Identification of Auction Models With an Unknown Number of Bidders

Unjy Song

Seoul National University

May, 2015

Abstract

This paper derives new identification results and proposes an estimation strategy for auction models with an unknown number of potential bidders. I first show that within the symmetric independent private values (IPV) model, the potential bidders' value distribution is nonparametrically identified and consistently estimated from observations of any pair of valuations of which rankings from the top or the bottom are known. The results hold even if the number of potential bidders follows a different distribution across auctions. If the distribution of the number of potential bidders is the same across auctions, it is also identified only from a pair of valuations. I then apply new identification results to establish identification of various auction models of eBay, ascending, second-price sealed-bid, and first-price sealed-bid auctions, including models with a binding reserve price and entry costs.

Key Words Auctions, identification, unknown number of bidders, ranked statistics

1 Introduction

This paper studies nonparametric identification and estimation of auction models when the number of potential bidders is unknown. The number of potential bidders is not easily available in many auctions because not all potential bidders take part in an auction. A seller may set a binding reserve price or a potential bidder may have to incur entry costs in order to participate in an auction. Moreover it is difficult to recognize existence of even participating bidders if they do not make a bid. In eBay auctions, for example, an auction price rises as a new bid is placed. Consequently one can never know the presence of bidders who intended to take part in auctions, but visited the auction sites only to find that the auction prices were already raised by their competitors over their own willingness-to-pay. In addition, auctioneers do not always make records of all the bidders or their bids.

Identification depends on observables and the model which yields a mapping between observables and model primitives of interest. In this paper I consider various auction models within the standard symmetric independent private values (IPV) framework. Model primitives of interest are the distribution of the number of potential bidders and their value distribution. With these two model primitives we can predict the outcomes of auctions. Regarding observables, I do not consider observables other than bid data which are available most commonly in auctions. I first derive new identification results and propose an estimation strategy that can be used for various auction formats to recover the potential bidders' value distribution without knowledge nor an assumption regarding the number of potential bidders. I show that the potential bidders' value distribution is nonparametrically identified and consistently estimated from observations of any pair of valuations of which rankings from the top or the bottom are known, for example the second- and third-highest valuations. The results hold even if the number of potential bidders follows a different distribution across auctions. If the distribution of the number of potential bidders is the same across auctions, it is also identified only from a pair of valuations. Next I apply new identification results to establish identification of various auction models of eBay, ascending, second-price sealed-bid, and first-price sealed-bid auctions. The new identification results contribute most to identification of auction models with entry where not all potential bidders take part in an auction. If not all potential bidders take part in the auction, there is a discrepancy between the number of potential bidders and the number of participating bidders. A typical endogeneity issue is raised when the exogenous number of potential bidders is inferred from the endogenous number of participating bidders. Since the new identification results do not require knowledge of the number of potential bidders, the results have strength to establish identification of auction models with entry without suffering from endogeneity problems or measurement errors.

Prior research has suggested solutions for the case where the number of potential bidders is unavailable, but almost all of the previous studies instead require observation of all the bids over the reserve price or at least knowledge of the number of potential bidder willing to pay the reserve price. Otherwise, the previous studies require a parametric distributional assumption or observables beyond bid data such as a measure of potential competition and its instrument. Laffont, Ossard, and Vuong (1995) propose a method to estimate the number of potential bidders and their value distribution only from transaction prices. But if the number of potential bidders is unavailable, identification in Laffont, Ossard, and Vuong (1995) counts on a parametric distributional assumption regarding potential bidders' valuations with an assumption that the number of bidders is constant. Although Bajari and Hortaçsu (2002a) model the unknown number of potential bidders as stochastic, they assume that the observed bidders are all potential bidders willing to pay the reserve price with a parametric distributional assumption regarding potential bidders' valuations. Paarsch (1997) also takes a parametric approach when he analyzed ascending auctions with a binding reserve price. Nonparametric approaches have been proposed by several papers, for instance, Athey and Haile (2002), Guerre, Perrigne, and Vuong (2000), Haile and Tamer (2002), and Hendricks, Pinkse, and Porter (2003). They all need knowledge of the number of potential bidders willing to pay the reserve price. Moreover if the binding reserve price is varied across auctions, the existing nonparametric estimation methods need an additional restriction on how the reserve price is set. An, Hu, and Shum (2010) exploit econometric results on models with misclassification error to develop a nonparametric estimation method with allowing the unknown number of potential bidders to vary across auctions. The method proposed in An, Hu, and Shum (2010) need a noisy measure of the number of potential bidders and its instrument. Hence their method does not always require knowledge of the number of potential bidders willing to pay the reserve price although An, Hu, and Shum (2010) use the number of potential bidders willing to pay the reserve price as a noisy measure of the number of potential bidders in their application and Monte Carlo simulations.

Given the results of the existing research, the econometric methods proposed in this paper are especially useful for nonparametric approaches to auctions where the number of potential bidders willing to pay the reserve price is not available. As explained before not all potential bidders willing to pay the reserve price, make a bid in many ascending auctions such as eBay auctions. Even in sealed-bid auctions not all potential bidders willing to pay the reserve price, make a bid if there are bid preparation costs as modeled in Samuelson (1985). In those auctions it is difficult to know the number of potential bidders willing to pay the reserve price. Moreover the nonparametric estimation method proposed in this paper is applicable to eBay, ascending, second-price sealed-bid auctions with varied reserve prices without an additional assumption regarding the reserve price.

The rest of the paper is organized as follows: The next section presents the basic model to be assumed throughout the paper. In Section 3, I develop econometric methods applicable for identification and estimation of various auction models. For that I newly define ranked statistics, variant of order statistics, and study their properties. In Section 4, I apply the properties of ranked statistics to derive new identification results of various auction models. Although the focus is on identification, I also mention an estimation method of the potential bidders' value distribution. The estimation method is applicable immediately to eBay, ascending, and second-price sealed-bid auctions but it needs more work to apply to first-price sealed-bid auctions. In Section 5, I conduct Monte Carlo experiments to show performance of the estimation method. Finally I make concluding remarks in Section 6.

2 Basic Model: Symmetric IPV Model

Throughout the paper, I represent random variables in upper case and their realizations in lower case. I consider an auction of single indivisible good. The number of potential bidders, N, is a random variable, with $p_n = \Pr(N = n)$, $n = 0, 1, ..., l$. Each potential bidder's valuation V^i $(i = 1, ..., n)$ is an independent draw from the absolutely continuous distribution $F(\cdot)$, having support on $[\underline{v}, \overline{v}]$, $\underline{v} > 0$. The random variable N and V^i are independent. Each bidder knows his valuation, the distribution $F(\cdot)$, and the probabilities $\{p_n\}_{n=0}^{l}$, but does not know how many competitors he is facing when he makes a bid. I do not specify a potential bidder's entry decision nor auction formats until Section 4. In any auction formats, I will consider symmetric, increasing, and differentiable Bayesian-Nash equilibria.

Model primitives of interest are $F(\cdot)$ and the distribution of N. A typical approach of structural analysis is started by assuming that the distribution of N is known or easily estimated from the observed number of bidders. Given the distribution of N, the potential bidders' valuations or their estimates are obtained from the observed bids by inverting the equilibrium bidding function. Then $F(\cdot)$ is estimated from the potential bidders' valuations or their estimates. However in many real-world auctions, the distribution of N is not easily estimated from the number of observed bidders because not all potential bidders make a bid. In this paper I analyze cases where the the distribution of N is neither known nor immediately estimated from the observed number of bidders. Therefore observables will be only part of potential bidders' bids.[1]

I consider a dataset which consists of T number of independent auctions. I assume that potential bidders' value distribution, $F(\cdot)$, is fixed in all auctions under consideration. Standard arguments extend the IPV model to the model where bidders' valuations are i.i.d. conditional on observable auction characteristics. Hence, to avoid unnecessary complexity, the model maintains the IPV assumption. But in the Monte Carlo simulations, bidders' value distribution is allowed to vary with observable auction characteristics.

[1] If all potential bidders' bids are observed, the number of potential bidders is obviously the number of observed bids.

3 Econometric Methods

Athey and Haile (2002) show that in the symmetric IPV model, one order statistic of potential bidder's valuation nonparametrically identifies the potential bidders' value distribution, when the number of potential bidders is known. Athey and Haile (2002) established their identification result by applying the known property of order statistics: the distribution of one order statistic characterizes the parent distribution when the sample size is known. The sample size is corresponding to the number of potential bidders in the auction model, and the parent distribution is corresponding to the potential bidders' value distribution. If we do not know the number of potential bidders, the identification issue eventually reduces to a statistical question of whether a parent distribution is uniquely determined by the distribution of its order statistics from a sample, of which the size is unknown. Since the order statistic is normally defined by its parent distribution and sample size, we need a new name for the order statistic of which the sample size is unknown to make a distinction. Hence I name the order statistic with an unknown sample size *ranked statistic* and study its properties regarding characterization of the parent distribution and the distribution of the sample size to ultimately study identification of auction models with an unknown number of potential bidders.

Definition 1 *Let X_1, $X_2, ..., X_N$ be N independent draws from an absolutely continuous distribution $F(\cdot)$ with an associated density f. A random variable, N and X_i $(1 \leq i \leq N)$ are independent. The support of $F(\cdot)$ is $[\underline{x}, \overline{x}]$, and $\underline{x}$ and $\overline{x}$ are allowed to take $-\infty$ and ∞ respectively. The sample size, N is known to take a nonnegative, finite integer but no realized value of N is known. Let $X^{(1:N)}$, $X^{(2:N)}, ..., X^{(N:N)}$ be a rearrangement of $\{X_i\}_{i=1}^N$ so that $X^{(N:N)} \geq X^{(N-1:N)} \geq ... \geq X^{(1:N)}$. Then I define high-ranked statistic and low-ranked statistic as follows.*

1. High-ranked statistic: the kth high-ranked statistic, $X_k^{(H)} = X^{(N-k+1:N)}$ where $k = 1, 2, .., N$

2. Low-ranked statistic: the kth low-ranked statistic, $X_k^{(L)} = X_k^{(k:N)}$ where $k = 1, 2, .., N$

Remark 1 *Although no realized value of N is known, existence of the kth high- or low-ranked statistic implies that the realized value of N is no less than k.*

To discover properties of the ranked statistics, I start by presenting the known results regarding the distribution of order statistics. The ith order statistic and jth order statistic ($1 \leq i < j \leq n$) from an i.i.d. sample of size n from an absolutely-continuous distribution $F(\cdot)$ have a joint probability density function (PDF)

$$g^{(i,j:n)}(x,y) = \begin{cases} \frac{n![F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j}f(x)f(y)}{(i-1)!(j-i-1)!(n-j)!}, & y > x \\ 0, & otherwise. \end{cases} \tag{1}$$

The *ith* order statistic has a PDF

$$g^{(i:n)}(x) = \frac{n!}{(i-1)!(n-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-i}f(x) \tag{2}$$

and the CDF

$$G^{(i:n)}(x) = \frac{n!}{(i-1)!(n-i)!}\int_0^{F(x)} t^{i-1}(1-t)^{n-i}dt \tag{3}$$

(See, for example Arnold, Balakrishnan, and Nagarajaet (1992)).

One order statistic identifies its parent distribution for every known sample size. Hence a natural starting point would be to see if one ranked statistic also identifies its parent distribution. The distribution of an order statistic or a ranked statistic is decided by the sample size and the parent distribution. Thus in the case of the order statistic of which the sample size is known, it is intuitive that the same distribution of the order statistic implies the same parent distribution. However, in the case of the ranked statistic of which the sample size is unknown, the same distribution of the ranked statistic may not imply the same parent distribution because the same distribution of a ranked statistic can be generated by different combinations of a parent distribution and a sample size. An example below shows that this is actually the case. Consider two underlying structures:

$$\text{(i) } F(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x < 1 \\ 1, & x \geq 1 \end{cases}, \; N = 3; \text{ and}$$

$$\text{(ii) } F(x) = \begin{cases} 0, & x \leq 0 \\ -\sqrt{2x^3 - 3x^2 + 1} + 1, & 0 < \; x < 1 \\ 1, & x \geq 1 \end{cases}, \; N = 2$$

where $F(x)$ is the parent distribution and N is the sample size. Both structures generate the same distribution of the second high-ranked statistic:

$$F(x)^{(N-1:N)} = \begin{cases} 0, & x \leq 0 \\ x^2(3 - 2x), & 0 < x < 1 \\ 1, & x \geq 1 \end{cases}.$$

Therefore, only the second high-ranked statistic cannot identify its parent distribution. A similar counter example can be constructed for the case where other ranked statistic is available.

Given this non-identification result, it is natural to see whether or not we can identify the parent distribution and/or the distribution of the sample size if we have another ranked statistic available. If we have more than one ranked statistic, their joint distribution is also decided by both the parent distribution and the unknown sample size. Hence it is still not obvious whether or not we can discriminate between changes in the parents distribution and changes in the distribution of the sample size from changes in the joint distribution of ranked statistics. In the remainder of this section, I will first show that a pair of ranked statistics nonparametrically identify their parent distribution and propose a consistent estimator of the parent distribution by using the semi-nonparametric maximum likelihood (SNP) estimation method. Next I will show that a pair of ranked statistics nonparametrically identify the distribution of the sample size as well.

3.1 Identification of the Parent Distribution

Theorem 1 *An arbitrary absolutely-continuous distribution $F(\cdot)$ is nonparametrically identified by a pair of high-ranked or low-ranked statistics from an i.i.d. sample, even when the distribution of the sample size is unknown and different across samples.*

Proof. Here I present only the proof for the case where a pair of high-ranked statistics are available, while the proof for the case where a pair of low-ranked statistics are available is provided in Appendix A. The lower limit of support of $F(\cdot)$ is denoted by $\underline{v}$, and $f(\cdot)$ is an associated density of $F(\cdot)$. It is allowed for $\underline{v}$ to take $-\infty$. The sample size is represented by a random variable N. Let Y denote the k_1th high-ranked statistic, which is the $(N-k_1+1)$th order statistic, and X denote the k_2th high-ranked statistic, which is the $(N-k_2+1)$th order statistic $(1 \leq k_1 < k_2 \leq N)$. For the sake of notational convenience, let $F(\cdot|x)$ denote the distribution of $F(\cdot)$ truncated from below at x, and $f(\cdot|x)$ denote an associated density.

Consider (X, Y) from the same sample. The density of Y conditional on X, $p_{(k_1|k_2)}(y|x)$, for $y \geq x$, is computed by applying Equation (1) and (2):

$$\begin{aligned}
p_{(k_1|k_2)}(y|x) &= \frac{(k_2-1)!}{(k_2-k_1-1)!(k_1-1)!} \\
&\times \frac{[F(y)-F(x)]^{k_2-k_1-1}[1-F(y)]^{k_1-1}f(y)}{(1-F(x))^{k_2-1}} \cdot I_{\{y\geq x\}} \\
&= \frac{(k_2-1)!}{(k_2-k_1-1)!(k_1-1)!} \\
&\times \frac{[(1-F(x))F(y|x)]^{k_2-k_1-1}[(1-F(x))(1-F(y|x))]^{k_1-1}f(y|x)(1-F(x))}{(1-F(x))^{k_2-1}} \cdot I_{\{y\geq x\}} \\
&= \frac{(k_2-1)!}{(k_2-k_1-1)!\{(k_2-1)-(k_2-k_1)\}!} \\
&\times F(y|x)^{k_2-k_1-1}(1-F(y|x))^{(k_2-1)-(k_1-1)}f(y|x) \cdot I_{\{y\geq x\}} \\
&= f^{(k_2-k_1:k_2-1)}(y|x) \cdot I_{\{y\geq x\}}.
\end{aligned} \tag{4}$$

Equation (4) means that the density of k_1th high-ranked statistic conditional on the k_2th high-ranked statistic is the same as the density of the (k_2-k_1)th order statistic of which the sample size is $(k_2 - 1)$ and of which the parent distribution is $F(\cdot|x)$ for all x. Now note that $\lim\limits_{x \to \underline{v}} p_{(k_1|k_2)}(y|x) = \lim\limits_{x \to \underline{v}} f^{(k_2-k_1:k_2-1)}(y|x) \cdot I_{\{y \geq x\}} = f^{(k_2-k_1:k_2-1)}(y)$. This implies that the density of k_1th high-ranked statistic conditional on the k_2th high-ranked statistic identifies the density of the $(k_2 - k_1)$th order statistic of which the parent distribution is $F(\cdot)$ and the sample size is $(k_2 - 1)$. Since the distribution of any order statistic with a known sample size identifies its parent distribution, the result follows. ■

The key insight into Theorem 1 is captured by Equation (4). The joint distribution of high-ranked statistics depends on both the sample size and the parent distribution. However, if the distribution of a high-ranked statistic is conditioned on a lower high-ranked statistic, that conditional distribution depends only on the parent distribution without depending on the sample size. Accordingly we can identify the parent distribution from the conditional distribution without knowledge of the sample size. Moreover the sample size does not have to follow the same distribution across samples for identification of the parent distribution. To see the intuition more concretely, consider a triplet (Z, Y, X) which are the first, the second, and the third high-ranked statistic from a sample which is drawn randomly from an absolutely continuous distribution $F(\cdot)$. We observe realized values of Y and X, but do not observe any realized values of Z. Consider an ordered sample which consists of Z and Y. The joint distribution of Z and Y depends on both $F(\cdot)$ and the size of the sample that (Z, Y, X) originally comes from. However if we analyze a sample of (Z, Y) conditional on that $X = x$, we can regard (Z, Y) as a random sample of which the size is *two* and the parent distribution is $F(\cdot)$ truncated from below at x whatever size of the sample from which the triplet (Z, Y, X) originally comes. Hence, regardless of the original sample size, the distribution of Y conditional on $X = x$ is the same as the distribution of the first order statistic of which the sample size is *two* and the parent distribution is $\frac{F(\cdot)-F(x)}{1-F(x)}$. Hence, the distribution of Y conditional on $X = x$ identifies $\frac{F(\cdot)-F(x)}{1-F(x)}$ for all x on the support of $F(\cdot)$. Letting x be the lower limit of the support of $F(\cdot)$ gives the identification of $F(\cdot)$. I give an example of Exponential distribution as follows.

Example 1 *Let* $F(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp(-\frac{x}{\theta}), & x \geq 0 \end{cases}$

with an associated density $f(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{\theta}\exp(-\frac{x}{\theta}), & x \geq 0 \end{cases}$.

Let's consider a pair of high-ranked statistics, the first and the second high-ranked statistics, Z *and* Y*. The joint probability density of* Z *and* Y *when the sample size is* n*:*

$$p(z, y) = \frac{n(n-1)}{\theta^2}\{1 - \exp(-\frac{y}{\theta})\}^{n-2}\exp(-\frac{y+z}{\theta}),$$

which depends on both the sample size, n *and the parameter of the parent distribution,* θ*. However, the probability density of the first high-ranked statistic conditional on the second high-ranked statistic is:*

$$p(z|Y = y) = \frac{1}{\theta}\exp(-\frac{z-y}{\theta}),$$

which depends only on the parameter of the parent distribution, θ*. Hence we can identify* θ *from the conditional density. In particular, the conditional expectation of the first high-ranked statistic on the second high-ranked statistic is:*

$$E(Z|Y = y) = \int_y^\infty z \cdot \frac{1}{\theta}\exp(-\frac{z-y}{\theta})dz = y + \theta.$$

Therefore the difference between the sample means of the first and the second high-ranked statistic is a consistent estimator of θ*; this estimator works even when the distribution of the sample size varies across samples because the conditional expectation does not depend on the sample size. Similarly, the conditional expectation of the first and the second high-ranked statistic on the third high-ranked statistic,* X *are:* $E(Y|X = x) = x + \frac{1}{2}\theta$*; and* $E(Z|X = x) = x + \frac{3}{2}\theta$*. Hence either the twice of the sample mean of the differences between the second and the third high-ranked statistics or two thirds of the sample mean of the differences between the first and the third high-ranked statistics is a consistent estimator of* θ*.*

3.2 Estimation of the Parent Distribution

For a consistent estimator of the parent distribution, $F(\cdot)$ having support on $[\underline{v}, \overline{v}]$, I employ the semi-nonparametric (SNP) method developed by Gallant and his coauthors.[2] Gallant and Nychka (1987) showed that simply replacing an unknown density with a Hermite form and applying the standard, finite dimensional maximum likelihood methods yields consistent estimators of model parameters and nearly all aspects of the unknown density itself, provided that the length of the series increases with the sample size. The rule for increasing series length can be data-dependent. The joint probability density of a pair of ranked statistics includes the unknown sample size. Hence I employ a partial likelihood using a conditional density instead of the full likelihood. In particular, I use $p_{(k_1|k_2)}(y|x)$ in Equation (4) as a likelihood function.

Below, as an example, I illustrate the estimation method for the case where observations of the second and the third high-ranked statistics are available. Song(2015a) analyzed that case for an empirical study of eBay auctions. Let (Y_t, X_t) denote the second and the third high-ranked statistics at samples $t = 1, 2, ..., T$. Let $C = \min_t X_t$. Since no information about $F(v)$ for $v < c$ can be discovered from a dataset of which the minimum is c, I treat $F^*(\cdot) = \frac{F(\cdot)-F(c)}{1-F(c)}$ as the model primitive of interest for given dataset.[3] Let $f^*(\cdot)$ denote an associated density of $F^*(\cdot)$. The density of Y_t conditional on X_t, $p^*(y_t|X_t = x_t)$, is calculated by substituting 3 for k_2, and 2 for k_1 in Equation (4):

$$p^*(y_t|X_t = x_t) = \frac{2[1 - F^*(y_t)]f^*(y_t)}{[1 - F^*(x_t)]^2} \quad \text{for } y_t \geq x_t \geq c.$$

The proof of Theorem 1 implies that $p^*(y|x)$ characterizes $F^*(v)$, when the lower limit of the support of $F^*(v)$ is identified. By construction, c is the lower limit of the support of

[2] See Gallant and Nychka (1987), Fenton and Gallant (1996a,b), and Coppejans and Gallant (2002).

[3] Since C is a consistent estimator of $\underline{v}$, a consistent estimator of $F^*(\cdot)$ becomes a consistent estimator of $F(\cdot)$. However it does not matter in practice whether we distinguish between $F^*(\cdot)$ and $F(\cdot)$. Because the lower part of the bidders' value distribution makes few effects on the auction price, estimation of $F^*(\cdot)$ is sufficient for study of most economic issues in auction markets.

$F^*(v)$. Accordingly, I consider the sample log likelihood function as follows:

$$L_T(\widehat{f}) = \frac{1}{T}\sum_{t=1}^{T} \ln \frac{2[1-\widehat{F}(y_t)]\widehat{f}(y_t)}{[1-\widehat{F}(x_t)]^2} \qquad (\widehat{F}(x) = \int_c^x \widehat{f}(t)dt).$$

Then $L_T(\widehat{f})$ is uniquely maximized at $\widehat{F}(v) = F^*(v)$.

Gallant and Nychka(1987) showed that any density that has a moment generating function can be used in a Hermite form and thus a natural choice would be the normal distribution truncated from below at c. But the appropriate density function will be different according to applications in hand. Below I illustrate specification of $\widehat{f}(x)$ and a sample log likelihood function when a truncated normal density function is chosen:

$$\widehat{f}(x) = \frac{[1 + a_1(\frac{x-\mu}{\sigma}) + ... + a_k(\frac{x-\mu}{\sigma})^k]^2 \phi(x;\mu,\sigma,c)}{\int_c^\infty [1 + a_1(\frac{x-\mu}{\sigma}) + ... + a_k(\frac{x-\mu}{\sigma})^k]^2 \phi(x;\mu,\sigma,c)dx}$$

where $\phi(x;\mu,\sigma,c)$ is the density of the Normal(μ,σ) truncated from below at c. An estimator, $\widehat{f_T}$, is the maximizer of $L_T(\widehat{f})$. Therefore,

$$\widehat{f_T}(x) = \frac{[1 + \widehat{a_1}(\frac{x-\widehat{\mu}}{\widehat{\sigma}}) + ... + \ \widehat{a_k}(\frac{x-\widehat{\mu}}{\widehat{\sigma}})^k]^2 \phi(x;\widehat{\mu},\widehat{\sigma},c)}{\int_c^\infty [1 + \widehat{a_1}(\frac{x-\widehat{\mu}}{\widehat{\sigma}}) + ... + \ \widehat{a_k}(\frac{x-\widehat{\mu}}{\widehat{\sigma}})^k]^2 \phi(x;\widehat{\mu},\widehat{\sigma},c)dx}$$

such that

$$(\widehat{a_1}, ..., \ \widehat{a_k}, \widehat{\mu}, \widehat{\sigma}) = \underset{a_1,..., \ a_k, \mu \in R, \ \sigma > 0}{\arg\max} \ L_T(\widehat{f}) = \frac{1}{T}\sum_{t=1}^{T} \ln \frac{2[1-\widehat{F}(y_t)]\widehat{f}(y_t)}{[1-\widehat{F}(x_t)]^2}.$$

In applications, the optimal series length, k^* can be chosen following the method proposed in Coppejans and Gallant (2002). They consider a cross-validation strategy, which employs the *ISE*[Integrated Squared Error] criteria. When $\widehat{h}(x)$ is a density estimate of $h(x)$, the ISE criterion is defined as follows:

$$\text{ISE}(\widehat{h}) = \int \widehat{h}^2(x)dx - 2\int \widehat{h}(x) \ h(x)dx + \int h^2(x)dx$$

In this subsection I proposed an estimation method of $F(\cdot)$ using only minimal data necessary for identification. If more than two ranked-statistics are available, it would enhance efficiency adjusting an estimation strategy to use all ranked-statistics available.

For example, Kim and Lee (2014) used the second, the third, and the fourth high-ranked statistics to estimate the parent distribution. In their estimation process, they considered the joint density function of the second and the third high-ranked statistics conditional on the fourth high-ranked statistic.

3.3 Identification of the Distribution of the Sample Size

The previous subsections show that we can identify and estimate nonparametrically the parent distribution only from a pair of ranked statistics even if the unknown sample size follows a different distribution across samples. Theorem 2 shows that as long as the sample size follows the same distribution across samples, that same distribution is also identified from a pair of ranked statistics.

Theorem 2 *Consider a pair of ranked statistics, the ith and the jth high[low]-ranked statistics (where $1 \leq i < j$), of which the parent distribution is $F(\cdot)$ and of which the sample size is represented by a random variable N. The sample size N takes a nonnegative integer and there exists a finite integer l such that $\Pr(N > l) = 0$. When $N \geq j$, a pair of the ith and the jth high[low]-ranked statistics exist; and when $N = i$, the ith high[low]-ranked statistic exists but the jth high[low]-ranked statistic does not. Assume that $F(\cdot)$ is absolutely continuous and i and j were chosen such that $\Pr(N \geq j) > 0$. Then $\{\Pr(N = n | N \geq i)\}_{n \geq i}$ are identified if either*

(a) the ith and the jth high-ranked statistics; or

(b) the ith and the jth low-ranked statistics

are observed.

Proof. As in Theorem 1, I present here only a proof for the case where a pair of high-ranked statistics are available while a proof for the case where a pair of low-ranked statistics are available is provided in Appendix A. Throughout the proof, note that $F(x)$ and its associated density $f(x)$ are identified for all x by applying Theorem 1.

Let $p_n = \Pr(N = n|N \geq i)$ where $n = i, ..., l$. Let X denote the ith high-ranked statistic and $G(x)$ denote the cumulative distribution function of X. Then

$$G(x) = \sum_{n=i}^{l} p_n F^{(n-i+1:n)}(x). \tag{5}$$

A. If $i = 1$, identification is easily established. Plugging $i = 1$ into Equation (5) yields:

$$G(x) = p_1 F(x)^1 ... + p_l F(x)^l.$$

The first high-ranked statistic identifies its cumulative distribution function, $G(x)$ for all x. Given identification of $F(x)$, the p_n, the coefficient on $F(x)^n$ is identified for $n = 1, ..., l$.

B. For a proof for the case where $i \geq 2$, I consider $g(x)$, an associated density of $G(x)$. By applying Equation (2) & (5) and rearranging terms lead to:

$$\begin{aligned}
g(x) &= \sum_{n=i}^{l} p_n f^{(n-i+1:n)}(x) \\
&= \sum_{n=i}^{l} p_n \cdot \frac{n!}{(n-i)!(i-1)!} F(x)^{n-i} (1 - F(x))^{i-1} f(x) \\
&= f(x)(1 - F(x))^{i-1} \sum_{n=i}^{l} p_n \cdot \frac{n!}{(n-i)!(i-1)!} F(x)^{n-i} \\
&= f(x) \left[\sum_{k=0}^{i-1} \binom{i-1}{k} (-F(x))^k \right] \left[\sum_{k=0}^{l-i} p_{i+k} \cdot \frac{(i+k)!}{k!(i-1)!} F(x)^k \right] \\
&= f(x) \left[\sum_{s=0}^{l-1} \left\{ \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^k \cdot I_{\{i \leq i+(s-k) \leq l\}} \cdot p_{i+(s-k)} \cdot \frac{\{i+(s-k)\}!}{(s-k)!(i-1)!} \right\} \cdot F(x)^s \right] \\
&= f(x)[(ip_i) \cdot F(x)^0 + \ \{i(i+1)p_{i+1} - (i-1)ip_i\} \cdot F(x) + \cdots \\
&\quad + \{(i-1)(-1)^{i-2} \frac{l!}{(l-i)!(i-1)!} p_l + (-1)^{i-1} \frac{(l-1)!}{(l-i-1)!(i-1)!} p_{l-1}\} \cdot F(x)^{l-2} \\
&\quad + \{(-1)^{i-1} \frac{l!}{(l-i)!(i-1)!} p_l\} \cdot F(x)^{l-1}]
\end{aligned} \tag{6}$$

Let $C(s)$ define as the coefficient on $F(x)^s$ in the right-hand side of Equation (6) for $0 \leq s \ \leq l-1$:

$$C(s) = f(x)\sum_{k=0}^{i-1} \left(\begin{array}{c} i-1 \\ k \end{array} \right) (-1)^k \cdot I_{\{i \leq i+s-k \leq l\}} \cdot p_{i+s-k} \cdot \frac{(i+s-k)!}{(s-k)!(i-1)!}. \tag{7}$$

The first high-ranked statistic identifies its probability density $g(x)$ for all x. Therefore $C(s)$'s, coefficients on $F(x)^s$ where $0 \leq s \leq l-1$, in Equation (6) are identified given identification of $F(x)$ and $f(x)$. Below I prove that identification of $\{C(s)\}_{0 \leq s \leq l-i}$ implies identification of $\{p_n\}_{i \leq n \leq l}$ by using mathematical induction:

(1) For $s = 0$

$C(0) = f(x) i p_i$; therefore identification of $C(0)$ implies identification of p_i given identification of $f(x)$.

(2) For $1 \leq s \leq l-i$

Dividing the summation in Equation (7) into two parts, $k = 0$ and $1 \leq k \leq i-1$, yields:

$$\begin{aligned} C(s) \quad = \quad & \underbrace{f(x) p_{i+s} \cdot \frac{(i+s)!}{s!(i-1)!}}_{(a)} \\ & + \underbrace{f(x)\sum_{k=1}^{i-1} \left(\begin{array}{c} i-1 \\ k \end{array} \right) (-1)^k \cdot I_{\{i \leq i+s-k \leq l\}} \cdot p_{i+s-k} \cdot \frac{(i+s-k)!}{(s-k)!(i-1)!}}_{(b)}. \end{aligned} \tag{8}$$

First consider identification of (b) in Equation (8). Given identification of $f(x)$, we only need to check identification of $I_{\{i \leq i+s-k \leq l\}} \cdot p_{i+s-k}$. If $I_{\{i \leq i+s-k \leq l\}} \cdot p_{i+s-k} \neq 0$, then $s-k \geq 0$. In addition, $s-k \leq s-1$ because k varies from 1 to $i-1$. Hence $1 \leq k \leq s$ whenever $I_{\{i \leq i+s-k \leq l\}} \cdot p_{i+s-k} \neq 0$. Accordingly, if $\{p_{i+s-k}\}_{1 \leq k \leq s}$ are identified, (b) is identified. If (b) is identified, identification of $C(s)$ implies identification of (a) in Equation (8) and thus identification of p_{i+s}, given identification of $f(x)$. In summary, if $C(s)$ and $\{p_{i+s-k}\}_{1 \leq k \leq s} = \{p_i, ..., p_{i+s-1}\}$ are identified, p_{i+s} is identified. Accordingly if $C(s)$ is identified, $\{p_{i+1}, ..., p_l\}$ can be identified sequentially by substituting $s = 1$ to $l-i$ into Equation (8).

Combining the cases that $s = 0$ and that $1 \leq s \leq l - i$, identification of $\{C(s)\}_{0 \leq s \leq l-i}$ implies identification of $\{p_n\}_{i \leq n \leq l}$ for $i \geq 2$. In conclusion, the ith and the jth high-ranked statistics identify $\{p_n\}_{i \leq n \leq l}$ for $i \geq 2$. ■

Remark 2 *Besides a pair of ranked statistics, if we observe the number of samples of which the size is less than i, we can identify* $\{\Pr(N = n)\}_{n \geq i}$ *and* $\sum_{n=0}^{i-1} \Pr(N = n)$.

To illustrate the result of Theorem 2 I consider an example where the second and the third high-ranked statistic, (Y, X) are observed. Plugging $i = 2$ into Equation (6) yields:

$$\begin{aligned} g(y) &= f(y)[(2p_2) \cdot F(y)^0 + \ (2 \cdot 3p_3 - 2p_2) \cdot F(y) + \cdots \\ &\quad + \{(l-1)lp_l - (l-2)(l-1)p_{l-1}\} \cdot F(y)^{l-2} - \{(l-1)lp_l\} \cdot F(y)^{l-1}]. \end{aligned}$$

Since $g(y)$, $f(y)$, and $F(y)$ are identified for all y, the coefficients on $F(y)^s$ where $0 \leq s \leq l - 1$ are identified; this implies identification of $[\ 2p_2,\ (2 \cdot 3p_3 - 2p_2), ..., \{l(l-1)p_l - (l-1)(l-2)p_{l-1}\},\ -\{l(l-1)p_l]$. Then $\{p_2,\ p_3, ..., p_l\}$ are identified sequentially.

Equation (6) suggests an estimation method for $\{p_n\}_{i \leq n \leq l}$ as well. Let Z_t and Y_t denote the second and the third high-ranked statistics at samples $t = 1, 2, ..., T$. Consistent estimators of $F(\cdot)$ and $f(\cdot)$ are proposed in the previous subsection. In addition, the kernel density estimator gives a consistent estimator of $g(z)$. We can then construct a pseudo data set:

$$\{\widehat{g}(z_t),\ \widehat{f}(z_t),\ \widehat{F}(z_t)^0,\ \widehat{F}(z_t)^1, \cdots, \widehat{F}(z_t)^{l'}\}_{t=1}^{T}$$

where l' is large enough such that $p_l = 0$ for $l \geq l'$. Then we will be able to estimate $\{p_n\}_{i \leq n \leq l}$ consistently by using Equation (6) as a moment condition.

4 Identification of Auction Models

In the previous section I defined ranked statistic and studied its properties. In this section I derive new identification results of auction models by using properties of ranked statistics. I first start by stating Corollary 1 & 2 and apply them to establish identification of various auction models. By applying Theorem 1, Corollary 1 is immediate as follows.

Corollary 1 *Assume the symmetric IPV model. Observation of a pair of high-ranked or low-ranked statistics of potential bidders' valuations nonparametrically identifies the potential bidders' value distribution, even if the number of potential bidders is unknown and follows a different distribution across auctions.*

The IPV assumption in empirical work practically means that bidders' valuations are independent and identically distributed conditional on auction characteristics observed by researchers. Regarding the number of potential bidders, note not only that we do not have to know the number of potential bidders, but also that its distribution does not have to be the same across auctions; this is very convenient in practice because we can pool bidding data from auctions with different number of potential bidders when we estimate potential bidders' value distribution. Even if there are unobservable auction characteristics to affect the number of potential bidders, it won't be an obstacle to estimation of potential bidders' value distribution. It has also a computational advantage when observable auction characteristics affect both the number of potential bidders and their valuations. In order to recover potential bidders' value distribution from observed bids, we typically specify an empirical model distinguishing between the effects of observable auction characteristics on the number of potential bidders and on their valuations; because both the number of potential bidders and their valuations affect the distribution of observed bids directly to make an effect on the mapping from the distribution of observed bids into the potential bidders' value distribution. However we do not have to control the effect of observable auction characteristics on the number of potential bidders when using an estimation method presented in Section 3.2.

Next Corollary 2 is straightforward from Theorem 2.

Corollary 2 *Consider the symmetric IPV model. Let N denote a random variable representing the number of potential bidders. Then $\{\Pr(N = n|N \geq i)\}_{n \geq i}$ are identified if either:*

(a) the ith and the jth high-ranked statistics of potential bidders' valuations; or

(b) the ith and the jth low-ranked statistics of potential bidders' valuations

are observed where $1 \le i < j$.

Remark 3 *Besides a pair of ranked statistics of potential bidders' valuations, if we observe the number of auctions with potential bidders less than* i, *we can identify* $\{\Pr(N = n)\}_{n \ge i}$ *and* $\sum_{n=0}^{i-1} \Pr(N = n)$.

Corollary 1 & 2 say that a pair of ranked-statistics of potential bidders' valuations nonparametrically identify most model primitives of interest. Now the question is whether we can obtain a pair of ranked-statistics of potential bidders' valuations or their estimates, which depends on observables and their relationship with model primitives. The rest of this section studies identification of various models of four auction formats - second-price sealed-bid, first-price sealed-bid, ascending, and eBay auctions. Among four standard auction formats, the descending auction is excluded from analysis because a pair of valuations won't be available in an descending auction where only the winning bid is made. I make a separate analysis of eBay auctions in spite of their ascending-bid format because eBay auctions have distinct features, especially in entry process. EBay bidders never come together and become aware of the existence of the auction to participate at different times in the course of the auction; therefore eBay bidders' entry decisions are made sequentially unlike the traditional ascending auction. I first start by analysis of models where every potential bidder takes part in an auction in Section 4.1 and move on models with entry in Section 4.2.

4.1 Auction Models without Entry

I consider the basic model presented in Section 2. In this subsection every potential bidder is assumed to take part in an auction. Accordingly all potential bidders make a bid in sealed-bid auctions, but which is not necessary in ascending auctions. I analyze eBay auctions in the next subsection where the entry process is considered. As mentioned in Section 1, not all potential bidders participate in eBay auctions by design and thus we need to take potential bidders' entry decision into account to analyze eBay auctions.

4.1.1 Second-price sealed-bid auctions

In the standard model of the second-price sealed-bid auction, it is the unique symmetric equilibrium for each bidder to bid his own valuation. Hence if a pair of bids per auction are available with their rankings, we can identify the standard model of second-price sealed-bid auctions by applying Corollary 1 and 2. The seller has more incentive to record top two bids - the winner's bid and the transaction price - than the other bids. If more than two bids with their rankings or the number of potential bidders is available, the model can be tested.

4.1.2 First-price sealed-bid auctions

In the first-price sealed-bid auction, each bidder's equilibrium behavior depends not only on his own valuation but also on the distribution of his competitors' bids. Hence we may need all potential bidders' bids even if we try to recover only a pair of valuations per auction. For example, Guerre, Perrigne, and Vuong (2000) drove an equation which expresses each bidder's valuation v_i as a function of his own equilibrium bid b_i, its distribution $G(\cdot)$, its density $g(\cdot)$, and the number of bidders, I as follows (Equation (3) on p. 529):

$$v_i = b_i + \frac{1}{I-1}\frac{G(b_i)}{g(b_i)}.$$

Song(2015b) considered a first-price sealed-bid auction model with an uncertain number of bidders[4] to show that each bidder's valuation can be consistently estimated from his own bid and top two bids per auction by developing the idea of Guerre, Perrigne, and Vuong (2000). Then by applying Corollary 1 and 2, we can identify the first-price sealed-bid auction model with an uncertain number of bidders from top two bids per auction. If the number of potential bidders is available, the model can be tested.

[4] As in the model presented in Section 2, each bidder knows the distribution of the number of his competitors but does not know the actual number of his competitors when he makes a bid. (See Krishna(2002), p.34) This model is different from the model analyzed in An, Hu, and Shum (2010). In their model, the researcher does not know the number of potential bidders, but a bidder knows the number of his competitors when he makes a bid.

4.1.3 Ascending auctions

The ascending auction has many variants. In some ascending auctions, the bids are called by the bidders themselves and the auction ends when no one is willing to raise the standing bid. In other ascending auctions, the auctioneer calls the bids, and a willing bidder indicates his assent by some slight gesture. (Milgrom and Weber, 1982) Hence it depends on the bidding model how to infer bidders' valuations from their bids. The dominant model of the ascending auction is the button auction model presented in Milgrom and Weber (1982). In their model, the auction ends as soon as the second-last bidder drops out from the auction and it is an equilibrium for each bidder to drop out when the auction price reaches his valuation. Hence every bidder's valuation is observed except for the highest bidder's; as a result, the button auction model is identified if a pair of dropout-prices are available per auction with their rankings. As in the second-price sealed-bid auction if more than two dropout-prices or the number of potential bidders is available the model can be tested. Kim and Lee (2014) proposes a nonparametric test of symmetric IPV framework by using the method proposed in Section 3.2 and applies to wholesale used-car auctions which use an ascending-bid format.

While the button auction model serves well in illustrating the complex strategic issues, as Haile and Tamer (2003) pointed out, the button auction model is often too abstract to be employed as an empirical model which should yield an exact mapping between observed bids and potential bidders' valuations. First of all, the unobserved dropout-price is not only highest bidder's unless a bidder is required to indicate in or out at every auction price. In most real-world auctions, many low-valued bidders who drop out early from the auction are not recognized. In those auctions, even the number of participating bidders is unobserved unless there are separate records of participating bidders. Second, bidders can re-enter to revise his previous bids in most auctions where bidders call out their bids. Hence even if a potential bidder's bid is observed, it does not have to be his valuation as envisioned in the button auction model especially when the bid is made early in the auction. In those auctions, the bids made toward the end of auction provide closer approximation of bidders' valuations than the early bids. Actually using only the winning bids for estimation has been a resolution to avoid misinterpretations of losing bids. (See, for example, Haile (2001)). Hence even if more than two dropout-prices are available, it is worth while to estimate the button auction model by using only top two dropout-prices for comparison.

4.2 Auction Models with Entry

A potential bidder may not take part in an auction with various reasons such as a binding reserve price, bid preparation costs, etc. If not all potential bidders take part in an auction, model primitives we can identify from bid data will be changed. In this subsection I discover what model primitives we can identify with incorporating entry processes without knowledge nor an assumption regarding the number of potential bidders. But I do not research on identification of an entry model, which cannot be done without further information beyond bid data.

There has been much research on estimation of entry models. For example, Athey, Levin, and Seira (2011) and Krasnokutskaya and Seim (2011) estimate a model of potential bidders' entry decisions. Gentry and Li (2014) give identification results on auctions with entry. They all need a measure of the number of potential bidders for their results. A measure of the number of potential bidders in auctions with entry is often subject to a measurement error and hence the methods developed in this paper compliment existing methods. In 4.2.1 and 4.2.2, I study identification of models with entry for three auction formats analyzed in the previous subsection and in 4.2.3, I analyze eBay auctions.

4.2.1 Auctions with a binding reserve price

I assume the basic model presented in Section 2 including an assumption that a potential bidder does not know the number of his competitors when he makes a bid. If a seller sets a binding reserve price, $r > \underline{v}$, no potential bidders with valuations below r would take part in the auction. Hence there will be no bid data from which we can identify either $F(v)$ for $v < r$ nor presence of potential bidders with valuations below r. However we can still apply Corollary 1 and 2 to identify essential part of the auction model. Consider following two auction environments Γ_r and Γ where the auction formats are the same except the reserve price:

■ Γ_r: Auction environments with a reserve price, r

• A random variable, N represents the number of potential bidders with $p_n = \Pr(N = n)$.

• A potential bidder's valuation is drawn randomly from $F(v)$.

■ Γ: Auction environments with no reserve price

• A random variable, N' represents the number of potential bidders in Γ_r with valuations no less than r. Then

$$\Pr(N' = n) = \sum_{i=n}^{\infty} p_i \left(\begin{array}{c} i \\ n \end{array} \right) (1 - F(r))^n F(r)^{i-n}.$$

• A potential bidder' valuation is drawn randomly from $F(v|V \geq r)$,

$$\text{where } F(v|V \geq r) = \begin{cases} \frac{F(v)-F(r)}{1-F(r)}, & v \geq r \\ 0, & v < r \end{cases}.$$

In Γ_r, a potential bidder with his valuation below r would not take part in an auction because of the reserve price. As a result, the sets of participating bidders are the same in Γ_r and Γ. In addition, the distributions of participating bidders' valuations are the same ex ante and so be the distributions of bids. Accordingly we can analyze bid data from Γ_r as if they were from Γ to identify $F(v|V \geq r)$ and the distribution of N' where N' represents the number of potential bidders with valuations no less than r. Hence Corollary 1 & 2 and identification results developed for each auction format in Section 4.1 are all applicable to auctions with a binding reserve price with modifications of model primitives of interest. While it is critical that a potential bidder knows the reserve price when he makes a bid for identification results, it does not matter whether or not the reserve price is known to the researcher because the reserve price, lower bound of bids, is identified from bid data.

This simple extension is owing to the fact that the identification results in Section 3 are established for the stochastic number of potential bidders. The number of participating bidders becomes stochastic with a binding reserve price even if the number of potential bidders is constant. Hence if the identification results were established for the constant number of potential bidders, their extensions to auctions with a binding reserve price could raise a more complicated issue of extension of the constant to the stochastic number of bidders.

One strength of the estimation method proposed in Section 3.2 is that it is easy to apply to auction data with varied reserve prices once a pair of valuations with their rankings are obtained. If identification of potential bidders' value distribution is established in an auction model with a binding reserve price, its extension to the auction model with varied reserve prices is straightforward. On the other hand, it can be involved to estimate auction models nonparametrically with varied reserve prices. If the reserve price is fixed, we do not have to know the level of $F(r)$ to estimate $F(v|V \geq r)$. However when we combine data from auctions with different reserve prices, the levels of $F(r)$'s come to matter if they affect the distribution of observables used for estimation, which is the case in all existing empirical auction literature. In that case we need a parametric assumption of $F(\cdot)$ or a model for $F(r)$, for instance a function of observables as in Guerre, Perrigne, and Vuong (2000). In contrast, the estimation method proposed in Section 3.2 uses a conditional likelihood whose value has no relationship with the levels of $F(r)$'s. Consequently the estimation method is easily applicable to bid data from eBay, ascending, and second-price sealed-bid auctions even with varied reserve prices. However, in first-price sealed-bid auctions where bidders' valuations are not directly observed from their bids, it may cause a difficulty with varied reserve prices in the process of nonparametric estimation of top two valuations from top two bids.

4.2.2 Auctions with entry costs

I introduce an entry process to the basic model presented in Section 2 to accommodate entry costs. Thus now the model has two stages. In the first stage, a potential bidder decides whether or not to incur entry costs to participate in the auction to be held in the second stage. The entry costs are the same for all potential bidders. In the second stage, the auction is conducted to those who has decided to participate in the auction.

I study identification of $F(v)$ and the distribution of N for given equilibrium behaviors generated by two representative entry models: (i) Samuelson(1985): a potential bidder's entry decision is made after learning his valuations; (ii) Levin and Smith (1994): a potential bidder's entry decision is made before learning his valuation. In both models the entry costs are moderate so that not all potential bidders participate in the auction and more than one potential bidders participate in the auction with a positive probability.

The entry model proposed in Samuelson(1985) generates an equilibrium behavior of selective entry: there is a break-even valuation v^* such that a potential bidder decides to participate in the auction if his valuation is no less than v^*. Samuelson(1985) analyzed only the first-price sealed-bid auction but the result in Gentry and Li (2014) (Proposition 1 on p.322) implies that the same kind of selective entry equilibrium exists in the second-price sealed-bid and the ascending auction as well.[5] Hence identification with Samuelson's entry model is similar to identification with a binding reserve price. Only difference is that v^* is unknown to the researcher, but v^*, the lower bound of bids, is identified. Accordingly, by modifying model primitives to $(F(v) - F(v^*))/(1 - F(v^*))$ and the distribution of N', where N' represents the number of potential bidders with valuations no less than v^*, Corollary 1 & 2 and identification results developed for each auction format in Section 4.1 are all applicable to auctions with Samuelson(1985)'s entry model.

The entry model proposed in Levin and Smith (1994) generates an equilibrium behavior of random entry: each potential bidder participates in an auction with the same probability q which is determined endogenously by the auction format and other market

[5] Gentry and Li (2014) study a more general model having Samuelson(1985)'s entry model as a special case.

factors to make a potential bidder indifferent between participating or not. As a result, potential bidders' entry decisions are not correlated with their valuations while the distribution of the number of participating bidders are determined endogenously. Introducing the entry process of Levin and Smith (1994) makes few differences in discussion of identification except that identifiable model primitives are not of potential bidders, but of participating bidders. Corollary 1 & 2 and identification results developed for each auction format in Section 4.1 are all applicable to identification of the distribution of the number of participating bidders and their value distribution. Since the entry is random, the participating bidders' value distribution is the same as the potential bidders', $F(v)$. Although the distribution of the number of participating bidders is different from that of potential bidders, its identification is not of statistical question. With Levin and Smith's entry model, it is impossible to distinguish between changes in q and changes in the distribution of the number of potential bidders without further information beyond bid data such as entry costs, a measure of potential competition and its instruments, etc.

4.2.3 eBay auctions

A big success of eBay has generated widespread interest in interpreting eBay auction data. While there exists a huge amount of empirical literature on eBay, structural analysis of eBay auction data has been limited. Several works have taken structural approaches, for instance, Bajari and Hortaçsu (2002a), Ackerberg, Hirano, and Shahriar (2006), Adams (2007), etc. However their estimation methods all depend on parametric distributional assumptions. In fact, to my best knowledge, there is no existing work to estimate an eBay auction model nonparametrically. The main obstacles to nonparametric approaches to eBay auctions are the fact that the number of potential bidders willing to pay the reserve price is not available and the reserve prices are varied across auctions. This paper complements the existing literature by proposing a nonparametric method to estimate an eBay auction model.

I consider only auctions in which a single item is sold. An eBay auction[6] starts as

[6] Here I briefly explain the eBay auction mechanism. For more detail of eBay auctions, see Lucking-

soon as a seller registers it to be held for as many days as the seller chooses. There are no pre-announcements of the auctions to be held. The auction price starts at the reserve price set by the seller and only a bid over the standing auction price is accepted. As the auction proceeds, the auction price is raised to the second-highest existing bid plus the minimum increments whenever a new bid is placed. All bids but the highest one are disclosed during the auction. The auction ends at the fixed closing time, which was known as the auction started. Once the auction has concluded, the winner pays the auction price posted at the closing time. Thus the winning price is the second-highest bid plus the minimum increments.

Song (2015a) proposes an eBay auction model within the symmetric IPV framework. In her model potential bidders are those who become aware of the existence of the auction among potential buyers. She showed that on the equilibrium every potential bidder places a bid of his valuation before the auction ends as long as the auction price has not raised over his valuation. As a result two highest-valued potential bidders, whose valuations cannot be lower than the auction price at any course of the auction, always place a bid of their valuations by the end of the auction. Unlike traditional ascending auctions, the highest-valued potential bidder places a bid. On the other hand, a lower-than-second-highest-valued potential bidder cannot place a bid of his valuation if the auction price has been already raised over his valuation before he does. Therefore not all potential bidders willing to pay a reserve price, place a bid. In addition, it is difficult to find information of virtually anonymous online bidders from secondary sources other than their bid data. As a result, the number of potential bidders willing to pay the reserve price is not available in eBay auctions. But by applying the Corollary 1 & 2, Song(2015a)'s eBay model is identified from top two bids. Although the highest bid is not revealed in public even after the auction ends, eBay has records of all bids. Hence if eBay would share, top two bids are available. For example, Adams and Hosken (2011) analyzed eBay auction data having all bids.

Previous research has proposed a different model of eBay auctions. Bajari and Hor-

Reiley(2000), Bajari and Hortaçsu (2002a,b), Lucking-Reiley, Bryan, Prasad and Daniel Reeve (2007).

tasçu (2002a) study eBay auctions within the common value paradigm. Ely and Hossain (2009) propose a model of multiple comcurrent auctions to explain the results from their field experiments. The estimation method should be adapted according to the choice of the model. However, whatever model is considered, it is still the case that observed bidders are not all potential bidders willing to pay the reserve price. Even if the bidding activities are concentrated on the very short time period toward the end of the auction so that every potential bidder places a bid on the webpage showing that the auction price is lower than his bid, there still exists the order of bidding times; therefore any bid lower than top two bids won't be accepted by the eBay system if the bid is transmitted later than top two bids. Hence this paper resolves one of fundamental obstacles to estimation of an eBay auction model.

5 Monte Carlo Experiments

To illustrate the performance of my estimation method, I conduct Monte Carlo experiments. The experiments are ascending auctions with varied reserve prices. I assume the button auction model where the second and the third high-ranked statistics of dropout-prices are available. For each experiment, artificial data of 600 auctions are generated.[7] The number of potential bidders, N_t $(t = 1, ..., 600)$, was first drawn from a Binomial distribution with trial number 50 and success probability 0.1. Potential bidders' valuations as many as N_t were then generated according to the equation:

$$\ln V_t^i = \alpha_1 X_{1t} + \alpha_2 X_{2t} + \nu_t^i \tag{9}$$

where $\alpha_1 = 1$, $\alpha_2 = -1$, $X_{1t} \backsim N(0, 1)$, $X_{2t} \backsim Exp(1)$, and $\nu_t^i \backsim f(\cdot) = Gamma(9, 3)$. For the sake of reference, $E(\nu_t^i) = 3$, and $Var(\nu_t^i) = 1$. The variables X_{1t} and X_{2t} represent observable auction characteristics. The random variable ν_t^i is bidder i's private information, the distribution of which is to be estimated. To reflect the fact that not all potential

[7] Six hundred may seem like a large number. I do not use data from all 600 auctions, as will become clear later on. Moreover, benefit of using auction data is that a huge number of observations are available in many real-world auctions such as eBay auctions.

bidders participate in the auction, reserve prices R_t, were set as follows:

$$\ln R_t = \alpha_1 X_{1t} + \alpha_2 X_{2t} + \omega_t$$

where $\omega_t \ = Y_t - 2,\ Y_t \backsim Gamma(9, 3)$. No realized value of ω_t is observed; Y_t and ν_t^i are independent. Both V_t^i and R_t depend on X_{1t} and X_{2t}, and thus V_t^i and R_t are positively correlated. The participating bidders are the potential bidders with valuations no less than R_t and the participating bidders drop out from the auction at their valuations.

In each experiment, a dataset consists of X_{1t}, X_{2t}, and the second and the third-highest among participating bidders' dropout-prices. I do not use the highest bidder's dropout-price which is not observed in ascending auctions. The auctions having fewer than three participating bidders are dropped. Hence, the number of the auctions used for estimation is less than 600, actually 412, on average, across 100 repetitions. I estimate α_1, α_2, and $f(\cdot)$ by varying the series lengths of the SNP estimator, k from 0 to 8. A researcher does not make a parametric distributional assumption on ν_t^i, though the specification in (9) is assumed to be known. In principle, the functional form of the effects of auction characteristics is also identified up to location (Athey and Haile, 2002), but its estimation requires a huge dataset.

<Table 1> documents the averages of various statistics for 100 repetitions. The Best SNP means the SNP estimate obtained by using the series length that minimizes a true ISE between 0 and 8 in each experiment. It would be ideal to obtain the performance of the SNP estimates by using the value of k^* chosen through the method described in the estimation section. But this would take a great deal of time, and furthermore, as <Table 1> shows, the SNP estimates are robust in terms of the choice of k in this simulation. The results in <Table 1> show that the estimators perform very well. I calculate the estimates of $E(\nu_t^i)$ and $STD(\nu_t^i)$ to examine the performance of the estimates of $f(\cdot)$. <Figure 1-3> in Appendix B show the graphs of estimates of $f(\cdot)$ and the density estimates of V_t^i evaluated at the median of $(X_{1t},\ X_{2t})$, along with the corresponding true densities. The

graphs also illustrate that the estimators perform very well.

< Table 1> Monte Carlo Experiment Results

	$\widehat{E(\nu_t^i)}$	$\widehat{STD(\nu_t^i)}$	$\widehat{\alpha_1}$	$\widehat{\alpha_2}$	$\widehat{SE}(\widehat{\alpha_1})$	$\widehat{SE}(\widehat{\alpha_2})$
True	3	1	1	-1		
Best SNP	3.057	1.021	1.015	-1.003	.091	.047
$k = 0$	2.950	1.058	1.015	-1.002	.101	.052
$k = 1$	2.989	1.040	1.014	-1.002	.100	.051
$k = 2$	3.081	1.013	1.015	-1.004	.091	.047
$k = 3$	3.115	1.007	1.016	-1.010	.090	.047
$k = 4$	3.027	1.014	1.012	-1.014	.088	.046
$k = 5$	3.099	1.000	1.018	-1.012	.081	.042
$k = 6$	3.057	1.009	1.010	-1.007	.075	.038
$k = 7$	3.078	1.001	1.067	-1.005	.068	.034
$k = 8$	3.054	1.003	1.009	-0.996	.061	.031

•$SE(\widehat{\alpha_1}), SE(\widehat{\alpha_2})$: BHHH estimators

6 Concluding Remarks

In this paper, I develop new econometric methods applicable to identify and estimate various auction models when the number of potential bidders is unknown. I focus on identification and estimation of model primitives of general interest - the distribution of the number of potential bidders and their value distribution - by using part of bid data which are most commonly available from auctions. Hence the proposed methods can be generally used and have potential to be extended if more observables are available in applications. If we have more observables, we may identify a new model primitive of interest, enhance efficiency of estimators, or develop a method to test a model.

The identification results depend critically on independency among bidders' valuations conditional on observable auction characteristics. Unlike most empirical auction methods, it won't cause a problem in estimation of potential bidders' value distribution if an unobservable auction characteristic affects the distribution of the number of potential bidders without generating correlations among bidders' valuations. However the methods proposed in this paper are not applicable if unobservable auction characteristics make bidders' valuations correlated conditional on observable auction characteristics whether or not unobservable auction characteristics affect the distribution of the number of potential bidders.

Appendix A: Proof of the Case Where a Pair of Low-ranked Statistics Are Available

1. Identification of the parent distribution (Proof of Theorem 1)

Proof. The upper limit of support of $F(\cdot)$ is denoted by $\overline{v}$, and $f(\cdot)$ is an associated density of $F(\cdot)$. It is allowed for $\overline{v}$ to take ∞. Let X and Y denote the k_1th and k_2th low-ranked statistics respectively where $1 \leq k_1 < k_2 \leq N$. For the sake of notational convenience, let $F(x|y) = \frac{F(x)}{F(y)}$ and $f(x|y) = \frac{f(x)}{F(y)}$. Note that we are considering a right-truncated distribution unlike the case of high-ranked statistics. The density of X conditional on Y, $p_{(k_1|k_2)}(x|y)$, for $x \leq y$, is computed by applying Equation (1) and (2):

$$\begin{aligned}
p_{(k_1|k_2)}(x|y) &= \frac{(k_2-1)!}{(k_1-1)!(k_2-k_1-1)!} \times \\
&\quad \frac{[F(y)F(x|y)]^{k_1-1}[F(y)(1-F(y|r))]^{k_2-k_1-1}F(y)f(x|y)}{F(y)^{k_2-1}} \cdot I_{\{x\leq y\}} \\
&= \frac{(k_2-1)!}{(k_1-1)!(k_2-k_1-1)!} F(x|y)^{k_1-1}[1-F(x|y)]^{k_2-k_1-1} f(x|y) \cdot I_{\{x\leq y\}} \\
&= \frac{(k_2-1)!}{(k_1-1)!(k_2-1-k_1)!} F(x|y)^{k_1-1}[1-F(x|y)]^{k_2-1-k_1} f(x|y) \cdot I_{\{x\leq y\}} \\
&= f^{(k_1:k_2-1)}(x|y) \cdot I_{\{x\leq y\}}.
\end{aligned}$$

Now note that $\lim_{y\to\overline{v}} p_{(k_1|k_2)}(x|y) = \lim_{y\to\overline{v}} f^{(k_1:k_2-1)}(x|y) \cdot I_{\{x\leq y\}} = f^{(k_1:k_2-1)}(x)$. This implies that the density of k_1th low-ranked statistic conditional on the k_2th low-ranked statistic identifies the density of the (k_2-k_1)th order statistic of which the parent distribution is

$F(\cdot)$ and the sample size is $(k_2 - 1)$. Since the distribution of any order statistic with a known sample size identifies its parent distribution, the result follows. ■

2. Identification of the distribution of the sample size (Proof of Theorem 2)

Proof. Throughout the proof, note that $F(x)$ and its associated density, $f(x)$ are identified for all x by applying Theorem 1. Let $p_n = \Pr(N = n | N \geq i)$ where $n = i, ..., l$. Let X denote the ith low-ranked statistic and $G(x)$ denote the cumulative distribution function of X. Then

$$G(x) = \sum_{n=i}^{l} p_n F^{(i:n)}(x). \tag{10}$$

A. If $i = 1$, identification is easily established. Plugging $i = 1$ into Equation (10) yields:

$$\begin{aligned} G(x) &= \sum_{n=1}^{l} p_n [1 - \{1 - F(x)\}^i] \\ &= \sum_{n=1}^{l} p_n - \sum_{n=1}^{l} p_n \{1 - F(x)\}^i \\ &= 1 - [p_1\{1 - F(x)\} + p_2\{1 - F(x)\}^2 + \cdots + p_l\{1 - F(x)\}^l\ ]\ . \end{aligned}$$

The first low-ranked statistic identifies its cumulative distribution function, $G(x)$ for all x. Given identification of $F(x)$, the p_n, the coefficient on $\{1 - F(x)\}^n$ is identified for $n = 1, ..., l$.

B. For a proof for the case where $i \geq 2$, I consider $g(x)$, an associated density of $G(x)$.

By applying Equation (2) & (10), and rearranging terms lead to:

$$
\begin{aligned}
g(x) &= \sum_{n=i}^{l} p_n f^{(i:n)}(x) \qquad (11)\\
&= \sum_{n=i}^{l} p_n \cdot \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}\{1-F(x)\}^{n-i} f(x)\\
&= f(x)F(x)^{i-1}\sum_{n=i}^{l} p_n \cdot \frac{n!}{(n-i)!(i-1)!}\{1-F(x)\}^{n-i}\\
&= f(x)\left[\sum_{k=0}^{i-1}\begin{pmatrix} i-1 \\ k \end{pmatrix}(-1)^k\{1-F(x)\}^k\right]\left[\sum_{k=0}^{l-i} p_{i+k}\cdot \frac{(i+k)!}{k!(i-1)!}\{1-F(x)\}^k\right]\\
&= f(x)\left[\sum_{s=0}^{l-1}\left\{\sum_{k=0}^{i-1}\begin{pmatrix} i-1 \\ k \end{pmatrix}(-1)^k \cdot I_{\{i\le i+(s-k)\le l\}}\cdot p_{i+(s-k)}\right.\right.\\
&\qquad\qquad\left.\left. \times\ \frac{\{i+(s-k)\}!}{(s-k)!(i-1)!}\right\}\cdot\{1-F(x)\}^s\right]\\
&= f(x)[(ip_i)\cdot\{1-F(x)\}^0 + \ \{i(i+1)p_{i+1}-(i-1)ip_i\}\cdot\{1-F(x)\}+\cdots\\
&\quad +\{(i-1)(-1)^{i-2}\frac{l!}{(l-i)!(i-1)!}p_l + (-1)^{i-1}\frac{(l-1)!}{(l-i-1)!(i-1)!}p_{l-1}\}\cdot\{1-F(x)\}^{l-2}\\
&\quad +\{(-1)^{i-1}\frac{l!}{(l-i)!(i-1)!}p_l\}\cdot\{1-F(x)\}^{l-1}]
\end{aligned}
$$

Equation (11) is the same as Equation (6) if the terms of $\{1-F(x)\}^s$ are all replaced by $F(x)^s$ where $0 \le s \le l-1$. Therefore by applying the same arguments as in the proof for the case where high-ranked statistics are available, the result follows. ■

Appendix B: Graphs of Monte Carlo Experiments

The 100 Best SNP are obtained through application of the series length that minimizes a true ISE in each experiment. <Figure 1> presents the graph of the density estimate of which true ISE is the smallest among 100 Best SNP estimates. <Figure 2> graphs the density estimate of which the true ISE is the 50th. Finally, <Figure 3> graphs the density estimate of which the true ISE is the biggest. Throughout these three figures, the left graph presents a density estimate of $f(\cdot)$, and the right graph presents a density estimate of V_t^i conditional on the median of $(X_{1t},\, X_{2t})$. It is important to remember that potential bidders' valuations were generated according to the equation:

$$\ln V_t^i = \alpha_1 X_{1t} + \alpha_2 X_{2t} + \nu_t^i \tag{7}$$

where $\alpha_1 = 1$, $\alpha_2 = -1$, $X_{1t} \backsim N(0,1)$, $X_{2t} \backsim Exp(1)$, and $\nu_t^i \backsim f(\cdot) = Gamma(9,3)$. The solid line represents the true density, and the dotted line represents an estimate.

Figure 1: The Best Performance

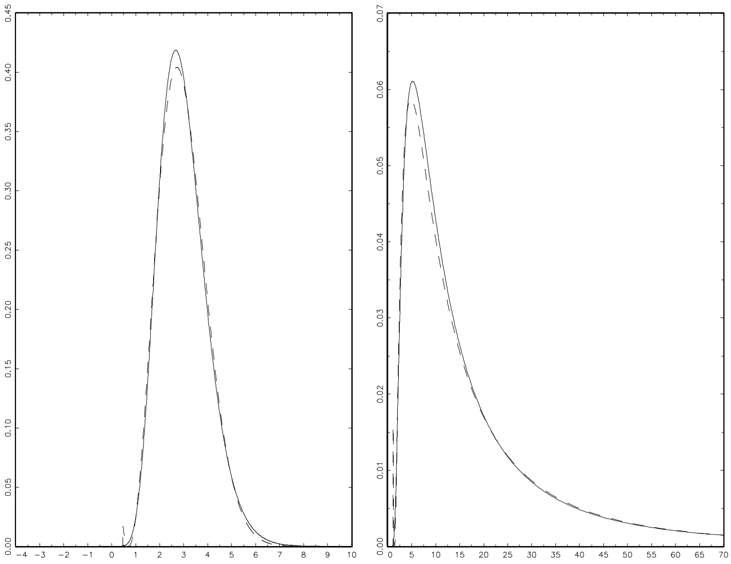


Figure 2: The Median Performance

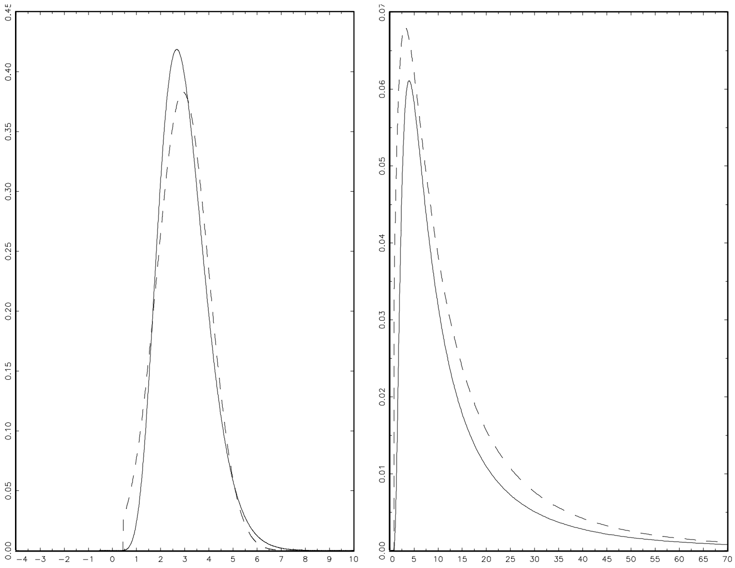
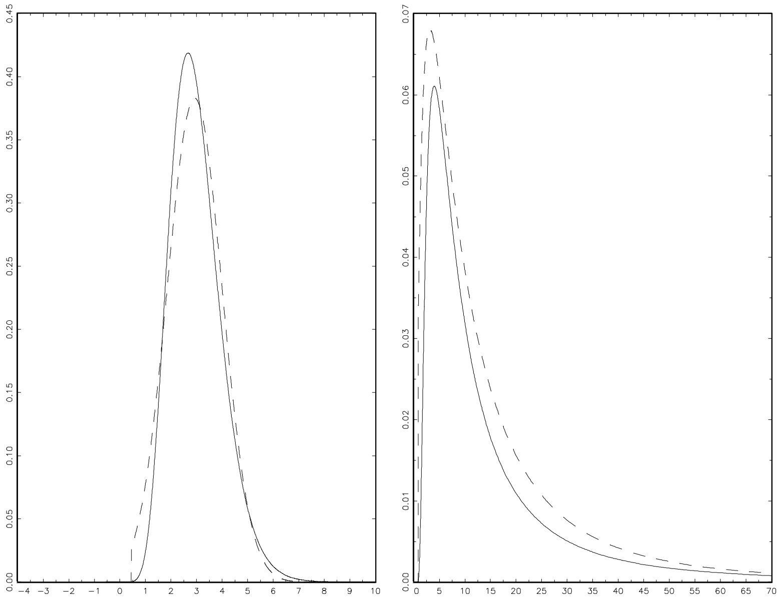


Figure 3: The Worst Performance



References

Ackerberg, Daniel, Keisuke Hirano, and Quazi Shahriar (2006), "The Buy-it-now Option, Risk Aversion, and Impatience in an Empirical Model of eBay Bidding," Working paper, University of Michigan.

Adams, Christopher P. (2007), "Estimating demand from eBay prices," *International Journal of Industrial Organization*, 25, p. 1213-1232.

Adams, Christopher P. and Laura Hosken (2011), "Vettes and lemons on eBay", *Quantitative Marketing & Economics*, 9, p. 109-127.

An, Yonghong, Yingyao Hu, and Matthew Shum (2010), "Estimating First-price Auctions With an Unknown Number of Bidders: A Misclassification Approach," *Journal of Econometrics*, 157, p.328-341.

Arnold, Barry C., N. Balakrishnan, and H. N. Nagaraja (1992), *A first course in order statistics*, New York: John Wiley & Sons.

Athey, Susan and Philip A. Haile (2002), “Identification of Standard Auction Models”, *Econometrica*, 70(6), p. 2107-2140.

Athey, Susan, Jonathan Levin, and Enrique Seira (2011): “Comparing Open and Sealed Bid Auctions: Evidence From Timber Auctions,” *Quarterly Journal of Economics*, 126 (1), p. 207-257.

Bajari, Patrick and Ali Hortaçsu (2002a), “Winner’s curse, Reserve Prices and Endogenous Entry: Empirical Insights from eBay Auction, ” *RAND Journal of Economics*, Summer 2003, p. 329-355.

Bajari, Patrick and Ali Hortaçsu (2002b), "Cyberspace Auctions and Pricing Issues: A Revise of Empirical Findings," *New Economic Handbook*, Edited by Derek C. Jones.

David Lucking-Reiley, Doug Bryan, Naghi Prasad, and Daniel Reeve (2007), “Pennies from eBay: the Determinants of Price in Online Auctions”, *Journal of Industrial Economics*, Vol. 55(2), p. 223–233.

Coppejans, Mark and A. Ronald Gallant (2002), "Cross-validated SNP density estimates, " *Journal of Econometrics*, 110, p. 27-65.

Fenton, Victor M. and A. Ronald Gallant (1996a), "Convergence Rates of SNP Density Estimators," *Econometrica*, 64(3), p. 719-27.

Fenton, Victor M. and A. Ronald Gallant (1996b), "Qualitative and asymptotic performance of SNP density estimators," *Journal of Econometrics,* 74, p. 77-118.

Gallant, A. Ronald and Douglas Nychka (1987), "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, p. 363-390.

Gentry, Matthew and Tong Li (2014), "Identification in Auctions with Selective Entry," *Econometrica*, 82(1), p. 315-344.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong (2000), "Optimal Nonparametric Estimation of First-price Auctions," *Econometrica*, 68(3), p. 525-574.

Haile, Philip A. and Elie Tamer (2003), "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy,* 111 (1), p. 1-51.

Jeffrey, C. Ely and Tanjim Hossain (2009), "Sniping and Squatting in Auction Markets," *American Economic Journal: Microeconomics*, 1(2), p.68-94.

Hendricks, Kenneth, Joris Pinkse, and Robert H. Porter (2003), "Empirical Implications of Equilibrium Bidding in First-Price, Symmetric, Common Value Auctions," *Review of Economic Studies*, 70(1), p.115-145.

Kim, Kyoo il and Lee, Joonsuk (2014), "Nonparmetric Estimation and Testing of the Symmetric IPV Framework with Unknown Number of Bidders," Working Paper, Michigan State University.

Krasnokutskaya, Elena and Katja Seim (2011), "Bid Preference Programs and Participation in Highway Procurement Auctions," *American Economic Review*, 101, p.2653-2686.

Krishna, Vijay (2002), *Auction Theory*, San Diego: Academic Press.

Laffont, Jean-Jacques, Herve Ossard, and Quang Vuong (1995), "Econometrics of First price Auctions, " *Econometrica*, 63, p. 953-980.

Levin, Dan and James L. Smith, (1994), "Equilibrium in Auctions with Entry," *American Economic Review*, 84(3), p. 585-599.

Lucking-Reiley, David (2000), "Auctions on the Internet: What's Being Auctioned, and How?", *Journal of Industrial Economics*, 48, p. 227-252.

Paarsch, Harry J. (1997), “Deriving an Estimate of the Optimal Reserve Price: An Application to British Columbian Timber Sales,” *Journal of Econometrics*, 78(1), p. 333-357.

Samuelson, William F. (1985), "Competitive Bidding with Entry Costs," *Economics Letters*, 17, p. 53-57.

Song(2015a), "Estimation of an eBay Auction Model with an Unknown Number of Bidders," Working Paper, Seoul National University.

Song(2015b), "Estimation of a First-price Auction Model with an Uncertain Number of Bidders," Working Paper, Seoul National University.