Copyright and Competition: Estimating Supply and Demand with Unstructured Data[*]

Sukjin Han[†] Kyungho Lee[‡]

January 27, 2025

Abstract

Copyright policies play a pivotal role in protecting the intellectual property of creators and companies in creative industries. The advent of cost-reducing technologies, such as generative AI, in these industries calls for renewed attention to the role of these policies. This paper studies product positioning and competition in a market of creatively differentiated products and the competitive and welfare effects of copyright protection. 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 focus on a stylized design product, fonts, and use data from the world's largest online marketplace for fonts. We use neural network embeddings to quantify unstructured attributes and measure the visual similarity. We show that this measure closely aligns with actual human perception. Based on this measure, we empirically find that competitions occur locally in the visual characteristics space. We then develop a structural model for supply and demand that integrate the embeddings. Through counterfactual analyses, we find that local copyright protection can enhance consumer welfare when products are relocated, and the interplay between copyright and cost-reducing technologies is essential in determining an optimal policy for social welfare. We believe that the embedding analysis and empirical models introduced in this paper

[*] We appreciate feedback from Steve Berry, Pierre Bodéré, Kevin Chen, Tim Christensen, Liran Einav, Phil Haile, Charles Hodgson, Alessandro Iaria, Michi Igami, Lorenzo Magnolfi, Katja Seim, Jesse Shapiro, Wujin Sim, Jann Spiess, Pietro Tebaldi and participants in SweRIE 2024, APIOC 2024, and seminars at Yale and Stanford.

[†] School of Economics, University of Bristol. vincent.han@bristol.ac.uk

[‡] Department of Economics, Yale University. kyungho.lee@yale.edu

can be applicable to a range of industries where unstructured data captures essential features of products and markets.

Keywords: Copyright, creative industries, unstructured data, embeddings, visual similarity, consumer demand, product positioning.

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 and serve as gatekeepers in various sectors of the economy. This ranges 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 engage in a human-like creative process with significantly low cost, generating high-quality images, texts, sounds, and videos with scale and efficiency never seen before. This recent transformative trend thus calls for renewed attention to the role of copyright policies (Samuelson, 2023; de Rassenfosse *et al.*, 2024).

The primary goal of this paper is to study competition in a market of creatively differentiated products and the role of copyright laws, especially in the context of low-cost technologies. A common feature of products with creative elements is that their key attributes are *unstructured* and thus high-dimensional. Examples of such unstructured attributes are images and text, which are often the focus of copyright protection. Therefore, quantifying these attributes and developing an economic model based on them are crucial steps towards achieving our research objectives. This is not a trivial undertaking. Products frequently possess complex unstructured attributes that are challenging to standardize, compare and analyze. As a result, mathematically characterizing copyright policies becomes a daunting task. Another challenge is that consumers may not value unstructured attributes as much as structured ones (e.g., product specifications), which could render the consideration of copyright of creative features less relevant.

To make progress on these challenges, we focus on a specific type creative product—fonts—which provides several advantages for our study. 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, mobile applications, 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 ensuring the welfare of market participants, and (iii) the optimal level of permissible similarity (i.e., optimal variety), particularly in the presence of cost-reducing technologies such as generative AI. We use data from the world's largest online font marketplace. The data includes information on nearly 33,000 fonts (created by font design firms known as foundries) and approximately 3,000,000 transactions spanning from 2014 to 2017. To achieve our goals, we initially represent font images as embeddings—low-dimensional normalized vectors—utilizing a state-of-the-art convolutional neural network (Schroff *et al.*, 2015; Han *et al.*, 2021) and show these representation aligns well with human perception. Given the embeddings, we characterize the competition of firms in the visual dimension as a spatial competition in the embedding space, namely in the visual characteristics space.

Visual similarity, computed using the embeddings, serves as a crucial metric in our policy analyses. As detailed below, it forms the basis for modeling copyright policy, enabling us to conduct counterfactual analyses by varying policy stringency. Visual similarity is also a practically relevant concept, as real-world copyright infringement judgments are typically based on this criterion (Lemley, 2009; Balganesh *et al.*, 2014). Furthermore, many policymakers are particularly interested in regulating output similarity in the context of generative AI, because regulating AI's inputs (i.e., training data) may hamper innovation and competition.[1]

We first conduct exploratory analyses to understand the nature of competition between products. Using the panel data and two distinct strategies to measure

[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).

changes among competitors deemed close based on embedding distance, we empirically demonstrate that firms in this market engage in *local competition* in the visual characteristics space. We find that business stealing has significant and lasting impacts on sales and revenue, especially when entry occurs near the focal product. This suggests that a copyright policy providing local protection in the characteristics space would directly influence market competition and have significant welfare implications.

To study competitive and welfare effects of copyright policy, we then develop an equilibrium model of demand and supply that integrate unstructured data. On the supply side, our model describes firms' location choices within the visual characteristics space as well as pricing and entry decisions. A copyright policy is modeled as imposing restrictions on the area of possible choices in the characteristics space, providing local protection to right holders. This modeling is made feasible though the embeddings we construct. On the demand side, we characterize consumers' heterogeneous preferences for visual attributes, focusing on recovering substitution patterns across different designs. The embeddings are used as product characteristics influencing the consumer utility. Our models are general and not specific to the font market or image data; they are designed to work with any unstructured data represented as embeddings and therefore applicable to other industries with similar features.

Overall, our demand-side estimation results show that consumers tend to prefer products with high quality and functionality, prices decrease utility, and visual attributes are important determinants of consumer substitution. In particular, the estimated model reveals that the degree of competition—as captured in consumers' substitution patterns—is effectively explained by the visual similarity measured through the embeddings. The estimated supply-side model indicates that the firm's development costs are low when mimicking close competitors and increase as products become more visually differentiated.

Using the structural model, we first assess how the stringency of copyright policy affects welfare. We find that as copyright policy becomes stricter (i.e., the protection boundary around each font expands) and infringers are removed, consumer surplus decreases, primarily due to the elimination of products with attributes preferred by consumers. However, when infringers are relocated outside the protection boundary, consumer surplus increases as the area with desirable product attributes is optimally filled through relocation.

The interplay between copyright policy and cost-reducing technologies is crucial in determining the optimal level of policy strictness. We demonstrate that stricter copyright protection increases both total consumer and producer surpluses when fixed costs are low, although the relationship exhibits non-monotonic patterns. As technology advances and reduces fixed costs, stricter protection can incentivize firms to explore diverse locations where consumer valuation is high yet business stealing is not substantial. Conversely, when fixed costs are high, stricter protection may be detrimental to social welfare, as it could hinder the entry of preferred products.

This paper is structured as follows. We first discuss relevant literature and our contributions to them. In Section 2, we introduce institution backgrounds such as fonts, copyright policy, and the marketplace. In Section 3, we explain data and stylized facts about the market. We also examine the visual characteristics space and interpret embeddings. In Section 4, we analyze the firms' spatial competition in the visual characteristics space. In Section 5, we describe structural models and discuss identification and estimation of them. In Section 6, estimation results are presented. Lastly, we show counterfactual simulation results in Section 7 and conclude in Section 8.

1.1 Related Literature

This research contributes to the literature employing high-dimensional unstructured data in the economics literature. 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. Bajari *et al.* (2023) and Compiani *et al.* (2023) also use images and text to analyze markets, focusing on estimating hedonic prices and demand, respectively. Han *et al.* (2021) use product images and construct embeddings to revisit market definitions and analyze mergers. The current paper develops supply and demand models that incorporate unstructured product attributes and conducts counterfactual analyses related to copyright policies.

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.

This paper also 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 used historical quasi-experimental variations to identify the effects of the copyright system on outcomes such as price, creation and quality.[2] Copyright protection is essential for the functioning of digital markets and new technology has generated policy challenges. Existing literature has paid attentions on piracy of digital products. For example, 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.

[2] 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.

2 Backgrounds

2.1 Product Similarity and Copyright Policy

The concept of substantial similarity is fundamental in copyright infringement cases, serving as a key criterion for establishing evidence 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 then used to assess whether the "total concept and feel" of one product is substantially similar to another. As these procedure rely on human perception, the subjectivity of similarity judgments has faced criticism in the legal literature (Lemley, 2009; Balganesh *et al.*, 2014).

Copyright issues have long been significant policy questions in the font industry (Carroll, 1994; Lipton, 2009; Manfredi, 2010; Evans, 2013).[3] As fonts are ubiquitous and serve as intermediate goods for numerous final products (see below), the enforcement of copyright policy in this industry has been carefully discussed. However, there remains a lack of consensus on the most appropriate copyright policies regarding protection levels and enforcement mechanisms.

2.2 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 digital and printed materials (e.g., book covers and interiors, banners, 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. In this market, the sellers are design firms known as foundries, which specialize in font production.

The market for fonts shares several characteristics with broader markets for creative goods. First, the key product attributes in this market are unstructured. This

[3] An illustrative example of this issue is a 2001 lawsuit in the UK. GreenStreet Technologies lost a High Court case to Linotype Library for including 122 infringing fonts in its collections, including Neue Helvetica (https://www.pinsentmasons.com/out-law/news/typeface-copyright-decision-in-uk-high-court).

Figure 1: Commercial Applications of Fonts

(a) Product Packages

(b) Books and Websites

Notes. These figures show examples of commercial applications of fonts. Other examples include apps or other digital products, brand logos, newspapers, posters, pamphlets, and store signs.

feature is related to the very reason copyright policies exist in this market. Among creative products, fonts possess arguably some of the simplest visual attributes, facilitating our analysis. Second, fonts have a relatively high fixed cost associated with font creation, which includes both the design of typefaces and the development of software. This aspect of high fixed cost associated creative production is common among markets with copyright protection (Waldfogel, 2012).

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.[4] The design process often begins with the creation of a default style, which serves as the foundation from which variations are developed. Glyphs are unique characters specific to each style, and the number of glyphs in a font family is often indicative of its functionality and quality. Figure 2 illustrates the family structure and provides examples.

2.3 MyFonts.com

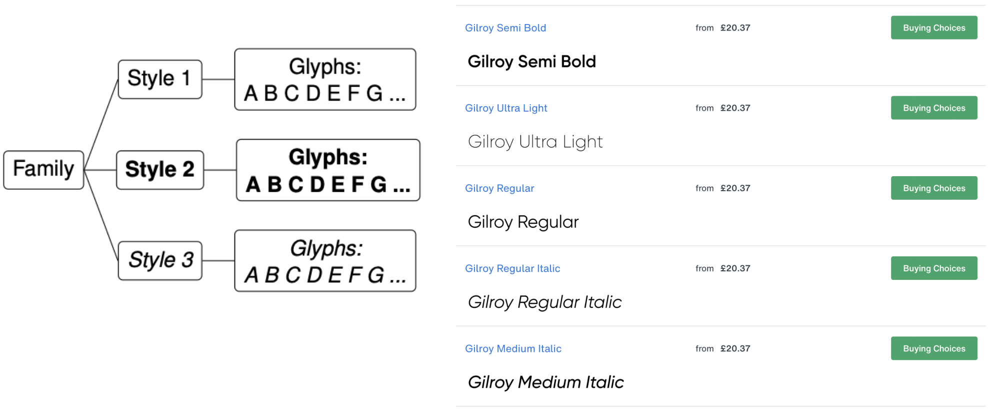
We consider MyFonts.com, the world's largest online font marketplace, which offers approximately 33,000 different fonts. This market is a superset of all major global online font stores, all of which, including MyFonts.com, are owned by Monotype Inc.[5] Panel (b) in Figure 2 presents an example of a webpage from MyFonts.com for a particular font family.

Related to our research questions, the platform owner, Monotype Inc., is highly concerned with preventing copyright infringement to maintain a well-functioning marketplace and foster competition. Their policy prohibits the acceptance and sale

[4] As detailed below, this structure becomes a key element in subsequent training data that enables us to use a triplet loss function in embedding construction.

[5] A recent *Freakonomics* podcast episode features MyFonts.com and explores the overall font industry: https://freakonomics.com/podcast/fonts/.

Figure 2: Font Family, Style and Glyphs



(a) Font Family Structure

Family
Style 1
Glyphs: A B C D E F G ...
Style 2
Glyphs: A B C D E F G ...
Style 3
Glyphs: A B C D E F G ...

(b) Example: Gilroy font family

Gilroy Semi Bold from £20.37 Buying Choices
Gilroy Semi Bold

Gilroy Ultra Light from £20.37 Buying Choices
Gilroy Ultra Light

Gilroy Regular from £20.37 Buying Choices
Gilroy Regular

Gilroy Regular Italic from £20.37 Buying Choices
Gilroy Regular Italic

Gilroy Medium Italic from £20.37 Buying Choices
Gilroy Medium Italic

Notes. Panel (a) illustrates a 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 unique characters in each style. Panel (b) presents Gilroy font family as an example.

of fonts that are suspected of plagiarizing existing products.[6] Plagiarism is mainly determined by comparing the shapes of new and existing fonts; if a new product is "nearly identical" to an existing one, Monotype regards the new font as plagiarized and prevents its listing in the marketplace.

3 Data and Embeddings

This section describes the datasets and embeddings we develop 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.[7] Each order contains information such as consumer

[6] See Monotype's policy on font plagiarism: https://foundrysupport.monotype.com/hc/en-us/articles/360029957811-Font-Plagiarism

[7] 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.

ID, transaction time, and revenue (i.e. subtotal). 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, last purchase date, and total marketplace expenditure.

In addition, we utilize firm and product data, which can be interconnected. This dataset provides important product-level information: list price, entry date, ownership, name, supported languages, number of glyphs per style, a family's default style, and product tags (i.e., short descriptive text assigned to each font family by font designers and consumer, such as "curly" and "geometric").[8] 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 (i.e., 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.[9] We define a market to be a combination of country, month and license type.[10]

3.2 Descriptive Facts about the Market

Table 1 presents descriptive statistics of the panel data. Overall, revenue, quantity, and prices are right-skewed, with large standard deviations. A notable feature is the sparsity of sales; while a few products are highly popular, the majority are sold only a few times.

List prices typically remain constant over time. From the analysis of variances (ANOVA) (Table A.2 in the Appendix), we find that product fixed effects account for approximately 99.2% of the variation in list prices. This suggests that once a price is set at the time of introduction, it remains largely unchanged over time. In contrast, other variables, such as revenue, quantity, and sales prices, vary over time; product fixed effects explain only 13.4%, 0.8%, and 65.2% of the variation in revenue, quantity, and sales prices, respectively.

The main consumers of the marketplace are small-sized businesses or individuals. The quantity units for desktop and web licenses correspond to the number of users who can install the software and the number of website views, respectively. Transactions involving one-user and five-user desktop licenses, as well as the 10,000-view web font license, account for approximately 87% of all transactions

[8] 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.

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

[10] We focus on desktop license that take majority of transactions for structural analysis. See 5.3 for more details.

Table 1: Descriptive Statistics of Panel Data

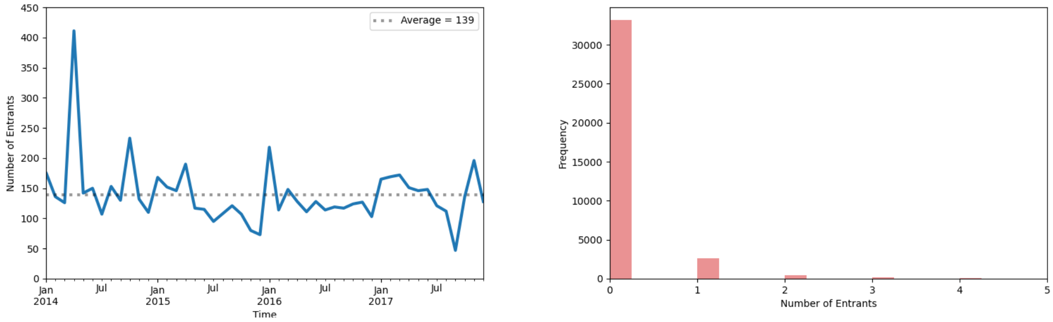
License	Variables (Unit)	Observations	Mean	Std. Dev.
Desktop	Revenue ($)	3,476,436	20.05	202.10
	Quantity (Users)	3,476,436	5.83	196.92
	Sales Price ($)	3,476,436	9.97	10.25
	List Price ($)	3,324,792	28.20	76.82
Web	Revenue ($)	989,196	17.50	205.88
	Quantity (1M Views)	989,196	3,244	191,322
	Sales Price ($)	989,196	12.44	12.31
	List Price ($)	943,176	28.34	50.55

Notes. This table contains descriptive statistics of panel data constructed from transaction records. The panel is four-way: product (font family), license type, country, and month. $ stands for United States Dollar.

(Table A.1 in the Appendix). This indicates that a significant portion of consumers are small-scale.

Differences between sales and list prices primarily arise from universal quantity discounts in the marketplace. For transactions without quantity discounts, the distributions of sales and list prices are similar, with sales prices slightly shifted to the left, possibly due to temporary discounts (Figure A.3 in the Appendix).

Figure 3: Descriptive Figures on Number of Entrants



(a) Monthly Trend

(b) Histogram

Notes. The panel (a) shows the monthly number of families that newly entered into the marketplace. The average number of entrants is 139, which is shown as the horizontal dot line. Panel (b) shows the histogram of the number of entrants across month-firm pairs.

We also examine the entry behaviors of firms, as shown in Figure 3. Panel (a) illustrates the monthly trend of entrants (i.e., new font families) listed on the platform. The trend remains relatively stable over time, fluctuating around the average level, except for a few outliers. On average, there are 139 monthly entrants. Panel (b) shows the distribution of the number of entrants across month-firm pairs, revealing that it is rare for a firm to introduce more than a single product. We leverage these findings for modeling supply-side behaviors in Section 5.2.

3.3 Images and Embeddings

In addition to the market data, we use font images as the main unstructured data, which are transformed into embeddings, a low-dimensional representation of images. More specifically, we construct embeddings for images of pangrams[11], because pangrams are what consumers typically see as product images. The embedding analysis is a machine learning method transforming high-dimensional (or unstructured) data into low-dimensional vectors of numerical numbers.[12] To construct embeddings, we follow Schroff *et al.* (2015) and Han *et al.* (2021) and use a convolutional neural network (CNN) with a triplet loss function.[13] The triplet loss function leverages the nested structure of font families by minimizing the resulting Euclidean distance between fonts within the same family and maximizing the distance between fonts from different families. This approach ensures that the resulting embedding distance corresponds to visual similarity, with fonts within the same family appearing closer together in the embedding space. Each embedding is constructed to be a 128×1 vector normalized to have a unit length and thus lies in a 128-dimensional hyper-sphere.[14] See Appendix A.1 for the details of embedding construction. The minimum and maximum Euclidean distances are 0.0002 and 0.9563, respectively, in our data.

3.4 Interpreting the Visual Characteristics Space

Our measure of visual similarity is crucial for understanding competition and welfare. In this section, we provide evidence that the constructed embeddings align well with interpretable human perception. First, we examine changes in shapes along the embedding distances to verify that the distance corresponds to visual similarity. Table 2 presents examples of fonts and images according to their distance from the focal font, Minion. As shown in this example, fonts are more visually distinct as the pairwise distance increases.

Second, we investigate how visual characteristics captured in the embeddings are aligned with human perception reflected in the product tags. To this end,

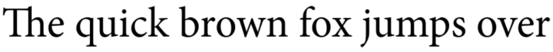
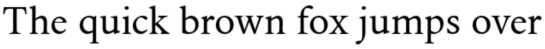
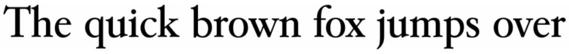
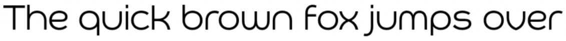
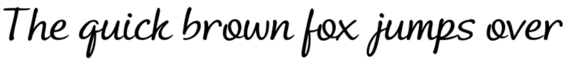
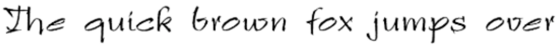
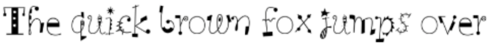
[11] A pangram is a sentence that uses every alphabet character at least once.

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

[13] 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).

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

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

Font Name	Distance	Pangram Shape
Minion	0.000	
Alia JY	0.057	
Garamond	0.081	
Bauhaus Bugler Soft	0.090	
Andrea Handwritting II	0.149	
Ruling Script	0.375	
Scruff	0.477	

Notes. This table displays examples of fonts alongside images of their corresponding pangrams (front sections). The pairwise Euclidean distances are calculated between the focal font (Minion) and each font listed in the table. The first column displays the names of the font families, the second column provides the calculated pairwise distances, and the third column exhibits the pangram images for each font. As a reference, in our data, the minimum and maximum Euclidean distances are 0.0002 and 0.9563, respectively.

we further reduce the dimension of the embeddings using Principal Component Analysis (PCA).[15] Figure 4 displays the scatter plot of Principal Components (PCs) 1 and 2 along with the sampled font shapes in various locations.[16] Through this visualization, we can confirm that similar designs are clustered together within the space, bolder shapes towards the right-hand side and more geometric shapes towards the top. The PCs, even in the two dimension, also have predictive power for the official product categories that classify shapes; see Figure A.4(a) in the Appendix.[17]

We then show that these findings are consistent with the information contained in the tags. Product tags (e.g., "bold," "serif," and "decorative") are created by both sellers and consumers, namely, they are human-labeled text. We run the Lasso regression of each of the first six PCs on tag dummies constructed from product tags.[18] Figure 5 presents the word clouds of tag dummies selected by Lasso (panels (a) and (b)), using the absolute value of the estimated coefficient as a weight, as well as top 5 tags (panels (c) and (d)). In panels (a) and (b), a tag dummy with

[15] Principal Components (PCs) have the well-known interpretation of capturing the largest orthogonal variations of the embeddings. The scree plot in Figure A.11 in the Appendix shows that most variations of the embeddings can be explained by just a few PCs, especially the first two.

[16] We display sampled fonts as showing all would be hard to visualize.

[17] In the figure (panel (b)), we also visualize the space using the Uniform Manifold Approximation and Projection (UMAP, McInnes *et al.* (2018)).

[18] We use the Lasso regression due to the large number of consumer tags (about 29,000) compared to the number of all products (about 33,000) and the variable selection feature, yielding interpretability. The details of running the Lasso regression are discussed in Appendix A.2.

a negative (positive) coefficient estimate is displayed on the left (right) side. The results suggest that PC 1 is associated with the "boldness" of the shape, consistent with Figure 4.[19] PC 2 seems to capture "display" features, namely, design features that are seen in short-form and large-format applications such as billboards or posters, headlines or headings in magazines or websites, and book covers. This is also consistent with Figure 4, because geometric shapes are common in display fonts.[20]

The interpretation of PCs 1 and 2 coincides with the patterns found in the pixel-level analysis. Figure 6 shows the pixel-level conditional mean and variance of product shapes (represented by the letter 'A') for a given range of PC values. Figure 6(a) suggests that as PC 1 increases, the font thickness also increases. This is confirmed by the low variance in the core of the letter in Figure 6(b) due to increased overlap of thick fonts. Though it initially appears that PC 2 also controls thickness in Figure 6(a), Figure 6(b) indicates otherwise; the increased size of pixel clouds is due to the increased variation of shapes. This is consistent with the fact that display fonts can come in more variety of shapes.

Finally, on a side note, each color in Figure 4 represents a different firm; it appears that no particular area is overwhelmingly occupied by a small number of firms. This aspect is taken into account when building a supply-side structural model in Section 5.2.

4 Spatial Competition in Characteristics Space

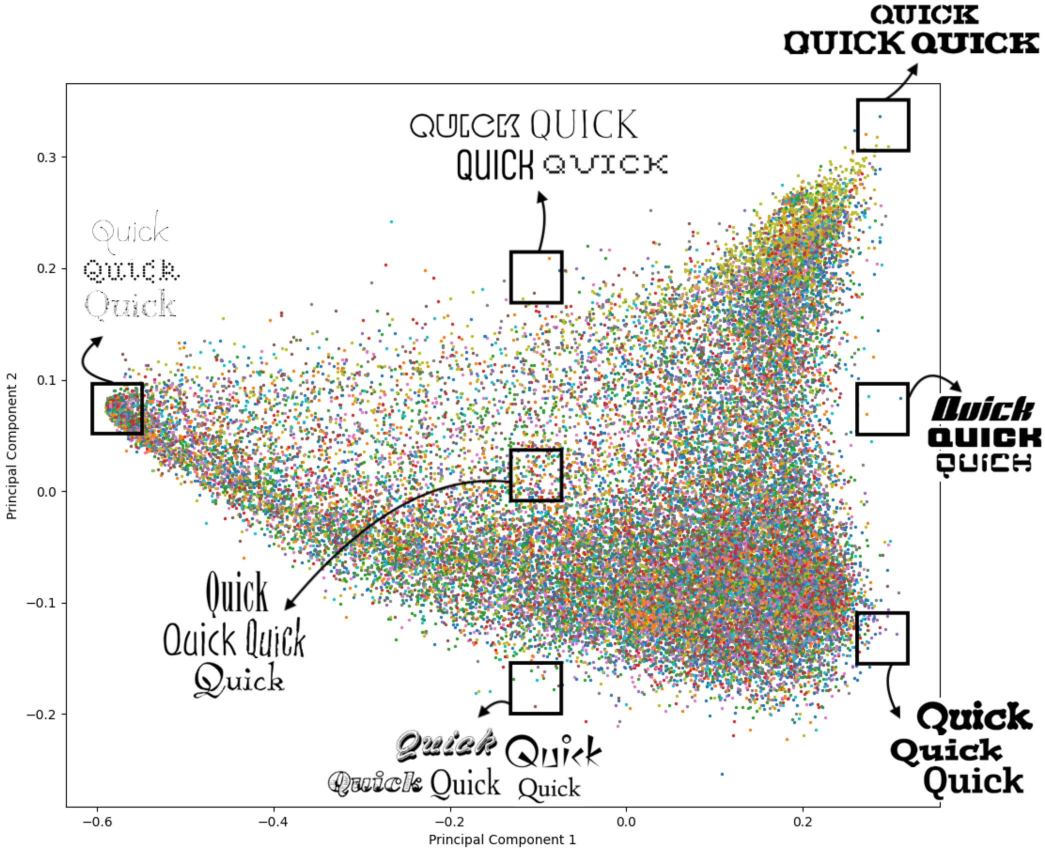
Given the neural network embeddings constructed from the image data, we can define the visual characteristics space of fonts as the subset of the embedding space. Then, each designer's design differentiation decision can be viewed as choosing a location in the characteristics space, potentially engaging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

As an initial exploratory 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 business stealing effects of entries of visually similar products on incumbents.

[19] In this exercise, we focus on the "regular" style, which are the representative style of a family. Therefore, "boldness" is a design feature inherent to the family, rather than a result of the "bold" style of the family.

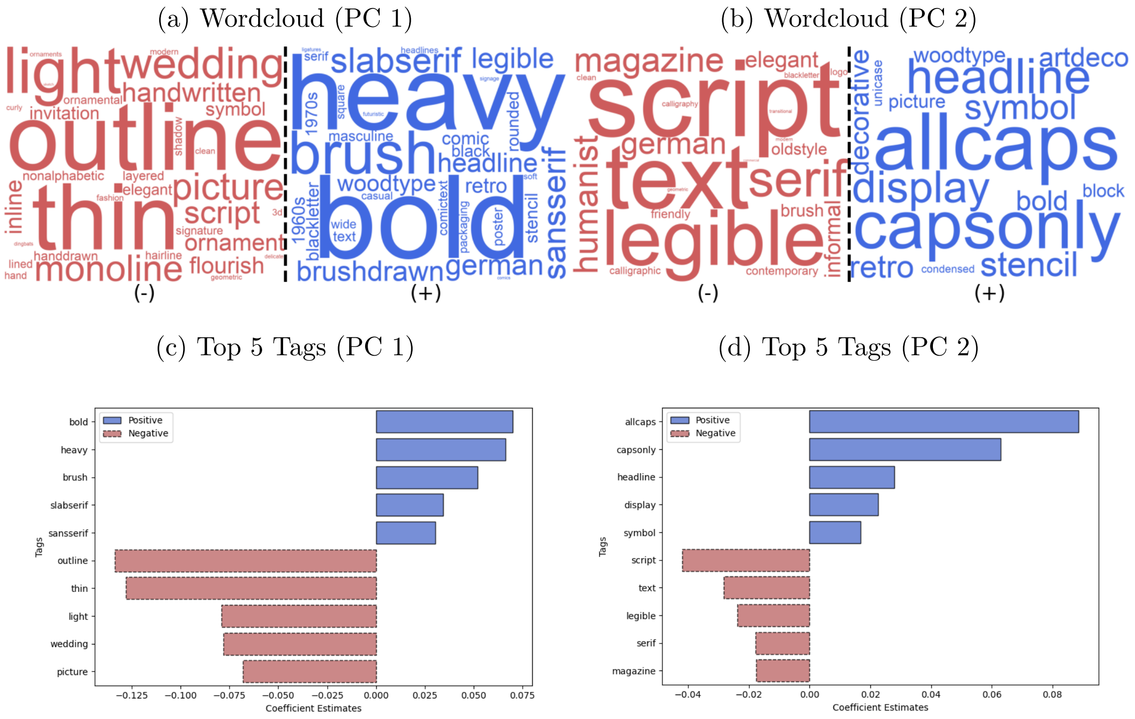
[20] We show the results for PC 3 to PC 6 in Figures A.5 and A.6 in the Appendix.

Figure 4: Scatter Plot of Principal Components 1 and 2 with Sampled Shapes



Notes. This figure presents a scatter plot of principal components 1 and 2. Shapes sampled in seven different regions are also displayed. Each color represents a different firm owning a font product.

Figure 5: Lasso Regression Results: Principal Components on Tags



(a) Wordcloud (PC 1)

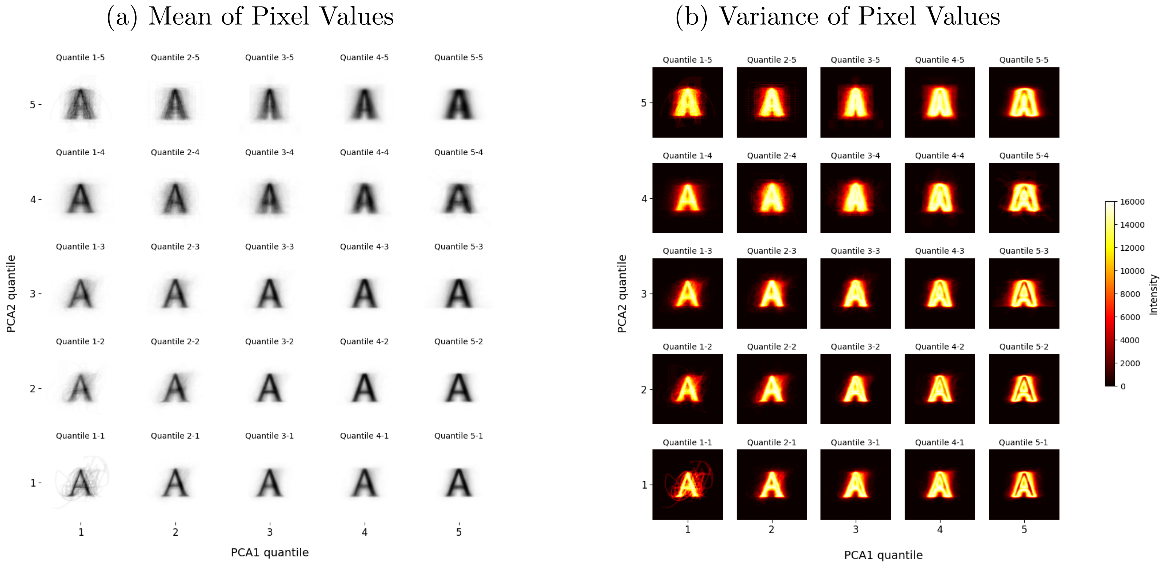
(b) Wordcloud (PC 2)

(c) Top 5 Tags (PC 1)

(d) Top 5 Tags (PC 2)

Notes. This figure presents the results of the Lasso regression for each principal component. Panels (a) and (b) display word clouds for principal components 1 and 2, respectively, with word sizes weighted by the coefficient estimates. Panels (c) and (d) show the top 5 coefficient estimates for principal components 1 and 2, respectively.

Figure 6: Pixel-Level Conditional Mean and Variance



(a) Mean of Pixel Values

(b) Variance of Pixel Values

Notes. This figure presents the means and variances of pixel values of the letter ‘A’, conditional on a range of values of principal components 1 and 2.

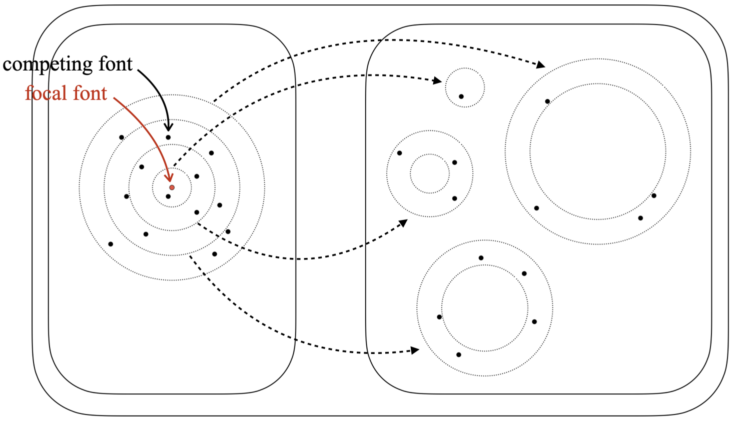
4.1 Number of Spatial Competitors and Market Outcomes

To understand the relationship between the number of spatial competitors and market outcomes, we define the number of competitors within an open ball of radius r around focal product j at time t as:

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

where $x_j^{emb} \in \mathbb{S}^{128}$ is the embedding in the 128-dimensional hypershere $\mathbb{S}^{128}$ and J_t is the set of products in period t. We then use $R_{jtr'}^{r} := (B_{jtr} - B_{jtr'})$ for $r > r'$ as a measure of the degree of spatial competition for a given distance range. The calculation of 1 is illustrated in Figure 7. Table A.3 in the Appendix presents the descriptive statistics of the number of spatial competitors. The number of competitors varies due to both cross-sectional and time-series variations, and there is significant dispersion in the number of competitors.

Figure 7: 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) located between two concentric circles with radii r and r', forming a radial area. We use the Euclidean pairwise distance.

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

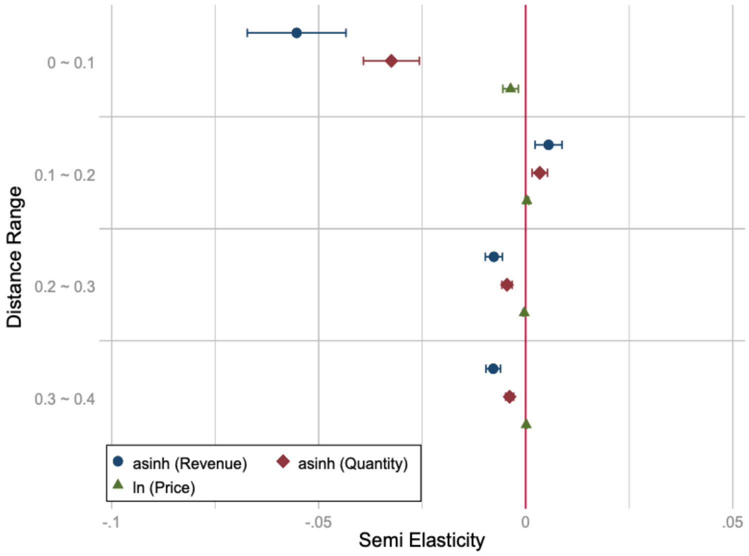
$$y_{jlct} = \gamma_{0.1} B_{jt0.1} + \sum_{r \in \{0.1,0.2,0.3,0.4\}} \gamma_r R_{jtr-0.1}^{r} + \alpha_j + \alpha_l + \alpha_c + u_{jlct}, \tag{2}$$

where y_{jlct} is the *arsinh* transformation of revenue and quantity, and the log of price.[21] Here, α_j, α_l, and α_c are product, license type and country fixed effect terms, and u_{jlct} is an error term. In this equation, γ_r captures the relationship

[21] 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,

between the number of additional competitors within a given distance range and market outcomes, controlling for fixed effects. We normalize $B_{jt0.1}$ and $R^r_{jtr-0.1}$ by dividing them by 100, which defines γ_r as the semi-elasticity for additional 100 products within a radius area.[22]

Figure 8: Spatial Regression Estimates (γ_r)



Notes. Coefficient estimates from regression (2) are presented. The radius of innermost ball is 0.1. 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 8 presents the regression results, which suggest that competition in the visual characteristics space significantly affects market outcomes and that such competition is local. The average elasticity of revenue and quantity in response to additional 100 competitors within the innermost ball is around -0.40 and -0.24, respectively.[23] Notably, these estimates are significantly larger than those for the outer rings. For prices, the coefficient estimates are near zero and not statistically significant, consistent with the fact that prices are not responsive in the market.[24]

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.

[22] Since the market comprises nearly 30,000 products, 100 products represent a relatively small portion of the total.

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

[24] When price is the dependent variable, the model may be too saturated to control for product-level fixed effects. Therefore, we use firm-level dummies instead of product dummies, which yields qualitatively similar results; see the Appendix B.

4.2 Business Stealing of Visually Similar Entrants

Motivated from the previous analysis which reveals that competition is local, we conduct an event study to estimate the causal business stealing effects of 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 a local entry. This analysis complements the previous analysis, which does not explain how the post-entry of a new product affects market outcomes. Also, the analysis in this section is free from the choice of distance cutoffs used in the previous analysis.

First, we define the treatment as an indicator for a new entry occurring within the five visually closest products. That is, the treatment indicates that there is a change in the membership of five closest competitors due to entry. Let T_j represent the first month when this treatment occurs for focal font j. 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 $s \in \mathbb{Z}$ and specify an event study design:

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

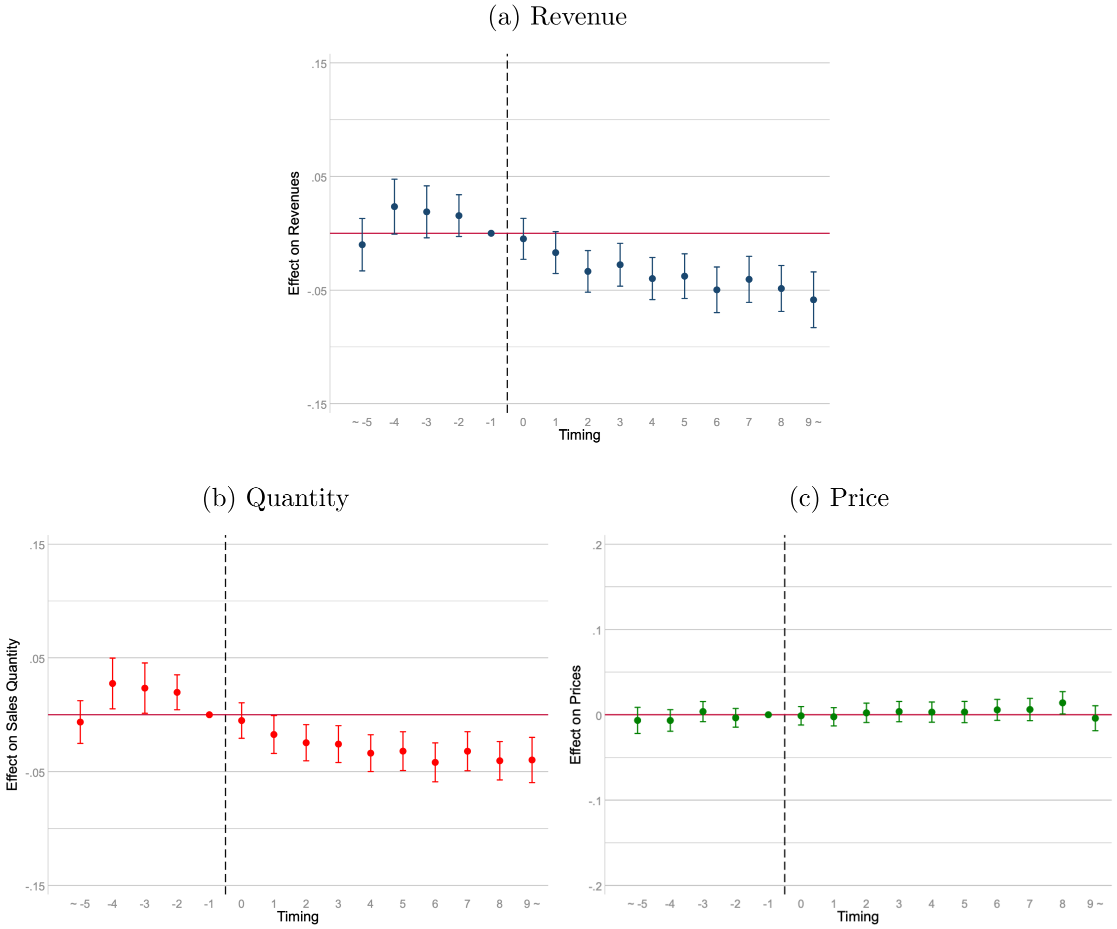
where α_f, α_l, α_c, and α_t are fixed effects of firm, license type, country, and time, respectively, and e_{jlct} is an error term. We define dependent variables as the $arsinh$ transformation of revenue and quantity, and the log of price. We normalize regression results by setting $\beta_{-1} = 0$,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

Figure 9 presents the results. 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, price is not responsive as expected. All the pre-trend coefficients are statistically insignificant, supporting the validity of the parallel trend assumption.

Due to the nature of online marketplace, products rarely exit from the market, while entry constantly occurs. This makes the treatment staggered. It is known that, with staggered adoption, OLS estimates may not be a convex combination of treatment effects on the 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.[25] The results, shown in Figure A.8 in the Appendix, are qualitatively similar, supporting that our estimates are robust to staggered timing of the treatment.

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

Figure 9: Event Study Estimates (β_s)



(a) Revenue

(b) Quantity

(c) Price

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

Results from several different specifications of model (3) show that the findings are robust.[26] First, we run the event study 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 A.9 in the Appendix, are very similar to our main findings in Figure 9. Second, we use an alternative definition of the treatment: a change in one of the *four* visually closest competitors due to a new entry. The estimation results, shown in Figure A.10 in the Appendix, are also qualitatively similar to the previous findings.

Overall, the empirical analysis of spatial competition suggests that competition is mainly local, with substantial business stealing occurring among nearby products. This suggests that a copyright policy that provides "local" protection in the characteristics space could significantly influence how the market functions. In subsequent analyses, we develop demand and supply models and apply them to evaluate the competitive and welfare effects of copyright policy.

5 Models

Guided by the empirical findings in the previous section, 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, especially in terms of the unstructured visual characteristics. The main model primitive to recover is the fixed costs of developing product design. On the demand side, the main objective is to characterize consumers' preferences over the visual characteristics, recovering substitution patterns among products. Our primary goal in building and estimating these models is to conduct counterfactual analyses to understand the role of similarity-based copyright policy and potential shifts in market fundamentals driven by technological advancements, such as the introduction of generative AI.

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

5.1 Discrete Choice Consumer Model

We introduce a discrete choice model to describe consumer behaviors. 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). The 128-dimensional characteristics, x_j^{emb}, enter the model after further

[26] We conduct robustness check for 10% randomly sampled observations due to the computational reason.

dimension reduction is applied. The indirect utility is specified (suppressing the country subscript c) as

$$U_{ijt} := \bar{\beta}^p p_{jt} + \bar{\beta}^{str} x_j^{str} + h(x_j^{emb})' \beta_i^{img} + \xi_{jt} + \epsilon_{ig(j)t} + (1-\rho)\bar{\epsilon}_{ijt} \tag{4}$$

and $U_{i0t} := \epsilon_{i0t}$, where $j = 0$ denotes the outside option that includes free open source fonts, x_j^{str} is the vector of structured characteristics, including glyph counts and a constant, x_j^{emb} is the vector of visual embeddings with $h(\cdot)$ being its dimension-reducing transformation (below), and p_{jt} is the sales price. In addition, we define nests by using the official product categories and tags that consumers use to browse products on MyFonts.com in order to account for the effects of the website's search system on consumer choices.[27] $\epsilon_{ijt} := \epsilon_{ig(j)t} + (1-\rho)\bar{\epsilon}_{ijt}$ is an i.i.d. shock, following the type I extreme value distribution. As $\rho \to 1$, substitutions would happen mostly within nests.

We assume that β_i^{img} is a random coefficient that follows the normal distribution $N(\bar{\beta}, \Sigma)$ where Σ is a diagonal matrix. To gain tractability of the random coefficient model, we assume h to have a relatively small dimension, specifying it as a principal component (PC) transformation.[28] According to the standard scree plot analysis (Figure A.11), roughly the first six PCs explain most of the variation of the embeddings, capturing about 99% of the total variation. Therefore, we choose to use a 6-dimensional vector of PCs, denoted by $x_j^{pca} := h(x_j^{emb})$.

5.2 Model for Entry and Product Positioning

For the supply side, we consider a multi-stage model in which each firm makes an entry decision in the first stage, followed by a product positioning decision subject to a copyright policy 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 (Hotelling, 1929; 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 f makes a decision on whether to introduce product k into the marketplace. We assume that each firm can introduce at most one product each period. Let $E_{ft,k}$ denote the entry decision ($E_{ft,k} = 1$ if k enters). Let J_{ft} be

[27] The official categories of fonts are coarse product categories defined by the industry; they are Serif, San Serif, Slab Serif, Script, Display, and Handwritten.

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

the portfolio of products offered by the firm f available at time t, which excludes product k that the firm currently considers launching or not.

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

$$\Pi_{ft,k} := \sum_{j \in J_{ft}} \pi_{jt} + 1\{E_{ft,k} = 1\} \left(\pi_{kt} - fc_{kt}\right), \tag{5}$$

where π_{jt} is the variable profit of product j (and similarly for π_{kt}) and fc_{kt} is the fixed costs of developing k. The variable profit of each product is expressed as

$$\pi_{jt} := M_t s_{jt}(p_{jt} - mc_{jt}), \tag{6}$$

where p_{jt} and mc_{jt} are the price and marginal cost of product j at time t, respectively, and s_{jt} and M_t are the market share and size at time t, respectively. The observed market share is mapped from the collection of characteristics via the demand function which is derived from the aggregated consumer choices in Section 5.1, incorporating the dimension-reduction restrictions discussed therein. We define the market size M_t as the number of active users registered in the marketplace.[29]

In addition, we model the fixed cost of development fc_{kt} in (5) as

$$\begin{aligned} fc_{kt} &= F(\boldsymbol{x}_t, \nu_k) && (7) \\ &:= \nu_{k0} + \sum_{\ell} \left[(\eta_{0\ell} + \nu_{k\ell}) x^{pca}_{k\ell} + \sum_{j \neq k} \left(\eta_{1\ell} d^{\ell}_{jk} + \eta_{2\ell} (d^{\ell}_{jk})^2 + \eta_{3\ell} (d^{\ell}_{jk})^3 \right) \right] && (8) \end{aligned}$$

where $d^{\ell}_{jk} := \|x^{pca}_{k\ell} - x^{pca}_{j\ell}\|_2$ for each PC dimension ℓ is the distance of incumbent product j from product k. The fixed cost is a function of the characteristics $\boldsymbol{x}_t$ across all products in the marketplace and i.i.d. random shocks $\nu_k := (\nu_{k0}, \nu_{k1}, ..., \nu_{k6})'$. We specify F to be a function of the characteristics of product k and the distances to its competitors.[30] This specification is intended to reflect a reduced fixed cost associated with the presence of visually similar products, partly due to the advantage of mimicking. However, the fixed cost does not necessarily decrease monotonically. One possible reason is that subtle differentiation from other competitors can be more costly, as it requires more sophisticated design strategies.

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}

[29] 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 at time t, s_{jt}, is calculated by $s_{jt} = \sum_l \sum_c s_{jlct} M_{lct} / M_t$. We consider a consumer to be active during the period between their first and last purchase.

[30] This specification is akin to the distance-based demand model as in Pinkse *et al.* (2002) and Magnolfi *et al.* (2022), but our problem is fundamentally different from theirs as we focus on supply-side behaviors.

is realized (and before the demand shock is realized). Upon entry, the firm chooses the optimal location of 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:

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

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

$$\sum_{j \in J_{ft} \cup \{k\}} \frac{\partial s_{jt}}{\partial p_{kt}} M_t (p_{jt} - mc_{jt}) + s_{kt} M_t = 0. \tag{9}$$

By solving the pricing equation (9), one can obtain the optimal pricing function $p^*_{kt}(\boldsymbol{p}_{-k,t}, \boldsymbol{x}_t, \boldsymbol{\xi}_t)$. 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.[31] We can recover marginal costs through the first order condition (9).[32]

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

$$x_k^{pca,*} = \arg \max_{x_k^{pca} \in \mathbb{S}^d} E_{\xi_{kt}} \left[\sum_{j \in J_{ft} \cup \{k\}} \pi_{jt} \right] - fc_{kt} \tag{10}$$
$$\text{s.t. } ||x_k^{pca} - x_{j'}^{pca}||_2 \geq \underline{\mathrm{d}} \text{ for all } j' \in J_{-ft},$$

where x_k^{pca} 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 constraint imposed by the copyright policy, forming a protective boundary of radius $\underline{\mathrm{d}}$ for incumbents. The specification of the similarity constraint is consistent with the modeling of local competition in the reduced-form analysis in Section 4, where the embedding distance between two products is used. The optimization problem in (10) is similar to that in Fan (2013), yet

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

[32] The wholesale price (or commission) between a firm and a platform is not observable. Therefore, although distributing fonts per se incurs minimal costs, we infer the marginal costs by using the pricing model instead of directly specifying them.

with key distinctions. First, the specification of similarity constraint using neural network embeddings is unique to our study, which enables us to model 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. The necessary conditions for optimality are written as

$$\sum_{j \in J_{ft} \cup \{k\}} E_{\xi_{kt}} \left[\frac{\partial \pi_{jt}}{\partial x_k^{pca}} + \sum_{j' \in J_{-ft}} \frac{\partial \pi_{jt}}{\partial p_{j'}} \frac{\partial p_{j'}}{\partial x_k^{pca}} \right] + \sum_{j' \in J_{-ft}} \left[\lambda_{kj'} \left(\frac{\partial \| x_k^{pca} - x_{j'}^{pca} \|_2}{\partial x_k^{pca}} - \underline{\mathrm{d}} \right) \right] = \frac{\partial F(\boldsymbol{x}_t, \nu_{kt})}{\partial x_k^{pca}}, \tag{11}$$

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 the fixed costs fc_{kt} if the expected net profit is greater than zero:

$$E_{\xi_{kt}} \left[\Pi_{ft,k}(E_{ft,k} = 1) \right] - \Pi_{ft,k}(E_{ft,k} = 0) \geq 0. \tag{12}$$

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 feature 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.[33]

In constructing the supply-side model, we assume that firms' forward-looking behaviors are not present, which is consistent with their observed market behaviors. One dynamic action that may be relevant for the copyright policy is entry deterrence: a firm can occupy an area in the product space to prevent the entry of competitors, leveraging copyright protection as a barrier. Nonetheless, we do not reflect this in our model, because it is unlikely that such a prevention motive is prevailing in this marketplace. This can be seen in Figure 4, where we observe no apparent areas predominantly possessed by particular firms.[34] Moreover, if one still were to develop a dynamic model, there are practical challenges. Since a firm's key strategic choice is differentiating products from those of competitors, the state

[33] These features also appear in Eizenberg (2014) and Wollmann (2018).

[34] This is consistent with the views of industry experts we interviewed. For example, Wujin Sim, the former director of Sandol, one of Korea's major font foundries, confirms that the preventive motive is minimal in font markets for at least two reasons: (i) the creative drive of designers is typically the main motive and (ii) labor-intensive font production has high costs.

space of a dynamic model is desired to incorporate the visual characteristics of *all* products in the market. Implementing this, however, is practically infeasible and extremely computationally burdensome, given the large number of products. One may consider reducing the dimensionality of the state space by limiting firms' considerations on competing products, similar to the approaches in Weintraub *et al.* (2008) and Benkard *et al.* (2015). Unfortunately, such simplification would not be ideal in our context as it restricts the visual differentiation behavior we hope to capture. Instead, our approach is to preserve the richness of firms' location choices, while adopting a static model that can be interpreted as reflecting the behavior of a myopic decision-maker.

5.3 Identification and Estimation

For the identification of parameters in the models for demand and product positioning, we use instrumental variables (IVs) and introduce related conditions. The key condition to identify the demand-side parameters is

$$E\left[\xi_{jt}|\boldsymbol{x}_t, \boldsymbol{z}_t\right] = 0, \qquad j = 1, ..., J, \tag{13}$$

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).[35] These variations in exchange rates generate an exogenous shift in prices across countries and over time periods. 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 as similarly in Eizenberg (2014). This also implies that own product characteristics are exogenous to the demand shock, hence can instrument themselves.

We use optimal IVs in the spirit of Amemiya (1977) and Chamberlain (1987).[36] We first construct differentiation IVs as in Gandhi and Houde (2019) and use them to attain estimates for approximating optimal IVs, following Berry *et al.* (1999). To be specific, we count the number of local competitors of each product j by own

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

[36] It is documented in the literature that using optimal IVs not only improves asymptotic efficiency but also may substantially increase the finite sample precision of the estimates (Reynaert and Verboven, 2014; Conlon and Gortmaker, 2020). Also, our practice of using differentiation IVs for approximating optimal IVs is known to enhance performance, especially against weak IV problem, in the literature (Gandhi and Houde, 2019).

and rival firms respectively as:

$$z_{jt\ell}^{\text{Local,Other}} = \sum_{j' \in J_{ft} \setminus \{j\}} 1\left(d_{jj'}^{\ell} < \text{SD}_{\ell}\right), \qquad z_{jt\ell}^{\text{Local,Rival}} = \sum_{j' \notin J_{ft}} 1\left(d_{jj'}^{\ell} < \text{SD}_{\ell}\right), \tag{14}$$

where $d_{jj'}^{\ell} = |x_{j'\ell}^{pca} - x_{j\ell}^{pca}|$ is the absolute difference between the PC ℓ of products j and j', and SD_{ℓ} is one standard deviation of the component ℓ. We use differentiation IVs to deal with a potential weak IV problem, which may arise from our "large" market setting in the sense of Armstrong (2016). Since there are many products in our marketplaces, using every product to construct "BLP instruments" might lead to weak identifying power. On the other hand, differentiation IVs are reported to be robust against this issue as these IVs are constructed based on *local* competitors' product characteristics instead of whole products in the marketplace (Gandhi and Houde, 2019). In addition to theoretical justification, our empirical findings of local competition in Section 4 naturally motivate the use of these IVs.

To mitigate potential bias from universally applied quantity discounts in the marketplace, we restrict our analysis to transactions involving desktop licenses at base quantity levels for estimation. These transactions are exempt from quantity discounts and represent the majority of all transactions.[37] In fact, we can reasonably assume that consumer choices regarding quantity are made independently of prices and product characteristics, as quantity here reflects the number of users. The size of consumers is likely to be exogenous and orthogonal to the font products themselves. Therefore, using this subsample should not introduce bias into the estimation.

Next, we discuss the supply-side estimation. First, we set $\underline{\text{d}}$ in (10) to 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 process, as the location choices of firms in the data become the interior solutions of (10) under the current copyright policy. This allows us to disregard the KKT multipliers due to the complementary slackness condition.

In the supply-side model, we want to identify the fixed cost function F in (7). Note that the variable profit function (6) is identified as long as the demand function is identified. Therefore, once we estimate the demand function, we can treat ξ_{jt} as residuals and calculate $E_{\xi_{kt}}\left[\partial \pi_{jt} / \partial x_k^{pca}\right]$ in (11).[38] Together with $E\left[\nu_{k\ell} | \boldsymbol{x}_t\right] = 0$ for

[37] As shown in Table A.1 in the Appendix, transactions involving a single user with a desktop license are not subject to quantity discounts and account for approximately half of all desktop license transactions.

[38] To reduce computational complexity, we additionally assume $\partial p_{j'} / \partial x_k^{pca} = 0$ for all $j' \neq k$ in (11). This assumption is supported by our previous findings. The descriptive statistics in Section 3 and the empirical results in Section 4 suggest that price adjustments of incumbent products are very rare. This implies that cross-product price responses with respect to visual characteristics are negligible in our model. Furthermore, we verified that computations under this assumption yield results very similar to those obtained with full price responses for a small sample. More

each ℓ, this in turn means that we can identify the "slope" of the fixed cost function F on the right hand side of (11).

Unlike the slope of F, however, we cannot point identify the constant term of F because the entry condition (12) is characterized as an inequality restriction. Thus, we take an approach to partially identify the constant term of the fixed cost by relying on the standard revealed profit rationale for firm f. Note that (12) provides the upper bound on the constant of F. The lower bound is assumed to be zero in the subsequent counterfactual analyses, where we examine various levels of fixed costs within the estimated bounds and report results for each specified cost level. Additional details are illustrated in Appendix A.4.

6 Structural Estimation Results

6.1 Demand-Side Results

Table 3: Fixed Coefficients Demand Estimation Results

Column	(1)	(2)	(3)	(4)
Model	OLS		IV (2nd Stage)	IV (1st Stage)
Variables	$\ln(s_j/s_0)$	$\ln(s_j/s_0)$	$\ln(s_j/s_0)$	Prices
Prices	-0.0196	-0.0207	-0.1658	-
	(0.0002)	(0.0002)	(0.0015)	-
Glyph Counts	0.0004	0.0003	0.0008	0.0034
	(0.0000)	(0.0000)	(0.0000)	(0.0001)
Ex Rate	-	-	-	0.2482
	-	-	-	(0.0024)
Constant	-8.0149	-6.8756	-5.5018	-3.8495
	(0.0084)	(0.9761)	(0.0282)	(0.2258)
Observations	225,658	225,658	225,658	225,658
PCs	Yes	No	Yes	Yes
Embeddings	No	Yes	No	No
R^2	0.0497	0.1128	-	0.1110
F stat	1192	203.8	1768	1174

Notes. This table shows results from OLS and IV regression models with fixed coefficients. Robust standard errors are in parentheses. All coefficient estimates are statistically significant at 1% level. The Cragg-Donald F statistic of IV regression is estimated to be 1266.

We first report regression results from fixed-coefficient linear models to check the validity and strength of instruments for prices. Table 3 shows the results. Columns (1) and (2) report simple OLS regression results, which respectively control for the 6-dimensional PCs and 128 dimensional neural net embeddings. The two regression

details on computing the derivatives are discussed in Appendix A.4.2.

results are qualitatively similar in terms of prices and glyph counts coefficients estimates, suggesting that PCs sufficiently control for the attributes captured in the embeddings. All the coefficient estimates are statistically significant at 1% level. The sign of the estimates all seem reasonable. As the number of glyphs (i.e., unique characters in a font family) is associated with functionality and the supported number of languages, the estimates are expected to be positive. The price coefficient estimates are all negative, although they would be biased due to the endogeneity between prices and unobserved preference shocks (e.g., quality). Columns (3) and (4) show IV regression results of the second and first stages, respectively. Again, the sign of estimates all seem reasonable. Our instruments appear to effectively address endogeneity. The estimated price coefficient of the IV regression in column (3) becomes more negative than those in columns (1) and (2). Because higher prices may be correlated with higher quality or taste shocks, this shift suggests that the instruments are valid. In addition, the Cragg-Donald F statistic is estimated to be 1266, indicating that the instruments are strong.

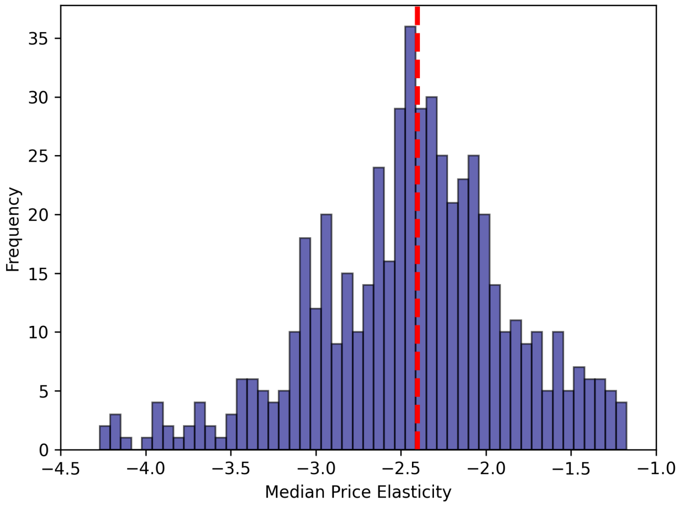
Table 4: Random Coefficents Demand Estimation Results

Variables/Parameters	$\bar{\beta}$	σ	ρ
Constant	-7.148	-	-
	(0.035)	-	-
Prices	-0.156	-	-
	(0.001)	-	-
Glyph Counts	0.001	-	-
	(0.000)	-	-
PC 1	5.292	9.500	-
	(0.082)	(0.096)	-
PC 2	-6.328	2.499	-
	(0.109)	(0.458)	-
PC 3	-11.823	7.652	-
	(0.177)	(0.209)	-
PC 4	-11.661	5.582	-
	(0.226)	(0.720)	-
PC 5	2.374	11.567	-
	(0.140)	(0.504)	-
PC 6	10.145	0.113	-
	(0.242)	(0.005)	-
Category & Tag	-	-	0.317
	-	-	(0.011)

Notes. This result shows estimation results of the random coefficient nested logit model with dimension reduction in (4). The number of observations and that of markets are 225,658 and 540, respectively. Heteroscedasticity robust standard errors are shown in parentheses. All coefficient estimates are statistically significant at 1% level.

Table 4 reports demand estimation results from our main specification, namely the random coefficient nested logit model with the dimension-reduced visual attributes in (4). For this specification, motivated from the previous estimation results, we include PCs as the product attributes, since using random coefficients on the 128-dimensional embeddings hampers the estimation process. Column "$\bar{\beta}$," "σ," and "ρ" show the estimates of mean, random, and nesting coefficients, respectively. All the estimates are statistically significant at 1% level. The signs of estimated price and glyph count coefficients are considered reasonable; an increase in price tends to decrease the mean utility, while a decrease in the number of glyphs also lowers the mean utility. The estimate for the nesting parameter (ρ) is 0.317, indicating some degree of substitution within the nest. The sign of the mean coefficient estimates ($\bar{\beta}$) of PCs is mixed. For instance, the positive and negative coefficient estimates of PC 1 and PC 2 indicate that consumers on average tend to prefer bolder fonts and display fonts, conditional on the other characteristics. However, the random coefficient estimates exhibit significant heterogeneity in preferences on product shape. Figure 10 displays the distribution of the median own-price elasticity estimates across markets, with each median calculated across products. The median elasticities are centered around -2.4. Overall, the demand estimation results appear economically meaningful and reasonable.

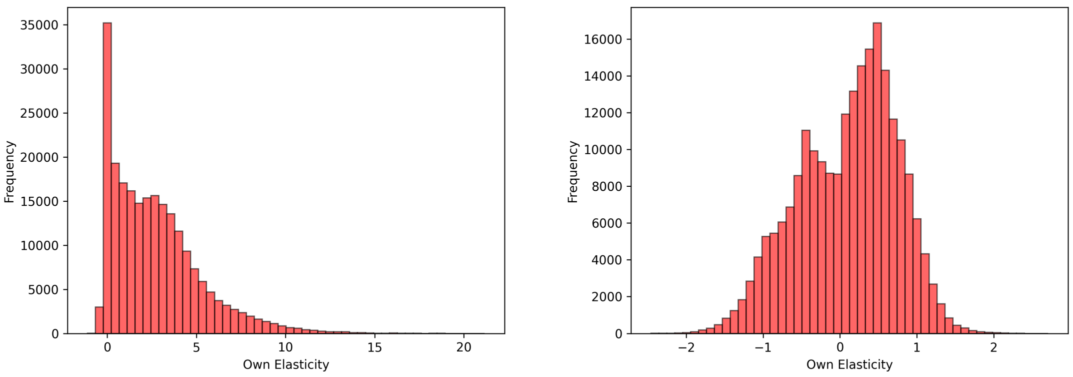
Figure 10: The Distribution of Median Own Price Elasticity



Notes. This figure shows median own-price elasticities across markets. A market is a country-month combination. The red dashed vertical line indicates the median value of the entire market.

Figure 11 shows the distributions across products and markets of own elasticities with respect to the PCs, referred to as *own shape elasticity*; we focus on the first two PCs, while the plots for the rest can be found in Figure A.12 in the Appendix. The results indicate substantial heterogeneity in preferences over product shapes.

Figure 11: The Distributions of Own Shape Elasticity



(a) Principal Component 1

(b) Principal Component 2

Notes. This figure shows the distributions of own elasticity with respect to PCs. The distribution is plotted across products and markets. Panels (a) and (b) correspond to the distributions with PCs 1 and 2, respectively. The distributions correspond to PCs 3 to 6 are shown in Figure A.12.

The own shape elasticities of the first PC are predominantly positive, suggesting that this component may capture design elements that generally enhance consumer utility.

Using the demand estimates, we examine the patterns of competition in the space of visual characteristics. Specifically, we calculate various measures of competition and display them across radial areas of different distances in the embedding space, following the idea of the reduced-form analysis in Section 4.1. We employ three measures: average diversion ratios, long-run diversion ratios, and cross-price elasticity. The aggregation M_{ct} of competition measure m at market (c, t) for a given baseline distance d can be written as:

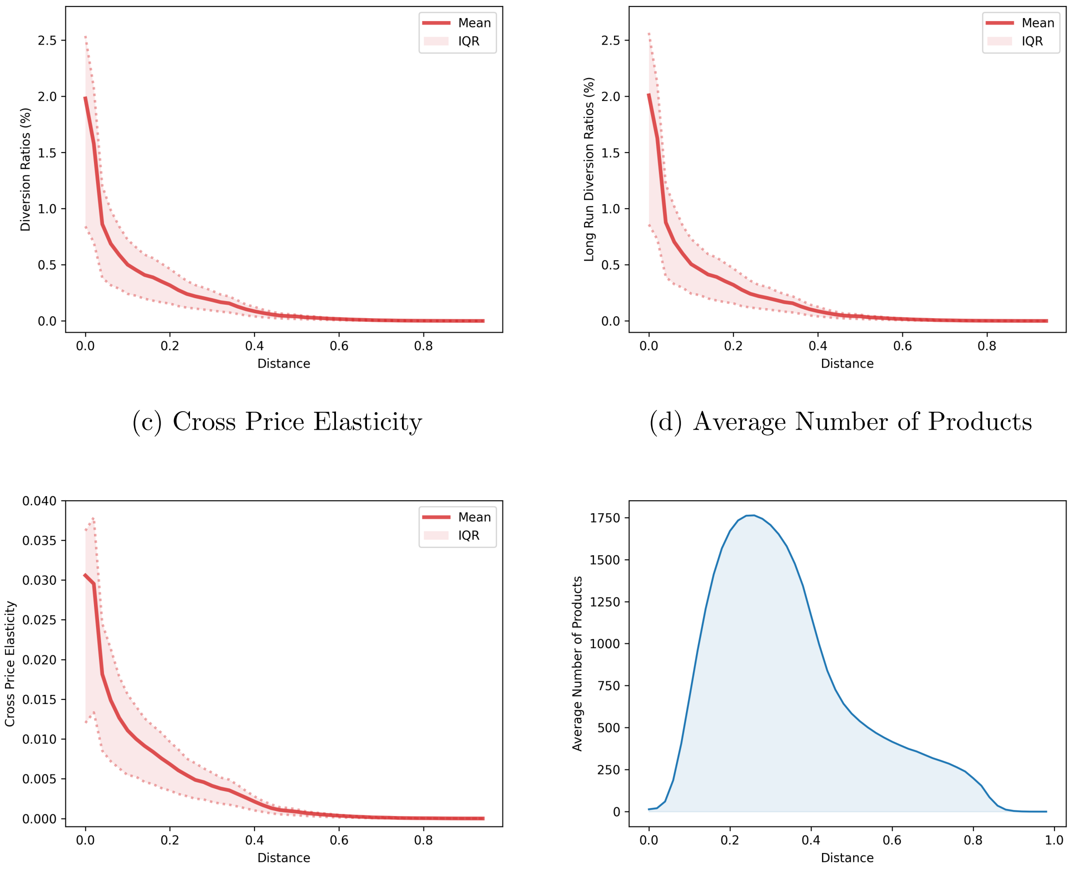
$$M_{ct}(m, d) := \frac{1}{|J_{ct}|} \sum_{j \in J_{ct}} \frac{\sum_{j' \in J_{ct} \setminus \{j\}} m_{jj'} 1\{x_{j'}^{emb} \in A_{jr}(d)\}}{\sum_{j' \in J_{ct} \setminus \{j\}} 1\{x_{j'}^{emb} \in A_{jr}(d)\}},$$

where the competition measure $m_{jj'}$ between product j and j' is either the price diversion ratio $(\frac{\partial s_{j'}}{\partial p_j} / \frac{\partial s_j}{\partial p_j})$, long-run diversion ratio diversion ratio $(\frac{s_{j'(-j)} - s_{j'}}{s_j})$, or cross-price elasticity $(\frac{p_j}{s_{j'}} / \frac{\partial s_{j'}}{\partial p_j})$, and $A_{jr}(d) := \{x \in \mathbb{S}^{128} : d \leq ||x_j^{emb} - x||_2 < d + r\}$. In the long-run diversion ratio, $s_{j'(-j)}$ denotes the equilibrium share of product j' when j is removed from the market.[39] We set $d = 0, 0.02, ..., 0.98$ and $r = 0.02$ and display the mean values and the inter-quantile ranges of $M_{ct}(m, d)$ along d.

Figure 12 presents the results. Each competition measure declines sharply as d increases, supporting the aspect of *local* competition in the visual characteristics

[39] For simplicity, we simulate shares holding the prices fixed as in the data.

Figure 12: Measures of Competition and Embedding Distances



(a) Prices Diversion Ratios

(b) Long Run Diversion Ratios

(c) Cross Price Elasticity

(d) Average Number of Products

Notes. Panels (a) to (c) plot the average price diversion ratios, long-run diversion ratios, and cross-price elasticity along the radial areas A_{jr}, respectively. As a reference, Panel (d) shows the average number of products along radial areas.

space. For instance, Panel (a) shows that, as price increases, consumers are more likely to switch to visually similar competing products. A similar pattern is observed in Panel (b), where consumers shift to close competitors when a product is removed from the market. The cross-price elasticity shown in Panel (c) follows a similar trend, although the magnitude of the decrease is relatively modest (e.g., a 1% price increase results in only about a 0.03% price increase in close competitors). These findings corroborate the presence of local competition and inactive pricing responses documented in Section 4. Furthermore, recall that we employ dimension reduction through the PCA for demand estimation, while the radial distances are calculated using the original embeddings. Figure 12 shows that, despite this additional transformation, the demand estimates remain economically meaningful, confirming that the PCs effectively retain the demand-relevant information of our embeddings. In Appendix A.3, we also find similar patterns for the substitution to outside goods, which includes free fonts—the main competitors of commercial font products.

Incorporating visual characteristics as observables in the demand model is crucial for capturing local competition. To validate this, we recalculate the same competition measures along radial areas in the embedding space by estimating the demand model *without* the visual attributes (Figure A.13 in the Appendix). This specification only incorporates the structured attributes of products via the constructed nests. The competition measures estimated without the PCs are substantially lower than those with the PCs, and they do not exhibit the sharp decline seen in Figure 12, remaining flat regardless of the distance increase. This result is not only inconsistent with the empirical findings of local competition in Section 4, but also contradicts industry experts' common understanding of competition.

6.2 Supply-Side Results

According to the supply-side estimation results, position in the characteristics space and distances to incumbents are significant determinants of fixed costs. This is evident from the slope estimates of the fixed-cost function shown in Table 5. Across all regressions involving different PCs, the estimates for the coefficients $\eta_{0\ell}$ on the distance to incumbents are statistically significant at the 1% level, indicating that location in the characteristics space is an important determinant of fixed costs. While the individual estimates of $\eta_{1\ell}$ to $\eta_{3\ell}$ are not always statistically significant, they are jointly significant at the 1% level in all regressions except for PC 4. To assess their joint significance, we conduct a Wald test with the null hypothesis $H_0 : \eta_{1\ell} = \eta_{2\ell} = \eta_{3\ell} = 0$. The results of the test are reported in the "F-stat" row.

Next, using the estimation results, we study how proximity to existing products in the characteristics space affects the cost of developing a new product. Specifically,

Table 5: Slope Estimation Results

Parameters	(1) $\partial F/\partial x_1^{pca}$	(2) $\partial F/\partial x_2^{pca}$	(3) $\partial F/\partial x_3^{pca}$	(4) $\partial F/\partial x_4^{pca}$	(5) $\partial F/\partial x_5^{pca}$	(6) $\partial F/\partial x_6^{pca}$
$\eta_{0\ell}$	3400.8	-2916.9	-6742.9	-5549.6	1699.0	5158.2
	(218.35)	(100.01)	(292.15)	(185.88)	(104.96)	(182.42)
$\eta_{1\ell}$	0.15	0.12	0.41	-0.04	0.10	-0.19
	(0.06)	(0.05)	(0.28)	(0.16)	(0.15)	(0.10)
$\eta_{2\ell}$	0.41	-0.34	-3.61	0.23	0.43	2.59
	(0.17)	(0.29)	(2.41)	(1.99)	(2.23)	(1.62)
$\eta_{3\ell}$	-0.22	0.57	10.85	-0.28	-0.98	-6.65
	(0.14)	(0.48)	(5.96)	(6.82)	(8.92)	(7.00)
R^2	0.33	0.06	0.07	0.00	0.25	0.01
F-stat	271.96	33.71	40.66	0.73	177.85	4.44
Observations	1,630					

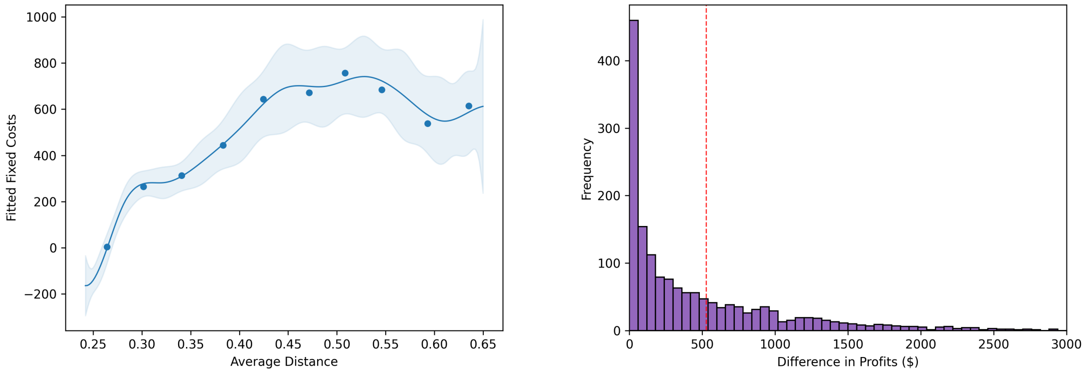
Notes. This table shows the estimated slopes of the fixed cost function. Heteroskedasticity robust standard errors are shown in the parentheses. F-statistics of Wald test on $H_0 : \eta_{1\ell} = \eta_{2\ell} = \eta_{3\ell} = 0$ v.s. H_1 : the negation of H_0 are shown in the F-stat row. The number of observations is 1,630, which correspond to the number of entrants.

we examine the relationship between the fitted values of fixed costs and the average distances to other competitors. Note that the shape of this relationship is identified from the slope coefficients estimated above. In Figure 13(a), the explained part of the estimated fixed costs is plotted against the average distance in the embedding space.[40] The results show an initial rise in fixed costs as the average distance increases, but the trend flattens around an average distance of 0.45. Beyond this point, average distances seem to have a relatively minor impact on fixed costs. This indicates that having close competitors appears to reduce development costs, although the relationship is not strictly monotonic.

Lastly, we present the estimates of the upper bound estimates on the fixed costs. Figure 13(b) shows the histogram across entrants of the differences between the expected variable profits upon entry and without entry in the entry condition (12), which exhibits reasonable patterns. The distribution is right-skewed with many values close to zero. This pattern is consistent with the observed product revenue distribution.

[40] Some fitted values can be lower than zero as we only focus on the explained part of the fixed costs.

Figure 13: Fixed Costs Estimation Results



(a) Fixed Costs and Distances

(b) Upper Bound Estimates

Notes. Panel (a) presents a binscatter plot of the estimated explained part of fixed costs against average distances. We compute the fitted values by using the estimated slope coefficients which are displayed in Table 5. The solid line represents a third-order local polynomial fit, with the shaded area indicating the 95% confidence band. Heteroskedasticity-robust standard errors are applied. The dots represent bin-by-bin averages with the evenly-spaced binning method (Cattaneo *et al.*, 2024). Panel (b) shows the histogram of the estimated differences between the expected and no-entry variable profits indicating the upper bound of the fixed costs, as defined in (12). The red dashed line represents the average value, 523. The standard deviation of the differences is 772. Both figures are based on 1,630 observations, corresponding to the number of entrants.

7 Counterfactual Policy Analyses

7.1 Enforcing Stricter Copyright Policies

In this section, we investigate the role of copyright policy in competition and welfare. As the first counterfactual analysis, we increase the degree of copyright protection to understand its impact on welfare. Given the similarity constraint in the product positioning equation (10), this can be done by increasing the protective boundary radius $\underline{\mathrm{d}}$. Setting a larger $\underline{\mathrm{d}}$ is interpreted as imposing stricter copyright protection in the market.

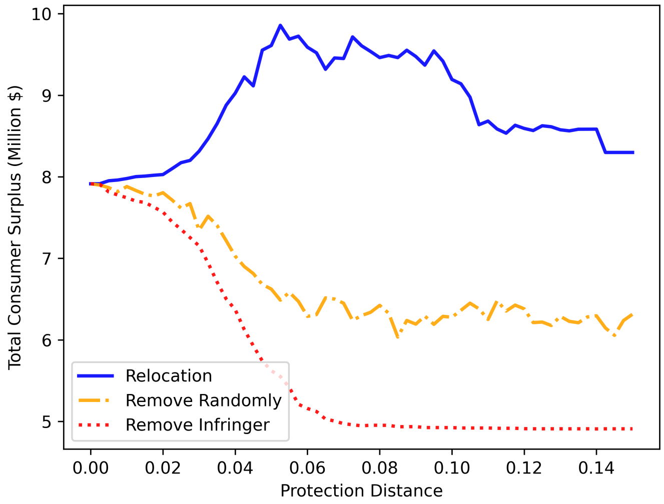
We perform two exercises: (1) a naïve simulation that removes entrants within a protective boundary (i.e., infringers) around existing products, and (2) a relocation simulation that pushes infringers outside the protective boundary, thereby rearranging the characteristic space. It is important to note that these exercises do not account for the counterfactual location choices of entrants under stricter copyright protection; counterfactual location choices are addressed in the next section. Nonetheless, they provide valuable insights. Specifically, the naïve simulation helps us understand the extent of monopolistic power that incumbents could exert as a result of increased protection and the corresponding reduction in consumer surplus due to fewer available choices in the marketplace. Additionally, the relocation simulation allows us to examine whether consumers might benefit from increased diversity in product attributes.

For these analyses, we impose the copyright policy from April 2014, the first period in our dataset, and simulate equilibrium prices and market shares based on the demand estimates and pricing model. The consumer surplus for each market ct under $\underline{\mathrm{d}}$ can be calculated as:

$$CS_{ct}(\underline{\mathrm{d}}) = \frac{1}{N_s}\sum_{i=1}^{N_s} \ln\left(1 + \sum_{j\in J_{ct}(\underline{\mathrm{d}})} \exp V_{ijct}(\underline{\mathrm{d}})\right) / (-\bar{\beta}^p), \tag{15}$$

where N_s is the number of simulated consumers and $V_{ijct}(\underline{\mathrm{d}}) := \bar{\beta}^p p_{jct}(\underline{\mathrm{d}}) + \bar{\beta}^{str} x_j^{str} + h(x_j^{emb})'\beta_i^{img} + \xi_{jct}$ and $J_{ct}(\underline{\mathrm{d}})$ is the consumers' choice set given $\underline{\mathrm{d}}$. Given (15), copyright protection can affect consumer surplus through two channels: (i) price increases due to monopolistic power would enter V_{ijt}, decreasing utilities; (ii) the reduced choice set would be reflected in $J_{ct}(\underline{\mathrm{d}})$.

Figure 14: Simulated Consumer Surplus with Copyright Protection



Notes. This figure shows changes in consumer surpluses of simulation exercises as protection distance increases. The orange dash-dot and red dot lines show the results of removing random products and infringers, respectively. The blue line presents those of pushing infringers to be outside of protection boundary.

We find that consumer surplus sharply decreases as infringers are removed with increasing protection levels. In Figure 14, the red dashed line represents the naïve simulation results. For example, under the most strict protection ($\underline{\mathrm{d}} = 0.15$), consumer surplus decreases by approximately 39% compared to the original copyright regime ($\underline{\mathrm{d}} = 0$). This change may not be solely driven by the reduced number of products; product attributes can also play an important role. To isolate this effect, we compare the change in consumer surplus from the naïve exercise with a random removal exercise. In the random removal, we eliminate a number of randomly selected products equal to the number of infringers in the naïve exercise. The random removal simulation shows a similar pattern, but consumer surplus decreases less than in the naïve exercise. Since the consumer surplus in (15) is a monotonic function of the utility value of each product, this suggests that infringers are often located in areas where utility levels are also high.[41]

In addition, consumers could benefit from the rearrangement of product locations induced by a stricter copyright policy. In Figure 14, the solid blue line illustrates an inverse U-shaped relationship between the increase in protection level and consumer surplus. Since the number of products remains constant across all copyright regimes, this result implies that a certain level of protection, such as $\underline{\mathrm{d}} = 0.05$, optimally fills the area with desirable product attributes, leading to a consumer surplus increase of approximately 24% compared to the original regime. However, as consumer surplus decreases beyond a certain $\underline{\mathrm{d}}$ level, it suggests that consumers may not necessarily prefer products with substantially differentiated attributes.

7.2 Interplay between Copyrights and Fixed Costs

In this section, we investigate the interaction between copyright protection and cost reductions driven by the industry's technological advances, such as generative AI that assists font designing. Using model estimates, we conduct simulation studies examining various combinations of fixed cost levels and protection distances ($\underline{\mathrm{d}}$). We then discuss the resulting welfare and market outcomes. These studies take into account for not only consumer decisions but also firms' optimal product positioning and pricing behaviors.

We first generate 250 potential locations by perturbing random embeddings, similar to the relocation analysis in Section 7.1, and compute the fixed cost associated with each location.[42] For the calculations, we specify three different levels of ν_{k0}, which we set to be identical across firms, and use the specified values of ν_{k0} along with the fixed cost function estimates.[43] This results in three levels of fixed costs, low, medium, and high, for investigation. The distributions of the fixed costs (shown in Figure A.14 in the Appendix) exhibit a reasonable pattern; the shape is similar to that of the upper bound estimates and revenue distributions, characterized by a long-tail. Additionally, we calculate the expected profit for each potential

[41] In Appendix A.5.2, we further decompose the decrease in consumer surplus into two component: price increases due to enhanced monopolistic power and diversity loss resulting from eliminating new entrants. We find the latter factor dominates.

[42] We assign the nest of the closest product as the nest for a potential entrant. This is because tags and search menus are based on shapes and functionalities that are reflected in our embeddings.

[43] There are some values of fixed costs that are below zero after adding ν_{k0}. We treat their development costs being negligible and set them to be zero.

location using the demand estimates. For exogenous characteristics and marginal costs of a given firm, we use the corresponding average values across products of that firm. Finally, we allow firms to search for the optimal location and decide on entry.[44]

Simplifications are necessary in simulating firms' decisions, as computationally solving the full model is extremely burdensome. Simulating a single market requires evaluating each combination of potential entrants, firms, and sampled demand shocks ξ_k to calculate the expected profits for all potential products and firms. Consequently, computation time increases rapidly as any of these elements grow. To address this issue, we implement the following three steps. First, we consider a random sample of 100 firms that sequentially decide on entry and product positioning. Second, we impute the expected value of entrants' demand shocks ξ_k to approximate the expectation, following a similar approach to Berry *et al.* (1999). To mitigate concerns about potential approximation errors, we compare simulated expected profits obtained using this imputation method with those based on the empirical distribution, confirming that they yield similar outcomes. Third, we focus our analysis to markets in April 2014, the first period in our data sample.

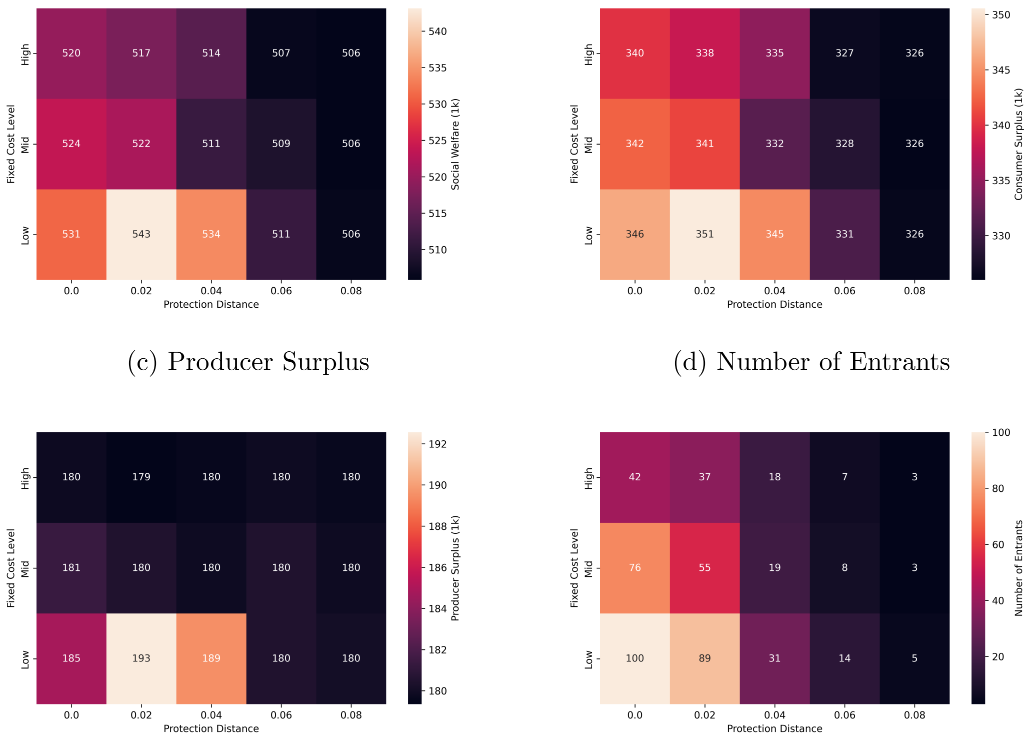
We use standard welfare measures. Producer surplus, PS_t, at time t is defined as the sum of total profits across firms (i.e., across old and new products), including the fixed costs associated with introducing new products k's: $PS_t := \sum_{f,k} \Pi_{ft,k}$ where $\Pi_{ft,k}$ is the total profit defined in (5). Then, social welfare is defined as $SW_t = PS_t + \sum_c CS_{ct}$, where CS_{ct} is the consumer surplus defined in (15).

Figure 15 presents the simulation results. As shown in Panel (a), the introduction of copyright protection increases social welfare when fixed costs are low. However, the relationship between the stringency of protection and welfare is not monotonic, and there exists an "optimal" level of permissible similarity. In Panel (a), social welfare is highest when the protection distance is 0.02 and diminishes as the distance increases. This increase in social welfare is driven by rises in both consumer and producer surpluses. Stricter protection results in more desirable products from the consumer's perspective, despite the smaller number of entrants shown in Panel (d). Additionally, total profits, or producer surplus, is optimized at the distance 0.02 because entrants are located in areas where the business-stealing effect is less pronounced, yet consumers find the products more appealing. In contrast, stricter protection may not be beneficial if fixed costs are significant. At both medium and high fixed cost levels in Figure 15, social welfare decreases as the level of protection increases.

These results suggest that the interplay between copyright policy and cost-

[44] The data show that entrants often have zero market share in multiple countries during their entry month. To compute the expected profits of a potential entrant, we calculate the probability of being visible in a given country and include the entrant in the choice set.

Figure 15: Simulation Results by Varying Fixed Costs and Protection Levels



(a) Social Welfare

(b) Consumer Surplus

(c) Producer Surplus

(d) Number of Entrants

Notes. This figure shows the counterfactual welfare and market outcomes under various combinations of fixed cost levels and protection distances. Panels (a) to (d) show social welfare, consumer surplus, producer surplus and number of entrants, respectively. A welfare outcome of each simulation is displayed in the corresponding cell, expressed in 1K USD.

reducing technologies is essential in determining the optimal level of policy strictness. As technology advances and reduces fixed costs, stricter protection can incentivize firms to explore diverse locations where they engage in less business stealing but achieve high profits due to increased consumer valuation.

8 Conclusion

In this paper, we study the role of copyright policy in a creative industry and its interaction with cost-reducing technologies. We combine a state-of-the-art embedding analysis for unstructured data with structural economic models to address the policy question. Our focus is on the global font marketplace, which has unique features well-suited to our research purposes. We document localized competition among firms in the characteristic space and the business-stealing effects caused by visually similar entrants. We then develop a model of supply and demand that captures firms' entry and positioning behavior within the visual characteristics space, as well as consumers' heterogeneous preferences for visual attributes. Our counterfactual analysis suggests that the stringency of copyright policy could significantly affect welfare through changes in product diversity and potential improvements driven by product relocation. Moreover, it highlights the importance of considering the interplay between copyright protection and technological advancements when determining the optimal level of policy stringency. We believe that the counterfactual policy analyses performed using the proposed empirical framework can offer a scientific reference for policymakers in making copyright infringement judgments.

The growing availability of unstructured data and machine learning tools is motivating new economic and policy questions. In certain contexts, a structural approach that integrates such data is essential for addressing both positive and normative aspects of an economy and its policies. The empirical models presented in this paper are not confined to our research setting; we believe they are broadly applicable to a wide range of industries where unstructured data can capture important features of products and markets. One important question in using embeddings for economic research is whether the embedding representation captures context-specific economically relevant features (e.g., substitution patterns, local competition), while maintaining general interpretability (e.g., distance, visual similarity). In this paper, we demonstrate how this question can be explored from various angles. It remains valuable to addresses this question more systematically in the current and various other contexts. Overall, we hope this paper serves as an illustration of how structural models can fruitfully leverage unstructured data, potentially inspiring important and exciting research questions more in the future.

References

AMEMIYA, T. (1977). The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica: Journal of the Econometric Society*, pp. 955–968.

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.

ARMSTRONG, T. B. (2016). Large market asymptotics for differentiated product demand estimators with economic models of supply. *Econometrica*, **84** (5), 1961–1980.

BAJARI, P., CEN, Z., CHERNOZHUKOV, V., MANUKONDA, M., VIJAYKUMAR, S., WANG, J., HUERTA, R., LI, J., LENG, L., MONOKROUSSOS, G. *et al.* (2023). Hedonic prices and quality adjusted price indices powered by ai. *arXiv preprint arXiv:2305.00044*.

BALGANESH, S., MANTA, I. D. and WILKINSON-RYAN, T. (2014). Judging similarity. *Iowa Law Review*, **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.

BENKARD, C. L., JEZIORSKI, P. and WEINTRAUB, G. Y. (2015). Oblivious equilibrium for concentrated industries. *The RAND Journal of Economics*, **46** (4), 671–708.

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

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

—, — and — (1999). Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review*, **89** (3), 400–431.

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

— and WALDFOGEL, J. (1999a). Mergers, station entry, and programming variety in radio broadcasting. *NBER Working paper: w7080*.

— 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.

Cattaneo, M. D., Crump, R. K., Farrell, M. H. and Feng, Y. (2024). On binscatter. *American Economic Review*, **114** (5), 1488–1514.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34** (3), 305–334.

Compiani, G., Morozov, I. and Seiler, S. (2023). Demand estimation with text and image data. *Available at SSRN: 4588941*.

Conlon, C. and Gortmaker, J. (2020). Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics*, **51** (4), 1108–1161.

de Rassenfosse, G., Jaffe, A. B. and Waldfogel, J. (2024). Intellectual property and creative machines. *NBER Working paper: w32698*.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *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. *Journal of Intellectual Property Law*, **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.

Gandhi, A. and Houde, J.-F. (2019). Measuring substitution patterns in differentiated-products industries. *NBER Working paper: w26375.*

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 Law Review*, **43**, 143.

Magnolfi, L., McClure, J. and Sorensen, A. T. (2022). Triplet embeddings for demand estimation. *Available at SSRN: 4113399*.

Manfredi, T. L. (2010). Sans protection: Typeface design and copyright in the twenty-first century. *USF Law Review*, **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. *The RAND Journal of Economics*, pp. 221–242.

McInnes, L., Healy, J. and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Morrow, W. R. and Skerlos, S. J. (2011). Fixed-point approaches to computing bertrand-nash equilibrium prices under mixed-logit demand. *Operations research*, **59** (2), 328–345.

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pinkse, J., Slade, M. E. and Brett, C. (2002). Spatial price competition: a semiparametric approach. *Econometrica*, **70** (3), 1111–1153.

Reynaert, M. and Verboven, F. (2014). Improving the performance of random coefficients demand models: The role of optimal instruments. *Journal of Econometrics*, **179** (1), 83–98.

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.

Schroff, F., Kalenichenko, D. and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

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

Weintraub, G. Y., Benkard, C. L. and Van Roy, B. (2008). Markov perfect industry dynamics with many firms. *Econometrica*, **76** (6), 1375–1411.

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.

A Additional Results and Details

A.1 Embedding Construction

To construct embeddings, we train convolutional neural network with triplet loss. Triplet i comprises anchor x_i^a, positive x_i^p, and negative x_i^n. An anchor is (crops of) a pangram image of a given font family (e.g., Helvetica), positives are (crops of) pangram images of the same or different styles of the same family (e.g., Helvetica Regular, Helvetica Light, Helvetica Bold, Helvetica Italic), and negatives are (crops of) pangram images of different families (e.g., Time New Roman). Then, a triplet-based loss function is defined as

$$L(f;\alpha) := \sum_{i}^{N} [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha]_+, \tag{A.1}$$

where $f(x) \in \mathbb{S}^{128}$ is the embedding of image x, $\|\cdot\|_2$ is the Euclidean norm, and α is a margin. We minimize this loss function using stochastic gradient descent (SGD).

We use approximately 20,000 pangrams of fonts to train the neural network. The training is an iterative process of improving the parameters of the network using small batches of images to estimate the gradient and then updating the parameters accordingly. As the gradient is evaluated at more batches, the parameters in the network are adjusted. There are 90,000 parameters. Each batch contains 270 cropped images (i.e., 90 triplets). The training of the network is completed when the loss function reaches below a certain threshold (e.g., 0.7). To ensure fast convergence while avoiding bad local minima, we focus on sampling semi-hard triplets, that is, triplets that violate $\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$ with $\alpha = 0$. The training takes approximately 24 hours with 4 GPUs (Nvidia 1080-TI).

A.2 Details of Lasso Estimation

In Section 3.3, to further interpret the principal component (PCs) we construct, we use Lasso to select tags that explain each PC. We choose the regularization parameter to be 0.002 for regression of PC 1, PC 2, and PC 3, and 0.0005 for that of PC 4, PC 5, and PC 6. To prepare the product tags for lasso regression, we first cleaned the tags by converting them to lowercase, removing non-alphanumeric characters, and eliminating common stopwords such as ‘and’, ‘or’, ‘the’, ‘a’, and ‘an’. We then filtered out tags with no votes. Finally, we merged the cleaned tag data with the principal component data. We use scikit-learn Lasso package for implementation (Pedregosa *et al.*, 2011).

A.3 Diversion to Outside Goods and Embedding Distances

In addition to the analysis of substitution patterns in relation to the embedding distance in Section 6.1, we investigate competition between a given product and outside goods, which includes free fonts—the main competitors of commercial font products.[45] To understand how the substitution to outside goods is affected by the availability of substitutes within the market, we calculate diversion ratios to outside goods for each product and examine their relationship with the (average) embedding distance to close competitors within the market. Figure A.1.(a) displays the binscatter plot of the diversion ratios to outside goods versus the distance to the nearest competitor. The results suggest that the more visually similar a product is within the marketplace, the more likely consumers are to opt for an alternative within the marketplace rather than leaving it entirely. The divergent ratios increase up to around 0.05, and then gradually plateau. Qualitatively similar results are found when examining the binscatter plots of the divergent ratios and the average distances to the 5 and 10 closest competitors. These figures are included in Figures A.1(b) and (c) below.

A.4 Details of Fixed Cost Estimation

A.4.1 Estimating the upper bound on the fixed cost

The entry condition defined in (12) requires taking the expectation with respect to ξ_{kt}. To address this, we randomly sample N_s $(= 30)$ demand shocks from the empirical distribution of entrants' shocks at the time of entry. For each sampled ξ^s_{kt}, we compute the variable profits and approximate the expectation as:

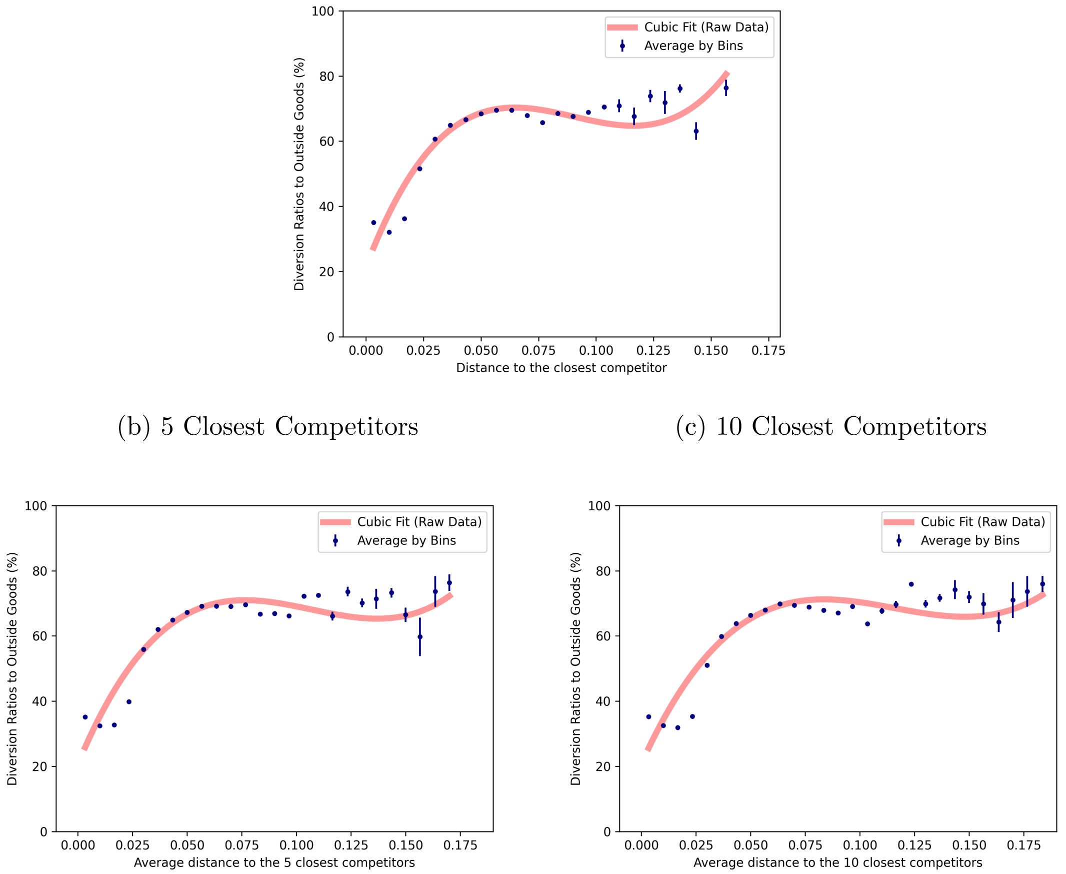
$$\mathbb{E}_{\xi_{kt}}\left[\sum_{j\in J_{ft}\cup\{k\}} \pi_{jt}(E_{ft,k}=1)\right] \approx \frac{1}{N_s}\sum_{s=1}^{N_s}\left[\sum_{j\in J_{ft}\cup\{k\}} \pi_{jt}(\xi^s_k)\right]. \tag{A.2}$$

To compute a variable profit, we simulate prices and shares while fixing all the characteristics, including ξ^s_{kt} and the marginal costs recovered from the pricing model. We use the fixed-point iteration method suggested by Morrow and Skerlos (2011), which is known to have a stable convergence property. In fact, we use the observed price as an initial point for the iteration and attain converged prices for every simulation. PyBLP is used for implementation (Conlon and Gortmaker, 2020).

We additionally compute the no-entry variable profit (i.e. $\sum_{j\in J_{ft}} \pi_{jt}(E_{ft,k}=0)$) by removing the product and following the same simulation method above.

[45] Monotype mentions that the largest competition is with the free font market (Link: Pricing in the type industry today).

Figure A.1: Diversion Ratios to Outside Goods and Distance to the Closest



(a) Closest Competitors

(b) 5 Closest Competitors

(c) 10 Closest Competitors

Notes. This figure presents a binscatter plot of diversion ratios to outside goods as a function of distances to the closest competitors or the average distances to the closest competitors. Each dot represents the bin-by-bin average, accompanied by a 95% confidence interval. The red solid line indicates the third-order polynomial fit of the raw diversion ratios to outside goods data. Panels (a) through (c) display the diversion ratios to outside goods against the distance to the closest competitor, the average distance to the five closest competitors, and the average distance to the ten closest competitors, respectively.

A.4.2 Estimating the slope of the fixed cost function

We use the FOC condition in (11) to estimate the slope of the fixed cost function. This process involves estimating the expected partial derivative of the profit function with respect to each principal component. We accomplish this using numerical differentiation and simulate the average to approximate the expectation. Specifically, we first randomly sample $N_s(=30)$ vectors $\xi^s_{kt} = \{\xi^s_{kct}\}_{c=1}^{12}$, compute $\frac{\partial \pi_{jt}(\xi^s_{kt})}{\partial x^{pca}_{k\ell}}$ for each ℓ by increasing $x^{pca}_{k\ell}$ by a small amount, h, and then differentiate while holding all the other characteristics fixed. Then the expectation is approximated as:

$$\mathbb{E}_{\xi_{kt}}\left[\frac{\partial \pi_{jt}(\xi^s_{kt})}{\partial x^{pca}_{k\ell}}\right] \approx \frac{1}{N_s}\sum_{s=1}^{N_s}\frac{\pi_{jt}(x^{pca}_{k\ell}+h,\xi^s_{kt}) - \pi_{jt}(x^{pca}_{k\ell},\xi^s_{kt})}{h} \text{ for each } \ell \tag{A.3}$$

This numerical differentiation, however, is computationally expensive because it requires simulating pricing responses for all products. Instead, we simplify the process by computing the numerical differentiation of the market share function and verify that this approximation closely matches the results obtained when considering the full pricing responses. That is, we approximate the expected derivative for given ξ^s_{kt} as:

$$\frac{1}{N_s}\sum_{s=1}^{N_s}\sum_{c}(p_{jct} - mc_{jct})\frac{s_{jct}(x^{pca}_{k\ell}+h,\xi^s_{kct}) - s_{jct}(x^{pca}_{k\ell},\xi^s_{kct})}{h}\frac{M_{ct}}{M_t}. \tag{A.4}$$

For a small sample, we confirm that (A.3) and (A.4) lead to very similar results.

A.5 More on Counterfactual Simulations

A.5.1 Details of Relocation Analysis

For the relocation exercise, we first generate potential shapes by locally perturbing the embeddings in the actual data. The main reason for local perturbation is that many points in the characteristic space (i.e., the manifold) do not represent actual font shapes, as they lie outside the support of the distribution of font shapes. Local perturbation addresses this issue by ensuring that potential shapes are closer to actual fonts. In addition, given the large number of products in the marketplace, this approach allows us to generate a diverse range of potential shapes.

Sets of potential locations are created by applying different levels of perturbation, which involve the following steps: First, we randomly sample $\check{N}$ actual font products. Then, different sets of Gaussian noises are generated from 128-dimensional multivariate normal distributions with varying diagonal covariance matrices. To be specific, let $\mathcal{X}$ be the set of actual embeddings. We generate the potential location $\check{x}^{emb} \in \mathbb{R}^{128}$ as:

$$\check{x}^{emb}_{g,level} = x^{emb} + e_{g,level},$$

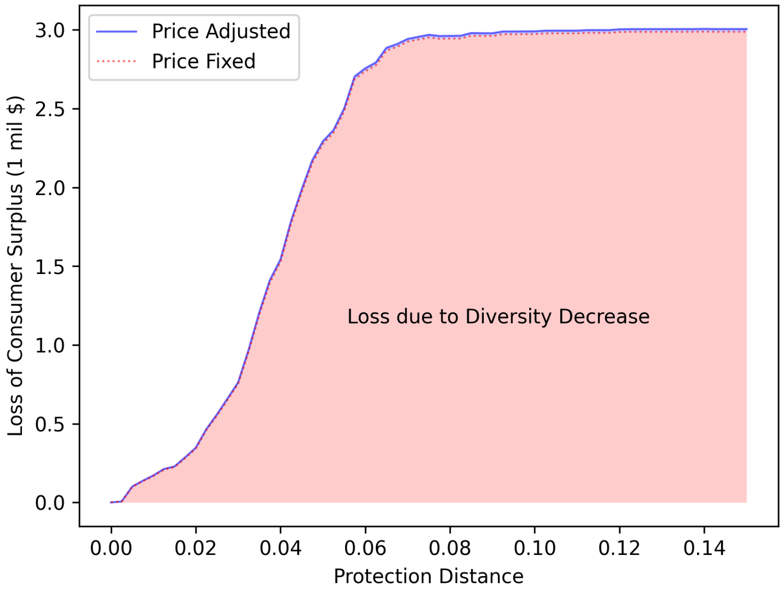
where x^{emb} is randomly sampled from $\mathcal{X}$, $e_g \sim N(0, \hat{\Sigma}^e_{level})$, $\hat{\Sigma}^e_{level} = diag(\hat{\sigma}_1, ..., \hat{\sigma}_{128})/c_{level}$, $\hat{\sigma}_\ell$ is the standard deviation of x^{emb}_ℓ for $\ell = 1, ..., 128$, and c_{level} is the constant governing the degree of perturbation. We set c_{level} to be one of $\{1, 2, 3, 45, 10, 20, 30\}$. As a result, we have in total $8\check{N}$ total potential locations.

A.5.2 Local Monopoly and Consumer Surplus

Following the counterfactual analysis in Section 7.1, to understand the degree of local monopolistic power granted by the copyright protection, we further decompose the decrease in consumer surplus into two channels: (i) price increases due to enhanced monopolistic power, and (ii) diversity loss resulting from eliminating new entrants. To do this, we first compute the consumer surplus in (15) by fixing the price under $\underline{\mathrm{d}} = 0$, while removing entrants according to the protective boundary. Then, we compare this price-fixed consumer surplus with the price-adjusted one shown above. The price-fixed simulation reflects only the loss from the reduced diversity, as this prevents firms from optimizing their prices. Also, the difference between the price-adjusted and fixed consumer surplus for each $\underline{\mathrm{d}}$ captures the loss of consumer surplus due to price increases resulting from the enhanced monopolistic power of incumbent products.

From this decomposition exercise, we find that most of the consumer surplus losses can be attributed to reduced diversity. In Figure A.2, the losses of consumer surplus, i.e. $Loss^{CS}_t(\underline{\mathrm{d}}) = CS_t(0) - CS_t$ for each $\underline{\mathrm{d}}$, are presented. Approximately 99% of the decrease in consumer surplus is due to the reduction in variety. This means that pricing is inactive, which is consistent with empirical findings in Section 4. The diversity here refers not only to the number of products, but also to the desirability of the product attributes, as reflected in the utilities V_{ijt} in (15). We demonstrate that to fully understand welfare changes due to copyright policy, it is crucial to consider the location choices of designers that result from policy changes, as these decisions directly affect the diversity of available products.

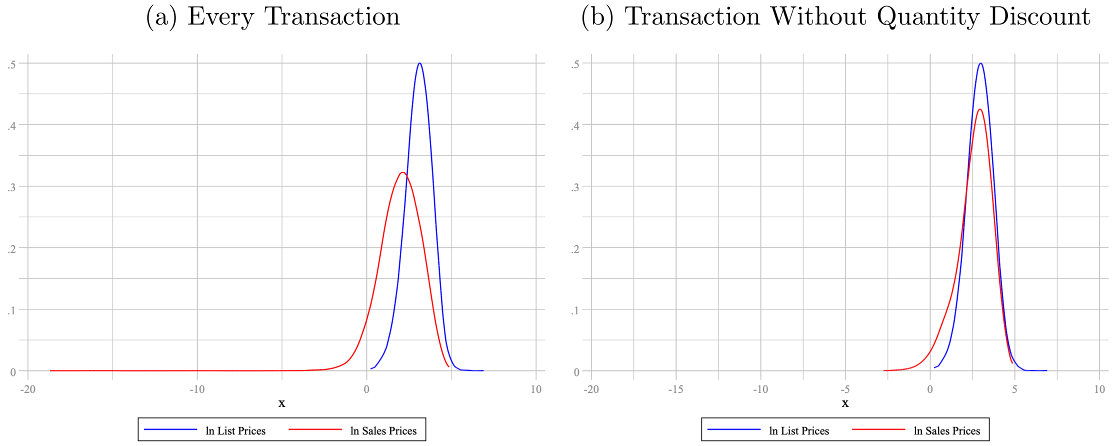
Figure A.2: Decomposition of Consumer Surplus Loss



Notes. This figure shows the analysis of changes in the loss of consumer surplus from naïve simulation.

B Additional Tables and Figures

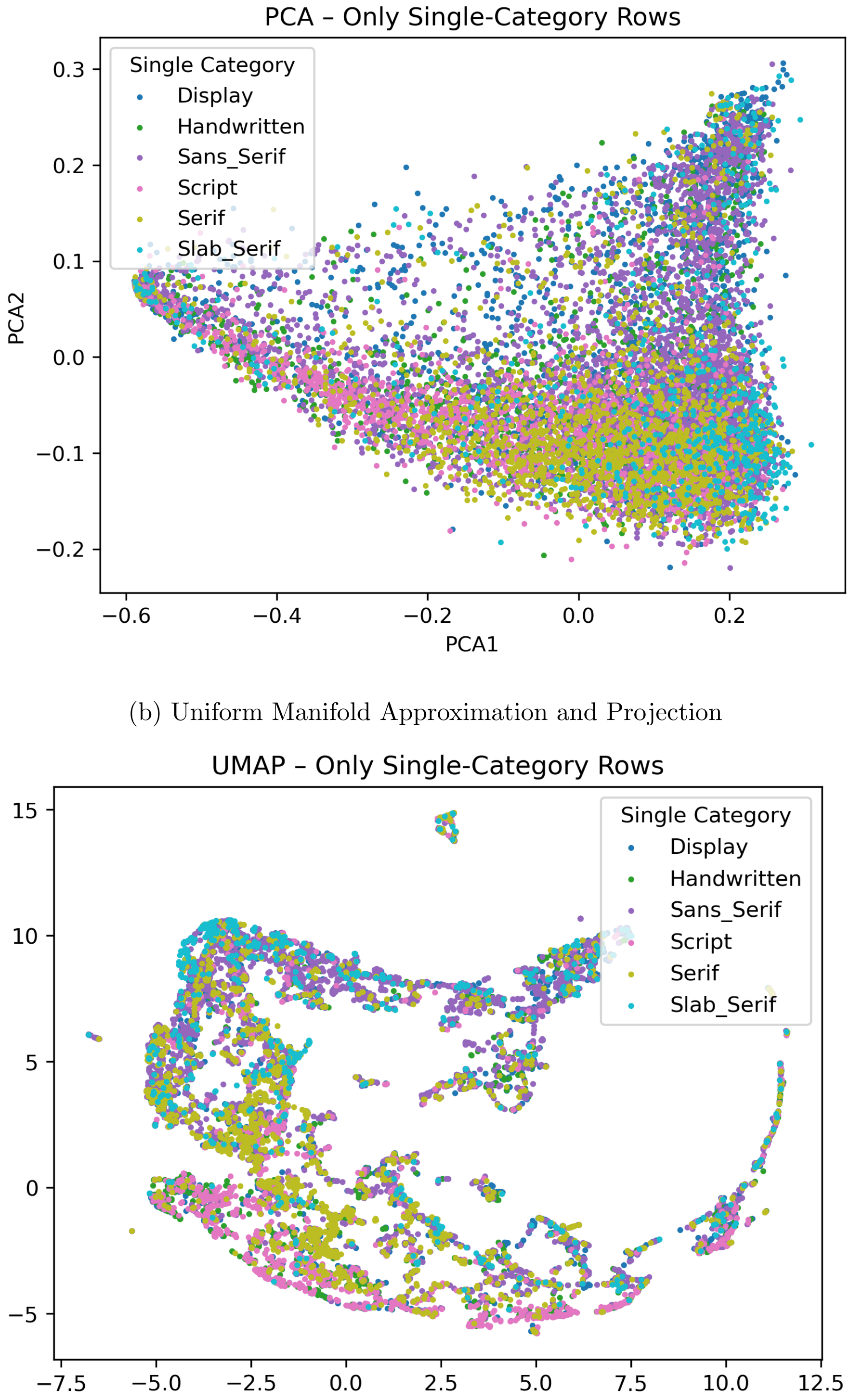
Figure A.3: Distributions of Log List and Sales Prices


(a) Every Transaction

(b) Transaction Without Quantity Discount

Notes. This figure shows the distributions of log list and sales prices. Panels (a) and (b) display the distribution of every transaction and transaction without quantity discount, respectively.

Figure A.4: Scatter Plots using PCA and UMAP



(a) Principal Component Analysis

(b) Uniform Manifold Approximation and Projection

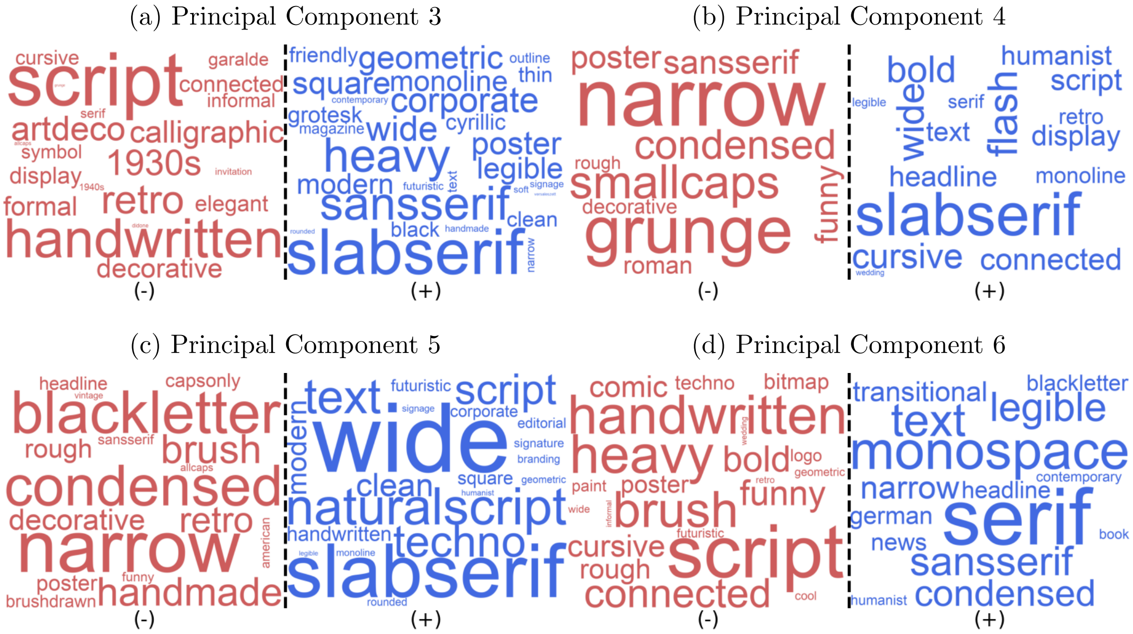
Notes. The color in the dots indicates different industry categories, information inferred from product tags. The categories are well separated by PCA and UMAP, with UMAP doing so more effectively due to its superior ability in clustering and maintaining global structure (McInnes *et al.*, 2018). However, for structural estimation, we utilize PCA for dimension reduction as it outperforms UMAP in terms of interpretability due to linearity.

Table A.1: Descriptive Statistics 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 A.5: Wordclouds of Selected Tags from Lasso Regression



(a) Principal Component 3

(b) Principal Component 4

(c) Principal Component 5

(d) Principal Component 6

Notes. This figure presents word clouds of selected tags from the Lasso regression for each principal component (PC), with word sizes weighted by the coefficient estimates. Panels (a) to (d) correspond to PCs 3 through 6, respectively.

Table A.2: 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		

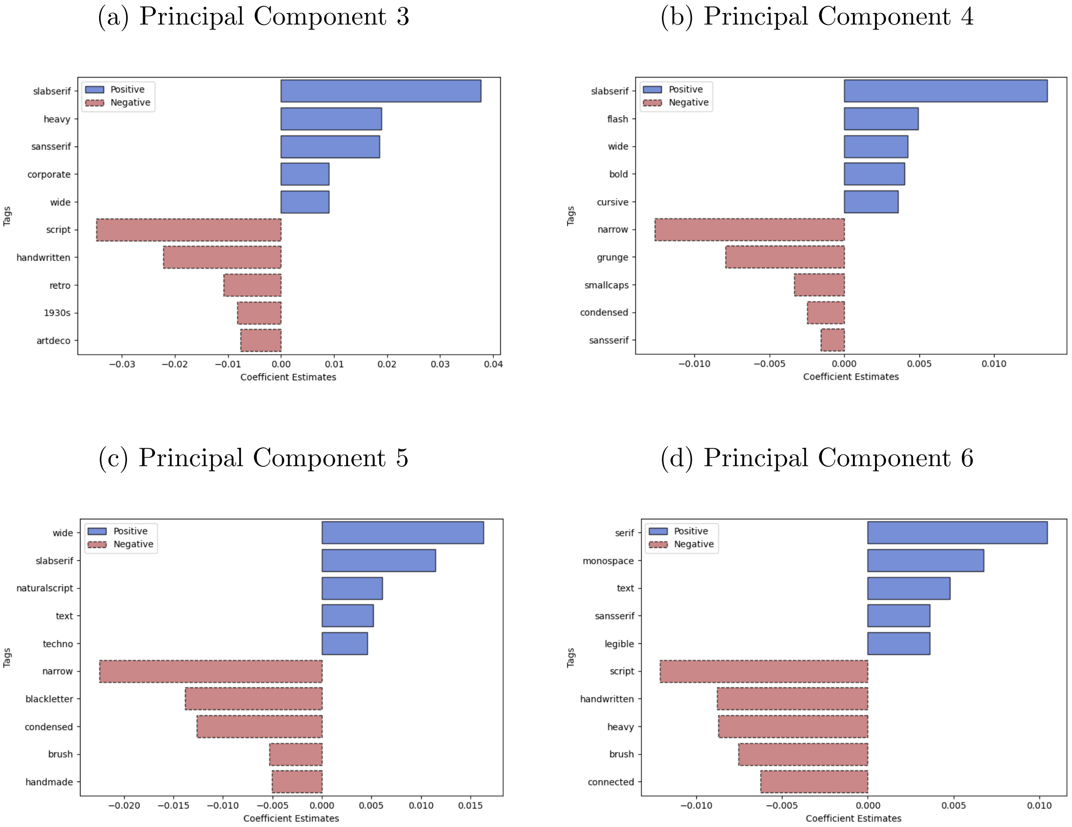
Notes. This table presents one-way ANOVA results of list price, revenue, quantity and sales price variables. SS stands for sum of square and DF stands for the degree of freedom. MS means model sum (=SS/DF). F Stats and P-value columns contain F statistics and its p-values, respectively. Due to computational constraints, we conduct the ANOVA on a randomly sampled 10% of observations.

Table A.3: Descriptive Statistics of Number of Spatial Competitors

Variables	$B_{0.1}$	$R^{0.2}_{0.1}$	$R^{0.3}_{0.2}$	$R^{0.4}_{0.3}$	$R^{0.5}_{0.4}$	$R^{0.6}_{0.5}$	$R^{0.7}_{0.6}$	$R^{0.8}_{0.7}$	$R^{0.9}_{0.8}$
Mean	700.1	5,423.1	7,846.9	6,930.2	3,862.7	2,346.5	1,774.6	1,349.3	47.0
S.D.	545.6	2,678.2	3,459.4	2,671.1	2,254.2	2,031.5	2,178.3	2,323.2	1,169.9
Min	0	0	20	1,176	1,109	87	0	0	0
Max	2,852	13,561	21,187	22,756	20,010	12,534	12,179	12,280	8,759

Notes. This table displays the descriptive statistics for the number of spatial competitors in the visual characteristics space. The number of observations is 4,462,308.

Figure A.6: Top 5 Positive and Negative Estimates from Lasso Regression


(a) Principal Component 3

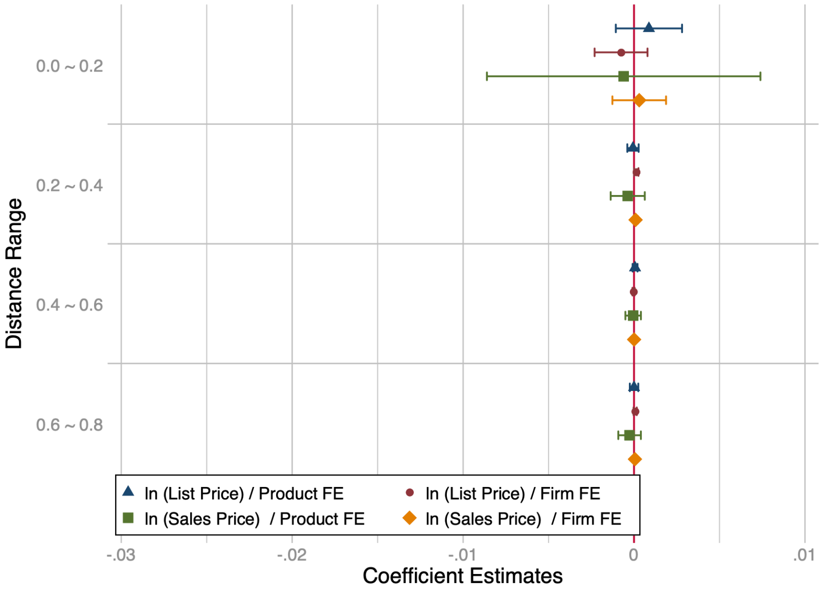
(b) Principal Component 4

(c) Principal Component 5

(d) Principal Component 6

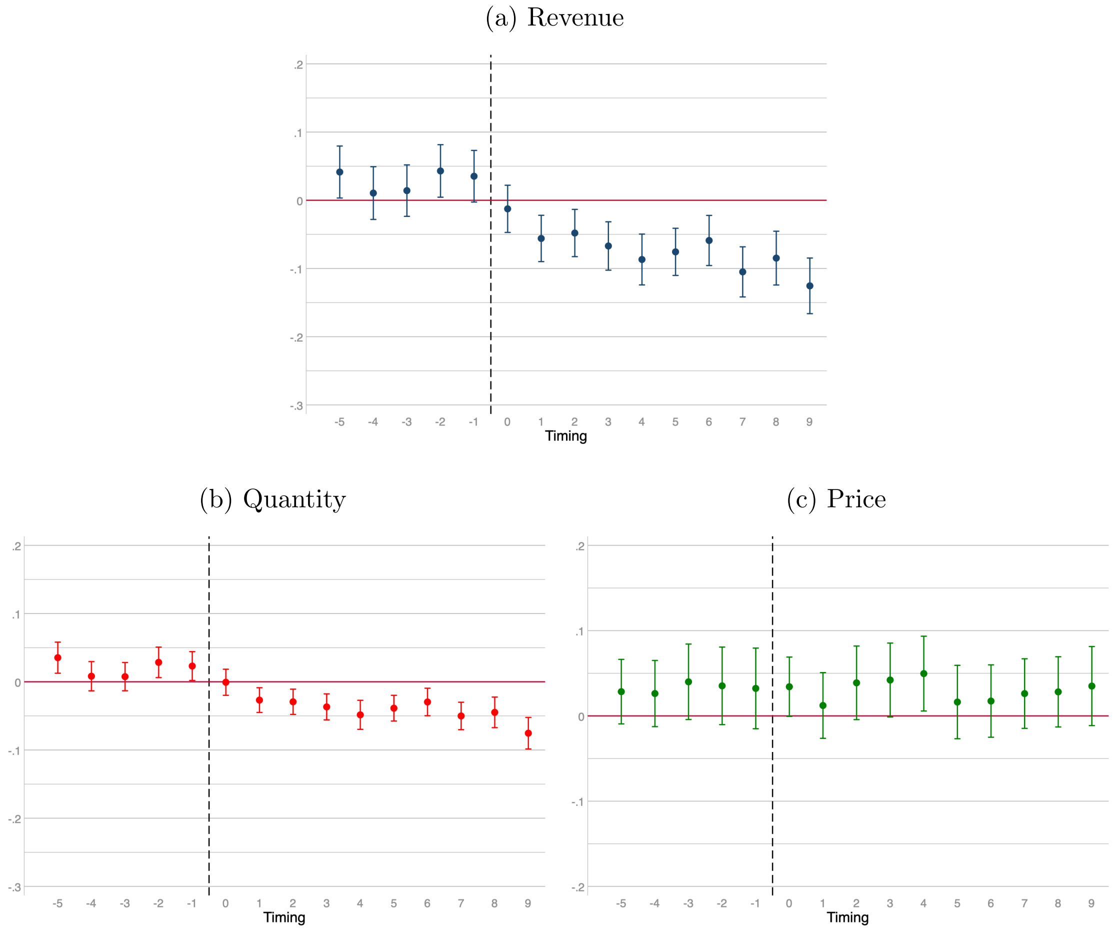
Notes. This figure presents the top 5 positive and negative coefficient estimates of tag dummies from the Lasso regression for each principal component (PC). Panels (a) to (d) correspond to PCs 3 through 6, respectively.

Figure A.7: 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 A.8: Event Study Design (β_s) by Using Borusyak *et al.* (2021)

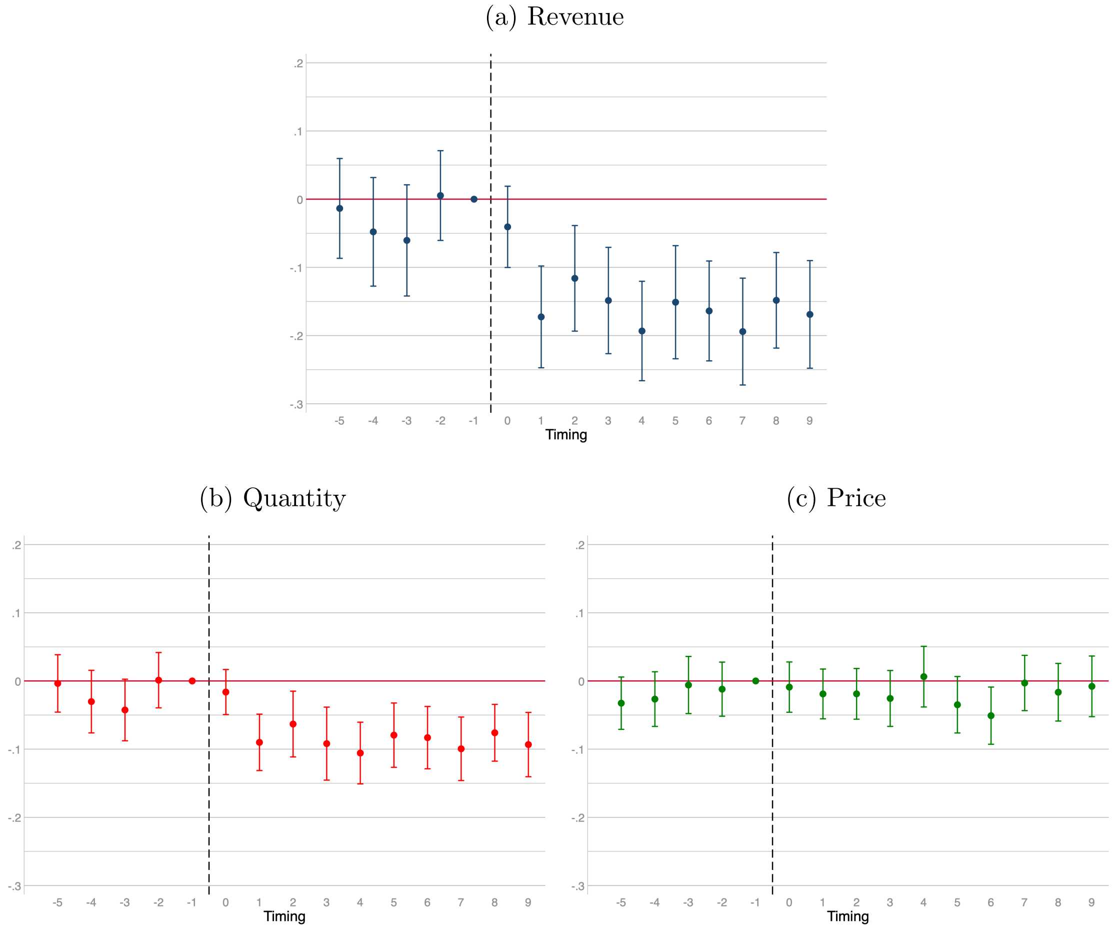

(a) Revenue

(b) Quantity

(c) Price

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

Figure A.9: Event Study Design (β_s): Additional Control Variables

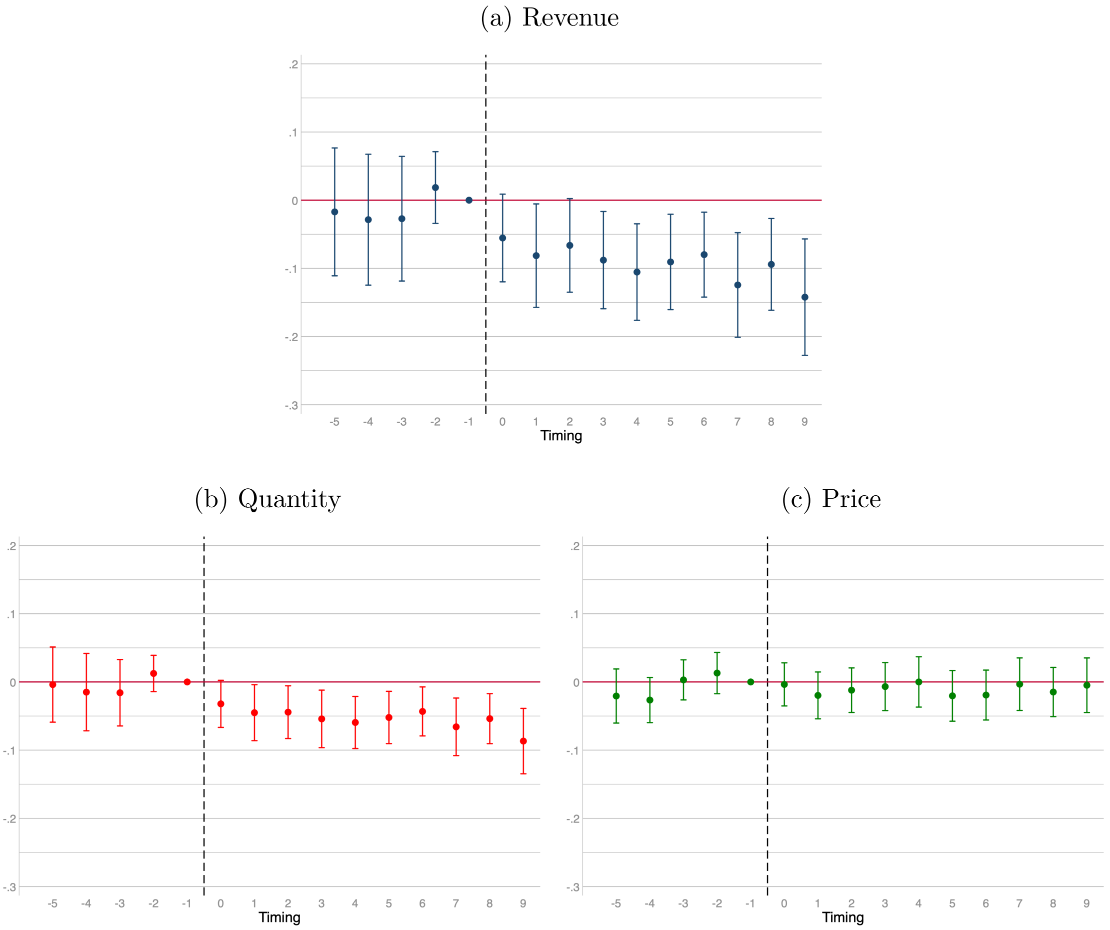

(a) Revenue

(b) Quantity

(c) Price

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. Panels (a) and (b) contain regression results for arsinh transformation of revenue and quantity as a dependent variable, respectively. Panel (c) shows the result for log of list prices as a dependent variable, respectively. Solid lines indicate 95% confidence intervals of estimates. Firm-level clustered standard errors are used.

Figure A.10: Event Study Design (β_s): Alternative Treatment Definition


(a) Revenue

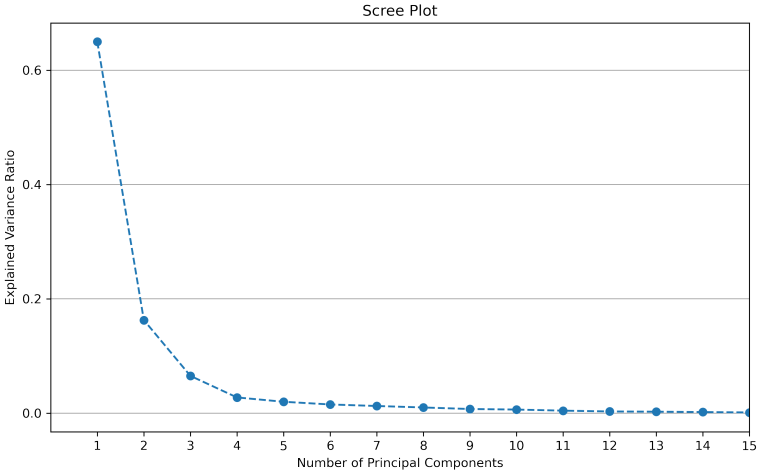
(b) Quantity

(c) Price

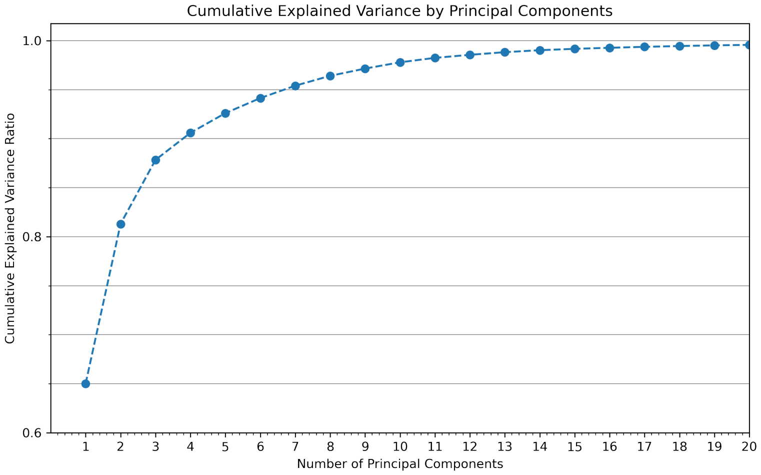
Notes. These figures show results of the event study regression in (3) with alternative treatment definition; in this figure the treatment is defined to be change in one of four closest competitors due to a new entry. Panels (a) and (b) contain regression results for arsinh transformation of revenue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

Figure A.11: Scree Plots

(a) Variance Ratio

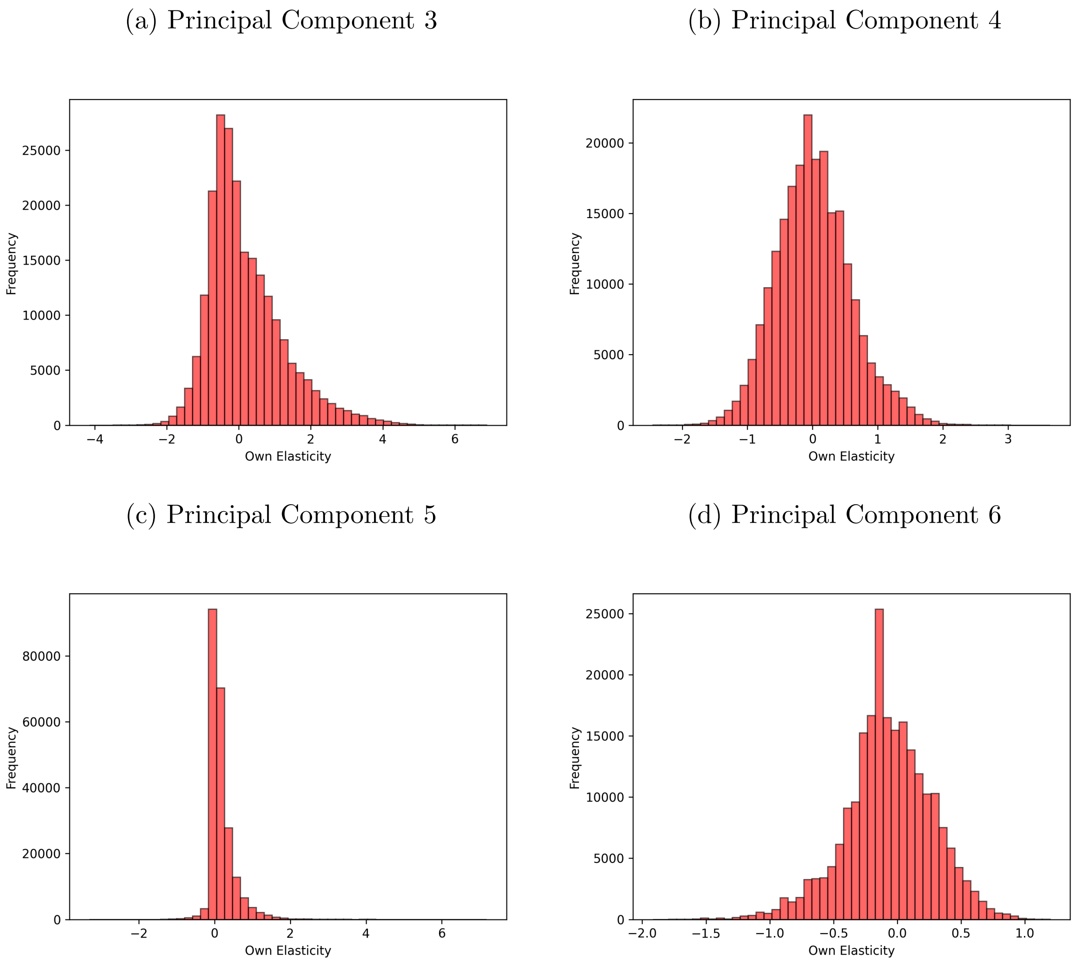


(b) Cumulative Variance Ratio



Notes. These figures show percentage of variance explained by each principal component. Panel (a) shows the explained variance ratio of each principal component. Panel (b) presents the cumulative explained variance ratio along the number of principal components.

Figure A.12: Distribution of Own Shape Elasticities (PCs 3 to 6)


(a) Principal Component 3

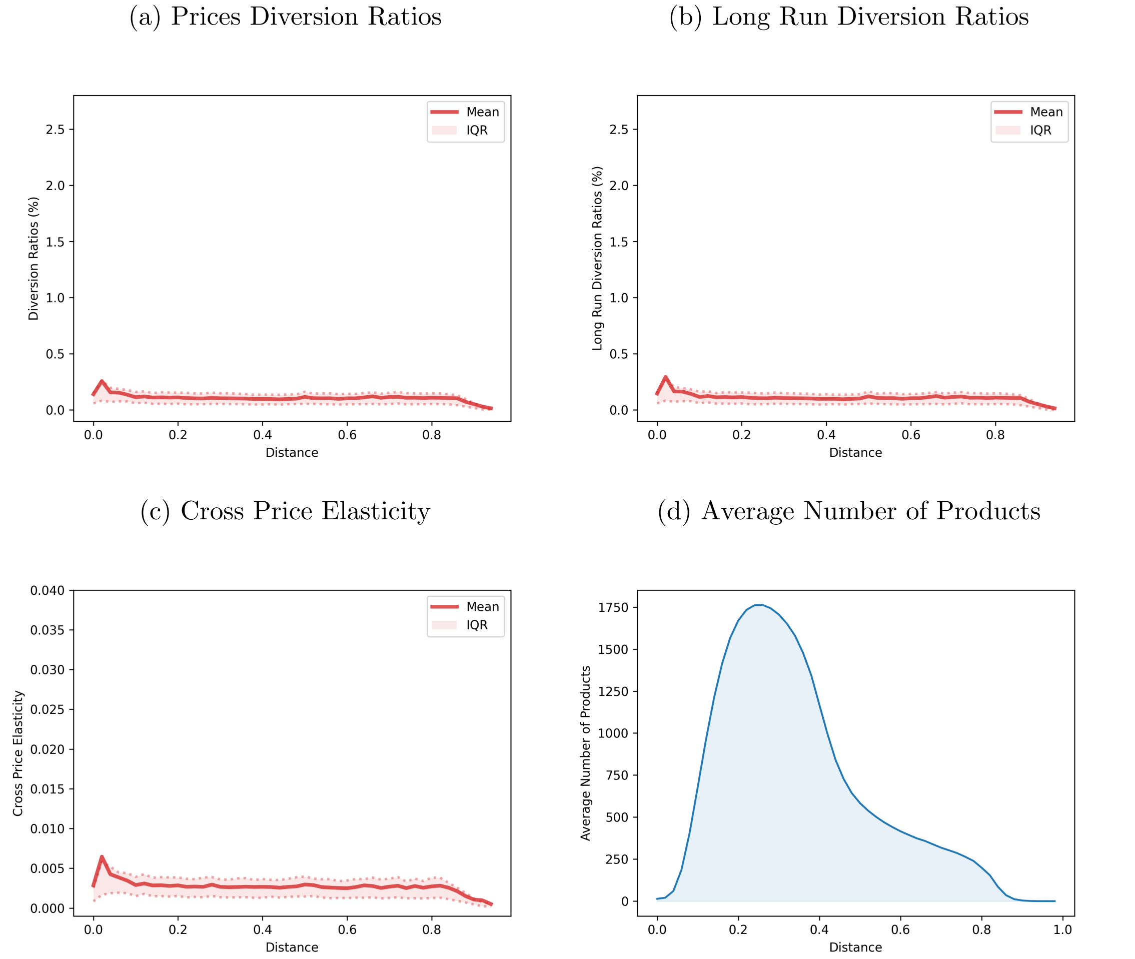
(b) Principal Component 4

(c) Principal Component 5

(d) Principal Component 6

Notes. This figure shows the distributions of own elasticity with respect to PCs. The distribution is plotted across products and markets. Panels (a) and (d) correspond to the distributions with PCs 3 to 6, respectively. The distributions correspond to PCs 1 and 2 are shown in Figure 11.

Figure A.13: Measures of Competition and Embedding Distances (Without Using Visual Characteristics)



(a) Prices Diversion Ratios

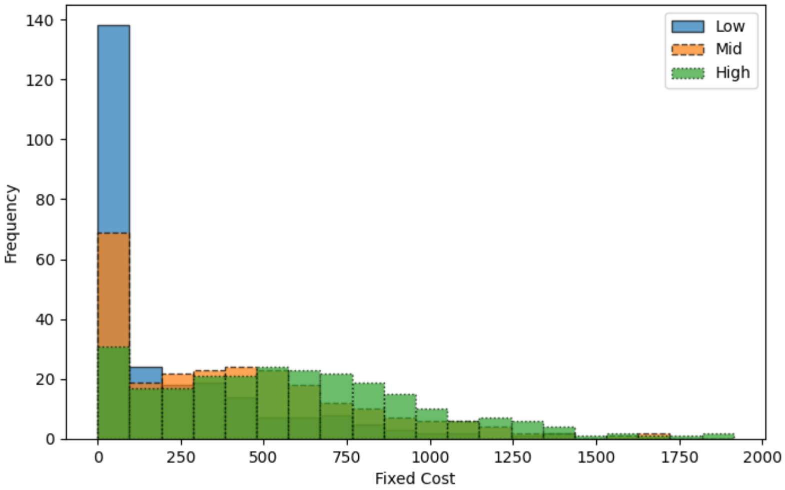
(b) Long Run Diversion Ratios

(c) Cross Price Elasticity

(d) Average Number of Products

Notes. Panel (a) to (c) plot the average price diversion ratios, long-run diversion ratios, and cross-price elasticity along the radial areas A_{jr}, calculated by using demand estimates without random coefficients on PCs, respectively. As a reference, Panel (d) shows the average number of products along radial areas.

Figure A.14: Histogram of Fixed Costs in the Simulation Exercises



Notes. This figure shows histograms of fixed costs (in dollars) by specified fixed cost levels in Section 7.2. The average values of low, medium and high fixed cost distributions are, 284, 529, and 692 dollars, respectively.