

Modeling Bid Arrivals in Online Auctions

Galit Shmueli[†]

Department of Decision and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742.

Ralph P. Russo

Department of Statistics & Actuarial Science, University of Iowa, Iowa City, IA 52242.

Wolfgang Jank

Department of Decision and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742.

Summary. We introduce a new family of non-homogeneous Poisson processes (NHPP) that are useful for modeling pure and contaminated self-similar processes which describe arrivals within a finite time period. Our motivation comes from the bid arrival process in online auctions. Modeling bid arrivals in online auctions is challenging since bidding dynamics change over the course of the auction. While the start of the auction typically sees an unusual amount of early bidding which is followed by a period of little activity, the auction end typically experiences an enormous amount of last minute bidding, also known as “sniping.” This observed heterogeneity in bidding dynamics commands a very flexible class of models. We address these modeling challenges by proposing a class of 3-stage non-homogeneous Poisson processes. We investigate the probabilistic and statistical properties of these models and illustrate their usefulness for fitting and interpreting real data from eBay.com.

KEY WORDS: Non-homogeneous Poisson process; online auction; bid data, self-similarity, bidding dynamics.

1. Introduction and Motivation

The bid arrival process in online auctions is central to understanding many phenomena related to the process of bidding, the outcome of auctions, and the dynamics of bidding. In the existing literature, many researchers have assumed that the bid or bidder arrivals follow a Poisson process (Zhang et al, 2002). Others have used a non-stationary Poisson Process (Vakrat & Seidmann, 2000). However, empirical studies of the bid process in online auctions suggest that the arrival of bids during a closed-end auction is closely related to self-similar processes (Roth & Ockenfels, 2000). In an attempt to investigate this phenomenon we collected data on 3651 bid times, placed in 189 Palm M515 online auctions on eBay.com. Figures 1 and 2 display the empirical CDFs for these bid arrivals. The CDF is plotted at several different resolution levels, “zooming-in” from the entire auction duration (of 7 days) to the last day, the last hour, the last 5 minutes, etc. until the very last minute. Notice that the CDF, as expected of a self-similar process, increases at the same rate independent of the scale, as can be seen in the first 4 or 5

[†]*Address for correspondence:* Galit Shmueli, Department of Decision and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742
E-mail: gshmueli@rhsmith.umd.edu

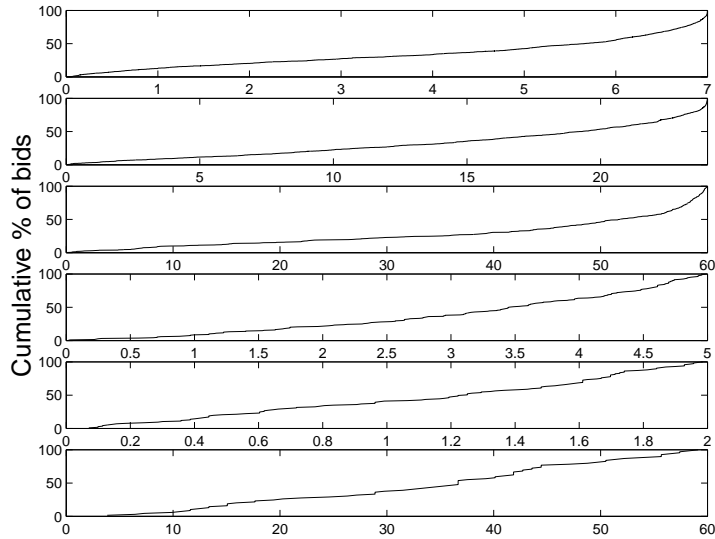


Fig. 1. Empirical CDF of number of bids in 189 Palm M515 auctions. The plots (top to bottom) are for the entire auction (7 days), the last day (24 hours), last hour (60 min), last 5 min, last 2 min and last 1 min (60 seconds) of the auction.

curves in Figure 1. Interestingly, however, for the last minute of the auction this pattern breaks down (see bottom curve in Figure 1). Self-similarity, it appears, is not prevalent throughout the entire auction duration!

Self similar processes appear to play an important role in online auctions, but they are also central in applications such as web, network and ethernet traffic. Unlike the latter areas where data can be collected at very small time intervals (e.g., in milliseconds), yielding very long and dense time-series, online auction data tend to be much sparser and shorter. One of the reasons for this is the short duration of auctions (only a few days). Careful inspection of online auction data raises a central concern: The “self-similarity” that is seen during the start and middle of auctions seems to break down at the last moments of the auction. Furthermore, it appears that the bid arrival process is not homogeneous, changing from the beginning of the auction, through the middle of the auction, and at the last moments of the auction (Roth & Ockenfels, 2000). This phenomenon has been observed and researchers have tried to explain it. However, there have been no attempts to model the self-similarity of the entire bid arrival process together with the breakdown.

In this paper we suggest a family of non-homogeneous Poisson processes (NHPP) that lead to models that are suitable for the empirical phenomena described above. We start by showing how a pure self similar process can be represented as a non-homogeneous Poisson process (NHPP₁) using a special intensity function (Section 2). We describe the properties of this process, its use for computing arrival probabilities and ratios, and describe practical issues such as estimation and simulation. Next, we introduce an improved model (NHPP₂) which has a dual intensity function. This model is aimed at capturing the last minute breakdown of self-similarity and naturally incorporates the phenomenon of “sniping” or last minute bidding. The beginning of the process is no longer a pure self similar process, but rather a contaminated one. We investigate this process, its properties, and estimation and simulation issues in Section 3. We present the most flexible model in Section 4. This model accounts for two frequently

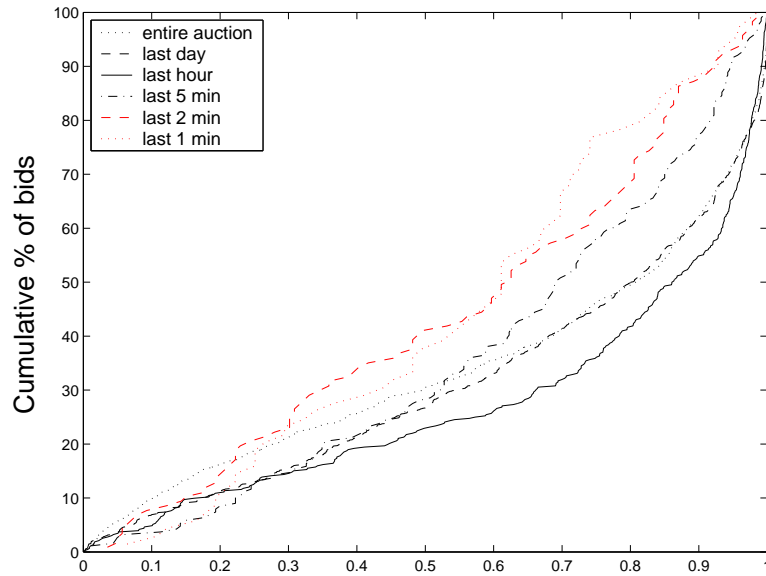


Fig. 2. Empirical CDFs of number of bids in 189 Palm M515 auctions overlaid

observed phenomena in the online auction literature: “early bidding” and “last minute bidding” (sniping). Empirical investigations of online auctions suggest that the bidding behavior is not homogeneous throughout the entire auction. Rather, after experiencing an initial surge in bidding activity (“early bidding”), the speed of the arriving bids slows down. It slowly picks-up again mid-way through the auction and steadily increases towards the auction end, only to culminate in a very intense rate of bid arrivals (“last minute bidding”). Our most flexible model formulation (NHPP₃) accounts for these different bidding dynamics.

We use simulated data as well as real online auction data from eBay.com to illustrate the different models and to show the contaminated self-similarity features of the bid arrivals and their breakdown at the beginning and end of the auction. In Section 5 we discuss the uses of the proposed bid arrival models for exploring and researching online auctions.

2. NHPP₁: A Non-Homogeneous Poisson Process That Leads to a Self Similar Process

2.1. Background: Online Auctions and Self Similar Processes

Our work is motivated by the empirical phenomenon that the bid arrival times in online auctions appear to have the self similarity property (Roth & Ockenfels, 2000). This finding is interesting, because self similarity is one of the main features of web and network traffic. Although the reasons behind self similarity of network traffic have not been clearly identified (Crovella & Bestavros, 1995), there have been several attempts to explain it. According to Mandelbrot (1969), and Willinger et al. (1995), self-similar traffic can be constructed by aggregating a large number of active and inactive sources where the lengths of the active and inactive periods are *iid*, independent of one another, and have infinite variances. This assumes that there is a non-negligible probability that the active and non-active periods can last a very long time. In the network traffic applications this could be achieved by a network of workstations, each of which is either silent or transferring data at a constant rate (Crovella & Bestavros, 1995). Crovella

& Bestavros (1995) explain the self similarity in terms of “file system characteristics and user behavior.” They show that for the active period the distribution of transfer times, the distribution of user requests for documents, and the underlying distribution of document sizes available on the Web are heavy tailed. For the inactive periods, inter-request times are also heavy tailed. In the online auction setting we can think of the behavior of individual bidders as an on/off behavior. Active periods occur when a user is submitting a bid, while an inactive period occurs when the user passively participates (e.g., by monitoring the website) but does not submit a bid. In practice whether the user is monitoring the auction or not is unknown. Crovella & Bestavros (1995) identify several types of inactive periods: not browsing the web, busy from the previous download job, or user is inspecting results from last download. They separate these types and show which are heavy tailed and which are not. In addition to those bidders who succeed in placing a bid, there are (very likely) more bidders that monitor the auction and/or attempt to place bids that are too low to get registered. Intuitively, non-active periods can last very long for some types of bidders. Bapna et al. (2003) divide bidders into evaluators, participators, and opportunists. Evaluators are bidders who place a single early bid, and opportunists are bidders who place a single late bid. These two types of bidders would add to the non-active periods. Finally, the heavy-tailed distribution of “user think times” also seems to be a feature of human information processing (Crovella & Bestavros, 1995). The implications of bid arrivals following a self-similar process and not the widely-assumed Poisson model are significant: The levels of activity throughout an auction with self similar bid arrivals would increase at a much faster rate than expected under a Poisson model. It would be especially meaningful towards the end of the auction, which has a large impact on the bid amount process and the final price.

2.2. Model Formulation

A non-homogeneous Poisson process differs from an ordinary Poisson process in that its intensity is not a constant but rather a function of time. We suggest a particular intensity function that leads to a self-similar process: Suppose bids arrive during $[0, T]$ in accordance with a non-homogeneous Poisson process $N(t)$, $0 \leq t \leq T$, with intensity function

$$\lambda(s) = c \left(1 - \frac{s}{T}\right)^{\alpha-1} \quad \text{some } 0 < \alpha < 1 \text{ and } c > 0 \quad (1)$$

so $\lambda(s) \rightarrow \infty$ as $s \rightarrow T$. That is, the bidding becomes increasingly intense as the auction deadline approaches. The r.v. $N(t)$ has a *Poisson*($m(t)$) distribution, where

$$m(t) = \int_0^t \lambda(s) ds = \frac{Tc}{\alpha} \left[1 - \left(1 - \frac{t}{T}\right)^\alpha\right]. \quad (2)$$

Given that $N(T) = n$, the joint distribution of the arrival times X_1, \dots, X_n is equivalent to that of the order statistics associated with a random sample of size n from a distribution whose pdf is shaped like λ on the interval $[0, T]$, namely

$$f(s) = \frac{\lambda(s)}{m(T)} = \frac{\alpha}{T} \left(1 - \frac{s}{T}\right)^{\alpha-1} \quad 0 < s < T. \quad (3)$$

$f(s)$ is a proper density for any $\alpha > 0$. Some special cases include uniform arrivals ($\alpha = 1$), and the triangular density ($\alpha = 2$).

The *CDF* and the survival function corresponding to (3) are

$$F(s) = \frac{m(s)}{m(T)} = 1 - \left(1 - \frac{s}{T}\right)^\alpha \quad 0 \leq s \leq T \quad (4)$$

and

$$R(s) = \left(1 - \frac{s}{T}\right)^\alpha \quad 0 \leq s \leq T \quad (5)$$

with the hazard function

$$h(s) = \frac{f(s)}{R(s)} = \frac{\alpha}{T-s} \quad \text{for } 0 \leq s < T. \quad (6)$$

Observe that for $0 < t \leq T$ and $0 \leq \theta \leq 1$ we have

$$\frac{R(T - \theta t)}{R(T - t)} = \theta^\alpha \quad (7)$$

which is not dependent on t .

This is equivalent to the formulation in Roth & Ockenfels (2000), who model the number of bids in the *last* minutes, and obtain the CDF

$$F^*(t) = \left(\frac{t}{T}\right)^\alpha. \quad (8)$$

The difference between the CDF in (4) and that in (8) results from time reversal. The former looks at the number of bids from the *beginning* of the auction until time s , while the latter looks at the number of bids in the *last* s time units of the auction. By setting $t = T - s$ in (8) and subtracting from 1, we obtain the same expression for the CDF as in (4).

Let F_e denote the empirical CDF corresponding to the $N(t)$ bid arrival times on $[0, T]$, and let $R_e = 1 - F_e$ denote the associated survival function.

2.3. Model Implications

Intensity Function Structure and Self-Similarity

One of the main features of a self similar process is that the distribution (the CDF, autocorrelation function, etc.) remains the same at any level of aggregation. Using the intensity function in (1), we show that this is also a property of NHPP₁: If we have m independent processes $N_j(t)$, $1 \leq j \leq m$, with intensity functions

$$\lambda_j(s) = c_j \left(1 - \frac{s}{T}\right)^{\alpha-1}, \quad 1 \leq j \leq m \quad (9)$$

(i.e., different c_j 's, but the same α), then all have the right form to possess the self similar property (7). The aggregated process $N(t) = \sum_{j=1}^m N_j(t)$ is a NHPP with intensity function

$$\lambda(s) = \sum_{j=1}^m \lambda_j(s) = c \left(1 - \frac{s}{T}\right)^{\alpha-1}, \quad c = \sum_{j=1}^m c_j \quad (10)$$

and also has property (7). Conversely, if we start with a NHPP $N(t)$ having intensity function $\lambda(s)$ given by (10) and randomly categorize each arrival into one of m "types" with respective probabilities $c_1/c, \dots, c_m/c$ then the resulting processes $N_1(t), \dots, N_m(t)$ are independent NHPP's with intensity functions as in (9). If we aggregate two processes with $\alpha_1 \neq \alpha_2$, the resulting process is a NHPP, but with an intensity function of the wrong form (even when $c_1 = c_2$) to satisfy (7).

Another self-similarity property that emerges from the intensity function is that when "zooming-in" into smaller and smaller intervals, the intensity function has the same form, but on a different time scale: Fix a time point $\beta T \in [0, T]$. The process $N_\beta(s) := N(s)$ on the shortened interval $[\beta T, T]$ is a NHPP with intensity function $\lambda_\beta(s) = \lambda(s)$ for $x \leq s \leq T$. This can be written as

$$\begin{aligned}
\lambda_\beta(s) &= c\left(1 - \frac{s}{T}\right)^{\alpha-1} \\
&= [c(1-\beta)^{\alpha-1}] \left(1 - \frac{s - \beta T}{T(1-\beta)}\right)^{\alpha-1} \\
&= \lambda(\beta) \left(1 - \frac{s - \beta T}{T(1-\beta)}\right)^{\alpha-1}
\end{aligned}$$

where $\lambda(\beta) \rightarrow \infty$ as $\beta \rightarrow 1$. Note that originally,

$$\lambda_0(s) = \lambda(0)\left(1 - \frac{s}{T}\right)^{\alpha-1}.$$

If we think of the fixed time point βT as the new *zero*, and we change time units so that $T(1-\beta)$ minutes on the old scale = T *shminutes*‡ on the new scale, the N_β process is defined on the interval $[0, T]$ and has intensity function

$$\lambda_\beta(s) = \lambda(\beta) \left(1 - \frac{s}{T}\right)^{\alpha-1} \quad 0 \leq t \leq T, \text{ time now measured in shminutes}$$

Thus, at any given point βT in $[0, T]$, the process of remaining bids is like the original process on $[0, T]$, but is on a reset and faster ($1/(1-\beta)$ faster) clock, and is more intense by a factor of $\lambda(\beta)/\lambda(0)$.

Probabilities and Ratios of Arrival

Let X be a random variable with CDF as in (4). The mean arrival time over the entire period $[0, T]$ is

$$E(X) = \int_0^T R(s)ds = T/(1+\alpha) \quad (11)$$

and the variance is

$$Var(X) = \frac{\alpha T^2}{(\alpha+2)(\alpha+1)^2}. \quad (12)$$

The probability that a bid will be placed within a time interval of length t depends on the location of the interval through the following function: For $0 < x < x+t < T$,

$$\begin{aligned}
p(x, t) &: = P(\text{receive a bid during } [x, x+t]) \\
&= 1 - \exp[m(x) - m(x+t)] \\
&= 1 - \exp\left\{\frac{cT}{\alpha} \left[\left(1 - \frac{x+t}{T}\right)^\alpha - \left(1 - \frac{x}{T}\right)^\alpha\right]\right\}.
\end{aligned}$$

In the context of *sniping*, or last minute bidding, a useful probability is that of the event where there are no bids after time x . This is given by

$$1 - p(x, T-x) = \exp\left\{-\frac{cT}{\alpha} \left(1 - \frac{x}{T}\right)^\alpha\right\} \rightarrow 1 \text{ as } x \rightarrow T. \quad (13)$$

‡We use the term “shminute/s” to signify a new unit of time.

Another quantity of interest is the conditional distribution of the next bid after time x , given that there *is* a next bid:

$$\begin{aligned}
 p^*(x, t) &: = P(\text{bid arrives during } [x, x+t] \mid \text{there is a bid after time } x) \\
 &= \frac{p(x, t)}{p(x, T-x)} \\
 &= \frac{\exp[\frac{cT}{\alpha}(1 - \frac{x}{T})^\alpha] - \exp[\frac{cT}{\alpha}((1 - \frac{x+t}{T})^\alpha)]}{\exp[-\frac{cT}{\alpha}(1 - \frac{x}{T})^\alpha] - 1}.
 \end{aligned} \tag{14}$$

Finally, a variable of special interest (Roth & Ockenfels, 2000) is the ratio of the number of arrivals within a fraction θ of the last t moments of the auction, and the number of arrivals within the last t moments. For $0 < t \leq T$ and $0 \leq \theta \leq 1$ define

$$\pi(t, \theta) := \frac{N(T) - N(T - \theta t)}{N(T) - N(T - t)} = \frac{R_e(T - \theta t)}{R_e(T - t)}. \tag{15}$$

We set $\pi(t, \theta) = 0$ if $R_e(T - t) = 0$. It can be shown that $\pi(t, \theta)$ converges uniformly in probability as $c \rightarrow \infty$ (see Appendix A for proof). This means that if the bidding is reasonably intense over the interval $[0, T]$, then there is a high probability – the higher the value of c , the greater the probability – that all the π functions ($\pi(t, \theta)$, $0 < t \leq T$) will be close to θ^α , for all t that are not too close to 0. Thus, the model does not guarantee convergence for small t . This accommodates, or at least does not contradict, the empirical evidence that self-similarity of bid arrival processes breaks down at the very last minute (Roth & Ockenfels, 2000). In their graphs of the empirical CDF (Roth & Ockenfels, 2000, pp. 30-31), the empirical $\pi(t, \theta)$ functions are displayed as functions of θ , for various choices of t . If t is not too small, these graphs all seem similar to $g(\theta) = \theta^\alpha$.

Regarding the behavior of $\pi(t, \theta)$ for fixed $t \in (0, T)$ and $\theta \in (0, 1)$, it can be shown that as $c \rightarrow \infty$

$$P\left(|\pi(t, \theta) - \theta^\alpha| > \frac{x}{\sqrt{c}}\right) \rightarrow P\left(|Z| > x \sqrt{\frac{t^\alpha T^{1-\alpha}}{\alpha \theta^\alpha (1 - \theta^\alpha)}}\right) \tag{16}$$

where Z is the unit normal variable (see appendix B for detailed proof).

2.4. Model Estimation

Under the NHPP₁ model, conditional on the event $N(T) = n$, the bid arrival times X_1, \dots, X_n are distributed as the order statistics of a random sample of size n from the distribution in (4). These variables, when randomly ordered, are equivalent to a random sample of size n from that distribution.

Given a set of n arrival times on the interval $[0, T]$ we want to assess whether the NHPP₁ model adequately describes the data, and if so, to estimate the parameter α that influences the intensity of the arrivals and controls the shape of the distribution of bid times. Due to the special form of the CDF in (4), which can be approximated by the empirical CDF $F_e(t)$, a log-log graph of $1 - F_e(t)$ vs. $(1 - t/T)$ should reveal the adequacy of the fit. If the model is reasonable we expect the points to fall on a line which has a slope of α . Then α can be estimated as the slope of the least squares line fitted to the n points. This type of plot is widely used for assessing self-similarity in general, the only difference being that we have a finite interval $[0, T]$ whereas usually the arrival interval is infinite $[0, \infty]$.

We present two alternative estimators of α : one based on the moment method and the other on maximum likelihood.

Moment Estimator

Using (11) we can derive the moments estimator (conditional on $N(t) = n$):

$$\hat{\alpha} = \frac{T}{\bar{X}} - 1 \quad (17)$$

This is quick and easy to compute. Although the estimator is biased for estimating α , it is consistent, since it is a continuous function of \bar{X} . Thus, for a very large sample we expect to get an estimate close to α .

Maximum Likelihood Estimator

Using the density in (3), the log-likelihood function of α , given n observations from NHPP₁, is:

$$\mathcal{L}(\alpha|x_1, \dots, x_n) = n \log \alpha/T + (\alpha - 1) \sum_{i=1}^n \log \left(1 - \frac{x_i}{T}\right) \quad (18)$$

This means that this distribution is a member of the exponential family. An important and useful implication is the existence of a sufficient statistic, namely $\sum_{i=1}^n \log \left(1 - \frac{x_i}{T}\right)$. This enables a great reduction in storage for the purpose of estimation. All that is required is the sufficient statistic and not the n arrival times.

The maximum likelihood estimate for α is then given by

$$\hat{\alpha} = -n \left[\sum_{i=1}^n \log \left(1 - \frac{x_i}{T}\right) \right]^{-1} \quad (19)$$

This is similar to the Hill estimator that is used for estimating a heavy tailed distribution. The two differences between this estimator and the Hill estimator are that in our case the arrival interval is finite ($[0, T]$), whereas in the general heavy tailed case the arrival interval is infinite ($[1, \infty]$). Secondly, the Hill estimator is used when the assumed process is close to a Pareto distribution only for the upper part of the distribution. Therefore the average in the Hill estimator is taken over only the k latest arrivals (Resnick, 1997), while in our case we assume that the entire bid arrival process comes from the same distribution and thus all arrival times are included in the estimator.

Note that the above estimator is conditional on $N(T) = n$. The unconditional likelihood is given by

$$L(c, \alpha) = P(N(T) = n) f(x_1, \dots, x_n) = \frac{(\exp(-\frac{Tc}{\alpha}))c^n}{n!} \prod_{k=1}^n \left(1 - \frac{x_k}{T}\right)^{\alpha-1} \quad (20)$$

which leads to the maximum likelihood estimates

$$\hat{\alpha} = -N(t) \left[\sum_{i=1}^{N(t)} \log \left(1 - \frac{X_i}{T}\right) \right]^{-1} \quad (21)$$

and

$$\hat{c} = \frac{N(T)}{T} \hat{\alpha} \quad (22)$$

The equivalent form of (19) and (21) stems from the fact that the parameter c affects the size $N(T)$ of the sample, but not the shape of the conditional distribution of the arrival times (see appendix D for a general version of this result). Asymptotics for $\hat{\alpha}$ as defined in (19) as $n \rightarrow \infty$, and (21) as $c \rightarrow \infty$, are given in Appendix C.

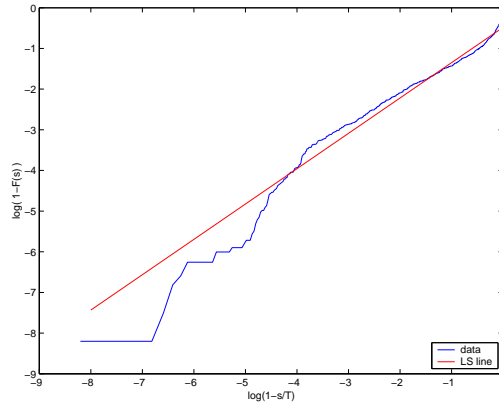


Fig. 3. Log-log plot of $1 - F_e(s)$ vs. $1 - s/T$ for Palm bid times

2.5. Simulation of $NHPP_1$

Although simulating self-similar processes on an infinite interval appears to be difficult from a practical point of view (Jeong et al., 1999), simulation on a finite interval is simple and efficient. We show a simple algorithm for simulating arrivals from $NHPP_1$, which create a self similar process. Since the CDF is easily invertible

$$F^{-1}(s) = T - T(1 - s)^{1/\alpha} \quad (23)$$

we can use it to simulate a specified number of $NHPP_1$ arrivals using the inversion method. To generate n arrivals, generate n $(0,1)$ -uniform variates u_k , $k = 1, \dots, n$, and transform each by plugging it into (23) in place of s :

$$x_k = T - T(1 - u_k)^{1/\alpha}. \quad (24)$$

2.6. Empirical & Simulated Results

For this study we collected bidding data for seven-day auctions of Palm M515 Personal Digital Assistant (PDA) units from Mid-March through June 2003 on eBay.com, resulting in 3561 bid times. Figure 3 displays $\log(1 - F_e(t))$ vs. $\log(1 - t/T)$ for these 3561 Palm M515 bid times. Had the data originated from a pure $NHPP_1$ -type self similar process, we would expect a straight line. From the figure it appears that the data follow a line for values of $\log(1 - t/T)$ in the approximate range $(-3, -0.25)$, which corresponds to the period of the auction between days 1.55 and 6.66. This includes about 42% of the arrivals. Figure 4 displays the same type of log-log plot but it uses only bid arrivals inside the interval $[1.55, 6.66]$. It can be seen that the bid arrivals in this interval are more consistent with the $NHPP_1$ model than are the bids from the entire 7-day auctions. The straight line is the least squares line fitted to the data. If we focus on bids that arrived only during this period (and scale them to a $[0, 1]$ interval), we obtain the following estimates for α : $\hat{\alpha}_{LS} = 0.91$, $\hat{\alpha}_{moments} = 1.03$, and $\hat{\alpha}_{ML} = 0.95$.

3. $NHPP_2$: Accounting for Last Moment Breakdown of Self-Similarity

The $NHPP_1$ model suggests that the rate of the incoming bids increases steadily as the auction approaches its end. Indeed, empirical investigations have found that many bidders wait until the very last possible moment to submit their final bid. By doing so, they hope to increase

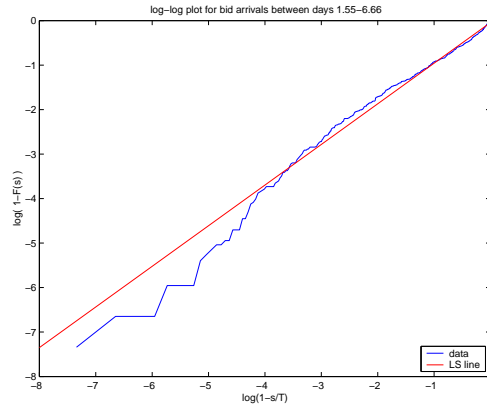


Fig. 4. Log-log plot of $1 - F_e(s)$ vs. $1 - s/T$ for Palm bid times between days 1.55-6.66

their chance of winning the auction since the probability that another competitor successfully places an even higher bid before closing is diminishing. This common bidding strategy, often referred to as “last minute bidding” or “bid sniping,” would suggest a steadily increasing flow of bid arrivals towards the auction end. However, empirical evidence from online auction data indicates that bid times over the last minute or so of closed-ended auctions tend to follow a uniform distribution. This has not been found in open-ended, or “popcorn” auctions, such as those on Amazon.com, where the auction continues 10 minutes after the last bid was placed. This phenomenon is seen in Figure 1, where the last-minute CDF looks different than the CDF on the other time scales. Such a phenomenon can occur if the probability of a bid not getting registered on the auction site is positive at the last moments of the auction, and increases as the auction comes to a close. There are various factors that may cause a bid not to get registered. One possible reason is the time it takes to manually place a bid (Roth & Ockenfels (2000) found that most last minute bidders tend to place their bids manually rather than through available sniping software agents). Other reasons are Hardware difficulties, internet congestion, unexpected latency, and server problems on eBay (see, for example, www.auctionsniper.com). Clearly, the closer to the end the auction gets, the higher the likelihood that a bid may not get registered successfully.

This increasing likelihood of an unsuccessful bid counteracts the increasing flow of last minute bids. The result is a uniform bid arrival process that “contaminates” the self-similarity of the arrivals until that point. In addition, it appears that there is no clear-cut line between the self-similar process at the beginning and the uniform process at the end. Rather, self-similarity appears to transition gradually into a uniform process. See, for example, the empirical CDF for the last 2 minutes of the Palm auctions in Figures 1 and 2.

3.1. Model Formulation

As before, we assume a Non-Homogenous Poisson Process, except that now the parameter α of the intensity function turns from α_1 to α_2 at the last d moments of the auction:

$$\lambda_2(s) = \begin{cases} c \left(1 - \frac{s}{T}\right)^{\alpha_1 - 1} & \text{for } 0 \leq s \leq T - d \\ c \left(\frac{d}{T}\right)^{\alpha_1 - \alpha_2} \left(1 - \frac{s}{T}\right)^{\alpha_2 - 1} & \text{for } T - d \leq s \leq T \end{cases} \quad (25)$$

Note that this intensity function is continuous, so there is no jump at time $T - d$. Also note that in this formulation the self-similar process transitions gradually into a uniform process if

$\alpha_2 = 1$. In such a case the intensity function flattens out during the final d moments, conveying a uniform arrival process during $[T - d, T]$. This seems to be the case in closed-ended online auctions, where the last minute breakdown of self similarity manifests itself as uniform arrivals of bids at the last minute of the auction (as reported in Roth & Ockenfels (2000) and can be observed in Figure 1).

In general, the beginning of the process is a contaminated self-similar process, and the closer it gets to the transition point (time $T - d$), the more contaminated it becomes. The random variable $N(t)$ which counts the number of arrivals until time t follows a Poisson distribution with mean value function

$$m_2(s) = \begin{cases} \frac{Tc}{\alpha_1} \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_1}\right) & \text{for } 0 \leq s \leq T - d \\ \frac{Tc}{\alpha_1} \left\{ \left[1 - \left(\frac{d}{T}\right)^{\alpha_1}\right] + \frac{\alpha_1}{\alpha_2} \left(\frac{d}{T}\right)^{\alpha_1} \left[1 - \left(\frac{d}{T}\right)^{-\alpha_2} \left(1 - \frac{s}{T}\right)^{\alpha_2}\right] \right\} & \text{for } T - d \leq s \leq T \end{cases} \quad (26)$$

Note that

$$m_2(T) = \frac{Tc}{\alpha_1} \left[1 - \left(1 - \frac{\alpha_1}{\alpha_2}\right) \left(\frac{d}{T}\right)^{\alpha_1}\right]. \quad (27)$$

The CDF of this process is then given by

$$F_2(t) = \frac{m_2(t)}{m_2(T)} = \begin{cases} \frac{1 - \left(1 - \frac{t}{T}\right)^{\alpha_1}}{1 - \left(1 - \frac{\alpha_1}{\alpha_2}\right) \left(\frac{d}{T}\right)^{\alpha_1}} & \text{for } 0 \leq t \leq T - d \\ 1 - \frac{\frac{\alpha_1}{T} \left(\frac{T-t}{d}\right)^{\alpha_2} \left(\frac{d}{T}\right)^{\alpha_1}}{1 - \left(1 - \frac{\alpha_1}{\alpha_2}\right) \left(\frac{d}{T}\right)^{\alpha_1}} & \text{for } T - d \leq t \leq T \end{cases}, \quad (28)$$

and the density by

$$f_2(t) = \begin{cases} \frac{\frac{\alpha_1}{T} \left(1 - \frac{t}{T}\right)^{\alpha_1 - 1}}{1 - \left(1 - \frac{\alpha_1}{\alpha_2}\right) \left(\frac{d}{T}\right)^{\alpha_1}} & \text{for } 0 \leq t \leq T - d \\ \frac{\frac{\alpha_1}{T} \left(\frac{d}{T}\right)^{\alpha_1 - \alpha_2} \left(1 - \frac{t}{T}\right)^{\alpha_2 - 1}}{1 - \left(1 - \frac{\alpha_1}{\alpha_2}\right) \left(\frac{d}{T}\right)^{\alpha_1}} & \text{for } T - d \leq t \leq T. \end{cases} \quad (29)$$

This is a proper density for any $\alpha_1 \neq 0, \alpha_2 > 0$, and $0 < d < T$. When $\alpha_1 = 0$, the form of the density is different than (29). This process describes a contaminated self-similar process. Of course, if $\alpha_1 = \alpha_2$ then NHPP₂ reduces to NHPP₁ which is a pure self-similar process. The contamination of the self-similarity means that as we get closer to $[T - d, T]$, the arrival process becomes less and less self-similar in a way that gradually changes into uniform arrivals in $[T - d, T]$. In Figure 1 it can be seen that the distribution during the last 2 minutes is somewhere between the uniform last-minute distribution and the earlier (almost) self-similar distribution. This transition can also be seen through the function $\pi(\theta, t)$, which changes form over three regions in the θ, t plane:

$$\pi_2(t, \theta) = \frac{1 - F_2(T - t\theta)}{1 - F_2(T - t)} = \begin{cases} \frac{(t\theta)^\alpha + (\alpha - 1)d^\alpha}{t^\alpha + (\alpha - 1)d^\alpha} & \text{for } t \geq \frac{d}{\theta} \\ \frac{\alpha t \theta d^{\alpha - 1}}{t^\alpha + (\alpha - 1)d^\alpha} & \text{for } d < t < \frac{d}{\theta} \\ \theta & \text{for } t \leq d. \end{cases} \quad (30)$$

3.2. Simulation of NHPP₂

To simulate the two-stage NHPP on the interval $[0, T]$ we can use the inversion method. The inverse CDF can be written in the form:

$$F_2^{-1}(s) = \begin{cases} T - T \left\{ 1 - s \left[1 - \left(1 - \frac{\alpha_1}{\alpha_2} \right) \left(\frac{d}{T} \right)^{\alpha_1} \right] \right\}^{1/\alpha_1} & \text{for } 0 \leq s \leq T - d \\ T - d \left\{ \frac{1-s}{1-F_2(T-d)} \right\}^{1/\alpha_2} & \text{for } T - d \leq s \leq T \end{cases} \quad (31)$$

The algorithm for generating n arrivals is then:

- (a) Generate n uniform variates u_1, \dots, u_n .
- (b) For $k = 1, \dots, n$ set

$$x_k = \begin{cases} T - T \left\{ 1 - u_k \left[1 - \left(\frac{d}{T} \right)^{\alpha_1} / F_2(T-d) \right] \right\}^{1/\alpha_1} & \text{if } u_k < F_2(T-d). \\ T - d \left\{ (1 - u_k) / [1 - F_2(T-d)] \right\}^{1/\alpha_2} & \text{if } u_k > F_2(T-d). \end{cases} \quad (32)$$

Note that

- (a) When $u_k = F_2(T-d)$ we get $x_k = T-d$, and
- (b) When $\alpha_2 = 1$, F_2 is linear on the interval $[T-d, T]$.

3.3. Fitting the NHPP₂ to data

3.3.1. Quick & Crude (CDF-Based) Estimation

Estimating the α Parameters

For estimating α_2 we can use the exact relation

$$\alpha_2 = \frac{\log R(t_2)/R(t'_2)}{\log(T-t_2)/(T-t'_2)} \quad (33)$$

where $R(t) = 1 - F_2(t)$ and t_2, t'_2 are within $[T-d, T]$. To estimate α_2 we pick reasonable values of t_2, t'_2 and use the empirical CDF. For the special case where the end of the auction is characterized by uniform arrivals, we have $R_e(t) \approx \text{const}(T-t)$ for $t \approx T$. Thus

$$\frac{\log(R_e(t_2)/R_e(t'_2))}{\log((T-t_2)/(T-t'_2))} \approx 1. \quad (34)$$

For estimating α_1 we cannot write an equation such as (33). However, it is possible to use an approximation which can be easily computed from the data. The approximation works for both intervals (i.e., for estimating α_1 and α_2), but we use it only for the first interval $[0, T-d]$. The calculations for the other period $[T-d, T]$ are the same. The idea is to choose an interval $[T-s, T-t]$ that we are confident lies within the period of interest. For example, in the Palm bid arrivals, we are confident that the first five days of the auction occur within the first period. Thus we choose an interval (or try several) that is contained in $[0, 5]$.

The mean value function for each of the two intervals $0 \leq y \leq T-d$ and $T-d \leq y \leq T$ is of the form

$$m_2(y) = \beta_j - \theta_j \left(1 - \frac{y}{T} \right)^{\alpha_j} \quad (35)$$

for $j = 1, 2$. For the first interval, fix $0 \leq T - s < T - t \leq T - d$. Writing α for α_1 , we have

$$\begin{aligned}
 \frac{N(T-t) - N(T-\sqrt{st})}{N(T-\sqrt{st}) - N(T-s)} &\approx \frac{E[N(T-t) - N(T-\sqrt{st})]}{E[N(T-\sqrt{st}) - N(T-s)]} \\
 &= \frac{m(T-t) - m(T-\sqrt{st})}{m(T-\sqrt{st}) - m(T-s)} \\
 &= \frac{\theta_1 \left[1 - \left(1 - \frac{T-t}{T}\right)^\alpha\right] - \theta_1 \left[1 - \left(1 - \frac{T-\sqrt{st}}{T}\right)^\alpha\right]}{\theta_1 \left[1 - \left(1 - \frac{T-\sqrt{st}}{T}\right)^\alpha\right] - \theta_1 \left[1 - \left(1 - \frac{T-s}{T}\right)^\alpha\right]} \\
 &= \frac{(ts)^{\alpha/2} - t^\alpha}{s^\alpha - (st)^{\alpha/2}} \\
 &= \frac{(s^{\alpha/2} - t^{\alpha/2})t^{\alpha/2}}{(s^{\alpha/2} - t^{\alpha/2})s^{\alpha/2}} \\
 &= \left(\frac{t}{s}\right)^{\alpha/2}. \tag{36}
 \end{aligned}$$

Taking logs in (36) we get

$$\alpha \approx 2 \frac{\log[N(T-t) - N(T-\sqrt{st})] - \log[N(T-\sqrt{st}) - N(T-s)]}{\log t - \log s} \tag{37}$$

This can be written in terms of $F_e(s)$, the empirical CDF:

$$\alpha \approx 2 \frac{\log[F_e(T-t) - F_e(T-\sqrt{st})] - \log[F_e(T-\sqrt{st}) - F_e(T-s)]}{\log t - \log s}. \tag{38}$$

The same approximation works on the interval $[T-d, T]$, but it is preferable to use the exact relation given in (33).

To learn more about the quick & crude estimates we generated 5000 random arrival times from NHPP₂ on the interval $[0, 7]$, with $\alpha_1 = 0.4$, $\alpha_2 = 1$, and $d = 5/10080$ (5 minutes) (see 3.2 for simulation algorithm). The left panel of Figure 5 shows that $\hat{\alpha}_1$ is close to 0.4 if a reasonable interval is selected. For an interval that excludes the last 20 minutes the estimate is 0.4. Using (33) we estimated α_2 . The right panel of Figure 5 shows $\hat{\alpha}_2$ as a function of the number of minutes before the auction close. The estimate seems to over-estimate the generating $\alpha_2 = 1$ value. Note that values of $T - t > 5$ are beyond the “legal” interval of the last 5 minutes. To refine the estimate, a value such as $\alpha_2 = 1.2$ can be used as an initial value in a maximum likelihood procedure.

Estimating d

The relation (28) between F_2 and d for $0 < t < T - d$ can be written in the form

$$d = T \left\{ \frac{1 - \frac{1 - (1 - \frac{t}{T})^{\alpha_1}}{F_2(t)}}{1 - \frac{\alpha_1}{\alpha_2}} \right\}^{1/\alpha_1} \tag{39}$$

Thus, using a “safe” initial value of t which we are confident is contained in the interval $[0, T - d]$, we can use (39), with F_e in place of F_2 to estimate d . Figure 6 illustrates the values of \hat{d} as a function of t . It can be seen that the estimate moves between 2-10 minutes, depending on the choice of t .

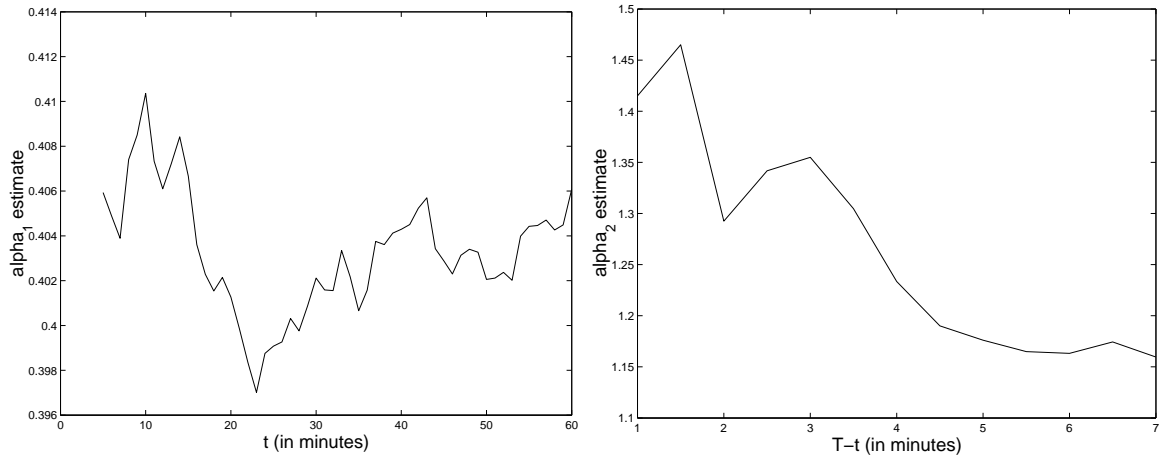


Fig. 5. Quick & crude estimates of α_1 as a function of t (with $s = 6.99$) (left), and of α_2 as a function of t (right), for the NHPP₂ simulated data with $\alpha_1 = 0.4, \alpha_2 = 1$.

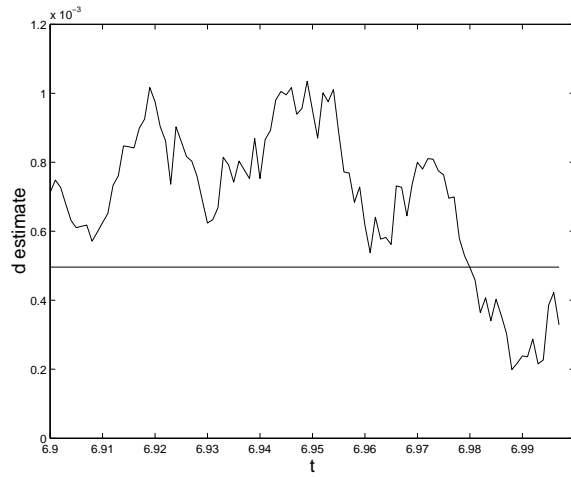


Fig. 6. Quick & crude estimate of d as a function of t .

3.3.2. Probability Plot for Special Case $\alpha_2 = 1$

Using the inverted CDF we can construct a probability plot. This can be used to quickly check the fit of the model to a given set of data. To draw the plot one must specify the duration of the auction (T) and the changepoint d .

3.3.3. Maximum Likelihood Estimation

Based on the density in (29), we obtain the likelihood function, conditional on $N(t) = n$ (see Appendix D for a note on unconditional estimation):

$$\begin{aligned} \mathcal{L}(x_1, \dots, x_n | \alpha_1, \alpha_2) = & \quad (40) \\ n \log \alpha_1 - n \log T + (\alpha_1 - 1) \sum_{i: x_i \leq T-d} \log \left(1 - \frac{x_i}{T}\right) + (\alpha_2 - 1) \sum_{i: x_i > T-d} \log \left(1 - \frac{x_i}{T}\right) + \\ & + n_2(\alpha_1 - \alpha_2) \log \frac{d}{T} - n \log \left(1 - \left(1 - \frac{\alpha_1}{\alpha_2}\right) \left(\frac{d}{T}\right)^{\alpha_1}\right) \end{aligned}$$

The two equations that need to be solved in order to obtain ML estimates for α_1 and α_2 are

$$\sum_{i: x_i \leq T-d} \log \left(1 - \frac{x_i}{T}\right) + n_2 \log \frac{d}{T} = n \frac{1 - (\alpha_2 - \alpha_1) \log \alpha_1}{\alpha_2(1 - (\frac{d}{T})^{-\alpha_1}) - \alpha_1} - \frac{n}{\alpha_1} \quad (41)$$

$$\sum_{i: x_i > T-d} \log \left(1 - \frac{x_i}{T}\right) - n_2 \log \frac{d}{T} = n \frac{-\alpha_1/\alpha_2}{\alpha_2(1 - (\frac{d}{T})^{-\alpha_1}) - \alpha_1} \quad (42)$$

where n_2 is the number of arrivals after $T - d$. In the uniform case ($\alpha_2 = 1$) where d is known (e.g., the “last minute bidding” phenomenon in online auctions), the maximum likelihood estimator of α_1 is the solution of the equation

$$\sum_{i: x_i \leq T-d} \log \left(1 - \frac{x_i}{T}\right) + n_2 \log \frac{d}{T} = n \frac{1 - (1 - \alpha) \log \alpha}{1 - \alpha - (\frac{d}{T})^{-\alpha}} - \frac{n}{\alpha} \quad (43)$$

This can be solved using an iterative gradient-based method such as Newton Raphson or the Broyden-Fletcher-Goldfarb-Powell (BFGP) method, which is a more stable quasi-Newton method that does not require the computation and inversion of the Hessian matrix (see, for example, Dennis and Schnabel, 1983). A good starting value would be the estimate obtained from the probability plot or the quick & crude method.

If d is unknown and we want to estimate it from the data, then a gradient approach can no longer be used. One option is to combine a gradient approach for α_1 and α_2 with an exhaustive search over a logical interval of d values. The size of the step in d should be relevant to the application at hand, and we expect that a small change should not lead to dramatic changes in $\hat{\alpha}_1, \hat{\alpha}_2$.

3.4. Empirical Results

To check for the presence of the “last-minute bidding” phenomenon in our Palm data, we started by fitting a probability plot. The left panel of Figure 7 shows a probability plot for the Palm data with $T = 7$ (days) and $d = 1/10080$ (the changepoint occurs one minute before the auction close). The plot contains several curves that correspond to different values of α . For these data it appears that the two-phase NHPP does not capture the relatively fast beginning! For example, with $\alpha = 0.5$ which yields the closest fit, the arrivals in the data seem to occur faster than expected under the NHPP₂ model during the first 2.5 days, then slow down more than expected until day 6 and then finally arrive at the expected rate until the end of the

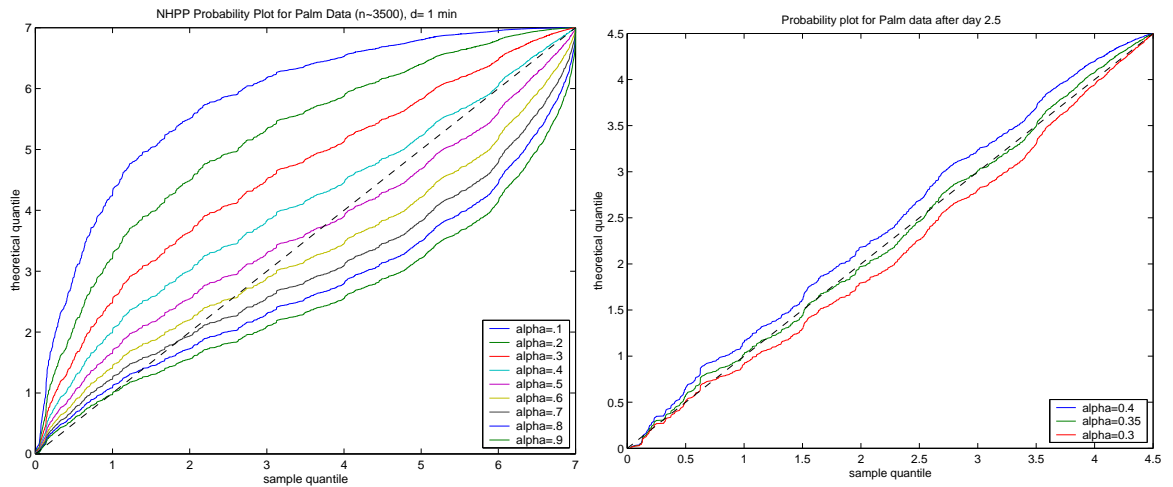


Fig. 7. Probability plots for Palm data. Left: $T = 7, d = 1/10080$, Right: $T = 4.5, d = 1/10080$

auction. To check that the first 2.5 days are indeed the cause of this mis-fit we drew only the last 4.5 days (Figure 7, right panel). Here the NHPP_2 model seems more appropriate and α can be estimated as 0.35.

Following these results we used only bids that were placed after day 2.5 (and shifted the data to the interval $[0, 4.5]$). We then obtained the quick & crude estimates $\hat{\alpha}_1 \approx 0.35$ and $\hat{\alpha}_2 \approx 1$. For estimating d we used the above estimates of α_1, α_2 and $t = 5/10080$ (5 minutes before the auction end), which is most likely contained within the first period $[0, 4.5 - d]$. This yields $\hat{d} = 2.8$ minutes. Figure 8 shows \hat{d} as a function of t . It appears that d is between approximately 1-3 minutes.

We also used a genetic algorithm to search for values of the parameters that maximize the likelihood (see Section 4). This yielded the estimates $\hat{\alpha}_1 = 0.37, \hat{\alpha}_2 = 1.1$, and $\hat{d} = 2.1$. A combination of a gradient method for estimating α_1, α_2 with an exhaustive search over d within a reasonable interval yielded the same estimates.

Finally, we compare our data to simulated data from an NHPP_2 with $d = 2.1$ minutes, $\alpha_1 = 0.37$, and $\alpha_2 = 1$. Figure 9 is a QQ-plot of the sorted Palm bid times (only those later than 2.5 days) and the sorted simulated data. It is clear that the fit is very good.

4. NHPP_3 : Accounting for Early and Last Moment Bidding

While our efforts so far in this paper have focused predominately on the auction-end and its last minute bidding activity, the beginning of the auction also features some interesting characteristics: early bidding! In fact, many empirical investigations of the online auctions have reported an unusual amount of bidding activity at the beginning of the auction followed by a longer period of little or no activity. The reasons for early bidding are not at all that clear. Bapna et al. (2003) refer to bidders who place a single early bid as evaluators but there may be other reasons why people place bids early in the auction. The next model incorporates early bidding and last minute bidding as a generalization of the NHPP_2 model. In order to model this early activity phase we formulate a 3-stage NHPP which has a continuous intensity function that

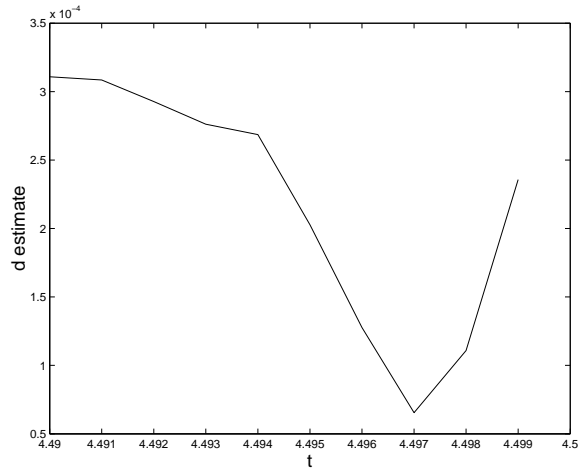


Fig. 8. Quick & crude estimate of d as a function of t .

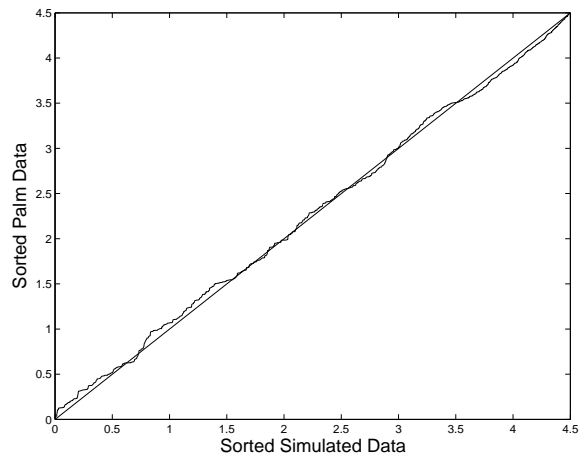


Fig. 9. QQ plot of Palm bid times vs. simulated $NHPP_2$ data on the interval $[0, 4.5]$ and parameters $\alpha_1 = 0.37$, $\alpha_2 = 1$, and $d = 2.1$ minutes.

switches the parameter α from stage to stage:

$$\lambda_3(s) = \begin{cases} c \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} \left(1 - \frac{s}{T}\right)^{\alpha_1 - 1} & \text{for } 0 \leq s \leq d_1 \\ c \left(1 - \frac{s}{T}\right)^{\alpha_2 - 1} & \text{for } d_1 \leq s \leq T - d_2 \\ c \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{s}{T}\right)^{\alpha_3 - 1} & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (44)$$

We expect α_3 to be close to 1 (uniform arrival of bids at the end of the auction) and $\alpha_1 > 1$ to represent the early surge in bidding. The random variable $N(t)$ which counts the number of arrivals until time t follows a Poisson distribution with mean

$$m_3(s) = \begin{cases} K \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_1}\right) & \text{for } 0 \leq s \leq d_1 \\ K \left(1 - \left(1 - \frac{d_1}{T}\right)^{\alpha_1}\right) + \frac{Tc}{\alpha_2} \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_1}\right) & \text{for } d_1 \leq s \leq T - d_2 \\ K \left(1 - \left(1 - \frac{d_1}{T}\right)^{\alpha_1}\right) + \frac{Tc}{\alpha_2} \left(1 - \frac{d_2}{T}\right)^{\alpha_1} + \frac{Tc}{\alpha_3} \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_3}\right) & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (45)$$

where $K = \frac{Tc}{\alpha_1} \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1}$.

The density function corresponding to this process is given by

$$f_3(t) = \begin{cases} C \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} \left(1 - \frac{t}{T}\right)^{\alpha_1 - 1} & \text{for } 0 \leq t \leq d_1 \\ C \left(1 - \frac{t}{T}\right)^{\alpha_2 - 1} & \text{for } d_1 \leq t \leq T - d_2 \\ C \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{t}{T}\right)^{\alpha_3 - 1} & \text{for } T - d_2 \leq t \leq T \end{cases} \quad (46)$$

where

$$C = c/m(T) = \frac{\alpha_1 \alpha_2 \alpha_3 / T}{\left(1 - \frac{d_1}{T}\right)^{\alpha_2} \alpha_3 (\alpha_1 - \alpha_2) + \alpha_3 \alpha_2 \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} + \left(\frac{d_2}{T}\right)^{\alpha_2} \alpha_1 (\alpha_2 - \alpha_3)}.$$

The CDF is given by

$$F_3(t) = \begin{cases} \frac{CT}{\alpha_1} \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} \left[1 - \left(1 - \frac{t}{T}\right)^{\alpha_1}\right] & \text{for } 0 \leq t \leq d_1 \\ \frac{CT}{\alpha_1 \alpha_2} \left[(\alpha_1 - \alpha_2) \left(1 - \frac{d_1}{T}\right)^{\alpha_2} + \alpha_2 \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} - \alpha_1 \left(1 - \frac{t}{T}\right)^{\alpha_2} \right] & \text{for } d_1 \leq t \leq T - d_2 \\ 1 - \frac{CT}{\alpha_3} \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{t}{T}\right)^{\alpha_3} & \text{for } T - d_2 \leq t \leq T \end{cases} \quad (47)$$

Note that for the interval $d_1 \leq t \leq T - d_2$ we can write the CDF as

$$F_3(t) = F_3(d_1) + \frac{CT}{\alpha_2} \left[\left(1 - \frac{d_1}{T}\right)^{\alpha_2} - \left(1 - \frac{t}{T}\right)^{\alpha_2} \right] \quad (48)$$

4.1. Simulation of NHPP₃

In order to simulate a three-stage NHPP on the interval $[0, T]$ we use the inversion method and follow the same logic as for NHPP₂. The inverse CDF can be written as:

$$F_3^{-1}(s) = \begin{cases} T - T \left\{ 1 - \frac{s \alpha_1}{CT} \left(1 - \frac{d_1}{T}\right)^{\alpha_1 - \alpha_2} \right\}^{1/\alpha_1} & \text{for } 0 \leq s \leq d_1 \\ T - T \left\{ \left(1 - \frac{d_1}{T}\right)^{\alpha_2} - \frac{\alpha_2}{CT} (s - F_3(d_1)) \right\}^{1/\alpha_2} & \text{for } d_1 \leq s \leq T - d_2 \\ T - T \left\{ \frac{\alpha_3}{CT} (1 - s) \left(\frac{d_2}{T}\right)^{\alpha_3 - \alpha_2} \right\}^{1/\alpha_3} & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (49)$$

The algorithm for generating n arrivals is then:

- (1) Generate n uniform variates u_1, \dots, u_n .
- (2) For $k = 1, \dots, n$ set

$$x_k = \begin{cases} T - T \left\{ 1 - \frac{u_k \alpha_1}{CT} \left(1 - \frac{d_1}{T} \right)^{\alpha_1 - \alpha_2} \right\}^{1/\alpha_1} & \text{if } u_k < F_3(d_1) \\ T - T \left\{ \frac{\alpha_2}{CT} (F_3(d_1) - u_k) + \left(1 - \frac{d_1}{T} \right)^{\alpha_2} \right\}^{1/\alpha_2} & \text{if } F_3(d_1) \leq u_k < F_3(T - d_2) \\ T - T \left\{ \frac{\alpha_3}{CT} u_k \left(\frac{d_2}{T} \right)^{\alpha_3 - \alpha_2} \right\}^{1/\alpha_3} & \text{if } u_k \geq F_3(T - d_2) \end{cases} \quad (50)$$

4.2. Estimation of NHPP₃ Parameters

4.2.1. Quick & Crude (CDF-Based) Estimation

Estimation of α Parameters

The quick & crude method described for estimating the α parameters in NHPP₂ works also for the NHPP₃. In each interval the mean of the Poisson process is in the form $m_3(y) = \beta_j - \theta_j \left(1 - \frac{y}{T} \right)^{\alpha_j}$, ($j = 1, 2, 3$), and therefore the same approximation works on each of the three intervals $[0, d_1]$, $[d_1, T - d_2]$ and $[T - d_2, T]$. The idea, once again, is to pick intervals $[T - t, T - s]$ that we are confident lie in the first, second, or third phases. Then, based on the bid times in each interval, the relevant α is

$$\alpha = 2 \frac{\log[F(T - t) - F(T - \sqrt{st})] - \log[F(T - \sqrt{st}) - F(T - s)]}{\log t - \log s} \quad (51)$$

and is estimated by plugging the empirical CDF F_e for F in the approximation. For α_3 we can use the exact relation

$$\alpha_3 = \frac{\log R(t_3)/R(t'_3)}{\log(T - t_3)/(T - t'_3)} \quad (52)$$

where $R(t) = 1 - F_3(t)$ and t_3, t'_3 are within $[T - d_2, T]$. To estimate α_3 we pick reasonable values of t_3, t'_3 and use the empirical survival function R_e .

To assess this method we simulated 5000 random observations from an NHPP₃ with parameters $\alpha_1 = 3, \alpha_2 = 0.4, \alpha_3 = 1$ and the changepoints $d_1 = 2.5$ (defining the first 2.5 days as the first phase) and $d_2 = 5/10080$ (defining the last 5 minutes as the third phase). We computed the quick & crude estimate for α_1 on a range of intervals of the form $[0.001, t_1]$ where $0.5 \leq t_1 \leq 5$. Notice that this interval includes values that are outside the range $[0, d_1 = 2.5]$. The left panel in Figure 10 illustrates the estimates obtained for these intervals. For values of t_1 between 1.5-3.5 days, the estimate for α_1 is relatively stable and close to 3. Similarly, the right panel in Figure 10 describes the estimates of α_3 , using (52), as a function of the choice of t_3 with $t'_3 = 7 - 1/10080$. The estimate is relatively stable and close to 1.

For estimating α_2 an interval such as $[3, 6.9]$ is reasonable. Figure 11 shows the estimate as a function of the interval choice. It is clear that the estimate is relatively insensitive to the exact interval choice, as long as it is reasonable.

Estimation of d_1 and d_2

Using functions of the CDF we can obtain expressions for d_1 and d_2 . Let t_1, t_2, t'_2 , and t_3 be such that $0 \leq t_1 \leq d_1, d_1 \leq t'_2 < t_2 \leq T - d_2$, and $T - d_2 \leq t_3 \leq T$. For d_1 we use the ratio

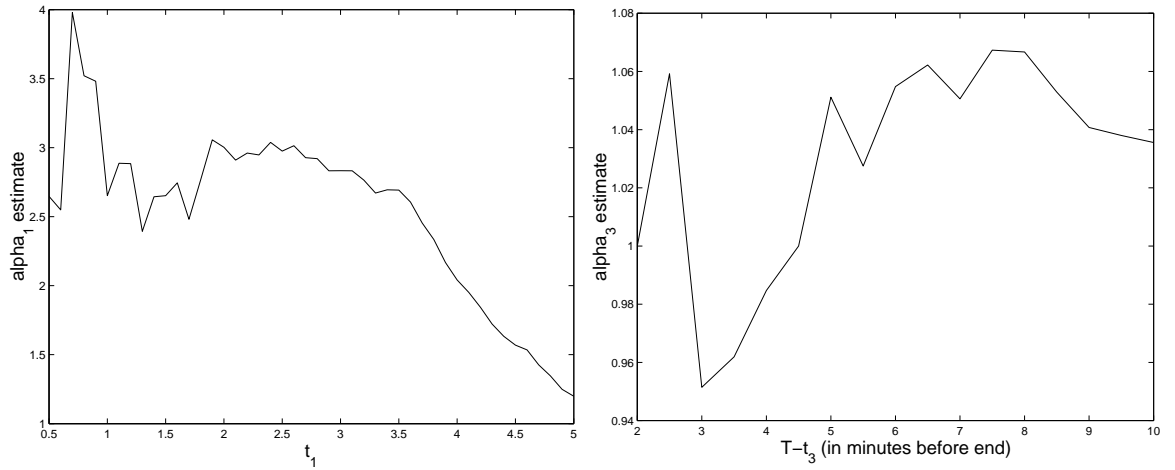


Fig. 10. Quick estimates of α_1, α_2 , and α_3 as a function of the input intervals, for simulated NHPP₃ data.

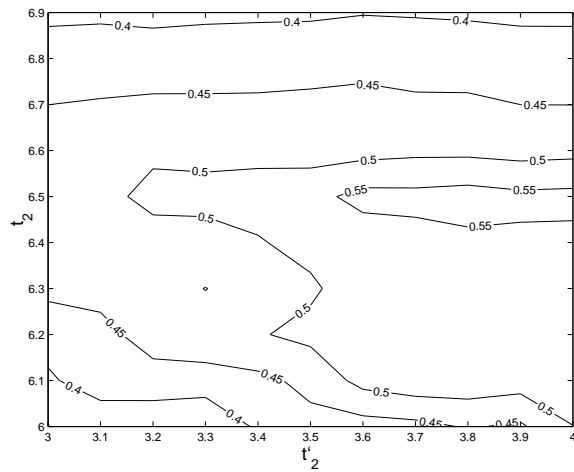


Fig. 11. Quick & crude estimate of α_2 as a function of $[t'_2, t_2]$ choice. $\hat{\alpha}_2$ is between 0.4-0.55 in the entire range of intervals. The more extreme intervals ($t'_2 < 3.4$ or $t_2 > 6.8$) yield $\hat{\alpha}_2 = 0.4$.

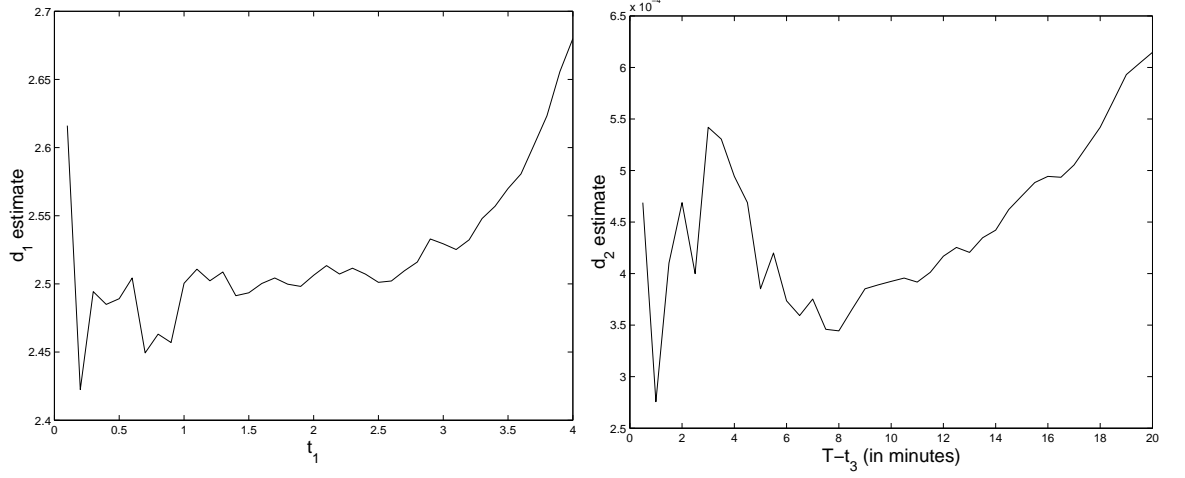


Fig. 12. Graphs of \hat{d}_1 vs. t_1 (left) and \hat{d}_2 vs. initial values of $T - t_3$ (right) for simulated data. The estimate for d_1 is stable at ≈ 2.5 . \hat{d}_2 using the last 2-5 minute interval is in the range of 4-5 minutes.

$\frac{F_3(t_2) - F_3(t'_2)}{F_3(t_1)}$ and for d_2 we use the ratio $\frac{F_3(t_2) - F_3(t'_2)}{1 - F_3(t_3)}$. These lead to the following expressions:

$$d_1 = T - T \left\{ \frac{\alpha_1}{\alpha_2} \cdot \frac{F_3(t_1)}{F_3(t_2) - F_3(t'_2)} \cdot \frac{(1 - t'_2/T)^{\alpha_2} - (1 - t_2/T)^{\alpha_2}}{1 - (1 - t_1/T)^{\alpha_1}} \right\}^{\frac{1}{\alpha_2 - \alpha_1}} \quad (53)$$

$$d_2 = T \left\{ \frac{\alpha_3}{\alpha_2} \cdot \frac{1 - F_3(t_3)}{F_3(t_2) - F_3(t'_2)} \cdot \frac{(1 - t'_2/T)^{\alpha_2} - (1 - t_2/T)^{\alpha_2}}{(1 - t_3/T)^{\alpha_3}} \right\}^{\frac{1}{\alpha_2 - \alpha_3}} \quad (54)$$

Thus we can estimate d_1 and d_2 by selecting “safe” values for t_1, t'_2, t_2 , and t_3 (which are confidently within the relevant interval) and using the empirical CDF at those points. Using this method we estimated d_1 and d_2 for the simulated data. We used the true values of the α parameters in (53). Using $t_1 = 1, t'_2 = 3, t_2 = 6$, and $t_3 = 7 - 2/10080$ yields $\hat{d}_1 = 2.5$ and $\hat{d}_2 = 4.73$ minutes.

4.2.2. Maximum Likelihood Estimation

Conditional on $N(T) = n$ (see Appendix D), the NHPP₃ likelihood function is given by

$$\begin{aligned} \mathcal{L}(x_1, \dots, x_n | \alpha_1, \alpha_2, \alpha_3, d_1, d_2) = & \quad (55) \\ n \log C + n_1(\alpha_2 - \alpha_1) \log \left(1 - \frac{d_1}{T} \right) + n_3(\alpha_2 - \alpha_3) \log \frac{d_2}{T} + (\alpha_1 - 1)S_1 + (\alpha_2 - 1)S_2 + (\alpha_3 - 1)S_3, \end{aligned}$$

where n_1 is the number of arrivals before time d_1 , n_3 is the number of arrivals after $T - d_2$, $S_1 = \sum_{i: x_i \leq d_1} \log \left(1 - \frac{x_i}{T} \right)$, $S_2 = \sum_{i: d_1 < x_i < T - d_2} \log \left(1 - \frac{x_i}{T} \right)$, and $S_3 = \sum_{i: x_i > T - d_2} \log \left(1 - \frac{x_i}{T} \right)$. In order to estimate $\alpha_1, \alpha_2, \alpha_3$ for given values of d_1, d_2 , the following three equations must be

solved (equating the first derivatives in $\alpha_1, \alpha_2, \alpha_3$ to zero).

$$S_1 = n_1 \log\left(1 - \frac{d_1}{T}\right) - \frac{n}{C} \frac{\partial C}{\partial \alpha_1} \quad (56)$$

$$S_2 = -n_1 \log\left(1 - \frac{d_1}{T}\right) - n_3 \log \frac{d_2}{T} - \frac{n}{C} \frac{\partial C}{\partial \alpha_2} \quad (57)$$

$$S_3 = n_3 \log \frac{d_2}{T} - \frac{n}{C} \frac{\partial C}{\partial \alpha_3} \quad (58)$$

where

$$\frac{\partial C}{\partial \alpha_1} = \frac{C^2 T}{\alpha_1^2} \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \left[\left(1 - \frac{d_1}{T}\right)^{-\alpha_1} \left(1 + \alpha_1 \log\left(1 - \frac{d_1}{T}\right)\right) - 1 \right] \quad (59)$$

$$\begin{aligned} \frac{\partial C}{\partial \alpha_2} &= \frac{C^2 T}{\alpha_1 \alpha_3 \alpha_2^2} \left\{ \alpha_3 \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \left[\alpha_2 \log\left(1 - \frac{d_1}{T}\right) \left(\alpha_2 - \alpha_1 + \alpha_2 \left(1 - \frac{d_1}{T}\right)^{-\alpha_1} \right) - \alpha_1 \right] + \right. \\ &\quad \left. + \alpha_1 \left(\frac{d_2}{T}\right)^{\alpha_2} \left[\alpha_3 + \alpha_2 \log \frac{d_2}{T} (\alpha_2 - \alpha_3) \right] \right\} = \quad (60) \end{aligned}$$

$$\begin{aligned} &= \frac{C^2 T}{\alpha_2^2} \left[\left(\frac{d_2}{T}\right)^{\alpha_2} - \left(1 - \frac{d_1}{T}\right)^{\alpha_2} - \frac{\alpha_2^2}{\alpha_1} \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \log\left(1 - \frac{d_1}{T}\right) \left(1 - \left(1 - \frac{d_1}{T}\right)^{-\alpha_1}\right) - \right. \\ &\quad \left. - \alpha_2 \left(\frac{d_2}{T}\right)^{\alpha_2} \log \frac{d_2}{T} + \alpha_2 \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \log\left(1 - \frac{d_1}{T}\right) + \frac{\alpha_2^2}{\alpha_3} \left(\frac{d_2}{T}\right)^{\alpha_2} \log \frac{d_2}{T} \right] \end{aligned}$$

$$\frac{\partial C}{\partial \alpha_3} = \frac{C^2 T}{\alpha_3^2} \left(\frac{d_2}{T}\right)^{\alpha_2} \quad (61)$$

Since the equations are non-linear in the parameters an iterative gradient method can be used (the second derivatives are given in Appendix E). If d_1 and d_2 are unknown and we want to estimate them from the data, then search algorithms such as genetic algorithms can be more efficient, more stable, and more easily programmable for finding a solution. Otherwise the likelihood needs to be computed for a grid of $d_1 \times d_2$ values. In addition, empirical evidence suggests that gradient methods tend to be unstable for solving this maximization problem. Therefore an exhaustive search over a reasonable grid of the parameter space or a stochastic search algorithm are good practical solutions.

Genetic Algorithm Search

An alternative to an exhaustive search in 5 dimensions or a computationally extensive hybrid of grid search for d_1 and d_2 combined with a numerical maximization procedure for estimating $\alpha_1, \alpha_2, \alpha_3$ is to use a stochastic search algorithm such as the genetic algorithm. Genetic algorithms are based on optimization strategies that are successfully being used by nature - known as Darwinian Evolution - and they utilize these strategies for application in mathematical optimization theory. Genetic algorithms have become popular for finding the global optima among a set of local optima but they are also very useful alternatives in situations where gradient methods struggle (e.g. if the derivatives are hard to come by).

We used a genetic algorithm for finding the values of $d_1, d_2, \alpha_1, \alpha_2, \alpha_3$ that maximize the likelihood function. We restricted the range of possible solutions to the hypercube $(\alpha_1, \alpha_2, \alpha_3, d_1, d_2) \in [0, 10] \times [0, 1] \times [0, 5] \times [0, 5] \times [0, 0.1]$. This yielded the estimates $\hat{\alpha}_1 = 3.06, \hat{\alpha}_2 = 0.39, \hat{\alpha}_3 = 1.01, \hat{d}_1 = 2.51, \hat{d}_2 = 4.68/10080$ for the simulated data. All of these estimates are completely in line with the quick & crude estimates, and very close to the values that were used to generate the data. The run time for this procedure was only a few minutes. The combined numerical maximization and grid search procedure did not converge.

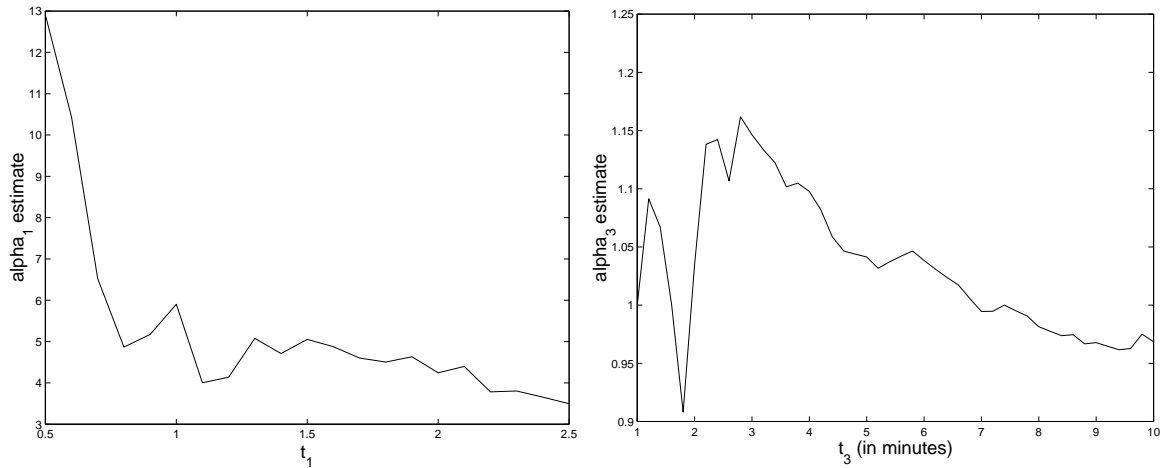


Fig. 13. Quick & crude estimates of α_1 as a function of t_1 (with $t'_1 = 0.001$) (left) and of α_3 as a function of t_3 (with $t'_3 = 0.5/10080$) (right). $\hat{\alpha}_1$ is stable around 5 for t_1 in the range 0.75-1.75 days. A shorter interval does not contain enough data. A longer interval leads to a drop in the estimate, indicating that $d_1 < 2$. $\hat{\alpha}_3$ is around 1.1 when t_3 is within the last 2-4 minutes.

4.3. Empirical Results

We use the quick & crude method to estimate the parameters for the 3651 Palm bid arrival times. From domain knowledge, we chose the first day for estimating α_1 , i.e., we believe that bids placed during the first day are contained within the first “early bidding” phase. Looking at the estimate as a function of the interval chosen (Figure 13, left panel), we see that the estimate is between 4-5 if we use the first 1-2 days. It is interesting to note that after the first two days, the estimate decreases progressively reaching $\hat{\alpha}_1 = 2.5$ on the interval $[0.01, 3]$, indicating that the changepoint d_1 is around 2.

The parameter α_3 was estimated using (52) with $t'_3 = 7 - 0.1/10080$ and a range of values for t_3 . From these, α_3 appears to be approximately 1. It can be seen in the right panel of Figure 13 that this estimate is relatively stable within the last 10 minutes. Also, notice that selecting t_3 too close to t'_3 results in unreliable estimates (due to a small number of observations between the two values).

Finally, we chose the interval $[3, 6.9]$ for estimating α_2 . This yielded the estimate $\hat{\alpha}_2 = 0.36$. Figure 14 shows the estimate as a function of the interval choice. Note that the estimate is stable between 0.2-0.4 for the different intervals chosen. It is more sensitive to the choice of t_2 , the upper bound of the interval, and thus an overly conservative interval could yield to large inaccuracies.

Using these estimates ($\hat{\alpha}_1 = 5, \hat{\alpha}_2 = 0.36, \hat{\alpha}_3 = 1.1$), we estimated d_1 and d_2 . Figure 15 shows graphs of the estimates as a function of the intervals selected. The estimate for d_1 (left panel) appears to be stable at approximately $\hat{d}_1 = 1.75$. The estimate for d_2 (right panel) appears to be around 2 minutes. From the increasing values obtained for $T - t_3 > 3$ minutes we also learn that $d_2 < 3$. Table 1 displays the above estimates and compares them to the two other estimation methods: An exhaustive search over a reasonable range of the parameter space (around the quick & crude estimates), and the much quicker genetic algorithm. It can be seen that all methods yielded estimates in the same vicinity.

Finally, to further validate this estimated model, we simulated data from an NHPP₃ with the above ML estimates as parameters. Figure 16 shows a QQ-plot of the Palm data vs. the simulated data. The points appear to fall on the line $x = y$, thus supporting the adequacy of

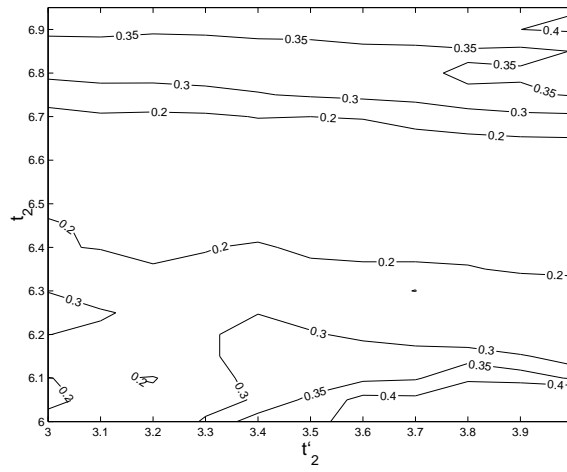


Fig. 14. Quick & crude estimate of α_2 as a function of $[t'_2, t_2]$. Shorter, “safer” intervals are at the lower right. Longer intervals, containing more data, are at the upper left. $\hat{\alpha}_2$ is between 0.2-0.4 for all intervals. For $t_2 > 6.9$ the estimate is approximately 0.35.

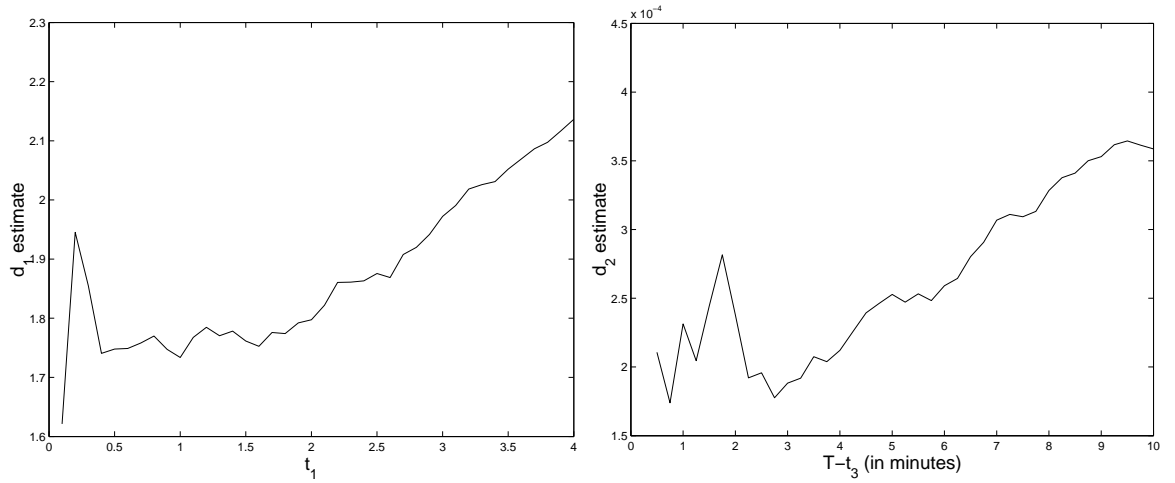


Fig. 15. Plots of \hat{d}_1 vs. t_1 (left) and \hat{d}_2 vs. initial values of $T - t_3$ (right) for Palm data. The estimate for d_1 seems stable at ≈ 1.75 . \hat{d}_2 is approximately 2 minutes.

Table 1. Estimates for five NHPP₃ parameters using the three estimation methods.

| | $\hat{\alpha}_1$ | $\hat{\alpha}_2$ | $\hat{\alpha}_3$ | \hat{d}_1 | \hat{d}_2 (minutes) |
|-------------------|------------------|------------------|------------------|-------------|-----------------------|
| CDF-based Q&C | 5 | 0.36 | 1.1 | 1.75 | 2 |
| Exhaustive search | 4.9 | 0.37 | 1.13 | 1.7 | 2 |
| Genetic Algorithm | 5.56 | 0.37 | 1.1 | 1.54 | 2.11 |

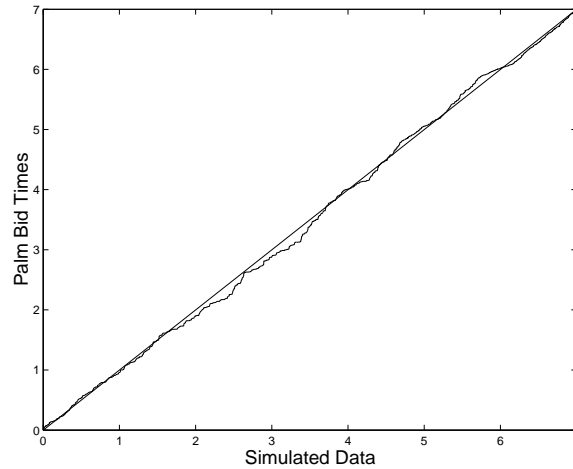


Fig. 16. Q-Q plot of Palm bid times vs. simulated data from an NHPP_3 with parameters $\alpha_1 = 4.9, \alpha_2 = 0.37, \alpha_3 = 1.13, d_1 = 1.7, d_2 = 2/10080$.

the estimated model for the Palm bid times.

The estimated model for the Palm data reveals the dynamics of these auctions over time: The “average” auction has three phases: the beginning takes place during the first 1.7 days, the middle continues until the last 2 minutes, and then the third phase kicks in. The bid arrivals in each of the three phases can be described by an NHPP process, but they each have different intensity functions. The auction beginning is characterized by an early surge of interest, with more intense bidding than during the start of the second phase. Then, the increase in bid arrival rate slows down during the middle of the auction. The bids do tend to arrive faster as the auction progresses, but at the very end, during the last 2 minutes of the auction we observe a uniform bid arrival process. Finally, it is interesting to note that in these data the third phase of bidding seems to take place within the last 2-3 minutes rather than the last 1 minute. Thus, we use the term “last-moment bidding” rather than “last-minute bidding”.

5. Discussion

The NHPP formulation that we suggest here is motivated by the need to model bid arrivals in online auction data. It is, nonetheless, very general and can be used for fitting data in other applications. The flexibility of the model is derived from the continuous intensity function, the large range of values that the α parameters can take, and by the ability to add more phases. The ability to model the bid arrival process has many important uses: First, it enables the exploration and formalization of observed phenomena like early and late bidding. Furthermore, such phenomena can then be explained as a function of bidder behavior. For example, it can be shown that certain bidder behavior dynamics, where each displayed bid is the minimum of a collection of (uniformly distributed) bid times contemplated by a shrinking population of bidders, leads to NHPP_1 (Russo & Shmueli, unpublished research). Thus, if a set of auctions is shown to follow a NHPP_1 model, it is possible that the bidder behavior occurring behind the scenes is of this type. Another bidder behavior of interest is collusion, where a buyer is actually an agent of the seller who participates in the auction in order to “run up the bid”. Kauffman & Wood (2000) hypothesize that colluders avoid bidding towards the end of the auction. In a sample of auctions infected by collusion we would therefore expect to see less activity than

usual towards the end of the auction.

A second use of a bid arrival model is in conjunction with the sequence of bid or price increments. Jank & Shmueli (2003) explore the bidding dynamics in online auctions by fitting smoothing splines to sets of bid times and bid amounts corresponding to an auction. The knots of the splines are determined by the bidding intensity, or the intensity of the bid arrivals. An NHPP₃ model can be used in this application, to determine favorable locations of the knots. One of the most researched questions is what factors affect the final price obtained in an auction. Several authors have shown that the final price is higher in auctions with more activity. An open question is whether there is a function relating a particular NHPP₃ model with an average final price.

Knowledge of the bid arrival process is of special importance for applications which determine the frequency of page updating. For example, if eBay users are monitoring an auction from a PDA which has costs attached to web connection, they must decide on a policy when to reconnect and update the information (Gal & Eckstein, 2001; Bright et al., 2004). In an auction that has the typical early and last moment phases of bidding, it is better for the user to update the information more frequently during these phases and not connect as much during the middle phase.

Finally, the bid arrival model can be useful for visualization tools that display the bids over an auction or a set of auctions. In order to determine the scale of the time axis and avoid over- and under-crowded areas on the display, the application must know "where the action is" and to what degree. An NHPP model, even if approximate, gives a sense of the scale of interest.

A. Proof of uniform convergence in probability of $\pi(t, \theta)$ as $c \rightarrow \infty$.

Let F be any strictly increasing and continuous *cdf* with $F(0) = 0$, and $F(T) = 1$ for some $T > 0$. Suppose that for some fixed $\delta \in (0, 1)$ and some increasing function $\phi : [0, 1] \rightarrow [0, 1]$,

$$\frac{1 - F(T - t\theta)}{1 - F(T - t)} = \phi(\theta) \quad \text{for } 0 \leq \theta \leq 1 \text{ and } \delta T \leq t \leq T, \quad (62)$$

We observe that (62) requires $\phi(0) = 0$ and $\phi(1) = 1$. Let U_1, U_2, \dots be a sequence of independent $U(0, 1)$ random variables, and G_n the empirical *cdf* corresponding to the first n of them. We construct an *i.i.d.*(F) sequence from the U_i 's as follows:

$$Y_1 = F^{-1}(1 - U_1), Y_2 = F^{-1}(1 - U_2), \dots \quad (63)$$

Let F_n denote the empirical *cdf* corresponding to the first n of the Y_i 's in (63), and define

$$\pi_n(t, \theta) = \frac{1 - F_n(T - \theta t)}{1 - F_n(T - t)} \quad \text{for } 0 \leq \theta \leq 1 \text{ and } \delta T \leq t \leq T \quad (64)$$

For convenience, set $\pi_0(t, \theta) = 0$. For $y \in [0, T]$ and $1 \leq k \leq n$,

$$Y_k \geq y \iff U_k \leq 1 - F(y) \quad (65)$$

so that $1 - F_n(y) = G_n(1 - F(y))$ and hence by (62),

$$\pi_n(t, \theta) = \frac{G_n(1 - F(T - \theta t))}{G_n(1 - F(T - t))} = \frac{G_n(\phi(\theta)(1 - F(T - t)))}{G_n(1 - F(T - t))} \quad (66)$$

From (66) it follows that

$$\sup_{\delta T \leq t \leq T, 0 \leq \theta \leq 1} |\pi_n(t, \theta) - \phi(\theta)| = \sup_{\delta T \leq t \leq T, 0 \leq \theta \leq 1} \left| \frac{G_n(\phi(\theta)(1 - F(T - t)))}{G_n(1 - F(T - t))} - \phi(\theta) \right| \quad (67)$$

As t varies over $[\delta T, T]$, we have $1 - F(T - t)$ varying over $[1 - F(T - \delta T), 1]$, which by (62) is the same as $[\phi(\delta), 1]$. Also, as θ varies over $[0, 1]$, we have $\phi(\theta)$ doing the same. By (67),

$$\pi_n := \sup_{\delta T \leq t \leq T, 0 \leq \theta \leq 1} |\pi_n(t, \theta) - \phi(\theta)| = \sup_{\phi(\delta) \leq \lambda \leq 1, 0 \leq \lambda \leq 1} \left| \frac{G_n(\lambda t)}{G_n(t)} - \lambda \right| \quad (68)$$

The Glivenko-Cantelli Theorem (Resnick (1998), p. 224, for example) applied to the U_i sequence says that

$$\Delta_n := \sup_{0 \leq t \leq 1} |G_n(t) - t| \rightarrow 0 \quad \text{almost surely} \quad (69)$$

For $t \in [\phi(\delta), 1]$ and $\lambda \in [0, 1]$ we have

$$\frac{G_n(\lambda t)}{G_n(t)} - \lambda \in \left[\frac{t\lambda - \Delta_n}{t + \Delta_n} - \lambda, \frac{t\lambda + \Delta_n}{t - \Delta_n} - \lambda \right] \subseteq \left[\frac{-2\Delta_n}{\phi(\delta) - \Delta_n}, \frac{2\Delta_n}{\phi(\delta) - \Delta_n} \right] \quad (70)$$

so that

$$\left| \frac{G_n(\lambda t)}{G_n(t)} - \lambda \right| \leq \frac{2\Delta_n}{\phi(\delta) - \Delta_n} \quad (71)$$

By (68), (69) and (71),

$$\pi_n \rightarrow 0 \quad \text{almost surely (and thus also in probability)} \quad (72)$$

Fix arbitrary $\varepsilon > 0$. By (72), there exists an $m_o \geq 1$ for which

$$P(\pi_n > \varepsilon) < \varepsilon \quad \text{for all } n \geq m_o \quad (73)$$

Let $N_c(t), 0 \leq t \leq T$, denote the NHHP having intensity function given in (1), let F be as in (4) and let $\pi(t, \theta)$ be as in (15). Conditional on the event that $N_c(T) = n$, the n arrival times (in randomized order) on the interval $[0, T]$ are distributed as a random sample of size n from the distribution F . Thus, with $\pi(t, \theta)$ as in (15) (note: we should define $\pi(t, \theta) = 0$ in (15) when $N(T) - N(T - t) = 0$), we have

$$\begin{aligned} P\left(\sup_{\delta T \leq t \leq T, 0 \leq \theta \leq 1} |\pi(t, \theta) - \phi(\theta)| > \varepsilon\right) &= \sum_{n \geq 0} P(\pi_n > \varepsilon) P(N_c(t) = n) \\ &\leq P(N_c(t) < m_o) + \sum_{n \geq m_o} P(\pi_n > \varepsilon) P(N_c(t) = n) \\ &\leq P(N_c(t) < m_o) + \varepsilon P(N_c(t) \geq m_o) \end{aligned} \quad (74)$$

Since ε is arbitrary and $P(N_c(t) < m_o) \rightarrow 0$ as $c \rightarrow \infty$, we conclude from (74) that

$$\sup_{\delta T \leq t \leq T, 0 \leq \theta \leq 1} |\pi(t, \theta) - \phi(\theta)| \xrightarrow{P} 0 \quad \text{as } c \rightarrow \infty \quad (75)$$

We observe that statement (68) also yields an invariance result: Fix $u \in (0, 1)$ and take $\phi(\theta) = \theta^\alpha$ and $\delta = \phi^{-1}(u)$ in (68) to obtain

$$\sup_{\phi^{-1}(u)T \leq t \leq T, 0 \leq \theta \leq 1} |\pi(t, \theta) - \phi(\theta)| \stackrel{D}{=} \sup_{u \leq t \leq 1, 0 \leq \lambda \leq 1} \left| \frac{G_{N(T)}(\lambda t)}{G_{N(T)}(t)} - \lambda \right| \quad (76)$$

where $N(T) \sim \text{Poisson}(m(T))$. Thus, the distribution of *LHS*(76) is invariant with respect to the collection of NHPP's having intensity functions of the form given in (1).

B. $\pi(\theta, t)$ for fixed t

Conditionally on $Y(t) = N(T) - N(T - t) = k$, the random variable $N(T) - N(T - \theta t)$ has a $\text{bin}(k, \theta^\alpha)$ distribution, so $\pi(t, \theta)$ is just a sample proportion. Since

$$Y(t) \sim \text{Poisson}(m(T) - m(T - t)) = \text{Poisson}\left(\frac{ct^\alpha T^{1-\alpha}}{\alpha}\right) \quad (77)$$

we have

$$\frac{\alpha Y(t)}{ct^\alpha T^{1-\alpha}} \xrightarrow{P} 1 \quad (78)$$

For $c > 0$, define the index set

$$S_c = \left\{k : \left|k - \frac{ct^\alpha T^{1-\alpha}}{\alpha}\right| \leq c^{2/3}\right\} \quad (79)$$

As $c \rightarrow \infty$,

$$\frac{k}{c} \approx \frac{t^\alpha T^{1-\alpha}}{\alpha} \quad \text{for all } k \in S_c \quad (80)$$

and by Chebychev,

$$P\{Y(t) \in S_c\} \geq 1 - \frac{t^\alpha T^{1-\alpha}}{\alpha c^{1/3}} \quad (81)$$

Thus, as $c \rightarrow \infty$, with $p_k = P(Y(t) = k)$,

$$\begin{aligned}
 P\left(|\pi(t, \theta) - \theta^\alpha| > \frac{x}{\sqrt{c}}\right) &= \sum_{k=0}^{\infty} \left(|\pi(t, \theta) - \theta^\alpha| > \frac{x}{\sqrt{c}} \mid Y(t) = k \right) p_k \quad (82) \\
 &\approx \sum_{k \in S_c} P\left(\left| \frac{\text{bin}(k, \theta^\alpha) - k\theta^\alpha}{\sqrt{k\theta^\alpha(1-\theta^\alpha)}} \right| > x \sqrt{\frac{k}{c\theta^\alpha(1-\theta^\alpha)}}\right) p_k, \text{ by (81)} \\
 &\approx \sum_{k \in S_c} P\left(\left| \frac{\text{bin}(k, \theta^\alpha) - k\theta^\alpha}{\sqrt{k\theta^\alpha(1-\theta^\alpha)}} \right| > x \sqrt{\frac{t^\alpha T^{1-\alpha}}{\alpha\theta^\alpha(1-\theta^\alpha)}}\right) p_k, \text{ by (80)} \\
 &\approx \sum_{k \in S_c} P\left(|Z| > x \sqrt{\frac{t^\alpha T^{1-\alpha}}{\alpha\theta^\alpha(1-\theta^\alpha)}}\right) p_k \text{ by the CLT} \\
 &= P\left(|Z| > x \sqrt{\frac{t^\alpha T^{1-\alpha}}{\alpha\theta^\alpha(1-\theta^\alpha)}}\right) P(Y(t) \in S_c) \\
 &\rightarrow P\left(|Z| > x \sqrt{\frac{t^\alpha T^{1-\alpha}}{\alpha\theta^\alpha(1-\theta^\alpha)}}\right), \text{ by (81)}
 \end{aligned}$$

C. Asymptotic distribution of MLE $\hat{\alpha}$ in NHPP₁

We show that the asymptotic distribution (as $n \rightarrow \infty$) of the MLE $\hat{\alpha}$ in (19) is

$$\sqrt{n} \left(\frac{\alpha}{\hat{\alpha}} - 1 \right) \rightarrow n(0, 1)$$

Suppose X has the distribution in (4). For $s > 0$,

$$P\left(-\log\left(1 - \frac{X}{T}\right) > s\right) = P(X > T(1 - e^{-s})) = e^{-s\alpha} \quad (83)$$

Thus,

$$-\log\left(1 - \frac{X}{T}\right) \sim \text{exponential with mean} = \frac{1}{\alpha} \text{ and variance} = \frac{1}{\alpha^2} \quad (84)$$

so that

$$Y_1 = -\alpha \log\left(1 - \frac{X_1}{T}\right) - 1, Y_2 = -\alpha \log\left(1 - \frac{X_2}{T}\right) - 1, \dots$$

is an *i.i.d.* sequence of random variables with *mean* = 0, and *variance* = 1. By the *CLT* for *i.i.d.* random variables,

$$\begin{aligned}
 \sqrt{n} \left(\frac{\alpha}{\hat{\alpha}} - 1 \right) &= \sqrt{n} \left(\frac{-\alpha \sum_{i=1}^n \log\left(1 - \frac{X_i}{T}\right) - 1}{n} \right) \\
 &= \frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \\
 &\rightarrow n(0, 1)
 \end{aligned}$$

If $\hat{\alpha}$ is defined as in (21), then since $P(N(T) \rightarrow \infty) = 1$ as $c \rightarrow \infty$, we have

$$\sqrt{N(T)} \left(\frac{\alpha}{\hat{\alpha}} - 1 \right) \rightarrow n(0, 1)$$

D. ML Estimation of the Unconditional Models NHPP_{1,2,3}

Let $N(s), 0 \leq s \leq T$, be a NHPP with an intensity function of the form

$$\lambda(s) = cg(\theta, s), \quad 0 \leq s \leq T$$

where c and $\theta = (\theta_1, \dots, \theta_k)$ are unknown parameters. Define $h(\theta) = \int_0^T g(\theta, s) ds$, so that $m(T) = ch(\theta)$. The pdf associated with λ is $f(\theta, s) = \lambda(s)/m(T)$, $0 \leq s \leq T$. Given a random sample x_1, \dots, x_n (non-random n) from this distribution, the likelihood and log-likelihood functions of θ are

$$L(\theta) = \prod_{i=1}^n f(\theta, x_i) \quad \text{and} \quad \mathcal{L}(\theta) = \log L(\theta)$$

On the other hand, given the value n of $N(T)$, and the arrival times x_1, \dots, x_n from the NHPP, the likelihood function of (c, θ) is given by

$$L(c, \theta) = \frac{e^{-m(T)} m(T)^n}{n!} \prod_{i=1}^n f(\theta, x_i) = \frac{e^{-ch(\theta)} (ch(\theta))^n}{n!} L(\theta)$$

The log-likelihood is thus

$$\mathcal{L}(c, \theta) = -ch(\theta) + n \log c + n \log h(\theta) - \log n! + \mathcal{L}(\theta)$$

The joint MLE of c and θ is the solution of the equations

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}(c, \theta)}{\partial c} = -h(\theta) + \frac{n}{c} \\ 0 &= \frac{\partial \mathcal{L}(c, \theta)}{\partial \theta_j} = -c \frac{\partial h(\theta)}{\partial \theta_j} + \frac{n}{h(\theta)} \frac{\partial h(\theta)}{\partial \theta_j} + \frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} \quad 1 \leq j \leq k \end{aligned} \quad (85)$$

Solving the first equation in (85) for c and plugging into the second we find that

$$\frac{\partial \mathcal{L}(c, \theta)}{\partial \theta_j} = \frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} \quad 1 \leq j \leq k$$

Hence, $L(c, \theta)$ and $L(\theta)$ yield the same MLE for θ . That is, if $\hat{\theta}_j = w_j(X_1, \dots, X_n)$ is the MLE of θ_j ($1 \leq j \leq k$) based on a random sample of non-random size n from the distribution with the pdf above, then the MLE of θ_j based on the arrival times $X_1, \dots, X_{N(T)}$ from the above NHPP is of the form: $\hat{\theta}_j = w_j(X_1, \dots, X_{N(T)})$. By the first equation in (85), the MLE of c is

$$\hat{c} = \frac{N(T)}{h(\hat{\theta})} \quad (86)$$

In NHPP₁, $\theta = \alpha$ and $h(\alpha) = T/\alpha$, so that by (19) and (86),

$$\hat{\alpha} = -N(T) \left[\sum_{i=1}^{N(T)} \log \left(1 - \frac{X_i}{T} \right) \right]^{-1} \quad \text{and} \quad \hat{c} = \frac{N(T)}{T} \hat{\alpha}.$$

E. Second derivatives of the log-likelihood function for the 3-stage NHPP

The second derivatives are given for using gradient methods of ML estimation such as Newton Raphson:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial^2 \alpha_1} &= -\frac{n}{C^2} \left(\frac{\partial C}{\partial \alpha_1} \right)^2 + \frac{n}{C} \frac{\partial^2 C}{\partial^2 \alpha_1} = \\ &= \frac{n}{C^2} \left(\frac{\partial C}{\partial \alpha_1} \right)^2 - \frac{n}{C} \left(\frac{2}{\alpha_1} + \log\left(1 - \frac{d_1}{T}\right) \right) \frac{\partial C}{\partial \alpha_1}\end{aligned}\quad (87)$$

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial^2 \alpha_2} &= -\frac{n}{C^2} \left(\frac{\partial C}{\partial \alpha_2} \right)^2 + \frac{n}{C} \frac{\partial^2 C}{\partial^2 \alpha_2} = \\ &= \frac{n}{C^2} \left(\frac{\partial C}{\partial \alpha_2} \right)^2 + \frac{2n}{\alpha_2 C} \frac{\partial C}{\partial \alpha_2} - \frac{nCT}{\alpha_2} \left[\frac{1}{\alpha_3} \left(\frac{d_2}{T} \right)^{\alpha_2} \log \frac{d_2}{T} \left(2 + (\alpha_2 - \alpha_3) \log \frac{d_2}{T} \right) - \right. \\ &\quad \left. - \frac{1}{\alpha_1} \left(1 - \frac{d_1}{T} \right)^{\alpha_2} \log\left(1 - \frac{d_1}{T}\right) \left(1 - \left(1 - \frac{d_1}{T} \right)^{-\alpha_1} \right) \left(2 + \alpha_2 \log\left(1 - \frac{d_1}{T}\right) \right) \right]\end{aligned}\quad (88)$$

$$\frac{\partial^2 \mathcal{L}}{\partial^2 \alpha_3} = -\frac{n}{C^2} \left(\frac{\partial C}{\partial \alpha_3} \right)^2 + \frac{n}{C} \frac{\partial^2 C}{\partial^2 \alpha_3} = \frac{n}{C^2} \left(\frac{\partial C}{\partial \alpha_3} \right)^2 - \frac{n\alpha_3}{2C} \frac{\partial C}{\partial \alpha_3}\quad (89)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_1 \alpha_2} = \frac{2}{C} \frac{\partial C}{\partial \alpha_1} \frac{\partial C}{\partial \alpha_2} + \log\left(1 - \frac{d_1}{T}\right) \frac{\partial C}{\partial \alpha_1}\quad (90)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_1 \alpha_3} = \frac{2}{C} \frac{\partial C}{\partial \alpha_1} \frac{\partial C}{\partial \alpha_3}\quad (91)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_2 \alpha_3} = \frac{2}{C} \frac{\partial C}{\partial \alpha_2} \frac{\partial C}{\partial \alpha_3} + \log\left(\frac{d_2}{T}\right) \frac{\partial C}{\partial \alpha_3}\quad (92)$$

References

- [1] Bapna R., Goes, P., & Gupta, A. (2003), "Analysis and Design of Business-to-Consumer Online Auctions", *Management Science*, **49**, 1, pp. 85-101.
- [2] Billingsley P., (1995), *Probability & Measure*, Wiley & Sons, 3rd Edition.
- [3] Bright, L., Gal, A., & Raschid, R.(2004), "Adaptive Pull-Based Data Freshness Policies for Diverse Update Patterns", *Technical Report*, UMIACSTR-2004-01, University of Maryland.
- [4] Crovella, M. E. & Bestavros, A. (1995), Explaining World Wide Web Traffic Self-Similarity, *Technical Report TR-95-015*, Computer Science Department Boston University.
- [5] Dennis, J. E. and Schnabel, R. B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Englewood Cliffs, NJ: Prentice Hall.
- [6] Gal, A. & Eckstein, J. (2001), "Managing Periodically Updated Data in Relational Databases: A Stochastic Modeling Approach", *Journal of the ACM*, **48**,6, pp. 1141-1183.
- [7] Jeong, HD.J., Pawlikowski, K. & McNickle, D.C. (1999), "Generation of Self-Similar Time Series for Simulation Studies of Telecommunication Networks", *Proceedings of the Proceedings of the First Western Pacific Workshop on Stochastic Models in Engineering, Technology and Management*, Christchurch, pp. 221-30.
- [8] Jank, W., & Shmueli, G. (2003), "Dynamic Profiling of Online Auctions Using Curve Clustering", *submitted for publication*.
- [9] Kauffman, R. J., & Wood, C. A. (2000), "Running Up the Bid: Modeling Seller Opportunism in Internet Auctions", *Proceedings of the 2000 Americas Conference on Information Systems*.
- [10] Leland, W.E. ,Taqqu, M.S. , Willinger, W., and Wilson, D.V. (1994), "On the self-similar nature of Ethernet traffic (extended version)", *IEEE/ACM Transactions on Networking*, **2**, pp.1-15.
- [11] Mandelbrot, B.B. (1969), "Long-run linearity, locally Gaussian processes, H-spectra and infinite variances", *Intern. Econom. Review*, **10**, pp. 82-113.
- [12] Resnick, S.I. (1997), "Heavy Tail Modeling and Teletraffic Data", *The Annals of Statistics*, **25**, 5, pp. 1805-1849.
- [13] Resnick, S. I. (1998), *A Probability Path*, Birkhäuser Publ.
- [14] Roth A.E. & Ockefels A. (2000), "Last Minute Bidding and The Rules for Ending Second-Price Auctions: Theory & Evidence from a Natural Experiment on the Internet", NBER Working Paper #7729.
- [15] Vakrat Y. & Seidmann, A (2000), "Implications of the Bidders Arrival Process on the Design of Online Auctions", *Proceedings of the 33rd Hawaii International Conference on System Sciences*, pp.1-10.
- [16] Vose M. D. (1999), *The simple Genetic Algorithm*, MIT Press.

- [17] Willinger, W., Taqqu, M.S., Leland, W.E., and Wilson, D. V. (1995), "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements", *Statistical Science*, Vol. 10, No. 1, pp. 67-85.
- [18] Zhang, A., Beyer, D., Ward, J., Liu, T., Karp, A., Guler, K., Jain, S., and Tang, H.K. (2002), "Modeling the Price-Demand Relationship Using Auction Bid Data", Hewlett-Packard Labs Technical Report HPL-2002-202 (<http://www.hpl.hp.com/techreports/2002/HPL-2002-202.pdf>)