Essays on Economics of Online Platforms

Shahryar Doosti

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Yong Tan, Chair

Ming Fan

Hyeunjung (Elina) Hwang

Program Authorized to Offer Degree:
Foster School of Business

University of Washington

Abstract

Essays on Economics of Online Platforms

Shahryar Doosti

Chair of the Supervisory Committee:
Michael G. Foster Professor Yong Tan
Information Systems and Operations Management

In this dissertation, I study the economics of online platforms. I explore economic consequences of the interaction between firms and users through online platforms. I investigate three aspects of such interactions through data-driven economic models: mobile economy, online crowd, and social media. I study the effect of platform design on demand and monetizing online traffic. First, I study the impact of emerging mobile ecosystem on users' interaction patterns. Specifically, I choose the online shopping platform in which users have two channels to purchase the products: PC and App. I investigate whether niche products are more likely to be seen and selected in mobile channel. I also look into customers' usage patterns to draw conclusions on how to increase sales of niche products. Second, I study the online crowdfunding and the effect of the campaign design on overall success. I build a structural model to estimate the demand for Kickstarter projects and make conclusions on the effect of reward scheme designs. Third, I explore the sponsored videos on Facebook. A sponsored content (e.g. video) features the sponsor's brand or product. I study the effect of the creator-sponsor association on viewership and user engagement for videos on Facebook. More specifically, I identify multiple aspects of creator-sponsor association and investigate their effect on the overall user engagement. I highlight the managerial implications of the findings in this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this opportunity to acknowledge the great impacts and influences in my doctoral studies. First, I wish to express my sincere appreciation to my PhD advisor Professor Yong Tan who has been a great influence through out my studies. He has been my mentor and more importantly a friend for me. We have spent numerous hours of discussing research, traveling and eating together. His focus on research rigor and methodological soundness has propelled me to deepen my knowledge and do research responsibly. I am grateful for the great insights he provided and I will be ever grateful to have his support and guidance. I also want to thank other members of my advisory committee, Professor Hema Yoganarasimhan, Professor Ming Fan, Professor Elina Hwang, and Professor Shan Liu for their insightful comments and support along the way. I thank Professor Kamran Moinzadeh, the chair of department, for providing the best environment for PhD students. I also appreciate the great classes with great teachers at UW that made me motivated in the field.

I would like to express my deep appreciation to Shawna Reimers, Jaime Banaag, Jessica Aceves, Beau Kirkeby and other staff of the Foster Business School. They always helped me through different stages of my studies and willingly went beyond their scope of responsibilities to help and support me. I also want to thank my dear friends that made this journey much more pleasant: Melissa Rhee, Majid Mirbagheri, Sareh Nabi, Mohsen Sharifani, Mohammad Arbabian, Amir Fazli, Behnaz Bojd, Eugene Pavlov, Arash Naderpour, Elnaz Jalilipour, Danial Vaezi, Omid Rafieian, and Aaravinda Garimella.

I thank my parents for their unconditional support. I could not imagine to be here had it not been for my parents Sakineh and Alireza Doosti. They are permanent source of

inspiration for me. I will be ever grateful that I have them. I also thank my brother and sister, Bardia and Mitra, for always being supportive and caring.

Last but not the least, I thank my wife and my best friend Pegah for her love and support. I definitely could not have made it this far without her. She has always been there for me with unconditional support. She always motivated me in every stage of my studies, and openheartedly inspired me with her love.

DEDICATION

to my wife and love of my life, Pegah

&

to my parents, Sakineh and Alireza

&

to the memory of my grandmother, Abali

Chapter 1

INTRODUCTION

Digital economy has taken over the traditional channels of monetary transactions. I find the consequences of such transformation in all aspects of customers' life including but not limited to shopping, communications, getting news, social relationships, and investing. It is hard to imagine an economy without online platforms. Online platforms have been facilitating in creating digital markets. They often provide two-sided markets that can bring sellers and buyers together. The notion of buyer and seller can be far more general than actual traders. For example, in case of crowd-based investments, investors can be referred as buyers and project owners as sellers. Despite the age of the digital economy does not reach far back in history, the pace of changes has been dramatically increasing and reshaping the industries and lifestyles. Hence, it is crucial to understand the economic consequences of the digital platforms and the role of interaction of online players. I study online platforms in three different contexts: Mobile economy, online crowd, and social media advertising. They represent three major streams in the internet era.

In Chapter 3, I study the mobile economy in the context of online shopping with having differentiated products and heterogeneous sellers on the platform. With the emergence of online channels, more niche products are sold compared to traditional channels, and it has resulted in more product variety. This Long Tail effect has distributed the sales among more products, reducing the importance of the mainstream top-selling items. As the online market grows, mobile applications are becoming the mainstream channel in retail and digital marketing. Using a rich dataset from an online marketplace, I study the distribution of product sales in online channels and compare the effect of the long tail in PC versus App. I show that the distribution of product sales in App is less concentrated than the one in

PC. Consequently, App increases consumer surplus due to buyers enlarged choice sets. I investigate the mechanism of the long tail effect and find that different shopping patterns in PC and App result in the long tail effect in App. Customers tend to spend more time searching the products in App. I find that more search time on products would increase the likelihood of choosing a niche item, and App has a positive moderating effect. This provides an important insight for retailers: adding mobile applications to their sales channels increases the variety of the products sold and generates value for them through increasing the customer surplus. Also, retailers and App developers can utilize the usage pattern to design Apps. My paper contributes to the literature by investigating the economics of App as sales channels and linking to customers behavior in online markets.

In Chapter 4, I study the mechanisms in online crowdfunding projects. Reward-based crowdfunding platforms are best known to facilitate the fund-raising process for entrepreneurs and small businesses. Entrepreneurs seek money from the backers to kick off their projects in exchange for rewards. While the competition among the projects to get more supporters grow, many projects fail to reach their fund-raising target. I use a structural demand estimation to understand the role of different aspects of reward scheme design and pricing. More specifically, I apply an aggregate level discrete choice demand model on Kickstarter projects. I characterize the features of the projects by utilizing the document embedding vectors from natural language processing methods. I treat the endogeneity of the price by applying latent instrument variables and finite mixture model for price. The results show that the price coefficient is biased towards zero in the absence of endogeneity treatment. On the reward scheme design, I show low-level rewards cannibalize the more expensive ones. I also find that low-level rewards have the least contribution in customer welfare. Campaign creators and online platforms benefit from the insights of the model for reward scheme design and pricing.

In Chapter 5, I investigate a specific form of advertising on social media: *video sponsorship*. Sponsored videos on social media have experienced a dramatic market growth. A sponsored content (e.g. video) features the sponsor's brand or product. With the advent of

accurate measures of user engagements on social media, there is not, yet, a clear strategy to improve the effectiveness of the sponsored contents. I study the effect of the creator-sponsor association on viewership and user engagement for videos on Facebook. More specifically, I identify multiple aspects of creator-sponsor association and investigate their effect on the overall user engagement. The association between the creator and the sponsor is categorized in three measures: 1. Theme and industry relevance, 2. Content similarity, and 3. Audience overlap. First, I find that sponsorship leads to 65% reduction in views. While the industry category similarity has no significant effect on the sponsored views and engagements, both content similarity and audience overlap have positive effect on the performance measures. The relevance impact is stronger in *Food & Drinks* (372% increase in views) and *Sports* (121% increase in views) categories. I also find creator-sponsor relevance has a positive moderation effect on the sponsor's presence. I conclude that the relevance between the creator and sponsor mitigates the negative effect of sponsorship on social media.

Chapter 2

LITERATURE REVIEW

I investigate different aspects of online platforms in three contexts. First, I overlay the related research on the effect of mobile channels in online retailing in Section 2.1. Second, I review the literature in the context of crowdfunding and demand estimation in Section 2.2. Lastly, in Section 2.3, I explore the extant research in social media advertising and sponsorship.

2.1 Mobile Economy and Long Tail

My research on mobile economy in online retailing relates to two streams of literature. First, it builds on an evolving body of literature of mobile commerce. With the emergent advancing mobile technologies such as 3G and 4G, now users can access the internet anytime and anywhere. More than 174 million U.S. consumers now own smartphones (Siwicki, 2014). Moreover, the prices of internet data for phones have been decreasing over time and people opt-in to unlimited data plans. Mobile users feel less barrier due to the growth and ubiquity of the mobile markets. I see a growing number of studies in the well-established literature of electronic commerce focusing on economics and behavioral patterns of mobile economy. Shim et al. (2008) study mobile usage patterns when people watch TV programs via their phone. They find that the highest usage time happens early in the morning and in the evening while people commute from home to work and back. O'Hara et al. (2007) find that people use their phones to spend time alone or pass time. Thus, I can expect mobile internet to bring different usage patterns which are not possible with laptops and personal computers. My study relates to both aspects of mobile economy. I investigate the economic consequences of rising mobile channels especially in competition with more conventional medium. I also

consider behavioral patterns of usage leading to those consequences.

Second, my study connects to the stream of literature addressing multi-channel retailing. There are numerous studies examine switching behavior between channels, customer loyalty, and choice models through multi-channel firms. Shriver and Bollinger (2015) build up a demand estimation model based on online and offline stores for a clothing chain retailer. While Shriver and Bollinger (2015) focus on generative models by structural modeling on individual decision making, others study the overall performance of multi-channel system in terms of the aggregate economic measures (Brynjolfsson et al., 2003, 2011). By the emergence of online channels, there are more studies comparing offline and online channels. It is estimated that 63% of sellers only sell on online marketplaces.[1] The online channels overall lead to a longer tail compared to offline stores (Brynjolfsson et al., 2003). That is caused by the lower search cost in internet channel due to utilizing recommendation systems (Brynjolfsson et al., 2011).

Online markets have gone through considerable changes since their birth. Most importantly, mobile channel has become the dominant channel in online marketplaces, since 70% of online transactions occur on a mobile device.[2] Although the shift toward mobile channels seems inevitable, there are little knowledge on the effects of such market transforms. In a study, Ghose et al. (2012) show that smaller screen sizes on mobile phones increase search and browsing cost. It suggests that mobile phones put more weight on top selling products rather than niche products due to the higher searching cost of niche products on phones. Moreover, mobile users have relatively shorter attention spans compared to PC users (Roto and Oulasvirta, 2005). Fang et al. (2015) found that location-based mobile promotions induce the impulsive, same-day purchases. On the other hand, it is shown that users have more access time on the mobile phones (Han et al., 2016). At the same time, surveys and anecdotes show that online users spend more time on smartphones than any other devices (Eadicicco, 2015). Putting these facts together, I may expect that online users leverage

[1] http://www.webretailer.com/lean-commerce/statistics-marketplace-seller-survey/

[2] http://www.snapretail.com/resource/retailer-stats-tactics/

the access time and overcome the search costs on mobile phones to collect more information about the products that they have not experienced. Table 2.1 presents a summary of literature on both streams.

2.2 Online Crowd and Demand Estimation

My paper on online crowdfunding is related to multiple streams of research. First, it contributes to the extensive literature of product line design and pricing in marketing and economics. In crowdfunding projects, entrepreneurs often offer multiple product choices in terms of rewards (Hu et al., 2015). These rewards may segment the market through quality or quantity attributes as a more is better property (Desai, 2001). This is called vertical differentiation representing second-degree price discrimination. There are numerous studies on the cannibalization effects in such markets (Moorthy, 1984). Most of the studies use analytical approaches with monopoly or oligopoly manufacturers (Sutton, 1986; Choi and Shin, 1992; Vandenbosch and Weinberg, 1995; Dutta et al., 1995; Cremer and Thisse, 1991). Balachander and Srinivasan (1994) study the role of private information and Bhargava and Choudhary (2001) investigate vertical differentiation in information goods. Using a multi-period model, Hu et al. (2015) study the product line design in crowdfunding through self-selection of the products by heterogenous customers. However, there is not much empirical evidence on vertical differentiation and cannibalization effect in online crowdfunding projects. I capture the competition in my model to study the product line design and pricing on crowdfunding projects.

My paper is also connected to the growing body of literature studying the crowdfunding platforms and customer behaviors. Agrawal et al. (2014) investigate the economic mechanism of crowdfunding platforms as emerging market for finance. Several papers have empirically investigated the behavioral effects such as the privacy (Burtch et al., 2015), the role of affiliations (Agrawal et al., 2015), and social influence (Burtch et al., 2013). There are also studies on the dynamics of fundraising as campaigns progress towards the deadline (Agrawal et al., 2011; Freedman and Jin, 2011; Zhang and Liu, 2012; Mollick, 2014). We, however,

Table 2.1: Summary of literature in mobile and multi-channel commerce

Area	Authors	Key Findings	Method
Behavioral Research on Mobile Commerce	Shim et al. (2008)	TV usage patterns: widespread user age group, peak viewing time, high indoor usage and longer viewing time. These results shaped strategic implications, furthering and enhancing a personalized media experience.	Secondary data analysis
	Ghose et al. (2012)	Smaller screen sizes on mobile phones increase search and browsing cost.	Natural Experiment
	Fang et al. (2015)	location-based mobile promotions induce the impulsive, same-day purchases	Field experiment
	Ghose and Han (2011)	An increase in content usage in the previous period has a negative impact on content generation in the current period and vice versa.	Simultaneous equation panel data model
	O'Hara et al. (2007)	Watching TV episodes on mobile are short, but mobile video allowed people to use different time periods for different purposes.	Explanatory study
	Roto and Oulasvirta (2005)	User's visual attention shifted away from the mobile browser usually between 4 and 8 seconds. In contrast, the continuous span of attention to the browser was longer in the laboratory.	Lab experiment
	Hoehle and Venkatesh (2015)	Mobile application usability are good predictors of continued intention to use and loyalty.	Conceptual model and survey
	Wu and Wang (2005)	Compatibility has the most important effect on behavioral intention to use and the second most important effect on the actual use. Cost is one of the important predictors of MC adoption intent, and this has a negative direct effect on behavioral intention to use.	Survey
	Jenkins et al. (2016)	Daily interruptions brought by personal computers and mobile devices cause increased stress and decreased productivity.	Empirical fMRI study
Economics of Mobile Commerce	Shriver and Bollinger (2015)	Channel complementarity through increased shopping frequency as the distance to retail outlets decreases, accompanied by increased substitution from online to retail.	Structural Modeling
	Brynjolfsson et al. (2003)	Increased product variety of online bookstores enhanced consumer welfare by $731 million to $1.03 billion in the year 2000, which is between 7 and 10 times as large as the consumer welfare gain from increased competition and lower prices in this market.	Empirical welfare estimation techniques
	Brynjolfsson et al. (2011)	The online channels overall lead to a longer tail compared to offline stores.	Empirical Model
	Ghose and Han (2014)	App demand increases with the in-app purchase option wherein a user can complete transactions within the app.	Structural Modeling
	Han et al. (2016)	Users baseline utility diverges substantially across app categories and their demographic characteristics explain a substantial amount of heterogeneity in baseline utility and satiation.	Structural Modeling
	Lin and Wang (2006)	Customer satisfaction, habit, trust and perceived value on mobile, affects customer royalty.	Survey and Structural modeling
	Xu et al. (2016)	Tablets play complementary roles for smartphones and act as the substitute for PC on e-commerce markets.	Natural experiment

consider the crowdfunding platfrom as a market and apply demand estimation techniques to understand the mechanisms.

My study is built on the ground of structural demand estimation methods. A very well appreciated approach is McFadden et al. (1977) discrete choice model. In McFaddens model, the product is nothing but the set of its features. There are numerous studies adopting and developing McFadden's framework (Bruno and Vilcassim, 2008; Dubé et al., 2012; Besanko et al., 1998; Hartmann, 2010). Many of the studies have utilized the individual level data using various approaches. Shriver and Bollinger (2015) developed a multi-channel demand model by utilizing the individual level data. Bajari et al. (2015) adopted a machine learning approach to estimate the demand. Some scholars also use hedonic approach to estimate demand Bajari and Benkard (2005). Using the aggregate level data, Berry et al. (1995) established a demand model for the auto industry in the US. Petrin (2002), and Nevo (2001) followed the same framework to study the demand in various industries. Since I have aggregate data from the crowdfunding platform, I use an approach similar to Berry et al. (1995). To the best of my knowledge, there is no study on investigating structural demand modeling of reward-based crowdfunding. I aim to contribute to the literature by tailoring an innovative model to get insights from the demand structure of these platforms.

There are several challenges to capture the effect of interest in formulating a structural model for crowdfunding projects. The first and foremost challenge is the enormous heterogeneity in the projects. Even though I restrict the model to the technology category, the projects might not be reduced to some of the very limited pre-defined characteristics. Hence, I employ novel natural language processing techniques to capture the contextual and semantic information embodied in the project description. The second challenge is that I do not observe any single campaign over time. That would add to the complexity of inferring about unobserved characteristics. In the next section, I outline my empirical setting and data description.

2.3 Social Media Sponsorship

The research on social media monetization through video sponsorship is related to multiple streams of research. I build on the sponsorship and online advertising literature. Effectiveness measurement has been one of the main domains of online advertising (Manchanda et al., 2006; Yang and Ghose, 2010; Zhang and Katona, 2012; Rishika et al., 2013; Prasad, 1976). For instance, Manchanda et al. (2006) discussed the effect of banner ads on purchases and customer returns. Gallagher et al. (2001) and Drèze and Hussherr (2003) studied changes in brand awareness, brand attitudes and purchase intentions as a function of exposure. Exposure effectiveness is also studied through advertising spending (Mela et al., 1998; Ilfeld and Winer, 2002). Another area that the advertising effectiveness is well-studied is sponsored keywords and sponsored listings (Edelman et al., 2007; Feng et al., 2007; Varian, 2007; Liu et al., 2010; Athey and Ellison, 2011). Baye and Morgan (2001), Baye et al. (2009), and Montgomery et al. (2004) have studied the impact of retailers rank during placement on click-through rates. I review the the extant literature on the role of relevance in sponsorship and ad avoidance.

2.3.1 Sponsorship

Sponsorship has transpired as an approach that has advantages over other forms of promotion (Meenaghan, 1991), because it is flexible and can improve the image of the sponsor (Hastings, 1984). Companies use sponsorship to boost the brand awareness (Rajaretnam, 1994). Event sponsorship has a significant impact on both consumer awareness and image (Otker and Hayes, 1987). However, online sponsorship is overlooked by scholars especially in the context of social media. In the prevalence of social media, more companies try to take advantage of the network structure and user connections (Miller and Tucker, 2013). Companies can form more sustainable relationships with customers through engaging users on social media (Rishika et al., 2013). Therefore, user engagement is very valuable to companies (Claussen et al., 2013). They seek various ways to engage customers to make the advertisements

more effective. User-generated contents (UGS) help companies to spread the word-of-mouth (WOM) in social media (Goldenberg et al., 2012; Fossen and Schweidel, 2016). Goh et al. (2013) found that UGS has a strong impact on purchase behavior and persuasiveness. Hence, targeting influential users can increase the content exposure (Trusov et al., 2010).

Online sponsorship on social media has grown significantly due to the surge in social media consumption. DÁstous and Bitz (1995) have categorized sponsored events in terms of four factors: the nature of the sponsorship (philanthropic versus commercial), its origin (preexisting versus event created by the sponsor), its frequency (continuous versus oneshot) and the strength of the link between the entity (or the event) and the sponsor (weak versus strong). The origin of the videos is likely to impact the users' perceptions. According to DÁstous and Bitz (1995), one critical issue with sponsor-born sponsorship is the lack of credibility. The genuine contents are perceived to be more credible. A credible sponsor improves both brand image and user engagement with the creator. In other forms of advertising, there are studies comparing organic contents with paid contents (Geerardyn et al., 2000). For example, Yang and Ghose (2010) have looked into paid search to compare organic and sponsored search advertising. Kolsarici and Vakratsas (2010) have shown that brand-based advertising is more effective than category-based (generic) advertising. Following the literature, one can expect the sponsored contents on social media draw less engagement among users due to lack of credibility of the sponsored contents. This affects the creators as well as sponsors. Creators may experience user churn out as a result of credibility concerns.

The sponsorship effect might depend on the type of creator. Creators are heterogeneous in terms of type, size, and content quality. As a result, the preferences of their audience might be different. MacKenzie et al. (1986) have shown that the attitude toward the ad would impact the advertising effectiveness.

Internet sponsorship is typically brief text ads that helps to identify sponsor's brand and product (Rodgers and Thorson, 2000). Yet, the degree of sponsor presence may affect brand perceived image. User engagement may suffer from stronger brand presence. I posit that users engage less in case of stronger presence of the sponsor. I measure strong sponsor

presence by identifying the contents with the brand name mentioned in their description. Presence of the sponsor, on the other hand, can have heterogeneous effects by the relevance of the sponsor and creator. Rodgers (2003) found that relevance in online sponsorship leads to stronger persuasiveness and long-term effect. Research on event sponsorship has revealed the positive effect of strong association between the company and the sponsored event (DÁstous and Bitz, 1995; Stipp and Schiavone, 1996; Stipp, 1998; Gwinner and Eaton, 1999). This could be also related to consumer trust in the sponsor (Xiao and Benbasat, 2011; Porter and Donthu, 2008).

In summary, the extant research has shown that brands can take advantage of social media through sponsorship of online contents that boost their image and purchasing intention. In addition, it is shown the effectiveness of sponsorship can be influenced by sponsor-sponsee relationship. To the best of my knowledge, the nature of such relationship on social media is not investigated.

2.3.2 Ad Avoidance

Traditional advertising (e.g. TV commercials) have been susceptible to the customer ad avoidance. That is, customer skip the advertising by zapping (switching) or zipping (fast forwarding) according to Danaher (1995). There is a rich literature on how the avoidance happens (Schweidel and Kent, 2011, 2010), and how it can be affected by other factors such as prior exposure to the brand (Siddarth and Chattopadhyay, 1998). For example, Schweidel et al. (2014) studied how product placement can affect the ad avoidance in TV commercials through synergy effect.

One of the sources of ad avoidance is consumer skepticism toward advertising (Baek and Morimoto, 2012). Obermiller and Spangenberg (1998) defined consumer skepticism as tendency toward disbelief of advertising claims. Obermiller et al. (2005) found that more skeptical audience, consume advertising less, but they are more positive toward emotional advertising rather than informational appeals. Foreh and Grier (2003) outlined two types of skepticism: Situational (distrust of creators motivation) and dispositional (extant suspicious

of other peoples motives). Many studies suggest that ad skepticism may be influenced by the context and implication strategies (Kelly et al., 2017; Obermiller and Spangenberg, 1998; Obermiller et al., 2005). While, aforementioned studies focused on more traditional media, Kelly et al. (2010) studied ad avoidance on social media. They found that a user is more likely to avoid the advertising if the advertising is not relevant to the user and the user is more skeptical toward the advertising message.

The current studies suggest that users are likely to exhibit ad avoidance behavior, especially, when advertising is not relevant. I want to explore whether online sponsorship mitigates or exacerbate the negative effect of ad avoidance. In addition, I study what strategies can effectively increase the exposure of the brands in existence of ad avoidance.

Chapter 3

MOBILE ECONOMY: THE EFFECT OF MOBILE APPS ON ONLINE RETAILING

In online markets, product availability has dramatically increased compared to traditional brick-and-mortar vendors with limited shelf spaces. This has led to more product variety, giving more chance to niche products. The economic consequence of such phenomenon called "Long Tail" effect (Anderson, 2006) is the dispersion of the sales among products. In other words, there is less weight on the demand curve for the mainstream products ("hits" or "bestsellers"). As a result, customers have wider choices with alternatives suited better for their needs. Thus, shopping of customized products and services will be as economically attractive as the dominant products. The long tail has had significant impacts on distribution and supply chain especially with online channels which have helped to remove constraints of limited physical shelf spaces. Many retailers have sought the opportunities brought by new technologies to fulfill their customers craving and give them more competitive advantage. On the other hand, social welfare increases due to a less concentrated distribution of product sales. In a pioneer study, Brynjolfsson et al. (2003) find that consumer surplus increases by $731 million through increased product variety of online bookstores. Overall, the efficiency of the market will improve because of diversity.

The effect of online versus offline channels on sales concentration has been studied in various domains (Brynjolfsson et al., 2011, 2003). It is estimated that two thirds of in-store shoppers go on their smartphones at some point to check the prices online before a purchase at the store (Hartjen, 2016). Even though online channels may encourage sellers and shoppers to choose the niche items, the effect of different online channels is unknown. Along the mainstream online medium (desktop and personal computers), a clear majority of retailers

have offered their mobile applications for smartphones available in different platforms such as iOS, Android, and Windows. There are many anecdotal evidences suggesting that mobile sales channels are becoming mainstream in retail and digital marketing (Kumar, 2016; Rosner, 2017). With more people adopting smartphones, the prices of internet data for phones have been decreasing over time and people opt-in to unlimited data plans. Mobile users feel less barrier due to the growth and ubiquity of the mobile markets. Although the shift toward mobile channels seems inevitable, there are little knowledge on the effects of such market transforms.

There are both supply side and demand side explanations for the long tail effect in online channels. On the supply side, retailers are less constrained by physical shelf space and on hand inventory at stores. In transition to online channels, they can more freely operationalize a richer portfolio of products. On the demand side, customers are more likely to search and explore thoroughly for items online than in-store shopping. All together makes online channels less concentrated on sales for narrowly-targeted goods.

The aim of this paper is to study the effect of the long tail on online channels specifically desktop (and personal) computers and mobile applications and derive inference on the customer behavior and managerial insights for e-commerce businesses. To disentangle the effects of supply and demand side, I restrict the scope of my study to the customer side by focusing on the products available in both channels as it removes the marketing and operational effect on the availability of products. Namely, I study these two research questions:

1. Do mobile apps bring a longer tail on product sales (less concentrated distribution of sales)?

2. Do differences in customers search patterns on two channels contribute to the long tail effect? In other words, would customers tendency to spend more time on their mobile phones affect the likelihood of sales of niche products in App?

The first question studies the causal effect of App on long tail. That is, whether the unique characteristics of online channel (App and PC) has any impacts on the distribution

of sales. Customers surplus will change by choice of channel and this will affect long-term competitive advantage of retailers. The second research question highlights how behavioral usage patterns come into play when different patterns lead to different product choices: the usage patterns that are driven by the channel characteristics instead of customer heterogeneity. The theoretical models in the literature may suggest the effect of interest in either way. One theory compares the search costs in two channels (Ghose et al., 2012). Thus, due to higher search costs in mobile Apps, they may lead to more concentrated sales distributions. Also, studies suggest that users have shorter attention spans in mobile devices. The alternative theory contrasts the difference in users access time to mobile and PC (Han et al., 2016; Böhmer et al., 2011). Due to higher mobility, users have more access to their mobile devices. I hypothesize that the online users succeed the search costs on mobile phones by higher access time to collect more information about the products. Whether the effect of access time overcomes the search costs is an empirical question. My hypothesis is that buyers leverage more access to mobile devices, therefore, allocate more search time in App and overcome high search cost on mobile devices with smaller screens.

To measure the distribution of products sales, I use the models following the power law which work suitably in this context (Brynjolfsson et al., 2010b). In other words, these models are based on aggregate sales data and have a log-linear form. To analyze the long tail effect, I calculate the aggregate sales at product level on each channel. I find that the distribution of product sales in App channel is less concentrated than the one in PC. I show that the effect holds by including control variables and using matching methods. As I use secondary observational data, I take several identification strategies to draw a causal relationship from the data. Next, by applying the instrumental variable approach, I show that the number of searches, as a measure of search time, does have a positive effect on customers choosing more niche product. We, thus, conclude that App results to more diverse sales distribution by increasing the product search time. I aim to contribute on the economic consequences of mobile applications in online retails. To the best of my knowledge, there has not been a study exploring the long tail on different online channels. Moreover, it is extremely difficult

and costly, if not impossible, to perform a randomized field experiment in this area. Random assignment of channel while keeping everything else intact might not be feasible other than in laboratory environment. Thus, the findings of my study implicate insights for the online retailers on how value is generated through additional channels along the more traditional media. I also contribute on studying the mechanism of the longer tail in App channel. It sheds light on how customers pattern of internet usage would be different over different channels.

The rest of the paper is structured as follows. In Section 3.1, I develop my hypotheses by building up on the theoretical models existing in the literature. Section 3.2 introduces the empirical settings and the model formulation regarding to the causal effect of App on long tail (research question 1) and presents the results of the corresponding models. Section 3.3 formulates the model for the mechanism in which long tail happens in App (research question 2) and reports the results. Section 3.4 discusses the results and Section 3.5 concludes by implications and conclusion remarks.

3.1 Theory Development

I posit that the usage pattern in two channels are different. This is supported by the past literature (Shim et al., 2008; O'Hara et al., 2007; Ghose et al., 2012; Roto and Oulasvirta, 2005; Fang et al., 2015; Han et al., 2016). The difference can be characterized in terms of access time, search time, search cost and ease of use, time of day use, and attention span. The theoretical models in the literature may suggest the effect of interest in either way. For example, due to higher search costs in mobile Apps, they may lead to more concentrated sales distributions. Also, studies suggest that users have shorter attention span in mobile devices. However, other theories embolden the difference in users access time to mobile and PC. I hypothesize that online users leverage the access time and succeed the search costs on mobile phones to collect more information about the products. Whether the effect of access time overcomes the search costs is an empirical question. I formulate my hypotheses by exploring how App channel tends to sell more niche products.

To investigate the mechanism of the long tail effect, customers are assumed to be rational and seeking to maximize their received value out of the product which they buy. With the lack of ex-ante information, they should either experience the product by determining the quality of brands by purchasing brands and using them or search for the product quality (Nelson, 1970). Search products are defined as products that customers would seek information through their characteristics to evaluate the quality of the item. I are studying the product categories of fishing rods and chargers and power supply, and I observe that customers spend a good amount of time searching for items in these categories. On average, customer browses 12 products before shopping in my setting. While the concept of search and experience products is a coarse classification and there is no standard metrics to categorize products into these two groups, the framework can give more insight on the mechanism leading to the long tail effect.

I build my hypotheses on how App channel is more likely to sell more niche products. I hypothesize that buyers leverage the fact that they have more access to mobile devices, therefore, allocating more search time on App overcomes high search cost on mobile devices. Thus, depending on whether the access time overcomes the search costs, the results could be either way, making it an empirical question. Recommendation systems also matter. Depending on the firm strategy and recommendation algorithm, they may favor top selling products or niche products. In my setting, the platform uses the same system and there is no difference regarding how products are suggested.[1] As a result, recommendation system would be ruled out from the analysis. Figure A.1 in the appendix shows sample screenshots from ranked products in both PC and App in a search for a fishing pole.

[1]Products are ranked similarly in both channels.

3.2 Empirical Settings and Data

3.2.1 Data

I utilize a rich data set from a large-sized multi-branded C2C marketplace providing a platform for small businesses and individual sellers to cater all variety of goods to customers. My dataset includes 33,301 products in two categories: fishing rods, and laptop power supplies. Moreover, the dataset includes all the transactions in two time-periods each lasting for two weeks in August and October 2014. There are 1,376,407 transactions in the first period and 1,159,670 records in the second one. I choose the products in fishing rod category as the main category for my analyses, since there are a greater number of products and transactions in that category. I use the other category as robustness checks. More than a third of sales occur on mobile channel and the rest on the conventional desktop computers (including laptops). I obtain customer data and historical information on products sales. There are a few outliers in the data for the quantity of purchase, representing mass purchases. These transactions include very small portion of observations (0.06%). Since most purchases (95%) are single unit purchases, I exclude the outlier transactions with massive purchase amount from the main data set. We, however, analyze the model with the entire data set as a robustness check. Table 2 shows the list of variables along with their definition and summary statistics. Note that the summary statistics in Table 3.1 refers to fishing products for both data periods.

My sample is randomly drawn from the category of fishing rods. There are three online channels wherein customers can purchase the products. The first channel is the conventional desktop computers and laptops, hereafter PC. There is a mobile channel which can be divided into two categories: mobile application, hereafter App, and mobile browser (web). Mobile browser would be the mobile version of retailers website usually optimized for smaller screens, whereas mobile applications are vendor-provided software downloaded from the app stores and installed on smartphones. Retailers App that I study offers the same functionality as the website, including searching and browsing for products, recommended items, adding items to cart, choosing wish lists and placing orders. In my data, mobile web only accounts for 3% of

Table 3.1: List of variables and summary statistics

Variable	Definition	Count	Mean (SD)	Min	Median	Max
listing_price	Original listing price of item in Chinese currency (¥)	835,182	353.76 (512.69)	0.1	218	14,860
payable_price	Final price after adjustments to be paid by customer	835,182	377.05 (570.34)	0.1	228	43,800
discount	Discount applied	835,182	216.17 (478.49)	0	89	39,420
quantity	Number of same items purchased in transaction	835,182	1.18 (1.34)	1	1	50
buyer_domestic	Whether buyer is domestic	720,613	0.98 (0.18)	0	1	1
buyer_stars	Ratings of the buyer	720,613	3.16 (2.16)	0	3	15
buyer_female	Whether buyer is female	450,365	0.31 (0.46)	0	0	1
buyer_age	Buyer's age	450,365	32.58 (10.78)	16	30	114
buyer_tenure	Buyer's tenure since joined the platform in days	720,613	1018.31 (853.94)	0	804	4,154
buyer_total	Buyer's total historic purchase	720,613	164.65 (930.12)	0	56	200,449
impression	Buyers product searches	424,836	8.43 (13.78)	1	4	1781
buyer_impression app/pc	Ratio of buyers searches in app to pc	470,958	0.37 (0.45)	0	0	1
seller_domestic	Whether seller is domestic	835,182	1.00 (0.014)	0	1	1
seller_stars	Ratings of the seller	835,182	10.76 (2.89)	0	11	18
seller_tenure	Seller's tenure since joined the platform	835,182	812.61 (626.74)	1	569	4,029
seller_total	Seller's total historic sales	835,182	68798.56 (157975.3)	0	14,367	4,151,548

all transactions while App includes almost 30% of shares. A survey on mobile industry shows that Apps dominate the browsers in mobile usage by 90% of the time (Khalaf, 2015). Thus, I drop transactions performed in mobile web and consider only two channels: PC (personal computers) and APP (mobile applications). I did not differentiate between smartphones and tablets as I do not know what specific device is used in the transactions. Xu et al. (2016) find that tablets play complementary roles for smartphones and act as the substitute for PC

on e-commerce markets, suggesting that merging smartphones and tablets in my study is in line with the previous literature.

3.2.2 Product Selection

The website offers the same set of products on both channels. Sellers have no power on differentiating their offerings over channels. Prices or other product characteristics may change over time, but that change would reflect on both channels. This feature helps us circumvent the product selection issue in the analyses. I also abstract away from supply side effects on sales since the customers in both groups face the same set of products and availability. There is a large extent of product differentiation: 17,465 products are available in the fishing rods category. As a robustness check I also restrict the product sample to those items which have been sold on both channels in my data collection period. By applying this restriction, the product set is reduced to 11,163 products. By comparing product sales in two channels, I observe more unique products are sold in App (17,465 in App, and 11,163 in PC) as a model free evidence for the long tail effect in App. Product prices are the same through all channels. However, there are discounts offered by the platform, and the discounts may vary over the channels. There are two types of discounts. First, seller-specific discounts that are applied by the seller. Sellers cannot discriminate channels by the discount they apply as it will be applied on two channels similarly. Hence, there will be no systematic difference in pricing by the sellers. The second type of discounts are those that employed by the platform. The discounts may be offered in terms of general coupons specifically to promote a channel (e.g. mobile). However, the coupons can be used for all the products on the specific channel. Hence, that will not impose any systematic difference in pricing strategies. To ensure that there is no systematic difference between pricing policies, I run a t-test on the payable prices to check if there is a significant difference over two channels. The t-statistic is 0.53 implying the prices are not significantly different over PC and App. Figure 3.1 presents the empirical distributions of final prices in App and PC.

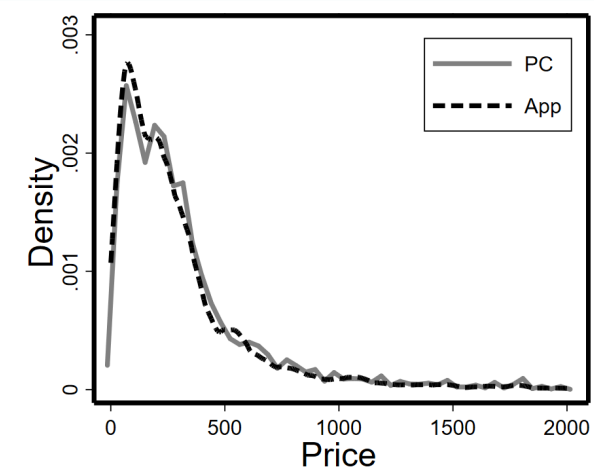


Figure 3.1: Kernel densities for final prices in PC and App.

3.2.3 Self-selection of Channel

One potential caveat to the analyses is that customers self-select the channel and might be systematically different. I try to disentangle the long tail effect for customer heterogeneity and channel effect as the long tail effect caused by channel characteristics is of interest in this paper. To address this issue, I take multiple approaches for the causal relationship. First, I restrict the sample to the customers who have browsed products on both PC and App. As using the App requires an installation of the application, customers must pay some initial costs in terms of time and effort to be able to use the App channel. I posit that once the buyers install the App, there is no switching cost between channels. By controlling over historic search behavior of the buyers on two channels, I believe that there is no systematic difference between the customers purchasing through PC and App in my experiment period. The only difference would be due to channel characteristics and interaction with products which is the effect of interest in this study. In addition, I use two alternative methods of matching: Mahalanobis distance matching (MDM) proposed and discussed by Cochran and Cochran and Rubin (1973) and Rubin (1976) and propensity score matching (PSM) method proposed by Rosenbaum and Rubin (1983) to address the potential selection of the

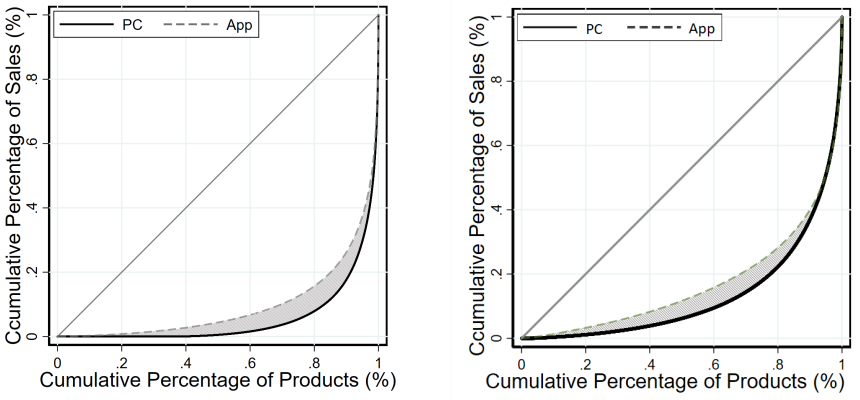
channel. Finally, I apply a control function (CF) approach (Heckman and Robb Jr, 1985) as a robustness check.

3.2.4 Aggregate-Level Model

To study the overall distribution of sales over the products, Brynjolfsson et al. (2010a) defined the long tail as an aggregate-level construct. In line to the literature, I calculate the aggregate sales at product level on each channel. I use Lorenz curve as a model free evidence to compare sales profile of PC and App. Figure 3.2 shows the Lorenz curve for PC and App for two product categories: fishing poles and chargers and power supplies. The curve closer to 45° line demonstrates more distributed sales and longer tail as a result. As Figure 2 demonstrates, the curve for App is closer to 45° line in both product categories and represents more distributed sales compared to PC. I also use Gini coefficients to quantify how far each curve is from the 45 line. Gini coefficients for fishing category are 0.81 and 0.83 for App and PC respectively. Lower Gini coefficient for App suggesting a less concentrated sales profile compared to PC channel. The gap in tail between PC and App represents 8.3% of total sales in App. The dollar value representation, for instance, is approximately ¥2,630,000 for two weeks in my data. Note that, this is just for a selected product category. Gini Coefficients for charger category are 0.66 and 0.73 for App and PC respectively. To check the significance of the difference, I follow the log-linear model as in Brynjolfsson et al. (2003). I obtain the rank of each product by sorting the products based on aggregate sales in a descending format. I define log rank as the log of rank introduced previously. I use the following model for both channels:

$$\log sales_j^c = \beta_0^c + \beta_1^c \log rank_j^c + \epsilon_{cj}. \tag{3.1}$$

Superscript c and subscript j represent channel and product respectively. Coefficient β_1^c measures how fast $\log sales_j^c$ decreases as a function of $\log rank_j^c$. Hence, a less negative coefficient (smaller in absolute value) would represent a longer tail. Next, to test if the dif-



(a) Fishing Poles (b) Chargers and Power supplies

Figure 3.2: Product sales concentration by Lorenz curves for PC and App.

ference is significant I introduce the *App* dummy showing whether that transaction occurred through the App channel. I pool all the aggregate data for both channels. I also introduce an interaction term for the *App* and $\log rank$ to check if App has any significant effect on the slop of increasing rank.

$$\log sales_j = \beta_0 + \beta_1 \log rank_j + \beta_2 App_j + \beta_3 \log rank_j \times App_j + \epsilon_j. \tag{3.2}$$

The coefficient of interest in Equation 3.2 is β_3 for the interaction term. If positive, it indicates that slope of rank in App is smaller than that of PC resulting in longer tail in App. The opposite would be true if the coefficient is negative.

3.2.5 Controlling for Buyers Demographics

Comparison of buyer and seller characteristics shows that the time of shopping is not statistically different (t-statistic=0.6) in two channels. In addition, the difference of buyer tenure time in the platform (time from the registration), buyer ratings, and all sellers characteristics are not significant over two channels. Figure 3.3 represents the distribution of tenure and

rating for buyer and seller in PC and App. Yet, there is a significant difference in buyer age and buyer gender (t-statistics are 82.0 and 44.3 for age and gender respectively). The

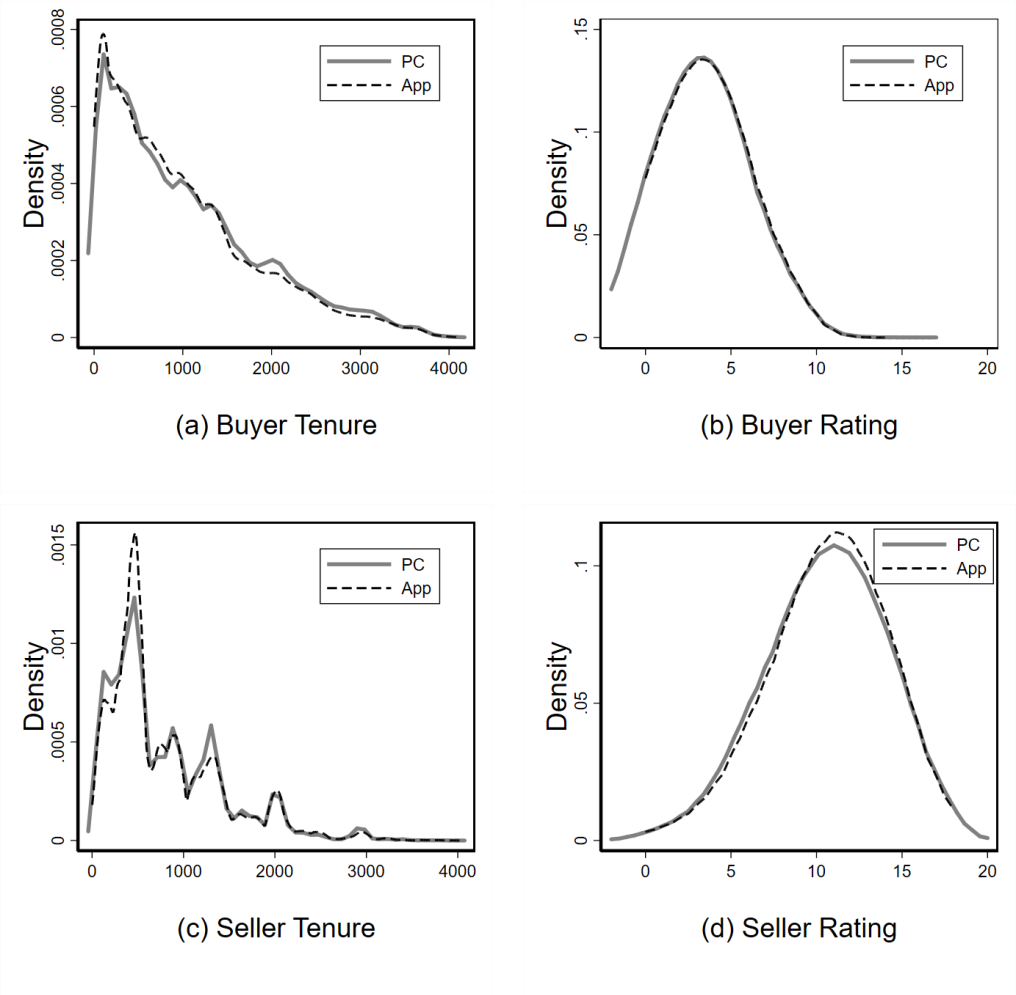


Figure 3.3: Distribution of tenure (days) and rating for buyer and seller in PC and App.

challenge of controlling for these demographics is that they are defined at the transactional level while my model is an aggregate level model. Therefore, I construct aggregate measures to control for age and gender. First, I use the dummy variable *female* to indicate the gender of buyers. Next, I create four age categories in which there is the approximately same number of buyers. Table 3.2 shows the age categories and their subsequent frequency. I can construct aggregate level data by variables App, age dummies, and buyer gender groups and the interaction terms. Then, I extend the model in Equation 3.2 by pooling transaction data

into the following model:

$$\log sales_j = \beta_0 + \beta_1 \log rank_j + \beta_2 App_j + \beta_3 \log rank_j \times App_j + \beta_4 female_j + \beta_5 \log rank_j \times female_j + \beta_6 Age_j^{25-29} + \beta_7 Age_j^{30-37} + \beta_8 Age_j^{38+} + \epsilon_j. \quad (3.3)$$

Coefficient of interest is still β_3 which captures the difference in long tail in two channels. The interpretation of the coefficient would be the same as in Equation 3.2.

Table 3.2: Age categories and frequencies

Buyer Age	Frequency	%
16-24	44,865	21%
25-29	54,806	26%
30-37	58,445	28%
38+	52,993	25%

3.2.6 Results

I estimate the models discussed in the Section 3.2.4 to test the first hypothesis. The question is whether App exhibits a less concentrated distribution of sales compared to PC. Table 3.3 summarizes the estimation results where the first two columns show the results for Equation 3.1 estimated for PC and App respectively. I see that the coefficient of interest is less negative in App ($\beta = -1.417$ for App and $\beta = -1.582$ for PC) showing that the slope in sales rank drops faster in PC. The results suggest that sales distribution is less concentrated in App. Also, the high R^2 represents a very good fit by the log-linear model. However, I cannot draw the causal conclusion based on this result as I do not know whether the difference is significant. Thus, models 3 and 4 demonstrate the pooled estimation results with the interaction term for App. Model 4 includes the products which have been purchased on both channels while other models use the products sold on any of channels (Note that all products are available in two channels). The coefficient of the interaction term of App and sales rank

is positive and significant (0.165). It shows that the difference slope of sales rank in PC and App is significant.

As a robustness check for the linear model, I also show the estimated results for quantile regression model in the fifth column. Model 6 summarizes the results for the set of customers who have used both channels prior my data. The effect is smaller compared to other models, but, it is statistically significant. The difference suggests that customer self-selection may exist. In the seventh column, the results of Equation 3 are shown. Controls for age and gender are used in the aggregate model. There is no significant difference on the main results showing that the findings are robust to the inclusion of the controls. Overall, the results of the all models are consistent and robust to model changes.

Table 3.3: Results of the long tail models

	(1) PC	(2) App	(3) Pooled	(4) Pooled (Overlap)	(5) Quantile	(6) Shared Buyers	(7) Controls
$\log rank$	-1.582***	-1.417***	-1.612***	-1.582***	-1.702***	-1.382***	-1.310***
	(0.006)	(0.005)	(0.006)	(0.006)	(0.004)	(0.007)	(0.004)
App			-1.640***	-1.853***	-2.550***	-0.576***	-1.240***
			(0.061)	(0.067)	(0.052)	(0.075)	(0.033)
$App \times \log rank$			0.188***	0.165***	0.246***	0.075***	0.116***
			(0.007)	(0.008)	(0.006)	(0.009)	(0.005)
$Female$							-1.276***
							(0.035)
$Female \times \log rank$							0.078***
							(0.005)
$Age^{(25-29)}$							0.391***
							(0.007)
$Age^{(30-37)}$							0.519***
							(0.007)
$Age^{(38+)}$							0.024***
							(0.007)
Constant	19.51***	17.66***	20.01***	19.51***	20.74***	16.69***	15.02***
	(0.051)	(0.044)	(0.47)	(0.048)	(0.038)	(0.055)	(0.027)
R^2	0.865	0.871	0.872	0.871		0.850	0.838

Standard errors in parentheses. DV is $\log sales$.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2.7 Endogeneity

I address the selection issue by restricting the customers who have purchased on both channels *prior* to my data collection (see the sixth column in Table 3.3). However, there might be still some degree of selection. I adopt different approaches to deal with the possible endogeneity by customers self-selection: mahalanobis distant matching (MDM), propensity score matching (PSM), and control function (CF). I utilize multiple approaches to show that the findings are robust to different endogeneity treatments as they have their own set of assumptions. On the one hand, matching methods, in general, leverage the implicit assumption that the differences are coming from observables: i.e. treatment assignment is independent of outcome given observables. Thus, matching provides balance in observed characteristics which reduces the endogeneity form the confounders and the model dependence aspect of results (King and Nielsen, 2016). I apply two methods of matching with different premises. First, I use MDM which approximates a more efficient fully blocked randomized experiment (details of the MDM method could be found in Carpenter (1977) and Rubin (1980)). Second, I apply a propensity score matching approach which attempts to achieve a completely randomized experiment. PSM has been widely used by many scholars from various fields (Abadie and Imbens, 2006; Dehejia and Wahba, 1999; Imbens, 2000; Lechner, 2001). On the other hand, control function approach attempts to model the selection problem through a set of statistical assumptions. I formalize an unrestricted control function approach which relaxes some of the assumptions in the restricted method. Technical details about the methods of matching and control function are presented in the Appendix.

I use buyer demographics and past behavior to identify the matched samples. For demographics, variables *buyer_age*, *buyer_female*, *buyer_domestic*, *buyer_total* and *buyer_stars* are used. I also use total number of impressions for each user in any channels and the ratio of App usage to PC usage to control for other unobservable factors that may reflect in past behavior. I keep the matched observations and conduct the aggregate analysis as in Equation 3.2. Table 3.4 reports the results of the aggregate models using the matched sample.

Through matching, I achieve high balance in my treatment (App) and control (PC) groups. Appendix D explores details about the matching results including sample balance, sensitivity analysis, and other settings.

Table 3.4: Matching results of long tail

	(1) Matched PC	(2) Matched App	(3) Matched Pooled	(4) PSM Pooled
$\log rank$	-1.614***	-1.425***	-1.614***	-1.523***
	(0.006)	(0.004)	(0.006)	(0.006)
App			-1.658***	-1.249**
			(0.061)	(0.067)
$App \times \log rank$			0.190***	0.135***
			(0.007)	(0.008)
Constant	20.04***	18.38***	20.43***	18.71***
	(0.052)	(0.037)	(0.047)	(0.050)
R^2	0.84	0.84	0.84	0.86

Standard errors in parentheses. DV is $\log sales$.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Furthermore, I estimate propensity score matching model as described earlier. I apply the nearest neighbor method in matching with logit propensity scores. The estimated results are shown in the fourth column of Table 3.4. The effect is down to 0.135 and statistically significant. Table D.3 in Appendix presents the propensity score from the logit model. I verify the overlap condition and balance of the sample through graphical and statistical measures. I report other technical details about PSM and differences in the two matching techniques in Appendix D.

Lastly, I present the results from the control function approach. I follow Vella and Verbeek (1999) by developing an unrestricted CF approach in which the variance of error term in App can be different from the variance of error term in PC. I formulate the selection approach by using an aggregate model at product level. That is, product sales are aggregated over channels and the demographic information of buyers are averaged through the aggregation

process. Then, I use the average demographic information for each product (and channel) to formulate the main regression model. Table 3.5 summarizes the results of restricted and unrestricted CF approaches. The parameter for the interaction term is positive and significant in both models suggesting that App results in longer tail in aggregate product sales. Note that the estimated parameter for Inverse Mills Ratio (IMR) suggests that the selection exists.

Table 3.5: Control function results of long tail

	(1) Restricted CF	(2) Unrestricted CF
$\log rank$	-1.532***	-1.523***
	(0.006)	(0.005)
App	-1.617***	-1.605***
	(0.062)	(0.062)
$App \times \log rank$	0.174***	0.170 ***
	(0.007)	(0.007)
Inverse Mills Ratio	0.030**	
	(0.009)	
σ_{App}		-0.050***
		(0.012)
σ_{PC}		-0.193***
		(0.017)
Constant	19.44***	19.41***
	(0.010)	(0.047)
R^2	0.89	0.89

Standard errors in parentheses. DV is $\log sales$.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2.8 Robustness Checks

As robustness checks, I replicated the estimations using the second batch of data sampled in October 2014. The data is in the same product category and only consists of 5% of customers who purchased during the first period. So, if the results are robust, I can expect the effect to be robust to the selection of the customers. I get similar results as my main findings with

the original data. I also implemented the analysis for another product category (charger and power supplies), and I got the results with similar signs, although the magnitudes are different: the long tail effect in charger category is stronger. Additionally, in CF approach, I use median as the aggregation measure for demographic information of buyers. The results suggest the main finding of the model is robust to the aggregation specification of users demographic information.

As the last step, I exploited different propensity score matching strategies. Brynjolfsson et al. (2011) propose another approach of using PSM to compare the likelihood of niche products on two channels. They define a cut-off by which products are being divided into niche and top products. Then, they use the propensity score matching to identify if the treatment (in this paper, App) has a significant effect on choosing niche products. I divide the products by median in both channels. Summary of product distribution after dividing is reported in Table 3.6. The results for the average treatment effect after using propensity score matching is shown in Table 3.7. The App has the coefficient of 0.0186 leading to increased likelihood of niche products sales in App. Finally, PSM is used while using all the buyers variables. The results are robust to such changes.

Table 3.6: Summary of niche products in each channel

	Average Unit Sales	Average Price (¥)
PC Channel		
Top 50%	58.85	249.02
	(3.68)	(5.23)
Bottom 50%	5.31	83.79
	(0.29)	(1.18)
App Channel		
Top 50%	35.26	262.69
	(2.91)	(5.70)
Bottom 50%	3.99	79.68
	(0.18)	(0.89)

Standard errors in parentheses.

Table 3.7: App effect on niche product sales

	Niche Product
App	0.0186***
	(23.2)
Constant	0.0615***
	(125.05)
N	423,526

t-statistics in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.3 Search Time Effect

I now explore the mechanism through which the long tail effect happens in App to address my second research question. The long tail effect shows that the overall distribution of sales is less concentrated in mobile channel implying that niche products in the tail are more likely to be purchased. My identification strategy and estimations show that this effect is resulted by channel characteristics rather than the unobserved heterogeneity of users; That is, usage patterns unique to mobile channels shape the distribution of sales. There are two important factors defining mobile user patterns (Ghose et al., 2012; Shim et al., 2008): First, search costs in phones are high regarding the smaller screens of smartphones compared to PC. The costlier browsing limits the ability of users to scroll or search in accessing products on the bottom of the ranking (aka *niche*). Second, mobile users have the luxury of more access time to the internet via the smartphones. Studies have shown that users spend more time using mobile Apps but in more interrupted sessions (Böhmer et al., 2011). These two aspects shape the user behavior in search and purchase of the products. While higher search costs and shorter users attention span in Apps can result in sales shorter tail, more access time may increase the likelihood of niche products sales. I argue that App users leverage the fact that they have more access time to their smartphones and this will let them overcome the

high cost of search costs in phones. That is, customers search more frequently to gain more information in lack of knowledge about brands quality and features.

It is worth noting that cross-channel switching behavior of customers might affect the results. That is, users search in App for browsing and go to PC for the shopping or vice versa. Since I have the browsing history of the customers in the period of August 2014, I can do the analysis for those who do not show switching patterns as a robustness check. By doing that, the results do not change in the sign showing the estimated effect is robust to cross-channel switching.

3.3.1 Product Search and Impressions

I employ the number of impressions (number of times an item is searched or browsed) in App and PC representing the frequency of access as a proxy for search time. As model-free evidence from data, products obtain more impressions in App. On average, there are 514.1 impressions for each product in PC and 742.3 in App (t-statistic=-5.2). That implies the same product in PC is visited 1.5 times more in Apps. The impressions are collected at the same time as my data. Figure 3.4 shows the distributions of impressions in App and PC. The distribution is truncated for the products with more than 1000 impressions. Furthermore, the number of impressions per buyer is 25.7 on average for PC, and 43.4 for App (t-statistic=-51.4). Hence, this difference suggests that even though PC channel accounts for more transactions, yet customers search more in App. I build a model to measure the effect of search time on the likelihood of niche product to be purchased. Following Brynjolfsson et al. (2011), I divide the products into the niche and top-selling groups by their overall sales and whether the aggregates sales are above the median. To test my hypothesis, I build a probit model in which the dependent variable is a binary outcome (0 niche, and 1 top-selling). I posit that as customers search more products, the likelihood of buying a niche product increases.

The independent variable, number of impressions, might be endogenous. That is, choosing a niche product is correlated with the number of searches for that product. However, the

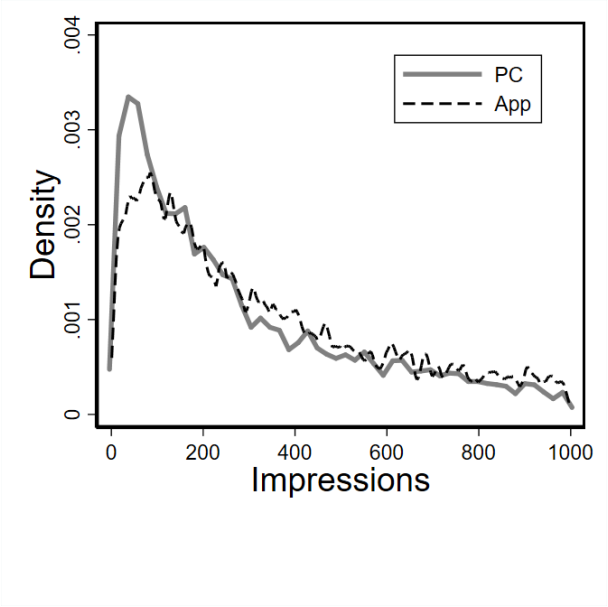


Figure 3.4: Distribution of product searches (impressions) for App and PC

endogeneity issue is expected to make the results downward biased. In other words, I expect top-seller products to be viewed and searched with a higher likelihood since they are either high quality or favored by the recommendation systems. Thus, the endogeneity issue is likely to make the results more conservative. As there might be other unobserved confounders, I introduce my identification strategy to address the potential endogeneity problem.

3.3.2 Instruments and Identification

To identify the effect of interest, I take advantage of the instrumental variables. I utilize three sets of instruments. The first instrument is defined as the number of searches for other products by the customer in the period of shopping. It will correlate with the number of searches for the product. That is, over the distribution of searches, customers characteristics act as shifters. Nevertheless, the number of searches for other products is uncorrelated with the sales or popularity of the target product since the two products might be in different

categories. This strategy helps us identify unbiased estimates for the variable of interest. Moreover, I use the aggregate number of historical searches of customers over two channels. This would be the sum of all product visits so far. Yet again, these variables are relevant instruments as they can shift the number of searches (impressions) of products. I argue for the validity of the instruments as they only change the likelihood of the target product being niche through the number of impressions. Note that these instruments are aggregated over channel making them unrelated to the target product. Table 3.8 summarizes the correlation structure of the instruments and the endogenous variable.

Table 3.8: Correlation structure of the instruments and log(impressions)

	Log(impressions)	Log(impressions others)	Previous PC searches	Previous App Searches
Log(impressions)	1			
Log(impressions others)	0.354	1		
Previous PC searches	0.095	0.381	1	
Previous App Searches	0.444	0.854	-0.0845	1

3.3.3 Model

Since there is enormous heterogeneity in customers number of searches for products, I use the log transformation of the number of impressions. I use the generalized method of moments (GMM) to estimate the probit model with IV. Equation 3.4 outlines the model specification:

$$\Pr(niche_{ij} = 1) = \beta_0 + \beta_1 \log impression_{ij} + \Theta X_j + \gamma Z_i + \epsilon_{ij}. \tag{3.4}$$

Equation 3.4 indicates the probability of customer i choosing product j as a niche product. Z_i is the vector of buyer demographics and usage patterns, and X_j is the vector of product characteristics. In the choice model analysis, I utilize text mining techniques to extract product features from the descriptions. I characterize 18 features which can be categorized in 6 subgroups including material, type, quality, feature, accessories, and length. Since I do

not have access to characteristics of the products, these extracted features can increase the power of the analysis. I control for these features in the model to show how search time will affect the sales distribution. Appendix C explains how I get these features and list the summary of extracted characteristics. As noted, a niche product is ranked below the median in the overall sales rank. I use different cut-offs in niche product definition as robustness checks. The coefficient of interest is β_1 which shows the effect of the number of searches on the probability of a niche product being chosen. The moment conditions are defined as follows:

$$\mathbb{E}\left(\log impression_i^{others} \times \epsilon_{ij}\right) = 0 \tag{3.5}$$

$$\mathbb{E}\left(Previous\ PC\ searches_i \times \epsilon_{ij}\right) = 0 \tag{3.6}$$

$$\mathbb{E}\left(Previous\ App\ searches_i \times \epsilon_{ij}\right) = 0 \tag{3.7}$$

I now explore whether the search time effect is moderated by the channel choice. Thus, building on the model in Equation 3.4, I include the interaction term of App and impressions presented in Equation 3.8. In this model, the coefficient of interest is β_3 which captures the interaction of channel and impressions on niche product selection.

$$\Pr\left(niche_{ij} = 1\right) = \beta_0 + \beta_1 \log impression_{ij} + \beta_2 App_{ij} + \beta_3 \log impression_{ij} \times App_{ij} + \Theta X_j + \gamma Z_i + \epsilon_{ij}. \tag{3.8}$$

With the potential endogeneity of the impressions, the interaction term may also be biased. Hence, I need to treat the endogeneity of the interaction term. However, conventional linear IV regression models cannot solve this problem and result in inconsistent estimates (Wooldridge, 2010). In addition, t-statistic from this regression is generally invalid, even asymptotically. The advantage of GMM model is that the efficiency of the model never falls by adding more non-linear functions of the exogenous variables to the IVs. Consequently, by including additional non-linear functions of exogenous instrument variables (e.g. through

interactions), I may consistently estimate the model through GMM model.[2] Therefore, I can identify the mechanism in which using App leads to the long tail effect.

Finally, I do further analyses on buyer demographics and product characteristics to study the moderation effects on the niche product selection through the search time. That is, I include interaction terms for variables including *buyer_female*, and *buyer_age* with impressions in the model and estimate it with GMM approach similar to the estimation of Equation 3.8.

3.3.4 Results

Table 3.9 presents the estimation results for Equation 3.4. First column indicates the results for the probit model without using the instruments. Second column presents the IV model results. As shown, the coefficient of $\log impression$ is positive and significant (0.136) which means that the number of impressions has a positive effect on the niche product selection. The Wald test of exogeneity is rejected by χ^2 value of 615.42 implying the variable of interest is indeed endogenous. By comparing the results in columns 1 and 2, I find that the effect of impressions is downward biased in the probit model as expected. The probit (without IV) shows that high searched products are less likely to be niche. This is likely caused by reversed causality. However, by utilizing the instrument variables, I get a positive coefficient for the impressions. This suggests that as customers searches increases, s/he is more likely to purchase a niche product. The caveat is that this is an average effect in both channels. To check whether the effect is different in App and PC, I use the model in Equation 3.8.

I present the estimation results of Equation 3.8 in Table 3.10. First column shows the model for the channel effect with App interaction term. Second column summarizes the results for the model with interaction of the product characteristics. The estimated coefficient for the App interaction term is 0.283 and statistically significant suggesting that increase in searches (impressions) in App will result in higher probability of a niche product being

[2] In theory, I never lose efficiency by adding more no-linear functions of IVs. However, this might not be true in practice, as adding too many instruments may be problematic. For more discussions on the asymptotic of this, see Wooldridge (2010), p. 268.

Table 3.9: Probit and GMM estimates of the effect of searches on niche products

	(1) Probit	(2) GMM-IV
log *impression*	-0.100***	0.136***
	(-19.86)	(13.03)
buyer impression app/pc	0.335***	0.043**
	(28.33)	(2.62)
buyer_domestic	0.099***	0.080**
	(3.70)	(3.05)
buyer_stars	0.074***	0.066***
	(25.91)	(23.27)
buyer_female	-0.161***	-0.117***
	(-15.26)	(-11.09)
buyer_age	0.0001	0.002***
	(0.29)	(3.40)
buyer_tenure	0.00003***	0.00002**
	(3.54)	(2.95)
Constant	-1.868***	-2.096***
	(-56.19)	(-62.66)
N	185,755	185,755

t-statistic in parentheses. The standard errors are heteroskedastic-consistent.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

purchased. Thus, my hypothesis regarding the customers usage pattern and the likelihood of choosing niche product is confirmed. In other words, the mechanism resulting in long tail in App is the increased search time (possibly through higher access time in mobile phones). I also find that gender and age play a role in selection of niche products. That is, women are more likely to choose a niche product as result of increased search compared to men. I observe similar effect for buyers age suggesting that older buyers have higher probability in choosing niche product due to item search. However, the coefficient for age is much smaller than that of gender suggesting the marginal effect of age interaction is less than the marginal effect of female interaction. On the product level, the features of *Long Rods* and *Adjustability* represent a positive moderating effect. This may suggest that those niche products with more

unique features will be more benefited by increased searches due to mobile usage patterns.

Table 3.10: Estimates of the channel effect

	(1) Channel Effect	(2) User and Product Effect
$\log impression$	0.119***	-0.648***
	(8.37)	(-7.50)
App	-0.595***	
	(-8.76)	
$App \times Logimpression$	0.283***	
	(9.93)	
buyer_impression app/pc	-0.091**	0.360***
	(-2.86)	(23.73)
buyer_domestic	0.077**	0.086**
	(2.76)	(2.77)
buyer_stars	0.073***	0.056***
	(23.78)	(17.06)
buyer_female	-0.126***	-0.285***
	(-11.17)	(-12.38)
buyer_age	0.002***	-0.016***
	(4.36)	(-5.39)
buyer_tenure	0.00003***	0.000005
	(4.36)	(0.71)
$buyer\_female \times \log impression$		0.110***
		(7.93)
$buyer\_age \times \log impression$		0.012***
		(6.11)
$HighLenghtPole \times \log impression$		0.101***
		(6.48)
$Supplemntary \times \log impression$		0.060***
		(3.59)
$Adjustable \times \log impression$		0.032
		(1.73)
Product Characteristics		✓
Constant	-2.154***	-0.255*
	(-56.05)	(-2.50)
N	185,755	185,755

z-statistic in parentheses. The standard errors are heteroskedastic-consistent.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.3.5 Robustness Checks

As a robustness check, I use the total number of impressions over two channels for customers as the independent variable instead of the separate number of impressions per channel. The result is consistent with the main results ($\beta = 0.100$, se=0.010). Moreover, in defining niche product, the issue of choosing a cut-off might seem arbitrary and one may argue that products cover a wide spectrum rather than only top and niche products. I apply different cut-offs for defining the niche products (40% and 60% ; 30% and 70%). The results remain consistent with the previous model. Table A8 in appendix presents the estimation results for the new threshold.

Lastly, the results presented in Tables 3.9 and 3.10 are performed on the unmatched sample. As robustness checks I run the same analyses on the matched sample. I find similar results on the analyses and conclude that the findings are robust to sample specifications. I present the details of estimation in Table F.2 in Appendix.

3.4 Discussion

In Section 3.2, I find that there is a causal relationship between channel selection (e.g. App) and sales distribution. I find that sales distribution in App is less concentrated compared to that of PC (long tail effect). That is, niche products are more likely to be purchased in App. Long tail can happen by various factors; Unobserved preferences of customers may lead to such effect in App. Through multiple identification strategies, I disentangle customer inherent differences and supply-side effects with the channel effect. Furthermore, by extensive robustness checks, I conclude that the results are not driven by my choice of functional forms or estimation approach. For instance, I analyze different selection of products: pooled set of products, and products sold on both channels. Also, I show that the findings are robust to selection of customers by running the model on the entire pool of customers and selected customers with historical purchases on both channels. Consequently, the findings suggest that the long tail effect is caused by channel characteristics. Thus, there is a specific feature

of App induces these behavioral patterns.

I find that search time plays a causal role on product selection. That is, buyers who spend more time (and more product searches) are more likely to purchase a niche product on both channels. Additionally, I show that App has a positive moderation effect on search instances. It means, buyers are more likely to purchase a niche product through App by more product searches. The finding has important managerial implications for retailers and App developers. First, the economic evolution of retailing may be impacted by new online channels like App. By changing the overall sales distribution of products, long tail phenomenon may require different paradigms in supply chain management. Second, the key finding in the purchasing behavior is that users try to spend more time to gain more information about the products. With additional information, buyers tend to have more options in their choice sets. The App developers can consider this to facilitate the process of information collection by customers. Also, marketers can leverage some of the findings as well. For instance, I find that, on average, women and older users are more likely to purchase a niche product by more searches. It could be used as a targeting method for marketing campaigns in retail industry.

The long tail and search time effects may be intensified for the products whose features and latent quality are important for customers (search products). One might expect different effects for the more conventional or fast-moving products calling for additional researches in this area. Also, I acknowledge that the studied mechanism might not be the only causal relationship that happens in practice, as there might be other features of App playing a role in long tail. However, I may not be able to identify other mechanisms by the data in hands.

3.5 Conclusion

With the emergence of online channels, more niche products are sold compared to brick-and-mortar channels, and it has resulted in more product variety. The "Long Tail" effect has distributed the sales among more products, reducing the importance of the mainstream products and bestsellers. As online market grows, more sellers become aware of the oppor-

tunities and benefits of multi-channeling and sell their goods through websites and mobile applications as well as offline stores.

Mobile applications are becoming the mainstream channel in retail and digital marketing. I study the distribution of product sales on online channels and compare the effects of the long tail in PC versus mobile Apps. Because searching for products on small screens is more time consuming than browsing through a PC, I hypothesized that online users allocate more search time on the Apps and overcome the cost of searching on small screens. I apply different models to show that the long tail effect consistently exists in App channel. I also investigate the mechanism of selecting more niche products in App. Previous studies have shown that mobile users spend more time on their phone leveraging the higher access time of mobile devices. That leads to an increased time of product search. Hence, I incorporate the search time variable to show whether it causes for longer tail. In the absence of randomized experiments, I take several identification strategies to draw a causal relationship from the observational data. I introduce an instrumental variable approach to identify the effect of number of searches on choosing the niche product.

My findings show that the distribution of product sales in App channel is less concentrated than the one in PC. This provides an important insight for retailers: introducing mobile App channel increases the variety of the products. This study also provides insights into the mechanism through which the longer tail happens: More searching compensates the high search costs and increases the value for the customers by providing more information about product details. Due to customers quest for more information about products through search time, my findings suggest that it is crucial for businesses to invest in technologies which make it easier for customers to gain more information about products and services.

My paper adds to the literature by answering questions about the overall sales concentration in mobile Apps. The economic implication of such phenomenon is that the customer surplus will increase through mobile channel by adding more variety to the choice sets. In addition, my study utilizes the customer behavioral theories to understand the mechanism of the long tail in App. It enriches the knowledge of how customers use their phone in the

online retailing context. Also, retailers and App developers can utilize the usage pattern to design Apps. My paper contributes to the literature by investigating the economics of Apps as sales channels and linking to customers behavior in online markets.

Future research can be done to address some of the limitations of the paper. For example, research on other product categories add insights on how the effect is different from the one I discussed. Also, due to data limitation, I cannot investigate any other effects for channels. Use of panel data might mitigate some of the barriers in identifying those effects. One interesting extension could be studying the supply side effect which requires the extensive information about the sellers. I acknowledge that the results might be context-specific, which future researches can shed light on the generalizability of the results.

Chapter 4

ONLINE CROWD: DESIGN OF CROWDFUNDING CAMPAIGNS

With the emergence of online platforms, crowdfunding has quickly become a major source of fundraising for creators, designers, and other entrepreneurs since its inception by eliminating the hassle of search for investors, financial supporters, and venture capitals. Among different types of platforms, reward-based crowdfunding provides the opportunity for the entrepreneurs to gain individual support in exchange for rewards to the backers. While the emphasis of the crowdfunding platforms has been on the novelty of the projects, the economics of the fundraising process is mostly overlooked by the scholars.

Many of the scholars in this domain have focused on consumer behavior in fundraising behavior. But there is not much knowledge on how these factors drive the demand for creative projects. I take a new perspective by considering the reward-based platforms as two-sided markets in which the entrepreneurs are the suppliers and the demand structure is characterized by the backers. The notion of market allows us to apply supply and demand theory in economics to infer about causal relationships that shape the market.

Despite the notable growth, crowdfunding remains a young field. The competitive landscape continually changes as the number of crowdfunding platforms is staggering. The growth in the crowdfunding market does not always come painlessly. Some of the entrepreneurs learn the tricks of the market in the hard way. Over 65% of the projects debuted on Kickstarter, failed to fulfill their goal by various reasons (over $ 347 million to unsuccessful projects). The rate goes beyond 80% for some sub-categories. There are, yet, important questions regarding the design of a project. Most experts assert that the innovation and the quality of the projects would not always translate to success of fundraising campaigns. One of the

important features of reward-based crowdfunding is the design of the reward scheme and pricing the products.

Platforms often guide creators on how to make a promising campaign. However, there is not much knowledge on the effect of reward scheme design and pricing on demand. A possible avenue to understand the mechanism is to look at the crowdfunding platforms as two-sided markets in which entrepreneurs can seek funding for their ideas. While, backers look for exciting opportunities to support. Hence, I can apply classic supply-demand economic theories and investigate the structural characteristics of the transactions. As a result, I treat the reward-based crowdfunding mechanisms as general online selling markets with sellers and buyers. There are various aspects in shaping the demand structure: project novelty, characteristics of offerings, pricing, campaign design and competition effects. I formulate the demand function based on the characteristics of this market. The goal is to estimate the demand of the technology projects, infer about switching patterns, and get insights about the effect of different reward levels on consumer welfare.

I focus on the demand side and employ aggregate level data for Kickstarter projects in technology category. I follow discrete choice framework and build a structural demand model. Since these fast-paced markets have unique characteristics, I must tailor the model specification to be suitable for crowdfunding platforms. The first difference between the online crowdfunding platform and the traditional markets is the notion of the product. It is typical to apply some sort of product aggregation in more mature markets. For instance, in the auto industry, scholars often aggregate products with over different trims (Berry et al., 1995; Petrin and Train, 2010). However, such approach is not feasible in crowdfunding. Different rewards of a project may offer totally various values. Thus, I do not perform any aggregation on the product-reward level. This makes the computation more intensive, but gives enough flexibility to understand the utility values for the rewards. The other difference is that the product features are not well-defined in crowdfunding platforms. That is, the information about projects (and rewards) may not be summarized in a structured fashion. Therefore, I use natural language processing techniques to convert the unstructured data

provided by project description into dense vectors representing the product features. Lastly, the competition structure may be different in such platforms as opposed to the traditional markets. There are products form different categories (therefore, different nature) to compete to attract pledged money from the pool of backers.

I use latent instrument approach to identify the model and address the potential endogeneity of the price. In this approach, I assume that the endogenous variable is comprised of two parts: exogenous and endogenous parts. I allow the endogenous part to correlate with the error term through unobserved characteristics of the project-reward. I validate my statistical assumptions though experiments using the data set. I also model heterogeneous users with different tastes (i.e. price sensitivity). I find that the price coefficient is indeed biased towards zero implying that there is positive correlation between price and the unobserved characteristics. I use different numbers of latent classes for price. I find that the model with five classes results the best fit and matching price distribution. I perform further analyses on the price elasticities and I conclude that technology sub-categories such as *apps*, *software*, and *web* provide the most market power (less own-price elasticity in absolute value) for the entrepreneurs. Through policy simulations, I show that the low-level reward tiers may cannibalize the higher rewards and lead to a decrease in the pledged money as a result. In addition, welfare analyses verify that the average customer surplus offered by the low-level reward tiers are lower compared to the other rewards. This will give more insights on how platforms should design their guidelines to increase the surplus. My results help campaign creators better understand the effect of campaign design on demand.

4.1 Empirical Setting and Data

The two major crowdfunding platforms have been running for about a decade, with Indiegogo (2008) first and Kickstarter (2009) following shortly thereafter. I use data from Kickstarter,[1] the leading reward-based crowdfunding platform worldwide. Kickstarter has been the most

[1] www.kickstarter.com

successful platform in terms of the raised funding.[2] It also leads its competitors in success rate by large margins.[3] For instance, the success rate in Kickstarter is 35% , whereas, it is around 10% in Indiegogo. There are several characteristics that makes Kickstarter more suitable for this study. First, Kickstarter has *all-or-nothing* policy meaning that the campaign creators can collect the pledged money only if the project is successful. Otherwise, the project fails at fundraising and the pledged money would go back to the backers. *success* is defined whether the campaign hits the project goal by the total money raised. According to Kickstarter, this policy would motivate the creators to set the goal as the minimum required to kick off the projects. If the target is too low, the creators might not have enough money to begin with and the project may fail. On the other hand, higher project goal would decrease the chance of *successful* fundraising. The funding goal and deadline cannot be changed once the project has launched. This feature likely prevents creators' strategic behavior in goal settings.

Second, Kickstarter has a project review policy by which they pre-screen the projects before the launch. The purpose of this policy is to ensure that the projects offer substantial value and have proper design. The platform requests for additional review process for approximately 60% of proposed projects.[4] This process is done manually by the staff members. They provide feedbacks and consultations to the reviewed projects. While, this policy has enabled the platform to offer quality projects and gain reputation in the growing competition, it requires insights on how likely a project can be successful.

Third characteristic of the platform is that the major part of the demand comes from the organic pledges directly from the platform. This feature rationalizes the idea of marketplace for the platforms. However, top successful projects enjoy the lift from social media, and local news more substantially. But that is not the case for most of the projects. For instance, while

---

[2]Projects in Kickstarter have raised over $ 3.5 billion as of January 2018 according to `www.kickstarter.com/help/stats`.

[3]Kickstarter is said to have as six-times market share as its closest competitor Indiegogo, according to `www.crowdfundbeat.ca/statistics-comparison-kickstarter-vs-indiegogo/` and `www.krowdster.co/blog/kickstarter-still-indiegogo.html`.

[4]`www.kickstarter.com/blog/how-projects-launch-on-kickstarter`

Facebook is the main source of promotion activities for Kickstarter projects, Kickstarter receives only 5% of its traffic redirected from Facebook according to alexa.com (ranking websites by traffic). In addition, Kickstarter has relatively loyal backers with 30% of the backers have supported more than one project. That is, perhaps, one of the reasons that why the platform experiences large organic pledges.

Kickstarter caters various categories including arts, crafts, dance, design, technology, etc. I collected data from technology design and products category, technology hereafter, which has multiple sub-categories. I focus on the technology category because the idea of marketplace matches the selected category, as there are often tangible products being traded in the platform. Other categories in the platform may not fully belong to the marketplace idea. The projects are highly diverse with broad topics and designs. The data includes all the projects in the Kickstarter starting from 2009 to the September 2017. I restrict the sample to those projects created in the United States. Hence, I focus on the US-based projects to remove the international interaction effect, if any. However, I allow for the projects, created in the US, to raise money from other countries that Kickstarter is active. After removing projects with missing data and non-English data, my sample includes 15,676 projects. In total, there are 105,821 reward tiers available in my data set. Figure 4.1 shows the number of projects per category and per year.

Table 4.1 provides the list of variables with their descriptions. There are 13,180 creators who, on average, created 1.6 projects. Creators are active members of the crowdfunding community in which they have backed 3.9 projects on average. More on the project level information, the overall success rate is 26% which is slightly lower than Kickstarter average success rate. As defined earlier, *success* is when a project hits its fundraising target. There is considerable heterogeneity in project goals. Most of the funded projects on Kickstarter have goals between $ 1,000 and $ 10,000. In technology category, however, goals are higher, as the median is $ 80,000. Figure 2 shows the histogram of the projects with goal below $ 10,000, along with histogram of the reward prices under $ 500. Each project has a deadline to achieve the fundraising goal. The deadline is set at the design stage and cannot be changed

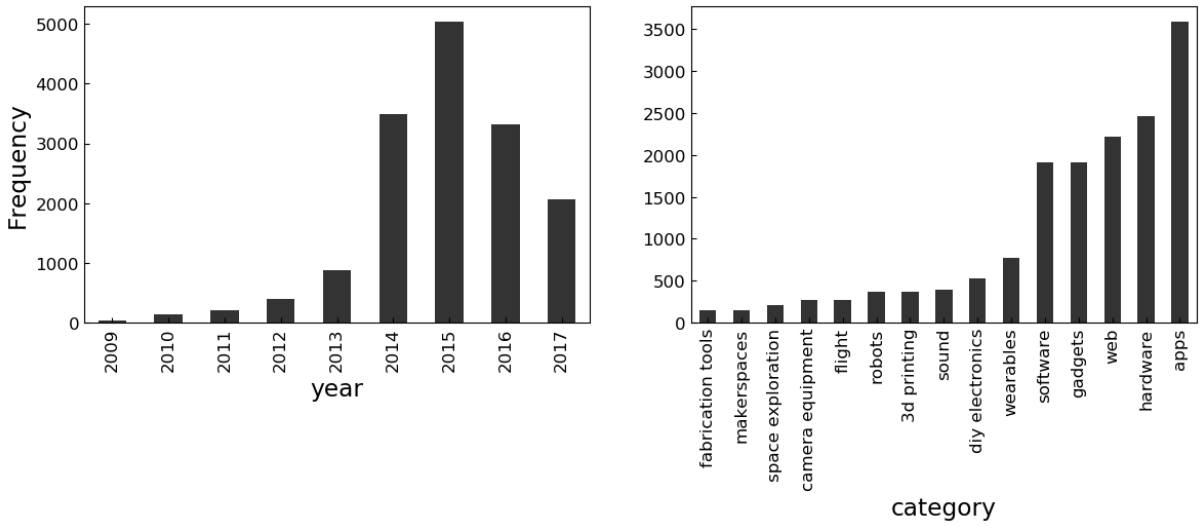


Figure 4.1: Number of projects over time and categories.

afterwards. Project duration can be set up to 60 days, although, the common practice, is a month.

Creators set the offerings for the reward tiers along with the price tag for each. There is a minimal guideline on the platform for the creators to design the reward schemes. Kickstarter suggests the creators to offer multiple reward levels including rewards with lower price tags to attract more backers to the project. As a result, the reward tiers often start with low-level rewards including some gratitude gestures as a form of "Thank you" cards, credits in the product website, and other ways of thanking for small supports. Creators often begin with very basic versions of their product as the initial rewards, and as they move up in the reward tiers, they include more premium versions with more features or often bundles of products. These practices are considered as vertical product differentiation. Thus, often the more expensive reward tiers offer the better product in terms of features, quality, or quantity. Creators may have some market power depending on the functionality of the products, as the innovative products might not have equivalent counterparts in other markets. Thus, pricing of the products can be very subjective due to no prior experience in the market. That is why

Table 4.1: List of variables and summary statistics

Variable	Definition	Mean	Min	Median	Max
goal	Target of the goal to reach in US dollars	83470.96	1	20000	10000000
raised	Pledged money ($)	24568.43	0	382.55	6225355
success	Whether the campaign hit the target	0.26	0	0	1
price	Reward level pledge money	283.03	1	85	25000
backers	Number of backers for each project	194.01	0	7	105857
reward backers	Number of backers for each reward level	27.89	0	1	30219
reward tiers	Number of reward levels of each project	6.75	0	6	64
duration	Project duration length in days	35.16	1	30	90
staff pick	Whether the project is picked by the platform	0.23	0	0	1
delivery time	Expected delivery time of the reward item (days)	102.75	0	75	2049

most creators stick to some natural price tags common in Kickstarter. The popular price tags are $ 5, $ 10, $ 25, $ 50, $ 200, and $ 500, as shown in Figure 4.2. Decision on the number of reward tiers is similar to that of a multi-product firm with vertically differentiated products. Offering low number of reward levels may lead to backer loss due to the less variability in the products and prices, whereas, too many levels prevent backers to reach to the higher rewards. In my data, the average number of reward tiers is 6.8, and the median is 6.

There are three more additional variables: *staff pick*, *delivery time*, and *shipping*. *staff pick* is one of the features of Kickstarter in which some projects are picked by staff members and receive a badge. Recently, this feature has changed name to *Project I Love*. In my data set, 9% of the project had received this feature. *delivery time* determines the expected delivery time of the reward item. On average, it takes 100 days to deliver the reward item. Note that this is the expected delivery time which is pre-announced. The real delivery time might vary. I do not have data on the actual delivery time. *shipping* shows the information and restrictions on shipping for those items needed the postal services.

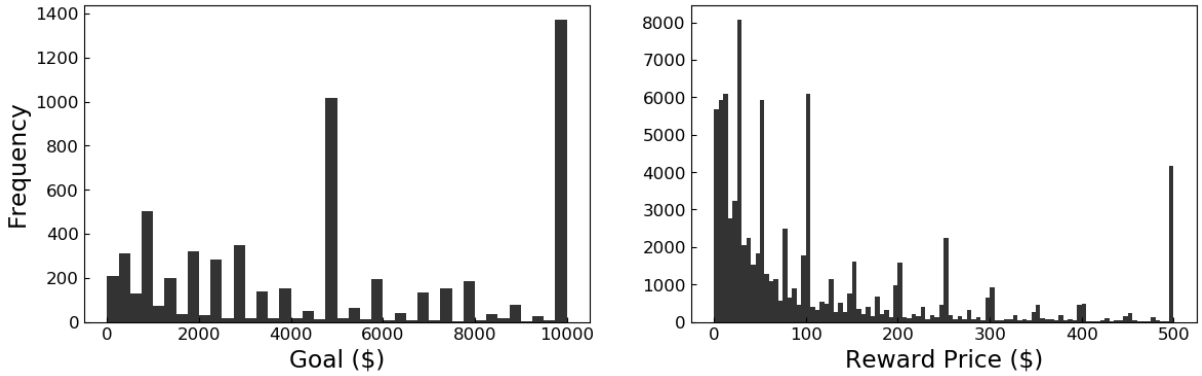


Figure 4.2: Left, histogram of projects with goal below $ 10,000. Right, histogram of reward prices under $ 500

4.1.1 Model-free Evidence

To explore the competition effect and pricing, I look at the market structure evolution over time. Figure 4.3 top row shows the average price of rewards along with the success rate over time. That suggests around years 2012, and 2013, Kickstarter has the highest success rate in the technology category, while, at the same time, the average prices go up during those years. Additionally, by comparing this graph to the number of available projects on Kickstarter, I see there is a positive spike in 2014 in terms of number of projects, suggesting that the demand for projects remains relatively constant while the number of projects (supply) grow dramatically. Also, I see that creators collectively respond to the competition effect by adjusting their prices. To look further into the competition within the platform, Figure 4.3 (bottom left) shows the number of projects per number of backers over time. This highlights the fact that competition might affect the pricing. Note that the gap between average goal and average pledged money in Figure 4.3 (bottom right) tends to increase during the times with surges in number of projects (2014 and 2015).

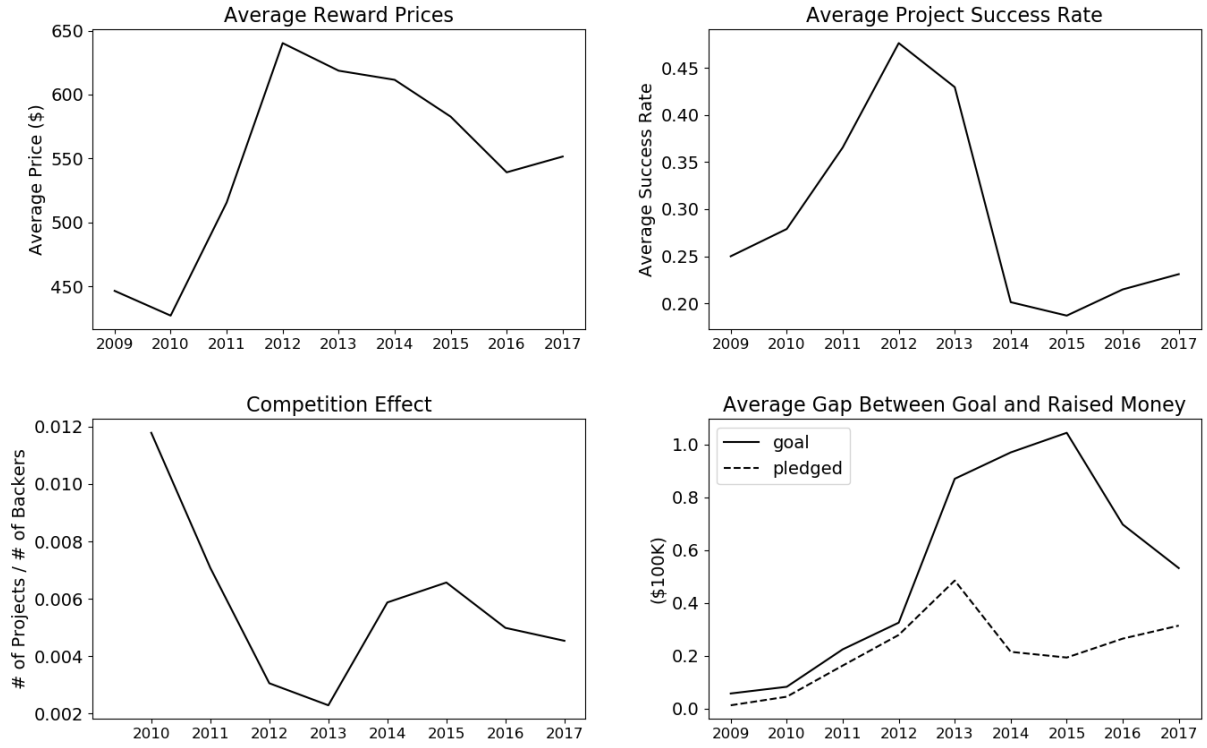


Figure 4.3: Market structure evolution over time

4.2 Model

I apply a structural approach using discrete choice models to estimate the demand for project-rewards. Since my data is aggregated data with no individual level data, I follow Berry et al. (1995), BLP hereafter, which is widely used in demand estimation with aggregate data. I consider Kickstarter as a marketplace in the model. As competition plays an important role in deciding which platform to choose, I abstract away from the creators' decisions on choosing the platform. Creators, however, do not often run parallel campaigns as it needs their focus and energy to successfully manage the fundraising process by addressing questions and concerns of potential backers.

In addition to typical assumptions presented by BLP and Nevo (2011), I consider different reward levels as different products with their own price tags. As I focus on the technology

category, the reward levels on the platform represent different offerings in terms of bundles, features and materials. Therefore, the unit of analysis is reward level. Rewards share common characteristics within a project while vary in the final product and price. Next, the notion of discrete choice implies that each backer would choose only one option (here, reward). Whereas, in practice, customers can choose more than one. This assumption is not restrictive, since each transaction is considered as one incident. Thus, a backer who selects multiple reward levels will be considered as multiple transactions. Last, I assume that all the *available* projects are in the choice set of potential customers. This assumption is reasonable for the loyal and returning backers who frequently come back to the platform to explore new projects. Other backers may not explore all the projects available. But, note that because of the chronological order of project presentations in the website, all the projects have chances to be on the choice sets on the aggregate level.

It is also worth noting that projects on crowdfunding platforms go through two stages. The first stage is the fundraising period, and the second stage is running the project towards the goal to deliver the committed rewards conditional on successful fundraising. There are different mechanisms in place to make sure that the creators would do their best on delivering the rewards. However, I do not model the of implementation stage of the projects. As small number of projects fail even after the successful fundraising due to various reasons. The backers may have some prior beliefs about these risks and account for those in their utility maximization process. Some part of risks can be associated with higher goal, price, and prior experience of the creator. Since it is a plausible scenario that there may still be some risks unexplained even after controlling for the mentioned variables, I believe my model can account for some of the risks involved. I take advantage of the information provided in the *Risks and Challenges* of the project in the platform.

Following frameworks in Lancaster (1971) and McFadden et al. (1973), products are described as collections of characteristics from which customers derive utility. In these frameworks, customers choose the product that maximize the utility gained from product characteristics. Product characteristics explain the mean utility level and drive the substitution

patterns. One of the challenges is that the products are very differentiated. There is not a set of characteristics, as in BLP, that can explain products. These projects can vary from an innovation in an existing product or completely new ideas and functions, yet, competing with one another. A possible solution to this problem is to include brand dummies. Brand dummies can be applied to those markets with few brands and repetition over time or markets. By including brand dummies, one cannot identify the effect of time-invariant characteristics. In my case, I cannot use brand dummies because I have many projects (brands) with no repetition over time. This prevents us from identifying the model with project dummies. Another challenge is that the most part of the information about the project and products is provided in campaign description which has un-structured format. To tackle those challenges, I utilize methods in natural language processing (NLP) to transform the data to structured format in the most efficient way. In the next section I provide details on how I achieve this goal.

4.2.1 Product Characteristics

Campaign owners carefully craft their product features in the project descriptions. That is the main source of information for crowdfunding projects. It contains multiple features for the product including material, design, functions, information about the production process, and risks and challenges. Another important aspect of description is the combination of tone, sentiments, and semantics of the text which has considerable effect on the purchase decision. In fact, there are several marketing firms specializing on how to design the campaign package and information presentation including the project description. There are numerous methods available in NLP to quantify the text. More generally, they can be divided into two categories of text representation: sparse and dense representation. In sparse representations, each word is converted into a very large sparse vector. Simple and intuitive methods such as Bag of Words, word co-occurrence matrices and TF-IDF are sample of sparse representations. As straightforward as these methods are, they cannot capture the complex nature of semantics in the data. Also, as large sparse matrices, it is more difficult to utilize them in most

econometric models. Hence, dense representations are more suitable for my purpose. In addition, dense vectors tend to generalize better and are more robust to noises in the data. By capturing synonymy and semantics, dense vectors often have more prediction power.

In the category of the dense vector representations, there are also several practices in this field, but they all share the same goal: to quantify text into a vector of real numbers in a way that it preserves the semantics and contextual information regarding the domain of the text. I follow a recent algorithm based on neural-net language model embeddings. Mikolov et al. (2013a) and Mikolov et al. (2013b) propose a method called *word2vec* which learns embeddings as part of the process of word prediction. It has resulted in an amazingly good performance in practice. As an extension of that, I use an algorithm which results in distributed document vectors (Le and Mikolov, 2014). This algorithm is also known as *doc2vec*. It is an unsupervised algorithm that learns fixed-length vector representations from variable-length pieces of texts. A document can be a sentence, paragraph, or in my case a project description. This approach has several advantages. First, compared to bag-of-words, it preserves the order of the words in a local context. Second, the objective function which is optimized while training these vectors, considers the contextual predictability of words, thus capturing semantic information in an unsupervised setting. In my analysis, each project description is represented as a dense vector which is trained to predict the words in the text. I need to train the model with the text from project description in the crowdfunding platform, because the trained vectors are context specific. I explain the technical specification in Appendix. The resulted vectors have interesting properties. For instance, the vector for `Paris` is closer to that of `France`. Hence, they tend to keep the relationships in summation or subtraction. One can form document vectors by adding or taking the average of all the vectors corresponding to the words in the document. As an extension to *word2vec*, document vector can be built with similar fashion.

I perform a similar process on reward descriptions to make numerical representations of the rewards. However, due to short length of reward descriptions, I use a different approach to construct the vector representations for rewards. I use set of word2vec pre-trained vectors

by Google. Then, I generate the reward vectors by averaging on all word vectors in the description. Consequently, for each reward, I concatenate the features from the project description and those from reward description. By this, I allow for correlation among reward levels within a project. I now use the resulted feature vector as my product characteristics in the econometric model. The downside of this method is the lack of interpretability. The dense vectors contain complex non-linear relationships and semantics which cannot be decomposed in a clear way.

As an example, I predict project sub-categories by the reward vectors to experiment the prediction ability of the model. Table 4.2 shows the results of sub-category prediction task by document vectors. The reported numbers are for test set. Both the logistic and random forest models have prediction power compared to the base model. Moreover, I use Principal Component Analysis (PCA) to reduce the dimension of the vectors by keeping the most information. I apply PCA for the top 20 dimensions of the data. The prediction power of the PCA model is comparable with the full model suggesting that they are good representative for my feature vector.

Table 4.2: Category prediction results by the trained document vectors

Model	Log Loss	
	Project Description	Project and Reward Description
Base	2.34	2.34
Logistic	1.33	1.09
Random Forest	1.94	1.58
PCA- Logistic (20)	1.49	1.36
PCARandom Forest (20)	1.57	1.2

I point out that the vector representation of the reward-specific information is very crucial to the model. There are considerable price variations within projects. To explain the price variation in the model, I need to leverage the information of the rewards beside the common project characteristics. In other words, the reward vector should be able to explain the

utility variations. I take advantage of a visualization method named t-distributed stochastic neighbor embedding (t-SNE) developed in machine learning to illustrate high-dimensional embedding data in a two (or three) dimensional space (Maaten and Hinton, 2008). This technique utilizes a nonlinear dimensionality reduction method to visualize high-dimensional vectors. Figure 4.4 shows a random sample of the reward vectors in a two-dimensional representation. Each point is colored by the price category. The graph suggests that there is good variation in the reward vectors correlated by price.

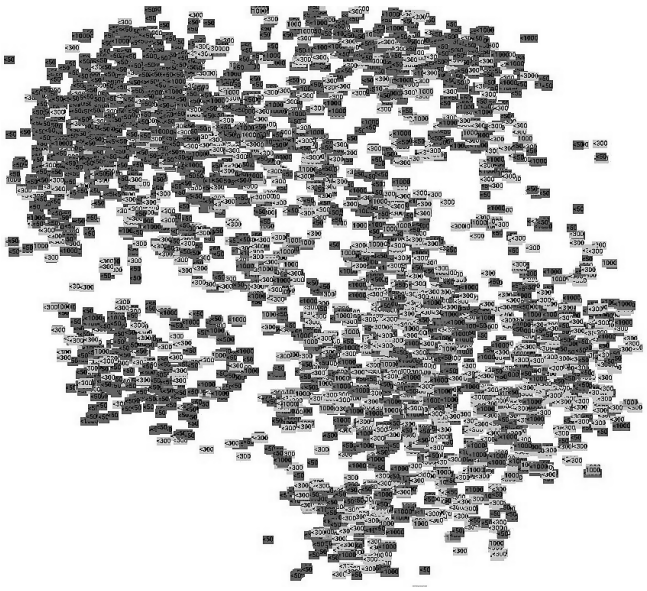


Figure 4.4: t-SNE graphical representation of reward vectors (darker color means lower price)

4.2.2 Choice Model and Utility Function

I model backer decision process of choosing a reward level (therefore a project). I implement a discrete choice model for backer's decision. I partition time into monthly time periods. Note that even though my data set is cross-sectional, I need to denote time in my model to specify the choice set at each time period. I formulate the utility function for backer i

choosing reward level j as follows:

$$u_{ijt} = X_{jt}\beta + \alpha\ Price_{jt} + \xi_{jt} + \epsilon_{ijt}, \tag{4.1}$$

where X_{jt} includes the feature vectors of reward j at time (or market) t, $Price_{jt}$ is the reward price tag, ξ_{jt} captures the unobserved characteristics of the product-reward, and ϵ_{ijt} is the idiosyncratic shock which is assumed to be independent and identically distributed (i.i.d.). Note that X_{jt} includes the trained project vector and reward vector as well as *staff pick*, log(*goal*) , *delivery time*, *duration*, and a constant term. I suppress the index for project for brevity. Note that the rewards can be correlated within a project through the shared characteristics. The utility of outside option is defined as following:

$$u_{it}^0 = \xi_{0t} + \epsilon_{it}^0. \tag{4.2}$$

Since I cannot identify ξ_{0t}, unobserved utility of outside option, I normalize ξ_{jt} by subtracting ξ_{0t} from the former. It is worth noting that I model the reward selection in a discrete choice fashion. Thus, I do not directly model success meaning that each project is in the choice set whether it is successful or not. Also, the total pledged money for each project would be determined by the aggregation of individual pledges. I define mean utility variable, δ, the utility of a project-reward which is the same for all the individuals across the platform. Note that I have omitted income in the utility function as it is canceled out by normalizing the utility levels.

$$\delta_{jt} = X_{jt}\beta + \alpha\ Price_{jt} + \xi_{jt}. \tag{4.3}$$

Following the common practice in industrial organization literature, I assume Type I Extreme Value distribution for the idiosyncratic shocks. This leads to a closed-form solution for choice probabilities. Since, all the backers are now assumed to be homogeneous, I can

write the market shares in the logit format as in Equation 4.4.

$$s_{jt} = \frac{e^{\delta_{jt}}}{1 + \sum_{j'} e^{\delta_{j't}}}. \tag{4.4}$$

By simple transform, I can rewrite Equation 4.4 as a function of observed market shares and outside option market share.

$$\log(s_{jt}) - \log(s_{0t}) = \delta_{jt}. \tag{4.5}$$

To estimate δ_{jt}, I use the observed market shares, s_{jt}, and the market share for the outside option, s_{0t}. To calculate the outside option market share, I need to estimate the overall size of the market at t. Nevo (2000) outlines criteria on how to choose the market size. The market size should not result in zero market share for the outside option. I derive the quarterly data for number of rewards selected in the technology category of Kickstarter. I interpolate the missing ones using the overall growth of Kickstarter. I break down the quarterly data into monthly market size for technology category. I do not differentiate the sub-categories, and do not consider different market size since they are presented together in the platform website. In other words, all the projects from sub-categories are pooled into the same market.

4.2.3 Price Endogeneity

If the creators know the unobserved characteristics, ξ, they can strategically set the prices in a way that prices are likely to be correlated with the unobserved characteristics. In this case, my results for price coefficients will be likely biased. Most scholars follow the instrumental variable approach proposed by BLP using cost shifters, other product characteristics in the same firm, and prices in different markets (Nevo, 2001; Petrin, 2002), while Petrin and Train (2010) take a control function approach using instruments. All those approaches require the existence of IV. In my study, I cannot use the classical BLP instruments because the projects only happen once. To account for the endogeneity of the price, I leverage the fact that there are natural price classes in crowdfunding categories. In the absence of IV, I follow latent

instrument approach (LIV) proposed by Ebbes et al. (2009). In this approach, I split the endogenous price into two exogenous and endogenous parts.

$$\delta_{jt} = X_{jt}\beta + \alpha\ Price_{jt} + \xi_{jt}, \tag{4.6}$$

$$Price_{jt} = \Pi s_{jt} + \upsilon_{jt}, \tag{4.7}$$

$$\begin{pmatrix} \xi_{jt} \\ \upsilon_{jt} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}),$$

where, Πs_{jt} is exogenous and uncorrelated with the unobserved characteristics, and υ_{jt} is endogenous and correlated with ξ_{jt}. I assume that ξ_{jt} and υ_{jt} follow a joint normal distribution with mean zero and the variance-covariance matrix $\mathbf{\Omega}$. Therefore, the correlation of price and ξ_{jt} is captured by the covariance term in $\mathbf{\Omega}$. LIV essentially helps to identify the model by set of statistical assumptions. Clearly if s and υ_{jt} both have normal distributions, the model cannot be identified. I explain more on the identification later in the paper. Park and Gupta (2012) take a similar approach in applying LIV in aggregate discrete choice model by estimating the exogenous part of price non-parametrically. They utilize copula model to estimate a non-parametric distribution of the exogenous part of price. However, they follow a different approach in estimating the model. I assume that the exogenous parts of price are discrete and unobserved instrument. The discrete instrument has n categories following a finite mixture distribution. I assume category indicators are unknown (ex-ante) and have a multinomial distribution with parameters $\lambda = (\lambda_1, ..., \lambda_n)$ representing the category sizes where $\sum_k \lambda_k = 1$. The assumption of latent class model for price is reasonable regarding to the structure of campaign designs. This is a common practice by crowdfunding campaign creators to provide multiple reward levels with corresponding prices. The different price levels are meant to cover potential heterogeneity in backers. Each backer may come with a different budget constraint. Consequently, creators provide rewards which matches different classes of backers.

4.2.4 Heterogeneity

In the homogeneous case, the mean utility can be calculated in closed-form due to the logit error structure. However, as mentioned by other scholars (Train, 2009; Berry, 1994; Dubé et al., 2012), this model suffers from independence of irrelevant alternatives (IIA) problem common in logit models. That is, the substitution patterns among rewards would be solely based on the current market shares. Additionally, the own price elasticity would be proportional to own price, implying lower the price, the lower the elasticity (in absolute value). Cross-price elasticities would depend on the market share ratios. Hence, the higher market power a project has, the more substitutions directed towards that project. Nevo (2000) discusses the potential problems of this model in details. There are two possible solutions for the shortcomings. First solution is to apply a nested structure by nested logit model. Since there is a natural hierarchy in my data (project and reward levels), it might be a good match. One of the benefits of the nested logit model is that it results in closed-form solutions for market shares. I perform nested logit model as a robustness check. Nested logit model may not, however, capture different individual tastes for attributes. Therefore, I utilize a more general random coefficient model with latent instrument variables (RCLIV) which can eliminate the IIA problem and modify the substitution patterns. In this model, I let the coefficients in Equation 4.1 follow normal distribution.

$$u_{ijt} = \delta_{jt} + \alpha_h \; Price_{jt} + \epsilon_{ijt}, \tag{4.8}$$

$$\alpha_i = \alpha + \sigma_\alpha w_i, \tag{4.9}$$

where, α_h is the random part of price coefficient. Equation 4.8 can be easily extended to allow for random coefficients in β.

4.2.5 Identification

The identification of a typical BLP model comes from the exogenous variables and instrument variables. Berry (1994) explores the identification of the model in details. I discuss the identification of the model with latent instrument variable. In general, LIV models are identified if the exogenous part of the endogenous variable does not follow the normal distribution Ebbes et al. (2009). Park and Gupta (2012) use a non-parametric approach to estimate the exogenous part. They show that if the non-parametric model is not normally distributed, the model is identified. In fact, if the distribution moves towards normal, the two parts are no longer identifiable, and the identification becomes weak. Due to the assumption of the finite mixture model as the exogenous part, I need at least two categories to identify the model. Note that the number of identifiable classes depend on the variation in the data and should be tested empirically. Moreover, I can use the typical two-step nested fixed point proposed by BLP while the LIV model handles the endogeneity. That is, recovering the mean utilities, δ, is not biased since the endogeneity is within the mean utility (correlation of the price and unobserved characteristics). Park and Gupta (2012) show that how these two steps can be separated.

4.2.6 Estimation

Moving from logit to the random coefficient model, there is no closed-form solution for choice probabilities. The choice probabilities are derived by numerical integration through simulations. The logit form is widely used in the field as it provides smoothness and non-zero probabilities.

$$\begin{aligned}\hat{s}_j(x_j, p_j, \xi_j; \theta) &= \int_{\alpha,\beta} \frac{exp(x_j\beta_i + \alpha_i p_j + \xi_j)}{1 + \sum_k^J exp(x_k\beta_i + \alpha_i p_k + \xi_k)} dF_{\alpha,\beta}(\alpha, \beta; \theta) \\ &\approx \frac{1}{n_s} \sum_r^{n_s} \frac{exp(x_j\beta_r + \alpha_r p_j + \xi_j)}{1 + \sum_k^J exp(x_k\beta_r + \alpha_r p_k + \xi_k)}.\end{aligned} \tag{4.10}$$

Park and Gupta (2009) propose a simulated likelihood approach in which the error for the endogenous part should be integrated out. They form the likelihood function based on a multinomial distribution. They acknowledge that the estimation approach is not neither efficient, nor stable. Because the likelihood tends to get smaller than machine precision rather quickly. They suggest sampling method to work around this problem. The most common way of estimation, however, is Nested Fixed Point (NFP) approach proposed by BLP. The advantage of the approach is that it separates the estimation process into two parts: linear, and non-linear estimation. That is, in the inner loop the values for δ are estimated by the system of equations in which the unknowns are the mean utility values. Berry (1994) prove that there is a fixed point in the systems of equations. Hence, they propose a contraction mapping approach to inverse the market shares to mean utility values with guaranteed convergence. As opposed to BLP, I use maximum likelihood estimation for the finite mixture model by modifying the estimation process. The modification is rather straightforward. As Equation 8 specifies, I can follow the inner loop of BLP to recover δ with no difference. However, I can no longer benefit from the linear estimation part in the inner loop of BLP due to the mixture model. Equation 11 shows the likelihood conditional on class k, and unconditional likelihood. This approach works because the endogenous price and the correlated unobserved characteristics both are inside δ. Therefor, there is no bias in estimation of the mean utility values. For the outer loop, I lose linearity of the model as I need to estimate a finite mixture model. Consequently, the computation burden is higher compared to BLP.

$$\begin{aligned}
\mathcal{L}_j(\delta_j, x_j, p_j | s_j = e_k) &= \mathcal{N}(\mu_{jk}, \mathbf{\Sigma}_k) \\
\mathcal{L}(\delta_j, x_j, p_j) &= \prod_{j=1}^{N} \sum_{k}^{N} \lambda_k \mathcal{L}_j(\delta_j, x_j, p_j | s_j = e_k).
\end{aligned} \tag{4.11}$$

I have two options for the likelihood functions: direct likelihood, and Expectation-Maximization (EM) method by Dempster et al. (1977). Note that both algorithms result

in local optima and need to be run several times with various initial points to make sure that local optima are indeed global optima. I apply EM method since it makes the optimization problem more efficient and, in my experience, more stable. The parameters to be estimated are the price coefficient mean and variance (α, σ_p^2) characteristics coefficients mean and variance (β_k, $\sigma_{x_k}^2$), price class means (π_s), price class population probability (λ_s), and variance-covariance matrix of the joint distribution of price and unobserved characteristics. Utilizing EM, I can rewrite Equation 4.11 as follows:

$$\begin{aligned} \log \mathcal{L}\left(\delta_j, x_j, p_j\right) &= \sum_j \sum_k q_{jk} \log \mathcal{L}_j\left(\delta_j, x_j, p_j | s_j = e_k\right), \\ q_{jk} &= \frac{\lambda_k \mathcal{L}_j\left(\delta_j, x_j, p_j | s_j = e_k\right)}{\sum_l \lambda_l \mathcal{L}_j\left(\delta_j, x_j, p_j | s_j = e_l\right)}, \\ \lambda_k &= \frac{1}{N} \sum_j q_{jk}. \end{aligned} \tag{4.12}$$

Following the assumption of the joint normality, EM utilizes an iterative method to update the model parameters: mean and variance. Generally, there are two steps involved in EM approach. First, the E-step (expectation) calculates the membership weight, q_{jk}. Second, M-step (maximization) updates the new mean and variance as presented in Equation 4.13. The iterative procedure continues until the error of the convergence is in my tolerance. EM optimization often trails as they reach the convergence. But, the likelihood improves after each iteration.

$$\begin{aligned} \mu_k^+ &= \frac{\sum_{j=1}^N m_k q_{jk}}{\sum_{j=1}^N q_{jk}}, \\ \Sigma_k^+ &= \frac{\sum_{j=1}^N \left(m_k - \mu_k^+\right)\left(m_k - \mu_k^+\right)^T m_k q_{jk}}{\sum_l \lambda_l L_j\left(\delta_j, x_j, p_j | s_j = e_l\right)}, \end{aligned} \tag{4.13}$$

where m_k is the mean of the estimated error terms defined in Equations 4.6 and 4.7. Note that I also allow the covariance matrix to vary by the latent class. In other words, the correlation of price and unobserved characteristics depends on the class. The latter relaxes

the assumption that the campaign creators are uniformly strategic for all the price categories. This is very important since one may expect the endogeneity of the price to be higher in more expensive rewards.

The EM estimation is very straightforward in the homogeneous model. To retrieve the model parameters for utility function in Equation 4.1, I require to take an additional step after EM convergence. Now I have two sets of equations as in Equation 6 and 7 with the estimated parameters for price class means, probabilities and covariance matrices. Thus, I use a Seemingly Unrelated Regression (SUR) method for efficiently estimating α and β. In the heterogeneous case, the difference is that I estimate through simulation. That is, I use a non-linear search method for the variance of coefficient distributions through estimation of mean utility variable. Within each search iteration, the process is similar to the homogeneous case, except I simulate using random coefficient distributions and calculating the mean at the end. The process is a numerical estimation of the integration in Equation 4.10. In contrast to BLP, I use the likelihood approach for the search for the additional parameters.

4.3 Results

I first estimate the logit demand model and OLS hedonic demand function. Table 4.3 presents the results for logit demand and hedonic estimations. Models (1)-(3) implement homogeneous logit demand without accounting for the endogeneity. In the first model no dummies and no product characteristics are used except of year dummies. I use product characteristics in the second and third models. In model (3), category dummies are also included. The coefficient of price is negative and significant in all the models. However, the price elasticities are slightly different. As there is a vast heterogeneity in the price of rewards, I normalized it by a standard deviation which is $ 905.7. Hence, this scaling should be considered in the interpretations of the price coefficient. $\log(goal)$ has a positive and significant effect on demand. *staff pick* has a strong positive effect as I expected. *delivery time* (presented in 90-day intervals) has a negative effect on demand. Next two models present OLS hedonic demand estimation. Model (4) is the regression of $\log(numberofbackers)$ on the covariates,

while model (5) uses the *number of backers* as the dependent variable. Interestingly models (5) results in positive effect for *delivery time* implying longer delivery time leads to more demand. Logit model even with no endogenous handling results in much better fit compared to the hedonic models. This is a direct result of structural model and assumptions in the aggregate demand estimation. Note that the inclusion of product characteristics (project-reward vectors) increases the fit of the models significantly (adjusted R^2 increases from 0.16 to 0.29).

Next, I perform the estimation of LIV model. I use the same set of variables used in the models with no endogeneity treatment. I need to assume the number of latent classes in advance. I estimate the model using 2, 3, 4, 5 and 6 price classes. Then, I use Akaike information criterion (AIC) and Bayesian information criterion (BIC) to compare the models and choose the appropriate model. I also estimate the model with higher number of classes. However, the estimation results are not stable due to empirical identification. Therefore, I conclude that the model is not identified for more than 6 classes. Table 4.4 reports the estimation results for the described models. Note that the price coefficient is consistently higher (in absolute value) when accounting for endogeneity suggesting the price coefficient is biased towards zero in the base logit demand model (column 3 in Table 4.3). I infer that the higher unobserved characteristics lead to higher price as expected. I discuss the implications of the results further in the next section. I report the estimates for the class means and unconditional (population) class probabilities as well as the correlation of price and unobserved characteristics calculated by the variance-covariance matrix. I observe, generally, that higher price classes have higher (and more positive) correlation with the unobservable characteristics. This may sound intuitive. I will discuss different implications of such results.

Table 4.5 represents the estimates for the latent class parameters. In case of two classes, many of the rewards are in the lower category with \$68, while the higher class with mean \$1264 has 18% of the rewards. For three latent classes, the rewards are divided into approximately \$26, \$179, and \$2017 categories. The results show the mixture distribution of price after dealing with endogeneity. The estimated results suggest that many of the re-

Table 4.3: Estimation results for logit and OLS demand models

	(1) Logit	(2) Logit	(3) Logit	(4) OLS	(5) OLS
price	-0.294***	-0.308***	-0.308***	-0.311***	-16.015 ***
	(-43.07)	(-48.10)	(-48.26)	(-50.62)	(-11.32)
log(goal)	0.113***	0.022***	0.017**	0.018***	7.350***
	(23.42)	(4.34)	(3.48)	(3.67)	(6.64)
staff pick	1.275***	1.005***	0.992***	0.940***	68.602***
	(76.42)	(59.53)	(58.11)	(57.23)	(18.14)
delivery time (quarters)	-0.039***	-0.046***	-0.047***	-0.040***	3.964**
	(-5.97)	(-7.62)	(-7.65)	(-6.74)	(2.94)
duration	0.004***	0.004***	0.004***	0.003***	0.567***
	(6.55)	(5.88)	(5.68)	(4.56)	(4.12)
constant	-5.156***	-3.881***	-3.554***	1.141***	-104.169*
	(-26.46)	(-21.29)	(-19.01)	(6.34)	(-2.52)
Category dummies			✓	✓	✓
Document Vectors		✓	✓	✓	✓
Year dummies	✓	✓	✓	✓	✓
Observations	59845	59845	59845	59845	59845
Adjusted R^2	0.16	0.29	0.29	0.28	0.03

t-statistics in parentheses.

Model 1-3 are the logit demand estimation in which the dependent variable is the mean utility δ. There are different sets of controls in these models. Models 4-5 are the OLS results. The dependent variable in Model 4 is log(backers) for rewards. In Model 5, the dependent variable is backers.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

wards belong to the medium to low level price categories. This is in accordance with the non-parametric distribution of the reward prices. Figure 4.5 shows the histogram of prices along with the estimated mixture distributions.

Finally, I estimate the heterogeneous model with the random coefficients. I consider two random coefficients in the model: *price*, and *constant* terms. Basically, it is a random intercept model plus a random coefficient for price. Table 4.6 summarizes the estimate results for 2, 3, 4, 5, and 6 latent classes. The estimates for the heterogeneous model are similar to

Table 4.4: Estimation results for LIV Logit models

	LIV Logit				
	(1)	(2)	(3)	(4)	(5)
	n=2	n=3	n=4	n=5	n=6
price	-0.983***	-0.683***	-0.648***	-0.507***	-0.473***
	(-71.72)	(-65.54)	(-63.69)	(-58.58)	(-57.33)
log(goal)	0.036***	0.030***	0.029***	0.024***	0.024***
	(7.26)	(6.15)	(5.95)	(4.92)	(4.79)
staff pick	1.030***	1.013***	1.010***	1.004***	1.003***
	(61.61)	(60.22)	(59.97)	(59.29)	(59.20)
delivery time (quarters)	-0.027***	-0.034***	-0.034***	-0.040***	-0.040***
	(-4.53)	(-5.60)	(-5.72)	(-6.61)	(-6.67)
duration	0.004***	0.003***	0.003***	0.003***	0.003***
	(6.07)	(5.69)	(5.65)	(5.58)	(5.56)
constant	-3.495***	-3.514***	-3.519***	-3.535***	-3.542***
	(-19.12)	(-19.10)	(-19.09)	(-19.08)	(-19.10)
Category dummies	✓	✓	✓	✓	✓
Document Vectors	✓	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓	✓
AIC	215817.4	216603.6	216828.0	217416.3	217416.2
BIC	216429.3	217215.5	217440.0	218028.2	218447.1
Observations	59845	59845	59845	59845	59845

t-statistics in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

those of the homogeneous model. Nevertheless, the overall fit is better in the heterogeneous model. The standard deviation of the random coefficient distribution of price is significant for models with 2, 3, 4, and 5 classes. Furthermore, the standard deviation for the intercept term is significant for all the models. Note that in calculation of standard errors, I use MLE Fisher score instead of GMM standard errors in BLP.

Similarly, I retrieve the parameters for class means, probabilities, and covariance matrices. Table 4.7 summarizes the estimated results for the models discussed above. The estimates

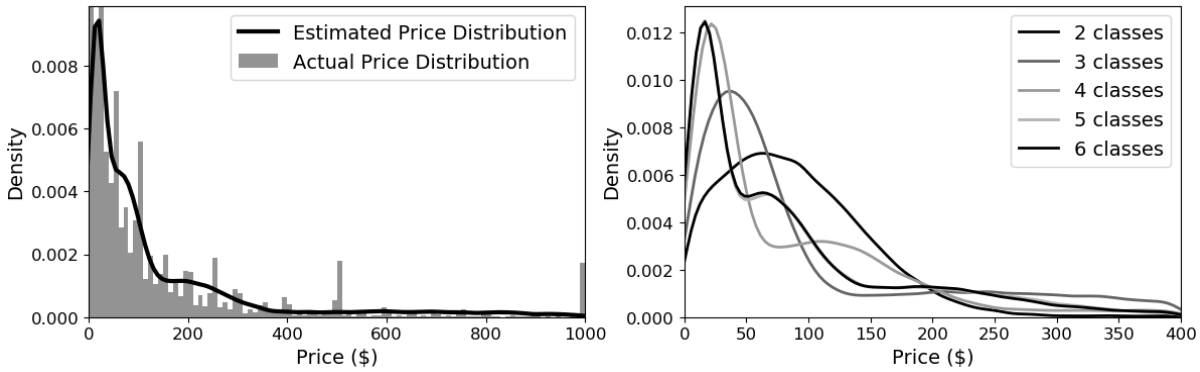


Figure 4.5: Actual vs. estimated price distribution.

for classes in the heterogeneous case is almost identical in many cases. Therefore, the interpretations regarding the reward prices in homogeneous case hold for the heterogeneous case.

Unlike the homogeneous model, five latent classes offer better fit. Hence, I perform further analyses on the substitution patterns, price elasticity, and customer welfare using the model with five latent classes. Figure 4.6 shows the model fits in terms of AIC and BIC for RCLIV model.

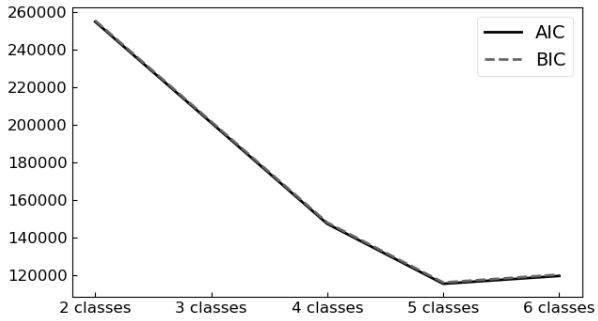


Figure 4.6: Akaike information criterion (AIC) and Bayesian information criterion (BIC) for estimated RCLIV.

Table 4.5: Estimation results for price latent classes (homogeneous)

Number of classes		Class Mean Price		Class Population Probability		Endogeneity Correlation
n=2	π_1	68.3	λ_1	0.82	ρ_1	-0.01
	π_2	1264.8	λ_2	0.18	ρ_2	0.8
n=3	π_1	25.6	λ_1	0.48	ρ_1	0.04
	π_2	178.9	λ_2	0.42	ρ_2	-0.12
	π_3	2016.7	λ_3	0.1	ρ_3	0.78
n=4	π_1	20.3	λ_1	0.4	ρ_1	0.0
	π_2	103.4	λ_2	0.37	ρ_2	-0.15
	π_3	452.9	λ_3	0.18	ρ_3	-0.08
	π_4	3318.0	λ_4	0.05	ρ_4	0.77
n=5	π_1	13.6	λ_1	0.3	ρ_1	-0.03
	π_2	62.6	λ_2	0.34	ρ_2	-0.18
	π_3	187.9	λ_3	0.2	ρ_3	-0.08
	π_4	600.8	λ_4	0.12	ρ_4	0.01
	π_5	3661.5	λ_5	0.04	ρ_5	0.76
n=6	π_1	13.5	λ_1	0.3	ρ_1	-0.04
	π_2	61.3	λ_2	0.34	ρ_2	-0.2
	π_3	178.8	λ_3	0.2	ρ_3	-0.15
	π_4	443.9	λ_4	0.07	ρ_4	-0.09
	π_5	914.3	λ_5	0.06	ρ_5	0.14

Table 4.6: Estimation results for Random Coefficient LIV Logit models

	LIV Logit				
	(1)	(2)	(3)	(4)	(5)
	n=2	n=3	n=4	n=5	n=6
price mean	-0.834***	-0.609***	-0.425***	-0.430***	-0.391***
	(-25.01)	(-24.61)	(-20.31)	(-21.03)	(-21.18)
price sd.	0.056*	0.015***	0.044*	0.056*	0.011
	(2.10)	(3.72)	(2.87)	(2.89)	(0.16)
log(goal)	0.066***	0.065***	0.054***	0.055***	0.040***
	(4.52)	(4.27)	(3.69)	(3.73)	(3.61)
staff pick	1.337***	1.391***	1.312***	1.313***	1.226***
	(32.45)	(32.20)	(31.86)	(31.89)	(31.85)
delivery time (quarters)	-0.052***	-0.063***	-0.066***	-0.065***	-0.065***
	(-3.82)	(-4.10)	(-4.62)	(-4.57)	(-4.62)
duration	0.003**	0.003**	0.003**	0.003**	0.004**
	(3.27)	(3.14)	(3.13)	(3.12)	(3.12)
constant mean	-4.554***	-5.960***	-4.578***	-4.577***	-2.891***
	(-7.80)	(-7.85)	(-7.86)	(-7.87)	(-7.89)
constant sd.	9.154***	9.725***	9.189***	9.193***	7.562***
	(7.26)	(45.67)	(21.84)	(26.30)	(116.81)
Category dummies	✓	✓	✓	✓	✓
Document Vectors	✓	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓	✓
AIC	254782.6	201007.1	147210.0	115266.5	119444.3
BIC	255348.4	201608.7	147847.4	115939.7	120153.3
Observations	59845	59845	59845	59845	59845

t-statistics in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3.1 Price Elasticity

I now explore own and cross-price elasticities. Price elasticity in the homogeneous logit demand model is very straightforward. However, there are many shortcomings regarding the analyses (Berry et al., 1995; Berry, 1994; Nevo, 2011), as the results are highly driven by

Table 4.7: Estimation results for price latent classes for RCLIV

Number of classes		Class Mean Price		Class Population Probability		Endogeneity Correlation
n=2	π_1	67.9	λ_1	0.82	ρ_1	-0.01
	π_2	1256.6	λ_2	0.18	ρ_2	0.8
n=3	π_1	37.9	λ_1	0.61	ρ_1	0.03
	π_2	235.1	λ_2	0.3	ρ_2	-0.02
	π_3	2228.1	λ_3	0.09	ρ_3	0.3
n=4	π_1	20.6	λ_1	0.41	ρ_1	0.02
	π_2	107.2	λ_2	0.37	ρ_2	-0.05
	π_3	466.5	λ_3	0.17	ρ_3	-0.06
	π_4	3378.6	λ_4	0.05	ρ_4	0.24
n=5	π_1	13.9	λ_1	0.32	ρ_1	0.03
	π_2	65.2	λ_2	0.33	ρ_2	0.02
	π_3	189.1	λ_3	0.19	ρ_3	0.01
	π_4	600.2	λ_4	0.12	ρ_4	-0.05
	π_5	3661.7	λ_5	0.04	ρ_5	0.24
n=6	π_1	13.9	λ_1	0.32	ρ_1	0.04
	π_2	64.7	λ_2	0.33	ρ_2	0.01
	π_3	184.4	λ_3	0.19	ρ_3	-0.04
	π_4	528.3	λ_4	0.11	ρ_4	-0.06
	π_5	1590.3	λ_5	0.03	ρ_5	0.15

the model assumptions. In addition, substitution patterns suffer the very well-known IIA problem. Thus, to infer the price elasticities, I use the heterogeneous model (RCLIV) that relaxes the key assumptions. It can potentially generalize the substitution patterns that will not necessary be driven by the functional form (Nevo, 2000). Henceforth, the own and cross-price elasticities are defined in Equation 4.14.

$$\eta_{jk} = \frac{\partial s_j p_j}{\partial p_k s_j} = \begin{cases} \frac{p_j}{s_j} \int \alpha_i s_{ij} \left(1 - s_{ij}\right) dF\left(v\right) & \text{if } j = k, \\ \frac{p_k}{s_j} \int \alpha_i s_{ij} s_{ik} dF\left(v\right) & \text{otherwise,} \end{cases} \tag{4.14}$$

where $s_{ij} = exp(x_j\beta_i + \alpha_i p_j + \xi_j)/(1 + \sum_k^J exp(x_k\beta_i + \alpha_i p_k + \xi_k))$ is the probability of individual i purchasing product j.

Recall that the unit of the analysis is reward level. Thus, I have more than 59,000 *"products"* for which I calculate the price elasticities. I summarize own-price elasticities by sub-category median measures. Table 4.8 presents the median own-price elasticity by sub-categories. If the own-price elasticity is greater (in absolute value), it suggests the market share will drop more in case of an increase in the reward price. I conclude that the entrepreneurs have higher market power in the categories such as *apps*, *software*, and *web*. However, categories such as *camera equipment*, *3d printing*, and *wearables* face the least market power in the Technology products.

Table 4.8: Sub-category median own-price elasticity

Sub-category	Own-price elasticity
apps	-0.014
software	-0.016
web	-0.016
space exploration	-0.024
makerspaces	-0.024
DIY electronics	-0.028
gadgets	-0.032
fabrication tools	-0.037
hardware	-0.037
robots	-0.04
flight	-0.041
wearables	-0.046
3d printing	-0.047
camera equipment	-0.051

To calculate the cross-price elasticity, it is infeasible to account for all the possible combinations. Therefore, I compare average cross-price elasticity for the rewards within the project and some randomly chosen rewards outside of the project. Henceforth, I divide the reward prices into seven pre-defined price buckets: under $25, between $25 and $50, between $50 and $75, between $75 and $100, between $100 and $150, between $150 and $300, and above $300. Figure 4.7 compares the cross-price elasticity for rewards within and between

the projects. That is, when the price of a reward increases, I expect the market share of the other rewards to increase. For more clarity, I construct the cross-price elasticities for other rewards within the same project and for rewards in other projects. I can compare different substitution patterns for each price category. Note that, in the mid-level rewards, the average cross-price elasticity is lower for rewards within the projects compared to other reward levels. This suggests that the mid-level rewards are not as good substitute as other reward tiers within a project.

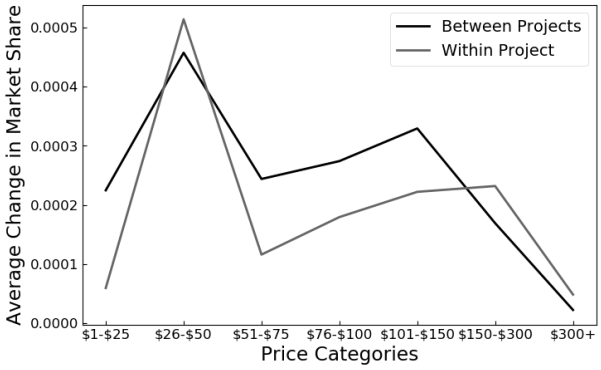


Figure 4.7: Average cross-price elasticity for rewards within and between projects.

4.3.2 Customer Welfare

One interesting aspect of discrete choice models is the possibility of analyses for customer welfare (Trajtenberg, 1989; Nevo, 2003). McFadden (1978) defines the inclusive value as the expected utility a customer receives form the choice set before observing idiosyncratic shocks. This is also referred as *social surplus*. I follow the notation in Nevo (2011) for the welfare definition. According to McFadden (1981) and Small and Rosen (1981), the inclusive value from the choice set A is defined as:

$$\omega_{iAt} = \log\left(\sum_{j \in A} exp\left(x_{jt}\beta_i + \alpha_i \; Price_{jt} + \xi_{jt}\right)\right), \tag{4.15}$$

which is equivalent to social surplus if I take average on individual specific coefficients. By dividing Equation 4.15 by α (absolute value), the welfare can be interpreted into monetary value.

I perform two welfare studies. First, I look at the reward tiers, and derive customer welfare by reward rank in the project. Then I cluster rewards by their price tags and compare the average welfare gain. Reward rank is defined by the ordered position of the reward in the project campaign page. Figure 4.8 presents the summary of the two welfare studies. The average customer welfare is lowest for the third reward item on the reward menu. Except a few kinks, rewards with higher rank in the page offer more welfare. That is, high utility rewards often placed in the bottom of the section in the campaign page. On the right panel in Figure 8, the welfare is presented by the price tags, i.e., the choice sets are those rewards with the same price. Since the price variable is a continuous variable, I divide the price tags into discrete classes described in the last section. The total surplus is the highest in the lower reward tiers. This is due to the more rewards in these categories. However, the average welfare is almost monotonically increasing by the price meaning high utility rewards, on average, are those with higher price tags. This is in line with the welfare results by reward rank. The welfare insights are interesting from platforms point of view. As Kickstarter faces more competition from other platforms or even conventional markets, it is important to know how it can respond by increasing customer surplus. My results suggest that Kickstarter projects bring the most value to the customers by offering the high-level reward tiers.

Second, I analyze the surplus by project sub-categories. Figure 4.9 demonstrates average customer welfare by categories. Categories *wearables*, *web*, *camera equipment*, and *gadgets* provide the most surplus, while the contributions of the other categories such as *hardware*, *software*, and *space exploration* are marginal. This finding is not obvious by just looking at the frequency of the projects in the sub-categories and their respective success rate. The most popular sub-category (in terms of number of available projects) is the *app* category. Also, the highest success rates happen in the *camera equipment*, *DIY electronics*, and *sound*.

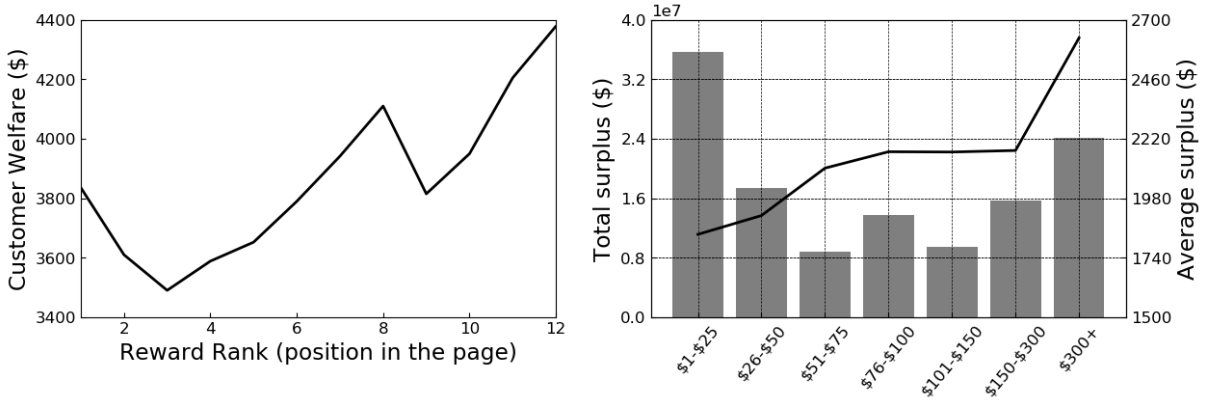


Figure 4.8: Customer welfare for reward tiers based on reward rank (left) and reward price (right).

Hence, these indicators are not proper measures for social surplus and consumer welfare contribution.

4.3.3 Policy Counterfactuals and Substitution Patterns

I use simulations to run counterfactual scenarios to understand the substitution dynamics. First, I show how the model performs to generate the pledged fund through simulations with no policy introduced. The simulated average pledged money for the projects is $32,348 compared to the actual amount $32,580. More on the project level, I also simulate project success rate.[5] Accuracy in success rate is 98.3%, suggesting the model fit is acceptable.

I now introduce three counterfactual simulations based on the rewards offered by projects. I remove different reward tiers for half of the projects selected randomly (treated group) while the rest of the projects will be intact (control group). In the first simulation experiment, I remove reward tiers under $25 for the treated group. I simulate new market shares for all projects and estimate the average pledged money. As the second simulation, I repeat the experiment with reward tiers between $50 and $75. Finally, I apply the experiment on the

[5]Recall that I do not incorporate *success* in the model. I use simulated pledged money and the goal to determine the success.

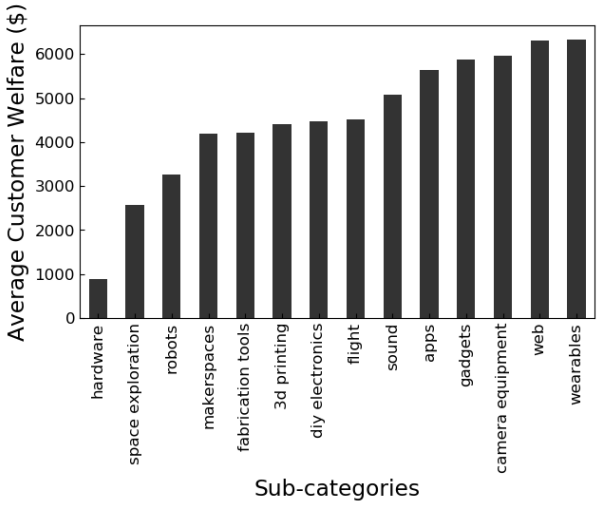


Figure 4.9: Customer welfare for technology sub-categories

higher reward tiers above $200. Table 4.9 summarizes the simulation results and compares them with the actual numbers for both groups.

I observe that the impact of the first policy (low-level reward tiers) on the treated group is very significant (29% increase in average pledged money), whereas the untreated group is not considerably affected. This finding suggests that backers switch to a higher reward level *within* a project possibly due to high correlation of rewards in a project. I infer that, on average, higher-level rewards are good substitute for low-level rewards in the same project.

Table 4.9: Simulation results for affected and unaffected groups

Group		Before Policy		After Policy		
		Actual	Simulated	Low Rewards	Middle Rewards	High Rewards
Treatment Group	Average pledged	$28,189	$27,769	$36,493	$20,060	$15,540
(affected)	Success rate	24.20%	23.30%	27.10%	19.10%	16.00%
Control Group	Average pledged	$37,305	$37,286	$38,468	$46,400	$39,538
(unaffected)	Success rate	36.10%	33.90%	35.00%	36.20%	34.80%

The simulated results for the second policy (mid-level reward tiers) suggest that the treated group suffers from the policy by a 29% decrease in average pledged dollars. The lost pledged money on the treated projects is transferred to the untreated group, showing the crucial role of mid-level rewards. Lastly, I see a drastic decrease in average pledged money for the treated group (45%) as a result of removing high-level rewards. However, there is no significance change in the other group. I conclude that the other reward levels in the projects (with lower price tags) are chosen as a result of the removed choices. Due to lower prices, this leads to the reduction in total raised money and success probability. Note that the correlation between the rewards shapes the substitution pattern in a way that, in absence of some reward levels, other rewards for the same project are likely to be selected.

4.3.4 Discussion

The estimated results suggest that the price coefficient is biased towards zero in the models with no endogeneity treatment. Positive correlation of price and unobserved characteristics (for high-level rewards) means higher price in the market as the higher unobserved characteristics (higher quality, or utility hidden to us as econometricians). This is consistent with the economic theory of the markets. Creators (as suppliers) tend to use their own market power in their markups. In more conventional markets, market power is often translated to the brand effect. In contrast, there is no brand effect in the context of online crowdfunding platforms as the life-cycle of a project is almost a month. However, market power can be interpreted in project uniqueness and quality. That is, when an idea of a project is novel and valuable, or the skill set of the creators makes backers trust creators' ability in reaching the project goal, prices of campaign rewards go up. This is very intuitive as the backers are more likely to choose those rewards.

I discuss that the correlation between the same-project rewards is high. The finding suggests that reward switching within a project is more likely. Due to the reward correlation, I realize that the low-level rewards force a cannibalization effect. That is, in absence of these rewards, backers are more likely to switch to higher reward levels leading to increased

funding. One explanation for this phenomenon is that the market segmentation through price discrimination fails to separate low-valuation and high-valuation backers. That is, low-valuation rewards become attractive to high-valuation customers. I can expect that the cannibalization effect gets worse as the competition increases. One caveat, however, is that this finding is apparent on average. Each individual campaign might have different nature and be unique. Additionally, customers are supposed not to have budget constraints coming to the platform. However, this can give us overall guidelines to design reward schemes.

I now discuss some of the implications of the assumptions and model. The reward levels with zero backers are excluded from the analysis as the effect on those rewards cannot be identified. That is, there is no variation in those projects-rewards. In more mature markets with less number of products and more markets, that is not a likely scenario. However, in fast-paced crowdfunding markets which there are numerous entrepreneurs kicking off their project, the projects are rather short-term concepts. In practice, I see considerable number of rewards with zero backers. The results represent the population of the projects-rewards which got at least one backer. Thus, the results may not be generalized to those projects. Intuitively, I do not know how *bad* those projects are. However, the results clarify the mechanisms by which those projects fail. As another implication, the assumption of normality for price coefficient implies that some individual might experience positive taste in price at the tails: some will have more demand for higher prices. However, based on the estimated standard deviation of the distribution, that is highly unlikely (less than 1%).

Another aspect of model is the choice of the price variable. Due to high price variations, the use of log transformation may seem reasonable. However, the distribution of the log(price) is very different with that of price. Since the distribution of log(price) is more similar to the bell-shaped normal distribution, the identification of latent price classes would be problematic in this case. Moreover, the interpretations of the price elasticity and welfare analyses would not be straightforward. Consequently, I use a normalized price variable instead of the log transformation.

The other implication of the join normality assumption for price and unobserved char-

acteristics is that the tails of the mixture normal distribution will result in negative prices. Thus, in practice it is a truncated mixture distribution. However, through my simulation experiments, I show only a fraction of prices are negative (less than 8%). Since the empirical consequences of this assumption are negligible and relaxing of it adds significant complications to the model, I forgo the adoption of a truncated mixture distribution. Lastly, the mean zero joint distribution assumes that the unobservable characteristics are centered around zero. To remedy this assumption, I take advantage of the project-reward characteristics by NLP techniques. Project-reward vectors explain some of the variation in the choice model.

4.4 Conclusion

My intentions of this research are twofold. First, I study the important aspects of campaign design in crowdfunding platforms. Given the growing market of crowdfunding, entrepreneurs face more design questions regarding the fundraising projects. Second, I contribute to the literature by adopting the demand estimation models and using natural language processing techniques to devise an innovative methodology to capture the unique features of the fast-paced markets. These strategies help us identify the intended results. I define product choices as the reward levels. Thus, I abstract away from the product aggregation methods typical in the demand estimation literature. Although this approach makes the assigns more computational burden, it gives the model enough flexibility to accommodate different aspects of reward scheme design and pricing.

I utilize a structural model approach for aggregate level data. I propose a mixed approach to handle the endogeneity of the price by using latent instrument variable method. I combine the BLP estimation approach and maximum likelihood estimation to address the endogeneity issue. I calculate the price elasticities and conclude that the typical demand models without handling the endogeneity lead to significant upward bias. The bias in price elasticities would result in perception that backers are less price sensitive than they really are. In fact, anecdotes support this finding that practitioners overlook the price sensitivity

in online crowdfunding platforms. Moreover, by analyzing the customer welfare, I find that effect of middle and high-level rewards can be significant. I finally calculate backers' welfare in the online crowdfunding platform per category and find that projects in wearable, web and gadget categories result in more surplus for customers (backers).

I also show the substitution patterns through reward levels. The insights have implications for entrepreneurs in designing the reward scheme and pricing. I find that, on average, low-level rewards do not contribute much to the customer surplus. Additionally, they cannibalize the rewards with higher prices. Entrepreneurs can benefit from these results in terms of designing the reward scheme and pricing.

There are several potential directions to extend the current research. One approach is to account for the competition between the platforms as they are fighting with each other to gain more overall market share. The interesting angle is that the success of projects would be entangled with the success of the platform. So, identifying potential high-quality products is of their interest to survive in the competition. Another possibility for researchers with access to timely data of the fundraising is to address the dynamic effects of fundraising progress. Backers might be heterogeneous in terms of what stage of fundraising they are more likely to support the project. Like the product adoption literature, there are early adopters and those who support later to learn more about the project fortune. In addition, the analyses with individual level data let the researcher to formulate more flexible models possibly accounting for user heterogeneity and budget constraints.

Chapter 5

SOCIAL MEDIA: EFFECTIVE STRATEGIES IN VIDEO SPONSORSHIP

The rapid growth of the online contents spurred the expansion of sponsorship which delivers commercial marketing messages to consumers. Social media has empowered content creators to utilize sponsorship to capitalize the audience interactions. For example, BuzzFeed reaches to 83% of all millennials per month through the online contents (Nielsen, 2018). The ubiquity of accessible online contents is not limited to media companies. During 2017-2018 basketball season, National Basketball Association (NBA) created 1.5 billion user interactions in Facebook, Twitter, Instagram, and YouTube with an estimated value of $490 million (NielsenSports, 2018). The value of sponsorship deals for social media contents are determined by expected impressions and engagement. While, content creators form various partnerships to monetize their contents, the performance of the sponsored contents is not identical across different sponsors.

It is projected that by 2019, total U.S. online ad spend would exceed offline ad spend, which includes television ads, print ads, and billboard posters[1]. Having the largest growth in online advertising, sponsorship is expected to surpass $24 billion in the United States.[2] Online advertisers, sponsors hereafter, seek various strategies to form partnerships with content creators, creators hereafter. Creators including brands and influential figures, try to monetize their property by channeling the attention of their audience to sponsors. Sponsorship can be manifested in various forms including celebrity endorsing, native advertising, product placement, and online sponsored contents. Brands high willingness to pay for sponsorship

[1]eMarketer, report on Social Media advertising and Video advertising in 2018

[2]IEG, ESP Properties, What Sponsors Want & Where Dollars Will Go in 2018

emerges on social media where instant measures of user interactions are available. They can leverage the social network aspect to enhance the word of mouth (WOM) (Susarla et al., 2012; Yoganarasimhan, 2012). In addition, user engagement is facilitated on social media platforms. Marketers may improve customer relationship by addressing users' feedbacks and concerns directly.

Extant research has mainly focused on event or sports sponsorship (Hastings, 1984; Otker and Hayes, 1987; Rodgers, 2003). In general, the stronger the associative link[3] is, the impacts on sponsor recall and attitude toward the sponsor are higher (Rodgers, 2003). That is, more *relevant* sponsor-sponsee matchups are more persuasive. However, there are two vital challenges to this conclusion. First, the definition of relevance seems arbitrary. In the multi-dimensional competing environment, it is hard, if not impossible, to define whether a partnership is relevant based on the corresponding industries. Second, this definition fails to explain the success of sponsorship campaigns of two seemingly unrelated companies[4] or the failure of two similar partners.

In this research, I study whether the association between creator and sponsor plays a role in exposure and engagement. That is, I define and explore various forms of creator and sponsor relationship and investigate their impact on content exposure and user engagement. I intend to answer the research question that what pair of creator/sponsor will generate the highest exposure and interaction. My empirical context is Facebook and I focus on video posts as the focal contents in my study. Video sponsorship has been the most popular format in social media advertising with $11.9 billion in 2017 spending.[5] The findings of my study highlight the importance of targeting in the sponsored contents. As important targeting is to online advertising, the literature on targeted sponsored contents are not as developed as search advertising or other forms of online ads. I elaborate various effective ways of targeting in sponsoring videos on social media.

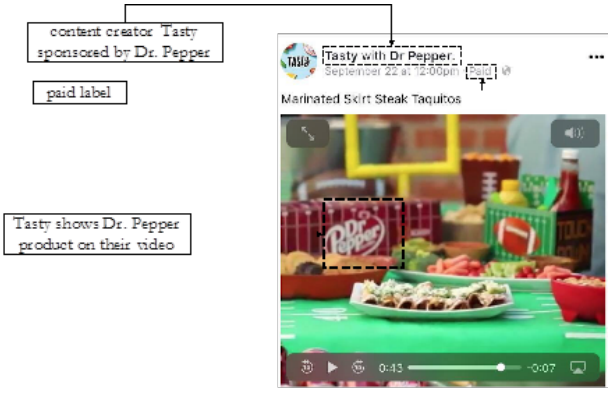
---

[3]An associative link refers to the perceived strength of the sponsor-sponsee matchup (Rodgers, 2003)

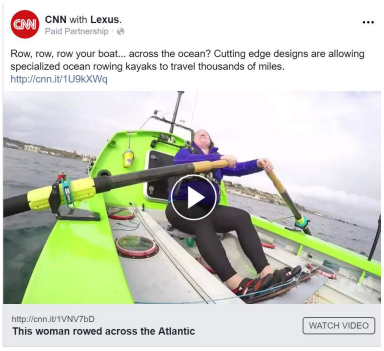
[4]Examples of this include the partnership of JetBlue-Boston Celtics and BudLight-NFL.

[5]IAB, Internet Advertising Revenue Report in 2017.

I use a unique data set from sponsored and non-sponsored videos on Facebook and leverage the branded content tool. The feature was introduced by Facebook to facilitate the creator-sponsor relationship. Many creators and sponsors have adopted sponsorship tools on Facebook. A sponsored content refers to a post on social media that features a business partner. Creators tag their business creators in their sponsored content posts and receive monetary incentives in return. The content is labeled as paid and the sponsor's name appears next to the time stamp. Figure 5.1 visually presents how the relationship of content creators and sponsors are featured on Facebook by two examples. The video in Figure 5.1a is published by the content creator *Tasty* and sponsored by *Dr. Pepper*.[6] Tasty shows and uses a Dr. Pepper product on their video. However, not all the creators feature the sponsored products in their videos and description. Figure 5.1b shows a video created by *CNN* and sponsored by textitLexus.[7] The video does not feature sponsor's product in this example.



(a) Content Creator (Tasty) and Sponsor (Dr. Pepper) while the product is featured in the video and description.



(b) Content Creator (CNN) and Sponsor (Lexus) with no featuring the sponsored product in the video and description

Figure 5.1: Samples of sponsored contents on Facebook. Source: Facebook pages of Tasty and CNN.

[6]https://www.facebook.com/buzzfeedtasty/videos/2002493509764636/

[7]https://www.facebook.com/cnn/videos/10154725684936509/

The link between the creator and sponsor may create synergy and help the sponsor's image as well as mitigate the credibility concerns. This association can be characterized in various ways. I formalize three forms of relevance between creator and sponsor: industry and theme, content, and audience. First, I use a binary variable indicating whether creator and sponsor are from the same category. The categories are defined by Facebook. Second, I define much more flexible measures to capture the similarity and relevance of two entities on Facebook. Thus, I define *Content Similarity* and *Audience Overlap*. For *Content Similarity*, I utilize published contents to measure how close a creator is to a given sponsor based on the published posts. By *Audience Overlap*, I introduce another aspect of creator-sponsor relationship that captures the relevancy regarding the audience usage patterns. That is, the relevance is measured by the ratio of creator's audience that also visit another creator. I elaborate on these metrics in the following sections.

One potential caveat in the analysis of sponsored and non-sponsored videos is that the contents are inherently different. This goes back to the aspect of the sponsorship about the origin of the content -whether it is sponsor-born. My identification strategy encapsulates using latent video topics and utilizing a control function (CF) approach (Heckman and Robb Jr, 1985) to address the potential selection. I employ a Natural language Processing (NLP) technique developed by Rosen-Zvi et al. (2004) to process the unstructured text data in video title and description. Consequently, I can identify heterogeneous latent topics of videos for my identification strategy. Alternatively, I use various matching methods to show the robustness of the findings.

I find that sponsored videos suffer by 65% decrease in viewership and gain 87% less user engagement compared to non-sponsored ones. Also, sponsor's presence represented by its name in the title or the description has no significant effect on viewership and engagement for sponsored videos. However, the effect of sponsor's presence is positively moderated by the sponsor-creator *Audience Overlap*. The results suggest that users do not react to stronger presence of sponsor in the branded contents. However, the effect is positively moderated by when creator and sponsor share higher user overlap. Both, *Content Similarity* and

Audience Overlap have strong positive impacts on viewership and engagement. The results of this research have important implications for both creators and sponsors. To fully monetize the videos, creators seek to maximize the value generated by the sponsored contents. The value is determined by the generated exposure and interactions. In addition, creators intend to find the right sponsors in terms of user experience. Hence, creators can benefit by choosing the right partner in terms of relevance measures. Furthermore, brands benefit by selecting the right creator in two ways: 1) they can raise awareness by uplifting viewership and engagement of the sponsored contents, and 2) they build stronger association to the content leading to higher persuasion.

In addition to managerial implications, this research also has policy implications for online advertising regulators. The prevalence of the sponsored contents on social media has called for regulators to enforce policies to prevent unfair and deceptive acts or practices. Unlike the conventional online advertising, sponsorship, however, sometimes fail to clearly identify their commercial nature. In the US, both the Federal Trade Commission (FTC) and the Federal Communications Commission (FCC) regulated sponsored contents. The FCCs *Sponsorship Identification Rule* states that if a broadcaster charges or accepts (or is promised) any money, service, or other valuable consideration in exchange for promoting products, services, or brands, then the broadcaster must disclose at the time of the broadcast: (1) that the programming is "sponsored", "paid", or "furnished", and (2) the identity of the sponsor (Volner and Sheridan, 2018). FTC has its own Enforcement Policy Statement on Deceptively Formatted Advertisements, i.e., its *Native Advertising Guide*. Ordinary television commercials for goods and services, especially short form ads, generally satisfy the rule without need for a specific disclosure if the ad mentions the sponsors corporate or trade name, and the name of its product Online contents, however, have followed various practices as regulation enforcement has not been consistent in social media contexts. Considering such regulations, the question is in what extent sponsorship changes users content consumption. This research helps to answer this question, thereby empowering regulators with the knowledge of sponsorship impacts on user engagement.

The remainder of the paper is organized as follows. In Section 5.1, I introduce the empirical context and explore the data. I define my constructs and measures on creator-sponsor association in Section 5.2. I formulate the model in Section 5.3 and layout the results in Section 5.4. Finally, I discuss the findings in Section 5.5 and conclude the paper with managerial implications and limitations of my approach in Section 5.6.

5.1 Empirical Settings and Data

5.1.1 Sponsored Content on Facebook

Sponsored content on Facebook refers to any post, including text, photos, and videos, which features or is influenced by a content creator's sponsor. Facebook's sponsored content policy implemented in 2016 requires all content creators to tag and reveal their sponsors when publishing any sponsored content. Sponsors include brands, advertisers, and marketers, who sponsor a creator's sponsored content on Facebook. Content creators include celebrities, influencers, public figures, and media publishers. When content creators tag their sponsor, "Paid Partnership" label appears on the branded content posts (Figure 1). The tag and paid label were implemented to improve transparency and consistency of sponsored content on Facebook. Figure 1 presents an example of a sponsored video on Facebook. The video is posted by a content creator, Tasty, and the video is sponsored by Dr Pepper. In the video, Tasty (i.e. creator) features and uses a Dr Pepper (i.e. sponsor) product. The video audience can recognize that the video is a branded content via the tag and label.

Sponsored content that I examine is different from promoted or advertised content where content creators pay to make their posts more likely to appear in their audience's news feed. The performance and success of promoted content depend on the platform's advertisement targeting algorithm. In contrast, sponsored content that I examine is organically shown to the audience who follow the creator's page or whose friend engages with a post.

5.1.2 Facebook Video Data

To examine the user viewership and engagement for sponsored videos, I collect data on videos posted on Facebook in 2016 and 2017. The data is obtained from a third-party aggregator, which is a global video analytic platform. Among creators who post sponsored videos, I identify top 1,000 creators with more than 10,000 total video views. Top 1,000 creators are defined in terms of the total number of sponsored videos uploaded in the platform in 2016 and 2017. After I restrict my sample to English videos published by creators in the United States, the final dataset includes 592 unique creators. I collect information on all videos – both sponsored and non-sponsored videos – posted by 592 creators in 2016 and 2017. The total of 239,557 videos is posted by the creators in my observation period, among which 43,945 videos are sponsored videos. 43,945 sponsored videos feature 5,573 unique sponsors. I additionally collect data on all 396,322 videos posted by sponsors. Table 5.1 summarizes the information on datasets that I used in the analyses.

Table 5.1: Summary of datasets used in the analyses.

Main dataset	
Number of observations (Videos)	239,557
Sponsored videos	43,945
Unique creators	592
Unique sponsors	5,573
Unique partnerships	9,619
Additional dataset for content analysis	
Number of observations (Videos) for sponsors	392,322
Additional dataset for audience analysis	
Unique creators in audience dataset	218,895

For all videos in my data, I observe their creator, video title, published time of the video, video description, video tag, and video duration. Importantly, the data includes information on the number of views each video receives; I observe number of total views 3 days after the video post-date (V3), 7 days after the video post-date (V7), 30 days after the video post-

date (V30), and number of total views as of my data collection date. Video views are an important measure of video's exposure and success in reaching audience. I additionally utilize the platform data API to retrieve information on video engagement rate 3 days after the video post-date (ER3), 7 days after the video post-date (ER7), and 30 days after the video post-date (ER30). The platform measures the engagement rate as the level of engagement (i.e. likes, shares, and comments) benchmarked across all videos on Facebook. For example, the engagement rate of 2.3 indicates that the video is 2.3 times more engaging than the average. Table 5.2 presents the descriptive statistics for 239,557 videos posted by 592 unique creators.

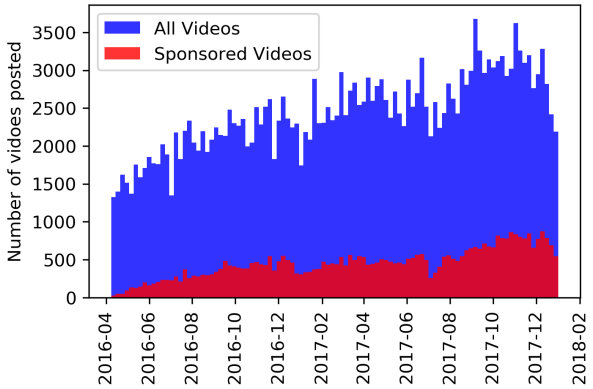


Figure 5.2: Number of videos posted on Facebook in 2016 and 2017 for selected creators.

There are twenty creator categories. The list of categories, the number of unique creators and sponsors, and number of sponsored and non-sponsored videos are listed in Table 5.3. Lastly, as a model-free evidence, I now compare the average of views for three time milestones: 3, 7, and 30 days. They are grouped by sponsored and non-sponsored videos in Figure 5.3. The figure suggests there is a difference between these two categories. However, it should be tested whether this difference is statistically significant and causal.

Table 5.2: Summary statistics of video creators

			Description	Count	Mean	Min	Median	Max
Video-level data	V30		Views by 30 days of publishing	239,557	329,053.4	688	63,833	113,691,344
	views		Total views by the data collection date	239,557	406,907.2	10,000	65,746	143,823,464
	ER30		Engagement rate by 30 days of publishing	239,557	0.99	0	0.75	102.4
	duration		Duration of the video in seconds	239,557	549.1	1	67	1,209,600
	sponsored		Dummy for sponsored video	239,557	0.18	0	0	1
Creator-level data	followers		Creator's followers by the data collection date	591	3,652,469	5,626	1,413,047	92,416,309
	monthly views		Creator's average monthly views	592	27,579,024	342	3,688,024	1,301,681,957
	monthly growth		Average monthly growth in views	592	0.032	0.00001	0.023	0.257
	uploads_90d		Number of video uploads in last 90 days	592	791.1	0	379	12,210
	Demographics	age 13-17	% of age between 13 and 18	495	6.5	0.0	5.5	32.7
		age 18-24	% of age between 18 and 25	495	30.5	0.0	31.5	54.0
		age 25-34	% of age between 25 and 35	495	31.3	0.0	31.7	53.3
		age 35-44	% of age between 35 and 45	495	14.2	0.0	13.4	46.7
		age 45-54	% of age between 45 and 55	495	8.9	0.0	7.9	40.5
		age 55+	% of age above 55	495	8.6	0.0	7.8	39.5
		female %	% of female	495	33.9	0.0	26.9	100.0
	Audience Location	AU	% of from Australia	586	2.58	0.04	1.76	94.50
		BR	% of from Brazil	586	2.13	0.01	1.25	25.22
		CA	% of from Canada	586	3.85	0.07	3.65	59.38
		DE	% of from Germany	586	1.29	0.04	0.92	12.90
		FR	% of from France	586	0.88	0.01	0.51	11.11
		GB	% of from the UK	586	4.20	0.05	3.07	100.0
		IN	% of from India	586	3.23	0.01	0.94	87.47
		IT	% of from Italy	586	0.86	0.01	0.55	17.97
		MX	% of from Mexico	586	2.20	0.01	1.38	58.21
		US	% of from the USA	586	64.74	3.12	67.40	100.0

The variables on the top section of the table are measured for each video. The variables in the bottom are defined at the creator level. They include number of followers, monthly views, number of uploads in the last 90 days and the demographic (age and gender) distribution.

Table 5.3: Summary creator categories with number of unique creators and sponsors videos.

	Creators	Sponsors	Sponsored Videos	Total Videos
Animals & Pets	3	38	104	1281
Beauty	2	23	62	119
Cars, Trucks & Racing	2	25	94	94
Education	1	14	31	193
Entertainment	230	3076	15292	120843
Family & Parenting	1	56	140	140
Fashion & Style	2	34	98	98
Film & Movies	9	91	295	2414
Food & Drink	44	602	2388	22484
Gaming	6	157	964	5098
General Interest	12	113	324	5880
Health, Fitness & Self Help	13	155	762	2893
Home & DIY	5	124	512	860
Kids Entertainment & Animation	3	57	117	743
Music & Dance	9	75	258	4650
News & Politics	30	346	1795	18883
People & Blogs	27	234	983	9333
Science & Tech	9	77	672	4196
Sports	181	1728	20086	62661
Travel	4	113	379	423

5.2 Creator-Sponsor Association

To examine the extent to which creator-sponsor association affect sponsored videos viewership and engagement, I construct two variables that measure creators and sponsors relationship: *Content Similarity* and *Audience Overlap*. First, I construct a variable *Content Similarity* which measures video topic similarity between creators and sponsors. Second, I retrieve from the platform API the cross-creator Audience Overlap measure, which is the ratio of a creator's audience that also watch another creator. For example, if creator A has an audience overlap of 0.6 with sponsor B, it indicates that 60% of creator A's audience is also sponsor B's audience.

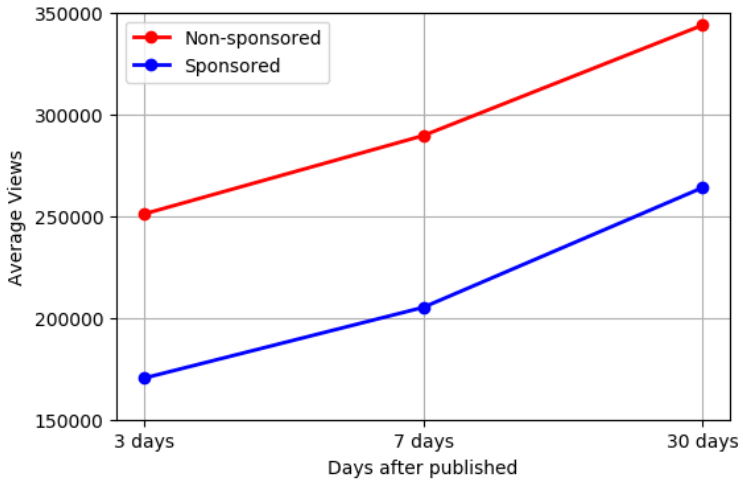


Figure 5.3: Model-free evidence showing the difference of views for sponsored and non-sponsored videos.

5.2.1 Video Topic and Video Content Similarity

The variable *Content Similarity* measures video topic similarity between creators and sponsors. To construct the variable, I employ a Natural Language Processing (NLP) technique to process a large amount of unstructured text data included in video title and description. More specifically, I use the author-topic model developed by Rosen-Zvi et al. (2004), which is an extension of the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). The LDA model is a widely used topic modeling algorithm, which seeks to discover latent topics in a collection of documents. The LDA model posits that each document is a mixture of topics and that each word in the document is attributable to one of the document's topics (Blei et al., 2003; Hofmann, 1999). The author-topic model extends the LDA to include authorship information, such that each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words (Rosen-Zvi et al., 2004). The author-topic model allows us to simultaneously model the content of documents (i.e. videos in my context) and the interests of authors (i.e. content creators in my context).

I identify video topics and measure content similarity by the following steps. First, for

all videos, I collect all text data included in video title and description. I posit that the text data can adequately capture video topic because the importance of choosing relevant video titles and description has been emphasized by many video creators are marketers (Hoben, 2018). Then, I pre-process and clean the text data. I convert words to lower case, remove stop words, and apply lemmatization. I provide additional details on text pre-processing procedure in Appendix H.3. Afterwards, I estimate the author-topic model using the Expectation Maximization (EM) (Hofmann, 1999). Appendix H.2 provides more information regarding the author-topic model and its estimation. I find that the author-topic model with 10 topics results in balanced topic distribution across documents.

Text Insights

Table 5.4 presents each topic with the list of top words. I label topics based on the distribution of words included in topics. For example, for the topic with "recipe," "chicken," "cake," "easy," "chocolate," "cheese," "food," "perfect," "delicious," and "butter" as top words, I label them as Cooking. For each video, I estimate the posterior distribution of topics; each video has different weights for the 10 topics. Each weight takes a value between zero and one and the sum of the weights across the topics equals to one. In my empirical model, I include the video topic distribution weights as control variables to account for video content.

I present visual representations of my author-topic model results through word clouds. Word cloud graphs show the frequent terms in each topic in which the size of each word is proportionate to the frequency of weight of that term. Figure 5.4 presents word clouds for 10 topics trained in my model.

A key advantage of the author-topic model is that it allows us to additionally estimate each content creator (i.e. author)'s distribution over topics. Content creators' distribution over topics measures their interest. For example, the distribution of topics for the content creator, NBA (National Basketball Association), shows that the creator produces more sport-related videos. The next step is to check how the identified topic distributions can differentiate videos. To do this task visually, I perform a probabilistic machine learning

Figure 5.4: Word cloud representation of terms in 10 topics.

Table 5.4: List of top words and topic label.

Label	Word 1	Word 2	Word 3	Word 4	Word 5
Lifestyle (Buzzfeed News)	life	look	friend	share	credit
Sports	game	football	win	night	week
Shows	watch	episode	game	live	tonight
Racing	world	race	redbull	bike	watch
Soccer	goal	fc	nycfc	football	club
News	live	watch	video	news	trump
Live	day	live	today	win	join
Workout	workout	time	video	week	tag
Cooking	recipe	chicken	cake	easy	chocolate
Social	video	use	instagram	channel	youtube

technique called *t-SNE* developed by Maaten and Hinton (2008). t-SNE transforms high dimension vectors into two dimensional data points to visualize clustering patterns. It has been shown that this technique performs favorably for high dimensional data that lie on several (and possibly related) manifolds (Maaten and Hinton, 2008). I perform this algorithm on estimated topic distributions for *all* creators.[8] The outcome of t-SNE is a data point in a two dimensional space for each creator. Figure 5.5 presents a graphical representation of creators by their distributions over topics. The circle size represents the number of videos posted by the creators. Each creator is colored by its dominant topic to visualize the relationship of the topics and videos. The videos are clearly clustered with their dominant topic and well-distanced with other videos. This suggests that the author-topic model can efficiently classify the videos based on the most important topic of the videos.

The model estimates the posterior distribution of the topics for creators. This distribution represents creator's interest and taste. For instance, the distribution for creator *NBA* shows the overall tendency of the creator to produce more sports related videos. The resulting vector is at the creator level and the same for all the videos for a creator. For example, in Figure 5.6, I show sample topic distributions for three creators *NBA*, *CNN*, and *Tasty*. As

[8]I select creators with more than 20 videos.

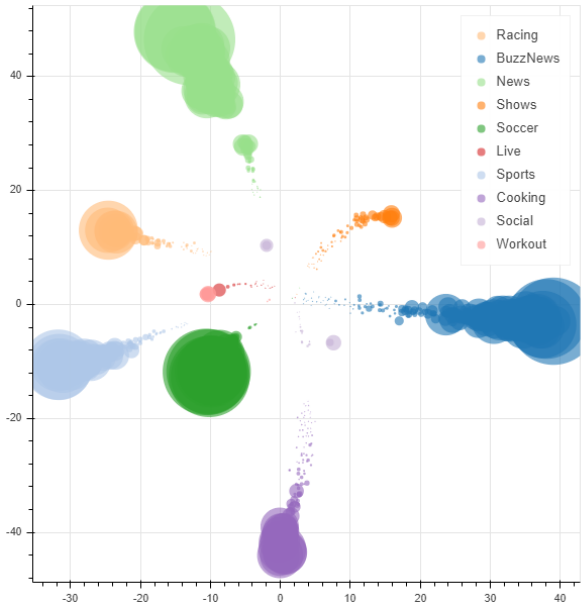


Figure 5.5: *t-SNE* representation of creators' distributions over topics. The circle size represents number of videos for creators. The points are colored by the dominant topic in the distribution.

the graph suggests, videos of *NBA* are more focused on *Sports* and *Live*, *CNN* has more videos on *News* and *Lifestyle*, and *Tasty* has focused on *Cooking*. All of these insights are in line with the actual contents of these creators.

Similarity Measure

I can calculate the similarity between any two authors i and j by using the cosine similarity between the topic vector of author i and the video topic vector of author j. Therefore, I incorporate data on videos posted by sponsors and construct the variable *Content Similarity* between a content creator and a sponsor (i.e. two authors). By definition, the variable *Content Similarity* is between 0 and 1 and Figure 5.9 presents the histogram of the estimated *Content Similarity* for sponsored videos.

Table 5.5 presents the list of creators who are most similar to NBA. Unsurprisingly other creators are sports-related. At the same time, it is important to note that two creators can

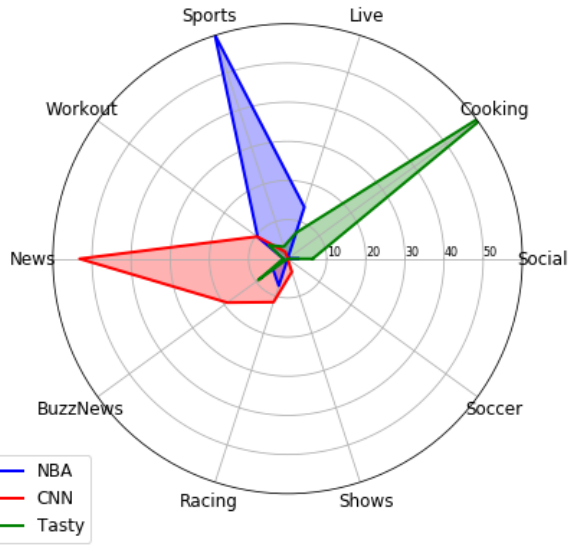


Figure 5.6: Topic distribution for *NBA*, *CNN*, and *Tasty*.

have a low similarity measure even though they are in the same industry. For example, my model indicates that Columbia Sportswear, an American manufacturer of outerwear and sportswear, and Manchester United, a British soccer club, have low similarity score despite the fact that both are in the sports industry. When I examine videos two creators have posted on Facebook, I find that two creators have a different video topic distribution. Figure 5.7 presents the video topic distribution for the two creators. Manchester United has higher weights on the topic Soccer (Topic 5) whereas Columbia Sportswear has a flatter distribution with greatest weight on topic Racing (Topic 4).[9]

5.2.2 Audience Overlap

I capture another aspect of similarity using the audience overlap of the creators. That is, how likely is that two creators are of interest of same users. It is worth noting that this approach captures an *ex-post* similarity, whereas content similarity proposes an *en-ante* closeness. An

[9]As another example, NBA, American basketball league, and Tottenham Hotspur, a London-based soccer club, have low similarity score.

Table 5.5: List of the most similar pages to *NBA*.

Creator	Similarity Score	Number of Videos
NBA	1	1870
San Antonio Spurs	0.912	181
FOX Sports	0.878	280
Los Angeles Lakers	0.872	985
NBC Sports	0.871	88
Chicago Bulls	0.870	143
NBATV	0.862	192

advantage of this measure is that it focuses on the primary users rather than the creators. In other words, two creators may provide totally different contents, say sports and food sections, yet be consumed by same users; i.e., they belong to the same consumption basket of users. This allows us to capture the complementary effects of content creators on social media.

To measure the user overlap for two creators, first I use the platform API to collect overlap measure for 6,165 creators (Among them, 592 are creators and 5,573 are sponsors). For each creator, the platform provides audience overlap measure for the top 2,000 other creators. That is, the top 2,000 Facebook pages that have the highest overlap with the creator of interest are listed.[10] Those Facebook pages, aka creators, in the overlap dataset may or may not include the creators in my dataset.[11] In total, there are 218,895 creators among which 6,165 entities are in my focal dataset. There are 9,619 creator-sponsor pairs in my data. Ideally, I want the audience overlap measure for all 9,619 pairs. However, the platform provides audience information on selected creator-sponsors as there is not enough data on every pair of creator and sponsors. Thus, the observed data is very sparse. This problem is known as the Netflix problem in which the data representing the ratings by users

[10]Note that some creators may have less than 2,000 pages to have overlap with.

[11]For example, consider I have creators 'A', 'B', and 'C' in my dataset. In the overlap data for 'A', I see overlap measures with 'B' and 'X'. Thus, the overlap data for 'A' includes 'B' (which is in my dataset), and 'X' (which is *not* in my dataset), but does not include 'C'.

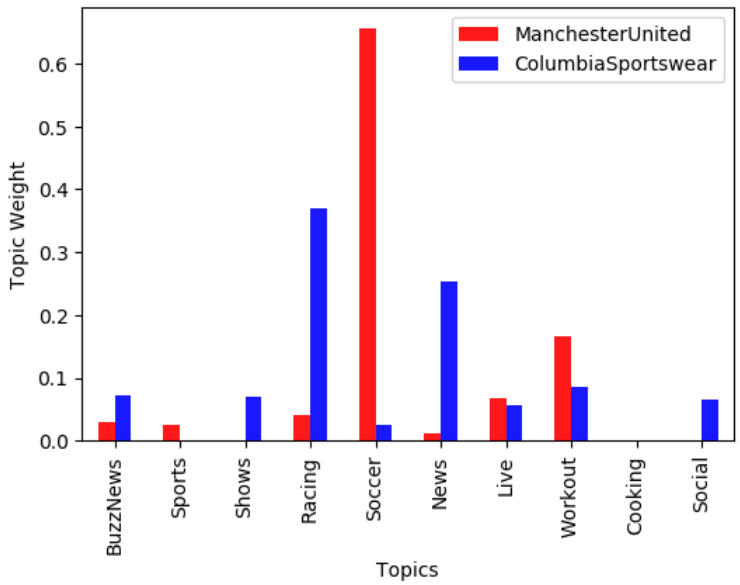


Figure 5.7: Comparison of *Manchester United* and *Columbia Sportswear*. They both belong to the *Sports* category while being not very similar based on the contents.

is very sparse. In the Netflix problem, there are a lot of users that do not rate the movies, yet I want to learn about their preferences.

Similarly, while I do not have data on all the possible switching (or preferences) behavior of users, I want to predict the consumption basket of users for the contents on Facebook. To overcome this empirical challenge, I employ *matrix completion* (Candès and Recht, 2009). After the matrix completion, I have audience information for 9,619 creator-sponsor pairs.

The key assumption that I rely is that there is an underlying structure in the patterns of users consume the contents on Facebook. In the following section, I present the classic matrix completion method: *nuclear norm low-rank minimization*.

Matrix Completion

Matrix completion is the process of predicting missing values for a matrix. One of the main applications of matrix completion methods is in *Collaborative Filtering* (Breese et al., 1998)

in which the matrix is the preference data set for individuals over topics, movies, and etc. From all the entries of data matrix M, I only observe a sparse set of Ω of observations ($M_{ij} \in \Omega$ where $(i, j) \subseteq \{1, ..., n_1\} \times \{1, ..., n_2\}$). The goal is to make an educated guess about missing values as Candès and Recht (2009) describe. In many cases, the matrix I wish to recover is known to be structured as a *low-rank* matrix. In the Netflix example, this assumption implies that the ratings are affected by a few factors. I use a very straightforward approach in matrix completion. The algorithm solves a minimization problem of finding the lowest rank matrix that can explain the non-missing data. Hence, the missing entries can be predicted by the estimated rank. Figure 5.8 depicts a visual representation of the matrix completion process. I present the technical details of this method as well as how the underlying assumptions are satisfied in my problem in Appendix I.

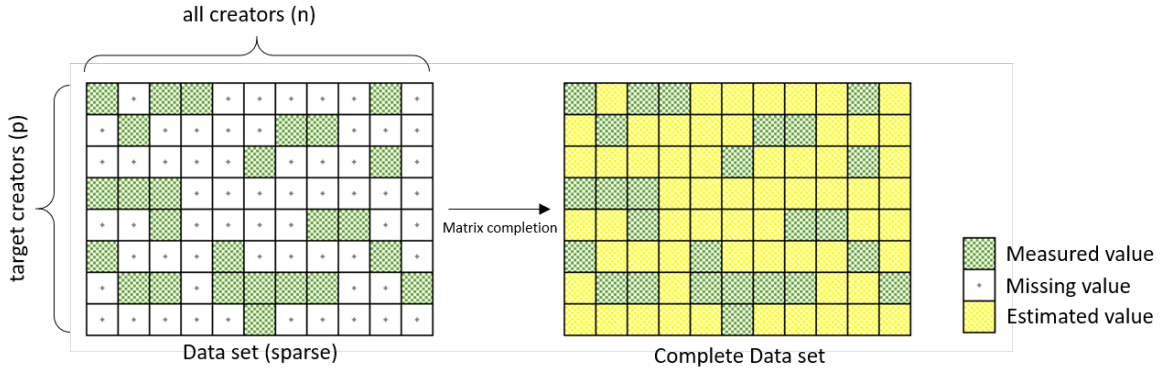


Figure 5.8: Matrix completion algorithms estimate missing values in a sparse data set.

I can form a matrix in which the rows correspond to my focal creators (and sponsors) and columns correspond to all the Facebook pages derived in the overlap data. Thus, the resulting matrix has the dimension of 6165×218895. I use the variable *Audience Overlap* introduced in the previous section. In the final matrix there are $1,337,342,286$ (99%) missing values. After implementing the algorithm, total number of predicted zeros are $2,574,534$. The prediction accuracy can be captured by RMSE for the test data set (10% of the original data set). I get test RMSE of 0.045 showing the matrix completion performs well in predicting

missing overlap data. Figure 5.10 shows the histogram of estimated audience overlap. As shown in the figure, there is a mass density for overlap at zero which is expected. There is a positive correlation between these two measures (0.33). Figure 5.11 depicts the relation between *Audience Overlap* and *Content Similarity*.

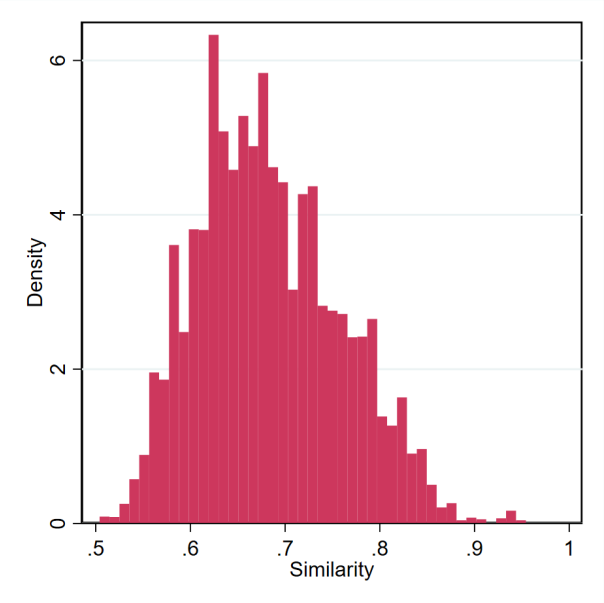


Figure 5.9: Histogram of estimated content similarity of creator and sponsor

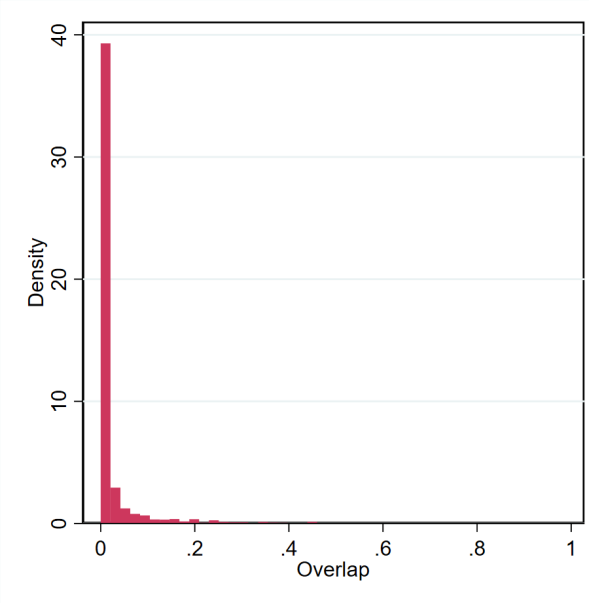


Figure 5.10: Histogram of estimated audience overlap of creator and sponsor

The measure *overlap* has positive correlation with *same_category* (correlation=0.222), but less than that of *similarity*. It shows that *overlap* has higher degrees of freedom, compared to *similarity*, to capture the dynamics of sponsor and creator. For example, *Seahawks*, Seattle-based NFL franchise, has a high audience overlap with *Delta*, an American carrier with a major hub in Seattle. As another example, *Now This*, a left-wing news aggregator, has a high overlap with *Pepsi*. Thus, overlap captures complex patterns of interests: political affiliations, geographical closeness, content relevance, complementary products, and so forth.

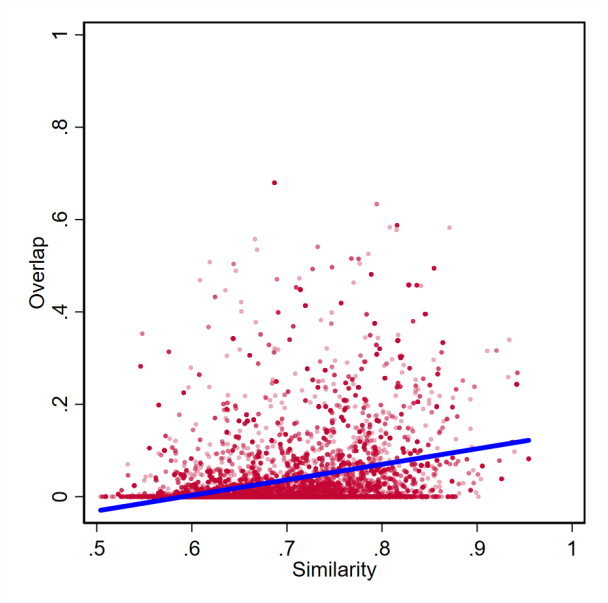


Figure 5.11: Correlation between audience overlap and content similarity with the fitted line.

5.3 *Empirical Model*

5.3.1 Effects of Sponsorship on Video Views and Engagement

To study the effect of sponsorship on video views and engagement, I first empirically examine how video views and engagement of sponsored videos differ from those of non-sponsored videos. To do so, I combine a fixed effects model with a control function approach. I estimate the following specification:

$$Y_i = \beta Sponsored_i + X_i\gamma + \delta_t + f_c + \varepsilon_i, \tag{5.1}$$

where y_i measures the views and engagement levels of video i uploaded by a creator c on date t. The outcome variables include V30 and ER30, which are important measures that reflect the popularity and success of video campaigns. The outcome variables are transformed into log forms to account for the dispersion of the outcome variables. $Sponsored_i$

is an indicator variable that equals one for sponsored videos and zero for non-sponsored videos. X_i is a vector of video characteristic and time-varying content creator characteristic controls. Video characteristic variables include video topics that I identify using an author-topic model technique. I additionally control for video duration, the number of videos uploaded by the creator in the last 90 days, and the number of content creators' followers. f_c indicates the creator-level time-invariant fixed effects. Content creators' time-invariant unobservable characteristics, including their overall reputation, popularity, experience, and video techniques would have heterogeneous effects on video viewership and engagement. Therefore, I employ a fixed effects model to account for content creators' time-invariant unobservable heterogeneity. In order to capture the seasonality and time trend in video activities, I include δ_t, which controls for the year, month, and the day of the week the video is uploaded. ε_i is an idiosyncratic random error term.

If sponsored videos are systematically different from non-sponsored videos, there can be a potential endogeneity problem. I highlight the fact that the endogeneity issue in sponsored contents is not as severe as the one in online advertising. As the level of targeting in vastly less than the online ads, I expect the bias to be significantly less than a typical advertising study (Gordon et al., 2019). Due to nature of sponsorship, individual targeting is not feasible. Nevertheless, I take several measures to address the potential endogeneity problem. First, the panel nature of the data where I observe all videos uploaded by the same content creators over time is an important source of identification because it allows us to examine how the same creators' video views and engagement change with the sponsorship status. The panel data and content creator fixed-effects model offer a partial solution to the endogeneity problem because they take into account time-invariant unobserved creator heterogeneity. Second, I include a wide range of control variables and incorporate rich information embedded in video title and description to control for other factors that might affect video outcomes. Third, to more robustly address the potential endogeneity problem, I employ a control function (CF) approach to account for sponsorship selection on observables (Heckman and Robb Jr, 1985; Petrin and Train, 2010).

In the control function approach, I estimate the expected sponsorship status (i.e. treatment status) conditional on observables:

$$Sponsored_i = \begin{cases} 1 & \text{if } Z_i\pi + v_i > 0. \\ 0 & otherwise \end{cases} \tag{5.2}$$

where Z_i indicates a vector of variables that affects whether a video is sponsored or not. The vector includes video characteristic, content creator characteristic, and time control variables. Under the assumption that the error terms ϵ_i in equation (1) and v_i in equation (2) are bivariate normal with mean zero, the conditional expectation of the error term ε_i can be formulated as:

$$\mathbb{E}\{\epsilon_i | Z_i, Sponsored_i\} = \sigma_{\varepsilon v}\left[Sponsored_i \cdot \lambda_i(Z_i\pi) - (1 - Sponsored_i) \cdot \lambda_i(-Z_i\pi)\right], \tag{5.3}$$

where $\sigma_{\varepsilon v}$ denotes the covariance between ε and v (Heckman, 1977a,b). The ratio $\lambda(Z_i\pi) = \phi(Z_i\pi)/\Phi(Z_i\pi)$ where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative distribution functions of the standard normal distribution, respectively, and I normalize $\sigma_v^2 = 1$. $\lambda(\cdot)$ is known as the inverse Mill's ratio (IMR).

I estimate the control function model using the Heckman two-step approach. First, I use the probit model for the sponsorship (i.e. treatment) status using maximum likelihood to obtain an estimate of π, or $\hat{\pi}$. Then using the estimate, I compute $\hat{gr} = Sponsored_i \cdot \lambda_i(Z_i\hat{\pi}) - (1 - Sponsored_i) \cdot \lambda_i(-Z_i\hat{\pi})$, which denotes the generalized residual (Gourieroux et al., 1987) of the probit model. In the second stage, I include the generalized residual and estimate Equation 5.1. The CF approach does not require Z_i to include at least one variable that is not included in X_i and is able to identify parameters through the non-linearity implied by joint normality. The coefficient of interest β. If β is negative, it shows the sponsorship has a negative effect on video performance. Following Vella and Verbeek (1999), I can relax the assumption that sponsored and non-sponsored videos have the same variance over views

and engagement. Thus, I define σ_t and σ_n to represent $\sigma_{\varepsilon v}$ for sponsored and non-sponsored videos respectively.

5.3.2 Effects of Creator-Sponsor Relationship on Sponsored Videos

To explore the effect of sponsor presence, I introduce the variable $featured$ which is a dummy variable indicating whether sponsor's name is mentioned in either the title or the description of the video.[12] Equation 5.4 formulates the moderating effect of $featured$.

$$\begin{aligned} Y_i = \beta Sponsored_i + \zeta\ featured + \alpha Sponsored_i \times featured + \\ X_i\gamma + \delta_t + f_c + \mathbb{E}\{\varepsilon_i | Z_i, Sponsored_i\}, \end{aligned} \tag{5.4}$$

where the parameter of the interest is α. If the effect of sponsorship (e.g. due to user annoyance) is negative, the effect of sponsorship is intensified by $featured$.

I now present the model for the impact of creator-sponsor association. The effect of interest is defined on the treated (sponsored) videos. Hence, I formulate Equation 5.5 to show the moderating effect of association on $featured$.

$$\begin{aligned} Y_i = \beta Association_i + \zeta\ featured + \alpha Association_i \times featured + \\ X_i\gamma + \delta_t + f_c + \mathbb{E}\{\varepsilon_i | Z_i, Sponsored_i\}, \end{aligned} \tag{5.5}$$

where $Association \in \{ContentSimilarity, AudienceOverlap\}$.

5.4 Results

I first explore the results on the impact of sponsorship on viewership and user engagement. Next, I present the effect of creator/sponsor similarity on sponsored contents.

---

[12] For example, creator 'AWE me' published a sponsored video featuring the sponsor, T-Mobile, in the description: *"Fear the Daywalker! Thanks T-Mobile for making this build possible."*

5.4.1 Effects of Sponsorship on Video Views and Engagement

Table 5.6 shows the results of three models for the linear regression without endogeneity treatment for $\log(V30)$. Model (1) includes time control variables but no fixed effect and creator characteristics. Model (2) presents the the results for the fixed effect model. As noted, using a fixed effect model, I cannot use time-invariant characteristics of the creators. However, I use creator's features in model (3) with a random effect model. The coefficient for *sponsored* is consistently negative and statistically significant across all models. In the fixed effect model, sponsorship results in 28.3% loss of viewership. I note that the fixed effect model improves model fit substantially (Adjusted R^2 from 0.021 to 0.426). This shows the that a significant part of variation in video views can be captured by creator fixed effect. Table 5.7 summarizes the effect of sponsorship on the engagement ratio ($ER30$). The model configurations are the same as Table 5.6. The results are similar to those of $\log(V30)$; Sponsorship has negative impact on user engagement.

Additionally, fixed effect model is consistent and robust to biases from time-invariant latent variables. However, random effect model is efficient. Thus, random effect model is superior to fixed effect model if it is consistent. Hence, I perform *Hausman* test on two similar random effect and fixed effect models to check whether the random effect model is unbiased. The null hypothesis in that the difference in coefficients is not systematic. The test statistic is 78.70 ($p-value < 0.001$) for the viewership model and 90.43 ($p-value < 0.001$) for the engagement model. They show that the null hypotheses are rejected, thus random effect model is biased. We, therefore, use fixed effect models hereafter.

Selection Treatment

I now show the results for the control function treatment for sponsorship on viewership and engagement. Table 5.8 summarizes the sponsorship results with accounting for the selection. Dependent variable for models (1)-(2) is $\log(V30)$ and for models (3)-(4) is $ER30$. IMR in Model (1) and Model (3) is the Inverse Mills Ratio derived by the control function (CF)

Table 5.6: Sponsorship effect on V30 (views on first 30 days)

	(1)	(2)	(3)
sponsored	-0.199*	-0.332***	-0.340***
	(-2.42)	(-5.43)	(-39.92)
log(duration)	-0.0471**	0.0129	0.00964***
	(-2.86)	(1.57)	(6.32)
log(monthly_growth)			-0.0162
			(-0.60)
uploads_90d			0.000003
			(0.09)
age_18_24			-4.709*
			(-2.23)
age_25_34			2.176
			(1.23)
age_35_44			-5.064
			(-1.66)
age_45_54			-6.246
			(-1.00)
age_55_plus			0.268
			(0.04)
female_percent			-0.275
			(-1.35)
Time Control	✓	✓	✓
Geo Distribution			✓
Creator Category			✓
Creator FE		✓	
Constant	11.49***	11.11***	4.513***
	(84.80)	(193.33)	(3.71)
Observations	239557	239557	224515
Adjusted R^2	0.006	0.546	

t statistics in parentheses.

Dependent variable is $\log(V30)$ (viewership in the first 30 days of exposure). Model (1) is the the linear model in which sponsored is the independent variable. Model (2) is fixed effect model capturing creator's time-invariant unonbservables. Model (3) is the random effect model with creator's characteristics. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5.7: Sponsorship effect on ER30 (Engagement Rate in 30 days)

	(1)	(2)	(3)
sponsored	-0.218***	-0.273***	-0.276***
	(-5.06)	(-10.27)	(-41.12)
log(duration)	-0.0264	0.0218*	0.0249***
	(-1.44)	(2.01)	(20.75)
log(followers)			0.0687**
			(3.13)
log(monthly_growth)			-0.0194
			(-0.83)
uploads_90d			-0.00008**
			(-3.13)
age_18_24			1.275
			(0.69)
age_25_34			-0.215
			(-0.14)
age_35_44			0.662
			(0.25)
age_45_54			5.377
			(0.99)
age_55_plus			-2.327
			(-0.38)
female_percent			-0.0374
			(-0.21)
Time Control	✓	✓	✓
Geo Distribution			✓
Creator Category			✓
Creator FE		✓	
Constant	1.311***	1.043***	1.430
	(11.84)	(16.81)	(1.35)
Observations	239557	239557	224515
Adjusted R^2	0.021	0.426	

t statistics in parentheses.

Dependent variable is ER30 (Engagement Ratio in 30 days). Model (1) is the the linear model in which sponsored is the independent variable. Model (2) is fixed effect model capturing creator's time-invariant unonbservables. Model (3) is the random effect model with creator's characteristics. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

approach. σ_t and σ_n in models (2) and (4) are the selection parameters for the relaxed control function representing potential different variation in treatment and control groups. The results are consistent with the those of Tables 5.6 and 5.7. However, the sponsorship effect is stronger after accounting for the selection. The effect of sponsorship are more negative in the CF models for both performance measures. Sponsorship results in 59% to 66% loss of viewership after accounting for selection. This suggests that the bias is downward. That is, those videos with higher likelihood of more viewership are also more likely to get sponsored. Thus, in the models with no selection treatment, the negative effect of sponsorship is biased toward zero. Also, the significance of Inverse Mills Ratio, σ_t, and σ_n suggests that there is indeed selection problem. It makes the previous results biased as explained.

5.4.2 Relevance Effect

Now I explore the impact of *relevance*[13] on viewership and engagement and its mediation role on sponsor featuring. First, I study whether hard-defined similarities, in terms of industry categories, have impact on views and engagement. In Table 5.9, I introduce three dummy variables: $same\_industry$, $same\_category$, and $featured$ in the fixed effect model. The first two variables capture whether the sponsor and creator are from the same industry and genre, respectively. The effect is not statistically significant within my confidence interval suggesting that there might not be a synergy effect between the creator and sponsor when they are in the similar section. The last variable captures whether the sponsor has been featured in the video description. Since, $featured$ is only defined for the sponsored videos, the estimated coefficient shows the interaction effect on the sponsorship. The effect of $featured$ is not significant. Similarly, the synergy and featuring effect also do not seem to positively or negatively affect the engagement in the sponsored videos. Note that only sponsored videos are included in the models as the variables of interest are not defined for non-sponsored videos. Hence, I can interpret the reported effect as an interaction effect with sponsorship. Models (2) and

[13]As discussed earlier, I introduce three measures of relevance: genre (category) accordance, content similarity, and audience overlap.

Table 5.8: Sponsorship effect with selection treatment (CF)

	(1) log(V30)	(2) log(V30)	(3) ER30	(4) ER30
sponsored	-0.893***	-1.072***	-0.867***	-1.233***
	(-3.68)	(-3.81)	(-6.15)	(-5.70)
log(duration)	0.0177*	0.0195*	0.0304**	0.0340**
	(2.17)	(2.34)	(2.79)	(3.18)
IMR	0.306*		0.312***	
	(2.57)		(4.40)	
σ_t		0.366**		0.435***
		(2.87)		(4.58)
σ_n		0.580**		0.871***
		(2.63)		(3.95)
FE (Time & Creator)	✓	✓	✓	✓
Constant	35.12	43.20	6.033	22.55
	(1.47)	(1.71)	(0.51)	(1.88)
Observations	220311	220311	220311	220311
Adjusted R^2	0.539	0.539	0.432	0.432

t statistics in parentheses

Dependent variable for models (1)-(2) is log($V30$) and for models (3)-(4) is ER30. IMR in Model (1) and Model (3) is the Inverse Mills Ratio derived by the control function (CF) approach. σ_t and σ_n in models (2) and (4) are the selection parameters for the relaxed control function representing potential different variation in treatment and control groups. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(4) in Table 5.9 report the results with selection treatment by CF approach. Note that when I restrict the sample to sponsored videos, Inverse Mills Ratio is not statistically significant anymore. The synergy effects are similar to the ones in models (1) and (3) in Table 5.9. But, featuring effects are negative with CF approach, however, they are not significant. I cannot identify whether this is because of lack of impact or existence of noise.

Next, I investigate the effect of other forms of relevance on sponsorship: Content similarity and audience overlap. I present the effect of content similarity on the sponsorship

and featuring sponsor in Table 5.10. Table 5.11 summarizes the results of audience overlap on sponsored videos as well. On one hand, I find that *Content Similarity* has a positive impact on both viewership and engagement in sponsored videos. *Content Similarity* leads to 101.2% increase in viewership. Note that the category similarity had no effect meaning that the similarity defined by content is more powerful in describing the sponsored video performance. However, the similarity does not seem to have a significant interaction effect on $featured$. It suggests, similarity effect does not change by featuring the sponsor. On the other hand, *Audience Overlap* has positive impact on viewership and engagement. *Audience Overlap* leads to 69.9% increase in viewership. Interestingly, audience overlap has a positive interaction with $featured$ on $\log(V30)$. It suggests that featuring a sponsor has a higher performance if the creators and sponsors have higher audience overlap. However, the interaction effect on $ER30$ is not statistically significant.

5.4.3 Heterogeneous Effect of Relevance and Similarity

As the next step, I explore how the effects of relevance and similarity are different across the creator categories. I preformed series of models on different samples divided by creator categories. I could not identify the effect of interest on all categories due to limited number of observations. Among all the models, I present the ones with the strongest similarity effect for $\log(V30)$ and $ER30$ in Tables 5.12 and 5.13. On the one hand, I find that the content similarity effect on viewership is the strongest in categories: *Food & Drink* (372% increase in views) and *Sports* (121% increase in views). This highlights the importance of similarity in these categories while I cannot conclude that finding on other categories. However, only *Sports* category remains as significant for user engagement. On the other hand, audience overlap has very strong positive impact on viewership in *Food & Drink* (535% increase in views) and *Sports* (96% increase in views) categories. Note that this does not imply that the effect do not exist in other categories. But, the noise might be high due to limited number of observations so I cannot tease out the effect of interest.

Table 5.9: Sponsor featuring and theme synergy on views and engagement on sponsored videos.

	(1) log(V30)	(2) log(V30)	(3) ER30	(4) ER30
log(duration)	0.0987***	0.104***	-0.0458***	-0.0307*
	(8.08)	(6.53)	(-3.57)	(-2.54)
same_industry	0.113	0.0676	-0.0140	-0.00694
	(1.56)	(0.98)	(-0.34)	(-0.16)
same_category	0.0374	0.0468	0.0695**	0.0809**
	(0.75)	(0.85)	(2.83)	(2.84)
featured	-0.0595	-0.0214	-0.0191	-0.0130
	(-1.15)	(-0.38)	(-1.10)	(-0.72)
IMV		0.0683		-0.216
		(0.23)		(-0.95)
FE (Time & Creator)	✓	✓	✓	✓
Constant	10.68***	10.64	1.253***	0.989
	(75.21)	(0.45)	(13.59)	(0.04)
Observations	43944	34339	43944	34339
Adjusted R^2	0.538	0.540	0.503	0.534

t statistics in parentheses

Models (1) through (4) represent fixed effect model including dummies for *same category*, *same industry*, and *featured*. Dependent variable for models (1)-(2) is log($V30$) and for models (3)-(4) is ER30. Models (2) and (4) are treated for the selection by by the control function (CF) approach and IMR is the Inverse Mills Ratio derived by the selection model. *same_industry* and *same_category* represent if creator and sponsor are both from the same industry and category respectively. *featured* is a dummy that shows whether the name of sponsor is mentioned in the description of the video. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.4.4 Matching

I now discuss the details of the matching methods I used in the paper. I use Mahalanobis distant matching (MDM) and propensity score matching (PSM) for the main analyses. I verify the validity and performance of the matching techniques by conducting several statis-

Table 5.10: Content similarity effect with selection treatment (CF)

	(1) log(V30)	(2) log(V30)	(3) ER30	(4) ER30
similarity	0.699*	0.684*	0.481*	0.488*
	(2.45)	(2.31)	(2.53)	(2.43)
featured		-0.0644		0.0289
		(-0.18)		(0.16)
featured × similarity		0.156		-0.0723
		(0.29)		(-0.27)
log(duration)	0.0899***	0.0902***	-0.0451*	-0.0452*
	(4.43)	(4.48)	(-2.43)	(-2.44)
IMV	-0.277	-0.264	-0.483	-0.489
	(-0.58)	(-0.56)	(-1.28)	(-1.29)
Constant	-20.33	-22.16	-33.29	-32.42
	(-0.54)	(-0.58)	(-0.97)	(-0.94)
Observations	17402	17402	17402	17402
Adjusted R^2	0.535	0.535	0.501	0.501

t statistics in parentheses

Dependent variable for models (1)-(2) is log($V30$) and for models (3)-(4) is ER30. IMR is the Inverse Mills Ratio derived by the control function (CF) approach. Models (2) and (4) are the interaction effect of *featured* and *similarity* on the treated (sponsored) group. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

tical tests and visual measures. I control for the overlap condition which is required for the matching algorithms. I also measure the balance for the observables in treatment and control groups after matching. These two matching algorithms apply two different paradigms to reach balanced samples. In MDM, I follow an exact match procedure; i.e., the distance between treatment and control groups are used to find the matched sample. But in PSM, an indirect match is pursued; i.e., PSM matches a low dimension representation of observables (aka propensity score).

Table 5.11: Audience overlap effect with selection treatment (CF)

	(1) log(V30)	(2) log(V30)	(3) ER30	(4) ER30
overlap	0.530***	0.479***	0.244***	0.228***
	(5.41)	(4.73)	(4.33)	(3.92)
featured		-0.0286		-0.0210
		(-1.21)		(-1.55)
featured × overlap		0.634*		0.185
		(2.01)		(1.02)
log(duration)	0.109***	0.108***	-0.0321***	-0.0322***
	(16.50)	(16.46)	(-8.44)	(-8.48)
IMR	0.0799	0.0679	-0.106	-0.113
	(0.40)	(0.34)	(-0.93)	(-0.99)
Constant	9.150	9.173	9.417	9.659
	(0.18)	(0.18)	(0.32)	(0.32)
Observations	29867	29867	29867	29867
Adjusted R^2	0.548	0.548	0.530	0.530

t statistics in parentheses

Dependent variable for models (1)-(2) is log($V30$) and for models (3)-(4) is ER30. IMR is the Inverse Mills Ratio derived by the control function (CF) approach. Models (2) and (4) are the interaction effect of *featured* and *overlap* on the treated (sponsored) group. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Matching algorithms rely on the so-called Conditional Independence Assumption (CIA) also known as *unconfoundedness* (Nannicini et al., 2007). That is, given observable characteristics, potential outcome and treatment assignment should be independent. In other words, the selection into treatment group is only driven by the factors that are observable. I try to reduce the confoundedness as much as possible by choosing various indicators of buyers and their past usage behavior. As there is no direct way to test this assumption, there are a few ways to check how the results are sensitive to the potential confounders.

Table 5.12: Content similarity per creator category

	log($V30$)		$Er30$
	Food & Drink	Sports	Sports
similarity	1.551**	0.792***	0.717***
	(2.83)	(4.53)	(6.18)
log(duration)	-0.0121	0.0764***	-0.110***
	(-0.24)	(8.05)	(-17.38)
IMR	-0.169	0.0597	0.0833
	(-0.48)	(0.62)	(1.30)
Fixed Effect	✓	✓	✓
Constant	14.08***	10.04***	1.002**
	(9.51)	(17.59)	(2.64)
Observations	931	7302	7301
Adjusted R^2	0.608	0.428	0.450

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

These methods are applied to measure the sensitivity of average treatment effect (ATE) to a bias. However, I do not directly measure ATE by the matching process. Instead, I build an aggregate model by the matched sample. Thus, the sensitivity analysis may not be as straightforward as a typical ATE problem. Therefore, I define a new model to measure the sensitivity of the results to the confounding factors. Note that the results of the sensitivity analysis might not be directly translated to my main results. However, it shows the level of sensitivity of the matched sample.

I define a model in which the dependent variable log($V30$) and $ER30$ defined earlier. The treatment is defined as whether the video is sponsored. I analyze the ATE of sponsorship on views by matched sample and measure the sensitivity of the results.

I follow the methods proposed by Rosenbaum (2002) and Ichino et al. (2008). First, I define the probability of niche product by a logit model as a function of observables, X_i,

Table 5.13: Audience overlap per creator category

	log($V30$)	
	Food & Drink	Sports
overlap	1.848**	0.674***
	(3.11)	(5.99)
log(duration)	0.0470	0.105***
	(1.13)	(14.44)
IMR	0.0587	0.0727
	(0.23)	(1.02)
Fixed Effect	✓	✓
Constant	14.31***	10.14***
	(12.26)	(23.84)
Observations	1546	12028
Adjusted R^2	0.611	0.426

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

and unobservables, ξ_i. Let δ capture the effect of the unobserved component on the niche product selection. If δ is zero, then there is no hidden bias in the process and the niche product probability will be solely determined by the observables.

Following Rosenbaum (2002), I derive a measure to test the significance of δ. The odds ratio between video i and video j being sponsored is $\exp\{\delta(\xi_i - \xi_j)\}$. Rosenbaum (2002) suggests using bounds as Equation 5.6 to test how the inference changes as δ and $\xi_i - \xi_j$ vary:

$$1/e^{\delta} \leq oddsratioofvideo_{ij} \leq e^{\delta} \tag{5.6}$$

Let $\Gamma = e^{\delta}$, thus $\Gamma = 1$ ($\delta = 0$) implies there is no hidden bias. The interpretation of this would be if unobservables cause the odds ratio of video to change by a factor of Γ, what is the probability that the ATE still excludes zeros.

In an alternative approach, Ichino et al. (2008) introduce a simulation-based method in which there is confounder U which can change the treatment assignment and viewership. The probability distribution can be given manually or can be driven by another variable as a potential confounder. I introduce a binary variable *high_growth* describing whether the creator has higher than median growth over the last month.

Mahalanobis Distant Matching

Mahalanobis distant matching (MDM) is an exact matching approach discussed by Cochran and Rubin (1973) and Rubin (1976). Under MDM, matched samples are chosen based on mahalanobis distance defined as:

$$\mathbb{X}_{MDM} = M(X|\sqrt{(X_i - X_j)S^{-1}(X_i - X_j)} < \delta) \tag{5.7}$$

where δ is called caliper, and adjusts how close the matches are, X is the original data, and S is the sample covariance matrix. MDM attempts to approximate fully blocked randomized experiment resulting in more efficient estimates (King and Nielsen, 2016). As a result of MDM, I achieve a balanced control and treatment group. Table 5.14 reports the resulted balance after matching. The null hypothesis is that the mean variables are the same in both treatment and control group. I observe that all the attributes used for matching are balanced after the process.

Table 5.15 presents the results for the sensitivity analysis for the hidden bias as discussed in previous section.

Propensity Score Matching

The steps for PSM is similar to MDM with a significant difference. I calculate propensity scores based on a logit model and use the scores to match. Table 5.16 reports the logit regression results for the propensity score for *sponsored*. I use nearest neighbor approach to match based on the scores. I allowed for multiple matches and used caliper=0.0001 for the

Table 5.14: Balance of attributes after matching

	Mean		t-test		
	Treated	Control	t	*p-value*	V(T)/V(C)
AU	0.034	0.034	0.45	0.651	1.00
BR	0.018	0.018	1.52	0.128	1.17
CA	0.037	0.042	-16.31	0.000	1.25
DE	0.014	0.014	-1.47	0.143	1.21
FR	0.007	0.007	-0.79	0.431	1.01
GB	0.039	0.040	-2.21	0.027	1.12
IN	0.02	0.02	0.12	0.906	1.01
IT	0.007	0.007	0.38	0.703	1.07
MX	0.014	0.013	2.21	0.027	1.12
US	0.659	0.642	6.19	0.000	1.30
age 18-24	0.325	0.325	0.00	1.000	1.00
age 25-34	0.309	0.309	0.00	1.000	1.00
age 35-44	0.127	0.127	0.00	1.000	1.00
age 45-54	0.082	0.082	0.00	1.000	1.00
age 55 plus	0.081	0.081	0.00	1.000	1.00
female	0.282	0.282	0.00	1.000	1.00

Table 5.15: Sensitivity analysis of hidden bias

Γ	*p-value*
1	0.0003
1.1	< 0.0001
1.2	< 0.0001
1.3	< 0.0001
1.4	< 0.0001
1.5	< 0.0001
1.6	< 0.0001
1.7	< 0.0001
1.8	< 0.0001
1.9	< 0.0001
2.0	< 0.0001

tolerance of the matched scores.

I use the simulated-based approach (Ichino et al., 2008) to measure the sensitivity of the results to a potential confounder. I use *high_growth* variable as discussed earlier. I use 100 simulations and bootstrapping for the standard errors. Table 5.17 summarizes the results for the simulation of the confounder. The results suggest that the ATE is not influenced by the confounder (Selection effect on odds ratio is 0.967).

Figure 5.12 shows the distribution of propensity score after matching. It represents a balanced distribution in control and treatment groups.

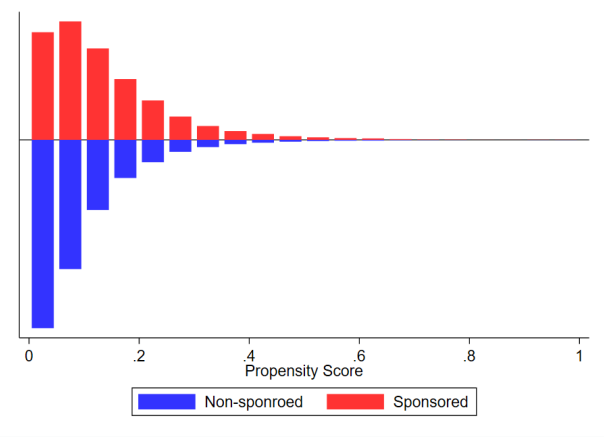


Figure 5.12: Histogram of propensity score for treated and untreated groups.

Robustness Checks

I measured the effect of sponsorship by using PSM approach. Table 5.18 presents the results of average treatment effect (sponsorship) on the matched sample. The results are consistent with the MDM approach.

Table 5.16: Sensitivity analysis of hidden bias

	Logistic Regression
Creator Category:	
Beauty	-0.058
	(0.255)
Education	-3.403
	(0.250)
Entertainment	-1.921
	(0.130)
Film & Movies	-7.847
	(0.279)
Food & Drink	-1.757
	(0.135)
Gaming	-2.239
	(0.140)
General Interest	-2.888
	(0.144)
Health & Fitness	-0.787
	(0.142)
Home & DIY	0.173
	(0.154)
Kids Entertainment	-1.670
	(0.174)
Music & Dance	-3.085
	(0.147)
News & Politics	-2.017
	(0.133)
People & Blogs	-2.410
	(0.138)
Science & Tech	-2.003
	(0.140)
Sports	-1.923
	(0.130)
log(followers)	0.106
	(0.005)
log(duration)	0.037
	(0.005)
Control for Audience and Topics	✓
Constant	277.762
	(94.30)
Observations	227,090
Adjusted R^2	0.251

Standard errors in parentheses. DV is the binary variable *sponsored*.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5.17: Simulation-based sensitivity analysis of hidden bias

ATE	S.E.	Outcome Effect	Selection Effect
-3.83e+05	85821.93	1.288	0.967

Table 5.18: Average treatment effect after applying PSM.

	Treated	Control	Difference
Unmatched	262131.5	517467.8	-255336.3
			(-19.5)
Average Treatment Effect of Treated	262131.5	560462.9	-298331.6
			(-16.16)

t-statistic is parentheses

The dependent variable is $V30$ (*first 30 days views*).

5.4.5 Alternative Approaches as Robustness Checks

Viewership for 3 and 7 days

I run all the models with $\log(V3)$ and $\log(V7)$ as the dependent variable replacing the variable $\log(V30)$. All the results follow similar patterns as the main findings so that the main conclusion would hold. This suggests that the findings of the paper is independent of the growth pattern in viewership. The same conclusion applies to $ER3$ and $ER7$ for the engagement rate.

5.4.6 Inclusion of Non-US Creators

In the main analyses, I restrict the sample to the creators in the United States to remove any geographical impact from the findings. I analyze the model with the pooled data of both US-based and Non-US based creators. Generally, the results are consistent with the main conclusions of the paper.

5.4.7 Audience Overlap

Additionally to robustness checks in Appendix 5.4.4, I perform the analysis with the original variable with no data imputations as an alternative measure for the predicted *overlap*. Results of the regression model with CF treatment are presented in Table 5.19. Overall, the results are consistent with the original findings, but the coefficients of interests are not statistically significant. With many entries missing, the data is much noisier compared to the constructed measure.

Table 5.19: Audience overlap (with no imputation) effect with selection treatment (CF) as robustness check.

	(1) log(V30)	(2) log(V30)	(3) ER30	(4) ER30
overlap_original	0.212	0.153	0.249*	0.247*
	(1.79)	(1.27)	(2.38)	(2.33)
featured		-0.0803		0.120**
		(-1.52)		(2.58)
featured × overlap_original		0.932**		0.225
		(3.07)		(0.84)
log(duration)	0.121***	0.121***	-0.0866***	-0.0873***
	(9.00)	(9.01)	(-7.32)	(-7.38)
IMR	0.227	0.231	0.0537	0.0171
	(0.55)	(0.56)	(0.15)	(0.05)
Constant	53.61	54.18	60.76	57.73
	(0.53)	(0.54)	(0.68)	(0.65)
Observations	7512	7512	7512	7512
Adjusted R^2	0.587	0.587	0.654	0.655

t statistics in parentheses

overlap_original represents the overlap variable with no data imputation. Dependent variable for models (1)-(2) is $\log(V30)$ and for models (3)-(4) is ER30. IMR is the Inverse Mills Ratio derived by the control function (CF) approach. Models (2) and (4) are the interaction effect of *featured* and *overlap* on the treated (sponsored) group. Standard errors are clustered by creators.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.5 Discussion

The results suggest that, on average, sponsorship has negative impact on both viewership and user engagement. I intend to shed light on how sponsorship can be more effective by the instantaneous engagement of the users so the creators and sponsors benefit more from the partnership. There are multiple explanations for this finding. First, the topics in sponsored and non-sponsored videos might be systematically different. That is, specific topics are more prone to be sponsored. Since the viewership is highly correlated with the topics, I can expect the selection in topics makes results biased. Another explanation is that the quality of the video contents are inherently different for sponsored and non-sponsored videos. For instance, organic videos are more likely to have genuine contents, hence, higher quality. Finally, there might be an issue of trust and credibility of the creator by sponsoring their videos. I need to disentangle these explanations with the effect of interest which is the sponsorship effect.

The perfect solution to interpret the causal effect is the randomized field experiment approach. However, I use observational ex-ante data. I use several approaches to account for such problem. First, I use creator fixed effect. By this treatment, I can control for all the time-invariant endogeneity. The findings are consistent with the previous models. Furthermore, I applied the author-topic model approach as an extension to LDA topic model to identify the distribution of the topics for each video and creator. Incorporating estimated distributions of topics in the main model leads to better model fit. That suggests that retrieved topics are good representative of the actual contents.

In addition, I use control function approach to treat for the selection of treated and untreated videos. In control function, theoretically, there is no need of a set of excluded variables. That is due to a non-linear selection model. However, I include estimated video distributions over a set of topics. I verify the results using two different matching approaches: PSM and MDM. I use numerous functional forms as robustness checks. The results remain consistent over all the model choices. However, I do not claim that there is no chance of endogeneity. Nevertheless, I believe these endogeneity treatments should capture a significant

part of it. However, a full treatment requires a carefully designed randomized field experiment. After accounting for selection, I retrieve the same results as before. It suggests that my findings are robust to the the endogeneity issue. I also point out that the endogeneity in sponsored contents is not as severe as in online advertising. In online advertising, especially in social media platforms such as Facebook, ads are targeted by individual preferences and characteristics. In sponsored contents, however, neither the brands nor the platform do not have direct control on who watches the video. This is due to the unique attribute of sponsorship in social media.

If the sponsorship impact is due to user skepticism, annoyance, and trust to the sponsors, I may expect that with more presence of the sponsor, the effect to be intensified. There are possible mechanisms for such interaction. Some users might be concerned about the credibility of the contents in case of sponsorship. The concept of trust comes along with the credibility of the creator-sponsor. Also, some may be simply annoyed by the sponsor presence. Nevertheless, these two mechanisms can be interrelated. To break the sponsorship effect, I posit that Facebook users will be negatively affected by stronger presence of the sponsor. The presence of the sponsor is captured by featuring its name in either the title or description. The results exhibit no significant effect on featuring sponsor's name or product. Consequently, users' viewership and engagement are not affected by featuring sponsors in the description.

Finally, my goal is to find out if stronger association between creator and sponsor leads to higher viewership and engagement. I measured three different aspects of association. First, I used platform-defined categories as an index of similarity. Examples of these categories are Food, Entertainment, Sports, and etc. This relationship is defined at the industry-level measure to control for a static link which is constant through the partnership. Clearly, this is a hard defined construct as it can be zero or one and it may not capture the dynamic association between the creator and sponsor. I find that this measure does not explain the viewership and engagement in sponsored videos. Therefore, coming from the same category, creators and sponsors experience no synergy effect. Second, I defined content similarity as

another index to capture the role of heterogeneous topics posted by creator. In this regard, similarity accounts for different variation of videos within the same creator. Lastly, I introduced third aspect of a link between sponsor and creator coming from their target audience. That is, how much they share customers (or users). This notion gives us the ability to explore the dynamics of sponsorship with taking potential users into account. I find that both content similarity and audience overlap have positive impact on sponsored videos in terms of viewership and engagement. This suggests that *relevant* partnerships are more effective in attracting users' attention and engagement. There are multiple potential explanations for this phenomenon. First, *relevant* partnerships would lead to familiar sponsors to users. Thus, the users are not annoyed by the existence of the sponsor. Additionally, users are more likely to be interested in the sponsored products. Lastly, due to the relevance of the sponsorship, users might not even notice the presence of the sponsored content. However, the last one is not an intended consequence of the sponsorship. Nevertheless, I cannot disentangle these possible scenarios. Furthermore, the audience overlap helps the viewership when the sponsor is featured in the video. This is very interesting as it layouts a strategy to increase the performance of the sponsored contents when the sponsor is featured. I also performed the analyses at the creator category to measure the heterogeneous effect on similarity and overlap. I find that in categories *Food & Drinks* and *Sports*, the impact of similarity and overlap is stronger. Creators in these categories can benefit more by strategically choosing more *relevant* sponsors.

As a robustness check, I focused on an alternative sampling approach. I constructed the relevance for the pair of creator and sponsor by the closeness of their audience. Thus, I restricted the sample to the videos with sponsors from auto industry for two reasons. The auto industry is the leading spender in sponsorship in the overall market in the US (both offline and online).[14] Moreover, by just having car manufacturers as sponsors, I can get the relevance data for a limited number of unique sponsors. The results remain consistent with

[14]IEG, ESP Properties, What Sponsors Want & Where Dollars Will Go in 2018

my main findings.

5.6 Conclusion

As more users spend more time on social network platforms compared to more conventional media such as TV, the online contents generated are growing enormously. Content creators try to make viral contents leveraging network aspects of online media consumption lifestyle. Those creators who are successful at gaining users' attention can monetize such advantage. However, this is not only limited to individual content creators. Social media has been the most dominant medium in the sponsorship market.

I utilize a rich data set on sponsored videos on Facebook to answer the questions on effectiveness of sponsorship on branded contents. More specifically, my goal is to study whether stronger association between creator and sponsor leads to higher viewership and user engagement. I show that sponsored videos receive fewer views and less engagement rate compared to non-sponsored videos.

To address the effect of creator-sponsor relationship on the sponsorship performance, I test the moderation effect of relevance on views. I conclude that the category similarity plays no role in video exposure and user engagement. However, creator and sponsor audience and content relevance have a positive impact on viewership and engagement. I also find that audience overlap has a positive interaction effect with the sponsor's presence. I studied heterogeneous effect of sponsorship in creator categories. Categories like *Food & Drink* and *Sports* have stronger effect for content similarity and audience overlap.

The contribution of the paper to the literature is twofold. First, it clarifies the effect of online sponsorship on social media. More specifically, it answers to an unanswered question in video sponsorship: Does similarity of creator and sponsor helps the viewership and engagement? Second, it suggests a new approach to make partnerships more effective. It highlights the role of targeting by the creator's audience. The latter, in particular, has significant implications for practitioners.

5.6.1 Managerial Implications

With the ubiquity of UGC platforms such as Facebook, YouTube, and Instagram, sponsorship has grown significantly. There are multiple aspects of partnership in sponsored contents that can affect the performance. I focus on the association between creator and sponsor. That is, the question is whether creators (and sponsors) should find more *relevant* (or similar) partner. The findings of my paper shows how creators and sponsors can leverage new measures and tools in online advertising to improve the performance of the sponsored videos. Two important aspects of this association is content similarity and audience overlap. Thanks to online platforms, both sides have access to information on all of these aspects. They can gauge their selection by following more similar partners. Especially, in *Food & Drink* and *Sports* categories, similarities are more crucial to the partnership.

5.6.2 Policy Implications

In addition to managerial implications, my research provides insights for the policy makers. As the regulation bodies (e.g. FCC and FTC in the US) start to impose stricter guidelines for sponsored contents on social media, it is important to know the effect of sponsorship. My research provides an answer to this question by quantifying sponsorship effect in terms of user response and reaction to the contents.

5.6.3 Limitations

I acknowledge that I do not study the return on investment for the video sponsorship. I addressed content exposure and user engagement as measures that are important for both sponsors and creators. Viewership has a direct effect on the final effectiveness of the advertising and sometimes determines the cost of sponsorship which is negotiated by the creator and sponsor. In addition, as I do not study the long-term effect of sponsored contents, it would be an interesting direction for future studies. As a result, one can incorporate the recall and return of customers as a response to sponsored contents.

I also note that my identification strategy may not entirely capture the endogeneity resulting from selection. This task has been an important challenge in the literature of online advertising and product placement. I took several steps in order to minimize the endogeneity problem. First, I used the fixed effect approach for creators. Fixed effect model resolve the endogeneity for the time-invariant sources. I also introduced a control function approach with using video descriptions and video specifics to identify the selection problem. Thus, the model implies that differences in the nature of sponsored and non-sponsored videos are reflected through the description. While this assumption might not be always true, I argue that this can be an appropriate one as creators often lay out video types and nature in the descriptions. I also discuss about the direction of any biases that are not addressed. As creators often try to monetize the viewership of their contents, it is reasonable to assume that those contents that are more likely to attract more views, are also more likely to be sponsored. In this case, the bias would be downward toward zero and that makes my findings conservative. It is clear that for those creators/sponsors who follow different strategies the bias might be upward.

Chapter 6

CONCLUDING REMARKS AND FUTURE RESEARCH DIRECTIONS

With the ubiquity of online platforms, it is increasingly important to understand the demand structure and latent mechanism that have economic consequences. In this dissertation, I explored different aspects of online platforms. First, I studied the demand from the online crowd and the proper response from the creators. Second, I investigated the concept of "choice in online shopping. The notion of choice can be translated to product diversity and fairness in giving chance to more fringe sellers and goods. That is a key driver in increasing customer welfare in online retailing. Lastly, I explored the monetization of the online contents which has been a steadily growing industry. I selected three different empirical settings which are suitable for the research questions laid out earlier.

Mobile applications are becoming the mainstream channel in retail and digital marketing. I study the distribution of product sales on online channels and compare the effects of the long tail in PC versus mobile Apps. My findings show that the distribution of product sales in App channel is less concentrated than the one in PC. This provides an important insight for retailers: introducing mobile App channel increases the variety of the products. This study also provides insights into the mechanism through which the longer tail happens: More searching compensates the high search costs and increases the value for the customers by providing more information about product details. my findings suggest that it is crucial for businesses to invest in technologies which make it easier for customers to gain more information about products and services.

In the analyses of online crowdfunding platforms, I find that the effect of middle and high-level rewards can be significant. I finally calculate backers' welfare in the online crowdfunding

platform per category and find that projects in wearable, web and gadget categories result in more surplus for customers (backers). I also show the substitution patterns through reward levels. The insights have implications for entrepreneurs in designing the reward scheme and pricing. I find that, on average, low-level rewards do not contribute much to the customer surplus. Additionally, they cannibalize the rewards with higher prices. Entrepreneurs can benefit from these results in terms of designing the reward scheme and pricing.

As more users spend more time on social network platforms compared to more conventional media such as TV, the online contents generated are growing enormously. Content creators try to make viral contents leveraging network aspects of online media consumption lifestyle. Those creators who are successful at gaining users' attention can monetize such advantage. However, this is not only limited to individual content creators. Social media has been the most dominant medium in the sponsorship market. My goal is to study whether stronger association between creator and sponsor leads to higher viewership and user engagement. I show that sponsored videos receive fewer views and less engagement rate compared to non-sponsored videos. I conclude that the category similarity plays no role in video exposure and user engagement. However, creator and sponsor audience and content relevance have a positive impact on viewership and engagement.

My intentions of this dissertation are twofold. First, I study the important aspects of design and demand in online platforms. Given the growing market of online platforms, businesses face more design questions regarding the effective strategies. These strategies help us identify the intended results. Second, I contribute to the literature by adopting novel approaches in modeling and using natural language processing techniques to devise an innovative methodology to capture the unique features of the fast-paced markets.

6.1 Main Lessons and Future Directions

The main lesson from this dissertation is that studies of economics of online platforms can bring us more managerial insights that are risen from data and rigorous analyses. The markets are constantly changing through technological innovations and scholars need to ad-

dress new questions that were not even defined previously. For example, any of the mobile economy, online crowdfunding, or social media monetizing is at most a decade long phenomenon. That highlights the importance of such studies in understanding the nature of markets, demand structures, and customer decision making processes. The other key lesson of the dissertation is that in the new era of online markets, small players such as small businesses, entrepreneurs, and individual content creators play much more significant roles in the entire industry. That gives leverage to small businesses, leading to more competition and innovation which are crucial for customer welfare.

For future directions, I am very looking forward to new opportunities brought by big data and advanced data-driven approaches. More specifically, the overlap of AI and causal inferences will be very interesting to explore. In this day and age, vast methodological approaches are common in driving business insights. Many of them are inspired by the success of machine learning and deep learning techniques. Although, recent achievements of deep learning are impressive, scholars, especially econometricians, should adopt these approaches and establish causal inferences.

The other aspect that I am interested is user's privacy and the monopolies by tech companies over customer data. The real case of companies such as "Google", "Facebook", and "Apple" makes one thinking about how privacy of the users will play a role in the market of technology products. First question is how users will response after realization of privacy breaches. In other words, *"can there be alternative markets in which user can monetize their own data?"* This is especially important when regulatory bodies lag behind the technological advances and business practices that leverage users' data. This is not limited to the United States. Even in Europe, which is the most user privacy advocating entity, there is not clear regulations on many non-trivial data usage or monopolies of data. The other question is whether the monopolistic markets resulted by data leverage would lead to less innovations, welfare, and efficient markets. For example, Google offer wide variety of services to online services and websites. Google utilize the enormous user data at their disposal to improve those services. However, it is not clear if there would be no monopoly over data, Google could

keep up with other competitors. This opens up interesting areas for scholar in information systems and digital economics.

BIBLIOGRAPHY

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.

Agrawal, A., Catalini, C., and Goldfarb, A. (2011). The geography of crowdfunding (= national bureau of economic research working paper series nr. 16820). *Cambridge, MA*.

Agrawal, A., Catalini, C., and Goldfarb, A. (2014). Some simple economics of crowdfunding. *Innovation Policy and the Economy*, 14(1):63–97.

Agrawal, A., Catalini, C., and Goldfarb, A. (2015). Crowdfunding: Geography, social networks, and the timing of investment decisions. *Journal of Economics & Management Strategy*, 24(2):253–274.

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. Hachette Books.

Athey, S. and Ellison, G. (2011). Position auctions with consumer search. *The Quarterly Journal of Economics*, 126(3):1213–1270.

Baek, T. H. and Morimoto, M. (2012). Stay away from me. *Journal of advertising*, 41(1):59–76.

Bajari, P. and Benkard, C. L. (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of political economy*, 113(6):1239–1276.

Bajari, P., Nekipelov, D., Ryan, S. P., and Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, 105(5):481–85.

Balachander, S. and Srinivasan, K. (1994). Selection of product line qualities and prices to signal competitive advantage. *Management Science*, 40(7):824–841.

Baye, M. R., Gatti, J. R. J., Kattuman, P., and Morgan, J. (2009). Clicks, discontinuities, and firm demand online. *Journal of Economics & Management Strategy*, 18(4):935–975.

Baye, M. R. and Morgan, J. (2001). Information gatekeepers on the internet and the competitiveness of homogeneous product markets. *American Economic Review*, 91(3):454–474.

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

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262.

Besanko, D., Gupta, S., and Jain, D. (1998). Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Science*, 44(11-part-1):1533–1547.

Bhargava, H. K. and Choudhary, V. (2001). Information goods and vertical differentiation. *Journal of Management Information Systems*, 18(2):89–106.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. (2011). Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*, pages 47–56. ACM.

Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc.

Bruno, H. A. and Vilcassim, N. J. (2008). Research notestructural demand estimation with varying product availability. *Marketing Science*, 27(6):1126–1131.

Brynjolfsson, E., Hu, Y., and Simester, D. (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, 57(8):1373–1386.

Brynjolfsson, E., Hu, Y., and Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49(11):1580–1596.

Brynjolfsson, E., Hu, Y., and Smith, M. D. (2010a). Research commentarylong tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research*, 21(4):736–747.

Brynjolfsson, E., Hu, Y. J., and Smith, M. D. (2010b). The longer tail: The changing shape of amazons sales distribution curve. *Available at SSRN 1679991*.

Burtch, G., Ghose, A., and Wattal, S. (2013). An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Information Systems Research*, 24(3):499–519.

Burtch, G., Ghose, A., and Wattal, S. (2015). The hidden cost of accommodating crowdfunder privacy preferences: a randomized field experiment. *Management Science*, 61(5):949–962.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Carpenter, R. (1977). Matching when covariables are normally distributed. *Biometrika*, 64(2):299–307.

Choi, C. J. and Shin, H. S. (1992). A comment on a model of vertical product differentiation. *The Journal of Industrial Economics*, pages 229–231.

Claussen, J., Kretschmer, T., and Mayrhofer, P. (2013). The effects of rewarding user engagement: the case of facebook apps. *Information Systems Research*, 24(1):186–200.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Cremer, H. and Thisse, J.-F. (1991). Location models of horizontal differentiation: a special case of vertical differentiation models. *The Journal of Industrial Economics*, pages 383–390.

Danaher, P. J. (1995). What happens to television ratings during commercial breaks? *Journal of Advertising Research*, 35(1):37–37.

DÁstous, A. and Bitz, P. (1995). Consumer evaluations of sponsorship programmes. *European Journal of Marketing*, 29(12):6–22.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Desai, P. S. (2001). Quality segmentation in spatial markets: When does cannibalization affect product line design? *Marketing Science*, 20(3):265–283.

Drèze, X. and Hussherr, F.-X. (2003). Internet advertising: Is anybody watching? *Journal of interactive marketing*, 17(4):8–23.

Dubé, J.-P., Fox, J. T., and Su, C.-L. (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267.

Dutta, P. K., Lach, S., and Rustichini, A. (1995). Better late than early: Vertical differentiation in the adoption of a new technology. *Journal of Economics & Management Strategy*, 4(4):563–589.

Eadicicco, L. (2015). Americans check their phones 8 billion times a day. `http://time.com/4147614/smartphone-usage-us-2015/`. [Online; accessed April-01-2019].

Ebbes, P., Wedel, M., and Böckenholt, U. (2009). Frugal iv alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics*, 24(3):446–468.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259.

Fang, Z., Gu, B., Luo, X., and Xu, Y. (2015). Contemporaneous and delayed sales impact of location-based mobile promotions. *Information Systems Research*, 26(3):552–564.

Feng, J., Bhargava, H. K., and Pennock, D. M. (2007). Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms. *INFORMS Journal on Computing*, 19(1):137–148.

Foreh, M. R. and Grier, S. (2003). When is honesty the best policy? the effect of stated company intent on consumer skepticism. *Journal of consumer psychology*, 13(3):349–356.

Fossen, B. L. and Schweidel, D. A. (2016). Television advertising and online word-of-mouth: An empirical investigation of social tv activity. *Marketing Science*, 36(1):105–123.

Freedman, S. M. and Jin, G. Z. (2011). Learning by doing with asymmetric information: evidence from prosper.com.

Gallagher, K., Foster, K. D., and Parsons, J. (2001). The medium is not the message: Advertising effectiveness and content evaluation in print and on the web. *Journal of Advertising Research*, 41(4):57–70.

Geerardyn, A., Fauconnier, G., and Pattyn, B. (2000). Hybrid forms in marketing communication: the increasingly fuzzy boundaries between information and commerce. *Media Ethics: Opening Social Dialogue, Peeters Publishers, Leuven, Belgium*, pages 329–355.

Ghose, A., Goldfarb, A., and Han, S. P. (2012). How is the mobile internet different? search costs and local activities. *Information Systems Research*, 24(3):613–631.

Ghose, A. and Han, S. P. (2011). An empirical analysis of user content generation and usage behavior on the mobile internet. *Management Science*, 57(9):1671–1691.

Ghose, A. and Han, S. P. (2014). Estimating demand for mobile applications in the new economy. *Management Science*, 60(6):1470–1488.

Goh, K.-Y., Heng, C.-S., and Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, 24(1):88–107.

Goldenberg, J., Oestreicher-Singer, G., and Reichman, S. (2012). The quest for content: How user-generated links can facilitate online exploration. *Journal of Marketing Research*, 49(4):452–468.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*.

Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. (1987). Generalised residuals. *Journal of econometrics*, 34(1-2):5–32.

Gwinner, K. P. and Eaton, J. (1999). Building brand image through event sponsorship: The role of image transfer. *Journal of advertising*, 28(4):47–57.

Han, S. P., Park, S., and Oh, W. (2016). Mobile app analytics: A multiple discrete-continuous choice framework. *MIS Quarterly*, 40(4):983–1008.

Hartjen, R. (2016). Retails main event: Brick & mortar vs. online. `http://retailnext.net/en/blog/brick-and-mortar-vs-online-retail/`. [Online; accessed April-01-2019].

Hartmann, W. R. (2010). Demand estimation with social interactions and the implications for targeted marketing. *Marketing science*, 29(4):585–601.

Hastings, G. B. (1984). Sponsorship works differently from advertising. *International Journal of Advertising*, 3(2):171–176.

Heckman, J. J. (1977a). Dummy endogenous variables in a simultaneous equation system.

Heckman, J. J. (1977b). Sample selection bias as a specification error (with an application to the estimation of labor supply functions).

Heckman, J. J. and Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, 30(1-2):239–267.

Hoehle, H. and Venkatesh, V. (2015). Mobile application usability: Conceptualization and instrument development. *MIS Quarterly*, 39(2):435–472.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Hu, M., Li, X., and Shi, M. (2015). Product and pricing decisions in crowdfunding. *Marketing Science*, 34(3):331–345.

Ichino, A., Mealli, F., and Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of applied econometrics*, 23(3):305–327.

Ilfeld, J. S. and Winer, R. S. (2002). Generating website traffic. *Journal of Advertising Research*, 42(5):49–61.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Jenkins, J. L., Anderson, B. B., Vance, A., Kirwan, C. B., and Eargle, D. (2016). More harm than good? how messages that interrupt can make us vulnerable. *Information Systems Research*, 27(4):880–896.

Kalofolias, V., Bresson, X., Bronstein, M., and Vandergheynst, P. (2014). Matrix completion on graphs. *arXiv preprint arXiv:1408.1717*.

Kelly, L., Kerr, G., and Drennan, J. (2010). Avoidance of advertising in social networking sites: The teenage perspective. *Journal of interactive advertising*, 10(2):16–27.

Kelly, S., Coote, L., Cornwell, T. B., and McAlister, A. (2017). Mellowing skeptical consumers: An examination of sponsorship-linked advertising. *International Journal of Sport Communication*, 10(1):58–84.

Khalaf, S. (2015). Flurrymobile. `http://flurrymobile.tumblr.com/post/127638842745/seven-years-into-the-mobile-revolution-content-is`. [Online; accessed April-01-2019].

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching. *Copy at http://j. mp/1sexgVw Download Citation BibTex Tagged XML Download Paper*, 378.

Kolsarici, C. and Vakratsas, D. (2010). Category-versus brand-level advertising messages in a highly regulated environment. *Journal of marketing research*, 47(6):1078–1089.

Kumar, H. (2016). Mobile commerce trends to buy into.

Lancaster, K. (1971). *Consumer demand: A new approach*. Columbia University Press.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, pages 43–58. Springer.

Lin, H.-H. and Wang, Y.-S. (2006). An examination of the determinants of customer loyalty in mobile commerce contexts. *Information & management*, 43(3):271–282.

Liu, D., Chen, J., and Whinston, A. B. (2010). Ex ante information and the design of keyword auctions. *Information Systems Research*, 21(1):133–153.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

MacKenzie, S. B., Lutz, R. J., and Belch, G. E. (1986). The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations. *Journal of marketing research*, pages 130–143.

Manchanda, P., Dubé, J.-P., Goh, K. Y., and Chintagunta, P. K. (2006). The effect of banner advertising on internet purchasing. *Journal of Marketing Research*, 43(1):98–108.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record*, (673).

McFadden, D. (1981). Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

McFadden, D., Talvitie, A., Cosslett, S., Hasan, I., Johnson, M., Reid, F., and Train, K. (1977). Demand model estimation and validation. *Urban Travel Demand Forecasting Project, Phase*, 1.

Meenaghan, T. (1991). The role of sponsorship in the marketing communications mix. *International journal of advertising*, 10(1):35–47.

Mela, C. F., Jedidi, K., and Bowman, D. (1998). The long-term impact of promotions on consumer stockpiling behavior. *Journal of Marketing research*, pages 250–262.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, A. R. and Tucker, C. (2013). Active social media management: the case of health care. *Information Systems Research*, 24(1):52–70.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, 29(1):1–16.

Montgomery, A. L., Hosanagar, K., Krishnan, R., and Clay, K. B. (2004). Designing a better shopbot. *Management Science*, 50(2):189–206.

Moorthy, K. S. (1984). Market segmentation, self-selection, and product line design. *Marketing Science*, 3(4):288–307.

Nannicini, T. et al. (2007). Simulation-based sensitivity analysis for matching estimators. *Stata Journal*, 7(3):334.

Nelson, P. (1970). Information and consumer behavior. *Journal of political economy*, 78(2):311–329.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy*, 9(4):513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69:307–342.

Nevo, A. (2003). New products, quality changes, and welfare measures computed from estimated demand systems. *Review of Economics and statistics*, 85(2):266–275.

Nevo, A. (2011). Empirical models of consumer behavior. *Annu. Rev. Econ.*, 3(1):51–75.

Nielsen (2018). Cutting-edge content from digital publishers keeps millenials coming back for more. `www.nielsen.com/us/en/insights/news/2018/cutting-edge-content-from-digital-publishers-keeps-millennials-coming-back-for-more.html`. [Online; posted January-01-2018].

NielsenSports (2018). Nba teams score a slam dunk with social media. `https://nielsensports.com/nba-teams-score-slam-dunk-social-media/`. [Online; accessed April-01-2019].

Obermiller, C., Spangenberg, E., and MacLachlan, D. L. (2005). Ad skepticism: The consequences of disbelief. *Journal of advertising*, 34(3):7–17.

Obermiller, C. and Spangenberg, E. R. (1998). Development of a scale to measure consumer skepticism toward advertising. *Journal of consumer psychology*, 7(2):159–186.

O'Hara, K., Mitchell, A. S., and Vorbau, A. (2007). Consuming video on mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 857–866. ACM.

Otker, T. and Hayes, P. (1987). Judging the efficiency of sponsorship, experiences from the 1986 soccer world cup. *European Research*, 15(4):53–58.

Park, S. and Gupta, S. (2009). Simulated maximum likelihood estimator for the random coefficient logit model using aggregate data. *Journal of Marketing Research*, 46(4):531–542.

Park, S. and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4):567–586.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy*, 110(4):705–729.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1):3–13.

Porter, C. E. and Donthu, N. (2008). Cultivating trust and harvesting value in virtual communities. *Management Science*, 54(1):113–128.

Prasad, V. K. (1976). Communications-effectiveness of comparative advertising: A laboratory analysis. *Journal of Marketing Research*, pages 128–137.

Rajaretnam, A. (1994). The long-term effects of sponsorship on corporate and product image: Findings of a unique experiment. *Marketing and Research Today*, 22(1):62–62.

Rishika, R., Kumar, A., Janakiraman, R., and Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. *Information systems research*, 24(1):108–127.

Rodgers, S. (2003). The effects of sponsor relevance on consumer reactions to internet sponsorships. *Journal of Advertising*, 32(4):67–76.

Rodgers, S. and Thorson, E. (2000). The interactive advertising model: How users perceive and process online ads. *Journal of interactive advertising*, 1(1):41–60.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Rosenbaum, P. R. (2002). Observational studies. In *Observational studies*, pages 1–17. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosner, J. (2017). Future of payments? phone will set the tone. `https://www.mobilepaymentstoday.com/articles/future-of-payments-phone-will-set-the-tone/`. [Online; accessed April-01-2019].

Roto, V. and Oulasvirta, A. (2005). Need for non-visual feedback with long response times in mobile hci. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 775–781. ACM.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, ii: Maximums on bias reduction for fixed sample sizes. *Biometrics*, pages 121–132.

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, pages 293–298.

Schweidel, D. A., Foutz, N. Z., and Tanner, R. J. (2014). Synergy or interference: the effect of product placement on commercial break audience decline. *Marketing Science*, 33(6):763–780.

Schweidel, D. A. and Kent, R. J. (2010). Predictors of the gap between program and commercial audiences: An investigation using live tuning data. *Journal of Marketing*, 74(3):18–33.

Schweidel, D. A. and Kent, R. J. (2011). Introducing the ad ecg: How the set-top box tracks. *J. Advertising Res*, 51(4):586–593.

Shim, J. P., Park, S., and Shim, J. M. (2008). Mobile tv phone: current usage, issues, and strategic implications. *Industrial Management & Data Systems*, 108(9):1269–1282.

Shriver, S. and Bollinger, B. (2015). Structural analysis of multi-channel demand. *Columbia Business School Research Paper*, 1(15-50).

Siddarth, S. and Chattopadhyay, A. (1998). To zap or not to zap: A study of the determinants of channel switching during commercials. *Marketing Science*, 17(2):124–138.

Siwicki, B. (2014). Hold the phone: 66% of time spent with e-retail is on mobile. `https://www.digitalcommerce360.com/2014/11/20/hold-phone-66-time-spent-e-retail-mobile/`. [Online; accessed April-01-2019].

Small, K. A. and Rosen, H. S. (1981). Applied welfare economics with discrete choice models. *Econometrica (pre-1986)*, 49(1):105.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Stipp, H. (1998). The impact of olympic sponsorship on corporate image. *International Journal of advertising*, 17(1):75–87.

Stipp, H. and Schiavone, N. P. (1996). Modeling the impact of olympic sponsorship on corporate image. *Journal of Advertising Research*, 36(4):22–28.

Susarla, A., Oh, J.-H., and Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, 23(1):23–41.

Sutton, J. (1986). Vertical product differentiation: some basic themes. *The American Economic Review*, 76(2):393–398.

Teh, Y. W., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360.

Train, K. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Trajtenberg, M. (1989). The welfare analysis of product innovations, with an application to computed tomography scanners. *Journal of political Economy*, 97(2):444–479.

Trusov, M., Bodapati, A. V., and Bucklin, R. E. (2010). Determining influential users in internet social networks. *Journal of Marketing Research*, 47(4):643–658.

Vandenbosch, M. B. and Weinberg, C. B. (1995). Product and price competition in a two-dimensional vertical differentiation model. *Marketing Science*, 14(2):224–249.

Varian, H. R. (2007). Position auctions. *international Journal of industrial Organization*, 25(6):1163–1178.

Vella, F. and Verbeek, M. (1999). Estimating and interpreting models with endogenous treatment effects. *Journal of Business & Economic Statistics*, 17(4):473–478.

Volner, I. D. and Sheridan, K. K. (2018). Fcc revives its own native advertising rule: Sponsorship identification. `http://www.surfline.com/surf-news/maldives-surf-access-controversy-update_75296/`. [Online; posted January-04-2018].

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wu, J.-H. and Wang, S.-C. (2005). What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & management*, 42(5):719–729.

Xiao, B. and Benbasat, I. (2011). Product-related deception in e-commerce: a theoretical perspective. *Mis Quarterly*, 35(1):169–196.

Xu, K., Chan, J., Ghose, A., and Han, S. P. (2016). Battle of the channels: The impact of tablets on digital commerce. *Management Science*, 63(5):1469–1492.

Yang, S. and Ghose, A. (2010). Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science*, 29(4):602–623.

Yoganarasimhan, H. (2012). Impact of social network structure on content propagation: A study using youtube data. *Quantitative Marketing and Economics*, 10(1):111–150.

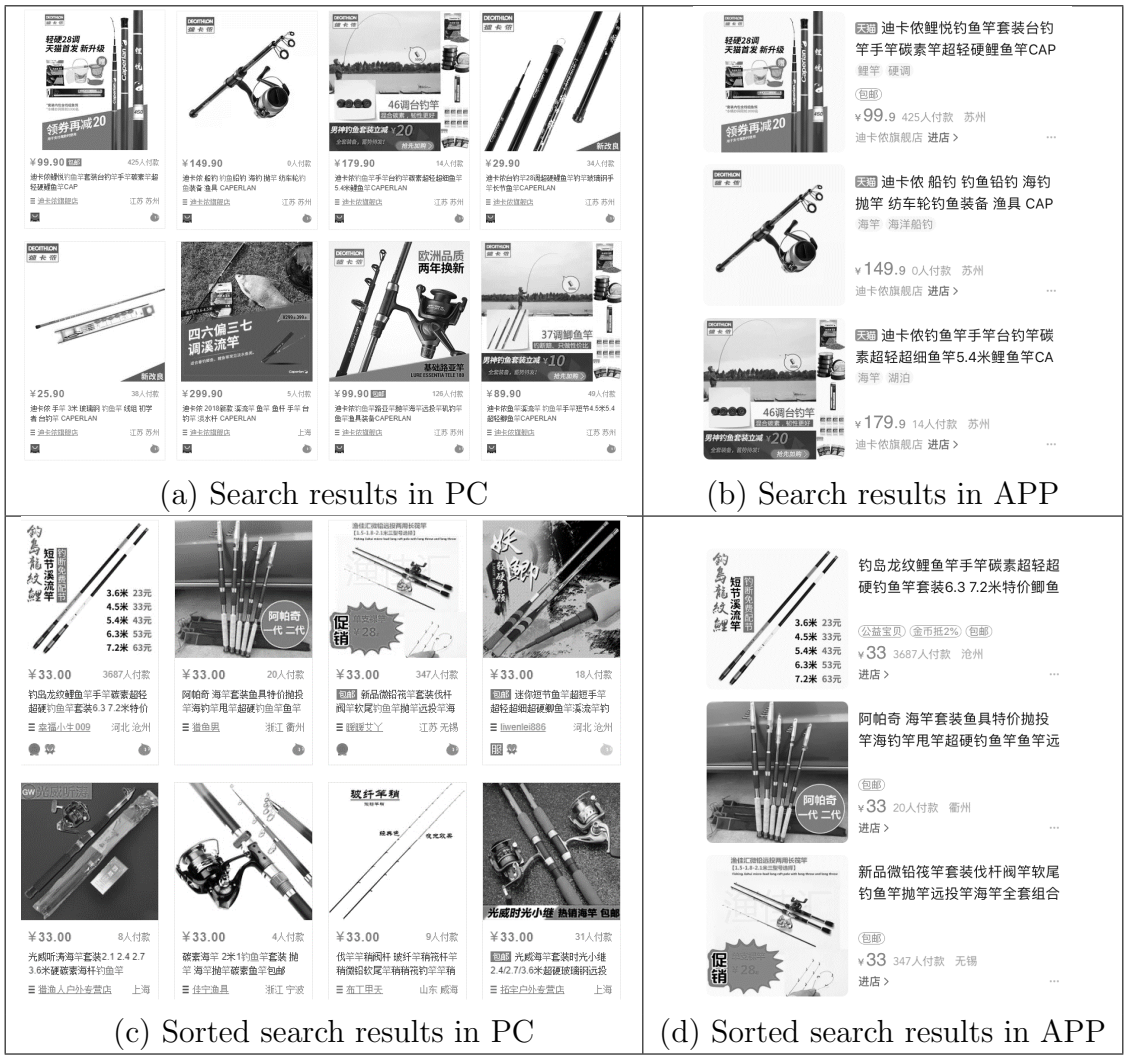
Zhang, J. and Liu, P. (2012). Rational herding in microloan markets. *Management science*, 58(5):892–912.

Zhang, K. and Katona, Z. (2012). Contextual advertising. *Marketing Science*, 31(6):980–994.

Appendix A

PRODUCT RANKINGS AND RECOMMENDATION SYSTEM IN APP AND PC

Figure A.1 shows and compares screenshots from sample product searches in App and PC. It suggests both organic and sorted searches result in the same rankings in both channels. However, the product searched lists are shown vertically in App due to screen size restriction.



(a) Search results in PC

(b) Search results in APP

(c) Sorted search results in PC

(d) Sorted search results in APP

Figure A.1: Sample rankings of search results in PC and App

Appendix B

LONG TAIL ROBUSTNESS CHECKS

I perform multiple analyses as robustness checks of my finding for the first research question.

B.1 Second Batch Analysis

As discussed in the paper, I estimate the long tail effect on the second batch of data collected in October 2014. Only a small percentage of buyers are shared among the two datasets. This feature adds to the generalizability of the findings. Table B.1 summarizes the main regression models for the second data.

B.2 Another Product Category: Chargers and Power Supplies

Table B.2 presents the results for the charger and power supply category. Overall, the results show a greater effect of long tail in App (compared to fishing rods).

Table B.1: Results of the long tail models for the second batch

	(1) PC	(2) App	(3) Pooled	(4) Quantile
$\log rank$	-1.640***	-1.530***	-1.640***	-1.730***
	(0.006)	(0.007)	(0.006)	(0.006)
App			-1.293***	-1.804***
			(0.077)	(0.069)
$App \times \log rank$			0.109***	0.171***
			(0.009)	(0.008)
Constant	19.83***	18.54***	19.83***	20.80***
	(0.053)	(0.055)	(0.053)	(0.047)
R^2	0.87	0.86	0.87	

Standard errors in parentheses. DV is $\log sales$.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.2: Results of the long tail models for charger category

	(1) PC	(2) App	(3) Pooled	(4) Quantile
$\log rank$	-1.285***	-0.978***	-1.285***	-1.458***
	(0.006)	(0.002)	(0.005)	(0.004)
App			-2.575***	-3.963***
			(0.044)	(0.034)
$App \times \log rank$			0.307***	0.484***
			(0.006)	(0.005)
Constant	14.66***	12.08***	14.66***	16.07***
	(0.047)	(0.019)	(0.034)	(0.027)
R^2	0.91	0.96	0.94	

Standard errors in parentheses. DV is $\log sales$.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix C

PRODUCT CHARACTERISTICS

As the platform runs a C2C business model, there are a larger number of differentiated products available on the website. These products may vary in unobservable way to us but known to customers. Most of the product attributes are described in the product descriptions in an unstructured format. To capture these features and address the product differentiation, I use *bag-of-words* of *n-gram* models in natural language processing (NLP) to extract the relevant information by frequency of word occurrences. I dropped the words with no added information like stop words and identifiers. I identified 18 features relevant to the fishing pole category. They describe length, material, finish and quality, origin, function, and additional features. They are formatted as dummy variables. Table C.1 describes descriptive statistics for the extracted features of products.

Table C.1: Summary statistics for the extracted product features by NLP method

Variable	Definition	Mean	St. Dev.
m2	Length (in meters)	0.27	(0.44)
m3	Length (in meters)	0.42	(0.49)
m4	Length (in meters)	0.52	(0.5)
m5	Length (in meters)	0.54	(0.5)
m6	Length (in meters)	0.42	(0.49)
m7	Length (in meters)	0.19	(0.4)
carbon	Material	0.72	(0.45)
bamboo	Material	0.01	(0.1)
fiberglass	Material	0.01	(0.08)
fine	Type	0.18	(0.38)
hard	Type	0.54	(0.5)
ultralight	Type	0.54	(0.5)
fishingkit	Accessories	0.66	(0.47)
desktop	Accessories	0.36	(0.48)
japan	Quality	0.04	(0.19)
sea	Feature	0.21	(0.41)
river	Feature	0.01	(0.09)
adjustable	Feature	0.03	(0.16)
Observations	835,919		

Appendix D

MATCHING

I now discuss the details of the matching methods I used in the paper. I use Mahalanobis distant matching (MDM) and propensity score matching (PSM) for the main analyses. I verify the validity and performance of the matching techniques by conducting several statistical tests and visual measures. I control for the overlap condition which is required for the matching algorithms. I also measure the balance for the observables in treatment and control groups after matching. These two matching algorithms apply two different paradigms to reach balanced samples. In MDM, I follow an exact match procedure; i.e., the distance between treatment and control groups are used to find the matched sample. But in PSM, an indirect match is pursued; i.e., PSM matches a low dimension representation of observables (aka propensity score).

Matching algorithms rely on the so-called Conditional Independence Assumption (CIA) also known as *unconfoundedness* (Nannicini et al., 2007). That is, given observable characteristics, potential outcome and treatment assignment should be independent. In other words, the selection into treatment group is only driven by the factors that are observable. I try to reduce the confoundedness as much as possible by choosing various indicators of buyers and their past usage behavior. As there is no direct way to test this assumption, there are a few ways to check how the results are sensitive to the potential confounders. These methods are applied to measure the sensitivity of average treatment effect (ATE) to a bias. However, I do not directly measure ATE by the matching process. Instead, I build an aggregate model by the matched sample. Thus, the sensitivity analysis may not be as straightforward as a typical ATE problem. Therefore, I define a new model to measure the sensitivity of the results to the confounding factors. Note that the results of the sensitivity

analysis might not be directly translated to my main results. However, it shows the level of sensitivity of the matched sample.

I define a model in which the dependent variable is a binary variable depicting a niche product if 1 and a top product otherwise. The treatment is defined as using the App while the untreated group uses PC. I analyze the ATE of App on niche by matched sample and measure the sensitivity of the results.

I follow the methods proposed by Rosenbaum (2002) and Ichino et al. (2008). First, I define the probability of niche product by a logit model as a function of observables, X_i, and unobservables, ξ_i. Let δ capture the effect of the unobserved component on the niche product selection. If δ is zero, then there is no *hidden bias* in the process and the niche product probability will be solely determined by the observables.

Following Rosenbaum (2002), I derive a measure to test the significance of δ. The odds ratio between buyer i using App and buyer j using PC is $\exp\{\delta(\xi_i - \xi_j)\}$. Rosenbaum (2002) suggests using bounds as Equation D.1 to test how the inference changes as δ and $\xi_i - \xi_j$ vary:

$$1/e^{\delta} \leq oddsratioofvideo_{ij} \leq e^{\delta} \tag{D.1}$$

Let $\Gamma = e^{\delta}$, thus $\Gamma = 1$ ($\delta = 0$) implies there is no hidden bias. The interpretation of this would be if unobservables cause the odds ratio of niche product to change by a factor of Γ, what is the probability that the ATE still excludes zeros.

In an alternative approach, Ichino et al. (2008) introduce a simulation-based method in which there is confounder U which can change the treatment assignment and viewership. The probability distribution can be given manually or can be driven by another variable as a potential confounder. I introduce a binary variable *fishingkit* describing whether the product includes a supply kit. I acknowledge that the fishing pole product categories might be inherently different by the fact that they are either fishing rod or fishing supply. Thus, *fishingkit* variable can be used to capture that difference.

D.1 Mahalanobis Distant Matching

Mahalanobis distant matching (MDM) is an exact matching approach discussed by Cochran and Rubin (1973) and Rubin (1976). Under MDM, matched samples are chosen based on mahalanobis distance defined as:

$$\mathbb{X}_{MDM} = M(X|\sqrt{(X_i - X_j)S^{-1}(X_i - X_j)} < \delta) \tag{D.2}$$

where δ is called caliper, and adjusts how close the matches are, X is the original data, and S is the sample covariance matrix. MDM attempts to approximate fully blocked randomized experiment resulting in more efficient estimates (King and Nielsen, 2016). As a result of MDM, I achieve a balanced control and treatment group. Table D.1 reports the resulted balance after matching. The null hypothesis is that the mean variables are the same in both treatment and control group. I observe that all the attributes used for matching are balanced after the process.

Table D.1: Balance of attributes after matching

	Mean		t-test		
	Treated	Control	*t*	*p-value*	V(T)/V(C)
Buyer_female	0.235	0.235	0	1	1
Buyer_age	29.06	29.07	0.15	0.885	1
Buyer_domestic	0.989	0. 989	0	1	1
Buyer_rating	4.263	4.263	0	1	1
Log(Buyer_total)	4.749	4.749	0.02	0.981	1
Log(Buyer_pc_impressions)	1.663	1.663	0	0.999	1
Log(Buyer_app_impressions)	2.162	2.161	0.02	0.988	1
Buyer_impressions App/PC	0.552	0.552	0.02	0.988	1

Table D.2 presents the results for the sensitivity analysis for the hidden bias as discussed in previous section.

Table D.2: Sensitivity analysis of hidden bias

Γ	*p-value*
1	0.0003
1.1	< 0.0001
1.2	< 0.0001
1.3	< 0.0001
1.4	< 0.0001
1.5	< 0.0001
1.6	< 0.0001
1.7	< 0.0001
1.8	< 0.0001
1.9	< 0.0001
2.0	< 0.0001
2.1	< 0.0001
2.2	< 0.0001
2.3	< 0.0001
2.4	< 0.0001
2.5	< 0.0001

D.2 Propensity Score Matching

The steps for PSM is similar to MDM with a significant difference. I calculate propensity scores based on a logit model and use the scores to match. Table D.3 reports the logit regression results for the propensity score for *App*. I use nearest neighbor approach to match based on the scores. I allowed for multiple matches and used caliper=0.0001 for the tolerance of the matched scores.

I use the simulated-based approach (Ichino et al., 2008) to measure the sensitivity of the results to a potential confounder. I use *fishingkit* variable as discussed earlier. I use 100 simulations and bootstrapping for the standard errors. Table D.4 summarizes the results for the simulation of the confounder. The results suggest that the ATE is not influenced by the confounder (Selection effect on odds ratio is 1.087).

Figure D.1 shows the distribution of propensity score after matching. It represents a balanced distribution in control and treatment groups.

Table D.3: Sensitivity analysis of hidden bias

	Logistic Regression
Buyer_female	-0.044*
	(0.017)
Buyer_age	-0.016***
	(0.001)
Buyer_domestic	0.189***
	(0.044)
Buyer_rating	0.119***
	(0.016)
Log(Buyer_total)	-0.049
	(0.021)
Log(Buyer_pc_impressions)	-0.880***
	(0.010)
Log(Buyer_app_impressions)	0.807***
	(0.012)
Buyer_impressions app/pc	1.563***
	(0.049)
Constant	-0.854***
	(0.067)
Pseudo R^2	0.61
N	423,526

Standard errors in parentheses. DV is the binary variable *sponsored*.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table D.4: Simulation-based sensitivity analysis of hidden bias

	ATE	S.E.	Outcome Effect	Selection Effect
Before				
	0.022	0.062		
After				
	0.028	0.063	0.389	1.087

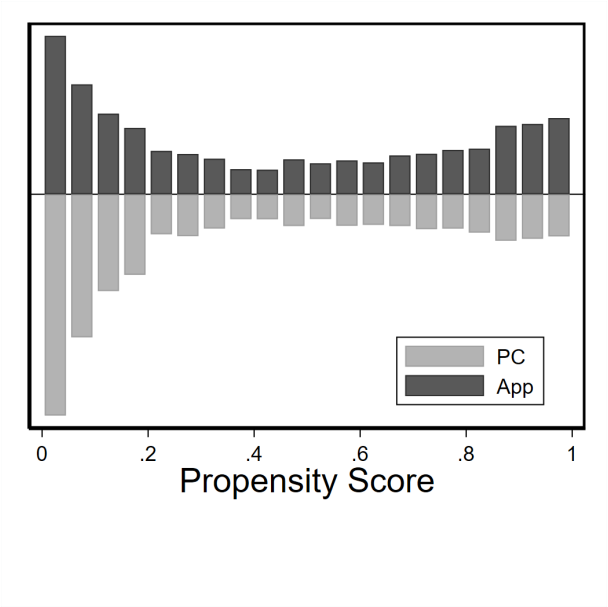


Figure D.1: Histogram of propensity score for treated and untreated groups.

Appendix E

CONTROL FUNCTION

I adopt the control function treatment proposed by (Heckman and Robb Jr, 1985). Let X_i be the set of observable characteristics. I have the following:

$$U_{ij}^{App} = X_i\beta^{App} + \epsilon_{ij}^{App}, \tag{E.1}$$

$$U_{ij}^{PC} = X_i\beta^{PC} + \epsilon_{ij}^{PC}, \tag{E.2}$$

where U_{ij}^{App} and U_{ij}^{PC} represent the utility of purchasing product j by customer i for App and PC respectively, and ϵ_{ij}^{App} and ϵ_{ij}^{PC} are zero mean error terms assumed to be independent of X_i. Let $App_{ij} = 1$ if the product is purchased through App, and $App_{ij} = 0$ otherwise. The observed sales are given by $AppU_{ij}^{App} + (1 - App)U_{ij}^{PC}$.

Since the treatment assignment (channel choice) is non-random, I consider the benefits acquired through the treatment (App), denoted by B_{ij}, in Equation E.3 following Vella and Verbeek (1999).

$$B_i = L_i\delta + \epsilon_{bi}, \tag{E.3}$$

where L_{ij} is a vector of exogenous variables, and δ is an unknown parameter vector. Assuming rational expectations, the decision to undergo the treatment can be written as following. I suppress the customer indexes for brevity.

$$App_i = I(U_j^{App} - U_j^{PC} + B_j > 0) = I(Z_j\pi + \epsilon_j > 0), \tag{E.4}$$

where $I(.)$ is an indicator function, $\epsilon_j = \epsilon_{ij}^{App} - \epsilon_{ij}^{PC}$, Z_j is a vector containing all elements found in X_j and L_j, and π is a vector of reduced-form parameters. In CF approach, I form

conditional expectation of U_j given App_j and Z_j (Heckman, 1977a,b).

$$\mathbb{E}\{U_j|Z_j, App_j\} = \alpha_{PC} + \alpha U_j + X_j\beta + \mathbb{E}\{\eta_j|Z_j, App_j\}. \tag{E.5}$$

The latter term in the above equation can be written as

$$\mathbb{E}\{\eta_j|Z_j, App_j\} = App_j\mathbb{E}\{\epsilon_j^{App}|Z_j, App_j = 1\} + (1 - App_j)\mathbb{E}\{\epsilon_j^{PC}|Z_j, App_j = 0\}. \tag{E.6}$$

Under the joint normality assumption, the two conditional expectations on the right side can be written as

$$\mathbb{E}\{\epsilon_j^k|Z_j, App_j\} = \sigma_\epsilon^k\lambda(Z_j\pi) \quad j = App, PC, \tag{E.7}$$

where

$$\lambda_i(Z_j\pi) = \mathbb{E}\{\epsilon_j|Z_j, App_j\} = (1 - App_j)\frac{-\phi(Z_j\pi)}{\Phi(-Z_j\pi)} + App_j\frac{\phi(Z_j\pi)}{1 - \Phi(-Z_j\pi)}, \tag{E.8}$$

is the generalized residual of the probit model (Gourieroux et al., 1987) . Note that this formulation is a *unrestricted* CF approach in which σ_ϵ^{App} are not required to equal σ_ϵ^{PC}. In addition, the CF approach does not require Z_j including at least one variable not found in X_i and is able to identify through the nonlinearity implied by joint normality. I expect α to be positive following my hypothesis about the effect of App on long tail.

Appendix F

SEARCH TIME ROBUSTNESS CHECKS

F.1 Niche Product Threshold

In my analysis for the effect of search time on the likelihood of niche product, I defined the niche products as the bottom half (median) based on the aggregate sales. I use different cutoffs for niche product definition. I use 60/40 and 70/30 splits. The results are consistent with the main findings. Table F.1 shows the estimation results for the 60/40 split. That is, niche products are the bottom 40% of products by aggregate sales.

F.2 GMM by Matched Sample

As another robustness check, I apply the GMM model in Equation 3.8 to a matched sample for treated and untreated groups. I use the same matched results discussed in Section 3.2 by MDM. Table F.2 reports the results of the channel effect with the matched sample. The coefficient of interest is the interaction term App and Log(Impression). The coefficient is 0.282 and statistically significant. This is consistent with the main result suggesting the finding is robust to the matching as well.

Table F.1: Estimates of the channel effect by 60/40 split

	GMM-IV
$\log impression$	0.109***
	(6.02)
App	-0.496***
	(-7.06)
$App \times \log impression$	0.264***
	(8.76)
buyer_impression app/pc	-0.079**
	(-2.30)
buyer_domestic	0.034**
	(1.13)
buyer_stars	0.073***
	(21.89)
buyer_female	-0.130***
	(-10.51)
buyer_age	0.002***
	(3.50)
buyer_tenure	0.00002***
	(2.72)
Constant	-2.223***
	(-53.65)
N	185,563

z-statistic in parentheses. The standard errors are robust.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table F.2: Estimates of the channel effect by matched sample

	GMM-IV
$\log impression$	0.119***
	-8.79
App	-0.592***
	(-8.71)
$App \times \log impression$	0.282***
	-9.89
buyer impression app/pc	-0.092**
	(-2.87)
buyer_domestic	0.076**
	-2.72
buyer_stars	0.073**
	-23.72
buyer_female	-0.126***
	(-11.15)
buyer_age	0.002***
	-4.37
buyer_tenure	0.00003***
	-3.85
Constant	-2.154***
	(-56.01)
N	185,563

z-statistic in parentheses. The standard errors are robust.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix G

TEXT ANALYSIS

I preprocess the text of project description by removing stop words, and lemmatizing. Following Mikolov et al. (2013b), I demonstrate the structure of neural network for *word2vec* if Figure G.1. It has a hidden layer with pre-set length. This is basically the size of the dense vector for embedding. Figure G.1 demonstrates an example of predicting task with a sample sentence: "*the cat sat on*". I this example, the algorithm takes the first three words in one-hot vector format and predicts the next word. In my case, I use a window size of 5 words around the target word. Finally, the task is to learn the weight matrix, W, through stochastic gradient descent. The weight matrix has dimensions matching the vocabulary size and the dense vector length. Each row in the matrix represents a corresponding word in my vocabulary. Thus, I can infer the word embedding by taking the row vectors of corresponding rows in the weight matrix.

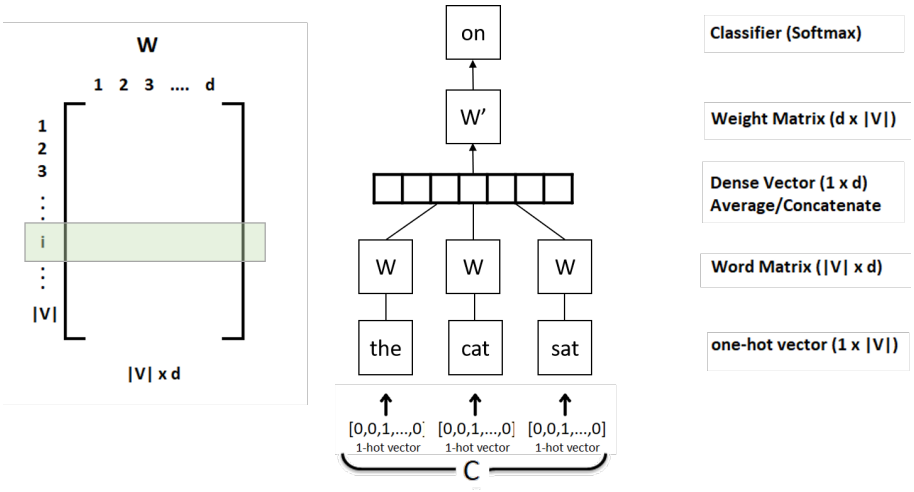


Figure G.1: *word2vec* neural network structure.

Similarly, I can build up a network structure including document identifiers such as Figure G.2. We, thus, have an additional weight matrix, D, corresponding to documents. Hence, each row in the weight matrix represents the embedding of a document. There are multiple variations within the class of doc2vec. I chose Paragraph Vectors Distributed Bag of Words (PD-DBOW) structure. In some applications, it is shown that this structure outperforms the other variations.

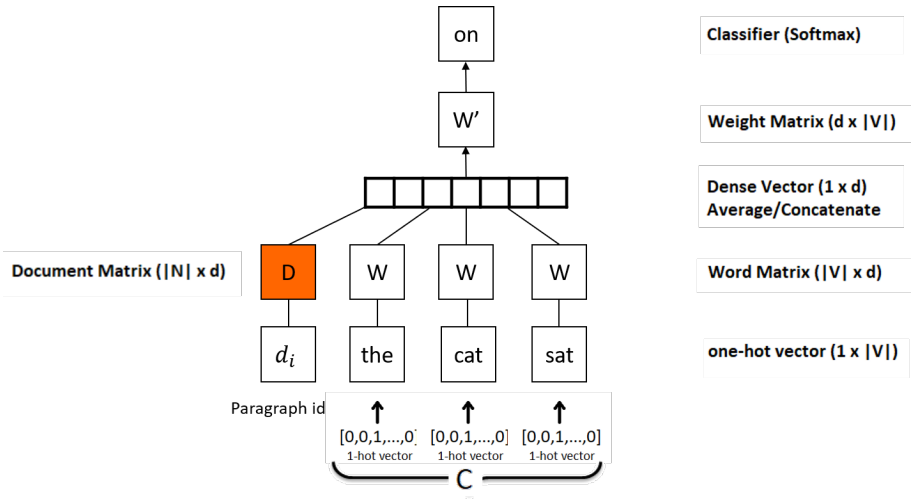


Figure G.2: *doc2vec* neural network structure. source: Le and Mikolov (2014)

As a product of the model, I can measure similarity of words by calculating the *cosine* similarity between two resulted dense vectors. I perform this task as a sanity check to see if the results of the model would make sense. As an example, I provide the most similar words to "headphone" in Table G.1. The results are all within the context of interest suggesting that the trained model captures the semantics and meanings of the project description.

Table G.1: Most similar words to “headphone” by cosine similarity.

Target Word	Similar Words
headphone	sound
	speaker
	earphone
	audio
	music
	earbuds
	microphone
	ear
	headset
	listening

Appendix H

LATENT DIRICHLET ALLOCATION

LDA is a generative statistics model in which the observed data are the words of each document and the hidden variables represent the latent topical distribution. I follow the notations in Blei et al. (2003) for clarity. Each document, d, is viewed as a mixture of various topics. This is a natural assumption in many contexts, as a single document cannot be identified with only one topic. The probability distribution is denoted by θ_d which is a $1 \times K$ vector. There are overall K topics. Each topic, k, has a distribution over words denoted by β_k which is a $1 \times V$ vector. V is the size of vocabulary in the corpus. That is, for the topic k the probability of each word is assigned is given by β_k. The i^{th} word in document d, denoted by $w_i^d \in \{1, ..., V\}$, belongs to topic z_i^d which is a latent variable. Let η be a scalar and α a positive K-vector. I let $Dir_V(\alpha)$ denote a V-dimensional Dirichlet with vector parameter α and $Dir_K(\eta)$ denote a K-dimensional symmetric Dirichlet with scalar parameter η.

The generative process manifested by LDA is as follows:

1. For each topic,
   (a) Draw a distribution over words $\beta_k \sim Dir_V(\eta)$.
2. For each document,
   (a) Draw a vector of topic proportions $\theta_d \sim Dir(\alpha)$.
   (b) For each word,
      i. Draw a topic assignment $z_i^d \sim Mult(\theta_d)$.
      ii. Draw a word $w_i^d \sim Mult(\beta_{z_i^n})$.

Figure H.1 demonstrates a graphical model representation of LDA with model parameters.

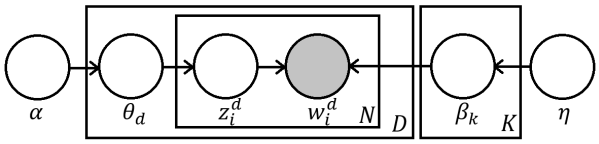


Figure H.1: Graphical model representation of the Latent Dirichlet allocation (LDA). Nodes denote random variables; edges denote dependence between random variables. Shaded nodes denote observed random variables; unshaded nodes denote hidden random variables.

The prior in LDA is coming from a Dirichlet distribution. The Dirichlet is used as a distribution over discrete distributions. The Dirichlet has density

$$p(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \Pi_i \theta_i^{\alpha_i - 1}, \tag{H.1}$$

where the parameter α is a positive K-vector, and Γ denotes the Gamma function. If all the components of the parameter is are equal, then Dirichlet would be *symmetric*.

The distribution of the *posterior* given the observed documents is given by Equation H.2. As this distribution is intractable because of the denominator, I use approximation methods. The topic probability of a term, the topic proportions of a document, and the topic assignment of a word are derived by the expected value of random variables $\beta_{k,v}$, $\theta_{d,k}$, and $z_{d,i,k}$, respectively. These quantities are conditioned on the observed corpus.

$$p(\theta_{1:D}, z_{1:D,1:N}, \beta_{1:K}|w_{1:D,1:N}, \alpha, \eta) = \frac{p(\theta_{1:D}, z_{1:D}, \beta_{1:K}|w_{1:D}, \alpha, \eta)}{\int_{\beta_{1:K}} \int_{\theta 1:D} \sum_z p(\theta_{1:D}, z_{1:D}, \beta_{1:K}|w_{1:D}, \alpha, \eta)}. \tag{H.2}$$

There are various techniques of approximations including mean field variational inference (Blei et al., 2003), collapsed variational inference (Teh et al., 2007), expectation propagation (Minka and Lafferty, 2002), and Gibbs sampling (Steyvers and Griffiths, 2007). I explain

the mean field variational approach in this section.

H.1 Mean Field Variational Inference

In this approach a simpler distribution containing free variational parameters is used to approximate an intractable posterior distribution in Equation H.2. The hidden variables in posterior distribution are dependent when conditioned on data. Thus, one must sum over all configurations of the interdependent N topic assignment variables z_i.

In mean field variational distribution for LDA, the variables are independent of each other:

$$q(\theta_{1:D}, a_{1:D,1:N}, \beta_{1:K}) = \Pi_k q(\beta_k|\lambda_k)\Pi_d \Big(q(\theta_d|\gamma_d)\Pi_i q(z_i^d|\phi_i^d) \Big), \tag{H.3}$$

where λ_k, γ_d, and ϕ_i^d are V-Dirichlet, K-Dirichlet, and K-Dirichlet distributions that describe β_k, θ_d, and z_i^d, respectively. Following Blei et al. (2003), the objective function becomes:

$$\mathcal{L} = \sum_k \mathbb{E}[\log p(\beta_k|\eta)] + \sum_d \mathbb{E}[\log p(\theta_d|\alpha)] + \sum_d \sum_i \mathbb{E}[\log p(z_i^d|\theta_d)] + \sum_d \sum_i \mathbb{E}[\log p(w_i^d|z_i^d, \beta_{1:K})] + \mathrm{H}(q), \tag{H.4}$$

where H denotes the entropy. See Blei et al. (2003) for details on how to compute this function. A coordinate ascent approach is applied for the optimization. Algorithm 1 demonstrates one iteration of mean field variational inference for LDA.

Algorithm 1: One iteration of mean field variation inference for LDA

for *each topic k and term v* **do**
  $\lambda_{k,v}^{(t+1)} = \eta + \sum_d \sum_i 1(w_{d,i} = v)\phi_{i,k}^{(t)}$.
for *each document d* **do**
  (a) Update γ_d:
  $\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_i \phi_{d,i,k}^{(t)}$
  (b) **for** *each word n, update $\phi_{d,i}$* **do**
    $\phi_{d,i,k}^{(t+1)} \propto \exp\{\Phi(\lambda_{d,k}^{(t+1)}) + \Phi(\lambda_{k,w_i}^{(t+1)}) - \Phi(\sum_v \lambda_{k,v}^{(t+1)})\}$
    where Φ is the digamma function, the first derivative of the $\log\Gamma$ function.

The estimated quantities to explore the corpus are derived directly from the variational

distribution. The per-term topic probabilities are

$$\hat{\beta}_{k,v} = \frac{\lambda_{k,v}}{\sum_{v'} \lambda_{k,v'}}. \tag{H.5}$$

The per-document topic proportions are

$$\hat{\theta}_{d,k} = \frac{\gamma_{d,k}}{\sum_{k'} \gamma_{d,k'}}. \tag{H.6}$$

The per-word topic assignment expectation is

$$\hat{z}_{d,i,k} = \phi_{d,i,k}. \tag{H.7}$$

H.2 Author-Topic Model

Rosen-Zvi et al. (2004) extend the Latent Dirichlet Allocation model (Blei et al., 2003) to a generative model for documents called *author-topic* model to include authorship information. Each author is associated with multiple documents. This approach can be extended to include other meta data associated with each documents. This generative model simultaneously models the content of documents and the association of meta data (e.g. interests of authors).

The author-topic model uses a topic-based representation to model both the content of documents and the interests of authors. A group of authors, a_d, decide to create the document d. For each word in the document an author is chosen uniformly at random. Then, a topic is chosen from a distribution over topics specific to that author, and the word is generated from the chosen topic. Figure H.2 shows the graphical representation of the model. In the model, as Rosen-Zvi et al. (2004) describe, x indicates the author responsible for a given word, chosen from a_d. Each author is associated with a distribution over topics, θ, chosen from a symmetric Dirichlet(α) prior. Note that the model allows for multiple authors. In my case, however, I use a single author setting as the creator of the video. The mixture weights corresponding to the chosen author are used to select a topic z, and is generated according to

the distribution ϕ corresponding to that topic, drawn from a symmetric Dirichlet(β) prior.

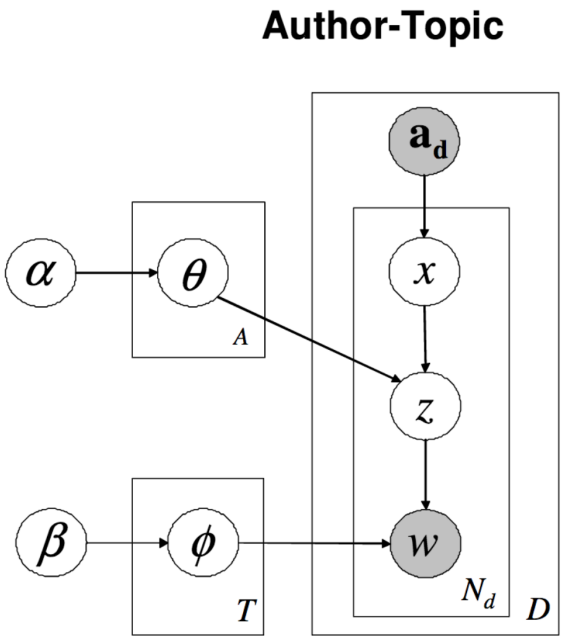


Figure H.2: Graphical model representation of the author-topic model. Nodes denote random variables; edges denote dependence between random variables. Shaded nodes denote observed random variables; unshaded nodes denote hidden random variables.

Note that the LDA is a special case of author-topic model. If each document is identified with a unique author (aka creator), the two models are equivalent. Rosen-Zvi et al. (2004) describe a Gibbs sampling algorithm to learn the model parameters. I use approximate variational inference as explained in H.1.

H.3 Text Preprocessing

In this section, I describe how I construct my textual data ready to implement topic model. For most videos, there are two parts with textual information: *Video Description* and *Video Title*. I use both texts for my topic model. I concatenate the texts from description and title. Note that this is not the case for all the videos; some videos do not have either title

or description, and some offer the same title as the description. In these cases, I use only one part either the description or title as it applies. Figure H.3 shows an example video with highlighted video description and video title. In this example, the video title is *"This woman rowed across the Atlantic."* and the video description is *"Row, row, row your boat... across the ocean? Cutting edge designs are allowing specialized ocean rowing kayaks to travel thousands of miles. http://cnn.it/1U9kXWq"*.

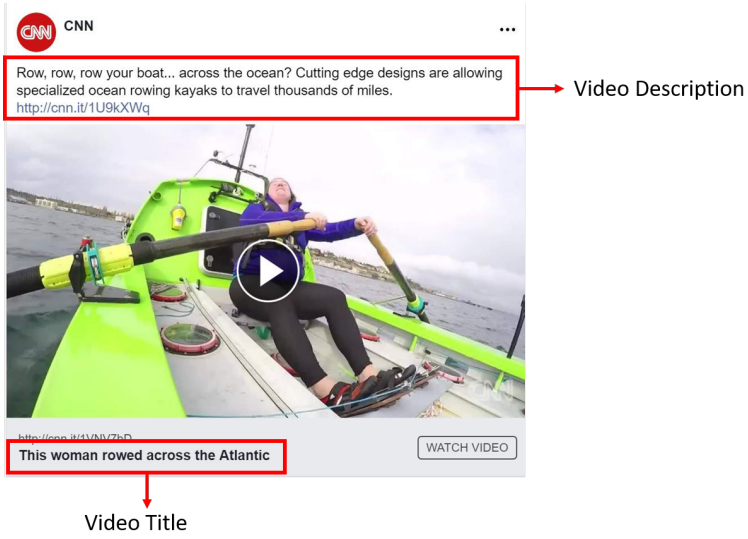


Figure H.3: There are two parts with textual information: *Video Description* and *Video Title*. I use both texts for my topic model. I concatenate the texts from description and title.

I describe several steps by which I preprocess the textual data.

1. **Concatenation.** First, I concatenate the video title and video description. In the example of Figure H.3, I will have: *This woman rowed across the Atlantic. Row, row, row your boat... across the ocean? Cutting edge designs are allowing specialized ocean rowing kayaks to travel thousands of miles. http://cnn.it/1U9kXWq"*. As mentioned earlier, in some cases, I skip this step.

2. **Removing hyperlinks and email addresses**. In the next step I remove hyperlinks

and email addresses. Note that I do *not* remove hashtags as they carry contextual information. As a result, I remove *"http://cnn.it/1U9kXWq"* from the sentence above.

3. **Common Cleaning.** I remove possessive *'s*, lower case letters, remove quotations, dots, and other punctuation.

4. **Removing stopwords.** The stopwords often do not carry significant contextual information. Thus, they add noise to the learning process. This is similar to *tf-idf* approach by penalizing too frequent terms. In above example, I have *"woman rowed atlantic row row row boat ocean cutting edge designs allowing specialized ocean rowing kayaks travel thousands miles"* as a result.

5. **Lemmatization.** I lemmatize the tokens in the sentences and remove the pronouns from my corpus. The processed tokenized terms in my example after lemmatization are: *"woman"*, *"row"*, *"atlantic"*, *"row"*, *"row"*, *"row"*, *"boat"*, *"ocean"*, *"cut"*, *"edge"*, *"design"*, *"allow"*, *"specialized"*, *"ocean"*, *"rowing"*, *"kayak"*, *"travel"*, *"mile"*.

6. **Removing very frequent and very rare words.** I remove those terms that occur too frequently (in more than half of the documents) and too rarely (with less than 100 occurrences).

Appendix I

MATRIX COMPLETION

Matrix completion is referred to the set of methods to fill a partially observed matrix of data (e.g. user characteristics and preferences). That is, the data matrix contains missing entries. There are various applications of matrix completion including movie or product recommendation (e.g. Netflix), and collaborative filtering (Breese et al., 1998). One of the matrix completion approaches assumes the matrix is *low-rank*. In low-rank matrix completion, I try to find the lowest rank that matches the observed matrix (Candès and Recht, 2009). The original problem is NP-hard, but there are tractable techniques to solve the problem. In the following section, I explain how the task of matrix completion is implemented.

I.1 Low-rank Solution

Suppose I wish to fill the missing values for a square $n_1 \times n_2$ matrix M of rank r. Matrix M can be represented by n^2 terms with only having $(2n - r)r$ degrees of freedom. When the rank of M is low (i.e. $r \ll n$), it implies M has much less information despite the possible large dimensions. From all the entries of matrix M, I only observe a sparse set of Ω of observations ($M_{ij} \in \Omega$ where $(i, j) \subseteq \{1, ..., n_1\} \times \{1, ..., n_2\}$). The goal is to make an educated guess about missing values as Candès and Recht (2009) describe. In many cases, the matrix I wish to recover is known to be structured as a *low-rank* matrix. In the Netflix example, this assumption implies that the ratings are affected by a few factors. The problem can be written in a rank minimization problem as in Equation I.1.

$$\begin{aligned} \min_{X \in \mathbb{R}^{n_1 \times n_2}} \quad & rank(X) \\ s.t. & X_{ij} = M_{ij} \ (i, j) \in \Omega \end{aligned} \tag{I.1}$$

The minimization task in Equation I.1 is a NP-hard problem and not feasible. To make the problem tractable, I can replace the $rank(X)$ with its convex surrogate known as *nuclear* or *trace* norm: $||X||_* = tr((XX^T)^{1/2}) = \sum_k \sigma_k$, where σ_k are singular values of X. This transformation leads to a semi-definite problem for which there are efficient algorithms to solve.

$$\begin{aligned} &\min_{X \in \mathbb{R}^{n_1 \times n_2}} ||X||_* \\ &s.t.\ X_{ij} = M_{ij}\ (i,j) \in \Omega \end{aligned} \tag{I.2}$$

The assumption is that Ω are sampled from the orthogonal model, i.e., sampled uniformly random. If Ω is uniformly distributed and $|\Omega|$ is sufficiently large, then minimizer of Equation I.2 is unique and coincides with the minimizer of Equation I.1 (Kalofolias et al., 2014). I provide evidence that why these assumptions are reasonable.

In my data, there is a pool of combined $9,161$ of focal creators and sponsors. To learn the patterns of browsing, I use the data from $218,895$ Facebook pages. Thus, my data matrix is 9161×218895 in dimension. Figure I.1 presents the distribution of number of unique creators that there is observed data for each creator. For the majority of the creators, I have a uniform distribution of number of observed entries.

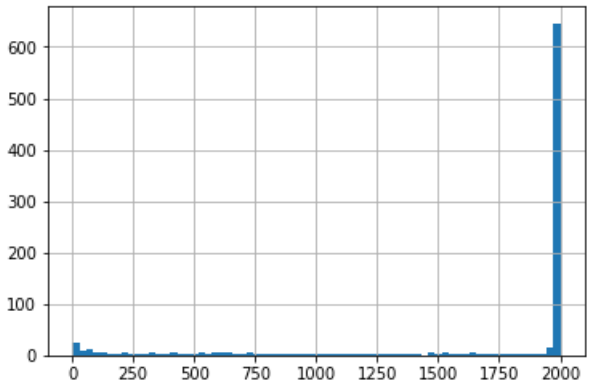


Figure I.1: The histogram of the number of observed unique Facebook pages associated with the focal page. The histogram suggests that the observed data follows two different patterns for creators and sponsors.

Candès and Recht (2009) argue that if Ω is from the orthogonal model, then there are C and c such that if

$$m \geq Cn^{5/4} r \log n, \tag{I.3}$$

the minimizer to the problem in Equation I.2 is unique and equal to M with probability at least $1 - cn^{-3} \log n$, where $n = \max(n_1, n_2)$, and r is the rank of M. In addition, if $r \leq n^{1/5}$, then the recovery is exact with probability at least $1 - cn^{-3} \log n$ provided that

$$m \geq Cn^{6/5} r \log n \tag{I.4}$$

This implies that for small values of the rank (e.g. $r = O(1)$, or $r = O(\log n)$), I need to have on the order of $n^{6/5}$ entries which is much smaller that n^2. Thus, under this hypothesis, the solution is unique, and equivalent of the original problem.

VITA

Shahryar Doosti is a researcher in *Information Systems* and *Digital Economics*. He has received his engineering and MBA degrees from Sharif University of technology, Iran. Prior to joining the Foster School of Business at the University of Washington for the degree in Information Systems, he worked in industry as a marketing manager for a few years. He gained practical insights on how businesses work during his tenure in the managerial roles. He became interested in online markets and data-driven approaches in the technology era. In pursuing of analytic and scientific approaches to answer economics of technologies, he joined the PhD program at UW and worked under supervision of professor Yong Tan. He mainly focuses on business analytics and causal inferences empowered by big data and statistical learning. He leverages unstructured data to gain insights on the interactions of different players in online platforms including businesses, customers, platforms, and regulators. As of July 2019, he will be joining the Argyros School of Business and Economics at Chapman University as an Assistant Professor.