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

NORMAN MATLOFF
University of California at Davis

Consider an Internet file for which data on last time of access/modification (A/M) of the file are collected at periodic intervals, but for which direct A/M data are not available. Methodology is developed here which enables estimation of the A/M rates, in spite of having only indirect data of this nature. Both parametric and nonparametric methods are developed. Theoretical and empirical analyses are presented which indicate that the problem is indeed statistically tractable, and that the methods developed are of practical value. Behavior of the parametric estimators is examined when these assumptions are violated, and these estimators are found to be robust against some such violations.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics—*Stochastic processes*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

General Terms: Measurement, Theory

Additional Key Words and Phrases: Access rate, asymptotic distributions, nonparametric density estimation, Poisson and renewal processes, statistical estimation, World Wide Web

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. Two questions that then arise concern the number of people who are sharing that information and the frequency with which the information is updated.

Consider for example an Internet site that distributes public-domain software, written by various authors, available on the 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 long-run average 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 which are frequently updated.

Author’s address: N. Matloff, Dept. of Computer Science, University of California at Davis, Davis, CA 95616; e-mail: matloff@cs.ucdavis.edu.

In this case, we are interested in modification rates instead of access rates.

Another application involving modification rates concerns the design of Web crawlers. The efficiency of a Web crawler would be much improved if it could do more frequent checks of sites which are known to have higher modification rates, and less frequent checks of sites with lower rates. This application is pursued in Cho and Garcia-Molina [**?**].

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

However, there is often related, publicly accessible information that *is* available, in the form of time of the last A/M transaction times for a file. For example, FTP typically offers last-access time, and HTTP Web servers can allow users to acquire last-modification times [**?**].

At first glance, last-A/M time data seems statistically insufficient for estimating A/M rates, as there is no direct relation between the data and the rates. However, this paper will develop methodology with which one actually can estimate the A/M rates from last-A/M time data.

2. ASSUMPTIONS AND NOTATION

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

$$\lambda = \lim_{t \to \infty} \frac{N(t)}{t}$$

when this limit exists. For now (this assumption will be broadened later) assume that A/M transactions to the given file occur as a (time-homogeneous) Poisson process, in which case the limit does exist.

Suppose we sample the process at n intervals of length τ, in each case recording the time of the last A/M transaction in the interval.

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

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

for $i = 1, 2, 3, \ldots$.

A problem that will become central to the issues addressed in this work is that some L_i may be 0. Let M_n 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 τ. In other words, $T_{i1}, \ldots, T_{iA_i}$ are the times of the transactions, as measured from the beginning of that interval. We are able to observe only M_n and for $i = 1, \ldots, M_n$ the values of $W_i = T_{iA_i}$. Estimation based on the W_i will be conditional on M_n, while estimation based on M_n itself will be unconditional.

3. TWO COMPETING ESTIMATORS OF AN A/M RATE

3.1 Estimation of λ via M_n

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

Remark 1. As pointed out in Cho and Garcia-Molina [**?**], this also covers the case in which we do not have the W_i at all, but do have M_n. For example, we may record a Web page at regular intervals, and thus by comparison of a new page to its previously recorded copy determine whether a modification had been made. In this kind of setting, we would know M_n but not the W_i.

Define

$$p = \mathrm{P}(L_i > 0) = 1 - e^{-\lambda\tau} \tag{1}$$

so that

$$\lambda = -\frac{1}{\tau}\ln(1-p). \tag{2}$$

The maximum likelihood estimator (MLE) of λ based on M_n is

$$\check{\lambda} = -\frac{1}{\tau}\ln(1-\check{p}) = -\frac{1}{\tau}\ln(1-M_n/n) \tag{3}$$

[**?**, Example 7.1.2].

As noted by Lehmann, $\check{\lambda}$ will not exist (or can be taken to be infinite) if $M_n = n$. With the very large sample sizes typical for the settings considered in the present work, this nonexistence problem is mainly of theoretical interest, but it could occur if λ or τ is very large. Cho and Garcia-Molina [**?**] present a modified estimator which is guaranteed to exist.

3.2 Estimation of λ via the W_i

3.2.1 *Derivation of the Likelihood Equation.* We wish to find the MLE of λ based on $W_1, \ldots, W_{M_n}$, conditional on M_n. We thus need the density g of the W_j. To this end, let Y denote the time, as measured from the epoch $(i-1)\tau$, of the occurrence of the last A/M event before $i\tau$, with $Y = 0$ if there are no events during $\Big((i-1)\tau, i\tau\Big]$. Let J_1, J_2 and J_3 denote the number of events in the intervals $\Big((i-1)\tau, (i-1)\tau+t\Big]$, $\Big((i-1)\tau+t, i\tau\Big]$, and $\Big((i-1)\tau, i\tau\Big]$, respectively, for $0 < t < \tau$. Then

$$\begin{aligned} P(Y \le t | J_3 > 0) &= \frac{P(J_2 = 0 \text{ and } J_3 > 0)}{P(J_3 > 0)} \\ &= \frac{P(J_2 = 0 \text{ and } J_1 > 0)}{P(J_3 > 0)} \\ &= \frac{e^{-\lambda(\tau-t)} - e^{-\lambda\tau}}{1 - e^{-\lambda\tau}}, \end{aligned}$$

where the last step uses the fact that J_1 and J_2 are independent.

Thus

$$g(w) = \frac{\lambda e^{\lambda w}}{e^{\lambda\tau} - 1}, \tag{4}$$

for $0 < w < \tau$. The conditional likelihood function of $W_1 = w_1, \ldots, W_m = w_m$ given $M_n = m$ is then

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

Maximizing the logarithm of this expression shows that the conditional MLE, $\hat{\lambda}$, must satisfy the equation

$$r(\hat{\lambda}) = \bar{W}, \tag{5}$$

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

$$r(t) = \frac{\tau}{1 - e^{-t\tau}} - \frac{1}{t}. \tag{6}$$

Let W have the density in Equation (4). Then $\bar{W}$ has mean

$$\mathrm{E}\bar{W} = \mathrm{E}W = \frac{\tau}{1 - e^{-\lambda\tau}} - \frac{1}{\lambda} = r(\lambda), \tag{7}$$

and variance

$$\begin{aligned} \mathrm{Var}(\bar{W}) &= \frac{1}{m} \cdot \mathrm{Var}(W) \\ &= \frac{1}{m}\left[\frac{\tau^2}{1 - e^{-\lambda\tau}} - \frac{2}{\lambda} r(\lambda) - (\mathrm{E}W)^2\right]. \end{aligned} \tag{8}$$

3.2.2 *Asymptotic Properties of* $\hat{\lambda}$. The estimator $\hat{\lambda}$ is a *strongly consistent* estimator of λ, meaning that $\hat{\lambda} \to \lambda$ as $m \to \infty$ with probability one. Furthermore, $\hat{\lambda}$ has an asymptotic normal distribution and is optimal in the sense that it achieves minimal asymptotic variance. (In the unconditional setting, the binomial-based $\check{\lambda}$ is well known to have these properties.)

One may easily verify the conditions for these properties [**?**, Section 4.2.2]: First, the density (4) must be thrice-differentiable with respect to λ, which is clearly the case. Second, the integrals, with respect to w, of that density and its logarithm must be differentiable under the integral sign with respect to λ, again clearly true. And finally, the partial derivative of the logarithm of (4), evaluated at $w = W$, i.e.

$$\frac{\partial}{\partial\lambda} \ln[g(w;\lambda)]\Big|_{w=W} = W + \frac{1}{\lambda} - \frac{\tau}{1 - e^{-\tau\lambda}} = W - r(\lambda),$$

must have finite variance, which we showed in (8).

3.2.3 *Solution of the Likelihood Equation.* Equation (5) has no closed-form solution. Thus iterative numerical methods must be used.

However, we will at least establish conditions under which the root exists and is unique. First, we show that $r(t)$ is a strictly increasing function of t. To see this, for convenience scale so $\tau = 1$, and write

$$r'(t) = \frac{\left(\frac{e^t-1}{t}\right)^2 - e^t}{\left(e^t - 1\right)^2}. \tag{9}$$

We need to show that the numerator is positive for $t > 0$. Using a Taylor series expansion for e^t, we have

$$\frac{e^t - 1}{t} = \sum_{i=0}^{\infty} \frac{t^i}{(i+1)!}. \tag{10}$$

Square this and then subtract the Taylor series for e^t. The squared quantity will consist of the sum of the squares of terms of (10),

$$\sum_{i=0}^{\infty} \frac{t^{2i}}{(i+1)!^2}, \tag{11}$$

and the sum of the cross products,

$$2 \sum_{i,j=0,i<j}^{\infty} \frac{t^i}{(i+1)!} \frac{t^j}{(j+1)!}. \tag{12}$$

For each $k = 0, 1, \ldots$, we will consider terms of degree k in the variable t, searching for a group of terms in the square of (10) whose sum is greater than or equal to the k-power term $\frac{1}{k!}t^k$ in the Taylor series for e^t. Here are the cases:

$k = 0$:
The term at $i = 0$ in (11) matches the term at $k = 0$ in the Taylor series for e^t.

$k = 1$:
The term at $i = 0, j = 1$ in (12) matches the term at $k = 1$ in the Taylor series for e^t.

$k > 1$:
The sum of the term at $i = 1$ in (11) and the term at $i = 0, j = 2$ in (12) is greater than $t^2/2!$.

So, the squared quantity in the numerator of (9) has terms corresponding to all those in e^t, plus more, so we have a strictly postive difference. Thus the numerator of (9) is indeed positive.

In addition, by a couple of applications to L'Hospital's Rule we find that $r(0)$ is equal to 0.5τ. Also, $r(t)$ goes to τ as $t \to \infty$. Since $r(t)$ is continuous, we see that it can take on any value between 0.5τ and τ. However, $\bar{W}$ can take on any value between 0 and τ. Thus the solution of the equation exists and is unique if $\bar{W} > 0.5\tau$, and otherwise the MLE is 0.0.

If desired, the approximate probability that the MLE will be nonzero, for any values of λ, τ and m of interest, can be calculated using the Central Limit Theorem and Equations (7) and (8).

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

4. STATISTICAL INFERENCE ON λ

In the Web applications of interest here, a rough point estimate of the A/M rate would often be sufficient, and formal statistical inference methods (confidence intervals, hypothesis testing) would not be 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.

4.1 The "Delta Method"

The "delta method" [**?**, Section 3.1] says, roughly, that a sufficiently smooth function of an asymptotically Gaussian-distributed sequence of random variables is itself an asymptotically Gaussian-distributed sequence. More precisely, suppose

$$\lim_{k\to\infty} \mathrm{P}\left(\frac{U_k - \theta}{\frac{1}{\sqrt{k}}\sigma(\theta)} \leq t \right) = \Phi(t) \text{ for all real } t,$$

where $\Phi(t)$ is the cumulative distribution function for the standard normal distribution. Then if h is continuously differentiable in a neighborhood of θ and $h'(\theta) \neq 0$, and $\sigma(\cdot)$ is continuous in a neighborhood of θ,

$$\lim_{k\to\infty} \mathrm{P}\left(\frac{h(U_k) - h(\theta)}{\frac{1}{\sqrt{k}}|h'(\theta)|\sigma(\theta)} \leq t \right) = \Phi(t) \text{ for all real } t. \tag{13}$$

The quantity $[h'(\theta)]^2\sigma^2(\theta)/k$ is then the *asymptotic variance* (AVar) of $h(U_k)$. The *estimated* square root of this quantity,

$$\mathrm{SE}(h(U_k)) = |h'(U_k)|\sigma(U_k)/\sqrt{k},$$

is known as the *standard error* of U_k. The standard error can be used for statistical inference purposes. For instance, it follows directly from (13) that as $k \to \infty$ an asymptotically valid 95% confidence interval for $h(\theta)$ based on $h(U_k)$ is

$$h(U_k) \pm 1.96|h'(U_k)|\sigma(U_k)/\sqrt{k}.$$

(The limit in Equation (13) remains valid if the standard error is used in place of the denominator in the fraction in that equation [**?**, Section 1.5.4].)

4.2 Inference via $\check{\lambda}$

In the notation above, take k to be n, take θ to be p, and take U_k to be $\check{p} = M_n/n$. Also, informed by Equation (2), then in Equation (13) take $h(t)$ to be $-\frac{1}{\tau}\ln(1-t)$. Since the Central Limit Theorem shows that $\check{p}$ is approximately normally distributed with mean and variance p and $p(1-p)/n$, then by the delta method $\check{\lambda}$ has an approximately normal distribution which has mean λ and variance

$$\frac{1}{n\tau^2} \cdot \frac{p}{1-p}.$$

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

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

4.3 Inference via $\hat{\lambda}$

Now, let us see what can be done in the case of $\hat{\lambda}$. Again, the main issue is what to take for the function h. Note first that since $\hat{\lambda}$ is a function of $\bar{W}$, we would ordinarily take h to be this function. In other words, h would be the functional inversion of Equation (5). 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 (5), "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.

Considering $\bar{W}$ to be a function of $\hat{\lambda}$ in Equation (5), rather than vice versa, and thus "applying the delta method in reverse," we have that

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

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

$$\mathrm{SE}(\hat{\lambda}) = \frac{1}{r'(\hat{\lambda})}\sqrt{\hat{\mathrm{Var}}(\bar{W})} = \frac{1}{\sqrt{m} \cdot r'(\hat{\lambda})}\sqrt{\frac{\tau^2}{1-e^{-\hat{\lambda}\tau}} - \frac{2}{\hat{\lambda}}r(\hat{\lambda}) - r^2(\hat{\lambda})}. \tag{15}$$

5. THE HOMOGENEOUS POISSON ASSUMPTION AND ALTERNATIVES

Up to this point, we have been assuming that the A/M transactions occur as a Poisson process, and that the process is time-homogeneous, meaning that λ does not vary through time. Let us now give these assumptions 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 homogeneous Poisson assumption, these intertransaction times are independent and identically distributed, with the common distribution being exponential. Now we continue to assume that the S_i are independent and have a common distribution, but we drop the assumption that that distribution is exponential.

We do continue to assume that that distribution is absolutely continuous. In other words, letting S denote a generic random variable having the distribution of each S_i, then there exists a nonnegative function f_S such that

$$P(S_i \leq s) = F_S(s) = \int_0^s f_S(u)\ du \text{ for all } s \geq 0.$$

It is also assumed that $E(S) < \infty$.

As before, let $N(t)$ denote the total number of A/M transactions that have occurred on or before time t, so that

$$N(t) = \max\{i : S_1 + \ldots + S_i \leq t\} \text{ for all } t \geq 0.$$

The set of random variables *N(t)* comprise a renewal process [**?**].

For any fixed-time multiple of τ, $i\tau$, consider the random variable Z_i, defined to be the time since the last renewal, called the *backward recurrence time.* From renewal theory, the (asymptotic, as $i \rightarrow \infty$) density function of Z_i is

$$b(t) = \frac{1 - F_S(t)}{\mathrm{E}(S)}. \tag{16}$$

Note, though, that $1/\mathrm{E}(S)$, being the reciprocal of the mean intertransaction time, is equal to the asymptotic mean number of A/M transactions per unit time [**?**, Section 5.3]. In other words

$$b(t) = \lambda[1 - F_S(t)]. \tag{17}$$

Note that this and the fact that S is nonnegative and has an absolutely continuous distribution implies that

$$\lambda = b(0). \tag{18}$$

This will be useful in later material.

5.2 Examining the W_i to Assess the Exponential Assumption

If the S_j are exponentially distributed, as we assumed earlier, then Equation (16) shows that the quantities Z_i are also (asymptotically) exponentially distributed. Thus we can investigate the appropriateness of the exponential assumption by applying 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 (as seen for example in Section 6), 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 the hypothesis at standard significance levels. Thus care should be used [**?**, Chapter 7], and the goodness-of-it assessment should be treated as exploratory only. Histograms or other nonparametric density estimation techniques can be used to plot the Z_i and check for roughly exponential shape.

5.3 The Behavior of $\check{\lambda}$ in the Nonexponential, Small-τ Case

How robust is the estimator $\check{\lambda}$ to the exponential assumption? In this section, we will investigate the behavior of $\check{\lambda}$ in the case in which the intertransaction distribution is nonexponential and τ is small.

We must first generalize our earlier definition of the quantity p. In the Poisson context of that equation, the probability of the i^{th} observation interval being nonempty, $p = \mathrm{P}(L_i > 0)$, was independent of i, due to the memoryless property of the exponential distribution. This is not the case in our more general setting here, but we can still define p to be the long-run probability of an interval being nonempty. Specifically, since the i^{th} interval will be nonempty if and only if $Z_i < \tau$, the material in Section 5.1 shows that the quantities $\mathrm{P}(L_i > 0)$ converge to an integral of $b(t)$, so we can define p as

$$p = \lim_{i\to\infty} \mathrm{P}(L_i > 0) = \int_0^\tau b(t)dt. \tag{19}$$

Recall that we are assessing the robustness of the estimator $\check{\lambda} = -\frac{1}{\tau}\ln(1-M_n/n)$, whose derivation assumes an exponential distribution, in the case in which S is not exponentially distributed but τ is small. As noted earlier, for nonexponential settings, A/M transactions in one observation interval are not independent of the ones in other intervals, so M_n is not a sum of independent random variables. Thus we should verify that $\lim_{n\to\infty} M_n/n$ exists and is equal to p:

Lemma 5.1.

$$\lim_{n\to\infty} \frac{M_n}{n} = p \text{ with probability 1.} \tag{20}$$

Proof. First define B_i to be the time of the i^{th} A/M transaction, mod τ:

$$B_i = (S_1 + \cdots + S_i) \bmod \tau,$$

for $i = 1, 2, \ldots$, and define $B_0 = 0$.

Also for $i = 1, 2, \ldots$, define Q_i to be the number of empty intervals "skipped over" by the i^{th} A/M transaction:

$$Q_i = \max\Big(\lfloor (B_{i-1} + S_i)/\tau \rfloor - 1, 0\Big).$$

For example, suppose $\tau = 1.0$ and $B_{15} = 0.2$. If $S_{16} = 0.5$, say, then this new A/M transaction will be in the same observation interval as the preceding one. If $S_{16} = 0.9$, then the new transaction will occur in the interval immediately following the last transaction. In both of these examples, Q_{16} will be 0. But if $S_{16} = 1.9$, the interval immediately following the last will be empty, and the new transaction will occur in the next interval after that, so that Q_{16} will be 1.

Note that the number of empty intervals which occur up to the n^{th} A/M transaction is

$$Q_1 + \cdots + Q_n + 1_{\{S_1 > \tau\}}, \tag{21}$$

where the indicator random variable $1_{\{S_1>\tau\}}$ is equal to 1 if $S_1 > \tau$ and 0 otherwise. This formulation of the count of empty intervals will be important below.

Due to the independence of the S_i, the pairs (B_i, Q_i), $i = 1, 2, \ldots$ form a discrete-time, continuous state space Markov process. Assume that the density of S satisfies conditions to make the process convergent to a stationary distribution. For instance, this will obtain if the support of S is a finite closed interval. (Rey-Bellet [**?**, Remark 8.2] shows that compactness of the state space, plus a continuous analog of irreducibility, implies convergence of a Markov process.) For convenience, we will assume this here.

For each $k = 0, 1, \ldots$, let C_{nk} denote the number of $Q_i = k$, for $i = 1, \ldots, n$. Then the quantities

$$\lim_{n\to\infty} \frac{C_{nk}}{n}$$

will converge to the second marginal component in the stationary distribution of the Markov process.

Thus since $Q_1 + \cdots + Q_n = \sum_k kC_{nk}$ we have

$$\lim_{n\to\infty} \frac{Q_1 + \cdots + Q_n}{n} = c \tag{22}$$

for some constant c, with probability one. Then (22) and the fact that $N(t) \to \infty$ as $t \to \infty$ with probability 1 imply that

$$\lim_{t\to\infty} \frac{Q_1 + \cdots + Q_{N(t)}}{N(t)} = c$$

with probability 1. Set $t = n\tau, n = 1, 2, \ldots$. Then consideration of (21) yields that

$$Q_1 + \cdots + Q_{N(n\tau)} + 1_{\{S_1>\tau\}} = n - M_n.$$

Thus

$$\begin{aligned}
\lim_{n\to\infty} \frac{M_n}{n} &= 1 - \lim_{n\to\infty} \frac{Q_1 + \cdots + Q_{N(n\tau)} + 1_{\{S_1>\tau\}}}{n} \\
&= 1 - \lim_{n\to\infty} \frac{Q_1 + \cdots + Q_{N(n\tau)}}{N(n\tau)} \cdot \frac{N(n\tau)}{n} \\
&= 1 - c\lambda\tau,
\end{aligned}$$

with probability 1, since standard renewal theory shows that

$$\lim_{u\to\infty} \frac{N(u)}{u} = \lambda \text{ w.p. } 1.$$

So, M_n/n converges to some constant. Then the relation

$$\mathrm{E}\left(\frac{M_n}{n}\right) = \frac{\mathrm{P}(L_1 > 0) + \cdots + \mathrm{P}(L_n > 0)}{n},$$

(19) and the Bounded Convergence Theorem yield (20). □

Then from (19),

$$\lim_{n\to\infty} \check{\lambda} = \lim_{n\to\infty} -\frac{1}{\tau} \ln\left(1 - \frac{M_n}{n}\right) = -\frac{1}{\tau} \ln(1-p) = u(\tau) \text{ w.p. } 1,$$

where

$$u(x) = \frac{-\ln\left[1 - \int_0^x b(t)dt\right]}{x}.$$

Now expanding at $\tau = 0$, we have

$$\lim_{n\to\infty} \check{\lambda} = u(0) + u'(0)\tau + o(\tau). \tag{23}$$

From (18) we see that $u(0) = \lambda$. To evaluate $u'(0)$, let $y(x) = \int_0^x b(t)dt$. Then

$$u'(x) = \frac{x \cdot \frac{y'(x)}{1-y} + \ln(1-y)}{x^2}, \tag{24}$$

so

$$u'(0) = \frac{z(x)}{2x}\Big|_{x=0},$$

where $z(x)$ is the derivative of the numerator in (24), i.e.

$$z(x) = x \cdot \frac{(1-y)y''(x) + [y'(x)]^2}{(1-y)^2}.$$

Thus

$$\begin{aligned} u'(0) &= \frac{(1-y)y''(x) + [y'(x)]^2}{2(1-y)^2}\Big|_{x=0} \\ &= \frac{b'(0) + [b(0)]^2}{2}. \end{aligned}$$

As previously noted, $b(0) = \lambda$, and from (17) we see that $b'(0) = -\lambda f_S(0)$. Thus

$$\lim_{n\to\infty} \check{\lambda} = \lambda + \frac{[-\lambda f_S(0) + \lambda^2]}{2} \cdot \tau + o(\tau) \tag{25}$$

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

$$\lim_{n\to\infty} \check{\lambda} \approx \lambda,$$

even without the Poisson assumption, thus greatly extending the applicability of this estimator.

Equation (25) also gives us some idea as to whether $\check{\lambda}$ will have a tendency to over- or underestimate λ in various non-Poisson cases. (Note that in the Poisson case, $f_S(0) = \lambda$, so the second term in (25) is 0, reflecting the fact that $\check{\lambda}$ is an exactly consistent estimator of λ in that setting.) If for instance S has a uniform distribution on $(0, c)$, we will have $\lambda = 2/c > f_S(0)$, resulting in $\check{\lambda}$ having a tendency to overestimate λ.

5.4 Two Data-Exploratory Approaches

Continue to assume that the A/M transactions occur as a renewal process. From renewal theory we know that the exponential assumption holds if and only if our renewal process has *independent increments,* meaning that it has the property that renewal counts in disjoint time intervals are independent. Thus even if we were to find a suitable nonexponential parametric model for the intertransaction times, say a gamma distribution, we would have a problem with standard statistical estimation and inference methodology; that methodology assumes that the W_i are independent, which would not be true.

We found in the previous subsection that the exponential assumption is not important for $\check{\lambda}$ if τ is small. For other nonexponential cases, we now present two exploratory tools for estimation of A/M rates. These will be based of a novel application of tools for nonparametric density estimation. Here Equation (16) will play a central role. Assume here that time has been scaled 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, in practice it is rare for nonparametric density estimation to be used to estimate a density at only one point, which is what we will do here: In light of Equation (18), our job is to estimate $b(0)$ from our data Z_i, without assuming a parametric family such as the exponential.

5.4.1 *A Graphical Approach.* The classic kernel nonparametric density estimator, applied here to the function $b(t)$, is

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

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 [?].

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

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

With this choice of K we would have

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

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

However, here $\#(-h,h) = \#(0,h)$ and for this reason 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 which has the value 0 outside of (-1,1),

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

[**?**]. So, in our case here,

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

Thus we will redefine (26) to be

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

to make the estimator consistent.

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. This is especially true for our situation, in which we wish to minimize mean squared error at a specific point (here $t = 0$) 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 [**?**], and we present Equation (27) in that spirit.

Note also that while one may need a large amount of data to make this work well, in the applications considered by this paper, we typically do have large amounts of data, as noted earlier.

5.4.2 *An Approach Based on Isotonic Inference Methodology.* Another approach would rely on the fact that Equation (16) shows that $b(t)$ is a nonincreasing function, suggesting the use of *isotonic inference* methodology [**?**], which takes into account ordinal relationships.

In particular, we could make use of nonparametric maximum likelihood estimators for unimodal densities [**?**], a class which of course includes monotonic densities. These estimators are *automatic*, meaning that they do not have a smoothing parameter like h above for which the user must choose a value. This would appear to solve the problem which arose in the previous section.

However, the classic estimator of this type is inconsistent at $t = 0$ [**?**]. A variation which overcomes this problem has been developed [**?**]. It is rather 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 $t = 0$, but this may be a promising direction to take, and should be the subject of future research.

5.5 Time Homogeneity Aspect of the Poisson Assumption

Even if the access pattern is Poisson, the rate λ might be time-varying instead of constant. For example, if the users of a Web page are disproportionately located in the U.S. and their usage is low during, say, early morning hours, then $\lambda(t)$ may be periodic with period 24 hours. How well do our estimators $\check{\lambda}$ and $\hat{\lambda}$ do in such a situation?

To investigate this, consider settings in which the accesses follow a nonhomogeneous Poisson process whose rate function $\lambda(t)$ has period τ [**?**, Section 6.3.1]. (We assume here that the period is known, e.g. 24 hours, and that τ has been chosen to match the period.) Then if X is the number of accesses during one period of $\lambda(t)$,

$$P(X = k) = \frac{1}{k!} e^{-m(\tau)} [m(\tau)]^k$$

and $EX = m(\tau)$, where

$$m(t) = \int_0^t \lambda(s)\ ds.$$

Note that this means that our A/M rate ν is now $m(\tau)/\tau$.

Let us first consider the behavior of $\check{\lambda}$. In analogy with (1) and (2), define

$$q = \mathrm{P}(L_i > 0) = 1 - \exp[-m(\tau)],$$

so that

$$\nu = -\frac{1}{\tau} \ln(1 - q).$$

From (3), we now can see the behavior of $\check{\lambda}$ in the periodic case we are examining here. The quantity $\check{p} = M_n/n$ will converge almost surely to q as $n \to \infty$ (unlike the situation in Section 5.3, events in different intervals are i.i.d. here), and thus $\check{\lambda}$ will converge to ν, just as desired. And the standard error given by (14) will remain valid as well, since M_n is still binomial, with parameter q.

However, the situation is quite different in the case of $\hat{\lambda}$. For our counterexample, take $\tau = 1$, and suppose $\lambda(t)$ has total mass 1 in a small interval centered at $t = 0.5$:

$$\lambda(t) = \begin{cases} \frac{1}{2\delta}, & t \in (0.5 - \delta, 0.5 + \delta), \\ 0, & \text{otherwise,} \end{cases}$$

for a small value of δ which we will choose below. In this setting, we will have $\nu = 1$ and

$$\bar{W} \in (0.5 - \delta, 0.5 + \delta).$$

Recall that $\hat{\lambda}$ is the solution of Equation (5), so in our case here,

$$r(\hat{\lambda}) \le 0.5 + \delta \qquad (28)$$

with probability 1.

From Section 3.2.3 we know that the unique solution of

$$r(u) = 0.5$$

is $u = 0$, and since the function r is continuous and strictly increasing, (28) shows that we can choose δ so that, say, $\hat{\lambda} \leq 0.1$ with probability 1. Yet $\nu = 1$. Thus $\hat{\lambda}$ will not be a consistent estimator of ν.

Fortunately, there is a way to work around this problem: One can simply increase the value of τ. If for example we collected data every 2 hours and suspect a daily pattern, we could do our analysis with τ set to, say, 480 hours instead of 2. The data in each set of 240 sampling intervals would be collapsed into one interval.

This has the effect of changing the nonhomogeneous Poisson process into an approximately homogeneous one. To see this, think of the effect on a nonhomogeneous Poisson process with period τ if we "compress" $\lambda(t)$ so that the period is τ/k, keeping τ constant. Formally, this would mean replacing $\lambda(t)$ by $\lambda(kt)$. For large k the behavior of the resulting nonhomogeneous Poisson process is approximately that of a homogeneous Poisson process.

Increasing the value of τ may result in some loss of information. This will typically be less of an issue, once again because the applications described here would tend to have large amounts of data.

6. EMPIRICAL ASSESSMENTS

6.1 Comparison of $\hat{\lambda}$ and $\check{\lambda}$ via Simulation

Intuitively, $\hat{\lambda}$ should typically be a superior estimator to $\check{\lambda}$, since the former is based on "richer" information than the latter (that is, last-A/M times rather than counts of nonzero intervals). However, such intuition must be tempered by the fact that if $\lambda\tau$ is small, the quantity M_n 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.

To investigate this, a simulation study was performed, calculating the mean squared errors (MSE) for $\hat{\lambda}$ and $\check{\lambda}$, $E[(\hat{\lambda} - \lambda)^2]$ and $E[(\check{\lambda} - \lambda)^2]$. The settings simulated had values of λ ranging from 0.4 to 10.0 in increments of 0.05, for $n = 50$ and $n = 200$ 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 $\frac{\sqrt{MSE}}{\lambda}$.

The figures confirm the intuitive speculation described above. For a sample size of 50, $\check{\lambda}$ performs better than $\hat{\lambda}$ for $\lambda < 3.7$, while for $n = 200$ the change point comes earlier, at approximately $\lambda = 3.0$. In other words, a sample size of $n = 200$ is large enough so that we will get a fairly large value of M_n even if λ is small.

6.2 Performance on Real Data

The author applied the methodology developed here to three of his Web pages, listed here with the corresponding numbers of accesses:

http://heather.cs.ucdavis.edu/˜matloff/unix.html	22853
http://heather.cs.ucdavis.edu/˜matloff/latex.html	14993
http://heather.cs.ucdavis.edu/˜matloff/chinese.html	4469

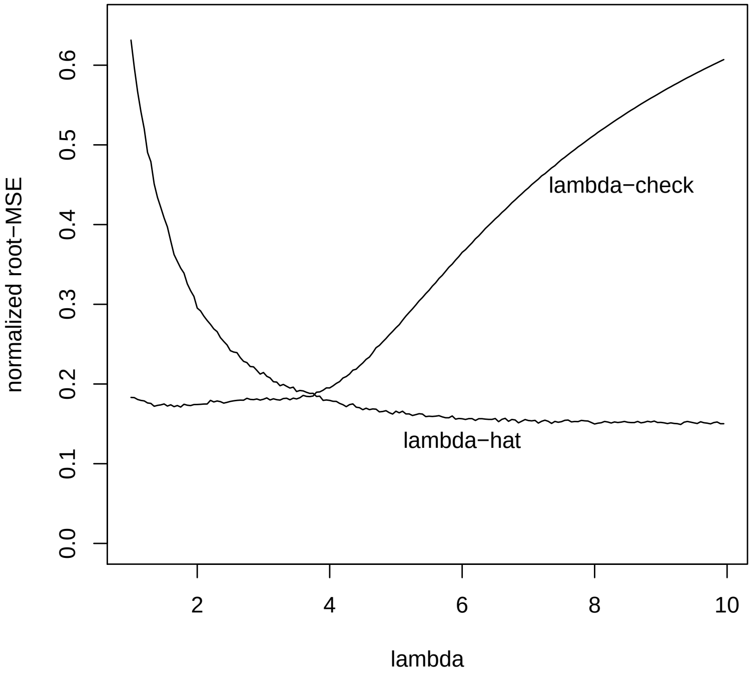

Fig. 1. MSE comparison, $n = 50$.

Since this was direct data T_{ij}, rather than time-of-last-A/M, the author could determine the true values of λ from the data (though technically these too were just estimates), and then compare them to the values of the estimators $\check{\lambda}$ and $\hat{\lambda}$ computed from the indirect data M_n and W_i. In other words, the data served as a good real-world test bed for the methodology. Again, keep in mind that here we are playing the role of an analyst who would only have access to M_n and the W_i.

First let us assess the quantities Z_i for an exponential distribution, as discussed in Section 5.2. A kernel-based density estimate for the case of the UNIX data set, using the R statistical package [**?**] with the default value for the smoothing parameter, is shown in Figure 3. The estimate is not monotone decreasing, as it would be if the parent population were to have an exponential distribution.

Moreover, Figure 4 suggests that the access times have a time-varying rate. This figure was generated from the UNIX data by converting the time of day of each A/M transaction to minutes since 12:00 a.m. of that day, and then computing a kernel density estimate from that data. As suspected, there tended to be fewer accesses in the middle of the night (U.S. time).

The corresponding graphs for the LaTeX and Chinese-software data sets, not shown here, were similar. Thus these data sets provide an opportunity to examine

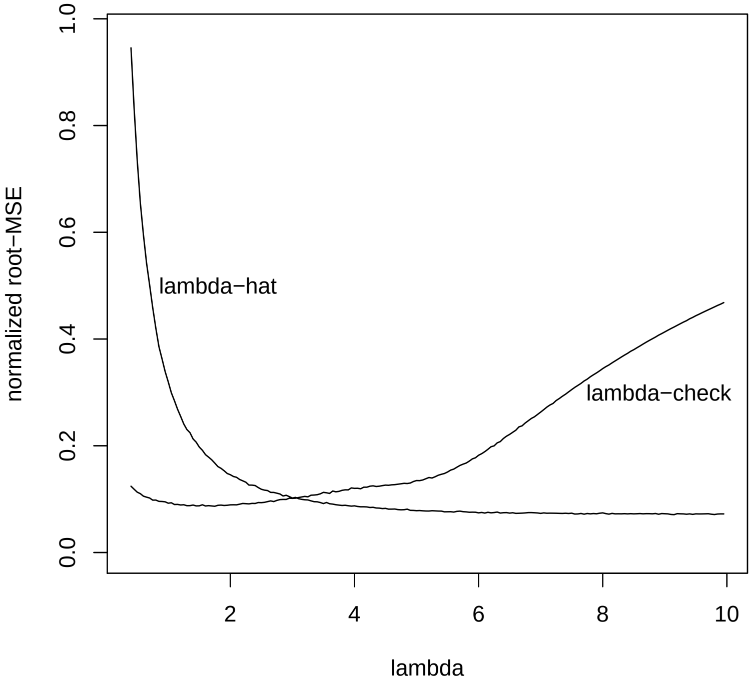

Fig. 2. MSE comparison, $n = 200$

the robustness of the estimators $\hat{\lambda}$ and $\check{\lambda}$ to the time-homogeneous and exponential assumptions.

Let us calculate $\hat{\lambda}$ and $\check{\lambda}$ on the UNIX data set, for various values of τ. The results are shown in Figure 5. (Time units are minutes.) First note that, similar to the simulation results in Section 6.1, $\hat{\lambda}$ tends to be a superior estimator relative to $\check{\lambda}$ only for larger values of τ. (The latter property may be due to the observation made in Section 5.5 regarding a strategy for dealing with nonhomogeneous Poisson processes.) In other words, $\max(\hat{\lambda}, \check{\lambda})$ seems to be the general estimator of choice, and that estimator does fairly well on these data sets, in spite of their departure from the Poisson assumption.

Second, we see that as predicted by the theoretical analysis in Section 5.3, $\check{\lambda}$ does quite well in the case of small τ.

Interestingly, we find similar results for the LaTeX and Chinese-software data sets, as seen in Figures 6 and 7, respectively.

7. COMPARISON WITH THE WORK OF CHO AND GARCIA-MOLINA

Another approach to this problem was taken by Cho and Garcia-Molina (CGM) [?]. Those authors, and the author of the present paper, became aware of each other's

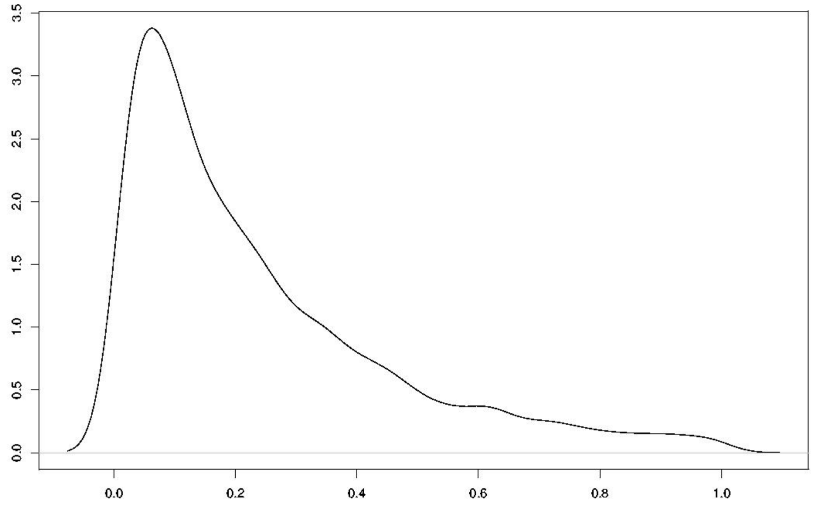

Fig. 3. Kernel estimate of Z density.

work in late 2002, after both papers had been submitted for publication. Thus the work on each of the two papers was done independently of the other. This section compares the results of the two papers. The important points of comparison are as follows.

Each of the two papers presents two Poisson-based estimators. CGM's first estimator is identical to $\check{\lambda}$ in the present paper. (As noted earlier, CGM also present a modification of this estimator.) Both papers made the Poisson assumption for this estimator, though they arrived at it from different approaches (CGM from a bias-reduction argument, the present paper using MLE).

Each paper again makes the time-homogeneous Poisson assumption for its second estimator. However, CGM's second estimator is different from the present paper's $\hat{\lambda}$. CGM's second estimator has an advantage in that it is simpler to compute than $\hat{\lambda}$, but $\hat{\lambda}$ has an advantage in being statistically optimal.

CGM present a version of their first estimator for the case of irregular (though deterministic) sampling intervals. The analysis of the present paper does not cover that case. However, in practice the sampling would typically be done via shell scripts which arrange to vist the site at regular intervals.

The present paper develops methodology for performing statistical inference, i.e. confidence intervals and hypothesis tests, using the estimators, an aspect not treated by CGM.

The most important differences between the two papers involve their coverage of cases in which the assumptions do not hold:

—*Theoretical Depth.* CGM do not include theoretical analyses for cases in which the assumption of a time-homogeneous Poisson process is violated. The present paper

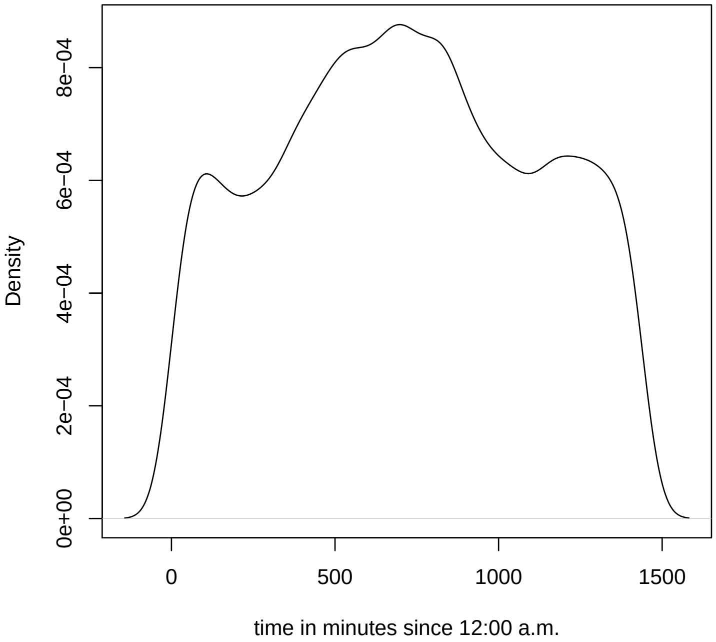

Fig. 4. Kernel estimate of time-of-day density.

develops a theoretical analysis proving that $\check{\lambda}$ is valid in non-Poisson renewal process settings if τ is small. The present paper also develops some theoretical analysis of the robustness of the homogeneous Poisson-based estimators in the nonhomogeneous Poisson case.

—*Estimation Methodology for the Non-Poisson Case.* CGM presents no estimators aimed specifically at the non-Poisson case. The present paper proposes two such estimators, though with some questions still to be answered.

—*Empirical Analysis.* Both papers perform investigations on real Web A/M data. CGM finds that their first estimator (the only one investigated) produces a bias of averaging about 15% on the various real data sets considered.

The present paper finds that on the real data the first estimator works well for small τ, as predicted by the theory, and the second estimator works well for large τ. Moreover, the present paper finds that the maximum of its first two estimators works well, with a bias of around 10%.

—*Overall Assessment of the Poisson Model.* CGM cite references which suggest that the Poisson model is a good one for Web modification rates, but do not assess the model. The present paper finds that the model is not very good for the access rate data it analyzes.

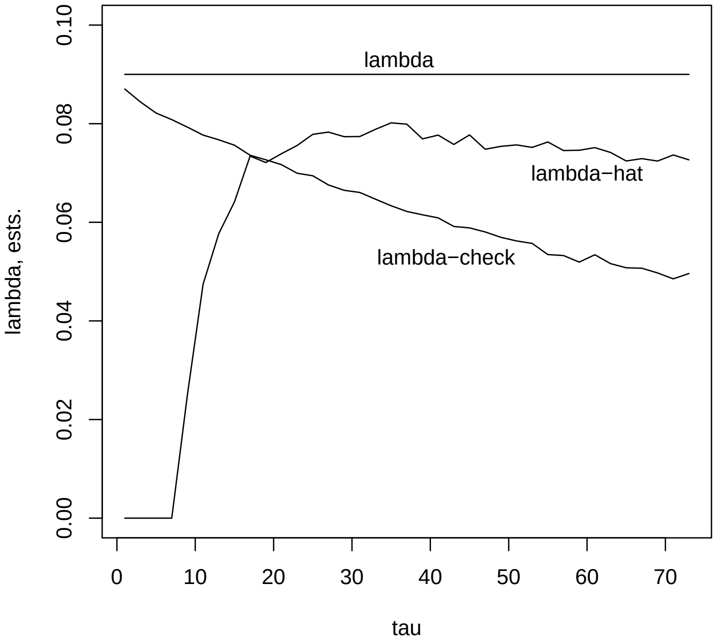

Fig. 5. Estimates on UNIX data.

8. CONCLUSIONS AND DISCUSSION

Four solutions—two parametric and two nonparametric—are proposed here for a problem which at first might seem to be fundamentally intractable, estimation of an A/M rate based on last A/M times within intervals. Although the derivation for the first two estimators, $\check{\lambda}$ and $\hat{\lambda}$, is based on a Poisson assumption for the data, theoretical analysis presented here shows that one of the estimators works well for the small-τ case without the Poisson assumption. The other estimator is statistically optimal under the Poisson assumption.

Tests on three sets of real, non-Poisson Web data presented here not only confirmed that $\check{\lambda}$ works well in the small-τ case without the Poisson assumption, but also show that $\hat{\lambda}$ works reasonably well on non-Poisson data in the case of large τ. The combined estimator max($\check{\lambda}$,$\hat{\lambda}$) seems to work very well across the range of τ studied.

The Poisson-based work done here sheds additional light on the work done independently by Cho and Garcia-Molina. The present paper goes much further in the non-Poisson case than do Cho and Garcia-Molina. A theoretical analysis of the behavior of $\check{\lambda}$ in the non-Poisson case for small τ is presented, in which it is found that this estimator is robust to the Poisson assumption. Then two nonparametric solutions are derived, based on a novel use of seemingly-unrelated statistical

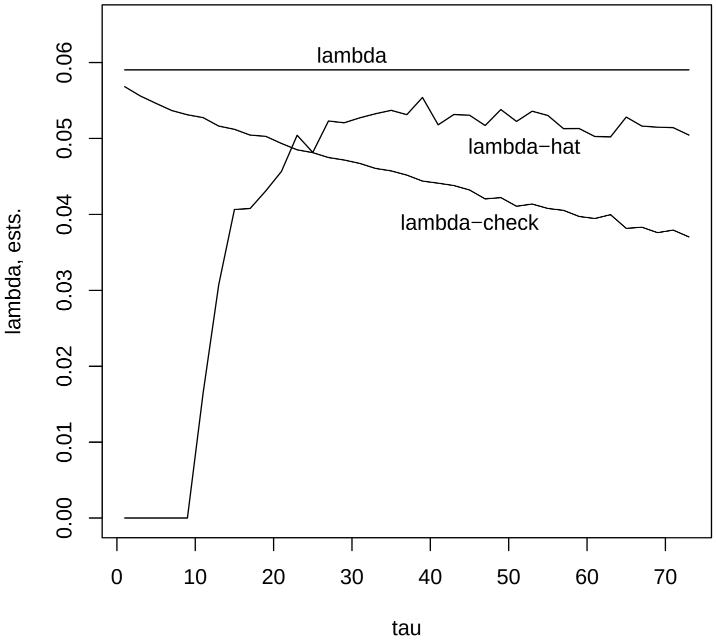

Fig. 6. Estimates on LaTeX data.

methodology. They appear to have promise, but future work is needed to fully develop their potential.

Another area of possible interest would be to investigate the case in which the A/M rate has a (nonperiodic) trend in time.

ACKNOWLEDGMENTS

The author wishes to thank the editor-in-chief for his extraordinarily careful and conscientious review.

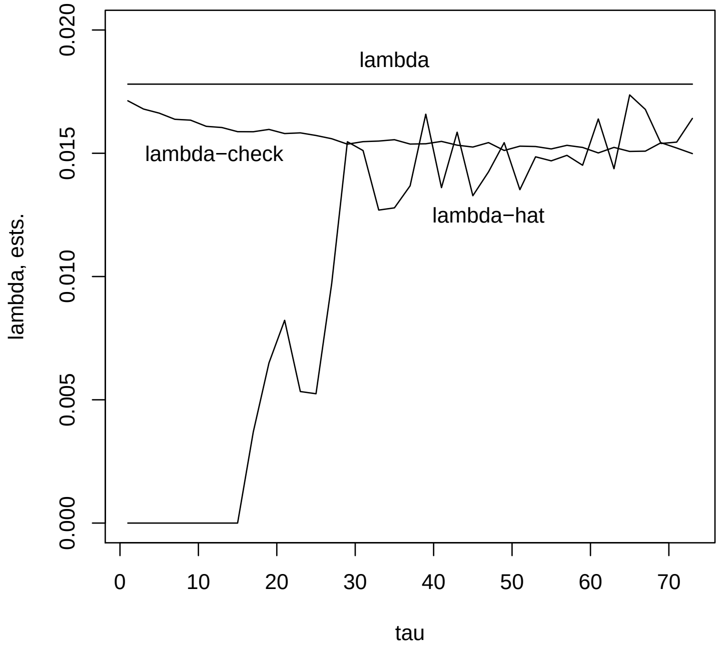

Fig. 7. Estimates on Chinese software data.