# *DRAFT 5/5/2008:* Estimation of Web Page Change Rates

Carrie Grimes[1], Daniel Ford[2]
Google[1]
Google[2]

**Abstract**

Search engines strive to maintain a "current" repository of all pages on the web to index for user queries. However, crawling all pages all the time is costly and inefficient: many small websites don't support that much load and while some pages change very rapidly others don't change at all. Therefore, estimated frequency of change is often used to decide how often to crawl a page. Here we consider the effectiveness of a Poisson process model for the updates of a page, and the associated Maximum Likelihood Estimator, in a practical setting where new pages are continuously added to the set of rates to be estimated. We demonstrate that applying a prior to pages can significantly improve estimator performance for newly acquired pages.

KEY WORDS: Poisson process, Empirical Bayes, Rate of change estimation

## 1. Introduction

Search engines crawl the web to download a corpus of web pages to index for user queries. Since the web updates all the time, the process of indexing new or updated pages requires continual refreshing of the crawled content. In order to provide useful results for time-sensitive or new-content queries, a search engine would ideally maintain a perfectly up-to-date copy of the web at all times. This goal requires that a search engine acquire new links as they appear and refresh any new content that appears on existing pages. As the web grows to many billions of documents and search engine indexes scale similarly, the cost of re-crawling every document all the time becomes increasingly high. One of the most efficient ways to maintain an up-to-date corpus of web pages for a search engine to index is to crawl pages preferentially based on their rate of content update [4].

If a crawler had perfect foresight, it could download each page immediately after the page updates, and similarly acquire new pages as they appear. However, no complete listing or update schedule exists today, although some limited collection of new-content sites such as news pages provide a feed of such structured data to search engines. For most of the web, a crawler must try to estimate the correct time scale on which to sample a web page for new content. The most common model for estimating expected rates of change has been based on the assumption that changes arrive as a Poisson process. However, this model also implies that a crawler cannot deterministically predict the correct time to resample a page. Instead, based on the estimated average time between changes, a crawler can establish a time between refreshes that achieves some benchmark for freshness.

In this paper, we investigate a specific setting for estimating web page change rates: where new pages are added to the web corpus regularly, and estimation must begin immediately to schedule the page for refresh crawling. In this setting, the estimator used for average rate of change must be able to handle cases with virtually no observations in a way that converges to the correct estimate. In addition, we establish a criterion for measuring the expected freshness of a page based on this model, and demonstrate that while the performance of the estimator may be optimized, an optimal estimator performance does not necessarily lead to the freshest corpus during the initialization period for a page.

## 2. Related Work

### 2.1 The Evolution of a Web Page

Several studies have been done on evolution of web pages over reasonably long time periods, in particular Cho and Garcia-Molina [2] and Fetterly et al [7], [6]. Cho and Garcia-Molina downloaded over 700,000 pages from a set of popular webservers on a daily basis over 4 months, and then compared the MD5 checksum of the page to previous versions of the page. Around 23% of pages over all domains in their study changed with an average frequency of one day (the minimum granularity measured), and an additional 15% changed within a week. Cho and Garcia-Molina also found that there were significant differences between the average rate of change for pages on common top-level domains, where `.com` pages were the fastest changing, and `.gov` the slowest.

Fetterly, et al.[7] examined a much larger number of web pages (151M) sampled from a breadth-first crawl and refetched approximately once a week over 10 weeks. To compare page content, they used a more sophisticated similarity metric based on a set of "shingles" of the syntactic content of the page. For each page, information about the start and duration of the download, the HTTP status code, and other details were recorded. This larger study of the web also shows significant differences in the size and permanence of web pages from different top level domains, with pages on the `.com` and `.net` being consistently less available than pages on `.jp` and `.edu`. The study also found, similar to Cho and Garcia-Molina, that `.com` pages change faster than `.gov` and `.edu`. Another important web phenomenon appeared in their study: a set of web pages served from the same webserver that seemed to have different content, based on automatic content change detection, but in fact formed of a set of disjoint adult phrases that updated frequently. This type of "apparent" change is an

extreme case of a persistent problem in assessing meaningful rates of change in page content. In this case, the content of the page actually changed (such content as there was), but in other cases the same automatically generated text problem may appear in surrounding content such as ads, links to commercial pages, or tables of contents. In all of these cases, the page appears to change, but the new content may not need to be crawled from a user perspective.

These two studies focused primarily on detected changes in the content of the page – either the raw file content or the text properties of the document, and in the status of the page. Other studies have used the "last-modified" date as recorded from the web server [5], [8] or changes in specific pieces of information on the page such as telephone numbers, HREF links, IMG tags or etc [1].

### 2.2 Estimating Average Rate of Change

Given that an understanding of when a page changes is important to maintaining a fresh corpus efficiently, but that many pages do not change frequently, several other works have sought to accurately estimate the average time between page updates based on periodic or variable-time samples. In almost all cases, a page is not crawled at the granularity at which changes could potentially occur, and therefore a fixed sample rate may be asymptotically biased for a given actual rate of change. Cho and Garcia-Molina [3] describe this problem of bias, and compute an asymptotically less biased estimator of rate of change for regularly sampled data. The same work also defines a Maximum Likelihood Estimator for the case where data is sampled irregularly, a question of practical importance for search engine crawlers because, as Fetterly et al [7] observed in their work, the delay between crawl request and page acquisition can vary widely for different web servers and for different retry policies by the crawler. This estimation model depends heavily on the assumption that the number of page changes which arrive in any time interval for a given page is a Poisson process, with a stationary mean that can be estimated from data. Matloff [8] derives an estimator similar to the MLE of Cho and Garcia-Molina but with lower asymptotic variance, and extends the work in a different direction: toward cases where the underlying distribution of changes is not Poisson.

Both of these papers focus primarily on the Poisson model or other smooth distributional models in the case where some training period of data has been observed. By contrast, our methods consider the case where a new page is added to the awareness of the crawler without any previous record, and the page must be scheduled for refresh in a sensible way. Because new pages created on the web frequently contain links to other new content, or indicate a hot new topic, crawling these pages correctly from the start is important for user facing content as well as acquisition of new content.

## 3. Estimators for Rate of Change

### 3.1 Definitions

For the $i$-th web page in the crawl, the number of changes in the $j$th crawl interval, of length $t_{ij}$, is $x_{ij}$ $Poisson(\lambda_i t_{ij})$. The change-period or average time between changes is $\Delta_i = 1/\lambda_i$. In the data observed by a crawler, we observe only $z_{ij} = Indicator_{x_{ij}>0}$, due to the fact that the crawler can only discern that the re-sampled page matches or does not match the page collected in the previous sample. Any additional changes during the time window between samples are invisible to the crawler. As a result, the collected data is censored no occurrence counts greater than 1 can be accurately observed for a single interval.

### 3.2 Simple Estimators

The simplest way to approach this problem is to assume that the crawler samples each page at equal intervals every time, such that the cij are equal for all i and all j. The nave estimator for $\lambda$ [9] is $\hat{\lambda_i} = (\sum_j z_{ij})/(\sum_j t_{ij})$. Figure 1 shows the problem with the nave estimator under censored data. The first row of the diagram shows the updates of a single URL over time, and the second two rows show crawl arrivals of crawlers that sample at a fixed interval of $C$ and $2C$ respectively over a total time of $T = 8C$. Given complete data, we would estimate $\hat{\lambda_i} = 7/T$. However, given censored data, Crawler 1 using the nave estimate will estimate $\hat{\lambda_i} = 6/T$, and Crawler 2 will estimate $\hat{\lambda_i} = 4/T$. As a result, the estimates are persistently biased with respect to the true value as a function of the true value and the size of the crawl interval.

### 3.3 Estimators for Censored Data

Cho and Garcia-Molina (2002) derive a Maximum Likelihood Estimate (MLE) for the case of a regular crawler with interval C that has significantly smaller bias than the nave estimator for larger ratios of $C/\Delta$ and a reasonably small sample size. Their estimator is created in terms of the ratio of $C/\Delta$ and is given by

$$\hat{r_i} = \frac{\Delta_i}{C} = -log\Big(\frac{\sum_j (1 - z_{ij}) + 0.5}{n + 0.5}\Big), \tag{1}$$

where $n$ is the number of intervals of length $C$ observed over the total time.

Cho and Garcia-Molina (2002) also propose an irregular sample MLE for the case of a fixed set of training samples on each page. For this estimate, we denote the length of crawl intervals where a change was observed as $t_{i,c(l)}$ for $l = 1, L$ (the total number of changed intervals observed), and $t_{i,u(k)}$ for $k = 1, K$ are the set of intervals for the $i$th page where no change was observed. Then solving for $\lambda_i$:

$$\sum_{l=1}^{L} \frac{t_{i,c(l)}}{e^{\lambda_i t_{i,c(l)}} - 1} = \sum_{k=1}^{K} t_{i,u(k)}. \tag{2}$$

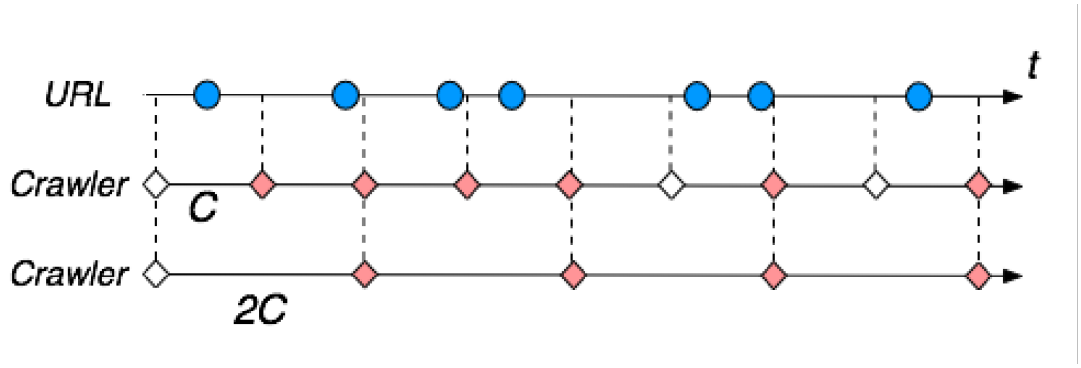


Figure 1: Illustration of two different fixed crawl cycles over time, where each blue circle is a page update, and a filled diamond is an observation of $z_{ij} = 1$

Asymptotic performance of this estimator depends on assumptions about the distribution of crawl intervals $t_{ij}$. The estimator is also undefined at two key points: if no changes have yet occurred and if changes have occurred in all sampled intervals. In these cases, Cho and Garcia-Molina recommend a sensible substitute: if the page has never changed, use the sum of all time intervals sampled so far, and if the page has always changed, use the minimum interval sampled so far.

There are two important situations where irregular samples are an important component of this estimation problem. First, a web crawler may be massively optimized, and have multiple threads competing for a fixed amount of bandwidth. Similarly, a single web server may be slow or reach a maximum allowable limit of crawler fetches. As a result, individual fetches of a web page may be delayed in an unpredictable way based on the overall system. The second situation is where we may intentionally alter the sampling intervals in an effort to establish the best estimate of rate of change or other factors.

## 4. Estimation in an Evolving Corpus

The methods considered so far assume that estimation is being done at the end of a set of training data. However, for our problem we want to consider the case where the corpus of web pages known by the estimator is continuously evolving. Old pages may drop out of the estimation, and more importantly, new pages appear. In this setting, we have three new goals: First, new pages are important and we want an estimation technique that performs well for brand-new content as well as for well-sampled content. Second, although we can't always rely on a perfectly regular sample, the sampling of pages is ongoing and we can control future samples at a per-page level. Finally, because a search engine index is large and continuously updating, we need an estimation mechanism that is computationally cheap - similar in computational cost to the implementation of the irregular sample MLE above (or cheaper).

### 4.1 Existing Estimators

If a page has never been seen before, it automatically triggers the "edge" case of the irregular-sample MLE given in Equation 2. In Figure 2, we simulate the estimator's path over 50 crawl sample epochs for a page with true $\lambda = 1/24$, using 50 different starting crawl intervals from 0 to 700 hours. The length of the sample interval for epoch $j + 1$ is defined as 80% of the estimate $\hat{Delta}_{ij}$, that is, the next sample interval is arbitrarily scaled to be 80% of the estimate made by after previous interval. Our first observation is that the number of samples to accuracy depends heavily on the initial interval. For the second observation, we highlight each estimate that encounters an edge case: blue points are "no change observed" and red are "all changes." We note that most of the first 10-15 epochs of estimation are in one of the edge cases. These percentages depend heavily on the relative size of the initial crawl interval to the actual rate of change of the page.

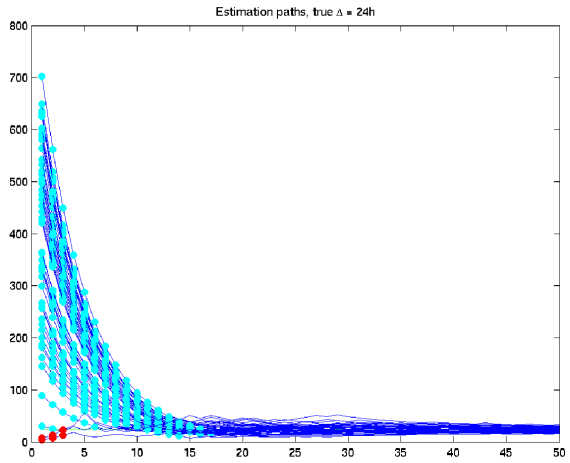


Figure 2: 50 simulated estimation paths using the irregular sampled MLE

### 4.2 Solving the Initial Interval Problem

Our first problem is to choose an initial interval such that the estimation sequence has the best chance of arriving at the correct solution, given the likely distribution of rates of change. In order to get a test corpus, we sample a set of 10,000 web

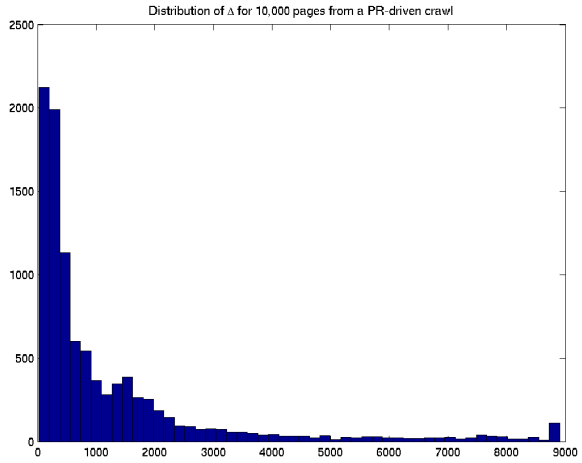


Figure 3: 50 simulated estimation paths using the irregular sampled MLE

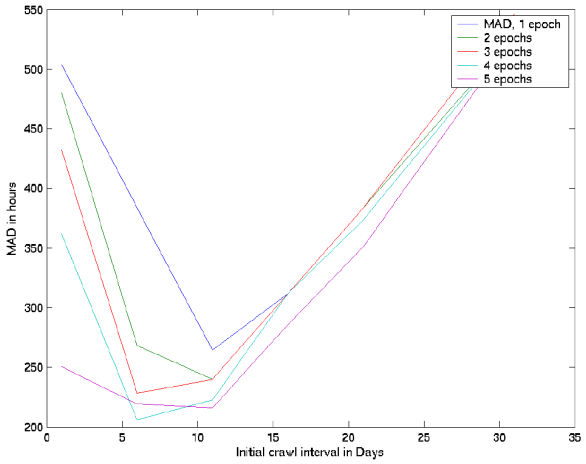


Figure 4: Median Absolute Deviation of initial 5 crawl epochs over a range of initial intervals

pages using a pagerank directed web crawler. For each page, we re-sample the page under a fixed interval refresh policy until we observe at least 25 changes. We apply the estimator given in 2 to estimate the rate of change at the end of that period, and plot the distribution of estimated $\hat{\lambda}$.

We use this sample of rates of change to test the optimal initial interval over a series of initial estimations. From the distribution of the sampled rates of change, we choose a set of 1000 $\hat{\lambda}_i s$, and use these as the true values of $\lambda_i$ to simulate a set of Poisson observations for each page over the first 5 crawl epochs, following the same sequential interval policy as in Figure 2. In this case, however, all pages start with the same initial interval, and for each page, we try a series of initial intervals between 0 and 25 days, and repeat the estimation. Over the set of 1000, the median absolute deviation between the estimated value and the true value is computed, and plotted in Figure 4. The results show a clear difference in aggregate behavior over the possible initial intervals. A value near 11 days gives the best aggregate results over the first 5 crawl epochs.

### 4.3 Applying a Prior Distribution

Although the initial interval with lowest MAD gives us a better guess for this distribution than choosing randomly, we can also apply prior information about the distribution directly. The MLE for irregular data (Equation 2)) is undefined for the edge case where no change is observed because the average changerate interval most likely to cause that result is an infinite change period. Therefore, the likelihood is monotonically increasing in $\Delta$, and has no defined maximum.

We can address this problem by applying a prior distribution over $\lambda$. Due to the censored data in the likelihood, where only $z_{ij}$ is observed, not $x_{ij}$, there isn't an easy analytic approximation to the posterior distribution, even if we choose a parametric prior. However, we can empirically apply the observed prior, $p(\Delta)$ shown in Figure 3 at a 1-hour discretization to produce an approximate posterior distribution, of $f(\lambda_i|z_{i,j})$. The resulting likelihood is shown on the right side of Figure 5 for the case where no change has been observed over a fixed number of crawl epochs.

The difficulty with this approach is that computing the posterior expectation, $\hat{\lambda_i}^* = E(\lambda_i \| z_{ij})$, requires additional storage of the computed posterior for each estimation step, and then a maximization computation. These steps make the process significantly more computationally expensive for each epoch of estimation.

### 4.4 Approximating the Prior

Although approximating and maximizing over the posterior likelihood is expensive we can approximate the solution in a different way: by inserting fake initial data points into each page sample that approximate the behaviour induced by the prior distribution. This method gives an additional advantage by removing any edge cases from the original MLE since at all times we have several initial data points.

To add a prior to the computation, two intervals are added to the data for every url: one "changed" interval and one "unchanged" interval. The interval lengths are chosen so that the posterior given just these two intervals is a best possible fit to the empirical change-rate distribution across all urls. This guarantees that the final posterior given any further data for a particular url behaves as if this approximation of the global empirical distribution were given as a prior.

The optimization of the interval lengths is performed with a Metropolis-Hastings Markov chain with simulated annealing, minimizing the L2 distance between the posterior and empirical distributions. On successive steps of the Markov chain, the interval lengths are randomly changed and this change either kept or undone. If the resulting posterior distribution is closer to the empirical distribution then the new interval lengths are kept. If the distributions are further apart then the change is kept with a probability depending on how much worse the match is. This probability of keeping a worse match allows escape from local minima. Adding simulated annealing reduces

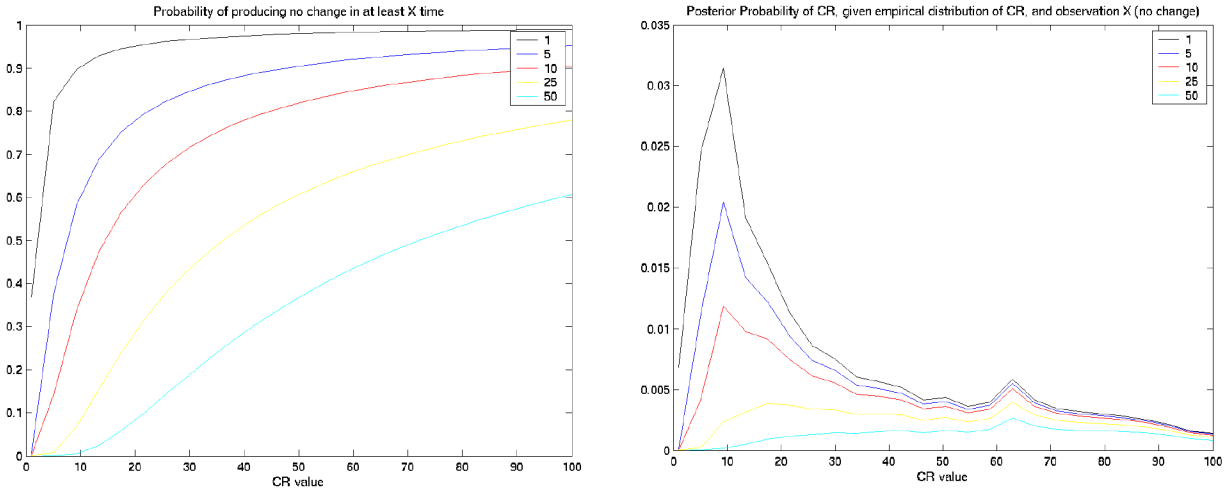


Figure 5: Left: likelihood Right: Posterior likelihood

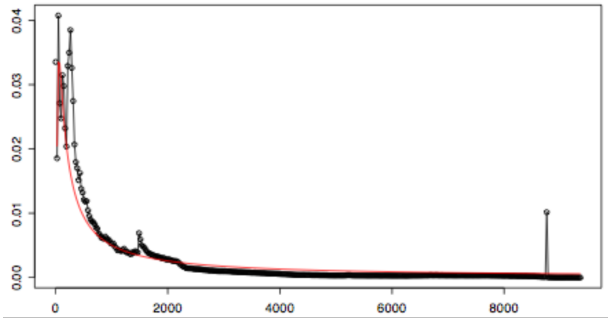

Figure 6: Black line, true prior from sample. Red line, approximated prior using two initial data points.

this probability towards zero over many steps, so that the final result is a local (and hopefully global) minima. Restarting the process many times and from different points gives a measure of confidence that the final answer is close to the global minima.

Figure 6 shows the final match between the initial sampled prior and the final induced prior using two data points.

## 5. Test Results

We test our approximated prior using a set of 1000 simulated change histories based on 1000 $\lambda$s sampled from an additional set of 10,000 actual page histories, created in the same way as the initial sample for the prior. The estimation is done in the following way:

1. For each page, estimate the initial crawl interval using the two "fake" data points, and the irregular-sample MLE
2. Re-crawl the page on the initial interval, and observe whether a change occurred
3. Update the estimate
4. Crawl the page again at interval $C_{j+1} = \hat{\Delta_i}, j$

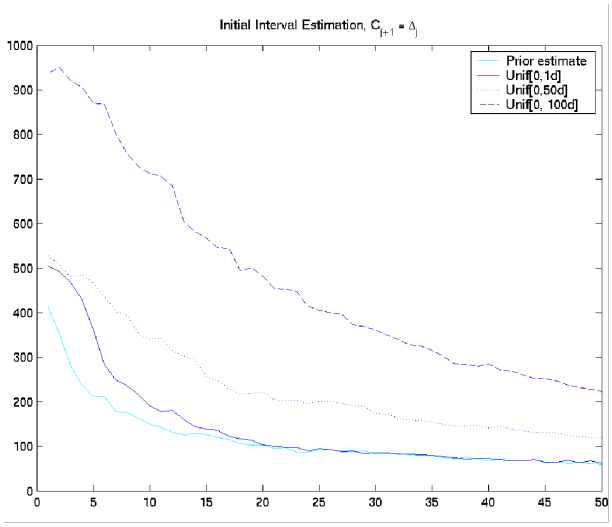


Figure 7: Comparison of the approximate prior estimate and choosing a uniformly sampled initial interval

5. repeat from step 2

We then compute the median absolute deviation over the 1000 simulated histories at each crawl epoch using all previous samples in the history. The approximated prior shows a MAD of 100 hours less at the initial interval than a uniformly sampled initial interval over 1 day or any larger interval.

## 6. Discussion and Extensions

Although the use of a prior distribution depends on the assumption that the new pages observed will correspond in aggregate to the same general distribution as the prior, this assumption can also be employed to our advantage. In many cases, additional information may be available about the page when it is first seen, such as the pagerank, the page type (news, blog, static content), or topic cluster of the page's content. That information can be used to more correctly tailor the ref-

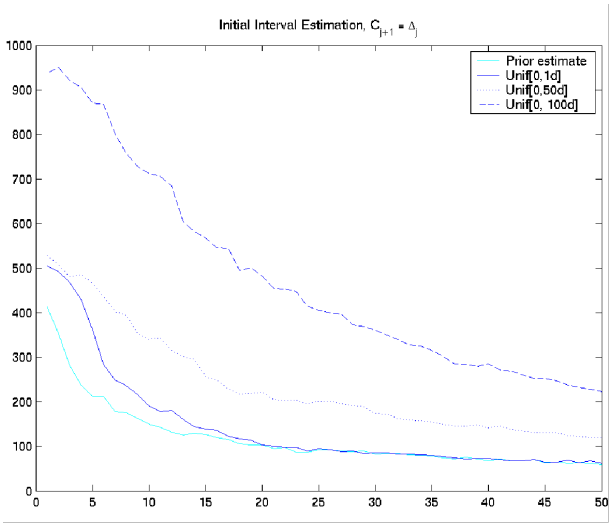


Figure 8: PLACEHOLDER

erence distribution, and the same mechanism can be followed to approximate the prior distribution.

## References


[1] B. E. Brewington and G. Cybenko. How dynamic is the Web? *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):257–276, 2000.

[2] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[3] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM Trans. Inter. Tech.*, 3(3):256–290, 2003.

[4] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.

[5] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USITS'97: Proceedings of the USENIX Symposium on Internet Technologies and Systems on USENIX Symposium on Internet Technologies and Systems*, pages 14–14, Berkeley, CA, USA, 1997. USENIX Association.

[6] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 669–678, New York, NY, USA, 2003. ACM Press.

[7] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. *Software: Practice and Experience*, 34(2):213–237, 2004.

[8] N. Matloff. Estimation of internet file-access/modification rates from indirect data. *ACM Trans. Model. Comput. Simul.*, 15(3):233–253, 2005.

[9] H. Taylor and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, 2004.