Estimation of Internet File-Access/Modification Rates from Incomplete Data

Norman Matloff
Department of Computer Science
University of California at Davis
Davis, CA 95616
USA
1-530-752-1953 (voice)
1-530-752-4767 (fax)
matloff@cs.ucdavis.edu

August 15, 2002

Abstract

Consider an Internet file for which last-time-of-access/modification (A/M) data is collected at periodic intervals, but for which full A/M data are not available. Methodology is developed here which enables estimation of the A/M rates, in spite of having only incomplete data of this nature.

Keywords: access rate; estimation; World Wide Web; renewal process; nonparametric density estimation

1 Introduction

One of the major functions of computer networks—ranging from databases on private local area networks to World Wide Web sites—is the sharing of information. A question that then arises concerns the number of people who are sharing that information.

Consider for example an Internet site that distributes public-domain software, written by various authors, available World Wide Web. In order to justify the time and funding the authors devote to these projects, it would be of interest to know how many users download the software, that is the mean number of downloads per unit time.

Another example arises with Web search engines. The user inputs one or more keywords, say "sailboats." The search engine will then produce a lengthy list of Web sites related to sailboats, ordered according to various criteria. One such criterion (possibly provided by the user as an option) might be frequency of modification; some users may be interested mainly in active sites that are frequently updated. In this case, we are interested in modification rates instead of access rates.

If we had direct data on A/M transactions, estimation of these and other simple rates would be straightforward (see Chapter 12 in Menasce and Almeida [1998]. However, such data may either be difficult to collect or else simply unavailable to the public.

For example, collection of some data of this type may require modification of server software, and a server administrator may not have the time to make such a modification, or may not have the source code at all (see Sec. 4.1.5 of Yeager and McGrath [1996]). For instance, in our Web search engine example above, the Web server software probably does not collect data on file-modification times, even though it does log external accesses.

Even worse, many server administrators would not be willing to divulge A/M data to the public, or would not have the time to do so. And even if they did, the logistics of dealing with thousands of Web site administrators would make our task infeasible.

However, there is related, publicly-accessible information that <u>is</u> available, in the form of time of the last A/M transaction times for a file. On a UNIX system with a file named, say x, the shell command

```
ls -ul x
```

will produce the time of last access to a file, while

```
ls -l x
```

will yield the time of last modification. Typically, as external users without accounts on the server machine, we will not have access to shell commands in the contexts we are considering here, but standard Internet client-side software does have commands that are similar to these. Even basic FTP versions have commands like these (with similar syntax). Similarly, http Web servers can allow users to acquire last-modification times (Hethmon [1997]).

At first glance, last-A/M time data seems insufficient for estimating A/M rates, as there is no direct relation between the data and the rates. To our knowledge, this problem has not been addressed in the literature. However, we will develop methodology here with that one actually can estimate the A/M rates from last-A/M time data.

2 Assumptions and Notation

Assume that A/M transactions to the given file occur as a Poisson process with intensity parameter λ. Suppose we sample the process at n intervals of length τ, in each case recording the time of the last file A/M transaction in the interval. By scaling time units, we may assume without loss of generality in our analysis that $\tau = 1.0$.[1]

Let L_i denote the (unobserved) number of file A/M transactions in the i^{th} interval, so that

$$P(L_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots \tag{1}$$

A problem that will become central to the issues addressed in this work is that some L_i may be 0. Let M denote the number of i for which $L_i > 0, i = 1, 2, \ldots, n$. In addition, let A_1 denote the value of the first nonzero L_i, A_2 the second one, and so on.

Define $T_{i1}, \ldots, T_{iA_i}$ to be the A/M transaction times to the file within the interval associated with A_i, mod 1.0; that is, these random variables consist of the noninteger portions of the A/M transaction times. We are able to observe only M and for i = 1,...,M the values of $W_i = T_{iA_i}$.

Estimation based on the W_i will be conditional on M, while estimation based on M itself will be unconditional.

It is important to note that our goals here are rather different from those in typical statistical applications, in the following senses:

- Formal statistical inference procedures, i.e., confidence intervals and hypothesis tests, are of only secondary importance in the Web applications discussed earlier. The focus would be on point estimators.
- Optimality of estimator accuracy is less of an issue than usual. In these types of applications one typically has very large amounts of data and thus an estimator need not be theoretically optimal. Instead, the major issue here is how to deal with the fact that we have no direct data on A/M times; how do we estimate overall A/M rates from last-A/M time data?
- Since formal inference procedures are not usually of interest, the analyst has more freedom to rely on *adaptive* methods, in this case choosing the value of τ after looking at the data.

3 Two Competing Estimators of an A/M Rate

3.1 Estimation of λ Via M

We begin with the simpler estimator, based only on M. It would seem more natural to estimate λ from the W_i, but if M is small, there will be too few W_i to get an accurate estimate from them, so we turn to using M itself.

[1] Note, however, that this is only to streamline our analysis. The choice of τ does matter, and guidelines will be discussed later.

Define

$$p = P(L_i > 0) = 1 - \exp -\lambda, \tag{2}$$

so that

$$\lambda = -\ln(1-p) \tag{3}$$

Maximum likelihood estimators (MLEs) are generally optimal, i.e., have minimum asymptotic variance (Cox and Hinkley [1974]). The MLE of p based on M is well known to be $\check{p} = M/n$ (see Sec. 10.2.2 of Trivedi [1982]), so from (3) the MLE of λ based on M is

$$\check{\lambda} = -\ln(1-\check{p}) = -\ln(1-M/n) \tag{4}$$

3.2 Estimation of λ Via the W_i

Let $f_k(w)$ and $F_k(w)$ denote the p.d.f. and c.d.f., respectively, of W_i, conditioned on $A_i = k$. Then for k > 0,

$$\begin{aligned} F_k(w) &= P(W_i \le w \,|\, A_i = k) && (5)\\ &= P(T_{i1} \le w, \ldots, T_{iA_i} \le w \,|\, A_i = k) && (6)\\ &= P(Y_{i1} \le w, \ldots, Y_{iA_i} \le w \,|\, A_i = k), \end{aligned}$$

where $Y_{i1}, \ldots, Y_{iA_i}$ are the *unordered* versions of the T_{ij} (the former variables can be thought of as a random permutation of the latter ones). By Theorem 6.1 of Trivedi [1982], the Y_{ij} are i.i.d. U(0,1), so we have that $F_k(w) = w^k$ and $f_k(w) = kw^{k-1}$ for w in (0,1) and k > 0.

Now for $0 < w < 1$let $G(w) = P(W_i \le w \,|\, A_i > 0)$, and set $g = G'$. Then

$$\begin{aligned} G(w) &= \frac{1}{P(A_i > 0)} \sum_{k=1}^{\infty} P(W_i \le w \, and \, A_i = k) && (7)\\ &= \frac{1}{1-e^{-\lambda}} \sum_{k=1}^{\infty} F_k(w) \cdot \frac{e^{-\lambda}\lambda^k}{k!} && (8)\\ &= \frac{e^{\lambda w} - 1}{e^{\lambda} - 1}. && (9) \end{aligned}$$

Thus

$$g(w) = \frac{\lambda e^{\lambda w}}{e^{\lambda} - 1}. \tag{10}$$

We wish to find the MLE of λ based on the W_i, conditional on the nonzero A_i. The conditional likelihood function of $W_1 = w_1, \ldots, W_m = w_m$ is

$$L(w_1, \ldots, w_m) = g(w_1)g(w_2)\ldots, g(w_m) = \frac{\lambda^m}{(e^{\lambda} - 1)^m} \cdot e^{\lambda \sum_{i=1}^{m} w_i}. \tag{11}$$

Maximizing this yields that the conditional MLE, $\hat{\lambda}$, must satisfy the equation

$$r(\hat{\lambda}) = \bar{W}, \tag{12}$$

where $\bar{W} = (W_1 + \ldots, W_m)/m$ and

$$r(t) = \frac{1}{1 - e^{-t}} - \frac{1}{t}.$$

3.2.1 Solution of the Likelihood Equation

Equation (12) has no closed-form solution. Thus iterative numerical methods must be used.

However, we can at least establish conditions under which the root exists and is unique, as follows. It is easily verified graphically that r(t) above is an increasing function of t, and varies continuously between 0.5 and 1. However, $\bar{W}$ can take on any value between 0 and 1. Thus the solution of the equation exists and is unique if $\bar{W} > 0.5$, and otherwise the MLE is 0.0.

Later, in determining the statistical accuracy of our estimator, we will again need to deal with the non-closed form of the MLE here, but will present a way to circumvent this problem.

4 Statistical Inference on λ

In the Web applications of interest here, a rough point estimate of the A/M rate is usually sufficient, and formal statistical inference methods (confidence intervals, hypothesis testing) is not needed. Nevertheless, in some cases inference methods may be of interest. For example, an analyst may be interested in investigating whether a Web page's current A/M rate has increased substantially from a past rate. In this section we develop machinery for conducting formal statistical inference. One additional benefit of the material in this section will be that it will help guide the user in choosing a good value of τ.

4.1 The "Delta Method"

The "delta method" (Serfling [1980]) says, roughly, that a function of an asymptotically Gaussian-distributed sequence of random variables is itself an asymptotically Gaussian-distributed sequence. In our context here, the relevant formulation is as follows.

Suppose that a random variable U_k, calculated from k observations from a distribution family parameterized by θ, is used to estimate θ and is distributed approximately $N(\theta, \sigma^2(\theta)/k)$. Then if q is a differentiable function, the quantity $q(U_k)$ is approximately distributed as

$$\lim_{k \to \infty} P\left(\frac{q(U_k) - q(\theta)}{\frac{1}{\sqrt{k}} |q'(\theta)| \sigma(\theta)} \leq t \right) = \Phi(t), \tag{13}$$

where $\Phi(t)$ is the cumulative distribution function for N(0,1).

The quantity

$$q'(\theta)^2 \sigma^2(\theta)/k$$

is then the *asymptotic variance* (AVar) of $q(U_k)$.

The approximate square root of this quantity,

$$SE(q(U_k)) = |q'(U_k)| \sigma(U_k)/\sqrt{k}$$

is known as the *standard error* of U_k. The standard error can be used for statistical inference purposes. For instance, an approximate 95% confidence interval for $q(\theta)$ based on $q(U_k)$ is

$$q(U_k) \pm 1.96 q'(U_k) \sigma(U_k)/\sqrt{k}$$

(The limit in Equation (13) remains valid if the standard error is used in place of the denominator in the fraction in that equation.)

4.2 Inference on $\check{\lambda}$

In the notation above, take k to be n, take θ to be p, and take U_k to be $\check{p} = M/n$. Then from Equation (2), take q(t) to be -log(1-t). Since the Central Limit Theorem shows that $\check{p}$ is approximately distributed as N(p, p(1-p)/n), then by the delta method $\check{\lambda}$ has an approximately normal distribution that has mean λ and variance

$$\frac{1}{n} \cdot \frac{p}{1-p}$$

Inference can then be done by replacing p in this expression by $\check{p}$. In other words,

$$SE(\check{\lambda}) = \sqrt{\frac{1}{n} \cdot \frac{\check{p}}{1-\check{p}}} \tag{14}$$

4.3 Inference on $\hat{\lambda}$

Now, let us see what can be done in the case of $\hat{\lambda}$. Note first that since $\hat{\lambda}$ is a function of $\bar{W}$, we would ordinarily take q() to be this function. In other words, q() would be the functional inversion of Equation (12). However, as noted earlier, we do not have this latter function in closed form.

We could find the approximate value of that function (actually, its derivative) during our iterative procedure to find $\hat{\lambda}$, but there is an easier approach. Instead, we use the delta method on the function r() in Equation (12), "pretending" that we do not know the asymptotic variance of $\bar{W}$ but do know that of $\hat{\lambda}$. Since we actually do know the asymptotic variance of $\bar{W}$, we can solve for what we do want. Here are the details.

The quantity $\bar{W}$ has mean

$$E\bar{W} = EW = \frac{1}{1-e^{-\lambda}} - \frac{1}{\lambda} = r(\lambda), \tag{15}$$

and variance

$$\begin{aligned}
Var(\bar{W}) &= \frac{1}{m} \cdot Var(W) && (16)\\
&= \frac{1}{m}\left(\frac{2}{\lambda^2} - \frac{1}{1-e^{-\lambda}} - (EW)^2\right) && (17)\\
&= \frac{1}{m}\left(\frac{1}{\lambda^2} - r(\lambda)[1+r(\lambda)]\right). && (18)
\end{aligned}$$

By the way, note that by comparing Equations (12) and (15), we see that $\hat{\lambda}$ is not only the MLE of λ, but also the Method of Moments Estimator of that quantity (Trivedi [1982]).

Due to the Central Limit Theorem, $\bar{W}$ has an approximately normal distribution with mean and variance as in (15) and (18). So, now considering $\bar{W}$ to be a function of $\hat{\lambda}$ in Equation (12), rather than vice versa, and thus "applying the delta method in reverse," we have that

$$Var(\bar{W} = AVar(\bar{W}) = r'(\lambda)^2 AVar(\hat{\lambda}),$$

so that the standard error of $\hat{\lambda}$ is

$$SE(\hat{\lambda}) = \frac{1}{r'(\hat{\lambda})}\sqrt{Var(\bar{W})} = \frac{1}{\sqrt{m} \cdot r'(\hat{\lambda})}\sqrt{\frac{1}{\hat{\lambda}^2} - r(\hat{\lambda})[1 + r(\hat{\lambda})]}. \qquad (19)$$

4.4 Comparison of $\hat{\lambda}$ and $\check{\lambda}$ Via Simulation

We performed a simulation study, calculating the mean squared errors (MSE) for $\hat{\lambda}$ and $\check{\lambda}$:

$$MSE(\hat{\lambda}) = E[(\hat{\lambda} - \lambda)^2],$$

$$MSE(\check{\lambda}) = E[(\check{\lambda} - \lambda)^2].$$

The settings simulated had values of λ ranging from 0.2 to 10.0 in increments of 0.2, for n = 50 and n = 500 sampling intervals of size $\tau = 1.0$. The MSE for each setting was based on 10,000 replications.

The results are shown in Figures 1 and 2, in the form of the square root of MSE, normalized by λ; in other words, what is plotted is

$$\frac{\sqrt{MSE}}{\lambda}$$

Intuitively, $\hat{\lambda}$ should typically be a superior estimator to $\check{\lambda}$, since the former is based on "richer" information than the latter (i.e., last-A/M times rather than counts of nonzero intervals). However, such intuition must be tempered by the fact that if λ is small, the quantity M might also be very small—in which case $\hat{\lambda}$ will be based on such a small sample that $\check{\lambda}$ may actually be the superior estimator.

This intuition is confirmed in Figures 1 (n = 50) and 2 (n = 500). For a sample size of 50, $\check{\lambda}$ performs better than $\hat{\lambda}$ for $\lambda < 3.7$, while for n = 500 the change point comes earlier, at approximately $\lambda = 3.0$. In other words, a sample size of n = 500 is large enough so that we will get a fairly large value of M even if λ is small.

5 The Poisson Assumption and Alternatives

Up to this point, we have been assuming that the A/M transactions occur as a Poisson process. Let us now give this assumption closer examination.

5.1 Formulation as a Renewal Process

Let S_i denote the time between the $(i-1)^{st}$ and i^{th} A/M transactions. Under the Poisson assumption, these inter-transaction times are independent and identically distributed, with the common distribution being exponential. We will now continue to assume that the S_i are i.i.d., but not necessarily with an exponential distribution.

Let N(t) denote the total number of A/M transactions that have occurred on or before time t, i.e.,

$$N(t) = max\{i : S_1 + \ldots, S_i \leq t\}$$

N(t) is a renewal process (Trivedi [1982]).

From here on, we will assume that the sampling interval width τ is large enough so that $P(L_i = 0)$ is negligible. It is important to keep in mind that we can arrange things this way, because we always have the option of increasing τ, even on an after-the-fact basis subsequent to collecting our data. After doing so, though, we still assume that we rescale time units so that $\tau = 1.0$, as before. We will now have that (except for negligible probability) $A_i = L_i$ for all i, and that M = n. Note also that W_i will have all of its mass on the unit interval (0,1).

The random variable $Z_i = 1 - W_i$ is the i^{th} *backward recurrence time*, meaning the time since the last renewal, measured from the i^{th} observation epoch i-1. From renewal theory, the (asymptotic) density function of this quantity is

$$b(t) = \frac{1 - C(t)}{E(S)} \tag{20}$$

where C is the cumulative distribution function of S.

Note that this implies that

$$b(0) = \frac{1}{E(S)} \tag{21}$$

The right-hand side here is the reciprocal of the mean inter-access time. By standard renewal theory that is equal to the asymptotic mean number of A/M accesses per unit time—exactly what we are trying to estimate, i.e. $b(0) = \lambda$. This will be useful later in this paper.

5.2 Examining the W_i to Assess the Exponential Assumption

If the S_i are exponentially distributed, as we have assumed earlier, then Equation (20) shows that the quantities Z_i are also exponentially distributed. Thus we can apply standard statistical goodness-of-fit assessment procedures to the quantities Z_i. It should be noted, though, that in our context there is typically a large amount of data, which means that if a formal goodness-of-fit hypothesis test is used, even slight departures from the exponential model will (misleadingly) result in rejection of they hypothesis at standard significance levels. Thus care should be used (see Chapter 7 of Matloff [1988]), and the goodness-of-it assessment should be treated as exploratory only.

5.3 Nonexponetial, Small-τ Case

Here we will investigate the behavior of $\check{\lambda}$ in the case in which τ is small.

Up to this point, for convenience we have been scaling time so that $\tau = 1$, but we will now drop this assumption. Modifying Equation (3) accordingly, we have that

$$\lim_{n \to \infty} \check{\lambda} = -\frac{1}{\tau} \ln(1-p) \tag{22}$$

In the Poisson context of Equation (3), the probability of the i-th observation interval being nonempty, $p = P(L_i > 0)$, was independent of i. This is not the case in our more general setting here, but p is still the long-run proportion of nonempty intervals.

To calculate p for this setting, Equation (20) yields[2] that

$$p = \int_0^\tau b(t)dt \approx \tau b(0) = \tau\lambda \tag{23}$$

Then

$$\lim_{n \to \infty} \check{\lambda} \approx -\frac{1}{\tau} \ln(1-\tau\lambda) = -\frac{1}{\tau}[-\tau\lambda + o(\tau)], \tag{24}$$

[2] The equation holds even without the conditions on τ imposed during the discussion preceding that equation.

In other words, for small τ the estimator $\check{\lambda}$ will be approximately *consistent* for λ, i.e.

$$\lim_{n\to\infty} \check{\lambda} \approx \lambda \tag{25}$$

even without the Poisson assumption.

5.4 A Large-τ Alternative Approach for Use in the Nonexponential Case

From renewal theory we know that the exponential assumption holds if and only if our renewal process has *independent increments,* i.e., the property that events in disjoint intervals of time are independent. Thus even if we were to find a suitable nonexponential parametric model, say a gamma distribution, we would have a problem with statistical estimation and inference procedures; that methodology assumes that the W_i are independent, which would not be true.

As an alternative to the exponential assumption, we now present a data-exploratory tool for estimation of A/M rates. This will consist of a novel application to renewal theory of tools for nonparametric density estimation. Here Equation (20) will play the central role. We resume our assumptions preceding Equation (21) that we have made τ large enough so that the probability of an empty interval is negligible and that we have then rescaled time so that $\tau = 1$.

Nonparametric density estimation is a refinement of the usual histogram methods taught in elementary statistics courses. It is used primarily as a tool for exploratory data analysis, with the aim being to answer questions about the overall shape of the density function, such as: Is the density unimodal or multimodal? Where does the bulk of the distribution lie? By contrast, it is rare that nonparametric density estimation has been used to estimate a density at only one point, which is what we will do here.

In light of Equation (21), our job now is to estimate b(0) from our data Z_i, without assuming a parametric family such as the exponential.

The classic kernel nonparametric density estimator, applied here to the function b(t), is

$$\widehat{b}(t) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{t - Z_i}{h})$$

where h is a *smoothing parameter* and the *kernel* K is chosen to be a mean-0 density function in its own right. The choice of K is up to the user, provided K satisfies certain regularity conditions (Simonoff [1996]).

The smoothing parameter is similar to the bin width in histograms. A large body of mathematical theory exists on this point; here we will assume that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. We will choose

$$K(t) = \begin{cases} 0.5, & -1 < t < 1 \\ 0, & \text{otherwise} \end{cases}$$

With this choice of K we would have

$$\widehat{b}(0) = \frac{0.5\#(0,h)}{nh} \tag{26}$$

where #(u,v) denotes the count of the number of Z_i in the interval (u,v).

However, kernel estimators are subject to serious bias problems near the boundary of a density's nonzero region. For a kernel estimator $\widehat{f}$ of a density f based on a kernel K that has the value 0 outside of (-1,1),

$$E[\widehat{f}(0)] = \int_{-1}^{0} K(u)du \text{ f}(0) + O(h)$$

(Simonoff [1996]).

So, in our case here,

$$E[\widehat{b}(0)] = 0.5b(0) + O(h)$$

Thus we must redefine (26) to be

$$\widehat{b}(0) = \frac{\#(0,h)}{nh} \tag{27}$$

It is up to the user to choose the value of the smoothing parameter h. Though some methods have been proposed for choosing h, no fully practical method has yet been developed—especially for our situation in which we wish to minimize mean squared error at a point (t = 0) rather than the usual criterion of integrated mean squared error. Thus nonparametric density estimation is used typically as a data-exploratory tool rather than a means of formal statistical inference (Simonoff [1996]), and we present (27) in that spirit.

Equation (20) shows that b(t) is a nonincreasing function. This would seem to allow us to make use of nonparametric maximum likelihood estimators for unimodal densities (Van der Vaart [1998]), which are *automatic*, i.e., do not have a smoothing parameter like h above for which the user must choose a value. However, the classic estimator of this type is inconsistent at t = 0 (Wegman [1975]). A variation that overcomes this problem was developed by Meyer [2001]. However, it is complicated to implement and again suffers from the fact that it is aimed at minimizing integrated mean squared error, rather than mean squared error at t = 0.

6 Numerical Example

7 Conclusions

References

[1] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical statistics.* London: Chapman and Hall.

[2] Hethmon, P. (1997). *Illustrated Guide to HTTP.* Manning.

[3] Matloff, N. (1988). *Probability modeling and computer simulation, applied to engineering and computer science.* Boston: PWS-Kent.

[4] Menasce, D. and Almeida, V. (1998). *Capacity planning for Web performance: metrics, models, and methods.* Englewood Cliffs, New Jersey: Prentice-Hall.

[5] Meyer, M. (2001). An alternative unimodal density estimator with a consistent estimate of the mode, *Statistics Sinica*, 11, 4, 1159-1174.

[6] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

[7] Simonoff, J. (1996). *Smoothing methods in statistics.* New York: Springer-Verlag.

[8] Trivedi, K. (1982). *Probability and statistics with reliability, queuing and computer science applications.* Englewood Cliffs, New Jersey: Prentice-Hall.

[9] Van der Vaart, A. (1998). *Asymptotic statistics*, New York: Cambridge University Press.

[10] Wegman, E. (1975). Maximum likelihood estimation of a probability density, *Sankhya (A)*, 37, 211-224.

[11] Yeager, N. and McGrath, R. (1996). *Web server technology: the advanced guide for World Wide Web information providers.* Belmont, California: Morgan Kaufmann.