powerlaw as the expected degree.

We show that in both models the expected degree and the PageRank of the first vertex, the root of the graph, follow the same powerlaw. However, the power undergoes a phase-transition as we vary the parameter of the model. This peculiar behavior of the root has not been observed in previous analysis and simulations of the two models.

1 Introduction

The structure of large-scale real networks has been studied extensively in recent years. A number of models have been proposed to model the World Wide Web graph or the Internet graph, usually as the outcome of a random process where the graph is grown one node at a time. Experimental studies by, Faloutsos, Faloutsos and Faloutsos [14], Albert, Jeong and Barabasi [2] and Broder et al. [10], have shown that the degree distribution follows a power law with a heavy tail. Several other properties of the Web graph have been studied, such as the diameter, number of small bipartite cliques and the distribution of PageRank.

*Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA 15213, USA.

A prominent model for the web graph is the Preferential Attachment model suggested by Barabasi et al. [5]. The first rigorous treatment of the model was given by Bollobas, Riordan, Spencer and Tusnády [8], which showed a power law for the degree distribution. Several extensions of the Preferential Attachment have been proposed and analyzed, e.g. a generalized version of Preferential Attachment by Cooper and Frieze [12], and the copying model by Kumar et al. [15]. For an overview of the history of such generative models see [17].

PageRank is an important ingredient in the ranking of search results used by Google [1]. Pandurangan et al. [18] studied the distribution of PageRank values on the Web and gave a generative model that grows the graph according to PageRank of each node. In another seemingly unrelated paper Blum et al. [7] considered a variation on the preferential attachment model where endpoints of new vertices are selected by taking a short directed random walk on the graph.

In fact we show that the two models are equivalent and proceed to analyze the expected in-degree and PageRank of vertices. To state the results we first give a more detailed description of the models.

1.1 Definitions and Results

The Random Surfer Web-Graph model is a sequence of directed graphs $G_t, t = 1, 2, \ldots$. The graph G_t has t vertices and t edges. The process is parameterized with a probability p and we let $q = 1 - p$. G_1 consists of a single vertex v_1 and a directed self loop. Given G_t we create G_{t+1} in the following manner:

1. add v_{t+1} to the graph
2. pick u uniformly at random from $v_1, \ldots, v_t$
3. with probability p add the edge (v_{t+1}, u) to the graph and stop
4. otherwise with probability q replace u by the unique vertex u points to and repeat step 3

We see that G_t will be a directed tree rooted at the first vertex, v_1, which we will refer to as the root.

For a directed graph the PageRank is defined as the stationary probability distribution of a random walk

on the graph. The random walk follows the outgoing edges of the graph, but will restart itself every time with probability $1-c$, and start at a new vertex, chosen uniformly at random. The probability c is called the decaying factor and influences the convergence of the computation of PageRank; typical values of c are about 0.85 [9].

The PageRank π_v for a graph G is given by the system of equations

$$\pi_v = \frac{1-c}{n} + c \sum_{u \in N^-(v)} \frac{\pi_u}{d^+(u)} \tag{1.1}$$

and $\sum_{v \in G} \pi_v = 1$. Here n is the number of vertices of G, $N^-(v)$ is the set of vertices pointing to v and $d^+(u)$ is the outdegree of u. For a more detailed discussion of PageRank see [16] and [9].

The PageRank-based selection model described by [18] follows the same scheme as the random surfer model except steps $2-4$ are replaced by the rule: pick u with probability π_u and add (v_{t+1}, u); where π is the PageRank distribution of G_t.

THEOREM 1.1. *The PageRank-based selection model gives the same distribution over graphs as the random surfer model, provided* $1-c=p$.

We note that this theorem is similar to a theorem in [13], however the context and motivations are different. This theorem is useful for simulation of the PageRank-based selection model since it allows us to sample from the model without actually computing the PageRank of each and every vertex. The random surfer model comes in handy since computing PageRank is a resource-intensive task.

It should be noted that when PageRank is used in practice, self-loops are removed and vertices with outdegree of 0, so called dangling nodes, are given edges to all other vertices. This would result in a different model, as the root would have an edge to every vertex in the graph. The random surfer model can be modified to account for this and the connection to PageRank still holds. However, this considerably complicates the analysis and is outside the scope of this paper.

THEOREM 1.2. *The expected in-degree in* G_n *of a vertex* u *that comes in at time* $t_u \geq 2$ *is*

$$\mathbf{E}[d^-_{G_n}(u)] = \frac{1}{q} \sum_{i=1}^{n-t_u} P_i(t_u, n)(pq)^i \frac{1}{i} \binom{2i-2}{i-1}$$

where

$$P_i(a,b) = \sum_{a \leq t_1 < \ldots < t_i < b} \frac{1}{t_1 \cdots t_i}$$

If $t_u = o(n)$ *then the expected degree is asymptotically equal to*

$$\Theta\left(\frac{(n/t_u)^{4pq}}{\ln^{3/2}(n/t_u)} \right) .$$

In their paper Blum et al. [7] gave a lower bound of $\left(\frac{n}{t_u}\right)^{pq}$ for the expected degree of a vertex. We obtain the correct exponent of $4pq$ for the expected degree of a vertex (other than the root). They obtained $\Theta((\frac{n}{t_u})^p)$ for the virtual degree of a vertex, but they could not relate the virtual degree to the actual degree of a vertex. We analyze the root separately. *What is special about the root?* Every random walk that reaches the root terminates at the root unlike at a normal vertex. This results in the root undergoing a phase transition in its expected degree which was not observed in their paper. This behavior of the root is summarized in the following theorem.

THEOREM 1.3. *The expected in-degree of the root undergoes a phase transition at* $p = \frac{1}{2}$.

The expected degree of the root is

$$E[d^-_{G_n}(v_1)] = \begin{cases} \tilde{\Theta}(n) & \text{for } p \leq \frac{1}{2} , \\ \tilde{\Theta}(n^{4pq}) & \text{for } p > \frac{1}{2} . \end{cases}$$

For $p \leq 1/2$, we observe that the root has degree $\tilde{\Theta}(n)$, where $\tilde{\Theta}(n)$ is shorthand for $O(n \ln^k n)$ for some k, while the first few vertices that come in at time $O(1)$ have degree $\tilde{\Theta}(n^{4pq})$. This shows that most vertices have an edge to the root and the tree is almost "star-like" with the root at the center of the star. This is the expected behavior, as every incoming vertex takes a random walk of expected length at least 1, thereby being more likely to be closer to the root. For $p > 1/2$, the expected degree of the root is comparable to any other vertex and is $\tilde{\Theta}(n^{4pq})$.

THEOREM 1.4. *The expected PageRank at time* n *of a vertex* u *that comes in at time* $t_u \geq 2$ *is*

$$\mathbf{E}[\pi_u(n)] = \frac{p}{n} + \frac{p}{n} \sum_{i=1}^{n-t_u} P_i(t_u, n)(pq)^i \frac{1}{i+1} \binom{2i}{i}$$

If $t_u = o(n)$ *then this is asymptotically equal to*

$$\frac{p}{n} \tilde{\Theta}\left(\left(\frac{n}{t_u} \right)^{4pq} \right)$$

A similar result was obtained for the PageRank for a slight modification of the Preferential Attachment model [4]. Power law results for PageRank have been observed in large webgraphs and verified by simulation [18, 6]. The approximations used in the analysis

in [18] are not tight enough to yield the correct power for the power law. Using the same approximation and treating the expected value as a deterministic value we would get a power law with a power $\frac{1}{4pq}+1$. However, to be able to justify such a claim we would need good concentration for the degree of a vertex, which we presently do not have.

THEOREM 1.5. *The expected PageRank of the root at time n is asymptotically given by*

$$E[\pi_{v_1}(n)] = \begin{cases} \frac{1}{n}\tilde{\Theta}(n) & \text{if } p \le \frac{1}{2}\ , \\ \frac{1}{n}\tilde{\Theta}(n^{4pq}) & \text{if } p > \frac{1}{2}\ . \end{cases}$$

When $p < \frac{1}{2}$ there is a sharp difference between the PageRank of the root and other vertices. This behavior can be compared to [11] where it is shown that a small set of "celebrity" sites accumulate a constant fraction of all links created. Their motivation was to look at the effect of search engine results on the growth of the Web.

2 Analysis of the Random Surfer model

Let $G_t, t = 1, \ldots, n$ be an evolving random surfer graph where $G = G_n$ is the final graph. Consider a vertex u, different from the root, that comes in at time $t_u \ge 2$. At any time $t \ge t_u$ the probability that a new vertex, added at time $t+1$, connects to u is dependent only on the subtree rooted at u. Furthermore if we know that the new vertex takes a random walk of length k, then it will connect to u if and only if it starts k levels below u in u's subtree.

Thus, we introduce the random vector $L(t) = [L_k(t)]_{k=0}^{n-1}$ where $L_k(t)$ is the number of vertices in the subtree of u at depth k. Here we define u to be at depth $k = 0$ and $L_0(t) = 1$ for all $t \ge t_u$. Now the probability that the new vertex connects to u given that it starts at depth k in u's subtree is pq^k and the probability that it starts the random walk at depth k is $\frac{1}{t}L_k(t)$. Thus summing over all levels, the probability that the new vertex connects to u is

$$\frac{p}{t}(L_0(t) + qL_1(t) + q^2L_2(t) + \ldots + q^{n-1}L_{n-1}(t)) \tag{2.2}$$

In order to find the expected degree of u we need to keep track of the size of each level of u's subtree.

Repeating the same argument as for (2.2) we can find the probability that level k acquires a new vertex that comes in at time $t+1$

$$\Pr[L_k(t+1) = L_k(t) + 1 \mid L(t)] \tag{2.3}$$

$$= \frac{p}{t}\left(L_{k-1}(t) + qL_k(t) + \ldots + q^{n-k}L_{n-1}(t)\right) \tag{2.4}$$

$$= \frac{p}{t}[A \cdot L(t)]_k \tag{2.5}$$

for $1 \le k \le n-1$ and $t \ge t_u$. Here A is the $n \times n$ matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & q & q^2 & \ldots & q^{n-2} & q^{n-1} \\ 0 & 1 & q & \ldots & q^{n-3} & q^{n-2} \\ \vdots & & & \ddots & & \\ 0 & 0 & 0 & \ldots & 1 & q \end{bmatrix} \tag{2.6}$$

For aesthetic reasons we will start counting rows and columns of A from 0 onwards. In general $A_{k,j} = q^{j-k+1}$ if $j - k + 1 \ge 0, k \ge 1$ and 0 otherwise.

2.1 The root The root vertex is treated differently than the general case, since it has a directed self loop. This means that once a random walk hits the root it will stop at the root eventually. As before we let $L_k(t)$ be the number of nodes at depth k. Given that a vertex starts at depth k it will connect to the root if its random walk is of length at least k. We see that the probability of connecting to the root is then

$$pq^k + pq^{k+1} + \ldots = pq^k(1 + q + q^2 + \ldots) = pq^k\frac{1}{1-q} = q^k$$

Therefore a new vertex that comes in at time $t+1$ connects to the root with probability

$$\frac{1}{t}(L_0(t) + qL_1(t) + q^2L_2(t) + \ldots + q^{n-1}L_{n-1}(t)) \tag{2.7}$$

For $k \ge 2$ the probability that the new vertex is at depth k is the same as in the case for general vertices, given by (2.4). We can write the probabilities in matrix form as in (2.5) as

$$\Pr[L_k(t+1) = L_k(t) + 1 \mid L(t)] = \frac{1}{t}[BL(t)]_k \tag{2.8}$$

where

$$B = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & q & q^2 & \ldots & q^{n-2} & q^{n-1} \\ 0 & p & pq & \ldots & pq^{n-3} & pq^{n-2} \\ \vdots & & & \ddots & & \\ 0 & 0 & 0 & \ldots & p & pq \end{bmatrix} \tag{2.9}$$

Note that B differs from pA only in the second row.

3 PageRank Model

We begin by proving Theorem 1.1

Proof. Initially both models start with a single vertex and a self loop. We show that the distributions over G_{t+1} are the same for both models if they start at the same graph G_t. Hence the distributions are the same by an inductive argument.

Conditional on starting at the same graph G_t, the distributions over G_{t+1} are determined by the probability that the new vertex v_{t+1} selects u as its endpoint. Therefore it is enough to show that both models give the same distribution over vertices.

Let P be the adjacency matrix of the graph G_t and note that in both the PageRank based selection model and the random surfer model each vertex has an out-degree of 1. We can write (1.1) with $c = 1 - p = q$

$$\pi = \frac{p}{t}\mathbf{1} + q\pi P \iff \pi(I - qP) = \frac{p}{t}\mathbf{1}$$
$$\iff \pi = \frac{p}{t}(I - qP)^{-1}\mathbf{1} = \frac{p}{t}\sum_{k=0}^{\infty} q^k P^k \mathbf{1}$$

where $\mathbf{1}$ is the vector of all ones. Now $[P^k\mathbf{1}]_u$ counts the number of paths of length k that end at the vertex u. So when u is not the root we have $[P^k\mathbf{1}]_u = L_k(t)$ where $L_k(t)$ is the number of vertices at depth k in the subtree of u. Then the PageRank of u is

$$\pi_u(t) = \frac{p}{t}\sum_{k=0}^{\infty} q^k L_k(t) = \frac{p}{t}\sum_{k=0}^{t-1} q^k L_k(t) \tag{3.10}$$

which is exactly the same formula as in (2.2).

For the root v_1 we have $[P^k\mathbf{1}]_{v_1} = L_0(t) + L_1(t) + \ldots + L_k(t)$, since we can take k steps or less to the root and then loop once the root is hit. So the PageRank of the root is

$$\begin{aligned}\pi_{v_1}(t) &= \frac{p}{t}\sum_{k=0}^{\infty} q^k \sum_{j=0}^{k} L_j(t) \\ &= \frac{p}{t}\sum_{j=0}^{\infty} L_j(t)\sum_{k=j}^{\infty} q^k = \frac{p}{t}\sum_{j=0}^{t-1} L_j(t)\frac{q^j}{1-q} \\ &= \frac{1}{t}\sum_{j=0}^{t-1} q^j L_j(t)\end{aligned}$$

Which is the same formula as in (2.7). Hence the distributions are the same.

4 Expected degree

We prove a stronger version of Theorem 1.2 which will be used in subsequent proofs.

THEOREM 4.1. *For a vertex u that comes in at time $t_u \geq 2$ the expected number of vertices at depth k in u's subtree is*

$$\mathbf{E}[L_k(n)] = \sum_{i=k}^{n-t_u} P_i(t_u, n) p^i q^{i-k} \frac{k}{i}\binom{2i-k-1}{i-1} \tag{4.11}$$

In particular, for $k = 1$ this implies the first part of Theorem 1.2.

Proof. From (2.5) we see that

$$\mathbf{E}[L(t+1)|L(t)] = L(t) + \frac{p}{t}AL(t) = (I + \frac{p}{t}A)L(t)$$

Taking expected values of both sides we get $\mathbf{E}[L(t+1)] = (I + \frac{p}{t}A)\mathbf{E}[L(t)]$. Solving the recurrence equation and noting that $L(t_u) = e_0$ where e_0 is the n-vector $[1, 0, \ldots]^T$ we get

$$\mathbf{E}[L(n)] = \left(\prod_{t=t_u}^{n-1}(I + \frac{p}{t}A)\right)e_0$$

Multiplying through the right hand side and grouping by powers of A,we get

$$\mathbf{E}[L(n)] = \sum_{i=0}^{n-t_u} P_i(t_u, n) p^i A^i e_0$$

where

$$P_i(t_u, n) = \sum_{t_u \leq t_1 < t_2 < \ldots < t_i < n} \frac{1}{t_1 \cdot t_2 \cdots t_i}$$

corresponds to the $\frac{1}{t}$ factors in the coefficient of A. Now the following Lemma finishes the proof

LEMMA 4.1. *For $i \geq 1, k \geq 1$ and $k \leq i$ we have*

1. $[A^i e_0]_k = q^{i-k}\frac{k}{i}\binom{2i-k-1}{i-1}$
2. $[B^i e_0]_1 = q^{i-1}\sum_{j=1}^{i-1}\frac{j}{i-1}\binom{2(i-1)-j-1}{i-2}p^{i-j-1}$, *for* $i \geq 2$
3. $[B^i e_0]_k = q^{i-k}\sum_{j=k-1}^{i-1}\frac{j}{i-1}\binom{2(i-1)-j-1}{i-2}p^{i+k-j-2}$, *for* $k \geq 2$

The proof of Lemma 4.1 is given in the Appendix. We now derive the asymptotic expression for the expected in-degree as stated in Theorem 1.2.

Proof. From Theorem 4.1 we have $\mathbf{E}[L_1(n)] = \frac{1}{q}\sum_{i=1}^{n-t_u} P_i(t_u, n)\frac{(pq)^i}{i}\binom{2(i-1)}{i-1}$. A simple upper bound on $P_i(t_u, n)$ can be obtained by dropping the restriction that we sum over distinct numbers, thus

$$\begin{aligned}P_i(t_u, n) &\leq \sum_{t_1,\ldots,t_i \in [t_u, n-1]} \frac{1}{i!}\frac{1}{t_1 \cdots t_i} \\ &= \frac{1}{i!}\left(\sum_{t_1=t_u}^{n-1}\frac{1}{t_1}\right)\cdots\left(\sum_{t_i=t_u}^{n-1}\frac{1}{t_i}\right)\end{aligned}$$

$$= \frac{(H_{n-1} - H_{t_u-1})^i}{i!} \leq \frac{\left(\ln \frac{n}{t_u} + 1\right)^i}{i!}$$
$$= \frac{\left(\ln \frac{en}{t_u}\right)^i}{i!}$$

where H_k is the k-th harmonic number.

Now let $\mu = 4pq \ln(en/t_u)$, $\delta = 4\sqrt{\frac{\ln \ln(en/t_u)}{\mu}}$ and $I = [\mu(1-\delta), \mu(1+\delta)]$. We see that most of the contribution to the sum comes from $i \in I$. In fact we have

$$\sum_{i \notin I} P_i(t_u, n) \frac{(pq)^i}{i} \binom{2(i-1)}{i-1}$$
$$\leq \sum_{i \notin I} \frac{(pq \ln(en/t_u))^i 2^{2(i-1)}}{i!}$$
$$= \frac{1}{4} \sum_{i \notin I} \frac{\mu^i}{i!} = \frac{e^\mu}{4} \sum_{i \notin I} \frac{e^{-\mu}\mu^i}{i!}$$
$$= \frac{e^\mu}{4} \Pr[|X - \mu| > \delta\mu]$$

where $X \sim Poi(\mu)$. By standard concentration results (see [3]) we can see that this probability is at most $2e^{-\delta^2\mu/4}$, thus the contribution is at most

$$e^\mu e^{-4 \ln \ln(en/t_u)} = \frac{(en/t_u)^{4pq}}{\ln^4(en/t_u)}.$$

For $i \in I$ we can approximate $\frac{1}{i}\binom{2(i-1)}{i-1}$ with $\frac{2^{2(i-1)}}{i^{3/2}\sqrt{\pi}}$ and get

$$\sum_{i \in I} P_i(t_u, n) \frac{(pq)^i}{i} \binom{2(i-1)}{i-1}$$
$$= \frac{(1+o(1))}{4} \sum_{i \in I} P_i(t_u, n) \frac{(4pq)^i}{i^{3/2}\sqrt{\pi}}$$
$$= \frac{1+o(1)}{4\sqrt{\pi} \ln^{3/2}(en/t_u)} \sum_{i \in I} P_i(t_u, n)(4pq)^i$$
$$= \frac{1+o(1)}{4\sqrt{\pi} \ln^{3/2}(en/t_u)} \sum_{i=0}^{n-t_u} P_i(t_u, n)(4pq)^i$$
$$= \frac{1+o(1)}{4\sqrt{\pi} \ln^{3/2}(en/t_u)} \prod_{t=t_u}^{n-1} (1 + \frac{4pq}{t})$$
$$= \Theta\left(\frac{(n/t_u)^{4pq}}{\ln^{3/2}(n/t_u)}\right).$$

Note that we can switch from summing over I to summing over the whole interval, since the contribution from outside I is $O(\ln(n/t_u)^{-5/2})$ of the final sum.

5 Root

When $p = 0$ the random surfer model gives a star so the root has degree $n-1$. When $p = 1$ the random surfer model is equivalent to selecting endpoints uniformly at random, the degree of the root is then polylogarithmic in n. For values of $p \in (0, 1)$ we show that the expected degree of the root is decreasing in p.

THEOREM 5.1. *Let $d_p(n)$ be the in-degree of the root in a Random-Surfer graph with parameter p. Then $d_{p_1}(n)$ stochastically dominates $d_{p_2}(n)$ for all $p_1 < p_2$, i.e. we can couple the two models s.t. inequality holds for every instance.*

Proof. We can generate an instance of a Random-Surfer graph with a sequence of pairs $(v_2, t_2), \ldots, (v_n, t_n)$ where v_k is picked uniformly at random from $[1, k-1]$ and t_k is taken uniformly at random from $[0, 1]$. Given p we construct the graph in the following way from the sequence. Vertex k begins by picking vertex v_k and then takes a random walk of length l towards the root, where l is the smallest number s.t. $t_k \leq 1-(1-p)^{l+1}$. Since t_k is uniformly distributed we see that l is geometric with parameter p.

If we use the same sequence to generate instances G_1 and G_2 with parameters $p_1 < p_2$, then the lengths of the random walks will be longer in G_1. Thus we see that vertices with the same label in both graphs will always be closer to the root in G_1 than in G_2. In particular this shows that a vertex connected to the root in G_2 must also be connected to the root in G_1.

Thus $d_{p_1}(n) \geq d_{p_2}(n)$ for these coupled instances.

This implies that $\mathbf{E}[d^-_{G_n}(v_1)]$ is decreasing as a function of p.

THEOREM 5.2. *For the root, v_1, the expected values for $L(n)$ are*

$$\mathbf{E}[L_1(n)] = \sum_{i=1}^{n-1} P_i q^{i-1} \sum_{j=1}^{i-1} \frac{j}{i-1} \binom{2i-j-3}{i-2} p^{i-j-1}$$

and for $k \geq 2$

$$\mathbf{E}[L_k(n)] = \sum_{i=k}^{n-1} P_i q^{i-k} \sum_{j=k-1}^{i-1} \frac{j}{i-1} \binom{2i-j-3}{i-2} p^{i+k-j-2}$$

where $P_i = P_i(1, n)$

Proof. As in the proof of Theorem 4.1 we have

$$\mathbf{E}[L_k(n)] = \sum_{i=k}^{n-1} P_i(1, n)[B^i e_0]_k$$

Using Lemma 4.1 and plugging in for $[B^i e_0]_k$ finishes the proof.

We now prove Theorem 1.3

Proof. Note that $pAx \le Bx$ for any vector x with non-negative coordinates. This implies

$$\begin{aligned}\mathbf{E}[L_1(n)] &= \prod_{t=1}^{n-1}(I+\frac{1}{t}B)e_0 \\ &\ge \prod_{t=1}^{n-1}(I+\frac{p}{t}A)e_0 = \tilde{\Theta}(n^{4pq}) \qquad (5.12)\end{aligned}$$

The last equality follows from Theorem 1.2.

Take $p=\frac{1}{2}$, then (5.12) gives $\mathbf{E}[L_1(n)] \ge \tilde{\Theta}(n)$. Since $\mathbf{E}[L_1(n)]$ is decreasing in p we have $\mathbf{E}[L_1(n)] \ge \tilde{\Theta}(n)$ for all $p \le \frac{1}{2}$. Now $L_1(n) \le n-1$ which implies $\mathbf{E}[L_1(n)] = \tilde{\Theta}(n)$ for $p \le \frac{1}{2}$.

Now assume $p > \frac{1}{2}$. By Theorem 5.2 we have

$$\begin{aligned}&\mathbf{E}[L_1(n)] \\ &= \sum_{i=1}^{n-1} P_i(1,n) q^{i-1} \sum_{j=1}^{i-1} \frac{j}{i-1}\binom{2(i-1)-j-1}{i-2} p^{i-j-1} \\ &\le q^{-1}\sum_{i=1}^{n-1} P_i(1,n)(pq)^i \sum_{j=1}^{i-1} 2^{2i-j-3}p^{-j-1} \\ &\le \frac{1}{8pq}\sum_{i=1}^{n-1} P_i(1,n)(4pq)^i \sum_{j=1}^{i-1}(2p)^{-j} \\ &\le \frac{1}{16p^2q(2p-1)}\sum_{i=1}^{n-1} P_i(1,n)(4pq)^i = \tilde{\Theta}(n^{4pq})\end{aligned}$$

This, with (5.12), shows that $\mathbf{E}[L_1(n)] = \tilde{\Theta}(n^{4pq})$.

6 PageRank

We now prove Theorem 1.4

Proof. Recall that by Theorem 1.1 the PageRank of a vertex u is equal to the probability that a new vertex connects to u at time $n+1$. Therefore we get the following for the expected value of PageRank

$$\begin{aligned}&E[\pi_u(n)] \\ &= E[\Pr(v_{n+1} \text{ picks } u \text{ at time } n+1)] \\ &= \frac{p}{n}\left(\sum_{j=0}^{n-t_u} q^j \mathbf{E}[L_j(n)]\right) \\ &= \frac{p}{n} + \frac{p}{n}\sum_{j=1}^{n-t_u} q^j \left(\sum_{i=j}^{n-t_u} P_i(t_u,n) p^i q^{i-j} \frac{j}{i}\binom{2i-j-1}{i-1}\right) \\ &= \frac{p}{n} + \frac{p}{n}\sum_{j=1}^{n-t_u}\sum_{i=j}^{n-t_u} P_i(t_u,n)(pq)^i \frac{j}{i}\binom{2i-j-1}{i-1} \\ &= \frac{p}{n} + \frac{p}{n}\sum_{i=1}^{n-t_u} P_i(t_u,n)(pq)^i \sum_{j=1}^{i}\frac{j}{i}\binom{2i-j-1}{i-1} \\ &= \frac{p}{n} + \frac{p}{n}\sum_{i=1}^{n-t_u} P_i(t_u,n)(pq)^i \frac{1}{i+1}\binom{2i}{i}\end{aligned}$$

The asymptotic expression is obtained in a similar manner as in the proof of Theorem 1.2.

We now prove Theorem 1.5

Proof. As in the proof of Theorem 1.4 we have

$$\begin{aligned}\mathbf{E}[\pi_{v_1}(n)] &= \mathbf{E}[\Pr(v_{n+1} \text{ picks the root at time } n+1)] \\ &= \frac{1}{n}\left(\sum_{k=0}^{n-1} q^j \mathbf{E}[L_k(n)]\right) \\ &= \frac{1}{n}\left(1 + qE[L_1(n)] + \sum_{k=2}^{n-1} q^k \mathbf{E}[L_k(n)]\right) \qquad (6.13)\end{aligned}$$

When $p \le \frac{1}{2}$ we have by Theorem 1.3 that $\mathbf{E}[L_1(n)] = \tilde{\Theta}(n)$. Since $\pi_{v_1}(n) \le 1$, (6.13) shows that $\pi_{v_1}(n) = \frac{1}{n}\tilde{\Theta}(n)$.

Now assume $p > \frac{1}{2}$ and consider

$$\begin{aligned}&\sum_{k=2}^{n-1} q^k \mathbf{E}[L_k(n)] \\ &= \sum_{k=2}^{n-1} q^k \sum_{i=k}^{n-1} P_i q^{i-k} \sum_{j=k-1}^{i-1} \frac{j}{i-1}\binom{2i-j-3}{i-2} p^{i+k-2-j} \\ &= \sum_{i=2}^{n-1} P_i (pq)^i \sum_{k=2}^{i} p^k \sum_{j=k-1}^{i-1} \frac{j}{i-1}\binom{2i-j-3}{i-2} p^{-2-j} \\ &= \sum_{i=2}^{n-1} P_i (pq)^i \sum_{j=1}^{i-1} \frac{j}{i-1}\binom{2i-j-3}{i-2} p^{-2-j} \sum_{k=2}^{j+1} p^k \\ &\le \frac{p^2}{q}\sum_{i=2}^{n-1} P_i (pq)^i \sum_{j=1}^{i-1} \frac{j}{i-1}\binom{2i-j-3}{i-2} p^{-j} \\ &\le \frac{p^2}{q}\sum_{i=2}^{n-2} P_i (pq)^i \sum_{j=1}^{i-1} 2^{2i-j-3} p^{-j} \\ &= \frac{p^2}{8q}\sum_{i=2}^{n-2} P_i (4pq)^i \sum_{j=1}^{i-1}(2p)^{-j} \\ &\le \frac{p^2}{8q(2p-1)}\sum_{i=2}^{n-2} P_i (4pq)^i = \tilde{\Theta}(n^{4pq})\end{aligned}$$

which by Theorem 1.3 is of the same magnitude as $\mathbf{E}[L_1(n)]$. Thus we have that $\mathbf{E}[\pi_{v_1}(n)] = \frac{1}{n}\tilde{\Theta}(n^{4pq})$ when $p > \frac{1}{2}$.

7 Conclusions and Remarks

Here we have given exact and asymptotic formulas for the expected in-degree and expected PageRank of vertices. We have shown that the expected in-degree and PageRank follow the same powerlaw. We have also established that the expected in-degree and PageRank of the root must undergo a phase-transition at $p = \frac{1}{2}$. One weakness of the analysis is that we only get the expected values but we have not been able to establish any concentration. A possible approach would be to use concentration results for vector-valued martingales but we were unable to get meaningful bounds for Lipschitz constants. We are interested in the following open questions

1. To obtain a powerlaw for the number of vertices with degree $\leq d$ and similar powerlaw for number of vertices with PageRank $\leq x$. We conjecture that the power will be $1 + \frac{1}{4pq}$ for both distributions.

2. Can we repeat the analysis for the model where a new vertex comes in with d edges? This would result in a directed acyclic graph and not a directed tree.

3. Can the variation of PageRank, where self-loops are removed and dangling nodes are given edges to all other vertices, be analyzed? In particular it would be interesting to see whether the phase transitions of degrees and PageRank remain the same or are they simply an artifact of our model.

The authors would like to thank Alan Frieze for discussions and suggestions.

References

[1] Google technology overview [http://www.google.com /intl/en/corporate/tech.html], 2004

[2] R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.

[3] N. Alon, J. H. Spencer, and P. Erdős. *The Probabilistic Method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley and Sons, Inc., 1992.

[4] K. Avrachenkov and D. Lebedev. Pagerank of scale-free growing networks. *Internet Math.*, 3(2):207-231, 2007.

[5] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[6] L. Becchetti and C. Castillo. The distribution of pageRank follows a power-law only for particular values of the damping factor. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 941-942. Edinburgh, Scotland. ACM Press.

[7] A. Blum, T.-H. H. Chan, and M. R. Rwebangira. A random-surfer web-graph model. In *ANALCO '06: Proceedings of the eigth Workshop on Algorithm Engineering and Experiments and the third Workshop on Analytic Algorithmics and Combinatorics*, pages 238-246, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics.

[8] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Struct. Algorithms*, 18(3):279-290, 2001.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117, 1998.

[10] 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 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking*, pages 309-320, Amsterdam, The Netherlands, 2000. North- Holland Publishing Co.

[11] S. Chakrabarti, A. Frieze, and J. Vera. The influence of search engines on preferential attachment. In *SODA '05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 293-300, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.

[12] C. Cooper and A. Frieze. A general model of web graphs. *Random Struct. Algorithms*,22(3):311-335, 2003.

[13] K. Csalogány, D. Fogaras, B. Rácz, and T. Sarlós. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Math.*, 2(3):333-358, 2005.

[14] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251-262, New York, NY, USA, 1999. ACM Press.

[15] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 57, Washington, DC, USA, 2000. IEEE Computer Society.

[16] A. Langville and C. Meyer. Deeper inside pagerank. *Internet Math.*, 1(3):335-380, 2003.

[17] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1(2):226-251, 2004.

[18] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. *In Computing and Combinatorics: 8th Annual International Conference*, pages 330-339, Singapore, August 15-17, 2002.

Appendix

We now prove Lemma 4.1

Proof. We prove *1.* by induction. First $A^1 e_0 = [0, 1, 0, \ldots]$ and

$$
\begin{aligned}
&[A^{i+1}e_0]_j \\
&= \sum_{l=j-1}^{n} q^{l+1-j}[A^i e_0]_l = \sum_{l=j-1}^{i} q^{l+1-j} q^{i-l} \frac{l}{i}\binom{2i-l-1}{i-1} \\
&= q^{(i+1)-j} \sum_{l=j-1}^{i} \left[\frac{l}{i}\binom{2i-l}{i} - \frac{l}{i}\binom{2i-l-1}{i}\right] \\
&= q^{(i+1)-j} \left(\sum_{l=j-1}^{i} \left[\frac{l}{i}\binom{2i-l}{i} - \frac{l+1}{i}\binom{2i-(l+1)}{i}\right] \right. \\
&\qquad \left. + \sum_{l=j-1}^{i} \frac{1}{i}\binom{2i-(l+1)}{i} \right) \\
&= q^{(i+1)-j} \left(\frac{j-1}{i}\binom{2i-j+1}{i} - \frac{i+1}{i}\binom{i-1}{i} \right) \\
&\quad + \; q^{(i+1)-j} \left(\frac{1}{i} \sum_{l=j-1}^{i} \left[\binom{2i-l}{i+1} - \binom{2i-(l+1)}{i+1}\right] \right) \\
&= q^{(i+1)-j} \left(\frac{j-1}{i}\binom{2i-j+1}{i} \right) \\
&\qquad + q^{(i+1)-j} \left(\frac{1}{i}\binom{2i-j+1}{i+1} - \frac{1}{i}\binom{i-1}{i+1} \right) \\
&= q^{(i+1)-j} \left(\frac{(j-1)(2i-j+1)!}{i\,(i+1-j)!i!} + \frac{(2i-j+1)!}{i\,(i+1)!(i-j)!} \right) \\
&= q^{(i+1)-j} \frac{j}{i+1}\binom{2(i+1)-j-1}{(i+1)-1} \; .
\end{aligned}
$$

Here we have used the relation $\binom{n}{k} = \binom{n+1}{k+1} - \binom{n}{k+1}$ twice to get a telescoping sum.

We prove *2.* and *3.* simultaneously by induction. First $[B^2 e_0] = [0, q, p, 0 \ldots]$ and

$$
\begin{aligned}
&[B^{i+1}e_0]_1 \\
&= q \cdot q^{i-1} \sum_{j=1}^{i-1} \frac{j}{i-1}\binom{2(i-1)-j-1}{i-2} p^{i-1-j} \\
&\quad + \sum_{l=2}^{i} q^l q^{i-l} \sum_{j=l-1}^{i-1} \frac{j}{i-1}\binom{2(i-1)-j-1}{i-2} p^{i+l-2-j} \\
&= q^i \left(\sum_{j=1}^{i-1} \frac{j}{i-1}\binom{2(i-1)-j-1}{i-2} p^{i-1-j} \right. \\
&\qquad \left. + \sum_{l=2}^{i} \sum_{j=l-1}^{i-1} \frac{j}{i-1}\binom{2(i-1)-j-1}{i-2} p^{i+l-2-j} \right) \; .
\end{aligned}
$$

Treating the expression as a polynomial in p, we group terms by the powers of p. The coefficient of p^α, $0 \le \alpha \le i-1$, in the first sum is

$$
\frac{i-1-\alpha}{i-1}\binom{2(i-1)-(i-1-\alpha)-1}{i-2} \; .
$$

The coefficient of p^α in the second sum for a fixed value of l is

$$
\frac{i+l-2-\alpha}{i-1}\binom{2(i-1)-(i+l-2-\alpha)-1}{i-2} \; .
$$

In the above expression, for the binomial term to be non-zero,

$$
2(i-1)-(i+l-2-\alpha)-1 \ge i-2 \;\; \Rightarrow \;\; l \le \alpha+1 \; .
$$

The coefficient of p^α from the second sum is

$$
\sum_{l=2}^{\alpha+1} \frac{i+l-2-\alpha}{i-1}\binom{2(i-1)-(i+l-2-\alpha)-1}{i-2} \; .
$$

Summing up the coefficient from the first and second sum, we get

$$
\begin{aligned}
&\frac{i-1-\alpha}{i-1}\binom{2(i-1)-(i-1-\alpha)-1}{i-2} \\
&+ \sum_{l=2}^{\alpha+1} \frac{i+l-2-\alpha}{i-1}\binom{2(i-1)-(i+l-2-\alpha)-1}{i-2} \\
&= \sum_{l=1}^{\alpha+1} \frac{i+l-2-\alpha}{i-1}\binom{2(i-1)-(i+l-2-\alpha)-1}{i-2} \\
&= \sum_{j=i-1-\alpha}^{i-1} \frac{j}{i-1}\binom{2(i-1)-j-1}{i-2} \\
&= \frac{i-\alpha}{i}\binom{2i-(i-\alpha)-1}{i-1} \; .
\end{aligned}
$$

The summation in the second last step can be simplified to the above expression using the techniques used in the proof of *1.*

The coefficient of p^α is

$$
\frac{i-\alpha}{i}\binom{2i-(i-\alpha)-1}{i-1} \; .
$$

The expression for the entry $[B^{i+1}e_0]_1$ is given by

$$
q^i \sum_{\alpha=0}^{i-1} \frac{i-\alpha}{i}\binom{2i-(i-\alpha)-1}{i-1} p^\alpha
$$

$$= q^i \sum_{j=1}^{i} \frac{j}{i} \binom{2i-j-1}{i-1} p^{i-j} \quad .$$

Let us consider $[B^{i+1}e_0]_k$ for $k \geq 2$.

$$[B^{i+1}e_0]_k$$
$$= \sum_{l=k-1}^{i} pq^{l+1-k} \times$$
$$\left(q^{i-l} \sum_{j=l-1}^{i-1} \frac{j}{i-1} \binom{2(i-1)-j-1}{i-2} p^{i+l-2-j} \right)$$
$$= q^{i+1-k} \sum_{l=k-1}^{i} \sum_{j=l-1}^{i-1} \frac{j}{i-1} \binom{2(i-1)-j-1}{i-2} p^{i+l-1-j} \; .$$

Treating the expression as a polynomial in p, we group terms by the powers of p. The coefficient of $p^\alpha,\ k-1 \leq \alpha \leq i$, for a fixed value of l is

$$\frac{i+l-1-\alpha}{i-1} \binom{2(i-1)-(i+l-1-\alpha)-1}{i-2}$$

For the binomial term to be non-zero,

$$2(i-1)-(i+l-1-\alpha)-1 \geq i-2 \quad \Rightarrow \quad l \leq \alpha$$

The coefficient of p^α is

$$\sum_{l=k-1}^{\alpha} \frac{i+l-1-\alpha}{i-1} \binom{2(i-1)-(i+l-1-\alpha)-1}{i-2}$$
$$= \sum_{j=i+k-2-\alpha}^{i-1} \frac{j}{i-1} \binom{2(i-1)-j-1}{i-2}$$
$$= \frac{i+k-1-\alpha}{i} \binom{2i-(i+k-1-\alpha)-1}{i-1} \quad .$$

The summation in the second last step can be simplified to the above expression using the techniques used in the proof of *1.*

The expression for the entry $[B^{i+1}e_0]_k$ is given by

$$q^{i+1-k} \sum_{\alpha=k-1}^{i} \frac{i+k-1-\alpha}{i} \binom{2i-(i+k-1-\alpha)-1}{i-1} p^\alpha$$
$$= q^{i+1-k} \sum_{j=k-1}^{i} \frac{j}{i} \binom{2i-j-1}{i-1} p^{i+k-1-j} \quad .$$