

# The BARISTA: A Model For Bid Arrivals in Online Auctions

Galit Shmueli

Department of Decision and Information Technologies, R. H. Smith School of Business,  
University of Maryland, College Park, MD 20742, USA. gshmueli@rhsmith.umd.edu

Ralph P. Russo

Department of Statistics & Actuarial Science, University of Iowa, Iowa City, IA 52242, USA. rrusso@stat.uiowa.edu

Wolfgang Jank

Department of Decision and Information Technologies, R. H. Smith School of Business,  
University of Maryland, College Park, MD 20742, USA. wjank@rhsmith.umd.edu

The arrival process of bidders and bids in online auctions is important for studying and modelling supply and demand in the online marketplace. A popular assumption in the online auction literature is that a Poisson process is a reasonable approximation. This approximation underlies theoretical derivations, statistical models and simulations used in field studies. However, there has been much evidence that the arrival of bids (or bidders) is affected by deadline and earliness effects. This leads to a three-stage process that is not captured by the homogenous Poisson model. An additional feature that has been reported by various authors is an apparent self-similarity in the arrival process. Despite the wide evidence for the changing bidding intensities and the self-similarity, there has been no rigorous attempt at developing a model that adequately approximates bid data. In this paper we introduce a non-homogeneous Poisson process that captures these features. We call this the BARISTA process (Bid ARrivals In STAgEs) because of its ability to generate different intensities at different stages, like espresso drinks with different intensities. We describe the properties of this model, show how to simulate bid arrivals from it, and how to use it for estimation and inference. We illustrate its power and usefulness by fitting simulated and real data from eBay.com.

*Key words:* Non-homogenous Poisson process; bidding frequency; self-similarity; bidding dynamics; sniping.

---

# 1. Introduction and Motivation

Empirical research of online auctions has been flourishing in recent years due to the important role that these auctions play in the marketplace, and the availability of large amounts of high-quality bid data from websites such as eBay, Yahoo!, OnSale, and uBid. Many of the theoretical results derived for traditional (offline) auctions have been shown to fail in the online setting for reasons such as globalism, computerized bidding and the recording of complex bids, longer auction durations, more flexibility in design choice by the seller, and issues of trust. A central factor underlying many important results is the number of bidders participating in the auction. Typically it is assumed that this number is fixed (Pinker et al. 2003) or fixed but unknown (McAfee & McMillan 1997). In online auctions the number of bidders and bids is not predetermined, and it is known to be affected by the auction design and its dynamics. Thus, in both the theoretical and empirical domains the number of bidders/bids plays an important role. Although the empirical domain enables the direct study of the bid arrival process, this process has not been treated in a rigorous way. In general, there has been strong evidence of two major features of the bid (and bidder) arrival processes in online auctions: (1) a non-homogenous intensity that possesses two or three distinct stages, and (2) a self similarity effect in the distribution of bid arrival times. We describe each of these features next.

## 1.1. Multi-stage arrival intensities

Time-limited tasks are omnipresent in the offline world: voting for a new president, purchasing tickets for a popular movie or sporting event, filing one's federal taxes, etc. In many of these cases arrivals are especially intense as the deadline approaches. For instance, during the 2001 political elections in Italy, more than 20 million voters cast their ballots between 13:00-22:00 (Bruschi et al. 2002), when ballots were scheduled to close at 22:00. Similarly, tax returns are typically filed in the last minute. For instance, about one-third of all returns are not filed until the last two weeks of tax season<sup>1</sup>. According to Ariely et al. (2003), deadline effects have been noted in studies of bargaining, where agreements are reached in the final moments before the deadline (Roth et al. 1998), among animals, when they respond more vigorously toward the expected end of a reinforcement schedule, and in human task completion where individuals become increasingly impatient toward the task's end. Furthermore, people use different strategies when games are framed as getting close to the end (even when these are arbitrary break points; Croson (1996)). In addition to the deadline effect, there is an effect of earliness where the strategic use of time moves transactions earlier than later, e.g. in the labor market (Roth & Xing 1994; Avery et al. 2001).

---

<sup>1</sup>[www.heraldstandard.com/site/news.cfm?newsid=14359378&BRD=2280&PAG=461&dept.id=480247&crfi=6](http://www.heraldstandard.com/site/news.cfm?newsid=14359378&BRD=2280&PAG=461&dept.id=480247&crfi=6)

Such deadline and earliness effects have also been observed in the online environment. Several researchers have noted deadline effects in internet auctions (Bajari & Hortacsu 2000; Borle et al. 2005; Ku et al. 2004; Roth & Ockenfels 2000; Wilcox 2000). In many of these studies it was observed that a non-negligible percent of bids arrive at the very last minute of the auction. This phenomenon, called “bid sniping” has received much attention, and numerous explanations have been suggested to explain its existence. Empirical studies of online auctions have also reported an unusual amount of bidding activity at the auction start followed by a longer period of little or no activity. Bapna et al. (2003) refer to bidders who place a single early bid as “evaluators”. Finally, “bid shilling”, a fraudulent act where the seller places dummy bids to drive up the price, is associated with early and high bidding (Kauffman & Wood 2000). The existence of these bid-timing phenomena are important factors in determining outcomes at the auction level as well as at the market level. They have therefore received much attention from the research community.

Despite the wide evidence for a varying intensity in bid/bidder arrivals, the existing online auction literature tends to favor the homogenous Poisson process. This assumption underlies various theoretical derivations, is the basis for the simulation of bid data, and is used to design field experiments. Bajari & Hortacsu (2000) specify and estimate a structural econometric model of bidding on eBay, assuming a Poisson bidder arrival process. Etzion et al. (2003) suggest a model for segmenting consumers at dual channel online merchants. Based on the assumption of Poisson arrivals to the website, they model consumer choice of channel, simulate consumer arrivals and actions, and compute relationships between auction duration, lot size, and the constant Poisson arrival rate  $\lambda$ . Zhang et al. (2002) model the demand curve for consumer products in online auctions based on Poisson bidder arrivals, and fit the model to bid data. Pinker et al. (2003), and Vakrat & Seidmann (2000) use a Poisson Process for modelling the arrival of bidders in going-going-gone auctions. They use the intensity function  $\lambda(t) = \lambda_a e^{-t/T}$ ,  $0 \leq t \leq T$ , where  $T$  is the auction duration, and  $\lambda_a$  is the intensity of website traffic into the auction. This model describes the decline in the number of new bidders as the auction progresses. However, their histograms describing the bidder arrival time distribution in 1-hour auctions as well as in 1-day auctions indicates a process with two or more stages. Haubl & Popkowski Leszczyc (2003) design and carry out an experiment for studying the effect of fixed-price charges (e.g., shipping costs) and reserve prices on consumer’s product valuation. The experiment uses simulated data that are based on Poisson arrivals of bidders. These studies are among the many that rely on a Poisson arrival process assumption.

## 1.2. Self-Similarity (and its breakdown)

While both the offline and online environments share the deadline and earliness effects, the online environment appears to possess the additional property of *self-similarity* in the bid arrival process<sup>2</sup>. Self similarity refers to the “striking regularity” of shape that can be found among the distribution of bid arrivals over the intervals  $[t, T]$ , as  $t$  approaches the auction deadline  $T$ . Self similarity is central in applications such as web, network and ethernet traffic. Huberman & Adamic (1999) found that the number of visitors to websites follows a universal power law. Liebovitch & Schwartz (2003) reported that the arrival process of email viruses is self-similar. However, this has also been reported in other online environments. For instance, Aurell & Hemmingsson (1997) showed that times between bids in the interbank foreign exchange market follow a power law distribution.

Several authors reported results that indicate the presence of self-similarity in the bidding frequency in online auctions. Roth & Ockenfels (2000) found that the arrival of last bids by bidders during an online auction is closely related to a self-similar process. They approximated the CDF of bid arrivals in “reverse time” (i.e., the CDF of the elapsed times between the bid arrivals and the auction deadline) by the power functional form  $F_T(t) = (t/T)^\alpha$ , over the interval  $[0, T]$ , and estimated  $\alpha$  from the data using OLS. This approximates the distribution of bids over intervals that range from the last 12 hours to the last 10 minutes, but accounts for neither the final minutes of the auction nor the auction start and middle. Yang et al. (2003) found that the number of bids and the number of bidders in auctions on eBay and on its Korean partner (auction.co.kr) follows a power law distribution. This was found for auctions across multiple categories. The importance of this finding, which is closely related to the self similarity property, is that the more bidding one observes up to a fixed time point, the higher the likelihood of seeing another bid before the next time point. According to Yang et al. (2003) such power-law behaviors imply that the online auction system is driven by self-organized processes, involving all bidders who participate in a given auction activity.

The implications of bid arrivals following a self-similar process instead of 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. The self-similar property suggests that the rate of 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 their chance of winning the auction since the probability that another competitor successfully places an even higher bid before closing is diminishing. This common

---

<sup>2</sup>This property was also found in the offline process of bargaining agreements, as described in Roth & Ockenfels (2000).

bidding strategy of “bid sniping,” (or, “last minute bidding”) 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-end auctions tend to follow a uniform distribution (Roth et al. (1998)). This has not been found in open-ended, or “going-going-gone” auctions, such as those on Amazon (which no longer holds auctions), Yahoo!, or uBid.com, where the auction continues several minutes after the last bid was placed.

Thus, in addition to the evidence for self similarity in online auctions, there is also evidence of its breakdown during the very last moments of a closed-end auction. Roth & Ockenfels (2000) note that the empirical CDF plots for intervals that range between the last 12 hours of the auctions and the last 1 minute all look very similar except for the last 1-minute plot. Being able to model this breakdown is essential, since the last moments of the auction (when sniping takes place) are known to be crucial to the final price. In the absence of such a model, we introduce a bid arrival process that describes the frequency throughout the *entire* auction. Rather than focusing on the last several hours and excluding the last moments, our model accommodates the changes in bidding intensity during the three main stages of the auction: the auction start, its middle, and the very last moments. It also estimates the changepoints between these three stages.

### 1.3. Motivating Example

In an attempt to investigate the bid arrival process in online auctions 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 for the purposes of examining the self-similarity property. 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. Note 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 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! 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 to not 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](http://www.auctionsniper.com)). Clearly, the closer to the end the auction gets, the higher the likelihood that a bid will not get registered successfully.

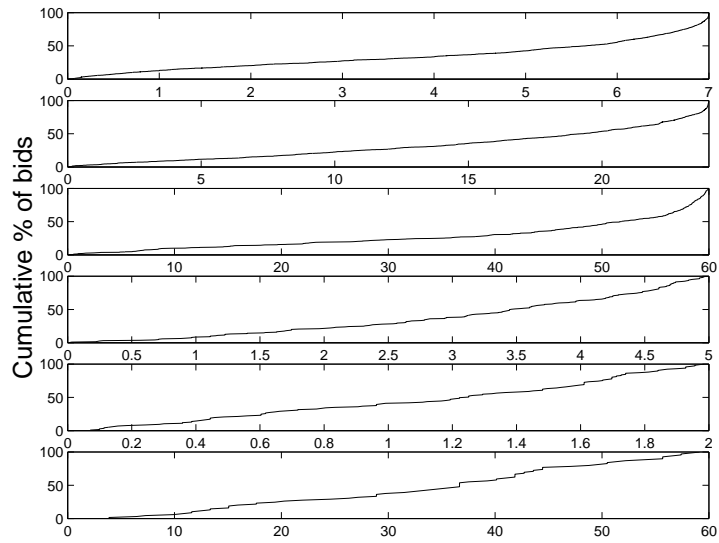


Figure 1: Empirical CDF of number of bids in 189 Palm M515 auctions. The plots (top to bottom) are for the entire auction (7 days), the last day (24 hours), last hour (60 min), last 5 min, last 2 min and last 1 min (60 seconds) of the auction.

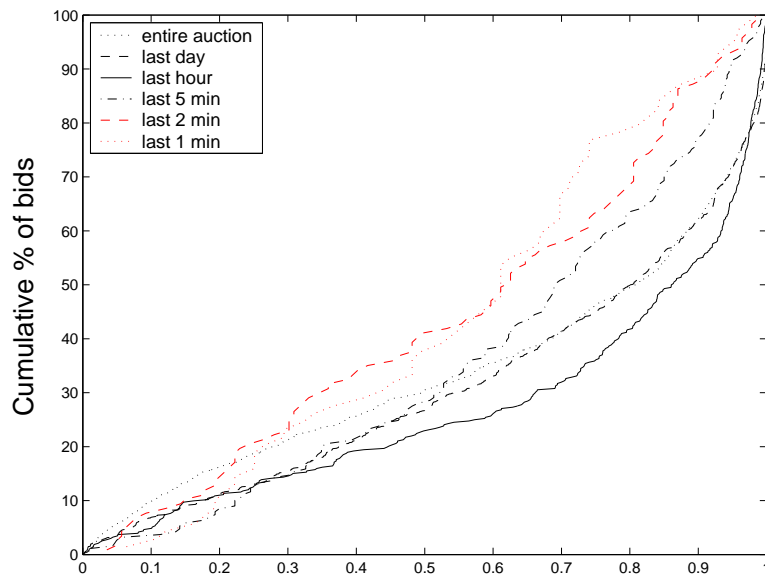


Figure 2: Empirical CDFs of number of bids in 189 Palm M515 auctions overlaid

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.

In the next section we introduce a flexible non-homogeneous Poisson process (NHPP) that captures the empirical phenomenon described above. It also accounts for two additional observed phenomena in the online auction literature: “early bidding” and “last minute bidding” (sniping).

The paper is organized as follows: Section 2 introduces the model and its properties, and describes two special cases where there are fewer stages in the auction. Section 3 describes a method for simulating data from this process and several methods for estimating the model parameters. In section 4 we use simulated data and the real data described above to illustrate the estimation and to show model fit. In Section 5 we discuss the meaning of self-similarity in the online auction setting and describe further practical applications where knowledge of the bid arrival process is advantageous.

## 2. The BARISTA: A three-stage Non-Homogeneous Poisson Process

We now describe a process that captures the two main features of arrivals in online auctions: the three stages and the self similarity (with its breakdown). We call this the BARISTA<sup>3</sup> (Bid ARrivals In STAgEs), because it generates different intensities of activity:

1. The *espresso* stage - early and intense: The intense bidding that occurs early in the auction,
2. The *macchiato*<sup>4</sup> stage - stained: The mid auction bid arrivals, characterized by self-similarity that is contaminated by the sniping at the auction end,
3. The *ristretto*<sup>5</sup> stage - short and very intense: The last moments of the auctions, characterized by very intense activity.

---

<sup>3</sup>A Barista is a person professionally trained in the art of espresso preparation.

<sup>4</sup>An espresso stained with a spot of milk foam

<sup>5</sup>An extra-intense espresso

## 2.1. 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 introduce a particular intensity function that captures the three-stage dynamics described above. Suppose bids arrive during  $[0, T]$  in accordance with a non-homogeneous Poisson process  $N(s)$ ,  $0 \leq s \leq T$ , with intensity function

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

Note that this intensity function is continuous, so there are no jumps at times  $d_1$  and  $T - d_2$ . The random variable  $N(s)$  which counts the number of arrivals until time  $s$  follows a Poisson distribution with mean

$$m(s) = \begin{cases} K \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_1}\right) & \text{for } 0 \leq s \leq d_1 \\ K \left(1 - \left(1 - \frac{d_1}{T}\right)^{\alpha_1}\right) + \frac{Tc}{\alpha_2} \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_1}\right) & \text{for } d_1 \leq s \leq T - d_2 \\ K \left(1 - \left(1 - \frac{d_1}{T}\right)^{\alpha_1}\right) + \frac{Tc}{\alpha_2} \left(1 - \frac{d_2}{T}\right)^{\alpha_1} + \frac{Tc}{\alpha_3} \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \left(1 - \frac{s}{T}\right)^{\alpha_3}\right) & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (2)$$

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

Given that  $N(T) = n$ , the collection of arrival times are equivalent to the order statistics of a random sample of size  $n$  from the distribution having distribution  $F(s) = m(s)/m(T)$ :

$$F(s) = \begin{cases} \frac{CT}{\alpha_1} \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} \left[1 - \left(1 - \frac{s}{T}\right)^{\alpha_1}\right] & \text{for } 0 \leq s \leq d_1 \\ \frac{CT}{\alpha_1 \alpha_2} \left[ (\alpha_1 - \alpha_2) \left(1 - \frac{d_1}{T}\right)^{\alpha_2} + \alpha_2 \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} - \alpha_1 \left(1 - \frac{s}{T}\right)^{\alpha_2} \right] & \text{for } d_1 \leq s \leq T - d_2 \\ 1 - \frac{CT}{\alpha_3} \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{s}{T}\right)^{\alpha_3} & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (3)$$

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

$$F(t) = F(d_1) + \frac{CT}{\alpha_2} \left[ \left(1 - \frac{d_1}{T}\right)^{\alpha_2} - \left(1 - \frac{s}{T}\right)^{\alpha_2} \right] \quad (4)$$

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 density function corresponding to this process is given by

$$f(s) = \begin{cases} C \left(1 - \frac{d_1}{T}\right)^{\alpha_2 - \alpha_1} \left(1 - \frac{s}{T}\right)^{\alpha_1 - 1} & \text{for } 0 \leq s \leq d_1 \\ C \left(1 - \frac{s}{T}\right)^{\alpha_2 - 1} & \text{for } d_1 \leq s \leq T - d_2 \\ C \left(\frac{d_2}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{s}{T}\right)^{\alpha_3 - 1} & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (5)$$

We expect  $\alpha_3$  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.

## 2.2. Properties of the BARISTA Process

The process described by (1) - (4) has two properties that lead to a wide family of processes:

### 2.2.1. An additive property

If  $N_k$ ,  $1 \leq k \leq m$ , are independent members of the BARISTA process having  $c$  parameters  $c_1, \dots, c_m$  and common  $(\alpha_1, \alpha_2, \alpha_3)$  and  $(d_1, d_2)$  parameters, then the aggregated process  $N = \sum_{1 \leq k \leq m} N_k$  is a BARISTA with parameters  $(\alpha_1, \alpha_2, \alpha_3)$ ,  $(d_1, d_2)$ , and  $c = \sum_{1 \leq k \leq m} c_k$ .

### 2.2.2. A regenerative property

An observer who counts only the bid arrivals occurring after time  $\beta T$ , some  $0 \leq \beta < 1$ , sees the process

$$N_\beta(s) := N(s) - N(\beta T), \quad \beta T \leq s \leq T$$

$N_\beta$  is an *NHPP* with intensity function  $\lambda_\beta = \lambda$ , restricted to the interval  $[\beta T, T]$ . We can write  $\lambda$  as

$$\lambda(s) = \begin{cases} c(1 - \beta)^{\alpha_2 - 1} \left(1 - \frac{d_1 - \beta T}{T(1 - \beta)}\right)^{\alpha_2 - \alpha_1} \left(1 - \frac{s - \beta T}{T(1 - \beta)}\right)^{\alpha_1 - 1} & 0 \leq s \leq d_1 \\ c(1 - \beta)^{\alpha_2 - 1} \left(1 - \frac{s - \beta T}{T(1 - \beta)}\right)^{\alpha_2 - 1} & d_1 \leq s \leq T - d_2 \\ c(1 - \beta)^{\alpha_2 - 1} \left(\frac{d_2}{T(1 - \beta)}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{s - \beta T}{T(1 - \beta)}\right)^{\alpha_3 - 1} & T - d_2 \leq s \leq T \end{cases}$$

Taking  $\beta T$  as the new *zero*, and recording time on a new (faster) clock where *one new minute (a shminute)* =  $(1 - \beta)$  minutes on the original clock, we have

$$\lambda_\beta(s) = \begin{cases} c_\beta \left(1 - \frac{d_{1,\beta}}{T}\right)^{\alpha_2 - \alpha_1} \left(1 - \frac{s}{T}\right)^{\alpha_1 - 1} & 0 \leq s \leq d_{1,\beta} \\ c_\beta \left(1 - \frac{s}{T}\right)^{\alpha_2 - 1} & d_{1,\beta} \leq s \leq T - d_{2,\beta} \\ c_\beta \left(\frac{d_{2,\beta}}{T}\right)^{\alpha_2 - \alpha_3} \left(1 - \frac{s}{T}\right)^{\alpha_3 - 1} & T - d_{2,\beta} \leq s \leq T \end{cases}$$

where  $c_\beta = c(1 - \beta)^{\alpha_2 - 1}$ ,  $d_{1,\beta} = \max(d_1 - \beta T, 0)$ , and  $d_{2,\beta} = \min(d_2, T(1 - \beta))$ . Thus,  $\lambda_\beta$  has of the same form as our original  $\lambda$  with  $\lambda = \lambda_0$  in the new notation.

### 2.3. Special Cases

In empirical studies,  $d_1$  appears to be small (a few hours or 1-2 days) and  $d_2$  very small (a few minutes) compared to  $T$  (several days). Thus, most of the BARISTA process is realized in the second (*macchiato*) stage, during which the process can be regarded as having *contaminated self-similarity*. The contamination is caused by the bid arrivals in the third stage, and increases as  $s \rightarrow T - d_2$ .

When  $d_1 = d_2 = 0$ , the BARISTA process reduces to the single-stage process ( $NHPP_1$ ) with an intensity function  $\lambda(s) = c(1 - \frac{s}{T})^\alpha$  and associated  $F$  function  $F(s) = 1 - (1 - \frac{s}{T})^\alpha$ ,  $0 \leq s \leq T$ . For  $(\theta, t) \in [0, 1] \times (0, T]$  we have

$$\frac{1 - F(T - t\theta)}{1 - F(T - t)} = \theta^\alpha \text{ (independent of } t)$$

and thus we have a *pure self-similar* process. The joint  $MLE$  of  $(\alpha, c)$  is obtainable in this case (see appendix A):

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

Since  $X \sim F \implies -\log \left( 1 - \frac{X}{T} \right) \sim \exp(\text{rate} = \alpha)$ , and  $\lim_{c \rightarrow \infty} \Pr(N(T) \rightarrow \infty) = 1$ , a conditioning argument on  $N(T)$  yields an asymptotic result:

$$\sqrt{N(T)} \left( \frac{\alpha}{\hat{\alpha}} - 1 \right) \xrightarrow{D} n(0, 1) \text{ as } c \rightarrow \infty$$

When  $d_1 = 0$ , the BARISTA process reduces to the two-stage process ( $NHPP_2$ ), with a single *changepoint* at  $d_2$ . This process is useful for modelling bid arrivals in auctions that lack the initial surge of early bidding. For more technical details on these special cases see Shmueli et al. (2004).

## 3. Fitting the BARISTA process to data: Parameter Estimation and Process Simulation

Simulated bid arrivals are useful in field experiments, in evaluation of model fit, and for quantifying sampling error. The method is simple to program and computationally efficient. Fitting the BARISTA process to data requires estimating the two changepoints and three  $\alpha$  parameters. We introduce three estimation methods that range in their computational intensiveness and accuracy<sup>6</sup>.

<sup>6</sup>Matlab code for the simulation and estimation procedures is available at <http://www.smith.umd.edu/ceme/statistics/code.html>

### 3.1. Process simulation

We use the inversion method in order to simulate data from the BARISTA process on the interval  $[0, T]$ .

The inverse CDF can be written as:

$$F^{-1}(s) = \begin{cases} T - T \left\{ 1 - \frac{s\alpha_1}{CT} \left( 1 - \frac{d_1}{T} \right)^{\alpha_1 - \alpha_2} \right\}^{1/\alpha_1} & \text{for } 0 \leq s \leq d_1 \\ T - T \left\{ \left( 1 - \frac{d_1}{T} \right)^{\alpha_2} - \frac{\alpha_2}{CT} (s - F_3(d_1)) \right\}^{1/\alpha_2} & \text{for } d_1 \leq s \leq T - d_2 \\ T - T \left\{ \frac{\alpha_3}{CT} (1 - s) \left( \frac{d_2}{T} \right)^{\alpha_3 - \alpha_2} \right\}^{1/\alpha_3} & \text{for } T - d_2 \leq s \leq T \end{cases} \quad (6)$$

The algorithm for generating  $n$  arrivals  $(x_1, \dots, x_n)$  is then:

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

$$x_k = \begin{cases} T - T \left\{ 1 - \frac{u_k \alpha_1}{CT} \left( 1 - \frac{d_1}{T} \right)^{\alpha_1 - \alpha_2} \right\}^{1/\alpha_1} & \text{if } u_k < F(d_1) \\ T - T \left\{ \frac{\alpha_2}{CT} (F_3(d_1) - u_k) + \left( 1 - \frac{d_1}{T} \right)^{\alpha_2} \right\}^{1/\alpha_2} & \text{if } F(d_1) \leq u_k < F(T - d_2) \\ T - T \left\{ \frac{\alpha_3}{CT} u_k \left( \frac{d_2}{T} \right)^{\alpha_3 - \alpha_2} \right\}^{1/\alpha_3} & \text{if } u_k \geq F(T - d_2) \end{cases} \quad (7)$$

### 3.2. Parameter Estimation

We describe three estimation methods each having a different tradeoff between computational intensity and accuracy, and with varying amounts of user input.

#### 3.2.1. Quick & Crude (CDF-Based) Estimation

The estimation of the  $\alpha$  parameters depends on the changepoints  $d_1$ ,  $T - d_2$  and vice-versa. As a crude start, we pick three intervals  $[T - t, T - s]$  that we are confident lie in the first, second, or third stages, and use those for estimating the  $\alpha$  parameters. We then use the  $\alpha$  estimates to obtain estimates for the changepoints.

In both cases the estimates are based on writing the parameters as a function of the CDF, and then plugging in the empirical CDF to obtain estimates.

## Estimation of $\alpha$ Parameters

From (3) it can be seen that in each interval the CDF of the Poisson process is in the form  $F(t) = \beta_j - \theta_j \left(1 - \frac{t}{T}\right)^{\alpha_j}$ , ( $j = 1, 2, 3$ ), and therefore the same approximation works on each of the three intervals  $[0, d_1]$ ,  $[d_1, T - d_2]$  and  $[T - d_2, T]$ . After choosing intervals  $[T - t, T - s]$  that we are confident lie in stage  $j$  (the first, second, or third stage), we have

$$\begin{aligned} \frac{F(T - t) - F(T - \sqrt{st})}{F(T - \sqrt{st}) - F(T - s)} &= \frac{\theta_j \left[1 - \left(1 - \frac{T-t}{T}\right)^{\alpha_j}\right] - \theta_j \left[1 - \left(1 - \frac{T-\sqrt{st}}{T}\right)^{\alpha_j}\right]}{\theta_j \left[1 - \left(1 - \frac{T-\sqrt{st}}{T}\right)^{\alpha_j}\right] - \theta_j \left[1 - \left(1 - \frac{T-s}{T}\right)^{\alpha_j}\right]} \\ &= \frac{(ts)^{\alpha_j/2} - t^{\alpha_j}}{s^{\alpha_j} - (st)^{\alpha_j/2}} = \frac{(s^{\alpha_j/2} - t^{\alpha_j/2})t^{\alpha_j/2}}{(s^{\alpha_j/2} - t^{\alpha_j/2})s^{\alpha_j/2}} = \left(\frac{t}{s}\right)^{\alpha_j/2}. \end{aligned} \quad (8)$$

The relevant  $\alpha$  is given by

$$\alpha_j = 2 \frac{\log [F(T - t) - F(T - \sqrt{st})] - \log [F(T - \sqrt{st}) - F(T - s)]}{\log t - \log s} \quad (9)$$

We then estimate  $\alpha_j$  by plugging the empirical CDF  $F_e = N(t)/N(T)$  for  $F$  in the approximation.

For  $\alpha_3$  we can use the exact relation

$$\alpha_3 = \frac{\log R(t_3)/R(t'_3)}{\log(T - t_3)/(T - t'_3)} \quad (10)$$

where  $R(t) = 1 - F(t)$  and  $t_3, t'_3$  are within  $[T - d_2, T]$ . To estimate  $\alpha_3$  we pick reasonable values of  $t_3, t'_3$  and use the empirical survival function  $R_e = 1 - F_e$ .

Obtaining standard errors for these estimators can be done by bootstrapping (see Efron & Tibshirani (1993) for details), due to the low computational effort involved in this estimation method.

To assess this method we simulated 5000 random observations from the BARISTA process 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 stage) and  $d_2 = 5/10080$  (defining the last 5 minutes as the third stage). The estimates and their standard errors are given in Table 1.

To study the robustness of the estimators to the choice of  $t$  and  $s$ , we computed the quick & crude estimate for  $\alpha_1$  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 3 illustrates the estimates obtained for these intervals. For values of  $t_1$  between 1.5-3.5 days, the estimate for  $\alpha_1$  is relatively stable and close to 3. Similarly, the right panel in Figure 3 describes the estimates of  $\alpha_3$ , using (10), 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  $\alpha_2$  an interval such as  $[3, 6.9]$  is reasonable. Figure 4 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.

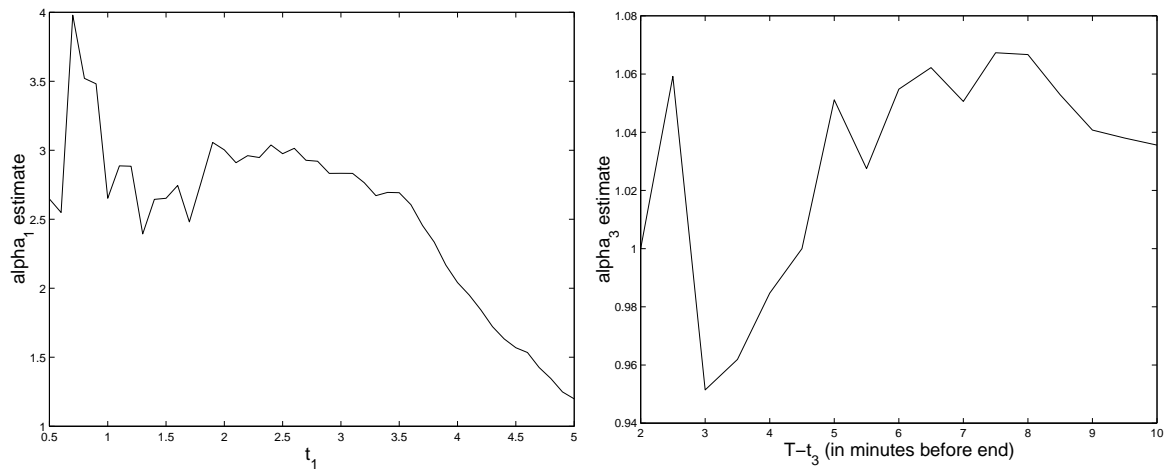


Figure 3: Quick estimates of  $\alpha_1, \alpha_2$ , and  $\alpha_3$  as a function of the input intervals, for simulated BARISTA process data.

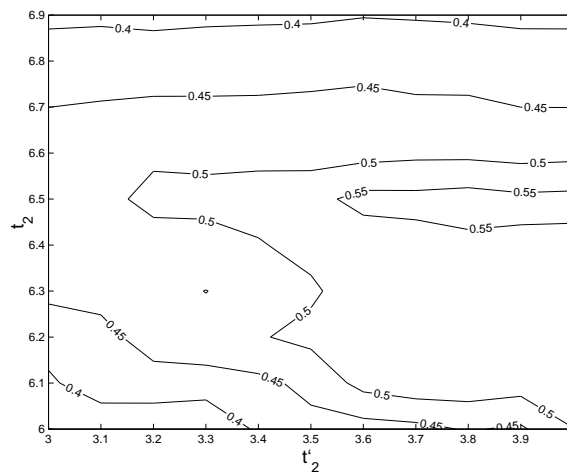


Figure 4: Quick & crude estimate of  $\alpha_2$  as a function of  $[t'_2, t_2]$  choice.  $\hat{\alpha}_2$  is between 0.4-0.55 in the entire range of intervals. The more extreme intervals ( $t'_2 < 3.4$  or  $t_2 > 6.8$ ) yield  $\hat{\alpha}_2 = 0.4$ .

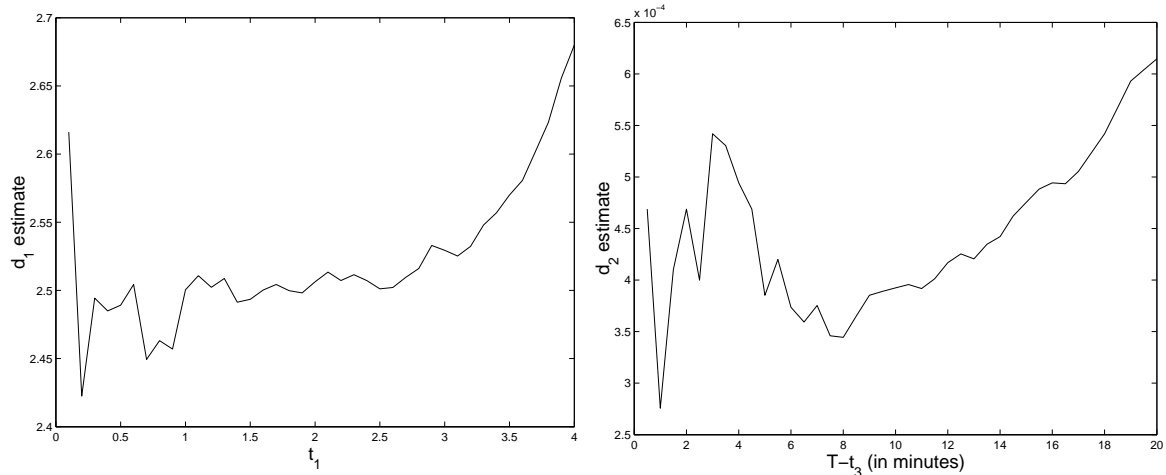


Figure 5: Graphs of  $\hat{d}_1$  vs.  $t_1$  (left) and  $\hat{d}_2$  vs. initial values of  $T - t_3$  (right) for simulated data. The estimate for  $d_1$  is stable at  $\approx 2.5$ .  $\hat{d}_2$  using the last 2-5 minute interval is in the range of 4-5 minutes.

### 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  $\frac{F_3(t_2) - F_3(t'_2)}{F_3(t_1)}$  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:

$$d_1 = T - T \left\{ \frac{\alpha_1}{\alpha_2} \cdot \frac{F(t_1)}{F(t_2) - F(t'_2)} \cdot \frac{(1 - t'_2/T)^{\alpha_2} - (1 - t_2/T)^{\alpha_2}}{1 - (1 - t_1/T)^{\alpha_1}} \right\}^{\frac{1}{\alpha_2 - \alpha_1}} \quad (11)$$

$$d_2 = T \left\{ \frac{\alpha_3}{\alpha_2} \cdot \frac{1 - F(t)}{F(t_2) - F(t'_2)} \cdot \frac{(1 - t'_2/T)^{\alpha_2} - (1 - t_2/T)^{\alpha_2}}{(1 - t_3/T)^{\alpha_3}} \right\}^{\frac{1}{\alpha_2 - \alpha_3}} \quad (12)$$

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  $\alpha$  parameters and the “safe” values  $t_1 = 1, t'_2 = 3, t_2 = 6$ , and  $t_3 = 7 - 2/10080$ . The estimates and their (bootstrap) standard errors are reported in table 1. Figure 5 shows the robustness of the estimates to the choice of the “safe” values. It can be seen that  $d_1$  estimates are stable between 2.4-2.6 even if we choose  $t_1$  slightly outside of the first interval  $[0, 2.5]$ .  $d_2$  estimates are between 3-5.5 minutes even when  $t_3$  is dislocated by a few minutes into the second interval.

### 3.2.2. Maximum Likelihood Estimation

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

$$\begin{aligned} \mathcal{L}(x_1, \dots, x_n | \alpha_1, \alpha_2, \alpha_3, d_1, d_2) = & \quad (13) \\ n \log C + n_1(\alpha_2 - \alpha_1) \log \left(1 - \frac{d_1}{T}\right) + n_3(\alpha_2 - \alpha_3) \log \frac{d_2}{T} + (\alpha_1 - 1)S_1 + (\alpha_2 - 1)S_2 + (\alpha_3 - 1)S_3, \end{aligned}$$

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).

$$S_1 = n_1 \log \left(1 - \frac{d_1}{T}\right) - \frac{n}{C} \frac{\partial C}{\partial \alpha_1} \quad (14)$$

$$S_2 = -n_1 \log \left(1 - \frac{d_1}{T}\right) - n_3 \log \frac{d_2}{T} - \frac{n}{C} \frac{\partial C}{\partial \alpha_2} \quad (15)$$

$$S_3 = n_3 \log \frac{d_2}{T} - \frac{n}{C} \frac{\partial C}{\partial \alpha_3} \quad (16)$$

where

$$\frac{\partial C}{\partial \alpha_1} = \frac{C^2 T}{\alpha_1^2} \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \left[ \left(1 - \frac{d_1}{T}\right)^{-\alpha_1} \left(1 + \alpha_1 \log \left(1 - \frac{d_1}{T}\right)\right) - 1 \right] \quad (17)$$

$$\begin{aligned} \frac{\partial C}{\partial \alpha_2} = & \frac{C^2 T}{\alpha_1 \alpha_3 \alpha_2^2} \left\{ \alpha_3 \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \left[ \alpha_2 \log \left(1 - \frac{d_1}{T}\right) \left( \alpha_2 - \alpha_1 + \alpha_2 \left(1 - \frac{d_1}{T}\right)^{-\alpha_1} \right) - \alpha_1 \right] + \right. \\ & \left. + \alpha_1 \left(\frac{d_2}{T}\right)^{\alpha_2} \left[ \alpha_3 + \alpha_2 \log \frac{d_2}{T} (\alpha_2 - \alpha_3) \right] \right\} = \quad (18) \end{aligned}$$

$$\begin{aligned} = & \frac{C^2 T}{\alpha_2^2} \left[ \left(\frac{d_2}{T}\right)^{\alpha_2} - \left(1 - \frac{d_1}{T}\right)^{\alpha_2} - \frac{\alpha_2^2}{\alpha_1} \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \log \left(1 - \frac{d_1}{T}\right) \left(1 - \left(1 - \frac{d_1}{T}\right)^{-\alpha_1}\right) - \right. \\ & \left. - \alpha_2 \left(\frac{d_2}{T}\right)^{\alpha_2} \log \frac{d_2}{T} + \alpha_2 \left(1 - \frac{d_1}{T}\right)^{\alpha_2} \log \left(1 - \frac{d_1}{T}\right) + \frac{\alpha_2^2}{\alpha_3} \left(\frac{d_2}{T}\right)^{\alpha_2} \log \frac{d_2}{T} \right] \\ \frac{\partial C}{\partial \alpha_3} = & \frac{C^2 T}{\alpha_3^2} \left(\frac{d_2}{T}\right)^{\alpha_2} \quad (19) \end{aligned}$$

Since the equations are non-linear in the parameters an iterative gradient method can be used (the second derivatives are given in Appendix B). 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). If the changepoints  $d_1$  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

Table 1: True and Estimated values (with standard errors) for the five BARISTA model parameters, by method.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{d}_1$	$\hat{d}_2$ (minutes)
Simulated (true) values	3	0.4	1	2.5	5
CDF-based Q&C	2.85 (0.06)	0.443 (0.001)	0.954 (0.0132)	2.5 (0.0036)	4.7 (0.13)
Genetic Algorithm	2.96 (0.008)	0.383 (0.005)	1.000 (0.009)	2.68 (0.004)	5.1 (0.35)

this maximization problem. A good starting value would be the estimate obtained from the probability plot or the quick & crude method.

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 (see, for example, Vose (1999)). 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). In addition, since they are much cheaper computationally, it is feasible to compute standard errors based on bootstrapping.

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 for the simulated data. We restricted the range of possible solutions to the hypercube  $(\alpha_1, \alpha_2, \alpha_3, d_1, d_2) \in [2.5, 3.5] \times [0.1, 0.7] \times [0.5, 1.5] \times [2, 3] \times [0, 0.001]$ . This yielded the estimates and standard errors given in table 1. All of these estimates are 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.

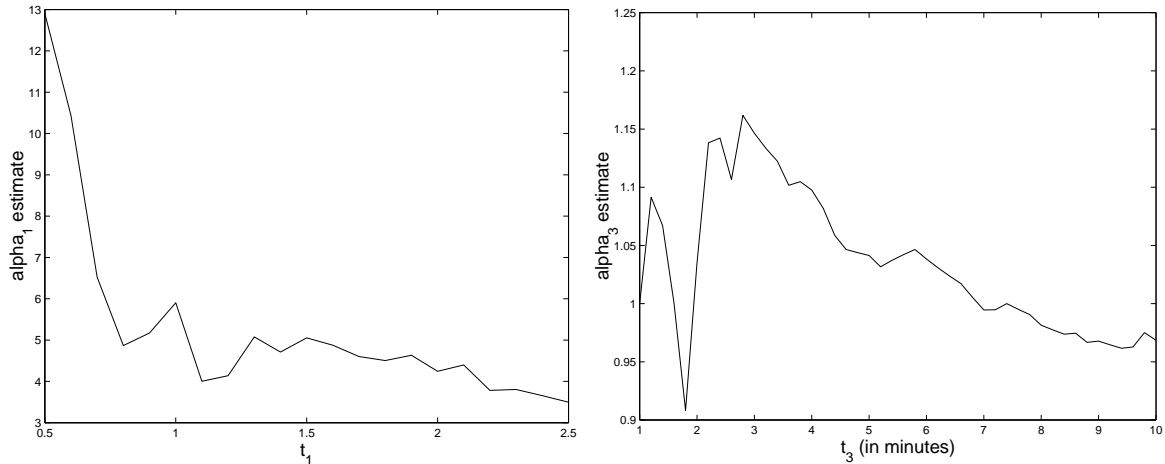


Figure 6: Quick & crude estimates of  $\alpha_1$  as a function of  $t_1$  (with  $t'_1 = 0.001$ ) (left) and of  $\alpha_3$  as a function of  $t_3$  (with  $t'_3 = 0.5/10080$ ) (right).  $\hat{\alpha}_1$  is stable around 5 for  $t_1$  in the range 0.75-1.75 days. A shorter interval does not contain 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. Empirical Results

We use the quick & crude method to estimate the parameters for the 3651 Palm bid arrival times. Based on previous empirical results, we chose the first day for estimating  $\alpha_1$ , i.e., we believe that bids placed during the first day are contained within the first “early bidding” stage. Looking at the estimate as a function of the interval chosen (Figure 6, 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  $\alpha_3$  was estimated using (10) with  $t'_3 = 7 - 0.1/10080$  and a range of values for  $t_3$ . From these,  $\alpha_3$  appears to be approximately 1. It can be seen in the right panel of Figure 6 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  $\alpha_2$ . This yielded the estimate  $\hat{\alpha}_2 = 0.36$ . Figure 7 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 ( $\hat{\alpha}_1 = 4.3, \hat{\alpha}_2 = 0.36, \hat{\alpha}_3 = 1$ ), we estimated  $d_1$  and  $d_2$ . Figure 8 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

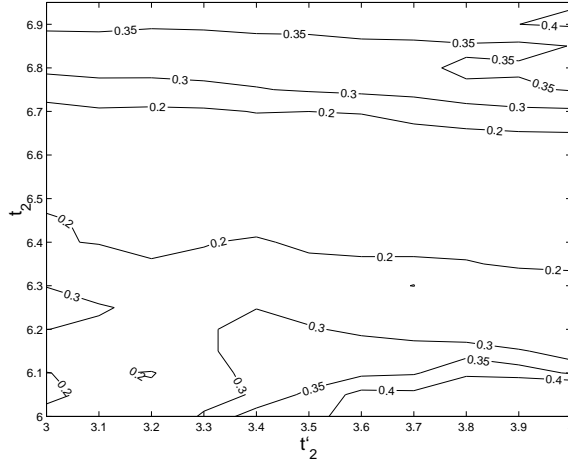


Figure 7: Quick & crude estimate of  $\alpha_2$  as a function of  $[t'_2, t_2]$ . Shorter, “safer” intervals are at the lower right. Longer intervals, containing more data, are at the upper left.  $\hat{\alpha}_2$  is between 0.2-0.4 for all intervals. For  $t_2 > 6.9$  the estimate is approximately 0.35.

Table 2: Estimates for five BARISTA model 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	4.3 (0.02)	0.36 (0.001)	1 (0.02)	1.8 (0.009)	3.3 (0.18)
Exhaustive search	4.9	0.37	1.13	1.7	2.0
Genetic Algorithm	5.55 (0.005)	0.35 (0.005)	1.1 (0.01)	1.55 (0.005)	2.0 (0.10)

increasing values obtained for  $T - t_3 > 3$  minutes we also learn that  $d_2 < 3$ .

Table 2 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. We restricted the range of possible solutions for the genetic algorithm to the hypercube  $(\alpha_1, \alpha_2, \alpha_3, d_1, d_2) \in [0, 10] \times [0, 1] \times [0, 5] \times [0, 5] \times [0, 1000min]$  and then reduced it to a tighter region. It can be seen that all methods yielded estimates in the same vicinity.

Finally, to further validate this estimated model, we simulated data from a BARISTA process with the above ML estimates as parameters. Figure 9 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 the estimated model for the Palm bid times.

The estimated model for the Palm data reveals the dynamics of these auctions over time: Indeed, the “average” auction has three stages: the initial *espresso* stage takes place during the first 1.7 days, the

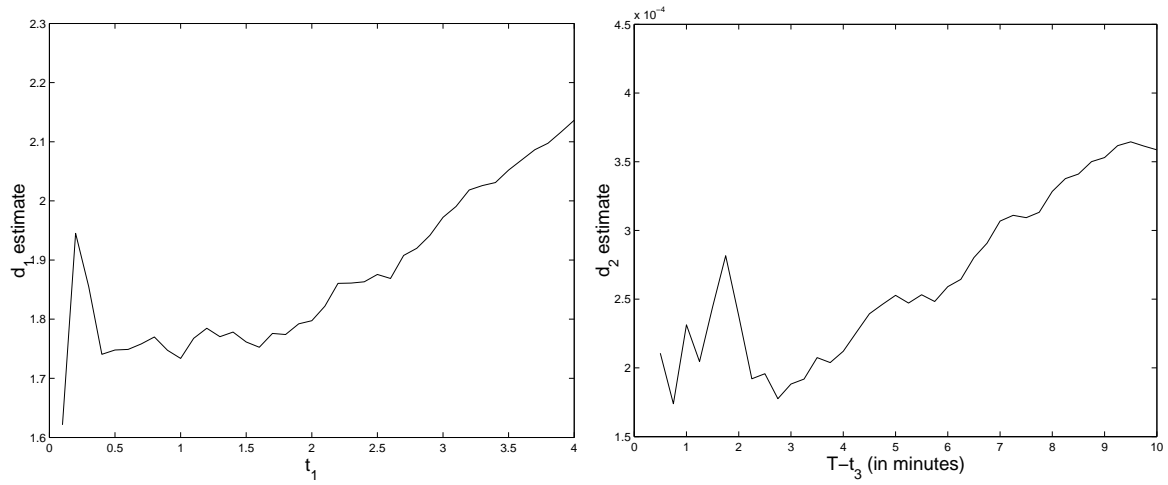


Figure 8: Plots of  $\hat{d}_1$  vs.  $t_1$  (left) and  $\hat{d}_2$  vs. initial values of  $T - t_3$  (right) for Palm data. The estimate for  $d_1$  seems stable at  $\approx 1.75$ .  $\hat{d}_2$  is approximately 2 minutes.

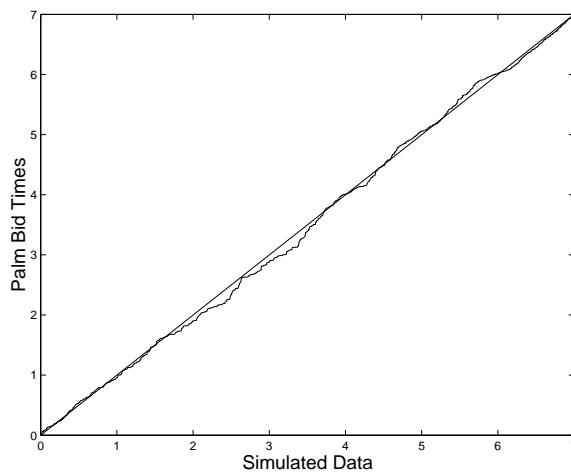


Figure 9: Q-Q plot of Palm bid times vs. simulated data from a BARISTA process with parameters  $\alpha_1 = 4.9, \alpha_2 = 0.37, \alpha_3 = 1.13, d_1 = 1.7, d_2 = 2/10080$ .

*macchiato* stage continues until the last 2 minutes, and then the third *ristretto* stage kicks in. The bid arrivals in each of the three stages 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 stage. 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 stage of bidding seems to take place within the last 2-3 minutes compared to the last 1 minute in Roth & Ockenfels (2000). 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  $\alpha$  parameters can take, and by the ability to add more stages.

### 5.1. Understanding Self Similarity

Why do the bid/bidder arrival processes in online auctions manifest self similarity? In general, power law distributions normally signal that a system is self-organizing and thus resistant to changes (Rodgers et al. 2003). Since this property is a major feature of network traffic, understanding the reasons there might shed light on the online auction environment. Although the causes underlying self similarity of network traffic have not been clearly identified (Crovella & Bestavros 1995; Roth & Ockenfels 2000), there have been several attempts to explain it. According to Mandelbrott (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 who 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).

## 5.2. The BARISTA as the basis for new research

The ability to model the bid arrival process via the BARISTA process can lead to new theoretical results as well as practical enhancements. First, it enables the exploration and formalization of observed phenomena like early and late bidding. The estimated intensity and changepoint parameters allow a quantification of such effects and their comparison across auctions of various types (e.g., the degree of sniping in hard-closing vs. “going-going-gone” auctions).

Second, the BARISTA model describes the aggregate bidding behavior of all the bidders in the auctions. The next step is to study the relationship between this aggregate process and individual 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 the pure self-similar process  $NHPP_1$  (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 third use of a bid arrival model is in conjunction with the sequence of bid or price increments. Jank & Shmueli (2005) explore the price dynamics in online auctions by representing the price curve during an

auction by a smoothed version of the bid history using smoothing splines (a type of piecewise polynomials). This smoothing step requires a specification of knots, which are the connecting points of the polynomials, and that should be determined by the bidding intensity, or the intensity of the bid arrivals. A BARISTA model can be used 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. With an actual model for bid arrivals there is now better ground for exploring the bivariate structure of bid timing and amount. One option is to find a function relating a particular BARISTA process with an average final price.

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

Finally, the bid arrival model can be useful for visualization tools that display the bids throughout an auction or a set of auctions (e.g., Shmueli and Jank 2005). In order to determine the scale of the time axis and to avoid over- and under-crowded areas on the display, the application must know “where the action is” and to what degree. An estimated BARISTA model, even if approximate, gives a sense of the scale of interest.

## A. ML Estimation of the Unconditional BARISTA Model

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  $\lambda$  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  $\theta$  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  $(c, \theta)$  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  $\theta$  is the solution of the equations

$$0 = \frac{\partial \mathcal{L}(c, \theta)}{\partial c} = -h(\theta) + \frac{n}{c} \tag{A.1}$$

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

Solving the first equation in (A.1) 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  $\theta$ . That is, if  $\hat{\theta}_j = w_j(X_1, \dots, X_n)$  is the *MLE* of  $\theta_j$  ( $1 \leq j \leq k$ ) based on a random sample of non-random size  $n$  from the distribution with the *pdf* above, then the *MLE* of  $\theta_j$  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 (A.1), the MLE of  $c$  is

$$\hat{c} = \frac{N(T)}{h(\hat{\theta})} \tag{A.2}$$

## B. Second derivatives of the log-likelihood function

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

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

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

$$\frac{\partial^2 \mathcal{L}}{\partial^2 \alpha_3} = -\frac{n}{C^2} \left( \frac{\partial C}{\partial \alpha_3} \right)^2 + \frac{n}{C} \frac{\partial^2 C}{\partial^2 \alpha_3} = \frac{n}{C^2} \left( \frac{\partial C}{\partial \alpha_3} \right)^2 - \frac{n\alpha_3}{2C} \frac{\partial C}{\partial \alpha_3}\quad (\text{B.3})$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_1 \alpha_2} = \frac{2}{C} \frac{\partial C}{\partial \alpha_1} \frac{\partial C}{\partial \alpha_2} + \log\left(1 - \frac{d_1}{T}\right) \frac{\partial C}{\partial \alpha_1}\quad (\text{B.4})$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_1 \alpha_3} = \frac{2}{C} \frac{\partial C}{\partial \alpha_1} \frac{\partial C}{\partial \alpha_3}\quad (\text{B.5})$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_2 \alpha_3} = \frac{2}{C} \frac{\partial C}{\partial \alpha_2} \frac{\partial C}{\partial \alpha_3} + \log\left(\frac{d_2}{T}\right) \frac{\partial C}{\partial \alpha_3}\quad (\text{B.6})$$

## References

- Ariely, D., A. Ockenfels, and A. E. Roth. 2003. An Experimental Analysis of Ending Rules in Internet Auctions. *RAND Journal of Economics*, forthcoming. (also, UCLA Department of Economics working paper).
- Aurell, E. and J. Hemmingsson. 1997. Bid frequency analysis in liquid Markets. *TRITA-PDC Report*. ISRN KTH/PDC/R-97/3-SE. ISSN 1401-2731. (<http://www.pdc.kth.se/payam/pub/AurellHemmingsson970208.ps>)
- Avery, C. N., C. Jolls, R. A. Posner, and A. E. Roth. 2001. The Market for Federal Judicial Law Clerks. *Univ. Chicago Law Review* **68** 793902.
- Bajari, P. and A. Hortacsu. 2000. Winner's Curse, Reserve Price and Endogenous Entry: Empirical Insights from eBay Auctions. *Working paper* Department of Economics, Stanford University.
- Bapna R., P. Goes, and A. Gupta. 2003. Analysis and Design of Business-to-Consumer Online Auctions. *Management Science* **49**(1) 85-101.
- Borle S., P. Boatwright, and J. B. Kadane. 2005. The Timing of Bid Placement and Extent of Multiple Bidding: An Empirical Investigation Using eBay Online Auctions. *Working paper*, Rice University.
- Bruschi, D., G. Poletti, and E. Rosti. 2002. E-vote and PKI's: a need, a bliss or a curse? in *Secure Electronic Voting*, D. Gritzalis ed., Kluwer Academic Publishers, ISBN 1-4020-7301-1.
- Bright, L., A. Gal and L. Raschid. 2004. Adaptive Pull-Based Data Freshness Policies for Diverse Update Patterns. *Technical Report*, UMIACSTR-2004-01, University of Maryland.
- Crosan, R. T. A. 1996. Partners and strangers revisited. *Economics Letters* **53** 25-32.
- Crovella, M. E. and A. Bestavros. 1995. Explaining World Wide Web Traffic Self-Similarity. *Technical Report TR-95-015*, Computer Science Department Boston University.
- Dennis, J. E. and R. B. Schnabel. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Englewood Cliffs, NJ: Prentice Hall.
- Efron, B. and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. Number 57 in Monographs on statistics and applied probability. Chapman & Hall, New York.
- Etzion, H., E. Pinker, and A. Seidmann. 2003. Analyzing the Simultaneous Use of Auctions and Posted Prices for On-line Selling. *Working Paper CIS-03-01*, Simon School, University of Rochester.

- Gal, A. and J. Eckstein. 2001. Managing Periodically Updated Data in Relational Databases: A Stochastic Modeling Approach. *Journal of the ACM* **48**(6) 1141-1183.
- Haubl, G. and P. T. L. Popkowski Leszczyc. 2003. Minimum Prices and Product Valuations in Auctions. *Marketing Science Institute Reports*, Issue 3, No. 03-117, 115-141.
- Huberman, B. A., and L. A. Adamic. 1999. Growth Dynamics of the WorldWide Web. *Nature* **401** 131.
- Jank, W. and G. Shmueli. 2005. Profiling Price Dynamics in Online Auctions Using Curve Clustering. *Smith School of Business working paper*, available at <http://www.smith.umd.edu/ceme/statistics/AuctionProfiling-Rev1-Jank&Shmueli.pdf>.
- Kauffman, R. J., and C. A. Wood. 2000. Running Up the Bid: Modeling Seller Opportunism in Internet Auctions. *Proceedings of the 2000 Americas Conference on Information Systems*.
- Ku, G., D. Malhorta, and J. D. Murnighan. 2004. Competitive Arousal in Live and Internet Auctions. *Working paper*, Northwestern University.
- Liebovitch, L. S. and I. B. Schwartz. 2003. Information flow dynamics and timing patterns in the arrival of email viruses. *PHYSICAL REVIEW E* **68** 017101-1 - 017101-4.
- Mandelbrot, B. B. 1969. Long-run linearity, locally Gaussian processes, H-spectra and infinite variances. *Intern. Econom. Review* **10** 82-113.
- McAfee, R. P., and J. McMillan. 1987. Auctions with stochastic number of bidders. *Journal of Economic Theory* **43** 119.
- Pinker, E., A. Seidmann, and Y. Vakrat. 2003. The design of Online Auctions: Business Issues and Current Research. *Management Science* **49** (11) 1457-1484.
- Rodgers, J. G., Y. J. Yap, and T. P. Young. 2003. Simple Models of Waiting Lists. *Advances in Complex Systems* **6** 215.
- Roth A.E., J. K. Murnighan, and F. Schoumaker. 1998. The Deadline Effect in Bargaining: Some Experimental Evidence. *American Economic Review*, **78**(4) 806-823.
- Roth A. E., and A. Ockefels. 2000. Last Minute Bidding and The Rules for Ending Second-Price Auctions: Theory and Evidence from a Natural Experiment on the Internet. *NBER Working Paper* #7729.
- Roth, A.E. and X. Xing. 1994. Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions. *American Economic Review* **84** 992-1044.

- Shmueli, G., R. P. Russo, and W. Jank. 2004. Modeling Bid Arrivals in Online Auctions. *Working paper*, Smith School of Business, University of Maryland.
- Shmueli, G. and W. Jank. 2005. Visualizing Online Auctions. *Journal of Computational and Graphical Statistics*. Forthcoming.
- Vakrat Y. and A. Seidmann. 2000. Implications of the Bidders Arrival Process on the Design of Online Auctions. *Proceedings of the 33rd Hawaii International Conference on System Sciences* 1-10.
- Vose, M. D. 1999. *The simple Genetic Algorithm*. MIT Press.
- Wilcox, R. T. 2000. Experts and Amateurs: The Role of Experience in Internet Auctions. *Marketing Letters* **11** 363-374.
- Willinger, W., M. S. Taqqu, W. E. Leland, and D. V. Wilson. 1995. Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements. *Statistical Science* **10**(1) 67-85.
- Yang, I., H. Jeong, B. Kahng, and A.-L. Barabasi. 2003. Emerging behavior in electronic bidding. *PHYSICAL REVIEW E* **68** 016102.
- Zhang, A., D. Beyer, J. Ward, T. Liu, A. Karp, K. Guler, S. Jain, and H. K. Tang. 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>)