UNIVERSITY OF CALIFORNIA,
IRVINE

The Impact of Social and Spatial Proximity on Consumer Choice in Digital Markets

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Management

by

Yi-Jen Ho

Dissertation Committee:
Professor Sanjeev Dewan, Chair
Associate Professor Vidyanand Choudhary
Assistant Professor Mingdi Xin

2016

DEDICATION

To

Dad and Mom

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the sincerest appreciation to my committee chair, Professor Sanjeev Dewan, who has been a tremendous adviser throughout my doctoral program. His guidance and encouragement are valuable to the completion of this dissertation. His great vision always inspiring. I am truly indebted to him for all his time and patience.

I would also like to thank my committee members, Professor Vidyanand Choudhary and Professor Mingdi Xi for their insightful and constructive feedback. I thank Professor Jui Ramaprasad at McGill University, who provides me with a precious opportunity to work with her together and help for job searching. I thank Professor Jiawei Chen, the proposal committee member, for giving me a lot of helpful suggestions on the research design. I would also like to acknowledge the helpful comments and suggestions of Professor Vijay Gurbaxani.

Due to the friendship and support of my fellow doctoral students, it could not be better for me to have such great experience during my Ph.D. study at University of California, Irvine. In particular, I want to thank Qiguang Wang, Qin Li, Federico Bumbaca, Jin Sik Kim, Jong-Yu (Paula) Hao, and Shengjun Mao.

Lastly, I need to thank my dearest parents, Chen-Ping Ho and Kuei-Feng Lee, coolest brother, Yi-Chun (Chad) Ho. Thank you for encouraging me to pursue a doctoral degree and support me to finish this long journal. With your love and advice, I finally find the way back to be confident and proud of myself. I Love them so much.

CURRICULUM VITAE

Yi-Jen Ho

2003	B.B.A. in Management Information Systems, National Central University
2008	M.S. in Management Information Systems, University of Arizona
2010-16	Research & Teaching Assistant, The Paul Merage School of Business, University of California, Irvine
2014	Instructor, The Paul Merage School of Business, University of California, Irvine
2016	Ph.D. in Management, The Paul Merage School of Business, University of California, Irvine

FIELD OF STUDY

Information Systems

PUBLICATIONS

Chen, H., X. Li, M. Chau, Y. Ho, C. Tseng. 2011. Using Open Web APIs in Teaching Web Mining. *IEEE Transactions on Education* **52**(4) 482-490.

Dewan, S., Y. Ho, J. Ramaprasad. 2015. Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Community. *Information Systems Research* (Forthcoming).

ABSTRACT OF THE DISSERTATION

The Impact of Social and Spatial Proximity on Consumer Choice in Digital Markets

By

Yi-Jen Ho

Doctor of Philosophy in Management

University of California, Irvine, 2016

Professor Sanjeev Dewan, Chair

Technology is closing the distance between users on the one hand and between users and businesses on the other. Social technologies help create social proximity and promote the sharing of information, while location-enabled services enable spatial proximity and allow businesses to leverage the precise dynamic location of users in their marketing strategies. This dissertation examines the impact of social and location-based technologies on consumer choice in the context of the music industry and mobile analytics. In the music context, I examines the interactions between social proximity and consumer choice in the context of an online music community. In the area of mobile analytics, my research studies how location-based services (i.e., local search and geo-fence marketing) impact consumer choice and transform business strategies. My research design applies a variety of empirical methods to highly granular data. My analysis finds robust evidence of the impact of social and spatial proximity on consumer choice. I discuss implications for design and marketing strategies for online communities, mobile local search engine and geo-fence advertising, such as the contexts studied in this dissertation.

Keywords: social influence, mobile, local search, geo-fence advertising, proximity

CHAPTER 1

Introduction

Technology is closing the "distance" between users on the one hand and between users and businesses on the other. Specifically, social-networking technologies help create social proximity by enabling social ties and promoting the sharing of news, information and opinions between users. At the same time, location-enabled mobile services such as local search and location-based marketing are enabling spatial proximity, allowing businesses to leverage the precise dynamic location of users in their marketing and distribution strategies. Both social and spatial proximity are transforming the ways in which consumers find products and services in digitally-enabled markets, and even how they consume them.

Understanding how ongoing technology innovations are changing the way consumers search for products and services, and make consumption choices, has remained an essential part of research in information systems (see, e.g., Bakos 1997; Brynjolfsson et al. 2010; Dewan and Ramaprasad 2012; Ghose et al. 2012). Each major innovation — from Web 1.0 in the 1990s, to Web 2.0 and social media in the 2000s, to location-enabled mobile technology today — has in turn expanded the bandwidth and scope of interaction between consumers and businesses on the one hand, and between consumers themselves on the other. In the era of social and mobile technology, it is fascinating to witness how technology continues to reshape consumer behaviors. Accordingly, my dissertation studies the impact of social and spatial proximity on user choice in two distinct contexts. Specifically, I examine the role of technology-enabled social proximity on user choice in an online community. Then, I investigate the impact of technology-enabled spatial proximity on user choice in two location-enabled mobile commerce contexts (i.e., mobile local search and mobile geo-fence advertising). The rest of this chapter discusses the introductions of these three research contexts.

Social Proximity

Social or peer influence has long been recognized as a driver of adoption and consumption decisions, going back to Katz and Lazarsfeld (1955), Arndt (1967) and Bandura (1971), but its importance has only been heightened recently with the proliferation of online social media and social networks (see, e.g., Godes et al.

2005, Brown et al. 2007, Chen et al. 2011, Aral and Walker 2011). In the music industry, the context we study here, social media have made sharing of consumption, tastes and preferences easier than ever before, and in a recent survey 54% of subjects indicated that they base their music purchasing decisions on positive recommendations from friends (The Nielsen Company 2012a). Per Nielsen Global Trust's survey (The Nielsen Company 2012b), 92% of consumers say that recommendations from people they know are the most trusted sources of information when making consumption decisions, followed by 70% of consumers who say that they trust consumer opinions posted online.

Despite this anecdotal evidence, we do not really know whether it is aggregate *popularity* information that matters, or information about consumption by friends in close social *proximity* — or both? We expect peer consumption information to influence the choice of music by users. In fact, there are two types of influence, one driven by aggregate peer consumption information and the other by music consumption in social network proximity. Our analysis covers both types of influence, where we call the effect of total favorites information *popularity influence* and the effect of friends' favorite information *proximity influence*. Our study is designed to measure each type of influence, as well as the interaction between the two. Specifically, our research questions are as follows: How does popularity influence affect music consumption choices? Is it more important for mainstream or niche music? How important is proximity influence in music consumption? What is the nature of interaction between the two types of influence? Are they complements or substitutes?

Recently, the role of social influence on consumer choices has been examined in a variety of contexts, such as movie sales (Moretti 2011), Facebook applications (Aral and Walker 2011), adoption of the iPhone 3G (de Matos et al. 2014), restaurant dining choices (Cai et al. 2009), software downloads (Duan et al. 2009), music subscription services (Bapna and Umyarov 2015), among others. In Chapter 1, we study the role of peer influence on consumption in an online music community — an mp3 blog aggregator — where users can listen to songs drawn from a large number of mp3 blogs. As a result of features introduced on the web site over time, users can listen to songs, favorite them, and use social networking features to follow other users and track their favoriting behavior. The site provides the total number of favorites

garnered by each song listed on the site and allows users to quickly look up which songs have been favorited by their "friends," allowing us to study both popularity and proximity influence. Prior work has looked at popularity influence (e.g., Chevalier and Mayzlin 2006, Dewan and Ramaprasad 2012, 2014, Chen et al. 2011) and proximity influence (e.g., Ma et al. 2010 and Egebark and Ekstrom 2011) individually, but has not studied them jointly in the same context, as we do here. Further, we are able to exploit exogenous feature implementations on the website which allow us to cleanly identify the two types of influence in a quasi-experimental framework.

The music context is ideal for the study of IT-enabled social influence, for a number of reasons. First, music is an experience good, so that consumers potentially value the opinions and actions of other consumers as signals of whether or not they would like the music themselves. Second, music is an information good, where discovery and consumption are increasingly becoming online activities, and in our case, these two activities occur on the very same website. Finally, the music industry has been transformed by technology and social networks in profound ways, so that understanding social influence in this context will foreshadow what we can expect for other information and experience goods, such as movies, software, and other digital media.

It is important to discuss the unit of social interaction in our setting, which is "favorite," akin to the "like" action on Facebook. Users can favorite songs and they can also favorite other users, giving them visibility into their friends' favoriting behavior. These two types of favoriting actions are illustrated in Figure 1.1. A directed arrow connecting two user nodes indicates that the first user has favorited the second; e.g., User A is following Users B, C and D.[1] The figure also shows which users have favorited each of the two songs 1 and 2. Thus, users can view two types of information for any song posted on the website: total favorites and friends' favorites, corresponding to what we call popularity influence and proximity influence, respectively.

[1] At the time of our study only 15-20% of the users were using the social networking features, while the remaining users were "isolates"; i.e., users who were using the site to sample music, but were not following other users.

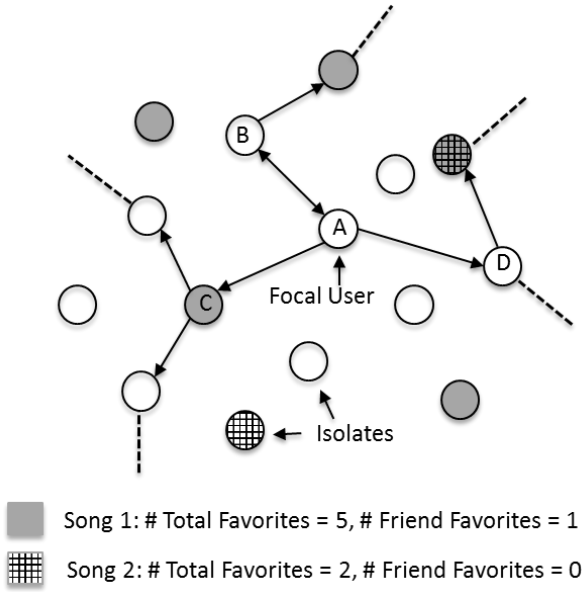


Figure 1.1. Illustrating Popularity and Proximity Social Interactions

Online social influence mediated by popularity or proximity has different implications for website design and marketing strategies. If proximity influence is important (as in the studies of de Matos et al. 2014, Aral and Walker 2011) then the website should incentivize the creation of social ties, provide visibility of social connections and actions, and encourage interaction and co-consumption. On the other hand, if popularity influence is important (as in Chen et al. 2011, Duan et al. 2009) then it would be a good idea to emphasize popularity statistics, for the overall population and also for sub-populations, based on demographics, listening preferences, etc. It might also make sense to provide information on multiple dimensions of popularity, such as the number of times a song has been listened to, saved to a playlist, or "liked." The popularity information could also combine internal and external (e.g., best seller lists or rankings) measures that are relevant to the online community in question. Finally, the interaction between the two types of influence also matters. If the two are substitutes then it would be important to understand which type of influence is more important for different types of users and music, so that the appropriate type of signal is prioritized, depending on the situation. If the two types of influence are complements then strategies to amplify the effect of one type of influence with the other might be useful. In general, design

and marketing strategies need to be linked to the types of social influence that are relevant to the context, as well as their interaction. This is the issue that broadly motivates this study.

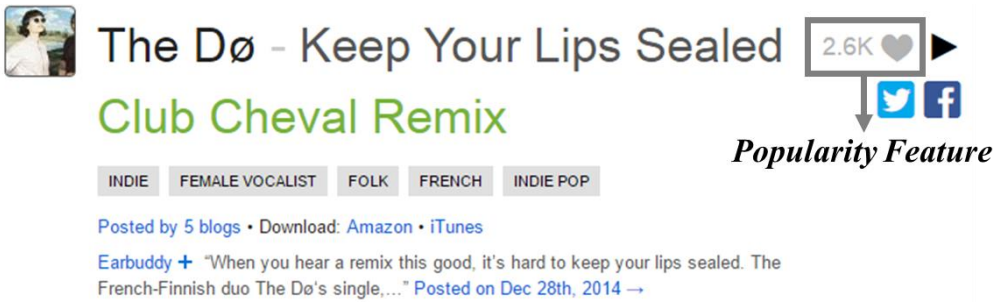


Figure 1.2. Popularity Feature on The Hype Machine (THM)

To study popularity influence, we exploit a natural experiment enabled by a newly implemented feature in an online music community, The Hype Machine (THM).[2] The popularity feature, illustrated in the screenshot of Figure 1.2, allowed users to observe all other users' music favoriting behavior in the aggregate, albeit anonymously — the feature was implemented on October 1, 2008. We deploy a difference-in-difference (DD) methodology to measure the impact of aggregate favorite data on other users' consumption decisions. In the second part of Chapter 1, we focus on proximity or social network influence. We deploy a variety of approaches to identify and measure proximity influence, including probit and hazard models, building on the work of Aral et al. (2009). Identifying proximity influence using observational data is challenging due to homophily, which may influence both the formation of social ties and music consumption decisions. To overcome the potential selection bias due to homophily, we use two matching techniques, propensity score matching and Euclidean distance matching, as we explain in more detail below. Finally, we develop a combined model to jointly estimate both types of influence using a two-dimensional quasi-experimental design including both popularity and proximity treatments.

To summarize our results, we find strong and robust evidence for popularity influence. Our difference-in-difference results confirm that being able to observe aggregate popularity information does

[2] The Hype Machine (THM), previously studied by Dewan and Ramaprasad (2012), is the largest mp3 blog aggregator. It tracks thousands of mp3 blogs and provides links to blog posts and mp3 tracks, for other users to stream but not to download.

have a causal impact on subsequent consumption choices. We further find that popularity influence is significant only for newly posted songs (due to the specific nature of the site), and it is more important for narrow-appeal music as compared to broad-appeal music, in line with the findings of Tucker and Zhang (2011). We also find consistent evidence of proximity influence, after accounting for homophily. Finally, our results suggest that popularity and proximity influence are substitutes for one another. Popularity influence is most effective when proximity influence is not available, either because the user is not connected to other social network users, or if none of a user's friends have favorited a song. Proximity influence, when available, tends to dominate popularity influence.

Spatial Proximity – Mobile Local Search

The advancement of mobile and geo-location technologies has redefined both the ways how consumers surf the Internet and how companies market to them. Consumers increasingly search for information using their mobile devices, often looking for location-based information pertaining to their present immediate surroundings. In fact, mobile devices have become the most used digital platform, accounting for some 60% of Internet usage in the United States (comScore 2014c). Eighty percent of consumers search for information with their geographic proximity[3] in mind, and 88% of them do so on mobile devices (Google 2014). Specifically, the Global Positioning System (GPS) service on smartphones enables consumers to access location-based information and services that are highly sensitive to the precise location of the user. To meet this rapidly growing demand for local search, a variety of location-enabled services are available either on mobile pages or in apps, including Google local, Yelp and YP, and contributes 7.1 billion dollars in revenue for mobile local advertising in 2014 (BIA/Kelsey 2014b).

Despite the importance of local search, we have little understanding of how consumers react to location-based information on their mobile devices. In fact, users consent to the carrier's request to use their current location in local search (even on smartphone devices equipped with a GPS chip). In our setting,

[3] Geographic proximity refers to the mileage between the user location and a physically entity. We use geographic proximity, local proximity and distance interchangeably throughout Chapter 1.

every time a user selects “Local Search” the provider asks whether or not it is OK to use the user’s current location. If the user accepts, the carrier can utilize the precise latitude/longitude information from the GPS chip on the device. Although the procedure of conducting local search is always the same, the user could be located in different areas which she may be associated with or not. We call these two search scenarios “Home” or “Away,” depending on whether or not the user conducts a local search in her familiar areas, such as home or office location. In Chapter 2, we conjecture that these two search scenarios may lead to different behavior and outcomes due to the extent of brand association. In particular, we study the following research questions: What are the impacts of three important factors, i.e., proximity, brand and screen position, on click-through rates (CTRs)? How do users make tradeoffs among these factors? How does click-through behavior vary between home and away searches?

The answers to these questions are crucial for practitioners to effectively implement location-based advertising, such as geo-fencing/geo-targeting and sponsor listings, in mobile local search. On the one hand, our findings on the tradeoff between brand popularity and distance provides useful guidance for optimizing geo-fencing boundaries. For an advertiser, we can quantify the impact of distance on click-through performance at the individual level (i.e., personalized geo-fencing) by evaluating click performance, brand characteristics and user utility. On the other hand, knowing the tradeoff between screen position and distance/brand popularity in local search, our findings provide implications on the insertion of sponsored links into search results. We advise that an advertiser insert sponsored listings in a situation where (1) it is a popular brand, (2) the position of its impression in search results is relatively low, and (3) the distance is in a reasonable range. This is, although a user prefers a popular store over a closer one, she still needs to see the impression of the popular store before stopping screening search results.

To exam these questions, we use a unique dataset of local search transactions collected from a nationwide wireless service provider. The data include click-through impressions corresponding to local searches conducted over several months in the first quarter of 2015. We are able to observe all the impressions seen by a user in response to a local search query, including which impressions were clicked upon, if any. In this study, we focus our analysis on the searches using keywords related to restaurant search,

including “restaurant”, “restaurants” and “food”. To profile users, we utilize mean shift cluttering to identity their home locations. In addition, our random-coefficients model in a hierarchical Bayesian framework can account for user heterogeneity to distill the impact of distance, brands and rank – and the interactions among them – on whether or not an individual local search impression is clicked upon. In the hierarchical setting, we further compare how users heterogeneously react to those factors, depending on whether the search is conduct at home or an away location. To estimate the model parameters, we develop a Markov Chain Monte Carol algorithm incorporating Metropolis-Hastings random walk.

Spatial Proximity – Mobile Geo-fence Advertising

The growing popularity of smart mobile devices with in-built geo-location (GPS) capabilities has led to the emergence of location-based advertisement practices, for a large variety of consumer products and services. Consumers spend an increasing amount of time on their mobile devices. Indeed, since February 2013 users are spending more time on average on their mobile devices than on their desktop computers; by the end of 2014 mobile accounts for 64% of user time as compared to 36% for desktop (comScore 2014b). Further, almost 90% of the time on mobile devices is inside mobile apps. Overall, it is estimated that users spend the equivalent of 23 days per year on their mobile phones (mobilestatistics.com 2015). A growing proportion of the mobile devices are smartphones equipped with Global Positioning System (GPS) chips, which allows advertisers to more readily exploit user location in their messaging campaigns.

The recent spike in mobile ad spending is well-documented. Advertisers spent $28.24 billion on mobile in the US market in 2015, an increase of nearly 60% over 2014 (eMarketer 2014), and BIA/Kelsey (2015) reports that mobile as a whole will account for at least 11.5 percent of total local media ad revenues in the next few years. Moreover, location-based mobile ad revenues in the US will grow from $6.8 billion in 2015 to $18.2 billion in 2019, representing a 28.5 percent compound annual growth rate (BIA/Kelsey 2014a). All together McKinsey Global Institute (2011) estimated that location-based services will account for $100 billion in revenues for service providers and over $700 billion of value to end users.

Not surprisingly, there is an increased research attention payed to location-based advertising, including location-based SMS (Luo et al. 2014), mobile local search (Dewan and Ho 2015), and mobile

couponing (Molitor et al. 2014). The past literature generally shows that the distance between the consumer and advertiser location has a negative impact on advertising performance. However, prior research has not examined the tradeoff between physical distance and competition; i.e., how is the impact of distance moderated by the extent of competition in the consumer's location vicinity, from similar products and services. Accordingly, in Chapter 2, we examine the separate impacts of distance and competition on location-based advertising performance, as well as the interaction between distance and competition.

We study these issues in the context of mobile geo-fencing campaigns, where advertisers send messages to smart device users in a pre-defined geographic area surrounding the business in question. For example, consumers who are within say five miles of a mall (based on their shared GPS coordinates) are targeted with ads by an advertiser in the mall, the logic being that proximate users are more likely to respond to the ads in a positive manner, all else equal. In this context, we address the following research questions: (1) What is the impact of *distance* and *competition* on click and conversion performance; (2) how does competition moderate the impact of distance (and vice versa), and (3) how are these effects different across the click-through and conversion stages of consumer decision making?

To examine these questions, we use a unique dataset of geo-fence advertising transactions collected from one of the largest location-based marketing companies. The data set covers all of the bids (in real-time advertising auctions) and resulting impressions originating from the marketing agency during the month of January 2015. The marketing agency bids in auctions on behalf of advertisers, and in many cases also runs location-based campaigns on the advertiser's behalf. Our data set includes latitude/longitude location data for both consumers and advertiser businesses, mobile device and operating system characteristics, publisher/app characteristics, and consumers' click and conversion response.[4] In addition, we obtain supplementary data from Yelp and Google to construct the index of the competition between an

[4] As we explain below, the conversion information is limited to whether or not the consumer took some further action on the landing page, such as calling the establishment or looking up directions on a map.

advertiser and its rivals. American Community Survey data (United State Census Bureau 2014) is also collected to approximate the demographics of consumers at the zip-code level.

In the initial analysis, we select the largest fast-food chain that engaged in a geo-fencing campaign through the advertising agency from which we sourced our data. Our results show that competition only negatively affects the advertising performance at the click-through stage but not in the conversion stage. In contrast, distance has negative impacts on consumer decision to convert but not to click. Quantitatively, adding one more competitor into a 5-mile radius area of the advertisers' location decreases the click-through rate by 7.5% while the conversion rate drops by 33.2% if the advertiser's store is one more mile away.

Our results have implications for both research and practice. For research, our analysis establishes the importance of local competition in advertising performance, and as a moderator for distance. We develop an index of micro-competition (in the geo-fence area) and demonstrate its relevance as a driver of location-based advertising performance. As it turns out, competition is more important than distance at the click-through stage, but distance is more important at the time of conversion. For practitioners, our results point to the importance of both distance and competition (as well as their interaction) in designing geo-fence campaigns and assessing their performance. The findings suggest that the decision of whether or not to target a particular device should account for both distance and competition. Given that the pricing model here is "pay for impression," our results on the impact of distance and competition on advertising performance can be used to optimize the ROI from geo-fencing campaigns. Further, the results also provide guidance on the design of the specific offer, by suggesting how the advertiser could "sweeten" the deals to offset the negative effects of distance and competition. Our results also shed light on the impact of user, device and app characteristics on advertising performance.

In sum, the purpose of these investigations is to provide a comprehensive understanding of how social or spatial proximity information shape consumers' choices. In the following sections, I review related literature, discuss data, specify model development, and describe research design for each of these three studies. I start with the first study of social proximity in Chapter 2, and discuss the other two studies focusing on spatial proximity in Chapter 3 and 4, respectively. Lastly, the three chapters are concluded.

CHAPTER 2

Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Community

Literature Review

This chapter draws from two main streams of work: literature examining word-of-mouth (WOM) and observational learning (OL) effects, and a second stream focused on studying influence in social networks. The first stream consists of studies that look at how individuals make decisions based on aggregate information on the preferences and actions of other peer customers — which we collectively call popularity influence. The second stream of literature examines the role that social network ties play on individual consumption decisions, what we call proximity influence. Below, we provide a brief review of the prior work that informs our analysis of each type of influence, starting with popularity influence.

Popularity Influence

It has long been recognized that consumers tend to be influenced by social interactions with other consumers, even without knowing them or their consumption intent. As noted by Chen et al. (2011) there are two distinct types of social interactions mediated by arms-length interaction and information exchange between consumers. The first type of social interaction hinges on consumer preferences and opinions, and has been labeled word of mouth (WOM) in the marketing literature, going back to Arndt (1967). The second type of social interaction is driven by the actions and decisions of other consumers, and is termed observational learning in the psychology and economics literatures (Bandura 1971, Bikhchandani et al. 1998). The importance of these types of social interactions has grown in the online arena, and has been the subject of considerable research interest, as we briefly summarize below.

Starting with research on online WOM, studies have examined the impact of both the volume (amount of information) and valence (net positive or negative opinion) of WOM in product review and reputation systems. The general conclusion is that volume and valence of WOM both affect product sales, though in some contexts valence is more important than volume (e.g., Mizerski 1982, Chevalier and

Mayzlin 2006), while in others volume matters relatively more (e.g., Liu 2006) due to increased awareness and number of informed consumers in the marketplace. Other research has also examined the impact of product, review, and reviewer characteristics, and a sampling of the interesting findings include: online reviews are more important for niche as opposed to popular books (Chen et al. 2006); negative reviews are more influential than positive reviews (Chevalier and Mayzlin 2006); featured reviews are more influential and non-featured reviews (Forman et al. 2008); and consumers not only use summary statistics and star ratings but also pay attention to the actual text of the reviews (Ghose and Ipeirotis 2011).

Observational learning is the process by which consumers make decisions based on aggregate consumption statistics of prior users. Whether or not the knowledge of aggregate consumption decisions has an effect on subsequent individual consumption has been examined in prior work in the context of books (Sorenson 2007), software adoption (Duan et al. 2009) and online music (Salganik et al. 2006). More recently, Chen et al. (2011) look at the effect of observational learning (OL) in the presence of word of mouth (WOM) effects, based on Amazon.com data, and find that not only do OL and WOM individually drive purchase decisions, but the interaction between the two processes is significant as well.

In our setting, the number of favorites for a song indicates how many users have favorited a song, so it is a measure of the volume of WOM. However, it is not known how many users listened to the song but did not favorite it, so they only have partial information on the valence of WOM. Further, in the absence of listening statistics, comparing the number of favorites across songs is an imperfect signal of which songs were listened to more than others[5] — which has the flavor of OL. We can conclude that the number of favorites is a hybrid of WOM and OL, and conveys both volume and valence, though neither perfectly. Despite its limitations, such a metric of social interaction is increasingly prevalent in online social media, most notably on Facebook and Twitter. Prior research has investigated the correlation between this type of social interaction and product sales and product quality. Specifically, Lee and Lee (2010) and Li and Wu (2013) find a positive impact of Facebook likes on the sale of Groupon vouchers. Moreover, Schöndienst

[5] The number of favorites for a song is a lower bound on the number of unique listens of the song.

at el. (2012) and Wang and Chang (2013) show that total number of likes result in a higher level of perceived product quality. We add to this literature by examining the relationship between this "liking" information and consumption choices in an online music community.

Proximity Influence

Social network influence is due to social proximity (contact and communication) between social network "friends." Brown and Reingen (1987) was one of the first studies to look at these "micro-level" interactions and how information spreads over ties in a social network in the offline world. Valente (1994) studies so called "relational models of diffusion," and discusses the role of specific types of people as network neighbors, arguing that an "individual's direct contacts influence his or her decision to adopt or not adopt an innovation." Factors such as opinion leadership (Katz and Lazarsfeld 1995) and the strength of ties (Granovetter 1973) are also related to influence and adoption.

When studying how social proximity affects actors' behaviors, a key challenge is to be able to separate social influence and homophily, where the latter refers to social correlation in actions due to the fact that people tend to befriend others who have similar tastes and preferences (e.g., Manski 1993). It is a challenge to distinguish real social network influence from correlated effects in that as observers, we do not know if two individuals who are socially tied to one another make the same adoption decision because they have the same taste, or because they were exposed to the same external "shock" at the same time (e.g. an advertisement), or because one influenced the other. Without knowing the social network structure, this reflection problem — where we cannot separate out the effect of the individual on the group, from the effect of the group on the individual — does hinder the identification of the endogenous effects. Fortunately, we have data on the underlying social network structure, and highly granular data on music consumption and favoriting behavior, which helps mitigate the identification issues stemming from the reflection problem. Such data are not often available in many settings.

There have been a variety of methods applied to find evidence of social network influence. Ma et al. (2010) construct a hierarchical Bayesian model to study the effects of peer influence and homophily on both the timing and choice of consumer purchases within a social network. Aral and Walker (2011) design

a randomization experiment on Facebook for quantifying social network influence. Tucker (2008), de Matos et al. (2014) and Lu et al. (2012) apply the intransitive triads instrumental variable approach to separate social influence from homophily. More recently, Belo and Ferreira (2013) use a randomization approach via the shuffle test of Anagnostopoulos et al (2008). By randomizing the timing of individuals' actions, they conclude that social network influence has both positive and negative effects on the diffusion of telecom-related products.

The method that we find most useful here is the one by Aral et al. (2009), who develop a propensity-score matching estimation framework to separate social influence from homophily. Briefly, they examine adoption of a mobile service application in an instant messaging social network. The key issue that motivates their analysis is that correlated behavior in product adoption, in the form of either assortative mixing (adopters tend to have adopter friends) or temporal clustering (a user adopts soon after a friend adopts), could be driven by both influence and homophily. Recall, peer-to-peer *influence* refers to the process by which a user causes their network friends to make similar choices, whereas *homophily* is the process by which similarities across network neighbors results in correlated choices — which could mimic contagion without any causal influence. As Aral et al. (2009) explain, homophily causes a selection bias because treatments are not randomly assigned — adopters are more likely to be treated because of similarity with their network neighbors. They show that propensity score matching helps to overcome this selection bias by linking up observations across the treatment and control groups with the same likelihood of treatment. We adopt a similar matched sample approach to identify proximity influence, and use probit and hazard models to estimate the magnitude of the influence.

Social Influence in Music Industry

For the reasons mentioned in the Introduction, there is great emerging interest in the role of IT-enabled social influence in the music industry. New music arrives to the marketplace at a growing pace and the growing Long Tail nature of the music market (i.e., increased consumption of niche music relative to mainstream music) is increasing the importance of social media in the process of music discovery and consumption. Accordingly, a number of studies have recently examined the impact of social media on music

consumption. For example, Dewan and Ramaprasad (2012) study the impact of music blogging on online sampling, and find that observational learning effects are stronger in the tail relative to the body of music sales distribution. Dhar and Chang (2009) find that the volume of user-generated content is predictive of music sales. Dewan and Ramaprasad (2014) study the interaction among social media (blog buzz), traditional media (radio play) and music sales and find that while blog buzz is positively related to album sales, it is negatively related to song sales, possibly due to the sales displacement effect of free online sampling. Using a randomized field experiment, Bapna and Umyarov (2015) find that peer influence exists in the diffusion of premium subscriptions in the online music community Last.fm.

In a study with similar objectives as ours, Salganik et al. (2006) looked at the impact of aggregate prior consumption decisions on the ultimate inequality and unpredictability in an artificial music market. They found that social influence due to observation of prior aggregate consumption decisions "contributes both to inequality and unpredictability in cultural markets," providing evidence that "collective behavior" plays a part in consumption decisions. The main difference between Salganik et al. (2006) and our study is that while the prior work created an artificial music market where individuals were not explicitly socially tied to one another, ours is based on real observational data from an online community where individuals are socially tied to one another. Further, we examine both popularity influence and proximity influence, whereas the prior study was restricted to just the observational learning component of popularity influence.

Data

We use a unique dataset provided by the online music community, The Hype Machine (THM). THM is the leading music blog aggregator, aggregating mp3s that are posted in their entirety on thousands of music blogs.[6] THM allows for users to create an account, stream (but not download) songs that are posted (by clicking on the "listen" link), and favorite songs and users. On October 1, 2008, THM implemented a popularity feature by adding a number next to each track indicating how many users of the site had favorited the song. While individuals could favorite songs prior to this, the number of favorites for a song was not

[6] The Hype Machine: http://hypem.com/

viewable by any other visitors to the site until the implementation of this feature. Figure 1.2 shows a screen shot of this popularity feature indicating that 242 Hype Machine members favorited the song "Heart Skipped A Beat." To measure the effect of popularity influence on music listening, we have obtained data on user behavior from before and after THM made this popularity information visible, providing an opportunity for a natural experiment.

THM also allows members to create a social network using a personal dashboard, to which they can add favorite tracks and favorite users. The act of "favoriting" a person is akin to following another user on Twitter in that it creates a unidirectional tie (as shown in Figure 1.1), which is not necessarily reciprocated. Figure 2.1 provides a screen shot of this feature, showing that this particular user has two favorite tracks and one favorite user. In order to construct the social network of users and measure proximity influence, we have obtained time-stamped data on members' user favoriting behavior.

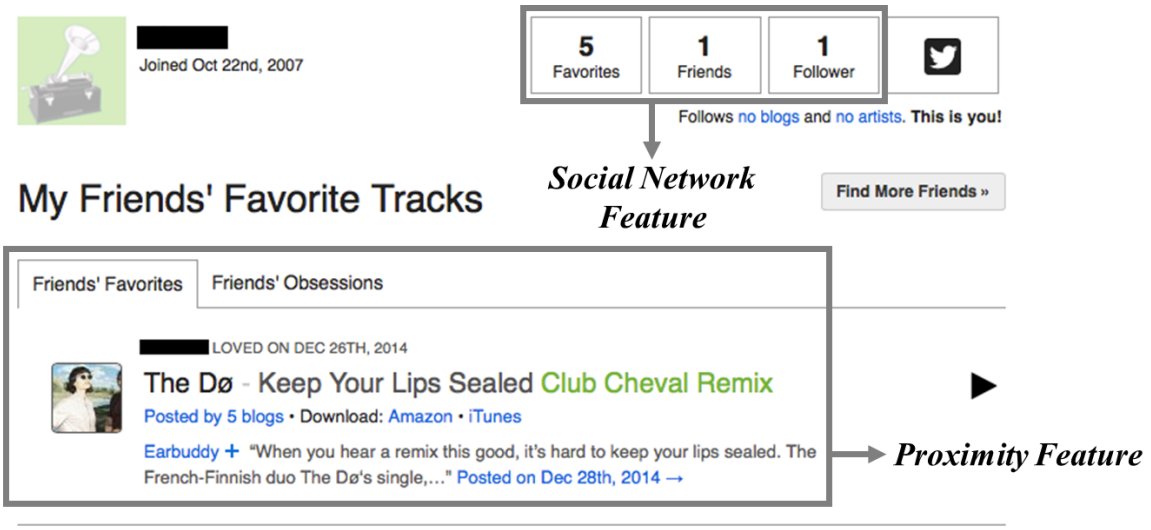


Figure 2.1. Popularity Feature on The Hype Machine (THM)

In order to estimate popularity and proximity influence, we use a detailed dataset that allows us to observe the entire history of users' listening and favoriting behaviors. THM has provided daily listen logs for September and October of 2008. These listen logs record each time any user listens to a song, along with the user ID and the details of the song, such as the artist, song title, and a posted timestamp. In addition, we have a separate dataset that contains the time-stamped log of members' favoriting of other users, which we

use to construct a member's social network, as well as song favoriting behavior. Finally, we have supplemented the data from THM with data on song characteristics collected from Amazon (Sales Rank) and the EchoNest (e.g., genre, artist popularity).

Empirical Methodology

In this section, we discuss the models we use to quantify popularity and proximity influence, including one that jointly estimates both influences in the same framework. Notation and variable descriptions are summarized in Table 2.1.

Table 2.1. Table of Variable Descriptions of Social Proximity Study

Variable	Definition
Popularity Influence	
$Listen_{jt}$	Total number of times Song *j* has been listened to at Time *t*
$PopTreatment_j$	Dummy variable; = 1 if Song *j* is treated (i.e., Song *j*'s total number of favorites are visible)
$After_t$	Dummy variable; =1 if time period *t* after the popularity treatment (i.e., after 10/1)
Proximity Influence	
$Listen_{ij}$	Dummy variable; = 1 if User *i* has listened to Song *j*
$ProxTreatment_{ij}$	Dummy variable; = 1 if User *i* has a friend who favorited Song *j* in the burn-in period
$Friends_i$	Total number of users that User *i* is following
Joint Model for Popularity and Proximity Influence	
$Listen_{gjt}$	Total number of times Song *j* has been listened to at Time *t* by Group *g*.
$PopTreatment_j$	Dummy variable; =1 if Song *j* is treated (i.e., Song *j*'s total number of favorites are visible)
$After_t$	Dummy variable; =1 if time period *t* is any day after the popularity treatment (i.e., after 10/1)
$ProxTreatment_{gj}$	Dummy variable; =1 if User *i* (in Group *g*) has a friend who favorited Song *j* in the burn-in period
Control Variables	
$PreFavorite_j$	Total number of favorites of Song *j* before the observation window of the study
$SalesRank_j$	Sales rank of Song *j* at Amazon.com
$Genre_j$	Genre of Song *j*

Model of Popularity Influence

For estimating popularity influence, we employ a difference-in-difference (DD) methodology (see, e.g., Card and Krueger 1994), exploiting the feature implementation in THM on October 1, 2008, that provided visibility of the total number of song favorites.[7] Given that the implementation of this feature is exogenous, as we discuss below, the DD model allows us to reliably measure the impact of the visibility of popularity information on music consumption. We compare a set of songs that experienced the implementation (the treatment group) to a set of songs that did not (the control group). Specifically, we define the songs posted on September 29, 2008 as the *treatment* group and the songs posted one week earlier, on September 22, 2008 as the *control* group. The latter group of songs, the group posted on September 22, 2009, was not affected by the feature implementation during the time period we examine. Figure 2.2 illustrates our DD experimental design, where T_1 = October 1, 2008, is the date of treatment (implementation of the popularity feature on THM) for the treatment group. Even though there was no such intervention for the control group, we create a dummy treatment event for the control group on T_0 = September 24, one week prior to the feature implementation. Similar to the event study literature in Finance, we use a short estimation window (+/- 1 day) to isolate the effect of the feature implementation on listening behavior.[8]

[7] Popularity information was visible to all users of the website, irrespective of whether they were registered to the site or not, and irrespective of whether they had social network friends or not.

[8] We thank an anonymous reviewer for suggesting that we look at the effect of popularity information visibility on songs earlier released earlier to the site. However, we do not find a significant popularity effect for older songs, due to the fact that such songs receive very little attention on THM, and therefore the popularity information is immaterial.

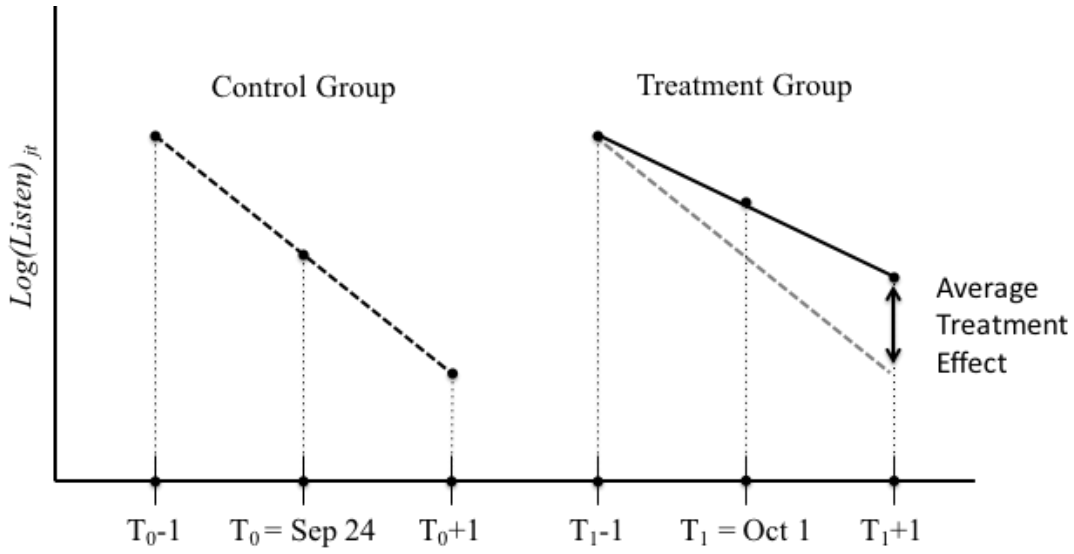


Figure 2.2. Difference-in-Difference Experimental Design for Popularity Influence

Ideally, the treatment and control groups should contain songs posted on the same date, with identical potential treatment dates. We are unable to construct coincident treatment and control groups, however, because all songs on the website were subject to the feature implementation at the same time — either all songs were treated or none were, depending on whether the date in question is after or before the date of feature implementation, respectively. It is for this reason that the treatment and control groups in our DD design include songs posted one week apart (*exactly* one week apart to avoid day-of-week differences). The time separation of the treatment and control groups is a cause for concern, however, because time shocks at different points in time could affect treatment and control groups differently, confounding the measurement of treatment effects. We believe, however, that this is not a serious concern, for the following reasons.

First, even though the samples are one week apart, the listening patterns are virtually identical, as shown in Figure 2.3. We graph the total number of listens in each hour of the pre-treatment period, T_1-1 and T_0-1 for the treatment and control groups, respectively, and show that they follow almost exactly the same pattern. This consistent pattern of listening behavior across the treatment and control dates provides us some assurance that there were no time-varying shocks that differentially affected the listening behavior of songs across the treatment and control groups. Second, the time of posting of a song on THM is

exogenous, because it is synchronized with the posting of the song on the original mp3 blog, rather than a decision made by THM. Further, the THM website did not publicize the fact that the favoriting feature was imminent, so mp3 blogs could not have anticipated the feature implementation. Third, the songs in the treatment and control groups are similar in terms of genre and popularity. To further increase the similarity of the samples, as we discuss in Section Results, we use coarsened exact matching (CEM) to match individual songs in the treatment and control groups on a one-to-one basis to make sure that the samples are balanced and the songs are similar to each other.[9] As we will see in Section 5, the results for the matched and unmatched samples are qualitatively similar. Still, we include treatment dummies in all of our DD specifications to absorb any systematic differences in listen frequency between the treatment and control groups.

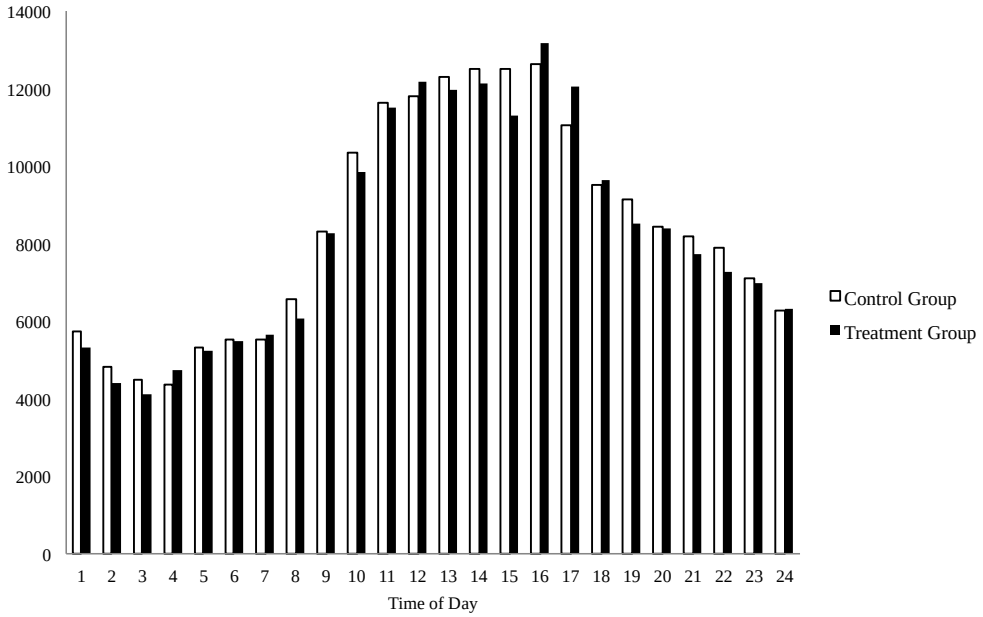
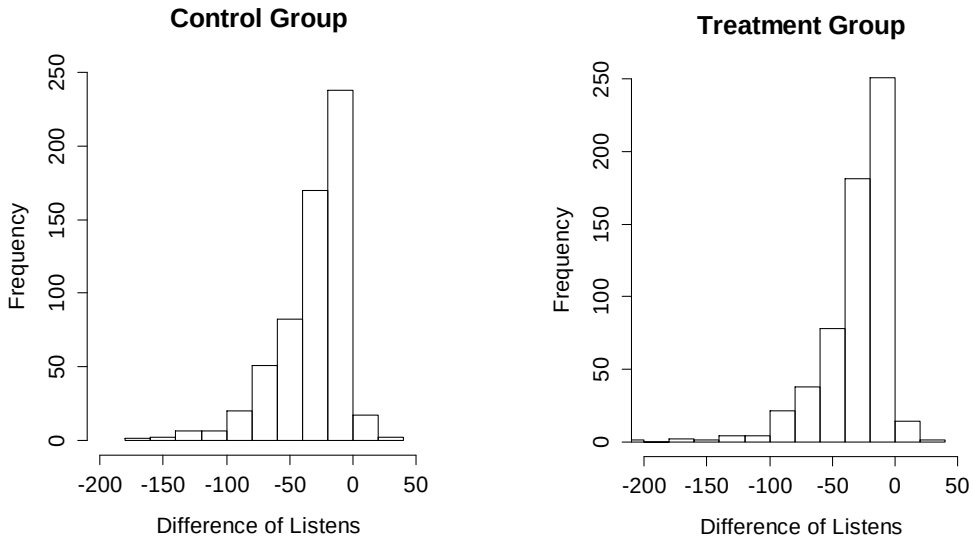


Figure 2.3. Distributions of Listens for DD Treatment and Control Subsamples

[9] Specifically, we match the two groups of songs on observable characteristics, including genre, the number of favorites prior to the feature implementation, and the Amazon Sales Rank. We employ one-to-one coarsened exact matching (CEM) in order to exactly match genres while not requiring our continuous variables to be exactly matched, but closely matched. A benefit of CEM is that the researcher can ensure balance in matching a priori through implementing bounds on the qualifications of a match for each variable that the groups are matched on. Each song in the treatment group is matched to one song in the control group, using the CEM procedure in Stata (Blackwell et al. 2009). The imbalance statistics produced by the CEM procedure indicate that the imbalance between the treatment and control groups was reduced due to matching.

Fourth, we find that our treatment and control samples satisfy the key identifying assumption of difference-in-difference estimation, which is that the treatment and control group have a common trend in the absence of treatment (Meyer 1995). (See Figure 2.2 for the importance of a common trend for being able to measure the average treatment effect.) This test is typically operationalized by comparing the trend in the dependent variable over the pre-treatment period, across the treatment and control samples (Card and Krueger 1994, Danaher et al. 2014). In our case, the general trend is one of declining listens over time, as newly posted songs lose novelty and move off the front page of the site. Because the songs are posted at different times during the posting dates (9/22 and 9/29 for the control and treatment samples, respectively) we characterize the pre-treatment trend by the difference in the average number of listens over the second 12-hour window and the first 12-hour window after posting: we expect this difference to be negative. Figure 2.4 displays the distribution of this difference measure (labeled Difference of Listens in the figure) for the control and treatment subsamples, along with a table of summary statistics and difference tests below the graphs. As shown in the figure the distributions are virtually identical, with both the difference of means t-test and the Kolmogrov-Smirnov test for equality of distributions being insignificant. This supports the assumption of a common pre-treatment trend for the treatment and control samples.

Finally, we conduct a variety of robustness checks (described in Section Results) with alternate treatment and control groups, drawn from different points in time, to show that the results are not sensitive to exactly when the songs are posted to THM. Overall, we believe that picking the treatment and control groups one week apart does not compromise the integrity of the difference-in-difference design. On the contrary, our design assures that treatment is exogenous, overcoming a major challenge in conventional difference-in-difference models. Indeed, our research design illuminates a practical approach for a quasi-experimental investigation of online feature implementations or policy changes that affect an entire community or website starting at a given point in time.



	Control Group	Treatment Group
Mean	-30.2202	-29.2315
Std. Dev	28.2601	26.8050
T test p-value	0.5294	
KS p-Value	0.5664	

Figure 2.4. Comparison of Pre-Treatment Trend for Treatment and Control Samples for the Popularity Influence Model

In order for us to measure the impact of popularity information, songs in the dataset must have had the opportunity to accumulate favorites, so we allow for an initial "burn-in period," from the time a song is posted on THM to one day prior to the treatment date. This requires us to look at the sample of songs posted two days prior to the treatment date in order to have a pre-treatment period. Then, the days T_1-1 and T_0-1 are the pre-treatment periods for the treatment and control group, respectively, while T_1+1 and T_0+1 are the corresponding post-treatment periods. Accordingly, our DD model specification is as follows, for Song *j* on Day *t*:

$$\log(Listens_{jt}) = \beta_0 + \beta_1 PopTreatment_j + \beta_2 After_t + \beta_3 \log(PreFavorites_j) + \beta_4 PopTreatment_j * After_t + \varepsilon_{jt}, \quad (1)$$

where $Listens_{jt}$ denotes the total number of listens of Song *j* on Day *t*. The regression covers the time periods running from one day before the feature implementation to one day after, for the treatment and control

groups; i.e., $t \in \{T_0 - 1, T_0 + 1, T_1 - 1, T_1 + 1\}$. $PopTreatment_j$ is a dummy variable indicating whether Song *j* is in the treatment group ($PopTreatment_j = 1$) or control group ($PopTreatment_j = 0$). $After_t$ is a dummy variable indicating whether the date *t* is the post-treatment period ($After_t = 1$) or pre-treatment period ($After_t = 0$). The control variable $PreFavorites_j$ is the number of favorites at the start of the pre-treatment period. The $PopTreatment_j$*$After_t$ interaction term characterizes the magnitude of popularity influence. We use Ordinary Least Squares (OLS) to estimate the regression.

Models for Proximity Influence

Turning to our models to measure proximity influence, we focus on how favoriting a song by a focal user impacts the listening behavior of her social ties. Following prior social network research, distilling social network influence from other drivers of correlation in behavior, such as homophily, is at the heart of our proximity influence analysis. We estimate a probit model and a hazard model, corresponding to how the probability of listening to a song, and the time to first listen, respectively, are affected by the favoriting behavior of friends in social network proximity. To conduct this analysis we follow Aral et al. (2009) and use propensity score matching (PSM) to control for potential homophily. Before specifying our proximity models, we describe our PSM procedure first.

Propensity Score Matching

For a given song, the treatment group consists of those users that have at least one friend who has favorited that song. The goal of PSM in our analyses is to match every user in the treatment group with a user in the control group (none of whose friends has favorited the song) who is homophilous to the user in the treatment group in terms of tastes, calculated based on the users' observable characteristics, and number of friends. In our case, we do not have data on consumer demographics and other characteristics, but we do observe perhaps the most relevant characteristic of all — actual song listening behavior. Much as a recommender system finds nearest neighbors (e.g., Adomavicius and Tuzhilin 2005), we find matches between the treatment and control group based on the relative song listening profiles of users. From the data on listening history of users, over the four-week period September 1-28, 2008, we construct a profile of each user on

THM. These profiles are constructed using data on over 80,000 songs, for which we collected supplemental data on genre and various measures of artist popularity from The Echo Nest.[10]

For each user-song pair, we constructed a weight based on the number of times the user listened to that particular song as a fraction of the total number of overall listens for that user. We then created a weighted average of each song characteristic in a vector of 28 song characteristics based on all of the songs that the user had listened to. The result of this allowed us to summarize a user's profile by a series of numbers (the weighted average of a song characteristic) each representing the user's taste towards a specific music characteristic. We then used these profiles to match each user in the treatment group to a user in the control group, using PSM as follows. Each song in our treatment and control groups are assigned a positive probability of being in the treatment group based on a logit model incorporating the characteristics of the users as characterized by their listening history as well the number of friends they have. To ensure overlap in the treatment group and control group, we constrained the group of matched observations to be within 0.1 propensity score of each other (i.e., caliper = 0.1). After finding a match for each user in the treatment group, we examined the distribution of propensity scores to ensure similarity between the treatment and control groups as advised by Lechner (2002). Looking at Figure 2.5, we see that the distributions of the propensity scores in both the treatment group and control group are almost identical to one another. Specifically, the boxplot shows that the two groups match on the minimum, maximum, and median as well as the first and third quartiles shown in the figure. The details of the PSM procedure are provided in the Appendix.

As a robustness check we also implemented a matching procedure at the user-song level, using Euclidean distance matching (EDM). For each song, this procedure matches each user in the treatment group (i.e., users that have at least one friend who has favorite the song) with a similar user (based on song listening profiles) in the control group who has a high likelihood of having a friend that might have favorited

[10] The Echo Nest has various measures of artist popularity that we collected and used to construct user profiles: artist hotttnesss, artist familiarity, and artist discovery.

the song. In this case, matching was based on minimizing the Euclidean distance between the song profile and friends' song listening profiles. The details of this procedure are also described in the Appendix. The tradeoff between PSM and EDM is that while PSM maximizes the control for homophily (by matching on user characteristics), EDM maximizes the likelihood of treatment (i.e., having a friend that has favorited the song, by matching at the user-song level) for the matching member in the control group. We estimate proximity influence by using both PSM and EDM, and comparing each to random matching, as we discuss in Section Results.

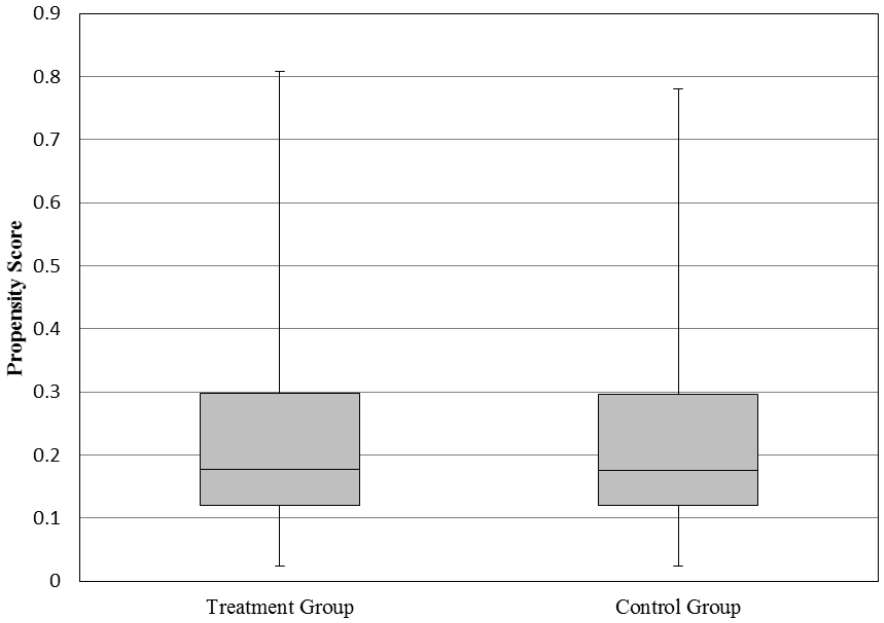


Figure 2.5. Distribution of Propensity Scores after Matching

Probit Model of Proximity Influence

To examine the impact of proximity influence on the likelihood of listening to a song, while controlling for other song characteristics, we implement a binary probit model. To do this, we look at songs posted on 9/22 and allow a 48-hour burn-in period after the time of song posting so that songs can acquire favorites. After this burn-in period, we track the users' listening choices for the following seven days in order to estimate the probit model. That is, we use a 2-day burn-in period followed a 7-day observation window for all of our proximity influence analyses. Using the matched treatment (*ProxTreatment = 1)* and control groups

(*ProxTreatment = 0),* under random matching and either PSM or EDM, combined with the song characteristics data, we estimate the following probit model:

$$\Pr(Listen_{ij} = 1) = \beta_0 + \beta_1 ProxTreatment_{ij} + \beta_2 Log(PreFavorites_j) + Genre_j + \varepsilon_{ij}, \quad (2)$$

where $Listen_{ij}$ is a binary outcome indicating whether User *i* listened to Song *j* or not. $ProxTreatment_{ij}$ is a dummy variable to capture the treatment of proximity influence; i.e., $ProxTreatment_{ij} = 1$ indicates that User *i* has at least one friend who has favorited Song *j*, while $ProxTreatment_{ij} = 0$ indicates that User *i* has no friend who has favorite Song *j*. We also include song-level controls $PreFavorites_j$ (for overall popularity of the song on THM) and $Genre_j$. β_1 is the coefficient of interest and it captures the impact of proximity influence on a focal user's listen decision. To isolate proximity influence from homophily we compare the estimate of β_1 under random matching with both PSM and EDM.

Hazard Model of Proximity Influence

Lastly, we investigate proximity influence by looking at the time to a user's first listen to a song. We apply a hazard model to estimate the impact of proximity influence on the duration of time before User *i* first listens to Song *j*. Similar to the probit model above, the hazard model compares the matched treatment and control groups. Similar to the probit model described in Section 4.2.2., we again use a seven-day observation window after the 48-hour burn in period after Song *j* was posted. We track User *i* until she listens to Song *j*. If User *i* did not listen to Song *j* within the observation window of seven days, we right-censor the observation. Specifically, the hazard rate, λ_{ij}, follows an exponential distribution,[11] and is related to the covariates of interest using the following simple parametric model:

$$\text{Log}(\lambda_{ij}) = \beta_0 + \beta_1 ProxTreatment_{ij} + \beta_2 Log(PreFavorites_j) + Genre_j + \varepsilon_{ij}, \quad (3)$$

where λ_{ij} is the hazard rate defined by whether and when *User i* listened to *Song j*. Similarly, $ProxTreatment_{ij}$ is a dummy variable coding the treatment and control groups, and β_1 is our coefficient

[11] Our results are robust to the choice of Weibull and Gompertz distributions for the hazard rate.

of interest. $PreFavorites_j$ and $Genre_j$ are song-level controls included in the regression. As before, we estimate the hazard model under random matching, compared with both PSM and EDM.

Combined Model for Popularity and Proximity Influence

To jointly estimate popularity and proximity influence, we need a model that can simultaneously capture the impact of the visibility of popularity information and friends' favorites on user listen decisions. We extend the difference-in-difference (DD) model of Section 4.1 to a difference-in-difference-in-difference (DDD) specification by adding a proximity influence treatment. That is, the DDD model is a two-dimensional treatment model, including both popularity treatment (represented by the *PopTreatment* indicator variable) and proximity influence treatment (represented by the *ProxTreatment* dummy variable). The third dimension in the DDD model is represented by the dummy variable $After_t$, which indicates whether the time period *t* in question is the pre-treatment period (for popularity influence) or the post-treatment period.

The DDD design has the songs posted on September 29, 2008 as the popularity-treatment group (*PopTreatment* = 1) and the songs posted on September 22, 2008 as the popularity-control group (*PopTreatment* = 0). On the other dimension, *ProxTreatment* divides users into two groups, where *ProxTreatment = 1* indicates users in the proximity-treatment group that have at least one friend who has favorited Song *j*, and *ProxTreatment* = 0 indicates users in the proximity-control group that do not have any friend who has favorited Song *j*. Users in the two proximity groups are matched by both random matching and propensity score matching (we do not use EDM here, as we discuss below). Accordingly, our DDD model specification is as follows:

$$\begin{aligned}\mathrm{Log}\left(Listens_{gjt}\right) = {} & \beta_0 + \beta_1 PopTreatment_j + \beta_2 After_t + \beta_3 ProxTreatment_{gj} + \\ & \beta_4 PopTreatment_j * After_t + \beta_5 PopTreatment_j * ProxTreatment_{gj} + \\ & \beta_6 After_t * ProxTreatment_{gj} + \beta_7 PopTreatment_j * After_t * ProxTreatment_{gj} + \\ & \beta_8 Log\left(PreFavorites_j\right) + \beta_9 Genre_j + \varepsilon_{ijt}, \end{aligned} \quad (4)$$

where g is the index of proximity treatment (0 or 1), and $Listen_{gjt}$ denotes the total number of listens of proximity treatment Type g of Song j at Time t, where $t \in \{T_0 - 1, T_0 + 1, T_1 - 1, T_1 + 1\}$. $PopTreatment_j$, and $ProxTreatment_{gj}$ are dummy variables to whether an observation is in the treatment or control group for popularity and proximity treatment, respectively. The dummy $After_t$ identifies whether the date t corresponds to the pre-treatment or post-treatment for popularity. β_3 represents the magnitude of proximity influence, β_5 captures the magnitude of popularity impact, and the coefficient on the three-way interaction term β_7 characterizes the nature of interaction between popularity and proximity influence ($\beta_7 > 0$ would indicate that the interaction is complementary, while $\beta_7 < 0$ would indicate that the interaction is one of substitutes). Equation (4) is estimated using OLS.

Descriptive Statistics

We start by providing descriptive statistics and correlations for the dataset used for estimating popularity influence (Table 2.2) followed by those for proximity influence (Table 2.3). The key dependent variable in estimating popularity influence is the total number of times a song is listened to on a given day ($Listens_{jt}$). Table 2.2a summarizes songs in the treatment and control group on the pre-treatment and post-treatment days. Overall, there are roughly 600 songs in both the treatment and control groups. On average, there were 47.54 listens per song per day on THM, with a standard deviation of 171.56. Some songs posted on THM did not get any listens, but the maximum number of listens in a day for a song was 3836. While users listen to a variety of songs, they appear to be more selective in their favoriting behavior ($PreFavorites_j$). On average, songs receive 1.33 favorites per day with a standard deviation of 3.46. Again, some songs do not receive any favorites, while the maximum number of favorites a given song in our dataset was 57. The pairwise correlations (Table 2.2b) indicate that listening and favoriting are significantly and strongly correlated with one another (0.71, $p < 0.01$) and Amazon sales rank of a song is negatively correlated with both the number of listens and number of favorites. This is expected as a higher sales rank corresponds to less popular songs.

Table 2.2a. Descriptive Statistics for Popularity Influence

Variable	N	Mean	Std. Dev.	Min	Max
$Listen_{jt}$	2,382	47.5369	171.5553	0	3836
$PreFavorites_j$	2,382	1.3283	3.4622	0	57
$SalesRank_j$	2,382	2,983,163	3,643,284	605	6,856,013

Table 2.2b. Correlations among Variables for Popularity Influence

	$Listen_{jt}$	$PreFavorites_{jt}$	$SalesRank_j$
$Listen_{jt}$	1		
$PreFavorites_j$	0.706^{***} (0.000)	1	
SalesRank	-0.129^{***} (0.000)	-0.165^{***} (0.000)	1

Notes: Standard errors are in parentheses. * indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$.

We now turn to the summary statistics for the relevant variables for the proximity influence analysis (Table 2.3a), which summarize data for a weeklong observation window from September 22, 2008 until September 29, 2008. Overall, we have a pool of over 800 users and roughly 200 songs to create the user-song pairs used in this analysis. From Table 2.3a, we see that the average likelihood of a user listening to an individual song is quite low — 0.0057 (0.57%), with a standard deviation of 0.0754. The likelihood that a user's friend has favorited a given song is even lower (as expected), 0.0015 (0.15%) with a standard deviation of 0.0392. These summary statistics demonstrate the sparseness of the data, making it challenging to estimate proximity influence. The number of total favorites of the average song is approximately 4.71 with a standard deviation of 9.30, and the Amazon Sales Rank is 3.9 million with a standard deviation of 4.3 million.[12] Turning to the pairwise correlations (Table 2.3b), we see the expected correlations — a positive and significant (though low in magnitude) correlation between *ProxTreatment* and *Listen*, as well as between *PreFavorites* and *Listen*. We also see a negative correlation, as expected, between *SalesRank*

[12] Recall that the Amazon Sales Rank information is collected in 2014, and thus represents a measure of quality as observed in the long-term. This explains the large values of sales rank, though there is still variation within our dataset.

and *ProxTreatment*, *PreFavorites*, and *Listen*. Generally, the correlations are relatively low and again reflect the sparseness of social correlations.

Table 2.3a. Descriptive Statistics for Proximity Influence

Variable	N	Mean	Std. Dev.	Min	Max
$Listen_{ij}$	159,583	0.0057	0.0754	0	1
$ProxTreatment_{ij}$	159,583	0.0015	0.0392	0	1
$PreFavorites_j$	159,583	4.7096	9.2960	0	88
$SalesRank_j$	159,583	3,925,858	4,297,366	1233	6,508,732
$OutDegree_i$	159,583	2.7868	3.6763	1	62

Table 2.3b. Correlations among Variables for Proximity Influence

	$Listen_{ij}$	$ProxTreat_{ij}$	$PreFavorites_j$	$SalesRank_j$	$OutDegree_i$
$Listen_{ij}$	1				
$ProxTreatment_{ij}$	0.027^{***} (0.000)	1			
$PreFavorites_j$	0.100^{***} (0.000)	0.057^{***} (0.000)	1		
$SalesRank_j$	-0.021^{***} (0.000)	-0.002 (0.503)	-0.223^{***} (0.000)	1	
$OutDegree_i$	-0.008^{***} (0.001)	0.041^{***} (0.000)	-0.001 (0.618)	0.001 (0.745)	1

Notes: Standard errors are in parentheses. * indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$

Results

We present our results in the following order: (i) popularity influence, (ii) proximity influence, and finally (iii) joint estimation of popularity and proximity influence.

Popularity Influence

Our results for the DD model (Equation 1) of popularity influence are presented in Table 2.4, where the dependent variable is *Log(Listens)*. Recall, the treatment sample is the set of songs posted on September 29, 2008, while the control sample consists of the songs posted one week earlier, on September 22, 2008. In Model 1, we consider all of the songs in both samples, and do not restrict the control group to match the

songs in the treatment group. In Model 2, the songs in the control sample are matched with songs in the treatment group, following the matching procedure described in Footnote 7. Model 3 is restricted to "isolates," that is, users who are not connected to others in the social network. This is an interesting group of users to study because it is subject solely to popularity influence and no proximity influence. Finally, in Model 4 the dependent variable is the number of unique listens. This case is interesting to consider because it is possible that popularity influence is restricted to the first listen of a song, rather than subsequent repeat listens of the same song. All models include genre fixed effects and logarithm of Amazon Sales Rank as control variables. We discuss the results of all four models together.

Table 2.4. Difference-in-Difference Results for Popularity Influence

	1 Unmatched	2 Matched	3 Matched, Isolates	4 Matched, Unique Listens
Constant	5.099*** (0.143)	5.338*** (0.197)	2.010*** (0.192)	1.929*** (0.182)
$PopTreatment_j$	-0.078 (0.054)	-0.107* (0.065)	-0.219*** (0.067)	-0.193*** (0.062)
$After_t$	-2.199*** (0.054)	-2.339*** (0.064)	-0.729*** (0.067)	-0.771*** (0.061)
$PopTreatment_j$*$After_t$	0.127* (0.076)	0.180** (0.090)	0.301*** (0.094)	0.292*** (0.087)
$Log(PreFavorites)_j$	0.943*** (0.032)	0.707*** (0.050)	0.476*** (0.045)	0.421*** (0.042)
Adjusted R^2	0.648	0.679	0.319	0.326
N	2382	1448	752	824

Notes: Models 1 does not match the control sample to the treatment sample of songs, while Model 2 matches the samples, as explained in Section 4. Model 3 uses the number of unique listens as the dependent variable. All models include genre fixed effects and log of Amazon Sales Rank as additional control variables. Standard errors are in parentheses. * indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$.

The *PopTreatment* variable has a negative sign, and it is significant in Models 2-4, indicating that the songs in the treatment have fewer listens, on average, than the songs in the control sample, all else being equal. Therefore it is a good idea to include the *PopTreatment* dummy variable to absorb such systematic differences across the two samples. The *After* variable is negative and significant, in all models, due to the

tendency of the number of listens to naturally decay over time. This may be because the novelty of newly introduced songs wears off over time, or the fact that songs get "buried" below newer songs added to the site. *Log(PreFavorites)* has the expected positive sign, reflecting the fact that more popular songs (as captured by the favoriting behavior of users on THM) get more listens on average.

The key variable of interest is the interaction term *PopTreatment*After*, which captures the average effect of the treatment on the number of listens, after the availability of song popularity information on the website. This interaction term is estimated to be positive, with varying degrees of significance in the three models. Interestingly, the magnitude and significance of the interaction term are highest in Models 3 and 4, consistent with the notion that popularity influence is strongest for isolates (users with no friends) and for the first listen of a song, as opposed to repeat listens. Overall, we find strong evidence of a causal link between the disclosure of song popularity information (in the form of the number of song favorites) and the number of user listens.

We can quantify the economic significance of popularity influence as follows. The estimate of the interaction term *PopTreatment*After* is 0.18 in our main baseline model, which is Model 2 in Table 2.4. Since the dependent variable is the logarithm of *Listens*, the magnitude of the interaction term implies that the availability of popularity information, after the corresponding feature implementation, increases the total listens of the average song by approximately 19.7% ($exp(0.18) = 1.197$). Given that the mean number of listens of the average song on the post-treatment date October 2 is 9.511, this implies that the availability of popularity information increases total listens of the average song by almost two. Thus, popularity influence is not only statistically significant, it is an economically significant effect as well.

For additional robustness, Table 2.5 considers alternative definitions of the treatment versus control samples, to make sure that the results are not driven by the specific dates we picked in our baseline results. In Model 1 the control group is taken to be songs posted *two* weeks prior to the treatment group; i.e., September 15 versus September 29 (in Table 2.4 the control group corresponds to songs posted *one* week prior). The *PopTreatment*After* term is positive and significant, consistent with our baseline results of Table 2.4. In Model 2 the treatment and control group are both after popularity information is available, so

as expected, the interaction term is not significant. Model 3 has the same control group as Model 1, but the treatment group is moved one week later to October 6, and we can see that the qualitative nature of the results are unchanged. In Model 4 both treatment and control are before the popularity feature implementation, so as expected the interaction term is insignificant. Finally, in Model 5 both samples are drawn after the feature implementation, and again the interaction term is not significant, as expected. Overall, we can conclude that the DD results are robust, and there are no "secular" effects in different weeks – validating our DD research design with the treatment and control samples drawn from neighboring weeks.

Table 2.5. Robustness of Popularity Influence Results to Alternative Scenarios

	1 09/29 vs. 09/15	2 09/29 vs. 10/06	3 10/06 vs. 09/22	4 09/22 vs. 09/15	5 10/06 vs. 10/13
Constant	1.573*** (0.041)	1.830*** (0.050)	1.468*** (0.039)	1.617*** (0.051)	1.749*** (0.043)
$PopTreatment_j$	-0.254*** (0.053)	-0.499*** (0.058)	0.354*** (0.063)	-0.155** (0.070)	0.143** (0.058)
$After_t$	-0.905*** (0.055)	-0.696*** (0.068)	-0.905*** (0.053)	-0.944*** (0.074)	-0.683*** (0.056)
$PopTreatment_j$*$After_t$	0.245*** (0.075)	-0.041 (0.082)	0.206** (0.089)	-0.026 (0.101)	0.009 (0.082)
$Log(PreFavorites)_j$	1.243*** (0.030)	1.208*** (0.029)	1.225*** (0.039)	1.131*** (0.078)	1.086*** (0.026)
Adjusted R^2	0.366	0.446	0.433	0.364	0.487
N	3438	2744	2608	3302	1968

Note: Each date corresponds to when the songs were added to THM. In each column the sample corresponding to the first date is taken to be the treatment group, while the sample for the second date is the control group. Standard errors are in parentheses. * indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$.

In Table 2.6 we examine the differential impact of song popularity information for broad versus narrow appeal songs, motivated by the work of Tucker and Zhang (2011). We characterize broad versus narrow appeal using two approaches. The first is based on Amazon Sales Rank (ASR) of the song, wherein songs with ASR < 130,000 (the top 20th percentile) are considered broad appeal, while songs with higher than 130,000 sales rank are considered narrow appeal. For robustness we also consider subsamples using 60,000 (the top 15th percentile) as the cut-off. Our second approach for distinguishing between-broad and narrow-appeal music is based on genre. Specifically, we include Pop, Rap, Hip-Hop, Dance and R&B in

the broad-appeal category, while the niche genres include Folk, Country, Classical and the various types of Rock music. Looking at the results in Table 2.6, we find that the sign and significance of the control variables are consistent with our baseline results of Table 2.4. As for the key *PopTreatment*After* interaction term we find that they are positive in sign, but significant only for the narrow-appeal song samples. This is consistent with the theory and findings of Tucker and Zhang (2011), in that popularity influence is more important for narrow-appeal music as compared to broad-appeal music.

Table 2.6. Examining Differential Popularity Influence

	1 Amazon Sales Rank		2 Amazon Sales Rank		3 Genre	
	<130,000	>130,000	< 60,000	> 60,000	Mainstream	Niche
Constant	4.723*** (0.851)	5.400*** (0.280)	4.338*** (1.238)	5.394*** (0.247)	4.702*** (0.313)	5.801 (0.248)
$PopTreatment_j$	0.064 (0.227)	-0.129** (0.066)	0.116 (0.302)	-0.120* (0.065)	0.010 (0.115)	-0.152* (0.078)
$After_t$	-1.828 (0.219)	-2.409*** (0.065)	-1.542*** (0.288)	-2.400*** (0.065)	-2.257*** (0.118)	-2.370*** (0.076)
$PopTreatment_j$*$After_t$	0.197 (0.311)	0.180** (0.092)	-0.099 (0.407)	0.202** (0.092)	0.080 (0.161)	0.220** (0.109)
$Log(PreFavorites)_j$	0.621 (0.134)	0.694*** (0.055)	0.771*** (0.163)	0.689*** (0.054)	0.753*** (0.079)	0.667*** (0.065)
$Log(SalesRank)_j$	-0.091 (0.076)	-0.156*** (0.019)	-0.083 (0.120)	-0.154*** (0.016)	-0.100*** (0.021)	-0.182*** (0.016)
Adjusted R^2	0.529	0.694	0.485	0.068	0.684	0.676
N	174	1274	104	1344	432	1016

Notes: All regressions for matched treatment and control samples, as explained in Section 4. The mainstream genres on The Hype Machine include Pop, Rap&Hip-Hop, Dance, and R&B, while the niche genres on The Hype Machine include the Rock genres, Folk, Country, and Classical. Standard errors are in parentheses. * indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$. The model estimated here is Model 2 from Table 2.3 (matched treatment and control samples), the main model we will use throughout our remaining analyses.

To get a sense for the economic significance of popularity influence for narrow-appeal music, note that the coefficient estimate for the *PopTreatment*After* interaction term is 0.18 in Model 1. That translates to an increase in listens by 19.7% per song per day. Given that the average number of listens per song per day for this subsample on the post-treatment date is 5.695, this means that the availability of popularity

information increases average listens per song per day to 6.817. With an average of 800 songs posted per day, this translates into an increase of almost 1,000 listens, on average.

Proximity Influence

We now turn to our results for proximity influence, where the analyses are conducted at the user-song level. As a preliminary step, we first compare the number of users in the treated group who listen to the song (n_+) and the number of users in the control group who listen to the song (n_-), based on the matched sample adoption ratio analysis of Aral et al. (2009). If having a friend who has favorited a song results in more listens, then the ratio n_+/n_- would be greater than one. Further, the magnitude of n_+/n_- ratio should reduce when going from random matching to propensity score matching, due to the fact that random matching reflects both homophily and proximity influence, whereas PSM eliminates the effect of homophily (Aral et al. 2009). We use the treatment and control groups as constructed by the PSM method described in the prior section, and compare the n_+/n_- ratio of the PSM-matched sample to the ratio of the random-matched sample.

The results of the (n_+/n_-) analysis are presented in Tables 2.7a and 2.7b. We consider random matching of users in the two groups as well as propensity score matching, wherein the control group is restricted to users who have a similar propensity to have a friend who had favorited the song as in the treatment group. We also do the same for Euclidean distance matching, which as described in the previous section, matches users on the propensity to be treated. Table 2.7a shows the results comparing random matching to propensity score matching. We find that n_+/n_- is equal to 10.67 under random matching, and declines to 4.57 under propensity score matching. The value of the ratio goes down because propensity score matching removes the homophily effect. Yet, the ratio is greater than one, suggesting the presence of proximity influence in this setting. Table 2.7b presents the results comparing random matching to Euclidean distance matching. Similarly, we find that n_+/n_- is equal to 11.67 under random matching, and declines to 5.83 under propensity score matching.

Table 2.7a. Estimating Proximity Influence Using Listen Ratios (PSM)

	Random Matching	*Propensity Score Matching*
n_+/n_-	32/3=10.67	32/7=4.57

Table 2.7b. Estimating Proximity Influence Using Listen Ratios (Euclidean Distance Matching)

	Random Matching	*Euclidean Distance Matching*
n_+/n_-	35/3=11.67	35/6=5.83

Table 2.8a. Estimating Proximity Influence (PSM)

	Probit Model		Hazard Model	
	Random Matching	*Propensity Score Matching*	*Random Matching*	*Propensity Score Matching*
Constant	-2.584*** (0.284)	-2.148*** (0.212)	-6.788*** (0.632)	-5.884*** (0.446)
ProxTreatment	1.208*** (0.259)	0.819*** (0.199)	2.464*** (0.604)	1.601*** (0.417)
Log(PreFavorites)$_j$	0.172** (0.066)	0.147** (0.061)	0.285** (0.116)	0.266** (0.110)
Genre Fixed Effects	Yes	Yes	Yes	Yes
LR Chi^2	53.61***	45.32***	65.18***	55.21***
Pseudo R^2	0.221	0.173	--	--
N	446	446	446	446

Notes: These regressions have 7-day observation window after 24 hour burn-in period. Standard errors are in parentheses.* indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$.

Table 2.8a and 2.8b, respectively, present the results for the probit and hazard models, comparing PSM (EDM, respectively) with random matching. In both tables we find that the *ProxTreatment* variable is positive and significant in both the probit and hazard models. Further, in each case, the magnitude of the coefficient goes down under PSM or EDM as compared to random matching. Specifically, in Table 2.8a, the coefficient of *ProxTreatment* in the probit model goes down from 1.208 under random matching to 0.819 under PSM. In Table 2.8b, the coefficient goes down from 1.061 under random matching to 0.963

under EDM. Since the most important consideration when estimating proximity influence is to be able to isolate it from homophily, we feel that PSM provides a more conservative estimation, since the coefficient on ProxTreatment declines by a larger amount. Accordingly, we use PSM as our primary matching method to account for homophily, and use it in favor of EDM in the joint model below as well. Under PSM (Table 2.8a), the marginal elasticities (or the percent increase in probabilities of listen, conditional on treatment) corresponding to the *ProxTreatment* estimates are 12% and 10.2% for random matching and PSM, respectively. This means that homophily and proximity influence together (under random matching) account for a 12.0% increase in the probability of listening to a new song, which can be separated into the components: 10.2% for proximity influence and 1.8% for homophily.

Table 2.8b. Estimating Proximity Influence (EDM)

	Probit Model		Hazard Model	
	Random Matching	*Euclidean Distance Matching*	*Random Matching*	*Euclidean Distance Matching*
Constant	-2.736*** (0.339)	-2.588*** (0.319)	-7.035*** (0.632)	-6.829*** (0.827)
ProxTreatment	1.061*** (0.212)	0.963*** (0.218)	2.966*** (0.340)	2.767*** (0.627)
Log(PreFavorites)$_j$	0.175** (0.087)	0.175** (0.081)	0.343** (0.173)	0.298** (0.145)
Genre Fixed Effects	Yes	Yes	Yes	Yes
LR Chi^2	51.80***	47.79***	73.83***	67.86***
Pseudo R^2	0.198	0.179	--	--
N	476	476	476	476

Notes: These regressions have 7-day observation window after 24 hour burn-in period. Standard errors are in parentheses.* indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$.

Combined Model of Popularity and Propensity Influence

Finally, we consider the results obtained in a combined model of popularity and proximity influence, as shown in Table 2.9. We build the model in stages, so that Model 1 has popularity influence alone, while Model 2 has proximity influence alone. Model 3 has variables for both popularity and proximity influence. We discuss just the key variables of interest — the control variables generally have the expected sign and significance. Starting with Model 1, we find that the interaction term *PopTreatment*Time* is not significant, probably due to the sparseness of total listens at the user-song granularity (recall that our original DD model

is at the aggregate song level). We also conduct a subsample analysis comparing the cases *ProxTreatment* = 0 and *ProxTreatment* = 1. We find that popularity influence is significant only when the user does not have a friend who has previously favorited the song. In other words, popularity influence is only important in the absence of proximity influence.

Turning to Model 2, we find that *ProxTreatment* is positive and significant, and its magnitude declines under propensity score matching, consistent with our earlier finding of proximity influence net of homophily. Looking at the subsamples based on *PopTreatment* (i.e., before and after the popularity information feature implementation) we find that the *ProxTreatment* variable has greater sign and significance in the absence of *PopTreatment*, consistent with the idea that the two types of influence are substitutes. The *ProxTreatment* variable is not significant for the case of *PopTreatment* = 1, but this regression itself is not significant, so we cannot draw a clear conclusion from it.

Finally, *ProxTreatment* remains significant in Model 3, which combines popularity and proximity treatment variables. Here, the most interesting coefficient is that of the three-way interaction *PopTreatment*After*ProxTreatment*, capturing the impact of proximity treatment on popularity influence – and vice versa. We find that this coefficient is negative and significant, consistent with our previous results suggesting that popularity influence and proximity influence are substitutes. Specifically, popularity influence is less important in the presence of proximity influence, echoing our findings from Model 1.

Table 2.9. Jointly Estimating Popularity and Proximity Influence

	Popularity Influence						Proximity Influence						Popularity & Proximity Influence	
	Full Sample		ProxTreatment = 0		ProxTreatment = 1		Full Sample		PopTreatment = 0		PopTreatment = 1		Full Sample	
	Random Matching	*P.S. Matching*	*Random Matching*	*P.S. Matching*	*Random Matching*	*P.S. Matching*	*Random Matching*	*P.S. Matching*	*Random Matching*	*P.S. Matching*	*Random Matching*	*P.S. Matching*	*Random Matching*	*P.S. Matching*
Constant	0.259 (0.169)	0.265 (0.171)	0.051 (0.136)	0.073 (0.182)	0.466 (0.343)	0.498 (0.318)	-0.017 (0.153)	-0.021 (0.156)	-0.198 (0.222)	-0.045 (0.167)	-0.122 (0.258)	0.264 (0.287)	0.116 (0.148)	0.103 (0.154)
$PopTreatment_j$	-0.011 (0.079)	-0.075 (0.083)	-0.110* (0.063)	-0.143** (0.069)	0.088 (0.113)	-0.008 (0.121)							-0.140 (0.089)	-0.171* (0.096)
$After_t$	-0.228*** (0.079)	-0.258*** (0.085)	-0.187*** (0.064)	-0.212*** (0.071)	-0.268** (0.114)	-0.304** (0.124)							-0.187** (0.093)	-0.212** (0.102)
$ProxTreatment_{gj}$							0.323*** (0.046)	0.313*** (0.048)	0.295*** (0.079)	0.277*** (0.082)	0.210** (0.071)	0.155* (0.082)	0.285*** (0.093)	0.253*** (0.097)
$PopTreatment_j$ * $After_t$	0.006 (0.102)	0.078 (0.108)	0.131* (0.074)	0.212** (0.090)	-0.120 (0.147)	-0.056 (0.157)							0.131 (0.121)	0.212** (0.102)
$PopTreatment_j$ * $ProxTreatment_{gj}$													0.258 (0.221)	0.192 (0.129)
$After_t$ * $PopTreatment_{ij}$													-0.082 (0.132)	-0.092 (0.145)
$PopTreat_j$ * $After_t$ * $ProxTreat_{gj}$													-0.251** (0.126)	-0.268** (0.134)
$Log(PreFavorites)_j$	0.041 (0.027)	0.054* (0.027)	0.061*** (0.022)	0.098*** (0.023)	0.021 (0.039)	0.009 (0.040)	0.041* (0.025)	0.055** (0.025)	0.075* (0.043)	0.115** (0.043)	0.004 (0.043)	0.001 (0.047)	0.041* (0.022)	0.054** (0.023)
Genre Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F	2.08**	2.30**	2.24**	3.70***	2.64***	2.28**	5.45***	5.74***	1.97*	3.51***	1.64	1.11	6.70***	6.42***
Adjusted R^2	0.071	0.083	0.152	0.273	0.197	0.152	0.210	0.212	0.092	0.203	0.077	0.016	0.353	0.338
N	160	160	80	80	80	80	160	160	80	80	80	80	160	160

Notes: Model 1 estimates popularity influence, Model 2 estimates proximity influence, and Model 3 jointly estimates both popularity and proximity influence. 1 day "before" window (9/30 and 9/23) vs. 1 day "after" window (10/2 and 9/25). Standard errors are in parentheses. * indicates $p < 0.10$, ** indicates $p < 0.05$ and *** indicates $p < 0.01$.

CHAPTER 3

Distances and Brands in Mobile Local Search Analytics

Literature Review

This chapter draws from three main streams of work: theory around economics of online search, a second stream focused on studying location-based services, and the literature on branding. The first stream consists of studies that look at how individuals make click-through/conversion decisions based on ranking effects and the information relevance on a result/landing page. The second stream of literature examines the role that location proximity information plays on consumption decisions. Last, this work is related to the literature on brand equity that focuses on understanding how brand effects influence consumer decisions.

First, it has long been recognized that consumers tend to pick the choices on the top of lists (Becker 1954). This concept is true in general across different contexts. Specifically, as documented by several studies (e.g., Ghose and Yang 2009; Agarwal et al. 2011), the higher position for an impression on a result page in an online search, the higher the click-through rate, and this is likely due to lower search costs (e.g. Yao and Mela 2011). Indeed, consumers need to exert additional effort to obtain the information from the less prominent listings (Brynjolfsson et al. 2010; Narayanan and Kalyanam 2015), and the ranking effect goes up as consumers' search cost increases. In the mobile setting, it is even true that consumers suffer from substantial search costs due to the small screen size (Ghose et al. 2012). Thus, we expect our analysis ought to provide results consistent with prior evidence on search rank effects – albeit in the somewhat different local search context.

Second, early work on the importance of location goes back to Hotelling (1927). In economic theory, transportation costs associated with geographic distance is one of key determinants of how a business optimizes strategies (Curry 1978; Stahl 1982) as well as how a consumer maximizes her utility (Hanson 1980; Mulligan 1983). With the evolution of e-commerce, researchers turned their focus to the comparison between online and offline retailers, once more highlighting the role that physical distance plays in decision making in both theory (Balasubramanian 1998) and empirical work (Chevalier and

Goolsbee 2003; Forman et al. 2009). Geographic proximity is salient in the mobile context. A growing literature studies the impact of distance/location on advertising in a verity of settings. Ghose et al. (2012) found that consumers are more location-selective in the mobile microblogging-browsing experience. Luo et al. (2014) also confirmed that consumers who received short message service (SMS) texts close to a movie location had a higher movie purchase intent. Echoing these findings, Molitor et al. (2014) found that mobile coupon redemption rates increase as consumers get closer to a retail store, conditional on the discount offered. Though mobile local search that we study is more pull-like demand from consumer side while location-based advertising is push-like demand from advertiser side, these two contexts have the same main ingredient – distance.

Last, we draw from literature on brand equity. Aaker (1991) formally defined brand equity as a set of assets (linked to a brand), that adds to the value provided by a product to a firm and to the firm's customers. In the literature, brand equity depends on how consumers perceive positive brand attributes and favorable consequences of brand use. Keller (1993) argued that high equity brands have high levels of brand awareness, positive brand imagine, returning in high purchase intent and great brand loyalty. Consumers with a strong, favorable brand attitude are more willing to pay premium prices for the brand (Starr and Rubinson 1978) or search more information about the brand (Simonson et al. 1988). In the economics view, consumers will either gain higher utility or have less uncertainty by choosing the product of a high-equity brand. Brand equity also matters in the online setting. Danaher et al. (2003) documented that brand loyalty for high equity brands converts to high online sales. Dewan and Hsu (2004) found that buyers were willing to pay a premium to the sellers with high reputation or low uncertainty (i.e. brand equity) derived from high review volume and valence in an online auction setting.

Currently, the literature on location-based service does not examine the role of brand equity, a key driver in consumer choice, nor does it examine how the impact of location-based services differ for "Home" and "Away" searches. Thus, in this study, we seek to close this gap by first examining differences in outcomes for the two different types of searches, and then asking whether and to what extent brand equity impacts consumers' decision making in location-based service settings. We posit that consumers are more

sensitive to distance in Away search than Home one. In addition, we conjecture that brand equity plays a critical role in the evaluation of options in a local search context. Specifically, in our context of mobile local search, we argue that consumers not only tend to choose a high-brand-equity business but make the tradeoff between the transportation cost and the utility gain from visiting a high-brand-equity merchant with favorable associations or less uncertainty. We examine these questions by using a dataset that captures consumer decision making across two search scenarios.

Data

We collect a unique data set provided by one of the largest wireless service providers (which shall remain unnamed in this study) in the United State. Most of the users use complimentary mobile devices provided by the company. These mobile phones are running on Android-based operating systems, and are pre-loaded with a web browser optimized for that platform.

A key application on this phone, is the "local search" application. When a user opens the browser and clicks on the local search icon (as shown in Figure 3.1a), the system asks her whether to use her current location by enabling the GPS service on her device and retrieving the latitude/longitude information. If she decides to use her current location (location-enabled search), the system automatically fills out field *Near*, which corresponds to the current device latitude/longitude location (as show in Figure 3.1b). After she submits the search, a result page shows a set of impressions in an ascending order of distances between the search location and the establishments as shown in Figure 3.1c). However, it is worthy to note that the distances are calculated in two different ways depending on whether the user has enabled location services or not. In a location-enabled search, the distance in the impression represents the exact mileage between the device and the establishment. If not, the system takes the center of the specified location (city or zip code) to calculate the distance to the establishment.

To investigate consumers' tradeoff between brand popularity and distance, we use a detailed dataset that allows us to observe the entire history of consumers' local search activities, including keyword, weather, job, and news search. In particular, the company has provided us with daily local search logs of the first quarter of 2015. The search logs record each time every user does a search, along with device series number

and the details of the search, such as search terms, locations, and impressions. We also observe the click-through decision of each impression to evaluate click performance. In addition, we collect supplementary brand characteristics from Yelp.

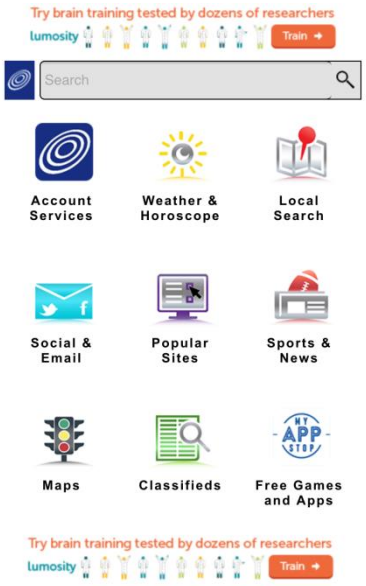


Figure 3.1a.
Portal Page

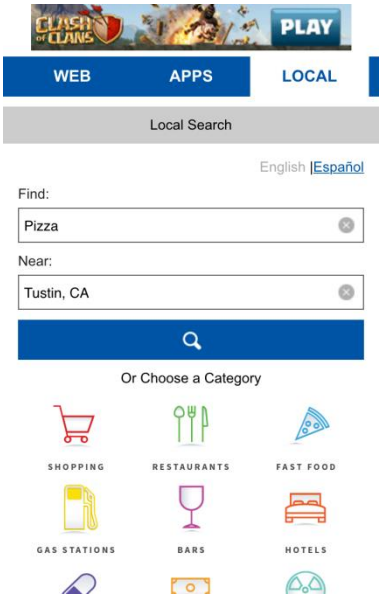


Figure 3.1b.
Local Search Page

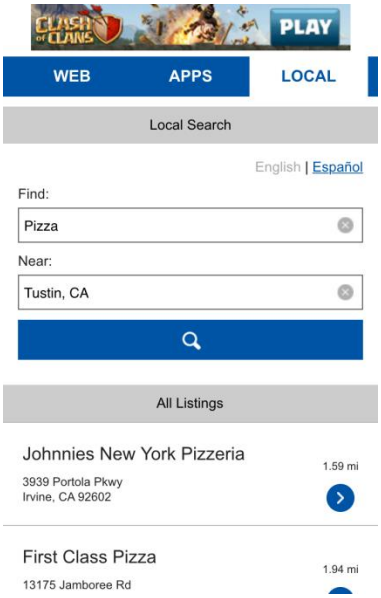


Figure 3.1c.
Result Page

Descriptive Statistics

Our local searches consist of sessions with the chosen keywords: "restaurant". We specifically choose these search keywords since "restaurant" is the top local searched category in our dataset and in general (comScore 2014a; Google 2013; YP 2012). The popularity/association of brands is captured in two levels in this study. We identify whether a restaurant belongs to a nationwide chain in a global perspective while we use Yelp review information to approximate the restaurant's local popularity. Table 3.1 summarizes descriptive statistics of the key variables in our model.

The dependent variable, $Click_{ijk}$, is a binary variable indicating User *i*'s click-through decision on Impression *k* in Search *j*. On average, the CTR is 3.4%, with a standard deviation of 0.182. Each impression comes with its rank of the search page, the proximity distance, and the restaurant title. To distinguish the two use cases, we use $Away_{ij}$, a dummy variable, for the case of Away and Home search, respectively. In order to Profile User *i*'s home location, we fully utilize her local search activities, mainly focusing on

mainly keyword and weather searches. First, we extract the latitude/longitude information of each search, and apply mean shift clustering, a nonparametric machine learning algorithm, on these coordinates. The object of mean shift clustering is to find the highest density region through kernel density estimation. Once the algorithm identifies the home coordinate, we second consider Search *j* as a Home search if the distance between the home and search coordinates within a pre-defined radius[13]. In the dataset, 58% of the local searches are considered as Away cases. Additionally, we also code a dummy variable ($Above_{ijk}$) to capture whether the impression is above the scroll of the screen. As shown in Table 3.1, the top two impressions are above-the-score while the rest of eight stand below the scroll. The distance variable widely ranges from zero to 51.24, and the restaurants in the data set are, on average, 5.02 miles away from the search locations.

Table 3.1. Descriptive Statistics of Local Restaurant Search

Variable	N	Mean	Std. dev.	Min	Max
$Click_{ijk}$	69,460	0.034	0.182	0	1
$Above_{ijk}$ *(above-the-scroll = 1; below-the-scroll = 0)*	69,460	0.200	0.400	0	1
$Distance_{ijk}$ *(in miles)*	69,460	5.023	18.603	0	51.24
$Nationwide_{ijk}$ *(nationwide-brand = 1; local-brand = 0)*	69,460	0.079	0.266	0	1
$ReviewVolume_{ijk}$	69,460	48.078	142.564	1	2,307
$ReviewValence_{ijk}$	69,460	3.273	0.866	1	5
$Price_{ijk}$ *($ = 0, $$ or more = 1)*	69,460	0.366	0.482	0	1
$Away_{ij}$ *(home = 0; away = 1)*	69,460	0.580	0.494	0	1

Moreover, we also use the restaurant titles to construct brand dummy variables of nationwide chains. We identify 278 nationwide brands in the dataset, and 7.9% of impressions come from nationwide chain restaurants. As to the information on Yelp, $ReviewVolume_{ijk}$ widely ranges from 1 to 2,307 while the average review valance of the restaurants is 3.27. $Price_{ijt}$, a dummy variable, indicates whether Impression *j* is an

[13] In this chapter, we have tried 5, 7.5 and 10 mile radius.

inexpensive or a pricey restaurant[14]. Overall, the restaurant data include 69,460 observations from 6,946 searches conducted by 357 distinct users.

Visualization

We first visualize how user click behavior varies with screen position, distance and brands. In Figure 3.2a, the impressions above-the-scroll have much higher CTR since a low ranked impression (higher in the search list) draws most of user attentions. Figure 3.2b illustrates that CTR falls with increased distance to search location. In general, the probability of click follows a downward trend over distance. We also plot the relationship between nationwide (vs. local) brands and click performance. In Figure 3.2c, nationwide chins slight perform better than local restaurants. In addition, popular restaurants (with high review volume) are more attractive to users, as reflected in CTR, than the less popular ones in Figure 3.2d.

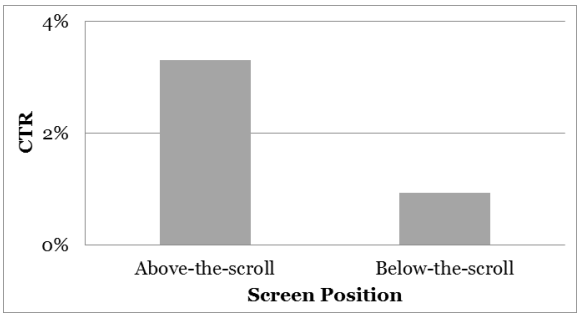


Figure 3.2a. CTR vs. Screen Position

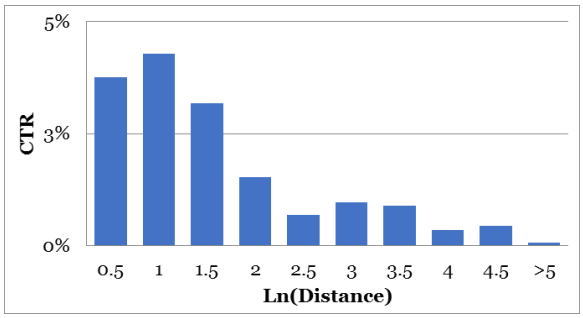


Figure 3.2b. CTR vs. Ln(Distance)

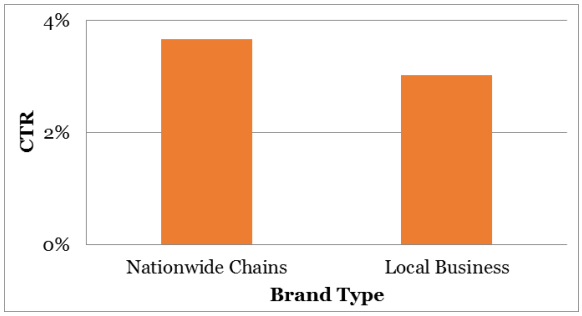


Figure 3.2c. CTR vs. Brand Type

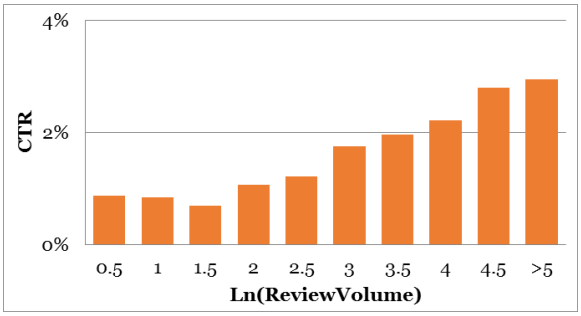


Figure 3.2d. CTR vs. Ln(ReviewVolume)

[14] If the restaurant has only one dollar-sign ($) on its related Yelp page, $Price_{ijk} = 0$. $Price_{ijk} = 1$, otherwise.

Empirical Methodology

Following random-utility theory, we model the click-through as a function of search case and impression-specific characteristics (i.e., above-the-scroll, distance and brands). To control for unobserved user heterogeneity, we specify our random-coefficients model in a hierarchical Bayes framework and use Markov Chain Monte Carlo (MCMC) methods to estimate random coefficients.

Model

Consider User *i* who looks at Impression *j* in Search *k*. She decides to click on that impression only when it provides her positive expected utility while gaining zero utility from not clicking it. The mapping between her latent utility u_{ijk} and the observed action y_{ijk} is:

$$y_{ijk} = \begin{cases} 1, & \text{if } u_{ijk} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Since we mainly investigate how the impression information impact on User *i*'s propensity to click an impression and how she makes tradeoffs between the search/distance cost and the brand, we can specify individual latent utility for restaurant search as the following equation:

$$\begin{aligned} u_{ijk} = {} & \beta_{i0} + \beta_{ij1} Above_{ijk} + \beta_{ij2} Ln(Dist)_{ijk} + \\ & \beta_{ij3} Nation_{ijk} + \beta_{ij4} Ln(RevVol)_{ijk} + \beta_{ij5} RevVal_{ijk} + \beta_{ij6} Price_{ijk} + \\ & \beta_{ij7} Above * Nation_{ijt} + \beta_{ij8} Above * Ln(RevVol)_{ijk} + \\ & \beta_{ij9} Above * RevVal_{ijk} + \beta_{ij10} Above * Price_{ij} + \\ & \beta_{i11} Ln(Dist) * Nation_{ijk} + \beta_{ij12} Ln(Dist) * Ln(RevVol)_{ijk} + \\ & \beta_{ij13} Ln(Dist) * RevVal_{ijk} + \beta_{ij14} Ln(Dist) * Price_{ijk} + \varepsilon_{ijk}\,, \end{aligned}$$

where ε_{ijk} is independent and identically distributed (iid) from type-I extreme value distribution. In the model, $\beta_{ij1} - \beta_{ij6}$ measure the main effects of screen position, distance, and brand. $\beta_{ij7} - \beta_{ij10}$ capture interplay between screen position and brand characteristics while $\beta_{ij11} - \beta_{ij4}$ quantify users' tradeoffs between distance and brand association. If we rewrite the utility function as: $u_{ijk} = \boldsymbol{X'}_{ijk}\boldsymbol{\beta}_{ij} + \varepsilon_{ijk}$, the probability of clicking follow a logistic specification:

$$\Pr(y_{ijk} = 1 \mid \boldsymbol{\beta}_{ij}) = \frac{\exp(\boldsymbol{X'}_{ijk}\boldsymbol{\beta}_{ij})}{1 + \exp(\boldsymbol{X'}_{ijt}\boldsymbol{\beta}_{ij})}.$$

In addition, we are interested in comparing how users behave differently across Home and Away search cases. We model the mean of the individual-level coefficients distribution to depend on the values of $Away_{ij}$. To complete the specification, we also need:

$$\beta_{ij\cdot} = \overline{\alpha}_{j\cdot} + \lambda_{ij\cdot} + (\overline{\delta}_{j\cdot} + \kappa_{ij\cdot}) * Away_{ij}$$

We allow our model to incorporate the heterogeneity at the moderating effect of Away search at the same time. Individual-level coefficients are normally distributed with respect to population means:

$$\beta_i{}' \sim \text{MVN}\left(\overline{\boldsymbol{\beta}}', \boldsymbol{\Sigma}\right),$$

where $\overline{\boldsymbol{\beta}}' = \begin{pmatrix}\overline{\boldsymbol{\alpha}}' \\ \overline{\boldsymbol{\delta}}'\end{pmatrix}$. Therefore, we can rewrite the distributions of individual-level deviations as:

$$\xi_i{}' \sim \text{MVN}\,(0, \boldsymbol{\Sigma}),$$

where $\xi_i{}' = \begin{pmatrix}\boldsymbol{\lambda}_i{}' \\ \kappa_i{}'\end{pmatrix}$. We implement a MCMC algorithm to estimate our hierarchical binary logit model (Rossi at el. 2005). Due to lack of conjugacy in the logistic model specification, our MCMC procedure is based on Metropolis-Hastings (MH) random walk. The full conditionals are:

$$\boldsymbol{\xi} \mid \boldsymbol{Y}, \boldsymbol{X}, \overline{\boldsymbol{\beta}}, \boldsymbol{\Sigma};$$
$$\boldsymbol{\Sigma} \mid \boldsymbol{\xi};$$
$$\overline{\boldsymbol{\beta}} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\xi}, \boldsymbol{\Sigma}.$$

Results

To improve the efficiency of the sampler, we use the following strategy. First, we run the MCMC for 50,000 iterations and treat the first 30,000 draws as "burn-in" samples, and then calculate posterior means and empirical variance of the remaining 20,000 draws. Second, we rerun the MCMC and use the posterior means calculated in the previous step as the initial values and the empirical variance as the diagonal elements of covariance matrix of the random-walk proposal distribution. Last, we run the chain for 60,000 iterations, keep every 10^{th} iterations and use the last 20,000 iterations to approximate the posterior distribution. We compute the posterior expectation and the significance level of parameter estimates based on the highest posterior density (HPD) intervals. We consider the estimates significant if the corresponding

HPD intervals do not contain zero. To make sure our MCMC chain reaches its stationary phase, we have 10 chains of different sets of initial values and perform the Gelman and Rubin test to diagnose the convergence. Gelman and Rubin (1992) suggest that the within-chain variance and the between-chain variance should be equal or close if chains converge. The Gelman and Rubin statistic is less than 1.12 across our analyses, confirming convergence.

Main Results

Our results for the restaurant search are presented in Table 3.2. In the hierarchical settings, we have two sets of coefficients for the two search cases. We start by discussing our results for Home search case. The positive, statistically significant coefficient of *Above* suggests that impressions benefit from being above-the-scroll, reflecting the findings of Ghose et al. (2012). *Log(Distance)* has a negative sign and significant, indicating that users tend to click on closer restaurants. This is consistent with market evidence (YP 2012) where a user's distance to the business location negatively impact their likelihood to click on an ad or impression for that business. Compared with local restaurants, nationwide chains are slightly more appealing to users. Interestingly, Yelp review volume is positive and significant, while valence is not significant. We believe this is due to the fact that review volume is a proxy for brand popularity. Also, Valence is marginally significant, suggesting users account for it in their clicking decisions. The price effects is also preserved in our results.

The coefficients of the interaction terms demonstrate how users make tradeoffs between brand popularity and the effects of above-the-scroll or distance proximity. For nationwide chains, the positive and significant coefficient of the interaction with *Above* means that the top rank amplifies the nationwide brand effects and boosts CTR. Similar logics are applied to other brand popularity measures. As to the relationship between distance and brands, it is worthy to highlight that the coefficient of *Log(Distance)*Nationwide* is positive and significant, suggesting that users tend to click on a nationwide chain restaurant impression even if it is further away than another local business. We see a consistent effect for different types of brand association.

Table 3.2. Results of Local Restaurant Search

	Main Effect (Home Search)	Moderating Effect (Away Search)
Above	0.313***	0.009
Ln(Distance)	-0.056**	-0.007**
Nationwide	0.034**	0.098**
Ln(ReviewVol)	0.049**	-0.031**
ReviewValence	0.095*	-0.090**
Price	-0.158**	0.019
*Above*Nationwide*	0.003*	0.017*
*Above*Ln(ReviewVol)*	0.006*	-0.005*
*Above*ReviewValence*	0.017*	-0.016*
*Above*Price*	0.009*	-0.002
*Ln(Distance)*Nationwide*	0.007**	0.021**
*Ln(Distance)*Ln(ReviewVol)*	0.011**	-0.009*
*Ln(Distance)*ReviewValence*	0.012**	-0.012**
*Ln(Distance)*Price*	0.004	-0.009
N	29,170	40,290

Note: ***, ** and * indicate that 99%, 95% and 90% HPD interval do not contain 0, respectively.

The second column lists the parameters estimated for the moderating effect of Away search case. In our hierarchical setting, we interpret them as how users behave differently between the two scenarios. In other words, the coefficients represent the additional effect of Away search, relative to the case of Home search. The intercept is positive and significant, indicating that users in Home search have a higher base CTR for the given factor. The effects of *Above* across the two cases are not statistically different, and the low rank position of impressions (i.e., top of the search list) has a positive impact on click performance in Away search mode as well.

Interestingly, distance matters slightly more in Away search case. The sensitivity of distance could be one of the key factors to drive this result. In fact, humans tend to be more cautious and conservative in uncertain situations and unfamiliar locations. Users are less likely to exert additional efforts to try unassociated restaurants. They, therefore, are more willing to click on the impressions of nationwide chains

when they are away from home. In addition, the positive sign of *Above*Nationwide* shows that Nationwide chain impressions garner more advantage of being above the scroll in Away search. Because lacking the information of local restaurants' popularity, users would tend to skip unknown diners even their impressions are above-the-score. Regarding to the tradeoffs between distance and brands, local businesses suffers while nationwide brands benefit from the fact that users to travel further for them. The overall impact of *Log(Distance)*Nationwide* on click-through rate remains positive (0.021-0.007).

Economic Significance

We can quantify the economic significance of above-the-scroll, distance and brand effects as follows. The dependent variable is the probability of clicks, and we use odds ratios to interpret the estimates. First, being above-the-scroll (i.e., the top two positions) increases CTR of an impression by 36.7% (*exp*(0.313)), holding other factors constant. This is consistent with the evidence from the online search literature (Ghose and Yang 2009; Agarwal et al. 2011). Users react to *Above* roughly the same in both search modes. Second, a one-percent increase in distance lowers the odds of clicking on an impression by 5.6% in Home search and by 6.3% (0.056+0.007) in Away search. The results share the same flavor as Hampton and Wellman's (2002) findings. The odds ratio between the cases is 1.125, indicating that CTR drops 12.5% more for one-percent increase in distance when users are located in the areas unfamiliar with. The odds ratio highlights that distance information is not only statistically significant but economically important. Last, the probability of clicking on an impression increases by 3.5% and 14.1% if it is associated with nationwide chains in Home and Away cases, respectively. On average, users are 0.3% (2%) more willing to click on an above-the-scroll impression of Nationwide in Home (Away) case. Regarding the distance-brand tradeoff, local businesses with 1% increase in review volume could promote users the odds of clicking by 1.1% in the base case.

Alternative Keyword – Grocery Search

In addition, we also look at grocery keyword search. The results of grocery search are summarized in Table 3.3. We start with a discussion of Home search case. CTR is positively associated with

above-the-scroll listings and negatively associated with distance, consistent with our results for restaurant search. Price has a negative effect on CTR, as expected. Interestingly, *Nationwide* dummy is insignificant, suggesting that users are more willing to support local grocery stores than chains. In addition, Yelp review volume is positive and significant, while valence is not significant. We believe this is due to the fact that review volume is a better proxy for brand popularity than valence. Overall, these results are consistent with what we found for grocery: users prefer a *familiar* grocery store over a *close* one. In the case of Away search, we find that distance has a stronger impact on CTR than in Home search again. Lastly, the moderating effect of Away search on users' tradeoffs in grocery search are consistent with restaurant search.

Table 3.3. Results of Local Grocery Search

	Main Effect (Home Search)	Moderating Effect (Away Search)
Above	0.340***	0.025
Ln(Distance)	-0.055***	-0.002*
Nationwide	0.016*	0.012****
Ln(ReviewVol)	0.037**	-0.035***
ReviewValence	0.125*	-0.118***
Price	-0.108**	0.000*
*Above*Nationwide*	0.010	0.014*
*Above*Ln(ReviewVol)*	0.011**	-0.009*
*Above*ReviewValence*	0.023	0.009
*Above*Price*	0.016*	0.007
*Ln(Distance)*Nationwide*	0.002**	0.004**
*Ln(Distance)*Ln(ReviewVol)*	0.015**	0.014*
*Ln(Distance)*ReviewValence*	0.016	0.009
*Ln(Distance)*Price*	0.004	0.009
N	15,520	12,460

Note: ***, ** and * indicate that 99%, 95% and 90% HPD interval do not contain 0, respectively.

CHAPTER 4

The Role of Distance and Competition in Location-Based Advertising

Literature Review

This chapter draws from two main streams of work: the economics of location-based services and the literature on competition. The first stream examines the role that location proximity plays on consumption decisions. The second stream includes studies that look at how competition has impacts on consumers' click and conversion decisions. Since the literature of location proximity is reviewed in Chapter 3, this section focuses on the studies of competition.

Competition is a critical factor affecting firm's profitability as long as the products that firms offer are substitutes for each other. In fact, a variety of perspectives are used to model competition between firms in economics. For example, Balasubramanian (1998) models the competition between shopping malls and direct channels (e.g., online stores). Another example is market concentration, such as Herfindahl index (HHI) and the concentration ratio (CR), which is commonly used to as a measure of competition in the literature of industrial organization. Moreover, competition could be also interpreted as how products are differentiated in terms of product characteristics (Berry 1994). Motivated by the spatial competition by Hotelling (1927), we use proximity-concentration to approximate competition between firms. The larger the number of competitor establishments that are located in the same area, and the closer they are to the focal establishment, the higher is the competition. Brainard (1993) empirically assesses proximity-concentration in Multinational Sales and Trade.

Research Framework

Though there is growing body of literature on location-based services, the role of competition has not been studied. This study tries to contribute to the literature by examining the impacts of competition and asking how competition could moderate distance – a common ingredient in research on location-based services. We are also interested in how effects of competition and distance vary between the click and conversion

stages of consumer decision making. Thus, our key dependent variables are click and conversion performance, which we explain on the basis of distance and competition, while controlling for user, device and app characteristics (see Figure 4.1).

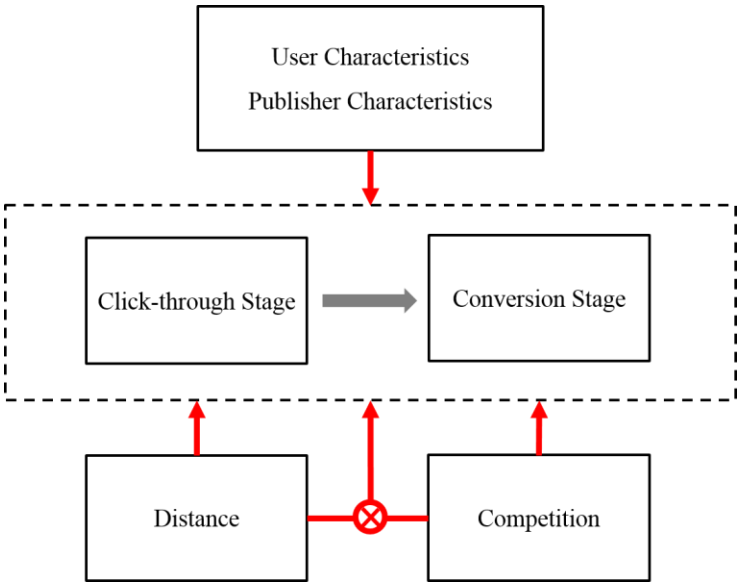


Figure 4.1. Research Framework of Geo-fence Advertising Study

We predict that distance would not play a critical role in the click-through stage, because distance information is usually not included in the ad impression. In contrast, we hypothesize that distance plays an important role in the conversion stage. To decide whether to visit the advertiser's store, a consumer would typically consider the impact of distance on purchase utility. We examine these questions by using a unique dataset that captures consumer decision making across the two stages of decision making (click and conversion) in the context of geo-fencing advertising.

Data

The mobile advertising ecosystem consists of both supply-side players, publishers (which in our case are mobile apps) exchanges (i.e., inventory aggregators), and demand side parties, marketing agencies and advertisers. Publishers, such as websites or apps, provide the mobile screen real estate for the display of impressions, while exchanges consolidate inventory to sell in a batch. Marketing agencies bid in real-time on impression auctions on behalf of advertisers (e.g., restaurant chains or shopping malls), and often also

run location-based campaigns on their behalf. For example, an advertiser's ad is immediately displayed in the ad space of a webpage or app which a consumer is browsing once an advertiser wins the auction. Then, the consumer makes the decision to click the ad in the first stage; if yes, she may take some action (such as click on a map link, make a call, or click on a hypertext link within the landing page), and we refer to it as conversion in the second stage.

We obtain unique and rich data from one of the largest location-based marketing agencies (which shall remain unnamed in this study) in the United States. The data consist of over million auctions handled by the agency during one-month period of January 2015. Each observation in our data set includes the consumer's location in terms of latitude and longitude, device and operating system characteristics, the centroid of the advertiser's campaign, and publisher characteristics. To investigate the impact of competition, we also collect data on competitors within 5 miles of the focal establishment, from Yelp and Google. In addition, the median income and population size at zip-code level are used to control for broad demographics that might affect consumer response to advertising.

Descriptive Statistics

In this chapter, our analyses focuses on one of the largest fast-food chains. We specifically choose this advertiser since geo-fence advertising is very common in the context of fast food restaurants (BIA/Kelsey 2014a; Thumbvista 2015). We randomly select 1 million winning bids of the advertiser in real-time advertising auctions. Since the distance measure between consumers and the advertiser's locations is one of the key variables of interest, we filter out the observations with imprecise location information (from sources such as Wi-Fi, IP lookup or cellular tower location), and only include the ones where location information is precisely obtained from GPS chips in smartphones. We end up with roughly 180 thousand observations. The rich data set covers most states in the U.S., as shown in Figure 4.2. There are over 20 thousand impressions displayed in Michigan, Texas, Virginia and Georgia.

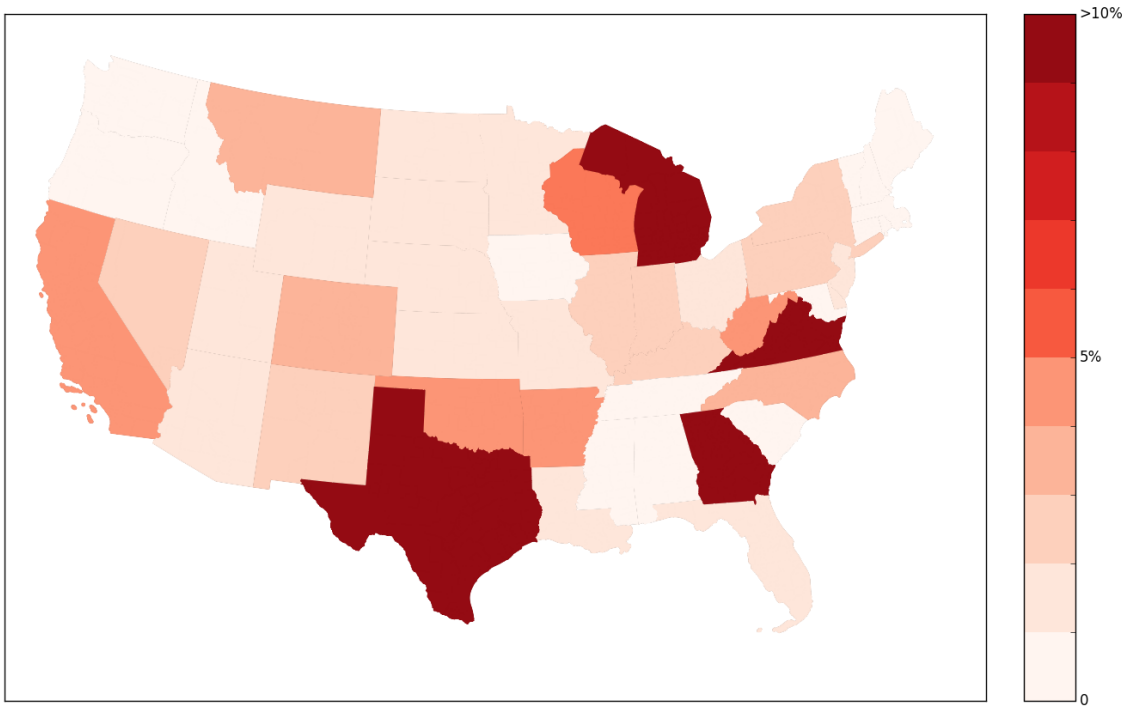


Figure 4.2. Heat Map of Geo-fence Advertising Campaigns

The extent of the competition between the advertiser and its local rivals in the geo-fence is captured through two metrics. The first one is simply the number of restaurant competitors in the geo-fence area (five miles around the focal establishment). The second metric emphasizes the proximity of the competition. For this purpose, we construct a competition index, $CompInd_i$, by incorporating the distance of the closest competitor to the focal establishment *i*:

$$CompInd_i = 1 - \frac{1}{1 + e^{\frac{1}{Distance_i^c}}}$$

where $Distance_i^c$ is the distance between the advertiser and its closest competitor. Therefore, $CompInd_i$ should be bounded between 0.5 and 1 since $Distance_i^c > 0$. Table 4.1 summarizes descriptive statistics of the key variables in our model. The dependent variables, $Click_{ij}$ and $Convert_{ij}$, both are binary variables indicating consumers click and conversion decisions on Impression *j* associated the advertiser's location *i*. On average, the click-through rate is 0.7%, with a standard deviation of 0.082, and the conversion rate is 2.1%, with a standard deviation of 0.144, conditional on $Click_{ij}$ = 1. $Distance_{ij}$ refers to the distance between consumer and advertiser's locations.

Table 4.1. Descriptive Statistics of Geo-fence Advertising

Variable	N	Mean	Std. dev.	Min	Max
$Click_{ij}$	180,870	0.007	0.082	0	1
$Convert_{ij}$ (conditional on $Click_{ij}$ = 1)	180,870	0.021	0.144	0	1
$Distance_{ij}$	180,870	1.619	1.242	0.000	5.000
$NoOfCompetitors_i$	180,870	7.453	2.855	1	18
$Distance_i^c$	180,870	1.368	1.251	0.000	5.000
iOS_{ij}	180,870	0.632	0.483	0	1
$MedIncome_{ij}$	180,870	49,960	18,009	316	178,284
$Population_{ij}$	180,870	26,697	16,410	1	111,086
$LargeImp_{ij}$	180,870	0.219	0.414	0	1
$CatEnt_{ij}$	180,870	0.409	0.492	0	1
$CatSocial_{ij}$	180,870	0.099	0.299	0	1
$CatGame_{ij}$	180,870	0.219	0.414	0	1

The average distance is 1.619 miles, and its range is from 0 to 5. There are 7.5 other restaurants, on average, within a 5-mile radius of the advertiser's focal establishment. $Distance_i^c$ ranges from 0 to 5 miles, with an average of 1.368 miles. Consumer characteristics are captured by the variables iOS_{ij}, (a dummy variable which is 1 for an iOS device, and 0 otherwise), $MedIncome_{ij}$ (median income in the geo-fence zip code) and $Population_{ij}$ (the population of the geo-fence zip code). 63.2% of consumers use mobile devices running Apple iOS operating systems. The average median income and population size across zip codes in the data are around 50 thousand dollars and 27 thousand people, respectively. As to publisher attributes, $LargeImp_{ij}$ is a dummy variable to indicate the size of impressions. Typically, the size is either 320*50 or 728*90 pixels, where $LargeImp_{ij}$ = 1. We use the three dummies to code the categories[15] of publishing apps. Specifically, there are three app categories – Entertainment, Social Networking and Gaming – accounting for 40.9%, 9.9%, and 21.9% of total observations.

[15] We follow Interactive Advertising Bureau's (IAB) Open RTB API specification to categorize apps.

Visualization

In Figure 4.3a, there is no clear pattern between the propensity of a click and distance. This can be caused either by high variation of other factors or by no explanation of distance itself. This observation seems to be somehow different from general intuition. Oppositely, the conversion rate drops as the distance between a consumer and the advertiser's location goes up, suggesting there could be negative impact of distance on the conversion performance.

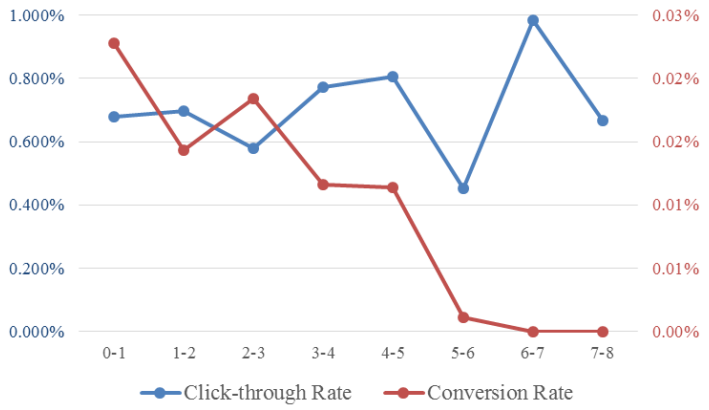


Figure 4.3a. Click-through Rate and Conversation Rate vs. Distance

Figure 4.3b illustrates that the click-through rate decrease with the completion measure, the number of competitors in a 5 mile area of the advertisers' locations. The rate drops from 0.9% to 0.2% when the number of competitors increases from 1 to 9. However, it is not clear whether competition also lowers the propensity of a click. To examine the abovementioned relationship, we incorporate these two key factors along with other variable into regression models introduced in the next section.

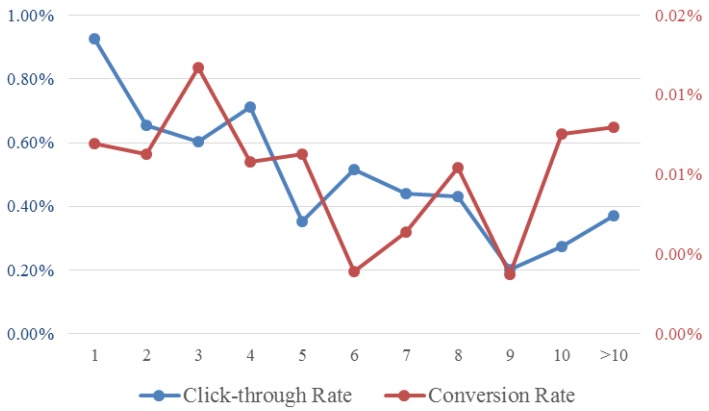


Figure 4.3b. Click-through Rate and Conversation Rate vs. Number of Competitors

Empirical Methodology

We propose a reduced-form model in this section to estimate the impacts of distance, competition, consumer characteristics and publishers' attributes on consumers' click and conversion decisions. In the click-through stage, the observed consumer's binary response (i.e., whether to click) on Impression *j* associated the location *i* is mapped as follows:

$$Click_{ij} = \begin{cases} 1, & if\ click; \\ 0, & otherwise. \end{cases}$$

Similarly, we have the following mapping for the conversion stage:

$$Convert_{ij} = \begin{cases} 1, & if\ convert; \\ 0, & otherwise. \end{cases}$$

Then, we have the following utility models, for click and conversion:

$$\begin{aligned} u_{ij} = {} & \beta_0 + \beta_1 Ln\big(distance_{ij}\big) + \beta_2 Ln(NoOfCompetitiors_i) + \beta_3 CompInd_i + \\ & \beta_4 Ln\big(distance_{ij}\big) * Ln(NoOfCompetitiors_i) + \beta_5 Ln\big(distance_{ij}\big) * CompInd_i + \\ & \beta_6 iOS_{ij} + \beta_7 Ln(MedIncome_{ij}) + \beta_8 Ln(Population_{ij}) + \\ & \beta_9 LargeImp_{ij} + \beta_{10} CatEnt_{ij} + \beta_{11} CatSocial_{ij} + \beta_{12} CatEnt_{ij} + \varepsilon_{ij}; \end{aligned}$$

$$\begin{aligned} v_{ij} = {} & \gamma_0 + \gamma_1 Ln\big(distance_{ij}\big) + \gamma_2 Ln(NoOfCompetitiors_i) + \gamma_3 CompInd_i + \\ & \gamma_4 Ln\big(distance_{ij}\big) * Ln(NoOfCompetitiors_i) + \gamma_5 Ln\big(distance_{ij}\big) * CompInd_i + \\ & \gamma_6 iOS_{ij} + \beta_7 Ln(MedIncome_{ij}) + \gamma_8 Ln(Population_{ij}) + \\ & \gamma_9 LargeImp_{ij} + \gamma_{10} CatEnt_{ij} + \gamma_{11} CatSocial_{ij} + \gamma_{12} CatEnt_{ij} + \tau_{ij}; \end{aligned}$$

where β_1 (γ_1) captures the impacts of distance, and β_2 and β_3 (γ_2 and γ_3) reflect the effects of the competition. β_4 and β_5 (γ_4 and γ_5) capture the interaction between distance and competition. If we assume that ε_{ij} and τ_{ij} are independently and identically distributed from type-I extreme value distribution, the click-through rate and conversion rate can be expressed as logistic regression:

$$\Pr\big(Click_{ij} = 1\big) \quad = \frac{exp^{(u_{ij})}}{1 + exp^{(u_{ij})}} \text{ and}$$

$$\Pr\left(Convert_{ij} = 1\right) = \frac{exp^{(v_{ij})}}{1 + exp^{(v_{ij})}}.$$

Results

Our results for consumers' click response to geo-fence advertising are presented in Table 4.2. We start with the models that only include the direct effects of distance and competition. In Model (1), the coefficient on distance is not statistically significant, suggesting that distance between consumer and the advertiser's locations does not impact on consumer's click decision. In Models (2) and (3), the two competition measure, $Ln(NoOfCompetitiors_i)$ and $CompInd_i$, are both negative and significant, indicating that higher micro-competition in the geo-fence zone lowers the consumer's propensity to click on ad impressions. The signs and significance of these variables are similar in Models (4) and (5). The coefficient estimates for the distance-competition interaction terms in Model (6)-(8), suggest that these two factors do not interact in consumers' click decisions.

The coefficient of iOS_{ij} is insignificant, indicating no systematic difference in click performance for iOS versus Android devices. $Ln(MedIncome_{ij})$ has a negative sign but insignificant while $Ln(MedIncome_{ij})$ is positive and significant, suggesting that click performance is better in high population areas but not in low income neighborhoods. Impressions of a larger size seems not more attractive to consumers to click. The categories of publishers' apps also play an important role affecting click performance. Due to their positive and significant coefficients, the restaurant advertiser can benefit from displaying ads on entrainment and gaming apps, rather than social media and other apps. This could be driven by the natural match between restaurant business and the consumer segment, such as millennials, interested in entertainment and gaming.

Table 4.3 summarizes the estimates of the conversion performance. Across Model (1)-(8), the coefficients of $Ln(Distance_{ij})$ are now significant. The negative signs show that consumer are sensitive to distance. Since distance incurs disutility, consumers may consider it more seriously when they do have high purchase intent. Therefore, it could be the reason distance matters only in the second stage, not in the

first one. YP (2012) also provides similar evidence for fast food markets. Also, it is surprising that both competition coefficients become insignificant though maintaining a negative sign, indicating competition does not significantly affect conversion performance. Moreover, it is worth noting that the interaction between distance and the number of competitors is not significant while the coefficient of $Ln(Distance_{ij}) * CompInd_i$ is marginally significant at 0.1 level. In other words, there is some evidence that a consumer is less likely to visit the advertiser's restaurant when it is farther from her and there is an alternative closer by.

Regarding consumer and publisher characteristics, consumers using Apple iOS devices have lower probability to actually visit the locations. This implies that the advertiser is not attractive to this specific consumer segment. Other demographics have no effect on the conversion rate. The size of impressions does not play a role, either. Gaming apps still allow the advertiser to enjoy better conversion performance while the others do not. The above results for the control variables are robust across all models in Table 4.2.

Table 4.2. Results of Click Response

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Intercept$	-6.387*** (0.920)	-6.588*** (0.913)	-6.142*** (0.920)	-6.582*** (0.914)	-6.561*** (0.919)	-6.511*** (0.932)	-6.495*** (0.911)	-6.491*** (0.0926)
$Ln(Distance_{ij})$	0.003 (0.065)			0.011 (0.065)	0.024 (0.069)	-0.061 (0.124)	-0.048 (0.121)	-0.042 (0.118)
$Ln(NoOfCompetitors_i)$		-0.090*** (0.013)		-0.090*** (0.013)	-0.077*** (0.015)	-0.072*** (0.025)	-0.070*** (0.022)	-0.072*** (0.024)
$CompInd_i$			-0.127*** (0.015)		-0.101*** (0.015)	-0.106*** (0.019)	-0.106*** (0.015)	-0.102*** (0.019)
$Ln(Distance_{ij}) * Ln(NoOfCompetitors_i)$						0.010 (0.054)		0.051 (0.093)
$Ln(Distance_{ij}) * Ln(CompInd_i)$							0.152 (0.110)	0.158 (0.175)
iOS_{ij}	0.718 (0.539)	0.360 (0.342)	0.404 (0.351)	0.359 (0.342)	0.370 (0.366)	0.359 (0.348)	0.363 (0.349)	0.374 (0.355)
$Ln(MedIncome_{ij})$	-0.059 (0.083)	-0.086 (0.082)	-0.109 (0.080)	-0.087 (0.083)	-0.092 (0.088)	-0.078 (0.084)	-0.054 (0.083)	-0.066 (0.080)
$Ln(Population_{ij})$	0.099*** (0.032)	0.103*** (0.032)	0.093*** (0.032)	0.102*** (0.032)	0.092*** (0.032)	0.103*** (0.030)	0.112*** (0.033)	0.115*** (0.032)
$LargeImp_{ij}$	0.032 (0.107)	-0.049 (0.114)	-0.070 (0.116)	-0.049 (0.115)	-0.071 (0.116)	-0.060 (0.114)	-0.031 (0.116)	-0.015 (0.116)
$CatEnt_{ij}$	0.190** (0.082)	0.182** (0.083)	0.188** (0.090)	0.181** (0.088)	0.179** (0.090)	0.180** (0.089)	0.085** (0.088)	0.100** (0.088)
$CatSocial_{ij}$	0.101 (0.122)	0.050 (0.122)	-0.053 (0.129)	0.049 (0.123)	-0.052 (0.130)	0.049 (0.131)	0.068 (0.129)	0.070 (0.124)
$CatGame_{ij}$	0.887*** (0.033)	0.684*** (0.126)	0.538*** (0.136)	0.683*** (0.126)	0.541*** (0.132)	0.583*** (0.136)	0.502*** (0.135)	0.526*** (0.124)
LR Chi^2	205.38	256.71	199.75	278.74	290.36	290.10	289.55	290.72
Pseudo R^2	0.014	0.018	0.015	0.019	0.024	0.023	0.022	0.023

Note: N = 180,870. Standard errors are in parentheses. * indicates p < 0.10, ** indicates p < 0.05 and *** indicates p < 0.01.

Table 4.3. Results of Conversion Response

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Intercept$	-5.623*** (0.584)	-5.612*** (0.567)	-5.694*** (0.570)	-5.532*** (0.553)	-5.504*** (0.548)	-5.523*** (0.550)	-5.513*** (0.541)	-5.526*** (0.543)
$Ln(Distance_{ij})$	-0.603*** (0.046)			-0.464*** (0.059)	-0.482*** (0.051)	-0.684*** (0.053)	-0.682*** (0.048)	-0.703*** (0.046)
$Ln(NoOfCompetitors_i)$		-0.039 (0.075)		-0.041 (0.068)	-0.036 (0.072)	-0.301 (0.070)	-0.287 (0.072)	-0.280 (0.073)
$CompInd_i$			-0.152 (0.265)		-0.109 (0.238)	-0.163 (0.240)	-0.175 (0.239)	-0.143 (0.238)
$Ln(Distance_{ij}) * Ln(NoOfCompetitors_i)$						-0.025 (0.132)		-0.023 (0.130)
$Ln(Distance_{ij}) * Ln(CompInd_i)$							-0.062* (0.036)	-0.057* (0.032)
iOS_{ij}	-1.040*** (0.106)	-0.899*** (0.117)	-1.000*** (0.121)	-0.879*** (0.139)	-0.857*** (0.140)	-0.682*** (0.143)	-0.699*** (0.139)	-0.689*** (0.141)
$Ln(MedIncome_{ij})$	-0.135 (0.532)	-0.207 (0.539)	-0.222 (0.537)	-0.130 (0.547)	-0.115 (0.532)	-0.150 (0.535)	-0.125 (0.529)	-0.136 (0.530)
$Ln(Population_{ij})$	-0.067 (0.170)	-0.074 (0.169)	-0.069 (0.169)	-0.070 (0.173)	-0.047 (0.172)	-0.018 (0.170)	-0.043 (0.171)	-0.038 (0.174)
$LargeImp_{ij}$	-0.364 (0.717)	-0.313 (0.706)	-0.347 (0.707)	-0.312 (0.726)	-0.292 (0.727)	-0.249 (0.721)	-0.251 (0.730)	-0.245 (0.729)
$CatEnt_{ij}$	0.285 (0.554)	0.288 (0.557)	0.202 (0.555)	0.312 (0.560)	0.324 (0.557)	0.310 (0.555)	0.327 (0.562)	0.329 (0.562)
$CatSocial_{ij}$	0.573 (0.734)	0.542 (0.732)	0.521 (0.735)	0.583 (0.756)	0.606 (0.756)	0.0631 (0.757)	0.636 (0.757)	0.653 (0.756)
$CatGame_{ij}$	1.105* (0.584)	1.121* (0.604)	1.100* (0.614)	1.179* (0.633)	1.189* (0.628)	1.106* (0.628)	1.127* (0.625)	1.128* (0.627)
LR Chi^2	4.70	7.47	7.75	8.74	8.36	8.10	9.55	9.56
Pseudo R^2	0.025	0.030	0.029	0.032	0.030	0.030	0.032	0.032

Note: N = 180,870. Standard errors are in parentheses. * indicates p < 0.10, ** indicates p < 0.05 and *** indicates p < 0.01.

Economic Significance

Based on the coefficient estimates in Model (8) in both Table 4.2 and 4.3, we can quantify the economic significance of distance, competition and other controls as follows. We mainly use odds ratios to interpret the estimates. In the click-through stage, having one more competitor nearby decreases the click-through rate by 7.5%, holding other factors constant. 1 unit increase in the competition index also lowers the rate by 10.7%. Last, comparing with other apps, entertainment and gaming apps increase the probability to click by 10.5% and 69.2%, respectively. In the conversion stage, 1 mile increase in distance results in 33.2% drop of conversion rate, and the effect will be amplified if there is 1 unit increase in the competition index. Apple users are less likely to take conversion action than other device users by 49.8%. Only gaming app could increase the conversion rate by 12.0%.

CHAPTER 5

Conclusion

In this dissertation, we discussed the impact of distance on consumer choice in digital markets. Different perspectives of distance are studied, including social distance, spatial distance, and mobile screen distance (i.e., the ranks/positions on mobile screens). Through rigorous empirical examination, we found that distance plays a critical role which fundamental affects consumer decision making process. In this chapter, we concluded the main findings of Chapter 2, 3 and 4 and point out the limitations and the future directions.

Social Proximity

We have examined the role of "favorites" as a mechanism for social interaction in an online music community, and jointly estimated popularity influence due to the total number of favorites for a song and proximity influence due to the favoriting behavior of social network friends in close social proximity. Applying a quasi-experimental design to highly granular data from a leading music blog aggregator we find robust evidence that both types of influence are statistically and economically significant. Quantitatively, we find that the availability of popularity information increases the number of listens for the average song by some 12%, and a full 21% for narrow-appeal music. This effect is significant for only newly posted songs, consistent with the nature of our site where older songs are not immediately visible and do not get much attention. Proximity influence (i.e., having a friend that has favorite a song) increases the likelihood of listening to a song by 10.2%, which appears to be more than five times as important as the effect of homophily in explaining correlated consumption. Finally, popularity and proximity influence are substitutes for one another, in that proximity influence, when available, tends to dominate the effect of aggregate song popularity information.

Our findings of significant popularity and proximity influence resonate with industry reports indicating that 92% of consumers say positive recommendations from people they know are the most trusted sources of information (The Nielsen Company 2012b). At the same time, when surveys indicate that 70% of consumers trust consumer opinions posted online (The Nielsen Company 2012b), our results suggest

that what might be driving the implied social influence might be both direct contact and communication between consumers, as well as distant observation of aggregate consumption statistics. Our results indicate that the engagement in online music communities would benefit from both the dissemination of popularity information, as well as the mobilization of social ties and co-consumption of music in online social networks.

These results have important managerial implications for the owners of online music communities, such as the one we study in this study. First, our results suggest that both popularity and proximity influence can be leveraged to increase music consumption and engagement, enabling better monetization of the website; e.g., through better online advertising or more profitable freemium pricing.[16] Marketing strategies should be tied to the type of user and music. To leverage popularity influence the website should make popularity information more salient, such as through the prominent display of daily, weekly or monthly most popular lists. This is more important for niche or narrow-appeal music as opposed to mainstream or broad-appeal music. To leverage proximity influence, users should be encouraged and incentivized to increase social ties and co-consumption of music, and rewarded for their own engagement and that of their friends. Indeed, music websites might be able to increase engagement further by proactively pushing relevant popularity and proximity information, rather than waiting for users to discover them on their own.

Users with many social network friends and activity should be continuously fed with updates from their friends in order to increase their engagement, not unlike the newsfeed feature of Facebook, along with other tactics to increase the virality of music co-consumption (see, e.g., Aral and Walker 2011). But popularity information would be important for socially active users as well, given the likely sparseness in the range of songs favorited in even the most active social network cliques. On the other hand, for users that are inactive socially popularity information is all that more important for music discovery. Here, based on the observational learning literature we can expect herd behavior and information cascades (Bikhchandani et al. 1998) and that initial conditions matter, leading to inequality in consumption (popular

[16] The freemium business model is common at music websites such as Last.fm, Spotify, etc.

songs will get more popular, while unpopular songs will get more unpopular) and to unpredictability of outcomes ("good" songs may not become popular, while "bad" songs may become viral hits), consistent with the findings of Salganik et al. (2006).

Our results should be generalizable to other experience goods such as online videos, books, software, and other digital content. They would also apply to other online communities where both popularity and proximity influence might be at play. In the music context, such communities include Last.fm, Spotify and YouTube. Outside the music context, popularity and proximity influence occur together in online gaming communities (such as Xbox and Blizzard Entertainment), online book clubs (for examples, see *The New York Times* 2013), online health and fitness communities (such as PatientsLikeme.com and nikeplus.com), among others. Both types of influence are also likely on mainstream social networking sites such as Facebook and Twitter, and we are not familiar with prior work that has simultaneously examined popularity and proximity influence, and their interactions, on such increasingly ubiquitous platforms. More broadly, it is important for online platforms to experiment with different features that may facilitate user interaction and engagement with the site.

Turning to limitations, while we have a high level of granularity in music listening and favoriting decisions, we do not have detailed user profiles (due to privacy concerns and/or lack of availability). This means that there are likely sources of unobserved heterogeneity underlying the variation in sampling behavior, which may add noise or bias to our empirical analysis. Seemingly, one shortcoming in our difference-in-difference design is the fact that the treatment and control groups are drawn from different (neighboring) weeks. However, as we discussed earlier, this is not a cause for serious concern. On the contrary, our approach improves the odds of truly exogenous treatment and provides a quasi-experimental approach to study the impact of global feature implementations that affect an entire website at a given point in time. Another limitation is the fact that at the time of our study, the social networking features on THM were relatively new, so that the data for the proximity influence analysis are quite sparse. With richer data we might be able to analyze the role of social ties and network structure more extensively, better leveraging the greater maturity of the community and its underlying social network. Overall, this work provides useful

and robust empirical regularities with respect to macro and micro social influences in online communities, and how they affect consumer behavior and profitable engagement strategies by the communities themselves.

Spatial Proximity – Mobile Local Search

In Chapter 3, we studied mobile local search, focusing on user click-through behavior as a function of distance, brands, and screen position. We implemented a hierarchical Bayes model on a unique data set of local search (impressions and click-throughs) transactions from a major mobile carrier. We find a number of robust empirical regularities. For the search of restaurant and grocery, CTR is positively related to brand popularity and screen position (e.g., above- or below-the-scroll); it is negatively related to distance. Quantitatively, national restaurant chain impressions have a 3.5% higher CTR than other local or independent restaurants; being above-the-scroll (i.e., the top two positions) increases CTR of an impression by 36.7%; increasing distance by one percent decreases CTR by 5.6%, all else being equal. We find a tradeoff between distance and brand popularity, in that users are willing to click on the impression for a market leader, i.e., nationwide brands or popular local businesses, even when it is relatively farther in distance. This tradeoff is giving more favors to national chain when users are located at less familiar areas. In contrast, local businesses suffer in Away search case since users lack the local information of them.

These results are relevant to both researchers and practitioners. On the research side, this is one of the earliest studies to empirically study local search, extending the research framework for general web search to the case of location enabled and disabled local search. Specifically, our analysis sheds light on the interactions between distance, search rank and brand popularity with respect to the impact on click-through performance. Even though our setting does not incorporate location-based advertising, our results provide useful benchmarks for consumer behavior in a location-based advertising setting.

Understanding the nature of this tradeoff is crucial for advertisers using such location-based advertising methods as geo-fencing and local search. On one side, our results on the tradeoff between brand characteristics and distance provides useful guidance for deciding how far to set geo-fencing boundaries. Evaluating the impact of distance on click-through performance in a location-based advertising setting, we

can explicitly calculate the optimal geo-fencing radius for a given advertiser. On the other side, our findings provide managerial implications on the tradeoff between screen position and brand popularity/distance. In local search context, we advise that a popular-brand advertiser insert sponsored ads when its impression position among search results is below a threshold once a consumer located in a geo-fencing area of interest.

While there is ample room to extend the scope of our research in future work, our initial results provide an encouraging first step for a comprehensive understanding of consumer behavior in mobile local search. Our future work will extend the analysis to different keywords and categories. Overall, this initial work provides useful and robust empirical regularities with respect to the effects of rank, distance, brands – and the interactions among them – in mobile local search, and how they affect consumer click-through decisions, with implications for profitable advertising strategies by the mobile platform itself.

Mobile Geo-fence Advertising

In this chapter, we have studied geo-fence advertising by looking at both consumer click-through and conversion behaviors. To understand what are the underling factors affecting consumer decisions, our analysis focuses on the distance between an advertiser establishment and a consumer, and the micro-competition between the advertiser and its rivals within the geo-fence zone. We obtained a rich dataset from a leading location-based marketing agency, and focused our attention to a leading fast food chain, in order to quantify the direct and interactive effects of distance and competition. We found a number of robust empirical regularities. The click-through rate is negatively related to competition but not to distance. Quantitatively, adding one more competitor into a 5-mile radius area of the advertisers' location decreases the click-through rate by 7.5%. On the contrary, distance has a negative impact on conversion rate while competition does not. The estimates showed that the conversion rate drops by 33.2% if the advertiser's store is one more mile away. Moreover, we discovered that population size positively affects click performance but has no effect on conversions. The performance of geo-fence advertising also depends on other consumer characteristics and the categories of delivery apps.

These results are relevant to both researchers and practitioners. On the research side, this is one of the earliest studies to empirically study geo-fence advertising. Specifically, our analysis sheds light on how

competition affect the click-through performance in location-based service. On the practice side, understanding the nature of distance and competition in different stages is crucial for advertisers. The competition between an advertiser and its rivals is critical but commonly ignored in the current practice of optimizing geo-fencing strategies. Evaluating the impact of competition on click-through performance, we can make advertiser better off. Moreover, our findings provide managerial implications on the matches among advertisers and consumer characteristics and publisher categories.

The preliminary results provide an encouraging first step for a comprehensive understanding of consumer behavior in geo-fence advertising. We plan to extend the study in the following two directions. First, we will extend the analysis to different advertisers. It will be very interesting to examine whether the effects of distance and completion on consumer behaviors vary across different advertisers – particularly for higher-valued purchases. Second, we plan to model click-through rate and conversion rate as a system of equations. Thus, the click-through stage and conversion stage can be modeled in a seemingly unrelated setup where two equations have correlated error terms. To sum up, this study provides useful and robust empirical regularities with respect to the effects of distance, competition, and the interactions between them.

REFERENCE

Aaker, D. A. 1991. *Managing Brand Equity*. New York: The Free Press.

Adomavicius, G., A. Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* **17**(6) 734-749.

Agarwal, A., Hosanagar, K., and Smith, M. D. 2011. Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets. *Journal of Marketing Research* **48**(6) 1057-1073.

Anagnostopoulos, A., R. Kumar, M. Mahdian. 2008. Influence and correlation in social networks. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Aral, S., D. Walker. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* **57**(9) 1623-1639.

Aral, S., L. Muchnik, and A. Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* **106**(51) 21544-21549.

Arndt, J. 1967. Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research* **4**(3) 291-29.

Balasubramanian, S. 1998. Mail versus Mall: A Strategic Analysis of Competition between Direct Marketers and Conventional Retailers. *Marketing Science* **17**(3) 181-195.

Balasubramanian, S., R.A. Peterson, S.L. Jarvenpaa, 2002. Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science* **30**(4) 348-361.

Bandura, A. 1971. *Social Learning Theory*. Englewood Cliffs, NJ: Prentice Hall.

Bapna, R., A. Umyarov. 2015. Do your online friends make you pay? A randomized field experiment in an online music social network. Working Paper, Carlson School of Management, University of Minnesota.

Becker, S. 1954. Why an Order Matter? *Public Opinion Quarterly* **18**(3) 271-278.

Belo, R., P.A. Ferreira. 2013. Diffusion and influence over cell phone networks. Working Paper, CMU Heinz School.

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

BIA/Kelsey. 2014a. Franchises spend nearly half of marketing budgets on digital, accessed at http://blog.biakelsey.com/index.php/2014/05/13/franchises-prove-to-be-more-sophisticated-than-the-average-smb/.

_____. 2014b. U.S. Local Media Revenues to reach $139.3 billion in 2015, accessed at http://www.biakelsey.com/Company/Press-Releases/140922-U.S.-Local-Media-Revenues-to-Reach-$139.3-Billion-in-2015.asp.

_____. 2015. Mobile will grab 11.5% of total local media revenues by 2019, accessed at http://www.biakelsey.com/Company/Press-Releases/150422-Mobile-Will-Grab-11.5-Percent-of-Total-Local-Media-Revenues-by-2019.asp.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives* **12**(3) 151-170.

Brainard, S.L. 1993. An empirical assessment of the proximity-concentration tradeoff between multinational sales and trade. *National Bureau of Economic Research* **No. w4580**.

Brown, J.J., P. Reingen. 1987. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research* **14**(3) 350-362.

Brown, J.J., A.J. Broderick, N. Lee. 2007. Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing* **21**(3) 2-20.

Brynjolfsson, E. Dick, A., and Smith, M. 2010. A Nearly Perfect Market? Differentiation versus Price in Consumer Choice. *Quantitative Marketing and Economics* **8**(1), 1-33.

Cai, H., Y. Chen, H. Fang. 2009. Observational learning: Evidence from a randomized natural field experiment. *American Economic Review* **99**(3) 864-882.

Card, D., A.B. Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review* **84**(4) 772-793.

CHD Expert. 2015. CHD Expert Evaluates the Pizza Industry in the United States, accessed at http://www.chd-expert.com/resource-center/chd-expert-evaluates-the-pizza-industry-in-the-united-states.

Chen, P., S. 2008. Dhanasobhon, M.D. Smith. All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon.com. Working paper, Heinz School, Carnegie Mellon University.

Chen, Y., Q. Wang, J. Xie. 2011. Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of Marketing Research* **48**(2) 238-254.

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

Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* **43**(3) 345-354.

Chiang, W. K., Chhajed, D., and Hess, J. D. 2003. Direct marketing, indirect profits: A strategic analysis of dual-channel supply-chain design. *Management Science* **49**(1) 1-20.

comScore. 2014a. 2014 local business search study.

comScore. 2014b. Media metrix multi-platform & mobile metrix.

comScore. 2014c. "The US mobile apps report."

Curry, L. 1978. Demand in the spatial economy: II Homo Stochasticus. *Geographical Analysis* **10**(4) 309-344.

Danaher, B., M.D. Smith, R. Telang, S. Chen. 2014. The effect of graduated response anti-piracy laws on music sales: Evidence from an event study in France. *The Journal of Industrial Economics* **LXII**(3) 541-553.

Danaher, P.J., I.W. Wilson, R.A. Davis. 2003. A comparison of online and offline consumer brand loyalty. *Marketing Science* **22**(4) 461-476.

De Matos, M.G., P.A. Ferreira, D. Krackhardt. 2014. Is viral marketing worth the trouble? Evidence from the Diffusion of the iPhone 3G over a Large Social Network. Working paper, Heinz School, Carnegie Mellon University, available at http://misq.org/peer-influence-in-the-diffusion-of-iphone-3g-over-a-large-social-network.html?SID=ii03luh9lmt6jtps3sn8rdagk1.

Dewan, S., Y. Ho. 2015. Distances and brands in mobile local search analytics. *Proceedings of 2015 Conference on Information Systems and Technology*.

Dewan, S., V. Hsu. 2004. Adverse selection in electronic markets: Evidence from online stamp auctions. *The Journal of Industrial Economics* **52**(4) 497-516.

Dewan, S., J. Ramaprasad. 2012. Music blogging, online sampling, and the long tail. *Information Systems Research* **23**(3) 1056-1067.

Dewan, S., J. Ramaprasad. 2014. Social media, traditional media, and music sales. *Management Information Systems Quarterly* **38**(1) 101-121.

Dhar, V., E.A. Chang. 2009. Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing* **23**(4) 300-307.

Duan, W., B. Gu, A.B. Whinston 2009. Informational cascades and software adoption on the Internet: An empirical investigation. *Management Information Systems Quarterly* **33**(1) 23-48.

Egebark, J, M. Ekstrom. 2011. Like what you like or like what others like? Conformity and peer effects on Facebook. Working paper, Research Institute of Industrial Economics, Stockholm, Sweden.

emarketer. 2015. Advertisers will spend Nnarly $600 billion worldwide in 2015, accessed at http://www.emarketer.com/Article/Advertisers-Will-Spend-Nearly-600-Billion-Worldwide-2015/1011691.

Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* **19**(3) 291-313.

Forman, C., A. Ghose, A. Goldfarb. 2009. Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Science* **55**(1) 47-57.

Gelman A., D. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistics Science* (7) 457–511.

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

Ghose, A., P.G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions* **23**(10) 1498-1512.

Ghose, A., P.G. Ipeirotis, B. Li. 2015. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science* **60**(7) 1632-1654.

Ghose, A., S. Yang. 2009. An empirical analysis of search engine advertising: sponsored search in electronic markets. *Management Science* **55**(10) 1605-1622.

Godes, D, D. Mayzlin, Y. Chen, S. Das, C. Dellarocas, B. Pfeiffer, B. Libai, S. Sen, M. Shi, P. Verlegh. 2005. The firm's management of social interactions. *Marketing Letters* **16**(3) 415-428.

Google. 2013. Mobile path to purchase.

_____. 2014. Understanding consumers' local search behavior.

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* **78**(6) 1360-1380.

Hampton, K., B. Wellman. 2002. Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City Community* **2**(3) 277–311.

Hanson, S. 1980. Spatial diversification and multipurpose travel: Implications for choice theory. *Geographical Analysis* **12**(3) 245-257.

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

Jeziorski, P., S. Moorthy. 2014. Brand effects in search advertising. Working paper, University of California, Berkeley.

Katz, E., P.F. Lazarsfeld. 1995. *Personal Influence*. The Free Press, New York, NY.

Keller, K.L. 1993. Conceptualizing, measuring, and managing customer-based brand equity. *The Journal of Marketing* 1-22.

Laja, P. 2013. *How to Build Websites that Sell: The Scientific Approach to Websites.* Hyperink.

Lechner, M. 2002. Some practical issues in the evaluation of heterogeneous labor market programmes by matching methods. *Journal of the Royal Statistical Society* **165**(1) 59-82.

Lee, K., B. Lee. 2011. An empirical study on quality uncertainty of products and social commerce. *Proceedings of the 13th International Conference on Electronic Commerce.*

Li, X., L. Wu. 2013. Measuring effects of observational learning and Social-Network Word-of-Mouth (WOM) on the Sales of Daily-Deal Vouchers. *Proceedings of the 46th Hawaii International Conference on System Sciences.*

Liu, Y. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing* **70**(3) 74-89.

Lu, Y., B. Gu, Q. Ye, Z. Sheng. 2012. Social influence and defaults in peer-to-peer lending networks. *Proceedings of 33th International Conference of Information Systems*.

Luo, X., M. Andrews, Z. Fang, Z. Phang. 2014. Mobile targeting. *Management Science* **60**(7) 1738-56.

Ma, L., R. Krishnan, A. Montgomery. 2010. Homophily or influence? An empirical analysis of purchase within a social network. Working paper, Heinz School, Carnegie Mellon University.

McKinsey Global Institute. 2011. Big data: the next frontier for innovation, competition, and productivity. 8.

Manski, C.F. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* **60**(3) 531-542.

Meyer, B.D. 1995. Natural and quasi-experiments in economics. *Journal of business & economic statistics* **13**(2) 151-161.

Mizerski, R.W. 1982. An attribution explanation of the disproportionate influence of unfavorable information. Journal of Consumer Research 301-310.

mobilestatistics. 2015. 23 days a year spent on your phone, accessed at http://www.mobilestatistics.com/mobile-news/23-days-a-year-spent-on-your-phone.aspx.

Molitor, D., P. Reichhart, M. Spann, A. Ghose. 2014. Measuring the effectiveness of location-based advertising: A randomized field experiment. Working paper, New York University.

Moretti, E. 2011. Social learning and peer effects in consumption: Evidence from movie sales. *Review of Economic Studies* **78**(1) 356-393.

Mulligan, G.F. 1983. Consumer demand and multipurpose shopping behavior. *Geographical Analysis* **15**(1) 76-81.

Narayanan, S., K. Kalyanam. 2015. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science* (published online).

The New York Times. 2013. Online book clubs: Talk that stays on the page. September 20, 2013.

The Nielsen Company. 2012a. Nielsen Music 360^0. Partial report available upon request and full report available upon license retrieval from The Nielsen Company.

The Nielsen Company. 2012b. Global consumers' trust in "earned" advertising grows in importance, accessed at: http://www.nielsen.com/us/en/press-room/2012/nielsen-global-consumers-trust-in-earned-advertising-grows.html.

Rossi, P.E., G.M. Allenby, R. McCulloch. 2005. *Bayesian Statistics and Marketing*. Wiley.

Salganik, M.J., P.S. Dodds, D.J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**(5762) 854-856.

Schöndienst, V., F. Kulzer, O. Günther. 2012. Like Versus Dislike: How Facebook's like-button influences people's perception of product and service quality. *Proceedings of 33th International Conference of Information Systems*.

Simonson, I., J., Huber, J. Payne. 1988. The relationship between prior knowledge and information acquisition order. *Journal of Consumer Research* (14) 566-78.

Sorenson, A.T. 2007. Bestseller lists and product variety. *The Journal of Industrial Economics* **55**(4) 715-738.

Stahl, K. 1982. Location and spatial pricing theory with non-convex transportation cost schedules. *The Bell Journal of Economics* 575-582.

Starr, M.K., J.R. Rubinson. 1978. A loyalty group segmentation model for brand purchasing simulation. *Journal of Marketing Research* (13) 378-383.

Thumbvista. 2015. Mobile ads drive foot traffic for restaurant franchises, accessed at http://www.thumbvista.com/2015/05/mobile-ads-drive-foot-traffic-for-restaurant-franchises.

Tucker, C. 2008. Identifying formal and informal influence in technology adoption with network externalities. *Management Science* **54**(12) 2024–2038.

Tucker, C., J. Zhang. 2011. How does popularity information affect choices? A field experiment. *Management Science* **57**(5) 828-842.

United State Census Bureau. 2014. American community survey.

Valente, T.W. 1994. *Network Models of the Diffusion of Innovations.* Hampton Press, Inc., Cresskill, NJ.

Wang, J., and C. Chang. 2013. The impacts of online lightweight interactions as signals. *Proceedings of 34th International Conference of Information Systems*.

Yao S., C. Mela C. 2011. A dynamic model of sponsored search advertising. *Marketing Science* **30**(3) 447-468.

YP. 2012. Local insights report, Q2 2012.

APPENDIX

Procedure for Matching

Procedure for Propensity Score Matching at User-Level

We implemented a propensity score matching (PSM) procedure at the user level, with the goal of matching users on their propensity to listen to any given song based on their taste. Each user *User* T_i in the treatment group (who has not listened to Song *j* but has at a friend favoriting it during the burn-in period), we find another *User* C_i in the control group, who: i) has similar tastes as T_i (based on matching the listen profiles), ii) has not listened to Song *j*, and iii) does not have any friend who has favorite the song during the burn-in period. The data construction procedure is detailed as follows:

0. Identify a set of active users during the observation window. Profile the listening behavior of users, during the window 9/1 to 9/21, by constructing a vector incorporating 28 music characteristics and the number of other users they have followed.
1. Determine the treatment group T:
    a. For each *Song j*, identify an active User *i* who has not listened to *Song j* but has at least one friend who has favorited *Song j* during the burn-in period
    b. Pool all such users into the treatment group T.
2. Determine the potential control group PC: pool active users not in T as the potential control group PC.
3. Determine the control group C: Match the propensity of listening to any given song based on the users' taste profiles, using a logit model to predict the propensity to be treated. Match each User T_i in T with a User PC_i in PC with the closest estimated propensity score. Last, we pool these matched users into the control group C.
4. Recover the user-song observations of T & C:
    a. For each Song *j*,
        - reconstruct the User-Song *j* pair if User T_i in the treatment group has a friend who has favorited that Song *j* (see step 1a)

- find the matching control group User C_i, who has a friend who has not favorited that Song j (see step 3)

b. Repeat Step 4a for all songs to construct both treatment and control groups

Procedure for Euclidean Distance Matching at User-Song Level

To supplement our PSM methodology, which accounts for homophily at the user-level, we also implemented EDM, which allowed us to match at the user-song level. The goal of conducting the matching at the user-song level is to match users not only on their likelihood to listen to a given song (which we done with user matching) but also on their likelihood of having a friend favorite the song (i.e. the likelihood of being treated). Therefore, for each song, every *User* T_i who has not listened to the song but has friends favoriting it (our treatment group), we find another *User* C_i who 1) has the similar tastes with T_i, 2) has not listened to the song either but 3) has a friend who is likely to favorite it (our control group). In this process of matching at the user-song level, we essentially have 238 unique treatment user-song pairs with a relatively large potential control group for each of these pair. However, we cannot use PSM to match at the user-song level as we did at the user-level. Recall that in PSM, we estimated the propensity scores for each user based on 28 song characteristics to match users according to their music tastes. At the user-song level of granularity, the user-song pairs have a small number of observations that are treated and thus estimating the logistic regression–the first step in propensity score matching–becomes intractable.

To mitigate this, we use Euclidean distance for the matching process, according to the procedure described below. Again, the goal of this procedure is: 1) to control for homophily, and 2) to match a focal treatment user whose friend has favorited a particular song to a control group user whose friend is likely to favorite that song but has not:

This matching procedure proceeds as follows:

0. Identify a set of active users during the observation window. Profile the listening behavior of users, during the window 9/1 to 9/21, by constructing a vector incorporating 28 music characteristics and the number of other users they have followed.

1. For *Song j*,
    a. Determine the treatment group T_j: the active users who have not listened to *Song j* but have at least one friend who has favorited *Song j* during the burn-in period.
    b. Designate a set of potential control group users PC_j : the rest of the active users have not listened to Song *j*, nor have any friend who has favorited *Song j*.
    c. Calculate the Euclidean distance between each user in the treatment group (T_j) and each user in the potential control group (PC_j) based on the vector of characteristics as described in step 0; we call this the user or the song's profile. For each user in the treatment group, select the three users from the potential control group with the shortest Euclidean distance as the "candidates" for the matched control group user.
    d. Determine the control group C_j: Calculate the Euclidean distance between *Song j*'s profile and each profile of these three candidates' friends. Pick the one of these three candidates whose friend's profile is the closest to the song's profile. Last, each user in T_j has a matched user in the control group (C_j).
    e. For the users in T_j and C_j, recover the set of user-Song *j* pairs, as before.
2. Repeating Step a – e for every song, pool T_j as the treatment group and C_j as the control group.