MODELING ONLINE AUCTIONS
STATISTICS IN PRACTICE

Advisory Editor
Marian Scott
University of Glasgow, UK

Founding Editor
Vic Barnett
Nottingham Trent University, UK

The texts in the series provide detailed coverage of statistical concepts, methods, and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field. Readers are assumed to have a basic understanding of introductory statistics, enabling the authors to concentrate on those techniques of most importance in the discipline under discussion.

The books meet the need for statistical support required by professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance, and commerce; public services; the earth and environmental sciences.

A complete list of titles in this series appears at the end of the volume.
To our families and mentors who have inspired and encouraged us, and always supported our endeavors:

Angel, Isabella, Waltraud, Gerhard and Sabina

– Wolfgang Jank

Boaz and Noa Shmueli, Raquelle Azran, Zahava Shmuely, and Ayala Cohen

– Galit Shmueli
CONTENTS

Preface ix
Acknowledgments xi

1 Introduction 1
  1.1 Online Auctions and Electronic Commerce, 3
  1.2 Online Auctions and Statistical Challenges, 4
  1.3 A Statistical Approach to Online Auction Research, 6
  1.4 The Structure of this Book, 6
  1.5 Data and Code Availability, 8

2 Obtaining Online Auction Data 9
  2.1 Collecting Data from the Web, 9
  2.2 Web Data Collection and Statistical Sampling, 18

3 Exploring Online Auction Data 31
  3.1 Bid Histories: Bids versus “Current Price” Values, 32
  3.2 Integrating Bid History Data With Cross-Sectional Auction Information, 36
  3.3 Visualizing Concurrent Auctions, 41
  3.4 Exploring Price Evolution and Price Dynamics, 44
  3.5 Combining Price Curves with Auction Information via Interactive Visualization, 57
  3.6 Exploring Hierarchical Information, 60
3.7 Exploring Price Dynamics via Curve Clustering, 63
3.8 Exploring Distributional Assumptions, 72
3.9 Exploring Online Auctions: Future Research Directions, 94

4 Modeling Online Auction Data 96

4.1 Modeling Basics (Representing the Price Process), 97
4.2 Modeling The Relation Between Price Dynamics and Auction Information, 132
4.3 Modeling Auction Competition, 157
4.4 Modeling Bid and Bidder Arrivals, 189
4.5 Modeling Auction Networks, 238

5 Forecasting Online Auctions 253

5.1 Forecasting Individual Auctions, 254
5.2 Forecasting Competing Auctions, 279
5.3 Automated Bidding Decisions, 291

Bibliography 301

Index 313
Our fascination with online auction research started in 2002. Back then, empirical research using online auction data was just beginning to appear, and it was concentrated primarily in the fields of economics and information systems. We started our adventure by scrutinizing online auction data in a statistically oriented, very simple, and descriptive fashion, asking questions such as “How do we plot a single bid history?,” “How do we plot 1000 bid histories without information overload?,” “How do we represent the price evolution during an ongoing auction?,” “How do we collect lots and lots of data efficiently?”. During that period, we started to collaborate with colleagues (primarily non-statistician) who introduced us to auction theory, to its limitations in the online environment, and to modern technologies for quickly and efficiently collecting large amounts of online auction data. Taking advantage of our “biased” statistical thinking, we started looking at existing questions such as “What auction factors affect the final price?” or “How can we quantify the winner’s curse?” in a new, data-driven way. Then, after having been exposed to huge amounts of data and after having developed a better understanding of how the online auction mechanism works, we started asking more “unusual” and risky questions such as “How can we capture the price dynamics of an auction?” or “How can we forecast the outcome of an auction in a dynamic way?” Such questions have led to fruitful research endeavors ever since. In fact, they have led to several rather unique innovations. First, since we posed these questions not only to ourselves but also to our PhD students, it has led to a new PhD specialization. Our graduating students combine a strong background in mathematics or statistics with training in problem solving related to business and electronic commerce. This combination is essential for competing and advancing in the current business analytics environment, as described in the recent monograph
Competing on Analytics by Davenport and Harris or many other articles and books on the same topic.

Our research has also led to the birth of a new research area called statistical challenges in eCommerce research. Our interactions and collaborations with many colleagues have led to a new annual symposium, now in its sixth year, that carries the same name (see also http://www.statschallenges.com). The symposium attracts researchers from different disciplines who all share the same passion: using empirical techniques to take advantage of the wealth of eCommerce data for the purpose of better understanding the online world.

There are several aspects that make online auction research a topic to be passionate about. Looking long and hard at data from online auctions, we have discovered surprising structures in the data, structures that were not straightforward to represent or to model using standard statistical tools. Two examples are the very unevenly spacing of event arrivals and the “semicontinuous” nature of online auction data. To handle such data and tease out as much information as possible from them, we adapted existing statistical methods or developed new tools and methodology. Our statistical approach to online auction research has resulted in a wide variety of methods and tools that support the exploration, modeling, and forecasting of auction data. While this book primarily focuses on auctions, the lessons learned can be applied more generally and could be of interest to any researcher studying online processes.

The book is intended for researchers and students from a wide variety of disciplines interested in applied statistical modeling of online auctions. On the one hand, we believe that this book will add value to the statistician, interested in developing methodology and in search of new applications—this book carefully describes several areas where statisticians can make a difference and it also provides a wealth of new online auction data. On the other hand, the book can add value to the marketer, the economist, or the information systems researcher, looking for new approaches to derive insights from online auctions or online processes—we are very careful in describing our way of scrutinizing online auction data to extract even the last bit of (possibly surprising) knowledge from them. To allow researchers to replicate (and improve) our methods and models, we also provide the code to our software programs. For this purpose, we make available related data and code at a companion website (http://ModelingOnlineAuctions.com) for those who are interested in hands-on learning. Our own learning started with data, and we therefore encourage others to explore online auction data directly.

WOLFGANG JANK AND GALIT SHMUELI

January 2010
ACKNOWLEDGMENTS

There are many people who helped and influenced us and we would like to thank them for making this book come to fruition:

First of all, we would like to thank our following students who have worked (and sometimes suffered) with us while developing many of the ideas and concepts described in this book: Phd students Shanshan Wang, Valerie Hyde, Shu Zhang, Aleks Aris, and Inbal Yahav, as well as Masters’ students Brian Alford, Lakshmi Urimi, and others who participated in our first Research Interaction Team on online auctions.

We also thank our many co-authors and co-contributors. The generation of new knowledge and ideas, and their implementation and application would have never seen the light of day without our close collaborations. A special thanks to Ravi Bapna (University of Minnesota), who introduced us early on to the research community studying online auctions and to the immense potential of empirical methods in this field. We thank our University of Maryland colleagues Ben Shneiderman and Catherine Plaisant from the Human-Computer Interaction Lab, PK Kannan from the Marketing department, and Paul Smith from the Mathematics department. We also thank Ralph Russo (University of Iowa), Mayukh Dass (Texas Tech University), N. D. Shyamalkumar (University of Iowa), Paolo Buono (University of Bari), Peter Popkowski Leszczyc (University of Alberta), Ernan Haruvy (University of Texas at Dallas), and Gerhard Tutz (University of Munich).

We express our deepest gratitude to our many colleagues who have helped shape ideas in our head through fruitful discussions and for their enthusiasm and support for our endeavors: Paulo Goes (University of Connecticut), Alok Gupta (University of Minnesota), Rob Kauffman (University of Arizona), Ramayya Krishnan (Carnegie Mellon University), Ram Chellapa (Emory University), Anindya Ghose (NYU), Foster Provost (NYU), Sharad Borle (Rice University), Hans-Georg
Mueller (University of California at Davis), Gareth James (University of Southern California), Otto Koppius (Erasmus University), Daryl Pregibon (Google), and Chris Volinsky (AT&T).

We also greatly thank our statistician colleagues Ed George (University of Pennsylvania), Jim Ramsay (McGill University), Steve Marron (University of North Carolina), Jeff Simonoff (NYU), David Steinberg (Tel Aviv University), Don Rubin (Harvard University), David Banks (Duke University), and Steve Fienberg (Carnegie Mellon University), who supported and encouraged our interdisciplinary research in many, many ways.

Finally, we thank Steve Quigley and the team at Wiley for their enthusiasm and welcoming of our manuscript ideas.
Online auctions have received an extreme surge of popularity in recent years. Websites such as eBay.com, uBid.com, or Swoopo.com are marketplaces where buyers and sellers meet to exchange goods or information. Online auction platforms are different from fixed-price retail environments such as Amazon.com since transactions are negotiated between buyers and sellers. The popularity of online auctions stems from a variety of reasons. First, online auction websites are constantly available, so sellers can post items at any time and bidders can place bids day or night. Items are typically listed for several days, giving purchasers time to search, decide, and bid. Second, online auctions face virtually no geographical constraints and individuals in one location can participate in an auction that takes place in a completely different location of the world. The vast geographical reach also contributes to the variety of products offered for sale—both new and used. Third, online auctions also provide entertainment, as they engage participants in a competitive environment. In fact, the social interactions during online auctions have sometimes been compared to gambling, where bidders wait in anticipation to win and often react emotionally to being outbid in the final moments of the auction.

Online auctions are relatively new. By an “online auction” we refer to a Web-based auction, where transactions take place on an Internet portal. However, even before the advent of Internet auctions as we know them today, auctions were held electronically via email messages, discussion groups, and newsgroups. David Lucking-Reiley (2000) describes the newsgroup rec.games.deckmaster where...
Internet users started trading “Magic” cards (related to the game *Magic: the Gathering*) as early as 1995. He writes

By the spring of 1995, nearly 6,000 messages were being posted each week, making `rec.games.tradingcards.marketplace` the highest-volume newsgroup on the Internet. Approximately 90 percent of the 26,000 messages per month were devoted to the trading of Magic cards, with the remaining 10 percent devoted to the trading of cards from other games.

Lucking-Reiley (2000) presents a brief history of the development of Internet auctions and also provides a survey of the existing online auction portals as of 1998. The first online auction websites, launched in 1995, went by the names of *Onsale* and *eBay*. Onsale (today *Egghead*) later sold its auction service to Yahoo! and moved to fixed-price retailing. Yahoo! and Amazon each launched their own online auction services in 1999. Within 10 years or so, both shut down their auction services and now focus exclusively on fixed-price operations. (At the time of writing, Yahoo! maintains online auctions in Hong Kong, Taiwan, and Japan.) Thus, from 1995 until today (i.e., 2010) the consumer-to-consumer online auction marketplace has followed the pattern of eCommerce in general: An initial mushrooming of online auction websites was followed by a strong period of consolidations, out of which developed the prominent auction sites that we know today: eBay, uBid, or Swoopo (for general merchandise), SaffronArt (for Indian art), or Prosper (for peer-to-peer lending).

Empirical research of online auctions is booming. In fact, it has been booming much more compared to traditional, brick-and-mortar auctions. It is only fair to ask the question: “Why has data-driven research of online auctions become so much more popular compared to that of traditional auctions?” We believe the answer is simple and can be captured in one word: data! In fact, the public access to ongoing and past auction transactions online has opened new opportunities for empirical researchers to study the behavior of buyers and sellers. Moreover, theoretical results, founded in economics and derived for the offline, brick-and-mortar auction, have often proven not to hold in the online environment. Possible reasons that differentiate online auctions from their offline counterparts are the worldwide reach of the Internet, anonymity of its users, virtually unlimited resources, constant availability, and continuous change.

In one of the earliest examinations of online auctions (e.g., Lucking-Reiley et al., 2000), empirical economists found that bidding behavior, particularly on eBay, often diverges significantly from what classical auction theory predicts. Since then, there has been a surge in empirical analysis using online auction data in the fields of information systems, marketing, computer science, statistics, and related areas. Studies have examined bidding behavior in the online environment from multiple different angles: identification and quantification of new bidding behavior and phenomena, such as bid sniping (Roth and Ockenfels, 2002) and bid shilling (Kauffman and Wood, 2005); creation of a taxonomy of bidder types (Bapna et al., 2004); development of descriptive probabilistic models to capture bidding and bidder activity (Shmueli et al., 2007; Russo et al., 2008), as well as bidder behavior in terms of bid timing and amount (Borle et al., 2006; Park and Bradlow, 2005); another stream of research focuses on the price evolution during an online auction. Related to this are studies on price dynamics
ONLINE AUCTIONS AND ELECTRONIC COMMERCE

(Wang et al., 2008a,b; Bapna et al., 2008b; Dass and Reddy, 2008; Reddy and Dass, 2006; Jank and Shmueli, 2006; Hyde et al., 2008; Jank et al., 2008b, 2009) and the development of novel models for dynamically forecasting auction prices (Wang et al., 2008a; Jank and Zhang, 2009a,b; Zhang et al., 2010; Jank and Shmueli, 2010; Jank et al., 2006; Dass et al., 2009). Further topics of research are quantifying economic value such as consumer surplus in eBay (Bapna et al., 2008a), and more recently, online auction data are also being used for studying bidder and seller relationships in the form of networks (Yao and Mela, 2007; Dass and Reddy, 2008; Jank and Yahav, 2010), or competition between products, between auction formats, and even between auction platforms (Haruvy et al., 2008; Hyde et al., 2006; Jank and Shmueli, 2007; Haruvy and Popkowski Leszczyc, 2009). All this illustrates that empirical research of online auctions is thriving.

1.1 ONLINE AUCTIONS AND ELECTRONIC COMMERCE

Online auctions are part of a broader trend of doing business online, often referred to as electronic commerce, or eCommerce. eCommerce is often associated with any form of transaction originating on the Web. eCommerce has had a huge impact on the way we live today compared to 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. eCommerce transactions include buying, selling, or investing online. Examples are shopping at online retailers such as Amazon.com or participating in online auctions such as eBay.com; buying or selling used items through websites such as Craigslist.com; using Internet advertising (e.g., sponsored ads by Google, Yahoo!, and Microsoft); reserving and purchasing tickets online (e.g., for travel or movies); posting and downloading music, video, and other online content; postings opinions or ratings about products on websites such as Epinions or Amazon; requesting or providing services via online marketplaces or auctions (e.g., Amazon Mechanical Turk or eLance); and many more.

Empirical eCommerce research covers many topics, ranging from very broad to very specific questions. Examples of rather specific research questions cover topics such as the impact of online used goods markets on sales of CDs and DVDs (Telang and Smith, 2008); the evolution of open source software (Stewart et al., 2006); the optimality of online price dispersion in the software industry (Ghose and Sundararajan, 2006); the efficient allocation of inventory in Internet advertising (Agarwal, 2008); the optimization of advertisers’ bidding strategies (Matas and Schamroth, 2008); the entry and exit of Internet firms (Kauffman and Wang, 2008); the geographical impact of online sales leads (Jank and Kannan, 2008); the efficiency and effectiveness of virtual stock markets (Spann and Skiera, 2003; Foutz and Jank, 2009); or the impact of online encyclopedia Wikipedia (Warren et al., 2008).

Broad research questions include issues of privacy and confidentiality of eCommerce transactions (Fienberg, 2006, 2008) and other issues related to mining Internet transactions (Banks and Said, 2006), modeling clickstream data (Goldfarb and Lu,
INTRODUCTION

2006), and understanding time-varying relationships in eCommerce data (Overby and Konsynski, 2008). They also include questions on how online experiences advance our understanding of the offline world (Forman and Goldfarb, 2008); the economic impact of user-generated online content (Ghose, 2008); challenges in collecting, validating, and analyzing large-scale eCommerce data (Bapna et al., 2006) or conducting randomized experiments online (Van der Heijden and Böckenholt, 2008); as well as questions on how to assess the causal effect of marketing interventions (Rubin and Waterman, 2006; Mithas et al., 2006) and the effect of social networks and word of mouth (Hill et al., 2006; Dellarocas and Narayan, 2006).

Internet advertising is another area where empirical research is growing, but currently more so inside of companies and to a lesser extent in academia. Companies such as Google, Yahoo!, and Microsoft study the behavior of online advertisers using massive data sets of bidding and bidding outcomes to more efficiently allocate inventory (e.g., ad placement) (Agarwal, 2008). Online advertisers and companies that provide services to advertisers also examine 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, 2008).

Another active and growing area of empirical research is that of prediction markets, also known as “information markets,” “idea markets,” or “betting exchanges.” Prediction markets are mechanisms used to aggregate the wisdom of crowds (Surowiecki, 2005) from online communities to forecast outcomes of future events and they have seen many interesting applications, from forecasting economic trends to natural disasters to elections to movie box-office sales. While several empirical studies (Spann and Skiera, 2003; Forsythe et al., 1999; Pennock et al., 2001) report on the accuracy of final trading prices to provide forecasts, there exists evidence that prediction markets are not fully efficient, which brings up interesting new statistical challenges (Foutz and Jank, 2009).

There are many similarities between the statistical challenges that arise in the empirical analysis of online auctions and that of eCommerce in general. Next, we discuss some of these challenges in the context of online auctions; for more on the aspect of eCommerce research, see, for example, Jank et al. (2008a) or Jank and Shmueli (2008a).

1.2 ONLINE AUCTIONS AND STATISTICAL CHALLENGES

A key reason for the booming of empirical online auctions research is the availability of data: lots and lots of data! However, while data open the door to investigating new types of research questions, they also bring up new challenges. Some of these challenges are related to data volume, while others reflect the new structure of Web data. Both issues pose serious challenges for the empirical researcher.

In this book, we offer methods for handling and modeling the unique data structure that arises in online auction Web data. One major aspect is the combination of temporal and cross-sectional information. Online auctions (e.g., eBay) are a point in case. 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 a time series. 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 only toward one type of data. Moreover, while methods for panel data can address some of these challenges, these methods typically assume that events arrive at equally spaced time intervals, which is not at all the case for online auction data. In fact, Web-based temporal data that are user-generated create nonstandard time series, where events are not equally spaced. In that sense, such temporal information is better described as a process. Because of the dynamic nature of the Web environment, many processes exhibit dynamics that change over the course of the process. On eBay, for instance, prices speed up early, then slow down later, only to speed up again toward the auction end. Classical statistical methods are not geared toward capturing the change in process dynamics and toward teasing out similarities (and differences) across thousands (or even millions) of online processes.

Another challenge related to the nature of online auction data is capturing competition between auctions. Consider again the example of eBay auctions. On any given day, there exist tens of thousands of identical (or similar) products being auctioned that all compete for the same bidders. For instance, during the time of writing, 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. While not all of these 10,000 auctions may sell an 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 bidders that seek immediate satisfaction, auctions that are 5 days away from completion may be less attractive than auctions that end in the next 5 minutes. Modeling differences in product similarity and their impact on bidders’ choices is challenging (Jank and Shmueli, 2007). Similarly, understanding the effect of misaligned (i.e., different starting times, different ending times, different durations) auctions on bidding decisions is equally challenging (Hyde et al., 2006) 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. (2008).

Another challenge to 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 enterprises such as Facebook, MySpace, and LinkedIn. 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.
1.3 A STATISTICAL APPROACH TO ONLINE AUCTION RESEARCH

In this book, we provide empirical methods for tackling the challenges described above. As with many books, we present both a description of the problem and potential solutions. It is important to remember that our main focus is statistical. That is, we discuss methods for collecting, exploring, and modeling online auction data. Our models are aimed at capturing empirical phenomena in the data, at gaining insights about bidders’ and sellers’ behavior, and at forecasting the outcome of online auctions. Our approach is pragmatic and data-driven in that we incorporate domain knowledge and auction theory in a less formalized fashion compared to typical exposés in the auction literature. We make extensive use of nonparametric methods and data-driven algorithms to avoid making overly restrictive assumptions (many of which are violated in the online auction context) and to allow for the necessary flexibility in this highly dynamic environment. The online setting creates new opportunities for observing human behavior and economic relationships “in action,” and our goal is to provide tools that support the exploration, quantification, and modeling of such relationships.

We note that our work has been inspired by the early research of Lucking-Reiley et al. (2000) who, to the best of our knowledge, were the first to conduct empirical research in the context of online auctions. The fact that it took almost 9 years from the first version of their 1999 working paper until its publication in 2007 (Lucking-Reiley et al., 2007) shows the hesitation with which some of this empirical research was greeted in the community. We believe though that some of this hesitation has subsided by now.

1.4 THE STRUCTURE OF THIS BOOK

The order of the chapters in this book follows the chronology of empirical data analysis: from data collection, through data exploration, to modeling and forecasting.

We start in Chapter 2 by discussing different ways for obtaining online auction data. In addition to the standard methods of data purchasing or collaborating with Internet businesses, we describe the currently most popular method of data collection: Web crawling and Web services. These two technologies generate large amounts of rich, high-quality online auction data. We also discuss Web data collection from a statistical sampling point of view, noting the various issues that arise in drawing data samples from a website, and how the resulting samples relate to the population of interest.

Chapter 3 continues with the most important step in data analysis: data exploration. While the availability of huge amounts of data often tempts the researcher to directly jump into sophisticated models and methods, one of the main messages of this book is that it is of extreme importance to first understand one’s data, and to explore the data for patterns and anomalies. Chapter 3 presents an array of data exploration methods and tools that support the special structures that arise in online auction data. One such structure is the unevenly spacing of time series (i.e., the bid histories) and their combination with cross-sectional information (i.e., auction details). Because many of the models presented in the subsequent chapters make use of an auction’s price...
evolution, we describe plots for displaying and exploring curves of the price and its dynamics. We also discuss curve clustering, which allows the researcher to segment auctions by their different price dynamics.

Another important facet is the concurrent nature of online auctions and their competition with other auctions. We present methods for visualizing the degree of auction concurrency as well as its context (e.g., collection period and data volume). We also discuss unusual data structures that can often be found in online auctions: semicontinuous data. These data are continuous but contain several “too-frequent” values. We describe where and how such semicontinuous data arise and propose methods for presenting and exploring them in Chapter 3.

The chapter continues with another prominent feature of online auction data: data hierarchies. Hierarchies arise due to the structure of online auction websites, where listings are often organized in the form categories, subcategories, and subsubcategories. This organization plays an important role in how bidders locate information and, ultimately, in how listings compete with one another.

Chapter 3 concludes with a discussion of exploratory tools for interactive visualization that allow the researcher to “dive” into the data and make multidimensional exploration easier and more powerful.

Chapter 4 discusses different statistical models for capturing relationships in auction data. We open with a more formal exposition of the price curve representation, which estimates the price process (or price evolution) during an ongoing auction. The price process captures much of the activity of individual bidders and also captures interactions among bidders, such as bidders competing with one another or changes in a bidder’s bidding strategies as a result of the strategies of other bidders. Moreover, the price process allows us to measure all of this change in a very parsimonious matter—via the price dynamics. Chapter 4 hence starts out by discussing alternatives for capturing price dynamics and then continues to propose different models for price dynamics. In that context, we propose functional regression models that allow the researcher to link price dynamics with covariate information (such as information about the seller, the bidders, or the product). We then extend the discussion to functional differential equation models that capture the effect of the process itself in addition to covariate information.

We then discuss statistical models for auction competition. By competition, we mean many auctions that sell similar (i.e., substitute) products and hence vie for the same bidders. Modeling competition is complicated because it requires the definition of “similar items.” We borrow ideas from spatial models to capture the similarity (or dissimilarity) of products in the associated feature space. But competition may be more complex. In fact, competition also arises from temporal concurrency: Auctions that are listed only a few minutes or hours apart from one another may show stronger competition compared to auctions that end on different days. Modeling temporal relationships is challenging since the auction arrival process is extremely uneven and hence requires a new definition of the traditional “time lag.”

Chapter 4 continues with discussing models for bidder arrivals and bid arrivals in online auctions. Modeling the arrival of bids is not straightforward because online auctions are typically much longer compared to their brick-and-mortar counterparts.
and hence they experience periods of little to no activity, followed by “bursts” of bidding. In fact, online auctions often experience “deadline effects” in that many bids are placed immediately before the auction closes. These different effects make the process deviate from standard stochastic models. We describe a family of stochastic models that adequately capture the empirically observed bid arrival process. We then tie these models to bidder arrival and bid placement strategies. Modeling the arrival of bidders (rather than bids) is even more challenging because while bids are observed, the entry (or exit) of bidders is unobservable.

Chapter 4 concludes with a discussion of auction networks. Networks have become omnipresent in our everyday lives, not the least because of the advent of social networking sites such as MySpace or Facebook. While auction networks is a rather new and unexplored concept, one can observe that links between certain pairs of buyers and sellers are stronger than others. In Chapter 4, we discuss some approaches for exploring such bidder–seller networks.

Finally, in Chapter 5 we discuss forecasting methods. We separated “forecasting” from “modeling” (in Chapter 4) because the process of developing a model (or a method) that can predict the future is typically different from retroactively building a model that can describe or explain an observed relationship.

Within the forecasting context, we consider three types of models, each adding an additional layer of information and complexity. First, we consider forecasting models that only use the information from within a given ongoing auction to forecast its final price. In other words, the first—and most basic—model only uses information that is available from within the auction to predict the outcome of that auction. The second model builds upon the first model and considers additional information about other simultaneous auctions. However, the information on outside auctions is not modeled explicitly. The last—and most powerful—model explicitly measures the effect of competing auctions and uses it to achieve better forecasts.

We conclude Chapter 5 by discussing useful applications of auction forecasting such as automated bidding decision rule systems that rely on auction forecasters.

1.5 DATA AND CODE AVAILABILITY

In the spirit of publicly (and freely) available information (and having experienced the tremendous value of rich data for conducting innovative research firsthand), we make many of the data sets described in the book available at http://www.ModelingOnlineAuctions.com. The website also includes computer code used for generating some of the results in this book. Readers are encouraged to use these resources and to contribute further data and code related to online auctions research.
2

OBTAINING ONLINE AUCTION DATA

2.1 COLLECTING DATA FROM THE WEB

Where do researchers get online auction data? In addition to traditional channels such as obtaining data directly from the company via purchase or working relationships, the Internet offers several new avenues for data collection. In particular, the availability of online auction data is much wider and easier compared to ordinary “offline” auction data, which has contributed to the large and growing research literature on online auctions. Because transactions take place online in these marketplaces, and because of the need to attract as many sellers and buyers, information on ongoing auctions is usually made publicly available by the website. Moreover, due to the need of buyers and sellers to study the market to determine and update their strategies, online auction websites often also make publicly available data on historical auctions, thereby providing access to large archival data sets. Different websites vary in the length of available history and the type of information made available for an auction. For example, eBay (www.eBay.com) makes publicly available the data on all ongoing and recently closed auctions, and for each auction the data include the entire bid history (time stamp and bid amount) except for the highest bid, as well as information about the seller, the auctioned item, and the auction format. In contrast, SaffronArt (www.saffronart.com), which auctions contemporary Indian art, provides past-auction information about the winning price, the artwork details, and the initial estimate of closed auctions, but the bid history is available only
during the live auction. On both eBay and SaffronArt websites, historical data can be accessed only after logging in.

When an online auction site makes data publicly available, we can use either manual or automated data collection techniques. Manual collection requires identifying each webpage of interest, and then copying the information from each page into a data file. Note that for a single auction there might be multiple relevant pages (e.g., one with the bid history, another with the item details, and another with the detailed seller feedback information). Early research in online auctions was often based on manually extracted data. However, this manual process is tedious, time consuming, and error prone. A popular alternative among eCommerce researchers today is to use an automated collection system, usually called a Web agent or Web crawler. A Web agent is a computer program, written by the researcher, that automatically collects information from webpages. Web agents mimic the operations that are done manually, but they do it in a more methodical and in a much faster way. Web agents can yield very large data sets within short periods of time.

Another automated mechanism for obtaining online auction data is using Web services offered by the auction website. A growing number of eCommerce websites offer users the option to download data directly from their databases. This is done via two computer programs that “talk” to each other: the researcher’s program requests information from the auction website’s server using an agreed-upon format.

Before describing in further detail the more advanced technologies for obtaining online auction data, we note that traditional methods for acquiring data still exist. We therefore start by describing traditional methods and then move on to more popular Web data collection techniques.

### 2.1.1 Traditional Data Acquisition

One option is to obtain data via a relationship with a company (e.g., via a consulting relationship). A second option is purchasing data directly from the auction house. eBay, for instance, sells data through third-party companies such as AERS (www.researchadvanced.com), Terapeak (www.terapeak.com), and HammerTap (www.hammertap.com). Such data can be richer than its publicly available data, but the cost and the type of bundling, which are geared toward businesses, can often be a limitation for academic researchers with no data acquisition funds. In addition, not every online auction website offers their data for sale.

A third, but currently still an unusual acquisition option, is direct download of data from an auction website. Prosper (www.prosper.com) is an online auction platform for loans, connecting individual borrowers with individual lenders (known as peer-to-peer lending). It offers a “Data Export” option where they “provide a complete snapshot of all listings, bids, user profiles, groups, and loans ever created on Prosper. This snapshot is updated nightly. [They] provide this data in a raw XML format. Additionally, [they] also provide the tools to import the XML data into the database of your choice” (see www.prosper.com/tools).

Finally, another traditional option is to use data that are shared by other researchers. Accompanying this book will be several data sets of online auctions that we have
collected and used in various studies. We have also greatly benefited from similar sharing of data by other colleagues. Online auction data do not pose confidentiality concerns, and therefore we hope that researchers will be inclined to share their collected data.

In the following, we describe two technologies that are the more widely used methods for obtaining online auction data: Web crawling and Web services.

2.1.2 Web Crawling

A Web crawler is a program or automated script that browses the Internet in a methodical, automated manner. Web crawling (also known as “screen scraping”) is a popular technique among online auction researchers for obtaining data. The researcher writes a set of computer programs that collect data from the relevant webpages and store them locally. Current popular programming languages for Web crawling are PHP, Perl, and Windows scripting languages. In general, any computer language that can access the Web can be used for Web crawling. In all cases, basic knowledge of HyperText Markup Language (HTML), which is the basic and predominant markup language in which webpages are written, is required. There are three general steps involved in collecting data via Web crawling:

1. **Identifying the Webpages of Interest (Creating the List of URLs):** This is done by browsing the auction website and identifying the structure of the URLs (the web address) of interest. For example, to search for all auctions for a particular book, use the website’s “search” function to obtain a results page that contains all the items of interest. The URL of this search result can then be the starting point for the Web crawler. Note that this page will include hyperlinks to the individual pages of each book auction.

2. **Collecting and Saving Webpages Locally:** Auction websites are highly dynamic. New bids arrive constantly, new auctions are being posted, new closed auctions are constantly added to the historical archives, while older auctions are no longer publicly available. Even other components, such as policies regarding bidding or selling rules, or how ratings are given, can and do change. Because of this dynamic nature, it is important to save all pages of interest to one’s local computer. The same information might no longer exist later on, or might be repackaged in a different format that would require new programming. The information might even change during the data extraction phase, such that earlier visited pages would contain different data than later visited pages. The advantages of locally saved pages include not only quicker data extraction but also the possibility of extracting further information from the pages in the future (e.g., for the purpose of further research or for revising a paper), as well as for data sharing.

3. **Parsing the Information from the Webpages into a Database:** This last stage relates to the extraction of the data from the locally saved pages. A program is written that identifies the fields of interest on a page and extracts only the
OBTAINING ONLINE AUCTION DATA

requested information into a database or some other file format (e.g., comma separated values (CSV) or Excel files). Identifying the location of the information of interest on a webpage requires careful examination of its HTML code, and the specification of character strings that uniquely identify the locations of interest. A more efficient approach uses Regular Expressions (RegExp), which is based on string matching, to find and extract data from strings (thereby avoiding the need for loops and instead extracting the information in one sweep). For more information on RegExp, see, for example, Goyvaerts and Levithan (2009) or www.regexp.info.

To illustrate these three stages, consider the collection of data from eBay on closed auctions for Amazon.com gift certificates. To identify the webpage that includes the list (and links) for the auctions of interest, we search for the keyword “Amazon” within eBay’s “Gift Certificate” category. The partial search results page is shown in Figure 2.1. The web address (URL) of this results

FIGURE 2.1  eBay search results page showing some of the closed auctions for Amazon.com gift certificates (obtained on December 15, 2009).
COLLECTING DATA FROM THE WEB

page is http://completed.shop.ebay.com/Gift-Certificates-31411/i.html?LH_Complete=1&_nkw=Amazon&_dmp=US_Gift_Certificates&_pgn=1. The URL includes an indication that these are closed auctions (“completed”), within the Gift Certificate category (in the United States), and that they include the keyword “Amazon.” The URL also indicates that this is the first results page, (pgn=1). If you were using a browser, then you could click the “Next” page or a specific page number to reach other results pages. However, noting that the page number is part of the URL (in this example and in many other cases) allows us to easily infer the URLs of the next search pages, which can be used in a faster construction of the pages to be visited by the Web crawler. The next step is therefore to save the search results pages locally. Then, the results pages are used to extract the URLs of the individual auctions. To see how the information can be extracted (or “parsed”), examine the partial HTML code for this same results page, as shown in Figure 2.2. The location of the information on the first three auction is marked, and the URL of each of the corresponding auction pages follows 2–3

FIGURE 2.2 Part of the HTML code generating the page shown in Figure 2.1. The first three auction titles are marked.

1 Viewing the HTML source code of a webpage can be done by using the browser’s “View Source” or “View Page Source” option.
OBTAINING ONLINE AUCTION DATA

In the next stage, all the individual auction pages are obtained and saved locally. In our example, there are at least two pages of interest for each auction, as shown in Figure 2.3. The first is the general “item description” page with information about the item, the closing time and date, seller information, and so on (top panel), and the second page of related information is the detailed bid history (bottom panel), which is hyperlinked from item description page. Note that the URLs of these two pages (http://cgi.ebay.com/ws/eBayISAPI.dll?ViewItem&item=300377726460 and http://offer.ebay.com/ws/eBayISAPI.dll?ViewBids&item=300377726460, respectively) are also informative, as they include the unique Item ID number. They can be used in the Web crawling script as well.

Finally, after the set of auction pages has been collected and saved locally, a separate program is written to extract the information from these pages. For instance, if we want to collect the time stamps and bid amounts from the bid history page, we must identify the location of these values within the HTML code, and then extract them accordingly. Figure 2.4 shows part of the HTML code for the bid history, where three bid amounts and their corresponding dates and times are marked.

Advantages and Challenges  Web crawling is currently the most popular method being used by academic researchers who study online auctions. The biggest advantage of Web crawling is its flexibility and automation. Once the researcher has written the script and has made sure that it works correctly, data extraction is usually quick and the quality of data is high. Web crawling is also still cheaper than purchasing data or using Web services. All these reasons contribute to the popularity of Web crawling in online auction research.

In some cases, Web crawling is the only way to obtain data. For example, if the website only displays data on ongoing auctions (and does not provide any tools for obtaining the data directly), then Web crawling during ongoing auctions can be used for collection. Bapna et al. (2006) describe a Web crawler that obtains information throughout the live auction: “to track the progress of the auction and the bidding strategies the same hypertext page has to be captured at frequent intervals until the auction ends.”

The challenges in Web crawling are technological and legal/ethical. From a technical point of view, Web crawling requires some knowledge of programming in a Web scripting language, and data quality relies on robustness of the website and the researcher’s programming skills. Extracting data from different websites, and sometimes even from different areas within a particular auction website, requires different programs. Also, changes in the website structure or access require updating the code accordingly. In terms of access, although many websites make their data publicly available, it does not mean that the data are freely available for download. When an auction site (or any website in general) does not want to allow Web crawling, it can proceed in several ways. One way is by implementing technological barriers that make the crawling difficult (e.g., by requiring a user login or information that is normally sent by the user’s browser (“cookies”), blocking IP addresses that send
FIGURE 2.3 Item description page (top) and bid history page (bottom) for item #3 in Figure 2.1.
many requests, or posing a question that only a human can solve (CAPTCHA)). Another approach is stating that crawling is not allowed. The latter is done by placing a file called robots.txt on the main page of the website (e.g., on eBay the URL is http://eBay.com/robots.txt) and/or by mentioning restrictions in the website’s “Terms and Conditions” page. There are several reasons why a website would restrict data crawling. One reason is to avoid overloading their servers, thereby potentially slowing down the service to the intended users. Web crawling therefore uses the website’s resources, which are intended to be for sellers and buyers. Restricting Web crawling is also intended against competitor websites or parties, who might use the data for purposes of competition. Although it is unlikely that Web crawling restrictions are aimed at academic researchers, there are ethical and legal issues to consider. Allen et al. (2006) describe the various issues related to Web crawling for academic purposes and suggest guidelines for acceptable use. Additional suggestions are given in Bapna et al. (2006).

2.1.3 Web Services

A growing number of websites are now providing archival data to interested users by allowing querying their database. Among the online auction websites, eBay