TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2008

 | ISBN 978-1-877640-23-0
DOI 10.1287/educ.1080.0042

# Statistical Challenges in eCommerce: Modeling Dynamic and Networked Data

***Wolfgang Jank, Galit Shmueli***
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742
{wjank@rhsmith.umd.edu, gshmueli@rhsmith.umd.edu}

***Mayukh Dass***
Rawls College of Business, Texas Tech University, Lubbock, Texas 79409, mayukh.dass@ttu.edu

***Inbal Yahav***
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742,
iyahav@rhsmith.umd.edu

***Shu Zhang***
Department of Mathematics, University of Maryland, College Park, Maryland 20742,
zhangshu@math.umd.edu

**Abstract** Empirical research in the field of electronic commerce (eCommerce) has been growing fast due to the availability of rich, high-quality data. eCommerce data originate from many different behavioral, social, or economic processes and interactions online that have not been observable and measurable in the offline world. This data-rich environment allows for the questioning of existing theories and the uncovering of new phenomena. However, eCommerce data and the new research questions associated with these data are often not supported by classic statistical machinery. New dependency structures arise due to factors such as online competition and user interaction. In this tutorial, we discuss three key aspects of eCommerce data: eCommerce process dynamics, competition between processes, and user networks. Each data structure raises new challenges for data representation, visualization, and modeling, and we describe each of them in detail. We also present three case studies that showcase the various statistical challenges and present some solutions.

**Keywords** online auctions; electronic commerce; dynamics; competition; networks; loyalty; functional data analysis; social network analysis; nonparametric methods; smoothing; forecasting

## 1. Introduction

Electronic commerce (eCommerce) has received an extreme surge of popularity in recent years. By eCommerce we mean any form of transaction using the Internet such as buying or selling goods, or exchanging information related to goods. eCommerce has had a huge impact on the way we live today compared with a decade ago: It has transformed the economy, eliminated borders, opened doors to innovations that were unthinkable just a few years ago, and created new ways in which consumers and businesses interact. Although many predicted the death of eCommerce with the "burst of the Internet bubble" in the late 1990s, eCommerce is thriving more than ever.

There are many examples of eCommerce transactions. They include buying, selling, or investing online; shopping on electronic marketplaces like Amazon.com or online auctions like eBay.com; Internet advertising (e.g., sponsored ads by Google, Yahoo!, and Microsoft); recording clickstream data and cookie tracking; transactions on e-bookstores and e-grocers;

Web-based reservation systems and ticket purchasing; marketing emails and message postings on Web logs; posting and monitoring downloads of music, video, and other online content; postings on user groups and other electronic communities; online discussion boards and learning facilities; auctioneering open source projects; and many, many more.

The public footprint of many Internet transactions has opened new opportunities for empirical researchers to study the behaviors of individuals, companies, organizations, and societies. Theoretical results, founded in economics and derived for the offline, brick-and-mortar world, have often proven not to hold in the online environment. Possible reasons are the worldwide reach of the Internet, anonymity of its users, virtually unlimited resources, constant availability, and continuous change. For this reason, and also due to the availability of massive amounts of publicly available high-quality Web data, empirical research is thriving.

Empirical eCommerce research covers many topics, ranging from very specific questions such as the impact of online markets for used goods on sales of CDs and DVDs (Telang and Smith [51]), the evolution of open source software (Stewart et al. [49]), or the optimality of online price dispersion in the software industry (Ghose and Sundararajan [18]), to more broad questions such as issues of privacy and confidentiality in eCommerce transactions (Fienberg [12]), how online experiences advance our understanding of the offline world (Forman and Goldfarb [13]), the economic impact of user-generated online content (Ghose [17]), and challenges in collecting, validating, and analyzing large-scale eCommerce data (Bapna et al. [5]).

Certain areas of eCommerce have attracted an especially large body of empirical research. One such area is that of online auctions. In one of the earliest examinations of online auctions (Lucking-Reiley [33]), empirical economists found that bidding behavior, particularly on eBay, often diverges significantly from what classical auction theory postulates. Since then, there has been an enormous surge in empirical analysis of online auction data in the fields of information systems, marketing, computer science, statistics, and others. Studies have examined bidding behavior in the new online environment from multiple different angles: identification and quantification of new bidding phenomena, such as sniping (Roth and Ockenfels [42]), shilling (Kauffman and Wood [31]), and price dynamics (Bapna et al. [3]); creation of a taxonomy of bidder types (Bapna et al. [6]); development of descriptive probabilistic models to capture bid frequency distributions (Shmueli et al. [45]), price dynamics during an auction (Wang et al. [54], Bapna et al. [3], Hyde et al. [23]), and bidder behavior in terms of bid timing and amount (Borle et al. [8], Park and Bradlow [36]); and development of novel models for dynamically forecasting auction prices (Wang et al. [53]) and quantifying economic value such as consumer surplus in eBay (Bapna et al. [4]); and more recently, online auction data are being used for studying bidder and seller networks and relationships (Yao and Mela [56], Dass and Reddy [10]), and competition between auctions (Hyde et al. [22], Jank and Shmueli [27]).

Internet advertising is another area where empirical research is growing, but currently more within the commercial world and less within academia. Companies such as Google, Yahoo!, and Microsoft study the behavior of online advertisers using massive data sets of bids and their results in order to more efficiently allocate inventory (e.g., ad placement) (Agarwal [1]). Online advertisers and companies that provide services to advertisers also use bid data. They study relationships between bidding and profit (or other measures of success) for the purpose of optimizing advertisers' bidding strategies (Matas and Schamroth [34]).

Another active and growing area of empirical research is that of prediction markets. Prediction markets, also known as information markets, idea markets, event futures, or betting exchanges, are increasingly used to aggregate the *wisdom of crowds* (Surowiecki [50]) from online communities to forecast the outcomes of events that are of interests to the public. Prediction markets have many interesting applications, e.g., in forecasting economic trends (e.g., HedgeStreet), natural disasters (the Hurricane Futures Market at University of Miami),

outcomes of political campaigns (e.g., the Iowa Electronic Markets (IEM)), sporting events (e.g., TradeSports), and Oscars and movie box offices (e.g., the Hollywood Stock Exchange). For example, since its establishment in 1988, the IEM has predicted the U.S. presidential elections more accurately than traditional polls 75% of the time. Prediction markets have also been used by an increasing number of major corporations, such as Hewlett-Packard, Intel, Microsoft, Google, Yahoo!, General Electric, Corning, Eli Lilly, and Goldman Sachs, to tap internal future-focused knowledge about sales, supplier behavior, project completion time, and new product release timing. Several empirical studies (Spann and Skiera [48], Forsythe et al. [14], Pennock et al. [37]) report on the accuracy of final trading prices to provide accurate forecasts. In more recent work, the effect of the trading shape and its dynamics is also documented using novel statistical approaches (Foutz and Jank [15]).

The availability of new eCommerce data sources also comes with new data challenges. Some of these challenges are related to data volume, whereas others reflect the new structure of eCommerce data. Both issues pose serious challenges for the empirical researcher. Here, we focus on the new data structure found in eCommerce. More specifically, eCommerce data often arrive as a *combination of temporal and cross-sectional* data. Consider data from online auctions (e.g., eBay) as an example. Online auctions feature two fundamentally different types of data, the bid history and the auction description. The bid history lists the sequence of bids placed over time and as such can be considered temporal information. In contrast, the auction description (e.g., product information, information about the seller and the auction format) does not change over the course of the auction and therefore is cross-sectional information. The analysis of combined temporal and cross-sectional data poses challenges because most statistical methods are geared towards only one type of data. Moreover, Web-based temporal data that are user generated are nonstandard time series where events are not equally spaced. In that sense, such temporal information is better described as processes. Moreover, many eCommerce *processes exhibit dynamics* that change over the course of their duration. On eBay, for instance, prices speed up early and slow down, only to speed up again towards the auction end. Classical statistical methods are not geared towards capturing the change in process dynamics and to teasing out similarities (and differences) across thousands (or even millions) of eCommerce processes.

Another challenge related to the process nature of the data is capturing *competition among eCommerce processes*. Consider again the example of eBay auctions. On any given day, there exist tens of thousands of same (or similar) products being auctioned that compete for the same bidders. For instance, a simple search under the keywords "Apple iPod" reveals over 10,000 available auctions, all of which vie for the attention of the interested bidder. Although not all of these 10,000 auctions may sell the identical product, some may be more similar (in terms of product characteristics) than others. Moreover, even among identical products, not all auctions will be equally attractive to the bidder due to differences in sellers' perceived trustworthiness or differences in auction format. For instance, to those bidders that seek immediate satisfaction, auctions that are five days away from completion may be less attractive than auctions that end in the next five minutes. Modeling differences in product similarity and their impact on bidders' choices is challenging (Jank and Shmueli [27]). Similarly, understanding the effect of misaligned (different starting times, different ending times, different durations) auctions on bidding decisions is equally challenging (Hyde et al. [22]), and solutions are not readily available in classical statistical tools. For a more general overview of challenges associated with auction competition, see Haruvy et al. [19].

Another level of dependence in eCommerce data that further challenges statistical modeling is the existence of *user networks* and their impact on transaction outcomes. Networks have become an increasingly important component of the online world, particularly in the "new Web," Web 2.0, and its network-fostering enterprizes such as Facebook.com,

MySpace.com, LinkedIn.com, etc. Networks also exist in other places (although less obviously) and impact transaction outcomes. On eBay, for example, buyers and sellers form networks by repeatedly transacting with one another. This raises the question about the mobility and characteristics of networks across different marketplaces and their impact on the outcome of eCommerce transactions. Answers to these questions are not obvious and require new methodological tools to characterize networks and capture their impact on the online marketplace.

In what follows, we elaborate on each of these three challenging data structures (process dynamics, process competition, and user networks) in more detail. For each, we present a small case study that illustrates the challenges and suggest some solutions. Our research (Jank and Shmueli [25]) has shown that functional data models are particularly useful for addressing challenges of combined temporal and cross-sectional eCommerce data. We thus start by giving a brief introduction to functional data models.

## 2. Functional Data Models

The technological advancements in measurement, collection, and storage of data have led to more and more complex data structures. Examples include measurements of individuals' behavior over time, digitized two- or three-dimensional images of the brain, and recordings of three- or even four-dimensional movements of objects traveling through space and time. Such data, although recorded in discrete fashion, can be thought of as continuous objects represented by functional relationships. This gives rise to the field of functional data analysis (FDA) where the center of interest is a set of curves, shapes, objects, or, more generally, a set of *functional observations*. This is in contrast to classical statistics where the interest centers around a set of data vectors. In that sense, functional data are not only different from the data structure studied in classical statistics, but actually generalize it.

### 2.1. The Price Curve and Its Dynamics

A functional data set consists of a collection of continuous functional objects. Despite their continuous nature, limitations in human perception and measurement capabilities allow us to observe these curves only at discrete time points. Thus, the first step in functional data analysis is to recover, from the observed data, the underlying continuous functional object (Ramsay and Silverman [39]). This is usually done with the help of data smoothing.

A variety of different smoothing methods exist. One very flexible and computational efficient choice is the penalized smoothing spline. Let $\tau_1, \ldots, \tau_L$ be a set of knots. Then, a polynomial spline of order $p$ is given by

$$f(t) = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + \sum_{l=1}^{L} \beta_{pl}(t - \tau_l)_+^p, \tag{1}$$

where $u_+ = uI_{[u \geq 0]}$ denotes the positive part of the function $u$. Define the roughness penalty

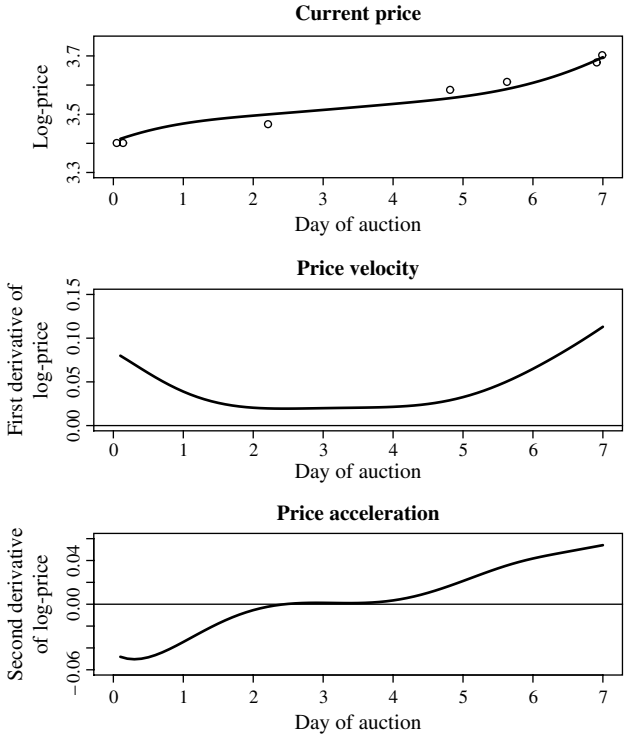
$$\mathrm{PEN}_m(t) = \int \{D^m f(t)\}^2 \, dt, \tag{2}$$

where $D^m f$, $m = 1, 2, 3, \ldots$, denotes the $m$th derivative of the function $f$. The penalized smoothing spline $f$ minimizes the penalized squared error

$$\mathrm{PENSS}_{\lambda, m} = \int \{y(t) - f(t)\}^2 \, dt + \lambda \mathrm{PEN}_m(t), \tag{3}$$

where $y(t)$ denotes the observed data at time $t$ and the smoothing parameter $\lambda$ controls the trade-off between data fit and smoothness of the function $f$. Using $m = 2$ in (3) leads to the commonly encountered cubic smoothing spline.

FIGURE 1. Auction dynamics.




*Notes.* Current price, price velocity (first derivative), and price acceleration (second derivative) for a selected auction are shown. The first graph shows the actual bids together with the fitted curve.

The process of going from observed data to functional data is now as follows. For a set of $n$ functional objects, let $t_{ij}$ denote the time of the $j$th observation $(1 \le j \le n_i)$ on the $i$th object $(1 \le i \le n)$, and let $y_{ij} = y(t_{ij})$ denote the corresponding measurements. Let $f_i(t)$ denote the penalized smoothing spline fitted to the observations $y_{i1}, \ldots, y_{in_i}$. Then, functional data analysis is performed on the continuous curves $f_i(t)$ rather than on the noisy observations $y_{i1}, \ldots, y_{in_i}$. For ease of notation we will suppress the subscript $i$ and write $y_t = f(t)$ for the functional object and $D^{(m)}y_t = f^{(m)}(t)$ for its $m$th derivative.

Consider Figure 1 for illustration. The *circles* in the top panel of Figure 1 correspond to a scatterplot of bids placed in an auction versus their timing. The continuous curve in the top panel shows a smoothing spline of order $m = 4$ using a smoothing parameter $\lambda = 50$.

One of our modeling goals is to capture the *dynamics* of an auction process. Although $y_t$ describes the *magnitude* of the current price, it does not reveal the dynamics of how fast the price is *changing* or *moving*. Attributes that we typically associate with a moving object are its *velocity* (or its *speed*) as well as its *acceleration*. Notice that we can compute the price velocity and price acceleration via the first and second derivatives, $D^{(1)}y_t$ and $D^{(2)}y_t$, respectively.

Consider again Figure 1. The middle panel corresponds to the price velocity, $D^{(1)}y_t$. Similarly, the bottom panel shows the price acceleration, $D^{(2)}y_t$. The price velocity has several interesting features. It starts out at a relatively high mark, which is due to the starting price that the first bid has to overcome. After the initial high speed, the price increase slows down over the next several days, reaching a value close to zero midway through the auction. A close-to-zero price velocity means that the price increase is extremely slow. In fact, there are no bids between the beginning of day 2 and the end of day 4, and the price velocity reflects that. This is in stark contrast to the price increase on the last day where the price velocity picks up pace, and the price jumps up.

The bottom panel in Figure 1 represents price acceleration. Acceleration is an important indicator of dynamics since a change in velocity is preceded by a change in acceleration. In other words, a positive acceleration *today* will result in an increase of velocity *tomorrow*. Conversely, a decrease in velocity must be preceded by a negative acceleration (or *deceleration*). The bottom panel in Figure 1 shows that the price acceleration is increasing over the entire auction duration. This implies that the auction is constantly experiencing forces that change its price velocity. The price acceleration is flat during the middle of the auction where no bids are placed. With every new bid, the auction experiences new forces. The magnitude of the force depends on the size of the price increment. Smaller price increments will result in a smaller force. On the other hand, a large number of small consecutive price increments will result in a large force. For instance, the last two bids in Figure 1 arrive during the final moments of the auction. Because the increments are relatively small, the price acceleration is only moderate. A more systematic investigation of auction dynamics has been done in other places (Jank and Shmueli [28], Bapna et al. [3]).

## 3. Dynamics of eCommerce Processes

Many eCommerce processes are governed by changing dynamics. By dynamics we mean "change" and the rate at which this change occurs. In online auctions, for instance, price increases only slowly during most of the auction duration, only to speed up rapidly towards the end (Bapna et al. [3]). In online virtual stock markets, prices of assets sometimes speed up towards the end of the trading period, whereas othertimes they slow down (Foutz and Jank [15]). Dynamics capture the rate at which this change occurs. In that sense, dynamics describe the rate at which information cascades through all the levels and participants of eCommerce processes.

Dynamics form an important component of our understanding of eCommerce processes. Knowledge of dynamics can lead to more accurate process predictions (Foutz and Jank [15], Wang et al. [53]) and they can also lead to a better understanding of what happens "behind the scenes." In the eCommerce world, many important agent decisions and interactions remain unobserved. For instance, in online auctions, we do not know the true strategy of bidders, we do not know how they interact with other bidders, and we do not know how they change their strategies in the face of competition. In that sense, many behavioral aspects of the auction process are unobserved. What we do observe, though, is how different bidders' behavior results in changing auction dynamics. Thus, the ability to measure and model dynamics is a first step towards a better understanding of what drives eCommerce processes.

There are many different ways of measuring and modeling dynamics. A recent stream of research (Jank and Shmueli [25, 28, 29], Bapna et al. [3], Wang et al. [53, 54]) has focused on taking advantage of the very flexible *functional data models* introduced in §2. Functional data models are flexible because they do not impose parametric restrictions; they also allow for a unified treatment of temporal and cross-sectional data and result in a natural way of gauging process dynamics. One powerful approach borrows ideas from physics by fitting differential equation models to functional data. Wang et al. [54] uses a novel test for multiple comparison of functional differential equation models and finds that dynamics are quite different, even for auction processes that sell the identical item. Jank and Shmueli [29] develop a novel *functional differential equation tree* to segment dynamics based on characteristics of the auction process. Functional differential equation models are an area of research that currently receives a tremendous amount of interest (Ramsay et al. [38]).

The existence of dynamics poses many challenges in eCommerce research. One of these challenges is explaining their causes. In the following, we describe a case study in the context of simultaneous auctions. We find that dynamics are partially explained by heavy bidder competition.

### 3.1. Case Study I: Explaining Price Dynamics in Simultaneous Online Auctions

Recent studies on online auctions have concluded that price dynamics (such as rate of change in price or price velocity) play an important role in providing insightful information about auction characteristics (Bapna et al. [3], Reddy and Dass [41]) and in improving prediction accuracy for price forecasts (Wang et al. [53]). One of the questions that remains unanswered, however, is what exactly cause dynamics? In this section, we shed light on this question in the context of simultaneous online auctions (SOA) of contemporary Indian art. Particularly, we show that direct inclusion of bidder competition in price forecasting models deems price dynamics information redundant. This in turn suggests that dynamics capture bidder competition in auctions. To do so, we proceed in two steps. We first illustrate the superiority of forecasting models with price dynamics, similar to the work of Wang et al. [53]; then we show that this superiority vanishes when bidder competition is included directly into the forecaster.

Many prior studies on dynamics are in the context of eBay and eBay-like online auctions. Such online auctions sell items independently of each other, meaning that beginning and ending times of auctions are not related. Recently, another auction format, namely SOA, is gaining popularity in the fine arts and collectibles market. Unlike eBay, whereas auctions are held for three to seven days, SOA have a shorter duration (typically two to three days). They sell multiple items simultaneously in a first-price ascending auction format. This means that auctions start and end at the same time for all the items. Because most items sold are highly complementary, bidders are frequently observed to compete against one another across one or more items. This leads to two types of dyadic bidder competition, namely *within-auction* competition and *between-auction* competition (Dass et al. [11]). In contract, bidders in eBay auctions rarely interact with each other in more than one auction at a time as they seldom have demand for similar products at the same time. Furthermore, eBay has recently changed their policy to mask the bidders' identities. Thus, even though some bidders may compete across multiple auctions, it is impossible for them to identify each other and thus realize between-auction competition. Furthermore, unlike eBay auctions, SOAs have a soft closing time where the auction automatically extends in case of a late bid. This auction design not only encourages bidders to bid early (Roth and Ockenfels [42]) but also discourages sniping in the last moments. Finally, SOAs are organized by only one seller, i.e., the auction house (similar to the Christie's and Sotheby's auctions), whereas eBay provides an auction platform for many different sellers.

We start our investigation by developing two dynamic models (DM-I and DM-II) based on functional data analysis to predict the price of an auction "in progress." The fundamental difference between these two models is that DM-I incorporates both the price and price velocity, whereas DM-II incorporates only price. Both these models also include static preauction information such as auction characteristics (opening bid), item characteristics (size of the item, type of artwork), and artist characteristics (artist reputation, previous auction history). We compare the prediction accuracy with an entirely static model (ST), which is based on the above static information and does not have any dynamic components. Another benchmark model is a simple dynamic model (DM-0) that includes, in addition to the static information, the price at the time of prediction.

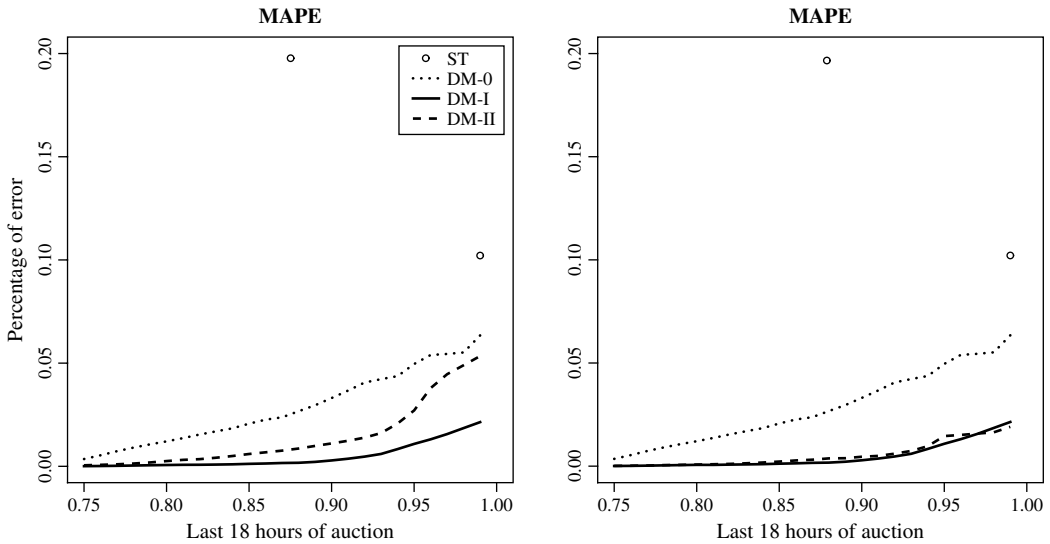
We build DM-I and DM-II using the framework of Wang et al. [53]. We first split the data set into two parts, a training set (70% of the auctions) and validation set (30% of the auctions). The modeling approach consists of four components. The first component recovers the underlying price curve and its dynamics from the observed data. Because bids arrive at unevenly spaced time intervals, we need the flexibility of FDA to obtain the underlying price curve and its first-order derivative (i.e., the rate of change in price or price velocity). This is done by interpolating the raw data and sampling at a common set of time points to account for the irregular spacing of the bid arrivals. Then, the underlying price curves

are recovered using penalized monotone curves (Ramsay and Silverman [39], Simonoff [46]). The second component fits a polynomial-trended linear regression model with autoregressive residuals to the price dynamics along with time-varying predictors. These two components are performed on the training set. The third component forecasts the price dynamics from the estimated model in the second component. Finally, in the fourth component, we forecast the actual price with DM-I and DM-II using the forecasted price dynamics from the third component together with other time-varying and time-invariant predictors. The last two components are performed on the validation set only. Note that the training set consists of auctions that are fully observed between the start and end; in contrast, auctions in the validation set are only partially observed; i.e., information is only available until time $T$, the time at which a forecast is desired. $T$ is flexible and can be set by the user. For more details, please see Wang et al. [53].

To compare the accuracy of all forecasters, we compute their mean absolute percentage error (MAPE) on the holdout sample (left panel in Figure 2). The results show that both dynamic models, DM-I and DM-II, outperform the static model ST and the simple dynamic model DM-0. Furthermore, in lines with prior findings, DM-I is also superior to DM-II, once again supporting the importance of price dynamics in the prediction process. (MAPE: DM-I < DM-II < DM-0 < ST).

Next, we explore the cause of price dynamics. Although no prior research has investigated this question in detail, some studies (Ariely and Simonson [2], Heyman et al. [20]) hint at the bidder rivalry as a possible answer. Based on this assertion, we quantify rivalry at the dyadic bidder level to measure within-auction and between-auction competition and incorporate them into our dynamic model to generate new DM-I and DM-II. Within-auction competition captures the rivalry intensity between two specific bidders in an auction. Operationally, it is computed as the maximum number of repeated sequential bids between two specific bidders in an auction. For every auction, we first determine the unique pairs of bidders participating in it. Then for each of these bidder pairs, we count the number of times the two bidders bid sequentially (i.e., $A > B > A$). The maximum number of such bids among all the bidder pair in an auction denotes the within-auction competition. Between-auction competition measures the spread of the rivalry between bidders across multiple auctions and, thus, measures the competitive reach of a bidder pair across several auctions. Like the

FIGURE 2. Forecasting accuracy.




*Notes.* Left panel: Models DM-I, DM-II, ST, and DM-0 without bidder competition components; right panel: same models with the bidder competition.

previous measure, we first determine the number of unique bidder pairs. Then, for all the bidder pairs, we count the number of auctions in which the pair is competing simultaneously. The between-auction competition for one auction is the average value across all pairs.

We now incorporate these two competition components into DM-I and DM-II and again forecast price on the validation set. The right panel in Figure 2 shows the predictive performance of these new models; we can see that the difference between DM-I and DM-II has vanished. This suggests that given direct competition information, the dynamic component loses its predictive usefulness. This also suggests that bidder competition may be a major source of the dynamics. It is further noted that the dynamic models also outperform the static models in this set of analyzes.

## 4. Competition Between eCommerce Processes

In this section, we focus on competition between eCommerce processes. An example is competition between simultaneous auctions that compete for the same bidders (to achieve the highest price); competition could also occur between sellers selling the same (or similar) products, or between auction platforms (e.g., eBay versus Yahoo! auctions). Competition can also occur in other eCommerce processes. For instance, virtual stock markets (e.g., www.hsx.com) trade stock for Hollywood movies. If two movies are scheduled to be released around the same time, then the amount of competition between the two movies (if any) could be reflected in the trading paths for the individual movie stocks.

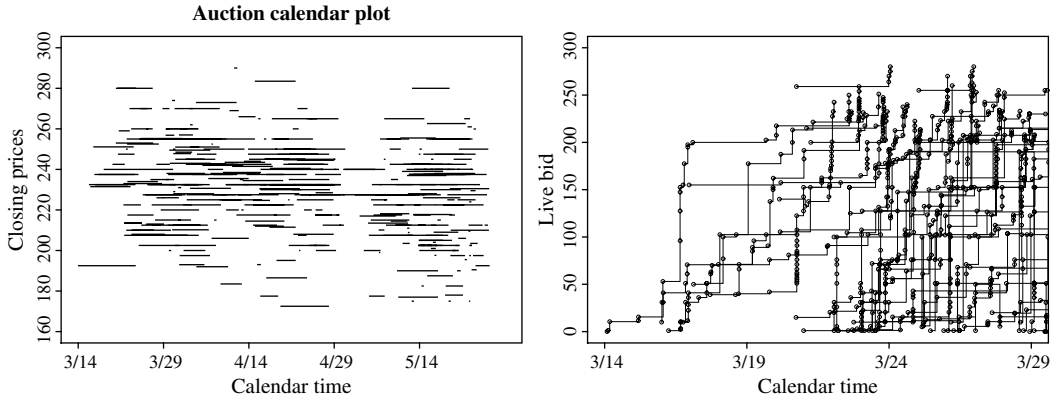
There are many other examples of competition between eCommerce processes. Statistically, it is hard to model competition for several reasons. First, finding a suitable measure for competition is not easy. One idea is to use the concept of correlation: the stronger the correlation between two processes, the more they compete with one another. However, it is very hard to define a statistical correlation measure for continuous processes, in part because eCommerce processes are not perfectly aligned and often only partially overlap. Online auctions are a point in case: when one auction ends, another one is about to start. Second, many processes are not perfectly identical (e.g., auctions for Microsoft Xbox gaming stations versus Sony Playstations). Dealing with varying degrees of similarity requires new statistical concepts for modeling (Jank and Shmueli [27]) or visualizing (Hyde et al. [22]) such data. In the following, we describe a case study in the context of eBay's online auctions. The study shows that the information from concurrent auctions matters (i.e., it matters in terms of the final price), and that by incorporating such information into a forecasting model one can achieve automated bidding systems that outperform classical bidding strategies (such as early bidding or last-moment sniping) in terms of efficiency and realized consumer surplus.

### 4.1. Case Study II: Competition Between Online Auctions

The left panel in Figure 3 shows a *calendar plot* (Jank and Shmueli [24]) of eBay auctions for a particular item (new Palm M515 handheld devices) over a certain period of time (March 14 until May 31). We can see that, at any point in time, many auctions are taking place. Because all auctions sell the identical item, bidders must decide between many competing auctions. What makes the bidder's decision crucial for his or her bottom line is that prices vary significantly (between \$170 and \$280, in this case). Thus, a poor decision on the bidder's part can lead to a large loss in consumer surplus.

Making informed bidding decisions in the face of auction competition is extremely hard due to information overload. Consider again Figure 3 (right panel). It shows a snapshot of the live price processes of competing auctions and represents the information that bidders see while bidding. At any point in time, bidders see many simultaneous auctions. Each of these auctions carries different information about the auction outcome (i.e., the final price); auctions that have only just begun carry less information about the final price compared

FIGURE 3. Information overload.



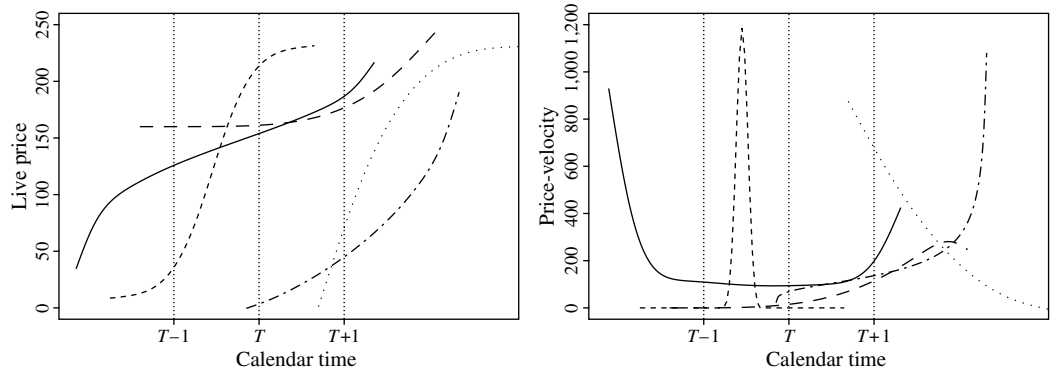

*Notes.* Left panel: Calendar plot of closing prices versus calendar time. Each line in the graph represents an auction, the length of the line corresponds to the length of the auction; the $y$-axis denotes the closing price, and the $x$-axis denotes calendar time. Right panel: A snapshot of the live price process of competing products during eBay auctions. The circles denote the arrival of new bids. The horizontal lines mark the time periods during which the price in an auction remains unchanged.

to auctions that are almost closing. It is hard for bidders to process all of this information "manually" and to pick the auction that results in the lowest projected price.

We propose a data-driven solution to this problem by modeling auction competition statistically. However, statistical models for competition are challenging for a variety of reasons. The price processes in Figure 3 form time series that are strongly misaligned and that are also of different length; modeling the information across misaligned time series of different length is not straightforward. Moreover, each price process experiences changes in price dynamics: prices usually move rather slowly during the earlier and especially middle parts of the auction, only to speed up heavily towards the auction end. Classical time series models do not account for changes in price dynamics. And last, one of the modeling challenges is to incorporate competition, that is, the effect of what happens in other auctions. None of these challenges is easily handled in classical time series models.

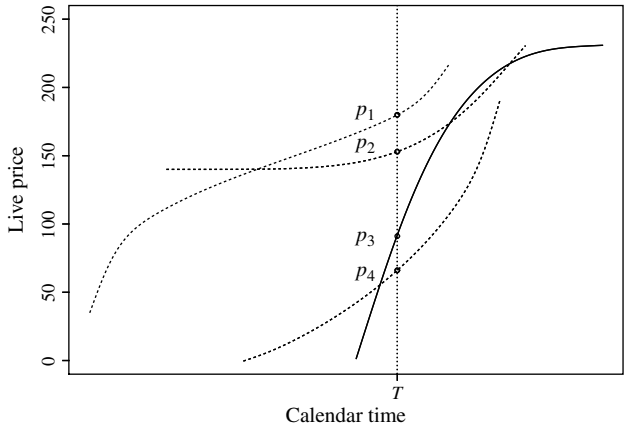
We propose an innovative approach based on functional data analysis (introduced in §2). In that approach, we use data smoothing to create smooth representations of the observed and noisy price processes. For an example of the resulting smooth price curves, consider the left panel of Figure 4. Notice that price curves are arranged over calendar time; thus, they are extremely misaligned. To cope with misaligned curves and curves of different length, we

FIGURE 4. Auction concurrency.




*Notes.* Smooth price paths (left panel) and corresponding velocities (right panel) are shown. The solid line corresponds to the focal auction; the broken line illustrate simultaneous, competing auctions.

FIGURE 5. Concurrency features.




*Notes.* The solid line is the focal auction and the broken lines correspond to simultaneous, competing auctions. $T$ corresponds to the time of decision making and $p_1$ through $p_4$ is the price information from all auctions available at that time.

evaluate each curve on a grid. Let $T$ be the time of a desired bidding decision. Then, our approach estimates a forecasting model from the information between time $T-1$ and $T$ and uses this model to predict price at time $T+1$. Moreover, because price curves are smooth, we can estimate price dynamics via their first or second derivatives. The right panel of Figure 4 shows the price velocities (i.e., first derivatives) for the price curves in the left panel.

One major component of our model is competition. That is, we want to capture the effect of what happens in other, simultaneous auctions. To that end, we must first define meaningful measures for competition. There are many different ways of defining competition measures and we explore several alternatives below. All measures are driven by the same general principle, which is illustrated in Figure 5. We define a focal auction (indicated by the solid line in Figure 5) as the auction for which a bidder wants to decide whether or not to bid on. At time $T$ of decision making, there are several other auctions that take place simultaneously (indicated by the dotted lines). One meaningful measure of competition is the level of price in other auctions. In our example, there are four different prices levels at time $T$, varying from high ($p_1$) to low ($p_4$). The price level in the focal auction at that time is $p_3$. Thus, a possible measure for the price competition is given by the *average price* in concurrent auctions (which we denote by *c.avg.price*), that is, by the average of $p_1$, $p_2$, and $p_4$. In a similar fashion, the *average price velocity* (*c.avg.vel*) in concurrent auctions would be the average of the corresponding price velocities, and so on. In this tutorial, we investigate several different competition features and their impact on the final price of the focal auction. (For a more comprehensive investigation, see Zhang and Jank [57].) The results are summarized in Table 1.

We observe several interesting effects in Table 1. First, *c.avg.price* has a positive effect, implying that if price is high in concurrent auctions (relative to the focal auction), then this

TABLE 1. Competition features and their price effect.

| Name | Explanation | Effect |
|---|---|---|
| *c.avg.price* | Average price of concurrent auctions | Positive |
| *c.avg.vel* | Average price velocity of concurrent auctions | Negative |
| *c.avg.acc* | Average price acceleration of concurrent auctions | Negative |
| *c.avg.t.left* | Average time left of concurrent auctions | Negative |

*Notes.* The first two columns show the name and a short explanation of the competition feature. The last column shows the effect on the final price of the focal auction.

has a positive effect on the price of the focal auction. In other words, high price levels in other, competing auctions will make bidders search for different auctions (with a lower price) and subsequently drive up the price in those auctions. The effects of the concurrent dynamics (*c.avg.vel* and *c.avg.acc*) are negative. This may indicate information cascading or herding: if the "buzz" for other auctions increases, then this will hurt the price of the focal auction. Similarly, the time that is left in other auctions (*c.avg.t.left*) negatively influences price in the focal auction, because bidders have more time to make informed decisions outside of the focal auction.

Using the competition features from Table 1 (together with additional information such as the auction format, information about the seller, the bidders, and the price dynamics), we build a forecasting model for price in competing auctions. Then, in the next step, we build a comprehensive bidding strategy around our forecasting model. The idea is based on maximizing consumer surplus, which refers to the difference between the bidders' willingness to pay and the price actually paid. We formulate an automated algorithm for selecting the optimal auction to bid on, and for determining the optimal bid time and amount. The optimal auction provides bidders with the highest surplus, and the optimal bid amount equals the predicted closing price. This strategy automates the entire decision-making process, and, in contrast to early or late bidding strategies, it frees the bidder of time constraints because bidding can occur immediately. (For more details, see Zhang and Jank [57].)

We conduct a simulation study to compare our automated bidding strategy to alternate bidding approaches such as *early bidding* (Bapna et al. [6]) or *last-moment sniping* (Roth and Ockenfels [42]). Early bidding is often used as a signal for the bidder's commitment and intends to deter other bidders. Last-moment sniping is a popular strategy because it does not allow enough time for other bidders to react. In our simulation, we assume that bidders' willingness to pay (WTP) is drawn from a uniform distribution that is symmetrically distributed around the market value. A bidder randomly draws from this distribution and then bids according to one of three strategies: Under early bidding, the bidder bids his WTP at the end of the first day of the auction.[1] Under last-moment sniping, the bidder waits until the last minute of the auction and then bids 1% over the current price. And finally, under our automated bidding strategy, the bidder bids the price predicted by our forecasting model. Note that in contrast to early bidding or last-moment sniping, the bidding decision in our strategy is free of time; i.e., the bidder bids *immediately* without having to wait for the end of the first auction day or for the last minute of the auction. Although this advantage may not mean much to bidders who use automated bidding agents such as www.cniper.com, not all eBay bidders use this technology. Moreover, such technology is only available for the eBay market and not for smaller, more specialized auction markets (such as those for Indian art mentioned on §3). We also want to point out that our simulations are set up such that, under each strategy, bidders will only bid if the current auction price is *less* than their WTP; otherwise, the bidder will move on to another auction, and so on.

The results are presented in Table 2. We find that, although snipers have the highest probability of winning, our automated and data-driven strategy results in a much higher surplus. It also results in considerably less time devoted to the process of bidding because no manual monitoring of the auction process is necessary. Moreover, early bidders have the lowest average winning probability and surplus. We also investigate the impact of the prediction window on the resulting surplus, and find that, as the width of the window increases, consumer surplus increases whereas the probability of winning decreases.

## 5. User Networks in eCommerce

Another challenge of eCommerce data is the existence of networks created by its users. For instance, customers that (repeatedly) buy from the same seller form a network with

[1] Note that most eBay auctions are between 3 and 10 days long.

TABLE 2. Winning probability and accrued surplus.

| Bidding rule | Prob(Win) | Avg(Surplus) ($) |
|---|---|---|
| Sniping | 0.92 | 18 |
| Early bidding | 0.62 | 0 |
| Automated bidding | 0.61 | 32 |

*Notes.* The table shows the result of three different bidding strategies: sniping, early bidding, and automated bidding according to the data-driven forecasting model. The second column shows the probability of wining an auction; the third column shows the resulting consumer surplus.

that seller. Bidders who intentionally bid on the same auction as other bidders (see, e.g., http://www.auctionshadow.com) form a different type of network with one another. More generally, users that visit each other's websites, post on each other's blogs, or "meet" online in one way or the other are all linked in a network. Although this data challenge is intrinsically related to the previous challenge of competition, we discuss it separately, partly due to the spacial nature of networked data.

Modeling user networks is complicated by the fact that data arrive in the form of ties and nodes, rather than in the more classical form of vectors and matrices. This new data structure motivates the need for new methods of analysis, which has given rise to the new field of *network analysis*, and in particular *social network analysis* (SNA) (e.g., Scott [44]). Social network analysis views social relationships in terms of ties and nodes. Nodes are the individual actors within the networks, and ties are the relationships between the actors. In its simplest form, a social network is a map of all of the relevant ties between the nodes being studied. The network can also be used to determine the social capital of individual actors. These concepts are often displayed in a social network diagram (e.g., Figure 6), where nodes are the points and ties are the lines.

The analysis of social networks has several objectives. On one hand, one wants to study *global* aspects of the entire network such as its *betweenness*, *closeness*, and the degree of *centrality*. All of these measures quantify the connectivity of an individual to other individuals in the network.

Centrality measures the number of incoming and outgoing interactions in a directed network (or the total interactions in the undirected case). There are two interpretation of this number. Freeman [16] suggests that individuals who have many ties to other individuals are more influential and thus less dependent on other individuals. Bonacich [7] suggests that two entities with the same centrality are not necessarily equally important. Instead, he suggests that, in addition, one needs to account for the centrality of its neighbors. The idea is that an individual that is connected to other individuals with few connections is more influential, because it makes these individuals more depended upon him. Bonachich's approach is commonly referred to as *The Bonachich Power* and is widely used in SNA literature.

The closeness of an individual measures the geodesic distances to other individuals within the same graph (Sabidussi [43]). The level of closeness of an individual is measured by the mean distance to all other individuals. A variation of closeness is called eigenvector centrality, which accounts for the scores of the connected entities, as well as their geodesic distances. Hence, connections to high-scoring individuals result in a higher score for a particular individual. Betweenness is another measure of the network and measures individuals that are connected on shorter paths than others. In other words, individuals with a low betweeness score have a faster reach of other network members (Freeman [16]).

In addition to global aspects of the network, one is also interested in studying *local* aspects, such as the existence and structural form of subgraphs (Lubbers and Snijders [32]). Moreover, research has also paid particular attention to longitudinal and dynamic aspects of networks (Snijders [47]).

User networks are important because they facilitate the flow of information between all network members. Hill et al. [21] study the effect of networks on viral marketing and find that "network neighbors" (i.e., those consumers linked to a prior customer) adopt a new product at a significantly higher rate. Watts and Dodds [55] study the role of centrality of a network participant on product diffusion. The authors simulate diffusion models and find that central participants are modestly more important than average participants in product diffusion. Interestingly, the authors also show that most social change is driven not by central network members, but by the easily influenced individuals who influence other easily influenced individuals. Similar applications of SNA in diffusion and promotion models are offered by Mayzlin [35], Valente [52], and others. An early application of SNA to online markets can be found in Dass and Reddy [9], who study bidder interactions in simultaneous auctions. The authors address the question of existence of "key" bidders that are more influential than others on the auction outcome. They formalize the interactions as a network where nodes correspond to bidders and an arc between two bidders corresponds to competition in an auction.

In the following case study, we investigate the existence and importance of user networks in online transactions. In particular, we investigate the impact of user networks on the outcome of an online auction. To that end, we derive a new measure of e-loyalty based on a combination of social network analysis and functional data models. Our preliminary analyses show that networks vary across product types and result in different types of bidder loyalty.

## 5.1. Case Study III: User Networks in Online Auctions and e-Loyalty

eBay is one of the largest consumer-to-consumer (C2C) online marketplaces and it enables trading between individuals and small businesses. Each day, several million items are up for sale and offered to several million potential buyers. Although much of the literature up to date considers events within one auction as independent of other auctions, in reality many auctions are interlinked. Bidders can participate in more than one auction at a time, they can purchase exclusively from one seller or from multiple sellers over time, and they can also choose among many competing products that are offered simultaneously. One result of this complex web of interactions is a seller–bidder network. In what follows, we study the effect of this network on the loyalty of bidders.

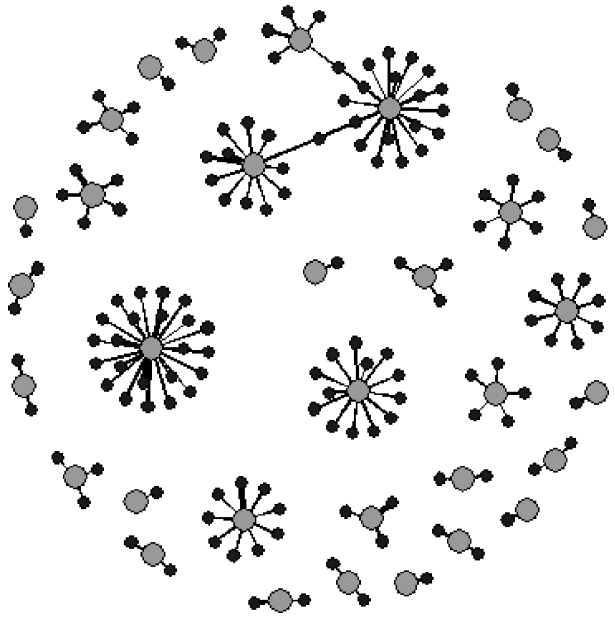
One of the main challenges in understanding seller–bidder interactions is to define a suitable measure that captures the different forms of e-loyalty. The goal is to map bidders' loyalty from an observed network to a continuous measure that ranges between multiple purchases from a single seller (high loyalty) and single purchases from multiple sellers (low loyalty). Using an innovative combination of social network analysis and functional data analysis, we define a unique measurement for both bidders' loyalty and sellers' perceived loyalty.

We start by defining user networks in online auctions, then define a measure of e-loyalty, and finally study the effect of loyalty on the outcome of an auction.

**5.1.1. User Network Definition.** We define a user network as the bipartite graph in which one set of nodes represents sellers (the large grey nodes in Figure 6), the other set represents bidders (small black nodes), and an arc between a bidder and a seller indicates that there is an interaction (a bid, in this case) between the two. The weight of the arc corresponds to the number of interactions in distinct auctions. Figure 6 shows the network of bidders and sellers for *Titleist* golf balls auctions that took place between August 9th, 2007 and October 2nd, 2007 on eBay.

**5.1.2. A Novel Measure of e-Loyalty.** One of the main challenges in understanding seller–bidder interactions is defining a suitable measure that captures the different forms of e-loyalty. Our goal is to map a bidder's loyalty to a continuous scale that ranges between multiple purchases from a single seller (i.e., high loyalty) to many single purchases from

FIGURE 6. e-Loyalty network for golf ball auctions.



*Notes.* The large grey nodes correspond to sellers; the small black nodes correspond to bidders. An arc represents an interaction between the two.

different sellers (i.e., low loyalty). To that end, we first derive the distribution of loyalty from the seller–bidder network. We then characterize each distribution by a single score using functional principal components analysis.

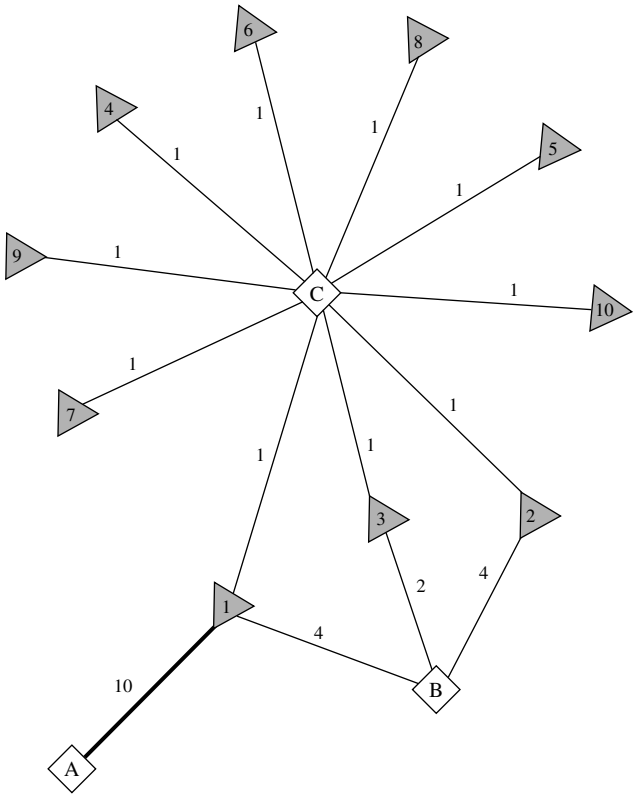
For illustration, consider the three bidders in Figure 7 (labeled "A," "B," and "C"). Each bidder has a total of 10 interactions with different sellers. Bidder A interacts 10 times with the same seller; thus, his loyalty is very high and exclusive to that seller. Bidder B interacts four times each with sellers 1 and 2, respectively, and two times with seller 3. We can consider bidder B's loyalty as moderate. In contrast, bidder C has a single interaction with each of 10 sellers. Bidder C is the least loyal of all three bidders.

We can summarize each bidder's loyalty by their corresponding *loyalty distributions*. Figure 8 shows the observed loyalty distributions (i.e., histograms, top panels) and their corresponding smooth representations (bottom panels). Smooth representations eliminate noise; they also allow for a very flexible and granular way of capturing and comparing different forms of loyalty distribution. We can see that very loyal bidders (i.e., such as bidder A) distribute all their loyalty to a single seller; thus, their loyalty distribution is very steep and shows only little variance. In contrast, bidders with little loyalty (e.g., bidder C) display distributions that are very flat and highly variable.

The next step is to summarize each distribution by a single number that captures similarities (and differences) across different loyalty distributions. This can be done via *functional principal components analysis* (fPCA), a functional version of principal components analysis (see Ramsay and Silverman [40]).

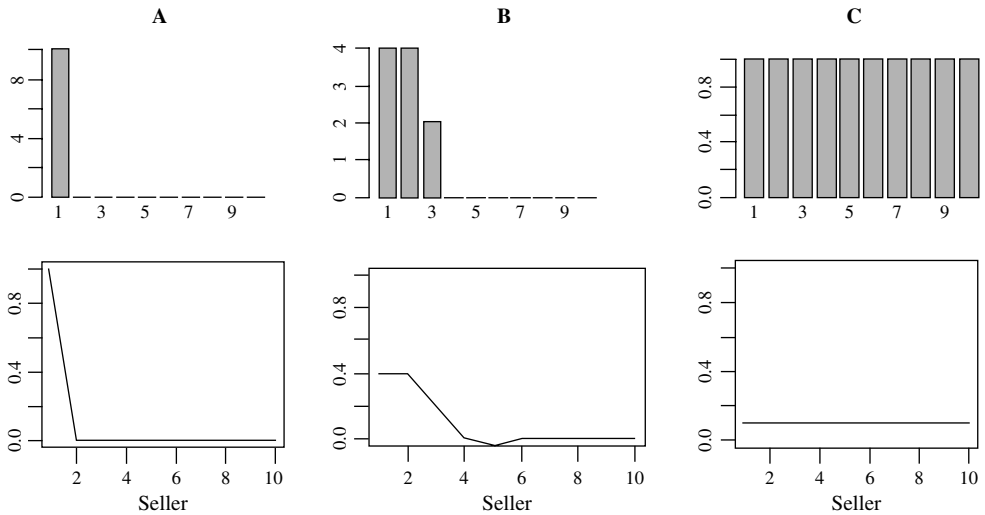
Functional principal components analysis is similar in nature to ordinary PCA; however, rather than operating on data vectors, it operates on functional objects (continuous curves, in our context). To make the discussion more concrete, let us assume that loyalty curves are measured at $t$ discrete time points. Let $\mathbf{Y}^s = (\mathbf{y}_1^s, \ldots, \mathbf{y}_n^s)$ denote the $n \times t$ matrix consisting of all smooth loyalty curves. Let $\mathbf{R} := \mathrm{Corr}(\mathbf{Y}^s)$ be the $n \times n$ correlation matrix obtained from $\mathbf{Y}^s$. Analogous to ordinary principal components analysis, $\mathbf{R}$ is decomposed into $\mathbf{P}^T\mathbf{\Lambda}\mathbf{P}$, where $\mathbf{\Lambda}$ is the $n \times n$ matrix of eigenvalues and $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]^T$ is the $n \times t$ corresponding

FIGURE 7. Illustration of observed seller–bidder network between bidders A, B, and C, and sellers 1–10.



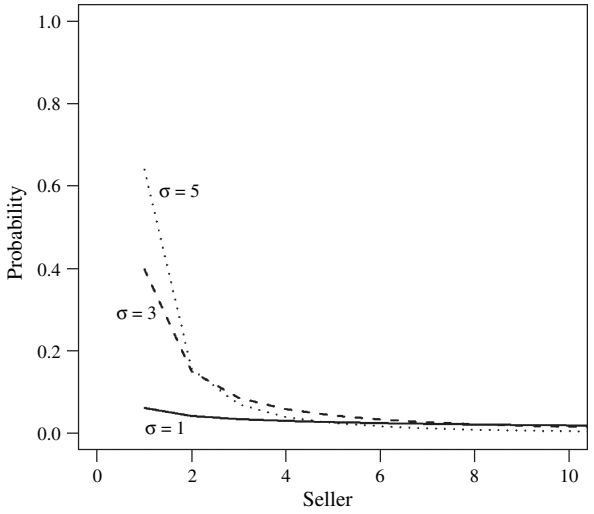

matrix of eigenvectors. In this simplified example, each $\mathbf{e}_i$, $i = 1, \ldots, n$ is a $t \times 1$ vector, but in reality $\mathbf{e}_i$ is a continuous function in time, i.e., $\mathbf{e}_i = \mathbf{e}_i(t)$. We can think of each eigenvector $\mathbf{e}_i$ as a loyalty-defining characteristic. Each eigenvector $\mathbf{e}_i$, $i = 1, \ldots, n$ captures $\lambda_i \times 100\%$ of the variability in $\mathbf{Y}^s$, where $\lambda_i$ is the $i$th eigenvalue or the $i$th diagonal element in $\mathbf{\Lambda}$.

FIGURE 8. Loyalty distribution.




*Notes.* The top panel shows the observed loyalty distribution for each bidder from Figure 7; the bottom panel shows a smooth representation of the standardized distribution using smoothing splines.

FIGURE 9. Simulated e-loyalty.




*Note.* Larger values of $\sigma$ correspond to higher levels of loyalty to the same seller.

The eigenvector corresponding to the largest eigenvalue is denoted as the first principal component (PC hereafter). Similarly, the second PC is the eigenvector that corresponds to the second highest eigenvalue, and so on. Common practice is to choose only those eigenvectors that correspond to the largest eigenvalues, i.e., those that explain most of the variation in $\mathbf{Y}^s$. By discarding those eigenvectors that explain no or only a very small proportion of the variation, we capture the most important characteristics of the observed data patterns without much loss of information. In our context, the first eigenvector captures 93% of the variation, so we summarize loyalty by the first eigenvector of the smooth loyalty curve.

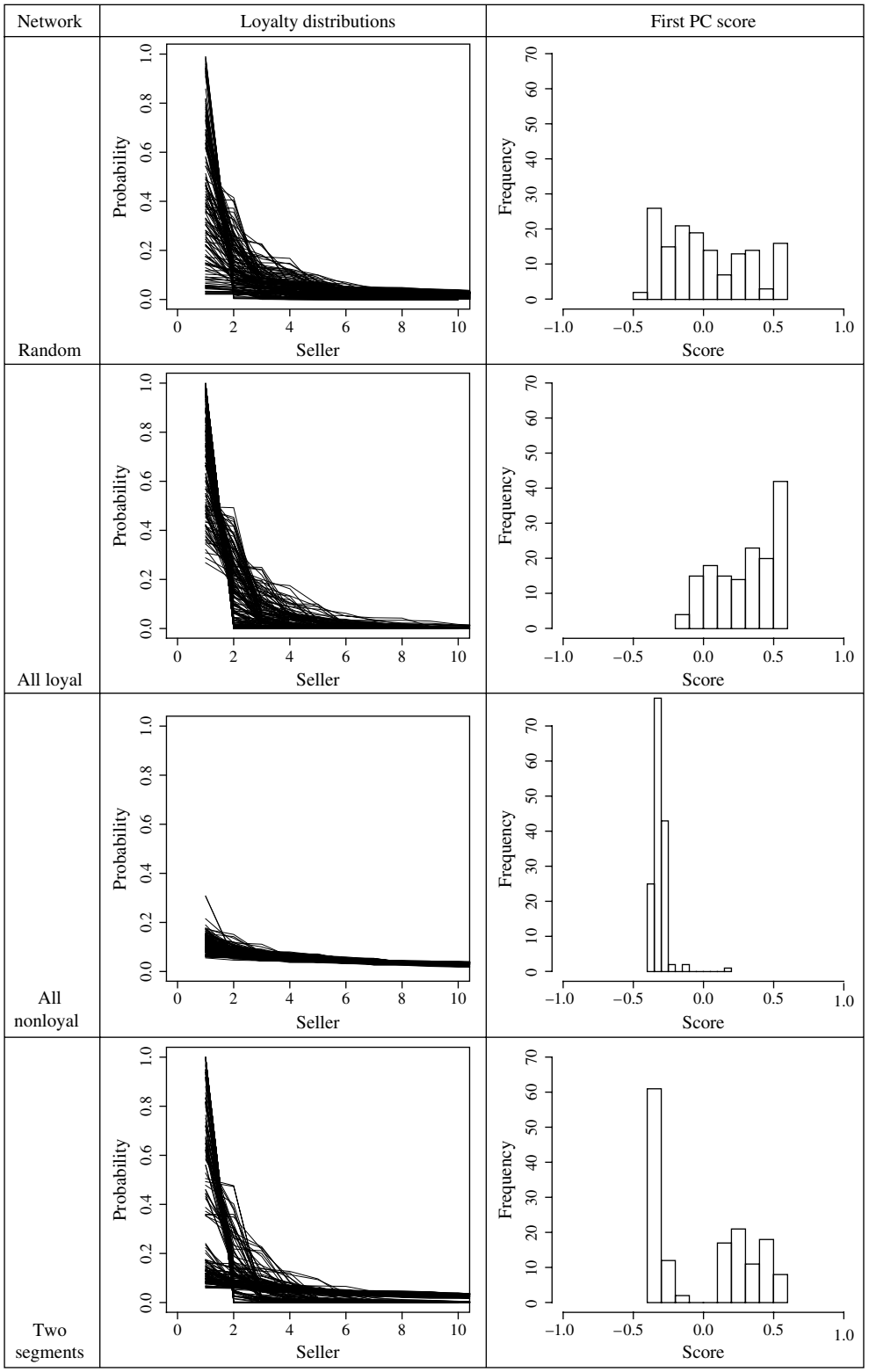
To better understand the derived loyalty measures, we first conduct a small simulation study. We simulate a network of 150 bidders interacting with 50 sellers. We assume that bidders' loyalties follow a log-normal distribution $LogNormal(\mu, \sigma)$ with a fixed[2] location parameter $\mu$ and a shape parameter $\sigma$ that varies between 1 and 5. Figure 9 illustrates the simulations. A small value of $\sigma$ (e.g., $\sigma = 1$) corresponds to a very low loyalty, whereas bidders with high loyalty to the same seller experience higher values of $\sigma$.

In our simulations, we consider four different types of networks. In the first network, loyalty of all bidders is *randomly distributed* between not loyal ($\sigma = 1$) and very loyal ($\sigma = 5$). In the second and third networks, all bidders are either completely loyal ($\sigma = 5$) or completely unloyal ($\sigma = 1$), respectively. Finally, in the forth network we consider a mixture of two different segments of bidders: some that are loyal and others that are unloyal bidders.

Figure 10 shows the results. The second panel displays the observed bidder loyalty distributions; the third panel displays the associated first functional principal component score (fPCS). The first fPCS captures the strength of loyalty: bidders that are extremely loyal have a very steep loyalty distribution (second row in Figure 10); hence, their fPCS score is high. Similarly, unloyal bidders (third row in Figure 10) have a very low fPCS score. We can see that the fPCS score efficiently captures differences in a bidder's loyalty distribution in one single number and characterizes loyal and unloyal bidders on a continuous scale.

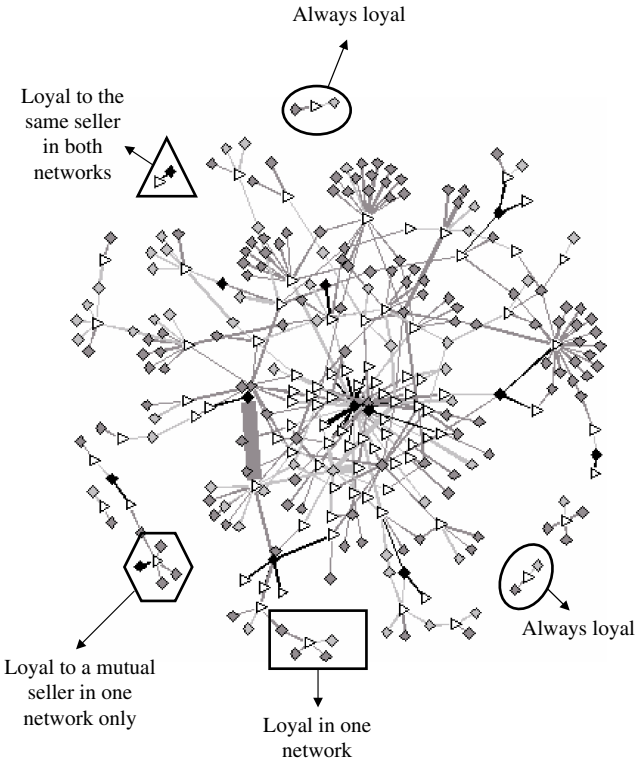
[2] The actual value of the location parameter does not have an effect on our simulation because we normalize the simulated data. By normalizing, we also do not differentiate between bidders that have the same level of loyalty but a different number of interactions.

FIGURE 10. Simulated e-loyalty distributions.




*Notes.* The middle column shows the observed bidder loyalty distributions; the right column shows the first functional principal components score.

FIGURE 11. e-Loyalty networks for *Parker* and *Cross* pens.



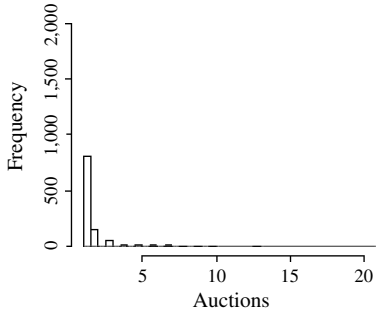
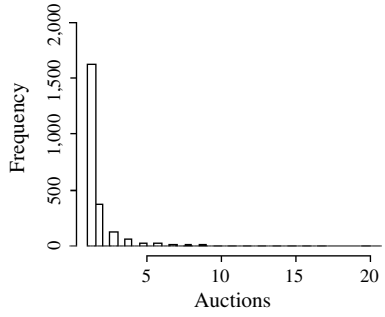
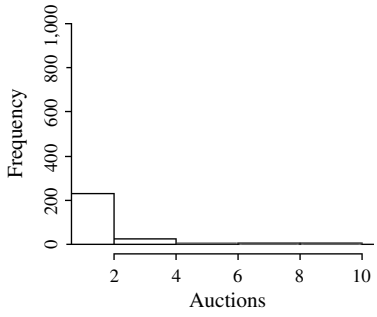
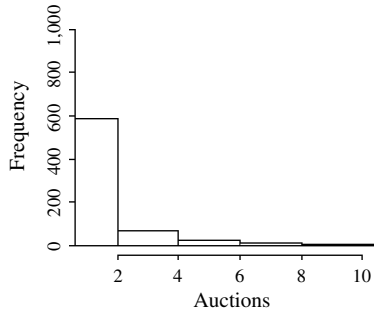
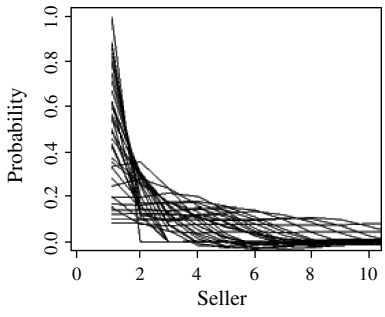
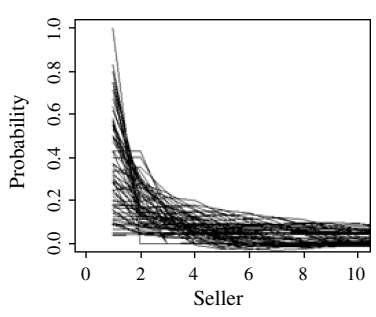
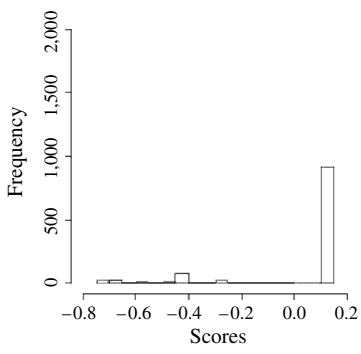
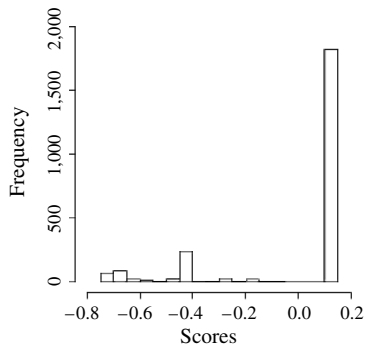

*Notes.* Sellers are denoted by squares, bidders by triangles; arcs denote an interaction between the two. Light (dark) grey arcs denote bidder–seller interactions exclusively for *Parker* (*Cross*) pens; black arcs denote an interaction between a bidder and a seller for *both* brands.

**5.1.3. Network Comparison.** We apply our loyalty measures to two different markets for roller-ball pens: auctions for *Parker* pens and *Cross* pens, collected from eBay in December 2007. Both are of similar value (around $30), with the Parker pen being slightly more expensive. Figure 11 shows the overlapping networks for both products. Bidders are denoted by white, triangular nodes; sellers are denoted by squares. Black squares denote sellers that sell both Parker *and* Cross pens; light and dark grey squares denote sellers that sell a single brand (either Parker (light grey) or Cross (dark grey) pens). The arcs are dyed similarly: black arcs represent interactions across both networks, whereas light grey (dark grey) arcs denote interactions exclusively in the Parker (Cross) network.

We can see that the loyalty networks are rather complex. Some bidders bid on only one brand (*Parker* or *Cross*), whereas others bid on both. For those bidders who bid on both brands, they can choose to interact with only one seller (who sells both brands), or with many different sellers. We can see that some bidders are always loyal to the same seller, regardless of the brand (i.e., across both networks), whereas others are loyal to one seller for *one* brand, but loyal to a *different* seller for the other brand. Yet, other bidders display only little loyalty and interact with a multitude of sellers. We investigate these initial observations in more detail below.

Figure 12 compares the two product networks. We can see that there is a significantly larger amount of *Cross* auctions compared with *Parker* auctions (4,041 versus 2,501), and, as a result, a larger number of bidders and sellers in the *Cross* network. Most bidders participate in one or two auctions (row 5 in Figure 12), but this number is a bit more right skewed for the *Cross* auctions. Similarly, most sellers have between two and four auctions (row 6) with a somewhat higher skew again in the *Cross* network. Row 7 displays the resulting loyalty

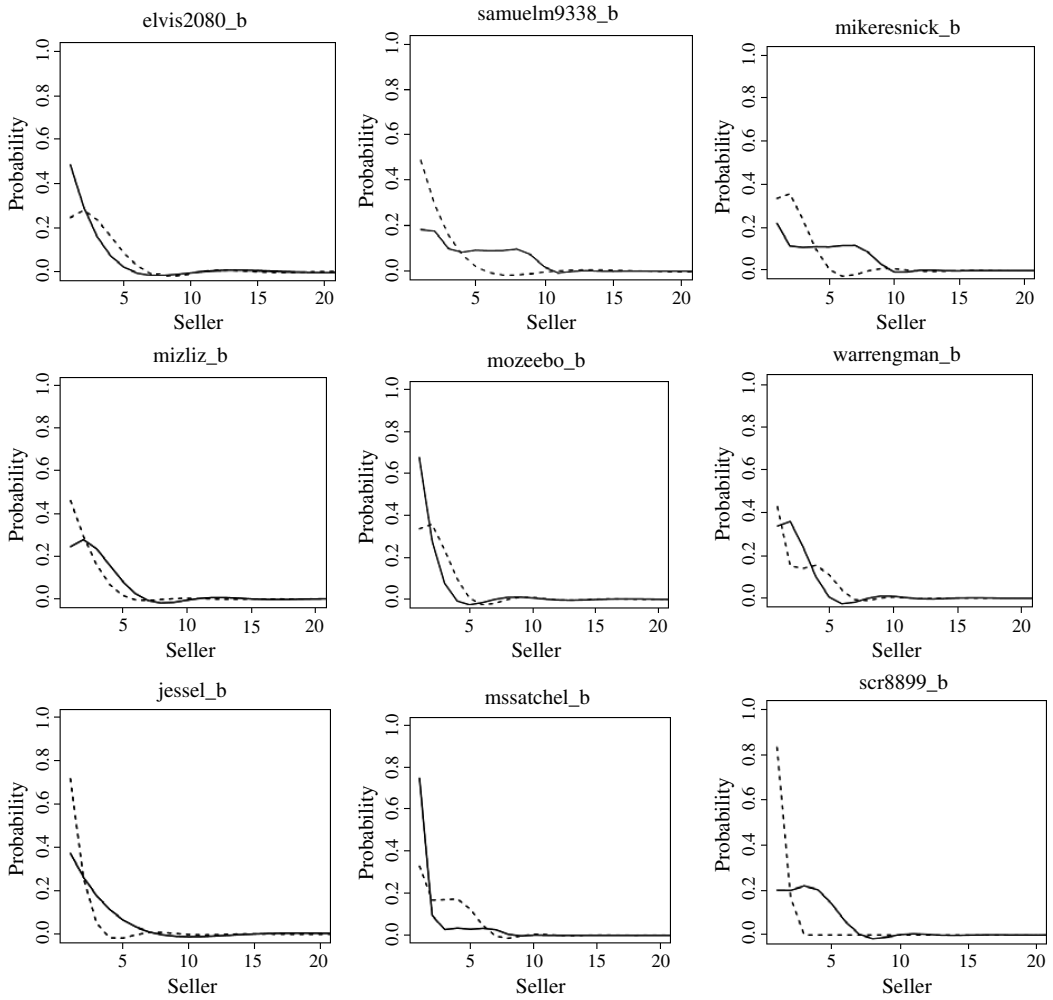
FIGURE 12. Comparison of brand networks.

| Measure | Parker | Cross |
|---|---|---|
| Number of sellers | 241 | 585 |
| Number of bidders | 1,101 | 2,354 |
| Number of auctions | 2,501 | 4,041 |
| Distribution of auctions per bidder |  |  |
| Distribution of auctions per seller |  |  |
| Loyalty curves |  |  |
| Loyalty distribution |  |  |



*Notes.* The left panel shows loyalty summaries for the *Parker* network; the right panel shows the corresponding summaries for the *Cross* network.

FIGURE 13. Comparison of bidders' loyalty curves in different product networks.




*Notes.* The solid lines correspond to loyalty in the *Cross* network; the dashed lines correspond to the *Parker* loyalty.

curves for each bidder, and row 8 shows the corresponding loyalty scores based on the first fPCA. We can see that most bidders show strong loyalty (cf. steep loyalty curves in row 7 or high loyalty scores in row 8). However, note that the *Cross* network shows a stronger left skew in the loyalty distribution, suggesting a larger proportion of nonloyal bidders for this product. We are quick to caution though that the (numerical) difference may merely be the result of the larger number of auctions and bidders for this product.

Although the above network differences may be entirely due to the different network sizes, we obtain a more careful picture by comparing the loyalty of individual bidders across the two product networks. Figure 13 shows the loyalty curves for bidders who participate at least three times in both the *Parker* as well as the *Cross* network (total of nine bidders). The solid lines in Figure 13 correspond to their corresponding loyalty in the *Cross* network, whereas the dashed lines correspond to the *Parker* loyalty. We can see that whereas some bidders (e.g., elvis_2080; mikeresnick) display equal loyalty (high or low) in both networks, others (e.g., scr8899) are extremely loyal in the *Parker* network but nonloyal for the other product.

## 6. Conclusion

This tutorial focuses on how novel statistical method and thought can be used to take advantage of the rich data found in online markets. Such data tend to have nonstandard structure in the sense that off-the-shelf statistical methodology is often not adequate, neither for capturing the entire information in the data, nor for answering new research questions based on the data. Three particularly challenging aspects of eCommerce data are dynamics, competition, and user networks. All three phenomena arise from the interaction between users, between processes, and between markets. Marked by different behavioral, economic, and social forces that are often are pulling into opposite directions, the online environment is becoming increasingly interactive, networked, and customized. These forces, together with their effect on the resulting transaction outcome, are now observable and measurable and as a result eCommerce data are richer in the different, often diverse pieces of information they contain. A main facet of this richness is the combination of several types of traditional data elements and the resulting new dependence structures that it brings about. Examples are combinations of cross-sectional and temporal data that are extremely common in eCommerce data sets; network data and their evolution over time; competition between users for a given good or even across markets, and more.

This tutorial illustrates that empirical research in eCommerce, although growing fast, involves many new statistical challenges. We regard these challenges as opportunities in that they call for the development of new statistical tools or the adaptation of existing statistical machinery. New ideas are needed for data representation, summarization, visualization, and modeling. Our case studies exemplify the many different challenges that arise in the context of online auctions, and we present some solutions and new research directions, many of which also apply to other eCommerce applications. We conclude on the note that statistics in eCommerce becomes more and more important to researchers and practitioners alike as exemplified by a recent special issue on the topic (Jank and Shmueli [26]), and an edited book (Jank and Shmueli [30]), or an annual workshop series that has steadily been growing out of its inaugural event in 2005 (http://www.smith.umd.edu/dit/statschallenges).

## References

[1] D. Agarwal. Statistical challenges in internet advertising. W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[2] D. Ariely and I. Simonson. Buying, bidding, playing, or competing? Value assessment and decision dynamics in online auctions. *Journal of Consumer Psychology* 13(1&2):113–123, 2003.

[3] R. Bapna, W. Jank, and G. Shmueli. Price formation and its dynamics in online auctions. *Decision Support Systems* 44:641–656, 2008.

[4] R. Bapna, W. Jank, and G. Shmueli. Consumer surplus in online auctions. *Information Systems Research*, forthcoming, 2008.

[5] R. Bapna, P. Goes, R. Gopal, and J. R. Marsden. Moving from data-constrained to data-enabled research: Experiences and challenges in collecting, validating and analyzing large-scale e-commerce data. *Statistical Science* 21(2):116–130, 2006.

[6] R. Bapna, P. Goes, A. Gupta, and Y. Jin. User heterogeneity and its impact on electronic auction market design: An empirical exploration. *MIS Quarterly* 28(1):21–43, 2004.

[7] P. Bonacich. Power and centrality: A family of measures. *The American Journal of Sociology*, 92(5):1170–1182, 1987.

[8] S. Borle, P. Boatwright, and J. B. Kadane. The timing of bid placement and extent of multiple bidding: An empirical investigation using eBay online auctions. *Statistical Science* 21(2):194–205, 2006.

[9] M. Dass and S. K. Reddy. Bidder networks and price dynamics in online auctions. Working paper, University of Georgia, Atlanta, 2007.

[10] M. Dass and S. K. Reddy. An analysis of price dynamics, bidder networks and market structure in online auctions. W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[11] M. Dass, S. K. Reddy, and R. Du. Dyadic bidder interactions and key bidders in simultaneous online auctions. Working paper, University of Georgia, Atlanta, 2007.

[12] S. E. Fienberg. Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. *Statistical Science* 21(2):143–154, 2006.

[13] C. Forman and A. Goldfarb. How has electronic commerce research advanced understanding of the offline world? W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[14] R. Forsythe, T. A. Rietz, and T. W. Ross. Wishes, expectations, and actions: A survey on price formation in election stock markets. *Journal of Economic Behavior & Organization* 39:83–110, 1999.

[15] N. Foutz and W. Jank. The wisdom of crowds: Pre-release forecasting via functional shape analysis of the online virtual stock market. Marketing Science Institute Report 07–114, Cambridge, MA, 2007.

[16] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks* 1(3):215–239, 1979.

[17] A. Ghose. The economic impact of user-generated and firm-published online content: Directions for advancing the frontiers in electronic commerce research. W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[18] A. Ghose and A. Sundararajan. Evaluating pricing strategy using e-commerce data: Evidence and estimation challenges. *Statistical Science* 21(2):131–142, 2006.

[19] E. Haruvy, P. Popkowski Leszczyc, O. Carare, J. Cox, E. Greenleaf, W. Jank, S. Jap, Y.-H. Park, and M. Rothkopf. Competition between auctions. *Marketing Letters*, ePub ahead of print May 10, 2008, http://www.springerlink.com/content/70l43453322417r8/.

[20] J. E. Heyman, Y. Orhun, and D. Ariely. Auction fever: The effect of opponents and quasi-endowment on product valuations. *Journal of Interactive Marketing* 18(4):8–21, 2004.

[21] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science* 22(2):256–276, 2006.

[22] V. Hyde, W. Jank, and G. Shmueli. Investigating concurrency in online auctions through visualization. *The American Statistician* 60:241–250, 2006.

[23] V. Hyde, W. Jank, and G Shmueli. A family of growth models for representing the price process in online auctions. W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[24] W. Jank and G. Shmueli. Visualizing online auctions. *Journal of Computational and Graphical Statistics* 14(2):299–319, 2005.

[25] W. Jank and G. Shmueli. Functional data analysis in electronic commerce research. *Statistical Science* 21(2):155–166, 2006.

[26] W. Jank and G. Shmueli. A special issue on statistical challenges and opportunities in electronic commerce research. *Statistical Science* 21(2):113–115, 2006.

[27] W. Jank and G. Shmueli. Modeling concurrency of events in online auctions via spatio-temporal semiparametric models. *Journal of the Royal Statistical Society—Series C* 56:1–27, 2007.

[28] W. Jank and G. Shmueli. Studying heterogeneity of price evolution in eBay auctions via functional clustering. *Handbook of Information Systems Series*: *Business Computing*. Elsevier, Amsterdam, forthcoming, 2008.

[29] W. Jank and G. Shmueli, eds. Modeling price dynamics in online auctions via regression trees. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[30] W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, 2008.

[31] R. J. Kauffman and C. A. Wood. The effects of shilling on final bid prices in online auctions. *Electronic Commerce Research and Applications* 4(2):21–34, 2005.

[32] M. Lubbers and T. Snijders. A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social Networks* 29(4):489–507, 2007.

[33] D. Lucking-Reiley. Using filed experiements to test equivalence between auction formats: Magic on the internet. *American Economic Review* 89(5):1063–1080, 1999.

[34] A. Matas and Y. Schamroth. Optimization of search engine marketing bidding strategies using statistical techniques. W. Jank and G. Shmueli, eds. *Statistical Methods in eCommerce Research*. John Wiley & Sons, New York, forthcoming, 2008.

[35] D. Mayzlin. The influence of social networks on the effectiveness of promotional strategies. Yale School of Management, Working paper, Yale University, New Haven, CT, 2008.

[36] Y.-H. Park and E. Bradlow. An integrated model for whether, who, when, and how much in internet auctions. *SSRN eLibrary*, 2005.

[37] D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen. The real power of artificial markets. *Science* 291(5506):987–988, 2001.

[38] J. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society—Series B* 69(5):741–796, 2007.

[39] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*, 2nd ed. Springer Series in Statistics, Springer-Verlag, New York, 2005.

[40] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

[41] S. K. Reddy and M. Dass. Modeling online art auction dynamics using functional data analysis. *Statistical Science* 21(2):179–193, 2006.

[42] A. E. Roth and A. Ockenfels. Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the internet. *The American Economic Review* 92(4):1093–1103, 2002.

[43] G. Sabidussi. The centrality index of a graph. *Psychometrika* 31(4):581–603, 1966.

[44] J. Scott. *Social Network Analysis: A Handbook*. Sage Publications, Thousand Oaks, CA, 2000.

[45] G. Shmueli, R. P. Russo, and W. Jank. The barista: A model for bid arrivals in online auctions. *Annals of Applied Statistics* 1(2):412–441, 2007.

[46] J. S. Simonoff. *Smoothing Methods in Statistics*, 1st ed. Springer-Verlag, New York, 1996.

[47] T. Snijders. The statistical evaluation of social network dynamics. M. Sobel and M. Becker, eds. *Sociological Methodology*. Basil Blackwell, Boston, 361–395, 2001.

[48] M. Spann and B. Skiera. Internet-based virtual stock markets for business forecasting. *Management Science* 49(10):1310–1326, 2003.

[49] K. Stewart, D. Darcy, and S. Daniel. Opportunities and challenges applying functional data analysis to the study of open source software evolution. *Statistical Science* 21(2):167–178, 2006.

[50] J. Surowiecki. *The Wisdom of Crowds*. Random House, New York, 2005.

[51] R. Telang and M. D. Smith. Internet exchanges for used digital goods. *SSRN eLibrary*, 2008.

[52] T. W. Valente. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory* 2(2):163–164, 1996.

[53] S. Wang, W. Jank, and G. Shmueli. Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economic Statistics* 26(2):144–160, 2008.

[54] S. Wang, W. Jank, G. Shmueli, and P. Smith. Modeling price dynamics in ebay auctions using principal differential analysis. *Journal of the American Statistical Association*, forthcoming, 2008.

[55] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research* 34(4):441–458, 2007.

[56] S. Yao and C. F. Mela. Online auction demand. *SSRN eLibrary*, 2007.

[57] S. Zhang and W. Jank. An automated and data-driven bidding strategy for online auctions. Technical report, Robert H. Smith School of Business, University of Maryland, College Park, 2008.