Identification and Impact of Online Deceptive Counterfeit Products: Evidence from Amazon

Ziyi Cao, Sanjeev Dewan, Jinan Lin
University of California, Irvine

Abstract: With the proliferation of third-party sellers, counterfeiting has become a serious source of friction in online marketplaces. The impact of counterfeit products in online marketplace remains largely unexplored, in part due to the difficulty of identifying fake items. By leveraging public sales and review data on Amazon, we apply natural language processing techniques to extract the likelihood of encountering a counterfeit product when purchasing from an Amazon ASIN listing. We focus on two product categories, men's fragrance and cell phone wireless chargers, both of which are known to be significantly impacted by online counterfeiting. We adopt a BLP-type choice model with random coefficients to investigate how the probability of a counterfeit product encounter affects user demand, and how the intensity of counterfeiting affects the demand for likely authentic products. Our results validate that consumers can leverage product reviews to avoid knockoffs; a higher propensity of counterfeit encounter significantly reduces consumer mean utility and that they are more cautious when purchasing high-end or popular items. We further find a substitution effect between likely counterfeit and likely authentic products; a 10% increase in one Likely Counterfeit product's price will cause a 0.11% average increase in the market share of each Likely Authentic product. Finally, the detection and disclosure of likely counterfeit products to consumers would protect the sales of likely authentic products, thereby improving consumer utility, but can increase customer churn.

Keywords: Online Product Counterfeiting, Deceptive Counterfeit Products, e-Commerce Platform, Amazon, Natural Language Processing, Random Coefficient Choice Model

1. Introduction

Counterfeit products account for over a half-trillion dollars of trade and are responsible for the loss of 750,000 jobs worldwide (Bressler and Bressler 2018). The problem is particularly acute in online marketplaces due to the inherent information asymmetries between buyers and sellers (Dewan and Hsu 2004, Dimoka et al. 2012, Kennedy 2020). According to a Wall Street Journal investigation (Berzon et al. 2019), thousands of items on Amazon were found to be deceptive or unsafe — even on Amazon Prime. Although Amazon has launched a series of anti-counterfeiting policies to deter knockoffs, the problem is exacerbated by the growing dominance of third-party sellers on Amazon, who account for over 55% of overall sales (Statista 2021). Counterfeiters have blended into the population of third-party sellers, and product recommendation algorithms often present a mix of genuine and counterfeit products to the online customer. Indeed, it is very difficult to distinguish genuine products from fake and unauthorized replicas — *deceptive counterfeit products* — which are practically indistinguishable in terms of price, description, pictures, packaging and delivery terms (Kennedy 2020).[1] This makes deceptive counterfeiting an insidious and costly problem online, but one that has not received much research attention.

Prior work has pointed out that counterfeiting harms a genuine producer's incentive to innovate and hurts its profits by taking away market share (Cho and Ahn 2010, Qian et al. 2015, Wang et al. 2018). On the contrary, some studies have shown that counterfeiting has an promotional effect for the original product and potentially expands market size and profits (Hui and Png 2003, Qian 2014). However, almost all of the prior work has focused on software piracy

[1] This is in contrast to nondeceptive counterfeit products, like a fake luxury handbag, where the seller does not hide the fact that the product is a knockoff, and the consumer is willing to buy a fake presumably because of the lower cost. These are not the focus of this study, which looks at deceptive counterfeit products where the user cannot easily tell whether the product is genuine or fake.

(Cho and Ahn 2010, Hui and Png 2003), or on the impact of non-deceptive counterfeit products where consumers consciously adopt cheaper fake versions (e.g., Qian 2014). In contrast, our focus is on deceptive counterfeit products in online e-Commerce platforms, where the overall impact on the demand for genuine products, on consumer utility, and on platform profits — can go both ways, positive or negative. Starting with the former, some consumers may not care if they have purchased a knockoff, if it functions well and comes at a reasonable price. Further, the spread of knockoffs may have a promotional effect for producers of the genuine products, increasing awareness and market size. The entry of third-party sellers can amplify these effects, driving increased platform transactions and profits. On the negative side, consumers who receive defective knockoffs suffer monetary loss. At the same time, genuine brands see increased price competition, spillover harm to reputation and loss of sales. The decreased confidence of both consumers and brands will impede transactions, negatively impacting platform profits as well. Accordingly, the net impact of deceptive counterfeiting is not at all clear — and one that serves to motivate this study.

In this work, we focus on Amazon.com, which is the most dominant e-Commerce platform in the U.S., and much of the rest of the world. Since deceptive counterfeit products are virtually indistinguishable from genuine ones in terms of description, price, and product characteristics, a key challenge is how to identify them in the first place. In this regard, we exploit the fact that the "proof of the pudding is in the eating"; i.e., consumers have a good idea about whether the product they purchased was fake or not after they actually consume it — many of whom go on to share their positive or negative experiences through the product review system. Still, a significant complication for any counterfeit identification effort is that multiple sellers on Amazon can (and do) sell under the same ASIN; individual reviews do not identify which seller the product came from. Thus, from a user's perspective, the prevalence of reviews indicating fake activity influences

their perception of the likelihood of encountering a fake product when purchasing from a given ASIN. We embrace these identification challenges and develop a probabilistic machine learning methodology to characterize the intensity of counterfeiting at the level of Amazon ASIN listings. The output of this methodology is the probability of encountering a fake product when purchasing from an ASIN, which in turn allows us to classify "likely authentic" and "likely counterfeit" products.

The data set that we use in this study are product and review data publicly available on Amazon, which we sampled through web-scraping. To that end, we crawled all the products listed under the same category, to characterize the entire market for that category on Amazon, keeping track of prices and sales ranks on a daily basis. Historical review data and information on reviewer activity are also captured. We use natural language processing (NLP) techniques to identify counterfeits based on online reviews. Specifically, we apply a semi-supervised hierarchical topic model on the review texts to generate most frequently mentioned topics along with consumer attitudes on product authenticity, quality, shipping and customer service. The percentage of each topic among all historic reviews is used to measure consumers' overall perception. We combine these variables with other numeric features as predictors to train machine learning classification models to identify the probability of a counterfeit encounter when purchasing from an ASIN listing. With the likelihood of counterfeit products probabilistically identified, we adopt a discrete choice model with random coefficients correlated to consumer preferences to estimate the impacts of counterfeits on market shares. We solve the endogeneity issue by constructing instrument variables for the counterfeiting probability using topic variables and multiple sellers' count and prices listed under the same ASIN. We also conduct two counterfactual experiments to explore the economic significance of detecting knockoffs.

Our results suggest that a higher propensity of being counterfeit significantly reduces consumer mean utility gained from purchasing under a listing and therefore its market share and the magnitude of the negative effect of counterfeiting probability is larger when the purchasing target is best sellers. When taking consumer heterogeneity in the sensitivity to knockoffs into account, 64.90% of consumers of men's fragrances category are negatively affected by the counterfeiting probability disclosed in online reviews and this proportion increases to 91.97% when it comes to cell-phone wireless chargers as a category of utilitarian good. Likely counterfeit products exhibit significant substitution effects with likely authentic products and reduce their profits; specifically, a 10% increase in one Likely Counterfeit product's price will cause a 0.11% average increase in the market share of each Likely Authentic product. And the substitution effect is especially strong for expensive brands, which are more inclined to become victims of counterfeiters. Also, our counterfactual experiments show consumers' total utility increases by 3% and market shares of Likely Authentic products increase by 6.5% if consumers gain more certainty on the authenticity of online products. At the meantime, although harms on consumers and original sellers can be mitigated by detection and disclosing of likely counterfeit products, the prevalence of knockoffs can hurt users' trust in the platform and aggravate customer churn.

Our work is among the first to explore the impact of deceptive counterfeits in online retail market. We contribute to tackle the challenge of identifying knockoffs in empirical research by applying natural language processing techniques on user generated reviews; we also develop a choice-modeling framework and solve the endogeneity issue to quantify economic impacts of counterfeiting activities on consumer utility, genuine manufacturers' profits and the platform welfare . We provide important implications for the e-commerce platform by examine the efficacy

and cost of largely conducting detection mechanisms and show the importance to keep a balance between present volume and future sustainability.

2. Literature Review

Prior literature has studied the impact of piracy and counterfeit sales, though much the work is theoretical in nature. In this regard, Sundararajan (2004) models the optimal strategy in pricing and technology protection in response to potential digital piracy. The analysis of Cho and Ahn (2010) suggests that the threat of counterfeiting will reduce firms' incentive to innovate, and to choose a lower quality level for information goods. Similar adverse effects have been investigated in the case of physical goods (Qian et al. 2015) and for the mobile app market (Wang et al. 2018). At the same time, there is contrary evidence of the potential positive impact of counterfeiting, in that illegal copies can increase market size for the original when taking network effects into account (Givon et al. 1995, Hui and Png 2003). In this vein, some analytical studies develop game-theoretic models to show that under strong network effects (Jain 2008), or in a monopoly market (Lahiri and Dey 2013), piracy of intellectual properties will strategically increase product quality and firm profits. Consistent with the theoretical conclusion, Lu et al. (2020) explores the movie industry and finds that piracy increases WOM volume, and post-release piracy is shown to have a positive effect on revenues. In the traditional retail scenario, Qian (2014) conducted a quasi-natural experiment to examine advertising and substitution effects of counterfeits and found the advertising effect dominates for high-end products. Generally, the impact of counterfeiting is complicated and heterogeneous across markets. However, most of the previous research either focuses on digital goods, on luxury goods, or on offline retailing. The counterfeiting issue in the setting of an online marketplace remains largely unexplored, in part due to the difficulty of

identifying knockoffs, and the lack of a ground truth. Given that most consumers intentionally purchase pirated information goods and counterfeit luxury goods, to save on price, the results from those settings do not readily translate to the case of deceptive online counterfeiting, where we expect the effects of counterfeiting on demand and consumer welfare to be quite different.

The detection of fake products in online platforms has been a popular topic outside of the IS literature. Work in computer science, specifically in deep learning, has developed various detection algorithms based on analyses of textual information and/or images posted by the seller. Specific approaches include comparing images posted by others combined with seller information on social network platforms (Cheung et al. 2018), processing microscopic images of physical products (Sharma et al. 2017), or identifying matches for specific products using crawled information (Chaloux et al. 2020). However, these methods are either embedded in devices or utilize photos of physical items as inputs, which do not transfer into consumers' perception of counterfeits in online marketplaces. As we discussed earlier, purveyors of deceptively counterfeit products online tend to thoroughly imitate authentic sellers' behavior by posting realistic pictures and descriptions, making it difficult to distinguish counterfeit from genuine products — short of purchasing and consuming the products.

In this context, we turn to user-generated review information to decipher product quality and authenticity, especially for products where consumers cannot reliably judge the genuineness of products before purchase. Plenty of research documents the significant effect of eWOM on sales in online retail (Chevalier and Mayzlin 2006), hotel booking (Lewis and Zervas 2016) and mediating consumer experience (Bai et al. 2020). In addition, textual information of reviews has also proven to have an economic impact. For example, Archak et al. (2011) applies natural language processing techniques to extract consumer opinions on product attributes from reviews

and explore their impact on sales. Ghose et al. (2011) uses text mining to generate consumer preferences on hotel features and develops a hotel ranking system based on a choice model. Yet none of them have used the topic features extracted from reviews for identification of fake products. We propose an approach which leverages semi-supervised topic modeling to capture consumers' feedback on product authenticity, and further use the outputs from topic modeling as predictors to train machine learning classification models to identify counterfeit products.

3. Data and Model-Free Evidence

3.1 Data and Variables

Our study is built on the most popular e-Commerce platform, Amazon, whose business has expanded from books to cover all categories of products such as electronics, grocery, and luxury goods. Our main analysis is focused on one product category which is known to be impacted by online counterfeiting — men's fragrances (see, e.g., Quora 2019, Steele 2019, Silcox 2021, Pieterse 2021). We also study the cell phone wireless charger category to validate and generalize our findings to a utility product category. We scrape data on all products in these categories and compile detailed product information and review information, both numeric and textual. Our analysis is at the level of unique Amazon Standard Identification Number (ASIN) — i.e., the listing level. In other words, each item in our data sample corresponds to one listing on Amazon.com, identified by a unique ASIN. Note that multiple sellers may be selling the same product under the same ASIN, which is particularly common in the men's fragrances category. It should also be noted that sometimes the same product is sold under different ASINs, in which case we consider the different ASIN listings and associated suppliers to be independent. In a scenario where there are multiple sellers under an ASIN, one of those sellers (primary) is placed in the so-called Buy

Box in the ASIN listing, whereas the other alternative sellers can be accessed from links on the side of the listing (see Figure 1). Which seller wins the Buy Box and becomes the featured seller is determined on a dynamic basis by Amazon, based on prices, shipment and other seller performance metrics such as the feedback rate, customer response time, etc. (Zeibak 2020). While the featured seller changes over time, and consumers can purchase the item from different sellers listed on the ASIN, the historical reviews are all aggregated under the same ASIN listing and there is no way to recover which review corresponded to which seller. Thus, when consumers see reviews indicating past fake product transactions, they can only draw inferences about the overall likelihood of encountering a fake product under an ASIN, and are not able to resolve which of the sellers is likely to be shipping counterfeit products. Accordingly, we use the ASIN-level predicted fake probability to capture the likelihood that a consumer would encounter a fake product when purchasing under a given ASIN.

[Insert Figure 1]

To ensure that our data set covers the entire product market, we collect all the products displayed in the search results, up to the last page. After manually removing a few items mistakenly included for they have similar matching names but do not belong to the focal category, we form the refined sample and scrape their prices, sales ranks and other time-varying features on a daily basis from December 2020 to April 2021. Coupon and discount information are also obtained to adjust the price paid by the consumer. Sales ranks are converted first to sales, and then into market shares — which are the dependent variables in our BLP-type empirical models, a la Berry et al. (1995), as we discuss in detail in Section 5. The men's fragrance product category which is used in our main analysis contains 5661 products in 52 daily (aggregated to 10 weekly) periods. And the category of cell phone wireless chargers contains 1120 products in 120 daily (aggregated to 17

weekly) periods. Considering that the market is rather competitive with long-tailed distribution of listing market shares, to achieve a better fit in the choice models, we use the 1037 highest-ranked products whose market shares in Amazon are greater than 0.01%, which add up to about 80% of the overall market. Summary statistics are shown in Table 1a. Similar processing is conducted on the cell-phone wireless charger category, and that sample contains 565 products in 17 weekly periods.

[Insert Table 1a]

Finally, we collect all historic reviews of each listing in our sample, including: the overall rating; the number of helpful votes that a review received; the number of pictures shared under a review; and the textual comments. We also collect details such as whether the review is linked with a verified purchase, a Vine voice, or posted by a "top reviewer." We applied NLP techniques on review texts to extract topics related to product quality and authenticity, which are key inputs to our machine learning models to estimate counterfeiting probability — see Table 1b for the sample of men's fragrances.

[Insert Table 1b]

3.2 Model-Free Evidence

Before conducting NLP techniques, we first use keyword matching to identify the review texts which that specifically complained about having bought a fake product, in part to warn to other users, so that we can gain an initial understanding of topic distributions and the potential of reviews to predict the probability of counterfeiting. In particular, we label the texts containing keywords like "fake", "counterfeit", "knockoff", "not real", etc. as *disclosing reviews* and examine the number and percentage of such reviews in each listing. The distribution of disclosing review count and ratio are plotted in Figure 2, from which we can tell that most products have less than 30

disclosing reviews, and they usually make up no more than 5% of all historic reviews. Although disclosing information is rather sparse in the review data, they serve as good indicators of suspicious knockoffs since a large proportion of listings never receive any (counterfeit) disclosing reviews at all.

[Insert Figures 2a, 2b]

After extracting topic variables from review texts using NLP techniques and manually labelling the training set, we are left with a correlation matrix and rating distribution as shown in Table 2. The labelled Fake Dummy is highly correlated with the number of disclosing reviews, percentage of one-star and three-star ratings, and topic variables such as positivity in scent, lasting power, prices as well as the fake topic.

[Insert Table 2]

Lastly, we highlight the variations of sales and prices for products that are likely genuine or likely counterfeit (based on a 50% predicted probability cutoff). As shown in Figure 3a, we find that likely counterfeits are more likely to have a higher price, as compared to likely genuine products. The pattern implies that counterfeiters target products with higher prices, i.e., counterfeiters tend to enter markets with medium or high prices. In terms of sales, we find that for products with lower sales, the composition of likely authentic products is higher relative to likely counterfeits. Both likely counterfeit and likely authentic products follow long-tailed distributions of sales (Figure 3b).

[Insert Figures 3a, 3b]

4. Counterfeit Identification

As discussed earlier, our focus is on deceptive counterfeit products, which are fake and unauthorized replicas of real goods, but are merchandised in a manner which makes them virtually indistinguishable from genuine products. Yet, the functionality of the fake products is generally inferior to that of the genuine ones. For example, in the category of fragrances, a listing of a fake product would look indistinguishable from one of a genuine product, with essentially identical description of important features such as scents, weights, ingredients, country of origin, pictures and videos. However, consumers of fake products would notice that the perfume seems diluted, has a strange and unexpected smell, or fades very fast. That is, we focus on counterfeit products which are deceptively sold to consumers online. Consumers are assumed to have a vertical preference for the authenticity of products; i.e., every consumer prefers a genuine product to a counterfeit one. The genuineness of the products can generally be discerned post-purchase, and this fact is likely to be revealed in user reviews — which is the key to our identification of counterfeits. And due to the existence of multiple sellers per listing, we focus on the probability of encountering a fake product when purchasing from a given ASIN listing, regardless of which listed seller they use.

The identification of counterfeit products in online markets has always been a challenge in empirical research due to the lack of ground truth. The most relevant and efficient method is contributed by Wang et al. (2018), who utilized text and image matching to cluster similar apps and differentiate copycats from the original mobile apps within each cluster by their launch dates. However, similar methods cannot be applied to the e-Commerce scenario. First, on giant e-commerce platforms like Amazon, the market composition is much more complicated. Within each defined category there may exist thousands of related products provided by different merchants and matching relationships between real and fake items are more likely to be many-to-

many than one-to-many. Second, based on our definition of online counterfeit products, sellers will pretend and imitate the behavior and signals of authentic sellers in terms of pricing and posting product information; therefore, it is difficult to differentiate genuine products from knockoffs within each matched cluster simply by leveraging one feature such as the price or launch date. Finally, a single ASIN might cover multiple sellers, and individual reviews cannot be linked to specific sellers; yet counterfeiting products can be sold only by one or some of all the participating sellers, which makes a binary identification at the listing level inaccurate and inefficient.

E-WOM (electronic word of mouth) has proven to be an important source of product quality, and a significant driver of online sales. Prior literature has extensively studied the impact of user-generated reviews on online sales, both in terms of numeric ratings (Chevalier and Mayzlin 2006) and textual reviews (Archak et al. 2011). However, to the best of our knowledge, reviews haven't been previously utilized to detect fake products. We apply NLP techniques to extract frequently mentioned topics including product authenticity and other aspects of quality and use these generated indicators to train machine-learning classification models for prediction, as we describe next.

4.1. Topic Extraction

Our original sample of men's fragrance contains 5661 products, for which 352,933 reviews are collected up to April 2021. We pre-process the raw text by case normalization, tokenization (including removing punctuation), POS (part of speech) tagging and lemmatization according to POS tags. Besides, we apply language detection to raw texts and keep only those written in English, which constitute over 85% of all reviews. Next, text vectorization is conducted to convert each pre-processed review text into a numeric vector. We choose the TF algorithm and keep the top

10,000 frequent tokens (words) in the vector, which is appropriate because review documents are relatively short and straightforward and contain no complicated semantic issues.

With textual data embedded into a numeric matrix, each row of which represents a piece of review, we train an anchored topic model on it to generate targeted topics and determine whether the documents contain corresponding information in each of the topics. Anchored CorEx (Correlation Explanation) is a semi-supervised hierarchical topic model which allows users to guide the topic generating direction by assigning anchored words for topics of interest (Gallagher et al. 2017). To decide which topics regarding product quality are most frequently mentioned by consumers either positively or negatively, we fit an LDA model to obtain some independent product features with relevant keywords. For example, in the men's fragrance category, the most relevant topics generated by the unsupervised models are consumer sentiments, attitudes on scents (with keywords nice, pleasant, classy and cheap, overpower, weird, etc.), attitudes on lasting power (with keywords long, throughout, all day or lost, flourish, etc.). Next, we randomly read a group of reviews to refine the construction of topics and anchor words, so that the topics of interest are better captured. We define 13 topics for men's fragrances related to counterfeit identification, as follows: explicit counterfeit (fake), overall sentiments, attitude towards price, scent, longevity, package and shipping (positive or negative). After fitting the model, we remove topics which are not accurately anchored or under which most of the pre-defined keywords were rarely captured; then we modify the model to include eleven topics in order to increase total correlation. We fit the refined model to extract indicators of texts under each topic and aggregate the review-level topic dummies into product-level percentage scores. For example, if one review states "This product is a cheap knock-off of the actual cologne. The box comes with white stickers on it to cover the actual serial info and the scent is off and does not last nearly as long", it will be tagged 1 under the topic

of “fake” and “last_negative”. The higher the product-level percentage scores under the topic of fake or other topics with negative sentiments towards a particular feature, the more likely the product will be deemed to be counterfeit.

For the purposes of further model improvement, we also calculate the topic dummies’ average values weighted by the log of helpful votes each review receives (denoted as *fake_w* for example), as well as the average of a subset including only reviews with at least one helpful vote (denoted as *fake_h*). User-specified anchored words are consistent with model-determined keywords, as listed in Table 3. The review-level topic indicators are aggregated to product level as predictors in the classification of fake products.

[Insert Table 3]

4.2. Classification Model for Likely Counterfeit Encounter

As we have discussed, we rely on Amazon review data to predict the likelihood of encountering a fake product when purchasing from a particular ASIN. Individual reviews cannot be linked to specific sellers, so consumers have to draw imprecise inferences at the ASIN level, about the likelihood of experiencing a fake product. We build a supervised classification model for predicting the probability of encountering a fake product. We manually constructed a training data set, wherein multiple human coders (graduate students in our case) browsed homepage information, rating distribution and textual reviews for all the reviews under a given ASIN, and generated a label for that ASIN, as either “Likely Authentic” or “Likely Counterfeit.” The interpretation of a Likely Counterfeit (Likely Authentic, respectively) is that there is a better than even perceived likelihood that a product purchased from this ASIN is going to be counterfeit (authentic, respectively). The labels were coded in the data set provided two coders independently agreed on

the label for each ASIN. ASINs for which there was no consensus among the coders were left out of the training data set.

Our approach to constructing the training data set is a departure from the traditional notion of a machine learning data set, where each instance has an objectively certain outcome, whereas in our case, the labels are the result of a subjective assessment. A number of factors contribute to an assessment that an ASIN should be coded "Likely Counterfeit." For one thing, an unusually high number of reviews that explicitly call the product fake — using some combination of keywords like "fake", "counterfeit", "knockoff" — were a sign of counterfeiting activity in the ASIN. Also, the coders looked at the distribution of ratings, and specifically, the proportion of 1-star ratings. Cases where there were more 1-star than 2-star ratings were examined more closely for counterfeiting activity. At the current stage, our training set for each category (men's fragrances and cell-phone chargers) contains 200 listings.

To reduce possible bias caused by topic modeling, we also leverage the metrics of "disclosing" reviews as supports, if it explicitly complains about having bought a fake product and contains one of more of the predefined keywords such as "fake", "counterfeit", "knockoff", etc. Correlation matrix shows consistency between the disclosing indicator and the fake indicator generated by topic modeling. A Likely Fake prediction can be discerned from both the absolute count and percentage of disclosing reviews. To this end, we define a dummy variable *counterfeit_10_01*, which is set to 1 if a product has at least 10 disclosing reviews that account for more than 1% of the total number of reviews; it is set to 0 otherwise.

We build six of the most commonly used machine learning classifiers (such as naïve Bayes and random forest) with multiple groups of predictors according to their correlations to the labeled counterfeit dummy. Among all the review-generated topics, overall positive and negative

sentiments (*like*, *dislike*), negative opinions on product authenticity (*fake*), negative opinions on the fragrance's scent and durability (*scent_negative*, *last_negative*) are most informative in prediction. Features of rating score distribution (i.e., *1-star ratings percentage, 3-star ratings percentage and 5-star ratings percentage*) are also highly correlated with the propensity of an ASIN being labeled Likely Counterfeit. We optimize the classifier by adjusting variables and the average accuracy is raised up to 83% in the second model displayed in Table 4. We select the random forest classifier with the highest accuracy to generate the variable of interest, which is counterfeit likelihood, for all the products in our sample.

[Insert Table 4]

5. Econometric Analysis

5.1. Discrete Choice Model Setup

We adopt a structural model of discrete choice with random coefficients, following Berry et al. (1995), to estimate the effect of perceived product authenticity on consumer choice and study what impact likely counterfeit products have on likely authentic ASIN sales, consumer utility and platform welfare. The BLP (Berry et al. 1995) model is a logit model estimating demand in differentiated product markets using aggregate data and allows for random coefficients of product characteristics and endogenous prices. We assume there is heterogeneity in both consumer preferences on prices and their capabilities of detecting a fake product therefore specify the coefficients of prices and counterfeiting probability as random. We also expand the model by allowing counterfeiting probability to be another endogenous variable besides the price.

Specifically, we define the utility of consumer i buying a product in ASIN j in market t as follows (we will use ASIN and product interchangeably):

$$u_{ijt} = \alpha_i P_{jt} + \gamma_i C_j + X_{jt}^{v}\beta^{v} + X_{j}^{inv}\beta^{inv} + \xi_{jt} + \varepsilon_{ijt} \quad (1)$$

where i represents the consumer, j indexes the product, and t represents an Amazon fragrance market t (week t in our setting). P_{jt} is the weekly-average price (adjusted by discounts) of product j in market t, and C_j is the probability of j being a Likely Counterfeit product, generated from the machine learning classification model.[2] It also represents consumer's perceived skepticism of the product's authenticity (buying from one of the sellers listed under the ASIN) after reading historic reviews online. X^v refers to time-varying product features such as rating valence and volume, numeric metrics extracted from the reviews such as average numbers of helpful votes and images. X^{inv} represents time-invariant product characteristics, such as the size and parfum concentration level in the case of fragrances. ξ_{jt} is the market-specific unobserved product attribute and ε_{ijt} has i.i.d. type I extreme value distribution. We aggregate our daily-level sales data into weekly averages to construct a 10-week panel, which corresponds to 10 markets.

A consumer will search the market, review product characteristics to find the one that best matches her preferences, read historical reviews including rating distribution and texts to determine if a product is authentic, and finally choose the product which maximizes her utility. We allow consumers to have heterogeneous tastes for prices and sensitivity to a counterfeit purchase. In other words, we include random effects of the price and the counterfeit probability variable. Note that although consumers are assumed to have vertical preferences on product authenticity and try to avoid knockoffs, they are not equally engaged in and familiar with the review community, which creates heterogeneity in their sensitivity and capability to identify the

[2] Yang et al. (2022) discusses potential bias resulting from the correlation in measurement error of the predicted covariate (C_j) and the regression error. However, the remedy suggested by them is not applicable here, as we do not have access to the ground truth, as they do.

probability of a product being counterfeit based on the user-generated reviews. Therefore, we model the consumer distribution as follows:

$$\begin{pmatrix}\alpha_i \\ \gamma_i\end{pmatrix} = \begin{pmatrix}\bar{\alpha} \\ \bar{\gamma}\end{pmatrix} + \Sigma v_i\,, \qquad v_i \sim N(0, I) \tag{2}$$

where v_i is consumers' unobserved preference for price and counterfeiting probability; in particular:

$$\alpha_i = \bar{\alpha} + \alpha_v v_i, \gamma_i = \bar{\gamma} + \gamma_v v_i \tag{3}$$

Accordingly, consumers select the product which generates the highest utility and market share is the sum of probabilities of being chosen by each consumer, outside option utility normalized as 0. Containing two components of consumer heterogeneity, i.e., ε_{ijt} and v_i, the market share function can be obtained in two stages. First, integrating out over the ε_{ijt} conditional on v_i gives us the logit model as Equation (4), following McFadden (1973). Second, integrate out over v_i to obtain the market share only conditioning on product attributes. For the second integration doesn't have a closed form, Monte Carlo simulation agent data is used as a substitute in the estimation process (Berry et al. 1995).

$$s_{ijt} = Pr(y_{it} = j) = \frac{\exp\left(X_{jt}^{v}\beta^{v} + \alpha_i P_{jt} + \gamma_i C_j + X_j^{inv}\beta^{inv} + \xi_{jt}\right)}{1 + \sum_{l=1}^{J}\exp\left(X_{lt}^{v}\beta^{v} + \alpha_i P_{lt} + \gamma_i\, C_l + X_l^{inv}\beta^{inv} + \xi_{lt}\right)} \tag{4}$$

5.2 Market Share and Instrument Variables

To obtain sales data and further calculate market shares, we follow the approach widely used in previous research (Chevalier and Goolsbee 2003, Ghose and Sundararajan 2006) to convert sales rank into a proxy of sales. Product sales rank is assumed to follow a Pareto distribution, where the probability that observation *s* exceeds a specific level *S* is $Pr\,(s > S)\; = (k/S)^{\theta}$. For a particular product, the probability of randomly drawing a more popular competing item is taken to be equal to the number of items that are ranked ahead of the given product, which can be modeled as

$(Rank-1)/(Total\ number\ of\ items) = (k/S)^{\theta}$. Taking logs of the two sides transforms the equation into $ln\ (Rank-1)\ = c - \theta * ln\ (Sales)$. Therefore, to convert rank data into sales rank, we only need to estimate the above log-linear regression model and obtain the linear coefficients c and θ. We conducted a simple experiment. First, two products are selected whose initial sales are ranked low enough to be approximated as 0. Then we purchased a few copies of these products and observed how sales rank changed. We repeated the purchase and tracked the updated sales ranks several times within two days, collected data point pairs and fit the aforementioned log-linear model. The estimated coefficient of men's fragrances product category is about 1.25, which perfectly falls between the suggested range of 0.9 to 1.3 in prior literature. In this way, we are able to estimate product weekly average market shares by feeding sales rank data into the model.

The traditional BLP model allows for endogenous prices and uses sums over product characteristics within or across brands as instrument variables. Here we select two characteristics of the men's fragrances category, size in ounces and volume (as a common measure of the concentration of alcohol and parfum in fragrances) and construct sums over characteristics of both non-rival goods (other goods under the same brand) and rival goods (goods of other brands) as instrumental variables for prices. In particular, we encode the volume from 0-4 according to the percentage of alcohol and parfum (i.e., after shave as 0, cologne as 1, eau de toilette as 2, eau de parfum as 3 and parfum as 4).

It should be noted that the distribution of fake products across ASINs is not random. Counterfeiters tend to target ASINs which are more expensive or popular, not only because they can make higher profits for higher cost products, but also because they have a greater chance to attract more consumers via a slight price reduction relative to authentic products. The figures we

show in Section 3.2 also support the above intuition. This selection issue will bias the estimation of counterfeiting impacts. Therefore, we extend the traditional BLP model by allowing the extra endogeneity (beyond price) in counterfeiting probability variable, and use a group of instrument variables to separate the effect of unobserved sellers' behavior.

We collect the number of multiple sellers (or buying options) and calculate the variance of their landing prices as the first group of instruments for the counterfeiting probability variable. We hypothesize that the more alternative sellers there are in one ASIN, the more likely is the entry of fake sellers for that ASIN. Another factor is that due to Amazon's anti-counterfeiting policy, hiding in a popular listing with many seller options reduces the risk of being reported by consumers, and consequently removed from the platform. Another group of instruments for counterfeiting probability is derived from topic-modeling variables such as the positive or negative attitude on fragrances' scent and lasting power. These topics by design are extracted from user-generated reviews and aggregate at a listing level to help predict the probability of being counterfeit, and at the same time they are not correlated with selection issues since they are independently generated from reviews, which makes them good instruments. Table 5 shows the correlations between instruments and endogenous variables.

[Insert Table 5]

5.3 Main Results

We first estimate a reduced-form 2SLS fixed-effects model to explore the effects of prices and counterfeiting on market shares and the efficacy of instrument variables. Next, following the implementation of Vincent (2015), we include consumer heterogeneity and estimate the BLP random coefficient logit model (Equations 1-4s) built on simulated data generated from Monte Carlo analysis. Last, we run a sub-sample analysis on a data set containing only top ranked

products whose market shares are greater than 0.1% to examine if there is a different pattern associated with the most popular products.

Table 6 shows results of the 2SLS models; Columns 1-3 report the estimation on the top 1096 products whose market shares are no less than 0.01% with 9520 listing-week observations with brand category fixed effects, week fixed effects. Columns 4-6 report estimations of the same models on top 148 products whose market shares are no less than 0.1% with 807 listing-week observations. The estimation of primary variables across different models is consistent. Demand in the online marketplace is negatively correlated with price. The coefficient of *Counterfeit* probability estimated on the larger data set is negative and significant, indicating that the perceived possibility of an item being counterfeit, based on past reviews, will negatively affect consumers' willingness to purchase and therefore the market share. We also find a significant positive coefficient on the log of number of ratings, suggesting a positive effect of popularity on consumers' purchasing decisions. The coefficients of *Image_count* and *Helpful_votes* are also significantly positive, indicating that the average quality of historical reviews can improve consumers' utility and product sales.

[Insert Table 6]

To better quantify counterfeiting effects, we go beyond the reduced form model above to estimate a random choice model. Specifically, we estimate a random coefficient (BLP) logit model (Equations 1-4), which generates results consistent with the reduced-form analysis. As shown in Table 7a with simulated agent data (Columns 1-2), a higher propensity of being counterfeit significantly reduces consumer mean utility and therefore the market share. The effect of price on consumer mean utility is also negative. On the other hand, the coefficients of rating and log of rating counts are both significantly positive, suggesting a better reputation and sales history have

positive impacts on consumer utility. Columns 3-4 show the estimation on the Top 20 percentile products ranked by average sales, whose market shares are no less than 0.1%. Comparing these results with what was found from the larger data set, it's clear that the effects of prices and some review metrics on mean utility are not significant anymore, while the magnitude of counterfeiting effect is greater, indicating consumers who seek to consume from best sellers attach relatively more importance to authenticity and historical sales record when they make purchasing decisions.

[Insert Table 7a]

Table 7b shows the estimated standard deviations of random coefficients based on consumers' heterogeneous preferences, with the four columns corresponding to the models presented in Table 6a, respectively. Looking at Column 2 (the complete model), we see that the mean effect of counterfeiting probability is -1.0073 and standard deviation is 2.6599, indicating about 64.90% of consumers are negatively affected by the counterfeiting probability disclosed in online reviews. However, Column 4, which captures the impact of counterfeiting on purchasers of best seller products, the proportion of consumers negatively affected by counterfeiting rises to 89.04%.

[Insert Table 7b]

5.4 Economic Significance of the Results

We now study economic impacts of counterfeiting on merchant profits and platform welfare, by generating elasticities and conducting a counterfactual experiment, based on the BLP model results.

5.4.1 Price Elasticity

First, to further explore the average effects of likely counterfeit products on likely authentic ones, we calculate the price elasticity between categories of products based on the random coefficient logit model. Products in our sample are labeled as Likely Counterfeit or Likely Authentic based

on the probability generated by the classification model using a 50% cutoff. Table 8 shows the own price-elasticity and cross price-elasticity between Likely Counterfeit and Likely Authentic items in ten markets. On average, the cross price-elasticity between Likely Counterfeit and Likely Authentic products is 0.011%, suggesting a 10 percent increase in one Likely Counterfeit product's price will cause a 0.11% average increase in the market share of each authentic product. Although the magnitude is small, considering this is a very competitive market which contains hundreds of knockoffs and authentic products, a substantial amount of market share will be taken away in total if only fake sellers lower their prices slightly.

[Insert Table 8]

The positive price-elasticity between Likely Counterfeit products and Likely Authentic products implies a significant substitution effect, which appears to dominate any potential promotional effect for genuine original merchants in the online marketplace. The finding is different from that documented by Qian et al. (2014) in the traditional retailing industry, in which advertising effects have been shown to dominate substitution effects for high-end products. The different outcomes have to do with the differences between non-deceptive versus deceptive counterfeit products. First, unlike the offline sales of luxury counterfeits which are targeted at a separate segment of consumers inclined to knowingly purchase cheap knockoffs, online deceptive knockoffs pretend to be authentic and are targeting a larger group of consumers. Second, unlike piracy of information goods or luxury copycats, prices of online knockoffs are not necessarily lower than that of authentic ones. Third, online consumers' perception of a counterfeit purchase happens only after the purchase is finished. Therefore, online counterfeiters take away larger market shares and profits from authentic manufacturers, and substitution effects dominate potential positive effects.

We also examine the heterogeneous substitution effects across different types of fragrance brands. In particular, we define seven fragrance categories horizontally according to their brand type, namely: designer fragrances (e.g., Paco Rabanne); luxury brand (e.g., Chanel); beauty brand (e.g., Lancome); price-friendly brand (e.g., Bod Man); high-end fashion brand (e.g., Davidoff); low-end fashion brand (e.g., Abercrombie & Fitch); and auto brand (e.g., Mustang). Among the seven categories, high-end fashion (2), designer's (3) and luxury (4) brands sell perfumes at a relatively higher price. Table 9 and Figure 4 show the average price elasticities across ten markets (weeks). The percentage of Likely Counterfeit products and substitution effect on the demand of authentic products are both high in the high-end fashion brands, designers' fragrance brands and luxury brands.

[Insert Table 9, Figure 4]

5.4.2 Counterfactual Experiments

To gain deeper insights into the impacts of counterfeiting, we conduct counterfactual experiments to explore the overall impact of counterfeiting on the market and platform welfare. We design the treatments as modifying product counterfeiting probabilities or the market structure by introducing simulated counterfeit products; then we simulate individual data to model their choice-making process in controlled or treated cases and examine how user utility and market shares would change if managerial interventions on the counterfeiting issue are applied.

We make two treatments, simulating two different scenarios. First, we polarize the counterfeiting probability of listings in our sample to approximate the case if the platform applies detection algorithms and disclose prediction results in a banner attached to each listing as a reference for consumers' judgment, which is, to increase those Likely Counterfeits while decrease that of Likely Authentics by 50% (normalized within the range zero to one), denoted as *Treatment*

I. Second, we generate 100 products by randomly drawing product characteristics based on the sample distribution, fix their counterfeiting probability at 0.8, and include them in the market competition of our real-world data set, denoted as *Treatment II*. This is to explore how the sales of Likely Authentic products will be affected when a significant amount of knockoffs flood in and intensify the competition.

To implement the experiments and compare results of the treated groups and the control (original) sample, we design a consumer simulation process following the logistic in model estimation, which provides computational foundation for the experiments. The general idea is to simulate unobserved individual characteristics ($\alpha_v v_i$ and $\gamma_v v_i$), plug them as well as estimated product coefficients in individual utility function (Equation (1)), and generate individual utility gained from each product to each consumer, based on which we can observe individual optimal choices, summarize total number of purchases for each product, and calculate the market share. Specifically, first, we draw the random part of the price ($\alpha_v v_i$) and counterfeiting probability coefficient ($\gamma_v v_i$) as suggested in Equation (3) for 300,000 consumers, α_v and γ_v estimated and given in Table 7b and v_i normally distributed by assumption. Second, we obtain product unobserved characteristics ξ_{jt}. Based on Equation (4), the market share of product j and outside alternatives are S_{jt} and S_{ot}, respectively.

$$S_{jt} = \frac{\exp\left(X_{jt}^{v}\beta^{v} + \bar{\alpha}P_{jt} + \bar{\gamma}C_{j} + X_{j}^{inv}\beta^{inv} + \xi_{jt}\right)}{1 + \sum_{l=1}^{J}\exp\left(X_{lt}^{v}\beta^{v} + \bar{\alpha}P_{lt} + \bar{\gamma}C_{l} + X_{l}^{inv}\beta^{inv} + \xi_{lt}\right)} \tag{5}$$

$$S_{0t} = \frac{1}{1 + \sum_{l=1}^{J}\exp\left(X_{lt}^{v}\beta^{v} + \bar{\alpha}P_{lt} + \bar{\gamma}C_{l} + X_{l}^{inv}\beta^{inv} + \xi_{lt}\right)} \tag{6}$$

Therefore, we can solve the product unobservable using market share data, observed product characteristics, and coefficient estimates as below.

$$\delta_{jt} = log\left(S_{jt}\right) - \log(S_{ot}) - \left(\bar{\alpha}P_{jt} + \bar{\gamma}C_{j} + X_{jt}^{v}\beta^{v} + X_{j}^{inv}\beta^{inv}\right) \tag{7}$$

Last, with random coefficients ($\alpha_v v_i$ and $\gamma_v v_i$) and product unobserved effect (ξ_{jt}) generated, leveraging the mean coefficient estimates ($\bar{\alpha}, \bar{\gamma}, \beta^v, \beta^{inv}$) from our structural econometric models, we are able to calculate individual-product pairwise utilities given product attributes including the price (P_{jt}), counterfeiting probability (C_j), and other characteristics (X_{jt}^v and X_j^{inv}), from original real-world data set or treated data sets. Individual discrete and optimal purchasing choices and aggregated market shares are also directly obtained from this utility matrix. Following the three steps above, we are able to simulate the entire process by which consumers choose the product that maximizes their utility, and to infer purchasing decisions and total utility, and ultimately summarize market shares. Utility and market shares of the original sample and treated samples are calculated based on the same group of simulated consumers.

After comparing with the decision matrix generated from the original data set, we document an 18.16% increase of average consumer utility in *Treatment I*, namely polarizing the counterfeiting probability as a proxy of platform providing detection reference to users. Also, consumers who choose the outside option and leave the platform decrease by 1.42%. *Treatment II* didn't indicate significant changes on the men's fragrances product category. These findings validate the importance of anti-counterfeiting policy in maintaining consumer trust and reducing customer churn. (Note that for the category of men's fragrances, estimation of individual heterogeneity in prices (α_v) is not significant, so we only include the random effect of counterfeiting probability in the first step of simulation for this category.)

6. Robustness Check Using Utility Goods

To explore the counterfeiting impact in another product category, and to validate the robustness of our methodology and findings, we conduct our entire analysis on a utility product — cell-phone

wireless chargers. This product is intrinsically different from perfumes, and is a rather utilitarian commodity, yet one that is subject to active counterfeiting. First, consumers' preference of wireless chargers is more homogeneous than that of fragrances. Further the preferences are much more vertical rather than horizontal, in that consumers essentially care about quality rather than other subjective non-functional product characteristics. As a result, a counterfeit wireless charger tends to hurt consumer utility explicitly for not satisfactorily performing the charging function. Second, when it comes to wireless chargers, people care less about brand value and personal taste, as they do for high-end fragrances. Last, as consumers have a higher level of acceptance to chargers of less-established brands, we anticipate less counterfeit activity in the charger category as compared to fragrances. On Amazon, we observe that about half of wireless-charger related ASINs list a single seller; those listings with multiple sellers, the average number of sellers is less than 5, which is substantially smaller than in the case of fragrances. This allows us to streamline the set of instruments by dropping the ones related to multiple sellers, as we explain below.

We define 15 topics for cell-phone wireless chargers: counterfeit warning (fake), overall sentiments, attitudes towards price, shipping and quality in terms of charging speed, connection, flexibility, lifespan, design, etc. (positive or negative). We extract topic indicators using anchored correlation explanation topic modeling. Variables on authenticity, overall sentiment, charging speed, lifespan, compatibility, shipping, and services as well as rating distribution metrics are selected to train a random forest classifier. 200 of the 565 listings in our data sample are labeled as Likely Counterfeit or Likely Authentic to construct the training data set, and the predicted probability is used as the variable of interest in the econometric model.

As mentioned earlier, when estimating the counterfeiting effect in cell-phone wireless chargers, one substantial difference from the prior analysis of fragrance product category is that

most chargers have only one seller per ASIN, rather than multiple. Therefore, we exclude the number of options and standard deviation of price from the set of instrument variables. Instead, we include topic variables regarding authenticity, charging speed, and lifespan as instruments for counterfeiting probability. Compared to fragrances where one listing is often linked to over a dozen sellers, wireless chargers can be considered a special and simpler case where the number of options is one. Accordingly, the predicted counterfeiting probability suggests the likelihood of a product (instead of a listing) being counterfeit. In addition, we define characteristics on multi-charging design (single charging, two-in-one, three-in-one, or four-in-one), and station design (pad, stand, or station) and generate sums of such characteristics over products within or across brands as the BLP-style instruments for endogenous prices. Lastly, we conducted purchasing experiments to convert charger sales rank into charger market share, as we did for fragrances.

The results for the charger category are reported in Table 10, which suggest that the counterfeit probability significantly hurts consumers' utility. A higher price reduces consumers' willingness to buy, while rating, images in reviews and free shipping have positive impacts on sales. Three-in-one and four-in-one charging stations are more popular than single or two-in-one chargers. While the results are quite consistent with what we found for men's fragrances, it is worth noting that the standard deviation of the individual price coefficient is significant in the case of cell-phone wireless chargers, indicating that user preferences for prices are more heterogeneous for the charger as compared to the than the fragrance category. Also, the individual coefficients of counterfeiting probability follow a normal distribution with estimated mean -2.788421 and estimated standard deviation 2.082199 (Column 4), suggesting that utility of 91.97% consumers are negatively affected by the likelihood of encountering a counterfeit product, which is substantially higher than the corresponding figure for fragrances.

[Insert Table 10a, 10b]

Finally, we conduct counterfactual experiments on the cell phone wireless charger sample, along the lines of what we discussed in the previous section. Among 1,700,000 choices made by 100,000 simulated consumers in 17 markets, Likely Authentic listings are selected 457,519 times as the optimal choice in the original sample and 567,198 times in the *Treatment I*. In other words, polarizing the perceived counterfeiting probability as a proxy of detection disclosed to users increases the market shares of Likely Authentic products by 6.45%. Similarly, market shares of Likely Counterfeit products in the first treatment group decreased by 7.92%; and 1.47% more customers leave the platform without purchasing anything, as compared to the original case. Average utility of consumers across markets increases by 0.3%. In *Treatment II* where 100 simulated Likely Counterfeit products enter the market and join the competition, 3.84% of the total market share is taken away from Likely Authentic products by Likely Counterfeit products. Results of counterfactual experiments validate the benefits of counterfeit identification for the welfare of authentic sellers and consumers, and also for a significantly positive impact on reducing customer churn and on platform welfare.

[Insert Table 11]

7. Discussion & Conclusions

The e-Commerce market has been disrupted by the proliferation of knockoffs, and this has become more critical with the growing presence of third-party sellers on platforms like Amazon. Inspite of the potential increase in overall revenue, counterfeiting hampers platform development with many brands leaving Amazon, such as Nike and PopSockets (Barkho 2020). The overall impact of online deceptive counterfeiting activities remains unclear to not only the platform, but also consumers

and original sellers. However, this area has been lack of study partly due to the difficulty of identification.

Our work is one of the only studies to explore the economic impact of deceptive counterfeits in online retail markets, and it contributes to the empirical literature on the identification and impact of online fake products. By applying NLP techniques on review data crawled from Amazon, we extract consumer opinions on product characteristics and develop a machine learning methodology to characterize the intensity of counterfeiting activity at listing level . We also develop a choice-modeling framework and solve the endogeneity issue of counterfeiting probability to quantify the economic impacts of counterfeiting activities on consumer utility and sales of likely authentic products . We design counterfactual experiments to explore how consumer utility, market shares and platform welfare change if the platform apply detection mechanisms and reveal the identification results to users. Our findings suggest that online customers can take the advantage of user-generated reviews to try to identify and avoid knockoffs; they tend to be more cautious when purchasing high-end products or best sellers; the inferred probability of encountering a fake product has a negative effect on consumer mean utility and a larger proportion of consumers are negatively affected by counterfeiting probability in utilitarian good categories; likely counterfeit products have significant substitution effects on the sales of likely authentic products in online marketplace, a 10% increase in one Likely Counterfeit product's price causing the market share of each Likely Authentic product to increase on average by 0.11%, and the effect is stronger for expensive brands; detecting and disclosing likely counterfeit products explicitly to platform users can improve both user utility and original sellers welfare, but disclosing the prevalence of counterfeiting may hurt consumer trust, increasing customer churn.

Our study provides managerial implications to online consumers, sellers, and e-Commerce platform operators as well. For consumers, our study confirms the potential of user-generated reviews in identifying knockoffs. Our method implies consumers should pay more attention to reviews with helpful votes and be sensitive to the disclosing reviews even if they are sparse or not at a noticeable place in first several pages. The percentage of four-star ratings are often ignored but proved to be useful and positively correlated with product credibility. We also provide evidence of uneven counterfeiting intensity among products of the same category. Obviously, consumers should expect more counterfeiters in the high-end and read reviews more comprehensively to avoid fake products and monetary loss. For original merchants, we emphasize the disturbance caused by counterfeiting in two ways. First, besides creating an exquisite and complete product page with all the information a consumer would need, sellers should also monitor the review section since negative reviews especially those doubting the authenticity due to a quality issue may make the product looks mixed up with knockoffs and mislead following consumers. They should not only respond to direct messages in a timely manner and try to solve problems for consumers before they leave a risky review, but also be careful to other sellers listed under the same ASIN for counterfeiters there can bring a negative spillover effect if they receive disclosing reviews. Second, even if they manage to keep a clean review record for their products and minimize the risk of being considered as knockoffs, original sellers still suffer from the counterfeiting issue for counterfeiters take away their market shares and profits. This finding points to the inevitability of integrating an anti-counterfeiting validation design into the manufacturing process. Although coming with a cost, it still seems an optimal strategy if original sellers can embrace the challenge and actively provide validation services such as a unique code attached to each authentic product by which consumers can check or scan after purchase to validate

its authenticity. From the perspective of platform operators, the counterfeiting problem seems less threatening yet more complicated. On the one hand, the existence of counterfeiters objectively increases market size on the seller side and subsequently the consumer side, network value therefore enhanced. Our results of counterfactual experiments show that applying harsh forbidding policies and revealing counterfeiter to customers may cause a sharp decrease of user base, revenues, and the platform's competitiveness in the short run. On the other hand, without appropriate intervention, deceptive knockoffs will steal a great amount of market share from original brands and even drive them out, for original sellers would rather give up limited profits at a sacrifice of brand reputation being jeopardized. As knockoffs take greater and greater shares on the platform, it will be much more difficult for consumers to find an ideal product which they have confidence in; matching efficiency will sharply decline, which hinders the platform from thriving or even threaten its survival. It is of vital importance that e-Commerce platform balances the volume and a healthy ecosystem of the marketplace. Operators should detect and constrain the activity of counterfeiters to an appropriate extent and block some of them when necessary but be careful of what's to disclose to users. Our identification approach is provided as an efficient way to identify knockoffs regularly and generally with a limited computational cost. Different actions can be determined according to the probability.

Our work has some limitations. First, our identification approach is built at listing level. Due to the fact that Amazon ask merchants to select an existing ASIN to list products they sell and apply to create a new ASIN only when no other merchants are selling the same products, it's very commonly seen multiple sellers under the same listing in the categories with many third-party sellers (the men's fragrances category in our case). Also, reviews do not suggest which seller corresponding purchases came from. Therefore, instead of predicting a product is fake or not, or

from counterfeiters or genuine sellers, we are only able to characterize the counterfeiting activities as the probability of encountering a knockoff if purchasing from an ASIN. Second, the training set is constructed by coders labelling the listings after reading their reviews, which inevitably introduce some subjectivity. Neither consumers, researchers nor sometimes even the platform itself can fully access the ground truth, which can only be confirmed by counterfeiters who has yet no reason to do so. We take measures to try to minimize the subjectivity in identification, but the only way to eliminate it and validate the true quality is to buy all the products from every seller under each listing, which is barely feasible. Having that said, the platform can better determine the true authenticity leveraging private information such as merchant activities, financial records, and inventory examination, etc.. With a more accurate and larger training set, the platform can make better use of the identification model and generate valuable results. Last, we only study two product categories because our econometric model requires us to specify product characteristic specific to that category. Although we select categories representing different types of goods, future study can expand the work to other online markets or generalize it to platforms beyond online retail markets.

References

Archak, N., Ghose, A. and Ipeirotis, P.G., 2011. Deriving the pricing power of product features by mining consumer reviews. *Management science*, *57*(8), pp.1485-1509.

Bai, X., Marsden, J.R., Ross Jr, W.T. and Wang, G., 2020. A Note on the Impact of Daily Deals on Local Retailers' Online Reputation: Mediation Effects of the Consumer Experience. Information Systems Research, 31(4), pp.1132-1143.

Barkho, G. 2020, Why some DTC brands are pulling out of Amazon, Modern Retail, < https://www.modernretail.co/startups/dtc-brands-leaving-amazon/>

Berry, S., Levinsohn, J. and Pakes, A., 1995. Automobile prices in market equilibrium. Econometrica: Journal of the Econometric Society, pp.841-890.

Berry, S. and Pakes, A., 2007. The pure characteristics demand model. International Economic Review, 48(4), pp.1193-1225.

Berzon, A., Shifflett, S., and Scheck, J. 2019, Amazon Has Ceded Control of Its Site. The Result: Thousands of Banned Unsafe or Mislabeled Products, The Wall Street Journal, <https://www.wsj.com/articles/amazon-has-ceded-control-of-its-site-the-result-thousands-of-banned-unsafe-or-mislabeled-products-11566564990>

Bressler, M.S. and Bressler L.A. 2018. Makers vs. Fakers: how counterfeit goods hurt competition, harm the economy, and kills consumers, available from Researchgate.net.

Chaloux, J.M., Hayes, J.P., Aggarwal, M. and Tsatsoulis, P.D., Accenture Global Solutions Ltd, 2020. Detection of counterfeit items based on machine learning and analysis of visual and textual data. U.S. Patent 10,691,922.

Chevalier, J. and Goolsbee, A., 2003. Measuring prices and price competition online: Amazon. com and BarnesandNoble. com. Quantitative marketing and Economics, 1(2), pp.203-222.

Cheung, M., She, J. and Liu, L., 2018, April. Deep learning-based online counterfeit-seller detection. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 51-56). IEEE.

Chevalier, J.A. and Mayzlin, D., 2006. The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3), pp.345-354.

Cho, W.Y. and Ahn, B.H., 2010. Versioning of information goods under the threat of piracy. Information Economics and Policy, 22(4), pp.332-340.

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

Coppola, D. 2021, Quarterly share of e-commerce sales of total U.S. retail sales from 1st quarter 2010 to 1st quarter 2021, Statista, <https://www.statista.com/statistics/187439/share-of-e-commerce-sales-in-total-us-retail-sales-in-2010/>

Decker, R. and Trusov, M., 2010. Estimating aggregate consumer preferences from online product reviews. International Journal of Research in Marketing, 27(4), pp.293-307.

Dewan, S. and Hsu, V. 2004. Adverse selection in electronic markets: evidence from online stamp auctions. Journal of Industrial Economics LII (4), pp. 497-516.

Dimoka, A., Hong, Y. and Pavlou, P.A., 2012. On product uncertainty in online markets: Theory and evidence. MIS quarterly, pp.395-426.

Droesch, B. 2021, Amazon dominate US ecommerce, though its market share varies by category, eMarketer, <https://www.emarketer.com/content/amazon-dominates-us-ecommerce-though-its-market-share-varies-by-category>

Gallagher, R.J., Reing, K., Kale, D. and Ver Steeg, G., 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. Transactions of the Association for Computational Linguistics, 5, pp.529-542.

Ghose, A., Ipeirotis, P.G. and Li, B., 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. Marketing Science, 31(3), pp.493-520.

Ghose, A. and Sundararajan, A., 2006. Evaluating pricing strategy using e-commerce data: Evidence and estimation challenges. Statistical Science, 21(2), pp.131-142.

Givon, M., Mahajan, V. and Muller, E., 1995. Software piracy: Estimation of lost sales and the impact on software diffusion. Journal of Marketing, 59(1), pp.29-37.

Hui, K.L. and Png, I., 2003. Piracy and the legitimate demand for recorded music. Contributions to Economic Analysis and Policy, 2(1), p.11.

Jain, S., 2008. Digital piracy: A competitive analysis. Marketing science, 27(4), pp.610-626.

Kennedy, J.P. 2020. Counterfeit Products Online, in The Palgrave Handbook of Cybercrime and Cyberdeviance, T. J. Holt, A. M. Bossler (eds.).

Lahiri, A. and Dey, D., 2013. Effects of piracy on quality of information goods. Management Science, 59(1), pp.245-264.

Lewis, G. and Zervas, G., 2016. The welfare impact of consumer reviews: A case study of the hotel industry. Unpublished manuscript.

Lu, S., Wang, X. and Bendle, N., 2020. Does Piracy Create Online Word of Mouth? An Empirical Analysis in the Movie Industry. Management Science, 66(5), pp.2140-2162.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior.

Pieterse, J., 2021. Are Fragrances On Amazon Real?, < https://fragrancetoday.com/are-fragrances-on-amazon-real/>

Qian, Y., 2014. Counterfeiters: Foes or friends? How counterfeits affect sales by product quality tier. Management Science, 60(10), pp.2381-2400.

Qian, Y., Gong, Q. and Chen, Y., 2015. Untangling searchable and experiential quality responses to counterfeits. Marketing Science, 34(4), pp.522-538.

Quora, 2019. How much perfume sold on Amazon is counterfeit?, https://www.quora.com/How-much-perfume-sold-on-Amazon-is-counterfeit

Sharma, A., Srinivasan, V., Kanchan, V. and Subramanian, L., 2017, August. The fake vs real goods problem: microscopy and machine learning to the rescue. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2011-2019).

Silcox, K., 2021. The Hidden Costs of Buying Fake Cosmetics on Sites Like Amazon, < https://money.com/fake-amazon-makeup-skincare/>

Smith, M.D. and Telang, R., 2009. Competing with free: The impact of movie broadcasts on DVD sales and Internet piracy. Mis Quarterly, pp.321-338.

Song, M., 2015. A hybrid discrete choice model of differentiated product demand with an application to personal computers. International Economic Review, 56(1), pp.265-301.

Steele, C., 2019. The Ugly Truth About Amazon Beauty, <https://www.pcmag.com/opinions/the-ugly-truth-about-amazon-beauty>

Sundararajan, A., 2004. Managing digital piracy: Pricing and protection. Information Systems Research, 15(3), pp.287-308.

Vincent, D.W., 2015. The Berry–Levinsohn–Pakes estimator of the random-coefficients logit demand model. The Stata Journal, 15(3), pp.854-880.

Wang, Q., Li, B. and Singh, P.V., 2018. Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. Information Systems Research, 29(2), pp.273-291.

Zeibak, 2020. How to Win the Amazon Buy Box in 2021, Amazon & Marketplaces, <https://tinuiti.com/blog/amazon/win-amazon-buy-box/>

Figures and Tables

Figure 1. Typical Amazon ASIN Listing with Multiple Sellers

Figure 2. Percentage of Disclosing Reviews

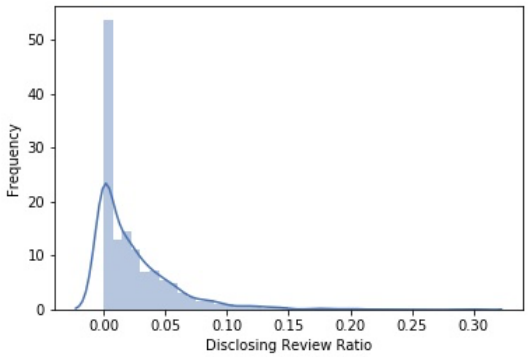

Figure 3. Price and Sales Distributions for Likely Authentic and Likely Counterfeit Products

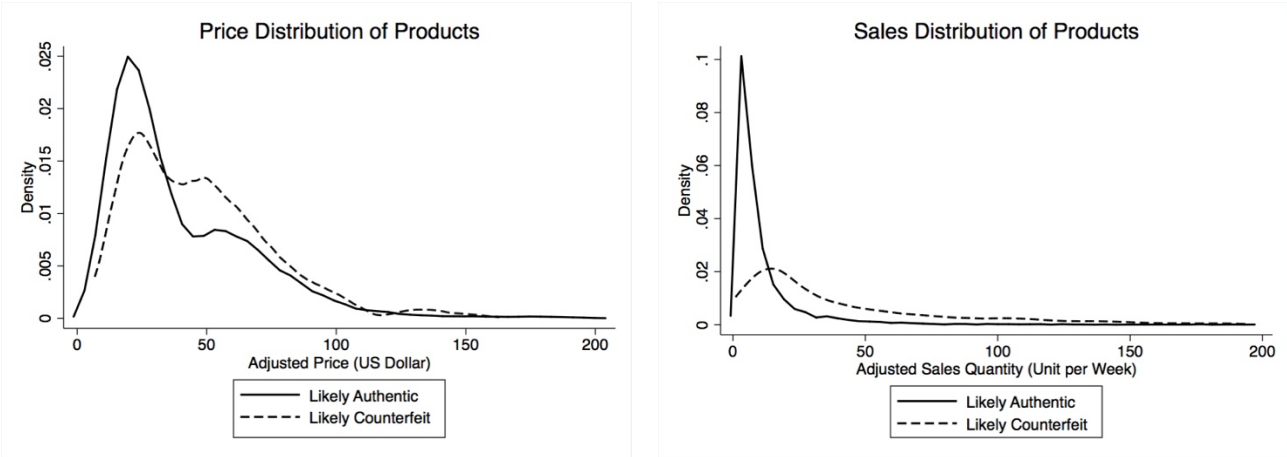

Figure 4. Cross-Price Elasticity and Counterfeit Percentage by Brand Category

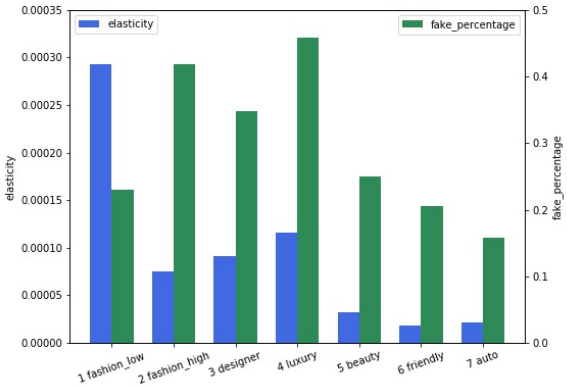

Table 1a. Summary Statistics of Sales Data

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
Week	9,520	5.390	2.815	1	10
Share	9,520	0.000221	0.00231	9.12e-08	0.0867
Price	9,520	42.12	32.51	2.560	380.8
Sales	9,520	35.73	381.6	0.0131	14,679
No. Ratings	9,520	874.5	2,188	5.667	48,266
Amazon's Choice	9,520	0.222	0.383	0	1
Rating	9,520	4.561	0.188	3	5
Rank	9,520	88,338	58,905	199.8	522,204

Note: Sales data is based on the weekly average amount; No. Rating and Rating are accumulative values in the current week.

Table 1b. Summary Statistics of Product Data

	VARIABLES	DEFINITION	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
Product Characteristics	Size	Perfume size in oz	1,037	3.628	1.800	0.0300	17
	Vol 1	=1 if Eau de Cologne	1,037	0.217	0.412	0	1
	Vol 2	=1 if Eau de Toilette	1,037	0.669	0.471	0	1
	Vol 3	=1 if Eau de Parfum	1,037	0.103	0.304	0	1
	Vol 4	=1 if Parfum	1,037	0.00675	0.0819	0	1
	Free Shipping	Level of shipping cost	1,037	2.777	0.493	0	3
	Free Return	=1 if free return	1,037	0.00771	0.0875	0	1
	No. Options	No. sellers listed per ASIN	1,037	13.96	12.10	1	63
	Price S.D.	Price S.D. within ASIN	1,037	6.334	6.245	0	86.03
	No. Disclosing	No. disclosing reviews	1,037	5.878	16.27	0	209
	Disclosing Ratio	% of disclosing reviews	1,037	0.0257	0.0354	0	0.300
	Counterfeit_10_01	=1 if No. Disclosing>=10 and Disclosing Ratio>0.01	1,037	0.135	0.342	0	1
	Counterfeit_5_01	=1 if No. Disclosing>=5 and Disclosing Ratio >0.01	1,037	0.236	0.425	0	1
Rating Distribution	Review_1_star_pct	% of one star reviews	1,037	0.102	0.0730	0	0.462
	Review_3_star_pct	% of three star reviews	1,037	0.0493	0.0413	0	0.333
	Review_5_star_pct	% of five star reviews	1,037	0.724	0.111	0.167	1
	Rating_1_star_pct	% of one star ratings	1,037	0.0407	0.0270	0	0.200
	Rating_3_star_pct	% of three star ratings	1,037	0.0502	0.0271	0	0.240
	Rating_5_star_pct	% of five star ratings	1,037	0.782	0.0757	0.410	0.950
Review Metrics (Average per review)	No. Helpful Votes	No. helpful votes	1,037	0.733	1.615	0.0357	48.47
	No. Images	No. images	1,037	0.0369	0.0787	0	1.400
	Summary Wordcount	No. words in summary	1,037	3.360	0.593	1.333	7
	Text Wordcount	No. words in text	1,037	19.67	8.788	4.143	96.18
Topic Variables	Fake		1,037	0.0421	0.0488	0	0.375
	Like		1,037	0.145	0.0738	0	0.600
	Dislike		1,037	0.0430	0.0409	0	0.333
	Price_positive		1,037	0.0865	0.0621	0	0.667
	Price_negative	% of reviews	1,037	0.0495	0.0463	0	0.500
	Scent_positive	With certain topic	1,037	0.0686	0.0507	0	0.400
	Scent_negative		1,037	0.0581	0.0452	0	0.300
	Last_positive		1,037	0.121	0.0707	0	0.667
	Last_negative		1,037	0.488	0.138	0	1
	Package_positive		1,037	0.00855	0.0147	0	0.143
	Shipping_positive		1,037	0.0240	0.0299	0	0.231
Classification Result	Counterfeit (0/1)	=1 if likely counterfeit	1,037	0.335	0.472	0	1
	Prob (Counterfeit=1)	Probability of Counterfeit=1	1,037	0.424	0.233	0.01000	0.990

Table 2. Correlation Matrix

	1.	2.	3,	4,	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1. No. Reviews	1.000															
2. No. Disclosing	0.766	1.000														
3. Disclosing Ratio	0.059	0.258	1.000													
4. Counterfeit_10_01	0.511	0.626	0.345	1.000												
5. Rating_1_star_pct	-0.030	-0.149	-0.095	-0.170	1.000											
6. Rating_3_star_pct	0.044	-0.151	-0.204	-0.202	0.620	1.000										
7. Rating_4_star_pct	-0.037	-0.165	-0.233	-0.245	0.111	0.415	1.000									
8. Vine	-0.023	-0.055	-0.088	-0.082	0.101	0.223	0.128	1.000								
9. Fake	0.034	0.208	0.792	0.276	-0.099	-0.224	-0.232	-0.095	1.000							
10. Price_positive	0.007	0.028	-0.065	0.035	-0.056	0.030	-0.011	0.035	-0.049	1.000						
11. Scent_positive	-0.008	-0.055	-0.146	-0.091	0.180	0.230	-0.088	0.072	-0.140	-0.139	1.000					
12. Fake_h	0.198	0.431	0.516	0.511	-0.128	-0.259	-0.323	-0.087	0.675	0.014	-0.154	1.000				
13. Dislike_h	0.137	0.158	0.155	0.160	0.252	0.072	-0.131	0.042	0.251	-0.009	-0.100	0.405	1.000			
14. Last_positive_h	0.048	0.070	-0.067	0.046	-0.132	0.060	0.130	0.123	-0.154	-0.021	-0.046	-0.039	-0.061	1.000		
15. Overall_h	-0.195	-0.286	-0.216	-0.363	-0.154	0.024	0.113	-0.008	-0.281	0.041	0.183	-0.544	-0.467	0.074	1.000	
16. Fake Dummy	0.277	0.368	0.222	0.387	0.042	-0.182	-0.271	-0.142	0.233	-0.106	-0.156	0.380	0.291	-0.174	-0.349	1.000

Table 3. Anchored Topic Model

Topic	Anchored Words	Model Generated Keywords
Fake	Fake, defective, knock, knockoff, counterfeit, diluted, water	Fake, water, knock, counterfeit, knockoff, defective, real, cool
Like	Love, like, amazing, best, great, awesome, satisfy, favorite, complement	Like, favorite, compliment, best, great, lot, love, satisfy
Dislike	Waste, disappoint, bad, dislike, poor, terrible	Bad, waste, disappoint, terrible, poor, dislike, money, total
Price_positive	Price, deal, sale, value, worth, bargain, good	Good, price, value, worth, deal, sale, bargain, size
Price_negative	Critique, expensive	Expensive, di, gio, aqua, creed, aventus, acqua, similar
Scent_positive	Scent, attract, crisp, nice, classy, pleasant, delicious, classic	Scent, nice, pleasant, classic, delicious, crisp, classy, attract
Scent_negative	Overpower, strange, strong, cheap, weird, disgust, much	Strong, cheap, overpower, weird, strange, disgust, offensively, cheerful
Last_positive	Last, long, throughout, all, day	Long, day, time, stay, doesn, lasting, father, valentine
Last_negative	Minute, away, lost, flourish	Minute, away, cologne, bottle, just, say, try, use
Package_positive	Package, fancy	Package, fancy, et, est, je, le, tr, pa
Shipping_positive	Shipping, fast	Fast, shipping, delivery, ship, super, described, service, quick

Table 4. Classification Model Accuracy

Model	(1) LR	(2) NB	(3) CART	(4) RF	(5) SVM	(6) LDA
Overall, Disclose Ratio, Counterfeit_5_01, Vine, Fake_h, Dislike_h, Scent_positive, Last_positive, Price_positive, Rating_1_star_pct, Rating_3_star_pct, Rating_4_star_pct	0.71	0.68	0.72	0.79	0.70	0.72
No. Disclosing, Counterfeit_10_01, Overall_h, Vine, Fake_h, dislike_h, Scent_positive, Price_positive, Last_positive_h Rating_1_star_pct, Rating_3_star_pct, Rating_4_star_pct	0.70	0.73	0.71	0.83	0.69	0.73

Note: LR = Logistic Regression, NB = Naïve Bayes, CART = Classification and Regression Trees, RF = Random Forest, SVM = Support Vector Machine, LDA = Linear Discriminant Analysis

Table 5. IV-Relevance: First Step Regression of 2SLS

VARIABLES	(1) Price	(2) Price	(3) Counterfeit Prob	(4) Counterfeit Prob
Size_other	-0.299***	-0.182***		0.00124
	(0.0878)	(0.0668)		(0.000982)
Size_rival	-0.0252	-0.0304		-0.000509***
	(0.0247)	(0.0195)		(0.000157)
Vol1_other	0.211	0.142		-0.00118
	(0.464)	(0.385)		(0.00471)
Vol1_rival	-0.0268	0.0300		0.00342***
	(0.262)	(0.229)		(0.00127)
Vol2_other	0.822***	0.501**		-0.00479
	(0.302)	(0.242)		(0.00339)
Vol2_rival	0.157**	0.137**		0.000990***
	(0.0729)	(0.0644)		(0.000374)
Vol3_other	5.613***	3.289***		0.0160***
	(0.595)	(0.476)		(0.00509)
Vol3_rival	-0.289	-0.0309		0.00352***
	(0.225)	(0.174)		(0.00108)
Vol4_other	5.785*	1.583		0.0324
	(3.250)	(2.927)		(0.0240)
Vol4_rival	-2.407	-2.116		0.00629
	(1.911)	(1.826)		(0.00517)
No. Options		-0.602***	0.00159**	0.00184***
		(0.0538)	(0.000642)	(0.000638)
Price S.D.		2.833***	0.00565***	0.00362***
		(0.225)	(0.00128)	(0.00114)
Scent_positive		-24.37*	-0.594***	-0.562***
		(13.05)	(0.123)	(0.124)
Last_positive		19.05*	-0.523***	-0.503***
		(10.47)	(0.0930)	(0.0928)
Scent_negative		-13.16	-0.0165	0.0358
		(15.40)	(0.152)	(0.155)
Last_negative		-1.553	-0.00992	-0.0368
		(4.984)	(0.0521)	(0.0520)
Price_positive		-6.517	-0.0375	-0.0170
		(9.355)	(0.114)	(0.114)
Constant	75.94***	61.31***	0.479***	0.498***
	(16.30)	(17.97)	(0.0352)	(0.0562)
Observations	10,047	10,047	9,520	9,520
R-squared	0.142	0.433	0.082	0.122

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6. 2SLS Estimation Results

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
Price	-0.000256*	-0.000361**	-0.000261*	-0.00119	-0.00235	-0.00138
	(0.000146)	(0.000179)	(0.000151)	(0.00169)	(0.00187)	(0.00190)
Counterfeit Prob	-0.106***	-0.111***	-0.121***	-0.709***	-0.920***	-0.622***
	(0.0365)	(0.0425)	(0.0391)	(0.194)	(0.216)	(0.196)
Log No. Ratings	0.0470***	0.0474***	0.0479***	0.416***	0.418***	0.414***
	(0.00369)	(0.00412)	(0.00387)	(0.0353)	(0.0360)	(0.0348)
Rating	-0.0228*	-0.0219	-0.0244*	0.413**	0.344	0.493***
	(0.0138)	(0.0149)	(0.0141)	(0.188)	(0.213)	(0.190)
No. Images	0.0869***	0.0935***	0.0857***	2.474***	1.824**	2.729***
	(0.0326)	(0.0335)	(0.0327)	(0.848)	(0.888)	(0.831)
No. Helpful Votes	0.00431***	0.00449***	0.00447***	-0.00771	-0.00268	-0.00642
	(0.00156)	(0.00157)	(0.00157)	(0.00633)	(0.00664)	(0.00620)
Text Wordcount	0.000107	0.000104	0.000101	0.00932	0.00506	0.0110
	(0.000308)	(0.000310)	(0.000310)	(0.0113)	(0.0116)	(0.0111)
Free Shipping			-0.00712			-0.306***
			(0.00552)			(0.0768)
Size		0.00184	0.00170		0.0283	0.0107
		(0.00148)	(0.00146)		(0.0187)	(0.0179)
Vol 1		0.0376				
		(0.0377)				
Vol 2		0.0359			0.286***	
		(0.0375)			(0.0970)	
Vol 3		0.0453			0.176	
		(0.0387)			(0.135)	
Vol 4		0.0366				
		(0.0488)				
Constant	-0.0777	-0.122	-0.0570	-4.251***	-4.185***	-3.941***
	(0.0627)	(0.0766)	(0.0640)	(0.951)	(1.057)	(0.957)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Tier FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,520	9,520	9,520	807	807	807
R-squared	0.034	0.032	0.031	0.155	0.131	0.189

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 7a. Results for BLP Choice Models with Random Coefficients: Fragrances

VARIABLES	(1)	(2)	(3)	(4)
Price	-0.0036994***	-0.0063561***	-0.0006103	-0.001654
	(0.0007941)	(0.0010277)	(0.0017829)	(0.0018782)
Counterfeit Prob	-1.030453*	-1.017308*	-2.277378**	-2.315536**
	(0.5935036)	(0.6114545)	(1.158386)	(1.04508)
Rating	0.8727193***	0.9302191***	-0.3877383**	-0.7098856***
	(0.0551215)	(0.0605114)	(0.1919291)	(0.2132161)
Log No. Ratings	0.5784206***	0.5862219***	0.6452466***	0.6430962***
	(0.0149595)	(0.015585)	(0.0396473)	(0.0387462)
No. Images	1.491585***	1.827954***	1.452147*	-0.5057678
	(0.1295981)	(0.1387974)	(0.7757194)	(0.9115698)
No. Helpful votes	0.0585438***	0.0681108***	0.0250802***	-0.0148516
	(0.0054414)	(0.0057258)	(0.0061424)	(0.0118649)
Text Wordcount		-0.0062778***		0.0257132***
		(0.0011679)		(0.006658)
Size		0.0179852***		0.0371665**
		(0.0053769)		(0.0186691)
Vol 1		0.2812084**		.
		(0.1370806)		(.)
Vol 2		0.022369		-0.0089874
		(0.1371093)		(0.0974715)
Vol 3		0.4600137***		-0.379501***
		(0.1420566)		(0.1353434)
Vol 4		0.4115168**		.
		(0.1804814)		(.)
Tier FE	Yes	Yes	Yes	Yes
Observations	9520	9520	807	807

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 7b. Standard Deviation of Individual Random Coefficients: Fragrances

VARIABLES	(1)	(2)	(3)	(4)
SD – Price	0.0000662	0.0000934	2.70e-07	7.53e-11
	(0.065769)	(0.0595901)	(0.073156)	(0.0697974)
SD – Counterfeit Prob	2.644236***	2.659966***	2.205202**	1.88428**
	(0.5992778)	(0.6010042)	(1.003856)	(0.9409133)

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 8. Price Elasticity in Ten Markets

	(1) Own Elasticity Fake on Fake	(2) Cross Elasticity Fake on Fake	(3) Cross Elasticity Fake on Real	(4) Own Elasticity Real on Real
Market 1	-0.323968***	0.000225***	0.000113***	-0.241899***
	(0.013164)	(0.000003)	(0.000002)	(0.000299)
Market 2	-0.320058***	0.000233***	0.000117***	-0.242565***
	(0.013451)	(0.000004)	(0.0000001)	(0.000321)
Market 3	-0.320543***	0.000218***	0.000107***	-0.241638***
	(0.012869)	(0.000003)	(0.000001)	(0.000297)
Market 4	-0.318547***	0.000223***	0.000109***	-0.239869***
	(0.011597)	(0.000003)	(0.000001)	(0.000278)
Market 5	-0.317793***	0.000206***	0.000103***	-0.239948**
	(0.011597)	(0.000003)	(0.000001)	(0.000276)
Market 6	-0.309141***	0.000226***	0.000116***	-0.243010***
	(0.012059)	(0.000004)	(0.000001)	(0.000294)
Market 7	-0.314103***	0.000205***	0.000105***	-0.240456***
	(0.012030)	(0.000004)	(0.000001)	(0.000299)
Market 8	-0.317099***	0.000199***	0.000103***	-0.241858***
	(0.011472)	(0.000004)	(0.000001)	(0.000299)
Market 9	-0.319383***	0.000210***	0.000109***	-0.243340***
	(0.011727)	(0.000004)	(0.000001)	(0.000307)
Market 10	-0.325582***	0.000255***	0.000128***	-0.246482***
	(0.013446)	(0.000007)	(0.000002)	(0.000416)

Table 9. Average Price Elasticity in Different Brand Categories

	(1) Own Elasticity Fake on Fake	(2) Cross Elasticity Fake on Fake	(3) Cross Elasticity Fake on Real	(4) Own Elasticity Real on Real
Fashion_low	-0.223566***	0.000598***	0.000293***	-0.159890***
	(0.009829)	(0.000027)	(0.000007)	(0.002876)
Fashion_high	-0.280193***	0.000148***	0.000075***	-0.262640***
	(0.007227)	(0.000001)	(0.000000)	(0.005673)
Designer	-0.282122***	0.000183***	0.000091***	-0.268660***
	(0.005500)	(0.000003)	(0.000000)	(0.007165)
Luxury	-0.404167***	0.000220***	0.000116***	-0.386510***
	(0.006469)	(0.000002)	(0.000000)	(0.005465)
Beauty	-0.554014***	0.000049***	0.000032***	-0.227800***
	(0.035741)	(0.000001)	(0.000000)	(0.007822)
Friendly	-0.151891***	0.000029***	0.000018***	-0.130790***
	(0.005192)	(0.000000)	(0.000000)	(0.001768)
Auto	-0.187630***	0.000033***	0.000021***	-0.199780***
	(0.012209)	(0.000004)	(0.000000)	(0.007912)

Table 10a. Results for BLP Choice Models with Random Coefficients: Chargers

VARIABLES	(2)	(3)	(4)	(5)
Price	-0.073162**	-0.0578887**	-0.0681249**	-0.0446224***
	(0.0307476)	(0.0262541)	(0.0284603)	(0.023549)
Counterfeit Prob	-2.402878***	-2.932925***	-2.330667***	-2.788421***
	(0.5320905)	(0.5321975)	(0.56752)	(0.572094)
Rating	1.459121***	1.265042***	1.437825***	1.251411***
	(0.1930702)	(0.1520193)	(0.1895784)	(0.1493571)
Log No. Ratings	0.8914176***	1.012251***	0.8806771***	1.016091***
	(0.0596195)	(0.0230472)	(0.0592359)	(0.022563)
No. Images	1.89168***	1.027836***	1.812292***	0.972098***
	(0.2899638)	(0.218348)	(0.2892375)	(0.215449)
Text Wordcount	-0.0106986***	-0.0090093***	-0.0104761***	-0.0095942***
	(0.003462)	(0.0026773)	(0.0034422)	(0.0026643)
No. Helpful votes	0.070377***	0.0622395***	0.0614253***	0.0510315***
	(0.0139089)	(0.0137529)	(0.0129173)	(0.0126382)
Log No. Q&A	0.323801***		0.3384376***	
	(0.065155)		(0.065379)	
Free Shipping	1.351283***	1.272033***	1.164859***	1.000808***
	(0.3685504)	(0.3279603)	(0.3208658)	(0.2788331)
3 in 1			0.2526764**	0.2531281**
			(0.1343489)	(0.1228166)
4 in 1			0.2622631	0.2734527*
			(0.1639447)	(0.1481719)
Pad			-0.2470385**	-0.1739426**
			(0.089228)	(0.0806567)
Stand			-0.0335347	-0.0489183
			(0.0913263)	(0.0788582)
Observations	4462	5176	4462	5176

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 10b. Standard Deviation of Individual Random Coefficients: Chargers

VARIABLES	(2)	(3)	(4)	(5)
SD – Price	0.0372305***	0.0283336**	0.0360726***	0.0234566
	(0.0144985)	(0.014462)	(0.0139713)	(0.0151741)
SD – Counterfeit Prob	2.617766***	2.313474***	2.47865***	2.082199***
	(0.4687269)	(0.3878497)	(0.4461563)	(0.3722704)

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 11. User Optimal Choices and Utility under Counterfactual Experiments

		Likely Authentic	Likely Counterfeit	Outside Option	Total Consumer	Utility
Control	Original Data Set	457519	737057	505424	1700000	15.5373
Treatment I	Polarized (Increase by 50% if Counterfeiting Probability >= 0.5, decrease the rest by 50%)	567198	602388	530414	1700000	15.5857
Treatment II	100 Likely Counterfeit products enter the market	392254	802861	504885	170000	16.5302