How dynamic is the web?*

Brian E. Brewington
George Cybenko
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire 03755-8000

brian.e.brewington@Dartmouth.edu
george.cybenko@Dartmouth.edu

January 29, 2000

Abstract

Recent experiments and analysis suggest that there are about 800 million publicly-indexable web pages. However, unlike books in a traditional library, web pages continue to change even after they are initially published by their authors and indexed by search engines. This paper describes preliminary data on and statistical analysis of the frequency and nature of web page modifications. Using empirical models and a novel analytic metric of "up-to-dateness", we estimate the rate at which web search engines must re-index the web to remain current.

Keywords: web dynamics, monitoring, document management

1 Introduction

Since its inception scarcely a decade ago, the World Wide Web has become a popular vehicle for disseminating scientific, commercial and personal information. The web consists of individual pages linked to and from other pages through Hyper Text Markup Language (HTML) constructs. The web is patently decentralized. Web pages are created, maintained and modified at random times by thousands, perhaps millions, of users around the world.

Search engines are an index of the web, playing the role of traditional library catalogs. However, a book or magazine does not change once it is published, whereas web pages typically do. Therefore, web search engines must occasionally re-visit pages and re-index them to stay current. This is a constant challenge considering that recent empirical studies by Lawrence and Giles [LG99] have estimated the size of the publicly-indexable web to be at least 800 million pages (and climbing). The size of the web is only one factor in the re-indexing problem; the rate at which pages change is equally important.

This paper starts with a description of our observational data on the rates of change for a large sample of web pages. Based on this data, we develop an exponential probabilistic model for the times between individual web page changes. We further develop a model for the distribution of the change rates defining those exponential distributions. These two estimates can be combined to answer questions about how fast a search engine must re-index the web to remain "current" with respect to a novel definition of currency. We introduce the concept of (α, β)-currency which defines our notion of being up-to-date by using a probability, α, that a search engine is current, relative to a grace period, β, for a randomly selected web page.

*This research was partially supported by AFOSR grant F49620-97-1-0382, DARPA grant F30602-98-2-0107 and NSF grant CCR-9813744. Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

Our observational data is based on statistics gathered from over two million web pages specified by over 25,000 users of a web clipping service [Inf95]. We have observed pages at a rate of about 100,000 pages per day, for a period of over seven months, recording how and when these pages have changed. The data indicate that the time between modifications of a typical web page can be modeled by an exponential distribution, which is parameterized by the rate of changes for the page. Our data further indicate that the reciprocal of that parameter, which is the expected time between changes, is well-modeled by a Weibull distribution across pages.

As a measure of how up-to-date a search engine is, we develop the precise concept of (α, β)-currency of a search engine with respect to a changing collection of web pages. Loosely speaking, the search engine data for a given web page is said to be β-current if the page has not changed between the last time it was indexed and β time units ago. In this context, β is the "grace period" for allowing unobserved changes to a web page. A search engine for a collection of pages is then said to be (α, β)-current if a randomly (according to some specified probability distribution) chosen page in the collection has a search engine entry that is β-current with probability at least α.

To get an intuitive feeling for this concept, we might say that a daily newspaper is (0.90, 1 day)-current when it is printed, meaning that the newspaper has at least 0.9 probability of containing 1 day current information on topics of interest to its readers (this reader interest is the specified probability distribution). Here 1 day current means that events that have happened within the last day, namely the grace period, are not expected to be reported and we "forgive" the newspaper for not reporting them. Similarly, hourly television news would be (0.95, 1 hour)-current and so on. The idea is that we are willing to "forgive" an index or source if it is not completely up-to-date with respect to the grace period, but we have a high expectation that it is up-to-date with respect to that time.

Our empirical analysis of web page changes is combined with existing estimates of the web's size to estimate how many pages a search engine must re-index daily to maintain (α, β)-currency of the entire indexable web. Using 800 million documents [LG99] as the size of the web, we show that a (0.95, 1 week)-current search engine must download and index at least 45 million pages a day, which would require a bandwidth of around 50 megabits/second (using an average page size of approximately 12 kilobytes and assuming uniform processing). A (0.95, 1 day)-current search engine must re-index at the rate of at least 94 million pages daily, or 104 megabits/second. Our results allow estimation of re-indexing rates in order to maintain general (α, β)-currency of a web index.

Previous work on web page change rates has addressed the effect changing pages have on cache consistency [DFKM97]. The metrics used there focus on the effect of dynamics on web caching, rather than on the web page change dynamics themselves. For example, [DFKM97] uses a web page "change ratio," defined as the number of accesses to a changed page divided by the total number of accesses.

Our work also concerns the performance of a search engine in maintaining a web index. In [CLW97], a formal proof is given for the optimal sample period for monitoring a collection of pages that change memorylessly, under certain sampling conditions. Optimality is measured by a sum of total time out-of-date for pages in the index, where each term is weighted by expected time between page changes. Our measures are similar in spirit, but introduce a temporal and probabilistic relaxation of what it means to be up-to-date, namely the concept of (α, β)-currency.

2 Collecting web page change data

Since early 1996, we have maintained a web clipping service called "The Informant"[1] that downloads and processes on the order of 100,000 web pages daily. The service monitors specific URLs for changes, and also runs standing user queries against one of four search engines[2] at specified intervals. Any of three events trigger a notification of a user by email. The user is notified by email if (1) a monitored URL changes, (2) new results appear in the top results returned by a search engine in response to a standing query, or (3) any of the current top search results shows a change. A change, for our purposes, is any alteration of the web page, no matter how minor.

[1] http://informant.dartmouth.edu
[2] AltaVista, Excite, Infoseek, and Lycos

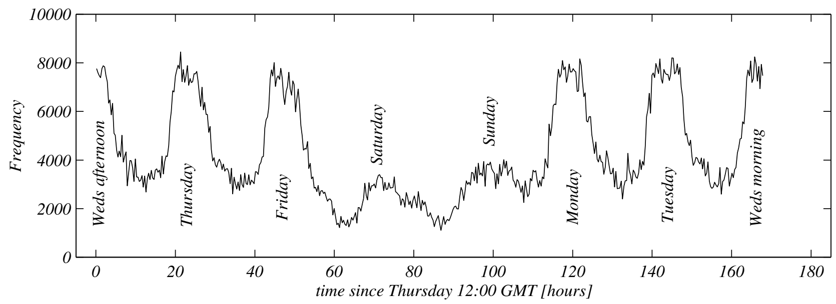


Figure 1: **Histogram: Last-modified times (GMT), mod 24×7 hours.** Peaks in modification frequency are clearly visible during US working hours, and diminish on weekends. Assumpitons of stationarity in page alteration probability will break down at this scale.

Beginning in March 1999, we started archiving HTML page summary information for all downloads. As of this writing, this has involved the download and processing of over 200 gigabytes of HTML data. The archived information includes the last-modified time stamp (if given), the time of observation (using the remote server's time stamp if possible), and stylistic information (number of images, tables, links and similar data). The Informant selects and monitors web pages in a very specific way, so conclusions from the data must be interpreted only after knowing our sampling methods.

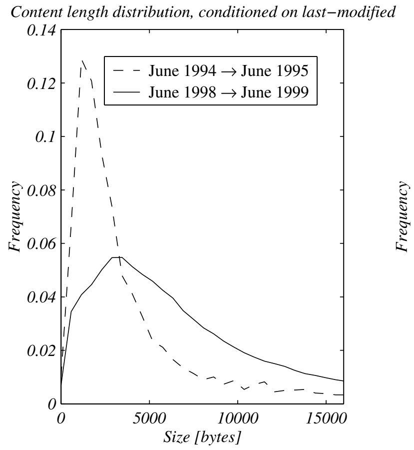
Since the Informant makes repeated observations of only those pages ranked high by search engines, this biases against those pages which are not relevant to our users' standing queries. Our sample is also biased towards the individual user-selected URLs which have been deemed worth monitoring. While neither of these is crippling, they do color our results by being slanted towards those pages that our users wish to monitor. We do not claim that this bias is a popularity bias, since our users' queries are not necessarily the same as those which are of general interest.

Another important consideration is the sample rate. Standing queries are run no more often than once every three days for any single user, and some users' queries are run once every seven days or more. Therefore, the only way a page is observed more than once every three days is if it is needed by a different user on each of those days. A number of popular sites (news sites, shareware distributors, proficient "keyword spammers") fall into this category. Moreover, to keep our service from annoying providers of popular content, we cache pages (and delete the cache prior to gathering each day's results), so no more than one observation is made of a single page per day. In addition, since we run our queries periodically and only at night, sample times for any given page are correlated.

Many monitored sites exhibit a partial overlap between users, resulting in observations being made at irregular intervals. For extremely fast-changing pages, it is quite possible that many changes will occur between observations, making direct observation of all such changes impossible. When `LAST-MODIFIED` information is given in the HTTP header, we can work around this by estimating change rates from ages. This will be discussed in greater detail in later sections.

While `LAST-MODIFIED` information is available for around 65% of our observations, the absence of such information does seem to indicate a more volatile resource. Specifically, not having this timestamp makes an observation of any given resource about twice as likely to show a modification. Therefore, estimates of change rates based solely on pages that provide a timestamp are lower bounds (slowest estimate). Timestamps also show, indirectly, that most webpages are modified during the span of US working hours (between around 8 AM and 8 PM, Eastern time). This is shown in Figure 1. This is where any assumption of stationarity in change probability will break down; modifications are less likely during the low times on this plot.

Not surprisingly, there is a correlation between the style of a webpage and its age. For example, in Figure 2, we show how the distribution of content-lengths and number of images depends upon age. Each plot shows two distributions, one using data from pages last modified between 6/94 and 6/95, and the other using pages between 6/98 and 6/99, to show how newer pages are frequently longer and have more images. Both distributions in the figure argue for the importance of space-saving technology (such as compression



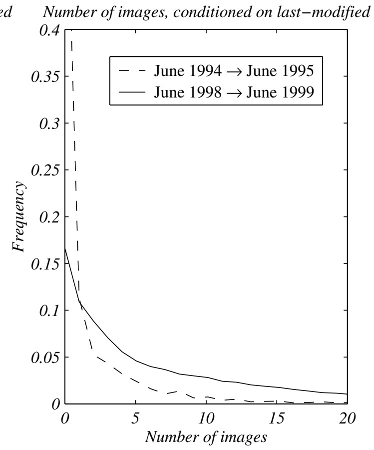


Figure 2: **Stylistic clues to webpage age.** On the left, we show two distributions of content-length, or the number of bytes in a webpage. One is for pages dated between 6/94 and 6/95, and the other is for pages last modified between 6/98 and 6/99. Widespread use of space-intensive scripting languages and stylistic elements (`<FONT>` tags, precise table and image sizing, and so forth) has driven the content length upwards. On the right, a similar trend is seen in the number of images, often used in more recently modified pages to make a more visually appealing presentation. Much of this reflects the shift from an academic-centric web to a commercial-centric one.

techniques written into the HTTP-1.1 standard, cascading style sheets (CSS), and use of Extended Markup Language (XML) where appropriate). Similar trends, sometimes much more pronounced, are seen in the usage of second-generation tags, such as the `<TABLE>` and `<FORM>` tags. While it might be feasible to use stylistic cues to estimate ages for pages which do not provide a timestamp, a far better solution is for content providers to include one along with an estimated expiration time. This potentially has many benefits, including better cache performance and fewer wasted observations by search engines (if honesty in expiration estimation is enforced).

A popular question regarding our data is, "What about dynamically-generated pages?" We can determine an upper bound on what percentage of pages are dynamic by looking at how many pages change on every repeat observation. Following [DFKM97], we can plot a cumulative distribution function of "change ratios" as in Figure 3. As mentioned in the introduction, a change ratio is defined by the number of changes observed, divided by the number of repeat accesses made. Obviously, this statistic depends heavily upon the sample rate, but it does give a feeling for the distribution of change rates. We have plotted change ratios corresponding to pages which had been observed six times or more. A unit ratio indicates a resource that always changes faster than the sample rate, meaning it may be totally dynamic, although it may just change very quickly. The plot shows that 4% of pages changed on every repeat observation (70% of these pages did not give a timestamp), while no change was observed for 56% of pages. The average page is observed 12 times over an average of 37 days, so this portion of pages that did not change would be much smaller if the monitoring was over a longer timespan.

The difference between a downloaded page's last-modified timestamp and the time at downloading is defined as the page's *age*. Recording the ages of the pages in the Informant database allows us to make several inferences about how those ages are distributed.

Estimates of the cumulative distribution function (CDF) and the probability density function (PDF) of page age are shown in Figure 4. A few observations about these plots give insight into the distribution of document ages. About one page in five is younger than eleven days. The median age is around 100 days, so about half of the web's content is younger than three months. The older half has a very long tail: about one page in four is older than one year and sometimes much older than that. In a few rare cases, server clocks are set incorrectly, making the timestamp inaccurate. The oldest pages that appear to have correct

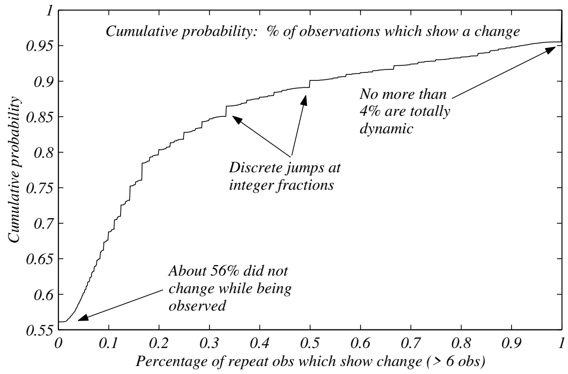


Figure 3: **Cumulative distribution of change ratios.** The "change ratio" for a page is defined as the number of changes observed divided by the number of repeat accesses. We have plotted the cumulative distribution of this statistic for pages which have been observed six times or more. This shows that no more than 4% of these pages are totally dynamic, while we have never observed any sort of change for 56% of pages. These values are very dependent upon the sampling scheme and are therefore not comprable to numbers taken from web cache-based studies.

timestamps are from around 1992, some of which are "archaeologically" interesting[3]. Our data on page age is similar to that found in an earlier study [DFKM97]; when the histograms in Figure 4 are altered so that the bins have the same size as in [DFKM97], our distribution matches their data for "infrequently-accessed" HTML pages.

Typical age observations are shown in Figures 5 and 6. Since pages are only observed for as long as they remain in any user's search results, many single pages are only monitored for a limited time. As such, no alterations are ever observed on about 56% of the pages we have monitored[4]. This type of behavior is often appears like the examples shown in Figure 5. When web pages are more dynamic, their age samples look more like the examples in Figure 6, where the pages have progressed through many changes and we have observed the ages over that time span. This usually produces distributions close to an exponential PDF. Some rapidly changing pages appear to be periodic, though the period is rarely larger than one day. Periodicity can be inferred from age distributions that appear to be approximately uniform. Still other pages are entirely dynamic, generated anew with each access, but these are not more than 4% of our collection.

3 Modeling the changes in a single page

To make further analysis possible, we model the changes in a single web page as a renewal process [Pap84]. A good example and analogy is a system of replacement parts. Imagine a light fixture into which we place a lightbulb. Whenever that bulb burns out, it is replaced immediately. We speak of the time between lightbulb failures as the "lifetime" of a bulb. At a specific instant, we define the time since the present lifetime began to be the "age" of the bulb. The analogy to web page changes is that a page's lifetime is the time between changes (where change is arbitrarily but unambiguously defined). The age is the time between a given instant and the most recent change prior to that instant. We diagram these concepts in Figure 7.

In this initial study, we assume that individual lifetimes are independent and identically distributed, and that the lifetime distribution of a particular page does not change over time (the distribution is stationary). Not surprisingly, the lifetime probability density, $f(t)$, is closely related to the age probability density, $g(t)$. The act of observing "the age is t units" is the same as knowing "the lifetime is no smaller than t units." Intuitively, this indicates that the PDF $g(t)$ should be proportional to the probability $1 - F(t)$ of a given

[3]These may not be around for long; before they disappear, see http://www.w3.org/Out-Of-Date/... hypertext/DataSources/WWW/Servers.html (a listing of web servers from 1992) or http://www.hcc.hawaii.edu/guide/... www.guide.html (a web guide from 1993)

[4]This statistic obviously depends upon the length of time we monitor a web page

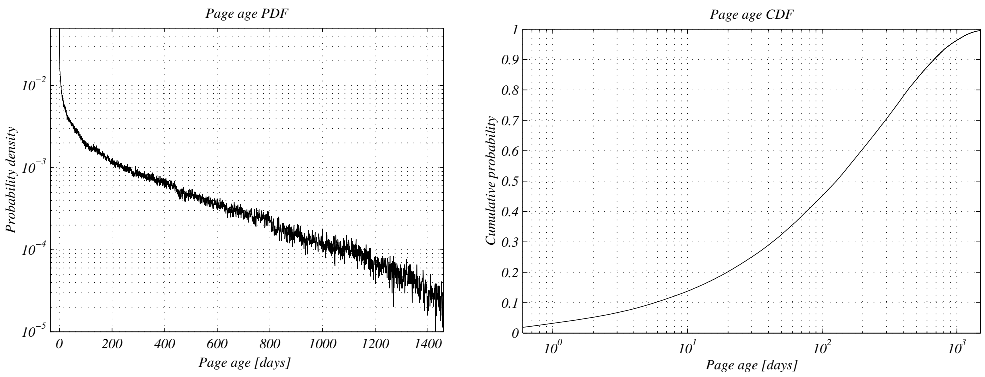


Figure 4: **Estimated distributions of web page ages.** Here we show estimates of the probability density function (PDF) and cumulative distribution function (CDF) of web page age. On the left, we estimate the PDF using a rescaled histogram of web page ages, using only one age observation per page. On the right, the corresponding CDF is formed by integrating the estimate of the PDF.

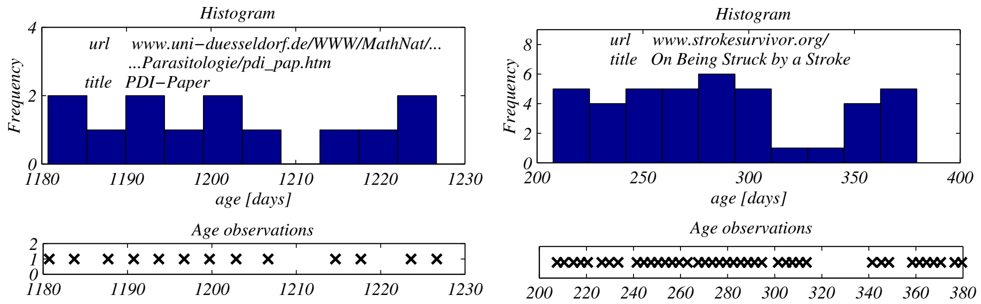


Figure 5: **Example age observations for relatively static pages.** Many of the pages we monitor do not change during the time they are observed, like the examples shown here. The upper plots are histograms, and the lower plots show the raw data. These examples show that many of the pages are quite old, and for some of them, the only change they will ever experience is their eventual disappearance.

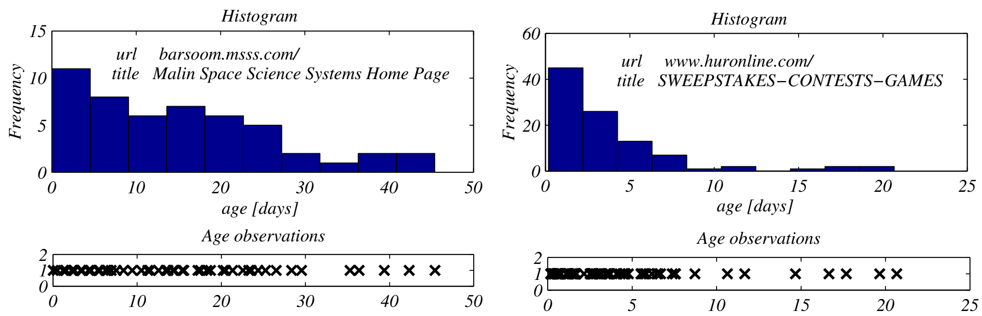


Figure 6: **Example age observations for changing web pages.** For some of our pages, we have observed a number of changes over a long timespan. The distribution of ages over this time is often approximately exponential, as can be seen in the histograms. The raw data is shown in the lower plots.

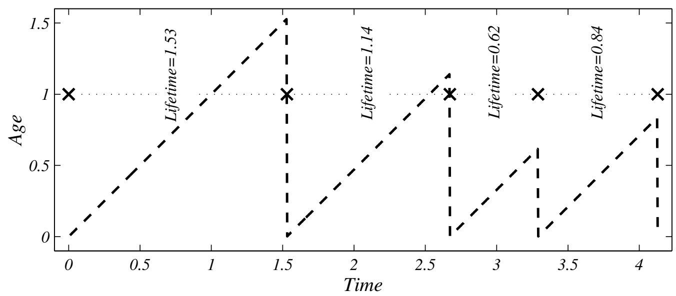


Figure 7: **Lifetimes vs. ages.** A single lifetime is represented by the time separating changes, which are denoted by $\times$'s in the graph. For each lifetime, the age (shown as a dashed line) increases linearly from 0 to the lifetime, then resets to 0 as the next lifetime begins.

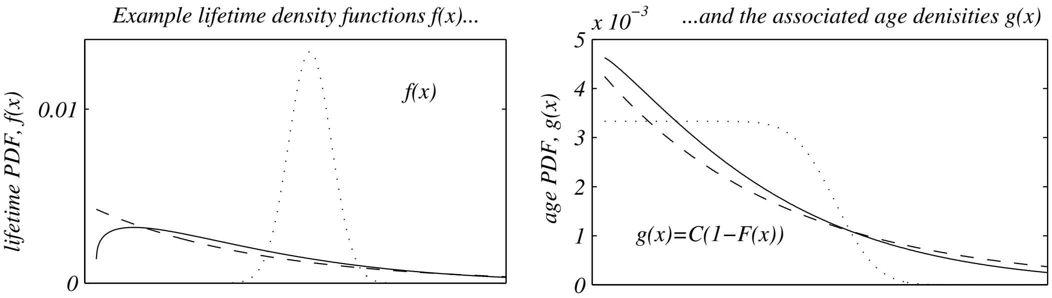


Figure 8: **Relationship between lifetime and age distributions.** On the left, we show three hypothetical lifetime distributions; a Gaussian (dotted), Weibull (solid), and an exponential (dashed). On the right, we show the corresponding age distributions. For the Gaussian, the age distribution is a renormalized and shifted complementary error function (erfc). For the exponential, the age and lifetime distributions are identical. The age distribution for the Weibull has a more general shape. Note that periodic lifetimes imply uniform age distributions.

lifetime exceeding t units, where $F(t)$ is the CDF corresponding to $f(t)$. To make $g(t)$ a proper probability distribution, the constant of proportionality is chosen so that $g(t)$ is normalized. This intuition proves correct and formal methods [Pap84] show that

$$g(t) = \frac{1 - F(t)}{\int_0^\infty [1 - F(t)]\, dt}. \tag{1}$$

Some examples of this relationship are shown in Figure 8.

Establishing the relationship of age to lifetime is useful, since it is difficult to sample the distribution $f(t)$ directly. Rather, it can be easier to estimate change rates using samples from the age distribution $g(t)$ and then use (1) to estimate $F(t)$ and then $f(t)$. Aliasing of $f(t)$ may happen when a page change is observed, since an observer can only conclude that one *or more* changes have occurred since the previous observation. In observing ages, there is no such difficulty. Avoiding the aliasing problem is not magic; we are merely making proper use of the fact that the filesystems on which the pages reside have sampled much faster than we can. Clearly, observation of a web page age requires the availability of the `LAST-MODIFIED` information, which restricts our analysis to a smaller sample.

The simplest possible page lifetime model, and a good one to use for this initial investigation, is one in which pages change memorylessly. Intuitively, this means that the probability of a page being altered in some short time interval is independent of how much time has elapsed since the last change was made. This is a common model used in queuing systems and statistical reliability theory [Pap84]. For such pages,

$f(t)$ is an exponential distribution with parameter λ. This distribution is a good choice, since much of our data on page changes show behavior like that shown in Figure 6. As for the more slowly-changing content, like the examples shown in Figure 5, it is certainly possible that these pages are not at all dynamic or that they change at a very low rate. We proceed with the assumption that all pages are dynamic, even if the only change they will ever experience is their disappearance. For these longer lifetimes, the best we can do is to obtain several (dependent) samples of the age distribution. Pages for which $f(t)$ is an exponential distribution also have exponentially distributed ages $g(t)$, since

$$\begin{aligned} 1 - F_c(t) &= 1 - \left(1 - e^{-\lambda t}\right) \\ &= e^{-\lambda t} \end{aligned}$$

implies

$$\begin{aligned} g(t) = \frac{1 - F_c(t)}{\int_0^\infty [1 - F_c(t)]\,dt} &= \frac{e^{-\lambda t}}{\int_0^\infty e^{-\lambda t}} \\ &= \lambda e^{-\lambda t}. \end{aligned} \tag{2}$$

This means we can estimate a page's lifetime PDF, assuming an exponential distribution, using only page age observations which we easily obtain from the data.

4 Dealing with a growing web

It is clear from the empirical page age distribution shown in Figure 4 that the majority of web pages are young. What is less clear is why. Different explanations can give rise to the same observed age distribution. One the one hand, a fixed population of pages whose change times are governed by identical exponential PDF's will produce an exponential age distribution when sampled collectively, as in (2). At the other extreme, an exponentially growing population of web pages in which changes are rare or even nonexistent will be skewed towards youth as well - there will be exponentially more pages in one generation relative to the previous generation.

The middle ground is an exponentially growing web in which each page changes at time intervals determined by an exponential. Such a model will also yield an exponential distribution of page ages when sampled.

Consider two very different models for the web. First, an exponentially-growing population of completely static web pages will produce an exponential distribution of observed page ages. To see this, note that the population at time t is given by an expression of the form $P_0 e^{\xi t}$ where P_0 is the initial population and ξ is the exponential growth rate parameter. An age distribution at time τ can be formed by reversing the sense of time, and normalizing by the population size:

$$g_{growing}(t,\tau) = \begin{cases} \frac{\xi e^{-\xi t}}{1-e^{-\xi\tau}} & t \in [0,\tau] \\ 0 & t \notin [0,\tau] \end{cases}. \tag{3}$$

This distribution will approach an exponential density with parameter ξ as τ gets large.

But an exponential distribution of page ages can arise for completely different reasons. Consider a fixed-size group of identical pages, each of which changes at time intervals governed by an exponential distribution. Each page undergoes many changes, with each change returning that page to age zero. Such a population also gives rise to essentially an exponential age distribution (see 2). In particular, the age distribution for such a population is

$$g_{dynamic}(t,\tau) = \begin{cases} \lambda e^{-\lambda t} & t \in (0,\tau) \\ \left(e^{-\lambda\tau}\right)\delta(t-\tau) & t = \tau \\ 0 & t \notin [0,\tau] \end{cases}. \tag{4}$$

As the time since the population's birth, τ, becomes large, the distribution of observed page ages will also approach an exponential distribution and will be hard to distinguish from that of a growing population of unchanging web pages. The hybrid model we use in this paper represents the middle ground - the web is growing *and* pages change according to exponential time distributions. These are reasonable working assumptions.

We now combine the effects of web growth and page change dynamics. The web has been growing for several years so that the time since creation of web pages is distributed approximately exponentially:

$$h(t_c) = \xi e^{-\xi t_c}. \tag{5}$$

where ξ is the growth rate and t_c is the time since creation of a page. We emphasize that t_c is not to be confused with our definition of the page's age, since age refers to the time since the last modification.

For an exponentially-growing population of dynamic pages, each of which has an exponential age distribution as described by (4), the aggregate age distribution $g(t, \lambda)$ will be a weighted average over time since creation, weighted by the number of pages created at the same time. Specifically,

$$\begin{aligned}
g(t,\lambda) &= \int_0^\infty g(t,\lambda,t_c)h(t_c)dt_c && (6)\\
&= \int_0^\infty \xi e^{-\xi t_c} e^{-\lambda t_c}\delta(t-t_c)dt_c + \int_0^\infty \xi e^{-\xi t_c}\left[U(t)-U(t-t_c)\right]\lambda e^{-\lambda t}dt_c && (7)\\
&= \xi e^{-(\xi+\lambda)t} + \int_0^\infty \xi e^{-\xi t_c}\lambda e^{-\lambda t}dt_c - \int_0^t \xi e^{-\xi t_c}\lambda e^{-\lambda t}dt_c\\
&= \xi e^{-(\xi+\lambda)t} + \lambda e^{-\lambda t} - \lambda e^{-\lambda t}\left(1-e^{-\xi t}\right)\\
&= (\xi+\lambda)e^{-(\xi+\lambda)t}. && (8)
\end{aligned}$$

This means that the age distribution of an exponentially growing population of objects with (identical) exponential age distributions remains exponential, with parameter given by the sum of the population growth and page change rate constants.

The age distribution for the entire population (namely the whole web) is yet another mixture, in which we take expectation of (8) with respect to a joint distribution of growth rate ξ and change rate λ. For simplicity we use the same growth rate for all change rates. Using a distribution over the inverse rate $\lambda = 1/x$, with this uniform growth rate ξ, we express the mixture as

$$g(t) = \int_0^\infty \left(\xi + \frac{1}{x}\right) e^{\left(\xi+\frac{1}{x}\right)t} w(x)dx. \tag{9}$$

The only factor remaining before this distribution can be matched to the data is the shape of the distribution $w(x)$ of inverse change rates. In our initial development, we use a generalized exponential (Weibull) distribution over the inverse change rate (which is also the mean change time), such that

$$w(t) = \frac{\sigma}{\delta}\frac{e^{-(t/\delta)^\sigma}}{\Gamma(1/\sigma)} \tag{10}$$

where δ is a scale parameter and σ is a shape parameter. See [MR94] for a discussion of Weibull distributions, as well as a more general discussion of this family of exponential distributions in [Fel71]. The shape parameter can be varied to change the shape from a very sharply-peaked distribution (for $\sigma < 1$) to an exponential (for $\sigma = 1$), to a unimodal distribution with maximum at some positive t (for $\sigma > 1$). The scale parameter δ adjusts the mean of the distribution.

To determine what values of ξ, σ, and δ best model the observations, we numerically evaluate (9) at a number of ages t. This is used to estimate the cumulative age distribution $G(t)$ at N points t_i. These estimates, $\hat{G}(t)$, are compared with samples from the empirical distribution $G(t)$ (as diagrammed in the left half of Figure 4) at points t_i. A sum of the squared error over all sample times t_i provides a scalar error function of the vector (ξ, σ, δ). This error function can be minimized:

$$SE_{age}\,(\xi,\sigma,\delta) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{G}(\xi,\sigma,\delta,t_i) - G(t_i)\right)^2. \tag{11}$$

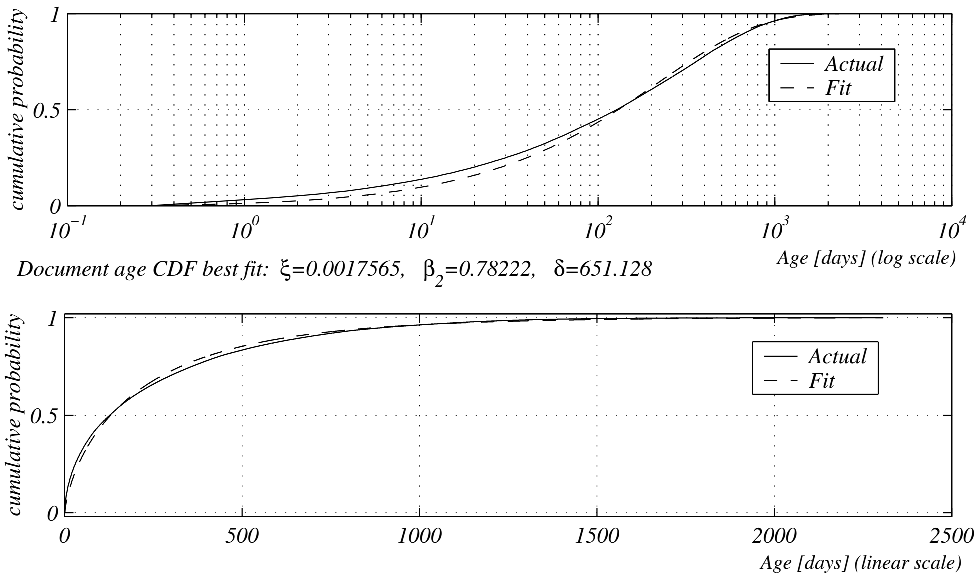


Figure 9: **Best-fit age CDF.** These plots show the distribution which results from a numerical optimization of (11), yielding the values $\xi = 0.00176$ (growth rate), $\sigma = 0.78$ (shape parameter), and $\delta = 651.1$ (scale parameter). The top plot uses a log scale to show the deviations in the fit for small age. The minimization was carried out using linearly-spaced points.

When this minimization is carried out numerically, the optimal values are found to be $\xi = 0.00176$, $\sigma = 0.78$, and $\delta = 651.1$. The fitted age distribution is shown in Figure 9. These parameters imply a steeper-than-exponential age distribution (since $\sigma = 0.78$) and a growth rate that implies a doubling time of around 390 days. This is not unreasonable, as [LG98] estimated a lower bound size of 320 million pages in December 1997, which increased in [LG99] to 800 million pages by February 1999. This would imply a growth constant over the 14 months of $\xi = 0.0022$, or a doubling time of 318 days. The difference in these estimates tells us to proceed with caution, understanding that estimates based on these results are somewhat uncertain. Moreover, the assumption of exponential growth in the number of *documents* is based on assertions of exponential growth in the number of web *hosts* (as in [Gra97] and [ISC99], for example). Growth rates have slowed appreciably, especially in the last year; other estimation methods prove more reliable.

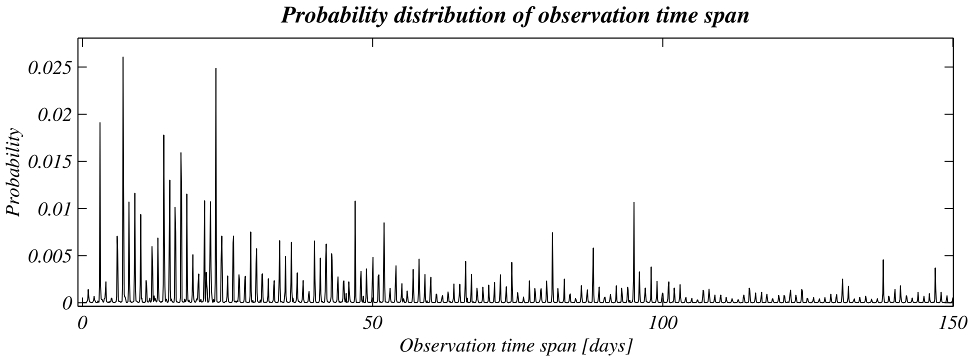
5 Estimating the change rate distribution using lifetimes

As mentioned previously, inferring change rates from observed lifetimes is somewhat tricky, since an observed change may only be the most recent of many changes that took place since the last observation. Moreover, changes that take a long time to happen are inherently more difficult to catch. For example, if one were to watch a calendar for three consecutive days, waiting for the month to change, there is a good chance that this event will not be observed. However, as the timespan gets longer it becomes more probable that a change will be seen. In the same way, it is necessary to account for the probability of observing a change, given the timespan of observation.

For a page which changes exponentially at rate λ, the probability that at least one change will be observed within a timespan τ is

$$\Pr(\text{change observed}|\tau, \lambda) = 1 - e^{-\lambda\tau}. \tag{12}$$

The pages in our collection are observed over many different timespans τ. Therefore, to determine the probability of observing changes for pages having change rate λ, we assume that change rate and timespan



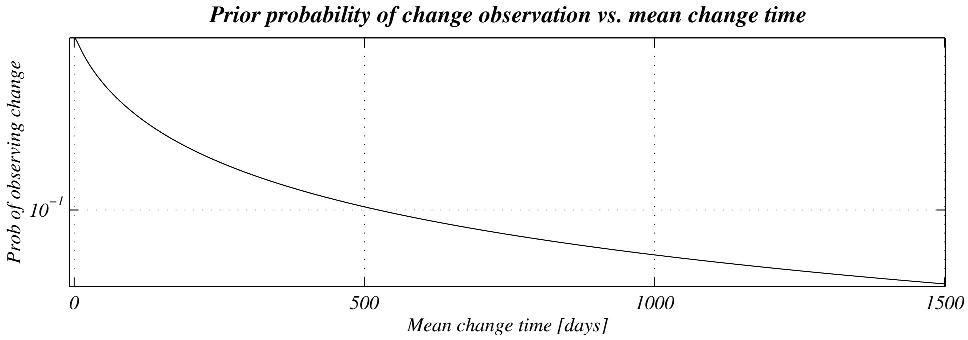


Figure 10: **Observation time distribution and induced finite time span bias.** The top plot shows the distribution of observation time spans, or the time difference between the first and last observation timestamps for individual pages. The spikes appear in this graph because we only run our checks at night, so timespans tend to cluster around 24-hour intervals. Using (13), these timespans translate into the probability of any mean change time being represented among our observed web page changes.

are independent and weight (12) with respect to the probability of all possible observation timespans τ_i (discretized):

$$\Pr(\text{change observed}|\lambda) = Z_{bias}(1/\lambda) = \sum_{i=1}^{i=N} \Pr(\tau_i)(1 - e^{-\lambda\tau_i}). \tag{13}$$

Possible timespans τ_i are distributed as shown in Figure 10. Combining this data with (13) allows us to compute Z_{bias} weighting each mean lifetime's probability of being among the observed data. The distribution of change rates sampled in our experiment is not the true rate distribution, but rather one that is weighted by (13). If the *actual* density of mean lifetimes is $f_{mean}(t)$, then the *observed* density of mean lifetimes is

$$f'_{mean}(t) = \frac{f_{mean}(t) Z_{bias}(t)}{\int_0^\infty f_{mean}(t) Z_{bias}(t) dt}. \tag{14}$$

These mean lifetimes are only seen through a mixture of exponential distributions, so the observed lifetimes should approximate the probability density

$$f_{observed}(t) = \int_0^\infty \lambda e^{-\lambda t} f'_{mean}(1/\lambda) d(1/\lambda). \tag{15}$$

As with the age-based estimates, we can form a mean squared-error function like (11) and fit the CDF corresponding to (15) to the observed lifetime distribution. We show the distribution of observed lifetimes

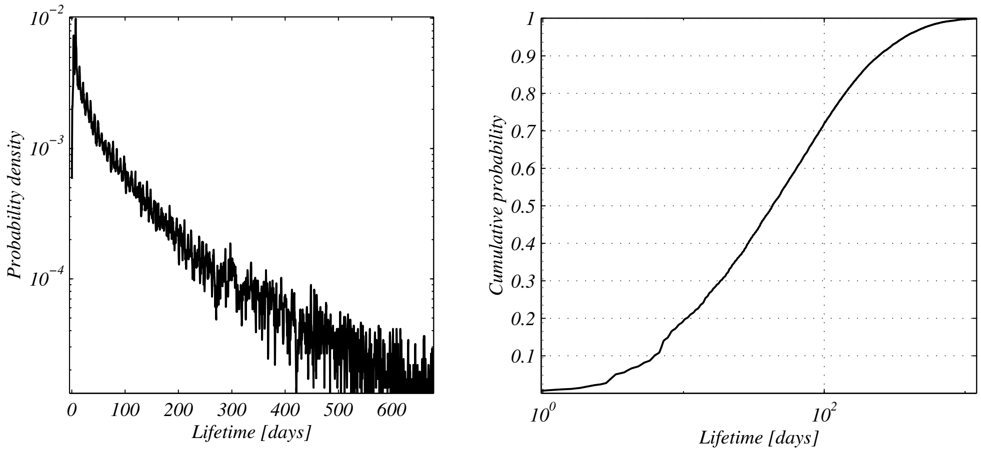


Figure 11: **PDF and CDF of observed lifetimes.** On the left, a rescaled histogram approximates the PDF of observed lifetimes, or differences in successive modification timestamps. On the right, we show the corresponding CDF.

in Figure 11. Using $F(t)$ as the cumulative lifetime distribution, and $\hat{F}(\sigma, \delta, t)$ as the estimator, the error function is

$$SE_{lifetime}(\sigma, \delta) = \frac{1}{N} \sum_{i=1}^{N} (\hat{F}(\sigma, \delta, t_i) - F(t_i))^2 \tag{16}$$

As before, we use a Weibull density (10) for the distribution of inverse rates (mean times) $\bar{t}$. This results in an error surface having a minimum at $(\sigma = 1.4, \delta = 152.2)$. An intensity plot of (16) is shown in Figure 12. The CDF and its estimator are overlaid in Figure 13, and the error in this fit is magnified in Figure 14. Using our estimates, the *mean* lifetime PDF and CDF are shown in Figure 15.

The lifetime-based estimates differ substantially from the age-based estimates, but are also more trustworthy, as can been seen by comparing the quality of the fit in Figures 13 and 9. There are two reasons for the difference. First, the assumption of exponential growth used for the age-based estimatation is probably a poor one, as true growth is much slower. Forcing exponential growth on a more slowly growing population forces the dynamics to be under-represented, driving our estimates away from their true value. The lifetime-based estimation is not perfect either, as change rates may not be independent of observation timespan. A change in a page might very well push it into or out of a user's set of search results. We count on the fact that in observing faster than the search engines, we can observe changes before these force a result from the top of the list. It is difficult to justify an assumption of any particular dependence, since this relationship is controlled by many unknown factors (re-indexing time for search engines used and result ranking strategy, for example).

6 How fast do search engines need to work

We now interpret our model of the constantly changing web in terms of web search engine performance. Our measure of performance is based on the intuitive concept of (α, β)-currency that we define below. Our web model and this new performance measure will allow us to estimate the speed at which pages must be re-indexed in order to maintain a given level of currency.

Recall from the introduction that a web page's index entry in a search engine is β-current if the web page has not changed since the last time the page was re-indexed and β time units ago. We are willing to forgive changes that have occurred within β time of the present. The grace period, β, relaxes the temporal aspect

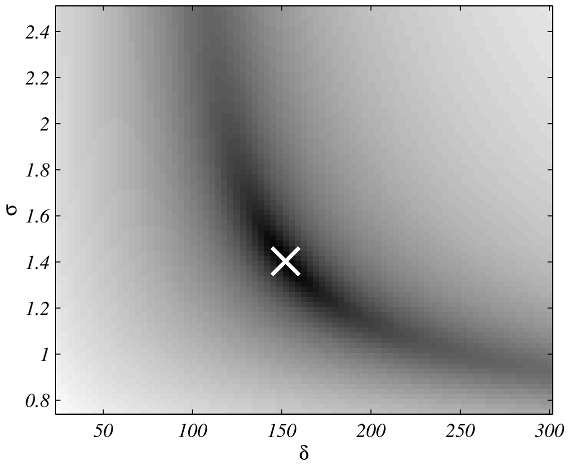


Figure 12: **Intensity plot of mean squared-error.** The minimum of (16) in the space of shape parameters σ and scale parameters δ is marked by a white "×" in the center of the dark patch, at ($\sigma = 1.4, \delta = 152.2$). The error function appears to be unimodal.

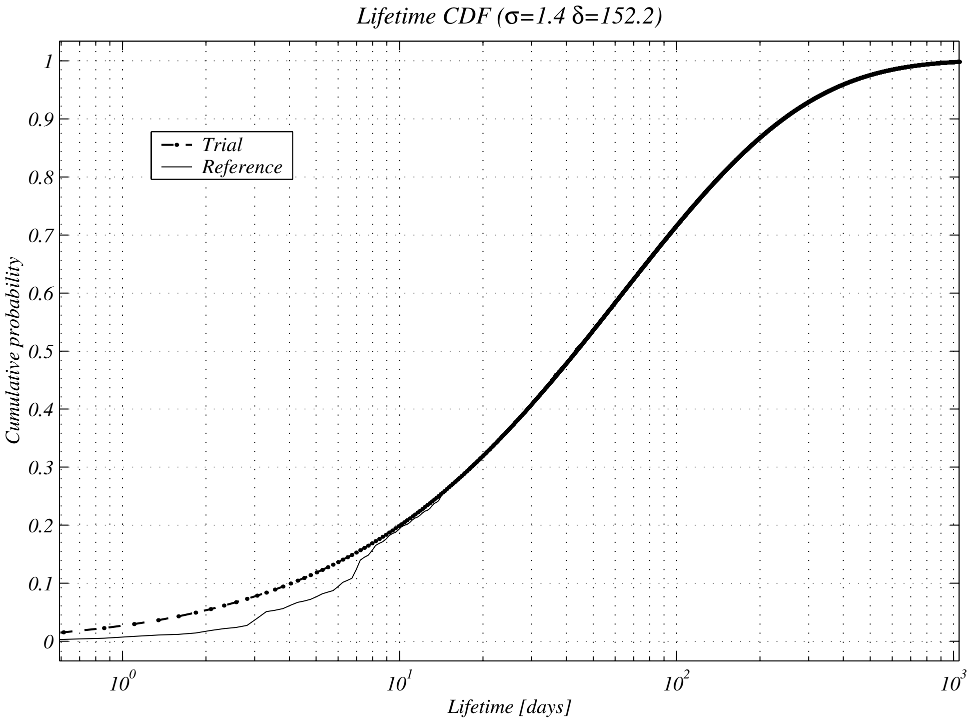


Figure 13: **Overlay of model lifetime CDF and observed CDF.** The minimum error distribution (marked "Trial" above) found by minimizing (16) is shown along with the observed lifetime distribution (marked "Reference"). Errors in the fit below around 8 days are due to aliasing, where multiple changes are masked and treated as a single, larger lifetime. Our estimates are only extrapolations in this region and may be inaccurate. The region above 8 days is an extremely precise fit; the two curves are nearly identical.

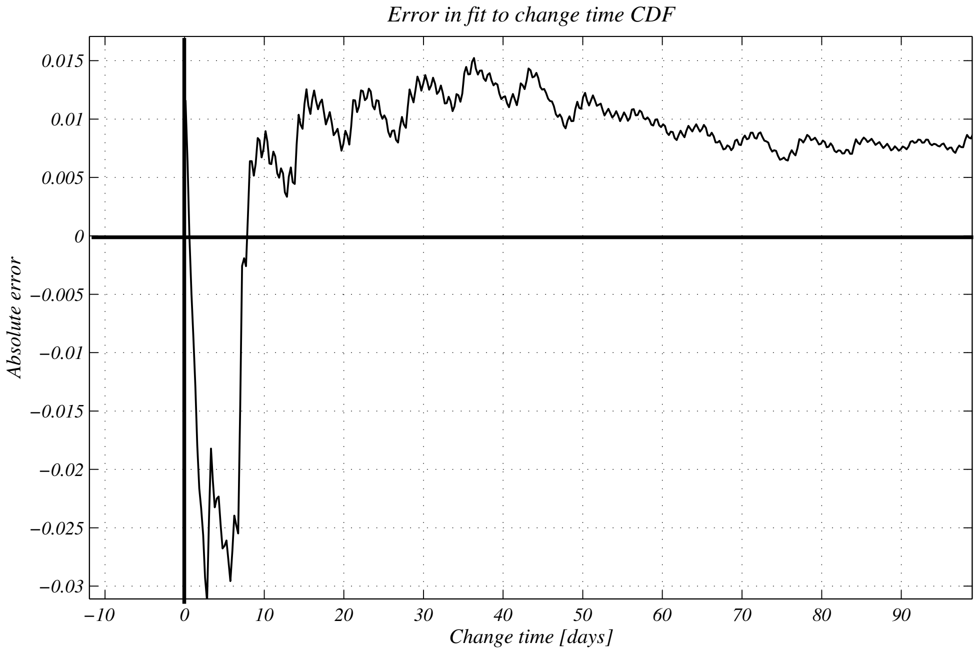


Figure 14: **Absolute error $\hat{F} - F$ vs. lifetime.** These are the errors in the fit shown in Figure 13; we have used a linear scale and just show the leftmost region. Note the large errors below around 8 days due to aliasing. The effect of diurnal and weekly trends, as plotted in Figure 1, is clearly visible in the long and short period ripples above 8 days. Slight improvements in our estimates could be had if we restricted the fit to samples above 8 days.

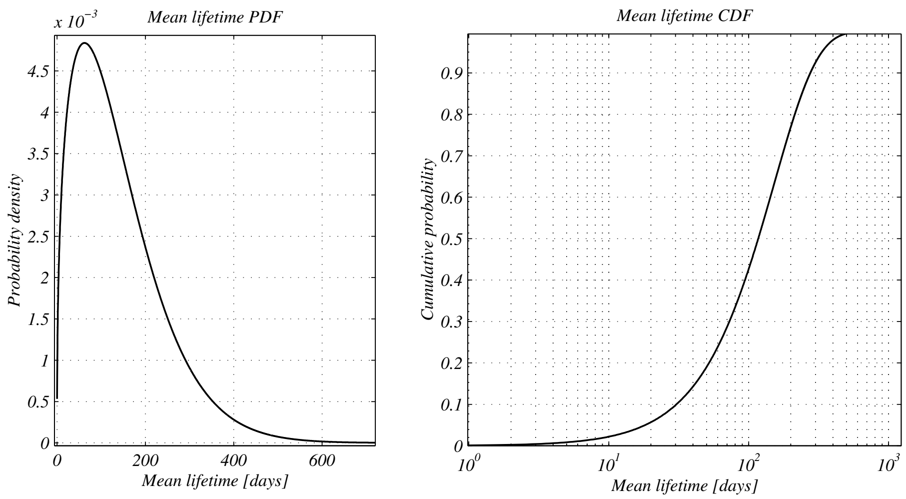


Figure 15: **Mean lifetime $(1/\lambda)$ estimated PDF and CDF.** Our lifetime-based population parameter estimation implies these distributions of mean lifetimes for the documents observed by the Informant. Note that these mean values are to be distinguished from the distribution of *observed* lifetimes. The average is around 138 days, the most likely value is 62 days, and the median is 117 days.

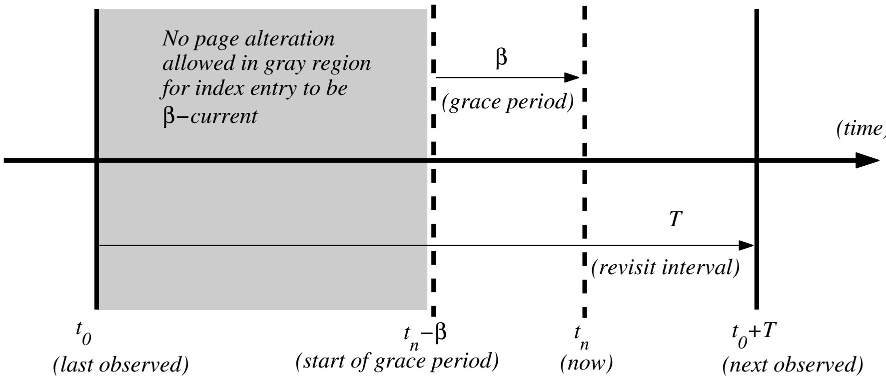


Figure 16: **Definition of "β-current".** This diagram shows what is meant when we say that an index entry is current with respect to a grace period, β. In order to be β-current, no modification can go unobserved up to β time units before the present.

of what it means to be current. The smaller β is, the more "current" our information about the page is. See Figure 16 for a graphical depiction of the concept.

To determine whether or not an index entry for a web page is β-current, we need to know the most recent time t_Δ at which the page changed. Assume that the page was last observed at time t_o. With this notation, the index entry corresponding to a page is β-current at time t_n if the page did not change between the last observation (at time t_o) and β units before the present, or time $t_n - \beta$ (assuming $t_o \leq t_n - \beta$). For $t_o > t_n - \beta$, the entry is by definition β-current because the most recent unobserved page change can occur either within the grace period or before we observed the page at t_o, but this includes all past time.

Combining these two cases, the probability that the search engine entry for a page is β-current at time $t_n > t_o + \beta$ is

$$\Pr(\text{ a fixed web page is } \beta\text{-current} \mid t_o, t_n) = 1 - \Pr\left(t_o \leq t_\Delta \leq t_n - \beta\right) \tag{17}$$

where these probabilities are understood to be for a fixed, given web page. We now compute the probability α that the search engine index entry for a randomly (according to some probability distribution) selected web page is β-current.

The above expression (17) for a single web page is stated in terms of a conditional probability. Given a prior distribution on the variables t_o and t_n, we can use Bayes' Theorem or the total probability theorem to eliminate them.

In our model, each web page has a change rate λ and an associated distribution of re-indexing times T (a periodic re-indexing system will have a single constant T_0). These parameters determine density functions which, together with the grace period β, specify the probability α of being β-current. First, define the probability Pr(a page is β-current $\mid \lambda, T, \beta, t_n)$ to be the probability of a single index entry being β-current given λ, T, β, and the time t_n at which the index is examined. Second, define the density $h(\lambda, T)$ to be the joint probability density for (λ, T). We assume that $h(\lambda, T)$ is independent of the time t_n, which is distributed according to a density $x(t_n)$. Using these densities and Bayes' Theorem, the probability α that the system is β-current is

$$\begin{aligned} \alpha &= \Pr(\text{The search engine is } \beta\text{-current}) \\ &= \iiint \left[\Pr(\text{a single page is } \beta\text{-current} \mid \lambda, T, t_n) x(t_n) dt_n\right] h(\lambda, T) d\lambda dT \end{aligned} \tag{18}$$

The integral is restricted to the first octant since no negative times or rates are allowed. In some settings, it is reasonable to assume a dependence between T and λ, since different re-visitation periods are desirable for sources with different change rates.

We will now evaluate (18) for a single, memorylessly-changing page. As before, this page has a change rate λ, and is observed periodically (every T time units). The probability that the next page change occurs in the time interval $[t_1, t_2]$, where the last observation or change (whichever occurred most recently) was at time $t_o \leq t_1 < t_2$, is

$$\int_{t_1-t_o}^{t_2-t_o} \lambda e^{-\lambda t} dt = e^{-\lambda(t_1-t_o)} - e^{-\lambda(t_2-t_o)}.$$

If $t_1 = t_o$, this reduces to $1-e^{\lambda(t_2-t_o)}$ so that the probability that a page change *did not occur* in the interval $[t_o, t_2]$ is the complement, $1 - \left(1 - e^{\lambda(t_2-t_o)}\right) = e^{\lambda(t_2-t_o)}$.

To evaluate (18) we need to specify the function $h(\lambda, T)$ as well as the distribution of times $x(t_n)$ over which we average the β-currency of the index. First, we consider the limits on the inner integral over t_n. Assuming as we have that all the web pages change memorylessly, it is sufficient to evaluate the inner integral in (18) over a single observation period T, since adding additional periods would only replicate the integral over one period.

For convenience, we choose an interval starting at $t_o = 0$, at which time an observation was last made, and extends until the time T at which the next observation occurs. Using this interval, the probability that the page does not change between $t_o = 0$ and $t = t_n - \beta$, and is therefore β-current, is

$$\Pr(\beta\text{-current } | \lambda, T, t_n) = e^{-\lambda(t_n-\beta)} \text{ for } \beta < t_n < T \tag{19}$$

by the above discussion. Further, note that the page is β-current with probability one in the interval $[t_n - \beta, t_n]$. Specifically,

$$\Pr(\beta\text{-current } | \lambda, T, t_n) = 1 \text{ for } 0 < t_n < \beta \tag{20}$$

Combining these, the expected probability of a single page being β-current over all values of the observation time t_n, using a uniform density $x(t_n) = 1/T$, is just an average value of the piecewise-defined $\Pr(\beta\text{-current } | \lambda, T, t_n)$ on the interval $t_n \in [0, T]$. This gives

$$\Pr(\beta\text{-current } | \lambda, T, \beta) = \int_0^\beta \frac{dt_n}{T} + \int_\beta^T \frac{1}{T} e^{-\lambda(t_n-\beta)} dt_n \tag{21}$$

$$= \frac{\beta}{T} + \frac{1 - e^{-\lambda(T-\beta)}}{\lambda T}. \tag{22}$$

In the first integral of (21), the probability of being β-current is one when $t_n \in [0, \beta]$, since this would force any change to be within β units of the present. We can clean up (22) by expressing β as a fraction ν of T (that is, $\beta = \nu T$) and setting $z = \lambda T$. With these changes, (22) becomes a function of the dimensionless relative rate, z, and the ratio of the grace period to the observation period, ν. When $z > 1$, a source is expected to change once or more prior to T, whereas $z < 1$ suggests fewer than one change expected before T. What fraction of these changes fall within the grace period β is loosely described by the parameter ν; some curves are shown for different choices of ν in Figure 17.

We note in passing some properties of the curves in Figure 17 that verify our intuition. First, note that the probability of being β-current goes to ν as the relative rate λT approaches infinity. High relative rate implies a web page which is observed much too slowly; the page changes many times between observations. As such, in the high rate limit, ν simply represents the percentage of these changes that occur during the grace period. For the case of low relative rate, where pages are sampled much faster than they change, the probability of a page being β-current approaches one, regardless of the grace period fraction ν.

Choosing a random web page to which we apply (22) is equivalent to selecting a value for λ. In our collections, as discussed earlier, we have observed that the mean time $\bar{t}$ between changes roughly follows a Weibull distribution, (10), which is given by

$$w(\bar{t}) = \frac{\sigma}{\delta} \frac{e^{-(\bar{t}/\delta)^\sigma}}{\Gamma(1/\sigma)}. \tag{23}$$

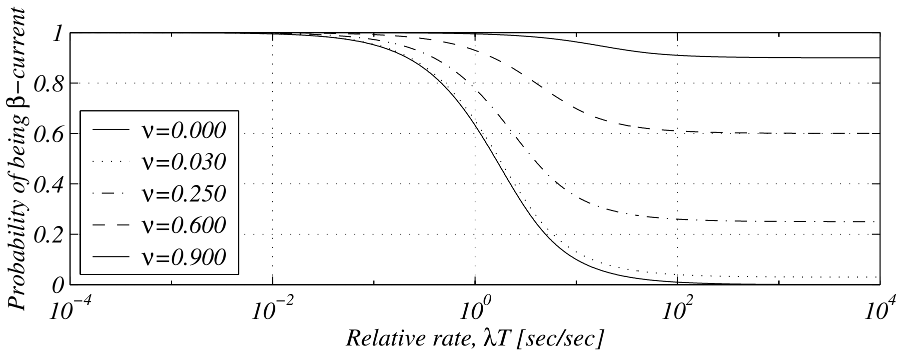


Figure 17: **Probability of β-currency vs. relative rate.** Expected value of $\Pr(\beta\text{-current} \,|(\lambda, T, \nu))$ as a function of relative rate $z = \lambda T$ and grace period percentage $\nu = \beta/T$

The change rate λ is the inverse of the mean time between changes, so we can replace λ in the integral with the change rate $1/\bar{t}$.

Using (23), along with the parameter values that resulted from our numerical optimization, we can determine the expected value of (22) over λ for our collection. This calculation for other collections or other demand distributions depends only on finding the distribution $w(\bar{t})$ of mean change times for those collections. Our analysis uses a simple periodic, round-robin re-indexing schedule, where the revistation time T is the same for all sources. Since we propose visiting each page every T time units, an accurate model for a real engine would need to account for the growth of the collection over time.

For this preliminary analysis, we assume a constant web size to avoid this difficulty. Using the Weibull distribution for inverse change rates, the expected probability α that a uniformly randomly selected page will be β-current in the search engine index is

$$\alpha = \int_0^\infty \left[\frac{\sigma}{\delta}\frac{e^{-(t/\delta)^\sigma}}{\Gamma(1/\sigma)}\right]\left[\frac{\beta}{T_0} + \frac{1 - e^{-(1/t)(T_0-\beta)}}{(1/t)T_0}\right] dt. \tag{24}$$

The integral (24) can only be evaluated in closed form when the Weibull shape parameter σ is 1; otherwise, numerical evaluation is required. The integral gives an α for every pair (T_0, β), defining a search engine "performance surface." This surface can be interpreted in a number of ways. For example, we can choose a probability α and determine all pairs (T_0, β) that give that probability. Using our parameter choices from the lifetime-based optimization of (16), we have evaluated the integral and plotted it in Figures 18 and 19, which show the level set for $\alpha = 95\%$. It is important to note that the revisitation times which result from this analysis are upper bounds since our analysis is based on the less volatile pages that provide timestamps.

From that plot, we can see that in order to maintain (0.95, 1-day)-current search engine, a re-indexing period of 8.5 days is necessary. For (0.95, 1-week)-currency, a re-indexing period of 18 days is necessary. Notice that these figures do not depend upon the number of documents in an index, so a re-indexing period defines a set of pairs (α, β), regardless of changes in the size of the index. Alternatively, we can estimate effective bandwidth requirements to maintain a given level of currency for a uniform index of a given size. By "uniform" we mean that no documents are given any sort of preference; all are re-indexed at the same rate. The effective bandwidth is not to be confused with the link bandwidth, it simply describes the overall processing rate, including download and analysis.

For example, an (0.95, 1-day) index of the entire web, using the estimate of 800 million pages from [LG99], would require a total effective bandwidth of (approximately)

$$\frac{800 \times 10^6 \text{ pages}}{8.5 \text{ days}} \times \frac{12 \text{ kilobytes}}{1 \text{ page}} = \frac{104 \text{ Mbits}}{\text{sec}} \text{ for (0.95, 1-day) currency of index of the entire web.}$$

A more modest index, closer to those actually in use, might have 150 million documents at (0.95, 1-week)

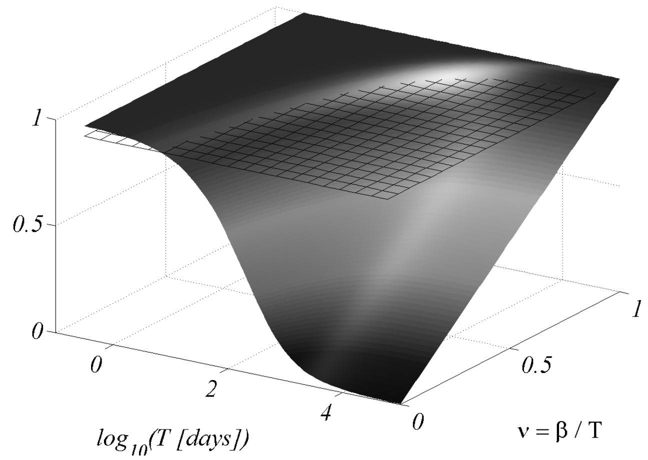


Figure 18: **Probability α as a function of ν and T_0:** Here, we plot the probability surface α as a function of the grace period fraction $\nu = \beta/T_0$ and fixed re-indexing period T_0. This surface results from using the more accurate lifetime-based population parameters, although this surface could be constructed for any population. The plane at $\alpha = 0.95$ intersects the surface in a level set, which is plotted in Figure 19 (with β values used instead of percentages ν).

currency, requiring an effective bandwidth of around

$$\frac{150 \times 10^6 \text{ pages}}{18 \text{ days}} \times \frac{12 \text{ kilobytes}}{1 \text{ page}} = \frac{9.4 \text{ Mbits}}{\text{sec}} \text{ for (0.95, 1-week) currency of index of around 1/5 of the web.}$$

Clearly, other re-indexing schemes exist where T is not constant but is a function of λ; see [CLW97] for some good discussion on possible schemes. When T is a function of λ, the integral (24) is modified by substituting in the function $T(1/\bar{t})$ and evaluating along the appropriate line in the $(T, \bar{t})$-plane. Additional modifications to this development might include the addition of a noise term to the observation period and choosing the grace period β as a function of the change rate λ.

7 Summary

This paper describes our efforts at estimating how fast the web is changing, using a combination of empirical data and analytic modeling. From here, we can begin to consider the "dynamics" of information, and how best to deal with observation of changing information sources over limited-bandwidth channels. Much work remains to be done. With a reasonable model of how the web is growing and how fast pages change, we can start to formulate scheduling problems for search engines. These scheduling problems will depend on what objective we are trying to optimize. This work has used a simple, deterministic periodic revisiting strategy. By allowing different revisit intervals for different pages, we can formulate a variety of scheduling problems, holding two of α, β and the communication resources (that is, server bandwidth) fixed for example. We have not gone into any detail about which changes are "important" and which changes are not, nor have we delved into the reliability and popularity of the web pages in question. These clearly bear heavily on a user's perception of how good a search engine performs. While we have such data available to us in our empirical database, we have not yet addressed this. How can we estimate the currency, in our formal terms of (α, β)-currency, of commercial search engines that only allow external probes? How do the different search engines compare in this sense? Indeed, the fast-changing and fast-growing web may soon force increased reliance on specialty search engines for the most volatile information sources.

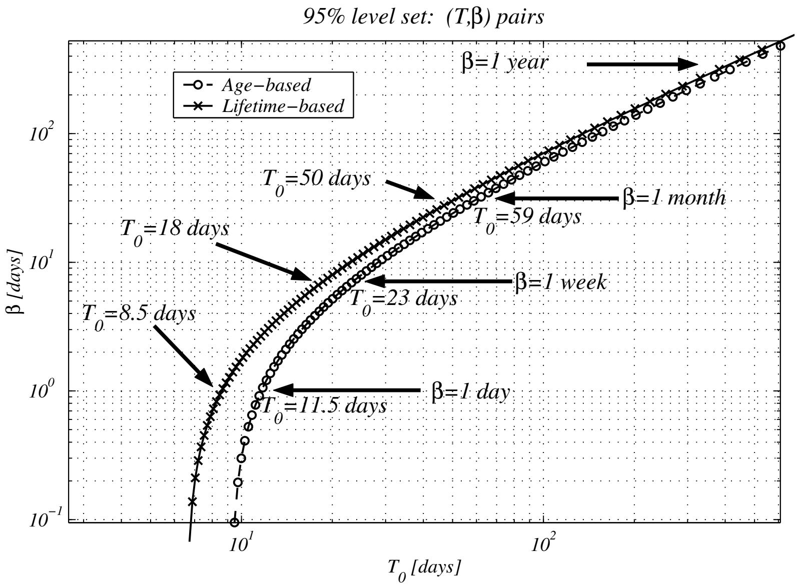


Figure 19: **Relating β and T_0: α=95% level set.** Here we have plotted two level sets of pairs (T_0, β) which yield a probability $\alpha = 0.95$ of being β-current. The two curves are derived from two different estimation methods, minimizing (11) or (16). The lifetime-based estimates are much more accurate. Regardless of the size of the collection, this data can be used to estimate how current an engine is when the indexing period T_0 takes on a value (in days) along the horizontal axis. As T_0 becomes large, relative check rate is too slow, and β approaches $0.95 \times T_0$.

References

[CLW97] E. G. Coffman, Z. Liu, and R. R. Weber. Optimal robot scheduling for web search engines. *Journal of Scheduling*, 1997. Available at http://www.inria.fr/mistral/personnel/Zhen.Liu/Papers/.

[DFKM97] Fred Douglis, Anja Feldmann, Balachander Krishnamurthy, and Jeffrey Mogul. Rate of change and other metrics: A live study of the world wide web. In *Proceedings of the USENIX Symposium on Internetworking Technologies and Systems*, December 1997. Available from http://www.research.att.com/˜anja/feldmann/papers.html.

[Fel71] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 2nd edition, 1971.

[Gra97] Matthew Gray. Internet growth summary. http://www.mit.edu/people/mkgray/net/internet-growth-raw-data.html, 1997.

[Inf95] Informant, 1995. http://informant.dartmouth.edu.

[ISC99] ISC, 1999. Internet Software Consortium; http://www.isc.org/dsview.cgi?domainsurvey/index.html.

[LG98] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 28:98–100, April 1998. Available by request at http://www.neci.nj.nec.com/homepages/lawrence/.

[LG99] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 1999.

[MR94] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers.* John Wiley and Sons, Inc., 1994.

[Pap84] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 2nd edition, 1984.