Using PageRank to Characterize Web Structure

Gopal Pandurangan* Prabhakar Raghavan† Eli Upfal‡

Abstract

Recent work on modeling the Web graph has dwelt on capturing the degree distributions observed on the Web. Pointing out that this represents a heavy reliance on "local" properties of the Web graph, we study the distribution of PageRank values on the Web. Our measurements suggest that PageRank values on the Web follow a power law. We then develop generative models for the Web graph that explain this observation, and moreover remain faithful to previously studied degree distributions. We analyze these models, and compare the analysis to both snapshots from the Web and to graphs generated by simulations on the new models. To our knowledge this represents the first modeling of the Web that goes beyond fitting degree distributions on the Web.

Keywords: Graph Structure, PageRank, Power Law, Web Measurement, Web Models.

*Department of Computer Science, Purdue University, West Lafayette, IN 47907-2066, USA. Corresponding author. E-mail: gopal@cs.purdue.edu; Ph: (765)494-0916; Fax: (765)494-0739.

†Verity Inc., 892 Ross Drive, Sunnyvale, CA 94089, USA. E-mail: pragh@verity.com

‡Department of Computer Science, Brown University, Providence, RI 02912-1910, USA. E-mail: eli@cs.brown.edu

The first and third authors were supported in part by the Air Force and the Defense Advanced Research Projects Agency of the Department of Defense under grant No. F30602-00-2-0599, and by NSF grant CCR-9731477. A preliminary version of this paper appeared in the Proceedings of the 8th Annual International Conference on Combinatorics and Computing (COCOON), Singapore, Lecture Notes in Computer Science 2387, Springer-Verlag, 330-339, 2002.

1 Introduction and Overview

There has been considerable recent work on developing increasingly sophisticated models of the structure of the Web [1, 5, 6, 7, 12, 19, 21, 20, 2, 15, 13, 23]. The primary drivers for such modeling include developing an understanding of the evolution of the Web, better tools for optimizing Web-scale algorithms, mining communities and other structures on the Web, and studying the behavior of content creators on the Web. In a recent paper, Henzinger [18] lists modeling the Web graph as one of the six important algorithmic problems that arise in Web search engines. Prior modeling has dwelt on fitting models to the observed degree distribution of the Web. One of the remarkable properties about the Web graph is that the degree distribution appears to follows a power law with a fixed exponent, regardless of the size of the graph. In fact, various measurement studies have shown [1, 5, 6, 7, 12, 19, 20] that the in-degree and out-degree distribution of the Web graph follow power laws with exponents 2.1 and 2.7 respectively. The power law property of the degree distribution is the key property that most Web models try to capture [1, 5, 6, 7, 12, 19, 21, 20, 2, 15, 13, 23].

While the previous approach to Web modeling is a significant step (both empirically and analytically), a troubling aspect of this approach is the heavy reliance on a single set of parameters – the *degree* distribution. Moreover, since degree distribution is a very "local" property of graphs [1] something that is well recognized from at least two distinct viewpoints: (1) as a ranking mechanism, ordering the Web pages in search results by in-degree (popularity of linkage) is relatively easier to spam (than PageRank [2]); (2) from a graph-theoretic standpoint, it is relatively easy to exhibit "very different" graphs that conform to the same degree distribution [3]. Indeed, the first of these reasons led to the PageRank function [11] used in the Google engine [17]. In this paper we present a more detailed approach to modeling, to explain the distributions of *PageRank* values on the Web. Our model augments the degree distribution approach, so that as a by-product we achieve previous models' success in explaining degree distributions.

We first review related background in Section 1.1; the reader familiar with this material may wish to skip ahead to Section 1.2.

[1] Adding an edge affects the degrees of only the two nodes concerned. This is unlike PageRank, where adding an edge between two nodes can affect the PageRank of many other nodes.

[2] In the sense as mentioned in the previous footnote. That is, it is easy to increase the in-degree of a particular page, but it is relatively more difficult to increase its PageRank (one has to add in-links from relatively high PageRank pages).

1.1 Background and Related Work

We now set the stage for discussing graph models of the Web, beginning with the standard view of the Web as a graph (Section 1.1.1). We next review the basics of the PageRank function [11] reportedly used in the Google search engine [17] (Section 1.1.2).

1.1.1 The Web as a Graph

View the Web as a *directed* graph whose nodes are html pages. Each hyperlink is a directed edge in the natural manner. The *in-degree* of a node is the number of edges (hyperlinks) into it; a simplistic interpretation of the in-degree of a page is as a popularity count. The *out-degree* of a node is the number of links out of it; this is simply the number of `href` tags on the page. The *degree distribution* of a graph is the function of the non-negative integers that specifies, for each $k \geq 0$, what fraction of the pages have degree k; there are naturally two degree distributions for a directed graph, the in-degree distribution and the out-degree distribution.

These distributions have been the objective of considerable prior study [1, 5, 6, 7, 12, 19, 20], on various snapshots of the Web ranging from the Web pages at a particular university to various commercial crawls of the Web. Despite the varying natures of these studies, the in-degree distribution appears to be very well approximated by the function $c_i/k^{2.1}$ where c_i is the appropriate normalization constant (so that the fractions add to one). Likewise, the out-degree distributions seem to be very well approximated by the function $c_o/k^{2.7}$. Such distributions are known as *power law* distributions.

Recent work of Dill et al. [14] provides some explanation for this "self-similar" behavior: that many properties of the Web graph are reflected in sub-domains and other smaller snapshots of the Web. Indeed, this will provide the basis for some of our experiments, in which we derive an understanding of certain properties of the Web by studying a crawl of the `brown.edu` domain. (This methodology was pioneered by Barabasi et al. [5, 6, 7], who extrapolated from the `nd.edu` domain of Notre Dame University. They made a prediction on the diameter of the undirected version of the Web graph, in which one ignores link directions.)

Other properties of the Web graph that have been studied (analytically or empirically) include connectivity [12], clique distributions [19] and diameter [10].

1.1.2 PageRank Primer

The *PageRank* function was presented in [11, 25] and is reportedly used as a ranking mechanism in the commercial search engine Google [17]. It assigns to each Web page a positive real value called its PageRank. In the simplest use of the PageRank values, the documents matching a search query are presented in decreasing order of PageRank. We now briefly

discuss the notion of PageRank and its practical implementation via the *decay* parameter.

The original intuition underlying PageRank was to visualize a random surfer who browsed the Web from page to page. Given the current location (page) q of the surfer, the successor location is a page reached by following a hyperlink out of page q uniformly at random. Thus each hyperlink is followed with probability proportional to the out-degree of q. In this setting, the PageRank of each page is the frequency with which, in the steady state, the page q is visited by such a surfer. Intuitively, the surfer frequently visits "important" pages such as `yahoo.com` because many pages hyperlink to it. Moreover, by calculations from elementary probability theory, the PageRank of a page q is increased if those pages that hyperlink to q have high PageRank themselves.

An immediate difficulty with this notion: some pages, or an (internally) connected cluster of pages may have no hyperlinks out of them, so that the random surfer may get stuck. To address this, Brin and Page [11] introduced the following device: at each step, with some probability, the surfer "teleports" to a completely random Web page, independent of the hyperlinks out of the current page. At least in consideration of the surfing behavior of early users of the Web (from the mid 1990's), such serendipitous teleporting followed by some depth-first exploration (before teleporting again) was reasonable. More important to the notion of PageRank, it removes the technical difficulty created by (connected clusters of) pages having no hyperlinks out of them.

Let the pages on the Web be denoted by $1, 2 \ldots, m$. Let $d_{out}(i)$ denote the number of outgoing links from page i, i.e., the out-degree of i. Let $In(i)$ denote the set of pages that point to i. Let $p (0 < p < 1)$ be the *decay factor* that represents the probability with which the surfer proceeds with the random walk, while $1 - p$ is the probability of teleporting to a random page amongst all m Web pages. Then the PageRank $r(i)$ of page i is given by ([11]):

$$r(i) = \frac{1-p}{m} + p * \sum_{j \in In(i)} \frac{r(j)}{d_{out}(j)}.$$

This represents a system of linear equations (one for each $i \in \{1, 2, \ldots, m\}$). We may rewrite this in matrix form, and the unique solution vector $r(i)$ can be expressed as the eigenvector of a matrix [11, 25] or as the stationary probability of a random walk [24] (thus $\sum_i r(i) = 1$).

While we will not get deeper into the mathematical underpinnings of PageRank here (we refer to [22] for an in-depth survey), it should be intuitively clear that the PageRank values of pages are global properties (in contrast to the more local nature of in-degree). One could in principle concoct examples in which the PageRanks of a few nodes could be "engineered", but fitting the distribution is relatively harder. This observation is one reason why we propose that the PageRank distribution is probably a more important characteristic to model than the degree distribution. Moreover, as we show below, our model captures the PageRank distribution while remaining faithful to the degree distribution.

1.2 Main Contributions and Guided Tour of the Paper

We review graph models in Section 2. We augment the current set of models by proposing a new model – which we call *PageRank-based selection* – in which attachment probabilities for new hyperlinks are based on the PageRanks of existing nodes (this model was also independently proposed in [15], but not motivated by the issues addressed here). The intent in proposing this model is to explain our empirical observations on PageRank distributions, described below. We also present a *hybrid selection model* that is a natural combination of previous models with our PageRank-based selection model.

In Section 3 we describe experiments on snapshots from the Brown University Web, as well as from the publicly available WT10g Web snapshot. Our first finding is that the PageRank distribution appears to follows a power law with exponent 2.1. This is interesting for several reasons: (1) PageRank is distributed as a power law; (2) it has the same exponent (namely, 2.1) as that observed for in-degree on many independent snapshots of the Web; (3) the distribution is (as already known for in- and out-degree distributions) relatively insensitive to the particular snapshot of the Web on which the measurement is made.

Section 4 adopts analytical as well as simulation-based approaches to validating our models and fitting model parameters. We first present heuristic analysis based on the "mean-field" approach [5, 6, 7] that the classical degree-based selection model as well as our new PageRank-based selection yield power laws for the PageRank distribution. The question then is whether the exponents predicted by the analysis match the observations. Given that these are parameterized models, we are able to find combinations of models and parameters that do indeed fit both the PageRank and degree distributions. We verify that these models do generate graphs with the correct distributions through simulations in which we generate multiple random graphs and measure their distributional properties (Section 4.3).

To our knowledge, these are the first results that capture global distributional properties in a model, validating empirical observations through analysis and simulation. Our new models simultaneously capture degree distributions — local properties studied in previous models.

2 Web Graph Models

In the *Erdös-Renyi* model of random graphs [8], each edge is directed from a node to another node that is chosen uniformly at random from all the other nodes in the graph. There is a wealth of research on such graphs, and many properties of such random graphs are well understood. For instance, an Erdös-Renyi random graph in which the average out-degree of each node is roughly 7 (as is the observed average out-degree of Web pages), the degree distributions are Poisson, and it is unlikely that there are any clique-like structures with

more than a handful of nodes. Given the many consistent observations of power law degrees on the Web graph, as well as the superabundance of clique-like structures [20], it is clear that the Web graph does not conform to the Erdös-Renyi model. Nevertheless, as we will see below, elements of random selection do play a role in models that are more faithful to the Web graph.

A number of research projects proceeded to develop models that better explained the power law behavior of degree distributions on the Web; see [26] for a survey of these. In all of these, the view is that of nodes and edges being added to the graph one at a time. As noted above, it does not suffice for such newly arriving edges to choose to point to a node (page) chosen uniformly at random, since this does not yield a power law distribution for degrees. The simplest model to overcome this problem uses the following device: each edge chooses the node to point to at random, but with non-uniform probabilities for choosing the various nodes. In particular, the edge points to a node q in proportion to the current in-degree of q. This yields Web graphs whose in-degree distributions have been shown to converge to the distribution $\approx 1/k^2$ [5, 6, 7].

However, as noted earlier, empirical studies have shown that in-degrees are in fact distributed as $\approx 1/k^{2.1}$ (rather than $1/k^2$). To help explain the exponent of 2.1, Kumar *et al.* [21] introduced the following more detailed process by which each edge chooses the node to point to. Some fraction of the time (a parameter they call $\alpha \in [0, 1]$) the edge points to a node chosen uniformly at random. The rest of the time (a fraction $1 - \alpha$), the edge picks an intermediate node v at random, and *copies* the destination of a random edge out of v. In other words, the new edge points to the destination of an edge e, chosen at random from the outgoing edges of a random node v. They then explain a number of empirical observations on the Web graph including the in-degree exponent of 2.1 and the large number of clique-like structures observed by [20]. In fact, they prove theorems that derive the exponent as a function of the parameter α. There is another way of viewing this model: a fraction α of the edges go to random nodes, while the remainder choose destination nodes in proportion to their current degrees. (A similar model incorporating both preferential and random components is also proposed in [23].) Thus, their model may be viewed as a generalization of the models of Barabasi and others, parameterized by α. We will henceforth refer to this model as the *degree-based selection model*. Could it be that this model would also explain the PageRank distributions we observe on the Web?

Before we address this question, we next introduce a new model inspired by the α model above. Suppose that each edge chose its destination at random a fraction $\beta \in [0, 1]$ of the time, and the rest of the time chose a destination in proportion to its *PageRank*. We will call this the *PageRank-based selection model*.

However, this now raises the question: if we could develop a model that explained ob-

served PageRank distributions, could it be that we lose the ability to capture observed degree distributions? To address this, we now present the most general model we will study. There are two parameters $a, b \in [0, 1]$ such that $a + b \leq 1$. With probability a an edge points to a page in proportion to its in-degree. With probability b it points to a page in proportion to its PageRank. With the remaining probability $1 - a - b$, it points to a page chosen uniformly at random from all pages. We thus have a family of models; using these 2-parameter models we can hope to simultaneously capture the two distributions we investigate – the PageRank distribution (representing global properties of the graph), and the in-degree distribution (representing local properties of the graph). We will call this the *hybrid selection model*.

3 Experiments

To set the context for exploring the models in Section 2, we study the distribution of PageRanks (as well as of the in- and out-degrees) on several snapshots of the Web.

3.1 Experiments on the Brown University Domain

Our first set of experiments was on the Web graph underlying the Brown University domain (`*.brown.edu`). Our approach is motivated by recent results on the "self-similar" nature of the Web (e.g., [14]): a thematically unified region (like a large subdomain) displays the same characteristics as the Web at large. The Brown Web consisted of a little over 100,000 pages (and nearly 700,000 hyperlinks) with an average in-degree (and thus out-degree) of around 7. This is very close to the average in-degree reported in large crawls of the Web [20]. Our crawl started at the Brown University homepage (`www.brown.edu` – "root" page) and proceeded in breadth-first fashion; any URL outside the `*.brown.edu` domain was ignored. We did prune our crawl – for example, URL's with /cgi-bin/ were not explored.

The graphs shown in Figures 1, 2 and 3 summarize our results on the in-degree, out-degree and the PageRank distributions in the Brown Web graph[3]. Our experiments show that the in-degree and out-degree distribution follows a power law with exponent 2.1 and 2.7 respectively. This is strikingly similar to the results reported on far larger crawls of the Web [12, 20]. For example [12] report exactly the same power law exponents on a crawl of over 200 million pages and 1.5 billion hyperlinks.

However, the most interesting result of our study was that of the PageRank distribution. We first describe our PageRank computation. As in [25], we first pre-process pages which do not have any hyperlinks out of them (i.e., pages with out-degree 0): [4] we assume that

[3]To avoid excessively "dark" plots resulting in large amounts of redundant data, all plots in this paper have been sub-sampled.

[4]Such a page with out-degree 0 is called a *rank leak* [25].

these have links back to the pages that point to them [4]. In our PageRank computation we set the decay parameter to 0.9; this is a typical value reportedly used in practice (e.g., [11] uses 0.85), and the convergence is fast (under 20 iterations). Similar fast convergence is reported in [25, 11]. However, varying the decay parameter does not significantly change our results, as long as the parameter is fairly close to 1. In particular, we get essentially the same results for decay parameter values down to 0.8.

The main result of our PageRank distribution plot is that a large majority of pages (except those with very small PageRank) follow power law with an exponent close to 2.1. That is, the fraction of nodes having PageRank r is proportional to $1/r^{2.1}$. In Section 4 we will give an analysis suggesting this PageRank distribution, based on various models from Section 2.

We also note that the distribution is almost flat for pages with very low PageRank. To check whether this is an anomaly, we repeated the experiments for the Brown Computer Science department subdomain (`*.cs.brown.edu`) and we got almost identical results (i.e., in-degree, out-degree and PageRank distributions follow power laws with almost identical exponents) even though `*.cs.brown.edu` is a much smaller graph (around 25,000 nodes); and a similar flattening at the top (corresponding to pages with very low PageRank) in the PageRank distribution. Comparing this pattern to the experiments on the WT10g corpus (next subsection) that captures a more generic subset of the Web, suggests that relatively structured domains, such as `brown.edu` and `cs.brown.edu`, have a smaller fraction of very "unimportant" pages than predicted by the power law distribution and observed in less structured corpora.

3.2 Experiments on WT10g Data

We repeated our experiments on the WT10g corpus [27], a 1.69 million document testbed for conducting Web experiments. The results are almost identical to those on the Brown Web; the in-degree, out-degree, and PageRank distributions follow power laws with exponent close to 2.1, 2.7 and 2.1 respectively. Figure 4 shows the plot of PageRank distribution of the WT10g corpus (we are not showing the in-degree and out-degree distribution plots as they are very similar to those of the Brown Web). The power law here appears much sharper than in the Brown Web. As noted above, a possible explanation is that unlike the Brown domain, the WT10g corpus is constructed by a careful selection of Web pages so as to characterize the *whole Web* [27].

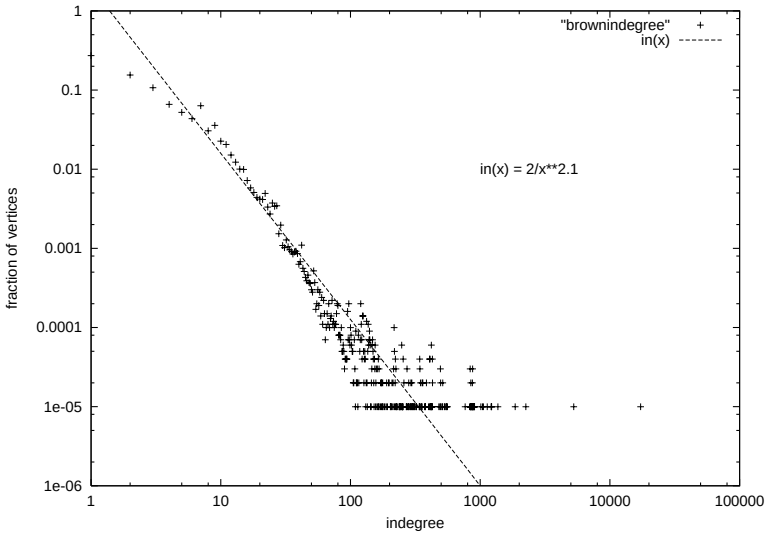

Figure 1: Log-log plot of the in-degree distribution of the Brown domain (`*.brown.edu`). The in-degree distribution follows a power law with exponent close to 2.1.

4 Fitting the Models: Analysis and Simulations

In this section we address some of the modeling questions raised in Section 2. Having obtained the empirical distributions in Section 3, we first give analytical predictions of the shape of the PageRank distributions for the degree-based and PageRank-based selection models of Section 2. The intent is to infer what choices of these model parameters would give rise to the distributions observed in our experiments. Finally, in Section 4.3 we generate random graphs according to these fitted models, to see if in fact they give rise to graphs that match the distributions observed on the Web.

4.1 Degree-Based Selection

Consider a graph evolving in a sequence of *time steps* – as noted in Section 2 such evolution is not only realistic in the context of the Web, it is also a feature of all Web graph models. A single node with r outgoing edges is added at every time step. (We assume that we start with a single node with a self-loop at time 0 [9]).) Each edge chooses its destination node independently with probability proportional to 1+in-degree[5] of each possible destination node. That is $\alpha = 0$, and attachment is solely based on degree with no random component. This model is essentially the one analyzed by Barabasi *et al.* and is a special case of the α model of Kumar *et al.*

Let $\pi^t(v)$ represent the PageRank of v at time step t. We can interpret the PageRank

[5]We assume that each incoming node has "weight" 1, else all growth is trivial.

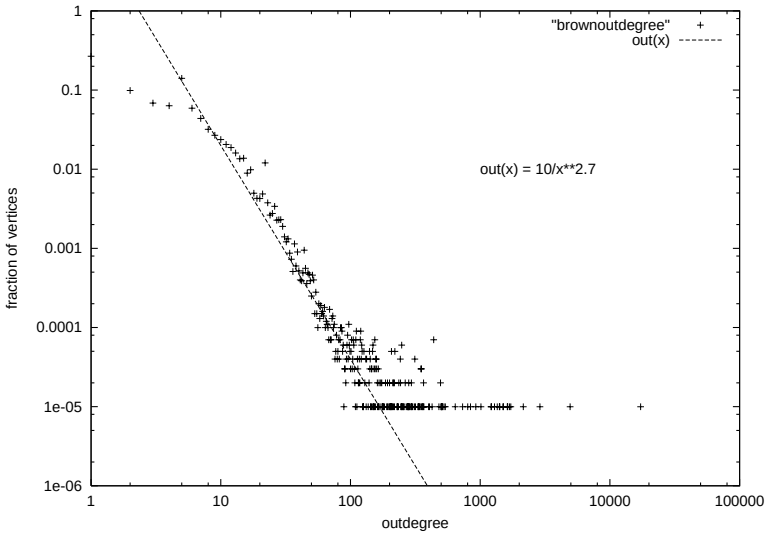

Figure 2: Log-log plot of the out-degree distribution of the Brown domain (`*.brown.edu`). The out-degree distribution follows a power law with exponent close to 2.7.

as the stationary probability of a random walk on the underlying graph, with the teleport operation (Section 1.1.2) being modeled by a "central" node c (see Figure 5). At each step, the surfer either decides to continue his random walk with probability p or chooses to return to the central node with probability $1-p$; from the central node he jumps to a random node in the graph. To write an expression for $\pi^t(v)$ it is useful to define $f^t(v)$, the "span" of v at time t: the *sum* of the in-degrees of all nodes in the network (including v itself) that have a path to v that does not use the central node (we also refer to the nodes contributing to the span as "span nodes"). Since each edge contributes a $1/r$ fraction (as mentioned earlier, each node has r outgoing edges) of the stationary probability of its source node we can bound $\pi^t(v)$ for the above random walk on a Markov chain by using the standard stationary equations [6] as follows:

$$\frac{f^t(v)\pi(c)p^H}{rt} \leq \pi^t(v) \leq \frac{f^t(v)\pi(c)}{rt} \tag{1}$$

where $\pi(c)$ is the stationary probability of the central node and H is the longest (directed) path in the network (ignoring link directions and the central node) [7]. For the upper bound we use the fact that the each edge in the span of v contributes $\pi(c)/rt$ to the stationary probability of v, ignoring the probability of jumping to the central vertex and assuming that

[6] The stationary probabilities π_j's of a irreducible, finite, and aperiodic Markov chain satisfy the equations: $\pi_j = \sum_k \pi_k P_{k,j}$ for all j (see e.g., [24]) where P is the transition probability matrix. In other words, every node k "contributes" P_{kj} fraction of its stationary probability to π_j. It is easy to see that the Markov chain induced by the directed graph considered here (Figure 5) is irreducible, finite, and aperiodic.

[7] Our model evolves essentially as a directed acyclic graph (DAG) with self loops and H is the height of the DAG.

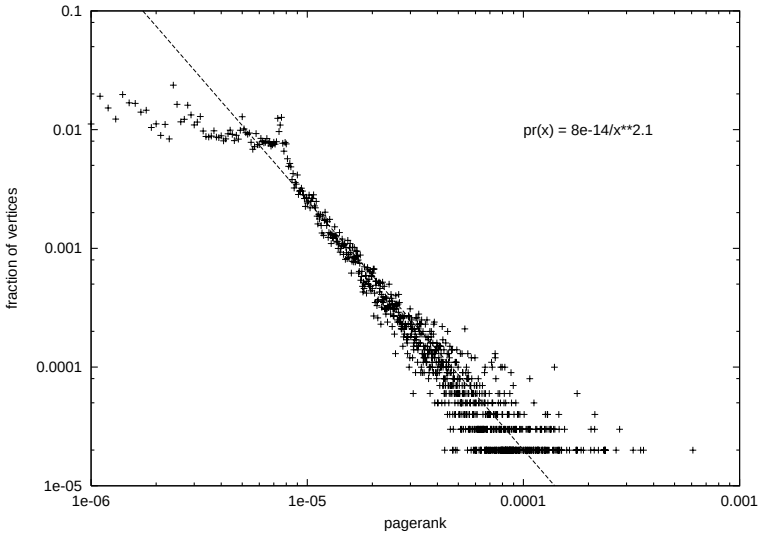

Figure 3: Log-log plot of the PageRank distribution of the Brown domain (`*.brown.edu`). A vast majority of the pages (except those with very low PageRank) follow a power law with exponent close to 2.1. The plot almost flattens out for pages with very low PageRank.

the entire contribution goes to $\pi^t(v)$; for the lower bound we do not ignore the probability of jumping, and hence the contribution of each edge goes down by a factor of p^H (in the worst case). We note two facts here. First, a simple calculation shows that $\pi(c) = (1-p)/(2-p)$, a constant independent of t; second, it can be shown that when t is sufficiently large, H at time t is at most logarithmic in the size of the graph (which is t) [10]. Thus if the decay factor is close to 1 [8] we can approximate $\pi^t(v)$ as

$$\pi^t(v) \approx \frac{f^t(v)\pi(c)}{rt} \tag{2}$$

We now proceed to calculate $f^t(v)$. We use the *mean-field* approach of Barabasi *et al.* [7] as follows. Assuming $f^t(v)$ to be continuous, we can write the differential equation for the rate of change of $f^t(v)$ (the span of v) with time:

$$\frac{d(f^t(v))}{dt} = \frac{f^t(v)}{t} \tag{3}$$

where the right hand side denotes the probability that an incoming edge connects to one of the span nodes of v (this also represents the rate of change of $f^t(v)$ since an incoming node

[8] Actually, we will obtain the same power law result, even if we do not assume this. For example, if we just assume that the decay factor is a constant, we will get a factor of $p^H \approx p^{\log t}$ and this will only affect the exponent of t in the denominator of Equation 2 and, as the reader can verify, it will not change the final power law result in Equation 6. We note that, however, the power law crucially depends on the (differential) Equation 3.

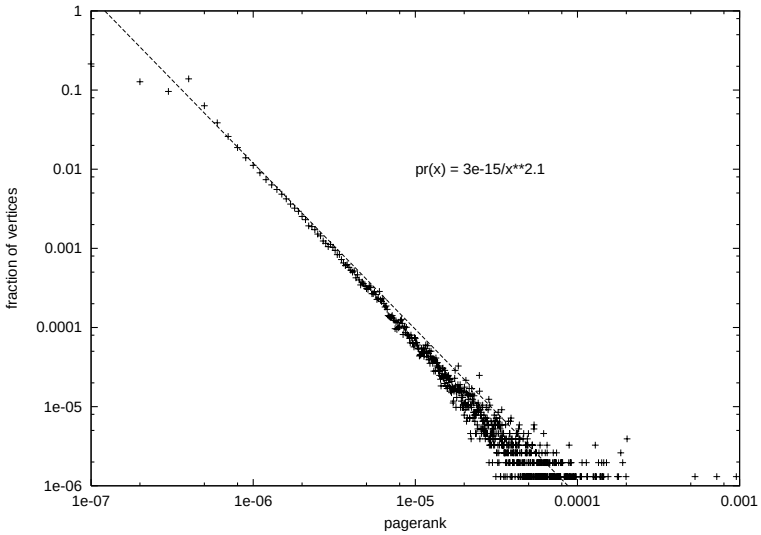

Figure 4: Log-log plot of the PageRank distribution of the WT10g corpus. The slope is close to 2.1. Note that the plot looks much sharper than the corresponding plot for the Brown Web. Also, the tapering at the top is much less pronounced.

connects with probability proportional to degree; if the incoming node connects to any span node of v it will increase the span of v). The solution to differential equation 3 with the initial condition that node v was added at time t_v is

$$f^t(v) = \frac{t}{t_v} \tag{4}$$

Combining Equations (2) and (4), we have

$$\pi^t(v) \approx \frac{\pi(c)}{rt_v} \tag{5}$$

Using the above equation,

$$\Pr(\pi^t(v) < \phi) = \Pr(t_v > \frac{\pi(c)}{r\phi})$$

Since nodes are added at equal time intervals, the probability density of t_v is $1/t$. Thus we obtain

$$\Pr(t_v > \frac{\pi(c)}{r\phi}) = 1 - \Pr(t_v \leq \frac{\pi(c)}{r\phi}) = 1 - \frac{\pi(c)}{rt\phi}$$

which yields that the probability density function F for $\pi^t(v)$ is:

$$F(\phi) = \frac{\partial(\Pr(\pi^t(v) < \phi))}{\partial\phi} \approx \frac{\pi(c)}{rt\phi^2} \tag{6}$$

implying that the PageRank follows a power law with exponent 2, independent of r and t. Simulations of this model (Figure 6) agree well with this prediction.

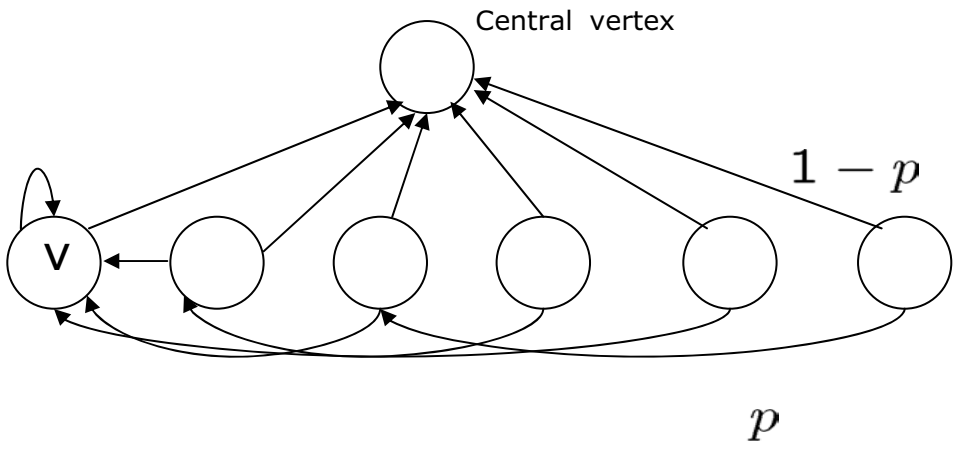

Figure 5: Random walk on the Web graph. From each of the (bottom) vertices the surfer decides to continue his random walk with probability p, or decides to go to the central vertex with probability $1 - p$. From the central vertex he jumps to a random vertex. The quantity $f^t(v)$ (the span of vertex v) is 6 in the above graph: it is the sum of the in-degrees of nodes which have a path to v including v itself.

As already mentioned in Section 2 the in-degree distribution of this model follows a power law with exponent 2, the same as the PageRank distribution derived above. This is striking given that in our empirical studies too, the in-degree and PageRank distributions had identical power laws. However, the empirically observed power laws have exponents of 2.1; thus the degree-based selection model does not quite match the in-degree and PageRank exponents observed in practice. Now a natural question is whether we can make it match both the distributions by changing α, i.e., by incorporating a random selection component in choosing nodes. The answer is yes; more on this in Section 4.3 below. But first we analyze PageRank-based selection.

4.2 PageRank-Based Selection

We show that power law emerges for the PageRank and degree distributions in this model as well, but the exponents are different from the degree-based model.

Our model and analysis is analogous to Section 4.1. A single node with r outgoing edges

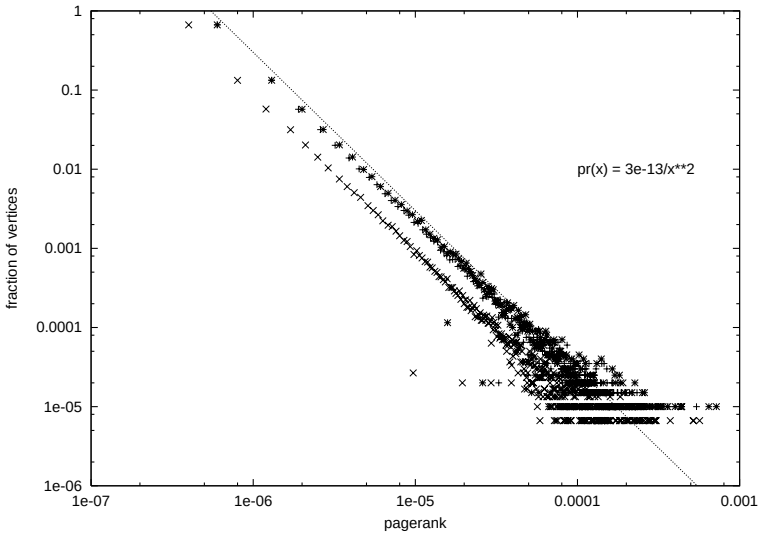

Figure 6: Log-log plot of degree-based selection with $\alpha = 0$. The number of nodes shown is 300,000 (+), 200,000 (*) and 100,000 (x). It clearly shows that the slope is 2, confirming the power law predicted by analysis.

is added at every time step, as before. Each edge chooses its destination node independently with probability proportional to the PageRank of each possible destination node. That is, $\beta = 0$, and attachment is solely based on PageRank with no random component (i.e., a incoming node chooses to connect with probability proportional to PageRank only). Using the same notations and arguments as before, we can show that Equation (2) holds. However, $f^t(v)$ follows a different differential equation from Equation (3). Instead we have

$$\frac{d(f^t(v))}{dt} \approx \frac{f^t(v)r}{2rt} \tag{7}$$

The reasoning is as follows. The probability that $f^t(v)$ increases by one is the probability that the incoming node chooses any one of the nodes in the span to connect to, which is proportional to the sum of the PageRanks of all the span nodes of v. To calculate this probability, we see that each directed edge contributes nearly *twice* to the sum (if p is sufficiently large) and the total PageRank is thus proportional to the sum of the degrees which is $2rt$.

The solution of the above differential equation with the initial condition that node v was added at time t_v is

$$f^t(v) = (\frac{t}{t_v})^{1/2} \tag{8}$$

The rest of the analysis is similar to the degree-based analysis. Combining Equations (2) and (8), we have

$$\pi^t(v) \approx \frac{\pi(c)}{r(tt_v)^{1/2}} \tag{9}$$

Using the above equation,

$$\Pr(\pi^t(v) < \phi) = \Pr(t_v > \frac{(\pi(c))^2}{r^2 t \phi^2})$$

Since nodes are added at equal time intervals, the probability density of t_v is $1/t$. Thus we obtain

$$\Pr(t_v > \frac{(\pi(c))^2}{r^2 t \phi^2}) = 1 - \frac{(\pi(c))^2}{r^2 t^2 \phi^2}$$

which yields that the probability density function F for $\pi^t(v)$ is:

$$F(\phi) = \frac{\partial(\Pr(\pi^t(v) < \phi))}{\partial \phi} \approx \frac{2(\pi(c))^2}{r^2 t^2 \phi^3} \tag{10}$$

i.e., predicting that the PageRank follows a power law with exponent 3. Analogously, we can show that the degree also a follows a power law with exponent 3. Simulations agree quite well with this prediction.

Thus, the PageRank-based selection model with $\beta = 0$ does not match the empirically observed in-degree and PageRank exponents. Can we hope to match the observations by varying β? Unlike the degree-based selection model, the answer is no; increasing β will only increase the power law exponent (above 3) for the in-degree distribution. This can be verified by experiments, as well as by a direct extension to the analysis above. We are thus left with the degree-based selection model and the hybrid selection model of Section 2 as candidates for explaining the observations.

4.3 Simulations of the Generative Models

An accurate model of the Web graph must conform with the experimentally observed in-degree, out-degree, and PageRank distributions. We simulated the degree-based and hybrid selection models defined in Section 2 under various parameters to find settings that generate the observed empirical distributions. We simulated graphs of size up to 300,000 nodes, and we varied the average number of new edges generated per new node generation (time step). In particular, to be "close" to the real Web's average out-degree (and in-degree), we focused on the range in which the average number of edges added per new node is around 7. We obtained essentially the same results for the power laws, irrespective of the size (from 10,000 nodes onwards) or the number of outgoing edges.

Our first step was fitting the out-degree distribution. Following Kumar *et al.*, we use the degree-based copying model with a suitable value of β to fit the out-degree distribution to a power law with exponent 2.7. At each time step, the incoming node receives edges from existing nodes. With probability β a node is chosen uniformly at random, with probability $1 - \beta$ the node is chosen proportional to the current out-degree distribution. Note that the

out-degree distribution is fixed independently of the in-degree distribution. We use $\beta = 0.45$ to get a power law exponent equal to 2.7.

We turn now to the problem of fitting the in-degree distribution. We first simulated the degree-based selection model. Setting $\alpha = 0$ (or $\beta = 0$), both the in-degree and PageRank distributions followed a power law with exponent 2. We observed that increasing α increases the exponents in the in-degree and PageRank distributions. In particular, setting $\alpha \approx 0.2$ brings both exponents to the empirical value of 2.1. This value is unique; by increasing or decreasing α we lose the fit. Thus, we found a setting of the parameters for which the degree-based selection model simultaneously fits all the three distributions.

Since degree-based selection model fits the empirical data, a natural question is whether PageRank-based selection is irrelevant in modeling the Web graph. To answer this, we experimented with the 2-parameter hybrid selection model proposed in Section 2. Surprisingly when $a = b \approx 0.33$, we could again simultaneously fit all three distributions. Thus, interestingly, we have an alternative model, with a substantial PageRank-based selection component, that fits the Web empirical data.

5 Conclusion and Further Work

We study the PageRank distribution on the Web graph, and use it to develop more accurate generative models for the evolution of the Web graph. Our first finding is that PageRank distribution on snapshots of the Web graph follows a power law distribution with the same exponent as the in-degree distribution. Unlike in-degree, PageRank is a global property of the graph, thus one expects to obtain more accurate modeling of the Web graph by fitting the models to the PageRank distribution.

Our study of PageRank distributions can also be of independent interest for Web search and ranking pages. For search engines employing PageRank and associated ranking schemes, it is important to understand whether, for instance, 99% of the total PageRank is concentrated in (say) 10% of the pages. This (especially in conjunction with query distribution logs) can have implications for compressing inverted indices and optimizing the available storage. Further work is needed to explore this.

We consider three possible models for the Web graph: degree-based selection model, PageRank-based selection model, and a hybrid model. Our analysis shows that the PageRank-based selection model cannot fit the empirical data. For the two other models we found settings of parameters under which the model fits simultaneously the in-degree and out-degree distributions and the PageRank distribution. A natural question for further study is whether one of these models describes the Web better than the other.

Another interesting question that arises from our work is whether PageRank is strongly

correlated with in-degree. We note that PageRank and in-degree follow power laws with almost identical exponents. Could it be that PageRank is highly correlated to in-degree, and thus the computational overhead (and ranking magic) of PageRank boils down to a simple popularity count by in-degree? Clearly one can concoct graphs for which the PageRank and degree distributions are highly correlated, just as one can concoct graphs for which they are not – but what happens on the true Web? Our preliminary experiments show not much correlation between the two properties on the Web graph as a whole. In general, a high in-degree of a node does not imply high PageRank and vice versa.

All models proposed and analyzed so far grow by making "global" choices: connections are chosen by various distributions, but from *all* the existing nodes. In practice, links between nodes cannot be fully explained just by the relative popularity of the nodes. While nodes are likely to link to important or popular nodes, these nodes are also likely to be in the same sub-community. Thus another challenging question is extending these simple models to capture the important notion of communities and sub-communities on the Web.

There are also important random-graph theoretic questions that remain to be solved: how to characterize precisely the PageRank distribution in the models proposed here. Our analysis is only approximate and it seems to work only for simplified versions of the degree-based and PageRank-based models. A key technical difficulty here seems to be in analyzing the stationary distribution of a dynamically changing directed graph.

Acknowledgments. We thank the anonymous referees for their useful comments and suggestions for improving the presentation of the paper. We are grateful to Joel Young for providing us with his Web crawler and for many hours of help.

References

[1] L. Adamic and B. Huberman. Power Law distribution of the World Wide Web, Technical Comment on [5], *Science*, **287**, 2000, 2115a.

[2] W. Aiello, F. Chung-Graham and L. Lu. Random evolution of massive graphs, *Handbook on Massive Data Sets, (Eds. James Abello et al.)*, Kluwer Academic Publishers, (2002), 97-122.

[3] W. Aiello, F. Chung-Graham, and L. Lu. A random graph model for massive graphs, in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* 2000, 171-180. Journal version: A random graph model for power law graphs, *Experimental Math.*, **10**, 2001, 53-66.

[4] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, **1**(1), 2001, 2-43.

[5] A. Barabasi and R. Albert. Emergence of Scaling in Random Networks, *Science* , **286**(509), 1999.

[6] A. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the World Wide Web, *Physica A*, **281**, 2000, 69-77.

[7] A. Barabasi, R.Albert and H. Jeong. Mean-field theory for scale-free random graphs, *Physica A*, **272**, 1999, 173-187.

[8] B. Bollobas. *Random Graphs*, Cambridge University Press, 2001.

[9] B. Bollobas, O. Riordan, J. Spencer, and G. Tusnady. The degree sequence of a scale-free random graph process, *Random Structures and Algorithms*, **18**(3), 2001, 279-290.

[10] B.Bollobas and O. Riordan. The diameter of a scale-free random graph, *Combinatorica*, **4**, 2004, 5-34.

[11] S. Brin and L. Page. The anatomy of a large-scale hypertexual Web search engine, In *Proceedings of the 7th WWW conference*, 1998.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, Andrew Tomkins, J. Weiner. Graph Structure in the Web, *Computer Networks and ISDN Systems*, **30**, 2000, 309-320.

[13] C. Cooper and A. Frieze. A general model of Web graphs, *Random Structures and Algorithms*, **22**, 2002, 311-335.

[14] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-Similarity in the Web, In *Proceedings of the 27th International Conference on Very Large Databases (VLDB)*, 2001.

[15] E. Drinea, M. Enachescu, and M. Mitzenmacher. Variations on Random Graph Models for the Web, *Harvard Technical Report TR-06-01*, 2001.

[16] D. Gibson, J.M. Kleinberg and P. Raghavan. Inferring Web communities from link topology, In *Proceedings of the ACM Symposium on Hypertext and Hypermedia*, 1998.

[17] Google Technology Overview. `http://www.google.com/intl/en/corporate/tech.html`

[18] M. Henzinger. Algorithmic Challenges in Web Search Engines, *Internet Mathematics*, **1**(1), 2003, 115-126.

[19] J. Kleinberg, S. Ravi Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. The Web as a graph: measurements, models and methods, In *Proceedings of the 5th Annual International Computing and Combinatorics Conference (COCOON)*, 1999.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for Emerging Cyber-Communities, In *Proceedings of the 8th WWW Conference*, 1999, 403-416.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic Models for the Web, In *Proceedings of the 41st Annual Symposium on the Foundations of Computer Science*, 2000.

[22] A. Langville and C. Meyer. Deeper Inside PageRank, *Internet Mathematics*, **1**(3), 2003, 335-380.

[23] M. Levene, T.I. Fenner, G. Loizou, and R. Wheeldon. A Stochastic Model for the Evolution of the Web, *Computer Networks and ISDN Systems*, **39**, 2002, 277-287.

[24] R. Motwani and P. Raghavan. *Randomized Algorithms*, Cambridge University Press, 1995.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing order to the Web, *Technical Report*, Computer Science Department, Stanford University, 1998.

[26] C.H. Papadimitriou. *Lecture Notes*, UC Berkeley, Available at `http://www.cs.berkeley.edu/ christos/games/powerlaw.ps`

[27] WT10g collection draft paper, `http://www.ted.cmis.csiro.au/TRECWeb/wt10ginfo.ps.gz`