Evaluating How Much Heterogeneity Can Explain Violations of the "Law of One Price"

Aaron Bodoh-Creed, Jörn Boehnke, and Brent Hickman

Abstract

We use a unique dataset consisting of posted price sales of new Kindle Fire tablets from the eBay "Buy It Now" market platform to assess the degree to which heterogeneity can explain the pricing of seemingly homogeneous products. By combining a rich set of data about each listing with machine learning techniques, we find that we can explain almost 42% of the price variation. An analysis using more traditional data and OLS techniques explains less than 13% of the observed price variation. We conclude that heterogeneity amongst seemingly homogenous products can play a significant role in explaining price variation even in markets for seemingly homogeneous products.

1 Introduction

The "Law of Pne Price" predicts that all exchanges of homogenous goods in a thick, frictionless market ought to take place at a single price. While in some markets, most notably security exchanges, the "Law of One Price" holds very well, in most consumer product markets the Law of One Price fails to describe the reality faced by everyday consumers. This fact is pithily summarized by Varian, who wrote "the law of one price is no law at all" (Varian [30]).

The law of one price is a result of a stylized model of perfect competition in which all potential buyers keen on purchasing a homogenous good observe the prices on offer from all of the sellers before buying the good from the seller offering the lowest price. Explanations for the failure of the law of one price must violate one of the assumptions of this simplistic model. Candidate causes include information asymmetries, consumer search frictions, and underlying product heterogeneity. One might predict that the law of one price would be more robust online as the internet has reduced search frictions, but empirical research has established that price dispersion online is of a comparable magnitude to that observed in

brick and mortar retailers even for products that appear homogenous (e.g., online book sales).

The existence of price dispersion in online markets is a well documented fact (e.g., Bailey [3], Brynjolfsson and Smith [11]). The theoretical literature has advanced by providing models of novel search frictions that explain price variation even in the presence of information clearinghouses (e.g., price aggregation sites) and other tools that have reduced the cost of obtaining a price quote. Although theoretical models of information asymmetries and price clearinghouses are not new (e.g., Varian [30]), the more recent literature has treated the price clearinghouses as important strategic actors in the affected markets (e.g., Baye and Morgan [4], Baye et al. [7]). A large branch of the more recent empirical literature on price dispersion has focused on tests of the various models (e.g., Baye, Morgan, and Scholten [5], [6], [9]). There are also studies that attempt to predict product prices and report statistics that describe their explanatory power. For example, Baye, Morgan, and Scholten [8] attempts to predict the price dispersion for online consumer electronics sales. Their regressions can explain 17% of the price variation using regressors capturing the attributes of the product and the retailers.

In a model with rational buyers and sellers, price variation can arise from three sources. First, it could be that units of a given product are heterogeneous in a fashion that is difficult to observe in the data. Second, it could be that sellers offer complementary services that buyers value (e.g., a long-term warranty). Third, and finally, price variation could be an endogenous outcome of market competition that is independent of product heterogeneity. The first two sources capture dimensions of product heterogeneity, and one interpretation of the prior literature is that either these sources do not explain much of the price variation or that data reflecting these dimensions of heterogeneity is difficult to collect.

The goal of our paper is to assess whether subtle aspects of product heterogeneity can explain price variation among seemingly identical goods.[1] To this end we have scraped a large set of posted-price listings for new Amazon Kindles from the eBay market platform. Few markets more closely resemble the canonical market of perfect competition than the posted-price market for new Kindles on eBay. This market is very thick with a large number of buyers and sellers trying to conduct exchange. The search features eBay provides make it essentially costless to obtain a price quote. Moreover, since the typical seller has only a single unit and few outside opportunities for sale, the information asymmetries sometimes used to explain price dispersion would not seem to apply. Finally, since new Kindles are

[1]To be clear, throughout this work, when we use the word "explain price variation" we mean that we are capable of predicting the price variation rather than provide a causal explanation.

intrinsically very similar, the sellers do not offer any complementary services, and eBay offers a strong customer satisfaction guarantee with uniform buyer protection measures, one would expect the buyers to view the sellers' products as nearly perfect substitutes. All of these features ought to suggest that price variation should be minimal in our data, and yet we find that the standard deviation of the price is equal to 21.2% of the mean.

Many of the obvious sources of product heterogeneity are ruled out in our setting. For example, one might have thought that the Kindles for sale on eBay are commonly bundled with accessories, but this turns out to be rare in our data and, when present, the accessories are of relatively low value. One might also assume that issues such as seller reliability induce significant price heterogeneity, but eBay's strong warranty against seller misbehavior should alleviate this concern.

The listings provide a wide array of features that might help explain why the price of that listing is unusually high or low. The first portion of each listing is a standardized description of the product provided by eBay as well as information about the seller's reliability (feedback score), shipping cost, and the posted sale price. The standardized information provided by eBay does an excellent job of concisely spelling out the technical features of the Amazon Kindle as well as eBay's definition of a "new product." After the standard information provided by eBay, each listing then displays information provided by the seller (e.g., photos, text descriptions, warranties the seller offers, etc.). We scraped any additional text information the seller provided about the product. Finally, we observed the number of photos the seller posted.

The natural language text varied widely from listing to listing. Because we have scrapped the entire content of the listing, we can potentially analyze almost everything that the buyers see about each listing. For example, listings have an average of 131 words of text written by the seller, but the standard deviation of the number of words is 280 words. Similarly, the listings had an average of 4.09 photos and the standard deviation of the number of photos is 4.39. Since we parse the content of the test using a bag-of-words approach, we are left with over 200 regressors that characterize each of the 1298 listings in our data set. To assess such a large set of potential regressors, we employ machine learning techniques both to assess which features are most important and how to combine them to form an effective price prediction model.

We break our explanation of the price variation into two parts. First, we aim to assess the importance of using the rich set of regressors we have collected. To make this assessment, we construct a basic data set that only includes regressors that are comparable to those employed in the prior work on price dispersion. We then judge the importance of our rich

set of regressors by comparing the explanatory power of price prediction models built using the basic data set relative to a prediction model built using the full data set. Second, we assess whether common econometric techniques such as multivariate regressions can explain the price dispersion when applied to our rich data set or whether a more sophisticated approach to prediction such as machine learning is required to exploit subtle patterns in the data to predict price dispersion.

We find that, in line with the previous literature described below, we can explain 13% of the price variation using an ordinary least squares (OLS) model estimated using our basic data set.[2] Surprisingly, an OLS model built using our full data set can only explain 14% of the price variation. In other words, the rich set of regressors improves our predictive power, but only slightly.

We then examine the predictive power of a model based on a regression forest (Breiman [10]). A regression forest predicts values by averaging the predictions of a large ensemble of regression trees . Much like a K-nearest neighbor or a kernel smoothed regression, an individual regression tree uses observations that are near the point of interest to generate a prediction. A single regression tree partitions the space of regressors to define what "near the point of interest" means using a data driven approach. Regression forests have proven popular due to their ability to capture complex interactions between large sets of regressors in a principled way that allows for relatively little input from the econometrician.

When we apply our regression forest techniques to the basic data set, we can explain 13% of the price variation, which is almost identical to the effectiveness of the OLS estimator. However, when we apply our regression forest methods to the full data set, we can explain over 40% of the price dispersion. The fraction of the explained price variation is also economically significant as it amounts to over 10% of the mean price of the new Kindles.

One criticism of our OLS approach is that we may have estimated an insufficiently flexible model. To explore this possibility, we estimated a linear model with a complete set of first, second, and third order interactions of our data, which results in a model with 6,463 non-redundant regressors. After using LASSO to choose our set of regressors, we find that the linear model explains 23% of the variation in prices in our preferred specification. Our conclusion is that while a more flexible linear model can (unsurprisingly) predict a greater degree of price variation, a linear model would have to be impractically flexible to begin to approach the effectiveness of machine learning techniques.

[2]This comparison with the previous literature is not intended as a model selection exercise for many reasons (e.g., the differing data sets). We wish to make the simpler point that there exists plenty of price variation that remains unexplained if one relies on OLS techniques as in the prior literature.

<Joern Stuff here>

The final takeaway is that while there is a significant degree of product heterogeneity, it requires both (1) a rich description of the products and (2) a flexible and principaled estimation tool (e.g., regression forests) to form an effective prediction model. Given the large heterogeneity even for new electronics, our results suggest that many products are relatively unique. Contrary to the quote from Varian above, the law of one price may be a law — just a fairly vacuous one.

2 Related Literature

Price dispersion has been viewed as a consequence of ignorance at least since Stigler [29]. Building on Stigler's original model of costly search, Diamond [14] proved that profit maximizing firms can act as monopolists if consumers face search costs. Although the model of Diamond [14] does not yield equilibrium price dispersion, it does show that large deviations from the perfectly competitive outcome are possible if consumers face even small search costs. Reinganum [25] shows that price dispersion can arise when consumers discover prices through a process of sequential search and firms have heterogeneous marginal costs. MacMinn [21] shows that price dispersion can also rise under this market structure when fixed sample search is used.

The relevance of these "quote cost" frictions have become questionable in light of the ease of price-comparison on the internet. In particular, the rise of price aggregator websites that provide price data from a large sample of retailers might lead one to think that the cost of obtaining a price quote has been reduced to the point of economic insignificance. However, it is equally true that there remains price variation even amongst firms that provide data to price aggregation websites that serve as information clearinghouses. Although theoretical models of information asymmetries and price clearinghouses are not new (e.g., Varian [30]), the more recent literature has treated the price clearinghouses as important strategic actors in the affected markets (e.g., Baye and Morgan [4], Baye et al. [7]).

Equilibrium price dispersion arises in these models as a result of the firms facing an incentive to exploit loyal customers by offering a high price as well as an incentive to offer low prices to attract customers through the price clearing house. These models do not provide a particularly convincing explanation of price dispersion on eBay since most of the sellers are individuals that lack any analog of the loyal customers assumed in these models. Therefore, it is perhaps even more surprising that we find the large degree of price dispersion on eBay even for a relatively homogenous product like a new Kindle.

In summary, there is little in the theoretical literature that would lead one to expect significant price dispersion in the eBay markets. One might, however, conjecture that if demand is uncertain, then heterogeneity in the reservation values/production costs of the sellers might generate price dispersion. Appendix B provides a stylized model of such a situation in a dynamic context and proves that significant price dispersion requires either unrealistically impatient sellers or an unrealistically high degree of heterogeneity in reservation values. The core intuition is that if demand is uncertain and sellers are patient, then it is optimal for each seller to price near the maximum possible sale price since, given the seller's patience, there is little harm in waiting for such a high demand period to occur.

As discussed above, the existence of price dispersion in online markets is a well documented fact (e.g., Bailey [3], Brynjolfsson and Smith [11]). A large branch of the more recent empirical literature on price dispersion has focused on tests of the various models. For example, Sorenson [28] shows that pharmaceutical products that necessitate repeated purchases have lower price variation, which makes sense given consumers have a strong incentive to find a low price for products that are purchased many times. Baye, Morgan, and Scholten [5] and [6] use data from a price comparison web site and data on the market structure across different products to test the implications of several different information clearinghouse models. Baylis and Perloff [9] find a combination of high quality, low priced firms competing with low quality, high priced firms in the online markets for scanners and digital cameras, which the authors interpret as support for the two price equilibrium predicted by Salop and Stiglitz [27]. In addition, some papers estimate a structural model and attempt to tease apart the sources of price variation based on the estimates of the structural primitives (e.g., Hong and Shum [19]).

The focus of our paper is to identify features of the listings that generate price dispersion. There are a few prior studies that attempt to predict product prices and report statistics that describe their explanatory power. Many of the estimated models have features that make it very difficult to interpret the explanatory power. Among the papers that are comparable to our project, Baye, Morgan, and Scholten [8] attempts to predict the price dispersion for online consumer electronics sales. Their regression can explain 17% of the price variation using regressors capturing the attributes of the retailers, and the explanatory power jumps to 72% when the regressions include firm specific dummy variables.

Even when seller dummy variables can explain a great deal of the price variation, it is unclear whether the dummy variables are capturing. For example, suppose that we conclude based on a regression explaining the prices of electronics that Best Buy, a brick and mortar electronics retailer that also has an online store, has consistently higher prices

than other electronics retailers. The higher prices at Best Buy could be because the products are slightly different (source one: product heterogeneity), it could be that Best Buy offers generous return policies (source two: heterogeneous retailers), or that Best Buy has a near monopoly over brick and mortar book sales in many regions that allows the firm to charge higher prices (source three: market competition). In other words, including dummy variables for individual sellers that are highly predictive does not shine much light on the underlying cause of the price variation.

Clay, Krishnan, and Wolfe [12] attempt to predict prices and achieve a high degree of explanatory power, but their regressions include time dummies. Time dummies explain a great deal of the price variation across our entire sample due to product depreciation, but this price variation is unrelated to the day-to-day price dispersion we are trying to explain. This makes it impossible to compare the explanatory power of these regressions with our analysis. Clay, Krishnan, and Wolfe [13] provides an analysis of the price dispersion of text books that explains 2.7% when regression do not include store-level dummy variables and 19.2% of the dispersion when the dummy variables are included. Pan, Ratchford, and Shankar [23] study the price dispersion across eight categories of retail products and can explain 22% or less of the price dispersion, with the notable exception being that their regressions explain 43% of the price variation of compact discs. Our general conclusion from the empirical literature is that price dispersion is in general quite difficult to explain without including regressors such as retailer specific dummy variables.

We would also like to highlight a handful of papers that have worked directly with eBay Buy It Now data. For example, Hui et al. [20] studies the interactions between the effects of reputational mechanisms and insurance against seller misbehavior on the prices received by sellers in Buy It Now and auction listings on eBay. Saeedi and Sundaresan [26] study a sample of Buy It Now and auction listings on eBay to understand the effect of a change in the reputation system on buyer and seller behavior. Other papers have studied the relationship between Buy It Now postings and auctions (e.g., Einav et al. [15], Einav et al. [16], Einav et al. [17]). Nosko and Tadelis **??** documents that buers' experiences with sellers spillover onto other sellers, and the authors propose a novel and more effective metric of interaction quality. Elfenbein, Fisman, and McManus **??** study the interaction of the value of quality certification and market structure. To the best of our knowledge, we are the first to use data from a platform like eBay to study price dispersion, emphasize the role of a set of contextual data (e.g., text or images) as rich as ours, or bring machine learning techniques to bear to explain price dispersion.

3 Data

We gathered our data from the eBay market platform, which uses a finely grained product classification system that dilenates between new and used products and different product specifications of the same product. For example, Kindles with different amounts of storage are treated as distinct kinds of products. The very specific product classification system encourages product heterogeneity within product categories.

Each data point is a single listing posted for a new, 1st generation Amazon Kindle Fire. We collected our posted price listings, known as "Buy It Now" listings on the eBay platform, from sellers located within the United States. We included only listings offering a single unit, and we eliminated listings with implausible prices (i.e., below $15 and above $250). Our scraping program captured listings that either were sold or removed from the platform without sale for the period from January 1, 2013 to September 30 2013, which yielded a total of 1298 listings. Since our goal is to assess the drivers of seller pricing decisions, we consider a listing to be a single data point even if the seller offers multiple units for sale in the listing. If a seller offers multiple listings across our sample, we treat these listing as distinct data points. There are 911 unique sellers in our data set, 5 of which have 10 or more listings. The vast majority of sellers have very few listings: 79.5% of sellers have a single listing and another 12.7% have 2 listings.

The seller is allowed to choose the duration of the listing from a discrete set of options: 1, 3, 5, 7, 10, or 30 days. In addition, the seller can choose to have the listing exist until it is otherwise canceled. Finally, if a listing's duration expires without a sale, then the user has the option of relisting it.

Posted price listings of Kindles on eBay follow a format with two general sections. At the top of each listing's webpage is a standardized description of the product for sale that is provided by eBay. This section includes information that is uniformly formatted across all listings such as the price chosen by the seller, the seller's reputation score, and the technical specifications of the Kindle. The second section of each listing is created by the seller and can vary wildly from post to post. The seller-created section can include photos, a textual description of the kindle (e.g., "8 GB of storage"), contact information for the seller, etc..

Figure 3 provides a time series plot of the median price of the listings as a function of date with the band describing the 25-75 spread in the distribution of prices on each day. All of the time series have been smoothed using a seven day moving average filter. Two features are of note. First, although at the beginning of our sample the Kindle is a relatively new product and the eBay market for Kindles is relatively immature, the market shows no sign of

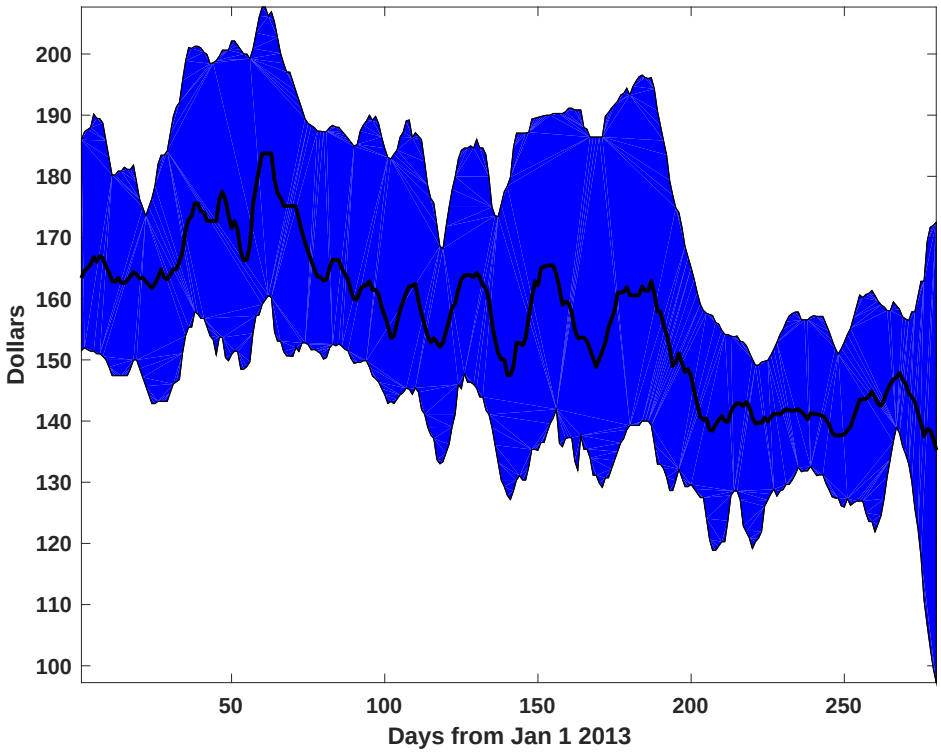

Figure 1: Price Trend Over Time

converging toward satisfaction of the law of one price by the end of our nine month sampling period. The persistence of price dispersion is well documented in other online markets (e.g., Baye, Morgan and Scholten [8]), so this is not terribly surprising.

The second feature to note in Figure 3 is the marked trend towards lower prices as the sample period persists. Again, this is not terribly surprising since electronics products depreciate as the anticipated release dates of newer versions approach. It is worth taking a moment to consider the ideal data set for our purposes and how this influences how we handle the time trend in our analysis. The ideal data set would be a snapshot of the prices offered in a very large market, which would allow us to hold all time-varying features fixed and isolate what about the product for sale (i.e., information that is reflected in the scrapped listing) caused the seller to believe that an unusually high or low price was warranted. Unfortunately, we do not have enough listings on any given day, so we collect listings across days. If we were to include time dummies or a time trend in our regressions, then we would be able to explain a great deal of price variation simply through these time-dependent variables. Since the focus of our research question is cross-sectional price variation rather than price variation over time, the appropriate course is to detrend our price variables. Figure 3 shows

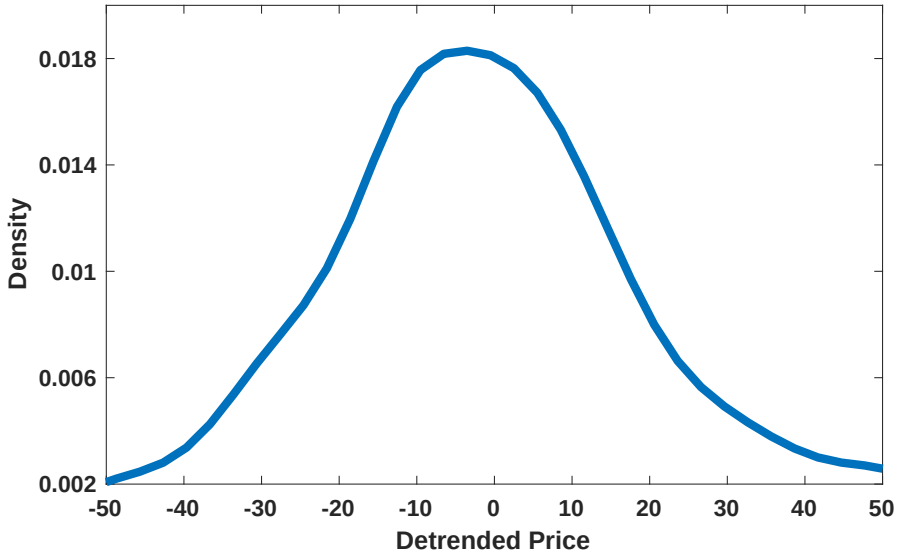

Figure 2: Detrended Price Distribution

the distribution of the detrended prices, which still displays substantial dispersion.

For our *basic data set* we only include variables present in the portion of a listing's webpage that is automatically generated by eBay. This data set is intended to reflect the data used in earlier papers that attempt to explain online price variation (e.g., Baye, Morgan and Scholten [8]). *Price* is either the price at sale or, for items that did not ultimately sell, the final price the seller offered before removing the unsold item from the site. *Shipping Price* is the price of shipping if a flat rate was included in the listing and 0 otherwise. *Shipping Calculated* is a dummy variable that is set to 1 if eBay automatically provided a forecasted price for shipping based on the seller's and the prospective buyer's locations, package and weight size estimates provided by the seller, and the seller's choice of shipping company. *Returns Allowed* is a dummy variable set to 1 if the seller accepts returns. *Seller Score* is a numeric value indicating the net positive feedback left by individuals that had purchased from this seller previously. *Relisted* is a dummy variable set to 1 if the seller chose to relist this item after the item did not during the listing's initial duration.[3]Variable names and summary statistics for the basic data set are included in Table 1.

We view our basic data set as an analog to the data used in previous studies of online price dispersion. As is typical of the prior work (e.g., Baye, Morgan, and Scholten [8]), we have information on price, shipping fees, and some aspects of seller reputation. There

[3]Whether an item is relisted is part of the metadata in a listing's .HTML code, which is not easily observed by buyers.

Variable	Mean	Median	Standard Deviation
Price	0	-$1.74	$30.23
Shipping Price	$3.71	0	$4.96
Shipping Calculated	0.437	0	0.496
Returns Allowed	0.303	0	0.460
ln(Seller Score)	4.78	4.72	2.02
Relisted	0.168	0	0.374

Table 1: Basic Data Set Summary Statistics

are a few major differences, however. First, prior work studied the pricing decisions of large retailers that sold many units to potentially very different consumer segments (e.g., brick and mortar versus online only). Our data consist mostly of small sellers with a single unit for sale. Second, the sellers all used the same infrastructure provided by eBay. This infrastructure includes, for example, eBay's money back guarantee for buyers, which can be triggered easily through the website and results in a rapid (less than five days) refund of the money paid to the seller. This reduces the scope for possible differences between the sellers. Third, and most usefully, with the exception of the content of the photos contained in the listings, we can observe all information contained in a listing.

Our *full data set* includes all of the information in the basic data set as well as information culled from the portion of the listing that can be customized by the sellers. These data include the number of characters, words, special characters, and fraction of upper case characters in the title and body of the listing. We also record the number of words and the percentage of upper case letters in the in the body of text. We capture the number HTML tags (e.g., sections of bold text), the number of font sizes, and the number of changes in font size in the seller's description of the product. These variables all reflect techniques that a seller might use to make a text eye catching. We record the number (although not the content) of the photos provided by the seller. We also record a categorical variable describing whether the listing started on the weekend (Saturday or Sunday), early in the week (Monday - Wednesday), or late in the week (Thursday or Friday). Finally, we record whether the listing was generated by eBay's mobile phone app. Summary statistics are provided in Table 2.

The natural language data was handled with a Bag of Words approach. First we separated each listing's text into sentences and words, and each word was stemmed using Porter's Stemming Algorithm (Porter [24]). The stemming algorithm is capable of identifying different forms of the same word. For example, the stemmer can identify "charges," "charged," and "charging" as sharing the same root "charge." Correctly stemming the text of the list-

Variable	Mean	Median	Standard Deviation
Number of Characters in Title	60.1	66	16.4
Number of Words in Title	10.6	12	2.68
Number of Special Characters in Title	15.0	16	4.63
% Uppercase Characters in Title	0.312	0.2632	0.168
Number of Words in Body	131	27	281
% Uppercase Characters in Body	0.122	0.0753	0.184
Number of HTML Tags in Body	102	16	201
Number of Font Sizes	1.58	1	1.451
Number of Font Size Changes	4.33	0	16.0
Number of Photos	4.09	1	4.39
Posted During Weekend	0.277	0	0.448
Posted During Early Week	0.429	0	0.495
Posted During Late Week	0.294	0	0.456
Posted with eBay Mobile	0.242	0	0.428

Table 2: Full Data Set Summary Statistics (without Bag of Words data)

ings. both removes potentially redundant features and correctly assess how frequently the words appear in each listings. We do not attempt to identify negations (e.g., "no returns") algorithmically as this is significantly more computationally difficult and subject to a greater error rate.

After the stems have been identified and the number of occurrences of each stem in each listing has been computed, we reduced the dimensionality of the natural language data in two steps. First, we formed a list of the 1,000 most frequently appearing elements of the bag of words. After eliminating articles and prepositions, we then manually reduced our set to 190 variables (i.e., word stems) that we thought represented potential sources of heterogeneity (e.g., "scratch") and appear in at least 5 of our listings.

Second, we used principal component analysis (PCA) to further reduce the dimensionality of the word frequency data. PCA analysis is a methodology for projecting a set of potentially correlated data vectors onto a set of orthogonal basis vectors. The first component is chosen by picking the linear combination of the regressors that has the highest variance in the data, the second components has the second highest variance, etc. In most cases, one can use the first few components to capture most of the variance in a set of data with much higher dimensionality. Our analysis used the first 10 principal components, which explains 80% of the variance in the 190 bag of words variables.[4] For a full list of the words, please see Appendix A.

It can be difficult to ascribe a straightforward meaning to many of the principal components, which is typical of this form of data reduction. However, by examining the loading factors of the first three components, which determine which variables have the most influence in defining the component, we can attribute meaning to these variables as described by Table 3. Intuitively, these components can be interpreted because their loading factors identify clusters of words that share a common theme or general meaning. The fact that the principal components with the most explanatory power appear to have reasonable interpretations gives some confidence that the PCA analysis reflects meaningful attributes of the listings. Together these components explain more than 60% of the variation in the bag of words data.

[4]We experimented with using up to the first 25 components, which explain 98% of the variation in the data. The difference in the results was negligible.

Component Name	% of Variance Explained	Words with High Loadings
Description of Item	43.6	"new", "read", "include", "content"
Shipping and Payment Information	11.9	"paypal", "return", "payment", "buyer"
Technical Specifications	7.3	"display", "connect", "gb", "charge"

Table 3: Interpretation of PCA components

4 Analysis of Price Variation

One ought to expect that we ought to be able to explain more price variation than the typical paper in the prior literature given the rich set of regressors in our data and the use of machine learning techniques designed to handle data sets like ours. The more interesting question is whether the additional explanatory power is due to the richer regressors, the machine learning techniques, or the combination of the two. To assess the importance of the richer data set, we compare the predictive power of a linear-in-variables model estimated on the full data set to the predictive power of the same kind of model estimated on the basic data set. To evaluate the importance of the machine learning algorithms, we compare the predictive power of a regression forest estimated on the full data set with the predictive power of a linear-in-variables model estimated on the same data set. To appraise the value of machine learning techniques on more traditional economic data sets, we also estimate our regression forest model on the basic data set. We use the R^2 statistic as our metric of the fraction of the price variation we have explained.

Given our large set of regressors, both ordinary least squares (OLS) and the regression forest techniques we employ are prone to overfitting. To address this problem we compute an out-of-sample version of the R^2 measure to get a more accurate sense of the predictive power of our models. Inspired by the machine learning literature, we compute our out-of-sample R^2 through 10-fold cross validation. The 10-fold cross validation procedure starts by randomly partitioning our data in 10 equally sized subsets that we denote $\{\mathcal{F}_1, ..., \mathcal{F}_{10}\}$. For each $k = 1, ..., 10$ we hold out $\mathcal{F}_k$ as a validation set and estimate our model on the remaining 9 subsets of data. We then compute the sum of squared error, SSE_k, and total sum of squares, TSS_k, in the validation set $\mathcal{F}_k$ using the estimated model. The *out-of-sample*

R^2 statistic is then[5]

$$R^2_{out} = 1 - \frac{\sum_{k=1}^{10} SSE_k}{\sum_{k=1}^{10} SSE_k}$$

The *in-sample* R^2 is computed by estimating the model on the full data set, forming a prediction for the price of each listing in the data set, and computing the R^2 based on these in-sample predictions. The in-sample R^2 is the statistic usually reported by economists.

The usual justification for treating the R^2 statistic as a measure of variance explained by the model is rooted in a decomposition of the variance in the data into a component capturing the variance in the model's predictions and a second component representing the variance in the predicted error. Since our primary results are based on out-of-sample tests, the orthogonality between the predictions and the predicted error required for this decomposition do not apply, even for the OLS models we consider. Nevertheless, we believe that the out-of-sample R^2 provides a useful summary of the predictive power of the data, so we focus on the R^2 statistic in our analysis.

4.1 Ordinary Least Squares

In order to assess the usefulness of standard econometric techniques for explaining price variation, we study the relationship between the price and the regressors we have collected using OLS. First we regress price against the regressors in the basic data set as a benchmark. This exercise is meant to be a proxy for the analysis done in papers like Baye, Morgan, and Scholten [8]. In line with the prior research, we are able to explain around 13% of the price variation using our basic data set. Next, we apply OLS to our full data set, which yields an R^2 of 0.14 Table 4.1 provides a summary of the results from our OLS investigation and includes both in- and out-of sample R^2. As asymptotic theory suggests, the in- and out-of-sample R^2 are essentially the same for the OLS model estimated on the basic data set since we can estimate the small number of regression coefficients very precisely. As expected, adding the information about the seller's textual description did provide an improved fit, but the improvement was very small and caused the OLS model to significantly overfit the sample.

The regression coefficients for the basic data set are listed in Table 5, and the coefficients

[5]We also computed the average out-of-sample R^2:

$$\overline{R^2_{out}} = 1 - \frac{1}{10}\sum_{k=1}^{10} \frac{SSE_k}{SSE_k}$$

The resulting values differed from R^2_{out} by less than 0.5%.

		R^2 Version	
		Out-of-Sample	**In-Sample**
Data Set	**Basic**	0.1286	0.1419
	Full	0.1456	0.3517

Table 4: OLS Predictive Power

Variable	**Coefficient**	**95% Confidence Interval**
Shipping Price	-1.159	[-1.576, -0.7410]
Shipping Calculated	-11.50	[-15.45,-7.553]
Returns Allowed	7.682	[4.034,11.33]
ln(Seller Score)	2.212	[1.285,3.139]
Relisted	11.48	[7.325, 15.62]
Constant	-7.389	[-13.09,3.139]

Table 5: OLS Coefficients

all had the expected sign. Somewhat surprisingly, the sellers seem to price as though purchase and shipping costs were interchangeable yielding a "shipping price" coefficient close to -1, although the confidence interval on that variables is fairly wide.

4.2 Regression Forest

Now we turn to our *regression forest* estimator, which was originally proposed by Breiman [10]. A regression forest is an ensemble estimator - in other words, it is the average of a large collection of underling regression tree models. Before describing how an ensemble of regression trees is constructed, let us describe the algorithm for creating a single regression tree. Consider a data set $\mathcal{D}$ with regressors $\boldsymbol{X}$ and regressand $\mathbf{Y}$. A regression tree partitions the space of possible realizations of $\boldsymbol{X}$ and assigns each element of the partition a value equal to the average of the regressands in that element of the partition. The prediction generated by a regression tree for a new realization of the regressors, x, is simply the value assigned to the element of the partition in which x falls. One can think of a regression tree as a form of nearest neighbor predictor where the "kernel" used is determined by the data.

The partition of the data set that defines a regression tree model can be represented graphically using a binary tree. An example of such a tree is displayed in Figure 4.2. Beginning at the root at the top of the diagram, each split represents a division of the sample into two (potentially unequally sized) subsets. Each leaf of the tree provides a prediction of the detrended price for listings within the data at that leaf. Figure 4.2 represents a tree with at least 100 listings in each leaf. The trees we use in our regression forest can have as

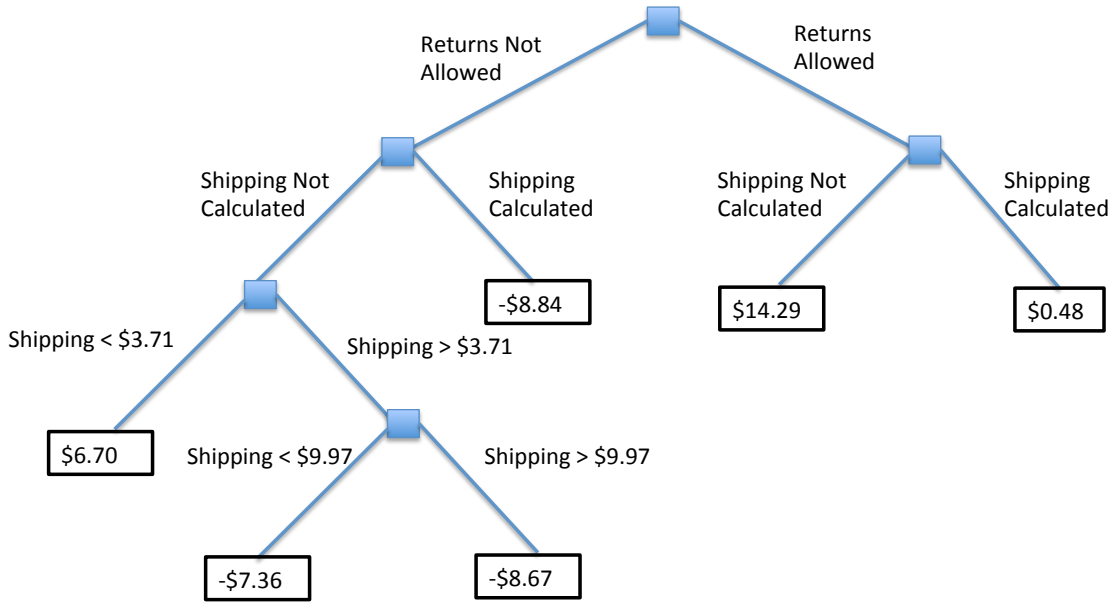

Figure 3: Simple Regression Tree

few as one datum at each leaf, so each tree in our forest is a much finer model (which can result in overfitting).

Now we formally describe the algorithm for growing a tree. Our regression forest is grown using *bootstrapped aggregation*, or bagging. To grow a single tree in the ensemble, a bootstrapped sample $\mathcal{B}$ is drawn from $\mathcal{D}$ that is of equal size to $\mathcal{D}$. A third of the explanatory variables are randomly chosen to be used as *splitting variables*, where the choice of a third is a commonly used heuristic for regression forest algorithms.

The root of the tree is a split of the bootstrap data set $\mathcal{B}$ into $\mathcal{B}_1 = (\boldsymbol{X}_1, \boldsymbol{Y}_1)$ and $\mathcal{B}_2 = (\boldsymbol{X}_2, \boldsymbol{Y}_2)$ such that $\mathcal{B}_1 \cap \mathcal{B}_2 = \varnothing$ and $\mathcal{B}_1 \cup \mathcal{B}_1 = \mathcal{B}$. These splits have the form $\mathcal{B}_1 = \{(x, y) \in \mathcal{B} : x_i \leq k\}$ where x_i is the realization of the i^{th} regressor, which must be one of the splitting variables, and k is a real number defining a *split point*. Since the predictions are the same within a leaf, it would make sense to divide the data set in such a way that the prices for the data within each leaf are as similar as possible. The *split criterion* we use implements this intuition by minimizing the following function at each node:

$$|\boldsymbol{Y}_1| \, Var(\boldsymbol{Y}_1) + |\boldsymbol{Y}_2| \, Var\,(\boldsymbol{Y}_2)$$

where $|\boldsymbol{Y}_i|$ refers to the cardinality of the set. For each regressand, the algorithm determines the value of k that minimizes the split criteria. The algorithm divides the data using the

		R^2 Version	
		Out-of-Sample	**In-Sample**
Data Set	**Basic**	0.1278	0.1969
	Full	0.4134	0.9126

Table 6: Regression Forest Predictive Power

splitting variable and the associated k that minimizes the splitting criterion.

The algorithm then recursively applies this splitting process on subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ until an entire tree is formed. Note that a splitting variable can appear multiple times in the same tree. The algorithm terminates when subsets contain a single data point. A prediction for a generic realization x' is the value at the leaf to which x' belongs. Obviously if the data point x' is in the training sample used to grow the tree, then the prediction will be perfect. This is the reason why it is so crucial to use out-of-sample tests to obtain realistic measures of the predictive accuracy of a regression forest

The intuition underlying the regression forest is that the average of many different models is likely to both provide a more accurate prediction and be less subject to overfitting. The use of a bootstrapped subsample and the choice of some, but not all, of the regressors as splitting variables in each tree is meant to create a diverse array of trees in the ensemble. The diversity of the trees, again, helps improve the fit of the ensemble and restrain the decision trees' tendency to overfit the data.

As in the case of the OLS analysis, we assessed the predictive power of the regression forest estimator when applied to both our basic and our full data set. The results are described in Table 4.2.[6]

When applied to our basic data set, the regression forest explains almost exactly the same amount of price variation as our OLS model. However, the regression forest explains three times more price variation than an OLS model when applied to the full data set. In short, explaining the price variation requires both our rich data set and the flexibility of the regression forest methods to explain the price variation.

4.3 LASSO

One might wonder whether a more flexible linear-in-variables model estimated on the full data set might perform as well as the regression forest. To test this conjecture, we estimated

[6]We could have used an "out of bag" estimate of R^2 in which an estimates is provided for a point (x, y) using only the trees in our forest that did not have (x, y) in their bag. The results are essentially the same as the out-of-sample computations provided here.

a linear model with interactions of up to third order of the variables in our full data set. Once we remove redundant regressors, we are left with 6,463 variables. We apply the LASSO algorithm to this data set to eliminate regressors, which is necessary for identification and reduces the potential for overfitting.

Denoting a single data point as (x_i, y_i) where x_i are the regressors we collected and y_i is the detrended price of the listing, we can describe the LASSO algorithm through the following optimization problem:

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - x_i'\beta\right)^2 + \lambda \|\beta\| \tag{1}$$

where λ is the LASSO penalty parameter and x_i is a vector of regressors. Since β enters the penalty parameter linearly, the solution to problem 1 retains only the regressors with a significant amount of predictive power. The amount of predictive power required for a variable to be retained in the solution is governed by λ. We refer to an OLS model incorporating the variables that have nonzero coefficients at the solution to to problem 1 as the model selected by the LASSO procedure.

LASSO algorithms typically provide results for a range of values of λ, and each possible λ is evaluated through a cross-validation process that is intended to penalize overfitting. Since the cross-validation procedure is slightly different than our algorithm for computing out-of-sample R^2 values, we provide an overview for completeness. The degree of overfitting is assessed through 10-fold cross validation of each value of λ. We first divide the data set into 10 equally sized subsets $\{\mathcal{F}_1, ..., \mathcal{F}_{10}\}$. For each $k = 1, ..., 10$ we hold out $\mathcal{F}_k$ as a validation set and estimate our model on the remaining 9 subsets of data. We then compute the sum of squared error, $SSE_k(\lambda)$, in the validation set $\mathcal{F}_k$. The sum of of squared error over the whole sample is

$$SSE(\lambda) = \sum_{k=1}^{10} SSE_k(\lambda)$$

Finally we compute the standard deviation of the cross-validation procedure, $SE(\lambda)$, which is equal to the standard error among $SSE_1(\lambda), ..., SSE_{10}(\lambda)$.

There are two standard rules of thumb for choosing λ. One obvious heuristic is to choose the value of λ that minimizes $SSE(\circ)$, and we denote this value as λ_{Min}. Since this choice often results in overfitting, the other common rule of thumb is to choose the largest value of λ, which we denote λ_{1SE}, such that $SSE(\lambda) \leq SSE\left(\underline{\lambda}\right) + SE\left(\underline{\lambda}\right)$. Table 4.3 documents the predictive power of the OLS models selected by the LASSO procedure. The

		R^2 Version	
		Out-of-Sample	**In-Sample**
Heuristic	$\boldsymbol{\lambda}_{1SE}$	0.2289	0.2539
	$\boldsymbol{\lambda}_{Min}$	0.2870	0.3775

Table 7: LASSO Predictive Power

primary takeaway is that using a more flexible model to predict prices using the basic data set does improve the explanatory power of the model, but there is still a large gap between the explanatory power of our regression forest predictor and a flexible OLS model.

4.4 Other Analysis Methods

We tried a variety of other machine learning methods to try and explore how much price variation we could explain, but we found that the performance was comparable to or worse than our simpler regression forest model. We experimented with individual neural networks, but we found that at best the neural networks performed at a level comparable to the regression forest in terms of out-of-sample R^2. We found that even moderately complex neural networks (e.g., those with two hidden layers) severely overfit the data, which made the out-of-sample R^2 quite poor. Moreover, the neural networks were also highly sensitive to the structure chosen.

We then tried using a bagged neural network, which consists of an ensemble of simple neural networks that are each trained using a bootstrapped sample of the data, and the final prediction of the model is the average of the predictions of the ensemble. One can think of the bagged neural network model wherein the regression trees are replaced by simple neural networks. Finally, we tried estimating a boosted gradient tree model, which uses a sequence of regression trees to fit the data. The first tree attempts to fit the raw data and each successive tree in the sequence tries to fit the residuals from the previous tree. Again, none of these methods had more predictive power than the regression forest.

5 Heterogeneous Marginal Effects

We would now like to investigate the nature of the heterogeneous effects the regression forest is modeling. In order to do this, we would like to use a machine learning algorithm that estimates the marginal effect of the regressors on the price rather than making a price prediction. Many machine learning algorithms, while they make excellent predictors, are not

capable of consistently estimating causal effects. However, recent papers at the intersection of machine learning and econometrics provide us some tools that we can employ.

Inspired by Athey, Tibshirani, and Wager [1], we estimated a forest of model trees.[7] A model tree is very similar to a regression tree except that the predictor at each leaf takes the form of a regression model, in our case an OLS model.[8] Before discussing the details of the model we estimate at each leaf, we need to describe the algorithm we use to build our trees.

In order to insure that our causal effects estimates are consistent, each tree in our forest is an *honest tree.* In a regression-tree algorithm, the full data set is used to both determine the tree's structure (model selection) and to estimate the values at each leaf (model estimation). When the same data is used for both model selection and estimation, this can lead to biases and the failure of properties such as consistency. Luckily, creating an honest tree is relatively easy - one needs to simply divide the data set into two subsets and use one subset for determining the structure of the trees (i.e., the split points) and the second subset to assign a value to the leaves of each tree.[9]

Formally we divide our data set $\mathcal{D}$ into two components, a ($\mathcal{S}$)election set and an ($\mathcal{E}$)stimation set. We apply the algorithm described above in Section 4.2 to data set $\mathcal{S}$ to determine the structure of the forest of trees. In particular, each tree is grown from a Bootstrap sample consisting of XX data drawn from $\mathcal{S}$, and the splitting points of each tree are determined as per the variance minimization criterion. However, there are two major differences that we explain in a moment. First, we do not allow the algorithm to create a split point on a variable if one of the resulting leaves has fewer than 50 data points. This lower bound insures that we will have enough data at each leaf to estimate an OLS model. Second, we do not allow splits that are based on the variables we include in the OLS model, which we justify below. In total our forest contains 1000 trees.

Once the split points of each tree have been computed, we move onto the estimation step using data set $\mathcal{E}$. Each tree within the forest is estimated in a three step process. First, we generate a bootstrap sample from $\mathcal{E}$. Second, we determine which leaf each of the bootstrap samples falls into. Second, we select one of the nonterminal nodes of the greatest depth in the tree. The third and final step of estimating a tree is to perform an OLS regression on the data at each leaf.

Our separation of the data used for estimation and splitting insures that our trees are

[7] We omit a detailed discussion of the formal requirements for our model tree forest to be consistent. The interested reader is encouraged to refer to Athey, Tibshirani, and Wager [1].

[8] One can think of the regression forest algorithm of Breiman [10] as a model forest when the estimated "model" is the coefficient on a constant variable.

[9] We make no claim that our honest trees are in any sense optimal.

honest. However, we also require that the OLS estimator be asymptotically consistent as the size of the tree grows. If we were to collect more data, we could simultaneously define more leaves (increasing our ability to detect heterogeneous causal effects) and increase the size of the leaves (increasing the precision of our OLS estimates). However, there is a tension between the size of the leaves and the amount of heterogeneity we can detect. As we add more leaves to the tree, it is likely that the regressor realizations within a leaf will become more similar, and the decrease in regressor variability would make our OLS estimates less precise. In the extreme, if we grow the leaves too quickly as our data set grows, our OLS models may not even be consistent.[10]

To solve this problem, we do not allow the trees to split on the variable we include in our OLS regressions. Our hope is that the resulting tree will have leaves that exhibit enough variation in the regressors to provide precise estimates of the marginal effects. Asymptotically, our trees will have an infinite set of data points at each leaf. The set of data at each leaf will exhibit a great deal of variation in our OLS regressors, which insures consistency of our estimates of the marginal effects.

The variables we included in our OLS regressions are "log seller score," "number of images," "number of words in the title," and "number of words in the body." Each forest provides up to 1000 estimates of the marginal effect of the regressors at each data point in $\mathcal{E}$.[11] For each data in $\mathcal{E}$ we compute the average marginal effect across the trees that included that data point in the bootstrap sample of $\mathcal{E}$. We then compute a CDF of the average marginal effect across the data points in $\mathcal{E}$. In order to determine confidence bounds, we repeated the estimation step 500 times, in effect producing bootstrap estimates of the confidence intervals for the quantiles of our CDF. Figures XX - YY provide the median value at each quantile rank from 0 to 1, and the whiskers represent the 10-90 confidence interval for the values at the respective quantile ranks.

6 Conclusion

The role of unobserved heterogeneity in a set of apparently homogenous goods is an obvious potential explanation for price variation. Simply put, two seemingly identical goods have different prices because they are not truly identical. The eBay setting provides a nearly ideal

[10] One can describe the rate at which the leaves can shrink as the number of data points in each leaf grows. Given the size of our data set, we would need large leaves to get enough variance in our regressors to estimate the OLS parameter, which motivates us to take this alternative approach.

[11] On average, just over 60% of the elements of $\mathcal{E}$ are used to estimate each tree.

environment for assessing the potential role of heterogeneity since the online platform allows us to observe the same information about the good as the would-be buyers. In principal, we can therefore detect whatever features of the object sellers expose to buyers to justify an unusually high (or low) price.

We are of course not the first to analyze online price variation. In order to replicate the earlier work in our online context, we started by trying to explain price variation using a basic data set that contains variables that we believe are analogous to what the previous studies used. We find that we can explain roughly 13% of the price variation using OLS techniques, which is inline with the prior work. This suggests that there is nothing intrinsic to eBay that makes the price variation easier or harder to explain. When we apply machine learning techniques to our basic data set, we can explain essentially the same amount of price variation as our OLS methods.

We can explain over 40% of the price variation when we use our full data set, but only when studied using regression forests. When analyzed using OLS techniques, we can only explain an additional 1% of the price variation. The takeaways from this are two fold. First, the posted price listings do convey a great deal of information regarding the heterogeneity of the listings. Second, one needs to use flexible models to adequately detect how variation in the observables drives variation in the prices. The explanatory power of either the richer data set or the more flexible estimation techniques has little power in isolation.

SAY SOMETHING ABOUT THE MARGINAL EFFECTS

We would like to close by discussing two caveats to our results. First, since there is data that we did not include in our prediction model, we probably underestimated the importance of heterogeneity for explaining price variation between seemingly homogenous goods. For example, we include the number of pictures as an explanatory variable, but we cannot use the content or quality of the images to predict price since we did not download individual images. Presumably we could explain a larger fraction of the price variation if we had more detailed information on either the photos or the text included in the listing.

Second, the eBay posted price market does not appear to share the features usually assumed by models that explain price variation as an equilibrium of market competition game played by sellers of a homogenous good. Viewed in this light, it is less surprising that a large fraction of the price variation is due to underlying product or listing heterogeneity. However, even in markets that do have features that support models of price variation of homogenous products (e.g., significant costs to obtain a price quote), our estimates suggest that one ought to expect significant price variation as a result of product heterogeneity regardless of the model of competition assumed.

There are two obvious directions to push this research agenda. First, one could gather more information from each listings and use yet more flexible analysis techniques to explain an even larger fraction of the price variation. Second, and more interestingly, one could extend this analysis to other products in other markets to try and understand the drivers of (previously) unobserved price variation. Our results suggests that future work should not be surprised to find significant heterogeneity in the market for mundane goods.

References

[1] Athey, S.; J. Tibshirani; and S. Wager (2017), "Solving Heterogeneous Estimating Equations with Gradient Forests," *mimeo*.

[2] Augenblick, N.; M. Niederle; and C. Sprenger (2015), "Working over Time: Dynamic Inconsistency in Real Effort Tasks," *The Quarterly Journal of Economics*, 130 (3) pp. 1067 - 1115.

[3] Bailey, J. 1998. "Electronic Commerce: Prices and Consumer Issues for Three Products: Books, Compact Discs, and Software," *Organization Economics Co-Operation Development*, 98 (4).

[4] Baye, M. and J. Morgan (2001) "Information Gatekeepers on the Internet and the Competitiveness of Homogenous Product Markets," *The American Economic Review*, 91 (3), pp. 454 - 474.

[5] Baye, M.; J. Morgan; and P. Scholten (2004) "Price Dispersion in the Small and the Large: Evidence From an Internet Price Comparison Site," *The Journal of Industrial Economics*, 52 (4), pp. 463 - 496.

[6] Baye, M.; J. Morgan; and P. Scholten (2004) "Temporal Price Dispersion: Evidence From an Online Consumer Electronics Market," *Journal of Interactive Marketing*, 18 (2), pp. 101 - 115.

[7] Baye, M.; J. Morgan; and P. Scholten (2006) "Information, Search, and Price Dispersion" in *Handbooks in Information Systems*, Elsevier.

[8] Baye, M.; J. Morgan; and P. Scholten (2006) "Persistent price dispersion in online markets" in *The New Economy And Beyond: Past, Present And Future*, Edward Elgar Publishing.

[9] Baylis, K. and J. Perloff (2002) "Price Dispersion on the Internet: Good Firms and Bad Firms," *Review of Industrial Organization*, 21, pp. 305 - 324.

[10] Breiman, Leo (2001) "Random Forests," *Machine Learning*, 45, pp. 5 - 32.

[11] Brynjolfsson, E. and M. Smith (2000) "Frictionless Commerce? A Comparison of Internet and Conventional Retailers," *Management Science*, 46 (4), pp. 563 - 585.

[12] Clay, K.; R. Krishnan; and E. Wolfe (2001) "Prices and Price Dispersion on the Web: Evidence from the Online Book Industry," *The Journal of Industrial Economics,* 49 (4), pp. 521 - 539.

[13] Clay, K.; R. Krishnan; E. Wolfe; and D. Fernandes (2002) "Retail Strategies on the Web: Price and Nonprice Competition in the Online Book Industry," *The Journal of Industrial Economics,* 50 (3), pp. 351 - 367.

[14] Diamond, P. (1971) "A Model of Price Adjustment," *Journal of Economic Theory*, 3, pp. 156-168.

[15] Einav, L.; C. Farronato; J. Levin; and N. Sundaresan (2013) "Sales Mechanisms in Online Markets: What Happened to Internet Auctions?" *mimeo.*

[16] Einav, L.; C. Farronato; J. Levin; and N. Sundaresan (2016) "Auctions versus Posted Prices in Online Markets',' *Journal of Political Economy, forthcoming.*

[17] Einav, L; T. Kuchler; J. Levin.; and N. Sundaresan (2015) "Assessing Sale Strategies in Online Markets Using Matched Listings," *American Economic Journal: Micoeconomics*, 7 (2) pp. 215 - 247.

[18] Elfenbein, D.; R. Fisman; and B. McManus (2015) "Market Structure, Reputation, and the Value of Quality Certification," *American Economic Journal: Micoeconomics*, 7 (4) pp. 83 - 108.

[19] Hong, H. and M. Shum (2006) "Using Price Distributions to Estimate Search Costs," *The RAND Journal of Economics*, 37 (2), pp. 257 - 275.

[20] Hui, X.; M. Saeedi; Z. Shen; and N. Sundaresan (2015) "Reputation & Regulations: Evidence from eBay," *mimeo.*

[21] MacMinn,, R. (1980) "Search and Market Equilibrium," *Journal of Political Economy*, 88 (2), pp. 308 - 327.

[22] Nosko, C. and S. Tadelis,, R. (2015) "The Limits of Reputation in Platform Markets: An empirical Analysis and Fied Experiment," *mimeo.*

[23] Pan, X.; B. Ratchford; and V. Shankar (2002) "Can Price Dispersion in Online Markets Be Explained by Differences in E-Tailer Service Quality?" *Journal of the Academy of Marketing Science*, 30 (4), pp. 433 - 445.

[24] Porter, M. (1980) "An algorithm for suffix stripping," *Program*, 14 (3), pp. 130 - 137.

[25] Reinganum, J. (1979) "A Simple Model of Equilibrium Price Dispersion," *Journal of Political Economy*, 87 (4), pp. 851 - 858.

[26] Saeedi, M. and N. Sundaresan (2016) "The Value of Feedback: An Analysis of the Reputation System," *mimeo.*

[27] Salop, S. and J. Stiglitz (1977) "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion," *The Review of Economic Studies*, 44 (3), pp. 493 - 410.

[28] Sorenson, A. (2000) "Equilibrium Price Dispersion in Retail Markets for Prescription Drugs," *Journal of Political Economy*, 108 (4), pp. 833 - 850.

[29] Stigler, G. (1961) "The Economics of Information," *Journal of Political Economy*, 69 (3), pp. 213 - 225.

[30] Varian, H. (1980) "A Model of Sales," *The American Economic Review*, 70 (3), pp. 651 - 659.

[31] Wager, Stefan and Susan Athey (2015) "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *mimeo.*

Stem	Count	Stem	Count	Stem	Count
gb	1201	hour	199	good	146
new	1139	case	197	faster	145
black	978	game	197	b	141
box	662	feature	194	mail	140
brand	635	store	193	paypal	139
model	521	thank	187	return	137
latest	464	app	184	popular	136
ship	453	work	183	audio	133
free	389	power	180	inform	131
seal	382	original	179	system	131
all	373	processor	177	view	130
open	367	device	172	factory	124
include	361	special	168	service	124
screen	336	charger	167	experience	122
display	333	condition	165	look	122
have	324	support	164	facebook	121
no	303	read	161	additional	119
not	294	purchase	160	perfect	116
usb	265	receive	160	available	112
question	251	connect	158	warranty	111
offer	250	email	155	bid	110
only	243	internal	155	provide	110
more	231	charge	152	perform	109
package	220	million	151	sell	109
touch	212	payment	150	ad	108
fast	209	contact	149	full	106
battery	205	access	147	great	104
content	205	set	147	enjoy	99
cable	201	technology	147	detail	98

A Bag of Words

The full set of word stems we parsed from the text of each listing is described in the following table along with the number of listings in which the word appears. For clarity, the table includes only one representative example of each word stem. Unsurprisingly, the most popular terms are descriptions of the Amazon Kindle that apply to all of the units available at that time. For example, the most popular tems, "gb", is used to refer to the 8 GB of storage. Once one reaches the rare terms such as "shipment," we are presumably considering listings by sellers with more experience that may in fact have standardized text regarding shipment terms, etc., that apply regardless of the good for sale.

Stem	Count	Stem	Count	Stem	Count
require	98	like	69	close	42
accept	97	pay	68	fluid	42
test	97	rate	68	response	42
price	95	left	67	treatment	41
accessory	94	cover	66	issue	40
best	94	number	61	complete	39
state	94	delivery	60	immediate	39
custom	93	description	60	guarante	38
buyer	92	must	60	insure	38
fully	92	except	58	paid	37
need	91	love	58	leather	35
feedback	89	policy	58	part	35
note	89	locate	56	combine	34
exchange	87	date	55	receipt	34
sale	87	tax	55	fedex	33
start	85	allow	54	friend	32
check	84	position	54	info	30
actual	82	photo	53	shipment	30
manufacturer	82	refund	53	describe	29
approximate	81	top	52	change	28
favorite	81	identify	51	damage	27
home	80	quickly	50	separate	25
design	78	beautifully	49	restock	24
well	77	cost	48	process	23
type	76	fee	48	clean	22
busy	75	off	48	scratch	22
back	73	quality	44	concern	17
first	70	sold	44	fair	17
seller	70	help	43	credit	15

Stem	Count	Stem	Count	Stem	Count
three	14	discount	10	invoice	8
win	13	law	9	appear	5
carefully	12	reserve	9		
inspect	12	wear	9		

B Heterogeneous Sellers and the Law of One Price

For brevity, we consider a model in which we assume the market is in steady-state with the distribution of buyers, sellers, and the respective prices stationary over time. We have in mind a market where a random number of buyers with single-unit demand enter the market each period and, since the goods are homogenous, buy from the lowest priced sellers. The demand curve, $D(p, \varepsilon)$, is exogenous, p is a price, and ε is an aggregate demand shock. We assume that there exists $\overline{p}$ such that $D(\overline{p}, \varepsilon) = 0$ for all ε.

The supply side of the market is generated by sellers who choose an offer price given the stationary equilibrium of the economy and their own reservation/production cost c. Each period sellers can choose a new price at which to offer their good, and (unlike buyers) if a seller fails to transact, then he or she automatically participates in the market the following period. Importantly, the sellers must choose their prices before observing the demand that period. Since the market is stationary, we can write the strategy of seller i as $p^*(c_i)$. If there are N_t sellers in period t, then the supply curve is

$$S(p) = \sum_{i=1}^{N_t} 1\{p^*(c_i) \leq p\}$$

Finally, the market clearing price is defined as

$$p_M(\varepsilon) = \min \{p : S(p) \geq D(p, \varepsilon)\}$$

which means the probability of a sale at an offer price of p is

$$\Pi(p) = \Pr\{p_M(\varepsilon) \geq p\}$$

Note that the sellers that successfully transact receiver their offer price, *not* the market clearing price. The market clearing price is merely an indicator of who transacts, not the particular price at which the transaction occurs.

The seller's problem is then

$$V(c) = \max_{p \geq 0} E_\varepsilon \left[\Pi(p, \varepsilon)\right](p - c) + \delta(1 - E_\varepsilon \left[\Pi(p, \varepsilon)\right])V(c) \tag{2}$$

where:

- Expected probability of sale at price p is $E_\varepsilon \left[\Pi(p, \varepsilon)\right]$
- Profit conditional on sale is $p - c$
- $\delta V(c)$ is the discounted continuation value

Consider an equilibrium of our game with discount factor δ, and let $\mathcal{P}(\delta, \rho)$ denote the set of prices p such that $p \leq p_M(\varepsilon)$ with positive probability. Let the minimum and maximum elements be

$$\begin{aligned} \underline{\mathcal{P}}(\delta, \rho) &= \inf \mathcal{P}(\delta, \rho) \\ \overline{\mathcal{P}}(\delta, \rho) &= \sup \mathcal{P}(\delta, \rho) \end{aligned}$$

Now we present our primary result on the law of one price. Our theorem says that (in effect) the distribution of prices accepted in equilibrium has to concentrate around a single price as $\delta \to 1$. The core intuition is that for two prices to frequently (i.e., with probability $\rho > 0$) be accepted in equilibrium, some seller must prefer to offer the low price to avoid delaying transactions. When agents get sufficiently patient, even small revenue improvements are worth waiting for, which leads to the concentration of the price distribution.[12]

Proposition 1 *For any $\rho > 0$, we have* $\operatorname*{Lim\,sup}_{\delta \to 1} \overline{\mathcal{P}}(\delta, \rho) - \underline{\mathcal{P}}(\delta, \rho) = 0$.

Proof. Suppose our result fails to hold for some $\rho > 0$. This means we can choose some $\gamma > 0$ such that for any $\delta' < 1$ there exists $\delta \in (\delta', 1)$ that satisfies $\overline{\mathcal{P}}(\delta, \rho) - \underline{\mathcal{P}}(\delta, \rho) > \gamma$. Choose $p, \widetilde{p} \in \mathcal{P}(\delta, \rho)$ such that $p - \widetilde{p} > \gamma$.

The payoff for an agent of type c that chooses a price p in every period can be written as:

[12]Note that formally each δ in the sequence is associated with a different equilibrium, so our proposition is properly interpreted as a statement about the properties of any sequence of equilibria that correspond to the sequence of δ. Without a sharper characterization of the equilibria, it is difficult to say whether the limit (as opposed to the limsup) even exists.

$$\sum_{t=0}^{\infty} \delta^t (1 - E_\varepsilon \left[\Pi(p,\varepsilon)\right]^t E_\varepsilon \left[\Pi(p,\varepsilon)\right] (p-c) = \frac{p-c}{1-\delta\left(1 - E_\varepsilon \left[\Pi(p,\varepsilon)\right]\right)}$$

As $\delta \to \infty$, the right hand side of the equality approaches $p - c$. Therefore, for $p - \widetilde{p} > \gamma$ to hold, there must exist c and $\widetilde{c}$ such that:

$$\frac{p-c}{1-\delta\left(1 - E_\varepsilon \left[\Pi(p,\varepsilon)\right]\right)} \geq \frac{\widetilde{p}-c}{1-\delta\left(1 - E_\varepsilon \left[\Pi(\widetilde{p},\varepsilon)\right]\right)} \tag{3}$$

$$\frac{\widetilde{p}-\widetilde{c}}{1-\delta\left(1 - E_\varepsilon \left[\Pi(\widetilde{p},\varepsilon)\right]\right)} \geq \frac{p-\widetilde{c}}{1-\delta\left(1 - E_\varepsilon \left[\Pi(p,\varepsilon)\right]\right)} \tag{4}$$

However, as $\delta \to \infty$, Equation 3 requires that $p - c \geq \widetilde{p} - c$ and Equation 4 requires that $\widetilde{p} - \widetilde{c} \geq p - \widetilde{c}$, which can only be the case if $p = \widetilde{p}$. As this contradicts our assumption that $p - \widetilde{p} > \gamma > 0$,we conclude that our theorem is correct. ■

Recalling that δ is discounted on a daily level, Proposition 1 implies that if agents have discount factors anywhere close to the conventional level and the market is roughly stationary, then the law of one price should hold even if the agents have very different production costs. Interpreted as a statement about elasticities, Proposition 1 implies that the elasticity should be 0 outside of a very narrow interval around the market clearing price.

It is worth taking a moment and considering what might interfere with our result. First, it could be that the agents do not have daily discount factors near the conventionally accepted level. For example, it could be that sellers have deadlines by which they need to sell an item. We view this as unlikely since it would imply a very serious credit constraint on the part of many sellers - after all Kindles, do not have a particularly high resale value. One might imagine that hyperbolic discounting might explain an unusually low discount factor, but recent evidence suggests that hyperbolic discounting is not significant when applied to monetary transfers unless the agent faces a binding credit constraint (Augenblick, Niederle, and Sprenger [2]).

One might also suppose Proposition 1 fails because the market is nonstationary. Some nonstationarity is to be expected - after all, the value of a new Kindle depreciates as newer models come closer to introduction as reflected in the downward time trend over the nine months of our sample (Figure 3). We could easily include in our model a time trend in the demand and the probability of sales. Roughly speaking, the slow drop in demand has an effect similar to that of the time discount factor. This means that unless the time trend is very steep, which it is not, one would expect a mild effect on the outcome.