MARKET FRICTIONS VERSUS SELLER SIGNALING: EXPLAINING DEVIATIONS FROM THE LAW OF ONE PRICE

AARON BODOH-CREED*, JÖRN BOEHNKE†, AND BRENT HICKMAN‡

Abstract. Significant price dispersion for homogeneous goods in e-commerce is a well-documented puzzle; we find robust empirical evidence that the puzzle is roughly half as large as previously thought. We use an approach motivated by economic theory, employing a rich new dataset encompassing a wealth of unstructured data from 14 product categories, and flexible machine-learning methods. We quantify sellers' marketing strategies through customization of the content, layout, and appearance of their product-listing web pages. Price variation systematically associated with these observable listing characteristics captures perceived listing heterogeneity rather than pricing noise generated by market frictions, allowing us to bound the role of market frictions in observed price dispersion. We find that seller-customized listing characteristics account for roughly half of observed cross-sectional price dispersion. These characteristics also predict buyer responses, with more information-rich listings associated with higher sale probabilities and shorter listing durations. Thus, a substantial portion of apparent deviations from the Law of One Price reflects economically meaningful differences in how sellers present otherwise physically identical products.

*Amazon.com, Inc., bodohcre@amazon.com
†Graduate School of Management, University of California, Davis, jb@ucdavis.edu
‡David Eccles Business School, University of Utah, brent.hickman@utah.edu

Keywords: Online Markets, Product Differentiation, Price Dispersion, Machine Learning, Law of One Price
JEL Subject Classification: D4, D43, L86.

1. Introduction

The "Law of One Price" (LOP) predicts that all exchanges of homogeneous goods in a thick, frictionless market ought to take place at a single price. While the LOP holds remarkably well in some instances (e.g., security exchanges), it fails to describe the reality of many retail markets. This failure is a puzzle: the internet has vastly simplified consumers' price comparisons across sellers of similar products, which one would expect to eliminate frictions, encourage fierce competition, and drive prices to a common level. Instead, online price dispersion is ubiquitous, even for products that appear homogeneous, such as new books (e.g., Bailey 1998; Brynjolfsson and Smith 2000).

If product listings are truly homogeneous, then economic theory offers two insights. First, in a simple model (Section 3.1), we show that in absence of market frictions, deviations from the LOP cannot form part of a pricing equilibrium. Second, market frictions may therefore play a substantial role in e-commerce markets by giving rise to various forms of endogenous price dispersion. However, this raises the question of whether homogeneity is a reasonable view of the market as perceived by consumers: due to seller marketing strategies, units of a given product could *appear* heterogeneous in ways difficult for researchers to detect. For example, sellers may have subtle ways of signaling value, reliability, or professionalism.

Our goal is to measure the extent to which online price dispersion reflects retailers strategically differentiating their products through marketing strategies, rather than the more commonly assumed market frictions. Our approach combines economic theory and descriptive analyses to produce a novel, tighter bound on the scope for market frictions to drive price dispersion. This distinction is important for several reasons. First, peer-to-peer platform markets are ubiquitous, including eBay and Amazon (retail), Upwork (freelance labor), Match.com (dating), and StubHub (event tickets), and recent work documents striking regularities in transaction patterns across online marketplaces (Bracci et al., 2022). Second, frictions cause waste in terms of user time and search effort; if this deters buyers from platform participation, it may also limit seller participation, since buyer-seller ratio is a known driver of expected profits (Bodoh-Creed et al. 2021). Third, frictions are known to support seller market rents (e.g., Sorensen 2000), and the division of surplus between buyers and sellers is an important consideration for platform design (Bhargava et al. 2020).

We analyze a rich dataset of HTML source code for posted-price product listings of 14 products on eBay. Our raw sample is an example of *unstructured data*, recently emphasized as an untapped source of inferential power in marketing (e.g., Balducci and Marinova 2018; Boehnke et al. 2025; Dew 2025). It allows us to observe essentially everything the buyer sees, including subtle distinctions between marketing strategies of different sellers. While each product listing contains some standardized content, eBay allows sellers wide latitude to customize prominent elements, including title, pictures, formatting, and text. The eBay market is also a good test case for studying LOP deviations, being thick with thousands of buyers and sellers. Moreover, many sources of product

heterogeneity can be ruled out: eBay has a transparent seller rating system and standard warranty policy, and our sampling procedure excludes accessory bundling. These features suggest consumers ought to view sampled listings as near-perfect substitutes, and that price competition should be fierce. Yet, the cross-sectional coefficient of variation (ratio of standard deviation/mean) of price offers is economically significant, ranging between 0.15 and 0.63 across our 14 products.

Though physically homogeneous, if product listings are subtly different, a portion of observed online price dispersion may not be attributable to market frictions. Using our bank of HTML pages scraped from eBay, we extract a high-dimensional set of listing characteristics. The style and content of the title section and the customizable description section are largely at each seller's discretion, and they vary widely. Seller-provided text and images may convey information about product condition or packaging, while listing presentation may itself signal seller professionalism and reliability. Consistent with this interpretation, more experienced sellers systematically employ richer and more stylized listings. We control for information content of the text using a bag-of-words (BoW) approach, part of 220 regressors characterizing each product listing. We then employ various machine learning methods to explain sources of price dispersion due to sellers' efforts to differentiate goods through marketing cues.

Our empirical analysis is rooted in economic theories that detail conditions under which price dispersion can plausibly arise in homogeneous goods markets. Existing models of search frictions generate price variation that is plausibly orthogonal to observable listing characteristics, as in mixed-strategy equilibria (see Section 3). Therefore, price predictability derived from our rich HTML data places an implicit bound on the role that search frictions could have played. Although our empirical model is predictive in nature, economic theory indicates that for our purposes we do not need to identify a causal demand system in order to parse between listing heterogeneity and market frictions as sources of price dispersion. One threat to our empirical strategy is that sellers might be able to design listings to manipulate eBay's search algorithm to affect the ordering of search results, thereby conflating the roles of listing features and search frictions. We test for this possibility and find no evidence consistent with search-algorithm manipulation (Section 4.1.1).

The literature has found it difficult to explain online cross-sectional price variation without seller fixed effects (e.g., Clay et al. 2002; Baye et al. 2006). A first problem is one of *omitted variables*, or failing to control for all information relevant to buyers (e.g., text and images). A second problem is one of *model specification*: cues consumers extract from this information may manifest in complex ways that are lost by imposing restrictive linear regression functional forms. We propose two solutions (Section 3.4): (*i*) enrich the data, and (*ii*) use more flexible models/methods (e.g., regression forest) to capture subtle interactions among listing traits. We separately evaluate the contribution of richer data and more flexible models by comparing their explanatory power across specifications.

Commensurate with the previous literature, we find limited explanatory power with linear regression estimated via ordinary least squares (OLS) and basic observables—between 0% and 15% of price variation, depending on the product. By contrast, a random forest estimated on our full dataset increases explanatory power to 48% on average across our 14 product categories, accounting for 26%-67% of price dispersion, or roughly 2-3 times as much predictive power as OLS applied to the basic dataset. Thus, a substantial share of what appears to be a violation of the Law of One Price may instead reflect economically meaningful differentiation among listings of physically identical products. Our results show that the combination of richer data *and* more flexible models is required to adequately explain price variation (and bound the role of market frictions), as neither suffices on its own. Alternatively, our empirical strategy can be viewed as a new approach for empiricists to supplement measures of product attributes in a flexible way (see also Quan and Williams 2021 and Blanchard et al. 2023).

We also analyze listing outcomes (Section 5) to assess whether seller-customized listing features are associated with demand responses consistent with their role in perceived product differentiation. We find two additional results. First, seller-customized product listing traits have substantial residual explanatory power for sale probability and listing duration (conditional on sale), after controlling for seller score, price, and basic listing covariates. Second, predicted marginal effects are nontrivial and conform to basic economic theory: holding style variables fixed, more information-rich listings are associated with higher sale probability and lower listing duration.

At the same time, significant unexplained price variation persists, suggesting that apparent product differentiation explains an important share—but not all—of deviations from the Law of One Price. Search frictions therefore still appear to play a meaningful role, though one roughly half as large as previously thought. A possible explanation is that, given buyer constraints on cognition (Monroe and Lee 1999) and market knowledge (Busse et al. 2017), a keyword search on the phrase *Amazon Kindle Fire HD* may inadvertently support price dispersion by overwhelming the user. While search engines solve the "needle-in-a-haystack" problem for consumers by sifting out *irrelevant* information, platform markets may create a new "needle-in-a-stack-of-needles" problem, providing consumers with prohibitively large quantities of *relevant* information.

1.1. **Related Literature.** Economic theory has long established that market frictions enable equilibrium price dispersion resulting from consumers' imperfect information and/or firms' strategic competition (e.g., Reinganum 1979; MacMinn 1980; Salop and Stiglitz 1977; Wilde and Schwartz 1979; Rosenthal 1980; Varian 1980; Baye and Morgan 2001; Burdett and Judd 1983). Search frictions on eBay have been documented by Backus et al. (2014), Dinerstein et al. (2018), Schneider (2016), and Bodoh-Creed et al. (2021). Other related work, including Lewis (2011), De Los Santos et al. (2017), and Lin and Wildenbeest (2020), have focused on characterizing search frictions in online markets for differentiated products such as cars, MP3 players, and health insurance. Our

empirical exercise focuses on a cleaner environment with homogeneous goods, where we nonetheless find high price dispersion. We also adopt a more modest goal of bounding the role that search frictions may plausibly play in driving price dispersion, by quantifying sellers' marketing strategies to create apparent differentiation. Although our analysis is descriptive, economic theory disciplines how we interpret the statistical evidence (see Sections 3.4 and 4.2), and key aspects of our estimates conform well to economic theory on market behavior (e.g., Sections 4.3 and 5). Zhao and Hastie (2021) discuss three conditions under which causal interpretations may be assigned to machine-learning estimates, which may be arguably satisfied in our setting: (*i*) a model with good predictive performance, (*ii*) *a priori* domain knowledge, and (*iii*) suitable visualization tools.

Several prior studies predict online prices with varying comparability to our results. Clay et al. (2001) explain price dispersion with time indicators, while our study focuses on *cross-sectional price dispersion* within de-trended data, which is still large. Clay et al. (2002) and Baye et al. (2006) find limited price predictability with standard models and basic controls, but substantially greater predictive power after including seller fixed effects. What these effects represent is often unclear; e.g., higher online prices by Best Buy could reflect product heterogeneity, generous return policies, or its large network of brick-and-mortar stores. We seek to understand *what* sets sellers apart by quantifying marketing strategies within a clean setting where customer service and return policies are standardized, each product type is physically homogeneous, and no sellers have physical stores.

Our work is also related to Hitsch et al. (2021), Clerides et al. (2023), and Aparicio et al. (2024), which study non-product based differentiation through pricing algorithms, advertisements, and promotion activities, and analyze a firm's ability to differentiate and shield itself from future competition. Lewis (2011) conducts an exercise similar to ours in that he examines whether the presence of phrases associated with vehicle quality (e.g., "dent") influence the final price received in an eBay Motors vehicle auction for heterogeneous vehicles.

2. Empirical Setting and Stylized Facts

This section describes the empirical setting, introduces the various data sources we use, and presents the central empirical puzzle motivating our analysis.

2.1. **The eBay Buy-It-Now Marketplace.** We collected posted-price ("Buy It Now") product listings, which constitute a majority of listings on eBay.[1] Sellers provide a listing title, one or more photographs, product specifications, condition, price, and shipping options. In addition to serving as a trading platform, eBay provides standardized product descriptions, payment infrastructure, seller reputation measures, and a user-friendly money-back guarantee. Buyers locate listings by entering keywords and selecting a product category. eBay then generates a page of search results, which

[1]Einav et al. (2018) finds evidence for increasing relative demand for posted-price listings over auctions on eBay over time. Since consumers can easily filter search results by selling format, we treat auctions and posted-price listings as separate markets.

buyers may reorder by price, browse sequentially, or click through to individual listing pages. During our sample period, conversations with eBay employees indicated that search rankings depended primarily on whether the listing title matched the buyer's search terms and the proximity of the listing's expiration date.

Each product listing webpage has three sections (see Online Appendix E for screenshots). The *title section* contains seller-written title text and photographs together with standardized information including price, shipping terms, product condition, and seller reputation. A second section provides standardized product specifications supplied by eBay. The final *description section* is created optionally by the seller and may include additional text, images, and HTML formatting. This distinction between standardized platform information and seller-created listing content will be central to our analysis. Because the products we study are physically homogeneous and eBay standardizes a portion of product information presented to buyers, systematic price differences associated with seller-customized listing features provide evidence on whether deviations from the Law of One Price reflect perceived listing differentiation rather than underlying product heterogeneity.

2.2. **Data.** Each observation is a single listing for a new, unused product from sellers located within the United States. Our Amazon Kindle Fire HD (8GB version) and Microsoft Surface listings were posted to the platform from December 28, 2012–September 30, 2013; other product listings were collected during January–March 2018 (see Appendix B.1 for further details on data collection).[2] Our raw data consist of HTML pages downloaded from eBay.com (plus associated images when available), enabling us to reconstruct the same layout and content observed by buyers during the sample period. We parsed the HTML source code to identify all external content (description and images), which were also downloaded. While we were able to collect all textual content, we were unable to download some images because their URLs were no longer active.

For our analysis we required that products satisfy two criteria: (*i*) they must be physically homogeneous, and (*ii*) several hundred active listings had to be available. Most candidate products on eBay (e.g., laptops, golf balls, tools, shoes, etc.) failed one of these criteria because listings differed substantially in physical characteristics or bundled accessories. Rather than combining heterogeneous products into a single dataset and attempting to control statistically for product characteristics, we adopt a cleaner single-UPC sampling strategy. This allows listing heterogeneity to be attributed credibly to seller marketing efforts rather than underlying product differences. Using this approach, we identified 14 suitable products spanning seven categories: wearables, mobile devices, computers, audio devices, home appliances, video games, and gaming hardware. Table 1 summarizes the products, sample sizes, and coefficients of variation. Details of the automated scraping procedure and manual data cleaning are provided in Appendix B.1.

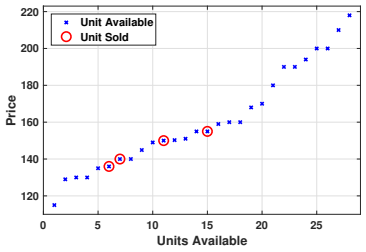
[2]A manual inspection of 200 randomly selected Kindle listings revealed no indications of usage prior to sale.

Category	Product	N	Coefficient of Variation	Category	Product	N	Coefficient of Variation
Wearables	Fitbit Charge 2	2,803	0.175	**Home Appliances**	Dyson Hairdryer	446	0.150
	Apple Watch	1,315	0.221		KitchenAid Mixer KSM150	463	0.256
Mobile Devices	Apple iPod Touch 6, 32GB	436	0.185		iRobot 690 (vacuum cleaner)	340	0.260
	Amazon Kindle Fire HD 7"	1,298	0.212	**Video Games**	Call of Duty, PlayStation 4	1,206	0.290
Computers	Microsoft Surface	922	0.210		Destiny 2, PlayStation 4	730	0.373
Audio Devices	Beats Solo 3	2,124	0.263	**Gaming Hardware**	Nintendo SNES Classic	5,064	0.440
	Apple AirPods	2,185	0.270		Xbox One Kinekt	251	0.630

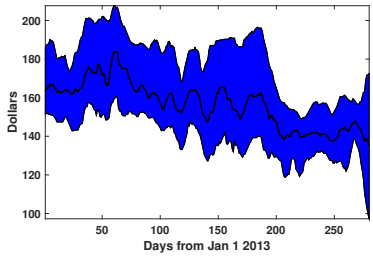
TABLE 1. Products Analyzed

Note: Coefficient of variation is defined as the ratio of de-trended (i.e., cross-sectional) price standard deviation to the de-trended mean.

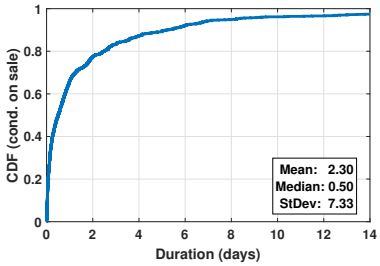
FIGURE 1. eBay Kindle Tablet Market



(A) Kindle Supply Curve: 06/23/2013



(B) Kindle Price Trend Over Time



(C) Conditional Listing Duration

Since our primary interest is sellers' pricing behavior, we include listings regardless of whether they ultimately resulted in a sale. We include only listings offering a single unit and eliminate a small number of listings with implausible prices. If a seller offered multiple listings during our sample period, we treat each listing as a separate observation.

2.2.1. *Outcome Variables.* Across all 14 products, *price* is defined as either the transaction price, if the item sold, or the seller's final posted price before removing the listing from the site. Because our objective is to explain cross-sectional price variation rather than predictable changes over time, we de-trend prices using product-specific linear time trends.[3] Figure 1 illustrates the Kindle market. Panel (A) depicts a typical daily market supply curve and Panel (B) plots the median raw price and interquartile range over time using a seven-day moving average.

Note that prices exhibit a gradual downward trend, reflecting depreciation as newer generations of the product approach release. Removing this predictable time trend allows our subsequent analysis to focus on persistent cross-sectional price dispersion. For Kindles we observe two additional outcomes (see footnote 14): a dummy for whether the listing item *sold* and *listing duration* (in days, see Figure 1, Panel (C)). These will be useful for demand analyses presented in Section 5.

[3]Higher-order polynomials produced negligible differences from the linear specification. Allowing all explanatory variables to interact with time likewise produced no meaningful changes in the results.

2.2.2. *Explanatory Variables.* Our *Basic Dataset* contains variables analogous to those examined in previous studies of online price dispersion (e.g., Baye et al. 2006). These include shipping price, an indicator for calculated shipping, an indicator for whether returns are accepted, an indicator for whether the listing was re-listed, and the logarithm of the seller's reputation score. Summary statistics appear in Table 2. Our *Full Dataset* augments these variables with seller-customized listing characteristics extracted from the HTML source code. These include measures of title length and formatting, seller-provided photographs, description length, HTML formatting, font usage, listing timing, mobile-app usage, and related presentation features. Summary statistics for the Kindle listing features appear in Table 3.

Variable	Mean	Median	Standard Deviation
Price (de-trended)	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
log(Seller Score)	4.78	4.72	2.02
Re-listed	0.168	0	0.374

TABLE 2. Basic Kindle Dataset Summary Statistics

2.2.3. *Natural Language Data.* Natural language data were processed using a Bag-of-Words (BoW) approach (see Gentzkow et al. 2019 for a survey). After stemming each listing's text using Porter's stemming algorithm (Porter 1980), we retained 193 word stems representing plausible sources of listing heterogeneity that appeared in at least five listings. Principal component analysis was then used to reduce the dimensionality of the resulting BoW data (see Appendix F.1). For each product we retained the leading principal components (PCs) explaining at least 90% of the total variance. Within the Kindle data this corresponds to the first 25 PCs.[4]

[4]Using the first 70 PCs (98% of the total variance) produced virtually identical results.

Variable Groups	Variable	Mean (Median)	StDev	Variable Groups	Variable	Mean (Median)	StDev
Title Length	Number of Characters in Title	60.1 (66)	16.4	**Descr. Length**	Number of Words in Description	131 (27)	281
	Number of Words in Title	10.6 (12)	2.68		% Uppercase Characters in Description	0.122 (0.075)	0.184
Title Style	Number of Special Characters in Title	15.0 (16)	4.63	**Descr. Style**	Number of HTML Tags in Description	102 (16)	201
	% Uppercase Characters in Title	0.312 (0.263)	0.168		Number of Font Sizes	1.58 (1)	1.451
Title Images	Number of Photos in Title	2.15 (1)	1.86		Number of Font Size Changes	4.33 (0)	16.0
	Number of Stock Photos in Title	0.439 (0)	0.509	**Misc.**	Dummy: Posted During Weekend	0.277 (0)	0.448
Descr. Images	Kilobytes of Photos in Description	16.0 (0)	80.1		Dummy: Posted During Early Week	0.429 (0)	0.495
	Dummy: 1 to 5 Photos in Description	0.103 (0)	0.304		Dummy: Posted During Late Week	0.294 (0)	0.456
	Dummy: 6 or more Photos in Description	0.063 (0)	0.243		Dummy: Posted with eBay Mobile	0.242 (0)	0.428

TABLE 3. Kindle Listing Feature Summary Statistics (without Bag of Words data)

% Variance Explained	Word Stems w/Top 10 Highest Factor Loadings (in descending order)	LLM-based Semantic Themes
PC_1: 43.6%	"read," "includ," "content," "free," "new" "app," "all," "more," "ship," "devic"	**ChatGPT 5.2:** "Content-centric value framing (tablet as a gateway to reading and digital media)" **Claude 3 Opus:** "Content Access & Functionality" or "Digital Content Ecosystem"
PC_2: 11.9%	"ship," "return," "not," "payment," "buyer" "paypal," "contact," "feedback," "onli," "packag"	**ChatGPT 5.2:** "Seller credibility, payment security, and post-purchase assurance" **Claude 3 Opus:** "Transaction & Logistics" or "Purchase Process & Policies"
PC_3: 7.3%	"display," "connect," "gb," "charg," "batteri" "usb," "processor," "hour," "type," "offer"	**ChatGPT 5.2:** "Hardware specifications and technical performance framing" **Claude 3 Opus:** "Technical Specifications & Hardware" or "Premium Features & Guarantees"
PC_4: 4.2%	"includ," "warranti," "display," "not," "million" "addit," "black," "touch," "technolog," "support"	**ChatGPT 5.2:** "Warranty-backed, branded quality framing" **Claude 3 Opus:** "Product Quality & Warranty/Support" 'or 'Premium Features & Guarantees"

TABLE 4. Interpretation of Kindle Dataset BoW Principal Components 1–4

If we examine word stems with the highest factor loadings within the leading PCs, it becomes possible to attach economic interpretations to them. Table 4 reports the ten highest-loading word stems for each of the first four Kindle PCs, in descending order of factor loadings, along with semantic interpretations produced independently by two leading large language models.[5] The two models agreed closely, interpreting the four components as corresponding broadly to content-based value framing, transactional assurances, technical specifications, and warranty or legitimacy signaling. This agreement, together with the semantic cohesion of the underlying word stems, suggests that the PCA procedure captures economically meaningful dimensions of listing content.

2.2.4. *Sources of Credible Signaling.* Consumers may perceive that seller-provided text and images convey information about product condition or packaging that is not fully captured by eBay's standardized descriptions. Images can verify attributes such as factory sealing, while text may disclose product characteristics affecting value. The credibility of these disclosures is reinforced by eBay's seller reputation score, which we include explicitly in our empirical analysis.

Beyond direct disclosure, listing presentation may itself signal seller professionalism and reliability. Richly formatted listings make extensive use of images and HTML formatting, improving visual organization and readability. These features are relatively inexpensive for experienced sellers to reproduce using reusable templates, but more difficult for inexperienced sellers to imitate. Some elements, such as links to other active listings or external websites, cannot be credibly copied at all.

Taking seller score as a measure of experience on eBay, Table 5 illustrates the relationship between listing design and seller experience by comparing sellers with scores above and below the median. Experienced sellers use listings that are both longer and more stylized. For example, their

[5]Both LLMs received identical prompts and inputs consisting of the ten highest-loading word stems, listed in descending order without their factor loadings. The prompts explained that the PCs were computed from BoW stem-count data for homogeneous Kindle Fire (8GB) listings and asked each model to assign a descriptive semantic theme to each PC.

descriptions contain nearly three times as many words and HTML tags, and they use roughly three times as many font-size changes. Mean differences across experienced and less-experienced sellers are highly statistically significant for every variable reported in the table.

Variable	Mean: Low Seller-Score	Mean: High Seller-Score
Number of Words in Title	9.96 (*s.e.* 0.11)	11.19 (*s.e.* 0.09)
Number of Special Characters in Title	14.25 (*s.e.* 0.19)	15.68 (*s.e.* 0.17)
% Uppercase Characters in Title	0.28 (*s.e.* 0.01)	0.34 (*s.e.* 0.01)
Number of Words in Description	68.22 (*s.e.* 6.28)	192.00 (*s.e.* 13.8)
% Uppercase Characters in Description	0.10 (*s.e.* 0.01)	0.15 (*s.e.* 0.01)
Number of HTML Tags in Description	52.32 (*s.e.* 5.58)	152.86 (*s.e.* 9.26)
Number of Font Size Changes	2.14 (*s.e.* 0.49)	6.51 (*s.e.* 0.73)

TABLE 5. Listing Characteristics in Kindle Dataset, Split by Seller Score

2.3. **The Empirical Puzzle: Price Dispersion in Homogeneous Goods.** Figure 1 illustrates the central empirical puzzle motivating our analysis. Despite describing a market for physically homogeneous products, the Kindle market exhibits substantial cross-sectional price dispersion and shows no indication of converging toward the Law of One Price over our nine-month sample period (Panel B). Panel (A) depicts a typical daily market supply curve for Kindles. Listings that ultimately sell tend to have lower prices than those that remain unsold, although many relatively inexpensive listings also fail to sell.[6] The existence of markedly different price quotes at any point in time for Kindles, is notable, given that they are for identical products and buyers commonly see listings side by side in search results. Table 1 shows that this is not merely an artifact of the market for mobile devices. We also see substantial price dispersion across each of the 14 products in our sample, with the cross-sectional standard deviation of price ranging between 15% and 63% of the mean. Weighting by sample size, the average (median) coefficient of variation across the 14 products we sampled is 0.294 (0.263).

These observations do not, by themselves, establish that the Law of One Price fails. Some degree of price dispersion could arise from heterogeneous seller reservation values, inventory costs, or from predictable intertemporal changes in product values. However, several features of our setting substantially limit the explanatory power of these mechanisms. First, our analysis focuses exclusively on physically homogeneous products. Second, prices have been de-trended to remove predictable intertemporal movements associated with product depreciation. Finally, the products we study are inexpensive consumer goods with negligible storage costs, suggesting that sellers should be relatively patient. Taken together, these considerations suggest that the stylized facts documented in this section warrant closer theoretical scrutiny. The remaining question is therefore whether the persistent price dispersion documented here reflects economically meaningful departures from the

[6]Some listings exceed Amazon's retail price of $199, but are less likely to sell. Malmendier and Lee (2011) document similar overpayment relative to contemporaneous retail prices in other eBay settings.

frictionless benchmark or instead arises from other mechanisms such as search frictions or perceived differentiation created by heterogeneous product listings.

3. Theoretical Background and Empirical Strategy

3.1. **A Frictionless Benchmark for Posted-Price Markets.** We begin with a model of a frictionless posted-price market for a homogeneous good. The model predicts that price dispersion ought to be minimal (Proposition 1) and is designed to underscore market features—including heterogeneous seller production costs, storage costs, and reservation values—that, by virtue of their inclusion, can be ruled out as sufficient explanations for the stylized facts documented in Section 2.

Sellers each have a single unit for sale and choose a take-it-or-leave-it posted price, p, for the good. They have heterogeneous reservation values, c, drawn from a distribution $F_C(c)$ on compact support $[\underline{c}, \overline{c}]$. These may represent sellers' values from retaining the good, production costs, and/or storage costs. We allow the market-clearing price to be random in equilibrium, and let $p_M(t, \delta)$ denote its realization in period t given time discount factor δ. A seller's item sells in period t if $p \leq p_M(t, \delta)$, which occurs with probability $\Pi_t(p) \equiv \Pr\{p \leq p_M(t, \delta)\}$.[7] Otherwise, the seller can continue to offer the good in future periods. Sellers have a common discount factor, δ; this assumption can be relaxed at the cost of additional notation.

A seller with reservation value c solves $\mathcal{V}_t(c; \delta) = \max_{p \geq 0} \{\Pi_t(p)(p - c) + \delta [1 - \Pi_t(p)] \mathcal{V}_{t+1}(c; \delta)\}$, where $\mathcal{V}_t(c; \delta)$ is the value function. The corresponding policy function is

$$p_t(c; \delta) = \arg\max_{p \geq 0} \{\Pi_t(p)(p - c) + \delta [1 - \Pi_t(p)] \mathcal{V}_{t+1}(c; \delta)\}.$$

The continuation value creates an opportunity cost of selling in a given period. Proposition 1 shows that as δ grows, this opportunity cost induces sellers to choose prices increasingly close to the top of the support of the market-clearing prices realized over time.

We assume that for seller behavior to be consistent with the realized market-clearing price, at least one seller must offer a price at or below $p_M(t, \delta)$:

$$\exists\, c \text{ such that } p_t(c; \delta) \leq p_M(t, \delta). \tag{1}$$

Definition 1. The market-clearing price $p_M(t, \delta)$ exhibits (γ, ρ, T)-price dispersion if there exist real numbers $\underline{\mathcal{P}}(\delta)$ and $\overline{\mathcal{P}}(\delta)$ such that, for all t,

$$\overline{\mathcal{P}}(\delta) - \underline{\mathcal{P}}(\delta) > \gamma, \tag{2}$$

$$\Pr\left[p_M(t + \tau, \delta) < \underline{\mathcal{P}}(\delta) \text{ for some } \tau \in \{1, \ldots, T\}\right] > \rho, \quad \textit{and} \tag{3}$$

$$\Pr\left[p_M(t + \tau, \delta) > \overline{\mathcal{P}}(\delta) \text{ for some } \tau \in \{1, \ldots, T\}\right] > \rho. \tag{4}$$

[7] We model the demand side of the market in reduced form, so departures from the posted-price homogeneous-good environment may be embedded within $\Pi_t(p)$. For example, if substitute products or listings using other selling mechanisms are available, $\Pi_t(p)$ may reflect shocks to those related markets.

Intuitively, price fluctuations are more extreme as γ increases, more common as ρ increases, and occur over shorter spells as T decreases. These conditions hold for some $\gamma, \rho > 0$ and $T = 1$ if $p_M(t, \delta)$ is i.i.d. across periods with nondegenerate support. They also hold for some $\gamma, \rho > 0$ and finite T if $p_M(t, \delta)$ takes values in a finite set and follows an aperiodic, irreducible Markov process.

For $\delta < 1$ sufficiently large, (γ, ρ, T)-price dispersion for any fixed $\gamma, \rho > 0$ and $T < \infty$ is inconsistent with optimal seller behavior. The intuition is that when sellers are sufficiently patient, even small revenue improvements are worth waiting for. Sellers therefore choose prices close to $\overline{\mathcal{P}}(\delta)$ regardless of c. But this behavior violates condition (1), which requires some sellers to post prices spanning $\left[\underline{\mathcal{P}}(\delta), \overline{\mathcal{P}}(\delta)\right]$ when the gap between the endpoints is nontrivial.

Proposition 1. ***(Law of One Price)*** *For any choice of $(\gamma, \rho) \in (0, 1)^2$ and $T < \infty$, there exists $\delta' < 1$ such that $p_M(t, \delta)$ cannot exhibit (γ, ρ, T)-price dispersion if $\delta \in (\delta', 1)$.*

A proof is provided in Appendix A. The proposition establishes that variation in the market-clearing price is incompatible with sufficiently patient sellers in a frictionless market for homogeneous goods. Interpreting δ as a daily discount factor, it is natural to view sellers of small, easily stored consumer goods as patient. Heterogeneity in production costs, storage costs, or reservation values therefore cannot, by itself, plausibly account for persistent price dispersion.

The eBay setting differs in this respect from commodity spot markets, where liquidity traders may need to transact quickly and storage costs may be substantial. Oil refiners, for example, may be effectively impatient because storing their products is costly. While impatient buyers on eBay are plausible (e.g., Coey et al. 2020), it is less clear why sellers would be eager to dispose of an inexpensive, easily stored product when waiting might yield a higher price. Nor is predictable nonstationarity likely to explain the observed dispersion. Product values may decline as newer models approach release, as illustrated for Kindles in Figure 1(B). Such depreciation makes future sales less attractive, much like a lower discount factor. However, unless the trend is steep, it should account for only a small fraction of the observed price dispersion. Moreover, our empirical analysis removes product-specific time trends before measuring cross-sectional price variation.

3.2. **Remaining Explanations for Price Dispersion.** The benchmark model rules out various explanations for the stylized facts documented in Section 2. If sellers are relatively patient within a frictionless market, heterogeneity in production costs, storage costs, or reservation values cannot, by itself, sustain persistent price dispersion. We therefore focus on two remaining explanations: market frictions and perceived differentiation created by seller customization of product listings.

A large theoretical literature shows that market frictions can support equilibrium price dispersion for homogeneous goods. One possibility is that consumers may be imperfectly informed because they face heterogeneous search costs or observe different numbers of price offers (e.g., Salop and Stiglitz 1977; Wilde and Schwartz 1979; Rosenthal 1980; Varian 1980; Baye and Morgan 2001). A

second possibility is that search frictions may interact with heterogeneous seller costs or reservation values (Reinganum 1979; MacMinn 1980). Third, even when buyers and sellers are ex ante homogeneous, stochastic search and mixed pricing strategies can generate dispersed equilibrium prices (Burdett and Judd 1983). Importantly, in each of these three cases, the resulting price variation reflects imperfect consumer information and/or strategic seller mixing behavior, rather than publicly observable differences among the product listings being served to buyers by the platform.

Alternatively, physically homogeneous products may appear differentiated because sellers vary the content and presentation of their listings. Seller-provided text and images may communicate information about product value, packaging, condition, authenticity, or reliability that is not fully captured by the platform's standardized fields. Listing style may also signal seller professionalism or competence. If consumers attach value to these signals, then sellers offering identical physical products may optimally charge different prices.

The scope for this mechanism depends partly on platform design, as seller customization varies considerably across e-commerce platforms. Amazon and Shein largely standardize listing content, whereas eBay and Alibaba combine standardized product information with seller-customized titles, photographs, and descriptions. Platforms such as Temu and Etsy rely even more heavily on seller-created content. eBay therefore occupies an attractive middle ground as an empirical case study: it hosts an ample supply of physically homogeneous products with some standardized listing content, while allowing sellers considerable latitude in how those products are presented to buyers. Moreover, its internal search engine provides a microcosm of broader online search (e.g., Google), presenting consumers with multiple listings that differ in design and information content for a given product. Together, these features make eBay a particularly useful setting for studying whether seller marketing efforts contribute to deviations from the Law of One Price.

3.3. **Empirical Implications.** Although market frictions and listing heterogeneity can both generate persistent price dispersion, they have different implications for the relationship between prices and observable listing characteristics. Perceived differentiation predicts that prices vary systematically with the content, layout, and appearance of product listings. By contrast, the common thread among the economic theory models discussed above is that price variation generated directly or indirectly by search frictions—e.g., through random search outcomes, heterogeneous search costs, or mixed-strategy pricing—is plausibly orthogonal to observable listing characteristics. This distinction motivates our empirical strategy.

Formally, for the i^{th} listing, let $\mathbf{X}_i$ denote a row vector containing the variables in our Basic Dataset, used in previous studies of online price dispersion: *shipping price*, *shipping is calculated*, *returns allowed*, log(*seller score*), and *re-listed*. Let $\mathbf{Z}_i$ contain the additional variables in our Full Dataset, including title variables $\mathbf{Z}_{1i}$, description and image variables $\mathbf{Z}_{2i}$, and the principal components of the BoW data $\mathbf{Z}_{3i}$, where $\mathbf{Z}_i \equiv [\mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \mathbf{Z}_{3i}]$. Finally, let y_i denote the *de-trended*

listing price. We then consider a basic pricing relationship of the form $y_i = f(\mathbf{X}_i) + e_i$, versus an augmented relationship of the form $y_i = \varphi(\mathbf{X}_i, \mathbf{Z}_i) + \varepsilon_i$.

The predictable components $f(\mathbf{X}_i)$ and $\varphi(\mathbf{X}_i, \mathbf{Z}_i)$ capture systematic price variation associated with observed seller and listing characteristics. The remaining residual variation e_i may reflect search frictions, but it may also include unobserved listing heterogeneity if f is mis-specified: $e_i = \varphi(\mathbf{X}_i, \mathbf{Z}_i) - f(\mathbf{X}_i) + \varepsilon_i$. Under the orthogonality argument above, price variation explained by the seller-customized variables $\mathbf{Z}_i$ reflects perceived listing heterogeneity rather than pricing noise generated by market frictions. Accordingly, the fraction of price variation explained by the Full Dataset provides a lower bound on the importance of observed listing differentiation. Conversely, the residual variation from the augmented model ε_i provides an upper bound on the fraction of price dispersion that can plausibly be attributed to market frictions.[8]

One concern is that some listing features we collected affect the order in which search results are served to users; this would violate orthogonality and conflate the roles of listing heterogeneity and market frictions. The "Best Match" algorithm eBay uses was based on keywords contained in the listing's title during our sample periods. Thus, if price dispersion were generated by seller manipulations of eBay's Best Match algorithm, a practice commonly known as *search engine optimization*, it would not be orthogonal to $\boldsymbol{Z}_i$.[9] In order for a given seller to effectively manipulate the search algorithm through strategic listing-title design, two conditions must be satisfied: (i) buyers collectively use a predictable set of keywords for their eBay searches, and (ii) many other sellers do *not* include these keywords in their listing titles. In Section 4.1.1 we test for such patterns in the data and find no evidence consistent with search-algorithm manipulation.

3.4. **Empirical Strategy.** Previous studies have found it difficult to explain online cross-sectional price variation without seller fixed effects (e.g., Clay et al. 2002; Baye et al. 2006), which are difficult to interpret. There are two possible reasons why the researcher may attribute too little explanatory power to the model and too large a role to the error term: *omitted variables* and *functional form mis-specification*. If the regression functions took a linear-in-parameters form, say $f(\mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and $\varphi(\mathbf{X}_i, \mathbf{Z}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}$, then the error terms would be related by $e_i = \mathbf{Z}_i\boldsymbol{\gamma} + \varepsilon_i$. Thus, an omitted variables problem would therefore hinge on whether $\mathbf{Z}_i$ had a meaningful impact on prices, which in the linear model is the same as $\boldsymbol{\gamma} \neq \mathbf{0}$. On the other hand, linear models like $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}$ do not allow complex interactions between the observables in signaling professionalism that provide scope for a seller to raise their price. In that case, a non-separable functional form for f or φ might be required to realize full explanatory power of the observables $(\mathbf{X}_i, \mathbf{Z}_i)$.

[8] To investigate the relative importances of listing heterogeneity and market frictions, note that we need not identify a causal demand system. Rather, our goal is more modest: we attempt to determine what combination of more ample observables and more flexible statistical models can exhaust predictive power for cross-sectional price differences.

[9] eBay maintains strict secrecy on the particulars of its search algorithm to discourage search engine optimization.

The flexible modeling approach we use to account for subtle interactions among our ample seller signaling covariates is *random forests* (see Appendix B.2 for full details).[10] Aside from asking *how much* price variation can be explained, the more interesting questions are *why*. Is the explanatory power due to rich observables, more flexible machine-learning (ML) methods, or a combination of the two? To assess the importance of our rich data, we compare the predictive power of linear models estimated on the full dataset and on the basic dataset, respectively. To evaluate the importance of the ML approach, we compare predictive power of random forest models and linear models estimated on both basic and full datasets.

Given our large set of regressors, both OLS and random forest techniques are prone to overfitting, so we measure explanatory power with an *out-of-sample* R^2 *statistic* using 10-fold cross validation (see Appendix B.3 for details). This design allows us to address the problems of omitted variables and functional form mis-specification encountered in the earlier literature. If the Full Dataset substantially outperforms the Basic Dataset, then omitted listing information is important. If the random forest substantially outperforms OLS when applied to the same data, then functional-form restrictions are important. If both improvements are necessary to explain a maximal share of price variation, then adequately measuring perceived listing differentiation requires both richer data and more flexible empirical methods.

4. Main Results

Table 6 provides R^2_{out} for the random forest and OLS models when estimated separately for each product using the basic and full datasets, for each of the 14 product samples. The first thing to note is that the random forest estimated using the full dataset predicts at least 26.4% of price variation, or at least three times as much price variation as the OLS model estimated on the basic dataset, which is analogous to techniques and data used by the prior literature. For seven products it predicts near or above half of cross-sectional price variation, and it explains 48% of price variation on average across all products we consider.

We get a sense of the flexibility of random forest methods relative to OLS by comparing the columns describing the predictive power of the random forest and OLS models estimated on the same dataset. When estimated on the full dataset, the random forest explains at least 60% more price dispersion than the OLS model, and often much more. To illustrate the importance of the full dataset versus the basic dataset, it suffices to make within-model comparisons. Across all products in Table 6, random forest explanatory power more than doubles, on average, when using the full dataset, relative to using only the basic dataset. The percent change comparison of explanatory power across datasets is substantially larger for the OLS model.

[10] We also estimated LASSO with up to third-order interactions (Appendix B.4 and Appendix C), boosted gradient trees, and neural networks (Appendix B.5); none outperformed random forests. Random forests require minimal tuning; boosting and deep nets demand extensive hyperparameter search, which tends to overfit in samples of our size.

Category	Product	R^2_{out} Rand Forest Full Dataset	R^2_{out} Rand Forest Basic Dataset	R^2_{out} OLS Full Dataset	R^2_{out} OLS Basic Dataset
Wearables	Fitbit Charge 2	0.569	0.331	0.212	0.027
	Apple Watch	0.338	0.148	0.044	0.018
Mobile Devices	Apple iPod Touch 6, 32GB	0.516	0.399	0.178	-0.023
	Amazon Kindle	0.415	0.198	0.190	0.130
Computers	Microsoft Surface	0.438	0.161	0.068	-0.012
Home Appliances	Dyson Hairdryer	0.422	0.174	0.262	0.036
	KitchenAid Mixer KSM150	0.485	0.295	0.282	0.054
	iRobot 690	0.632	0.512	0.328	0.116
Audio Devices	Beats Solo 3	0.383	0.133	0.074	0.009
	Apple AirPods	0.672	0.496	0.321	0.148
Video Games	Call of Duty, PlayStation 4	0.264	0.121	0.086	0.000
	Destiny 2, PlayStation 4	0.585	0.432	0.336	0.149
Gaming Hardware	Nintendo SNES Classic	0.582	0.325	0.084	0.032
	Xbox One Kinekt	0.371	0.104	-0.029	-0.007

TABLE 6. Price Predictability

Note: product categories appear in ascending order (from top to bottom) of sample-weighted average coefficient of variation.

In short, explaining price variation requires both the rich dataset and the flexibility of the random forest model. The patterns we find in Table 6 indicate that, not only do users derive a value signal from various pieces of information, but these different pieces may also interact in complex and subtle ways (within the flexible random forest model) to create a perception of value.

4.1. **Robustness Checks.** Online Appendix B presents various robustness checks using our Kindle data. First, perhaps a more flexible linear-in-parameters model estimated on the full dataset might perform better than OLS and more comparably to ML methods. We estimate a LASSO model (Appendix B.4) on the full dataset with first- second- and third-order interactions; the predictive power is still far lower than a regression forest. Second, we repeat our main analysis using other ML methods such as neural networks and boosted gradient trees, but these did not substantially improve upon the explanatory power of the random forest method when applied to our data (see Appendix B.5). Third, we assess the impact of sellers with multiple listings (Appendix B.6), which could inflate our predictive power, and find that this is not a significant driver of our main results. Fourth, in Appendix B.8 we repeat our analysis restricting attention to the subset of listings that resulted in a sale, and show that this distinction is not driving our main results either.

4.1.1. *Testing for Search-Algorithm Manipulation.* Appendix B.7 executes various robustness tests within the Kindle sample to probe for the possibility that sellers are able to effectively use listing design to systematically guarantee themselves more prominent placement within search results. First, we find that the most common title keywords tend to appear either in virtually all listing titles (e.g., "GB"), or in too few to facilitate widespread manipulation (e.g., "sealed," "factory," or "screen"). An exception is the keyword "new," which appears in 61% of listing titles. Second, we implement two tests for search-algorithm manipulation proposed by Backus et al. (2014). We

find that covariate explanatory power is slightly *higher* among listings with the keyword "new" in the title, and among listings with titles that are longer than the median title length. This is the opposite of what one would expect if search algorithm manipulation was present (see Appendix B.7 for further discussion). Thus, patterns in our data are consistent with our central assumption that seller-customized listing traits are plausibly orthogonal to price variation driven by search frictions.

4.2. **Sources of Predictive Power.** The complex structure of a regression forest makes it difficult to assess the importance of individual regressors. How then can we determine which variables drive predictive power, whether they contain distinct information, or whether prediction relies on interactions among variables? To address these questions, we use the variable-importance measure proposed by Breiman (2001). The basic idea is simple: we "garble" the information contained in a variable by randomly permuting its values across listings, and then measure how much predictive power the (pre-estimated) model loses. If a variable contains information that the random forest uses for prediction, scrambling that information should cause out-of-sample R^2 to fall. Variable importance is the difference between the original R^2_{out} and resulting R^2_{out} after garbling the variable; the larger the decline in predictive power, the more important that variable is to the model.[11]

For comparison, we also compute partial-R^2_{out} for three variable groups (see Appendix G). Partial-R^2 measures the loss in predictive power when a variable group is excluded and the random forest is re-grown using the remaining information. Variable importance instead holds the fitted forest fixed and measures the loss in predictive power when the information contained in a variable group is garbled. Thus, partial-R^2 allows the model to adapt to a reduced information set, while variable importance measures how heavily the original fitted model relies on the garbled information.

Results for the Kindle data are reported in Table 7. Partial-R^2 is similar across the three variable groups, suggesting substantial substitutability. In contrast, variable importance reveals differences in how heavily the full random forest relies on each information source. Description variables are more important than title variables, and the BoW variables are most important of all. Garbling the full set of BoW components reduces R^2_{out} by more than twice the reduction from garbling all description variables. Later BoW components retain nontrivial importance, indicating that the random forest extracts useful predictive information well beyond the leading PCs.

We draw two conclusions from this analysis. First, the data-adaptive model loses relatively little predictive power when restricted to title, description, or BoW variables alone. This suggests that different listing features carry substantial redundant information from a seller signaling perspective: interactions across variable groups are less crucial for accurate price predictions. Second, although the full model uses information from all three groups, it relies most heavily on the BoW variables,

[11] Formally, when computing R^2_{out} via cross-validation, we randomly permute the values of variables in $\boldsymbol{Z}_l$, $l = 1, 2, 3$, across listings within each hold-out sample. We then recompute out-of-sample R^2 using the permuted data and average across the 10 folds. We repeat this procedure 100 times and denote the resulting average by pR^2_{out}.

Variable Group	Subset	Variable Name(s)	Permuted R^2_{out}	Variable Importance	Partial R^2_{out}
Title Variables	Volume of Information Conveyed by Title	Number of Characters/Words and Images	0.403	0.013	—
	Title Text Style	Number of special characters, % Upper case	0.408	0.008	—
	All Title Variables	—	0.391	0.026	0.227
Description Variables	Descr. Images	Kilobytes of Photos, Dummy for 1 - 5 Photos, Dummy for 6 or More Photos	0.394	0.023	—
	Descr. Length	Number of Words in Description	0.394	0.023	—
	Volume of Info. Conveyed by Descr.	Number of Words and Size and Number of Photos in Descr.	0.350	0.067	—
	Descr. Text Style	Number of Font Sizes, Number of Font Size Changes Number of HTML Tags, % Upper Case Characters	0.376	0.041	—
	All Descr. Variables	—	0.271	0.145	0.243
Bag of Words Variables	Component 1	Description of Item	0.384	0.033	—
	Component 2	Shipping and Payment Information	0.406	0.011	—
	Components 3 - 15	—	0.279	0.138	—
	All BoW Components	—	0.095	0.322	0.263

TABLE 7. Variable Importance: Kindle Listings

with description variables playing a secondary role. This highlights the importance of analyzing seller-provided text when explaining price variation.

4.2.1. *Other Products.* Table 14 (Appendix C) reports partial-R^2 measures and Table 15 (Appendix C) reports variable importances of the three variable groups for non-Kindle products. Although all three sections of the listing convey a significant amount of information, the random forest often places a significantly greater importance on one section. Which one is most important differs between products: for 12 out of 14 products, the description was either nearly as important or more important than the title section, and for 6 of 14 products, BoW variables are the most important.

4.3. **Honest Model Forest.** The random forest predicts prices well, but its complex structure makes it difficult to interpret how individual listing characteristics affect those predictions. To recover interpretable marginal effects, we use an *honest model forest* (Wager and Athey, 2018). A *model tree* resembles a random forest tree, but replaces the constant prediction within each leaf with a statistical model—OLS in our case. Each leaf therefore corresponds to a subset of similar listings, with marginal effects given by the leaf-specific OLS coefficients.

We interpret the honest model forest as a best locally linear predictor, allowing us to average marginal effects across listings or examine how they vary across leaves. A key feature of our implementation is that the forest cannot split on the variable whose marginal effect is being estimated, preserving within-leaf variation in that variable. Full implementation details are provided in Online Appendix D. Although our analysis is descriptive, the estimated marginal effects align closely with

economic intuition about seller incentives. As argued by Zhao and Hastie (2021), when predictive performance is strong, domain knowledge is available a priori, and appropriate visualization tools are used, ML-based estimates may provide credible evidence about underlying causal relationships.

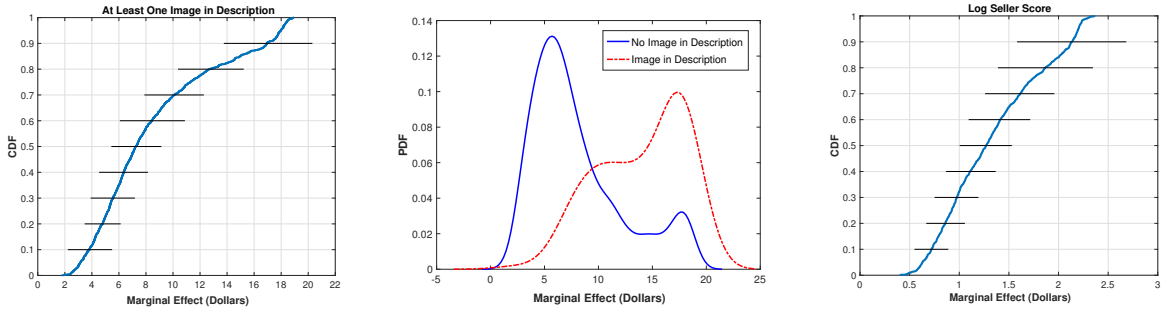


FIGURE 3. Marginal Effects: Including ≥ 1 Image in Description, and Seller Score

The first two panels of Figure 3 examine heterogeneity in the predicted price effect of including an image in the description. The left panel reveals substantial heterogeneity: marginal effects vary widely across listings, with statistically significant differences between the first and ninth deciles.[12] The middle panel provides suggestive evidence that sellers tailor listing design to where it is most valuable. We divide the sample between the 216 listings containing at least one image in the description and the 1,082 listings without one. The average marginal effect of an image is \$14.02 among listings in which sellers actually include one, compared with \$8.06 among listings without an image (mean difference p-value$< 1\%$). Thus, sellers are more likely to include an image precisely when the other characteristics of the listing imply a larger price premium from doing so. While our data do not permit a direct analysis of seller motives or consumer search, these patterns suggest that seller marketing strategies are tied to expectations for boosting demand, consistent with profit-maximizing sellers tailoring listing design to optimize value signaling.

Turning to the right panel of Figure 3, our analysis contributes to the related literature—e.g., Bajari and Hortaçsu (2004), Cabral and Hortaçsu (2010), and Nosko and Tadelis (2015)—by allowing for heterogeneous marginal effects of seller reputation. We find substantial heterogeneity: the predicted marginal effect of a one-log-point increase in seller score ranges from just under \$0.50 to just below \$2.50 across listings.[13] This heterogeneity suggests that different signals of seller reliability may partially substitute for one another. A seller with a sparse listing may benefit substantially from an increase in reputation, whereas a seller with a low reputation score can potentially compensate by constructing a richer listing with images and stylized text. More generally, the results

[12]We compute confidence intervals by forming 1,000 bootstrap samples of the estimation set and re-estimating the trees for each sample. Of the 1,298 listings, 332 have marginal effects above the average and 592 below the average at the 90% confidence level; average effects in the first and ninth deciles differ at the 99% confidence level.

[13]Of the 1,298 listings, 310 have marginal effects above the average and 502 below the average at the 90% confidence level.

suggest that buyers may infer professionalism and reliability from multiple aspects of the listing rather than from seller reputation alone.

5. Demand Analysis

The correlations between seller reputation score and volume of design elements in Table 5 suggest that more experienced sellers purposefully choose listing attributes in a homogeneous-good market with anticipated demand responses in mind. If so, then the price variation explained by our main predictive model reflects proactive efforts to signal either product utility or seller quality. In this section we assess whether this interpretation is a reasonable view of the data by analyzing demand responses. Two additional outcome variables, listing *duration* and an indicator for whether the item *sold*, are cleanly observable within the Kindle sample.[14] These allow us to investigate how publicly observable listing features translate into realized demand outcomes.

5.1. **Predicting Kindle Demand.** More concretely, we examine whether listing characteristics have residual explanatory power for sale probability and listing duration (conditional on sale) after controlling for seller characteristics and price. Economic theory suggests that seller traits (proxies for reliability) and posted price should influence these demand outcomes, so we include them as baseline controls. Let $\boldsymbol{X}^+$ contain our *Basic* variables $\boldsymbol{X}$ (see Table 2) plus a cubic polynomial $(price, price^2, price^3)$, and a full set of *price*-$\boldsymbol{X}$ interactions. Let $\boldsymbol{Z}_l^+$ contain $\boldsymbol{Z}_l$ (see Tables 3 and 4) plus *price*-$\boldsymbol{Z}_l$ interactions for l =1 (title), 2 (description), and 3 (BoW), with $\boldsymbol{Z}^+ \equiv [\boldsymbol{Z}_1^+, \boldsymbol{Z}_2^+, \boldsymbol{Z}_3^+]$.

We estimate a sequence of predictive demand models using the two-stage procedure proposed by Belloni and Chernozhukov (2013): model selection via cross-validated LASSO, followed by parameter estimation using the selected covariates.[15] Our objective here is not to maximize predictive performance over $\boldsymbol{X}^+ \cup \boldsymbol{Z}^+$, but rather, to test whether the blocks of listing controls $(\boldsymbol{Z}_1^+, \boldsymbol{Z}_2^+, \boldsymbol{Z}_3^+)$ retain explanatory power, measured by partial-R^2, after conditioning on $\boldsymbol{X}^+$.

Results are reported in Table 8. For listing duration, the combined listing variables account for additional explanatory power (partial-$R^2_{\boldsymbol{Z}^+} = 0.086$) that exceeds the baseline from seller characteristics and price alone ($R^2_{\boldsymbol{X}^+} = 0.077$). Each variable group adds between 45% and 55% of baseline explanatory power, when added to the model with only the $\boldsymbol{X}^+$ controls. For sale probability, adding the full set of listing controls augments predictive power by 62%. Once again, individual listing control blocks significantly increase explanatory power by between 17% and 37% when added to baseline controls. These results suggest that listing features indeed translate into demand, and demonstrate that the results in Section 4 are not an artifact of uninformed or irrational sellers.

[14] This is possible because of the length of the Kindle sample period (especially important for observing duration) and because all sampled Kindle pages were single-unit listings. Multi-unit listings (common in many product categories) obscure sale probability and listing duration within the raw scraped HTML code for each product page, by making it difficult to tell when each unit sold.

[15] For the duration (sale probability) outcome, the model selection step optimizes a linear LASSO (LASSO-Logit) model via 10-fold cross validation, followed by OLS (Logit) on the selected variables, with a constant term and *price* always included. This process is repeated 100 times for robustness to 10-fold partition randomness, with R^2 measures and model predictions averaged across the 100 models.

Outcome	$R^2_{X^+}$ Baseline Model	Partial-$R^2_{X^+\cup Z^+}$ (Full)	Partial-$R^2_{Z_1^+}$ (Title)	Partial-$R^2_{Z_2^+}$ (Description)	Partial-$R^2_{Z_3^+}$ (BoW)	Avg SD Partial Effect: Z_3 (BoW)
Duration, Cond. on Sale (OLS, $N = 995$)	0.0773	0.0861	0.0420	0.0347	0.0376	-1.2888
Sale Prob. (Logit, $N = 1{,}298$)	0.2727	0.1703	0.0464	0.0631	0.1016	0.13721

TABLE 8. Predicting Demand Outcomes: Kindle Tablets

Notes: The *Baseline Model* includes only controls in $\boldsymbol{X}^+$. $R^2_{\boldsymbol{X}^+\cup\boldsymbol{Z}^+}$ for duration (sale prob.) is 0.1568 (0.3965). Partial-$R^2_{\boldsymbol{Z}_l^+}$ for OLS is computed as $(R^2_{\boldsymbol{X}^+\cup\boldsymbol{Z}_l^+} - R^2_{\boldsymbol{X}^+})/(1-R^2_{\boldsymbol{X}^+})$, $l=1,2,3,\emptyset$. Partial-$R^2_{\boldsymbol{Z}_l^+}$ for Logit uses the McFadden pseudo-R^2 formula: $1 - Dev_{\boldsymbol{X}^+\cup\boldsymbol{Z}_l^+}/Dev_{\boldsymbol{X}^+}$, $l=1,2,3,\emptyset$, where $Dev_{\boldsymbol{W}} \equiv \sum_{i=1}^{N}\left\{sold_i \times \log(\hat{\Pr}[sold|\boldsymbol{W}_i]) + (1-sold_i)\times\log(1-\hat{\Pr}[sold|\boldsymbol{W}_i])\right\}$ denotes deviance of the model estimated with regressors $\boldsymbol{W}$. *Avg SD Partial Effect* is the mean predicted change in the outcome—within the full $\boldsymbol{X}^+ \cup \boldsymbol{Z}^+$ model—from increasing each of the BoW variables by a standard deviation.

Economic theory does not deliver sharp predictions for the signs of effects within the style-related variable groups $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. These elements may reflect targeted salesmanship and/or idiosyncratic communication style. Even if they purely reflected the former, it is unclear how sellers optimally trade off time to sale against profits—e.g., signaling high quality, charging a premium price, and waiting longer for a sale—without information on storage or opportunity costs. This is especially true when items can be costlessly relisted. By contrast, a cleaner prediction arguably applies to the BoW controls $\boldsymbol{Z}_3$, which proxy for information content. Holding style elements $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ fixed—e.g., layout, HTML tags, images, etc.—more information-rich listings should increase all Kindle consumers' expected surplus. The final column of Table 8 confirms that a SD increase in information content $(\boldsymbol{Z}_3)$ is associated with a predicted duration reduction of approximately 1.29 days and a 13.7 percentage point increase in sale probability. While these estimated effects are not necessarily causal, their magnitudes are plausible and their signs support an interpretation of our main results in which sellers purposefully design listings in anticipation of demand responses.

6. Conclusion

The eBay platform provides a nearly ideal environment for assessing the potential role of listing heterogeneity in generating price dispersion. We study a large sample of transactions involving physically identical products, but with listings that differ in the information sellers provide and the manner in which they present it. We find that these differences are quantitatively important: observed listing characteristics can account for roughly 48% of the raw price dispersion in our data. Thus, even for products that are physically identical, a substantial portion of apparent violations of the Law of One Price reflects differences in how sellers present their products to buyers.

Reaching this conclusion requires both richer data and more flexible empirical methods than are typically employed in studies of price dispersion. Using the basic dataset and conventional OLS methods explains only a relatively small share of price variation. Expanding the information set substantially increases predictive power, while allowing for nonlinearities and interactions through

machine-learning methods increases it further. The resulting predictive power does not appear to reflect arbitrary statistical correlations: listing characteristics also predict buyer responses, increasing sale probability and reducing listing duration, and the effects of information content generally have economically sensible magnitudes. Taken together, these results suggest that listing characteristics capture economically meaningful dimensions of differentiation that buyers value.

Our findings therefore qualify, rather than overturn, the conventional interpretation of price dispersion as evidence of market frictions. Even after accounting for a rich set of listing characteristics, nontrivial residual price dispersion remains, leaving considerable scope for search costs and other frictions to sustain price differences across sellers. At the same time, our results suggest that the magnitude of the Law-of-One-Price puzzle has been overstated when physically identical products are treated as economically homogeneous. Sellers have many partially substitutable ways of differentiating their listings, and accounting for this seller-induced differentiation substantially narrows the amount of price dispersion left for market frictions to explain.

Why might substantial price dispersion persist in an online market where search is seemingly so easy? One possible explanation is that the sheer scope of modern electronic markets creates new sources of frictions. Internet search technology solves the "needle-in-a-haystack" problem by allowing buyers to quickly sift out masses of irrelevant information. However, when a keyword search such as "Kindle tablet 8GB" returns a large number of relevant results, leaving the user with an entire stack of needles to sift through, it may still be costly or difficult to digest all information relevant to her needs (Reutskaja et al. 2020). Understanding these remaining search frictions, and how platform design might mitigate them, is an important direction for future research.

References

Aparicio, D., Z. Metzman, and R. Rigobon (2024): "The Pricing Strategies of Online Grocery Retailers," *Quantitative Marketing and Economics*, 22, 1–21.

Backus, M., J. Podwol, and H. Schneider (2014): "Search Costs and Equilibrium Price Dispersion in Auction Markets," *European Economic Review*, 71, 173–192.

Bailey, J. (1998): "Electronic Commerce: Prices and Consumer Issues for Three Products: Books, Compact Discs, and Software," *OECD Digital Economy Papers*, 32.

Bajari, P. and A. Hortaçsu (2004): "Economic Insights from Internet Auctions," *Journal of Economic Literature*, 42, 457–486.

Balducci, B. and D. Marinova (2018): "Unstructured Data in Marketing," *Journal of the Academy of Marketing Science*, 46, 557–590.

Baye, M. and J. Morgan (2001): "Information Gatekeepers on the Internet and the Competitiveness of Homogeneous Product Markets," *The American Economic Review*, 91, 454–474.

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

——— (2006): "Persistent Price Dispersion in Online Markets," in *The New Economy And Beyond: Past, Present And Future*, Edward Elgar Publishing.

Belloni, A. and V. Chernozhukov (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521 – 547.

Bhargava, H. K., O. Rubel, E. J. Altman, R. Arora, J. Boehnke, K. Daniels, T. Derdenger, B. Kirschner, D. LaFramboise, P. Loupos, et al. (2020): "Platform data strategy," *Marketing Letters*, 31, 323–334.

Blanchard, S. J., T. J. Noseworthy, E. Pancer, and M. Poole (2023): "Extraction of visual information to predict crowdfunding success," *Production and Operations Management*, 32, 4172–4189.

Bodoh-Creed, A., J. Boehnke, and B. Hickman (2021): "How Efficient Are Decentralized Auction Platforms?" *The Review of Economic Studies*, 88, 91–125.

Boehnke, J., E. Pontikes, and H. K. Bhargava (2025): "Decoding Unstructured Text: Enhancing LLM Classification Accuracy with Redundancy and Confidence," Working paper, UC Davis Graduate School of Management.

Bracci, A., J. Boehnke, A. ElBahrawy, N. Perra, A. Teytelboym, and A. Baronchelli (2022): "Macroscopic properties of buyer–seller networks in online marketplaces," *PNAS Nexus*, 1, pgac201.

Breiman, L. (2001): "Random Forests," *Machine Learning*, 45, 5–32.

Brynjolfsson, E. and M. Smith (2000): "Frictionless Commerce? A Comparison of Internet and Conventional Retailers," *Management Science*, 46, 563–585.

Burdett, K. and K. Judd (1983): "Equilibrium Price Dispersion," *Econometrica*, 51, 955–969.

Busse, M. R., A. Israeli, and F. Zettelmeyer (2017): "Repairing the Damage: The Effect of Price Knowledge and Gender on Auto Repair Price Quotes," *Journal of Marketing Research*, 54, 75–95.

Cabral, L. and A. Hortaçsu (2010): "The Dynamics of Seller Reputation: Evidence from eBay," *The Journal of Industrial Economics*, 58, 54–78.

Clay, K., R. Krishnan, and E. Wolff (2001): "Prices and price dispersion on the web: evidence from the online book industry," *The Journal of Industrial Economics*, 49, 521–539.

Clay, K., R. Krishnan, E. Wolff, and D. Fernandes (2002): "Retail strategies on the web: Price and non–price competition in the online book industry," *The Journal of Industrial Economics*, 50, 351–367.

Clerides, S., P. Courty, and Y. Ma (2023): "Store Expensiveness and Consumer Saving: Insights from a New Decomposition of Price Dispersion," *Quantitative Marketing and Economics*, 21, 65–94.

Coey, D., B. Larsen, and B. Platt (2020): "Discounts and Deadlines in Consumer Search," *American Economic Review*, 110, 3748–3785.

De Los Santos, B., A. Hortaçsu, and M. R. Wildenbeest (2017): "Search with Learning for Differentiated Products: Evidence from E-Commerce," *Journal of Business & Economic Statistics*, 35, 626–641.

Dew, R. (2025): "Adaptive preference measurement with unstructured data," *Management Science*, 71, 3996–4012.

Dinerstein, M., L. Einav, J. Levin, and N. Sundaresan (2018): "Consumer Price Search and Platform Design in Internet Commerce," *American Economic Review*, 108(7), 1820–1859.

Einav, L., C. Farronato, J. Levin, and N. Sundaresan (2018): "Auctions versus Posted Prices in Online Markets," *Journal of Political Economy*, 126, 178–215.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as Data," *Journal of Economic Literature*, 57, 535–574.

Hitsch, G., A. Hortaçsu, and X. Lin (2021): "Prices and Promotions in US Retail Markets," *Quantitative Marketing and Economics*, 19, 289–368.

Lewis, G. (2011): "Asymmetric Information, Adverse Selection, and Online Disclosure: The Case of eBay Motors," *American Economic Review*, 101, 1535–1546.

Lin, H. and M. R. Wildenbeest (2020): "Nonparametric Estimation of Search Costs for Differentiated Products: Evidence from Medigap," *Journal of Business & Economic Statistics*, 38, 754–770.

MacMinn, R. (1980): "Search and Market Equilibrium," *Journal of Political Economy*, 88, 308–327.

Malmendier, U. and Y. H. Lee (2011): "The Bidder's Curse," *American Economic Review*, 101, 749–787.

Monroe, K. and A. Lee (1999): "Remembering Versus Knowing: Issues in Buyers' Processing of Price Information," *Journal of the Academy of Marketing Science*, 27, 207–225.

Nosko, C. and S. Tadelis (2015): "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment," Mimeo.

Porter, M. (1980): "An Algorithm for Suffix Stripping," *Program*, 14, 130–137.

Quan, T. and K. R. Williams (2021): "Extracting Characteristics from Product Images and its Application to Demand Estimation," Working paper, Yale School of Management.

Reinganum, J. (1979): "A Simple Model of Equilibrium Price Dispersion," *Journal of Political Economy*, 87, 851–858.

Reutskaja, E., S. Iyengar, B. Fasolo, and R. Misuraca (2020): "Cognitive and Affective Consequences of Information and Choice Overload," in *Routledge Handbook of Bounded Rationality*, ed. by R. Viale, Routledge, chap. 43.

Rosenthal, R. (1980): "A Model in Which an Increase in the Number of Sellers Leads to a Higher Price," *Econometrica*, 48, 1575–1579.

Salop, S. and J. Stiglitz (1977): "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion," *The Review of Economic Studies*, 44, 493–510.

Schneider, H. (2016): "The Bidder's Curse: Comment," *American Economic Review*, 106, 1182–1194.

Sorensen, A. T. (2000): "Equilibrium Price Dispersion in Retail Markets for Prescription Drugs," *Journal of Political Economy*, 108, 833–862.

Varian, H. (1980): "A Model of Sales," *The American Economic Review*, 70, 651–659.

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wilde, L. and A. Schwartz (1979): "Equilibrium Comparison Shopping," *The Review of Economic Studies*, 46, 543–553.

Zhao, Q. and T. Hastie (2021): "Causal Interpretations of Black-Box Models," *Journal of Business & Economic Statistics*, 39, 272–281.

Online Supplement to accompany

Market Frictions Versus Seller Signaling: Explaining Deviations from the Law of One Price

by Aaron Bodoh-Creed, Jörn Boehnke, and Brent R. Hickman

Online Appendix A. Proofs

Proposition 2. *For any choice of* $(\gamma, \rho) \in (0,1)^2$ *and* $T < \infty$*, there exists* $\delta' < 1$ *such that* $p_M(t, \delta)$ *cannot exhibit* $(\gamma, \rho, T)-$*price dispersion if* $\delta \in (\delta', 1)$*.*

Proof. Suppose for a contradiction there exists $(\gamma, \rho) \in (0,1)^2$ and $T < \infty$ such that for any choice of δ' one can find $\delta \in (\delta', 1)$ where $p_M(t, \delta)$ exhibits $(\gamma, \rho, T)-$price dispersion. Condition (1) implies that some agent finds it optimal to choose $p_t(c; \delta) \leq \underline{\mathcal{P}}(\delta)$ at some point in the next T periods. However, by deviating to $\overline{\mathcal{P}}(\delta)$ for the next T periods, this agent could improve her profit by at least γ with probability at least ρ. A permanent deviation is optimal if the following condition holds:

$$\frac{\rho(p_t(c;\delta) + \gamma - c)}{1 - (1-\rho)\delta^T} > p_t(c;\delta) - c \tag{5}$$

The left hand side is a lower bound on the benefit of permanently deviating to a price of $\overline{\mathcal{P}}(\delta)$, and the right hand side is an upper bound on the payoff from choosing a price of $p_t(c; \delta)$. Equation (5) must hold for δ sufficiently large, which leads to a contradiction. □

Online Appendix B. Estimator Details and Robustness Checks

B.1. **Data Collection Procedures.** We briefly explain here our algorithm for collecting the Kindle listings, but the processes for the other products was very similar. For over two years our algorithm continuously iterated through all listings inside eBay's "iPads-Tablets-eBook-Readers" category and noted when a listing was posted to the eBay platform. For all listings containing the words "amazon fire" in the title, the algorithm downloaded the entire HTML source code once the listing was completed (regardless of whether it resulted in a sale or not).

We searched in the "iPads-Tablets-eBook-Readers" category using the keywords "Amazon Fire." This returned all active listings that included either of these words. We did not include the word "Kindle" because this tended to return listings for much older black and white Kindle e-readers, which were more numerous on the eBay site at the time. After experimenting with various keyword combinations for our scraping algorithm, we found that the search phrase "Amazon Fire" provided the best balance of specificity and breadth. This combination of search terms and keywords allowed us to capture virtually all listings of Kindle Fire HD tablets on eBay during our sample period.

B.2. **Random Forest.** A random forest is an ensemble estimator, or an average of a large collection of underlying regression tree models (Breiman 2001). Before describing how an ensemble of regression trees is constructed, let us describe the algorithm for creating a single regression tree. A

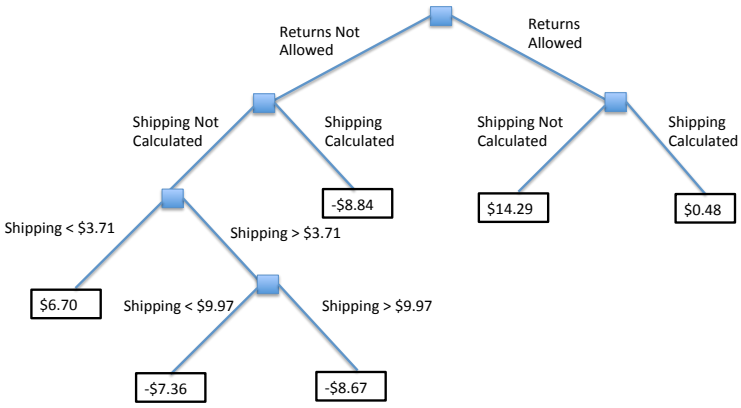


FIGURE 4. Simple Regression Tree

regression tree partitions the space of possible regressor values and assigns each element of the partition a prediction value equal to the average of the outcome variables in that bin of the partition. The prediction generated by a regression tree for a data point is simply the value assigned to the element of the partition containing that data point. A regression tree is a form of nearest-neighbor predictor where all data within the same partition is considered as being "near" each other.

The partition of the dataset 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. This simple example employs three of our regressors — "Shipping is Calculated," "Returns Allowed," and "Shipping" — as splitting variables whose values are used to partition the space of data. Beginning at the root (top of the diagram), each node of the tree represents a splitting of the sample into two (potentially unequally sized) subsets. Each leaf of the tree provides a prediction of the de-trended price for listings within the data at that leaf. Once a tree is grown, it exists in the form of a set of variable and cutoff choices that define the splits in the tree and establish a complete partition of the space of possible values for the regressors, along with a predicted value within each bin of the partition.

Now we formally describe the algorithm for growing a tree. Let $\mathbf{V}_i$ denote the set of variables used to represent listing $i = 1, 2, \ldots, I$, including $\mathbf{X}_i$ and/or $\mathbf{Z}_i$, and denote the dataset as $\mathcal{D} = (\boldsymbol{V}, \boldsymbol{Y})$ where $\boldsymbol{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_I]^\top$ denotes a full matrix of regressor realizations and $\boldsymbol{Y} = [y_1, \ldots, y_I]^\top$ is a vector of prices for each listing. Our random forest is grown using *bootstrapped aggregation*, also known as "bagging." To grow a single tree in the ensemble, a bootstrapped sample $\mathcal{B}$ is generated from $\mathcal{D}$ (equal in size to $\mathcal{D}$). A certain fraction of the explanatory variables are randomly chosen to be used as *splitting variables*, and the choice of a one-third fraction is a commonly used rule of thumb for random forest algorithms.

The root of the tree is a split of the bootstrap dataset $\mathcal{B}$ into subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ such that $\mathcal{B}_1 \cap \mathcal{B}_2 = \varnothing$ and $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B}$. These splits have the form $\mathcal{B}_1 = \{(\mathbf{V}_i, y_i) \in \mathcal{B} : \mathbf{V}_{i,j} \leq k\}$ where $\mathbf{V}_{i,j}$ in the i^{th} listing's realization of the j^{th} regressor, which must be one of the splitting variables, and

	R^2 VERSION	
	Out-of-Sample	**In-Sample**
DATASET **Basic**	0.1976	0.3074
Full	0.4151	0.9144

TABLE 9. Random Forest Predictive Power

k is a real number defining a *split point*. Now, let $\mathcal{Y}_1$ and $\mathcal{Y}_2$ denote the sets of price realizations contained in $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively. The *split criterion* we use minimizes the following function at each node: $|\mathcal{Y}_1| Var(\mathcal{Y}_1) + |\mathcal{Y}_2| Var(\mathcal{Y}_2)$, where $|\mathcal{Y}_l|$ refers to the cardinality of set $l = 1, 2$. The algorithm divides $\mathcal{B}$ to form $\mathcal{B}_1$ and $\mathcal{B}_2$ by choosing the splitting variable j and cutoff k that together minimize the split criterion. It then recursively applies this splitting process on subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ until an entire tree is formed. The intuition is that since the predictions are the same within a leaf, the algorithm divides the dataset in such a way that the prices of the listings within each leaf are as similar as possible. Note that a splitting variable can appear multiple times in the same tree. The algorithm terminates when the subsets contain a single data point.

A prediction for a generic realization $\mathbf{V}_i$ is the value at the leaf to which $\mathbf{V}_i$ belongs. Since each regression tree partitions the full support of the regressors, the tree generates a prediction for any possible realization of $\mathbf{V}_i$. The predictions of a regression tree will be perfect for the bootstrap dataset used to estimate the tree (i.e., $\mathcal{B}$), which is the source of the random forest's overfitting. In the full random forest, not all data are used to estimate any given tree since we estimate each one using a bootstrap sample. Moreover, we compute R^2_{out} using cross validation, which means that the data used to estimate the random forest is distinct from the holdout dataset used to evaluate the tree's predictive power.

Once many such trees have been grown using many different bootstrapped samples, the prediction of the random forest is the average of the predictions of the trees in the forest. Unless otherwise stated, our random forests include 1,000 trees. 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 reduce the correlation of the trees in the ensemble, which increases the estimator accuracy out-of-sample. Using Breiman (2001) terminology, the decorrelation of the trees reduces the generalization error.

To illustrate the importance of using out-of-sample metrics of predictive power for the random forest, Table 9 describes the in- and out-of-sample fit of the estimated trees. As one can see, the random forest overfits in-sample even when using the small number of regressors in the basic dataset. The problem of overfitting is, unsurprisingly, more pronounced when using the much larger number of regressors in the full dataset.

B.3. **Out-of-Sample R^2 with Cross-Validation: Details.** The procedure for computing our cross-validated out-of-sample R^2 randomly partitions our data into 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 union of the remaining nine subsets of data. We then compute the sum of squared errors, SSE_k, and total sum of squares, TSS_k, in the validation set $\mathcal{F}_k$ using the estimated model. For a fixed partition of the data, the out-of-sample R^2 statistic is $R^2_{out} = 1 - \frac{\sum_{k=1}^{10} SSE_k}{\sum_{k=1}^{10} TSS_k} = 1 - \frac{\sum_{k=1}^{10} \sum_{y_i \in \mathcal{F}_k} (y_i - \widehat{y}_i)^2}{\sum_{k=1}^{10} \sum_{y_i \in \mathcal{F}_k} (y_i - \overline{y}_i)^2}$, where y_i is the price of the i^{th} listing in our data, $\widehat{y}_i$ is a predicted price for that listing, and $\overline{y}_i$ is the average price in the validation set.[16] We repeat the cross-validation process with 100 random 10-fold partitions of the data, and average the resulting R^2_{out} to compute the statistics we report.[17]

B.4. **LASSO.** One might wonder whether a more flexible linear-in-variables model estimated on the full dataset might perform as well as the random forest. To test this conjecture, we estimated a linear model with higher-order terms and interactions of the variables in the full Kindle dataset up to third order. Once we remove redundant regressors, we are left with a cubic complete polynomial with 6,463 terms. We use LASSO to choose which of the regressors to include in our model.[18]

Denote a single data point as $(\mathbf{V}_i, y_i)$ where $\mathbf{V}_i$ are the regressors we collected (including higher-order terms and interactions) and y_i is the de-trended price of the listing. We can describe the LASSO algorithm through the following optimization problem:

$$\min_{\boldsymbol{\alpha}} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{V}_i' \boldsymbol{\alpha}\right)^2 + \lambda \left\|\boldsymbol{\alpha}\right\|_1, \tag{6}$$

where $\boldsymbol{\alpha}$ contains the regressor coefficients and λ is the LASSO penalty parameter. Since $\|\boldsymbol{\alpha}\|_1$ enters the objective function linearly, the solution to Equation 6 sets $\alpha_j = 0$ for the regressors with the least amount of predictive power, which amounts to removing that regressor from the model. The 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 (6) above as the model selected by the LASSO algorithm.

LASSO algorithms typically solve equation (6) for a range of values of λ, and each possible λ is evaluated through a cross-validation process that penalizes overfitting. Since the cross-validation procedure is slightly different than our algorithm for computing out-of-sample R^2 values, we provide an overview here for completeness. The degree of overfitting is assessed through 10-fold cross

[16] We also computed the *average out-of-sample* R^2, $\overline{R^2_{out}} = 1 - \frac{1}{10} \sum_{k=1}^{10} \frac{SSE_k}{TSS_k}$. The resulting values differed from R^2_{out} by less than 0.5% in the Kindle data. Breiman (2001) proposed using out-of-bag measures to assess goodness of fit of a random forest. We prefer our R^2_{out} metric since it can be computed for the OLS model as well.

[17] The *in-sample* R^2 statistic typically reported by economics and marketing researchers, is computed by estimating the model on the full dataset, forming a prediction for the price of each listing in the same dataset, and computing the R^2 based on these in-sample predictions. This measure reflects a high degree of "over-fitting" (i.e., artificially inflated R^2 values) when the random forest both selects and estimates a flexible model on the same data. Table 9 in Online Appendix B.2 reports both in- and out-of-sample R^2 for the random forest to demonstrate this problem.

[18] Appendix C includes a LASSO analysis of the Surface data with qualitatively similar results.

validation of each value of λ. We first divide the dataset 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 solve equation (6) for a given λ on the union of the remaining nine subsets of data. We then compute the sum of squared errors, $SSE_k(\lambda)$, in the validation set $\mathcal{F}_k$. The sum of squared errors 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 variance of $SSE_1(\lambda), ..., SSE_{10}(\lambda)$. The computation of $SE(\lambda)$ is the primary difference between the LASSO cross validation procedure and the algorithm used to compute R^2_{out}.

A common heuristic for choosing λ is to select the largest value such that $SSE(\lambda) \leq SSE(\lambda_{Min}) + SE(\lambda_{Min})$, where λ_{Min} minimizes $SSE(\lambda)$.[19] We find that the LASSO model estimated on the full dataset has an out-of-sample R^2 equal to 0.330. The primary takeaway is that using a more flexible linear-in-parameters model to predict prices does increase explanatory power, but there is still a large gap between the explanatory power of our random forest predictor and a very flexible OLS model.

B.5. **Other Analysis Methods.** We tried a variety of other ML methods to explore how much price variation we could explain, but we found that the performance was comparable to or worse than our simpler random forest model. We experimented with individual neural networks, but we found that even moderately complex neural networks (e.g., those with two hidden layers) severely overfit the data. We also 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. 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 fits the residuals from the previous tree. Again, none of these methods had more predictive power than the random forest.

There are other methods one can use to reduce the dimensionality of text data. For example, one could use topic modeling techniques such as latent Dirichlet allocation (LDA) to algorithmically define topics and ascribe a metric of each topic's influence on the text in each listing.We experimented with this approach, but found that the topics were not strongly associated with any particular words. This is not terribly surprising since LDA techniques are known to work relatively poorly when the text in each data point are short, which is the case in our data since the median

[19] One might assume that it is obviously optimal to choose the value of λ that minimizes $SSE(\circ)$. Typically this is not recommended since it can result in overfitting. The above rule of thumb generally allows for a great deal of flexibility while also promoting out-of-sample predictive power at the same time.

number of words in the description is only 27. Simply put, one needs larger samples of text to be able to attribute topics to any given listing's text.

B.6. **Sellers with Multiple Listings.** If a seller posts an identical listing twice, then one copy of the listing could appear in the set of data used to estimate the random forest and the second copy could appear in the set of data used to compute the out-of-sample predictive power. If that listing is in the bootstrap sample of data used to estimate a regression tree, then that tree will make a perfect prediction of the price for the copy of the listing in the validation dataset. This could cause us to overestimate the predictive power of our models. Of course, this problem would apply to all methodologies we used in our study, so its effect on the relative comparisons between models is unclear. Regardless, we looked into this problem as a check on robustness.

When we examined a sample of the listings posted by sellers with multiple listings, we found that the listings tend to vary significantly. The titles and descriptions often have different lengths, the number and sizes of the pictures in the title and the description vary, and the style of the text changes. In addition, sellers make price adjustments between listings. In order to err on the side of caution, we repeated our analysis after eliminating duplicate listings from our dataset, and we refer to our dataset with the duplicates removed as the "de-duped" dataset.

We defined "duplicate listing" coarsely so as to cast a broad net in identifying potentially problematic data points. As a first pass, we deemed two listings as duplicates if they matched on all of the variables other than "Relisted" and the log seller score rounded up to the nearest 0.1. We found that duplicate listings accounted for only 11.8% of our dataset, leaving us with 1145 listings after eliminating them. With this reduced dataset, the R^2_{out} for the random forest dropped to 0.3622. The reduction in the R^2_{out} is due to the joint effects of using a smaller dataset and mitigating the potential contamination of the validation set. To get a sense for the effect of the reduction in the size of our data alone, we eliminated 11.8% of our full sample at random. Importantly, the randomly selected sample can contain duplicate listings. We found that randomly eliminating data resulted in an R^2_{out} for the random forest equal to 0.3982. In summary, since the random elimination of data and selectively eliminating redundant data points yielded a similar R^2_{out} value to, we conclude that our results are not significantly inflated by contamination of the cross-validation sets.

We also experimented with increasing the size of the leaves of our random forest from 1 to as many as 10 listings. Larger leaves mean that none of the trees will make a perfect prediction of the price of any listing, which helps alleviate any issues which may arise from the presence of identical listings in the estimation and validation sets. When we include at least 5 listings in each leaf the R^2_{out} drops only slightly to 0.397, and including 10 listings in each leaf pushes the R^2_{out} to 0.369. Some of this modest drop may be due to the lessened effect of the contamination of the validation set, but some of it is due to a decrease in model flexibility caused by the larger leaves.

B.7. **Testing for Search-Algorithm Manipulation.** We seek to separate listing heterogeneity from other sources of price variation such as search frictions. If some listings are served to buyers first, based *predictably* on listing features and are thus more likely to sell, these sellers might charge higher prices than listings appearing later in search results.[20] Thus, our separation of listing heterogeneity and search frictions as sources of price dispersion is threatened if sellers can manipulate listing features we collect in order to game eBay's search algorithm. At the time our data were collected, eBay used its proprietary "Best Match" algorithm—designed with the primary objective of *preventing* search engine optimization—for ordering search results.

We computed several robustness checks to assess the potential impact of algorithm manipulation in our data. First, we generated a separate BoW for the title within the Kindles dataset; this allows us to isolate subsets of listings where particular words appear in the title. A challenge is to find common words that (*i*) buyers plausibly search for a large fraction of the time and (*ii*) are present in enough (but not all) listings to afford a reasonable application of ML techniques. When we restrict our sample to listings that include the word "new," we are left with 789 listings. We considered other words such as "sealed" (123 listings), "factory" (43 listings), or "screen" (25 listings), but the resulting datasets were small. If sellers were including these extra terms in their title text to manipulate the search algorithm, then we would expect words like these to appear more often. Others, such as "GB", were used in nearly all listings, in this case 1,170 of the 1,298 listings in our Kindles sample. The standard deviation of price in the restricted sample of listings with "new" in the title is $27.70, which is 96% of the standard deviation of the price variation in the entire sample, so there remains significant price dispersion to potentially explain.

Following Backus et al. (2014), by restricting attention to the 789 listings that include "new" in the title, we examine listings that were served to buyers in an order that is independent of (or at least less correlated with) the features we collect. If our predictive power is driven by search engine optimization, we would expect that controlling for the word "new" in the title would reduce our R^2_{out} measure of predictive power. We also analyzed a random subsample of 789 listings from the full dataset to obtain a benchmark for the R^2 we ought to expect given our prior results and the reduced size of the selected sample of listings with "new" in the title. The results of our random forest are displayed in Table 10.

		R^2 VERSION	
		Out-of-Sample	**In-Sample**
DATASET	**Listing Titles using "New"**	0.3743	0.6493
	Random Subsample	0.3318	0.6252

TABLE 10. Random Forest Predictive Power for Listings with "New" in the Title

[20] E.g., if title text length is predictive of price because long titles include many potential search terms, then our analysis would attribute price variation to listing heterogeneity when it is really due to the presence of search frictions (and sellers' power to manipulate frictions).

		R^2 VERSION		
		Number of Listings	**Out-of-Sample**	**In-Sample**
DATASET	**Listings With Long Titles**	660	0.409	0.912
	Listings With Short Titles	638	0.357	0.907
	Random Subsample	660	0.345	0.910

TABLE 11. Random Forest Predictive Power: Listings w/ Long vs Short Titles

After conditioning on use of the word "new" in the title, we still find R^2_{out} values of over 0.37, as compared to the random subsample where the R^2_{out} is 0.33. Not only does predictability remain comparable across the two subsamples, but in the subsample with "new" in the title it is actually slightly higher. This is the opposite of what one would expect if our observed listing features captured common search engine manipulations, and spuriously inflated our ability to predict non-friction price dispersion.

We also execute a second test (also following Backus et al. (2014)) based on the number of words in the title. The idea is that titles with more words are more likely to contain words that buyers use in search queries and thus appear earlier in the search results provided by eBay's Best Match algorithm. The core idea is similar as before: suppose the price predictability we infer from our analysis is driven by sellers' ability to push their listing towards the top of the search results (and hence successfully sell at a higher price relative to those that appear lower in the list) by manipulating the features we collect (e.g., using the word "new" in the title or using a title with many words). Moreover, suppose that this is the only channel through which features of the listing affect price, meaning the price dispersion is generated by search frictions alone. In this case, listings that appear near the same point in the list of search results (e.g., other listings with titles of the same length) ought to have prices that are less predictable than the prices of the total sample.

Our analysis breaks the listings into two sets. The first set includes 660 listings having titles with a number of words that is strictly greater than the median in the full sample, which ought to appear near the top of the search results if search algorithm manipulation is happening. The second set of 638 listings includes those with a number of words in the title that is strictly less than the median across the full sample. If the argument in the preceding paragraph is correct, then both sets of listings should exhibit less price predictability (based on observed covariates) than a random subsample (chosen from the full set of listings) of the same size. Also note that there is significant price dispersion within both the above-median and below-median subsamples. The listings with long titles have price dispersion equal to 21.6% of the full-sample mean, while the listings with short titles have price dispersion equal to 21.2% of the full-sample mean.

To construct our comparison set, we drew 10 random subsamples of 660 listings each from the full dataset, ran our analysis, and computed the R^2_{out} within each subsample. The R^2_{out} for the

		R^2 VERSION	
		Out-of-Sample	**In-Sample**
MODEL TYPE	**OLS**	0.045	0.205
	Random Forest	0.271	0.894

TABLE 12. Predictive Power Restricted to Listings that Result in a Sale: Kindle Tablets

subsets of listings with long and short titles are displayed in Table 11, and the comparison result is the average R^2_{out} across the 10 random subsamples we drew.

The R^2_{out} in the samples of long and short title listings are both somewhat higher than the R^2_{out} for the random subsample, in contrast to what one would expect if search algorithm manipulation was common in the data-generating process and driving our results.[21] We conclude from this that the heterogeneity we detect does not reflect the sellers attempting to use search engine optimization to alleviate search frictions. This is consistent with our central assumption that our set of observable controls is plausibly orthogonal to price variation driven by search frictions.

B.8. **Focusing on Sold Listings.** As pointed out by Baye et al. (2004), price dispersion is compatible with Bertrand competition among firms with heterogeneous costs. In such equilibria, only the lowest-priced firm sells, and only the two lowest-cost firms have incentives to price aggressively; higher-cost firms can set arbitrary prices but never transact. This raises a natural question: should listings that never result in a sale be treated as relevant market participants? As a robustness check, we therefore reassess the predictive performance of the OLS and random forest models after restricting attention to the 965 listings that result in a sale. These listings are $31.06 cheaper on average than those that don't sell, but substantial price dispersion remains—the coefficient of variation is still 0.19—and the distributions for sold and unsold listings overlap substantially.

Ex ante, one should not expect out-of-sample R^2 values as high as those in Table 9 since the sold subsample is roughly 3/4 of the full dataset, and a price standard deviation among sold listings is roughly 90% of the overall standard deviation. Moreover, this exercise is conservative: some currently unsold listings would plausibly sell if re-posted with longer durations. Nevertheless, Table 12 shows that the loss in predictive power is modest, while the performance gap between the random forest and OLS models remains pronounced. Since the sold-listing subsample exhibits substantial and predictable price dispersion, we conclude that our main results are not driven by the presence of implausibly high-priced listings that never had a realistic chance of selling.

ONLINE APPENDIX C. EXPLORING EXTERNAL VALIDITY: SURFACE & OTHER PRODUCTS

The Microsoft Surface data provide an additional check using a product sampled during the same period as the Kindle data. The Surface is a tablet computer sufficiently powerful to serve

[21] Note that the results of Table 11 are based on random forests grown with substantially less data, and are therefore noisier than our baseline results using the full sample.

	DATASET	
MODEL TYPE	**Basic**	**Full**
OLS	0.068	-0.012
Random Forest	0.161	0.438
LASSO	0.064	0.234

TABLE 13. Surface R^2_{out}

as a replacement for a laptop, making it substantially more expensive than the Kindle. The 922 eBay listings for Surfaces that we collected have a mean price of $968 with a standard deviation of $203. Thus, although Surface prices are roughly six times higher than Kindle prices in our dataset, the dispersion of detrended prices as a fraction of the raw mean is both sizeable and similar across the two products, at roughly 21%. As reported in Table 6, the Surface also yields comparable predictive power under the random forest model. Table 13 supplements those results with the LASSO specification used in our Kindle robustness analysis. As with the Kindle data, allowing a flexible linear-in-parameters model with higher-order interactions improves substantially upon OLS, but still falls well short of the random forest: R^2_{out} rises from -0.012 under OLS to 0.234 under LASSO, compared with 0.438 under the random forest. Thus, the Surface results reinforce the conclusion that the predictive gains from our preferred specification reflect more than simply adding higher-order interactions to a linear model.

		TITLE		DESCRIPTION	
Product	**R^2_{out} Random Forest**	**Volume of Design Elements**	**Style**	**Volume of Design Elements**	**Style**
Kindle	0.415	0.172	0.198	0.159	0.367
Surface	0.438	0.182	0.200	0.132	0.178
Apple AirPods	0.672	0.450	0.519	0.463	0.354
Apple iPod Touch 6, 32GB	0.516	0.258	0.253	0.307	0.293
Apple Watch	0.338	0.168	0.279	0.203	0.215
Beats Solo 3	0.383	0.279	0.273	0.209	0.248
Call of Duty, PlayStation 4	0.264	0.127	0.128	0.104	0.151
Destiny 2, PlayStation 4	0.585	0.358	0.415	0.278	0.321
Dyson Hairdryer	0.422	0.214	0.241	0.223	0.248
Fitbit Charge 2	0.569	0.393	0.392	0.317	0.396
iRobot 690	0.632	0.310	0.308	0.302	0.329
Xbox One Kinekt	0.371	0.284	0.281	0.137	0.191
KitchenAid Mixer KSM150	0.485	0.236	0.294	0.307	0.294
Nintendo SNES Classic	0.582	0.432	0.477	0.270	0.427

TABLE 14. Predicting the Price Variation of Other Products, Partial R^2_{out}

Tables 14 and 15 extend the decomposition exercises of Section 4.2 across all 14 products. Table 14 shows that both the volume and style of seller-customized content generally have substantial predictive power in both the title and description sections, although their relative importance varies across products. Table 15 likewise shows substantial heterogeneity in which portions of the listing

Product	R^2_{out} RF	VARIABLE IMPORTANCE OF Title	Description	Bag of Words
Kindle	0.415	0.026	0.145	0.322
Surface	0.438	0.009	0.049	0.525
Apple AirPods	0.672	0.369	0.233	0.516
Apple iPod Touch 6, 32GB	0.516	0.106	0.441	0.179
Apple Watch	0.338	0.117	0.379	0.072
Beats Solo 3	0.383	0.298	0.246	0.130
Call of Duty, PlayStation 4	0.264	0.122	0.255	0.048
Destiny 2, PlayStation 4	0.585	0.279	0.165	0.420
Dyson Hairdryer	0.422	0.136	0.311	0.167
Fitbit Charge 2	0.569	0.365	0.448	0.174
iRobot 690	0.632	0.135	0.315	0.276
Xbox One Kinekt	0.371	0.268	0.220	0.123
KitchenAid Mixer KSM150	0.485	0.142	0.394	0.138
Nintendo SNES Classic	0.582	0.325	0.289	0.497

TABLE 15. Predicting the Price Variation of Other Products, Variable Importance

the random forest relies upon most heavily: for 12 of 14 products, the description is nearly as important as or more important than the title, while Bag-of-Words variables are the most important group for 6 of 14 products. Thus, the broader product sample reinforces the main conclusion from Section 4.2: predictive power is not driven by a single universal listing feature or channel, but by different combinations of seller-provided information across products.

ONLINE APPENDIX D. HONEST TREE FOREST IMPLEMENTATION

The first step of building an honest model forest predictor is to divide our dataset $\mathcal{D}$ into two equally sized subsets, a ($\mathcal{S}$)election set and an ($\mathcal{E}$)stimation set. Our separation of the data used for model selection and model estimation ensures that our trees have the "honest" property. Once we have used the ($\mathcal{S}$)election set to define the structure of the trees, the ($\mathcal{E}$)stimation set is being used to estimate a set of OLS models. If the OLS models are consistent individually, an issue we discuss later, then our forest estimator will have the usual desired asymptotic properties (e.g., consistency, normality). The OLS model at each leaf was quite simple and included only the regressor of interest and a constant term.[22]

To determine the structure of the trees in our model forest, we apply the algorithm described in Section B.2. Each tree is grown from a Bootstrap sample of size $I/2 = 649$ drawn from dataset $\mathcal{S}$ (i.e., half of the total sample size), and the splitting points of each tree are determined using the variance minimization criterion. However, there are two major differences from the algorithm in Section B.2. First, we do not allow the algorithm to create a split point on a variable if one of the resulting leaves has fewer than 30 data points; this ensures enough data at each leaf to estimate an

[22]Including all regressors of interest in the model at each leaf did not substantively change the results.

OLS model.[23] Second, we randomly choose a third of our variables to use as splitting variables, but we do not allow splits that are based on the variable we include in the OLS models we estimate at the leaves (i.e., the regressor of interest for a particular forest). Each forest we estimated contained 500 trees in total.

Once the split points of each tree have been computed, we move onto the estimation step using dataset $\mathcal{E}$. Each tree within the forest is estimated in a three step process. First, we generate a 649-element bootstrap sample from $\mathcal{E}$. Second, we determine which leaf contains each element of the bootstrap sample. Third, we perform an OLS regression on the data at each leaf.[24]

We require that the OLS estimator be asymptotically consistent as the size of the dataset grows. If we were to collect more data, we could simultaneously define more leaves (increasing our ability to detect heterogeneous marginal 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, the regressor realizations within each leaf will become more similar, and the decrease in regressor variability could make our OLS estimates less precise. In the extreme, if we add leaves too quickly as our dataset grows, our OLS models may not be consistent.

To solve this problem, we do not allow the trees to split on the variables we include in our OLS regressions. Asymptotically, our trees will have an infinite set of data points at each leaf, and the data in each leaf will be similar across the variables we allow the tree to use to define splitting points. Since the regressors we include in the model at each leaf are not used to define splitting points, there will be enough variation in those variables to accurately estimate the model at each leaf. Each tree in our random forest provides a single estimate of the marginal effect of each regressor at each data point (i.e., listing), so a random forest of 500 trees provides 500 estimates of the marginal effect at each data point. We aggregate the estimates of the trees for each datum by averaging the estimates of the trees.

We compute 95% confidence intervals for the marginal effects of the listings at each decile by forming 1000 bootstrap samples of our estimation set and re-estimating the trees with each bootstrap sample. We do not bootstrap the tree growing process. We provide confidence bounds of the value realized for the listing at each decile of the distribution of marginal effects. We do it this way because our later discussion will focus on differences between the marginal effects at different quantile ranks of the distribution. To clarify, consider an alternative statistic that we could have generated, but did not. We could have computed the CDF of the marginal effects for each bootstrap run and then

[23] Using fewer, larger leaves would have resulted in more accurate estimates within each leaf, but the smaller number of leaves would have made detecting the heterogeneous marginal effects more difficult.

[24] It is possible that one of the leaves will be empty. To account for this possibility, we execute a pruning algorithm. If a leaf is found to be empty, we eliminate the split that created the leaf, merging the data contained in each of the leaves created by the split. If the new leaf resulting from the merger is also empty, we recursively apply our algorithm until a nonempty leaf is formed.

presented the 50^{th} and 950^{th} largest values at each decile. The listing that occupied each quantile rank would vary between bootstrap runs. In addition, the standard errors would appear deceptively small.

Online Appendix E. Webpage Description

The webpage for a listing that a potential buyer sees once he or she clicks on an item in the search results page has a format with three sections. Figure 5 displays an example of the first section of the listing, which we refer to as the *title section*. The title section includes a brief text description of the item written by the seller and one or more photos that are provided by the seller. The title section also has a standardized format that includes the price chosen by the seller, shipping information, the item's condition (e.g.,"New"), and a seller reputation score. The seller reputation score is equal to the total number of positive buyer ratings minus the number of negative ratings.

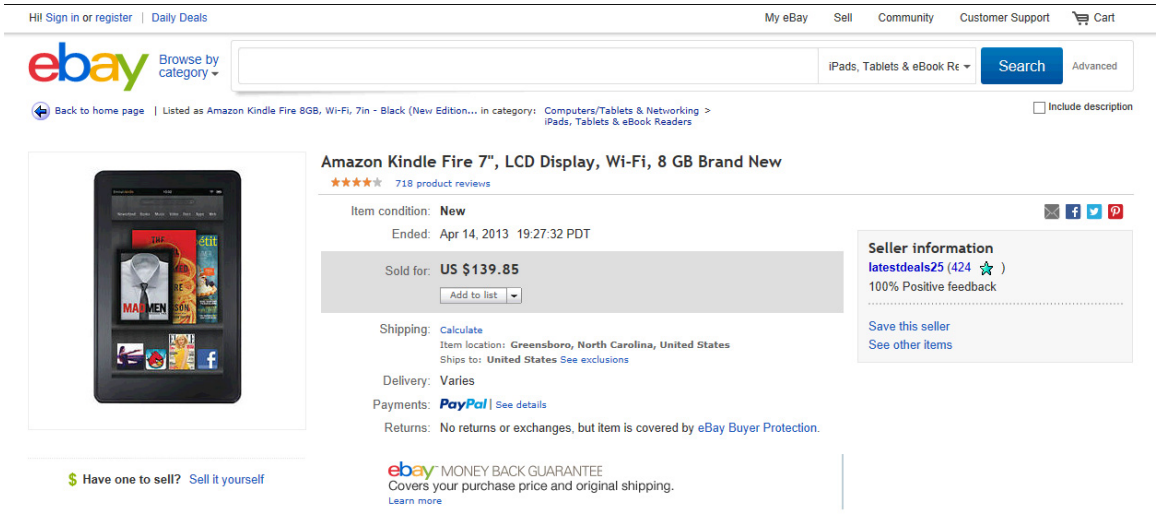
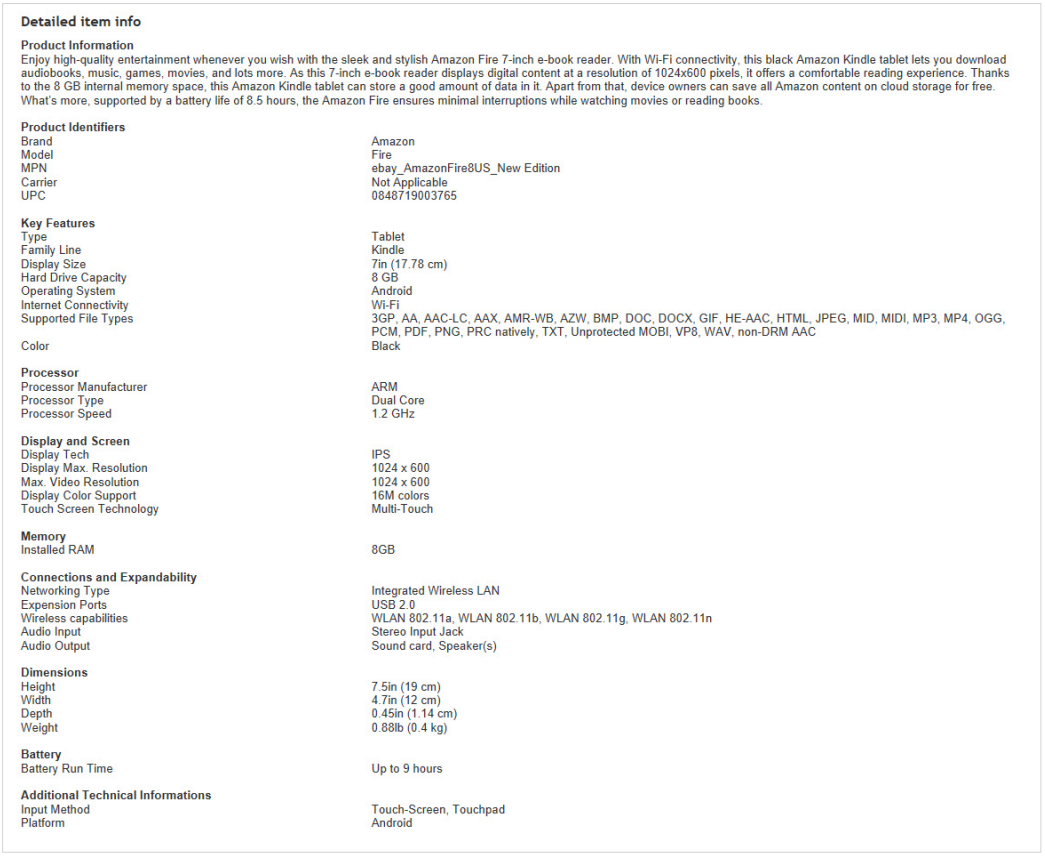


Figure 5. eBay's Standardized Listing Information

The second section of each listing is a box that provides a standardized, uniformly formatted set of information about the product that is provided by eBay, an example of which is depicted in Figure 6. The box describes the precise definition of the condition of the product and detailed technical specifications of the product (e.g., CPU processor speed). Since the section is standardized across our sample, it does not play a role in our analysis. However, the existence of this section shows that there can be little ambiguity about either the product being sold or the product's condition.

A fairly elaborate example of the third section of each listing, which we refer to as the *description section*, is provided in Figure 7. This third section is created entirely by the seller and is optional, with about 16% of listings in our sample having no content in the description. The seller has the ability to provide a large amount of text and images, and the text can be formatted using HTML



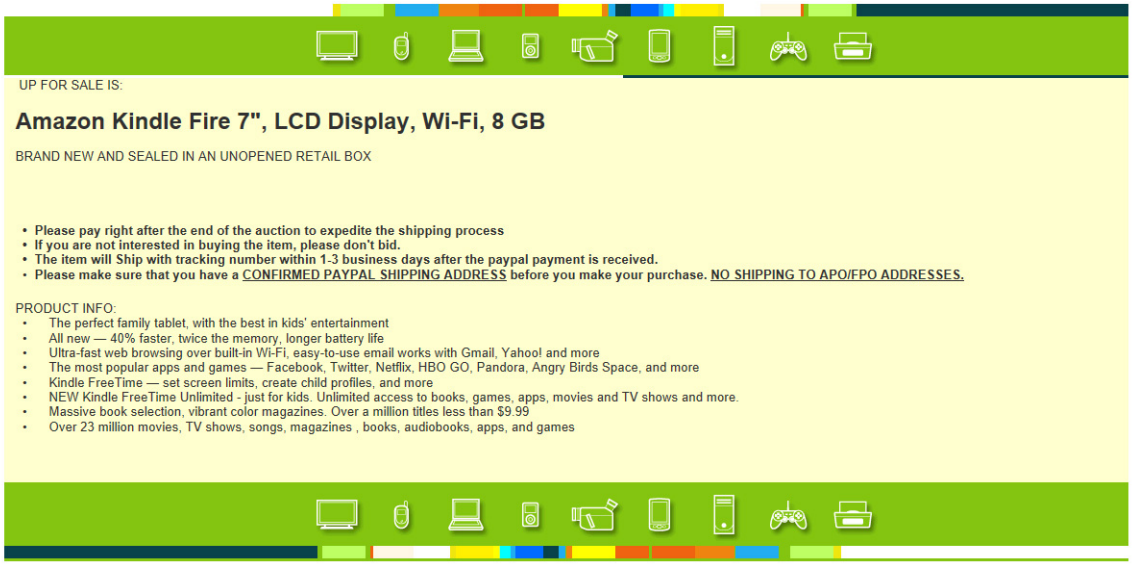
Detailed item info

Product Information

Enjoy high-quality entertainment whenever you wish with the sleek and stylish Amazon Fire 7-inch e-book reader. With Wi-Fi connectivity, this black Amazon Kindle tablet lets you download audiobooks, music, games, movies, and lots more. As this 7-inch e-book reader displays digital content at a resolution of 1024x600 pixels, it offers a comfortable reading experience. Thanks to the 8 GB internal memory space, this Amazon Kindle tablet can store a good amount of data in it. Apart from that, device owners can save all Amazon content on cloud storage for free. What's more, supported by a battery life of 8.5 hours, the Amazon Fire ensures minimal interruptions while watching movies or reading books.

Product Identifiers	
Brand	Amazon
Model	Fire
MPN	ebay_AmazonFire8US_New Edition
Carrier	Not Applicable
UPC	0848719003765
Key Features	
Type	Tablet
Family Line	Kindle
Display Size	7in (17.78 cm)
Hard Drive Capacity	8 GB
Operating System	Android
Internet Connectivity	Wi-Fi
Supported File Types	3GP, AA, AAC-LC, AAX, AMR-WB, AZW, BMP, DOC, DOCX, GIF, HE-AAC, HTML, JPEG, MID, MIDI, MP3, MP4, OGG, PCM, PDF, PNG, PRC natively, TXT, Unprotected MOBI, VP8, WAV, non-DRM AAC
Color	Black
Processor	
Processor Manufacturer	ARM
Processor Type	Dual Core
Processor Speed	1.2 GHz
Display and Screen	
Display Tech	IPS
Display Max. Resolution	1024 x 600
Max. Video Resolution	1024 x 600
Display Color Support	16M colors
Touch Screen Technology	Multi-Touch
Memory	
Installed RAM	8GB
Connections and Expandability	
Networking Type	Integrated Wireless LAN
Expension Ports	USB 2.0
Wireless capabilities	WLAN 802.11a, WLAN 802.11b, WLAN 802.11g, WLAN 802.11n
Audio Input	Stereo Input Jack
Audio Output	Sound card, Speaker(s)
Dimensions	
Height	7.5in (19 cm)
Width	4.7in (12 cm)
Depth	0.45in (1.14 cm)
Weight	0.88lb (0.4 kg)
Battery	
Battery Run Time	Up to 9 hours
Additional Technical Informations	
Input Method	Touch-Screen, Touchpad
Platform	Android

FIGURE 6. eBay's Standardized Description



UP FOR SALE IS:

Amazon Kindle Fire 7", LCD Display, Wi-Fi, 8 GB

BRAND NEW AND SEALED IN AN UNOPENED RETAIL BOX

- **Please pay right after the end of the auction to expedite the shipping process**
- **If you are not interested in buying the item, please don't bid.**
- **The item will Ship with tracking number within 1-3 business days after the paypal payment is received.**
- **Please make sure that you have a CONFIRMED PAYPAL SHIPPING ADDRESS before you make your purchase. NO SHIPPING TO APO/FPO ADDRESSES.**

PRODUCT INFO:

- The perfect family tablet, with the best in kids' entertainment
- All new — 40% faster, twice the memory, longer battery life
- Ultra-fast web browsing over built-in Wi-Fi, easy-to-use email works with Gmail, Yahoo! and more
- The most popular apps and games — Facebook, Twitter, Netflix, HBO GO, Pandora, Angry Birds Space, and more
- Kindle FreeTime — set screen limits, create child profiles, and more
- NEW Kindle FreeTime Unlimited - just for kids. Unlimited access to books, games, apps, movies and TV shows and more.
- Massive book selection, vibrant color magazines. Over a million titles less than $9.99
- Over 23 million movies, TV shows, songs, magazines , books, audiobooks, apps, and games

FIGURE 7. Seller's Customized Description

tags (e.g., bolded text). The challenge for analyzing the information in the description is condensing the many features of the text and images into data amenable to statistical analysis.

Online Appendix F. 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 count of the number of listings in which the word appears. For clarity, the table includes only one representative example of each word stem. It is important to note that the following table pertains only to seller-customized content used as controls in our empirical analysis.

Stem	Count	Stem	Count	Stem	Count	Stem	Count	Stem	Count	Stem	Count
gb	1201	hour	199	good	146	require	98	cover	66	insure	38
new	1139	case	197	faster	145	accept	97	number	61	paid	37
black	978	game	197	b	141	test	97	delivery	60	leather	35
box	662	feature	194	mail	140	price	95	description	60	part	35
brand	635	store	193	paypal	139	accessory	94	must	60	combine	34
model	521	thank	187	return	137	best	94	except	58	receipt	34
latest	464	app	184	popular	136	state	94	love	58	fedex	33
ship	453	work	183	audio	133	custom	93	policy	58	friend	32
free	389	power	180	inform	131	buyer	92	locate	56	info	30
seal	382	original	179	system	131	fully	92	date	55	shipment	30
all	373	processor	177	view	130	need	91	tax	55	describe	29
open	367	device	172	factory	124	feedback	89	allow	54	change	28
include	361	special	168	service	124	note	89	position	54	damage	27
screen	336	charger	167	experience	122	exchange	87	photo	53	separate	25
display	333	condition	165	look	122	sale	87	refund	53	restock	24
have	324	support	164	facebook	121	start	85	top	52	process	23
no	303	read	161	additional	119	check	84	identify	51	clean	22
not	294	purchase	160	perfect	116	actual	82	quickly	50	scratch	22
usb	265	receive	160	available	112	manufacturer	82	beautifully	49	concern	17
question	251	connect	158	warranty	111	approximate	81	cost	48	fair	17
offer	250	email	155	bid	110	favorite	81	fee	48	credit	15
only	243	internal	155	provide	110	home	80	off	48	three	14
more	231	charge	152	perform	109	design	78	quality	44	win	13
package	220	million	151	sell	109	well	77	sold	44	carefully	12
touch	212	payment	150	ad	108	type	76	help	43	inspect	12
fast	209	contact	149	full	106	busy	75	close	42	discount	10
battery	205	access	147	great	104	back	73	fluid	42	law	9
content	205	set	147	enjoy	99	first	70	response	42	reserve	9
cable	201	technology	147	detail	98	seller	70	treatment	41	wear	9
						like	69	issue	40	invoice	8
						pay	68	complete	39	appear	5
						rate	68	immediate	39		
						left	67	guarantee	38		

F.1. **Principal Components Estimator Details.** Principal components analysis (PCA) is a methodology for projecting a set of data points onto a set of orthogonal basis vectors, usually referred to as *components*. To fix ideas, each BoW datum is a 193-dimensional vector, call it $\mathbf{w}_i = [w_{1,i}, \ldots, w_{193,i}]^\top$, where $w_{k,i}$ indicates the number of times the k^{th} word stem occurred in the seller's description in the i^{th} listing page, minus the mean count of the k^{th} word stem across all listings. The first principal component is chosen by picking a vector of weights or *factor loadings*, $\boldsymbol{\pi}_1 = [\pi_{1,1}, \ldots, \pi_{1,193}]^\intercal$, to construct a linear combination of the regressors, $PC_{1,i} = \mathbf{w}_i^\intercal \boldsymbol{\pi}_1$, that has the highest possible variance subject to the constraint that $\boldsymbol{\pi}_1$ has unit length. The other components, each with their own factor loading vector $\boldsymbol{\pi}_j$, $j \geq 2$, are constructed to be orthogonal to all previous components, and to have the highest possible variance subject to the same unit

length constraint. The orthogonality and unit-length constraints together imply that the variances of successive principal components will be monotone decreasing. This often means that one can use the first few components to capture most of the variance in a set of data with much higher dimensionality.

Online Appendix G. Partial R^2 Metrics

We now assess the predictive power of different sets of variables using the same groupings as in Table 3. Within the variables describing the title and description, we also differentiate between variables describing the volume of information (e.g., the number of images or the length of the text, etc.) from those that describe the style of the text (e.g., percentage of uppercase letters, the number of HTML tags used, etc.). To begin with, we use the regression tree model estimated on the basic dataset ($R^2_{out} = 0.1990$) as a baseline.

We discussed reasons why these different classes of variables could be credible signals of the value of the underlying product or of properties of the seller in Section 2.2, but for context we briefly review that discussion first. The volume of text and image variables can be understood as tools to convey information about the product (e.g., verify the box is factory sealed). Style variables could indirectly convey information about the seller's reliability or professionalism. There are a number of features of the eBay platform that would allow long, stylized listings to be credible signals to would-be buyers. First, sellers that list a large number of goods often use a fixed listing template; second, it is not easy for a technically unsophisticated user to scrape images, tags, and other formatting features from a professional seller's listing; and third, experienced sellers often include links to other items or external websites, further depressing the incentive to copy such listings.

We measure predictive power of a variable group by computing an out-of-sample partial-R^2, or the fraction of price dispersion explained by that variable group, beyond what is explained by the basic dataset. Let SSE^{Basic}_{out} denote out-of-sample sum of squared errors for a random forest model estimated on the basic dataset. For a set of additional variables $\boldsymbol{Z}$, let $SSE^{\text{Basic\&}\boldsymbol{Z}}_{out}$ denote out-of-sample SSE of a random forest model estimated on the basic dataset plus $\boldsymbol{Z}$. The partial-R^2 of the variables in $\boldsymbol{Z}$, controlling for the basic dataset, is thus $PR^2_{\boldsymbol{Z}|\text{Basic}} \equiv \frac{SSE^{Basic}_{out} - SSE^{\text{Basic\&}\boldsymbol{Z}}_{out}}{SSE^{Basic}_{out}}$, using 10-fold cross-validation similarly as before. Table 5 depicts results, along with the out-of-sample R^2 for the model containing the basic dataset and $\boldsymbol{Z}$, for comparison with results from Section 3.

We first consider the seller-supplied title section of a listing. There are three components of the title section: the number of title images, the length of the title's text, and the style of the title's text. The total volume of information conveyed by the title section is characterized by the number of characters and words in the title and the number of images. Table 16 lays out each set of variables as well as the predictive power of each set. When we examine the different channels for conveying information, we find that the length of the title has more predictive power than the number of

Variable Group	Subset	Variables	R^2_{out}	$PR^2_{Z\|\text{Basic}}$
Title Variables	Title Images	Number of Images in Title	0.269	0.089
	Title Text Length	Number of Characters/Words	0.317	0.149
	Volume of Information Conveyed by the Title	Number of Characters/Words and Images	0.335	0.172
	Title Text Style	Number of special characters, % Upper case	0.356	0.198
	All Title Text Variables	Number of Characters/Words Number of special characters, % Upper case	0.371	0.216
	All Title Variables	All of the Above	0.380	0.227
Description Variables	Description Images	Kilobytes of Photos, Dummy for 1 - 5 Photos, Dummy for 6 or More Photos	0.267	0.087
	Description Length	Number of Words in Description	0.291	0.116
	Volume of Information Conveyed by the Description	Number of Words and Size and Number of Photos in Description	0.325	0.159
	Description Text Style	Number of Font Sizes, Number of Font Size Changes Number of HTML Tags, % Upper Case Characters	0.367	0.211
	All Description Text Variables	Number of Words in Description Number of Font Sizes, Number of Font Size Changes Number of HTML Tags, % Upper Case Characters	0.389	0.239
	All Description Variables	All of the Above	0.393	0.243
Bag of Words Variables	Component 1	Description of Item	0.300	0.127
	Component 2	Shipping and Payment Information	0.383	0.119
	Components 3 - 15	N/A	0.402	0.030
	All BoW Components	N/A	0.408	0.263

TABLE 16. Partial R^2_{out}, Kindle Listings: Title, Description, and BoW Variables

images. Although the volume of information conveyed by the title has a significant amount of predictive power, the style of the title text has slightly more predictive power. In fact, the style of the title section, a signal of seller professionalism, has almost as much predictive power as the entire set of title variables combined.

We now turn to the seller-supplied description section. Data from this section are significantly richer than title data for the simple reason that the seller has considerably more space to write and latitude to elaborately format text using HTML tags. Images can be used both to describe the item and format portions of the description. Total volume of information conveyed by the description section is measured as the number of words, and the number and size of images. The predictive power of description section variables (middle section of Table 16) show a pattern similar to title variables. First, description style variables are more important volume of information; in fact, style variables carry almost as much predictive power as all description section variables

Title Variables	Description Variables	BoW Variables	$PR^2_{\mathcal{V}\|\text{All Other Variables}}$
√	√	√	0.252
√	√		0.009
√		√	0.047
	√	√	0.079
√			0.002
	√		0.001
		√	0.035

TABLE 17. Partial R^2_{out} - √: Variables Removed from Full Dataset of the Kindle Listings

combined. Second, variables characterizing the length of the description are more important than image variables, but the difference is less significant than for the title section.

Finally, we turn to the principal component controls derived from the BoW data. The lower section of Table 16 presents the incremental predictive power of adding additional components to the model. For example, adding the second component has a predictive power of 0.119 relative to a model including only the basic dataset and the first component. Each row reports numbers from including all components up to the one indicated in the row, on top of basic data plus components in previous rows (we include our interpretation of the components from Table 4 for reference). We find that BoW information about the product (component 1) and about the seller (component 2) carry almost the same amount of information, while added predictive power of the successive components drops off quickly. This suggests that, although the BoW contains significant information, their collective predictive value can be captured by a parsimonious set of interpretable components.

We now complement the analysis above by considering what happens when we start with the full dataset and remove variables characterizing either the title section, the description section, or the BoW principal components. Here, the predictive power of a given set of variables, as measured by the partial R^2, reflects the amount of predictive power lost if we remove those variables from the model, while retaining others. Results are summarized in Table 17, where check marks indicate the sets of variables whose predictive power is being assessed (i.e., *removed* from the full dataset).

The title and description variables have negligible predictive power after controlling for the other variables in the full dataset, while the BoW variables have only a small amount of predictive power. In fact, the predictive power of any two of these sets of variables is not high after controlling for the other variables in the full dataset. We only obtain a significant loss of predictive power when we remove all three sets of variables from the full dataset, but then the only variables that are being controlled for are those in the basic dataset and the miscellaneous variables described in Table 3. The main lesson we draw from this exercise is that there is significant redundancy or substitutability across our different sets of variables in terms of their capacity for conveying information from sellers

to buyers. For example, if the seller provides a detailed title and uploads images for the title section, then she need not also provide an elaborate description section at the bottom of the listing.