More is Better, or Not? An Empirical Analysis of Buyer Preferences for Variety on the E-Market*

Senay SOKULLU†

August 6, 2020

Abstract

This paper examines the effect of the number of sellers of a good on buyers' demand using an extensive dataset from an e-commerce platform (priceminister.com) in France. Accounting for seller characteristics constitutes variety amongst the same product. Although a wide variety might be preferred by the buyers, it can also introduce a search cost. Using a flexible semiparametric specification, I find that the demand of buyers is not monotonically increasing in the number of sellers (variety), contrary to what has been assumed in the literature. I illustrate the consequences of misspecification of these network effects by a counterfactual simulation.

Keywords: Nonparametric IV Regression, Two-Sided Markets, E-Commerce, Online Platforms

JEL Classification: C14, C30, L14

*I would like to thank Ozlem Bedre-Defolie, Gregory Jolivet, Bruno Jullien, Andrew Rhodes, Marc Rysman, Helene Turon, Frank Windmeijer as well as the participants at seminars (University of Southampton and University of Groningen) and conferences (Bristol IO Workshop 2017, 2nd Economics of Platforms Workshop, EARIE 2018, CRESS 2019) for useful discussions and suggestions. All errors are mine. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

†University of Bristol, School of Economics, *Priory Road Complex, Priory Road, Bristol BS8 1TU, UK*; E-mail: *senay.sokullu@bristol.ac.uk*

1 Introduction

E-commerce platforms are now as popular as traditional stores. One may buy almost anything from her seat just with a click or a touch on the screen. There are many other advantages accompanying the minimization of the physical effort. One of these advantages is that the variety of the products is much larger than one can possibly find in a traditional store. On the one hand this gives the buyers the opportunity to buy their most preferred good, on the other hand there is a cost associated with the search amongst these variety of goods. In this paper, using a dataset from one of the largest e-commerce website from France and assuming that different seller characteristics constitutes variety amongst the same product, I examine the effect of number of sellers (variety) on consumer preferences. I find that consumers' demand is not monotonically increasing in the number of sellers and consumers' benefit from variety starts to decrease after reaching a threshold.

E-commerce websites are clearly two-sided platforms with buyers and sellers on each side, see Rochet and Tirole (2003); Jullien (2010); Hagiu and Jullien (2007). Hence there are two-sided network externalities between the buyers and the sellers. In other words, the buyers are more likely to join the platform if there are more sellers and vice versa, as the benefit gained by each side depends on how well the platform is doing on the other side. In most of the two-sided platforms literature, two-sided network externalities are assumed to be linear. More precisely, the utility of the agent on one side is assumed to be a linear function of the number of the agents on the other side as in Armstrong (2006). This assumption has been retained in most of the empirical two-sided platforms literature (see Rysman, 2004; Argentesi and Filistrucchi, 2007) until recently, Sokullu (2016) showed that the network effects are nonlinear and nonmonotone in the local daily newspaper industry in the US. In the e-commerce platforms literature although linearity has not been assumed, it is assumed

that the utility of buyers is monotonically increasing in the number of sellers (Hagiu, 2009). Although this assumption sounds plausible, the increased number of sellers for a given good might also result in increased search costs which would decrease the utility if it is too high or increased number of sellers may also decrease the utility of buyers further as a result of limited attention. Indeed, in consumer search literature it is shown that search costs still exist in e-commerce platforms, see Jolivet and Turon (2018); Dinerstein, Einav, Levin, and Sundaresan (2018). Hence the network externality exerted on buyers by sellers may be nonlinear and nonmonotone as was shown in Sokullu (2016) in the case of readers vs. advertisers in the newspaper industry. E-commerce platforms differ from newspapers in the sense that one can actually observe the transaction between the two sides. In this paper, I build a model which allows for flexible effects of product variety on buyers' demand when the variety is partly introduced by the existence of different sellers on an e-commerce website.[1] Using the advances in the literature of semi/non-parametric instrumental variables models, I propose to estimate a semi-parametric logit model where the effect of variety on the indirect utility of the buyers is specified nonparametrically.

Optimal platform pricing and optimal provision of variety have already been examined in the theoretical e-commerce platforms literature. As expected, results depend highly on the assumptions on the shape of two-sided network externalities. For example, Hagiu (2009) states that the ability of the platform to price the sellers higher and let the buyers to enter the platform for free stems form the fact that the buyers prefer product variety. On another note Anderson and Bedre-Defolie (2017) examine the optimal provision of variety and assume that the indirect utility of buyers increases in the number of variants offered by the platform. They then show that the optimal pricing problem of the platform is equivalent to the problem of a multi-product monopolist who maximises his profits by selecting variety as well as prices. This equivalence result cannot be obtained without the assumption of monotonically increasing utility in variety. Thus in this paper, I check empirically the assumptions used in

[1]Throughout the paper, I will use number of sellers and variety interchangeably. In online platforms even if the goods are exactly the same, seller characteristics might introduce variety.

the theoretical literature. I believe that the results I obtain will be useful for the analysis related to optimal platform decisions in e-commerce platforms literature. With the increased use of e-commerce platforms, the need for regulation is likely to appear. The results of this paper will also be relevant to any analysis which will be done for regulatory/policy purposes.

The empirical analysis in this paper uses a unique dataset from one of the largest e-commerce platforms from France, Priceminister.com. It is an administrative dataset recording all transactions and all adverts posted on the website. I observe all of the adverts for a particular product at a particular date. In other words I observe variants of goods for a particular product. The variation may come from price, the condition of the good, the reputation of the seller, etc. I adopt a semiparametric logit model to exploit the rich information given in the data set with flexible functional forms. The challenge in the estimation comes from the fact that I have a partially linear model with endogenous variables entering the equation both linearly and nonlinearly. Nonparametric estimation of these models have already been studied by Florens, Johannes, and Bellegem (2012) and Blundell, Chen, and Kristensen (2007) among others. I follow Florens et al. (2012) to estimate the effect of number of sellers on demand. My main result shows that the buyers' demand does not monotonically increase in the number of sellers, it indeed starts decreasing once a threshold is reached. Moreover, the consumers prefer smaller size professional sellers and they prefer recent items either used or new.

This paper is related to three strands of literature. First of all, e-commerce websites being online platforms links this paper to the two-sided markets literature. Starting with seminal papers by Rochet and Tirole (2003) and Armstrong (2006) two-sided markets have been studied extensively. One common feature of the previous theoretical and empirical literature on two-sided markets is that the network externalities have been assumed to be either linear or monotonic, see Rochet and Tirole (2003); Armstrong (2006); Rysman (2004) and Argentesi and Filistrucchi (2007). Anderson and De Palma (2009) moves away from the monotonic network effects in their article where they examine the information congestion in

the advertisement market. They state that the consumer benefit may not be monotonically increasing in the number of advertisements as a result of limited attention. Hence, when the number of senders (advertisements) increases, information congestion might happen as the receivers (consumers) are not able to look at all the messages as a result of limited attention. The first best outcome, in other words the no-congestion case, can be obtained via a subsidy or tax depending on the level of transmission costs. Indeed one can view the e-platforms same as in the advertisement market, where the senders are the sellers and the receivers are the buyers and the transmission cost is the price the platform is charging to the sellers. With this analogy between two markets, it is easy to see that limited attention may lead buyers' utilities not to be monotonic in the number of sellers. Though in the theoretical literature on online platforms, the assumption of monotonic network effects has been crucial in obtaining results on optimality, such as in the debate of quantity versus quality (Hagiu, 2009) or in the derivation of optimal pricing strategy of the platform (Anderson and Bedre-Defolie, 2017). To the best of my knowledge, this is the first empirical paper in the e-commerce literature which confronts monotonic network effects assumption.

This paper can also be related to consumer search on online platforms although it does not fully explain the mechanism behind the non-monotonic buyer utility in the number of sellers. Using data from e-Bay, Dinerstein et al. (2018) show that search costs exist in online marketplaces. As a result, an increased number of sellers increases price variation. Using revealed preference analysis, Jolivet and Turon (2018) show that some transactions on e-commerce platforms can only be explained in the presence of search costs. Kim, Albuquerque, and Bronnenberg (2010) consider a sequential consumer search model and apply it to a dataset on camcorders from Amazon.com. First, their model shows that the cross-price elasticities between the viewed and non-viewed products are almost zero. Second, they find that the consumers search on average 14 products. A more related paper to my work in the search literature is Chen and Zhang (2013). They have a theoretical result similar to mine which concludes that consumer welfare is not monotonic in entry and has an inverse

U-relationship with the entry cost. In the low levels of entry cost, there are too many sellers entering the market hence although it is easier to get a better match, the low expected quality offsets this positive effect coming from the match and decreases consumer welfare. When the entry cost is high, only high quality sellers can enter the market but the search options are lower and this effect offsets the positive high quality benefit and decreases the consumer welfare. One can think that the entry cost in an online platform is small and fixed, hence as the number of sellers increases the consumers' benefit may decrease.

Finally, as I use nonparametric instrumental variable techniques to recover the effect of network effects on the buyers' utility, this paper can also be related to the nonparametric econometrics literature. Nonparametric IV models have been studied extesnively since the 2000s. Darolles, Fan, Florens, and Renault (2011) considers a nonparametric model where the right hand side of the equation is a nonparametric function of an endogenous variable. Later Horowitz (2011) and Fève and Florens (2010) examine these techniques from a more applied perspective, such as obtaining confidence bands, selection of smoothing parameters etc. In my model, the decisions to join the platform on both sides as well as the prices are taken simultaneously hence I need to control for endogeneity of prices and the number of sellers in the estimation of buyers' demand equation. I use excluded variables, which affect sellers' decisions, as instruments and follow closely Florens et al. (2012) to estimate the semiparametric logit model I introduce. Florens et al. (2012) examines the nonparametric estimation of a partially linear endogenous model and show that the $\sqrt{N}$-asymptotic normality can be achieved for the parametric part under some regularity conditions.

The rest of the paper is structured as follows. In Section 2, I introduce the model. Some descriptive statistics about the data is presented in Section 3. Then in Section 4, I define the estimation strategy while the empirical analysis is presented in Section 5. Section 6 concludes.

2 The Model

In this section I introduce the model which I use for the empirical analysis. In Section 5, I only estimate the demand of buyers and use the other structural equations pertaining to sellers and platform behaviour to motivate my instruments.

2.1 Buyers

The main aim of this paper is to recover the effect of number of sellers, hence variety, on buyers' preferences. I adopt a logit model with a nonparametric component.[2,3] A buyer i, $i = 1, ..., I$ who would like to buy the product j has two options; she can buy it from a seller on the online platform or she can choose the outside option which includes buying it from a traditional store.[4] If she decides on buying it from the online platform, she searches for the product on the platform and sees a list of available items. She then chooses the seller s, $s = 1, ..., S_j$ from whom she would like to buy the product. The list of products appears as a result of search contains the same good sold by different sellers which might have different attributes such as price, if the good is new or used, if it is used then its condition, the reputation of the seller, etc. Hence each good sold by a different seller can be counted as a variant of the product. One can then write the indirect utility of buyer i who decides to buy the good j from seller s on the online platform as:

$$U^i_{js} = \bar{U}_{js} + \epsilon^i_{js}$$

[2]In this model the consumers are choosing a book amongst the variants (in terms of seller and book characteristics) of the same book. Hence, in this case, the restrictions of the logit model on substitution across goods is less concerning.

[3]A nested logit model would have a first nest of buying the book from the platform or not and then selection of the seller conditional on the decision of buying the book from the platform. Hence, the parameter on the group share would measure the substitutability of the books sold on the platform, i.e., if the books sold on the platform are more/less substitutable compared to the books sold in the traditional stores. Given that I am comparing the same book sold by different sellers and stores, one would not expect the coefficient on the group share to be statistically significant. I have estimated a nested logit model and found this parameter insignificant.

[4]Outside option also includes other e-commerce websites such as Amazon, eBay and Fnac. None of these websites are "pure" platforms (Hagiu, 2007), neither they have the particular fraud-safe system Priceminister.com has, see Section 3. Hence I included them in the outside option.

where $\bar{U}_{js}$ is the mean utility of buying good j from seller s on the platform and ϵ^i_{js} is the consumer specific unobservable effect which is assumed to follow an Extreme Value Type I distribution.[5]

The mean utility level of buying good j from seller s, $\bar{U}_{js}$ is specified as:

$$\bar{U}_{js} = X'_{js}\beta - \alpha p_{js} + \phi(v_j) + \xi_{js}$$

where X_{js} is a vector of observable seller and good attributes, p_{js} is the price of good j sold by seller s, v_j number of the sellers of product j on the platform and ξ_{js} is the error term capturing seller and good characteristics which are unobservable to the econometrician such as the quality of the picture of the good on the website, the reviews about the seller, etc. As already explained, each good with a different combination of attributes is counted as a different one, hence v_j is the number of the elements of the set which is composed of product j with different attributes. In other words, there are v_j different seller-book pair with different features for book j.

ϕ function in the mean utility may capture the search cost due to variety and can be motivated by structural models such as Draganska and Jain (2005) and Richards and Hamilton (2015). Although one might think that the buyers on the platform have a preference for variety, it has also been shown that there are search costs related to this variety as well as the problem of limited attention, see Dinerstein et al. (2018); Jolivet and Turon (2018); Anderson and De Palma (2009). A consumer who wants to buy the product that she prefers most, has to go over all of the adverts on the platform to find it. The contribution of variety to consumer utility can be formalised following Kim, Allenby, and Rossi (2002), Draganska and Jain (2005) and Richards and Hamilton (2015). Conditional on the desired book, each buyer maximises her utility by choosing to buy it on the platform as the platform is likely to offer her best match. So, presence of book j sold by seller s on the platform provides

[5]Note that I assume that the consumer already knows which product she would like to buy, so the choice of product is not modeled.

an additional utility $\hat{u}_{js}$ to the buyer. Then the contribution of the variety to the indirect utility can be written as:

$$\phi(v_j) = \hat{u}_{js} Pr(js \in Platform) - c(v_j) \tag{1}$$

where $c(v_j)$ is the cost of going through different options. If one assumes that the probability of finding book j sold by seller s on the platform is equal to $\frac{1}{v_j}$ and the utility of finding the book does not depend on any seller attributes, hence same for each seller given j, i.e. $\hat{u}_{js} = \hat{u}_j$, then $\phi(v_j)$ is given by:

$$\phi(v_j) = \hat{u}_j \frac{1}{v_j} - c(v_j) \tag{2}$$

$\phi(v_j)$ function captures the utility from finding the best match as well as the cost of the search. Draganska and Jain (2005) and Richards and Hamilton (2015) assume specific functional forms for cost function $c(v_j)$ and hence for $\phi(v_j)$. By specifying $\phi(v_j)$ nonparametrically, I allow the effect of number of sellers to be nonlinear and non-monotonic in a flexible way. In other words, my specification does not assume that the indirect utility of the consumers is linear or nonlinear in a specific way in the number of sellers, the benefit from variety may decrease due to increased search costs, decreased expected quality or limited attention.

Another interpretation of ϕ can be obtained following Ackerberg and Rysman (2005) where they study the identification issues appearing in discrete choice models due to the restriction on the effect of number of parameters on unobservable characteristics space. The authors state that the logit errors imply that all products are equidistant from each other in the unobserved characteristics space and that there is no crowding out effect, i.e., the distance stays the same when new products/firms added to the market. This property leads to unintuitive identification results. Ackerberg and Rysman (2005) propose to solve this issue by a simple adjustment where they add a function of the number of products to the random

utility and they show that this adjustment can actually be motivated by a structural model. The ϕ function in my model can also be seen as the adjustment term given in Ackerberg and Rysman (2005).

I assume that the function ϕ is unknown so it is the object of interest along with the parameters α and β.

Given the model above, following Berry (1994), one can obtain the estimation equation:

$$ln(s_{js}) - ln(s_0) = X'_{js}\beta - \alpha p_{js} + \phi(v_j) + \xi_{js} \tag{3}$$

where s_{js} is the market share of the seller s and s_0 is the market share of the outside option which is buying the good from a traditional store. The mean utility of the outside option is normalised to zero, $U_0 = 0$. The market shares are measured as:

$$s_{js} = \frac{q_{js}}{M}$$

where q_{js} is the number of buyers who buy the product j from seller s and M is the total market size.

2.2 Sellers

The sellers are charged by the platform only if they sell the good they listed. There are no listing fees. Given the per transaction fee t_{js} the profit function of seller s for good j can be written as:

$$\pi_s = (p_{js} - t_{js} - c_{js})D(p_{js}, p_{-js}) - K_s$$

where c_{js} is the marginal cost of seller s, $D(p_{js}, p_{-js}) = Ms_{js}$ is the demand for book j sold by seller s and K_s is the fixed cost. Given this profit function, the seller s enters the platform if:

$$\pi_s \geq 0 \tag{4}$$

Hence the seller's problem can be written as

$$\max_{p_{js}} \pi_s \quad subject \quad to \quad \pi_s \geq 0$$

His optimal price is given by:

$$p_{js} = -\frac{D(p_{js}, p_{-js})}{\partial D(p_{js}, p_{-js})/\partial p_{js}} + t_{js} + c_{js} \tag{5}$$

Given the demand function specification (Equation 3), Equation 5 can be rewritten as:[6]

$$p_{js} = t_{js} + c_{js} + \frac{1}{\alpha(1 - s_{js})} \tag{6}$$

Note that anything that shifts the marginal cost of the seller or per transaction fee, t_{js}, affects his pricing decision as well as his entry decision. Hence, in Section 5, where I estimate the model, the instruments for price and variety are constructed based on these cost shifters.

2.3 The Platform

As already mentioned, the platform charges only a per transaction fee t_{js} to sellers. It also gives them a subsidy for delivery cost, d_{js}, as a tool to attract sellers. This delivery subsidy is determined exogenously by using the postal service tariff in France.[7] Moreover, I assume that the platform maximises its profits per product. Then the profit of the platform for product j can be written as:

$$\pi_{pl} = \sum_{s=1}^{n} (t_{js} - d_{js} - c_{pl}) D(p_{js}, p_{-js}) - K_{pl} \tag{7}$$

[6]Under the assumption of symmetric sellers, the equilibrium price of each seller will be a function of t_{js}, d_{js} and the number of sellers on the platform v_j. In this paper, I am interested in recovering the effect of v_j on buyers' demand using flexible forms and I do not study the equilibrium properties of the theoretical model. As already mentioned, sellers' and platform's behaviour are only introduced to motivate the instruments.

[7]It is written on the website:"*correspond as possible as the tariff applied by La Poste for the articles with the same form, size and weight.*"

where K_{pl} is the fixed cost of the platform and c_{pl} is its marginal cost. Given Equation 7, the profit maximising transaction fee is given by:

$$t_{js} = d_{js} + c_{pl} + \frac{1}{\alpha(1 - s_{js})} - \frac{\sum_{k=1,k\neq s}^{n}(t_{jk} - d_{jk} - c_{pl})s_{jk}(1 - s_{jk})}{(1 - s_{js})} \tag{8}$$

Three things are worth mentioning in this equation. First of all, the platform internalises the effect of a change in transaction fee for seller s on other sellers. Secondly, here I assume that the platform charges a different transaction fee to each seller. In reality, transaction fees are given by a menu. However, the pricing scheme of the platform I am studying is highly nonlinear, it almost amounts to charging each seller a different price. Hence in the model I assume that this is the case. Third, the exogenously given delivery subsidy d_{js} affects the per transaction fee (t_{js}) and hence the profits of sellers thus affecting their decision to join the platform or not.

3 Data

The data I am using come from one of the largest e-commerce websites in France, www.priceminister.com, which had 11 million registered users in 2010 with over 120 millions of products for sale, see Jolivet and Turon (2018).[8] Priceminister.com was acquired by Rakuten in 2010 and by August 2018 the brand Priceminister disappeared and was replaced by Rakuten France.

The website is a perfect example of the platform model, which allows the transaction between sellers and buyers by giving them affiliations. It does not sell any of its own products as Amazon does, nor it allows for auctions like eBay. Moreover, the sellers can be professional or non-professional and the goods sold can be brand new or used. One can find many different

[8] As stated in Jolivet, Jullien, and Postel-Vinay (2016), the website was rated the first among e-commerce websites in a survey in France in 2010. The other main e-commerce websites in France were Amazon, Ebay and Fnac. However, it is only Priceminister.com which acts like a pure platform, hence in the structural model I include the other e-commerce websites in the outside option.

goods on the platform from books, CDs, games and DVDs to shoes and TVs. In this paper I focus on books.[9]

Priceminister.com did not charge a listing fee to the sellers but only a variable transaction fee which was calculated according to a menu published in the website.[10] Moreover, as already mentioned in Jolivet et al. (2016), buyers are free to choose the delivery method for each transaction where priceminister.com imposes a fixed shipping cost scale. Given the choice of the buyers, this delivery cost is then transferred to the seller. The seller then can minimise its delivery cost as long as he uses the method complying with the buyer's choice. Hence, I represent this scheme in Section (2.3) as the delivery subsidy.

Once a buyer goes on the website and searches for the book she would like to buy, the website returns a page of available items which have different prices and product features such as its condition ('as new', 'very good', 'good'), seller's score, seller's status (professional or not), etc. Hence, the more alternatives the website returns as a result of search, the more information the consumer needs to go through to be able to find her favourite item. The items are sorted by price by default however the buyer has the option to sort them differently.

One feature that has made Priceminister.com quite particular is the way it has fought fraud. As explained in Jolivet et al. (2016), once a buyer purchases a good from a seller, the payment first goes to Priceminister.com and it is held there until the buyer confirms the receipt of the item and gives feedback. Only after this Priceminister.com closes the transaction and transfers the payment to the seller. This feature also ensures that all buyers rate sellers.

The data are composed of two administrative datasets from the website. The first one has information on all transactions that took place between 2001 and 2008 while the second one has information on all adverts posted on the website between 2001 and 2008. So, any advert posted on the website, even if it hasn't sold appears in the second dataset while the

[9]To check the robustness of my results, I also estimate the model using samples of CDs and DVDs. The results are presented in Appendix D.

[10]See Appendix A.

first dataset has the records of sold items only. For each transaction I can observe the price, product ID, seller ID, advert ID, seller and product characteristics as well as the commission the seller pays to the website and taxes on different items related to transaction such as the commission tax, shipping cost, etc.

To be able to estimate the demand function given by equation 3, I need to have data over seller-book combination. For this purpose, I take a snapshot of the data from September 2007 and use transactions realised during this period for BOOKs. Taking the snapshot also allows me to control for the number of adverts at a given time, more easily.[11] The number of sellers, in other words the number of active adverts is created by counting the active adverts during September 2007 using the advert dataset.

The estimation of the demand equation needs some data selection. First of all, I drop the observations where the number of alternatives is equal to zero, i.e. when the only active advert is the advert of the transaction as then there is only one product on the screen and no search cost or limited attention problem exists in this situation. Moreover, Jolivet and Turon (2018) show that when there are only two adverts for a product, 91% of the transactions can be explained by a full information model, i.e. can be rationalised without a search cost. Hence, I also drop all observations with two active adverts only. Moreover I only included the observations where a seller-book pair appears at least twice.[12] After the selection process, I am left with 1806 observations, consisting of 1503 different books sold by 1070 different sellers. 83.44 % of the books are used while the 16.56 % are new and 74% of the sellers in my sample are not professional sellers, see Table 1.

Table 2 shows the summary statistics of the continuous variables I am using. The mean price of a book that is sold on the platform is 8.61 Euros with 8.86 standard error. One issue

[11] I aggregate the data over seller-book in September 2007. Aggregation over longer time period would result in having adverts which are irrelevant.

[12] Although nonparametric methods require large sample sizes to perform better, the method I use in this paper necessitates inversion of a square matrix whose size is equal to the sample size, so when the sample is too large, there is a computation problem. This last selection gives me a smaller sample where I do not run into any computation problems. I run parametric regressions with the larger dataset where this last restriction is not imposed and the results on the ϕ function are robust. These robustness results are available upon request.

with using books is that the prices of new books in France are regulated, hence I may not get price variation. This is not the case in my data as most of the sold books are used ones. I also checked the price dispersion per book. To do it, I computed the standard deviation of the price for each book and then look at the distribution of this variable. The minimum value it takes is larger than zero, meaning there is always a price variation between the sellers for each book. Figure 5 in Appendix A shows the distribution of the standard deviation of the prices per book. The mean number of active adverts (sellers) during a transaction is around 14, and this number goes up to 285. However, as Figure 1 shows, this number is less 100 for most of the observations. Table 3 shows roughly the distribution of number of sellers. As can be seen from the table for most of the observations I have in my data (90.5%), the number of sellers is less than 30. Moreover, when the distribution of number of sellers for used and new books are considered, there are no significant differences. In other words, there is no pattern such that there are more sellers for new books than for used ones, see Table 4. The average seller has a reputation score of 4.5 over 5 and she/he has sold around 3250 items. Also, most of the reputation scores in the sample are between 4 and 5. For example, in 94% of the observations the reputation is 4 and above and in 87% of the observations, it is 4.5 and above, see Figure 6 in Appendix A.[13]

4 Estimation Method

Equation 3 is a semi-parametric demand function with a linear parametric part ($X'_{js}\beta - \alpha p_{js}$) and a nonparametric function ($\phi(v_j)$). As a result of simultaneity the price and the number sellers in this demand equation are endogenous. Hence, I propose to estimate the model given by Equation 3 by nonparametric instrumental variables (NPIV) regression.

Nonparametric estimation techniques have been studied extensively in the last 2 decades and the number of applications that use these techniques has risen, especially with the

[13]The fact that the reputation scores are highly skewed is also reported in Nosko and Tadelis (2015) where the data is from eBay.

availability of software packages for practitioners. One of the reasons of the popularity of nonparametric methods is that they give flexibility in the modelling. Like in the model given in Equation 3, a nonparametric specification of the number of sellers allows us not to put any ad hoc shape restrictions on the effect of variety on demand.

The model I propose to estimate is a partially linear model where the parametric part include both exogenous and endogenous regressors and the nonparametric function is a function of an endogenous variable. The case where all the regressors that enter the model (both parametrically and nonparametricaly) are endogenous has already been studied by Florens et al. (2012). It has been shown that $\sqrt{N}$-convergence for the parametric part can be obtained with NPIV estimation under some regularity conditions. In this paper I include both endogenous and exogenous variables in the parametric part of the model while the nonparametric function is a function of an endogenous variable and hence the estimation still follows from Florens et al. (2012) and it is straightforward. For the sake of exposition, denote $Y_{js} = ln(s_{js}) - ln(s_0)$. Then one can rewrite Equation 3 as:

$$Y_{js} = X'_{js}\beta - \alpha p_{js} + \phi(v_j) + \xi_{js} \tag{9}$$

Assume that there is a valid vector of instruments, Z_{js}, for the price and the number of sellers, such that:[14,15]

$$E[\xi_{js}|Z_{js}] = 0 \quad and \quad Cov(p_{js}, Z_{js}) \neq 0 \quad and \quad Cov(v_j, Z_{js}) \neq 0$$

Moreover, let us denote $W_{js} = [X'_{js} \quad p_{js}]'$. Then using the set of instruments one can write:

$$E[Y_{js}|X_{js}, Z_{js}] = E[W'_{js}\gamma|X_{js}, Z_{js}] + E[\phi(v_j)|X_{js}, Z_{js}] \tag{10}$$

[14]The instruments used in the estimation are introduced in Section 5.

[15]Note that the conditions given for the valid set of instruments are relevant only if parametric IV methods are used. Indeed for NPIV a different assumption, i.e., *completeness*, is needed for the instrument relevance: $E[m(p, v)|X, Z] = 0 \rightarrow m(p, v) = 0 \quad a.s.$

where $\gamma = (\beta, \alpha)$. Equation 10 is the main identifying equation and the estimation follows from this relation. More precisely, estimation is done by replacing conditional expectations by their empirical counterparts and solving for objects of interest, $\phi(.)$ and γ. I describe the estimation process in detail in Appendix B.

5 Empirical Analysis

In this section, I estimate buyers' demand function given in Equation 3 by nonparametric and parametric methods. I then illustrate the effect of misspecification with a simple counterfactual simulation.

5.1 Identification

I estimate the model given in Equation 3 using both the nonparametric and parametric methods, the NPIV estimation approach described in Section 4 and Generalised Method of Moments (GMM), respectively. Equation 3 is a partially linear logit model and there are two endogenous variables on the right hand side. One of them is the price of the book j sold by seller s and the other is the number of sellers, i.e. variants, v_j of book j during the time period I am considering. As a result, I need at least one instrument for each of these variables.

The first instrument I use is the tax rate on the platform's variable commission. In the data, I observe variable commission and the tax paid on variable commission. I then derive the tax rate by dividing the latter by the former. Tax rate is a cost shifter for the platform which affects t_{js} via Equation 8 which in turn affects pricing and entry decisions of the sellers via Equations 4 and 6. Hence one expect this variable to be correlated with p_{js} and v_j. Although, the true tax rate is fixed (5%), the variable I calculate is a noisy observation of the true tax rate and it varies between 0 and 0.08. The noise stems from the

rounding which is not correlated with the unobserved characteristics of the product.[16] First, we know that the tariff for variable fee is nonlinear. Assume that this fee is given by some function of the price, i.e. $f_{js} = g(p_{js})$, where f_{js} is the variable fee. But because of rounding, the fee I observe in the data is given by $f_{js} = g(p_{js}) + \omega_{js}$. Second, the true tax rate, τ is equal to 5%, but again because of rounding, the variable tax paid on variable fee is a noisy observation, such that tax paid$= \tau(g(p_{js}) + \omega_{js}) + \eta_{js}$. Finally, the variable I construct is given by $\hat{\tau}_{js} = \frac{\tau(g(p_{js})+\omega_{js})+\eta_{js}}{g(p_{js})+\omega_{js}} = \tau + \delta_{js}$ where $\delta_{js} = \frac{\eta_{js}}{g(p_{js})+\omega_{js}}$. Hence, it varies across observations. At the same time, as p_{js} enters δ_{js} nonlinearly and there are two noises, ω_{js} and η_{js}, one would expect that the validity condition, $Corr(\delta_{js}, \xi_{js}) = 0$, is satisfied, i.e., δ_{js} is expected to vary independently of unobserved product quality.[17,18] To sum up, the first instrument varies across observations because of the noise and is expected to be correlated with p_{js} and v_j while it is expected to be independent of the unobservables in the demand equation.[19]

The second instrument I use is the delivery subsidy paid by the platform to sellers. As shown in Equation 8, delivery subsidy affects the transaction fee the platform charges which then affect the entry and pricing decisions of the sellers via Equations 4 and 6. This subsidy is determined exogenously to match the rates charged by French postal service, La Poste, as much as possible. It varies between the products in different choice sets and within the same choice set as a result of size of the item and the delivery method. As regards to validity, rates do not vary between locations in France, Andorra and Monaco. Moreover, all delivery

[16] In Appendix A, I present the histogram of the tax rate, it shows clear variation in this variable. Moreover, I regress the tax rate on observed book and seller characteristics to see if there is any correlation. The tax rate seems to be independent of the exogenous book and seller characteristics. Regression results are available upon request.

[17] I show that validity condition could hold using simulated ξ_{js} and the data. First, I know g function and the true tax rate, so I calculate ω_{js}, η_{js} hence δ_{js} from the data. I then generated ξ_{js} correlated with p_{js} using $\xi_{js} = 0.015 * p_{js} + \mathcal{N}(0, 2.25)$, which resulted in $Corr(\xi_{js}, p_{js}) = 0.11$ and $Corr(\xi_{js}, \delta_{js}) = 0.02$. Moreover, the coefficient of ξ in the regression of δ on ξ is highly insignificant (p-value=0.85), whereas it is highly significant in the regression of p on ξ.

[18] To give a simple example in support of this argument -i.e. if (x, y) are correlated, $(f(x), y)$ may not be - let x be a random variable taking values $(-1, 0, 1)$ with probability $(1/3, 1/3, 1/3)$. One can show that $Cov(x, x) = Var(x) \neq 0$ whereas $Cov(x, x^2) = 0$.

[19] I show that both validity and relevance conditions hold during the parametric estimation, shown in Tables 6 and 7.

methods are available for every item and the prices for different methods are fixed for buyers, hence the delivery method does not affect the decision of picking one seller over the other. Finally, as this subsidy is determined ex-ante to match the rates of La Poste, one does not expect it to be correlated with the unobservables in the demand equation, $Cov(\xi_{js}, d_{js}) = 0$.[20]

I use both nonparametric and parametric methods to estimate the demand equation given in (3). Unfortunately, there are not any statistical tests developed to test the strength and validity of instruments in a NPIV setting. To get a brief idea about the relevance condition, I can simply check and see if the instruments are correlated with the endogenous variables.[21] However, in Section 5.2, where I estimate the model using parametric methods, I test both for the weak instruments and validity. According to results of these tests, my instruments are valid and not weak, see Tables 6 and 7.

5.2 Nonparametric Estimation of the Buyers' Demand

I define the market size to be the population of France who are between 20 and 60.[22] Other seller-book observable characteristics I use are seller's size (the number of transactions she/he has done), seller's reputation, a dummy variable indicating if the seller is a professional or not and a dummy variable indicating if the book is new or used as well as a variable capturing how recent that particular book is. The densities of the continuous exogenous variables are estimated by Gaussian kernels and the densities of the discrete variables are estimated by Aitchison and Aitken kernels.[23] Using bootstrap methods, I also obtained 95% confidence

[20] As delivery cost changes with the size of the good, delivery subsidy could be correlated with ξ_{js}, if size is one of the unobservable characteristics of the books which effects demand of buyers. However, it is unlikely that size of a book effects the decision of buyers especially in my model where the buyers already know which book to buy before going to the platform. Hansen's test results in Tables 6 and 7 support this claim as I fail to reject the exogeneity of the instruments.

[21] For instance, correlation coefficient between price and delivery subsidy is equal to 0.27 and correlation coefficient between the number of variants and tax rate is equal to 0.17. Moreover, Babii and Florens (2017) show that in the case completeness (nonparametric counterpart of relevance condition) fails, one may still obtain an estimate which is consistent for the best approximation of parameter of interest and in many cases, this best approximation coincides with the structural parameter.

[22] I also use number of internet users as the market size and the results are robust and available on request.

[23] Nonparametric IV methods require regularization as it is explained in Appendix B. The optimal regularization parameter in this estimation is selected following Fève and Florens (2010).

intervals for both the parameters and the nonparametric function. The estimation is done using author-written code in Matlab.

The results are presented in Figures 2 to 4 and in Table 5. Figure 2 shows the estimated ϕ function with the full sample whereas the estimated functions with observations where the number of sellers are less than 70 and 30 are presented in Figures 3 and 4, respectively. The first thing to notice is that all figures look qualitatively the same. It can also be seen that Figure 3 is a closer version of Figure 2 and Figure 4 is a closer version of Figure 3. The reason for estimating Equation 3 with three different samples is to check the robustness of the results. It is well known that nonparametric methods cannot reveal the information as well on the support of data where there are fewer observations. As for 91% of the observations the number of sellers are less than 30, I wanted to focus on the area where most of the transactions take place by using the smaller sample.[24] A closer look at the Figures 2 and 3 shows that the estimate of $\phi(v_j)$ is the same at low levels of sellers. For this reason, I focus on the results obtained with observations with less than 30 sellers.

The estimate of the ϕ function implies that number of sellers have almost zero effect until around 12 where the positive effect becomes significant and the utility starts increasing. However, this positive externality from sellers to buyers starts decreasing when the number of sellers reaches around 17. As I already mentioned, the confidence intervals are obtained via bootstrap methods and it can be seen that between 12 to 20 the effect is significantly positive. So, we can conclude that although consumers prefer a higher number of sellers, after some threshold this positive preference starts to weaken. In other words, the number of sellers have little effect for low numbers and then buyers' utility starts increasing up to a threshold level. After that level, increase in the number of sellers does not increase the utility; the utility starts decreasing.

If one would like to be conservative in the interpretation of the results for the fact that

[24]Figure 8 in Appendix A shows the plot of the estimation result with full sample and the distribution of active adverts (number of sellers) on each other. As can be seen from the figure, there are very few observation points after 30 adverts.

the curvature might not be estimated correctly and/or curvature is the oscillation resulting from small smoothing parameters, etc., one can still conclude that buyers' demand is not monotonically increasing in variety as its effect is insignificant after around 20 variants.[25] My results is in line with the findings of Kim et al. (2010). Using data on camcorders from Amazon, Kim et al. (2010) find that consumers on average view 14 products. As they do not view all the existing products on the website, it is not surprising to have insignificant and decreasing network effects after a threshold. The threshold of 17 implied by my results is also not far from 14, average number of products viewed, found in Kim et al. (2010).

Table 5 shows the parameter estimates as well as the 95% confidence intervals. The consumers prefer sellers who are professionals and whose size is smaller. Moreover, more recent books are preferred either they are new or used. One interesting result in Table 5 is the coefficient on the reputation. It is estimated to be negative and significant. Using data from eBay, Nosko and Tadelis (2015) show that the feedback measures are highly skewed and they claim that this fact introduces problems for buyers in interpreting feedback scores. For instance, a buyer might think that a score of 98% is excellent. However, when the distribution of feedback score is considered, a seller with 98% is placed below the 10th percentile. Hence, my result can also be explained by the imperception of buyers due to the skewed distribution of reputation. One other reason for this negative effect might stem from the fact that we cannot observe the reviews the buyers left, hence cannot control for them. For sellers whose reputations are already between 4 and 5, reviews might be determinant in the buyers' decisions.

5.3 Parametric Estimation of the Buyers' Demand

In the previous section I showed that the buyers' demand is not monotonically increasing in the number of sellers. Although nonparametric methods are flexible in terms of functional

[25] As a robustness check, I estimate two more demand functions. In the first one, variety is given by the number of sellers whose reputation is above the median reputation and in the second one it is given by the number of sellers whose size is above the median size. I present the results in Appendix A2; non-monotonicity prevails in both estimations.

form specification, they might be opaque to non-economists when they need to be used to conduct welfare analysis. Hence in this section, using the results obtained in Section 5.2 as a guide to find a parametric specification for the effect of variety, I estimate the buyers' demand equation given in (3) using the Generalised Method of Moments. Using the results of this section and of the nonparametric estimation, I then illustrate the effect of misspecification numerically.

I used two different specifications for the ϕ function. In the first one (NL), following the nonparametric estimation results, ϕ is specified as a second order polynomial whereas in the second specification (L), the ϕ function is specified as a linear function of variety:

$$\begin{aligned} ln(s_{js}) - ln(s_0) &= X'_{js}\beta - \alpha p_{js} + \varphi_1 v_j + \varphi_2 v_j^2 + \xi_{js} \qquad (NL) \\ ln(s_{js}) - ln(s_0) &= X'_{js}\beta - \alpha p_{js} + \varphi_1 v_j + \xi_{js} \qquad (L) \end{aligned}$$

I include the same variables in X_{js} and used the same instruments as in Section 5.1 as well as their nonlinear combinations and some additional excluded variables such as dummies showing the condition of the book. I estimate the equations by Generalised Method of Moments (GMM) in STATA. Moreover, I used Windmeijer (2018)'s robust Cragg-Donald test of underidentification to see if the estimation results suffer from weak IV problem.

The estimation results from the two parametric specifications, NL and L are presented in Tables 6 and 7. First, as can be seen in Table 6, the effect of the number of sellers on buyers' demand is nonlinear and nonmonotonic as the coefficients on both variety and square of variety are found to be significant at 10% level. More specifically, the effect of variety increases the demand of buyers at a decreasing rate. Coefficient on price has the expected sign and it is significant at 5% level. Coefficients on all variables except the new dummy have the same sign as in the non-parametric estimation results. As in the nonparametric estimation, only the coefficients on new and professional dummies are found to be insignificant. Moreover, the magnitudes of the coefficients are quite similar. Hansen's

J-statistic has a p-value of 0.2584 hence I fail to reject the null that the instruments are valid. P-value of robust Cragg-Donald test shows that I can reject the null of weak instruments at 6% level.

As regards to the linear specification, it clearly shows the effect of misspecification as it is found that buyers prefer less variety. Coefficients on price, variety, reputation and age are estimated to be significant whereas others are estimated to be insignificant. Comparing with the nonparametric estimation results, all the coefficients have the same sign, except variety and professional dummy. For this specification too, Hansen's J-test and robust Cragg-Donald test suggest that the instruments are valid and not weak (p-values are 0.2381 and 0.0003, respectively).

The estimation results from linear and nonlinear parametric specifications clearly show that in an anti-trust analysis, one may end-up with erroneous conclusions if the network effect function is misspecified. In the next section, I illustrate this with a counterfactual simulation exercise.

5.4 An Illustration

In this section, I illustrate the effect of misspecification by a counterfactual simulation. Using nonparametric and linear parametric estimation results, I simulate different scenarios and compute the changes in the quantity of books demanded, the prices charged by sellers, the sellers' revenue and the sign of the change in sellers' mark-ups.[26] More precisely, I solve the demand equation (3) and pricing equation (6) simultaneously to obtain new equilibrium values of demand and price under each scenario for each model. To keep things simple, I treat the number of sellers exogenous and did not consider the entry and exit decisions of sellers. I believe that the results are still informative about the effect of misspecification.

In all scenarios, I assume that there is an exogenous shock which shifts number of sellers and/or prices. In the first scenario, I assume that the number sellers increases by 1 for each

[26]The simulation results using nonlinear parametric model are presented in Appendix C.

book in my sample whereas in the second scenario I assume that this number increases by 25 for each book. In the third and fourth scenarios, I assume that on top of the change in the number of sellers, the price charged by the platform to sellers also changes. Hence, in the third scenario I assume that the number of sellers increases by 1 and the transaction fee paid to the platform increases by 50%. Finally in the fourth scenario I assume that the number of sellers increases by 25 and the transaction fees decrease by 20%.

Obtaining the effect of a change in variety in the demand function is quite straightforward under parametric specification. However, it is less so under nonparametric specification. Nonparametric estimation of the ϕ function returns a value $\hat{\phi}$ for each value of v_j in the data. To compute the $\hat{\phi}$ after a shock in v_j, i first calculated v_j^{new} which is given by $v_j + 1$ for scenarios 1 and 3 and $v_j + 25$ for scenarios 2 and 4. I then match these v_j^{new} values to the corresponding $\hat{\phi}$ using the estimation results and also taking into account the fact that the effect of variety is estimated to be insignificant after around $v_j = 20$, in other words, in the simulation the effect of variety is taken to be equal to zero, $\hat{\phi} = 0$, for $v_j > 20$.

The results are presented in Table 8. For the sake of exposition, I report the mean percentage changes in demand (Δq), price (Δp) and revenue (Δr) and sign of the change in mark-ups. First, as linear parametric estimation predicts that consumers' utitily decreases with variety, under all scenarios the predicted change in quantity is negative. When nonparametric estimation results are used, all scenarios lead to an increase in quantity. Moreover, as I do not account for seller entry/exit, both parametric and nonparametric simulation results predict the same change in prices. The structural model predicts a decrease in revenue in scenarios 1, 2 and 4 and an increase in scenario 3 under parametric simulation whereas it predicts an increase in revenue in all scenarios except 4 under nonparametric specification. Finally, parametric model predicts a decrease in seller's mark-up in all scenarios and nonparametric model predicts the opposite.

This simulation exercise shows that one can arrive at different conclusions under different specifications of ϕ function. These different conclusions can have important implications in

policy analysis. To fix ideas, assume that this exercise were done for market power analysis after an exogenous change which increases the number of sellers on the platform. Under linear specification we would conclude that the market power of the sellers decreases, whereas the conclusion would be different under a nonlinear specification.

6 Conclusion

In this paper I examine the effect of number of the sellers of a good on demand of buyers on an e-commerce platform. I model the demand for book j sold by seller s using a multinomial logit model where the number of sellers enters the mean utility in a nonparametric function. Using nonparametric estimation techniques, I find that the demand of buyers does not increase monotonically in the number of sellers/variants. After reaching a threshold, having more sellers for a product on the platform decreases the mean utility of the buyers and hence the demand. I illustrate the importance of misspecification by a counterfactual simulation which shows that linear and nonlinear specification of the effect of number of sellers on demand of buyers would give different conclusions. The correct specification of this effect has important implications in empirical anti-trust analysis of e-commerce platforms. Thus the optimal platform pricing models should take this effect into account.

7 Tables and Figures

Table 1: New vs. Used Books; Pro vs. Non-Pro Sellers

	Used	New	Pro	Non-Pro
Freq.	1,507	299	471	1,335
Percent	83.44	16.56	26.08	73.92

Table 2: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
No. of sellers	13.80	22.53	3	285
Price	8.61	8.86	0.9	135.45
Reputation	4.50	1.19	0	5
Size	3252.42	13975.88	0	94478

Table 3: Distribution of Number of Sellers per Product

	No. of observations	Percentage
< 5	738	41%
< 10	1248	69%
< 20	1536	85%
< 30	1636	90.5%
< 50	1708	94.6%
< 70	1756	97.2%
< 90	1781	98.6%

Table 4: Number of sellers by product's condition

	Mean	Std. Dev.	Min	Max
Used	14.88	22.75	3	277
New	8.38	20.57	3	285

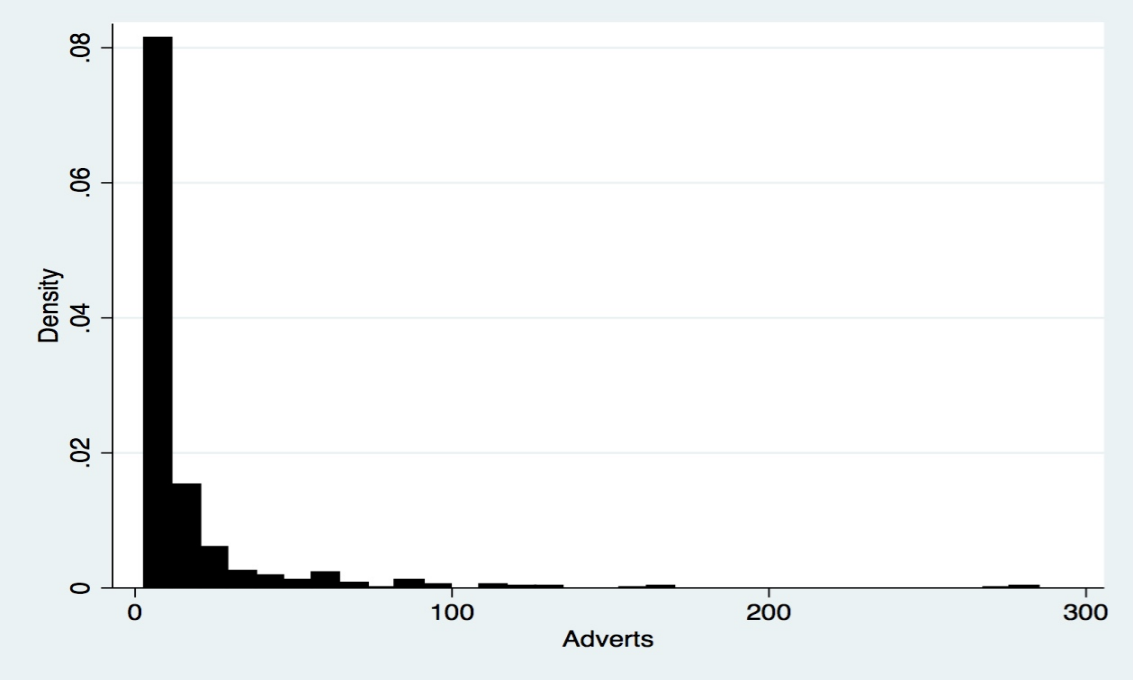

Figure 1: *Histogram of no. of active adverts*

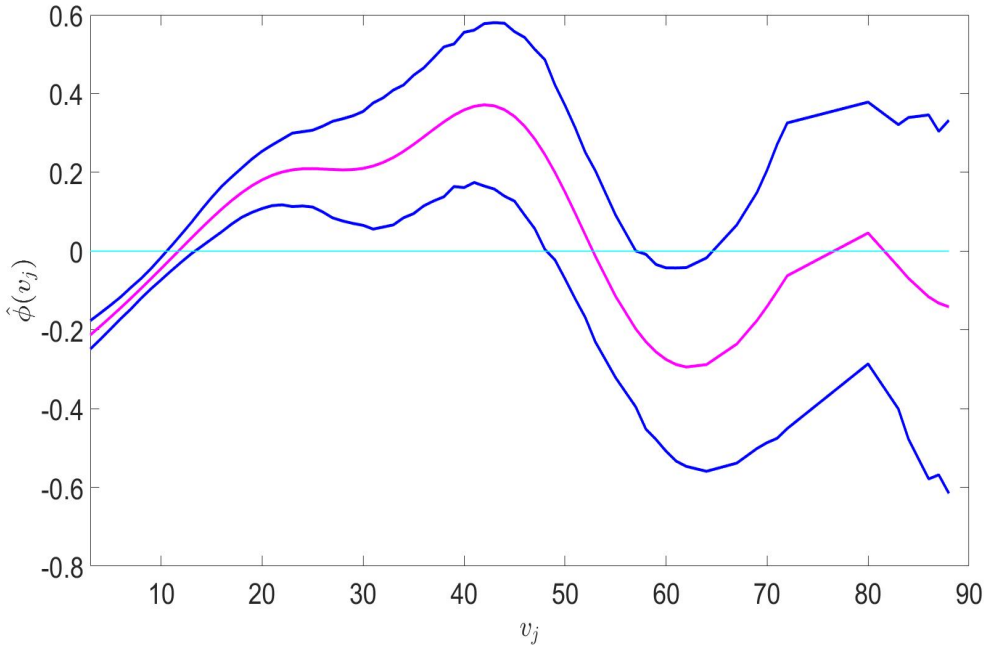

Figure 2: *Estimation results with full sample*

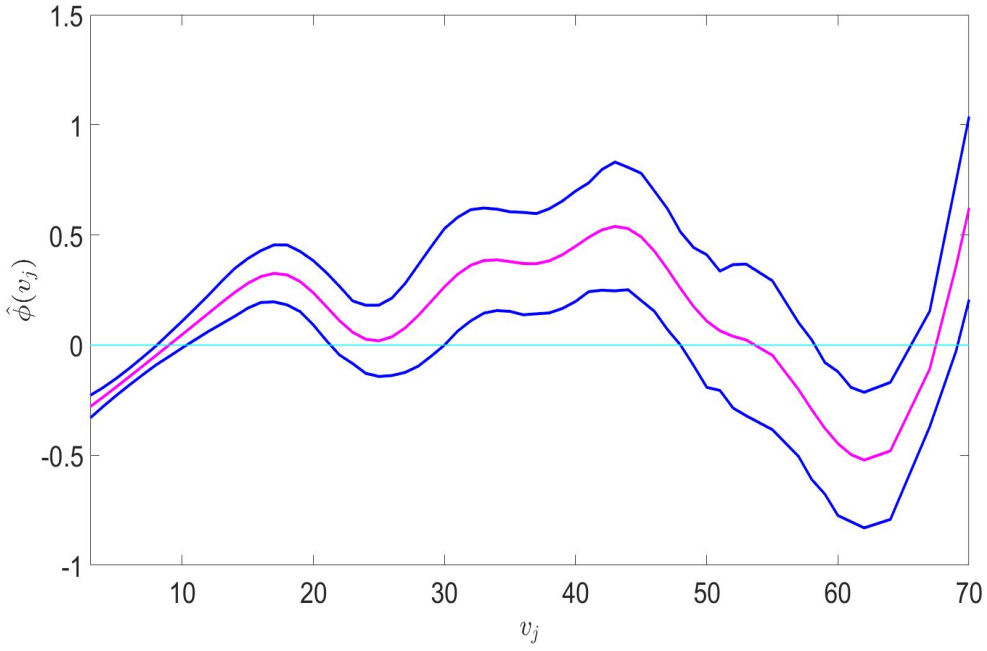

Figure 3: *Estimation results with less than 70 sellers*

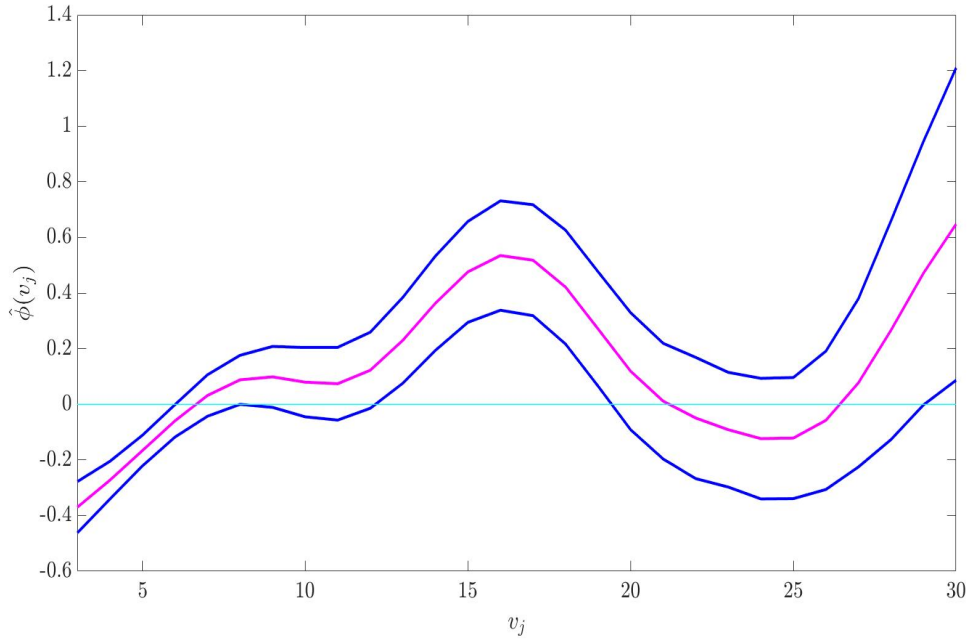

Figure 4: *Estimation results with less than 30 sellers*

Table 5: Estimation results

Parameter	**Estimate**	95% **Confidence Interval**
Price	0.0128	[0.0114 0.0142]
Reputation	-0.1422	[−0.2512 − 0.0332]
Size	0.0256	[0.0219 0.0293]
New	0.0290	[−0.3166 0.3745]
Professional	0.0395	[−0.0658 0.1448]
Age	-0.6289	[−0.6305 − 0.6274]

Table 6: Estimation results with Nonlinear ϕ

Parameter	**Estimate**	**Std. Error**	95% **Confidence Interval**
Price	0.0183	0.0067	[0.0053 0.0314]
Reputation	-0.1613	0.0499	[-0.2590 -0.0635]
Size	0.0237	0.0107	[0.0027 0.0448]
New	-0.0571	0.1028	[-0.2586 0.1443]
Professional	0.0398	0.0770	[-0.1123 0.1919]
Age	-0.6673	0.0285	[-0.7231 -0.6114]
Variety	0.2833	0.1705	[-0.0508 0.6174]
Variety^2	-0.0124	0.0067	[-0.0255 0.0007]

Hansen's J p-value: 0.2584

Robust Cragg-Donald test p-value: 0.0566

Table 7: Estimation results with Linear ϕ

Parameter	**Estimate**	**Std. Error**	95% **Confidence Interval**	
Price	0.0213	0.0053	[0.0109	0.0317]
Reputation	-0.1269	0.0395	[-0.2043	-0.0494]
Size	0.0147	0.0085	[-0.0019	0.0313]
New	-0.0757	0.0785	[-0.2295	0.0782]
Professional	-0.0174	0.0584	[-0.1318	0.0971]
Age	-0.6129	0.0091	[-0.6299	-0.5941]
Variety	-0.0501	0.0197	[-0.0886	-0.0115]

Hansen's J p-value: 0.2381

Robust Cragg-Donald test p-value: 0.0003

Table 8: Simulation Results

	Nonlinear NP ϕ				**Linear** ϕ			
	Δq	**Δp**	**Δr**	**Δm**	**Δq**	**Δp**	**Δr**	**Δm**
$\Delta v_j = 1$	0.096	≈ 0	0.96	+	-0.049	≈ 0	-0.049	−
$\Delta v_j = 25$	0.061	≈ 0	0.061	+	-0.71	≈ 0	-0.71	−
$\Delta v_j = 1$ & $\Delta t_{js} = 50\%$	0.084	0.21	0.31	+	-0.065	0.21	0.13	−
$\Delta v_j = 25$ & $\Delta t_{js} = -20\%$	0.066	-0.085	-0.026	+	-0.71	-0.085	-0.74	−

$\Delta x = (x_1 - x_0)/x_0$ where x_0 is the initial value.

References

Ackerberg, D. A. and M. Rysman (2005): "Unobserved Product Differentiation in Discrete-Choice Models: Estimating Price Elasticities and Welfare Effects," *The RAND Journal of Economics*, 36, 771–788.

Anderson, S. and O. Bedre-Defolie (2017): "Optimal Variety and Pricing in a Trade Platform," Tech. rep., mimeo.

Anderson, S. P. and A. De Palma (2009): "Information congestion," *The RAND Journal of Economics*, 40, 688–709.

Argentesi, E. and L. Filistrucchi (2007): "Estimating market power in a two-sided market: The case of newspapers," *Journal of Applied Econometrics*, 22, 1247–1266.

Armstrong, M. (2006): "Competition in Two-Sided Markets," *RAND Journal of Economics*, 37, 668–691.

Babii, A. and J.-P. Florens (2017): "Is completeness necessary? penalized estimation in non-identified models," *arXiv preprint arXiv:1709.03473*.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242–262.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 6 of *Handbook of Econometrics*, chap. 77.

Chen, Y. and T. Zhang (2013): "Entry and welfare in search markets," *The Economic Journal*.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565.

Dinerstein, M., L. Einav, J. Levin, and N. Sundaresan (2018): "Consumer Price Search and Platform Design in Internet Commerce," *American Economic Review*, 108, 1820–1859.

Draganska, M. and D. C. Jain (2005): "Product-line length as a competitive tool," *Journal of Economics & Management Strategy*, 14, 1–28.

Fève, F. and J.-P. Florens (2010): "The practice of non-parametric estimation by solving inverse problems: the example of transformation models," *The Econometrics Journal*, 13, S1–S27.

Florens, J.-P., J. Johannes, and S. V. Bellegem (2012): "Instrumental regression in partially linear models," *Econometrics Journal*, 15, 304–324.

Hagiu, A. (2007): "Merchant or two-sided platform?" *Review of Network Economics*, 6.

——— (2009): "Two-Sided Platforms: Product Variety and Pricing Structures," *Journal of Economics and Management Strategy*, 18, 1011–1043.

Hagiu, A. and B. Jullien (2007): "Designing a Two-Sided Platform: When To Increase Search Costs?" Idei working papers, Institut d'Economie Industrielle (IDEI), Toulouse.

Horowitz, J. L. (2011): "Applied nonparametric instrumental variables estimation," *Econometrica*, 79, 347–394.

Jolivet, G., B. Jullien, and F. Postel-Vinay (2016): "Reputation and prices on the e-market: Evidence from a major french platform," *International Journal of Industrial Organization*, 45, 59–75.

Jolivet, G. and H. Turon (2018): "Consumer search costs and preferences on the internet," *The Review of Economic Studies*, 86, 1258–1300.

Jullien, B. (2010): "Two-Sided B2B Platforms," IDEI Working Papers 652, Institut d'conomie Industrielle (IDEI), Toulouse.

Kim, J., G. M. Allenby, and P. E. Rossi (2002): "Modeling consumer demand for variety," *Marketing Science*, 21, 229–250.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010): "Online demand under limited consumer search," *Marketing science*, 29, 1001–1023.

Nosko, C. and S. Tadelis (2015): "The limits of reputation in platform markets: An empirical analysis and field experiment," Tech. rep., National Bureau of Economic Research.

Richards, T. J. and S. F. Hamilton (2015): "Variety pass-through: An examination of the ready-to-eat breakfast cereal market," *Review of Economics and Statistics*, 97, 166–180.

Rochet, J.-C. and J. Tirole (2003): "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 1, 990–1029.

Rysman, M. (2004): "Competition Between Networks: A Study of the Market for Yellow Pages," *Review of Economic Studies*, 71, 483–512.

Sokullu, S. (2016): "A Semi-Parametric Analysis of Two-Sided Markets: An Application to the Local Daily Newspapers in the USA," *Journal of Applied Econometrics*, 31, 843–864.

Windmeijer, F. (2018): "Testing over-and underidentification in linear models, with applications to dynamic panel data and asset-pricing models," *University of Bristol Department of Economics Working Paper*.

Appendices

A Data

A.1 Per Transaction Fee Tariff

The platform charges seller a variable fee and a fixed fee per transaction. The tariff is given as follows:

- **Fixed fee**
  - EUR0-5 → EUR0.05
  - EUR5-10 → EUR0.10
  - EUR10-15 → EUR0.20
  - >EUR15 → EUR0.40
- **Variable fee**
  - EUR0-50 → 22%
  - EUR50-100 → 18%
  - EUR100-300 → 12%
  - EUR300-600 → 8%
  - >EUR600 → 4%

Given this tariff for example, for a good sold at EUR300, it charges the seller $0.22 \times 50 + 0.18 \times (100 - 50) + 0.12 \times (300 - 100) + 0.40$ =EUR44.40.

A.2 Figures

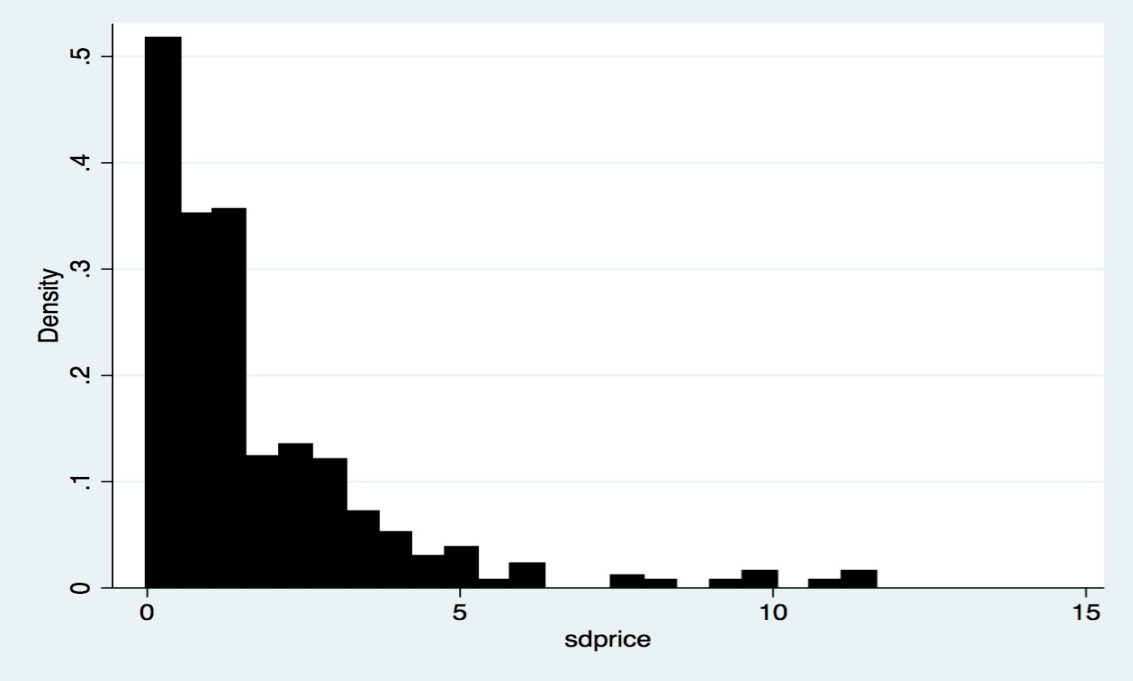

Figure 5: *Histogram of price variation*

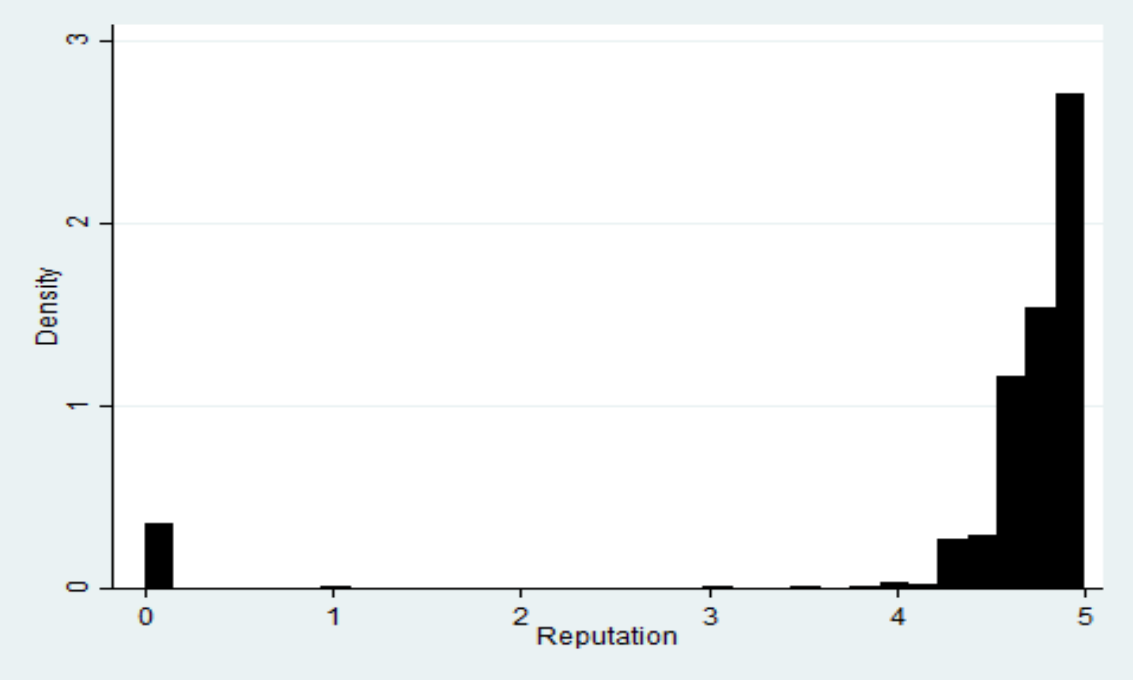

Figure 6: *Histogram of reputation*

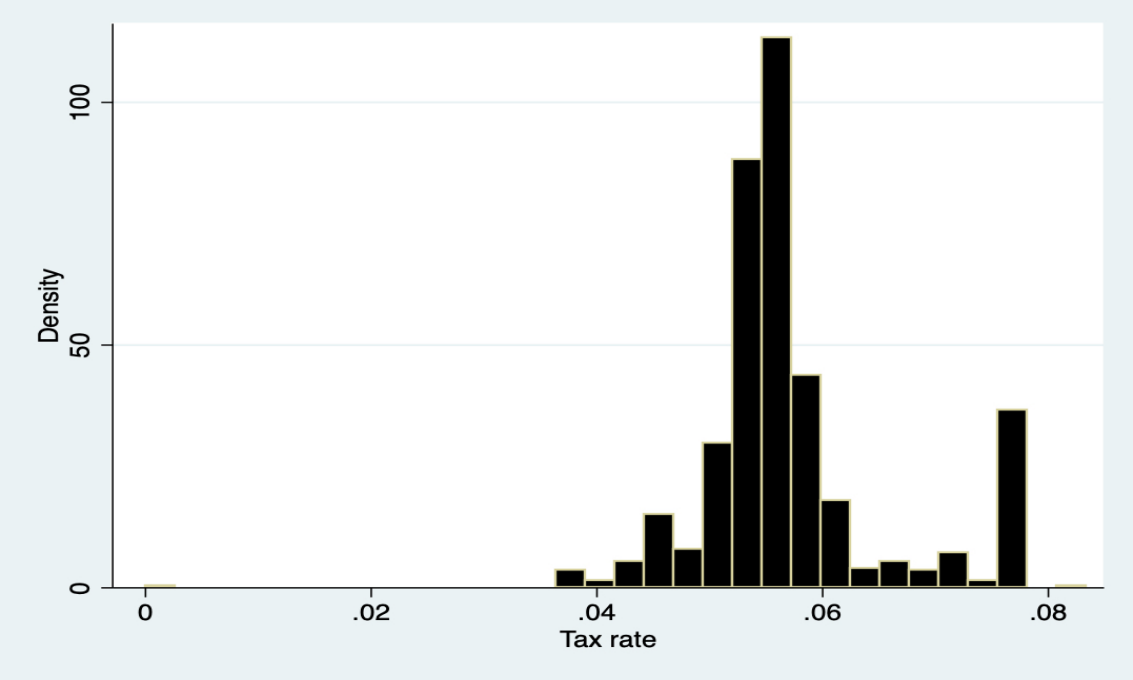

Figure 7: *Histogram of tax rate*

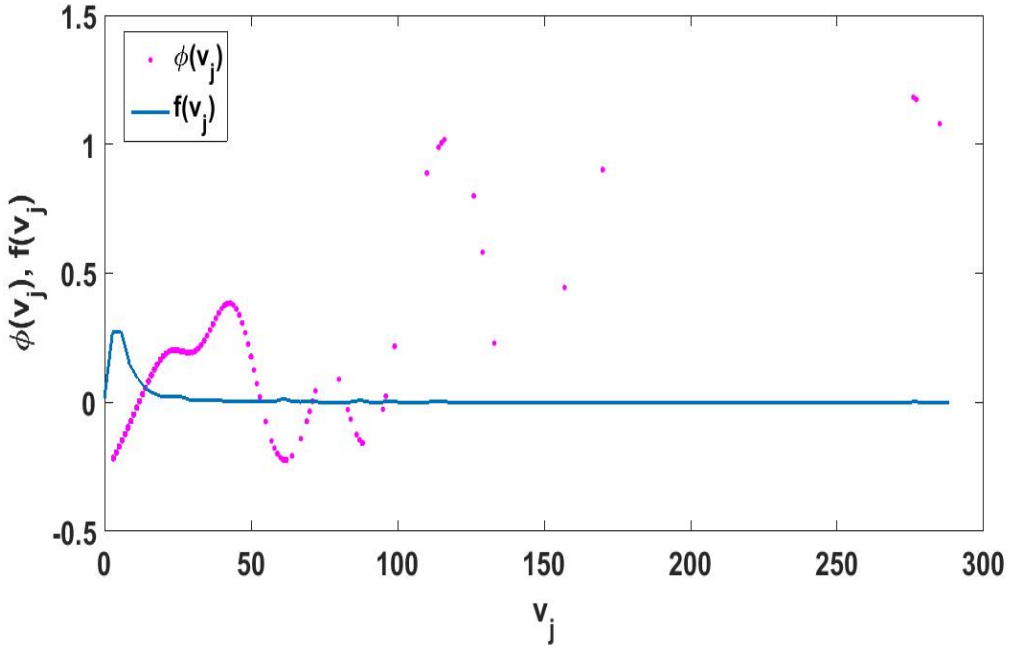

Figure 8: $\hat{\phi}(v_j)$ *vs. density of* v_j, $f(v_j)$

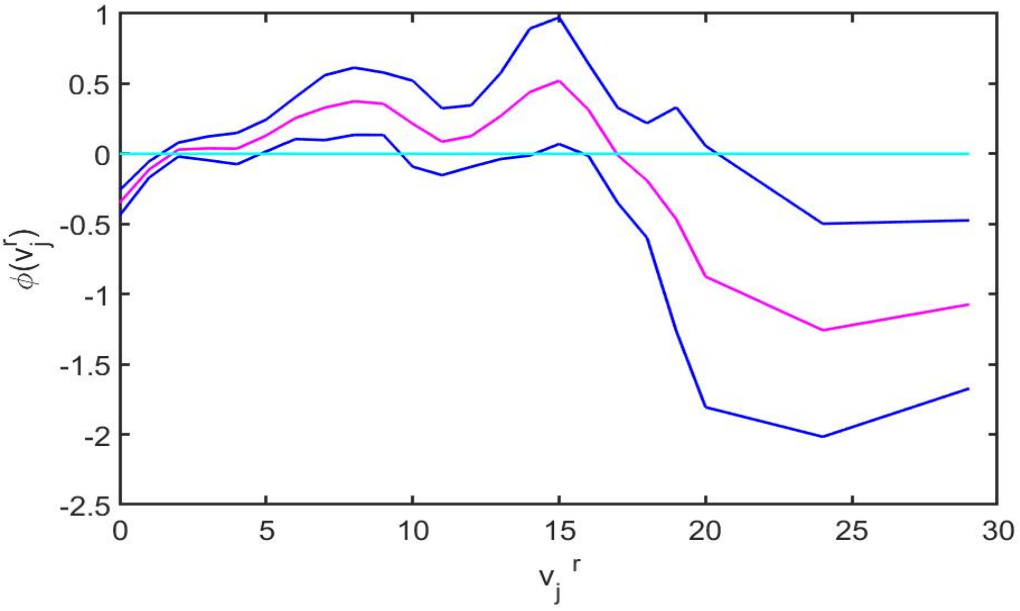

Figure 9: *Estimation results where variety is given my number of high reputation sellers*

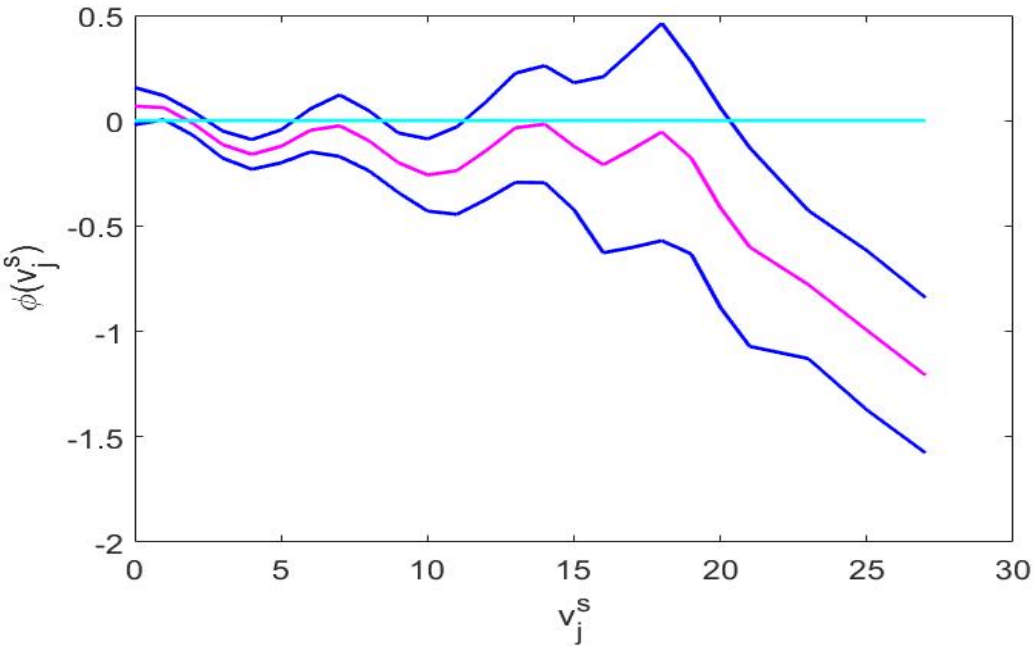

Figure 10: *Estimation results where variety is given my number of big sellers*

B Details of Non-Parametric Estimation Method

I would like to estimate demand which is given by Equation (3):

$$Y_{js} = X'_{js}\beta - \alpha p_{js} + \phi(v_j) + \xi_{js} \tag{11}$$

where $Y_{js} = ln(s_{js}) - ln(s_0)$. In the presence of a valid vector of instruments, one can write:

$$E[Y_{js}|X_{js}, Z_{js}] = E[W'_{js}\gamma|X_{js}, Z_{js}] + E[\phi(v_j)|X_{js}, Z_{js}] \tag{12}$$

where $W_{js} = [X'_{js} \quad p_{js}]'$ and $\gamma = (\beta, \alpha)$. Assume that (Y, P, V, X, Z) generate a random vector, Ξ, which has a cumulative distribution function F. Then for each F, define subspaces of the variables as $L^2_F(Y), L^2_F(P), L^2_F(V), L^2_F(X)$ and $L^2_F(Z)$ which belong to a common Hilbert space. That is, $L^2_F(Y)$ denotes the subspace of L^2_F of real valued functions depending on Y only. In the sequel, I use the notation L^2_Y to denote the $L^2_F(Y)$. Let T_v be an operator defined as follows:

$$T_v : L^2_V \mapsto L^2_{X\times Z} : T_v(\phi(V)) = \mathbb{E}[\phi(V)|X, Z].$$

Moreover, define the operator T_w as follows:

$$T_w : \mathbb{R}^k \mapsto L^2_{X\times Z} : T_w\gamma = \mathbb{E}[W'\gamma|X, Z]$$

One can then rewrite the Equation (12) as the following:

$$r = T_w\gamma + T_v\phi$$

where $r = E[Y|X, Z]$. If I denote the adjoint operators of T_v and T_w by T^*_v and T^*_w respectively, the normal equations can be written as:

$$T^*_w r = T^*_w T_w\gamma + T^*_w T_v\phi \tag{13}$$

$$T_v^* r = T_v^* T_w \gamma + T_v^* T \phi \tag{14}$$

Using Equation (13), one can get an expression for γ and replacing it back in Equation (14) gives the solution for ϕ function:

$$\gamma = (T_w^* T_w)^{-1} T_w^* (r - T_v \phi)$$

$$T_v^* (I - P_w) r = T_v^* (I - P_w) T_v \phi \tag{15}$$

where $P_w = T_w (T_w^* T_w)^{-1} T_w^*$. Note that, one can solve for ϕ by inverting the operator $T_v^*(I - P_w)T_v$. However, this operator is infinite dimensional and it has infinitely many eigenvalues around zero, when it is inverted, the solution for ϕ becomes unstable, in other words, it is ill-posed. Indeed this is a well-known problem in the NPIV literature and it is suggested to stabilise the solution by using some penalization or truncation techniques. The parametric counterpart of this problem arises when we have multicollinearity which then leads to nearly singular second moment matrix of the covariates. Following Florens et al. (2012) I obtain a stable solution for ϕ using penalization, in other words, I regularise the ill-posed problem. This is equivalent to use Ridge regression in a highly collinear parametric model.[27] The Tikhonov regularized solution of ϕ is given by the following:

$$\phi_\alpha = (\alpha I + T_v^* (I - P_w) T_v)^{-1} T_v^* (I - P_w) r \tag{16}$$

where α is a strictly positive regularization parameter, in other words, the penalty, which approaches to zero as the sample size tends to infinity. The estimates of ϕ and γ can then be obtained by replacing the operators by their estimators.

Assume that an i.i.d data of random variables $\{y_i, p_i, w_i, v_i, z_i, \}_{i=1}^n$ exists and define the

[27]For more details on ill-posed problems in econometrics, see Horowitz (2011), Darolles et al. (2011),Carrasco, Florens, and Renault (2007) among others.

$(n \times n)$ matrix A_{wz}, whose $(i,j)th$ element is given by:

$$A_{wz}(z)(i,j) = \frac{K_w\left(\frac{w_i - w_j}{h_w}\right) K_z\left(\frac{z - z_j}{h_z}\right)}{\sum_j K_w\left(\frac{w_i - w_j}{h_w}\right) K_z\left(\frac{z - z_j}{h_z}\right)},$$

and define the $(n \times n)$ matrix A_v whose (i,j)th element is given by:

$$A_v(i,j) = \frac{K_v\left(\frac{v_i - v_j}{h_v}\right)}{\sum_j K_v\left(\frac{v_i - v_j}{h_v}\right)}$$

where K_s is the kernel function and h_s is the bandwidth for the random variable $s = w, z, v$. Moreover $\hat{P}_w$ matrix is estimated by $\hat{P}_w = A_{wz}W(W'A_{wz}W)^{-1}A_{wz}W'$ where W is $n \times k$ martix. Then the estimator of $\hat{\phi}$ is given by:

$$\hat{\phi}_\alpha = (\alpha I + A_v(I - \hat{P}_w)A_{wz})^{-1}A_v(I - \hat{P}_w)A_{wz}y. \tag{17}$$

Once the estimate of ϕ is obtained, it can be replaced back in the first normal equation and an estimate of γ can be obtained. One challenge in this estimation is the selection of the smoothing parameters h_w, h_z, h_v and regularization parameter, α. This problem is well studied in the literature and it has been shown that whichever the way the bandwidths are selected, the regularization parameter adopts itself when we use a data-based selection method for it, see Fève and Florens (2010). In the estimation, I use a rule of thumb to select the bandwidths first and then given the bandwidths I use a data-based selection rule for α.

C Illustration: Nonlinear Parametric Model vs Nonparametric Model

Below I present results of the simple counterfactual simulation where I compare the predictions from nonlinear parametric model and nonparametric model. For scenarios 1 and

3 where the number of sellers increase by 1 for each book on the platform, the two models predicts qualitatively the same outcome. This is not the case for scenarios 2 and 4 where I assume that the number of sellers increase by 25. It should be noted that nonparametric estimation results show that ϕ function becomes insignificant after around 20 adverts, $v_j = 20$ and the simulation with nonparametric structural model takes this into account. Nonlinear parametric model predicts that increasing number of sellers keeps decreasing the utility of buyers once the maximum benefit is reached, and hence predicts a decrease in average quantity.

Table 9: Simulation Results

	Nonlinear NP ϕ				**Nonlinear P ϕ**			
	Δq	**Δp**	**Δr**	**Δm**	**Δq**	**Δp**	**Δr**	**Δm**
$\Delta v_j = 1$	0.096	≈ 0	0.96	+	0.31	≈ 0	0.31	+
$\Delta v_j = 25$	0.061	≈ 0	0.061	+	-0.49	≈ 0	-0.49	$-$
$\Delta v_j = 1$ & $\Delta t_{js} = 50\%$	0.084	0.21	0.31	+	0.29	0.21	0.56	+
$\Delta v_j = 25$ & $\Delta t_{js} = -20\%$	0.066	-0.085	-0.026	+	-0.488	-0.085	-0.53	$-$

$\Delta x = (x_1 - x_0)/x_0$ where x_0 is the initial value.

D Estimation results using samples of CD's and DVD's

I estimate the model given by equation 3 also using samples consist of CDs and DVDs only. As it is done in the case of books, a screenshot of data is taken from September 2007. The results are very similar to those presented in Section 5.1. Figure 11 shows the estimation result with the sample of CDs and Figure 12 shows the estimation result obtained with the sample of DVDs.

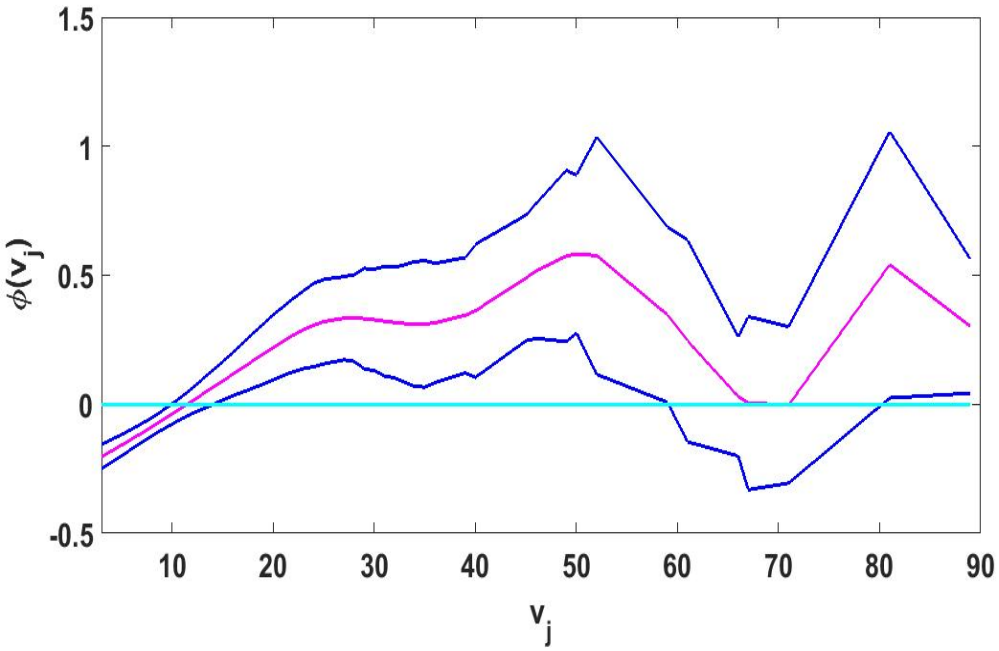

Figure 11: *Estimation results with CD's*

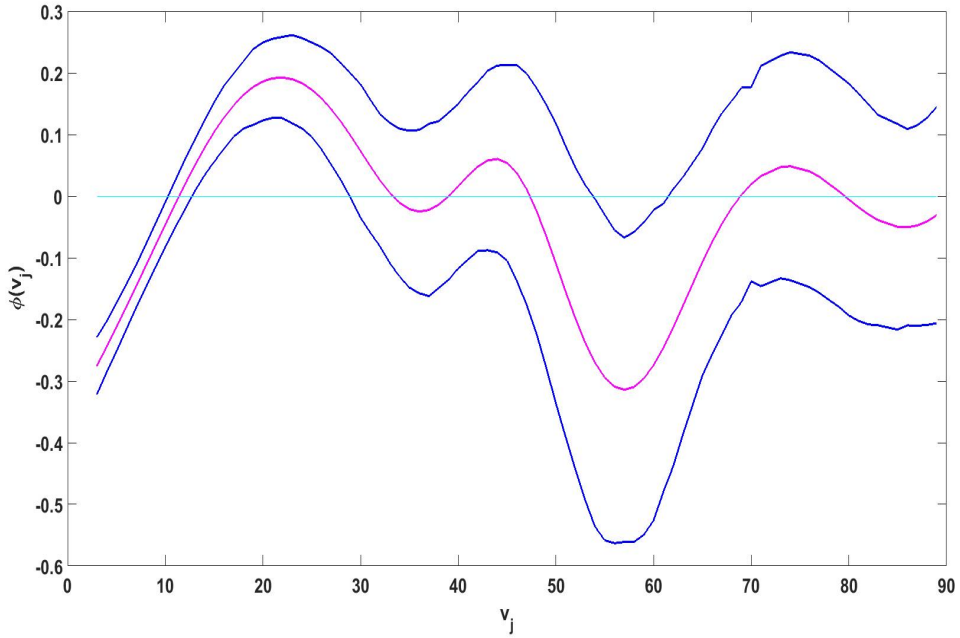

Figure 12: *Estimation results with DVD's*

It should be noted that Figure (11) is very similar to figure obtained by the whole sample

of the books. One can see that the effect of number of sellers is insignificant after some point. We observe the same thing in Figure(12). Up to 30 sellers, again the figure is very similar to that obtained with sample of books where only the observations with less than 30 variants are used. After 30 sellers, the effect becomes insignificant. Hence one can conclude that the nonmonotonic network effect is still there when samples from different goods are used.