Optimal Product Similarity in Copyright Policy*

Sukjin Han† Kyungho Lee‡

March 4, 2024

Abstract

Copyright policies play a pivotal role in protecting the intellectual property of creators and companies in creative industries. The advent of generative AI calls for renewed attention to the role of these policies. The goal of this paper is to study entry and competition in a market of creatively differentiated products and the role of copyright laws, especially in the face of technologies substituting human creation. A common feature of products with creative elements is that their key attributes (e.g., images and text) are *unstructured* and thus high-dimensional. We use neural network embeddings to quantify unstructured attributes and building a structural model based on them. We focus on a stylized design product, fonts, and use data from the world's largest online font marketplace. In this market, we aim to understand (i) the anatomy of competition among design products in terms of visual attributes, (ii) the role of copyright policy in protecting originality, and (iii) the optimal level of visual similarity (i.e., optimal variety).

Keywords: Copyright, creative industries, optimal product variety, visual data, convolutional neural network.

1 Introduction

Copyright policies play a pivotal role in protecting the intellectual property of creators and companies in the modern knowledge-based economy. These policies grant monopoly rights to creators. They serve as gatekeepers in various sectors of

*This is a preliminary draft. Please do not circulate.

†University of Bristol

‡Yale University

the economy, from sectors as obvious as cultural industries (e.g., books, movies, music, illustrations) to less obvious ones like design industries (e.g., garments, automobiles, furniture, mobile applications). In recent years, these creative industries have witnessed the advent of a disruptive technology, namely, generative artificial intelligence (AI). Generative AI has begun to effortlessly engage in a human-like creative process, generating high-quality images, texts, sounds, and videos with scale and efficiency never seen before. It is expected that, in every one of these industries, markets will be crowded with AI generated products, competing with human-generated products. This recent transformative trend therefore calls for renewed attention to the role of copyright policies (Samuelson, 2023).

The primary goal of this paper is to study entry and competition in a market of creatively differentiated products and the role of copyright laws, especially in the face of technologies substituting human creation. A common feature of products with creative elements is that their key attributes are *unstructured* and thus high-dimensional. Examples of unstructured attributes are images and text, which are in turn key creative objects typically protected by copyright laws. Therefore, quantifying unstructured attributes and building an economic model around them are crucial steps to achieve our research goal.

Among creative products, we focus on a particular design product—fonts—for several reasons. First, fonts are differentiated products where visual attributes mostly describe the characteristics of the product, highly predictive of its value and functionality. This is a feature unique to this particular product. Second, copyright issues have been important policy questions in this industry (e.g., Carroll (1994); Lipton (2009); Manfredi (2010); Evans (2013)) as well as the introduction of AI-assisted design of fonts (e.g., Zeng *et al.* (2019); Wang *et al.* (2020)). Third, the product's visual information is one of the simplest among all design products. As described below, font images are monochrome and standardized, facilitating our dimension reduction procedure. Fourth, due to the visual simplicity, it is easy to interpret the unstructured attribute and the associated copyright policy within our economic model. This aspect is useful for our counterfactual analyses. Fifth, fonts are ubiquitous and serve as intermediate goods for many final products (e.g., websites, printed materials), and the fonts market is large with frequent productions and transactions. Therefore, policies in this market have implications beyond the font market, to markets for final products. Finally, we view fonts as stylized products that capture an essential aspect that many products in the market have in common, namely, design attributes and associated copyrights.

In this stylized market, we aim to understand (i) the anatomy of competition among design products in terms of visual attributes, (ii) the role of copyright policy in protecting originality, and (iii) the optimal level of visual similarity (i.e., optimal variety). To this end, we first characterize images of fonts as embeddings, low-dimensional vectors in the Euclidean space, using a state-of-the-art convolutional neural network (Han *et al.*, 2021). Given the embeddings, we characterize the competition of firms in the visual dimension as a spatial competition in the Euclidean space. We build entry and price models of firms accordingly. The entry model describes the firms' choices of location considering the later stage's pricing decisions. A copyright policy is modeled as imposing restrictions on the area of possible choices. We also model demand by a discrete choice model of heterogeneous consumers, in which the embeddings as product characteristics influence the utility. The purpose of this structural model is to conduct counterfactual analyses, such as assessing the effects of the stringency of the copyright policy and the effects of entry of AI-generated fonts. Visual similarity is a crucial metric in our counterfactual exercises, and it is practically relevant for several reasons. First, AI-generated fonts may be indistinguishable from those created by humans. Second, given the concerns that regulating AI's inputs (i.e. training data) may hamper innovation and competition, many policymakers of the copyright policy are interested in the regulation of output similarity.[1]

The data we use comes from the world's largest online marketplace for fonts. The data contain information about nearly 28,000 fonts (produced by font design firms called foundries) and 2,400,000 transactions over the past six years. Using this data and panel data models, we first document that firms in this market engage in *local competition*. Employing two distinct empirical strategies, we show that business stealing exerts significant and lasting effects on sales and revenue. This is particularly pronounced when entry occurs near the focal product, with "nearness" defined by the Euclidean distance derived from the embeddings. Next, we turn to estimating the structural model of supply and demand described above. Overall, our demand-side estimation results show that consumers tend to prefer fonts with high quality and functionality, prices decrease utility, and visual attributes are important in explaining consumer substitution. The supply-side

[1]For example, in July 2023, the Japanese government introduced copyright policy guidelines pertaining to generative AI. These guidelines permit the use of copyrighted materials as training inputs for AI *without* permission. Instead, the guidelines enforce copyright policies through the application of existing similarity-based criteria for determining copyright infringement (https://www.natlawreview.com/article/japanese-government-identified-issues-related-ai-and-copyrights).

estimation and counterfactual analyses are in progress.

1.1 Related Literature

This paper relates to the literature on the welfare trade-off engendered by property rights, which is a classic economic problem (Romer, 2002; Stiglitz, 2007). Studies have utilized historical quasi-experimental variations to identify the effects of the copyright system on outcomes such as price, creation and quality. For instance, Li *et al.* (2018) find that the UK Copyright Act of 1814, which resulted in a differential increase in copyright length, increased prices. Biasi and Moser (2021) exploit the weakening of copyrights during World War II, highlighting that such dilution led to the creation of follow-on science, manifested as increased citations. Giorcelli and Moser (2020) show that the adoption of copyright policy in Italy, induced by Napoléon's victories, leads to the creation and longevity of new operas.

Copyright protection is essential for the functioning of digital markets and new technology has generated policy challenges. Existing literature has paid considerable attentions on piracy of digital products. Oberholzer-Gee and Strumpf (2007) study the effect of file sharing on revenues in the music industry and conclude that file sharing resulted in a significant decline in music sales. Rob and Waldfogel (2006) use a sample of college students and report reduced expenditures on albums but an increase in consumer welfare due to downloading. Waldfogel (2012) shows that the quality of music was not degraded due to the introduction of Napster. To the best of our knowledge, our paper is the first attempt to address the question of permissible similarity for copyright protection in economics, a key concept in the copyright policy.

Our study is also related to the literature on optimal product variety. It is theoretically well-documented that free entry may lead to social inefficiency (Dixit and Stiglitz, 1977; Spence, 1976a,b; Mankiw and Whinston, 1986; Anderson *et al.*, 1995). This conclusion has motivated empirical researchers to examine inefficiency in markets and policy tools for achieving socially optimal levels. Berry and Waldfogel (1999a, 2001) empirically demonstrate market inefficiency in the radio broadcasting market. They conclude that market concentration reduces entry yet increases product variety, using the 1996 Telecommunications Act as a quasi-experimental variation for relaxation of ownership restrictions. Berry *et al.* (2016) also report the existence of such inefficiency by extending the empirical model of Berry and Waldfogel (1999a) with vertical differentiation of radio stations. Sweeting (2013) studies dynamic product positioning of radio stations and shows that high fees for music performance rights quickly decrease the number of music stations. We contribute to this literature by treating allowable similarity under copyright protection as a policy tool to enhance social welfare.

This research also contributes to the literature employing high-dimensional unstructured data in the economics literature by introducing several new approaches based on embedding analysis. For instance, Gentzkow and Shapiro (2010) use text data in the U.S. daily newspapers to construct an index of ideological slant in news. Based on the index they estimate consumer demand for newspapers and find that ideological preferences significantly influence newspaper demand. Gentzkow *et al.* (2019b) measure political polarization by congressional speech textual data. In addition, Hoberg and Phillips (2016) propose product classification by creating a product location space, similar to the visual characteristics space in this paper, via 10-K product descriptions of firms. Gentzkow *et al.* (2019a) provide a survey about textual data and its application in economics. In the realm of image data, Glaeser *et al.* (2018) use Google Street View data and predict economic outcomes of neighborhoods. Zhang (2018) report that images of Airbnb properties affect consumer demand. Han *et al.* (2021) use font images and a convolutional neural network to construct embeddings that are used to conduct a merger analysis. Using their embeddings, this paper illustrates how state-of-the-art computer vision methods are useful in addressing policy-relevant questions in economics.

This paper extends the literature on product positioning by considering entry decisions in a high-dimensional characteristics space. Papers like Berry (1992), Mazzeo (2002) and Seim (2006) introduce models of firms' entry choices, utilizing cross-sectional variations in the number of firms across markets. Moreover, Jia (2008) studies the entry decisions of large retailers in each location and their welfare implications for nearby small retailers. Holmes (2011) study Wal-Mart's location choices as a single-agent dynamic problem and trade-offs between the benefits of economies of density and cannibalization. Fan (2013) proposes a merger analysis including firms' static product differentiation using U.S. newspaper market data. Eizenberg (2014) studies the impact of upstream innovation, i.e. increases in CPU performance, on downstream product configuration choices in the computer market. Wollmann (2018) investigates model-level entry and exit in the U.S. truck market, showing the importance of changes in product offerings in terms of welfare analysis.

Lastly, this paper contributes to the literature on similarity judgment in copyright policy, which extends beyond the scope of the economics literature. The subjectivity of similarity judgments by humans has faced criticism in the legal literature (Lemley, 2009; Balganesh *et al.*, 2014). Scheffler *et al.* (2022) propose a computational framework to determine whether two products are substantially similar by quantifying the "novelty" of the products. In contrast, we introduce an economic framework to analyze the impact of similarity-based copyright policies on markets. This framework employs economic models to understand market behaviors and assess welfare implications, in conjunction with machine-measured metrics to quantify high-dimensional characteristics.

2 Institutional Background

2.1 Fonts

Fonts are recognized as software goods in the digital marketplace. The software delivers typefaces to the user and is purchased through downloads. The main consumers of fonts are designers, who use them in a wide array of commercial design projects. Examples of such design outputs include printed materials (e.g., book covers and interiors, advertising posters), packaging and store signs, as well as websites and mobile applications (Figure 1). In this sense, fonts serve as intermediate goods for final products and they are among the most ubiquitous objects encountered in daily life. Fonts are downloaded by consumers under specific types of licenses. For instance, a desktop license (used for printed materials) specifies the number of users who can install and use the font, while a web font license (used for websites) is based on the number of online views the font receives.

Figure 1: Commercial Applications of Fonts

(a) Product Case/Box (b) Book and Website

Notes. These figures show commercial applications of fonts.

The market for fonts shares several characteristics with broader markets for creative goods. First, the key attributes of products in this market are unstructured and thus high-dimensional. Second, fonts are distributed at a markedly low marginal cost. However, the low distribution cost is offset by a relatively high fixed cost associated with font creation, which includes both the design of typefaces and the development of software. This feature of low marginal cost but high fixed cost is common among markets with copyright protection (Waldfogel, 2012).

In the font industry, fonts are organized into a hierarchical or nested structure that includes family, styles, and glyphs. A font family is a set of font styles that share common design traits, while individual styles within the family, such as italic or bold, introduce variation to the base design. The design process often begins with the creation of a default style, which serves as the foundation from which variations are developed. Glyphs are characters in each style. Figure 2 illustrates the family structure and provides examples.

2.2 Copyright Policy

The concept of substantial similarity plays a crucial role in copyright infringement cases, serving as a basic criterion for establishing proof of copying (Lemley, 2009; Balganesh *et al.*, 2014). According to Lemley (2009), court procedures for determining copyright infringement involve gathering and aggregating information from "ordinary observers" — consumers of copyrighted products — and experts knowledgeable about the characteristics of such products. This information is used to evaluate whether the "total concept and feel" of one product is substantially similar to another. In this paper, we employ neural networks to measure the similarity between typefaces of fonts, effectively using machines as observers for policy and market analysis.

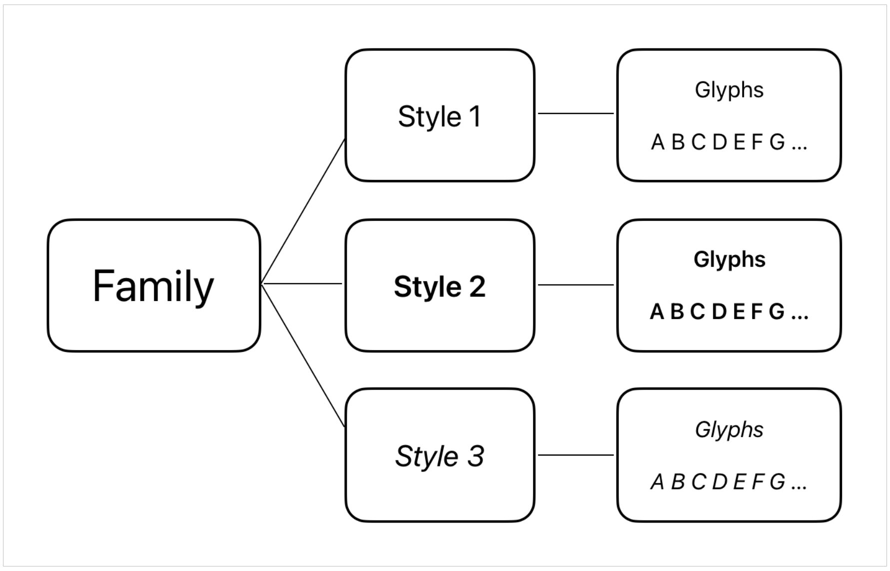
One interesting aspect of the font market is the variation in copyright laws between countries. In the United States, fonts are protected under copyright law as software, granting them a distinct legal standing. Meanwhile, in Europe, both the design aspects and the software codes of fonts are subject to copyright protection. This lack of consensus on copyright policies, with varying degrees of stringency, motivates our research question regarding the optimal copyright policy.

2.3 Myfonts.com

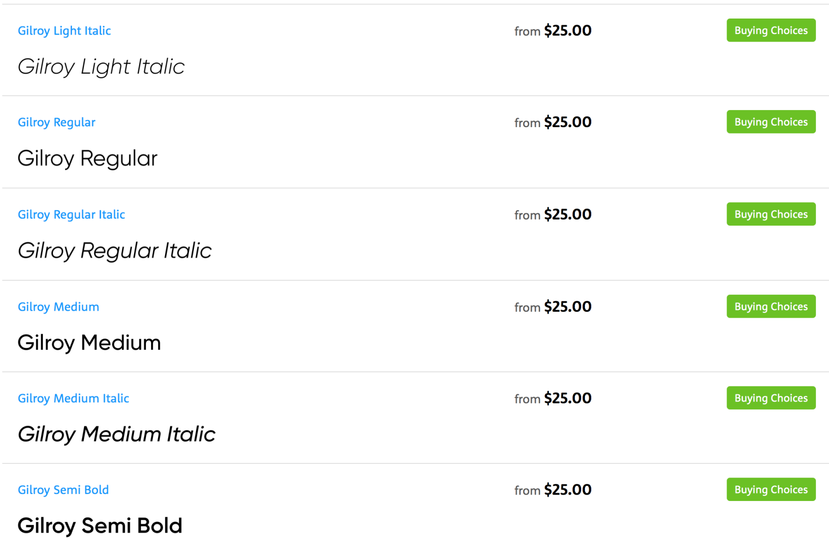
We consider MyFonts.com, the world's largest online font marketplace, which offers approximately 30,000 different fonts. This market is a superset of all major global online font stores, all of which, including MyFonts.com, are owned by Monotype

Figure 2: Font Family, Style and Glyphs

(a) Font Family Structure



(b) Example: Gilroy font family



Notes. Panel (a) illustrates a hierarchical or nested structure of a font family, styles and glyphs. A family is a set of font styles, while individual styles within the family, such as italic or bold, are variation to the default style design. Glyphs are characters in each style. Panel (b) presents Gilroy font family as an example.

Inc. Panel (b) in Figure 2 presents an example of a webpage from MyFonts.com for a particular font family.

3 Data

This section describes the datasets used for our empirical analyses. The first dataset is panel data that we construct from market data, which includes transaction records. The second dataset is embedding data that we construct from image data. The market and image data are sourced from Myfonts.com.

3.1 Market Data

Our transaction data spans from the second quarter of 2014 to the end of 2017 and can be likened to "scanner data" from retail shopping. Each order, identified by a unique ID, consists of one or more SKUs, each corresponding to either a font family or a single font style.[2] Each order contains detailed information such as consumer ID, transaction time, revenue (i.e. subtotal), and tax paid. We also have information on registered consumers in the marketplace, which can be linked to transactions via consumer ID. This includes country, city, registration date, and total marketplace expenditure.

In addition, we utilize firm and product data, which can be interconnected. This dataset provides important product-level information, such as list price, entry date, ownership, name, consumer-generated tags, supported languages, number of glyphs per style, and a family's default style.[3] We also observe changes in list prices through new SKUs for the same family. Furthermore, we observe designers associated with each font creation.

From the transaction data, we construct panel data structured by product (font family), license type, country and month. We only consider transactions involving desktop and web license types, which represent approximately 99% of total transactions. Also, we focus on 12 countries that contribute the most to total sales and primarily use the Roman alphabet.[4]

[2] While the data begins in 2012, transactions involving desktop licenses are not recorded from 2012 to 2014. As desktop licenses account for the majority of transactions, we limit our data period from the second quarter of 2014 to the end of 2017, using earlier data for auxiliary purposes. Approximately 80% of all transactions involve desktop license fonts.

[3] List prices are recorded at the SKU level, not the family level. As list prices vary with the number of styles, we calculate a per-style list price for each family.

[4] These countries are Australia, Austria, Canada, Finland, France, Germany, Italy, Netherlands, Sweden, Switzerland, United Kingdom, and United States of America.

3.2 Descriptive Analysis

Table 1 shows descriptive statistics of the panel data. Overall, revenue, quantity and prices are right-skewed with large standard deviations.

Table 1: Descriptive Statistics of Panel Data

License	Variables (Unit)	Observations	Mean	Std.	Min	Max
Desktop	Revenue ($)	381,216	19.99	213.33	0	47,498
	Quantity (Users)	381,216	1.75	32.61	0	12,815
	Sales Price ($)	381,216	10.01	10.36	0.01	160
	List Price ($)	366,852	26.15	26.41	1.25	1,000
	Zero Sale	381,216	0.84	0.36	0	1
Web	Revenue ($)	105,996	17.36	204.54	0	20,574
	Quantity (10K Views)	105,996	4.54	221.45	0	20,000
	Sales Price ($)	105,996	12.48	11.56	0.00	125
	List Price ($)	102,000	25.49	17.40	1.39	199
	Zero Sale	105,996	0.86	0.35	0	1

Notes. This table contains descriptive statistics of panel data that is constructed from transaction records. The panel is four-way: product (font family), license type, country and month. $ stands for United States Dollar. Zero Sale variable is a dummy indicating whether the product is not purchased by consumers in the market (i.e., a combination of license type, country and month).

There are some notable facts about the marketplace. First, list prices typically do not change over time. We analyze list prices further to understand their variation. We find that product fixed effects can explain about 99.2% variations of list prices. However, this is not the case for other variables such as revenue, quantity, and sales prices; product fixed effects only account for 13.4%, 0.8 %, and 65.2% of the variations in revenue, quantity and sales prices, respectively. See Table A2 in the Appendix for details.

In addition, universal quantity discounts exist in the marketplace: the greater the quantity purchased in each transaction, the cheaper the price is. However, the quantity units of most transactions are small and are not subject to quantity discounts. In Table A1, we show fraction of transactions by quantity purchased in each transaction. Quantity discount is not applied for 1-user and 5-users desktop license and 10,000 views web font license, which takes about 87 % of whole transactions. Because the quantity units for desktop and web licenses are the number of users who can install the software and the number of website views, it appears that the main consumers of the marketplace are small-sized.

To see how price distributions are affected by quantity discount, we compare distributions of log of sales and list prices. In Figure A1 distributions of sales prices and list prices are presented. For those transactions without quantity discount, the distribution of sales prices and list prices are similar; sales prices are slightly shifted left possibly due to temporary discounts. This is shown in Figure A1.(b)

3.3 Image and Embedding

In addition to the market data, this paper uses font images (i.e., images of pangrams[5]) and their embeddings, which are the low-dimensional representation of images. The embedding analysis is a machine learning method transforming high-dimensional (or unstructured) data into low-dimensional vectors of numerical numbers.[6] This paper uses embedding data from Han *et al.* (2021), consisting of 128×1 vectors, each representing the image of a font (pangram). Han *et al.* (2021) compute the embedding of font images using a convolutional neural network (CNN) that minimizes a triplet loss function.[7] The embedding vector is normalized to have a unit length and thus lies in a 128-dimensional hyper-sphere. For the distance between embeddings, we use the Euclidean distance.[8]

Based on the neural network embeddings, we calculate the Euclidean pairwise distances, with the minimum and maximum distances being 0.03 and 1.37, respectively, in our data.[9] Table 2 presents examples of fonts and images according to their distance from the base font, Minion. The visual difference becomes more apparent as the pairwise distance increases. For example, the closest font, Alia JY, is visually very similar to the base font, Minion. Conversely, the typeface of Vodka, which is 0.73 distance away from the base font, is handwritten, making it clearly distinguishable from Minion.

[5]A pangram is a sentence that uses every alphabet character at least once.

[6]For example, the embedding analysis can be applied to text data for measuring the similarity of words or sentences.

[7]The CNN model is adept at capturing the visual properties of images because the method preserves the local features among pixels, as demonstrated in (Krizhevsky *et al.*, 2017; Simonyan and Zisserman, 2014). The triplet loss function is utilized to leverage information that fonts belong to the same family are visually similar. This way, fonts within the same family are grouped closer in the resulting embedding space. To evaluate embedding performance, Han *et al.* (2021) conduct visualizations of images in 2-dimensional space, nearest neighbor analysis, and measure similarity between embeddings of image and consumer tags, which are words describing the shape of the font, created by consumers.

[8]As the length of the embedding is normalized to one, the Euclidean distance and the cosine similarity distance have a one-to-one relationship.

[9]We only use a subsample of images for our analysis and are currently working on extending it to the full image sample.

Table 2: Examples of Fonts and Images by Distance from Base Font (Minion)

Font Name	Distance	Pangram Shape
Minion	0.00	The quick brown fox jumps over
Alia JY	0.08	The quick brown fox jumps over
Garamond	0.19	The quick brown fox jumps over
Bauhaus Bugler Soft	0.28	The quick brown fox jumps over
Andrea Handwirtting II	0.44	The quick brown fox jumps over
Vodka	0.73	The quick brown fox jumps over

Notes. This table presents examples of fonts alongside their corresponding pangram images. The pairwise Euclidean distances are calculated between the base font (Minion) and each font listed in the table. The Font Name column displays the names of the font families, while the Distance column provides the calculated pairwise distances. The Shape column exhibits the pangram images for each font.

4 Spatial Competition in Characteristics Space

Given the neural network embeddings constructed from the image data, we can define the visual characteristics space of fonts as a subset of the Euclidean space. Then, each designer's design differentiation decision can be viewed as choosing a location in the characteristics space, potentially competing spatially with other designers. This conceptual framework is the basis for the paper's empirical analyses.

As an initial empirical analysis, we investigate the nature of spatial competitions among designers. We presents evidence for the effects of competition in the characteristics space on firm outputs. First, we seek descriptive relationship between the degree of spatial competition and market outcomes, such as revenues, sales quantity and prices. Second, we quantify the causal effects of business stealing from entries of visually similar products on incumbents.

4.1 Number of Spatial Competitors and Market Outcomes

We first investigate how the number of spatial competitors affects market outcomes. To start, we count the number of product entrants at time t within an open ball of radius r around product j:

$$B_{jtr} := \sum_{j' \in J_t} 1\{||x_{j'}^{emb} - x_j^{emb}||_2 < r\} \text{ for } r \in \mathbb{R}, \tag{1}$$

where subscript j, l, and t denote the font product, license type, and monthly time, respectively, and J_t is the set of products in period t. We then use $(B_{jtr} - B_{jtr'})$ for $r > r'$ as a measure of the degree of spatial competition for a given distance range. Figure 3 visualizes this approach.

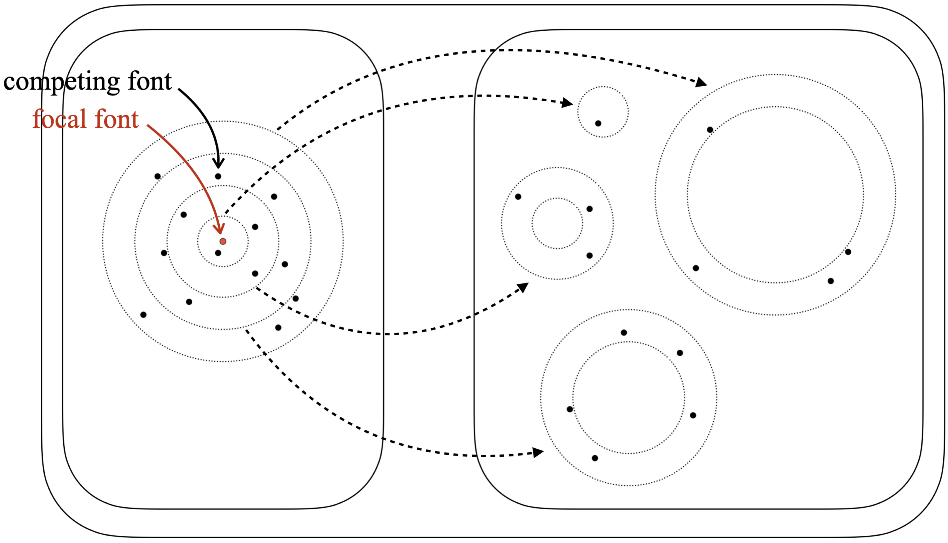
Table 3 presents the descriptive statistics for the number of spatial competitors. The variation in the number of competitors arises from both cross-sectional and time-series variations. There is considerable dispersion in the number of spatial competitors.

Given the number of spatial competitors, we write the regression equation for market outcomes as

$$y_{jlct} = \sum_r \gamma_r (B_{jtr} - B_{jtr-0.2}) + \alpha_j + \alpha_l + \alpha_c + \alpha_t + u_{jlct}, \tag{2}$$

where y is the *arcsinh* transformation of revenue and quantity, and the log of

Figure 3: Counting Competitors on Visual Characteristics Space



Notes. For each focal product j (the red dot on the left-hand side), we count the number of competitors (the black dots) between two adjacent balls with radius r and r'. We use the Euclidean pairwise distance.

Table 3: Descriptive Statistics of B_{jtr}

Variables	$B_{jt0.1}$	$B_{jt0.2}$	$B_{jt0.3}$	$B_{jt0.4}$	$B_{jt0.5}$	$B_{jt0.6}$	$B_{jt0.7}$	$B_{jt0.8}$	$B_{jt0.9}$
Mean	0.4	10.4	97.3	397.4	999.9	1,788.0	2,503.5	3,005.4	3,296.8
Std.	2.9	16.8	99.5	293.9	559.1	722.9	681.4	510.9	303.1
Min	0	0	0	0	0	5	17	42	100
Max	38	93	522	1,382	2,433	3,096	3,493	3,716	3,741

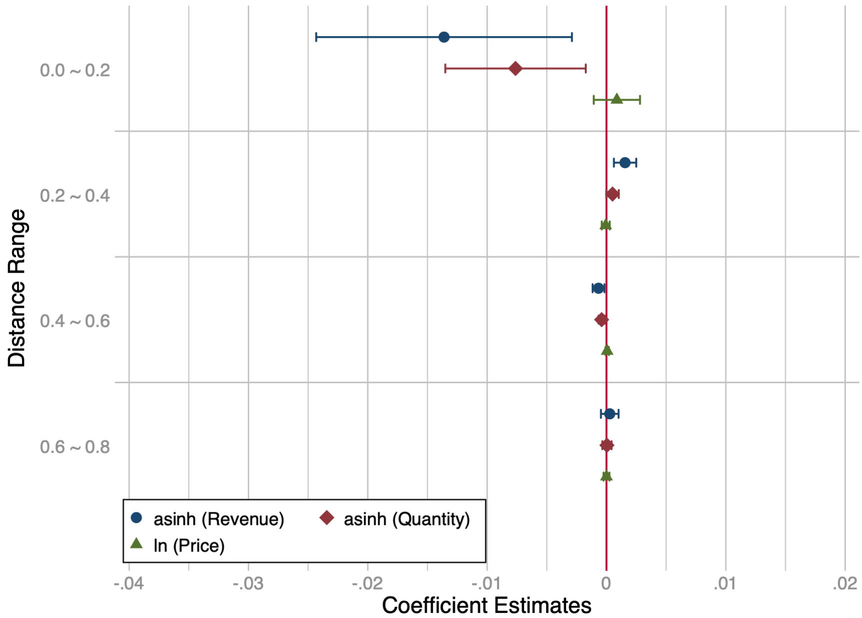
Notes. This table displays the descriptive statistics for the number of spatial competitors in the visual characteristics space. The number of observations is 486,744.

price.[10] Here, α_j, α_l, α_c, and α_t are fixed effects for product, license type, country of the market, and time, respectively, and u_{jlct} is an error term. We consider two regression models with different radius specifications: one where r is set to $0.2, 0.4, 0.6$, and 0.8 and another where r is set to $0.1, 0.3, 0.5, 0.7$, and 0.9. We refer to the first and second regression specifications as the 0.2-ball model and 0.1-ball model, respectively. In these models, γ_r captures the relationship between the number of additional competitors within a given distance range and market outcomes, controlling for fixed effects.

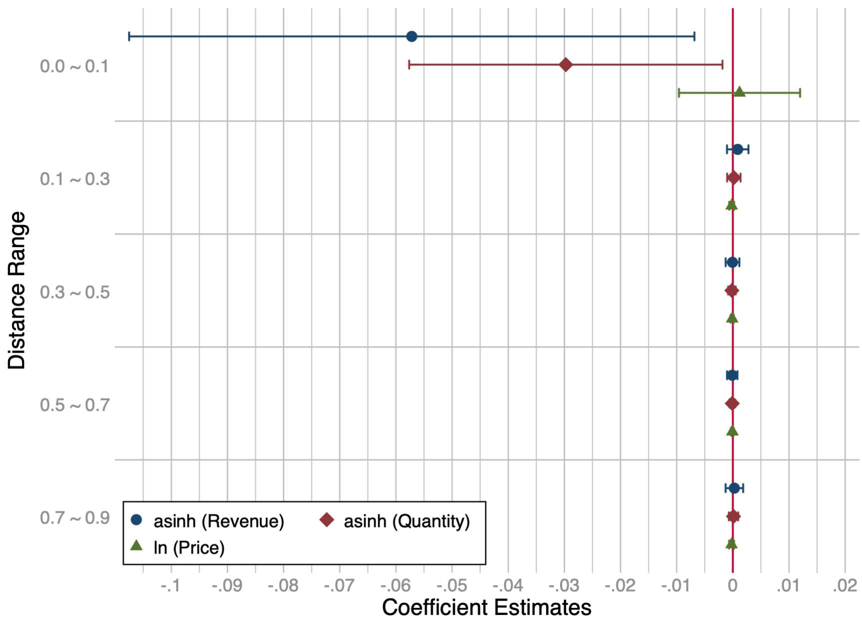
[10] We also consider an alternative specification by taking the log after adding 1 to revenue and quantity to accommodate zero-sale products. The results are qualitatively similar. Additionally, to address potential bias due to universal quantity discounts in the marketplace, we use the list price per style of product j as the price variable.

Figure 4: Spatial Regression Estimates (γ_r)

(a) 0.2 radius innermost ball and outer rings



(b) 0.1 radius innermost ball and outer rings



Notes. Coefficient estimates from regression (2) are presented. Panel (a) and (b) display results under specifications where the radii of inner balls are 0.2 and 0.1, respectively. Round-, diamond-, and triangle-shaped dots represent estimates for the revenue, quantity, and price variables, respectively. Solid lines indicate 95% confidence intervals. Standard errors are clustered at the product level.

Figure 4 presents the regression results of model (2). Panel (a) shows results for the 0.2-ball model. Local competition in the characteristics space is both statistically and economically significant. These results suggest that the elasticity of revenue and quantity in response to an additional competitor within the innermost ball is around -1.32 and -0.84, respectively.[11] Notably, for both revenue and quantity, the coefficient estimates for the innermost ball, $\gamma_{0.2}$, are significantly larger than those for the outer rings. For prices, the coefficient estimates are near zero and not statistically significant. When price is the dependent variable, the model may be too saturated to control for product-level fixed effects. Therefore, we also use firm-level dummies instead of product dummies, which yields qualitatively similar results; see the Appendix.

Panel (b) shows results for the 0.1-ball model. It is noteworthy that, for revenue and quantity, the coefficient estimates for the innermost ball, $\gamma_{0.1}$, are about three times larger than those for the 0.2-ball model; the estimated elasticity of revenue and quantity in response to an additional competitor in the innermost ball is about -5.69 and -3.36, respectively. The coefficient estimates of outer rings are not statistically significant at 5% level and are close to zero. The findings in Figure 4 support that competition in the visual characteristics space significantly affects market outcomes and that such competition is local.

4.2 Business Stealing Effect of Visually Similar Product

Next, we analyze the causal effects of business stealing due to the entry of visually similar products. We are particularly interested in determining whether and to what extent the profits of an incumbent are reduced by such an entry. This analysis aims to provide evidence on the crucial role visual attributes play as a dimension of product substitution. This has potential implications for optimal product variety in this marketplace, which we explore in subsequent sections. Moreover, this analysis complements the previous analysis, which does not explain how pre- and post-entry of a new product affects market outcomes. Also, the analysis in this section is free from the distance cutoff (e.g., setting 0.2 as the distance range) used in the previous analysis.

First, we define the treatment as a change in one of the five visually closest competitors due to a new entry. Let T_j represent the first month when the treat-

[11] We approximate the elasticity by calculating $\frac{\partial y}{\partial B_r}\frac{B_r}{y} = \gamma_r \times \sqrt{1+\frac{1}{y^2}}$ for the revenue and quantity variables under the *arcsinh*-linear specification, following Bellemare and Wichman (2020). The elasticity approaches γ_r as y increases.

ment occurs. If the treatment never happens, we set $T_j = \infty$. We then define the event dummies as $E^s_{jt} := 1\{t - T_j = s\}$ for $s \in \mathbb{Z}$. Given these dummies, we specify an event study design:

$$y_{jlct} = \sum_{s=-5}^{9} \beta_s E^s_{jt} + \alpha_f + \alpha_l + \alpha_c + \alpha_t + e_{jlct}, \tag{3}$$

where α_f, α_l, α_c, and α_t are fixed effects of firm, license type, country, and time, respectively, and e_{jlct} is an error term. We define dependent variables as the *arcsinh* transformation of revenue and quantity, the log of price, and a zero sale dummy. We normalize regression results by setting $\beta_{-1} = 0$, following the standard event study exercises.

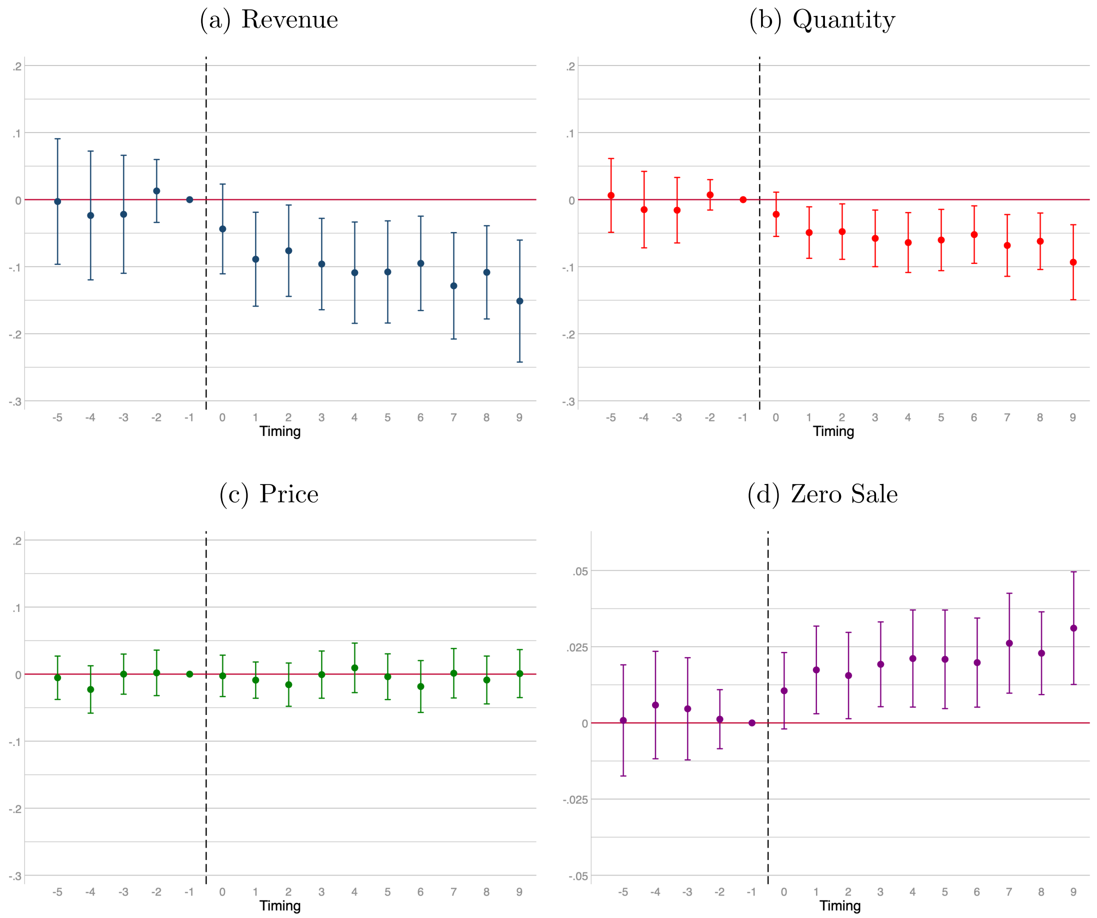
Figure 5 presents the results of the regression (3). We find that the business stealing effects are significant and enduring for revenue and quantity, suggesting strong substitution between visually similar products after new entries. The substantial loss of revenue due to entry is mostly driven by a decrease in sales quantity, as shown in Panel (b). An increase in zero sale, as seen in Panel (d), also suggests substantial substitution due to new entries. However, it is shown that price is not responsive. All the pre-trend coefficients are statistically insignificant, supporting the validity of the parallel trend assumption.

One potential threat for identifying causal effects is the staggered adoption of treatment in our setting. It is known that, with staggered adoption, OLS estimates may not be a convex combination of treatment effects on treated in different timing (Goodman-Bacon, 2021). To circumvent this problem, we employ the method by Borusyak *et al.* (2021), which imputes the unobserved potential outcome value under the linear fixed effect model and then takes an average to calculate the causal effect.[12] The results, shown in Figure A3 in the Appendix, are qualitatively similar, supporting that our estimates are robust to staggered timing of the treatment.

Furthermore, we conduct several different specifications to check the robustness of our model (3). First, we run the event study design with additional control variables, including the age of the product measured by months after entry into the marketplace, the log of glyphs, and interaction dummies between time and image cluster. Results, shown in Figure A4 in the Appendix, are again very similar to our main findings in Figure 5. Second, we try an alternative definition of treatment. In this exercise, we define the treatment as a change in one of

[12] We take a simple average of all treated observations.

Figure 5: Event Study Estimates (β_s)



Notes. These figures display the results of the event study regression as described in (3). Panels (a) and (b) present regression results for the arcsinh transformation of revenue and quantity, respectively, as the dependent variables. Panels (c) and (d) show results for the log of list prices and the zero sale dummy as dependent variables, respectively. Solid lines represent the 95% confidence intervals of the estimates. We use firm-level clustered standard errors.

the *four* visually closest competitors due to a new entry. The estimation results, shown in Figure A5 in the Appendix, are also qualitatively similar to the previous findings.

5 Models

To justify the above empirical findings and understand the role of copyright policy and the optimal level of variety, we now build empirical models for demand and supply in the font market. On the supply side, our model aims to capture the entry decision-making process given the high-dimensional visual characteristics. The main model primitive to recover is the fixed costs associated with entry decisions. On the demand side, we aim to characterize consumers' preferences for the high-dimensional visual characteristics. The main model primitive to recover is consumers' substitution patterns. Our primary goal in building these models is to conduct counterfactual analyses that help us understand the impacts of similarity-based copyright policy and potential shifts in market fundamentals due to the introduction of generative AI .

Throughout the section, the subscript i, j, l, c, and t denote consumer, font product, country of the marketplace, license type, and time, respectively. We define the market m as a combination of license type, country, and time.

5.1 Demand Model

First, we consider the demand system of J goods, following the model in Berry and Haile (2014). The market here is defined as the combination of license type l, country c, and time t. We suppress the subscripts l and c for notational simplicity. We denote the collection of characteristics as

$$\chi_t = (\boldsymbol{x}_t, \boldsymbol{p}_t, \boldsymbol{\xi}_t)\,,$$

where $\boldsymbol{x}_t = (\boldsymbol{x}_t^{\prime str}, \boldsymbol{x}_t^{\prime emb})'$ is the observable characteristics vector and $\boldsymbol{p}_t$ is the vector of prices, and $\boldsymbol{\xi}_t$ is the vector of demand shocks unobservable to the analyst. The system is written as

$$\sigma(\cdot) = (\sigma_1(\cdot), \cdots, \sigma_J(\cdot)) : \mathcal{X} \to \Delta^J,$$

where $\mathcal{X}$ is the support of χ_t and Δ^J is the J-dimensional unit simplex. In Berry and Haile (2014), the nonparametric identification of the demand system hinges on the the inversion of the demand system and the use of conditional moment restrictions with valid instruments. Similar to their approach, we first impose an index restriction for the inversion. Specifically, we let the structured attributes including the number of glyphs enter the demand functions in a linear fashion:

$$\delta_{jt} = x_{jt}^{str} + \xi_{jt}.$$

We denote the vector of indices as $\boldsymbol{\delta}_t = (\delta_{1t}, \cdots, \delta_{Jt})'$. Let the observed market share be s_{jt}. Then the demand system maps the indices, prices, and product characteristics to the observed market share as

$$s_{jt} = \sigma_j(\boldsymbol{\delta}_t, \boldsymbol{p}_t, \boldsymbol{x}_t^{emb}), \qquad j = 0, ..., J. \tag{4}$$

In addition, we impose the assumption of connected substitutes introduced in Berry *et al.* (2013):

Assumption 1. Products $(0, 1, ..., J)$ are connected substitutes in $-\boldsymbol{p}_t$ and in $\boldsymbol{\delta}_t$.

Intuitively, Assumption 1 means that goods are allowed to be weak substitutes to each other with respect to the prices and linear indices. This would reasonably hold in our empirical setting; an increase in price of competing products would (weakly) increase the market share of a given product. In addition, an increase in the number of glyphs results in better functionality of the font, whose impact would be similar to that of a decrease in price. The assumptions of index restriction and connected substitutes allow the inversion of the demand system as

$$\delta_{jt} = x_{jt}^{str} + \xi_{jt} = \sigma_j^{-1}(\boldsymbol{s}_t, \boldsymbol{p}_t, \boldsymbol{x}_t^{emb}), \qquad j = 1, ..., J,$$

which is equivalent to the following specification of nonparametric regression:

$$x_{jt}^{str} = \sigma_j^{-1}(\boldsymbol{s}_t, \boldsymbol{p}_t, \boldsymbol{x}_t^{emb}) - \xi_{jt}. \tag{5}$$

The equation (5) allows us to cast the problem of identification and estimation of the inverse demand functions as a nonparametric instrument variable (NPIV) regression problem. Let $\boldsymbol{z}_t = (z_{1t}, ..., z_{jt})$ be a vector of instruments. We impose the conditional moment restriction and completeness condition as follows:

Assumption 2. $E\left[\xi_{jt} | \boldsymbol{x}_t, \boldsymbol{z}_t\right] = 0$ a.s. for all $j = 1, ..., J$ and $t = 1, ..., T$

Assumption 3. For all functions $B(\boldsymbol{s}_t, \boldsymbol{p}_t, \boldsymbol{x}_t^{emb})$ with finite expectation, if $E\left[B(\boldsymbol{s}_t, \boldsymbol{p}_t, \boldsymbol{x}_t^{emb}) | \boldsymbol{x}_t, \boldsymbol{z}_t\right] = 0$ a.s., then $B(\boldsymbol{s}_t, \boldsymbol{p}_t, \boldsymbol{x}_t^{emb}) = 0$ a.s.

These assumptions are standard in the NPIV literature (e.g. Newey and Powell, 2003) and employed in Berry and Haile (2014). Under these assumptions, we can point identify the inverse demand functions σ_j^{-1}.

Next, we turn to the estimation problem. An obvious challenge in estimating (5) is dimensionality, arising from two separate sources: (i) the large number of products and (ii) the high dimensionality of visual attributes. This is because σ_j^{-1} is a function of the shares, prices, and embeddings for a given products and competing products in the marketplace. The issue of the curse of dimensionality, akin to that in nonparametric regression, is well-discussed in Compiani (2022).

We impose a restriction of local competition on (5), motivated by the empirical findings in Section 4. Specifically, we assume that the variables of competitors in σ_j^{-1} only concern K visually nearest competitors:

$$x_{jt}^{str} = \sigma_j^{-1}(\boldsymbol{s}_t^j, \boldsymbol{p}_t^j, \boldsymbol{x}_t^{j,emb}) - \xi_{jt}, \tag{6}$$

where the superscript j indicates that each vector contains own and nearby competitors' variables; for instance, $\boldsymbol{s}_t^j$ is the vector of shares of product j and j's K nearest competitors. This restriction would substantially reduce the dimensionality of inputs in the nonparametric function.

In addition, to reduce the dimension of product characteristics, which is 128-dimension for each product, we use neural network in estimating σ_j^{-1} and utilize its dimension reduction feature. There is the large literature documenting how neural nets can be viewed as dimension-reduction techniques (e.g. Fodor, 2002; Wang *et al.*, 2014).

As an alternative model specification and estimation approach, we also consider the standard parametric specification followed by GMM (e.g., Berry, 1994; Berry *et al.*, 1995). Still, to deal with the unstructured product attributes in our setting, we use a two-step approach: (i) applying the dimension-reduction algorithm (i.e., producing embeddings) and (ii) estimating the random coefficient logit model. In Appendix A, we provide the detail of this approach. Then, we compare the estimation results with those from above.

5.2 Entry Model

For the supply-side, we consider a multi-stage model in which a firm makes an entry decision in the first stage, followed by a positioning decision in the second stage, and a pricing decision in the third stage. This model serves as an empirical counterpart to the theory of spatial location choice, pioneered by Hotelling (1929) and Salop (1979), but with some key differences. First, we do not make any assumptions on the topology of spatial competition, such as linear or circular shapes. Instead, we model space competition in a characteristics space constructed from the neural network embeddings. Second, we do not assume symmetry among firms and their equilibrium outcomes. Instead, we aim to estimate model primitives that reflect firm heterogeneity.

In each period t, a firm (font designer) f makes a decision on whether to introduce product k into the marketplace. We assume a firm can introduce at most one product each period. Let E_{kt} denote the entry decision ($E_{kt} = 1$ if k enters). Let J_{ft} be a portfolio of products offered by the firm f available at time t, which excludes product k that the firm is currently considering to launch or not.

The total profit $\Pi_{ft,k}$ of firm f at time t is specified as

$$\Pi_{ft,k} = \sum_{j \in J_{ft}} \pi_{jt} + 1\{E_{kt} = 1\} \left(\pi_{kt} - F_{kt} - sc_f\right), \tag{7}$$

where π_{jt} is the variable profit of product j (and similarly for π_{kt}). F_{kt} and sc_f are fixed costs of developing product k and firm-specific sunk costs for entry, respectively. The variable profit of each product is expressed as

$$\pi_{jt} = M_t s_{jt}(p_{jt} - mc_{jt}) \tag{8}$$

with s_{jt} and M_t denoting the market share and size at time t, respectively. The observed market share is mapped from the collection of characteristics via the demand function (4), with restrictions for dimension-reduction discussed there. p_{jt} and mc_{jt} are the prices and marginal costs of product j at time t, respectively. We define the market size M_t as the number of users registered in the marketplace.[13]

Fonts are digital goods that can be easily downloaded and installed in computers and on webpages at low costs.[14] Thus, we may reasonably assume that the cost of distribution is mainly driven by wholesale fees (or commissions) paid by suppliers to the platform (i.e., MyFonts.com, owned by Monotype). Therefore, we

[13] A market size for each license type and country, M_{lct}, is calculated by counting the number of registered users at country c (including consumers who only purchase free fonts) and multiplying the fraction of each license type's sales to it. The overall market share, s_{jt}, is calculated by $s_{jt}M_t = \sum_l \sum_c s_{jlct} M_{lct}$.

[14] The impact of digitization on marginal costs is pointed out by, e.g., Waldfogel (2012).

specify the marginal costs as being proportional to the price level p_{jt}:

$$mc_{jt} = c_f p_{jt},$$

where $c_f \in [0, 1]$ is a commission rate paid by the firm f, which owns product j. According to the official Monotype Partner Policy, the commission is set at 50% for third-party sellers.[15] We calibrate this value for third-party sellers and set it at 0% for Monotype's private label brands in the marketplace. This yields the variable profit function as

$$\pi_{jt} = M_t s_{jt}(1 - c_f)p_{jt}. \tag{9}$$

In addition, we model the fixed development cost F_{kt} and sunk cost sc_f in (7) as

$$F_{kt} + sc_f = g(\boldsymbol{x}_t, \nu_{kt}) \tag{10}$$

with the characteristics $\boldsymbol{x}_t$ across all products in the marketplace and i.i.d. random shocks ν_{kt}. For instance, we can specify g to be a function of only the characteristics of product k and a fixed number of nearest competitors, analogous to the dimension-reduction approach in the demand model.

In each period t, each firm makes a sequence of decisions along the following timeline: in the first stage, the firm makes an entry decision after the cost shock ν_{kt} is realized (and before the demand shock is realized). Upon entry, the firm chooses the optimal location of the product k subject to similarity constraints imposed by a copyright policy. Lastly, the unobserved demand shock (ξ_{kt}) is realized and the firm conducts pricing.

We specify the model in a backward fashion. In the final stage, a firm solves the pricing problem for given product characteristics and unobserved demand shocks:

$$\boldsymbol{p}^*_{ft} = \arg \max_{p_{jt} \in \{p_{jt}: j \in J_{ft} \cup \{k\}\}} \sum_{j \in J_{ft} \cup \{k\}} s_{jt} M_t (1 - c_f) p_{jt},$$

where we suppress the arguments of s_{jt} for simplicity. The standard first-order condition with respect to the price of product k is given by

$$\sum_{j \in J_{ft} \cup \{k\}} \frac{\partial s_{jt}}{\partial p_{kt}} M_t (1 - c_f) p_{jt} + s_{kt} M_t (1 - c_f) = 0, \tag{11}$$

[15] The policy can be accessed here: Monotype Foundry Partner Policy

which yields

$$p_{kt}^* = \left[-\frac{\partial s_{kt}}{\partial p_{kt}} \right]^{-1} \left(s_{kt} + \sum_{j \in J_{ft}} \frac{\partial s_{jt}}{\partial p_{kt}} p_{jt} \right). \tag{12}$$

By solving the pricing equation (12), we can obtain the optimal pricing function $p_{kt}^*(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t)$, where the elements of $(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t)$ are determined by the dimension-reduction restriction on (5).[16] There are some notable features. First, the optimal price is a nonlinear function of prices and observed and unobserved characteristics of all (possibly neighboring) products in the market through demand, which means that the pricing equation effectively captures competition in the marketplace.[17] Another important feature is that the price is mainly determined by the markup, generated by product differentiation. Interestingly, the equation (12) implies that if the cross price elasticity measures of own products is close to zero, the own price elasticity of the product k should be minus one: $(\partial s_{kt}(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t)/\partial p_{kt})/(p_{kt}/s_{kt}) = -1$.

In the second stage, firm f decides the positioning of product k given the optimal price p_{kt}^*. The demand shock ξ_{kt} is not realized yet, hence the firm chooses the location of k to maximize the expected profit as

$$x_k^{emb,*} = \arg \max_{x_k^{emb} \in \mathbb{S}^d} E_{\xi_{kt}} \left[\sum_{j \in J_{ft} \cup \{k\}} \pi_{jt} \right] - F_{kt} \tag{13}$$
$$\text{s.t. } ||x_k^{emb} - x_{j'}^{emb}||_2^2 \geq \underline{\mathrm{d}} \text{ for all } j' \in J_{-ft},$$

where x_k^{emb} is the embedding vector of product k lying on the d-dimensional hypersphere $\mathbb{S}^d$, $J_{-ft} := J_t \setminus \{J_{ft} \cup \{k\}\}$ is the set of products sold by j's competitors at time t, and $\underline{\mathrm{d}}$ is the similarity contraint imposed by the copyright policy. This optimization problem is similar to that of Fan (2013), yet with notable distinctions. First, the specification of similarity constraint is unique to our study, which concerns copyright policies. Second, unlike Fan (2013), we consider the maximization of expected net profit. Third, we model fixed costs as dependent on the characteristics of competing products, reflecting the cost-benefit consideration of emulating similar products.

[16]The price in USD is a default setting in MyFonts.com. According to the platform, if a firm sets prices in USD, prices in other countries are automatically set by daily exchange rates unless a firm specifies otherwise.

[17]We assume the optimal pricing equation gives a single pricing rule.

The necessary conditions for optimality are written as

$$\sum_{j \in J_{ft} \cup \{k\}} E_{\xi_{kt}} \left[\frac{\partial \pi_{jt}}{\partial x_k^{emb}} + \sum_{j' \in J_{-ft}} + \frac{\partial \pi_{jt}}{\partial p_{j'}} \frac{\partial p_{j'}}{\partial x_k^{emb}} \right] \\ + \sum_{j' \in J_{-ft}} \left[\lambda_{kj'} \left(\frac{\partial ||x_k^{emb} - x_{j'}^{emb}||_2^2}{\partial x_k^{emb}} - \underline{\mathrm{d}} \right) \right] = \frac{\partial g(\boldsymbol{x}_t, \nu_{kt})}{\partial x_k^{emb}}, \tag{14}$$

where $\lambda_{kj'}$ is a Karush–Kuhn–Tucker (KKT) multiplier for the similarity constraint imposed on k with respect to $j' \in J_{-ft}$. We assume that the expectation and partial differentiation are interchangeable.

Finally in the first stage, firm f pays sunk costs sc_f if the expected net profit is greater than zero:

$$E_{\xi_{kt}} \left[\Pi_{ft,k}(E_{kt} = 1) \right] - \Pi_{ft,k}(E_{kt} = 0) \geq 0. \tag{15}$$

This follows a revealed profit approach, which has been used by many studies in the entry game literature (e.g., Bresnahan and Reiss, 1991; Berry, 1992; Berry and Waldfogel, 1999b; Seim, 2006). The distinctive part is that we consider a product-level entry instead of a firm-level entry. Also, we do not consider reduced-form parametric specification of the profit function; instead, the profit function is determined by demand and supply-side primitives. These features can also been found in Eizenberg (2014) and Wollmann (2018).

5.3 Identification

For the identification of parameters in the demand and entry models, we use instrumental variables (IVs) and introduce conditions on them, among others. The key condition to identify the demand-side parameters is

$$E \left[\xi_{jt} | \boldsymbol{x}_t, \boldsymbol{z}_t \right] = 0, \qquad j = 1, ..., J, \tag{16}$$

where $\boldsymbol{z}_t$ is the vector of IVs across all products in the market. Valid IVs should generate shifts in prices across markets and be exogenous from the unobserved demand shock ξ_{jt}. To this end, we first use the monthly average spot exchange rates from Federal Reserve Economic Data (FRED).[18] These variations in exchange rates generate an exogenous shift in prices across countries and over time periods.

[18]Exchange rates include Euro, Britain Pound Sterling, Australian Dollar, Canadian Dollar, Swedish Krona, and Swiss Franc to USD.

Second, we also use the characteristics of competitors as IVs, namely, the "BLP instruments." According to the timing assumption of the supply-side model, the product characteristics are chosen exogenously to unobserved demand shocks. A similar idea is used in Eizenberg (2014).

To mitigate potential bias from universally applied quantity discounts in the marketplace, we restrict our analysis to transactions at base quantity levels for estimation. These transactions are exempt from quantity discounts and represent the majority of all transactions. This is supported by Table A1 in the Appendix; transactions involving one user with a desktop license and 10,000 views with a webfont license are not subject to quantity discounts and constitute 53.5% of all transactions.

Furthermore, by assuming that consumers' choice of quantity are made independently of prices and product characteristics, using the subsample should not introduce bias in estimation. This assumption remains valid even if the choice of the product itself is influenced by other unobserved factors. Specifically, for desktop and web licenses, the quantity is defined as the number of users who can use the font software and the number of website visitors, respectively. It is reasonable to argue that the number of users for desktop licenses and website visitors is likely determined independently of the font products themselves.

Next, we discuss the supply-side estimation. First, we set $\underline{d}$ in (13) be the minimum value of pairwise distances across all products, which can be regarded as the radius of the protective boundary under the current copyright regime. This simplifies the estimation step; the location choices of firms in the data become the interior solutions of (13) under the current copyright policy, allowing us to ignore the KKT multipliers due to the complementary slackness condition. In addition, the descriptive statistics in Section 3.2 and the empirical findings in Section 4 suggest that price adjustments of incumbent products are very rare. This implies that the cross products price response with respect to visual characteristics is negligible in our model. Therefore, we may assume $\partial p_{j'}/\partial x_k^{emb} = 0$ for all $j' \neq k$ in (13).

In the supply-side model, we want to identify the fixed cost function g and the sunk cost sc_f in (10). Notably, the variable profit function (8) is identified as long as the demand function is identified, given the specification of marginal costs in (5.2). Therefore, once we identify the demand function, we can treat ξ_{jt} as observables and calculate $E_{\xi_{kt}}\left[\partial \pi_{jt}/\partial x_k^{emb}\right]$ in (14). This means that we can identify the "slope" of the fixed cost function g on the right hand side of (14).

However, unlike the slope of g, we cannot point identify the sunk cost sc_f because the entry condition (15) is characterized as an inequality rather than equality. Thus, we take an approach to partially identify the sunk cost by relying on the standard revealed profit rationale for firm f. Note that (15) provides the upper bound on sc_f. In order to attain the lower bound, we simulate the location and price of potential products that was not introduced by the firm. This simulation is possible since the location choice and pricing decision do not require the knowledge of the sunk cost.

6 Estimation Results

6.1 Demand-Side

Table 4 presents estimation results for the demand-side model using the nested logit model specification discussed in Appendix A. Column (1) contains OLS estimates and Column (2) shows 2SLS estimates. First, the estimated price coefficients are more sensible with IVs; the OLS estimate is -0.003, whereas the 2SLS estimate is -0.034. Given the existence of unobserved factors that may be correlated with prices (e.g., the higher the quality, the higher the price), it seems that the IVs are effective in dealing with price endogeneity.

The nesting parameter estimates are economically significant and are also corrected by the IVs; the OLS estimate is 0.258 whereas the 2SLS estimate is 0.546. As the nesting parameter estimates are away from zero, we can conclude that within-visual-group substitution is significant in this marketplace. This is consistent to our earlier analysis of local competition in the visual characteristics space in Section 4. That is, competition between visually similar products is important for explaining substitution between products.

We also check the validity of our IVs in the first-stage, which is shown in Columns (3) and (4) of Table 4. The estimates are reasonable; the exchange rates introduce strong variation in prices and the number of designers significantly affects the nest shares. Also, it is likely that our estimation does not suffer from weak instruments as the R^2 is large. The large value of Cragg-Donald F statistic in Column (2) also corroborates this result by rejecting the null hypothesis of weak instruments.

Overall, our estimation result is reasonable, and the signs of estimates are as expected. Consumers tend to prefer fonts with a larger number of glyphs, which are correlated with the number of supported languages and characters. In

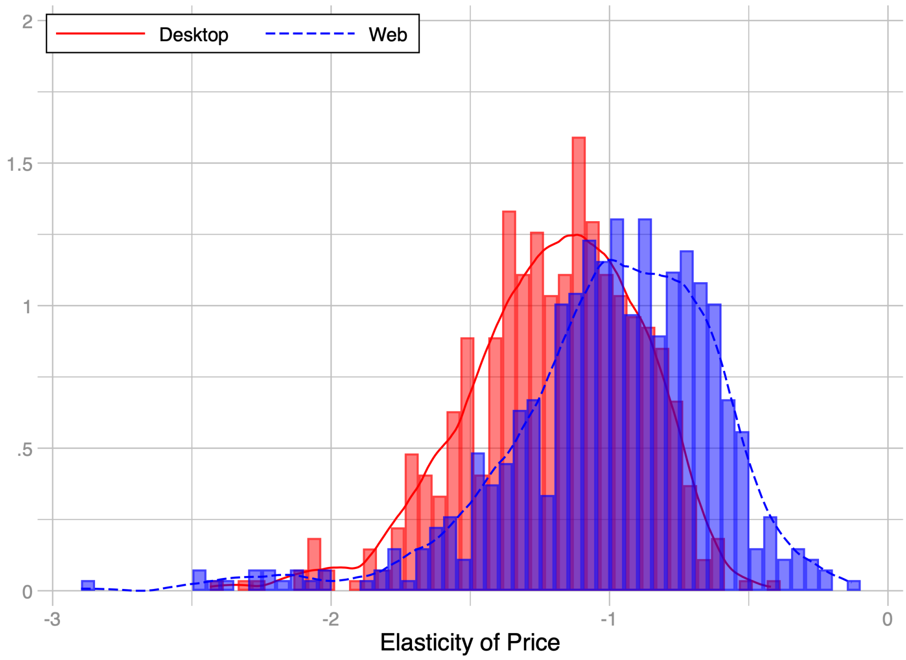
Table 4: Nested Logit Demand Estimation Results

	(1)	(2)	(3)	(4)
Method	OLS	2SLS	First-Stage	
Variables	$\ln(s_j/s_0)$		Price	$\ln(s_{j\vert k(j)})$
Price	-0.003	-0.034		
	(0.0001)	(0.005)		
$\ln(s_{j\vert k(j)})$	0.258	0.546		
	(0.007)	(0.114)		
Exchange Rate			0.258	-0.000
			(0.029)	(0.000)
Number in Cluster			0.033	-0.014
			(0.036)	(0.001)
ln(Glyph)	0.199	0.558	11.540	-0.001
	(0.012)	(0.057)	(0.409)	(0.006)
Firm FE	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
License FE	Yes	Yes	Yes	Yes
Observations	39,886	39,886	39,886	39,886
R^2	0.812	-	0.353	0.243
Cragg-Donald F-Stat	-	38.63	-	-

Notes. Heteroskedasticity robust standard errors are presented in parantheses.

addition, prices decrease utility, and visual attributes are important in explaining substitution. All the estimates are both economically and statistically significant.

Figure 6: Mean Own Price Elasticity



Notes. This figure shows mean own price elasticity of markets. A market is country and month combination. Demand estimation results for the nested logit formulation are used to calculate elasticity. The red solid line and blue dashed line indicate kernel density estimates of mean own price elasticity in desktop and web font market, respectively. The red and blue bars show its histograms in desktop and web font markets, respectively.

Figure 6 presents the distributions of mean own price elasticity estimates. Even though the pricing equation (12) is *not* imposed in demand estimation, the price elasticity estimates are consistent with the implication of the pricing equation (12) of our supply-side model. That is, the model implies that the own price elasticity should be close to minus one if the pricing response of cross products $\partial s_{jt}/\partial p_{kt}$ is close to zero. Interestingly, both distributions of price elasticity in the desktop and web font markets are dispersed around minus one.

6.2 Supply-Side

This part is in progress.

7 Counterfactual Analysis

7.1 Enforcement of Stricter Similarity Constraints

This part is in progress. We plan to vary $\underline{\mathrm{d}}$.

7.2 Reduced Fixed Costs by Generative AI

This part is in progress. We plan to change the parameters in $g(\cdot)$.

References

Anderson, S. P., De Palma, A. and Nesterov, Y. (1995). Oligopolistic competition and the optimal provision of products. *Econometrica: Journal of the Econometric Society*, pp. 1281–1301.

Balganesh, S., Manta, I. D. and Wilkinson-Ryan, T. (2014). Judging similarity. *Iowa L. Rev.*, **100**, 267.

Bellemare, M. F. and Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, **82** (1), 50–61.

Berry, S., Eizenberg, A. and Waldfogel, J. (2016). Optimal product variety in radio markets. *The RAND Journal of Economics*, **47** (3), 463–497.

—, Gandhi, A. and Haile, P. (2013). Connected substitutes and invertibility of demand. *Econometrica*, **81** (5), 2087–2111.

—, Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pp. 841–890.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica: Journal of the Econometric Society*, pp. 889–917.

— (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pp. 242–262.

— and Haile, P. A. (2014). Identification in differentiated products markets using market level data. *Econometrica*, **82** (5), 1749–1797.

— and Waldfogel, J. (1999a). Mergers, station entry, and programming variety in radio broadcasting.

— and — (1999b). Public radio in the united states: does it correct market failure or cannibalize commercial stations? *Journal of Public Economics*, **71** (2), 189–211.

— and — (2001). Do mergers increase product variety? evidence from radio broadcasting. *The Quarterly Journal of Economics*, **116** (3), 1009–1025.

Biasi, B. and Moser, P. (2021). Effects of copyrights on science: Evidence from the wwii book republication program. *American Economic Journal: Microeconomics*, **13** (4), 218–260.

Borusyak, K., Jaravel, X. and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.

Bresnahan, T. F. and Reiss, P. C. (1991). Entry and competition in concentrated markets. *Journal of political economy*, **99** (5), 977–1009.

Carroll, T. J. (1994). Protection for typeface designs: A copyright proposal. *Santa Clara Computer & High Tech. LJ*, **10**, 139.

Compiani, G. (2022). Market counterfactuals and the specification of multiproduct demand: A nonparametric approach. *Quantitative Economics*, **13** (2), 545–591.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American economic review*, **67** (3), 297–308.

Eizenberg, A. (2014). Upstream innovation and product variety in the us home pc market. *Review of Economic Studies*, **81** (3), 1003–1045.

Evans, E. N. (2013). Fonts, typefaces, and ip protection: Getting to just right. *J. Intell. Prop. L.*, **21**, 307.

Fan, Y. (2013). Ownership consolidation and product characteristics: A study of the us daily newspaper market. *American Economic Review*, **103** (5), 1598–1628.

Fodor, I. K. (2002). *A survey of dimension reduction techniques*. Tech. rep., Lawrence Livermore National Lab., CA (US).

Gentzkow, M., Kelly, B. and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, **57** (3), 535–74.

— and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, **78** (1), 35–71.

—, — and Taddy, M. (2019b). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, **87** (4), 1307–1340.

Giorcelli, M. and Moser, P. (2020). Copyrights and creativity: Evidence from italian opera in the napoleonic age. *Journal of Political Economy*, **128** (11), 4163–4210.

Glaeser, E. L., Kominers, S. D., Luca, M. and Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, **56** (1), 114–137.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, **225** (2), 254–277.

Han, S., Schulman, E. H., Grauman, K. and Ramakrishnan, S. (2021). Shapes as product differentiation: Neural network embedding in the analysis of markets for fonts. *arXiv preprint arXiv:2107.02739*.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, **124** (5), 1423–1465.

Holmes, T. J. (2011). The diffusion of wal-mart and economies of density. *Econometrica*, **79** (1), 253–302.

Hotelling, H. (1929). Stability in competition. *The Economic Journal*, **39** (153), 41–57.

Jia, P. (2008). What happens when wal-mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica*, **76** (6), 1263–1316.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, **60** (6), 84–90.

Lemley, M. A. (2009). Our bizarre system for proving copyright infringement. *J. Copyright Soc'y USA*, **57**, 719.

Li, X., MacGarvie, M. and Moser, P. (2018). Dead poets' property—how does copyright influence price? *The RAND Journal of Economics*, **49** (1), 181–205.

Lipton, J. D. (2009). To (c) or not to (c)? copyright and innovation in the digital typeface industry. *UC Davis L. Rev.*, **43**, 143.

Manfredi, T. L. (2010). Sans protection: Typeface design and copyright in the twenty-first century. *USFL Rev.*, **45**, 841.

Mankiw, N. G. and Whinston, M. D. (1986). Free entry and social inefficiency. *The RAND Journal of Economics*, pp. 48–58.

Mazzeo, M. J. (2002). Product choice and oligopoly market structure. *RAND Journal of Economics*, pp. 221–242.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, **71** (5), 1565–1578.

Oberholzer-Gee, F. and Strumpf, K. (2007). The effect of file sharing on record sales: An empirical analysis. *Journal of political economy*, **115** (1), 1–42.

Rob, R. and Waldfogel, J. (2006). Piracy on the high c's: Music downloading, sales displacement, and social welfare in a sample of college students. *The Journal of Law and Economics*, **49** (1), 29–62.

Romer, P. (2002). When should we use intellectual property rights? *American Economic Review*, **92** (2), 213–216.

Salop, S. C. (1979). Monopolistic competition with outside goods. *The Bell Journal of Economics*, pp. 141–156.

Samuelson, P. (2023). Generative ai meets copyright. *Science*, **381** (6654), 158–161.

Scheffler, S., Tromer, E. and Varia, M. (2022). Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pp. 37–49.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, **37** (3), 619–640.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spence, M. (1976a). Product differentiation and welfare. *The American Economic Review*, **66** (2), 407–414.

— (1976b). Product selection, fixed costs, and monopolistic competition. *The Review of economic studies*, **43** (2), 217–235.

Stiglitz, J. E. (2007). Economic foundations of intellectual property rights. *Duke LJ*, **57**, 1693.

Sweeting, A. (2013). Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry. *Econometrica*, **81** (5), 1763–1803.

Waldfogel, J. (2012). Copyright research in the digital age: Moving from piracy to the supply of new products. *American Economic Review*, **102** (3), 337–342.

Wang, W., Huang, Y., Wang, Y. and Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 490–497.

Wang, Y., Gao, Y. and Lian, Z. (2020). Attribute2font: Creating fonts you want from attributes. *ACM Transactions on Graphics (TOG)*, **39** (4), 69–1.

Wollmann, T. G. (2018). Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review*, **108** (6), 1364–1406.

Zeng, Z., Sun, X. and Liao, X. (2019). Artificial intelligence augments design creativity: a typeface family design experiment. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International*

Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21, Springer, pp. 400–411.

Zhang, S. (2018). *How much is an image worth? Airbnb property demand estimation leveraging large scale image analytics*. SSRN.

A Discrete Choice Consumer Model

In addition to nonparametrically estimate the demand system using neural network, we consider a parametric discrete choice model for consumers. In order to capture heterogeneous preferences on the visual attributes of fonts, we consider the random coefficient logit formulation of the indirect utility in the spirit of Berry *et al.* (1995), while applying machine learning techniques to (further) reduce the 128-dimension of characteristics, x_j^{emb}. We specify the indirect utility model (suppressing the subscripts for license type l and country c) as

$$U_{ijt} = \beta^p p_{jt} + \beta^{str} x_j^{str} + h(x_j^{emb})' \beta_i^{emb} + \xi_{jt} + \epsilon_{ijt}, \tag{17}$$
$$U_{i0t} = \epsilon_{i0t},$$

where $j = 0$ denotes the outside option that includes free open source fonts, x_j^{st} is the vector of structured characteristics, including glyph counts, x_j^{emb} is the vector of visual embeddings with $h(\cdot)$ being its dimension-reducing transformation, p_{jt} is the sales price, and ϵ_{ijt} is an i.i.d. error term following the Type I extreme value distribution.

The first approach to dimension reduction is the principal component analysis (PCA). Specifically, we specify the function $h(\cdot)$ in (17) to be a PCA transformation.[19] Then, β_i^{emb} is the random coefficents following the normal distribution $N(\bar{\beta}, \Sigma)$.

The other approach is a clustering-based method. We use a $K-$means clustering algorithm for specifying $h(\cdot)$ in (17) and then use the nested logit formulation by treating each cluster as a nest that shares the same taste shock. Let $k(j) \in \{1, ..., K\}$ for fixed K be an image cluster containing the font j, constructed by the K-means clustering algorithm.[20] In this analysis, β_i^{emb} would be random coefficients on the dummy variables indicating nests, with products in each nest sharing a common random coefficient. Moreover, the nesting parameter is common to all nests and consumers. This specification is identical to estimating a nested logit model with the same nesting parameter (Berry, 1994).

[19]Note that the PCA construction is unsupervised. To incorporate information on demand responses into dimension reduction, we can alternatively use the parital least squares, a supervised altarnative to the PCA (Hastie *et al.*, 2009).

[20]We set the number of cluster to be 500. We conduct additional robustness checks using different numbers of clusters.

The preference shock ξ_{jt} (unobserved to the analyst) is specified as

$$\xi_{jt} = \xi_f + \xi_l + \xi_c + \xi_t + \Delta\xi_{jt},$$

where ξ_f, ξ_l, ξ_c, and ξ_t are firm, license type, country, and time preference shocks affecting the mean utility level of j, respectively. For instance, ξ_f would capture the brand effects of the firm on the mean utility of good j, and ξ_c and ξ_t can capture the preferences on product j across countries and time, respectively. To deal with unobserved shocks, we include firm, country and time fixed effect terms. $\Delta\xi_{jt}$ is a deviation from these preference shocks.

B Additional Tables and Figures

Figure A1: Distributions of Log of List and Sales Prices

(a) Every Transaction (b) Transaction Without Quantity Discount

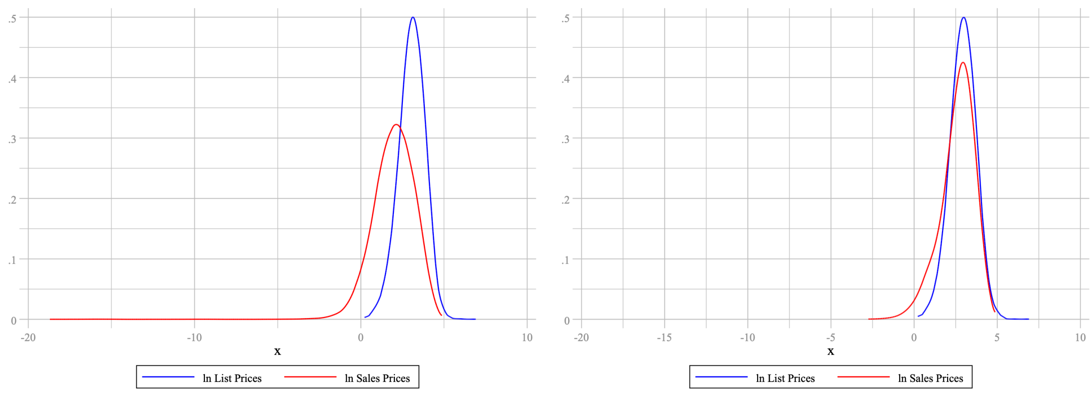
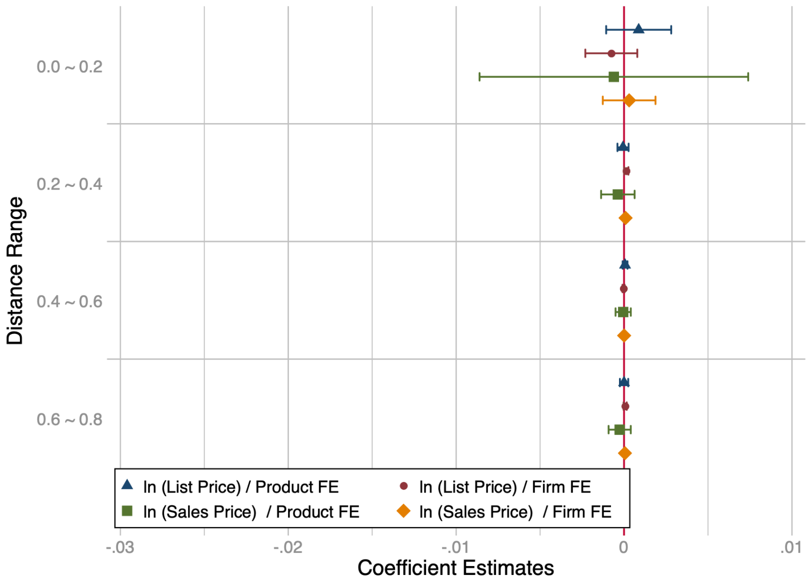


Table A1: Descriptive Table on Quantity Units in Transactions

License Type	Quantity	Frequency	Percent (%)	Cumulative (%)
Desktop	1 user	970,882	42.06	42.06
	5 users	781,630	33.86	75.91
	Others	101,596	4.40	80.31
Web	10k views	266,155	11.53	91.84
	250k views	104,101	4.51	96.35
	Others	84,207	3.65	100.00
Total		2,308,571	100.00	100.00

Notes. This table shows the number and fraction of transactions in each license type and quantity. Transactions of 12 countries and 2 license types (Desktop and Web) are used.

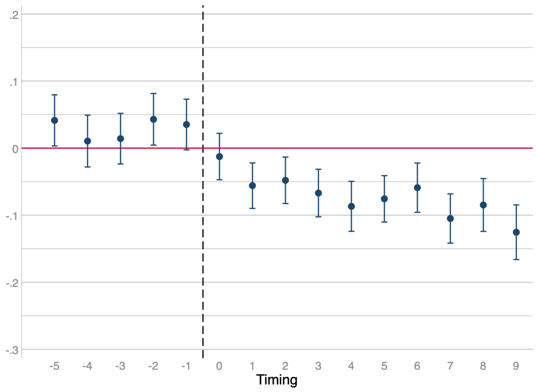
Figure A2: Spatial Regression Results (γ_r) of Prices



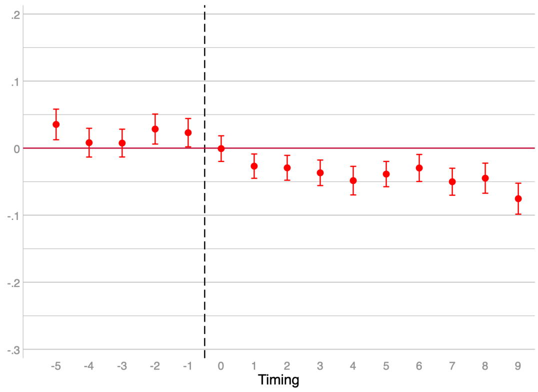
Notes. This figure present coefficient estimates of regression equation (2) for different price measures and fixed effect specifications. The triangle and circle dots indicate results from the log of list prices with product and firm fixed effects, respectively. Similarly, the square and diamond dots indicate results from sales prices. The solid lines show 95% confidence intervals. Standard errors are clustered at the product level.

Figure A3: Event Study Design (β_s) by Borusyak *et al.* (2021)

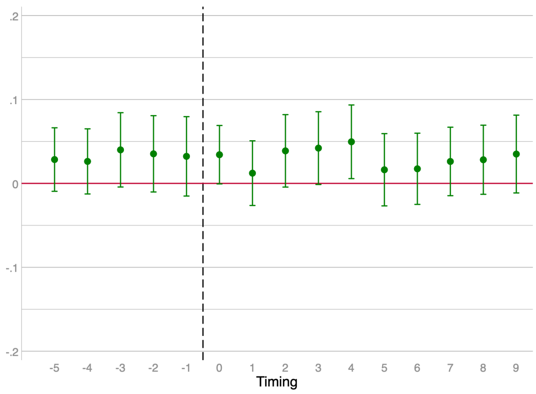
(a) Revenue



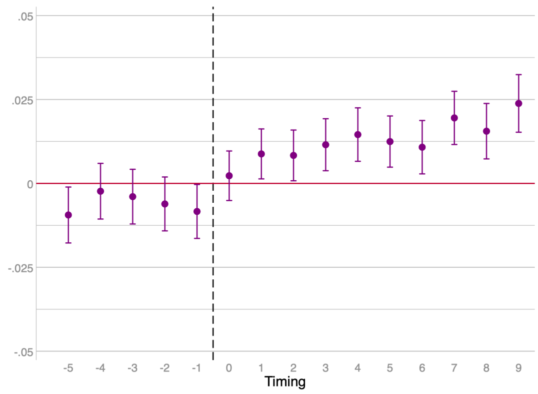
(b) Quantity



(c) Price



(d) Zero Sale



Notes. These figures show results of the event study regression in (3), which is implemented via the method by Borusyak *et al.* (2021). Panel (a) and (b) contain regression results for arcsinh transformation of revenue and quantity as a dependent variable, respectively. Panel (c) and (d) show results for log of list prices and zero sale dummy as a dependent variable, respectively. Solid lines indicate the 95% confidence intervals of estimates.

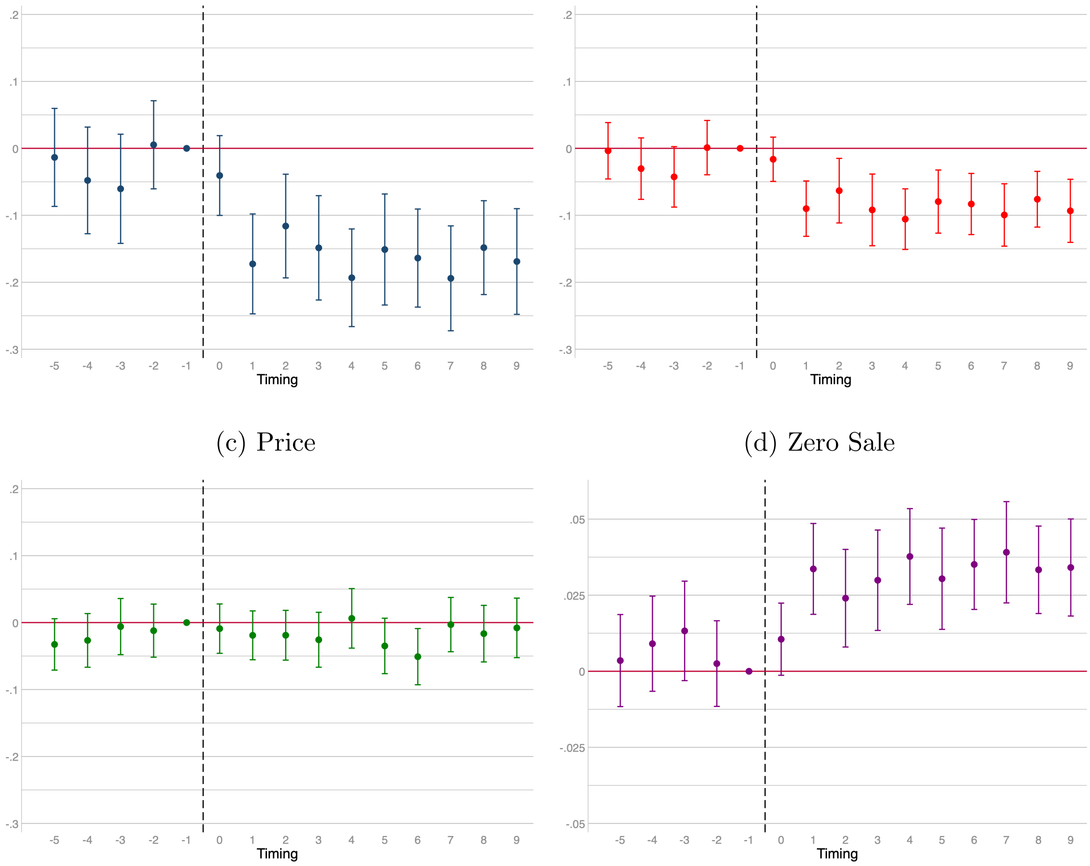
Figure A4: Event Study Design (β_s): Additional Control Variables

(a) Revenue

(b) Quantity

(c) Price

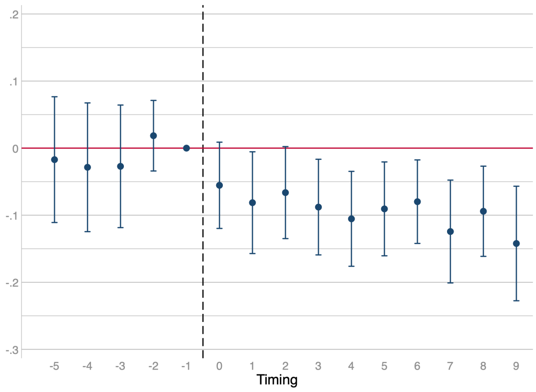
(d) Zero Sale



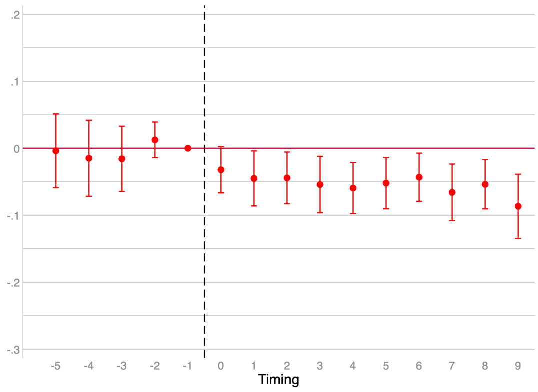
Notes. These figures show results of the event study regression in (3) with additional control variables; we include 500 image cluster and time interaction dummies, month after the introduction to the marketplace and log of glyphs. Image clusters are attained by K means clustering algorithm. Panel (a) and (b) contain regression results for arcsinh transformation of revenue and quantity as a dependent variable, respectively. Panel (c) and (d) show results for log of list prices and zero sale dummy as a dependent variable, respectively. Solid lines indicate 95% confidence intervals of estimates. Firm-level clustered standard errors are used.

Figure A5: Event Study Design (β_s): Alternative Treatment Definition (1 of 4)

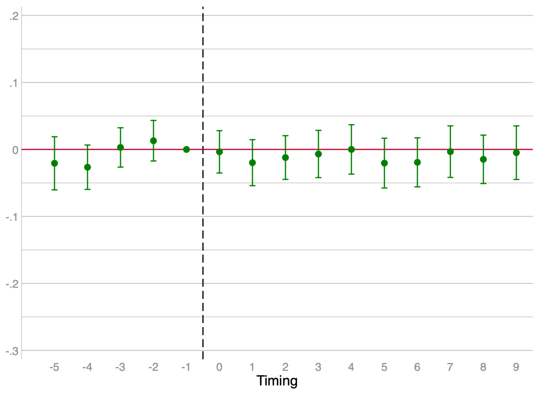
(a) Revenue



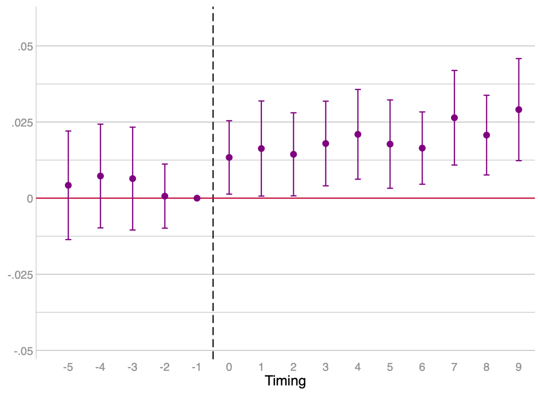
(b) Quantity



(c) Price



(d) Zero Sale



Notes. These figures show results of the event study regression in (3) with alternative treatment definition; in this figure the treatment is defined to be change in one of four closest competitors due to a new entry. Panel (a) and (b) contain regression results for arcsinh transformation of revenue and quantity as a dependent variable, respectively. Panel (c) and (d) show results for log of list prices and zero sale dummy as a dependent variable, respectively. Solid lines indicate 95% confidence intervals of estimates. Firm-level clustered standard errors are used.

Table A2: One-Way ANOVA Results (Factor: Product Dummy)

Variables	Source	SS	DF	MS	F Stats	P-value
List Prices	Factor	2.9×10^8	2,575	1.1×10^5	2.3×10^4	< 0.0001
	Residual	2.2×10^6	466,276	4.8		
	Total	2.9×10^8	468,851	611.7		
Observations				468,852		
R^2				0.9922		
Revenue	Factor	2.9×10^9	2,659	1.1×10^6	28.24	< 0.0001
	Residual	1.9×10^{10}	484,552	3.9×10^4		
	Total	2.2×10^{10}	487,211	4.5×10^4		
Observations				487,212		
R^2				0.1342		
Quantity	Factor	4.5×10^7	2,659	1.7×10^4	1.46	< 0.0001
	Residual	5.6×10^9	484,552	1.1×10^4		
	Total	5.6×10^9	487,211	1.2×10^4		
Observations				487,212		
R^2				0.0079		
Sales Prices	Model	3.6×10^7	2,659	1.4×10^4	341.69	< 0.0001
	Residual	1.9×10^7	484,552	39.9		
	Total	5.6×10^7	487,211	114		
Observations				487,212		
R^2				0.6522		

Notes. This table presents one-way ANOVA results of list price, revenue, quantity and sales price variables. SS stands for sum of square and DF stands for the degree of freedom. MS means model sum (=SS/DF). F Stats and P-value columns contain F statistics and its p-values, respectively.