

# Using PageRank to Characterize Web Structure

Gopal Pandurangan<sup>1</sup>, Prabhakar Raghavan<sup>2</sup>, and Eli Upfal<sup>1</sup>

<sup>1</sup> Computer Science Department, Brown University  
Box 1910, Providence, RI 02912-1910, USA  
{gopal,eli}@cs.brown.edu\*

<sup>2</sup> Verity Inc., 892 Ross Drive, Sunnyvale, CA 94089, USA  
pragh@verity.com

**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 (used in the Google search engine) on the Web. This distribution is of independent interest in optimizing search indices and storage. We show that PageRank values on the Web follow a power law. We then develop detailed 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 analyses 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.

## 1 Introduction

There has been considerable recent work on developing increasingly sophisticated models of the structure of the Web [1, 3, 4, 9, 13, 14]. 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. Prior modeling has dwelt on fitting models to the observed degree distribution of the Web. While this represents a significant step (both empirically and analytically), a weakness of this approach is the heavy reliance on a single set of parameters – the degree distribution. Moreover, the degree distribution is a very “local” property of graphs, 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 very easy to spam and thus not reliable; (2) from a graph-theoretic standpoint, it is easy to exhibit “very different” graphs that conform to the same degree distribution. Indeed, the first of these reasons led to the PageRank function [8] used in the Google engine.

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 de-

---

\* Supported in part by NSF grant CCR-9731477 and NSF ITR grant CCR-0121154.

gree distribution approach, so that as a by-product we achieve previous models' success in explaining degree distributions.

Our study of PageRank distributions is also 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) has implications for compressing inverted indices and optimizing the available storage.

## 2 Background and Related Work

**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 objects of considerable prior study [1, 3, 4, 9, 13, 14], 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  $c/k^{2.1}$  where  $c$  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. [10] 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. [3, 4], 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 [9], clique distributions [13] and diameter [7].

**PageRank.** The *PageRank* function was presented in [8, 17] and is reportedly used as a ranking mechanism in the commercial search engine Google [12]. 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.

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 [8] introduced a *decay* parameter  $p$ : at each step, with probability  $p$  the surfer proceeds with the random walk, and with probability  $1 - p$ , the surfer “teleports” to a completely random Web page, independent of the hyperlinks out of the current page. We refer to [8, 17] for details on the mathematics of PageRank and its practical implementation using the decay parameter.

### 3 Web Graph Models

The classical random graph models of *Erdős-Renyi* [5] do not explain the power law properties of the degree distribution nor the the superabundance of clique-like structures [14] in the Web graph. Thus, it is clear that the Web graph does not conform to the Erdős-Renyi model. One of the first models to explain the power law property was proposed by Barabasi *et al.* which has two key features: (1) nodes and edges are added to the graph one at a time (*uniform growth*) and (2) each incoming node chooses to connect to a node  $q$  in proportion to the current in-degree of  $q$  (*preferential attachment*). This model yields Web graphs whose in-degree distributions have been shown to converge to the distribution  $\approx 1/k^2$  [3, 4].

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.* [15] 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  $v$ . Kumar *et al.* offer the following behavioral explanation for this process: some fraction of the time a content creator creating a page refers to a random new topic and thus creates a link (edge) to a random destination. The remainder of the time, the content creator copies a hyperlink off an existing page (in this case  $v$ ), having decided that this is an interesting link. They then explain a number of empirical observations on the Web graph includ-

ing the in-degree exponent of 2.1 and the large number of clique-like structures observed by [14]. Their model can be viewed as a generalization of the models of Barabasi and others, parameterized by  $\alpha$ . 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  $\alpha$  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*. Following the behavioral motivation of Kumar *et al.*, this can be thought of as a content-creator who chooses to link to random pages some fraction of the time, and to pages highly rated by a PageRank-based engine such as Google the remainder of the time. In other words, content creators are more likely to link to pages that score high on PageRank-based search results, because these pages are easy to discover and link to. This is not implausible from the behavioral standpoint, and could help capture the PageRank distributions we observe (just as in-degree based linking helped explain in-degree distributions in prior work). We will call this the *PageRank-based selection model*.

However, this now raises the question: if we could develop a model that explained observed 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*.

## 4 Experiments

To set the context for exploring the models in Section 3, we study the distribution of PageRanks (as well as of the in- and out-degrees) on several snapshots of the Web.

**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., [10]): 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 [14]. 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.

Our experiments show that the in-degree and out-degree distribution follows a power law with exponent 2.1 and 2.7 respectively. The plots are strikingly similar to the ones reported on far larger crawls of the Web (see [9, 14]). For example [9] report exactly the same power law exponents on a crawl of over 200 million pages and 1.5 billion hyperlinks.

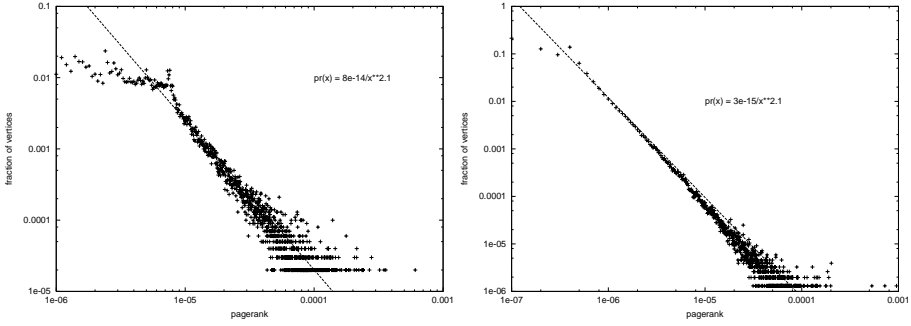
However, the most interesting result of our study was that of the PageRank distribution. We first describe our PageRank computation. As in [17], we first pre-process pages which do not have any hyperlinks out of them (i.e., pages with out-degree 0): we assume that these have links back to the pages that point to them [2]. This is intuitively more justifiable than just dropping these pages: we expect surfers to trace back their trail when they reach a dead end. In our PageRank computation we set the decay parameter to 0.9; this is a typical value reportedly used in practice (e.g., [8] uses 0.85), and the convergence is fast (under 20 iterations). Similar fast convergence is reported in [8, 17]. 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 (Figure 1) is that a large majority of pages (except those with very small PageRank) follow a power law with an exponent close to 2.1. That is, the fraction of nodes having PageRank  $r$  is proportional to  $1/r^{2.1}$ . This appears to be the same as the in-degree exponent; more on this later. In Section 5 we will give an analysis suggesting this PageRank distribution, based on various models from Section 3.

**WT10g data.** We repeated our experiments on the WT10g corpus [18], a recently released, 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 1 shows the plot of PageRank distribution of the wt10g corpus. The power law here appears much sharper than in the Brown Web. Also, unlike the Brown Web, the plot has slope 2.1 across almost the entire spectrum of PageRank values, except for those with very low PageRank values; 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* [18].

## 5 Fitting the Models: Analysis and Simulations

In this section we address some of the modeling questions raised in Section 3. Having obtained the empirical distributions in Section 4, we first give analytical predictions of the shape of the PageRank distributions for the degree-based and PageRank-based selection models of Section 3. The intent is to infer what choices of these model parameters would give rise to the distributions observed in our experiments. Finally, in Section 5.3 we generate random graphs according to



**Fig. 1.** Log-log plot of the PageRank distribution of the Brown domain (left) and the WT10g (right). A vast majority of the pages (except those with very low PageRank) follow a power law with exponent close to 2.1.

these fitted models, to see if in fact they give rise to graphs that match the distributions observed on the Web.

### 5.1 Degree-Based Selection

Consider a graph evolving in a sequence of *time steps* – as noted in Section 3 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 [6].) Each edge chooses its destination node independently with probability proportional to  $1+\text{in-degree}^1$  of each possible destination node. This model is essentially the one analyzed by Barabasi *et al.* and is a special case of the  $\alpha$  model (where  $\alpha = 0$ ) of Kumar *et al.*

Let  $\pi^t(v)$  represent the PageRank of  $v$  at time step  $t$ . We can interpret the PageRank as the stationary probability of a random walk on the underlying graph, with the teleport operation (Section 2) being modeled by a “central” node  $c$ . 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 of the stationary probability of its source node (using the standard stationary equations (see [16])), we can bound  $\pi^t(v)$  for the above random walk as follows:

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

<sup>1</sup> We assume that each incoming node has “weight” 1, otherwise there won’t be any non-trivial growth.

where  $\pi(c)$  is the stationary probability of the central node and  $D$  is the diameter of the network (ignoring link directions). We note two facts here. First, a simple observation shows that  $\pi(c)$  is a constant, independent of  $t$ ; second, it can be shown that when  $t$  is sufficiently large, the diameter of the graph at time  $t$  is logarithmic in the size of the graph (which is  $t$ ) [7]. Thus if the decay factor  $p$  is sufficiently close to 1, we can approximate  $\pi^t(v)$  as

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

We can estimate  $f^t(v)$ , using the “mean-field” approach of Barabasi *et al.* [4]. Treating  $f^t(v)$  as continuous, we can write the differential equation for the rate of change of  $f^t(v)$  with time:  $\frac{d(f^t(v))}{dt} = \frac{f^t(v)}{t}$ , where the right hand side denotes the probability that an incoming edge connects to one of the span nodes of  $v$ . The solution to the above equation with the initial condition that node  $v$  was added at time  $t_v$  is  $f^t(v) = t/t_v$ . Using this in equation (2), and assuming that nodes are added at equal time intervals, we can show that the probability density function  $F$  for  $\pi^t(v)$  is:  $F(\phi) \approx \pi(c)/rt\phi^2$ , implying that the PageRank follows a power law with exponent 2, independent of  $r$  and  $t$ . Simulations of this model (shown in Figure 2) agree well with this prediction.

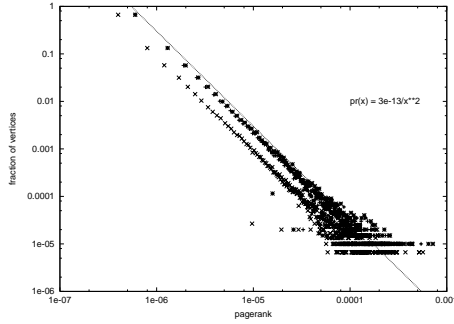
As already mentioned in Section 3, the in-degree distribution of this model follows a power law with exponent 2, the same as the PageRank distribution derived above. However, the empirically observed power laws of both PageRank and in-degree have exponents of 2.1; thus the degree-based selection model 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  $\alpha$ , i.e., by incorporating a random selection component in choosing nodes. The answer is surprisingly<sup>2</sup> yes; more on this in Section 5.3.

## 5.2 PageRank-Based Selection

We show that power law emerges for the PageRank and degree distributions in this model (we assume  $\beta = 0$ , i.e, the node selection is based only on Pagerank), but the exponents are different from the degree-based model.

Using the same argument as before, we can show that Equation (2) holds. However,  $f^t(v)$  here follows a different differential equation than the one in the previous analysis:  $\frac{d(f^t(v))}{dt} \approx \frac{f^t(v)r}{2rt}$ . 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

<sup>2</sup> Surprising because, it is not the case that PageRank and in-degree distributions are related – as suggested by the the similarity of the power law exponents of the two distributions. It follows from our analysis above, that even when nodes are selected uniformly at random (i.e.,  $\alpha = 1$ ), a power law (with a small exponent) emerges for the PageRank; but the degree distribution is Poisson.



**Fig. 2.** 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.

large) and the total PageRank is thus proportional to the sum of the degrees which is  $2rt$ .

Plugging the solution of the above differential equation in Equation (2), we can show that the probability density function  $F$  for  $\pi^t(v)$  in this model is:  $F(\phi) \approx (\pi(c))^2 / r^2 t^2 \phi^3$ , i.e., predicting that the PageRank follows a power law with exponent 3. Analogously, we can show that the degree also follows a power law with exponent 3. Simulations also 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  $\beta$ ? Unlike the degree-based selection model, the answer is no; increasing  $\beta$  will only increase the power law exponent (above 3) for the in-degree distribution. This can be verified by experiments. We are thus left with the degree-based selection model and the hybrid selection model of Section 3 as candidates for explaining the observations.

### 5.3 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 3 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  $\beta$  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  $\beta$  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  $\alpha = 0$ , both the in-degree and PageRank distributions followed a power law with exponent 2. We observed that increasing  $\alpha$  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  $\alpha$  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 3. Surprisingly when  $a = b \approx 0.33$ , we could again simultaneously fit all three distributions. Thus we have an alternative model, with a substantial PageRank-based selection component, that fits the Web empirical data. As mentioned in Section 3, this model is plausible from the behavioral standpoint.

## 6 Conclusion

We present experimental and analytical studies of PageRank distribution on the Web graph, and use it to develop more accurate generative models for the evolution of the Web graph. We consider three possible models: degree-based selection, PageRank-based selection, 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 is investigating the relationship between PageRank and in-degree which may shed new insight into Web structure.

## Acknowledgments

We are very 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 [3], *Science*, **287**, 2000, 2115a.

2. Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, **1**(1), 2001, 2-43.
3. A. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, **286**(509), 1999.
4. A. Barabasi, R. Albert and H. Jeong. Mean-field theory for scale-free random graphs. *Physica A*, **272**, 1999, 173-187.
5. B. Bollobas. *Random Graphs*. Academic Press, 1990.
6. 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.
7. B. Bollobas and O. Riordan. The diameter of a scale-free random graph. *preprint*, 2001.
8. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th WWW conference*, 1998.
9. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, Andrew Tomkins, J. Weiner. Graph Structure in the Web. In *Proceedings of the 9th WWW Conference*, 2000.
10. 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.
11. 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.
12. Google Inc. <http://www.google.com>
13. 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.
14. 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.
15. 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 (FOCS)*, 2000.
16. R. Motwani and P. Raghavan. *Randomized Algorithms*, Cambridge University Press, 1995.
17. 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.
18. WT10g collection draft paper.  
<http://www.ted.cmis.csiro.au/TRECWeb/wt10ginfo.ps.gz>