Methods for Sampling Pages Uniformly from the World Wide Web

Paat Rusmevichientong[1]*, David M. Pennock[2], Steve Lawrence[2], C. Lee Giles[3]
[1] Department of Management Science and Engineering, Stanford University
[2] NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
[3] School of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16801
paatrus@stanford.edu, {dpennock,lawrence}@research.nj.nec.com, giles@ist.psu.edu

Abstract

We present two new algorithms for generating uniformly random samples of pages from the World Wide Web, building upon recent work by Henzinger *et al.* (Henzinger *et al.* 2000) and Bar-Yossef *et al.* (Bar-Yossef *et al.* 2000). Both algorithms are based on a weighted random-walk methodology. The first algorithm (DIRECTED-SAMPLE) operates on arbitrary directed graphs, and so is naturally applicable to the web. We show that, in the limit, this algorithm generates samples that are uniformly random. The second algorithm (UNDIRECTED-SAMPLE) operates on undirected graphs, thus requiring a mechanism for obtaining inbound links to web pages (e.g., access to a search engine). With this additional knowledge of inbound links, the algorithm can arrive at a uniform distribution faster than DIRECTED-SAMPLE, and we derive explicit bounds on the time to convergence. In addition, we evaluate the two algorithms on simulated web data, showing that both yield reliably uniform samples of pages. We also compare our results with those of previous algorithms, and discuss the theoretical relationships among the various proposed methods.

Introduction

The World Wide Web is a massive network of information, reflecting ideas and trends from throughout society. Accurate characterizations of the web's content and structure provide a window into the interests of millions of individuals and organizations around the globe. However, obtaining good estimates of these and other quantities is difficult, due to the web's sheer size (about 10^9 pages (Inktomi/NEC January 19 2000; Lawrence & Giles 1999) and growing), and its distributed and dynamic nature. Exhaustive enumeration of all web pages is a technically challenging and costly task, and any results become rapidly outdated (Brin & Page 1998; Kahle 1997). A more reasonable approach is to infer statistics based on a random sample of web pages.

Generating a uniform sample of web pages is itself a nontrivial problem. Several methods have been proposed, though no standard methodology has emerged. Lawrence and Giles (Lawrence & Giles 1998) queried major search engines to estimate the overlap in their databases. They infer a bound on the size of the indexable web and estimate the search engines' relative coverages. The same authors (Lawrence & Giles 1999) employ a sampling approach based on random testing of IP addresses to determine characteristics of hosts and pages found on the web, and to update their size estimate for the web. Bharat and Broder (Bharat & Broder 1998) propose another methodology based on generating random queries.

Other methods make explicit use of the hyperlink structure of the web. We can view the web as a directed graph whose nodes correspond to web pages and edges correspond to hyperlinks. A sample of web pages can be generated by crawling the web graph according to a specified policy. Henzinger *et al.* (Henzinger *et al.* 2000) present a crawling policy that is based on reweighting the results of a random crawl in order to approximate a uniform sample. Their algorithm (which we refer to as PAGERANK-SAMPLE) makes use of the PageRank of each web page. PageRank is a measure of the popularity of a web page, and is used in part by Google (`http://www.google.com/`) to rank search results (Brin & Page 1998). The authors explain how their method can ideally yield a nearly uniform sample, though they point out a few practical and theoretical barriers, discussed further in the **Comparative Experiments** Section. In their experimental results, the generated sample still appears biased toward web pages that have large numbers of inbound links.

Bar-Yossef *et al.* (Bar-Yossef *et al.* 2000) propose an alternative random-walk method for sampling web pages uniformly. Their method (REGULAR-SAMPLE) assumes that the web is an undirected graph, so that hyperlinks can be followed backward as well as forward. Since the web is not truly undirected, in practice the algorithm must query a search engine for the inbound links to each page. Their experiments seem to indicate that, when applied to the actual web, this type of approximation results in a biased sample.

In this paper, we develop a new algorithm for generating uniform samples from the web, also built upon a random-walk methodology. Our approach (DIRECTED-SAMPLE)

*Part of this work was conducted while visiting NEC Research Institute.

works on arbitrary directed graphs, so it is directly applicable to the web. We prove that, in the asymptotic limit, the algorithm generates a uniform sample of web pages—that is, the expected number of occurrences in the sample is the same for all pages on the web. To our knowledge, this is the first algorithm with provable asymptotic convergence on the real web. In the **Algorithm** UNDIRECTED-SAMPLE **and Analysis** Section, we modify the algorithm to operate on an undirected graph (UNDIRECTED-SAMPLE), where we assume knowledge of the inbound links to every page. In this setting, we show that the problem and analysis is considerably simplified. In addition to guaranteeing convergence, we can derive formal bounds on the time to convergence.

In the **Empirical Results and Comparisons** Section, we test our DIRECTED-SAMPLE algorithm on a simulated directed web graph, showing that it generates near-uniform samples that are independent of the number of links to or from a page. We directly compare our UNDIRECTED-SAMPLE algorithm with the PAGERANK-SAMPLE and REGULAR-SAMPLE algorithms, using simulated web data under undirected modeling assumptions. In these experiments, both UNDIRECTED-SAMPLE and REGULAR-SAMPLE outperform PAGERANK-SAMPLE. We isolate the key approximation in PAGERANK-SAMPLE that we believe leads to biased samples, and we show that a modified PAGERANK-SAMPLE can be thought of as a special case of DIRECTED-SAMPLE. Finally, we conclude with a summary and discussion of future research directions.

Setup and Notation

Broder *et al.* (Broder *et al.* 2000) find that the web can be divided into four main components: a central strongly connected core, a set of pages that can be reached from the core but do not connect back to it, a set of pages that connect to the core but cannot be reached from it, and a set of pages disconnected from the core. We seek to generate a uniform sample from among those pages in the core and some of the pages reachable from the core. Almost any sampling method based on a random walk cannot hope to sample pages that are not reachable from the starting page. We believe that, for the purposes of statistical testing, this subset of the web is sufficiently representative of most publicly accessible web pages of interest.

We view the web as a directed graph where web pages are nodes and hyperlinks are edges. Denote the set of all web pages as $\mathcal{S} = \{1, 2, \ldots, n\}$ and assume without loss of generality (w.l.o.g.) that page 1 is within the strongly connected core (e.g., `http://www.yahoo.com/`). Let $\mathbf{X}$ be an $n \times n$ matrix defined by

$$X_{ij} = \begin{cases} 1, & \text{if there is a link from page } i \text{ to page } j, \text{ OR} \\ & \text{if } i = j, \text{ OR} \\ & \text{if } i \text{ has no outgoing links and } j = 1. \\ 0, & \text{otherwise.} \end{cases}$$

The matrix $\mathbf{X}$ can be thought of as a connection matrix whose non-zero elements denote connections between various pages in the web. Note that we assume w.l.o.g. that each node is connected to itself (i.e., the diagonal elements of $\mathbf{X}$ equal 1), and that "dead end" pages without hyperlinks connect back to page 1. We also make the following assumption:

Assumption 1 *There exists* $n \geq 1$ *such that* $(\mathbf{X}^n)_{ij} > 0$ *for all* $i, j \in \mathcal{S}$.

In other words, every page can be reached from every other page by traversing edges. This assumption holds for all pages in the strongly connected core and some pages reachable from the core.

Let $\mathbf{P}$ denote a matrix obtained by normalizing each row of $\mathbf{X}$ so that it sums to one:

$$P_{ij} = \frac{X_{ij}}{\sum_{s \in \mathcal{S}} X_{is}}, \quad \forall i, j \in \mathcal{S}.$$

We can then think of $\mathbf{P}$ as the transition probability matrix associated with a random walk in which a hyperlink is chosen uniformly at random. Assumption 1 implies that the associated random walk is irreducible, and since the diagonal elements of $\mathbf{P}$ are positive, it is also aperiodic. Thus, it follows from a standard result in stochastic processes that there exists a stationary distribution $\pi = (\pi(1), \ldots, \pi(n))$ associated with $\mathbf{P}$ such that

$$\pi(i) = \sum_{j \in \mathcal{S}} \pi(j) P_{ji}, \quad \forall i \in \mathcal{S}.$$

Algorithm DIRECTED-SAMPLE and Analysis

The stationary distribution π represents the asymptotic frequency of visiting each page under the random walk with transition probability $\mathbf{P}$. Thus, for a sufficiently long walk, we would expect that the frequency that a web page i is visited is $\pi(i)$. So, for each web page i, if we include it in our sample with probability $1/\pi(i)$, then the expected number of occurrences of each web page in our sample should be the same, yielding a uniformly random sample. Unfortunately, we do not know the true value of $\pi(i)$, so we must estimate it. For each web page i collected in our initial crawl, we compute an estimate of the stationary probability $\pi(i)$ by performing another random walk and recording how often the walk visits page i.

The formal definition of the algorithm follows. Note that there are five design parameters: s_0, N, K, M, and β.

Algorithm DIRECTED-SAMPLE

1. Start at some web page s_0. Crawl the web according to transition probability $\mathbf{P}$ for N time periods.

2. After N time periods, continue to crawl the web for an additional K time periods. Let $X_1, \ldots, X_K$ denote the web pages visited during this crawl, and let $\mathcal{D}_N$ denote the collection of *distinct* web pages collected during this crawl. Note that $|\mathcal{D}_N| \leq K$ since some web pages might be repeated.

3. For each page $l \in \mathcal{D}_N$, compute an estimated stationary probability $\tilde{\pi}(l)$ as follows. Crawl the web according to $\mathbf{P}$ for M time periods starting at page l. Let $Z_0^l, Z_1^l, \ldots, Z_M^l$ denote the sequence of web pages that are visited during the crawl, where $Z_0^l = l$. Then, let

$$\tilde{\pi}(l) = \frac{\sum_{r=1}^{M} 1_{\{Z_r^l = l\}}}{M},$$

where $1_{\{Z_r^l = l\}}$ is an indicator random variable which is 1 if Z_r^l is equal to l, and 0 otherwise.

4. For $1 \leq l \leq K$, include page X_l in the final sample set $\mathcal{D}$ with probability $\beta/\tilde{\pi}(X_l)$, where β is some positive number such that $0 < \beta < \min_{l=1,\ldots,K} \tilde{\pi}(X_l)$.

The following proposition establishes the algorithm's soundness: for sufficiently large N and M, the returned sample set $\mathcal{D}$ is a uniform sample. Specifically, in the asymptotic limit, the expected number of occurrences of page i in $\mathcal{D}$ is the same for all pages $i \in \mathcal{S}$.

Proposition 1 *For any web page $i \in \mathcal{S}$, let*

$$W_i = \text{number of occurrences of page } i \text{ in } \mathcal{D},$$

where $\mathcal{D}$ is the final set of web pages returned after step four of the algorithm. Then, for all $i \in \mathcal{S}$,

$$\lim_{N,M\to\infty} E[W_i] = \beta K.$$

Proof. Recall that the variables $X_1, \ldots, X_K$ denote the collection of web pages gathered in step two of the algorithm. For $1 \leq l \leq K$, let a random variable Y_l be defined by

$$Y_l = \begin{cases} X_l & \text{with probability } \beta/\tilde{\pi}(X_l) \\ \triangle & \text{otherwise,} \end{cases}$$

where $\triangle$ represents the event that X_l is not included in our sample. Thus, the collection of random variables $\{Y_l : Y_l \neq \triangle\}$ corresponds to the set of web pages in our final sample $\mathcal{D}$. If s_0 denotes the starting page for our crawl in step one, it follows that

$$\begin{aligned} \lim_{N,M\to\infty} E[W_i] &= \lim_{N,M\to\infty} E\left[\sum_{l=1}^{K} 1_{\{Y_l = i\}}\right] \\ &= \lim_{N,M\to\infty} \sum_{l=1}^{K} \frac{\beta}{\tilde{\pi}(i)} \Pr\{X_l = i\} \end{aligned}$$

where the last equality follows from the fact that $Y_l = i$ if and only if $Y_l \neq \triangle$ and $X_l = i$. Since X_l corresponds to the state at time $l + K$ of our random walk,

$$\lim_{N,M\to\infty} E[W_i] = \lim_{N,M\to\infty} \sum_{l=1}^{K} \frac{\beta}{\tilde{\pi}(i)} \left(\mathbf{P}^{N+l}\right)_{s_0 i}.$$

It is a standard result in stochastic processes that

$$\lim_{N\to\infty} \left(\mathbf{P}^{N+l}\right)_{s_0 i} = \lim_{M\to\infty} \tilde{\pi}(i) = \pi(i),$$

which implies that

$$\lim_{N,M\to\infty} E[W_i] = \beta K.$$

■

We do not yet have a bound on time to convergence. Very large values of N and M may be required to produce a near-uniform sample, potentially affecting the practicality of running DIRECTED-SAMPLE on the entire web. Note that the parameter N is the "burn-in" time. During this phase, we need only perform a memoryless crawl; we need not record any information along the way. The purpose of burn-in is to eliminate any dependence on s_0 and to induce a near-stationary distribution on the random variables $X_1, \ldots, X_K$, which are instantiated in step two. Parameter M is the number of pages traversed in order to estimate the stationary distribution over $X_1, \ldots, X_K$. Because web pages with a high number of inbound links (e.g., `http://www.yahoo.com/`) are visited often, the algorithm produces accurate estimates of their stationary probabilities even with a relatively small M. However, web pages that have small numbers of inbound links (the vast majority of pages) occur very infrequently during the crawl, and thus M may need to be very large to obtain good estimates for these pages. In the next section, we see that we can eliminate this estimation phase (step three) altogether, if we can assume access to the inbound links to every page.

We are currently pursuing techniques to reduce and bound the time to convergence for DIRECTED-SAMPLE. We are also examining the potential use of the algorithm in conjunction with a focused crawler (Chakrabarti, van den Berg, & Dom 1999; Diligenti *et al.* 2000), in order to generate uniform samples of topic-specific subsets of the web without traversing the entire web.

Algorithm UNDIRECTED-SAMPLE and Analysis

In the DIRECTED-SAMPLE algorithm, each page i collected in step two is included in the final sample with probability inversely proportional to its estimated stationary probability $\tilde{\pi}(i)$. The quality of the resulting sample depends significantly on how accurately we are able to estimate the true stationary probability $\pi(i)$.

In this section we show that, if we assume the web is an undirected graph, then the algorithm becomes greatly simplified and expedited. This assumption is equivalent to requiring knowledge of all pages that point to a given web page. The assumption is partially justified by the fact that many search engines allow queries for inbound links to web pages. Note, however, that the search engines' databases are neither complete nor fully up to date, and performing many such queries can be time consuming and costly.

In this undirected setting, let $\overline{\mathbf{X}}$ be an $|\mathcal{S}|$-by-$|\mathcal{S}|$ matrix defined by

$$\overline{X}_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } i \leftrightarrow j \\ 0 & \text{otherwise.} \end{cases},$$

where $i \leftrightarrow j$ means that there is either a hyperlink from i to j or from j to i. Thus, the matrix $\overline{\mathbf{X}}$ denotes the connection matrix between various web pages on the web, ignoring directionality. In this case, Assumption 1 holds for all pages in the strongly connected core, all pages reachable from the core, and all pages that connect to the core.

Consider a transition probability matrix $\overline{\mathbf{P}}$ defined by

$$\overline{P}_{ij} = \frac{\overline{X}_{ij}}{\sum_{s \in \mathcal{S}} \overline{X}_{is}}, \quad \forall i, j \in \mathcal{S}.$$

The matrix $\overline{\mathbf{P}}$ corresponds to a transition probability where, at any given web page, one hyperlink is chosen uniformly at random to follow (including from among those which point to the page).

For any web page $i \in \mathcal{S}$, let $d(i)$ denote the *degree* of i, or the sum of the number of links from page i and the number of links to i. The degree $d(i)$ is the total number of connections associated with page i. Application of the DIRECTED-SAMPLE algorithm to the case of an undirected graph relies on the following lemma.

Lemma 1 *If $\overline{\mathbf{P}}$ is irreducible, then the associated stationary distribution $\bar{\pi}$ of $\overline{\mathbf{P}}$ is given by*

$$\overline{\pi}(i) = \frac{d(i)+1}{|\mathcal{S}| + \sum_{s \in \mathcal{S}} d(s)} = \frac{d(i)+1}{|\mathcal{S}| + 2|\mathcal{E}|}, \quad \forall i \in \mathcal{S},$$

where $|\mathcal{E}|$ denotes the total number of edges in the graph.

Proof. Since $\overline{P}_{ii} > 0$ for all $i \in \mathcal{S}$, it follows that $\overline{\mathbf{P}}$ is both irreducible and aperiodic. It follows that a stationary distribution exists and it is unique. Thus, it suffices to show that

$$\overline{\pi}(i) = \sum_{j \in \mathcal{S}} \overline{\pi}(j) \overline{P}_{ji}, \quad \forall i \in \mathcal{S}.$$

Note that

$$\begin{aligned} \sum_{j \in \mathcal{S}} \overline{\pi}(j) \overline{P}_{ji} &= \sum_{j: j=i \text{ or } j \leftrightarrow i} \frac{d(j)+1}{|\mathcal{S}| + \sum_{s \in \mathcal{S}} d(s)} \cdot \frac{1}{d(j)+1} \\ &= \sum_{j: j=i \text{ or } j \leftrightarrow i} \frac{1}{|\mathcal{S}| + \sum_{s \in \mathcal{S}} d(s)} \\ &= \frac{d(i)+1}{|\mathcal{S}| + \sum_{s \in \mathcal{S}} d(s)} \\ &= \overline{\pi}(i) \end{aligned}$$

The desired conclusion follows. ■

The above lemma provides us with an explicit formula for the stationary distribution of $\overline{\mathbf{P}}$. Thus, we can use the same basic algorithm presented in Section , but without step three, since we do not need to estimate the stationary probability. This eliminates the potentially very long crawl of M steps required in the directed case. The formal definition of the algorithm follows.

Algorithm UNDIRECTED-SAMPLE

1. Start at some web page s_0. Crawl the web according to transition probability $\overline{P}$ for N time periods.
2. After N time periods, continue to crawl the web for an additional K time periods. Let $X_1, \ldots, X_K$ denote the web pages visited during this crawl.
3. For $1 \leq l \leq K$, page X_l will be included in our sample with probability $\beta / (d(X_l) + 1)$, where β is some positive number such that $0 < \beta < 1 + \min_{i=1,\ldots,K} d(X_l)$.

The proof of Proposition 1 immediately extends to the current case. Moreover, by assuming that the web is modeled as an undirected graph, we are able to put a bound on the deviation of our sample from a truly-random sample.

Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ denote the (undirected graph) of the web, where $\mathcal{S}$ denotes the set of web pages, and $\mathcal{E}$ denotes the collection of hyperlinks (assuming that they are undirected). Let $d_* = \max_{x \in \mathcal{S}} d(x)$ denote the maximum degree, and let γ_* denote the diameter of the web. Also, let Γ denote the collection of shortest paths between any two points in the web graph, and let

$$b = \max_{e \in \mathcal{E}} |\{\gamma \in \Gamma : e \in \gamma\}|.$$

That is, b denotes the maximum number of paths in Γ that have a single edge in common. Thus, b measures the "bottleneck" in the web graph.

Proposition 2 *For each $i \in \mathcal{S}$, let*

$$W_i = \text{number of occurrences of page } i \text{ in sample.}$$

Then,

$$\left| \frac{E[W_i]}{K} - \frac{\beta}{|\mathcal{S}| + 2|\mathcal{E}|} \right| \leq \left(\frac{\beta}{d(i)+1} \sqrt{\frac{2|\mathcal{E}| + |\mathcal{S}| - d(s_0) - 1}{d(s_0)+1}} \right) \lambda_*^N,$$

where

$$\lambda_* \leq \max \left[\left(1 - \frac{2|\mathcal{E}|}{(d_*+1)^2 \gamma_* b} \right), \left(1 - \frac{2}{(d_*+1)} \right) \right].$$

Proposition 2 implies that for a sufficiently large value of N, the expected number of occurrences of each web page in the sample approaches $\beta K / (|\mathcal{S}| + 2|\mathcal{E}|)$. It also shows that the rate of convergence is exponential in the parameter λ_*, which depends on graphical properties of the web, such as its maximum degree, diameter, and bottleneck. In the future, we plan to compute estimates of these quantities for the real web.

The proof of Proposition 2 relies on the following results from Diaconis and Stroock (Diaconis & Stroock 1991) on geometric bounds for eigenvalues of Markov chains.

Lemma 2 *Let* $\mathbf{P}$ *be a transition probability for a reversible Markov chain on a state space* X *with* $|X| = m$ *and stationary distribution* π. *Assume that* $\mathbf{P}$ *is irreducible with eigenvalues* $1 = \lambda_0 > \lambda_1 \geq \ldots \lambda_{m-1} \geq -1$. *Then, for all* $x \in X$,

$$\sum_{y \in X} \left| P_{xy}^n - \pi(y) \right| \leq \sqrt{\frac{1 - \pi(x)}{\pi(x)}} \lambda_*^n,$$

where $\lambda_* = \max(\lambda_1, |\lambda_{m-1}|)$.

Proof. See Proposition 3 in (Diaconis & Stroock 1991). ■

Lemma 3 *Let* (X, E) *be a connected graph. If* $\mathbf{P}$ *denotes a random walk on this graph, and* λ_1 *denotes the second largest eigenvalue of* $\mathbf{P}$, *then*

$$\lambda_1 \leq 1 - \frac{2|E|}{d_*^2 \gamma_* b}$$

where $d_* = \max d(x)$ *is the maximum degree,* γ_* *is the diameter of the graph, and*

$$b = \max_{e \in E} |\{\gamma \in \Gamma : e \in \gamma\}|,$$

where Γ *denotes the collection of shortest paths between any two points in the graph.*

Proof. See Corollary 1 in (Diaconis & Stroock 1991). ■

Lemma 4 *Let* (X, E) *be a connected graph which is not bipartite. For each* $x \in X$, *let* σ_x *be any path of odd length from* x *to* x. *Let* Σ *be the collection of such paths, one for each* $x \in X$. *If* $\mathbf{P}$ *denotes a random walk on this graph, and* $\lambda_{\min}$ *denotes the smallest eigenvalue of* $\mathbf{P}$, *then*

$$\lambda_{\min} \geq -1 + \frac{2}{d_* \sigma_* b_*},$$

where d_* *is the maximum degree,* σ_* *is the maximum number of edges in any* $\sigma \in \Sigma$, *and*

$$b_* = \max_{e \in E} |\{\sigma \in \Sigma : e \in \sigma\}|.$$

Proof. See Corollary 2 in (Diaconis & Stroock 1991). ■

Since each node in our web graph has a self loop, we can choose Σ in the above lemma to be a collection of self loops, one for each $x \in X$. Then,

$$\lambda_{\min} \geq -1 + \frac{2}{d_*}.$$

This is the bound that we will use in our proof of Proposition 2, which follows.

Proof. Recall that $X_1, \ldots, X_K$ denote the collection of web pages collected during step two. For $1 \leq l \leq K$, let a random variable Y_l be defined by

$$Y_l = \begin{cases} X_l & \text{with probability} \beta/(d(X_l)+1) \\ \triangle & \text{otherwise,} \end{cases}$$

where $\triangle$ represents the event that X_l is not included in our sample. Thus, the collection of random variables $\{Y_l : Y_l \neq \triangle\}$ corresponds to the set of web pages in our final sample. Thus,

$$W_i = \sum_{l=1}^{K} 1_{\{Y_l = i\}},$$

which implies that

$$E[W_i] = \sum_{l=1}^{K} \frac{\beta}{d(i)+1} \Pr\{X_l = i\}.$$

Thus,

$$\begin{aligned} &\left| \frac{E[W_i]}{K} - \frac{\beta}{2|\mathcal{E}| + |\mathcal{S}|} \right| \\ &= \left| \frac{1}{K} \sum_{l=1}^{K} \left(\frac{\beta}{d(i)+1} \Pr\{X_l = i\} - \frac{\beta}{2|\mathcal{E}| + |\mathcal{S}|} \right) \right| \\ &= \frac{\beta}{K(d(i)+1)} \left| \sum_{l=1}^{K} \left(\Pr\{X_l = i\} - \frac{d(i)+1}{2|\mathcal{E}| + |\mathcal{S}|} \right) \right| \\ &\leq \frac{\beta}{K(d(i)+1)} \left| \sum_{l=1}^{K} \left(\Pr\{X_l = i\} - \overline{\pi}(i) \right) \right| \end{aligned}$$

where the last equality follows from Lemma 1. Since s_0 is our starting web page, and X_l corresponds to the web page that is visited at time $N + l$ of our crawl, it follows from Lemma 2 that

$$|\Pr\{X_l = i\} - \overline{\pi}(i)| \leq \sqrt{\frac{1 - \overline{\pi}(s_0)}{\overline{\pi}(s_0)}} \lambda_*^N, \quad \forall l = 1, \ldots K,$$

where $\lambda_* = \max\{\lambda_1, |\lambda_{\min}|\}$ and λ_1 denotes the second largest eigenvalue of the transition probability $\overline{\mathbf{P}}$ and $\lambda_{\min}$ denotes the smallest eigenvalue of $\overline{\mathbf{P}}$. Thus,

$$\begin{aligned} \left| \frac{EW_i}{K} - \frac{\beta}{2|\mathcal{E}| + |\mathcal{S}|} \right| &\leq \frac{\beta}{d(i)+1} \sqrt{\frac{1 - \overline{\pi}(s_0)}{\overline{\pi}(s_0)}} \lambda_*^N \\ &= \left(\frac{\beta}{d(i)+1} \sqrt{\frac{2|\mathcal{E}| + |\mathcal{S}| - d(s_0) - 1}{d(s_0)+1}} \right) \lambda_*^N, \end{aligned}$$

where the last equality follows from the expression of the stationary probability given in Lemma 1. It follows from Lemmas 3 and 4 that

$$\lambda_* \leq \max\left[\left(1 - \frac{2|\mathcal{E}|}{(d_*+1)^2 \gamma_* b} \right), \left(1 - \frac{2}{(d_*+1)} \right) \right],$$

from which the desired result follows. ■

Empirical Results and Comparisons

In this section, we present experimental evaluations of the DIRECTED-SAMPLE and UNDIRECTED-SAMPLE algorithms, including comparisons with previous algorithms. We describe how a modified PAGERANK-SAMPLE algorithm can be viewed as a special case of DIRECTED-SAMPLE, and we isolate the key approximation step in PAGERANK-SAMPLE that we believe leads to biased samples.

DIRECTED-SAMPLE

We test the DIRECTED-SAMPLE algorithm on a simulated web graph of 100,000 nodes, 71,189 of which belong to the primary strongly connected component. The graph was generated using Pennock *et al.*'s (Pennock *et al.*) extension of Barabási and Albert's (Barabási & Albert 1999) model of web growth. The model generates undirected graphs with edge distributions almost identical to that of the real web. We converted the resulting undirected graph into a directed graph by assigning directionality to edges at random. We ran the DIRECTED-SAMPLE algorithm with parameters $N = 5 \times 10^5$, $M = 2 \times 10^6$, and $K = 5 \times 10^5$, resulting in a sample of size $|\mathcal{D}| = 2057$. Because our method is based on a random walk, we expect the numbers of inbound and outbound links to be the most likely source of sample bias. Figure 1 compares the distribution of inbound links in the sample set $\mathcal{D}$ to the true distribution over the entire graph. Figure 2 displays the same comparison for outbound links. The likelihood that a particular page occurs in the sample appears to be independent of the number of links to or from that page.

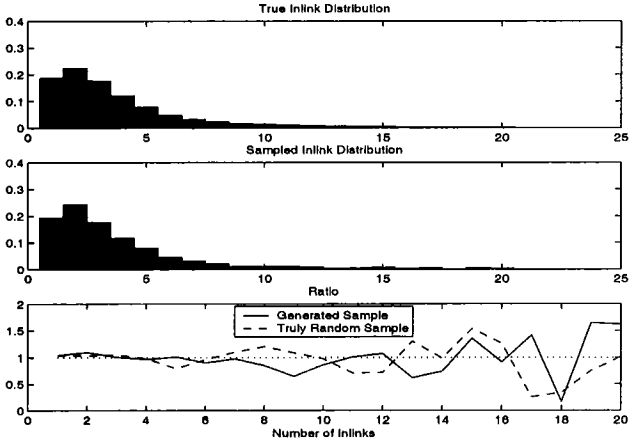

Figure 1: (a) Distribution of inbound links for the simulated web graph. (b) Distribution of inbound links for pages returned by the DIRECTED-SAMPLE algorithm. (c) Ratio of the sampled distribution to the true distribution, for both DIRECTED-SAMPLE and a truly random sampling algorithm.

Figure 3 shows a histogram of the node ID numbers in the sample. All nodes in the graph are grouped into ten equally-spaced buckets. Figure 3(a) shows the proportion of sampled nodes chosen from each bucket. If the sample is uniform, then the proportion in each bucket should be about the same. Figure 3(b) plots the ratio of the proportion in each bucket to the true expected value under uniform sampling. From these figures, there does not appear to be any systematic bias in our sampling.

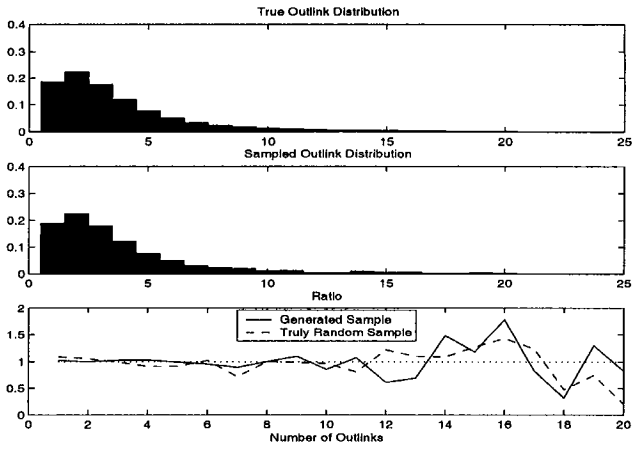

Figure 2: Comparison of sampled and actual outbound link distributions for the DIRECTED-SAMPLE algorithm.

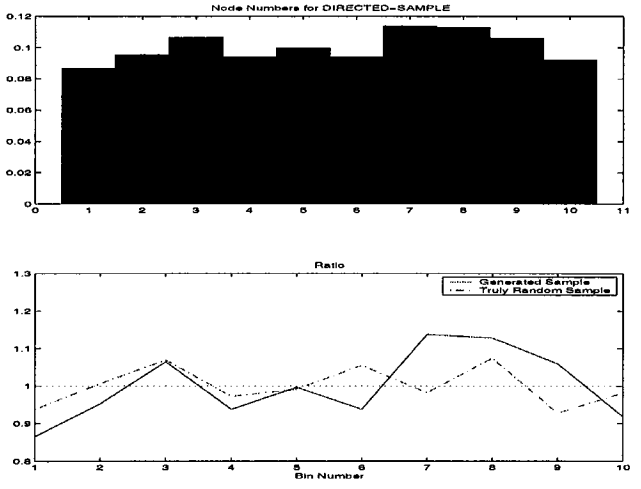

Figure 3: Distribution of node numbers for samples generated by the DIRECTED-SAMPLE algorithm.

UNDIRECTED-SAMPLE

In this section, we present empirical results of the UNDIRECTED-SAMPLE algorithm. The experiments were performed on an undirected graph of 100,000 nodes (71,189 in the main connected component), generated according Pennock *et al.*'s (Pennock *et al.*) model of web growth. The burn-in time N is set at 2000, and our starting node s_0 is set to node number 50,000. The generated sample size is $|\mathcal{D}| = 10,000$ nodes. With the choice of $\beta = 1.999$, about one in every five web pages visited is accepted into our sample.

Figure 4 shows the distribution of the number of edges in the graph versus that of our sample. The bottom portion of Figure 4 shows the ratio between the proportion of nodes

in the sample having a certain number of edges versus the true proportion. Overall, there seems to be no systematic bias toward nodes with small or large degrees. Figure 5 shows the histogram of the node numbers generated under the algorithm.

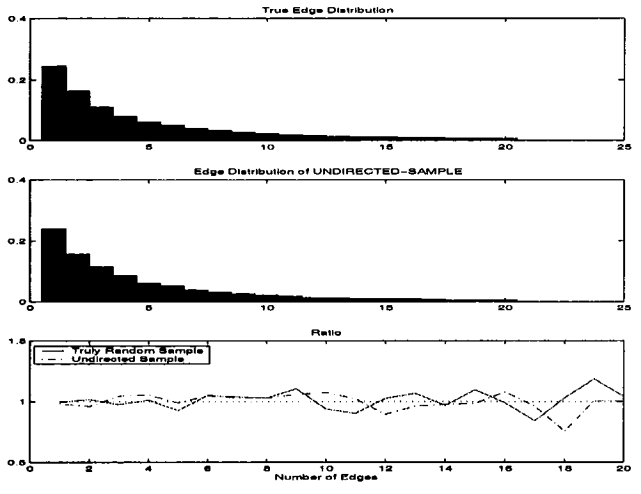

Figure 4: Distribution of the number of edges for samples generated by the UNDIRECTED-SAMPLE algorithm.

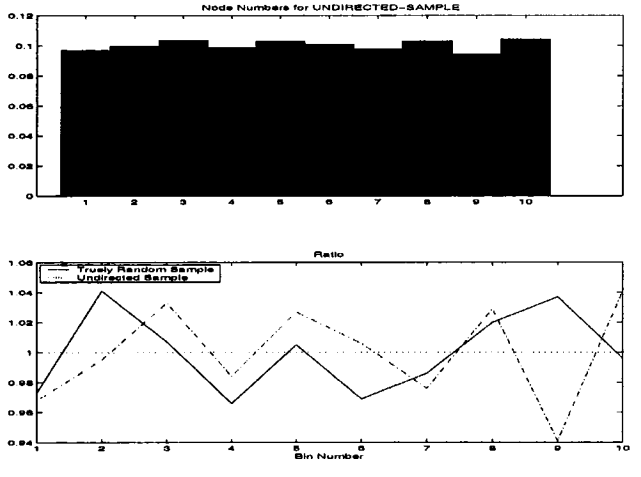

Figure 5: Distribution of node numbers for samples generated by the UNDIRECTED-SAMPLE algorithm.

Comparative Experiments

In this section, we compare the results of UNDIRECTED-SAMPLE with those of PAGERANK-SAMPLE (Henzinger *et al.* 2000) and REGULAR-SAMPLE (Bar-Yossef *et al.* 2000).

Let n denote the total number of nodes in the graph. The PAGERANK-SAMPLE algorithm conducts a random walk on the graph: at each node i, with probability $1 - d$ a neighbor of i is chosen at random; with probability d, any node in the graph is chosen at random. Since choosing a node at random from among all nodes is not feasible, Henzinger *et al.* (Henzinger *et al.* 2000) approximate this step by choosing from among all nodes visited so far. The final sample is generated by choosing from among the visited nodes with probability inversely proportional to their PageRank.

The REGULAR-SAMPLE algorithm conducts a random walk on a dynamically-built graph constructed such that every node has the same degree (i.e., so that the graph is regular). The construction is performed by assuming that the degree of every node in the original graph is bounded by some constant d_{max}. The new graph is built by adding the appropriate number of self loops at each node, so that every node has the same degree. Because each node in this graph has the same degree, the associated stationary distribution for the random walk on this graph is uniform.

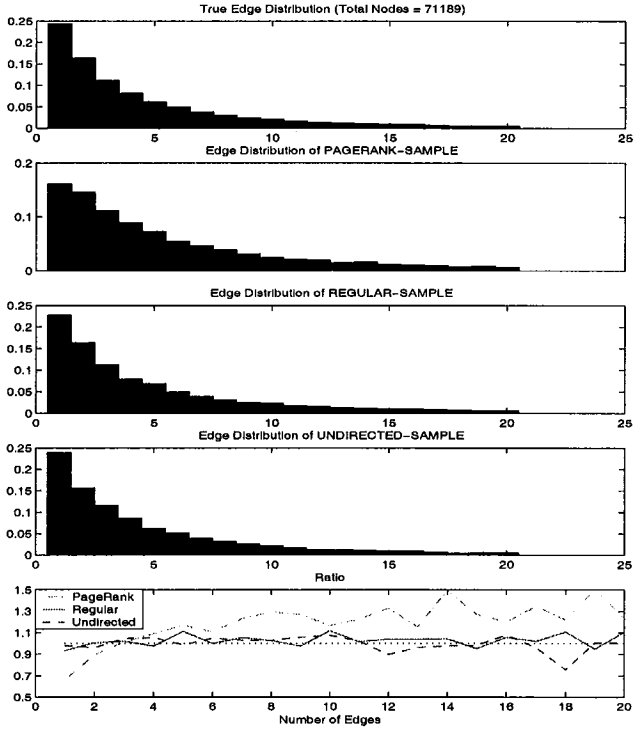

Figure 6: (a) Distribution of edges for the simulated undirected graph. (b) Distribution of edges for pages returned by the PAGERANK-SAMPLE algorithm. (c) Distribution of edges for pages returned by the REGULAR-SAMPLE algorithm. (d) Distribution of edges for pages returned by the UNDIRECTED-SAMPLE algorithm. (e) Ratio of the sampled distribution to the true distribution, for all three algorithms.

We used a burn-in time of 2000 for UNDIRECTED-SAMPLE and REGULAR-SAMPLE, and an initial seed set size of 1000 for PAGERANK-SAMPLE. We ran the UNDIRECTED-SAMPLE and PAGERANK-SAMPLE al-

gorithms until samples of size 10,000 were collected. The REGULAR-SAMPLE algorithm generated a sample of 2,980,668 nodes, 14,064 of which were unique (the vast majority of nodes were repeats due to self loops). We assumed knowledge of the true $d_{max} = 937$ for the REGULAR-SAMPLE algorithm. Figure 6 shows the true edge distribution for the simulated web graph, and the sampled distributions for all three algorithms. We see that both UNDIRECTED-SAMPLE and REGULAR-SAMPLE produce what appear to be uniform samples, without any noticeable bias based on the number of edges incident onto a node. On the other hand, PAGERANK-SAMPLE does appear to exhibit a consistent bias toward pages with large numbers of edges.

We should note that the original idea underlying the PAGERANK-SAMPLE algorithm can be seen in some sense as a special case of our DIRECTED-SAMPLE algorithm. The idealized crawling policy of PAGERANK-SAMPLE corresponds to a transition probability P_H given by

$$(P_H)_{ij} = \begin{cases} \frac{1-d}{d_{out}(i)} + \frac{d}{|\mathcal{S}|} & \text{if } j \text{ is a child of } i \\ \frac{d}{|\mathcal{S}|} & \text{otherwise.} \end{cases}$$

where $d_{out}(i)$ denotes the number of outlinks of web page i. Once the random walk is performed, each page i that is visited will be included in the sample with probability that is inversely proportional to the PageRank of i, denoted by $PR(i)$. It turns out that $PR(i)$ is the stationary probability of page i under the transition probability P_H. So, the idea underlying PAGERANK-SAMPLE is very similar to DIRECTED-SAMPLE, though employing an alternative crawling policy and associated stationary distribution.

We believe that the source of bias in PAGERANK-SAMPLE stems from the approximation step required. As Henzinger *et al.* (Henzinger *et al.* 2000) note, we cannot actually conduct a random walk on the web graph according to transition probability matrix P_H. Recall that under P_H, with probability d, a web page is chosen at random from among all pages. However, this is not feasible. Indeed, if we could choose a web page at random, then there would be no need for the algorithm in the first place. So, the authors approximate this step by randomly choosing from among the pages visited so far. We conjecture that this is the primary source of error contributing to the bias in the resulting sample.

Conclusion

We presented two new algorithms (DIRECTED-SAMPLE and UNDIRECTED-SAMPLE) for uniform random sampling of World Wide Web pages. Both algorithms generate samples that are provably uniform in the limit. There are tradeoffs between the two algorithms. The DIRECTED-SAMPLE algorithm is naturally suited to the web, without any assumptions, since it works on any directed graph, though it may take a long time to converge. The UNDIRECTED-SAMPLE algorithm converges faster, and we can formally bound its convergence time, but the algorithm requires an assumption that hyperlinks can be followed backward as well as forward. Empirical tests verify that both algorithms appear to produce unbiased uniform samples. On simulated web data, the UNDIRECTED-SAMPLE algorithm performs as well as the REGULAR-SAMPLE algorithm and better than the PAGERANK-SAMPLE algorithm—two methods recently proposed in the literature. We discuss the theoretical connections between DIRECTED-SAMPLE and PAGERANK-SAMPLE, highlighting what we believe is the key approximation step in PAGERANK-SAMPLE that leads to biased samples.

References

Bar-Yossef, Z.; Berg, A.; Chien, S.; and Fakcharoenphol, J. 2000. Approximating aggregate queries about web pages via random walks. In *Proceedings of the 26th International Conference on Very Large Data Bases.*

Barabási, A., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Bharat, K., and Broder, A. 1998. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of the 7th International World Wide Web Conference.*

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World Wide Web Conference.*

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; and Tomkins, A. 2000. Graph structure in the web. In *Proceedings of the 9th International World Wide Web Conference.*

Chakrabarti, S.; van den Berg, M.; and Dom, B. 1999. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the 8th International World Wide Web Conference.*

Diaconis, P., and Stroock, D. 1991. Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability* 1(1):36–61.

Diligenti, M.; Coetzee, F.; Lawrence, S.; Giles, C. L.; and Gori, M. 2000. Focused crawling using context graphs. In *26th International Conference on Very Large Databases, VLDB 2000,* 527–534.

Henzinger, M.; Heydon, A.; Mitzenmacher, M.; and Najork, M. 2000. On near-uniform URL sampling. In *Proceedings of the 9th International World Wide Web Conference,* 295–308.

Inktomi/NEC. January 19, 2000. Web surpasses one billion documents. Inktomi/NEC press release, `http://www.inktomi.com/`.

Kahle, B. 1997. Preserving the Internet. *Scientific American.*

Lawrence, S., and Giles, C. L. 1998. Searching the World Wide Web. *Science* 280(5360):98–100.

Lawrence, S., and Giles, C. L. 1999. Accessibility of information on the web. *Nature* 400(6740):107–109.

Pennock, D.; Flake, G.; Lawrence, S.; Giles, L.; and Glover, E. Winners don't take all: A model of web link accumulation. in preparation.