Essays on merchant-generated contents and merchant-to-merchant spillovers in e-commerce platform

Zhang, Xueli

2021

Zhang, X. (2021). Essays on merchant-generated contents and merchant-to-merchant spillovers in e-commerce platform. Doctoral thesis, Nanyang Technological University, Singapore. <https://hdl.handle.net/10356/147299>

<https://hdl.handle.net/10356/147299>

<https://doi.org/10.32657/10356/147299>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 17:52:52 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**ESSAYS ON MERCHANT-GENERATED CONTENTS AND
MERCHANT-TO-MERCHANT SPILLOVERS IN E-COMMERCE
PLATFORM**

ZHANG XUELI

NANYANG BUSINESS SCHOOL

2021

**ESSAYS ON MERCHANT-GENERATED CONTENTS AND
MERCHANT-TO-MERCHANT SPILLOVERS IN E-COMMERCE
PLATFORM**

ZHANG XUELI

Nanyang Business School

A thesis submitted to the Nanyang Technological University in partial
fulfilment of the requirement for the degree of Doctor of Philosophy

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

15/02/2021

.....

Date

.....

Zhang Xueli

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate with amendments, changes and improvements as suggested by me as the Supervisor. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

15/02/2021

.....
Date

.....
Sadat Reza

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

15/02/2021

.....

Date

.....

Zhang Xueli

Acknowledgments

I wish to express my greatest gratitude to my supervisor, Prof Sadat; co-supervisor, Prof Yupeng; my mentors, Prof Hongyan and Prof Liu; my wife, Lu; and my family.

Table of Contents

Summary	1
Essay 1: Optimal Product Quality Display and Pricing Strategies for Online Retailers in the Presence of Product Return Options.....	5
Abstract	5
1. Introduction	6
2. Literature Review	9
3. Model Setup	12
4. Analysis of Store's Optimal Pricing and Quality Display Strategies.....	18
5. Conclusion.....	26
Appendices	28
A1. Derive the expected utility ex-ante of purchasing the product.	28
A2. Derive the net demand function.....	29
A3. Derive the relationship between the net demand, return rate, and return amount.	33
A4. Derive the optimal pricing and quality display strategies	36
A5. Derive the relationship between the optimal solutions and d, t respectively.	42
A6. Derive the relationship between return amount, return rate, and d, t respectively under the optimal pricing and product display strategies.....	43
A7. Derive the relationship between the optimal profit and d, t respectively	44
References	44
Essay 2: Seller Show Vs Buyer Show: Utilizing Online Images Posted by Customers and Stores to Predict Product Returns	47
Abstract	47
1. Introduction	48
2. Literature Review	50
2.1. Product Returns	50
2.2. User-Generated Contents.....	52
2.3. Image Mining	54

3. Data	55
3.1. Data Source.....	55
3.2. Variable Operationalization.....	57
3.3. Descriptive Statistics	58
4. Empirical Analysis	61
4.1. Preliminary Processing of the Image Data	61
4.2. Features Extraction and Distance Calculation.....	63
5. Predictive Model: Gradient Boosted Regression Trees (GBRT)	72
6. Conclusion.....	75
References	75
Essay 3: Partial Identification of Spillovers in E-Marketplaces Utilizing	
Consumer Clickstream Data	80
Abstract	80
1. Introduction	80
2. Literature Review	83
3. Conceptual Framework and Identification	88
3.1. Conceptual Framework.....	88
3.2. Identification of the Positive Spillover Effect	91
3.3. Partial Identification of the Negative Business-Stealing Effect	92
3.4. Non-Parametric Estimation of the Positive Spillover and the Negative Business-Stealing Effects	95
4. Data and Preliminary Analyses	97
4.1. Data Description and Processing	97
4.2. Preliminary Analysis	99
5. Empirical Analyses	102
5.1. Estimation of the Positive Spillover and the Negative Business- Stealing Effects.....	102
5.2. The Effect of Cashback Rates on the Two Effects.	106
6. Conclusion.....	111
Appendix	113
References	114

Summary

Unlike in the brick-and-mortar environment, online shoppers are not able to tangibly evaluate product qualities using multiple sensory organs. The only way for online shoppers to evaluate product qualities is to view the product information provided by online merchants. Therefore, the merchant-generated contents, such as product images, videos, textual descriptions of attributes, prices, etc., play an important role in forming shoppers' quality expectations and will directly influence shoppers' purchase decisions in the context of online shopping. Besides, various merchants cluster together in e-marketplaces providing complementary or substitutive products or services, so they will influence each other positively or negatively. This dissertation will focus on issues related to merchant-generated content and merchant-to-merchant spillovers in e-marketplaces.

In Essay 1, I develop an analytic model to examine the optimal product quality display and pricing strategies of an online retailer in the presence of return options. It seems that online merchants usually have incentives to exaggerate their product quality. Because by doing so, they can attract more consumers to purchase by increasing the perceived quality, and consumers cannot verify the authenticity of the merchant-posted information immediately in the context of online shopping. However, with the presence of return options, online merchants should reconsider it. Consumers might return the product if they find the real information is inconsistent with the information that they received when they made the purchase decision. If the displayed quality is too low, few consumers will buy it; if the displayed quality is too high, consumers will probably return it. Therefore, it is a trade-off between the increasing demands and increasing returns when determining the level of quality displayed. Besides, the pricing strategy will also affect consumer purchase and return decisions.

I consider a model in which an online retailer offers a product with both vertical and horizontal attributes, sets a price, and claims the quality. In the model, consumers make the purchase and return decisions sequentially. Consumers are heterogeneous due to their different trusts to the retailer and different preferences for the product. I derive the retailer's optimal strategies as a function of the real product quality, consumer's disutility factor for the product mismatch, and the disutility factor for the exaggeration, and I identify how the optimal strategies, profits, and product returns are affected by consumer's exaggeration disutility factors and mismatch disutility factors.

In Essay 2, I examine whether online user-generated contents have any predictive power on product returns. On one hand, merchant-generated contents play an important role in forming users' quality expectations. On the other hand, user-generated contents, such as text reviews, images, and videos, posted online after usage reflect the users' realized quality levels. Therefore, for product categories where both merchant-generated and user-generated contents are available, the differences between these two types of contents can reflect the differences between the expected and realized quality levels, which in turn is potentially a factor that affects consumers' product return decisions.

Product return has a significant impact on firms' profitability. Evidence shows that for the United States retailers in 2018, total customer returns account for more than \$369 billion, which is close to the market cap value of Facebook (Retail, 2018). Hence, it is important for firms to be able to accurately predict product returns and consider them for profit optimization. Since the reason for consumer return behavior is that real product quality is lower than merchant-displayed quality, the differences between merchant-generated and user-generated contents can be used to predict product returns. However, no matter merchant-generated or user-generated contents, they are mostly presented in unstructured formats, such as texts, images, voices, and videos, etc. How to extract useful attributes for analyses is a challenge for researchers.

I develop an AI system to extract features from both seller-posted and buyer-posted images and investigate the predictive power of the difference between these two types of images on product returns. Specifically, first, I use well-established image-processing techniques to extract basic image attributes, such as entropy, cool color ratio, lightness, and saturation, which are defined as “shallow features”. Second, I use a residual learning framework, ResNet-50, combining with a special neural network structure, Siamese neural network, to recognize images, and choose the second-to-last layer of the network, containing 512 features, as “deep features”. Then I calculate the Euclidean and cosine distance between seller-posted images and buyer-posted images using the extracted features, and I employ a state-of-art machine learning algorithm – gradient boosted regression tree –to predict product return rate.

In Essay 3, I examine merchant-to-merchant traffic spillover in e-marketplaces. In e-marketplaces, various online merchants are clustering together and providing similar or different product information. They separately make decisions that what kinds of product information should be displayed to online shoppers, but they will influence each other positively or negatively. On one hand, merchants could be beneficial to each other because of demand-side externalities, making them tend to cluster together. On the other hand, there could be business-stealing effects due to the presence of competitors and substitutive products, making them stay away from each other. These two opposite effects exist simultaneously among merchants in marketplaces. How to empirically identify them separately has not been solved in the previous literature.

I develop an empirical framework to separately identify these two opposite effects in e-marketplaces using consumer clickstream data. To do this, using consumer clickstream data, I separate the traffic into four types: initial traffic, indicating the consumers who initially visit the merchant when they land in the site; spillover traffic, indicating the consumers who move to the merchant but initially visit another merchant; diverted traffic, indicating the consumers who

initially visit the merchant but later move to other merchants; return traffic, indicating the consumers who initially visit the merchant, and then move to other merchants, but finally come back to the merchant. The additional sales brought by the spillover traffic measure the positive spillover effect, while the lost sales due to the diverted traffic measure the negative competitive effect. The two effects could be separately identified by seven parameters: initial visiting probabilities; three buying probabilities of the initial traffic, spillover traffic, and return traffic, respectively; three traffic transferring proportions of the diverted traffic in the initial traffic, the spillover traffic in others initial traffic, and the return traffic in the diverted traffic, respectively. Since the empirical context of this research is a cashback company based in Asia, I investigate the effect of cashback rates and their interaction effect with merchant categories and weekends on the seven latent parameters.

The next three chapters will discuss the literature review, theoretical framework, empirical studies, contributions, and managerial implications for the three works in detail, respectively.

Essay 1: Optimal Product Quality Display and Pricing Strategies for Online Retailers in the Presence of Product Return Options

Abstract

Unlike in the brick-and-mortar environment, online shoppers are not able to tangibly evaluate product qualities using multiple sensory organs. The only way for online shoppers to evaluate product quality is to view the product information provided by online merchants, which will directly influence consumers' purchase and return decisions. Therefore, product quality display is an important strategic decision for online retailers. This research develops an analytic model to examine the optimal product quality display and pricing strategies of an online retailer in the presence of return options. Specifically, an online retailer sets a price and claims product quality to consumers, offering a product with exogenously given quality. Consumers make the purchase and return decisions sequentially, and they are heterogeneous due to their different trusts in the retailer and different preferences for the product. I derive the retailer's optimal strategies as a function of the real product quality, the consumer's disutility factor for the product mismatch, and the disutility factor for the retailer's exaggeration. The main results show that (1) the exaggeration disutility factor negatively affects both the optimal pricing and product display strategies, but the mismatch disutility factor does not influence them under certain conditions. (2) Both the return amount and return rate are positively related to the mismatch disutility factor but negatively related to the exaggeration disutility factor. (3) The optimal profit decreases as the mismatch disutility factor increases; and interestingly, the optimal profit first decreases and then increases as the exaggeration disutility factor increases.

Keywords: Quality display; pricing; product return; analytic model.

1. Introduction

Product quality display is a very important strategic decision for online retailers. Unlike shopping in physical stores that consumers can identify product qualities by touching, smelling, tasting, or even trying, the only way for online shoppers to evaluate the product qualities is to view the product information provided by online retailers. The information provided by retailers will directly influence consumers' purchase decisions in the context of online shopping. For example, consider a consumer who intends to buy an apparel product from a particular online store. Before she makes her purchase decision, she will browse through the information posted by the online store to reduce (but unlikely resolve) her uncertainty on product quality (e.g., the material of the product) and product fit. This inference through the posted information then determines her willingness to pay for the product, which, if higher than the product price, will lead to a purchase. Therefore, it is an essential question for online retailers to figure out how to strategically display product quality.

Pricing strategy is another important decision for online stores. In particular, how will it, in combination with the decision on the product quality display, affect consumers' purchase decisions as well as return decisions? For example, if an online retailer upward-twists the product quality, which price level should it set to maximize its profits? Intuitively, the online store should set a higher price for the high displayed quality to get a higher marginal profit. However, if it sets a higher price, there will be less immediate demand. Furthermore, because of the upward-twisted display quality, it is more likely consumers will return the purchased product after they receive the product, which will further drive down the net sales.

As main shopping platforms are gradually transferring from traditional physical stores to online markets, it is important to re-examine stores' optimal product quality display and pricing strategies in the context of online shopping. Accordingly, this paper intends to develop an analytical model to answer two

fundamental questions: (1) what is the online store's optimal product display strategy, and (2) how should the online store choose its optimal price with the presence of return options?

Intuitively, online stores should have strong incentives to display upward-twisted product quality to attract consumers. By displaying up-twisted product quality, online stores can increase consumers' inferred product quality and attract more consumers to purchase their products. However, with the presence of return options, online retailers should reconsider it. Consider again the example of purchasing an apparel product online. Suppose the consumer has purchased the product. After she receives the product, she can figure out the true quality and product fit, which, if contradicts too much from her previous inference, may lead to product return. Hence, online retailers need to reconsider whether it is still beneficial for them to upward-twist product quality since consumers can potentially return the product, very often at the full refund, if they find the real product quality is (much) lower than the displayed quality or/and the product fit is sufficiently low. It is a trade-off between increasing immediate demand and reducing potential returns when determining the level of quality displayed.

Following the framework proposed by Wood (2001), the online shopping process is separated into two stages: (1) consumers make purchase decisions whether to buy or not to buy the products; (2) consumers make return decisions whether to return or keep the items. In the first stage, consumers will browse the displayed product quality information and make purchase decisions. In the second stage, after consumers get the delivery and identify the real product quality and product fit, they will make return decisions.

There are two sources of heterogeneity in consumers. First, consumers are heterogeneous in using the displayed quality to infer product quality: some consumers think the store will display the real quality, while others think the store will exaggerate. Therefore, their formed expected product quality will be

different even when they are facing the same displayed quality information. Second, given the same product, consumers are heterogeneous in their individual tastes, hence, different in terms of product fit. Combining the expected quality, product fit, and price, consumers will make purchase decisions. Furthermore, consumers are rational, so they will consider the probability of returning the product when they make their purchase decisions. Consumers will “punish” the online store’s dishonest quality display as well as incur disutility associated with product unfit. If the real product quality is much lower than the expected quality, or if the product does not fit them well enough, they will return the item.

The analyses show several interesting results. First, consumer’s disutility factor for retailer’s exaggeration has a negative effect on both the optimal pricing and product display strategies, and the disutility factor for product mismatch does not influence the optimal pricing and product display strategies under the assumption that the return fee is zero and when the product is very unfit consumers will probably return it. Second, both the return amount and return rate are positively related to the mismatch disutility factor but negatively related to the exaggeration disutility factor. Third, the online store’s optimal profit will first decrease and then increase as the disutility factor for the exaggeration increases. The increasing part is quite counterintuitive. One explanation might be that firms can please consumers by displaying the quality lower than the real quality.

This research provides rich managerial implications. First, I propose that if online stores offer a free return policy, they only need to consider consumers’ exaggeration disutility factor but not the product mismatch factor. To maximize the profit, when consumers have a large exaggeration disutility factor, the optimal price and displayed quality should be small; otherwise, stores could increase their optimal price and the displayed quality. Second, product mismatch disutility factor and exaggeration disutility factor have opposite

effects on product return. To minimize product return, online retailers should target consumers with high exaggeration disutility factor and low mismatch disutility factor. Third, a high disutility factor for the store exaggeration is not always a bad thing for online stores. When it is large, online stores can utilize the large effect by largely decreasing the displayed quality to gain more profit.

The rest of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 sets up the model. Section 4 analyzes the online store's optimal decision on quality display and pricing. Section 5 concludes the research.

2. Literature Review

This research draws on and contributes to two streams of marketing literature: exaggerated advertising and product return.

Different from the classic signaling literature (George, 1970; Kirmani & Rao, 2000; Spence, 1978) that advertising serves as credible information for customers because of the high costs and only firms with high-quality products can afford it, the exaggerated advertising literature focuses on the content of the advertising and investigates whether it benefits the company if it overstates the function of the product.

Some researchers argue that exaggerated advertising is useful to companies only when it could successfully dupe buyers (Hoffmann et al. 2013; Preston, 1996). Consumers sometimes are credulous and pay limited attention to the content of advertisings, so they might choose to believe what the advertiser claims. Then the company could extract more consumers and gain more demands by exaggerating its product quality. Crawford and Sobel (1982) also indicate that communication could be credible if and only if the sender's incentive to mispresent is small. When the exaggeration in advertising is small, it will be hard for consumers to detect the biased message, which also supports

the argument that puffery is useful to the company only when consumers trust it. However, Calfee and Ringold (1994) show that 70% of consumers think advertising is often unrealistic, and it is just a means for companies to persuade them to buy what they do not want. Now that puffery cannot dupe most of the consumers, then why does it extensively exist in real-world business?

Even when consumers think advertising is often untruthful, it could still be beneficial if consumers cannot verify the authenticity of the content. Chakraborty and Harbaugh (2014) claim that although consumers know that most advertisements are mere puffery, they will know the relative advantages of the product when the company talks up one strength but does not talk up another potential strength, so the puffery could still be informative. Malär et al. (2011) propose that consumers might choose to believe it even when they notice the message is biased, as long as the message appeals to their “ideal self”. These studies are talking about how exaggerated advertising could be credible when it is unverifiable. However, consumers can verify the product quality in most cases, for example, they can directly go to the physical store to test the product.

Companies will still have opportunities to benefit from exaggerating their products when advertising is verifiable. Nelson (1974) states that companies can exaggerate the utility of the actual good by any amount up to the search cost to gain more sales. Marks and Kamins (1988) investigate the effects of the sequence of exposure to product sampling and advertising on consumers’ attitudes. They find that when consumers are exposed to the advertising in advance of product sample distribution, they will have higher purchase intention than the other way around. The reason is that consumers appear to accept the claim initially and then not accept or discount the claim after realizing its lack of credibility. Even when consumers can verify the product quality, puffery still could be beneficial if it happens before the verification.

The analytical research on whether firms should exaggerate when advertising is verifiable is sparse. Gardete (2013) considers a cheap-talk model that allows consumers to verify the authenticity of advertising content by paying a search cost. He finds that advertising can be both informative to consumers and profitable for firms even if the incentives to exaggerate exists. They claim that a firm with a low-quality product may not want to exaggerate its claims considerably because doing so will scare away potential consumers who can only afford low-quality products, while a firm with a high-quality product may not want to understate its true quality level because that statement would attract consumers unwilling to pay enough for quality. Unlike Gardete (2013) that consumers pay search costs to verify the content of advertising, the current research investigates companies' optimal quality display strategy in the presence of return options. When consumers find the real quality is lower than the store-displayed quality, they have options to return the product. Companies need to consider the trade-offs that exaggeration could increase demands but also increase returns when they make quality display decisions. Besides, I do not restrict the analysis to search good, and I consider both vertical and horizontal product characteristics.

This research also contributes to the product return literature (Shulman et al., 2009, 2010; Shulman et al. 2011). Shulman et al. (2009) develop an analytical model to investigate a seller's optimal pricing, restocking fee policy, and information provision strategies. Consumers are uncertain about both vertical and horizontal product quality. They find that it is not always a good strategy to provide product fit information to consumers. Shulman et al. (2010) extend the model to a bilateral monopoly to examine the impact of reverse channel structure on the optimal pricing and return policy. They find a counterintuitive result that the return issue will be more severe when the product is returned to the channel member who has a greater salvage value of the returned product. Shulman et al. (2011) further extend the model to a condition that two competing firms sell horizontally differentiated products.

They find that the optimal restocking fee in a competitive condition is higher than that in a monopoly condition. Comparing to Shulman, Coughlan, et al., the current analysis will focus on online shopping and split the process into two stages: consumer's purchase decision to buy or not to buy; return decision to keep or to return. Besides, both academic (Bower & Maxham III, 2012; Janakiraman et al., 2016) and industrial evidence shows that taking a free return policy is beneficial to online retailers because a convenient product return process is a signal of good quality product and will increase customer satisfaction. Therefore, I treat the restocking fee as a constant variable. Instead, the quality display strategy is an important decision variable for companies in this research.

3. Model Setup

Consider an online store that offers a product with quality q_r , located at $x = 0$ on a Hotelling unit line. The marginal cost of producing a product is c . The store has two decision variables: the product quality displayed to customers q_s , and the product price p .

Consumers are heterogeneous in two dimensions. On the one hand, consumers are heterogeneous with respect to their trust in the displayed product image, hence their inference of product quality is different. Their expected product quality after they observe the displayed quality is $\alpha_i q_s$, where α_i is the discount factor and uniformly distributed on $[0,1]$. On the other hand, consumers are heterogeneous with respect to their individual tastes x_i , which is uniformly distributed on $[0,1]$. Before purchase, consumer i does not know the distance between her taste x_i and the location of the product, or $|0 - x_i|$. But they know the tastes of all consumers follow a uniform distribution on $[0,1]$. Product fit and product quality are defined as two orthogonal attributes of products in our model. For example, for a T-shirt, the product quality q_r could be the quality of the fabric, sewing, color, or design details; while for the

product fit $|0 - x_i|$, it could be the size of the T-shirt. Consumers might return a product either due to low quality or because of unfitness. Thus, these two attributes will vertically influence consumers' utilities.

Besides, previous literature (Wolinsky, 1983) shows that product prices could be used as a cue to quality in terms of consumer's perspectives. Typically, consumers would expect a higher (lower) quality from a higher (lower) price. Therefore, a consumer's expected product quality q_e inferred from the displayed quality q_s will also be impacted by product price p . I capture the price-quality relationship by setting α_i as a function of price p , where $\frac{\partial \alpha_i(p, q_s)}{\partial p} > 0$. The higher the price is, the larger $\alpha_i(p, q_s)$ is, resulting in higher expected product quality.

Adapted from the model in Shulman et al. (2009), consumer i 's pre-purchasing utility function is defined as:

$$U_i = \alpha_i(p, q_s) - p - t|0 - x_i| \quad (1)$$

, where $\alpha_i(p, q_s)$ is her expected product quality given the store-displayed quality and price; $t > 0$ measures the disutility per unit deviation from consumer i 's taste. Consumers are rational, which means when they make purchase decisions, they will anticipate whether to return or keep the product if they buy the product and try it. For consumer i , if her utility of keeping the product is larger than her utility of returning it, she will keep it; otherwise, she will return it. If a consumer returns the product, she will lose a return fee f and gets the price p back.¹

¹ Alternatively, the following utility function could be considered $u = f(|x - x_0|, p)$, $x_0 = g(q)$, where x is the consumer's taste, x_0 is the location of the product revealed by the marketer, q is the information presented by the marketer, which can be exaggerated, accurate or understated. However, I opted for the one in (1) since the former is unable to differentiate the product quality and product fit and assumes they are correlated. In real life, product quality and product fit are two vertical attributes. In the T-shirt example, a T-shirt could have high quality but in the wrong size. In this case, the model proposed in the current study is more appropriate.

In the current study, I assume that consumers are rational, as other game-theoretical research related to this topic (e.g. Shulman et al., 2009, 2010, 2011). Consumer's expected utility of purchasing the product is strictly negatively related to the price. For example, when consumers face two products with the same displayed quality information but different prices, they will rationally choose the product with a lower price. Therefore, under the assumption of rationality, consumers' perceived product quality will not be influenced by price p , and thus, I set $\alpha_i(p, q_s) = \alpha_i q_s$.

Consumer i 's expected utility ex-ante of purchasing the product is (see the appendix for details)²:

$$E_i(U_{purchase}) = \frac{(\alpha_i q_s - p + f)^2}{2t} - f \quad (2)$$

Customer i will buy the product if and only if her expected utility is greater than zero, i.e., $E_i(U_{purchase}) > 0$, which gives:

$$\alpha_i > \frac{\sqrt{2ft} + p - f}{q_s} \equiv \hat{\alpha} \quad (3)$$

Hence, the online store's initial product demand is:

$$q_b(q_s, p) = 1 - \hat{\alpha} = 1 - \frac{\sqrt{2ft} + p - f}{q_s} \quad (4)$$

Intuitively, the initial demand function $q_b(q_s, p)$ satisfies the following: the higher the displayed quality (q_s), or the lower the price (p), or the smaller the unfit disutility factor (t) is, or the smaller the return fee (f) is³, the higher the initial demand will be.

Next, I discuss consumers' return decisions if they buy the product. After getting the product, consumer i will observe the real product quality q_r and

² It is an interior solution, with condition $t > q_s + f - p$.

³ Combining the initial demand greater than zero and t 's condition gives that the return fee is negatively related to the initial demand.

figure out the product fit $|0 - x_i|$. Consistent with the observation of a relevant market phenomenon, there usually is a punishment or reward for the store's exaggeration or modesty, respectively, from the consumers' side. Besides, many empirical studies consider the impact of endowment effects when they investigate product returns. For example, Wood (2001) finds that the endowment effect suggests some surprising benefits of return policy leniency to the retailer. Consumers feel an increase in value for products they own, which will decrease the probability of returning the product. Therefore, considering this possible punishment or reward and the endowment effect, consumer i 's utility function after she gets the product could be written as:

$$U_i = q_r - d(\alpha_i q_s - q_r) - tx_i - p + e \quad (5)$$

, where $d > 0$ measures consumer i 's disutility for the online store's exaggeration when she realizes the difference between the real product quality and the expected quality; e indicates the increased value because of the endowment effects.

For the return decision given purchase, if the utility of keeping is larger than the utility of returning, i.e., $U_i > -f$, she will keep the product, which gives,

$$\alpha_i < -\frac{t}{dq_s} x_i + \frac{(1+d)q_r - p + f + e}{dq_s} \quad (6)$$

Combining (3) and (6) then gives the net demand function. For simplicity, I now assume the return fee as zero, i.e., $f = 0$, and discuss how $f > 0$ potentially affects our results later. Note that this assumption of free return is reasonable given the prevalent practice in real-world business. Also, because e is exogenously given in the model as a constant term, it will only influence the results numerically but not analytically, which means it will not change the relationship between the optimal solutions (q_s^*, p^*) and other analytical factors (e.g. t, d). Mathematically, it will not change the sign of $\partial p^*/\partial t$, $\partial p^*/\partial d$, $\partial q_s^*/\partial t$ and $\partial q_s^*/\partial d$. Therefore, involving e in the model will not impact the

interpretation of the results in the current study, and thus, for simplicity and without the loss of generality, I set $e = 0$.

The net demand function is derived as (see the appendix for details):

$$D(q_s, p) = \begin{cases} 0, & \text{if } p > q_r, q_s > p \\ \frac{[(1+d)(q_r - p)]^2}{2dtq_s}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r, \frac{(1+d)q_r - p}{d} < q_s \\ \frac{(1+d)(q_r - p) - 0.5t}{dq_s}, & \text{if } p < q_r - \frac{t}{(1+d)}, \frac{(1+d)q_r - p}{d} < q_s \\ \frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)}, p < q_s < \frac{(1+d)q_r - p}{d} \\ \frac{-2dpt + 2dq_s t - (p + d(q_s - q_r) - q_r + t)^2}{2dq_s t}, & \\ \text{if } p < q_r - \frac{t}{(1+d)}, \frac{(1+d)q_r - t - p}{d} < q_s < \frac{(1+d)q_r - p}{d} \\ \frac{q_s + p}{q_s}, & \text{if } p < q_s < \frac{(1+d)q_r - t - p}{d} \end{cases} \quad (7)$$

Figure1 demonstrates the different parameter regions in the above demand function, where the mark “○” indicates a zero demand for the online store, and numbers 2-6 sequentially correspond to the parameter regions in (7) one by one. Here, I only consider the regions with positive net demand. Furthermore, if a product is extremely unfit (i.e., $x = 1$), consumers will most likely return the product. Accordingly, I will not consider case 3 and case 5. Hence, the following analysis mainly focuses on Case 2 and Case 4.

Figure 1. The net demand functions in each region

Then the net demand function is reduced to:

$$D(q_s, p) = \begin{cases} \frac{[(1+d)(q_r - p)]^2}{2dtq_s}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ \frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases}$$

(8)

Equation (8) shows the relationships between the net demand and p, q_s .

The net demand will first increase and then decrease as q_s increases, and it will decrease as p increases (see the appendix for details). Intuitively, when q_s increases, the initial demand will diminishingly increase, and the return amount will also increase. In the beginning, the growth rate of the initial demand is larger than that of the return amount, so the net demand will increase first. However, when q_s continuously increases, the growth rate of the initial demand becomes smaller than that of the return amount, so the net demand will decrease then. As p increases, consumers are less likely to buy the product and more likely to return it so the net demand will decrease.

The return rate function is $r(q_s, p) = 1 - D(q_s, p)/(1 - p/q_s)$, or

$$r(q_s, p) = \begin{cases} 1 - \frac{(1+d)^2(p-q_r)^2}{2d(q_s-p)t}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ 1 - \frac{2(1+d)q_r - (2+d)p - dq_s}{2t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases}$$

(9)

Intuitively, the return rate will increase as the price and the displayed quality increase (see the appendix for details). When the price or the displayed quality increase, consumers are more likely to return the product due to the more gains from returning or the larger disutility from the exaggeration.

The return amount function is defined as:

$$R(q_s, p) = \begin{cases} (1 - \frac{p}{q_s})(1 - \frac{(1+d)^2(p-q_r)^2}{2d(q_s-p)t}), & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ (1 - \frac{p}{q_s})(1 - \frac{2(1+d)q_r - (2+d)p - dq_s}{2t}), & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases}$$

(10)

The return amount will increase as q_s increases and will first increase and then decrease as p increases (see the appendix for details). As q_s increases, the initial demand and the return rate will increase, so the return amount will increase. As p increases, the initial demand will decrease, but the return rate will increase. When p is low, the initial demand is high, so the return amount (which equals the product of the initial demand and the return rate) will increase; while when p is high, the initial demand is low, so the return amount will decrease even the return rate is increasing.

4. Analysis of Store's Optimal Pricing and Quality Display Strategies

The online store's profit function is $\pi(q_s, p) = (p - c) \left(1 - \frac{p}{q_s}\right) + (s - p)R(q_s, p)$. Assuming the salvage value equals the cost, i.e., $s = c$, then the profit function is simplified to:

$$\pi(q_s, p) = (p - c)D(q_s, p) \quad (11)$$

The cost c comes with the demand function D in a product term ($-cD$), and thus it will influence the results of the optimization problem. However, similarly as the endowment effect parameter e , it will only influence the results numerically but not analytically, which means it will make the value of optimal solutions (q_s^*, p^*) larger or smaller, but it will not change the relationship between the (q_s^*, p^*) and other analytical factors (e.g. t, d). This is because the cost c is related to the real quality q_r , which is exogenously given in the model. Mathematically, for example, p^* is a function of (q_r, t, d, c) and could be written as $p^*(q_r, t, d, c)$. Because c and q_r are given exogenously, they will not change the sign of $\partial p^*/\partial t$ and $\partial p^*/\partial d$. In other words, cost c will not influence the results analytically and will not impact the interpretation of the results. Therefore, without the loss of generality and for simplicity, I assume $c = 0$.⁴ Then Equation (11) can be further simplified to:

$$\pi(q_s, p) = \begin{cases} p \frac{[(1+d)(q_r - p)]^2}{2dtq_s}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ p \frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases} \quad (12)$$

Maximizing the online store's profit in (12) jointly over q_s and p leads to the following optimal solution (see the appendix for details):

$$p^* = (2q_r + 2dq_r - q_r\sqrt{2d(1+d)})/(4+2d),$$

$$q_s^* = q_r\sqrt{(1+d)/2d}. \quad (13)$$

Correspondingly, the online store's optimal profit is:

⁴ Since c is related to the real quality q_r , which is exogenously given in the model, it is independent to the decision variables. Then it will not influence the results mathematically if setting $c = 0$.

$$\pi^* = q_r^2(1+d)^2(2+3d-2\sqrt{2}\sqrt{d(1+d)})/2(2+d)^2t \quad (14)$$

Recall that I have only considered Case 2 and Case 4, which means that I implicitly assume that the online store is better off with strategies in Case 2 and/or Case 4 than other cases. This requires $t < q_r(1+d)(1 + \sqrt{d(1+d)/2})/(2+d)$. Alternatively, when $t < q_r(1+d)(1 + \sqrt{d(1+d)/2})/(2+d)$, or consumer's disutility factor for product mismatch is low enough, the store will adjust its strategy to let those consumers who have the highest unfit to the product ($x = 1$) still keep it, which is Case 5.

I summarize the assumptions under which the optimal solutions are generated.

A1: The store offers a policy of free return.

A2: When the product is very unfit, consumers will definitely return it no matter how good the quality is.

A3: Consumers are rational, so psychological factors such as the price-signaling effect and the endowment effect will not influence consumers' decisions.

Equation (13) gives a direct answer to the questions I raised at the beginning: (1) what is the online store's optimal product display strategy, and (2) how should the online store choose its optimal price with the presence of return options? First, the optimal quality display strategy is negatively related to consumer's exaggeration disutility factor d , or $q_s^* = q_r\sqrt{(1+d)/2d}$. When $d < 1$, or consumers are less sensitive to store's exaggeration, the optimal strategy is to display products with higher quality than the real quality; while when $d > 1$, or consumers are more uncomfortable on store exaggeration, online stores should display a lower quality than the real quality to maximize their profits. Intuitively, if consumers do not feel uncomfortable when they find that the real quality is lower than the displayed quality, then stores can display a

higher quality than then real quality to attract more consumers; while if consumers' benefits will largely increase when they find that the real quality is higher than the displayed quality, stores can even display a lower quality to cater consumers. Second, the optimal price p^* depends on the value of the exaggeration disutility factor d but not the product mismatch disutility factor t , or $p^* = (2q_r + 2dq_r - q_r\sqrt{2d(1+d)})/(4+2d)$. p^* is negatively related to the exaggeration disutility factor d , because when consumers feel more uncomfortable with store exaggeration, they are more likely to return the product. Thus, the store has incentives to decrease prices to reduce the return probability. I will clarify it more in the following propositions.

The following proposition, with its proof in the appendix, summarizes how the optimal pricing and quality display strategies vary as these two consumer's disutility factors increase.

Proposition 1. Under the assumptions A1 to A3, the optimal price and quality displayed strategies will not be impacted by the mismatch disutility factors but will be negatively related to the exaggeration disutility factors. When the exaggeration disutility factor is large, the optimal displayed quality should be lower than the real quality; otherwise, it should be higher. Mathematically,

$$(1a) \frac{\partial p^*(q_r, d, t)}{\partial t} = 0, \text{ and } \frac{\partial p^*(q_r, d, t)}{\partial d} < 0; (1b) \frac{\partial q_s^*(q_r, d, t)}{\partial t} = 0, \text{ and } \frac{\partial q_s^*(q_r, d, t)}{\partial d} < 0;$$

$$(1c) \text{ when } d \leq 1, q_s^* \geq q_r; \text{ otherwise, } q_s^* < q_r.$$

Proposition 1 shows that, first, counterintuitively, consumer's disutility factor for product mismatch t does not influence the optimal pricing and product display strategies. This is because of the assumption that the return fee f is zero and the assumption that when the product is very unfit consumers will definitely return it. When the return fee is zero, the mismatch disutility factor will not influence the initial demand because consumers can return the product without any cost. Furthermore, when the disutility factor is large enough, consumers with high-level mismatch will definitely return the product, which means that the online store can do nothing (decreasing p , or q_s) to prevent them

from returning it. Therefore, even when t continuously increases, it will not influence the store's optimal pricing and quality display strategies.

Second, consumers' disutility factor for punishing online store's exaggeration d has a negative effect on both the optimal pricing and product display strategies. When the disutility factor increases, consumers are more likely to return the product. Thus, the store has incentives to decrease both the displayed quality and price to reduce the return rate.

Third, when the disutility factor for punishing the exaggeration is low, the store will post the quality level higher than the real product quality to attract more consumers but will not suffer much punishment from the bluffing; however, when the disutility factor is high, the store will not display high quality, and to the opposite, they will utilize the high disutility factor by posting quality lower than the real quality to please consumers.

The managerial implication for Proposition 1 is that if online stores offer a free return policy, they only need to focus on consumers' exaggeration disutility factor but not the product mismatch factor. This is because consumers can return the product without any cost, and thus, the mismatch disutility factor will not influence stores' optimal profits. To maximize the profit, when consumers have a large exaggeration disutility factor, the optimal price and displayed quality should be small; when consumers have a small disutility factor, the optimal price and displayed quality should be large.

Since product return is an important issue for online retailers, next, I discuss the relationship between return amount R^* , return rate r^* and d, t respectively under the optimal pricing and product display strategies. The following proposition, with its proof in the appendix, shows the findings.

Proposition 2. The optimal return amount and return rate are positively related to the mismatch disutility factor but negatively related to the exaggeration

disutility factor. Mathematically, (2a) $\frac{\partial R^*(q_r, d, t)}{\partial t} > 0$, and $\frac{\partial R^*(q_r, d, t)}{\partial d} < 0$;

(2b) $\frac{\partial r^*(q_r, d, t)}{\partial t} > 0$, and $\frac{\partial r^*(q_r, d, t)}{\partial d} < 0$.

Proposition 2 shows that, first, both the return amount and return rate are positively related to the disutility factor for the product mismatch. As the mismatch disutility factor increases, consumers will be more uncomfortable about the mismatch and are more likely to return the product, and the factor will not influence the initial demand, so the return amount will increase.

Second, both the return amount and return rate are negatively related to the disutility factor for the exaggeration. When the exaggeration disutility factor increases, consumers are also more likely to return the product, but the store can adjust the displayed quality closer to the real product quality or even lower than the real quality to reduce consumers' discomforts (when displayed quality is larger than the real quality) or increase consumers' pleasant surprises (when displayed quality is lower than the real quality). Therefore, the return rate and return amount will decrease instead.

The managerial implication for Proposition 2 is that product mismatch disutility factor t and exaggeration disutility factor d have opposite effects on product return. To minimize product return, online retailers should target consumers with high exaggeration disutility factor and low mismatch disutility factor.

Proposition 3 shows how the online store's optimal profit varies as the two consumer's disutility factors increase.

Proposition 3. The optimal profit is negatively related to the mismatch disutility factor. When the exaggeration disutility factor is small, it will negatively impact the optimal profit; while when it is large, it will positively impact the optimal profit. Mathematically, (3a) $\frac{\partial \pi^(q_r, d, t)}{\partial t} < 0$; (3b) when $0 < d < -2 + \sqrt{5}$, $\frac{\partial \pi^*(q_r, d, t)}{\partial d} < 0$; when $d > -2 + \sqrt{5}$, $\frac{\partial \pi^*(q_r, d, t)}{\partial d} > 0$.*

Proposition 3 shows that first, the disutility factor for the product mismatch t has a negative effect on the online store's optimal profit. As the mismatch disutility factor increases, although the store will not adjust its optimal price and displayed quality (from Proposition 1), which means that the initial demand will not change, the return amount will increase (from Proposition 2). Therefore, the net demand will decrease, and in turn, the profit will decrease given the price staying the same.

Second, the optimal profit will first decrease and then increase as the disutility factor for the exaggeration increases. When the disutility factor for the exaggeration is low, the online store will optimally display a high quality to attract more consumers. As it increases, the online store will decrease the displayed quality and price to decrease consumers' discomforts from the exaggeration (from proposition 2), so the optimal profit will decrease. However, as it continues to increase, the store will utilize the increasing effect of the disutility factor to either decrease the discomforts or increase the pleasant surprises by largely decreasing the displayed quality, so the optimal profit will increase.

The managerial implication here is that a high disutility factor for the store exaggeration d is not always a bad thing for online stores. When it is large, online stores can utilize the positive effect by largely decreasing the displayed quality to gain more profit.

Furthermore, when the online store sets the price and determines the displayed quality, an interesting question is what the relationship between the price and displayed quality should be to optimize the profit. For example, when the displayed quality is high, should the price also be high, or should it be low? During solving the profit optimization problem, I notice that,

$$q_s = \sqrt{p(-dp - 2p + 2dq_r + 2q_r)/d}. \quad (15)$$

Because $t > q_r(1 + d) \left(1 + \sqrt{\frac{d(1+d)}{2}}\right) / (2 + d) > \frac{1+d}{2+d} q_r$, $p = \frac{1+d}{2+d} q_r$ is within the boundary condition. Then q_s is monotone increasing on $[q_r - \frac{t}{1+d}, \frac{1+d}{2+d} q_r]$ and decreasing on $[\frac{1+d}{2+d} q_r, q_r]$ w.r.t p , which is described in the following proposition.

Proposition 4. When the price is low, the optimal price should be positively related to the optimal displayed quality; while when the price is high, the optimal price should be negatively related to the optimal displayed quality.

Mathematically, $\frac{\partial q_s}{\partial p} > 0$, when $p \in [q_r - \frac{t}{1+d}, \frac{1+d}{2+d} q_r]$; $\frac{\partial q_s}{\partial p} < 0$, when $p \in [\frac{1+d}{2+d} q_r, q_r]$.

Proposition 4 shows that to optimize the profit, the displayed quality should first increase and then decrease as the price increases. Suppose the price and displayed quality start from small values. When the store increases its price, consumers will accept it since the price is small, or the negative effect of the price on the net demand is small; meanwhile, the store will also increase the displayed quality to increase the net demand. However, as the price continues to increase, consumers will not accept it, and the net demand will largely decrease. What is worse, the positive effect of the displayed quality on the net demand will vanish or even turn to a negative effect when it is large. Therefore, the store will decrease the displayed quality to increase the net demand.

The managerial implication here is that if online stores set a high price, they should also display high quality. However, when the price is very high, the optimal displayed quality should decrease as the price continuously increases.

If online stores that always exaggerate exist in the market, which means that condition $q_s > q_r$ always holds, then consumers will always take a discount on the displayed product quality $\alpha_i q_s$. The specific value of the discount factor α_i depends on individuals' history purchase experience. Then it

turns to be a dynamic game between stores and consumers, which I will discuss in our future study.

On one hand, merchant-generated content plays an important role in consumer's purchase and return decisions; on the other hand, another type of content, user-generated content, such as reviews, may also influence consumer's purchase and return behavior. When consumers view two different information (store-displayed vs. consumer-displayed information) from the same store, they will feel uncertain and lack trust, leading to low sales and high returns. I will investigate it empirically in Essay 2.

5. Conclusion

This research examines an online retailer's optimal pricing and quality display strategies in the presence of product return options and identifies how the optimal strategies, profits, and product returns are affected by consumer's exaggeration disutility factors and mismatch disutility factors. I develop several interesting results with important managerial implications for marketing practice.

First, the net demand will increase and then decrease as the displayed quality increases, and it will decrease as the price increases. For most online retailers, the net demand or sale is an important key performance indicator. When they increase the displayed quality, the sale will first increase and then decrease because of the return option. The optimal displayed quality should be slightly larger than the real quality, and the specific magnitude depends on consumers' exaggeration disutility factors. Online retailers could survey to get the factor or infer it from consumers' purchase history.

Second, to maximize the profit, the optimal price and displayed quality are negatively related to the consumer's exaggeration disutility factor and have nothing to do with the mismatch disutility factor under the conditions that the

return fee is zero and when the product is very unfit consumers will definitely return it. In real-world business, these conditions are satisfied in most cases, so online retailers could decide the optimal pricing and product display strategies depending on the value of the exaggeration disutility factor. When the factor is large, the optimal price and displayed quality should be small; otherwise, they should be large.

Third, the optimal profit is negatively related to the mismatch disutility factor, and first negatively but then positively related to the exaggeration disutility factor. A high value of the exaggeration disutility factor is not always a bad thing for online retailers, in contrast, they can utilize it to increase the profit by decreasing the displayed quality. The return amount is positively related to the mismatch disutility factor but negatively related to the exaggeration disutility factor. In my model, the salvage value equals the cost, but in real life, it could be smaller than the cost, then the product return is a more important issue for retailers. My results suggest that online retailers should target consumers with high exaggeration disutility factor and low mismatch disutility factor.

I acknowledge some limitations of this paper. I investigate the online retailer's optimal pricing and quality display strategies under a monopoly condition. Future research can study competition between two retailers selling two products with different qualities. Furthermore, it is interesting to study the case in which consumer-generated content will influence retailers' pricing and quality display strategies and the dynamic game between consumers and retailers. I hope this research can inspire more interest in this area.

Appendices

A1. Derive the expected utility ex-ante of purchasing the product.

The condition where customer i is indifferent between keeping the product and returning it is given by,

$$U_i = \alpha_i q_s - p - tx_i = -f \quad (A1.1)$$

If $x_i \in [0, x_1]$, where $x_1 = (\alpha_i q_s - p + f) / t$, customer i will keep the product; if $x_i \in [x_1, 1]$, customer i will return it. Note that customer i is uninformed about the value of $|0 - x_i|$ before purchase and only knows that it is uniformly distributed on $[0, 1]$. Therefore, the probability that consumer i will keep the product is:

$$\phi_{ik}(q_s, p; \alpha_i, t) = \frac{\alpha_i q_s - p + f}{t} \quad (A1.2)$$

, and the average utility derived from keeping the product will be equal to:

$$U_{ik} = \alpha_i q_s - p - t \frac{x_1}{2} = \frac{\alpha_i q_s - p + f}{2}. \quad (A1.3)$$

The probability that consumer i will return the product is:

$$\phi_{ir}(q_s, p; \alpha_i, t) = 1 - \frac{\alpha_i q_s - p + f}{t} \quad (A1.4)$$

, and the average utility derived from returning the product will be equal to:

$$U_{ir} = -f. \quad (A1.5)$$

I may now write the expected utility ex-ante of purchasing the product for consumer i as:

$$\begin{aligned} E_i(U_{purchase}) &= E_i(U_{purchase} | keep) * P(keep) + E_i(U_{purchase} | return) * P(return) \\ &= U_{ik} \phi_{ik}(q_s, p; \alpha_i, t) + U_{ir} \phi_{ir}(q_s, p; \alpha_i, t) \end{aligned}$$

When the two probabilities (keep and return) are non-negative, i.e.

$$0 < \frac{\alpha_i q_s - p + f}{t} < 1 ,$$

$$E_i(U_{purchase}) = \frac{(\alpha_i q_s - p + f)^2}{2t} - f \quad (A1.6)$$

When $\frac{\alpha_i q_s - p + f}{t} > 1$, the returning probability is zero, and the expected

utility of purchasing is: $E_i(U_{purchase}) = \frac{(\alpha_i q_s - p)^2 - f^2}{2t} .$

When $\frac{\alpha_i q_s - p + f}{t} < 0$, the keeping probability is zero, and consumers will never buy the product.

Since I am interested in the condition that both keeping and returning might happen, I assume $t > q_s + f - p$. Under this condition,

$$E_i(U_{purchase}) = \frac{(\alpha_i q_s - p + f)^2}{2t} - f .$$

Q.E.D.

A2. Derive the net demand function.

Since I set the return fee $f = 0$, the purchase condition is:

$$\alpha_i > \frac{p}{q_s} \quad (A2.1)$$

, and the keeping condition is:

$$\alpha_i < -\frac{t}{dq_s} x_i + \frac{(1+d)q_r - p}{dq_s} \quad (A2.2)$$

I put the boundary conditions that customers will buy and keep the product into a unit box.

a. When $\frac{(1+d)q_r - p}{dq_s} < \frac{p}{q_s} < 1$, or $p > q_r$ and $q_s > p$, no overlap area

between two boundary conditions, which means there is no net demand.

b. When $\frac{(1+d)q_r - t - p}{dq_s} < \frac{p}{q_s} < \frac{(1+d)q_r - p}{dq_s} < 1$, or $q_r - \frac{t}{(1+d)} < p < q_r$

and $\frac{(1+d)q_r - p}{d} < q_s$, the overlap area is a triangle.

The area of the triangle can be calculated as $\frac{[(1+d)(q_r - p)]^2}{2dtq_s}$, which is the

net demand function, representing the customers who will buy and keep the product.

c. When $\frac{p}{q_s} < \frac{(1+d)q_r - t - p}{dq_s} < \frac{(1+d)q_r - p}{dq_s} < 1$, or $p < q_r - \frac{t}{(1+d)}$ and $\frac{(1+d)q_r - p}{d} < q_s$, the overlap area is a trapezoid.

The area of the trapezoid can be calculated as $\frac{(1+d)(q_r - p) - 0.5t}{dq_s}$.

d. When $\frac{(1+d)q_r - t - p}{dq_s} < \frac{p}{q_s} < 1 < \frac{(1+d)q_r - p}{dq_s}$, or $p > q_r - \frac{t}{(1+d)}$ and $p < q_s < \frac{(1+d)q_r - p}{d}$, the overlap area is a trapezoid.

The area of the trapezoid can be calculated as $\frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}$.

e. When $\frac{p}{q_s} < \frac{(1+d)q_r - t - p}{dq_s} < 1 < \frac{(1+d)q_r - p}{dq_s}$, or $p < q_r - \frac{t}{(1+d)}$ and $\frac{(1+d)q_r - t - p}{d} < q_s < \frac{(1+d)q_r - p}{d}$, the overlap area is a pentagon.

The area of the pentagon can be calculated as:

$$\frac{-2dpt + 2dq_s t - (p + d(q_s - q_r) - q_r + t)^2}{2dq_s t}$$

f. When $\frac{p}{q_s} < 1 < \frac{(1+d)q_r - t - p}{dq_s} < \frac{(1+d)q_r - p}{dq_s}$, or $p < q_s < \frac{(1+d)q_r - t - p}{d}$, the overlap area is a rectangle.

The area of the rectangle can be calculated as $\frac{q_s + p}{q_s}$.

To sum up, I obtain the net demand function:

$$D(q_s, p) = \begin{cases} 0, & \text{if } p > q_r, q_s > p \\ \frac{[(1+d)(q_r - p)]^2}{2dtq_s}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r, \frac{(1+d)q_r - p}{d} < q_s \\ \frac{(1+d)(q_r - p) - 0.5t}{dq_s}, & \text{if } p < q_r - \frac{t}{(1+d)}, \frac{(1+d)q_r - p}{d} < q_s \\ \frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)}, p < q_s < \frac{(1+d)q_r - p}{d} \\ \frac{-2dpt + 2dq_s t - (p + d(q_s - q_r) - q_r + t)^2}{2dq_s t}, & \\ \text{if } p < q_r - \frac{t}{(1+d)}, \frac{(1+d)q_r - t - p}{d} < q_s < \frac{(1+d)q_r - p}{d} \\ \frac{q_s + p}{q_s}, & \text{if } p < q_s < \frac{(1+d)q_r - t - p}{d} \end{cases}$$

Q.E.D.

A3. Derive the relationship between the net demand, return rate, and return amount.

The net demand function is:

$$D(q_s, p) = \begin{cases} \frac{[(1+d)(q_r - p)]^2}{2dtq_s}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ \frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d} \end{cases}$$

When $q_r - \frac{t}{(1+d)} < p < q_r$ and $\frac{(1+d)q_r - p}{d} < q_s$, F.O.C. gives:

$$\frac{\partial D(q_s, p)}{\partial p} = -\frac{(1+d)^2(q_r - p)}{dq_s t} < 0$$

$$\frac{\partial D(q_s, p)}{\partial q_s} = -\frac{(1+d)^2(p - q_r)^2}{2dq_s^2 t} < 0$$

When $p > q_r - \frac{t}{(1+d)}$ and $p < q_s < \frac{(1+d)q_r - p}{d}$, F.O.C. gives:

$$\frac{\partial D(q_s, p)}{\partial p} = -\frac{-(2+d)p + q_s + q_r + dq_r}{q_s t} < 0$$

$$\frac{\partial D(q_s, p)}{\partial q_s} = -\frac{(2+d)p^2 + dq_s^2 - 2(1+d)pq_r}{2q_s^2 t}$$

$$\frac{\partial D(q_s, p)}{\partial q_s} = 0 \text{ gives, } q = \sqrt{\frac{p[2(1+d)q_r - (2+d)p]}{d}}, \text{ which yields:}$$

when $p < q_s < q$, $\frac{\partial D(q_s, p)}{\partial q_s} > 0$; when $q < q_s < \frac{(1+d)q_r - p}{d}$, $\frac{\partial D(q_s, p)}{\partial q_s} < 0$.

Q.E.D.

The return rate function is:

$$r(q_s, p) = \begin{cases} 1 - \frac{(1+d)^2(p - q_r)^2}{2d(q_s - p)t}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ 1 - \frac{2(1+d)q_r - (2+d)p - dq_s}{2t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases}$$

When $q_r - \frac{t}{(1+d)} < p < q_r$ and $\frac{(1+d)q_r - p}{d} < q_s$, F.O.C. gives:

$$\frac{\partial r(q_s, p)}{\partial p} = \frac{(1+d)^2(p - q_r)(p - 2q_s + q_r)}{2d(p - q_s)^2 t} > 0$$

$$\frac{\partial r(q_s, p)}{\partial q_s} = \frac{(1+d)^2(p - q_r)^2}{2d(p - q_s)^2 t} > 0$$

When $p > q_r - \frac{t}{(1+d)}$ and $p < q_s < \frac{(1+d)q_r - p}{d}$, F.O.C. gives

$$\frac{\partial r(q_s, p)}{\partial p} = \frac{2+d}{2t} > 0$$

$$\frac{\partial r(q_s, p)}{\partial q_s} = \frac{d}{2t} > 0$$

Q.E.D.

The return amount function is:

$$R(q_s, p) = \begin{cases} 1 - \frac{p}{q_s} - \frac{(1+d)^2(p-q_r)^2}{2dq_s t}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ \frac{(q_s - p)((2+d)p + d(q_s - 2q_r) - 2q_r + 2t)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases}$$

When $q_r - \frac{t}{(1+d)} < p < q_r$ and $\frac{(1+d)q_r - p}{d} < q_s$, F.O.C. gives:

$$\frac{\partial R(q_s, p)}{\partial q_s} = \frac{2dtp + (1+d)^2(p-q_r)^2}{2dtq_s^2} > 0$$

$$\frac{\partial R(q_s, p)}{\partial p} = \frac{-dt + (1+d)^2(-p + q_r)}{dtq_s}$$

$$\frac{\partial R(q_s, p)}{\partial p} = 0 \text{ gives, } p_1 = \frac{(1+d)^2 q_r - dt}{(1+d)^2}, \text{ which yields:}$$

when $q_r - \frac{t}{(1+d)} < p < p_1$, $\frac{\partial R(q_s, p)}{\partial p} > 0$; when $p_1 < p < q_r$, $\frac{\partial R(q_s, p)}{\partial p} < 0$.

When $p > q_r - \frac{t}{(1+d)}$ and $p < q_s < \frac{(1+d)q_r - p}{d}$, F.O.C. gives

$$\frac{\partial R(q_s, p)}{\partial q_s} = \frac{(2+d)p^2 + dq_s^2 - 2p(q_r + dq_r - t)}{2q_s^2 t} > \frac{d(q_s^2 - p^2)}{2q_s^2 t} > 0$$

$$\frac{\partial R(q_s, p)}{\partial p} = \frac{-(2+d)p + q_s + q_r + dq_r - t}{q_s t}$$

$$\frac{\partial R(q_s, p)}{\partial p} = 0 \text{ gives, } p_2 = \frac{q_s + q_r + dq_r - t}{2+d}, \text{ which yields:}$$

$$\text{when } q_r - \frac{t}{(1+d)} < p < p_2, \frac{\partial R(q_s, p)}{\partial p} > 0; \text{ when } p_2 < p < q_r, \frac{\partial R(q_s, p)}{\partial p} < 0$$

Q.E.D.

A4. Derive the optimal pricing and quality display strategies

The profit function is:

$$\pi(q_s, p) = \begin{cases} p \frac{[(1+d)(q_r - p)]^2}{2dtq_s}, & \text{if } q_r - \frac{t}{(1+d)} < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s; \\ p \frac{(-p + q_s)(-dp - dq_s + 2dq_r - 2p + 2q_r)}{2q_s t}, & \text{if } p > q_r - \frac{t}{(1+d)} \text{ and } p < q_s < \frac{(1+d)q_r - p}{d}. \end{cases}$$

$$(1). \text{ When } \max(q_r - \frac{t}{(1+d)}, 0) < p < q_r \text{ and } \frac{(1+d)q_r - p}{d} < q_s.$$

Since the profit is decreasing in q_s , the optimal value is obtained at the lower

bound which is $q_s = \frac{(1+d)q_r - p}{d}$. Then I substitute it into the profit function to

solve the maximization problem with respect to the price only:

$$\pi(p) = p \frac{[(1+d)(q_r - p)]^2}{2t[(1+d)q_r - p]}$$

F.O.C. gives us:

$$\frac{\partial \pi}{\partial p} = \frac{(1+d)^2(q_r - p)(2p^2 - 3(1+d)pq_r + (1+d)q_r^2)}{2(p - (1+d)q_r)^2 t}$$

$$\frac{\partial \pi}{\partial p} = 0 \text{ yields:}$$

$$p_1 = \frac{3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}}{4},$$

$$p_2 = \frac{3(1+d)q_r + q_r\sqrt{(1+d)(1+9d)}}{4},$$

$$p_3 = q_r.$$

Since $\frac{\partial p_2}{\partial d} > 0$, $p_2(d > 0) > p_2(d = 0) = q_r$. p_2 is out of the range, so I

consider p_1, p_3 .

a. If $t < (1+d)[\frac{q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)}}{4}]$, p_1 is out of the boundary,

and the profit is monotone decreasing in $[q_r - \frac{t}{(1+d)}, p_r]$. Then

$$p^* = p_1^* = q_r - \frac{t}{(1+d)}, \quad q_s^* = q_{s1}^* = \frac{d(1+d)q_r + t}{d(d+1)}, \quad \pi_1^* = \frac{t[(1+d)q_r - t]}{2[d(1+d)q_r + t]}.$$

b. If $(1+d)[\frac{q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)}}{4}] < t < (1+d)q_r$, p_1 is between the

boundary. The profit is monotone increasing in $[q_r - \frac{t}{(1+d)}, p_1]$ and decreasing

in $[p_1, q_r]$. Then $p^* = p_2^* = \frac{3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}}{4}$,

$$q_s^* = q_{s2}^* = \frac{q_r(d+1) + q_r\sqrt{(1+d)(1+9d)}}{4d},$$

$$\pi_2^* = \frac{[(1+d)(q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)})]^2 [3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}]}{32t[(1+d)q_r + q_r\sqrt{(1+d)(1+9d)}]}.$$

c. If $t > (1+d)q_r, q_r - \frac{t}{(1+d)} < 0$, the boundary condition becomes $0 <$

$$p < q_r, \text{ which is satisfied, } p^* = p_2^* = \frac{3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}}{4},$$

$$q_s^* = q_{s2}^* = \frac{q_r(d+1) + q_r\sqrt{(1+d)(1+9d)}}{4d},$$

$$\pi_2^* = \frac{[(1+d)(q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)})]^2 [3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}]}{32t[(1+d)q_r + q_r\sqrt{(1+d)(1+9d)}]}.$$

(2). When $p < q_s < \frac{(1+d)q_r - p}{d}$ and $q_r - \frac{t}{(1+d)} < p < q_r$, which is

equivalent to $\max(q_r - \frac{t}{(1+d)}, 0) < p < \min(q_s, q_r, (1+d)q_r - dq_s)$.

F.O.C. gives us:

$$\frac{\partial \pi}{\partial p} = \frac{p(d+2)(p-q_s) + p(dp + dq_s - 2dq_r + 2p - 2q_r) + (p-q_s)(dp + dq_s - 2dq_r + 2p - 2q_r)}{2q_s t}$$

$$\frac{\partial \pi}{\partial q_s} = \frac{-p(q_s(dp + dq_s - 2dq_r - d(p-q_s) + 2p - 2q_r) + (p-q_s)(dp + dq_s - 2dq_r + 2p - 2q_r))}{2q_s^2 t}$$

$$\frac{\partial \pi}{\partial p} = 0 \text{ and } \frac{\partial \pi}{\partial q_s} = 0 \text{ gives us:}$$

$$p_4 = \frac{2q_r + 2dq_r - q_r\sqrt{2d(1+d)}}{4+2d}, q_4 = q_r\sqrt{\frac{1+d}{2d}}$$

I compare p_4 and q_4 with their boundaries. It's easy to check $p_4 < q_r$, $p_4 < q_4$, $p_4 < (1+d)q_r - dq_4$. p_4 satisfies the right boundary condition.

For the left boundary condition:

a. When $t < (1+d)(1 + \sqrt{\frac{d(1+d)}{2}}) \frac{q_r}{2+d}$, $p_4 < q_r - \frac{t}{(1+d)}$. Then the optimal solution should be the boundary conditions. I compare the optimal solution at each boundary.

a.1 $p = q_r - \frac{t}{(1+d)}$. I get the optimal solution is:

$$p^* = p_3^* = q_r - \frac{t}{(1+d)}, \quad q_s^* = q_{s3}^* = \sqrt{\frac{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)}{d(1+d)^2}}$$

,

$$\pi_3^* = \frac{(q_r + dq_r - t)(dq_r + d^2q_r + t - \sqrt{d} \sqrt{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)})}{(1+d)^2 t}$$

.

a.2 $q_s = \frac{(1+d)q_r - p}{d}$. The optimal solution is the same as the first case,

when $q_r - \frac{t}{(1+d)} < p < q_r$ and $\frac{(1+d)q_r - p}{d} < q_s$.

Comparing the solutions in a.1 and a.2 gives the optimal solution:

$$p^* = p_3^* = q_r - \frac{t}{(1+d)}, \quad q_s^* = q_{s3}^* = \sqrt{\frac{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)}{d(1+d)^2}}$$

$$, \pi_3^* = \frac{(q_r + dq_r - t)(dq_r + d^2q_r + t - \sqrt{d} \sqrt{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)})}{(1+d)^2 t}$$

b. when $(1+d)(1+\sqrt{\frac{d(1+d)}{2}})\frac{q_r}{2+d} < t < (1+d)q_r$, $p_4 > q_r - \frac{t}{(1+d)}$, and

the optimal solution is: $p^* = p_4^* = \frac{2q_r + 2dq_r - q_r\sqrt{2d(1+d)}}{4+2d}$,

$$q_s^* = q_{s4}^* = q_r\sqrt{\frac{1+d}{2d}} \quad \pi_4^* = \frac{q_r^2(1+d)^2(2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2t}.$$

c. When $t > (1+d)q_r$, $q_r - \frac{t}{(1+d)} < 0$, the boundary condition becomes

$0 < p < \min(q_s, q_r, (1+d)q_r - dq_s)$. p_4, q_4 is the optimal solution,

$$p^* = p_4^* = \frac{2q_r + 2dq_r - q_r\sqrt{2d(1+d)}}{4+2d}, \quad q_s^* = q_{s4}^* = q_r\sqrt{\frac{1+d}{2d}}$$

$$\pi_4^* = \frac{q_r^2(1+d)^2(2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2t}.$$

Combining the solutions of case 2 and case 4, I have:

(1). When $t > (1+d)(1+\sqrt{\frac{d(1+d)}{2}})\frac{q_r}{2+d}$, I compare

$$\pi_4^* = \frac{q_r^2(1+d)^2(2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2t} \quad \text{and}$$

$$\pi_2^* = \frac{[(1+d)(q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)})]^2[3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}]}{32t[(1+d)q_r + q_r\sqrt{(1+d)(1+9d)}]}.$$

$$\pi_4^* > \pi_2^*$$

(2). When

$$(1+d)\left[\frac{q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)}}{4}\right] < t < (1+d)(1+\sqrt{\frac{d(1+d)}{2}})\frac{q_r}{2+d}, \text{ I}$$

compare

$$\pi_3^* = \frac{(q_r + dq_r - t)(dq_r + d^2q_r + t - \sqrt{d}\sqrt{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)})}{(1+d)^2t} \quad \text{and}$$

$$\pi_2^* = \frac{[(1+d)(q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)})]^2 [3(1+d)q_r - q_r\sqrt{(1+d)(1+9d)}]}{32t[(1+d)q_r + q_r\sqrt{(1+d)(1+9d)}]}.$$

$$\pi_3^* > \pi_2^*$$

(3). When $t < (1+d)[\frac{q_r - 3dq_r + q_r\sqrt{(1+d)(1+9d)}}{4}]$, I compare

$$\pi_1^* = \frac{t[(1+d)q_r - t]}{2[d(1+d)q_r + t]} \text{ and}$$

$$\pi_3^* = \frac{(q_r + dq_r - t)(dq_r + d^2q_r + t - \sqrt{d}\sqrt{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)})}{(1+d)^2t}.$$

$$\pi_3^* > \pi_1^*$$

To sum up,

$$(1) \text{ when } t > (1+d)(1 + \sqrt{\frac{d(1+d)}{2}}) \frac{q_r}{2+d},$$

$$p^* = p_4^* = \frac{2q_r + 2dq_r - q_r\sqrt{2d(1+d)}}{4+2d}, \quad q_s^* = q_{s4}^* = q_r\sqrt{\frac{1+d}{2d}}$$

$$\pi_4^* = \frac{q_r^2(1+d)^2(2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2t};$$

$$(2) \text{ when } t < (1+d)(1 + \sqrt{\frac{d(1+d)}{2}}) \frac{q_r}{2+d}, \quad p^* = p_3^* = q_r - \frac{t}{(1+d)},$$

$$q_s^* = q_{s3}^* = \sqrt{\frac{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)}{d(1+d)^2}}$$

$$, \pi_3^* = \frac{(q_r + dq_r - t)(dq_r + d^2 q_r + t - \sqrt{d} \sqrt{(q_r + dq_r - t)(d(1+d)q_r + (2+d)t)})}{(1+d)^2 t} ;$$

$$(3) \text{ when } t = (1+d)(1 + \sqrt{\frac{d(1+d)}{2}}) \frac{q_r}{2+d},$$

$$p_3^* = p_4^* = \frac{q_r(2+2d - \sqrt{2d(1+d)})}{2(2+d)}, q_{s3}^* = q_{s4}^* = q_r \sqrt{\frac{1+d}{2d}}, \pi_3^* = \pi_4^*.$$

Note that it is necessary to compare this case with case 5 because when t is very small, the firm will adjust its strategy to let the consumers who are extremely unfit to the product also keep it, then the condition in case 4 will not hold. Finally, I get:

$$p^* = p_4^* = \frac{2q_r + 2dq_r - q_r \sqrt{2d(1+d)}}{4+2d}, q_s^* = q_{s4}^* = q_r \sqrt{\frac{1+d}{2d}}$$

$$\pi_4^* = \frac{q_r^2(1+d)^2(2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2 t}, \text{ with } t > (1+d)(1 + \sqrt{\frac{d(1+d)}{2}}) \frac{q_r}{2+d}.$$

Q.E.D.

A5. Derive the relationship between the optimal solutions and d, t respectively.

$$p^* = \frac{2q_r + 2dq_r - q_r \sqrt{2d(1+d)}}{4+2d}, q_s^* = q_r \sqrt{\frac{1+d}{2d}}$$

F.O.C. gives:

$$\frac{\partial p^*}{\partial t} = 0, \frac{\partial p^*}{\partial d} = \frac{(-2\sqrt{2} - 3\sqrt{2}d + 4\sqrt{d(1+d)})q_r}{4\sqrt{d(1+d)}(2+d)^2} < 0$$

$$\frac{\partial q_s^*}{\partial t} = 0, \frac{\partial q_s^*}{\partial d} = -\frac{q_r}{2\sqrt{2}\sqrt{d^3(1+d)}} < 0$$

When $d \leq 1$, $q_s^*(d) \geq q_s^*(d=1) = q_r$; when $d > 1$, $q_s^*(d) < q_s^*(d=1) = q_r$.

Q.E.D.

A6. Derive the relationship between return amount, return rate, and d, t respectively under the optimal pricing and product display strategies.

The return amount and return rate functions are:

$$R^* = \frac{(2+2d-\sqrt{2}\sqrt{d(1+d)})(2t-q_r-dq_r)}{2(2+d)t}, \quad r^* = \frac{2t-(1+d)q_r}{2t}$$

F.O.C. gives:

$$\frac{\partial R^*}{\partial t} = \frac{(1+d)(2+2d-\sqrt{2}\sqrt{d(1+d)})q_r}{2(2+d)t^2} > 0$$

$$\frac{\partial R^*}{\partial d} = \frac{(1+d)(7\sqrt{2}d+2\sqrt{2}d^2-4d\sqrt{d(1+d)}+2(\sqrt{2}-6\sqrt{d(1+d)}))q_r+(-4\sqrt{2}-6\sqrt{2}d+8\sqrt{d(1+d)})t}{4\sqrt{d(1+d)}(2+d)^2t}$$

$$\frac{\partial r^*}{\partial t} = \frac{(1+d)q_r}{2t^2} > 0, \quad \frac{\partial r^*}{\partial d} = -\frac{q_r}{2t} < 0$$

Since $-4\sqrt{2}-6\sqrt{2}d+8\sqrt{d(1+d)} < 0$, the numerator of $\frac{\partial R^*}{\partial d}$ is monotone

decreasing w.r.t. t , and $t > (1+d)(1+\sqrt{\frac{d(1+d)}{2}})\frac{q_r}{2+d}$, then I obtain:

$$\begin{aligned} & (1+d)(7\sqrt{2}d+2\sqrt{2}d^2-4d\sqrt{d(1+d)}+2(\sqrt{2}-6\sqrt{d(1+d)}))q_r+(-4\sqrt{2}-6\sqrt{2}d+8\sqrt{d(1+d)})t \\ & < \frac{(1+d)(-20\sqrt{d(1+d)}+d(14\sqrt{2}-26\sqrt{d(1+d)}+d(15\sqrt{2}+2\sqrt{2}d-4\sqrt{d(1+d)})))}{2+d}q_r < 0 \end{aligned}$$

Therefore, $\frac{\partial R^*}{\partial d} < 0$.

Q.E.D.

A7. Derive the relationship between the optimal profit and d, t respectively

The optimal profit function is:

$$\pi^* = \frac{q_r^2 (1+d)^2 (2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2 t}$$

F.O.C. gives:

$$\frac{\partial \pi^*}{\partial t} = -\frac{q_r^2 (1+d)^2 (2+3d-2\sqrt{2}\sqrt{d(1+d)})}{2(2+d)^2 t^2} < 0$$

$$\frac{\partial \pi^*}{\partial d} = \frac{(1+d)(-2\sqrt{2}-2\sqrt{2}d^3+10\sqrt{d(1+d)}+d^2(-11\sqrt{2}+3\sqrt{d(1+d)}))+d(-11\sqrt{2}+15\sqrt{d+d^2}))q_r^2}{2\sqrt{d(1+d)}(2+d)^3 t}$$

When $0 < d < -2 + \sqrt{5}$, $\frac{\partial \pi^*(q_r, d, t)}{\partial d} < 0$; when $d > -2 + \sqrt{5}$,

$$\frac{\partial \pi^*(q_r, d, t)}{\partial d} > 0.$$

Q.E.D.

References

- Anderson, E. T., et al. (2009). "The option value of returns: Theory and empirical evidence." *Marketing Science* 28(3): 405-423.
- Bower, A. B. and J. G. Maxham III (2012). "Return shipping policies of online retailers: normative assumptions and the long-term consequences of fee and free returns." *Journal of Marketing* 76(5): 110-124.

- Calfee, J. E. and D. J. Ringold (1994). "The 70% majority: Enduring consumer beliefs about advertising." *Journal of public policy & marketing* 13(2): 228-238.
- Chakraborty, A. and R. Harbaugh (2014). "Persuasive puffery." *Marketing Science* 33(3): 382-400.
- Crawford, V. P. and J. Sobel (1982). "Strategic information transmission." *Econometrica: Journal of the Econometric Society*: 1431-1451.
- Davis, S., et al. (1998). "Return policies and the optimal level of "hassle"." *Journal of Economics and Business* 50(5): 445-460.
- Gardete, P. M. (2013). "Cheap-talk advertising and misrepresentation in vertically differentiated markets." *Marketing Science* 32(4): 609-621.
- George, A. (1970). "The Market for Lemons: Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics*.
- Hess, J. D., et al. (1996). "Controlling product returns in direct marketing." *Marketing Letters* 7(4): 307-317.
- Hoffmann, F., et al. (2013). "Hypertargeting, limited attention, and privacy: Implications for marketing and campaigning." *Limited Attention, and Privacy: Implications for Marketing and Campaigning (May 1, 2013)*.
- Janakiraman, N., et al. (2016). "The effect of return policy leniency on consumer purchase and return decisions: A meta-analytic review." *Journal of Retailing* 92(2): 226-235.
- Kirmani, A. and A. R. Rao (2000). "No pain, no gain: A critical review of the literature on signaling unobservable product quality." *Journal of Marketing* 64(2): 66-79.
- Malär, L., et al. (2011). "Emotional brand attachment and brand personality: The relative importance of the actual and the ideal self." *Journal of Marketing* 75(4): 35-52.

- Marks, L. J. and M. A. Kamins (1988). "The use of product sampling and advertising: Effects of sequence of exposure and degree of advertising claim exaggeration on consumers' belief strength, belief confidence, and attitudes." *Journal of Marketing Research* 25(3): 266-281.
- Nelson, P. (1974). "Advertising as information." *Journal of political economy* 82(4): 729-754.
- Preston, I. L. (1996). *The great American blow-up: Puffery in advertising and selling*, Univ of Wisconsin Press.
- Shulman, J. D., et al. (2009). "Optimal restocking fees and information provision in an integrated demand-supply model of product returns." *Manufacturing & Service Operations Management* 11(4): 577-594.
- Shulman, J. D., et al. (2010). "Optimal reverse channel structure for consumer product returns." *Marketing Science* 29(6): 1071-1085.
- Shulman, J. D., et al. (2011). "Managing consumer returns in a competitive environment." *Management Science* 57(2): 347-362.
- Spence, M. (1978). Job market signaling. Uncertainty in economics, *Elsevier*: 281-306.
- Wolinsky, A. (1983). Prices as signals of product quality. *The review of economic studies*, 50(4), 647-658.
- Wood, S. L. (2001). "Remote purchase environments: The influence of return policy leniency on two-stage decision processes." *Journal of Marketing Research* 38(2): 157-169.
- Zhang, S., et al. (2019). Can Lower-quality Images Lead to Greater Demand on AirBnB?, *Working Paper, Carnegie Mellon University*.

Essay 2: Seller Show Vs Buyer Show: Utilizing Online Images Posted by Customers and Stores to Predict Product Returns

Abstract

Product returns have become an increasingly important issue for firms to address, as main shopping platforms are gradually transferring from traditional brick-and-mortar stores to online markets, and firms often offer lenient return policies (such as free return). The ability to better predict the product returns is economically significant for the e-commerce platforms, since such returns typically constitute 2%-5% of the revenue, depending on the product category. In this article, I develop an AI system to extract features from both seller-posted and buyer-posted images and investigate the predictive power of the difference between these two types of images on product returns. Seller-posted images reflect consumers' expected product quality, while buyer-posted images reflect the real product quality. A larger difference between these two types of images reflects a larger difference between the expected and real product quality, resulting in a high return rate. Specifically, first, I use well-established image-processing techniques to extract basic image attributes, such as entropy, cool color ratio, lightness, and saturation, which are defined as "shallow features". Second, I use a residual learning framework, ResNet-50, combining with a special neural network structure, Siamese neural network, to recognize images, and choose the second-to-last layer of the network, containing 512 features, as "deep features". Third, I calculate the Euclidean and Cosine distances between seller-posted images and buyer-posted images using the extracted features, and I employ a state-of-art machine learning algorithm – gradient boosted regression tree – to predict product return rate leveraging the calculated distances. The results show that the estimated differences between the two types of images have statistically significant predictive power for product returns by increasing the prediction accuracy by 31.25%.

Keywords: Product returns; user-generated content; computer vision; deep learning.

1. Introduction

Product returns have become a more and more important issue for firms to address, as the main shopping platforms are gradually transferring from traditional brick-and-mortar stores to online markets, and firms often offer lenient return policies (such as free return) (Minnema, Bijmolt, Gensler, & Wiesel, 2016). It has a significant impact on firms' profitability. Evidence shows that for the United States retailers in 2018, total customer returns account for more than \$369 billion, which is close to the market cap value of Facebook (Retail, 2018). Hence, firms need to be able to accurately predict product returns and consider them for profit optimization. In this paper, I investigate the predictive power of the differences between seller-posted and buyer-posted images on product returns at one of the largest B2C e-commerce companies in Asia.

Despite its importance, there is limited academic research on product returns, in both the marketing and computer science community, likely due to the unavailability of product returns data. In the marketing literature, most of the previous research focuses on designing the optimal return policy, e.g., lenient vs. strict. A recent work (Sahoo, Dellarocas, & Srinivasan, 2018) quantifies the impact of online product reviews on product returns, which is closely related to my work. They frame the problem under an information and expectation management perspective and show that by reducing product uncertainty, product reviews affect the probability of product returns. However, they only explore text reviews in the empirical study.

In computer science literature, there is only a handful of research and mostly focus on end-of-life returns (Ma & Kim, 2016; Toktay, van der Laan, & de Brito, 2004; Urbanke, Kranz, & Kolbe, 2015). Two very recent works (J. Li, He, & Zhu, 2018; Y. Zhu, Li, He, Quanz, & Deshpande, 2018) predict product returns using graph-based algorithms on graphs constructed from purchase and return history. However, they do not utilize the rich product and consumer

attribute information, such as product characteristics and reviews posted by consumers, and suffer from cold start problems. Also, in J. Li et al. (2018), the prediction happens after the consumer puts together the shopping basket. Online retailers can only take interventions such as incentivizing the customer not to issue a return, instead of taking a more proactive action at an earlier stage in the consumer purchase path and preventing him/her from putting the product in the shopping basket.

Given the rise of visual communication and the fact that more and more consumers upload images when reviewing products, this offers companies an invaluable information source that they can consider leveraging to help understand and manage product returns. The online purchase decision and return decision can be framed as two stages (Wood, 2001): (1) When consumers go shopping online, they will browse the images posted by sellers. Combining with other descriptive information provided by the sellers on the websites, they will form an expectation about the product. If the expected quality is good enough based on their shopping experience, then they will buy them. (2) After they receive the product, they will observe the real product quality. If the real quality is lower than the expected quality, then probably they will return the product. I propose that store-posted images can reflect customers' expected product quality, while customer-posted images can reflect the real product quality. The differences between seller-generated images and user-generated images reflect the differences between the expected and realized product quality. I predict that the larger the differences between the two kinds of images the more product returns will be. I further investigate the predictive power of the differences between these two kinds of images on product returns.

The main contributions of the current research are that (1) I propose an artificial intelligence framework to automatically calculate the differences between seller-posted images and buyer-posted images and use the extracted attributes to predict product return; (2) I prove the value of image information (both store and customer posted images) on predicting product return in online

shopping platforms. After involving the two types of images, the prediction accuracy is increased by 31.25%. (3) I bring insights into the synergetic effect of store-posted and consumer-posted images on consumer's shopping behavior.

The rest of the paper is organized as follows. In the next section, I review the literature related to this research. In Section 3, I describe the data and conduct a preliminary analysis. Section 4 presents the methods that I use to extract features from images and investigates the correlation between the difference and return rate. In section 5, leveraging the image features, I use a state-of-art machine learning method to predict product returns by involving image information. Finally, I conclude the paper in Section 6.

2. Literature Review

This paper draws on and contributes to three streams of research in the marketing literature: product returns, user-generated content, and image mining.

2.1. Product Returns

Previous literature on product returns mainly focuses on the impacts of different product return policies. In practice, retailers often offer different return policies, and their implications and consequences have been studied in the literature. Some retailers offer lenient return policies, while others offer strict policies (Davis, Hagerty, & Gerstner, 1998). By imposing strict return policies, retailers can avoid loss by reducing the number of inappropriate returns (Hess, Chu, & Gerstner, 1996). However, a very costly return process will make the company hardly retain current customers and lose a lot of potential customers through the negative word of mouth (Petersen & Kumar, 2009). On the other hand, Bechwati and Siegal (2005) show that a lenient product return process is a signal of good quality products and can increase the satisfaction of customers to the company, which will bring more profits for the company in the future. However, lenient return policies will increase the probability of returning products. Even worse, zero-cost return policies will lead to customers'

“deshopping” behaviors, which are defined as deliberately returning goods for reasons other than actual faults in the product (King & Dennis, 2003). Such dishonest behavior will make the return issues worse and it is difficult for retailers to identify such dishonest consumers.

Given the trade-off between lenient and strict return policies, what is the optimal return policy that companies should launch? Davis et al. (1998) argue that if a product cannot be consumed in a short time and it has a high salvage value for returned merchandise, the retailer should offer a low hassle return policy. More generally, Anderson, Hansen, and Simester (2009) claim that product return choices can be considered as an option that has measurable value for customers, and they develop a structural model that enables a firm to balance the costs and benefits of different returns policies. Using their model, a firm can optimize its return policies across categories and customers. But they do not consider the multi-retailer situation and the variation in competitors’ return policies.

For online shopping platforms, they usually contain multiple product categories and multiple retailers. Besides, customer return issues are even more severe due to the intangible online shopping environment and customers’ limited abilities to identify the quality of the product when they browse the products online. Bower and Maxham III (2012) suggest that online retailers should decrease product return cost as much as possible or offer a free-return policy based on findings from two longitudinal field studies over four years. Janakiraman and Ordóñez (2012) conduct a meta-analysis of 21 papers to study the effect of lenient policies on sales and returns, and they find that leniency increases product sales more than product returns. From previous literature, it indicates that, overall, lenient return policies are better than strict policies for online retailers.

Consistent with the literature, most online retailers take a policy of free product returns. However, product returns bring a huge amount of loss to online retailers. What are the reasons for consumer returns? Petersen and Kumar

(2009) state that products received as gifts, products purchased on sale, products purchased in new distribution channels, are less likely to be returned; while products purchased during holidays, purchased in new categories are more likely to be returned. Product return rates also increase with the price paid (Anderson, Hansen, Simester, & Wang, 2008) and when deadlines are shorter (Janakiraman & Ordóñez, 2012). But those factors are not the fundamental reasons that customers return the products.

The fundamental reasons that customers return the products are the exposure to disconfirming information related to what information customers received when they made the purchase decision (Bechwati & Siegal, 2005), i.e. if customers are exposed to disconfirming information, the customer's return probability will be high. Using this theoretical framework, many phenomena could be explained. For example, Anderson et al. (2008) find product return rates increase with the price paid since the high product price will lead to a high product quality expectation, which will be more likely to lead to expectation disconfirmation. Minnema et al. (2016) find that customers' reviews have a positive effect on product return rate since the positive customer reviews increase the expectation of good quality. This paper adopts the theoretical framework combined with user-generated content and product information to predict product returns for online retailers.

2.2. User-Generated Contents

With the development of the internet and computer science, tremendous user-generated content is stored on various online platforms. Over the past decade, extensive studies have demonstrated the power of user-generated content (UGC) in marketing research. Several studies examine the relationship between user-generated content and sales. For instance, Chevalier and Mayzlin (2006) find that customers' ratings and the length of customers' reviews on books can influence book sales, and one-star reviews have a greater impact than

five-star reviews. Y. Liu (2006) shows that the volume and valence of customers' reviews on movies offer significant explanatory power for box office revenue, but the explanatory power mainly comes from the volume. Similarly, Xinxin Li and Hitt (2008) claim that early buyers' reviews can affect long-term consumer purchase behavior, and consumers do not discount early reviews when they search for books' quality information. More recently, Babić Rosario, Sotgiu, De Valck, and Bijmolt (2016) conduct a meta-analysis of 96 studies covering 40 platforms and 26 product categories and found that UGCs are positively correlated with sales on average, although the effect size varies across platforms and product categories.

Some studies look more deeply into the impact of UGCs. F. Zhu and Zhang (2010) find that the impact of customers' reviews on sales is moderated by product and customer characteristics. For less popular games and games whose players have more internet experience, the sales are less likely to be influenced. Forman, Ghose, and Wiesenfeld (2008) show that if reviewer identities are disclosed, the impact of reviews on subsequent sales will be greater. Compared with traditional marketing communication tools, UGCs have substantially longer carryover effects on sales and produce higher response elasticities (Trusov, Bucklin, & Pauwels, 2009). Connecting UGCs to product returns, Sahoo et al. (2018) argue that product reviews can reduce the probability of product returns by reducing product uncertainty.

The majority of UGCs' studies extract information by measuring UGCs' valence and volume. However, UGCs are stored in unstructured formats (e.g., texts, images, audios, and videos), which should be multifaceted. It is unreasonable and inefficient to compress the complex contents to just a number. As sophisticated natural language processing (NLP) and machine learning techniques have been developed, researchers have been exploring ways to extract more useful information from these unstructured contents. Hu and Liu (2004) mine product features that have been commented on by customers using an approach called "part-of-speech (POS) tagger", which can identify

frequently mentioned nouns and noun phrases in the text contents. Netzer, Feldman, Goldenberg, and Fresko (2012) construct the market-structure of sedan cars and diabetes drugs by investigating user-generated content from consumer forums. Tang and Guo (2015) extract features from texts by pre-defining a keyword list and matching the text with the list. X. Liu, Singh, and Srinivasan (2016) use a text mining technique called stochastic singular value decomposition combined with cloud computing services to identify the features of TV shows from Tweets.

Although marketing research has made progress in mining unstructured contents, most of the work solely focuses on text, in which a significant portion of user-generated content is overlooked. Compared to text content, visual content is ubiquitous in modern life, and it can reflect more intangible attributes of the brands (Keller & Lehmann, 2006).

2.3. Image Mining

The marketing literature on image mining is sparse. Xin Li, Wang, and Chen (2014) utilize well-established image-processing techniques to study the impact of product photos on consumers' purchase intention. They derive entropy, cool-color ratio, saturation, and lightness from product photos as factors that will influence consumers' purchases. But those variables are the features of photos and cannot reflect product attributes well.

With the rapid development of deep learning techniques, an increasing number of scholars have applied the convolutional neural network (CNN, or ConvNet) (Krizhevsky, Sutskever, & Hinton, 2012) to marketing research to extract features from online images. L. Liu, Dzyabura, and Mizik (2018) apply deep learning methods to identify how brands are portrayed on social media by mining images posted on Instagram. They find key differences between the perceived brand's attributes derived from the consumer-created brand images and those derived from the firm-created brand images. Dzyabura, El Kihal, and

Ibragimov (2018) show that by incorporating product image features, the predictive accuracy of the product return prediction model can be improved by 37%. But they only consider store-posted images. Zhang, Mehta, Singh, and Srinivasan (2019) argue that lower-quality images on Airbnb will lead to more demand in the future by reducing customers' expectations and lowering the dissatisfactory gap. In the current research, I consider both store-posted product photos and customer image reviews.

3. Data

3.1. Data Source

The data for this research comes from a large e-commerce company in Asia, which is a massive business-to-customer online retailer with tremendous transaction volume and revenue. As of the first quarter of 2018, the platform has 301.8 million active users. The online shopping platform covers 100 first-class product categories (e.g., apparel, shoes, household appliances, mobile phone, etc.), 1441 second-class product categories (e.g. women's shoes, women's wear, kitchen appliance, mobile communications, etc.), and 14817 third-class product categories (e.g. high heels, dress, washing machine, electric kettle, mobile phone shell, etc.).⁵

Because of the tremendous transaction volume, I only select the data in August 2018. In total, it has generated more than 400 million valid orders, with more than 70 billion RMB in revenue and the returned products worth 800 million RMB⁶. The return rate is around 1.1711%, which shows that product returns indeed brought a huge amount of loss to the company.

Since the probabilities of returning products are different for different product categories, I investigate each first-class product category. I keep the

⁵ All the data are stored as tables in a Hadoop distributed file system, and I utilize Hive query language (HQL) to select and manipulate the data.

⁶ All the sales information is presented vaguely because of the NDA.

product categories that have more than 10 million RMB sales. Fig 1 presents the sales and returns information in each first-class product category. From the left part, household appliances, computers, and cellphone categories have the highest sales, which are more than 10 billion RMB. For the right part, apparel and outdoor categories have more product returns than cellphone categories, which shows that larger sales do not always imply larger returns.

To compare normalized returns across different categories, I calculate the product return rate, which is defined by the ratio of return amounts to sale amounts, in each category and sort them in descending order. Fig 2 presents the product return rate in each category. I find that consumable product categories (e.g. food, cosmetics, personal care, cleaning supplies) have smaller return rates than inconsumable product categories (e.g. apparel, digitals, etc.), which is consistent with the past research (Bonifield, Cole, & Schultz, 2010). Besides, I find that apparel and outdoor goods have a larger return rate than household appliances, computers, and digital products. Product return decisions are based on customers' expectations of the products' quality and the uncertainty of the expected quality. The uncertainty comes from the customers' limited ability to identify product quality. There will be more uncertainty in the fashion products whose perceived quality is mainly derived from product images than functional products whose perceived quality is mainly derived from the text information. Therefore, the return rate of fashion products will be higher than that of functional products.

Fig1. Sales and returns information by first-class product categories in August 2018

Fig2. Ordered product returns by first-class product categories in August 2018

3.2. Variable Operationalization

In this paper, I focus on four sets of variables: product-related information, customers' text review information, customers' image review information, and product images posted by stores.

For product-related information, I collect product sales amount, sales quantity, refund amount, price, freight fee, return freight fee, purchase user number, and duration for each stock keeping unit (SKU).

For customers' text review information, I gather comment number, rating score, standard deviation, "good" number, "mid" number, and "bad" number for each SKU.

For product image information, I have the images posted by stores and customers.

Detailed definitions of these variables are shown in Table 1.

3.3. Descriptive Statistics

Due to the massive transaction volume and the difficulty of processing a huge number of images, instead of studying all categories, I focus on the "Dress" category. Although it is only a third-class product category, the number of products (SKUs) sold in August is huge. There were 300 thousand kinds of dress products sold in August, with a total of 200 million RMB in sales amount and 30 million RMB in return amount. The product return rates reach up to 10.92%. Among the 300 thousand kinds of dresses, 297,826 products have no freight fee; 110,748 products have text reviews by customers; 4,620 products have image reviews; only 167 products have video reviews.

Table 2 reports the descriptive statistics for the variables of interest. On average, the return rate of products with image reviews is lower than that of products without image reviews (Image M = 11%, No-image M = 15%). Observation suggests that image reviews can help customers better assess the quality of a product, which will reduce uncertainty on the perceived quality and hence reduce the probability of returning the product. However, the causality should be inferred by solving some endogenous issues. The first endogeneity is simultaneity. Products with higher return rates may tend to have more image

reviews. A high return rate could indicate that customers are not satisfied with the products, which would lead them to post more image reviews to complain about it. The second endogeneity is omitted variables. Some confounding factors may influence both product return rate and image reviews.

Unobservable product attributes are possible omitted variables. For example, some fashion products have high product return rates for their low practical value and high uncertainty in the perceived quality; meanwhile, people are likely to post images when they buy fashion products for their conspicuous consumption features. These issues are outside the scope of this paper and I will discuss them in the future study.

In addition, products with image reviews have larger sales amount ($M = 5489.91$ RMB), more return amount ($M = 329.63$ RMB), more purchasing customers ($M = 39.47$), more text reviews ($M = 20.61$), and higher variation in text reviews ($M = 0.19$) than products with no image reviews.

The main research question of this paper is to utilize the differences between images posted by stores and those posted by customers to predict product return rates. In the next section, I will discuss how to process and extract useful information from images, and then incorporate the extracted image features into the predictive model.

Table 1. Variable operationalization

Variable	Description and measure
Product-related information	
Sales amount	Aggregate sales amount (in CNY) for each SKU
Sales quantity	Aggregate sales quantity for each SKU
Refund amount	Aggregate refund amount for each SKU
Return rate	Refund amount divided by sales amount for each SKU
Price	Average price for each SKU

Freight fee	Average freight fee for each SKU
Return freight fee	Average return freight fee for each SKU
Purchase user number	The number of users who purchase the SKU
Duration	Days that the SKU exists in the shopping platform until 2018-09-01
Customers' text review information	
Comment number	Total number of reviews by customers for each SKU posted
Rating score	Average rating score by customers for each SKU
Standard deviation	The standard deviation of the rating scores for each SKU
Good number	The number of good reviews (5, 4 scores) by customers for each SKU
Mid number	The number of mid reviews (3 scores) by customers for each SKU
Bad number	The number of bad reviews (1, 2 scores) by customers for each SKU
Image information	
Image number	The number of image reviews posted by customers for each SKU
Video number	The number of video reviews posted by customers for each SKU
Images posted by customers	All the images posted by customers that are displayed on each SKU's page
Images posted by stores	All the images posted by stores that are displayed on each SKU's page

Table 2. Descriptive statistics

	Mean	Std	Min	Max	No-img	Img
Return rate	0.15	0.31	0.00	1.00	0.15	0.11
Sales amount	792.18	5488.76	0.07	506698.00	725.91	5489.91

Sales quantity	5.04	37.54	1.00	3668.00	4.53	41.30
Refund amount	86.19	285.88	0.00	30887.00	82.76	329.63
Price	196.36	227.45	0.01	11039.00	196.53	184.97
Freight amount	0.96	3.15	0.00	123.00	0.97	0.62
Return freight	0.01	0.25	0.00	20.00	0.01	0.02
Purchase user number	4.82	35.76	1.00	3236.00	4.33	39.47
Duration	143.50	193.23	1.00	2012.00	143.27	159.73
Comment number	1.97	23.86	0.00	1658.00	1.70	20.61
Score	4.81	0.70	1.00	5.00	4.81	4.69
Standard deviation	0.07	0.32	0.00	2.00	0.07	0.19
Good number	5.92	41.39	0.00	1657.00	5.28	20.26
Mid number	0.05	0.25	0.00	10.00	0.04	0.14
Bad number	0.06	0.30	0.00	16.00	0.06	0.21
Image number	0.03	0.31	0.00	29.00	0.00	1.81
Video number	0.00	0.02	0.00	1.00	0.00	0.04

4. Empirical Analysis

4.1. Preliminary Processing of the Image Data

To measure the quality of each product as objectively as possible, I download all the images that are displayed on the online shopping platform on Sep 1st, 2018 in the dress category. I get 400 million dresses that have store-posted images, among which 700 thousand dresses have customer-posted images. Totally I have 3,162,602,079 store-posted images and 2,189,113 customer-posted images. For the statistical validity, I select dresses that have more than 5 image reviews (I also try 10, and 20 as robustness checks). Finally, I get 1,063,431 customer-posted images and 3,122,857 store-posted images.

Since customers uploaded many unrelated images to the online shopping platform, I need to filter this kind of image. I utilize a neural network named “single-shot multi-box detector (SSD)” (W. Liu et al., 2016) to detect clothes in images. If there are no clothes in the image, then I drop that image.

To train the SSD neural network, I use the “DeepFashion” dataset (<http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html>) as the training dataset. The DeepFashion dataset contains 800,000 clothes images, which are annotated with 50 categories, clothing landmarks, and 1,000 attributes which are categorized into five groups (texture, fabric, shape, part, and style). I spent around one week training the SSD neural network and then used the well-trained network to filter the customer-posted images. To be fair, I also filtered the images posted by stores. Next, I will extract features from images.

Fig 3. An example of unrelated images posted by customers

Fig 4. The structure of the single-shot multi-box detector

4.2. Features Extraction and Distance Calculation

4.2.1 Deep Learning Methods

To calculate the distance between seller-posted images and consumer-posted images, I first utilize a deep neural network to extract features from online product images and then calculate the distance based on the extracted features. Deep learning techniques have been widely applied to many fields, including speech recognition, natural language processing, and computer vision, etc. A deep neural network is a type of artificial neural networks with many layers, which can be expressed as linear or non-linear functions. To analyze visual imagery, the convolutional neural network (CNN, or ConvNet) proposed by Krizhevsky et al. (2012) is mostly applied, which allows the network to be deeper with fewer parameters than the normal neural network. CNNs are inspired by the organization of the animal visual cortex, applying a convolution operation to the input and passing the result to the next layer. CNNs recognize an object gradually. In the first few layers, it detects relatively simple things such as edges or textures; then in the following layers, it detects angels or motifs; in the deeper layers, it detects very complex objects, which sometimes even cannot be recognized by humans.

With the rapid development of deep learning techniques, more layers and parameters are required to meet the needs of increasingly complex tasks, making the deep neural networks more difficult to train. He, Zhang, Ren, and Sun (2016) present residual learning (ResNet) framework which enables us to train very deep networks. Sometimes even networks of over 100 layers. Comparing to a normal ConvNet that several layers are stacked and are trained forward layer by layer, a ResNet has additional shortcut connections that directly connecting the input of n th layer to a deeper $(n + x)$ th layer. By this kind of “residual block”, a ResNet can learn extra residuals of features learned from the input of that layer.

In this research, I will use ResNet 50 as the base model. Besides, to identify the differences between the store-posted and customer-posted images, it is better to apply a special neural network structure called Siamese neural network (Koch, Zemel, & Salakhutdinov, 2015), which is good at learning the similarity between two inputs. It consists of two identical neural networks with the same structure and parameters. When I would like to find out whether there is a similarity between the two images, I put two images to the two identical neural networks respectively and get two encodings that could be two latent feature representations of the two images. Then I compare these two encodings to check whether they are the same. To train the Siamese neural network, I need three inputs for each calculation of loss function: the anchor image, a positive image that is in the same category as the anchor image, and a negative image that is in a different category as the anchor image. The loss function is given by:

$$triplet\ loss = \max(d(a, p) - d(a, n) + margin, 0) \quad (1)$$

, where “a” represents the anchor image, “p” represents a positive image and “n” represents a negative image, function $d(\cdot, \cdot)$ measures the differences between two inputs. This loss function is known as a triplet loss function. By minimizing the average triplet loss, I can make two similar images have similar encodings calculated by the Siamese neural network, and make sure that the

more different images are the more different encodings are. With only triplet loss in the loss function, the neural network will be trained inefficiently, since it is very easy to make $d(a, p) < d(a, n) - \text{margin}$. So, I also include cross-entropy loss in the loss function. The cross-entropy loss is given by:

$$\text{cross_entropy} = -\sum_{c=1}^M y_{i,c} \log(p_{i,c}) \quad (2)$$

, where $y_{i,c}$ is a dummy variable that indicates whether the image i is correctly classified into the category c , $p_{i,c}$ is the probability that image i will be classified into the category c .

To sum up, the neural network structure is ResNet50 – (Linear 1024 to 512) – (Linear 512 to 20), and the 512-dim vector is regarded as the latent feature representations of images. The loss function is defined as:

$$\text{loss_function} = \text{cross_entropy_loss} + \text{weight} * \text{triplet_loss} \quad (3)$$

To accelerate the training of the model, I use the reduced form of “DeepFashion” dataset as the training dataset, with 3*224*224 dimensions for each image, 289,222 number of clothes images and 20 categories. Since for each training image it has 1,000 attributes in five aspects, texture, fabric, shape, part and style, it is confident to say that the 512-dim encodings can describe a dress well.

4.2.2. Cosine and Euclidean Similarities

After I get the 512-dim encodings for each image, I calculate the average value of it for each SKU. For each product, it has a 512-dim “deep” feature extracted from the store-posted images and customer-posted images respectively, and then I calculate the difference between them. Let f_i^s denote the features extracted from the store-posted images of product i ; let f_i^c denote the features extracted from the customer-posted images of product i . Note that f_i^s and f_i^c are 512-dim vectors. Then I apply Cosine and Euclidean similarities as the distance metrics, which are given by:

$$\text{Cosine similarity: } \cos(f_i^s, f_i^c) = \frac{f_i^s \cdot f_i^c}{\|f_i^s\| \|f_i^c\|} = \frac{\sum_t f_{it}^s \cdot f_{it}^c}{\sqrt{\sum_t f_{it}^{s^2}} \sqrt{\sum_t f_{it}^{c^2}}} \quad (4)$$

$$\text{Euclidean similarity: } Euc(f_i^s, f_i^c) = \sqrt{\sum_t (f_{it}^s - f_{it}^c)^2} \quad (5)$$

, where f_{it}^s, f_{it}^c are the elements of the vector f_i^s, f_i^c .

Note that the Euclidian similarity and Cosine similarity do not consider the direction of the difference. If consumers find the real product quality is better than the expected quality, they will definitely not return the product. In this scenario, a larger difference between the two types of images might indicate a smaller return rate. However, prior literature shows that exaggerated advertising is ubiquitous in real life (Calfee and Ringold, 1994; Gardete, 2013; Nelson, 1974) and stores can benefit from the exaggerated behavior (Chakraborty and Harbaugh, 2014; Malar et al, 2001). In the current study, I also checked the two types of images manually and found that typically seller-posted images are better than consumer-posted images.

Besides, in the current study, I cannot simply use $(q_i^s - q_i^c)$ to measure the difference between the two types of images to capture the direction of the distance. For the extracted attributes derived from deep learning methods, they are not directly related to product quality but just a semantic feature that describes the visual content of an image by correlating low-level features such as color, gradient orientation with the content of an image scene. Therefore, it is meaningless to separately consider the attributes and use one attribute to minus another, but the distance matters.

4.2.3. Basic Image Attributes

Apart from the features extracted from the deep learning method, the basic attributes of images, e.g. entropy, cool-color ratio, lightness, contrast, and saturation etc. might also influence consumers' perceptions and then impact their shopping behaviors. To this end, I also extract these basic attributes from product photos by using well-established image-processing techniques and control them in our analyses.

Following the approach proposed by Xin Li et al. (2014), I extract the entropy, cool-color ratio, saturation, lightness, and contrast from product images.

Shannon (1948) introduced the concept of entropy, which is a measure of image information content. To calculate an image's entropy, I first convert a color image into a gray-scale image and calculate the image histogram. Then the image can be represented by a vector of gray levels. Next, I compute:

$$Entropy = -\sum_{k \in K} p_k \log_2(p_k) \quad (6)$$

, where K is the set of gray levels and p_k is the probability associated with gray level k . An image with relatively low entropy is an image with relatively low information content.

Each pixel has a color type, cool or warm. I use the percentage of cool color pixels to measure the relative coolness or warmness. To calculate the cool color ratio, I first convert each image from RGB color space to HSL (hue, saturation, lightness) color space, which is more closely align with the way human vision perceives color-making attributes. Then I have the hue of each pixel, which is a degree on the color wheel from 0 to 360 (0 is red, 120 is green, 240 is blue). Generally, yellows, oranges, and reds are considered warm, while blues, greens, and violets are considered cool. The hue values between 90 and 270 are annotated as cool colors. Then I compute:

$$cool\ ratio = \frac{the\ number\ of\ coll\ color\ pixels}{total\ number\ of\ pixels} \quad (7)$$

Saturation, lightness, and contrast are usually viewed as three classic measures to assess color-related aesthetics (Xin Li et al., 2014). Color saturation is used to describe the intensity of color in the image, where 0% means a shade of gray and 100% is the full color. Lightness is a representation of a color space's brightness, where 0% is black, 100% is white. I can directly get saturation and lightness from an image's HSL color space. Since saturation and lightness are two pixel-level variables, I calculate the average value for

each image. In contrast, it is the difference in lightness that makes an object distinguishable, which is defined by the below equation:

$$Contrast = std(\frac{L - \min(L)}{\max(L) - \min(L)}) \quad (8)$$

, where L is the matrix for lightness in the HSL color space.

I consider the above basic image features as “shallow features” since they can only capture shallow attributes of an image, and the features extracted by the deep learning method as “deep features” because they can capture the texture, fabric, shape, part, and style attributes of products.

Note that both “shallow features” and “deep features” are not directly connected to the quality of products. Therefore, it is meaningless to consider them separately in the analyses. The idea is that seller-posted images reflect consumer’s expected product quality, while buyer-posted images reflect real product quality. A larger difference between these two types of images means a larger difference between the real quality and expected quality, resulting in a high return rate. Therefore, although it is very hard to directly measure the product quality quantitatively from images, I only need to calculate the difference between the seller-posted and buyer-posted images to predict product return.

4.3. The Relationship between the Differences and Product Return

I propose that if the distance between seller-posted and buyer-posted images is large, then the product return rate will be high. To test it, I regress product return rates on the differences between features extracted from store-posted images and customer-posted images.

$$y_i = \beta_0 + \beta_1 d(q_i^s, q_i^c) + X \cdot \gamma \quad (9)$$

, where y_i represents the return rate of product i ; $d(q_i^s, q_i^c)$ denotes the distances between the attributes extracted from product i ’s images posted by stores and by customers; X is a set of covariates related to product returns, for example, product sales, number of purchase users, return freight fee, duration, customers’ ratings, number of customer reviews and so on.

Note that sales appear on both sides of the model, which seems to be inappropriate here. The reason why I used the return rate but not the return amount as the dependent variable is that the company considers the return rate as one of the key features of products but not the absolute return amount, and they aim at accurately predicting the return rate. In this case, the return rate is a more interesting variable to look at; meanwhile, sales will be definitely one of the key factors that can be used to predict return rate, so I set sales as one of the independent variables.

Before I run the regression, I do some data preprocessing. First, I delete dresses with less than 5 sales quantities; second, I cut out outliers with return rates more than 100%; third, I normalize all the independent variables to 0 mean value and 1 standard deviation for they have different units.

The result of the regression is in Table 3. The first column of Table 3 indicates that the image differences are positively correlated with product return rate, both in terms of ‘deep’ feature and ‘shallow’ features. The coefficient on the deep-difference is 0.16, significant at 1%, which means the bigger the differences between deep features extracted from store-posted images and customer-posted images, the larger the product return is. The lightness-difference, saturation-difference, contrast-difference are also positively correlated with product return rate, and the coefficients are 0.10, 0.13, 0.06 respectively. In column (2), I add sales information to the model. It shows that the coefficient of the deep difference stays still significantly positive. Besides, price is positively correlated with the return rate, which is consistent with the prediction. A higher price will increase customers’ expectations of products’ quality, which leads to a higher product return rate. Duration and return freight is positively correlated with the return rate, which is a little counterintuitive. The longer a product exists in the market, the less uncertainty in its quality will be, so the return rate should be lower. Return freight is a part of the cost of returns, so the higher the return freight is, the higher the cost of returns will be, and the return rate should also be lower. In column (3), I add both sales

information and text reviews to the model. Although the coefficient magnitude is attenuated by sales information and text reviews, it is still significantly positively correlated with the return rate. Being consistent with previous user-generated content (UGC) literature reviews volume and valence are negatively correlated with return rate with significance level at 1%, and the coefficients are -0.0647 and -0.0661. The variance of customers' reviews measures the variation of customers' opinions on products, which indicates the level of uncertainty in products' quality. Consistent with the predictions, reviews variance is positively correlated with return rate with significance level at 1%.

For the above regression, I use Euclidean Distance as the metrics of the feature difference between store-posted and customer-posted images. For robustness check, I also use Cosine Distance as the metrics, which is a more common metric to measure the similarity between two images. Table 4 shows that the deep difference is still positively correlated with the product return rate with significance level of 1% in all three models.

Table 3. The relationship between the differences and product returns (Euclidean Distance)

	(1)	(2)	(3)	(4)
Diff deep	0.1851*** (0.018)	0.1597*** (0.019)	0.101*** (0.016)	0.0612*** (0.016)
Diff entropy		-0.048** (0.021)	-0.017 (0.018)	-0.0161 (0.017)
Diff cool ratio		0.0215 (0.018)	0.0341** (0.015)	0.0323** (0.015)
Diff lightness		0.0972*** (0.021)	0.0761*** (0.017)	0.0866*** (0.017)
Diff saturation		0.1316*** (0.018)	0.0628*** (0.016)	0.0762*** (0.015)
Diff contrast		0.0608*** (0.018)	0.0307** (0.015)	0.0248 (0.015)
Sales quantity			-0.8644*** (0.097)	-0.8377*** (0.095)
Purchase user num			0.3259*** (0.097)	0.37*** (0.096)

Price			0.1314*** (0.016)	0.1685*** (0.016)
Duration			0.031** (0.016)	0.0478*** (0.015)
Return freight			0.0349** (0.015)	0.0287* (0.015)
Comment num				-0.0647*** (0.021)
Score				-0.0661** (0.028)
Variance				0.1137*** (0.028)
Constant	-2.1638*** (0.018)	-2.1638*** (0.018)	-2.1638*** (0.015)	-2.1638*** (0.015)
Adj. R-squared:	0.03	0.054	0.337	0.362
No. Observations:	3326	3326	3326	3326

Note: *p<0.1, **p<0.05, ***p<0.01, and standard error in parentheses.

Table 4. The relationship between the differences and product returns (Cosine Distance)

	(1)	(2)	(3)	(4)
Diff deep	0.1632*** (0.018)	0.139*** (0.019)	0.0909*** (0.016)	0.0506*** (0.016)
Diff entropy		-0.0491** (0.021)	-0.0173 (0.018)	-0.0165 (0.017)
Diff cool ratio		0.0291 (0.018)	0.0388** (0.015)	0.0354** (0.015)
Diff lightness		0.1028*** (0.021)	0.0789*** (0.017)	0.0892*** (0.017)
Diff saturation		0.1353*** (0.019)	0.0652*** (0.016)	0.0775*** (0.015)
Diff contrast		0.0538*** (0.019)	0.0258* (0.016)	0.0222 (0.015)
Sales quantity			-0.8667*** (0.097)	-0.8388*** (0.096)
Purchase user num			0.3252*** (0.097)	0.3706*** (0.096)
Price			0.1313*** (0.016)	0.1684*** (0.016)
Duration			0.0297* (0.016)	0.0469*** (0.015)

Return freight			0.0349**	0.0287*
			(0.015)	(0.015)
Comment num				-0.0669***
				(0.021)
Score				-0.0667**
				(0.028)
Variance				0.1148***
				(0.028)
Constant	-2.1638***	-2.1638***	-2.1638***	-2.1638***
	(0.018)	(0.018)	(0.015)	(0.015)
Adj. R-squared:	0.023	0.049	0.335	0.361
No. Observations:	3326	3326	3326	3326

Note: *p<0.1, **p<0.05, ***p<0.01, and standard error in parentheses.

5. Predictive Model: Gradient Boosted Regression Trees (GBRT)

In this section, I use a state-of-art machine learning technique – gradient boosted regression tree – to predict product return by involving image information in the model. Comparing to traditional machine learning methods (e.g. lasso, ridge, support vector machine, etc.), GBRT is good at handling mixed type (continuous and discontinuous) data, and its predictive power usually outperforms other methods. It is also robust to outliers in output space via robust loss functions.

To have a larger training dataset, I use data in June, July, and August. For the statistical validity, I select products that have more than 5 image reviews (I also try 10, and 20 as robustness checks). Following Dzyabura et al.’s paper (2018), I train the models on 80% of observations (K_{train}), and report the R-squared predictive accuracy of the remaining 20% of observations (K_{test}):

$$R_{model}^2 = 1 - \frac{\sum_{i \in K_{test}} (RR_i - \widehat{RR}_i)^2}{\sum_{i \in K_{test}} (RR_i - \overline{RR})^2} \quad (10)$$

, where RR_i is the real return rate of product i ; $\widehat{RR}_i$ is the predicted return rate of product i according to the model; $\overline{RR}$ is the mean value of the return rate of products; K_{test} is the test dataset.

Table 5 shows the results of the gradient boosted regression tree. In column (1), I only put product information as the input, and the predictive accuracy is 0.16. In column (2), I add image information into the model, and the predictive accuracy is 0.21, which is improved by 31.25%. In column (3), the model involves text review information, and the predictive accuracy is 0.52, which is improved by 147.62%. This result shows that text reviews have overwhelming predictive power than the other two types of information, which is not what I expected.

Fig 5 gives the deviance for training and test dataset as the boosting iteration increases. Fig 6 gives the importance of the input variables. The deep feature extracted from images is in the fifth place, while the comment number is in the first place, which is consistent with the result above that text reviews have overwhelming predictive power than images.

Table 5. The results of GBRT for predicting product return rate

	Variables	(1)	(2)	(3)
Product information	Price	Y	Y	Y
	Duration	Y	Y	Y
	Return freight	Y	Y	Y
Image information	Diff deep		Y	Y
	Diff entropy		Y	Y
	Diff cool ratio		Y	Y
	Diff lightness		Y	Y
	Diff saturation		Y	Y
	Diff contrast		Y	Y
Comment information	Comment num			Y
	Score			Y
	Variance			Y
R-square		0.16	0.21	0.52
Improved percentage			31.25%	147.62%

Fig 5. The deviance for training and test dataset

Fig 6. Variable importance for GBRT

6. Conclusion

This paper investigates the predictive power of the difference between store-posted and customer-posted images on product return rates. To quantify the difference, I use both well-established image-processing techniques and deep learning methods to extract image features. Specifically, I first extract basic image attributes, such as entropy, cool color ratio, lightness, and saturation, which are defined as “shallow features”. Then, I use a residual learning framework, ResNet-50, combining with a special neural network structure, Siamese neural network, to recognize images, and choose the second-to-last layer of the network, containing 512 features, as “deep features”. After that, I calculate the Euclidean and Cosine distance between seller-posted images and buyer-posted images using the extracted features, and I employ a state-of-art machine learning algorithm – gradient boosted regression tree – to predict product return rate. The results show that the estimated differences between the two types of images have statistically significant predictive power for product returns by increasing the prediction accuracy by 31.25%.

I acknowledge some limitations of this paper. First, I only investigate the correlation between the distance and product returns, but not the causal effect. Second, I discuss the problem at the product level, but not the individual level. Third, I only examine the predictive power of images in the dress category, I will develop an AI system that is suitable for all the product categories. In future studies, I will improve the work from these three aspects.

References

- Anderson, E. T., Hansen, K., & Simester, D. (2009). The option value of returns: Theory and empirical evidence. *Marketing Science*, 28(3), 405-423.

- Anderson, E. T., Hansen, K., Simester, D., & Wang, L. K. (2008). How are demand and returns related. *Theory and empirical evidence*, 11, 70.
- Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, 53(3), 297-318.
- Bechwati, N. N., & Siegal, W. S. (2005). The impact of the prechoice process on product returns. *Journal of Marketing Research*, 42(3), 358-367.
- Bonifield, C., Cole, C., & Schultz, R. L. (2010). Product returns on the internet: a case of mixed signals? *Journal of Business Research*, 63(9-10), 1058-1065.
- Bower, A. B., & Maxham III, J. G. (2012). Return shipping policies of online retailers: normative assumptions and the long-term consequences of fee and free returns. *Journal of Marketing*, 76(5), 110-124.
- Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345-354.
- Davis, S., Hagerty, M., & Gerstner, E. (1998). Return policies and the optimal level of “hassle”. *Journal of Economics and Business*, 50(5), 445-460.
- Dzyabura, D., El Kihal, S., & Ibragimov, M. (2018). Leveraging the power of images in predicting product return rates.
- Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3), 291-313.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Hess, J. D., Chu, W., & Gerstner, E. (1996). Controlling product returns in direct marketing. *Marketing Letters*, 7(4), 307-317.

- Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.
- Janakiraman, N., & Ordóñez, L. (2012). Effect of effort and deadlines on consumer product returns. *Journal of Consumer Psychology*, 22(2), 260-271.
- Keller, K. L., & Lehmann, D. R. (2006). Brands and branding: Research findings and future priorities. *Marketing Science*, 25(6), 740-759.
- King, T., & Dennis, C. (2003). Interviews of deshopping behaviour: an analysis of theory of planned behaviour. *International Journal of Retail & Distribution Management*, 31(3), 153-163.
- Koch, G., Zemel, R., & Salakhutdinov, R. (2015). *Siamese neural networks for one-shot image recognition*. Paper presented at the ICML Deep Learning Workshop.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Paper presented at the Advances in neural information processing systems.
- Li, J., He, J., & Zhu, Y. (2018). *E-tail product return prediction via hypergraph-based local graph cut*. Paper presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.
- Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4), 456-474.
- Li, X., Wang, M., & Chen, Y. (2014). *The Impact of Product Photo on Online Consumer Purchase Intention: an Image-Processing Enabled Empirical Study*. Paper presented at the PACIS.
- Liu, L., Dzyabura, D., & Mizik, N. (2018). *Visual listening in: Extracting brand image portrayed on social media*. Paper presented at the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). *Ssd: Single shot multibox detector*. Paper presented at the European conference on computer vision.
- Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science*, 35(3), 363-388.
- Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3), 74-89.
- Ma, J., & Kim, H. M. (2016). Predictive model selection for forecasting product returns. *Journal of Mechanical Design*, 138(5).
- Minnema, A., Bijmolt, T. H., Gensler, S., & Wiesel, T. (2016). To keep or not to keep: Effects of online customer reviews on product returns. *Journal of retailing*, 92(3), 253-267.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521-543.
- Petersen, J. A., & Kumar, V. (2009). Are product returns a necessary evil? Antecedents and consequences. *Journal of Marketing*, 73(3), 35-51.
- Retail. (2018). Consumer Returns in the Retail Industry.
- Sahoo, N., Dellarocas, C., & Srinivasan, S. (2018). The impact of online product reviews on product returns. *Information Systems Research*, 29(3), 723-738.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3), 379-423.
- Tang, C., & Guo, L. (2015). Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (eWOM) communication. *Marketing Letters*, 26(1), 67-80.
- Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31(2), 198-215.

- Toktay, L. B., van der Laan, E. A., & de Brito, M. P. (2004). Managing product returns: the role of forecasting. In *Reverse Logistics* (pp. 45-64): Springer.
- Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of Marketing*, 73(5), 90-102.
- Urbanke, P., Kranz, J., & Kolbe, L. (2015). Predicting product returns in e-commerce: the contribution of mahalanobis feature extraction.
- Wood, S. L. (2001). Remote purchase environments: The influence of return policy leniency on two-stage decision processes. *Journal of Marketing Research*, 38(2), 157-169.
- Zhang, S., Mehta, N., Singh, P. V., & Srinivasan, K. (2019). Can Lower-quality Images Lead to Greater Demand on AirBnB? .
- Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2), 133-148.
- Zhu, Y., Li, J., He, J., Quanz, B. L., & Deshpande, A. A. (2018). *A Local Algorithm for Product Return Prediction in E-Commerce*. Paper presented at the IJCAI.

Essay 3: Partial Identification of Spillovers in E-Marketplaces Utilizing Consumer Clickstream Data

Abstract

Two opposite spillovers simultaneously exist among firms in marketplaces. One is the positive spillover effect, which is captured by demand-side externalities, making firms tend to cluster together. The other one is the negative business-stealing effect because of the presence of competitors, hindering firms from co-locating in the same marketplace. Separate identification of these two effects empirically using observational data requires complete footprints of an individual's shopping journey. In the brick-and-mortar space, such footprints are not easily traceable. Hence, the extant literature primarily focuses on the net effect of agglomeration. This paper develops an empirical framework to separately identify these two effects under the context of a cashback website. Specifically, I argue that the positive spillover effect is pointed identified, while the negative business-stealing effect is partially identified, or its bounds are identified. Then I non-parametrically estimate these two effects and explore their heterogeneities in different types of merchants. Besides, I examine how cashback rates will impact them separately. I find that cashback rates are positively related to the spillover effect and negatively related to the business-stealing effect. The effects of cashback rates are heterogenous in different merchant categories.

Keywords: Spillovers; business-stealing effect; cashback websites; clickstream data.

1. Introduction

Various stores and retailers cluster together in marketplaces providing complementary or substitutive products and services. On one hand, retailers could be beneficial to each other because of demand-side externalities. For

instance, the reputation of competitors may attract more consumers to the marketplace (Chung & Kalnins, 2001), or the variety of products may make consumers more likely to find their optimal products (Schuetz, 2015; Stephen & Toubia, 2010), or the intensified competition indicates a credible commitment to low prices (Dudey, 1990). On the other hand, retailers' revenue may be harmed by others due to the presence of competitors and substitute products. Existing literature has shown the benefits of agglomeration by testing merchant's entry or expansion decisions (Myles Shaver & Flyer, 2000; Schuetz, 2015; Shen & Xiao, 2014; Vitorino, 2012). A question is naturally raised. How exactly will a merchant's benefits be influenced by others after they affiliate in a marketplace? This paper empirically examines the positive spillover and the negative business-stealing effects separately among merchants in marketplaces and explores what merchants can do to increase their benefits in various merchant categories.

Separate identification of the positive and negative spillovers empirically using observational data requires complete footprints of an individual's shopping journey. In the brick-and-mortar space, such footprints are not easily traceable. Hence, the extant literature primarily focuses on the net effect of agglomeration (e.g. Shen & Xiao, 2014). Vitorino (2012) shows that firms tend to cluster together despite the business-stealing effects. Some researchers claim that firms can benefit from agglomeration because their neighbors can be complementary to them in terms of product categories, price, or quality, etc. (Clapp, Ross, & Zhou, 2019; Picone, Ridley, & Zandbergen, 2009; Schuetz, 2015). Other researchers argue that even firms cluster to their competitors with similar products they can also benefit from others by demand expansion and market learning effect (Li, Shen, & Bart, 2018; Shen & Xiao, 2014; Toivanen & Waterson, 2005; Yang, Lu, & Lu, 2014).

The aforementioned empirical studies used aggregate level data and investigated the net effect of agglomeration. It was infeasible to separately examine the positive spillover and the negative business-stealing effects among

merchants when only aggregate level data is available. Fortunately, as the development of e-commerce and big data techniques, it allows researchers to track and record consumers' online clickstream among e-retailers. This research develops an empirical framework to look into the spillovers among merchants in an e-marketplace using consumer clickstream data.

The context of this research is a regional cashback e-commerce platform based in Asia. Cashback shopping, as a novel shopping mode, is becoming increasingly popular worldwide. It allows consumers to reimburse a certain portion of the payments if they initiate the transaction through the websites of cashback companies. There are many successful well-known cashback companies, like Ebates.com and Fatwallet.com in the United States, Quidco.com and TopCashBack.com in the United Kingdom, and Fanli.com in China.

To separately identify the positive spillover and the negative business-stealing effects, I classify the traffic into four types: initial traffic, indicating the consumers who initially visit the merchant when they land in the site; spillover traffic, indicating the consumers who move to the merchant but initially visit another merchant; diverted traffic, indicating the consumers who initially visit the merchant but later move to other merchants; return traffic, indicating the consumers who initially visit the merchant, and then move to other merchants, but finally return to the merchant. By looking into these four traffic types and according purchasing probabilities, I argue that the spillover effect is pointed identified, while the business-stealing effect is partially identified, or its bounds are identified. I then use the bootstrap method to non-parametrically estimate these two effects and explore their heterogeneities in two different scenarios: high vs. low market-share merchants, merchants in the "General Marketplace" category vs. other categories.

Furthermore, I explore what merchants can do to increase their benefits or sales. Given that the context of this research is a regional cashback e-commerce

platform, I investigate the effect of cashback rates on these two effects. I also investigate the category and weekend interaction effect of cashback rates. The results show that cashback rates are positively related to the spillover effect and negatively related to the business-stealing effect. Cashback rates have a larger impact on the spillover effect in categories of “Fashion” and “Home & Lifestyle” and have a larger impact on the negative business-stealing effect in categories of “Dining & Food Delivery”, “General Marketplace”, and “Travel”.

This research provides useful managerial implications for both cashback websites and affiliated merchants. First, I show that overall, merchants can benefit from other merchants’ traffic in a marketplace. Second, I separately identify the positive spillover and the negative business-stealing effects and non-parametrically estimate them. Third, I explore the heterogeneities of these two effects in different merchant types. Fourth, I derive the average partial effects of cashback rates on these two effects in each category, so I can give specific instructions to the cashback platform on how to increase merchants’ benefits by manipulating cashback rates. Besides, the theoretical framework and the empirical models in the current research could be extended to any other online marketplaces.

The rest of the paper is organized as follows. In Section 2, I will review the literature related to firm agglomeration and cashback websites. In Section 3, I will introduce the conceptual framework and identification strategy. In Section 4, I describe the data and conduct a preliminary analysis. In Section 5, I show the estimation results of the spillover and the business-stealing effects and investigate the effect of cashback rates on them. Finally, I will conclude the paper in Section 6.

2. Literature Review

This research draws on and contributes to two streams of literature: firms’ agglomeration and cashback platform.

Firms' agglomeration has been extensively studied in previous literature (Clapp et al., 2019; Glaeser, Kallal, Scheinkman, & Shleifer, 1992; Hausman & Leibtag, 2007; Myles Shaver & Flyer, 2000; Quigley, 1998; Schuetz, 2015). It is first mentioned in the Hotelling model (1929) that firms tend to locate together. The model indicates that if there are two competing firms located along a unit line, and consumers are uniformly distributed on the line, then the optimal locating strategy for two firms is locating at the same halfway point. The situation in the Hotelling model is simplified with two firms selling similar products in a linear city. In more complex situations, the following studies showed that firm agglomeration intensively exists in the real world. By using the store configurations of all U.S. regional shopping centers, Vitorino (2012) shows that firms have incentives to collocate even when potential business-stealing effects exist. Toivanen and Waterson (2005) find that the probability of entry will increase with the presence of competitors by using McDonald's and Burger King's store-opening data in the United Kingdom. Similarly, Shen and Xiao (2014) investigate the reason why McDonald's and KFC are usually located together in China by studying the entry and expansion decisions of two firms.

From theoretical perspectives, Koçak and Özcan (2013) summarize four mechanisms explaining why firms are usually co-located with competitors. (1) From strategy perspectives, firms tend to enter each other's markets to retaliate and deter their competitors (Baum & Korn, 1999; Fuentelsaz & Gómez, 2006; Gimeno, 1999). (2) According to theories of agglomeration spillovers, on the manufacturing side, firms can get better availability of suppliers, labors, and knowledge by learning from each other (Marshall, 1890); on the retailer side, firms can share better environments and get greater consumer traffics by market expansion effect (Li et al., 2018). (3) According to the density dependence theory, it's inefficient for firms to attract resources when the density of firms in a market is low, so firms tend to cluster to gain more density (Carroll & Hannan, 2000). (4) According to the new institutional theory, firms need to

emulate competitors' entry decisions when facing uncertainty in the market (Bastos & Greve, 2003). This research will focus on the second mechanism, agglomeration spillovers, in the context of the cashback website.

The agglomeration effect among firms happens on both supply and demand sides. On the supply side, or the manufacturing side, firms can learn from nearby firms on how to improve product qualities and increase productivities (Glaeser et al., 1992). The agglomeration effect on the supply side has been empirically tested by many studies. Furman, Kyle, Cockburn, and Henderson (2006) find that knowledge spillovers significantly exist among pharmaceutical firms by using their research activities' geographic distribution. Lu and Shin (2018) show that firms can benefit from their competitors by sharing innovation using an analytical model. Besides, agglomerated manufacturers can also benefit from sharing pooled laborers and other resources (Marshall, 1920). On the demand side, or the retailer side, clustering firms can attract more consumers by providing various products so that consumers will be more likely to find their optimal product characteristics (Konishi, 2005), and by providing a credible commitment to low prices (Dudey, 1990). More studies deeply investigate why retailers tend to cluster together.

Some researchers claim that firms can benefit from agglomeration because their neighbors can be served as complementary functions. Schuetz (2015) points out that if firms cannot differentiate their products, and if consumers can compare prices, then firms will locate far away from their competitors instead of clustering together. He finds that new Big Box stores tend to locate near other Big Box stores that sell products in other sectors. Picone et al. (2009) also show that retailers are more likely to strategically cluster if they have a greater ability to differentiate their products. Besides, if firms are complementary in prices, they also tend to cluster. Clapp et al. (2019) look at the opening and closing probabilities of department stores and find strong negative competitive effects from the firms at the same price level, but they do not find effects from

other price level firms. These studies show that firms tend to cluster with their complimentary firms instead of substitutive firms.

Other researchers argue that even firms cluster with their competitors with similar products they can also benefit from others. Shen and Xiao (2014) investigate the entry and expansion decisions of McDonald's and KFC in China and find that KFC can benefit from MacDonald's by demand expansion effect while KFC has a positive effect on MacDonald's because of market learning effect. Toivanen and Waterson (2005) also show that firms have incentives to enter the market with rivals because of the expected increasing market size. Li et al. (2018) provide evidence that in a two-sided market the more merchants using the platform in the past, the more firms will enter the platform in the future, which reflects the observational learning effect.

To my best knowledge, although many studies empirically test the agglomeration effect by examining firm's entry and expansion decisions, merely no one investigates how will firms influence each other after they join the cluster, and no one directly tests the underlying mechanism that firms can bring more traffic to each other, because it is hard to collect the data on consumer traffics in physical stores. Fortunately, by using consumers' online clickstream data, I can track consumer traffic flows between each merchant on the cashback platform. The current research provides an empirical framework to examine the positive spillover effect and negative business-stealing effect among stores.

For online shops, the agglomeration effect also exists among retailers if they are displayed together on the same platform. What is more, because it is easier for consumers to transfer from one store to another, the effect should be more significant. Yang et al. (2014) look at the paid-search advertising market and find a significantly positive effect of the number of entrants on the basic click volume. Stephen and Toubia (2010) also proved that by allowing sellers to be connected, they can generate more value because it makes shops more

accessible to customers. On one hand, affiliated merchants generate traffic spillovers to each other within and across categories; on the other hand, I must admit that more merchants mean more intensified competition. How much traffic each store gains from other stores and losses due to the competition might depend on its identity and the interaction with the other stores. Chu and Manchanda (2016) investigate a large online C2C platform, Alibaba Group's Taobao.com, and show that there is no direct positive network effect on both buyers' and sellers' growth, but a positive cross-network effect on both sides of the platform.

This research also contributes to the growing literature on the cashback platform. Cashback shopping, as a novel online shopping mode, allows consumers to reimburse a certain portion of the payments if they initiate the transaction through the websites of cashback companies. Its capability of bringing more sales to online merchants has been studied in previous literature. By analyzing panel data from a large cashback company, Vana et al. (2018) explore the effect of the reimbursed money from the platform on future sales, and they find that it will induce consumers to make more consumptions via the cashback website and increase the transaction size in the future because they usually fail to treat cashback payments as a fungible resource. Zhou et al. (2017) indicate that the nature of cashback websites is to allow online merchants to exercise third-degree price discrimination. They investigate the value of cashback websites in both centralized and decentralized settings and find that they can always bring more profits to online merchants because of the price discrimination function. Ho et al. (2017) also suggest that especially for some undistinguished merchants, it is beneficial for them to offer cashback deals to price-sensitive consumers. These studies have shown that it is more profitable for merchants with the presence of cashback websites comparing to without it. This research provides additional insights that merchants in the cashback platform could benefit from others' traffic.

In summary, the main contribution of the paper can be summarized below:

- (1) For the literature related to spillovers among retailers, I directly test the underlying mechanism that firms can benefit from others' traffic, and I separately identify the positive spillover and the negative business-stealing effects using consumer online clickstream data.
- (2) I explore the heterogeneities of these two effects in different merchant types.
- (3) I further investigate the direct effect of cashback rates and the interaction effect with categories and weekends on these two effects.
- (4) For the cashback literature, I provide additional insights into online merchants' profitability in the presence of cashback websites. Apart from its price discrimination function and reimbursed money as a windfall that will induce more consumption in the future, I propose that merchants in the cashback platform as an agglomeration can bring more consumer traffic to each other and then induce more demands.

3. Conceptual Framework and Identification

3.1. Conceptual Framework

Consider the case of two merchants, M_1 , M_2 operating in an online marketplace. The consumer shopping process is framed as follows. First, consumers visit M_1 , M_2 with probabilities θ_1 , θ_2 respectively, or other options (e.g. visit neither of them, exit the site) with probability $1 - \theta_1 - \theta_2$. Second, if consumers visit M_1 , they can choose to stay at M_1 with probability γ_{11} , or move to M_2 with probability $1 - \gamma_{11}$. Third, if they stay at M_1 , they buy with probability δ_{11} or exit with probability $1 - \delta_{11}$. Alternatively, if they move to M_2 , they have two options, stay at M_2 with probability $1 - \gamma_{121}$, or return to M_1 with probability γ_{121} . Fourth, if they stay at M_2 , they buy with probability δ_{12} or exit with probability $1 - \delta_{12}$. Alternatively, if they go back to M_1 , they buy from M_1 with probability δ_{121} or exit with probability $1 - \delta_{121}$. The whole process is depicted in Figure 1.

Figure 1. The footprint of a shopping journey.

The above process has exhausted all the possibilities given that I only allow one purchase in a session. For example, if a consumer first visit M_1 , then six independent events might happen. First, she purchases from M_1 without visiting any other merchants (A). Second, she does purchase from M_1 and exit without further browsing (B). Third, she visits M_1 then moves to M_2 and purchases from M_2 (C). Fourth, she visits M_1 then moves to M_2 without purchasing and exit (D). Fifth, she visits M_1 , moves to M_2 , returns to M_1 and purchases from M_1 (E). Sixth, she visits M_1 , moves to M_2 , and returns to M_1 without purchasing and exit (F). The according probability of each event is given by:

$$A=(M_1, \text{Buy}) = \theta_1 \gamma_{11} \delta_{11}$$

$$B=(M_1, \text{No buy or exit}) = \theta_1 \gamma_{11} (1 - \delta_{11})$$

$$C=(M_1, \text{Move to } M_2, \text{No return to } M_1, \text{Buy}) = \theta_1 (1 - \gamma_{11}) (1 - \gamma_{121}) \delta_{12}$$

$$D=(M_1, \text{Move to } M_2, \text{No return to } M_1, \text{No buy or exit}) = \theta_1 (1 - \gamma_{11}) (1 - \gamma_{121}) (1 - \delta_{12})$$

$$E=(M_1, \text{Move to } M_2, \text{Return to } M_1, \text{Buy}) = \theta_1 (1 - \gamma_{11}) \gamma_{121} \delta_{121}$$

$$F=(M_1, \text{Move to } M_2, \text{Return to } M_1, \text{No buy or exit}) = \theta_1 (1 - \gamma_{11}) \gamma_{121} (1 - \delta_{121})$$

According to the process described above, the probability of buying at M_1 can be written as:

$$P(Buy_1) = \theta_1 \gamma_{11} \delta_{11} + \theta_2 (1 - \gamma_{22}) (1 - \gamma_{212}) \delta_{21} + \theta_1 (1 - \gamma_{11}) \gamma_{121} \delta_{121} \quad (1)$$

The first item on the right side of Equation (1) indicates the buying probability of consumers who initially visit M_1 and stay at M_1 ; the second item is the buying probability of consumers who initially visit M_2 but move to M_1 and stay at M_1 ; the third item describes the buying probability of consumers who initially visit M_1 , then move to M_2 but finally go back to M_1 .

Similarly, the probability of buying at M_2 can be written as:

$$P(Buy_2) = \theta_2 \gamma_{22} \delta_{22} + \theta_1 (1 - \gamma_{11}) (1 - \gamma_{121}) \delta_{12} + \theta_2 (1 - \gamma_{22}) \gamma_{212} \delta_{212} \quad (2)$$

Then I extend the framework to the case with M merchants affiliating in a cashback website. For Merchant m , its buying probability is:

$$P(Buy_m) = \theta_m \gamma_{mm} \delta_{mm} + \sum_{m'} \theta_{m'} \gamma_{m'm} (1 - \gamma_{m'mm'}) \delta_{m'm} + \sum_{m'} \theta_m \gamma_{mm'} \gamma_{mm'm} \delta_{mm'm} \quad (3)$$

, where θ_m is the probability that consumers initially visit m ; γ_{mm} is the probability that consumers will stay at m ; $\theta_{m'} \gamma_{m'm}$ is the probability that consumers initially visit other merchants m' but move to m , and $\delta_{m'm}$ is the corresponding buying probability; $\theta_m \gamma_{mm'} \gamma_{mm'm}$ is the probability that consumers initially visit m , then move to other merchants m' , but finally return to m , and $\delta_{mm'm}$ is the corresponding buying probability.

To separately identify and estimate the positive spillover and the negative business-stealing effects, for each merchant m , I classify consumer traffic into four types: initial traffic, indicating the consumers who initially visit merchant m when they land in the site; spillover traffic, indicating the consumers who move to merchant m but initially visit other merchants m' ; diverted traffic, indicating the consumers who initially visit merchant m but then browse out; return traffic, indicating the consumers who initially visit merchant m , and then move to other merchants, but finally return to merchant m . Besides, I classify

sales into four types: initial sales, indicating the purchases from the initial traffic; spillover sales, indicating the purchases from the spillover traffic; diverted sales, indicating the purchases at other merchants due to the diverted traffic; return sales, indicating the purchases from the return traffic. In the next two sections, I will separately identify the positive spillover and the negative business-stealing effects.

3.2. Identification of the Positive Spillover Effect

I define the positive spillover effect on merchant m as the increased sales due to consumers move from other merchants to merchant m . To identify the spillover effect, I investigate the effect of a spillover traffic to merchant m on m 's sales.

I introduce a few notations. Let $Y_{imp} = 1$ if an individual i chooses to purchase from merchant m and $Y_{imp} = 0$ otherwise. Let $Y_{im'm} = 1$ if an individual i chooses to move from other merchants m' to merchant m and otherwise $Y_{im'm} = 0$.

Let the set of individuals who move from other merchants m' to merchant m and stay at merchant m be $J_{m'm}$, and the set of individuals who do not move to merchant m be $J_{m'm'}$. Then the average treatment effect on the treated group (ATET) of a spillover traffic is represented by:

$$E[\delta_s^m | i \in J_{m'm}] = E[Y_{imp} = 1 | Y_{im'm} = 1, i \in J_{m'm}] - E[Y_{imp} = 1 | Y_{im'm} = 0, i \in J_{m'm}] \quad (4)$$

The first term represents the expectation of purchasing from merchant m if an individual i moves from other merchants m' to merchant m ; the second term is the counterfactual component indicating the expectation of purchasing from merchant m if an individual i moves from other merchants m' to merchant m had instead not moved to merchant m . The ATET of a spillover traffic is given by the difference in the purchase expectations between these two groups.

Normally, for $i \in \mathcal{J}_{m'm}$, we never observe $E[Y_{imp} = 1 | Y_{im'm} = 0, i \in \mathcal{J}_{m'm}]$. However, if an individual does not visit merchant m , then her probability of purchasing from merchant m must be zero. Then, the value of the counterfactual part becomes zero and the ATET of a spillover traffic is reduced to:

$$E[\delta_s^m | i \in \mathcal{J}_{m'm}] = E[Y_{imp} = 1 | Y_{im'm} = 1, i \in \mathcal{J}_{m'm}] \quad (5)$$

Note that the expected value of a binary indicator is simply the probability of the indicator taking the value 1. Then the positive spillover effect is given by:

$$\begin{aligned} & \sum_{m'} P[Y_{im'}] \cdot P[i \in \mathcal{J}_{m'm}] \cdot E[Y_{imp} = 1 | Y_{im'm} = 1, i \in \mathcal{J}_{m'm}] \\ &= \sum_{m'} P[Y_{im'}] \cdot P[i \in \mathcal{J}_{m'm}] \cdot P[Y_{imp} = 1 | Y_{im'm} = 1, i \in \mathcal{J}_{m'm}] \\ &= \sum_{m'} \theta_{m'} \gamma_{m'm} (1 - \gamma_{m'mm'}) \delta_{m'm} \quad (6) \end{aligned}$$

, where $P[Y_{im'}]$ or $\theta_{m'}$ is the probability that consumers initially visit other merchants m' ; $P[i \in \mathcal{J}_{m'm}]$ is the probability that consumers move to merchant m and stay at m ; $\gamma_{m'm}$ is the probability that consumers move to m from other merchants m' ; $(1 - \gamma_{m'mm'})$ is the probability that consumers stay at m ; and $\delta_{m'm}$ is the corresponding buying probability. Because each parameter in Equation (6) can be observed, the positive spillover effect is pointed identified.

3.3. Partial Identification of the Negative Business-Stealing Effect

I define the negative business-stealing effect as the loss of sales due to the consumers who move out from merchant m to other merchants. To identify the business-stealing effect, I investigate the effect of a diverted traffic to merchant m on m 's sales.

Similarly, I introduce a few notations. Let $Y_{imp} = 1$ if an individual i chooses to purchase from merchant m and $Y_{imp} = 0$ otherwise. Let $Y_{imm'} = 1$

if an individual i chooses to move from merchant m to other merchants m' and otherwise $Y_{imm'} = 0$.

Let the set of individuals who move from merchant m to other merchants m' and stay at merchant m' be $J_{mm'}$, and the set of individuals who do not move to another merchant m' be J_{mm} . Then the average treatment effect on the treated group (ATET) of a diverted traffic is represented by:

$$E[\delta_s^m | i \in J_{mm'}] = E[Y_{imp} = 1 | Y_{imm'} = 1, i \in J_{mm'}] - E[Y_{imp} = 1 | Y_{imm'} = 0, i \in J_{mm'}] \quad (7)$$

The first term represents the expectation of purchasing from merchant m if an individual i moves from merchant m to other merchants m' ; the second term is the counterfactual component indicating the expectation of purchasing from merchant m if an individual i moves from merchant m to other merchants m' had instead not moved to other merchants m' . The ATET of a diverted traffic is given by the difference of the purchase expectations between these two groups.

For $i \in J_{mm'}$, we never observe $E[Y_{imp} = 1 | Y_{imm'} = 0, i \in J_{mm'}]$.

Therefore, the ATET of a diverted traffic and the business-stealing effect is not point identified. The condition under which the above effect is identified is if the choice of moving to other merchants m' is completely random or the choice is completely determined by the observed characteristics of the individuals. That is, if the condition “selection on observables” or “unconfoundedness” holds. This is a strong assumption that may not be tenable in our context. While the observable individual characteristics may determine to some extent the choice probability, it is fathomable that an individual self-selects into choosing whether to move to other merchants m' depending on her unobserved taste and preference.

No assumption bounds: Although the ATET of a diverted traffic is not point identified, it is bounded above and below since Y_{imp} is a binary outcome that can only take values 0 and 1. Besides, for an individual i , the probability of

her purchasing from merchant m is zero conditionally on that she moves to other merchants m' , or $E[Y_{imp} = 1 | Y_{imm'} = 1, i \in \mathcal{J}_{mm'}] = 0$. Then without further assumption, I have the following:

$$E[\delta_s^m | i \in \mathcal{J}_{mm'}] \in (-1, 0) \quad (8)$$

Bounds with weak assumptions: In a series of papers Manski (1990, 1997, 2003) showed that under monotone treatment response (MTR) and monotone treatment selection (MTS) assumptions the bounds can be tighter than the one in (8). These assumptions are weaker than the ‘unconfoundedness’ assumption typically invoked in applying many other popular frameworks to estimate treatment effect (such as the propensity score method). The MTR assumption suggests that the treatment effect (depending on the context) cannot be negative. In our context, this assumption suggests that the ATET of a diverted traffic cannot be positive and it is naturally satisfied. Violation of this assumption would essentially mean that an individual moving to other merchants m' will increase the probability of purchasing from merchant m . This does not seem plausible in our context.

I also posit that the MTS holds in our context. This assumption suggests that those who opt for the treatment generally have higher average outcomes. For example, in Manski and Pepper’s paper (2000), they examined the effect of education on earnings, tightened the bound of treatment effect by assuming that average earnings of the high education group would be higher than the average earnings of the low education group had they chosen to go for higher education. In our context similar assumption seems plausible. Invoking this assumption suggests that for individuals who move to other merchants m' , the probability of purchasing from other merchants m' is larger than the probability of purchasing from merchant m had they stayed at merchant m . This leads to the following:

$$E[Y_{imp} = 1 | Y_{imm'} = 0, i \in \mathcal{J}_{mm'}] \leq E[Y_{im'p} = 1 | Y_{imm'} = 1, i \in \mathcal{J}_{mm'}] \quad (9)$$

, where $Y_{im'p} = 1$ indicates that individual i purchases from other merchants m' .

Therefore, with MTR and MTS assumptions the bounds of the ATET of a diverted traffic are:

$$E[\delta_s^m | i \in J_{mm'}] \in (-E[Y_{im'p} = 1 | Y_{imm'} = 1, i \in J_{mm'}], 0) \quad (10)$$

Note that the expected value of a binary indicator is simply the probability of the indicator taking the value 1. Then the bounds of the business-stealing effect are given by:

$$\begin{aligned} & (0, \sum_{m'} P[i \in J_{mm'}] \cdot E[Y_{im'p} = 1 | Y_{imm'} = 1, i \in J_{mm'}]) \\ &= (0, \sum_{m'} P[i \in J_{mm'}] \cdot P[Y_{im'p} = 1 | Y_{imm'} = 1, i \in J_{mm'}]) \\ &= (0, \sum_{m'} \gamma_{mm'} \cdot (1 - \gamma_{mm'm}) \cdot \delta_{mm'}) \quad (11) \end{aligned}$$

, where $P[i \in J_{mm'}]$ is the probability that consumers move to other merchants m' from merchant m and stay at m' ; $\gamma_{mm'}$ is the probability that consumers move to other merchants m' from merchant m ; $(1 - \gamma_{mm'm})$ is the probability that consumers stay at other merchants m' ; $\delta_{mm'}$ is the probability of purchasing from other merchants m' given that consumers move from m . Because each parameter in Equation (11) can be observed, the business-stealing effect is partially identified.

To sum up, the positive spillover effect is pointed identified by $\sum_{m'} \theta_{m'} \cdot \gamma_{m'm} \cdot (1 - \gamma_{m'mm'}) \cdot \delta_{m'm}$; the negative business-stealing effect is partially identified by $(0, \sum_{m'} \gamma_{mm'} \cdot (1 - \gamma_{mm'm}) \cdot \delta_{mm'})$.

3.4. Non-Parametric Estimation of the Positive Spillover and the Negative Business-Stealing Effects

The parameters in the spillover and the business-stealing effects can be directly calculated using observed traffic and sales data. To estimate these two effects, I introduce a few notations. For merchant m , let T_{m1} denote its initial

traffic; let T_{m2} denote its spillover traffic from other merchants m' ; let T_{m3} denote its diverted traffic; let T_{m4} denote its return traffic; let S_{m1} denote its initial sales indicating the purchases from the initial traffic; let S_{m2} denote spillover sales, indicating the purchases from the spillover traffic; let S_{m3} denote diverted sales, indicating the purchases at other merchants due to the diverted traffic; let S_{m4} denote return sales, indicating the purchases from the return traffic. Then I have:

$$\begin{aligned}
T_{m1} &= N \cdot \theta_m \\
T_{m2} &= N \cdot \sum_{m'} \theta_{m'} \cdot \gamma_{m'm} \cdot (1 - \gamma_{m'mm'}), \text{ for } m' \neq m \\
T_{m3} &= N \cdot \sum_{m'} \theta_m \cdot \gamma_{mm'}, \text{ for } m' \neq m \\
T_{m4} &= N \cdot \sum_{m'} \theta_m \cdot \gamma_{mm'} \gamma_{mm'm}, \text{ for } m' \neq m \\
N &= \sum_m T_{m1}, \forall m
\end{aligned} \tag{12}$$

Suppose the buying probability $\delta_{m'm}$ obeys a normal distribution with the mean value $\overline{\delta_{m'm}}$ and the buying probability $\delta_{mm'}$ obeys a normal distribution with the mean value $\overline{\delta_{mm'}}$. Then the expectation of the positive spillover effect and the expectation of the upper bound of the negative business-stealing effect are given by:

$$E_m[\text{spillover}] = \sum_{m'} \theta_{m'} \cdot \gamma_{m'm} \cdot (1 - \gamma_{m'mm'}) \cdot \overline{\delta_{m'm}} = T_{m2} \cdot \overline{\delta_{m'm}} / N. \tag{13}$$

$$\begin{aligned}
E_m[\text{stealing}] &= \sum_{m'} \gamma_{mm'} \cdot (1 - \gamma_{mm'm}) \cdot \overline{\delta_{mm'}} \\
&= (T_{m3} - T_{m4}) \cdot \overline{\delta_{mm'}} / (N\theta_m) \tag{14}
\end{aligned}$$

According to the law of large numbers (LLN), the average of the results obtained from a large number of samples should be close to the expected value. In our context, $\overline{\delta_{m'm}} \approx S_{m2}/T_{m2}$, and $\overline{\delta_{mm'}} \approx S_{m3}/(T_{m3} - T_{m4})$.

Furthermore, I use the bootstrap method to get the mean value and standard error of the positive spillover effect and the upper bound of the negative business-stealing effect for each merchant and each day.

4. Data and Preliminary Analyses

4.1. Data Description and Processing

The data I use for this research comes from a regional cashback e-commerce platform based in Asia. The data spans from Apr 2019 to Jun 2019 with 6,957,634 observations. Each observation records a customer's visiting history, including account id, click time, merchant name, merchant category, whether to make a purchase, purchase amount, cashback amount, etc.

To utilize the consumers' clickstream data, I define a session as follows. For each consumer, once she makes a purchase, I consider it as the end of the session; then for her next click, I consider it as the beginning of a new session. Besides, following the clickstream literature (Montgomery et al, 2004), if consumers do not have any activities in a certain time period, then the session is considered to be ended, and the next viewing marks the beginning of a new session. In the current research, I use the 90 percentiles of the gap between the visit time and purchase time, which is 489 minutes, as the threshold. If consumers do not have any activities in 489 minutes, then I consider it is the end of the session.

To examine the positive spillover and the negative business-stealing effects separately, I classify consumer traffic into four types: initial traffic, indicating the consumers who initially visit the merchant when they land in the site; spillover traffic, indicating the consumers who move to the merchant but initially visit another merchant; diverted traffic, indicating the consumers that initially visit the merchant but later move to other merchants; return traffic, indicating the consumers who initially visit the merchant, and then move to

other merchants, but finally return to the merchant. In a session, I ignore the intermediate merchants that the consumer visits and only consider the first and the last merchants since I assume that purchases can only happen at the first or the last merchants according to the definition of a session. Table 1 gives an example of consumers' clickstream data and four types of traffics. For Session 1, the consumer first visits M1, and then move to M2, and then move to M3, so M1 has 1 initial traffic and 1 diverted traffic, and M3 has 1 spillover traffic.

From the definition of four types of traffic, we see that the initial traffic is important to merchants since I treat it as a kind of endowment to merchants. If a portion of the initial traffic moves to other merchants, then it is considered as a loss to that merchant. However, it is possible that consumers just randomly click the first merchant. In that case, we cannot say that it is a loss to that merchant if consumers move to others. To rule out this possibility, I examine the difference between the purchase probability of the initial traffic and that of the indirect traffic (i.e. the spillover traffic) by Student's t-test. The result (Appendix Table 1) shows that the purchase probability of the initial traffic is significantly greater than that of the indirect traffic, which indicates the importance of the initial traffic.

Table 1 Examples of four types of traffics

Session	Clickstream	Initial	Spillover	Diverted	Return
1	M1, M2, M3	M1	M3	M1	0
2	M2	M2	0	0	0
3	M3, M2, M1, M3	M3	0	M3	M3
4	M1, M2, M3, M2	M1	M2	M1	0
Total		Initial	Spillover	Diverted	Return
M1		2	0	2	0
M2		1	1	0	0
M3		1	1	1	1

After generating four types of traffics, I classify sales into four types: initial sales, indicating the purchases from the initial traffic; spillover sales, indicating the purchases from the spillover traffic; diverted sales, indicating the purchases at other merchants due to the diverted traffic; return sales, indicating the purchases from the return traffic. Table 2 provides summary statistics at the merchant level. On average, a merchant has 49.5 units of sales each day, with 2715.1 SGD revenue, and pays out 85.8 SGD cashback with an average cashback rate of 1.7%. There are 130.4 initial traffics, 24.2 spillover traffics, 28.6 diverted traffics, 4.4 return traffics for each merchant in each day. The detailed definition of each variable is described in Table3.

Table 2. Descriptive statistics

Variable	Obs	Mean	Std.Dev.	Min	Max
Sales	19468	49.533	231.563	0	4260
Revenue	19468	2715.145	10200.27	0	175000
Cashback amount	19468	85.795	313.815	0	8084.13
Cashback rate	19468	.017	.025	0	.144
Initial traffic	19468	130.405	492.913	0	13411
Diverted traffic	19468	28.61	94.473	0	3562
Spillover traffic	19468	24.158	77.55	0	2126
Return traffic	19468	4.423	18.125	0	765
Initial sales	19468	41.158	201.874	0	3471
Diverted sales	19468	6.705	22.812	0	956
Spillover sales	19468	6.698	26.616	0	869
Return sales	19468	1.682	7.14	0	189

4.2. Preliminary Analysis

Shen & Xiao (2014) claim that there are positive spillovers among competitors. They capture the spillovers by the effect of the rival's stock of outlets on the number of new outlets. Similarly, Schuetz (2015) captures the spillovers among competitors by the effect of the other retail sectors' employment density on the number of new employees per square mile. In the current analysis, there is not much variation in the number of merchants during three months, but I have the information on the number of consumer traffic and

merchant cashback rate. Following their ideas, to understand how merchants affect each other in a marketplace, I first investigate the correlation between a merchant's sales and others' weighted average traffic and others' weighted average cashback rate.

Table 3. Variable description

Variable	Description and measure
Sales	The total number of transactions at each merchant and each day
Revenue	The total purchase amounts at each merchant and each day (in SGD)
Cashback amount	The total amount of money paid back to consumers by each merchant and each day (in SGD)
Cashback rate	The average value of cashback rates for each merchant and each day
Initial traffic	The number of consumers who initially visit the merchant when they land in the cashback site
Spillover traffic	The number of consumers who move to the merchant of interest but initially visit another merchant
Diverted traffic	The number of consumers who initially visit the merchant of interest but later move to other merchants
Return traffic	The number of consumers who initially visit the merchant of interest and then move to other merchants but finally return
Initial sales	The number of purchases among initial traffic
Spillover sales	The number of purchases among spillover traffic
Diverted sales	The number of purchases at other merchants among the diverted traffic
Return sales	The number of purchases among return traffic

The correlation coefficients are obtained by estimating below regression equation.

$$y_{it} = \alpha_0 + \alpha_1 y_{i,t-1} + \alpha_2 Traffic_{it} + \alpha_3 \sum_{j \neq i} \omega_{ij} Traffic_{jt} + \alpha_4 CashbackRate_{it} + \alpha_5 \sum_{j \neq i} \tau_{ij} CashbackRate_{jt} + \mu_c + \gamma_w + \varepsilon_{it} \quad (15)$$

, where y_{it} is merchant i 's sales at day t ; $Traffic_{it}$ is the number of consumers who initially visit merchant i ; $\sum_{j \neq i} \omega_{ij} Traffic_{jt}$ is the weighted average number of the initial traffics of other merchants; $CashbackRate_{it}$ is merchant i 's average cashback rate at day t ; $\sum_{j \neq i} \tau_{ij} CashbackRate_{jt}$ is the

weighted average cashback rate of other merchants. The cashback rate and the weighted average of others' cashback rates are timed by 100. Besides, I control the lagging sales since it will influence the current sales through the accumulation effect. I generate the standard error by bootstrap method for bias correction given that the error term has serial correlation issues. The model also includes merchant category and weekday fixed effects, to control for time and category-specific characteristics. I set $\omega_{ij} = \tau_{ij} = 1/n$, where n is the number of merchants in the platform.

Because of the nature of the sales that it could not be negative, the dependent variable is left-censored at zero. Therefore, the preferred specification is a Tobit model, adjusting for left censoring at zero.

The estimation results are shown in Table 4. I gradually put the traffic variables, cashback rate variables, category dummies, and weekday dummies in the regression, showing that the results are robust. Column (4) includes all the variables of interest and covariates. The coefficient on other traffics is 0.137, significant at the 1% level, suggesting that the sales will increase 0.137 unit when others average traffics increase 1 unit. The coefficient on other cashback rates is -14.49, significant at the 1% level, indicating that the sales will decrease 14.49 units when others' average cashback rates increase 1%. Besides, the coefficients on its own traffic, cashback rate, and lagging sales are 0.194, 12.91, and 0.617 respectively, significant at the 1% level, which is as expected.

The results of the preliminary analysis show that merchant sales are significantly positively correlated to other merchants' traffic and negatively correlated to other merchants' cashback rates. In the next section, I will show the estimation results of the positive spillover and the negative business-stealing effects and investigate the impact of cashback rates on these two effects.

Table 4. The correlation between a merchant's sales and others' weighted average traffic and others' weighted average cashback rate.

D.V. = Sales	(1)	(2)	(3)	(4)
--------------	-----	-----	-----	-----

Lagging Sales	0.606*** (0.0424)	0.622*** (0.0485)	0.616*** (0.0392)	0.617*** (0.0416)
Others' traffics	0.205*** (0.0207)	0.197*** (0.0233)	0.195*** (0.0193)	0.194*** (0.0196)
Own traffic	0.185*** (0.0633)	0.210*** (0.0683)	0.217*** (0.0831)	0.137 (0.0871)
Cashback rate		12.23*** (1.076)	12.92*** (1.107)	12.91*** (0.912)
Others' cashbacks		-15.50*** (5.307)	-15.55*** (4.818)	-14.49*** (5.070)
Constant	-75.46*** (10.32)	-83.98*** (10.17)	-91.14*** (13.52)	-88.80*** (13.63)
Category FE	N	N	Y	Y
Weekday FE	N	N	N	Y
Observations	16224	16224	16224	16224

Note: *p<0.1, **p<0.05, ***p<0.01, and standard error in parentheses. All coefficients are estimated from Tobit models, adjusting for left censoring at zero. Variables cashback rate and mean value of others' cashback rates are timed by 100. The standard errors are generated by the bootstrap method for bias correction.

5. Empirical Analyses

5.1. Estimation of the Positive Spillover and the Negative Business-Stealing Effects

In section 3, I propose that the positive spillover effect is pointed identified by $\sum_{m'} \theta_{m'} \cdot \gamma_{m'm} \cdot (1 - \gamma_{m'mm'}) \cdot \delta_{m'm}$; the negative business-stealing effect is partially identified by $(0, \sum_{m'} \gamma_{mm'} \cdot (1 - \gamma_{mm'm}) \cdot \delta_{mm'})$. The positive spillover effect and the upper bound of the negative business-stealing effect for merchant m could be estimated by:

$$E_m[\text{spillover}] = T_{m2} \cdot \overline{\delta_{m'm}}/N.$$

$$E_m[\text{stealing}] = (T_{m3} - T_{m4}) \cdot \overline{\delta_{mm'}}/(N\theta_m)$$

, where $\overline{\delta_{m'm}} \approx S_{m2}/T_{m2}$, and $\overline{\delta_{mm'}} \approx S_{m3}/(T_{m3} - T_{m4})$; T_{m2} is the spillover traffic; T_{m3} is the diverted traffic; T_{m4} is the return traffic; S_{m2} is the spillover sales; S_{m3} is the diverted traffic.

I use the bootstrap method with 200 iterations to get the mean value and standard error of the positive spillover effect and the upper bound of the negative business-stealing effect for each merchant and each day. I also obtain the increased sales due to the spillover traffic and the potential loss because of the diverted traffic for each merchant and each day. I report the estimation results of 10 merchants on April 1, 2019 as an example in Table 5.

Table 5. The estimation results of the spillover and the business-stealing effect

Date	Merchant	Category	Spillover effect	Gains in sales	Stealing effect	Loss in sales
2019/4/1	M1	Travel	0.0029 (0.00002)	73.8508 (0.62250)	0.0739 (0.00059)	67.8010 (0.57476)
2019/4/1	M2	Travel	0.0026 (0.00002)	65.5473 (0.56906)	0.0877 (0.00064)	83.3632 (0.65606)
2019/4/1	M3	Dining & Food Delivery	0.0029 (0.00002)	73.5274 (0.58365)	0.0783 (0.00070)	56.0547 (0.51327)
2019/4/1	M4	Travel	0.0018 (0.00002)	46.5224 (0.43947)	0.0696 (0.00068)	42.0746 (0.44032)
2019/4/1	M5	Dining & Food Delivery	0.0028 (0.00002)	71.6567 (0.58529)	0.0865 (0.00069)	81.8607 (0.66546)
2019/4/1	M6	Transportation	0.0067 (0.00003)	169.5821 (0.85044)	0.0154 (0.00015)	56.5025 (0.54124)
2019/4/1	M7	General Marketplace	0.0047 (0.00003)	119.1990 (0.82372)	0.0402 (0.00024)	123.2736 (0.73173)
2019/4/1	M8	General Marketplace	0.0032 (0.00003)	81.3333 (0.64663)	0.0477 (0.00029)	133.2189 (0.83946)
2019/4/1	M9	General Marketplace	0.0053 (0.00003)	135.6119 (0.85419)	0.0387 (0.00024)	98.7214 (0.62943)
2019/4/1	M10	General Marketplace	0.0008 (0.00001)	21.0697 (0.31587)	0.0430 (0.00060)	28.5473 (0.40610)

Note: Standard errors are included in parentheses.

We can see from Table 5 that for some merchants their gains are larger than loss, e.g., “M1”, “M3”, “M6”, etc., while for some other merchants they lose more, e.g., “M2”, “M7”, “M8” etc. It is interesting to investigate which type of merchants have a larger positive spillover effect and which type of merchants have a larger business-stealing effect. I explore two scenarios: “General Marketplace” category vs. other categories, and high vs. low market-share merchants.

“General Marketplace” is a special category in the e-marketplace in our context since it is a marketplace per se. More various products and services are provided by the merchants in category “General Marketplace” than other categories, and they further reduce the search cost. Consumers are more likely to move to such merchants, e.g. Amazon, Taobao, Lazada, etc., to find the products they want, and they are less likely to move out from this type of merchants given that they can buy almost anything they want in the merchants. Therefore, the positive spillover effect should be larger on the merchants in the category “General Marketplace”, and the negative business-stealing effect should be smaller on them. Figure 2 shows some evidence of it. The right-hand side figure shows that the mean value of the business-stealing effect is significantly smaller on the merchants (Mean = .0659) in the “General Marketplace” category than those (Mean = .0974) in other categories. The left-hand side figure indicates that the mean value of the spillover effect is significantly larger on the merchants (Mean = .0015) in the “General Marketplace” category than those (Mean = .0002) in other categories.

Figure 2. Two effects on merchants in “General Marketplace” and other categories

High market-share merchants are more popular than low market-share merchants. Typically, high market-share merchants offer better products or lower prices, so consumers are more willing to visit and purchase from them

rather than some other unknown merchants. If a consumer starts from a high market-share merchant, then she will directly buy from it and will not move to other merchants; if a consumer starts from a low market-share merchant (attracted by ads or promotions), then she probably will move to a high market-share merchant to at least make a comparison. Also, for the across category movement, consumers are more likely to move to the high market-share merchants because of their reputation. Therefore, the business stealing effect will be larger on the low (vs. high) market-share merchants, and the spillover effect will be larger on the high (vs. low) market-share merchants. Figure 3 proves this point. The right-hand side figure shows that the mean value of the business-stealing effect is significantly smaller on the high market-share merchants (Mean = .0519) than the low market-share merchants (Mean = .0986). The left-hand side figure indicates that the mean value of the spillover effect is significantly larger on the high market-share merchants (Mean = .0015) than the low market-share merchants (Mean = .0002). In our context, high market-share merchants indicate the merchants that account for 80% of market shares, and the rests are the low market-share merchants.

Figure 3. Two effects on high vs low market-share merchants

5.2. The Effect of Cashback Rates on the Two Effects.

Having shown that the positive spillover effect and the business-stealing effect are identifiable/partially identifiable from the observed data, I now explore what merchants can do to increase sales. Given that the context of this research is a regional cashback e-commerce platform, I investigate the effect of cashback rates on these two effects.

Since the two effects are fractions which are ranging from zero to one (zero and one included), linear panel data models are not well suited here because it is difficult to impose a bounded effect of cashback rates on the two effects. I utilize the fractional response models for panel data developed by Papke and Wooldridge (2008) to deal with it. The fractional probit model I estimate has the form:

$$E(y_{it}|x_{i1}, x_{i2}, \dots, x_{iT}) = \Phi(\varphi_t + x_{it}\beta + \bar{x}_i\xi) \quad (16)$$

, where Φ is the standard normal cumulative distribution function; y_{it} is the two effects; x_{it} is the cashback rate; $\bar{x}_i = T^{-1} \sum_{t=1}^T x_{it}$ is the time averages; the presence of φ_t emphasizes that a different intercept in each year is allowed. Estimating Equation (16) by adding the time averages of cashback rate and using pooled Bernoulli quasi-MLE is identical to the fixed effects model (Mundlak's 1978, Chamberlain 1980). Besides, the directly estimated coefficients are not identified and need to be scaled to give marginal effects or average partial effects.

For the endogenous issues, first, for the concerns regarding omitted variable bias, I control the time-invariant unobserved variables by adding the time averages of merchant-level characteristics in the model. For example, for some popular merchants, like Taobao, Amazon, they can have higher cashback rates and higher spillover effect comparing to unknown merchants. I also include the weekday dummies to rule out part of the time fixed effect. For example, merchants tend to offer higher cashback rates, and consumers are

more likely to move to a certain type of merchants on weekends comparing to workdays. Second, for the concerns regarding simultaneous bias, consumers make purchase and visit decisions after they view the cashback information, so the buying and visiting probabilities are not likely to affect cashback rates in the current period. Although I have considered these factors that might induce endogenous issues, it is still reasonable to think cashback rates could be correlated with time-varying unobservable variables, in addition to being correlated with merchant-level heterogeneity and weekday heterogeneity. Therefore, I utilize the third-order lagged term of cashback rate as an instrument variable and apply the control function approach introduced by Heckman and Robb (1985) to estimate the model.

The legitimation of using the third-order lagged term of cashback rate as an instrumental variable deserves a discussion. First, to test the power of the third-order lag, I run three regressions. I gradually add the first-order, second-order, and third-order lagged terms of cashback rates into the regressions step by step as independent variables. I find that the adjusted R-square increases from 0.4226 to 0.5295, and coefficients on 1st, 2nd, and 3rd lagged terms are all significant, which means the third-order lag has the additional explanatory power. Second, to test the validity of the third-order lag, I conduct the Arellano-Bond test (Arellano & Bond, 1991) for the autocorrelation of the error term of Regression (11). The results show that there is no autocorrelation at the third-order lag, which means the lagged term has no direct causal effect on the unobserved confounders. In this case, the “lagged IV” could mitigate the endogeneity issue by reducing both bias and the RMSE (Wang & Bellemare, 2019). The estimated results of direct results are reported in Column (1) of Table 6 and Table 7.

Next, I investigate the interaction effect of cashback rates with merchant categories and with weekends on the two effects. The fractional probit model I estimate is as follows:

$$E(y_{it}|x_{i1}, x_{i2}, \dots, x_{iT}) = \Phi(\varphi_t + x_{it}\beta + \bar{x}_i\xi + x_{it} \cdot i.categories \cdot \Gamma_1 + \bar{x}_i \cdot i.categories \cdot \Gamma_2 + x_{it} \cdot i.weekends \cdot \gamma_1 + \bar{x}_i \cdot i.weekends \cdot \gamma_2) \quad (17)$$

, where *i.categories* is a set of dummy variables indicating merchant categories; *i.weekends* is a dummy variable indicating weekends or workdays. The model is estimated by the pooled Bernoulli quasi-MLE and the control function approach after considering the lagging term of the cashback rate as an instrument variable. I also estimate the linear fixed effect model with instrument variables as a robust check. The results are reported in Column (2) and (3) respectively of Table 6 and Table 7.

Table 6. The effect of cashback rates on the spillover effect

Model:	Fractional probit		Fractional probit		Linear
Estimation method:	Pooled QMLE		Pooled QMLE		Fixed effects
	(1)		(2)		(3)
	Coefficient	Margins	Coefficient	Margins	Coefficient
Cashback rates	1.944*** (0.684)	0.00243*** (0.000917)	0.475 (1.036)	0.00280*** (0.000834)	0.000209 (0.000188)
Beauty × Cashback rates			2.383 (1.479)	0.00488 (0.00348)	0.00184* (0.00103)
Dining × Cashback rates			1.186 (1.607)	0.00417 (0.00352)	0.000799 (0.000766)
Fashion × Cashback rates			3.218** (1.309)	0.00317* (0.00185)	0.000968 (0.000651)
General × Cashback rates			-0.113 (1.140)	0.00159 (0.0136)	0.000524 (0.00101)
Home × Cashback rates			8.562*** (2.397)	0.00788* (0.00422)	0.00910 (0.00757)
Electronics × Cashback rates			2.641 (1.746)	0.00145 (0.00131)	0.000359 (0.000256)
Travel × Cashback rates			2.004 (1.339)	0.00335* (0.00197)	0.00348* (0.00192)
Other Services				0.0000351 (0.000216)	
Weekends × Cashback rates			-1.005 (0.699)	0.00167** (0.000836)	0.0000452* (0.0000254)
Workdays				0.00350***	

				(0.00101)	
Constant	-3.091***		-3.126***		0.000262***
	(0.212)		(0.233)		(0.0000266)
Observations	13114	13114	13114	13114	13114

Note: (i) *p<0.1, **p<0.05, ***p<0.01, and standard error in parentheses. (ii) The estimated coefficient of the fractional probit model can only give the sign and significance level of the effect but not the effect size. (iii) The margins in the first row give the marginal effect of cashback rates on the positive spillover effect. (iv) The margins for each interaction effect give the average partial effect of cashback rate on the positive spillover effect at that specific category. (v) The standard errors in all the models are clustered by merchants.

The results of the estimation of cashback rates' effect on the positive spillover effect are reported in Table 6. Column (1) reports the direct effect. The coefficient on cashback rates is 1.944, significant at the 1% level, but the magnitude of the coefficient is not directly indicating the partial effect. The average partial effect of cashback rate on the positive spillover effect is given by the margins column, which is 0.00243, indicating the spillover effect will increase by 0.00243% when cashback rates increase by 1%. Column (2) reports the interaction effect. The results show that the cashback rate has a larger effect on the spillover effect in the category of "Fashion" and "Home & Lifestyle". This is as expected since consumers in the category of "Fashion" and "Home & Lifestyle" are often more sensitive to price so the cashback rate will have a larger impact on them. Likewise, the magnitude of the coefficient is not directly comparable to the estimate from the linear model. The margin column reports the average partial effect of cashback rate on the spillover effect in each category. Besides, there is no significant interaction effect with weekends. Column (3) reports the estimated results of the linear fixed-effect model. The coefficients are insignificant, indicating that the results are biased and inefficient from the linear model.

Table 7. The effect of cashback rates on the business-stealing effect

Model:	Fractional probit		Fractional probit		Linear
Estimation method:	Pooled QMLE		Pooled QMLE		Fixed effects
	(1)		(2)		(3)
	Coefficient	Margins	Coefficient	Margins	Coefficient
Cashback rates	-2.710***	-0.466***	-5.926*	-0.417***	-0.420
	(0.567)	(0.0988)	(3.261)	(0.106)	(0.341)

Beauty × Cashback rates		0.0152	-0.957**	-0.889
		(3.767)	(0.395)	(0.602)
Dining × Cashback rates		5.341*	-0.0579	0.338
		(3.185)	(0.0809)	(0.342)
Fashion × Cashback rates		3.069	-0.429***	-0.101
		(3.163)	(0.132)	(0.359)
General × Cashback rates		5.830*	0.0224	0.361
		(3.293)	(0.151)	(0.389)
Home × Cashback rates		1.081	-0.707**	-0.538
		(3.867)	(0.342)	(0.524)
Electronics × Cashback rates		3.826	-0.297**	0.134
		(3.284)	(0.136)	(0.355)
Travel × Cashback rates		5.538*	-0.0247	0.344
		(3.320)	(0.173)	(0.372)
Other Services			-1.283*	
			(0.680)	
Weekends × Cashback rates		0.852	-0.338	0.138
		(1.465)	(0.223)	(0.160)
Workdays			-0.448***	
			(0.125)	
Constant	-1.514***	-1.262***		0.0991***
	(0.0499)	(0.0313)		(0.00414)
Observations	13412	13412	13412	13412

Note: (i) * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, and standard error in parentheses. (ii) The estimated coefficient of the fractional probit model can only give the sign and significance level of the effect but not the effect size. (iii) The margins in the first row give the marginal effect of cashback rates on the negative business-stealing effect. (iv) The margins for each interaction effect give the average partial effect of cashback rate on the negative business-stealing effect at that specific category. (v) The standard errors in all the models are clustered by merchants

The results of the estimation of cashback rates' effect on the negative business-stealing effect are reported in Table 7. Column (1) reports the direct effect. The coefficient on cashback rates is -2.710, significant at the 1% level, but the magnitude of the coefficient is not directly indicating the partial effect. The average partial effect of cashback rate on the negative business-stealing effect is given by the margins column, which is -0.466, indicating the spillover effect will decrease by 0.466% when cashback rates increase by 1%. Column (2) reports the interaction effect. The results show that the cashback rate has a larger effect on the business-stealing effect in the category of “Dining & Food Delivery”, “General Marketplace”, and “Travel”. Likewise, the magnitude of the coefficient is not directly comparable to the estimate from the linear model.

The margin column reports the average partial effect of cashback rates on the business-stealing effect in each category. Besides, there is no significant interaction effect with weekends. Column (3) reports the estimated results of the linear fixed-effect model. The coefficients are insignificant, indicating that the results are biased and inefficient from the linear model.

6. Conclusion

Merchants in a marketplace will influence each other. On one hand, a merchant can benefit from others due to demand-side externalities; on the other hand, merchants will lose part of their sales because of the business-stealing effects. These two kinds of opposite effects exist simultaneously in the marketplace. How to empirically identify them, and, in the context of the cashback platform, how cashback rates will affect these two effects are questions worth investigating.

The current research separately identifies the two effects by using consumer online clickstream data. I classify consumer traffic into four types: initial traffic, spillover traffic, diverted traffic, and return traffic using consumer clickstream data. I argue that the positive spillover effect is pointed identified, while the negative business-stealing effect is partially identified. I non-parametrically estimate these two effects and further investigate the effect of cashback rates on them. Since the nature of the dependent variable are fractional data, I utilize a fractional response probit model instead of a linear model. Besides, although I have controlled the time averages of merchant-level characteristics and weekdays dummies, it is still possible that cashback rates are correlated with time-varying unobservable variables. Therefore, I utilize the lagging term of cashback rate as an instrument variable and apply the control function approach to estimate the model. The results show that cashback rates have significant positive impacts on the positive spillover effect and have significant negative impacts on the negative business-stealing effect.

I also investigate the category and weekend interaction effect of cashback rates. The results show that cashback rates have a larger impact on the positive spillover effect in categories “Fashion” and “Home & Lifestyle” and have a larger impact on the negative business-stealing effect in categories “Dining & Food Delivery”, “General Marketplace”, and “Travel”. Besides, I derive the average partial effects of cashback rate in each category, which could give rich managerial instructions to the cashback platform.

In future studies, I plan to extend the current research from two aspects. First, I will build a nested multinomial logit model to investigate consumer click and purchase behavior. Second, I would like to build an AI system to automatically choose cashback rates and promotion information to maximize merchants’ revenues in real-time data.

Appendix

Table A1. The results of the t-test for the difference between the purchase probability of the direct traffic and that of the indirect traffic

Variable	Obs	Mean	Std.Err.	Std.Dev.	[95% Conf. Interval]
direct	12,352	0.237127	0.002129	0.236585	0.232954 0.2412993
indirect	12,352	0.141177	0.001882	0.209189	0.137488 0.1448665
diff	12,352	0.0959496	0.002202	0.2447255	0.0916334 0.1002658

Table A2. The results of the first stage of the IV approach

	(1)	(2)	(3)
L.cashback_rates	0.651*** (0.00578)	0.404*** (0.00736)	0.312*** (0.00793)
L2.cashback_rates		0.376*** (0.00737)	0.276*** (0.00802)
L3.cashback_rates			0.246*** (0.00794)
Constant	0.00665*** (0.000183)	0.00446*** (0.000190)	0.00349*** (0.000198)
Observations	17304	15892	14959

Note: (i) *p<0.1, **p<0.05, ***p<0.01, and standard error in parentheses.

Table A3. The results of the Arellano-Bond test for zero autocorrelation in errors

Order	spillover	p-value	stealing	p-value
1	-2.8913	0.0038	-7.1753	0.0000
2	-.21673	0.8284	-4.6853	0.0000
3	1.5389	0.1238	-1.2899	0.1971

Note: (i) H0: no autocorrelation.

References

- Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, 58(2), 277-297.
- Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2016). Consumer behavior on cashback websites: Network strategies. *Journal of Business Research*, 69(6), 2101-2107.
- Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. *Journal of Business Research*, 88, 407-414.
- Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841-890.
- Bastos, P. V., & Greve, H. R. (2003). Interorganizational learning and the location of manufacturing subsidiaries: Is chain migration also a corporate behavior? *Advances in Strategic Management*, 20, 159-192.
- Baum, J. A., & Korn, H. J. (1999). Dynamics of dyadic competitive interaction. *Strategic management journal*, 20(3), 251-278.
- Carroll, G. R., & Hannan, M. T. (2000). Why corporate demography matters: Policy implications of organizational diversity. *California Management Review*, 42(3), 148-163.
- Chamberlain, G., (1980). Analysis of variance with qualitative data. *Review of Economic Studies* 47, 225–238.
- Chu, J., & Manchanda, P. (2016). Quantifying cross and direct network effects in online consumer-to-consumer platforms. *Marketing Science*, 35(6), 870-893.

- Chung, W., & Kalnins, A. (2001). Agglomeration effects and performance: A test of the Texas lodging industry. *Strategic management journal*, 22(10), 969-988.
- Clapp, J. M., Ross, S. L., & Zhou, T. (2019). Retail Agglomeration and competition externalities: Evidence from openings and closings of multiline department stores in the US. *Journal of Business & Economic Statistics*, 37(1), 81-96.
- Dudey, M. (1990). Competition by choice: The effect of consumer search on firm location decisions. *The American Economic Review*, 1092-1104.
- Fuentelsaz, L., & Gómez, J. (2006). Multipoint competition, strategic similarity and entry into geographic markets. *Strategic management journal*, 27(5), 477-499.
- Fujita, M., & Thisse, J.-F. (1996). Economics of agglomeration. *Journal of the Japanese and international economies*, 10(4), 339-378.
- Furman, J. L., Kyle, M. K., Cockburn, I. M., & Henderson, R. (2006). *Public & private spillovers, location and the productivity of pharmaceutical research* (0898-2937). Retrieved from
- Gimeno, J. (1999). Reciprocal threats in multimarket rivalry: Staking out 'spheres of influence' in the US airline industry. *Strategic management journal*, 20(2), 101-128.
- Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., & Shleifer, A. (1992). Growth in cities. *Journal of political economy*, 100(6), 1126-1152.
- Hausman, J., & Leibtag, E. (2007). Consumer benefits from increased competition in shopping outlets: Measuring the effect of Wal-Mart. *Journal of Applied Econometrics*, 22(7), 1157-1177.
- Heckman, J. J., & Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, 30(1-2), 239-267.

- Ho, Y.-C., Ho, Y.-J., & Tan, Y. (2017). Online cash-back shopping: Implications for consumers and e-businesses. *Information Systems Research*, 28(2), 250-264.
- Koçak, Ö., & Özcan, S. (2013). How does rivals' presence affect firms' decision to enter new markets? Economic and sociological explanations. *Management Science*, 59(11), 2586-2603.
- Konishi, H. (2005). Concentration of competing retail stores. *Journal of Urban economics*, 58(3), 488-512.
- Li, H., Shen, Q., & Bart, Y. (2018). Local market characteristics and online-to-offline commerce: An empirical analysis of Groupon. *Management Science*, 64(4), 1860-1878.
- Lu, M. Y., & Shin, J. (2018). A model of two-sided costly communication for building new product category demand. *Marketing Science*, 37(3), 382-402.
- Manski, C.F., (1990). Nonparametric Bounds on Treatment Effects, *American Economic Review* 80 (2), 319-323.
- Manski, C.F., (1997). Monotone Treatment Response, *Econometrica* 65(6), 1311-1334.
- Manski, C.F., (2003). Partial identification of probability distributions. Springer Science & Business Media.
- Manski, C.F., Pepper, J.V., (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling, *Econometrica* 68 (4), 997-1012.
- Marshall, A. (1890). Principles of Economics, 8th edn (1920). London, Mcmillan.
- Marshall, A. (1920). Industry and trade. In: London: macmillan.
- Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing science*, 23(4), 579-595.

- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69-85.
- Myles Shaver, J., & Flyer, F. (2000). Agglomeration economies, firm heterogeneity, and foreign direct investment in the United States. *Strategic management journal*, 21(12), 1175-1193.
- Picone, G. A., Ridley, D. B., & Zandbergen, P. A. (2009). Distance decreases with differentiation: Strategic agglomeration by retailers. *International Journal of Industrial Organization*, 27(3), 463-473.
- Quigley, J. M. (1998). Urban diversity and economic growth. *Journal of Economic perspectives*, 12(2), 127-138.
- Schuetz, J. (2015). Why are Walmart and Target next-door neighbors? *Regional Science and Urban Economics*, 54, 38-48.
- Shen, Q., & Xiao, P. (2014). McDonald's and KFC in China: Competitors or Companions? *Marketing Science*, 33(2), 287-307.
- Stephen, A. T., & Toubia, O. (2010). Deriving value from social commerce networks. *Journal of marketing research*, 47(2), 215-228.
- Toivanen, O., & Waterson, M. (2005). Market structure and entry: Where's the beef? *RAND Journal of Economics*, 680-699.
- Vana, P., Lambrecht, A., & Bertini, M. (2018). Cashback Is cash forward: Delaying a discount to entice future Spending. *Journal of marketing research*, 55(6), 852-868.
- Vitorino, M. A. (2012). Empirical entry games with complementarities: An application to the shopping center industry. *Journal of marketing research*, 49(2), 175-191.
- Wang, Y., & Bellemare, M. F. (2019). Lagged Variables as Instruments. *Working paper*.

- Yang, S., Lu, S., & Lu, X. (2014). Modeling competition and its impact on paid-search advertising. *Marketing Science*, 33(1), 134-153.
- Zhou, Y.-W., Cao, B., Tang, Q., & Zhou, W. (2017). Pricing and rebate strategies for an e-shop with a cashback website. *European Journal of Operational Research*, 262(1), 108-122.