ONLINE RETAILING IN SPATIALLY DISPERSED OFFLINE MARKETS

Jeonghye Choi

A DISSERTATION

in

Marketing

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

*Signature*_____

David R. Bell, Associate Professor of Marketing

Graduate Group Chairperson

*Signature*_____

Eric Bradlow, Professor of Marketing, Statistics, and Education

Dissertation Committee

Leonard M. Lodish, Professor of Marketing

Edward I. George, Professor of Statistics

Christophe Van den Bulte, Associate Professor of Marketing

Raghuram Iyengar, Assistant Professor of Marketing

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, David R. Bell, for his continued support and advice. I would like to extend my gratitude to my dissertation committee members: Leonard M. Lodish, Christophe Van den Bulte, Raghuram Iyengar, and Edward I. George. I would also like to thank my family and friends for their support throughout my studies.

ABSTRACT

ONLINE RETAILING IN SPATIALLY DISPERSED OFFLINE MARKETS

Jeonghye Choi

Advisor: David R. Bell

This dissertation comprises three essays that study online demand coming from local offline markets. In the first essay, I study two social influence effects reflected in physical proximity and in demographic similarity, respectively, on online demand evolution. As these effects can be time-varying, I specify their dynamics using a polynomial smoother embedded within the Bayesian framework. Using new buyers at Netgrocer.com in Pennsylvania, I find that the proximity effect is especially strong in the early phases of demand evolution, whereas the similarity effect becomes more important with time. In the second essay, I study social influence effects emanating from two types of buyers in the installed base—search buyers, those acquired by online search, and WOM buyers, those acquired by offline word-of-mouth (WOM)—on online demand evolution. Since Internet retailers acquire buyers from multiple locations over time, I allow time-varying parameters to also vary across counties. Using data on new buyers at Childcorp.com, I find that WOM buyers are on average of "better quality", however, substantial variation

in the temporal parameter paths over counties suggests that a third of the markets are better able to breed social influence from search buyers. In the third essay, I examine how online demand in a location is affected by the relative size of the target population holding the absolute size constant. I hypothesize that in regions where this target group is in the minority online category sales will be higher (H_1) and will be relatively price-insensitive (H_2). I further conjecture that online sales of niche brands, relative to popular brands, will be even more responsive to preference minority status (H_3). Finally, I show that niche brands in the tail of the Long Tail sales distribution (Anderson 2006) will draw a greater proportion of their total sales from high preference minority regions (H_4). Sales data from Childcorp.com supports all four hypotheses. This dissertation concludes with a short chapter, briefly discussing the key findings and describing areas for future research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Traditional retailers serve a fixed trading area and "location" is a primary determinant of success. Conversely, Internet retailers can attract consumers over time from a wide geographic area, which means that even spatially-separated consumers can easily utilize the same Internet retailing service. This raises important questions about how demand for Internet retail service is likely to evolve over time and space. In recent years the topic of online demand evolution has received increasing attention from scholars in diverse fields including economics, information systems, marketing, and sociology. Research on shopping behavior on the Internet is, however, still quite limited in comparison to the numerous studies on shopping behavior offline. It is important to understand online retailing, not only because of its ubiquity and scope, but also because it raises issues that are very different to those in offline retailing. In this dissertation, I examine demand evolution for Internet retail service from two different perspectives. First, I study how social influence from the installed customer base drives new online demand over time and space (essays 1 and 2). Second, I show how the "preference minority status" of target customers in local markets affects the spatial distribution of online demand (essay 3).

A large body of research assumes that social influence plays an important role in generating demand. Studies directly relevant to the first essay offer two key findings.

First, all else equal, social influence is more likely when agents are geographically proximate (Bell and Song 2007; Manchanda, Xie, and Youn 2008). Second, the likelihood of social influence is higher among agents who are "similar" (Albuquerque, Bronnenberg, and Corbett 2007; Yang and Allenby 2003). In the first essay in Chapter 2, I contribute to the literature by uncovering the *time-varying* effects of social influence as captured by physical proximity and demographic similarity in driving new demand for Internet retail service. The model is applied to new buyers at Netgrocer.com acquired from its inception in May 1997 to January 2001 and is calibrated on forty-five months of data that span all 1,459 zip codes in Pennsylvania. I find that the proximity effect is especially strong in the early phases of demand evolution, whereas the similarity effect becomes more important with time. Over time, new buyers are increasingly likely to emerge from new zip codes beyond the "core set" of zip codes that produce the early new buyers, and spatial concentration declines. I explore managerial implications stemming from the findings through a hypothetical "seeding" experiment.

I study the social influence effect based on physical proximity further in the second essay as the proximity effect is dominant in big markets with large online demand.[1] Prior studies relevant to the second essay find that different acquisition modes bring different qualities of buyers to a firm (Lewis 2006; Villanueva, Yoo, and Hanssens 2008). This implies that the *mix* of buyer types in the installed base has a critical influence on future demand for Internet retail service to the extent that different types of buyers exert

[1] The first and second essays are different in at least three aspects. First, the two essays examine two different Internet retailers, Netgrocer.com and Childcorp.com. Second, each examines different local markets. While the first essay focuses on the zip codes in Pennsylvania, the second essay looks into the zip codes in metropolitan areas throughout the 48 contiguous states. Third, the data periods in two essays are four years apart. Netgrocer.com data are from May 1997 to January 2001 whereas sales data at Childcorp.com are collected from its inception in January 2005 to March 2008.

different social influence. The two main acquisition modes for Internet retailers are online search (hereafter, "search") and offline word-of-mouth (hereafter, "WOM"), and thus I construct two social influence measures based on physical proximity to each type. In the second essay in Chapter 3, I contribute to the literature by contrasting, both *spatially* and *temporally*, the effects of social influence from the existing buyers acquired by search and WOM (hereafter, "search buyers" and "WOM buyers"). The model is applied to new buyers at Childcorp.com and is calibrated on 13 quarters of data from the first quarter in 2005 to the first quarter in 2008 that span 4,532 zip codes in metropolitan areas.[2] I find that not all customers are created equal in terms of their ability to bring future new customers to the firm. Customers acquired by WOM are on average of "better quality" in the sense that the fixed (or average) social influence parameter paths for the installed WOM customers are larger than those for the installed search customers. However, substantial variation in the temporal parameter paths over counties suggests that not every market favors WOM acquisitions. The superiority of one acquisition channel over the other varies markedly over space.

The third essay in Chapter 4 examines how online demand in a target region is affected by the relative, rather than absolute, size of the target population. Local stores face trading area and retail space constraints, so the products they offer tend to cater to the tastes of the local majority. Consumers whose preferences are dissimilar to the majority in trading area—*preference minorities*—are more likely to be relatively underserved by offline stores. I explain why Internet retailers draw more sales from regions that contain them, holding the *absolute* number of target buyers constant, and

[2] For reasons of confidentiality I refer to this leading Internet retailer by the *nom de plume*, "Childcorp.com".

why members of the preference minority are relatively price-insensitive. Furthermore, the effect is exacerbated for niche products, relative to popular products. I show that niche products in the tail of the Long Tail (Anderson 2006) draw a *greater proportion* of their total online demand from high preference minority regions.[3] I discuss implications for retailing theory and practice.

The three essays in Chapter 2 to Chapter 4 are followed by a general discussion and conclusion in Chapter 5 including the key findings and the implications for both researchers and practitioners, and possible future research opportunities stemming from this dissertation.

[3] "The Long Tail" (Anderson 2006) refers to the phenomenon where an Internet firm offers an almost *unlimited product assortment* as the product stocking constraint is relaxed, and thus small sales levels over a large number of products account for substantial sales in aggregate.

Chapter 2

Bayesian Spatio-Temporal Analysis of Imitation Behavior across New Buyers at an Online Grocery Retailer

2.1 Introduction

The Internet has reduced customer access costs for firms and facilitated long-range connections among consumers. While "location" is a primary determinant of success for traditional retailers (e.g., Huff 1964), Internet retailers are not subject to the constraint of physical location. They can attract consumers over a wide geographic area which means that even physically-separated consumers can easily utilize the same Internet retailing service. This raises important questions about how demand at Internet retailers is likely to evolve not only through time but also over space. In particular, how consumers may imitate their peers in their adoption behavior, and ultimately, what firms might do to expedite the demand process.

A large body of research assumes, in general, that imitation behavior plays an important role in generating demand (see, e.g., Bass 1969; Hauser, Tellis, and Griffin 2006). Studies that are directly relevant to our research offer two key findings. First, all

else equal, imitation among agents is more likely when they are geographically proximate. Researchers have found consumption externalities for prescribing physicians (Manchanda, Xie, and Youn 2008), competitive effects among retailers in new brand rollout (Bronnenberg and Mela 2004), and possible emulation in trial behavior for an Internet retailer (Bell and Song 2007). Second, the likelihood of imitation is higher among agents who are "similar". These include academics with overlapping research interests (Rosenblat and Mobius 2004), firms with comparable cultural profiles (Albuquerque, Bronnenberg, and Corbett 2007), and individuals with common overall socio-demographic characteristics (Yang and Allenby 2003).

We contribute to the literature by analyzing the space-time diffusion process as a function of both factors (i.e., proximity and similarity), identifying the relative importance of each over time, and relating our findings to an Internet retailer's new buyer acquisition strategy.

Figure 2.1 motivates the underlying phenomenon. It shows the cumulative number of new buyers at Netgrocer.com in each zip code in the state of Pennsylvania recorded in fifteen-month intervals from the inception of the service in May 1997 through January 2001. Three interesting patterns appear. First, the evolution of new buyers seems to have started from two distinct locations and spread to nearby areas (these "hot spots" are Philadelphia and Pittsburgh, the two major cities in Pennsylvania). Second, the pool of buyers within smaller disaggregate "neighborhoods" intensifies over time. Third, as time progresses, the adopting group expands throughout Pennsylvania such that later areas of sales are physically distant from earlier ones; as a result, the spatial concentration of the new buyers decreases over time. To analyze the data in Figure 2.1, we formulate a

dynamic Bayesian spatio-temporal Poisson model (e.g., Knorr-Held and Besag 1998), and specify the adoption rate for each region at each time period as a function of imitation effects based on proximity and similarity, along with other locally-defined covariates. We utilize a conventional distance-based proximity measure and a demographic similarity metric that mirrors approaches in Rosenblat and Mobious (2004) and Albuquerque, Bronnenberg, and Corbett (2007). To produce efficient estimates of the time-varying coefficients for these variables, we embed a polynomial smoother within our Bayesian model using a random walk prior (Angers and Delampady 1992; Kalyanam and Shively 1998; Wahba 1978; Wedel and Zhang 2004).

Applying our model to the spatio-temporal evolution of new buyers at Netgrocer.com yields three new insights into how demand evolves for an Internet retailer that is geographically unconstrained. First, we find that geographic proximity has the stronger initial impact on the rate at which new buyers are acquired, but that its relative importance weakens with time. Long term viability is therefore unlikely to be secured through local appeal in "hot spots" alone. Second, imitation based on demographic similarity, independent of geographic proximity to the preceding buyers is relatively unimportant early on but as time progresses it accounts for a greater number of new buyers that emerge from spatially-dispersed places. That is, places that lack sufficient density to be served through conventional means, but that on average share characteristics with regions containing earlier adopters. This provides a rationale for the decline in spatial concentration of new buyers. The temporal ordering of the importance of the two components—geographic proximity first and demographic similarity second—holds controlling for differences in observed local characteristics (including access to the

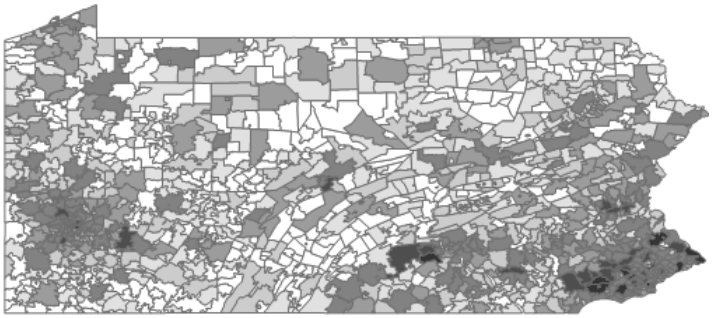
Internet), and unobserved heterogeneity in the adoption rate. Third, we follow Libai, Muller, and Peres (2005) and use "market seeding" to illustrate possible managerial implications stemming from these results. An initial focus on populous regions should be balanced against acquisition of more remote and dispersed customers.

Figure 2.1: Spatio-Temporal Evolution of New Buyers in Pennsylvania

(a) Cumulative Number of New Buyers in July 1998

(b) Cumulative Number of New Buyers in October 1999

(c) Cumulative Number of New Buyers in January 2001



Our research is subject to the following caveats. First, for reasons of parsimony, data availability, and managerial value, we focus on region-level behavior, rather than individual behavior, per se. Second, the main purpose of the model is to provide a descriptive analysis of proximity and similarity effects. We do not attempt to build a forecasting model as this would require a substantially different approach. Lastly, the seeding analyses using the imitation coefficients are the best-case scenario given the data and intended to be illustrate the potential benefit of the *proximity-and-similarity-based* strategy.

The paper is organized as follows. The next section summarizes extant literature that employs geographic proximity and demographic similarity as proxies for imitation behavior in spatial demand analysis. The following section describes the data and key summary statistics, and the subsequent section specifies our Bayesian spatio-temporal model. We then report our substantive empirical findings. The concluding section outlines a hypothetical seeding experiment and discusses implications for Internet retailers and for future research.

2.2 Background Literature

We first present selected empirical evidence from articles in marketing, economics, and sociology that develop proxy measures of geographic proximity and demographic similarity (e.g., among individuals, regions, and firms), and find evidence for imitation behavior.

Geographic Proximity. Proximity-based imitation or "the local neighborhood effect" is largely viewed as arising from either direct social interactions or local emulation among near neighbors. Standard empirical approaches incorporate measures that proxy for imitation, or, more broadly, social interactions among physically close individuals. In Goolsbee and Klenow (2002), the proportion of local households owning computers is used to show that individuals in areas with a high proportion of computer ownership are more likely to become first time buyers, even after controlling for personal traits and local environments. Forman, Ghose, and Wiesenfeld (2008) find that online book sales in a local market are not only associated with the overall disclosure level of user identity-descriptive information, but also amplified when disclosure comes from reviewers residing in the same locality.

In addition to being measured through proportions, proximity effects can also be investigated using information on pair-wise distances or contiguity. Bronnenberg and Mela (2004) employ measures of this sort and find emulation effects among local retailers—new product rollout is influenced by product decisions made by local competitors. Bell and Song (2007) find that new trials of an Internet retailer are related to prior trials in proximate regions.

None of the above studies measures imitation or social interaction directly. Instead, the observed prior behavior of physically close "neighbors" is used to create measures that, in turn, influence the probability of later action by another individual, firm, or region of interest. Statistically significant effects, in the presence of other controls, are taken as corroborating evidence. Our approach follows this precedent and uses spatially-derived proxies to account for the geographic proximity effect.

Demographic Similarity. Fischer (1978) suggests that a resident of Los Angeles has a greater chance of coming into contact with someone from Chicago than with someone from Springfield, even though both Illinois locations are approximately the same physical distance from Los Angeles. This underscores the idea that the propensity for individuals to interact with each other and/or to imitate each other might not be accounted for solely by physical distances. In line with this idea, many researchers have extended the "neighborhood" construct in ways that depart from a specification based on physical locations. Van Alstyne and Brynjolfsson (2005), for example, point out that "neighborhoods" can be shaped by many dimensions including interests, preferences, and member characteristics. One might expect that individuals agglomerating either in online communities or through their revealed preferences for certain online businesses, should exhibit some homogeneity along demographic lines (e.g., by criteria such as occupation, education levels, income, or ethnic grouping).

The way in which "similarity" is measured is an important empirical and conceptual issue. Agrawal, Kapur, and McHale (2008), for instance, define individual-level social proximity using a co-ethnicity indicator and find a substitution effect between social and spatial proximities. Social proximity provides greater benefit for inventors who are not co-located, whereas spatial proximity does for those who are. Rosenblat and Mobius (2004) define economists' "types" according to academic interests and find that the Internet led to narrower collaborations, e.g., labor economists are now less likely to write with economic historians and more likely to co-author with labor economists who are physically distant. Yang and Allenby (2003) study automobile choice and define individuals who share similar demographic profiles as "demographic neighbors." A

model that accounts for choices by both types of neighbors (i.e., demographic and geographic) is preferred to ones that account for either alone.

Other studies have investigated region-level similarity. Conley and Topa (2002) examine spatially-clustered unemployment rates in Chicago. Social networks are defined separately for physical distance, race and ethnicity, and occupation, using Euclidean distances of the corresponding regional compositions across census tracts. The effects of physical distance and occupation are significant, whereas the effect of race and ethnicity is not. Albuquerque, Bronnenberg, and Corbett (2007) study of ISO certification diffusion across countries and find that diffusion of ISO9000 is driven by proximity and trade-based similarity, whereas diffusion of ISO14000 is driven by proximity and cultural similarity. Building on these studies and following Rosenblat and Mobius (2004) and Van Alstyne and Brynjolfsson (2005) in particular, we define our similarity measure according to region "types" based on socio-demographic characteristics.

Summary. Prior research demonstrates that geographic proximity and demographic similarity drive imitation behavior. These studies say relatively little, however, about how such effects evolve over time. Since different forms of imitation exert different degrees of influence at various stages of the adoption cycle, analyzing their effects in static rather than intertemporal settings may not provide a complete picture of their influence. In this paper, we aim to focus on the temporal aspects of geographic proximity and demographic similarity and understand their dynamic influences in driving adoptions of the online retailer.

2.3 Data

New Buyers. We obtained monthly transaction data for new buyers at Netgrocer.com in the state of Pennsylvania from the inception of the service in May 1997 through the end of January 2001. During this period orders were shipped from a warehouse in New Jersey via FedEx, and customers were charged a fixed shipping fee. The customer file records the order month and shipping zip code for each transaction. To understand how demand evolution varies over space and over time, we consider the number of new buyers after aggregating spatially and temporally. The final (bottom right) map in Figure 2.1 shows considerable spatial dispersion in the distribution of cumulative new buyers. Figure 2.2 panel (b) highlights the time dimension of the raw data. It shows that while the overall number of new buyers across zip codes is generally increasing through the forty-five month period, there is substantial variability in the overall trend.

Table 2.1: Summary Statistics for the Number of New Buyers

	Mean	**Std Dev**	**Min**	**Max**
May, 1997	.001	.037	.000	1.000
Oct, 1997	.008	.090	.000	1.000
Mar, 1998	.055	.282	.000	3.000
Aug, 1998	.198	.631	.000	8.000
Jan, 1999	.065	.322	.000	5.000
June, 1999	.083	.350	.000	6.000
Nov, 1999	.212	.602	.000	7.000
Apr, 2000	.220	.607	.000	5.000
Sep, 2000	.219	.823	.000	22.000
Jan, 2001	.235	.802	.000	19.000

Next, we consider the space-time path of the raw data in Figure 2.1 in more detail. Table 2.1 shows summary statistics for the number of buyers per zip code in five-month intervals. The mean number of new buyers per zip code increases over time, but so does the variability across zip codes. That is, the spatial concentration of new customers appears to decrease over time. To examine this more formally, we compute the Getis-Ord G^* statistic (Getis and Ord 1992) each month. The decay in localized concentration of demand supports the observation that, over time, the distribution of new buyers is expanding over space. The considerable spatial and temporal variation in the raw data underscores that when building our model, we must carefully control for regional and temporal baseline effects to accurately measure the demand effects due to imitation.

Regional Characteristics. The data for the imitation proxy variables and the direct measures of regional heterogeneity are assembled from three sources: (1) the 2000 US Census, (2) ESRI retailing statistics (esri.com), and (3) the Federal Communications Commission (FCC) broadband access survey. To create empirical measures of local presence for supermarkets and general merchandisers (e.g., Wal-Mart) we count the number located within the focal zip code, and the first- and second-order contiguous neighbors. We then compute store density by store type, based on the land area. Since warehouse clubs are less common we use a binary indicator for presence within the focal, first-, or second-order contiguous zip codes. Table 2.2 provides descriptions and summary statistics for all zip code level variables. For ease of exposition, the variables are classified as pertaining to region-level: (1) local environment, (2) household characteristics, (3) access to retail services, and (4) access to the Internet.

The FCC estimates the number of Internet service providers (ISPs) in each region, however, these data are known to be approximate. Some ISPs fail to report their services and others report a presence in zip codes on the basis of a single customer. Moreover, the data were collected at four discrete time periods only (December 1999, June 2000, December 2000, June 2001), three of which are covered by our transaction data. Following the suggestion of Wand (2003), we therefore employ a low-rank thin plate spline smoother to improve the FCC data, and provide the complete details in Appendix in Section 2.7.1.[4] In addition, since the timeframe of the broadband access data does not coincide perfectly with the Netgrocer.com data, we impute part of the missing data using a linear interpolation (see also Bell and Song 2007).

We assess and verify the appropriateness of our approach with reference to additional external sources, including prior literature and alternative data collected in the Current Population Survey (CPS).[5] Application of linear interpolation and spatial smoothing creates a zip code and time-period specific measure which we call "Broadband Access." This control for access to the Internet is important in order to help rule out the alternative hypothesis that space-time evolution of Netgrocer.com new buyers simply mimics the diffusion of Internet access.

[4] We also estimated our model using non-smoothed broadband data and obtained qualitatively similar results.

[5] Household level Internet usage data were collected as supplementary data in the Current Population Survey (CPS) from 8,162 national zip codes in October 1997, December 1998, August 2000, and September 2001. Although the CPS data match nicely with the time period for the Netgrocer.com data, they include only 670 (46%) of the zip codes in Pennsylvania. Therefore, we utilize the spatially-smoothed "Broadband Access" variable derived from the FCC data as this measure can be constructed for all 1,459 zip codes in Pennsylvania. The average zip code-level correlations between the CPS data and smoothed "Broadband Access" are 0.95 for the total US sample of 8,162 zip codes and 0.97 for the 670 Pennsylvania zip codes. This suggests that the interpolated "Broadband Access" variable reflects the temporal growth pattern of household-level Internet usage present in the CPS data. Moreover, Bell and Song (2007) demonstrate that a measure constructed from the FCC data is empirically superior to one developed from the CPS data alone.

Table 2.2: Summary Statistics for Zip Code Characteristics

Variable	Description	Mean	Std Dev
Local Environment			
Population	Total population	8391.600	11149.400
Population Density	Population density	1298.799	3117.592
Population Growth	Annual population growth rate from 2000 to 2004	0.004	0.011
Home Value	% of homes valued at $250,000 or more	0.060	0.108
Urban Housing	% of houses with 50 units or more	0.018	0.056
Land Area	Area in square miles	30.607	35.511
Household Characteristics			
Asian	% of Asians	0.008	0.016
Black	% of Blacks	0.038	0.112
White	% of Whites	0.938	0.130
College	% with bachelors and/or graduate degree	0.370	0.144
Elderly	% aged 65 and above	0.156	0.041
Wealthy	% of households earning $75,000+	0.165	0.118
Access to Retail Services			
Density General	Density of general stores within the second order neighboring zip codes	0.107	0.251
Density Supermarket	Density of supermarkets within the second order neighboring zip codes	0.224	0.393
Presence Warehouse	Presence of warehouse clubs within the second order neighboring zip codes	0.245	0.430
Access to the Internet			
Broadband Access	Number of high-speed Internet service providers		
Dec, 1999		1.784	1.320
June, 2000		2.060	1.749
Dec, 2000		2.940	2.665
June, 2001		2.840	2.773

Finally, it is important to note that during the period of data collection, Netgrocer.com was not involved in any significant marketing activities in Pennsylvania; thus, this dataset offers a unique opportunity for us to assess imitation effects across space and time, free of explicit marketing interventions. While we cannot therefore comment on the relationship between local marketing efforts and demand, we can assess the equally important relationship between local characteristics and demand—a relationship of increasing interest (see Forman, Ghose, and Goldfarb 2008; Pauwels and Nelsin 2008; Waldfogel 2007).

2.4 Measures and Model

2.4.1 Measures of Proximity and Similarity

Competition between Netgrocer.com and offline alternatives is local so region-level (zip-code) sales are of particular managerial relevance and data that describe regions are widely available and generally reliable. Hence, our proximity and similarity measures are defined with respect to regions (see also Avery et al. 2008; Brynjolffson, Hu, and Rahman 2008) Moreover, individual-level neighbor covariate information is neither available nor practical to work with. In our model specification, exogenous definition of "neighbors" at the region (zip code) level, and influence from the lagged cumulative behavior of neighbors are used in order to help mitigate the well-known "reflection problem" (Manski 1993; 2000). Manski (1993) emphasizes that in order to claim imitation effects, two alternatives—contextual (exogenous) effects and correlated

effects—should be ruled out. With respect to contextual effects, it is unlikely that some unique exogenous feature of neighboring regions is systematically influencing trial of new buyers in the focal region. Correlated effects—where the number of new buyers in the focal region is influenced by a similarity in institutional constraints—are also unlikely given our controls for Internet access, retail store availability, and so forth.[6]

We apply standard approaches from the literature that define neighborhood relationships through the use of weighting matrices (Anselin 1988; Bell and Song 2007; Bronnenberg and Mela 2004; Yang and Allenby 2003). Specifically, we employ two such matrices: the matrix *G* captures across-region *geographic* proximity and the matrix *D* captures across-region *demographic* similarity. For ease of exposition, assume there is a finite number of zip codes, *n*, such that all pair-wise relations can be summarized by an $n \times n$ weighting matrix, *G* (*D*), in which each nonnegative element, G_{ij} (D_{ij}) denotes the degree of geographic (demographic) "closeness" of region *j* to region *i* . Each weighting matrix is symmetric and row-normalized (row-normalization takes into account relative closeness among neighbors). We also assume that the neighbor relationships do not change over time, as is standard in the previous literature.

Geographic Proximity (*G*). Our measure of across-region proximity is assumed to be an inverse function of the physical distance in miles, d_{ij} ,

[6] Manski (1993, p. 532-537) provides relevant conditions for identification and estimation of endogenous effects. Possible correlated effects are unlikely for the following reasons. First, Netgrocer.com did not conduct significant marketing activities during the data period. Second, access to the Internet and to local retailers is controlled for in our model. Also, regional and temporal baselines account for region and time specific shocks. While spatially (and/or demographically) correlated tastes might drive results of imitation behavior, our rich data and specification make this more unlikely than in much of the existing literature. Thus, we have made progress toward addressing the reflection problem but we cannot entirely rule it out. We thank an anonymous reviewer for these observations.

$$(2.1) \qquad G_{ij} = \begin{cases} \exp(-\Delta d_{ij}), i \neq j \\ 0 \qquad\quad , i = j \end{cases}$$

Following Yang and Allenby (2003), we further assume Δ to be equal to one.[7] The distance-based measure helps control for the fact that different zip codes vary greatly in land area and number of contiguous neighbors. Alternative proximity matrices based on shared boundaries and contiguity information are considered in Appendix in Section 2.7.2.

Demographic Similarity (*D*). Unlike with the measures of physical proximity, there is no single widely-used and straightforward approach to defining similarity. We presume that shared socio-demographic characteristics across regions serve as a proxy for similarity (see Conley and Topa 2002). In other words, if the characteristics of two regions are alike, these regions are more likely to imitate each other, everything else constant. We therefore focus on observable characteristics that previous studies have shown to be correlated with levels of imitation; namely, education, income, age, and ethnicity, and their corresponding subcategories (e.g., Howard, Raine, and Jones 2001; Katz, Rice, and Aspden 2001; Van Alstyne and Brynjolfsson 2005).

The US Census reports zip-level information on the percentages of residents in the following educational attainment categories: (1) below high school completion, (2) completed high school, but no university degree, (3) university degree holder, and (4) graduate degree holder. Similarly for income: (1) below the poverty line, (2) medium

[7] This assumption is made for reasons of computational tractability and consistency with the previous literature (e.g., Claude 2002; LeSage and Pace 2005; Yang and Allenby 2003). In order to demonstrate that our empirical findings are robust to this assumption, we defined two additional proximity measures with an inverse function of *half* (Δ=0.5) and *twice* (Δ=2) the geographic distance, and re-estimated the models. Both measures provide consistent model estimates and thus the same qualitative insights. We thank an anonymous reviewer for suggesting this check.

income, and (3) income in excess of $75,000 per year. Age categories are: (1) up to 20 years old, (2) 21 to 40, (3) 41 to 65, and (4) more than 65 years old. Ethnicity is reported for each region according to the percentage of Asians, Blacks, Hispanics, and Whites living there. Following Rosenblat and Mobius (2004) and Van Alstyne and Brynjolfsson (2005), we define "profile vectors" that measure the extent of overlap between two regions. The socio-demographic profile vectors have a total of fifteen elements (four for education, age, and ethnicity, and three for income). Pair-wise similarity measures are defined as

$$D_{ij} = \begin{cases} \sum_k \min(v_{ik}, v_{jk}), & i \neq j \\ 0, & i = j \end{cases} \tag{2.2}$$

where v_{ik} is the k-th element of the socio-demographic vector of region i; i.e., we sum the minimum values, based on the element-wise comparisons across two socio-demographic vectors for all $k = 1, 2, \ldots, 15$ elements of their socio-demographic profile. As in the case of physical proximity, two alternative measures of demographic similarity are defined in Appendix in Section 2.7.2.

2.4.2 A Bayesian Spatio-Temporal Model of New Buyers

Given the sparseness of the adoption data (see Table 2.1 and Figure 2.1) our model must take into account significant sampling error in order to accurately estimate the role of imitation behavior. Towards this end, we specify our model in two levels, as is standard in Bayesian generalized linear models (Gelman et al. 2003). In the first level, we assume that the number of new buyers in zip code i at time t follows a Poisson

distribution with (latent) rate parameter λ_{it}; we then model λ_{it} as a function of imitation behavior and other controls. Formally, we specify

(2.3) $$y_{it} \sim Poisson(\lambda_{it})$$

where y_{it} denotes the number of new buyers in zip code *i* during month *t*.

We justify the Poisson assumption in equation (2.3) on both theoretical and empirical grounds. In Appendix in Section 2.7.3 we outline a mathematical argument (adapted from Knorr-Held and Besag 1998 and Ross 1996) that the Poisson approximation is valid under the assumption that adoption is sparse and within-period imitation is limited. We show empirically in the next section using a posterior predictive check (Gelman et al. 2003), that the Poisson distribution provides an excellent fit to the raw data. Finally, the Poisson distribution has been used in other instances where events are rare, e.g., to model the spread of new species (Wikle and Hooten 2006), or the number of new patients infected by a rare disease (Knorr-Held and Besag 1998).

Next, in the second level, latent adoption rates λ_{it} are modeled as a function of region-level characteristics, temporal baseline effects, and geographic and demographic imitation effects

(2.4) $$\log(\lambda_{it}) = \log(n_{it}) + \gamma_i + \zeta_t + \beta_t^W z_{it} + \beta_t^G G_{(i)} \vec{z}_t + \beta_t^D D_{(i)} \vec{z}_t + \varepsilon_{it}$$

(2.5) $$\gamma_i = \vec{x}_i' \vec{\tau} + \tilde{\gamma}_i, \quad \tilde{\gamma}_i \sim N(0, \sigma_\gamma^2)$$

(2.6) $$\beta_t^W, \beta_t^G, \beta_t^D \geq 0 \quad \forall t$$

where n_{it} denotes the number of people in region *i* yet to try the service at time *t*, and serves as an offset variable (Agresti 2002; Rabe-Hesketh and Skrondal 2005). μ_i and ζ_t

are regional and temporal baselines, respectively. The regional baseline, γ_i, is comprised of two terms: observed heterogeneity explained by $\vec{x}_i'\vec{\tau}$, a vector of standardized region-level characteristics and the corresponding coefficients vector, and remaining unobserved heterogeneity captured by $\tilde{\gamma}_i$.[8] $G_{(i)}$ and $D_{(i)}$ denote the *i*-th rows of the matrices *G* and *D*, respectively, while z_{it} as denotes the (log-) cumulative number of buyers in region *i* prior to time *t*. The coefficients, β_t^W, β_t^G and β_t^D, denote the strength of *within*-region imitation (*W*), *across*-region imitation due to *geographic proximity* (*G*), and *across*-region imitation due to *demographic similarity* (*D*), respectively. The error terms, ε_{it} are assumed to be independent and normally distributed with mean 0 and variance σ_ε^2, allowing for over-dispersion.

We are interested in the final three terms for imitation in equation (2.4). $\beta_t^W z_{it}$ represents the within-zip code imitation effect due to prior buyers in the same zip code. The row vector $G_{(i)}$ ($D_{(i)}$) measures geographic (demographic) "closeness" of region *i* to all other regions (see equations (2.1) and (2.2)). Post-multiplication by the vector of neighbors' cumulative and lagged numbers of new buyers (i.e., $G_{(i)}\vec{z}_t$ and $D_{(i)}\vec{z}_t$) produces a scalar variable that captures the aggregate time-varying influence of geographic and demographic neighbors on region *i* at time *t*. The parameters β_t^G and β_t^D capture imitation effects based on geographic proximity and demographic similarity, respectively.

[8] One can specify that these random effects are spatially correlated, e.g., using a CAR (Conditional AutoRegressive) formulation (Cressie 1993). Albuquerque, Bronnenberg, and Corbett (2007), however, found that incorporating spatially-correlated errors do not improve their model's performance. Thus, we retain the i.i.d. specification. One can also specify a more general model with demographically correlated random effects, e.g., as a joint distribution across zip codes with correlation in demographic space. We thank the AE for this observation.

Given the nature of our data we are unable to disentangle—except in an ex post analysis of marginal effects—whether current users within a region are propagating positive or negative information about Netgrocer.com. Since only non-perishable branded products (e.g., paper products, canned food, etc.) were sold during the data collection period, potential new buyers should have been able to assess product quality ex ante. Prices were also known. Hence, negative information was most likely to relate to delivery, which was handled by Federal Express. Therefore, we postulate the more cumulative buyers there are, the greater the number of new buyers that will emerge. Equation (2.6) reflects this restriction which assumes that all three imitation coefficients are non-negative. These restrictions are of a theoretical nature only; they play no role in the actual empirical application. The estimated imitation coefficients are bounded far away from making this restriction irrelevant (in the Conclusion we sketch an extension of our model that could accommodate both positive and negative influence).

2.4.3 Prior Specification and Smoothing

The main substantive goal of this research is to understand the relative magnitudes of proximity- and similarity- based imitation effects, and how they vary over time. From a model estimation standpoint, our goal is to obtain efficient estimates for β_t^W, β_t^G and β_t^D. To this end, we embed a "polynomial smoother", commonly used in Frequentist nonparametric statistics, into our Bayesian model (Angers and Delampady 1992; Kalyanam and Shively 1998; Wahba 1978; Wedel and Zhang 2004). A smoother allows us to take observations from neighboring time periods into account when making

inference about a certain time period. When making inference about an estimate at time *t*, we take into account information from periods *t*-1, *t*-2, …, and also *t*+1, *t*+2, … in polynomially decreasing weights, thereby allowing us to borrow strength from other periods to improve estimation efficiency. The smoother produces estimates that vary smoothly over time, which is consistent with our intuition about how imitation coefficients should in fact evolve. It also provides several key statistical advantages (see Appendix in Section 2.7.4).

We specify a Gaussian random walk prior on our time-varying coefficients. For $t > 1$,[9]

(2.7) $$\zeta_t \sim N(\zeta_{t-1}, \sigma_\zeta^2)$$

(2.8) $$\beta_t^W \sim N(\beta_{t-1}^W, \sigma_W^2)$$

(2.9) $$\beta_t^G \sim N(\beta_{t-1}^G, \sigma_G^2)$$

(2.10) $$\beta_t^D \sim N(\beta_{t-1}^D, \sigma_D^2)$$

Standard proper conjugate priors are specified for all the other parameters in the model. An MCMC procedure is used to sample from the posterior distributions (see Appendix in Section 2.7.5).

[9] Giving these temporal parameters independent diffuse normal distributions (i.e., $N(0,100^2)$) is undesirable for two reasons. First, since these parameters measure the strength of imitation over time, one would expect them to vary smoothly over time, instead of jumping around in a rather haphazard manner. Second, the independence assumption of the prior distributions fails to "borrow strength" across the different time periods when estimating these parameters, and hence reduce estimation efficiency (Rossi and Allenby 2003). This latter aspect is particularly important for our data, which are fairly sparse with small numbers of buyers over space and time (see Table 2.1).

2.5 Empirical Findings

We first compare our model to reduced models and demonstrate the adequacy of our model in describing both the spatial and temporal dimensions of the data. We then present time-varying imitation parameter estimates (see Appendix in Section 2.7.6 for other control variables), interpret them, and discuss implications for market seeding and why the spatial concentration of new buyers declines over time.

2.5.1 Model Fits and Validation

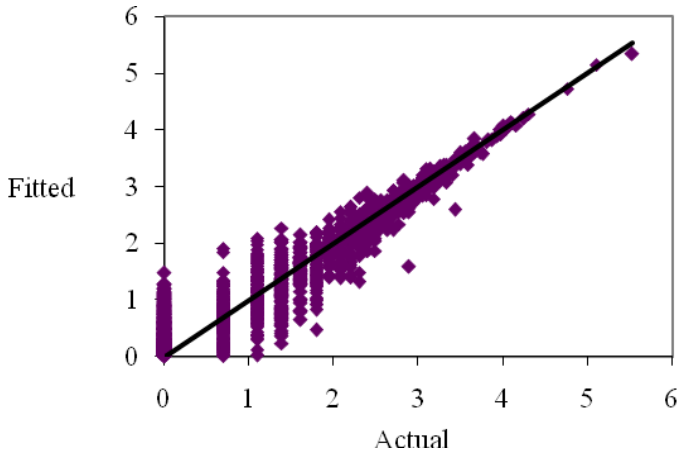
The full model is compared, using marginal log-likelihood (Chib 1995; Chib and Jeliazkov 2001), to reduced models that "turn off" imitation effects based on proximity and similarity. The marginal log-likelihood for the full model with the proximity and similarity effects is -70,324, which is higher than for the model with neither effect (i.e., $\beta_t^G = \beta_t^D = 0$), the model with proximity only (i.e., $\beta_t^G = 0$), and the model with similarity only (i.e., $\beta_t^D = 0$).[10] To assess overall fit to the raw data (y_{it}) we also compare the actual distribution of y_{it} to the posterior predictive distribution of $\hat{y}_{it}$ (Gelman et al. 2003). Figure 2.2 panels (a) and (b) indicate an adequate model fit on the spatial and temporal dimensions after aggregating over time and space, respectively. Importantly,

[10] We also compared the full model and three reduced models using the procedure in Newton and Raftery (1994) and obtained the same qualitative results. We thank an anonymous reviewer for suggesting Chib (1995).

accurate spatial fit is obtained not only in the regions with high demand, but also in the spatially-distant regions with relatively sparse sales.

Figure 2.2: Aggregate Model Fits over Space and Time

(a) Fitted Versus Actual Number of New Buyers in Log Transformation by Zip Code (aggregated over time (months))



(b) Fitted Versus Actual Number of New Buyers over Time (aggregated over space)

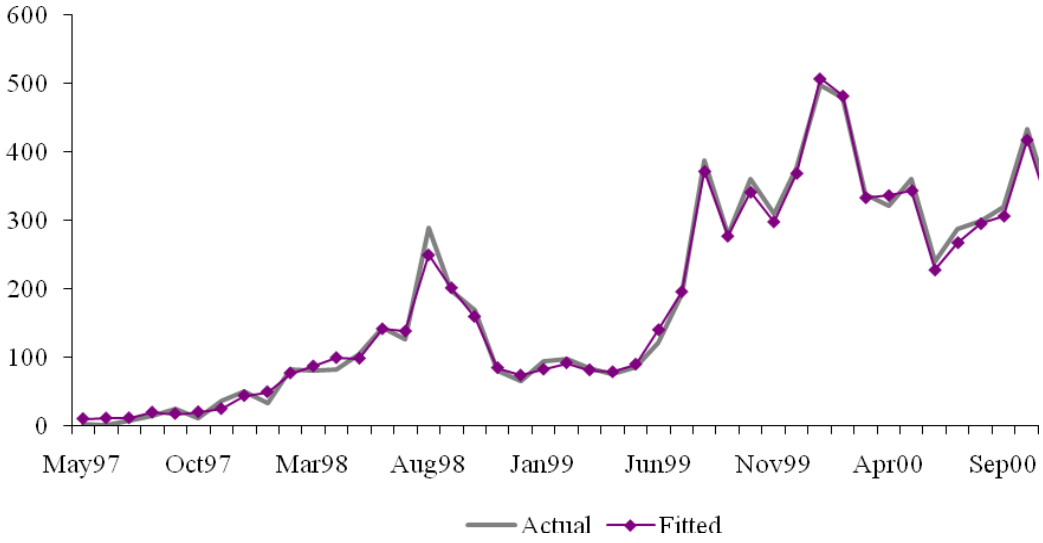
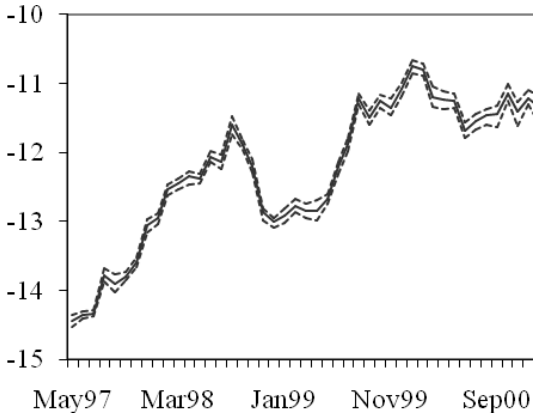
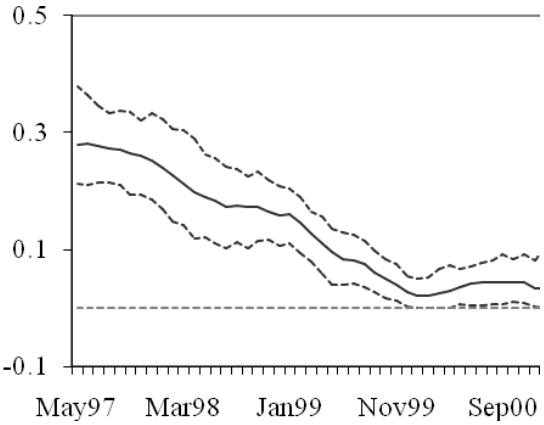


Figure 2.3: Posterior Means and 95% Posterior Intervals for the Temporal Baseline (ζ_t) and the Time-Varying Imitation Parameters (β_t^W, β_t^G and β_t^D)

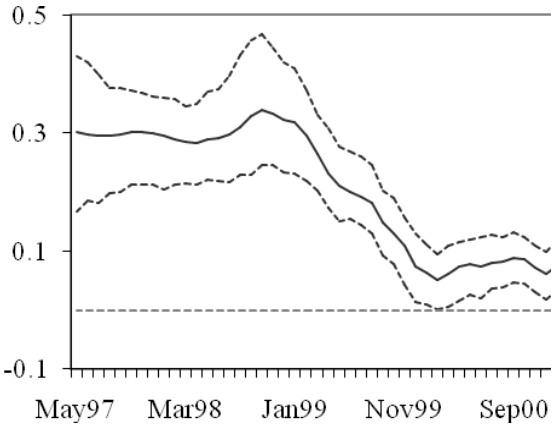
(a) Temporal baseline, ζ_t



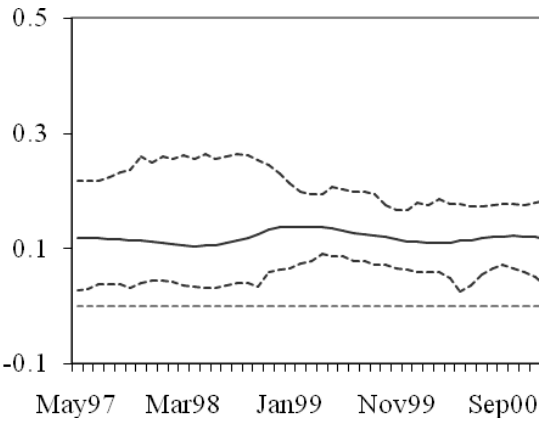
(b) Imitation Parameter for Within-Region Geographic Proximity, β_t^W



(c) Imitation Parameter for Across-Region Geographic Proximity, β_t^G



(d) Imitation Parameter for Across-Region Demographic Similarity, β_t^D



2.5.2 Parameter Estimates and Interpretation

Time-Varying Coefficients of Imitation (β_t^W, β_t^G *and* β_t^D). The posterior means and 95% posterior intervals for these parameters together with the temporal baseline ζ_t are shown in Figure 2.3. There is significant non-stationarity in the imitation parameters; β_t^W and β_t^G tend to decay over time while β_t^D stays somewhat constant. The decay in β^W and β^G is consistent with the decreasing imitation parameter estimate in the Bass model as a data window is extended (Van den Bulte and Lilien 1997; Van den Bulte and Joshi 2007). The decay in the two proximity coefficients offsets the increase in log-cumulative new buyers in the focal region (z_{it}) and contiguous regions ($\vec{z}_t$). The relative constancy of the similarity coefficient indicates that new buyers continue to emerge from disparate and physically-distant regions. One interpretation is that new-buyer acquisition through proximity "taps out" while new-buyer acquisition through similarity holds at a "steady" rate of accumulation. An Internet retailer's survival may depend on the ability to acquire similar types of customers from a wide-ranging area.

Further insights come from examining how the marginal effects of imitation vary across space and time. The marginal effect of imitation at region *i* at time *t* can be assessed by looking at the model-based expected number of new buyers E(y_{it}) compared to the expected number of new buyers (under the full model) with the imitation coefficients (β_t^W , β_t^G and β_t^D) set equal to 0. To assess the marginal effect of imitation across *space* we aggregate the 1,459 zip codes to their corresponding county, which results in 67 different counties. Figure 2.4 shows the expected number of buyers in each

county under the full model, versus the expected number when the imitation coefficients are set to 0. The gap between the two expected values indicates the marginal effect of imitation in that county. The location of each county on the *x*-axis is given by its rank in terms of number of new buyers. To avoid clutter we identify by name only the top six counties (Philadelphia is the number one county and Allegheny, which includes Pittsburgh, is the number two county). The marginal effect of imitation is not uniform, but varies significantly even among the well-performing counties. For example, while Philadelphia shows more than a 40% contribution of imitation behavior to the total number of buyers, Allegheny shows only 30%. This could be because Allegheny is more spatially-isolated from other well-performing areas, i.e., Philadelphia, Montgomery, Chester, Delaware, and Bucks, and therefore less likely to be subject to imitation effects based on proximity.

Figure 2.4: Expected Number of New Buyers in Each County

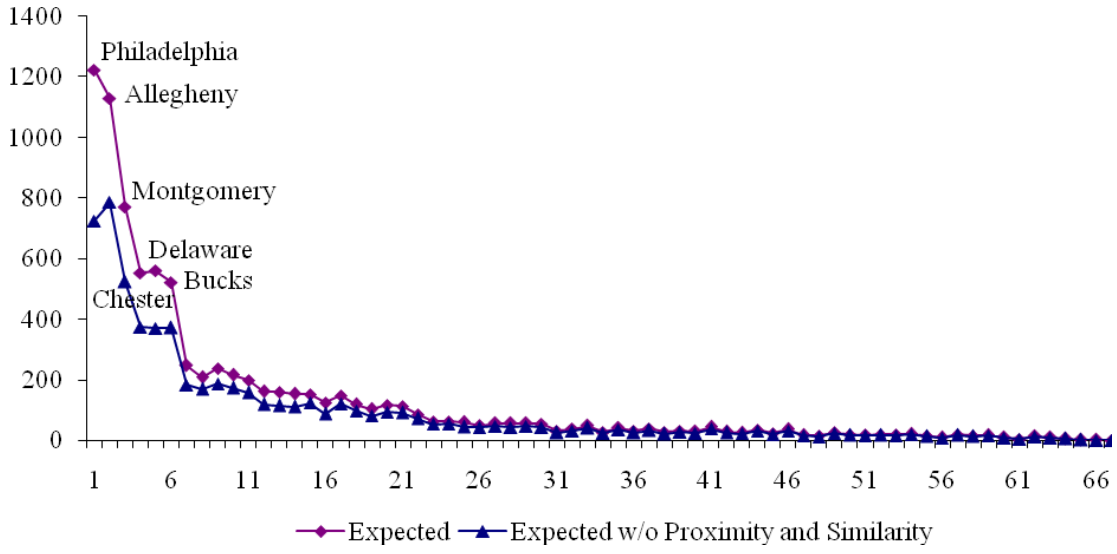
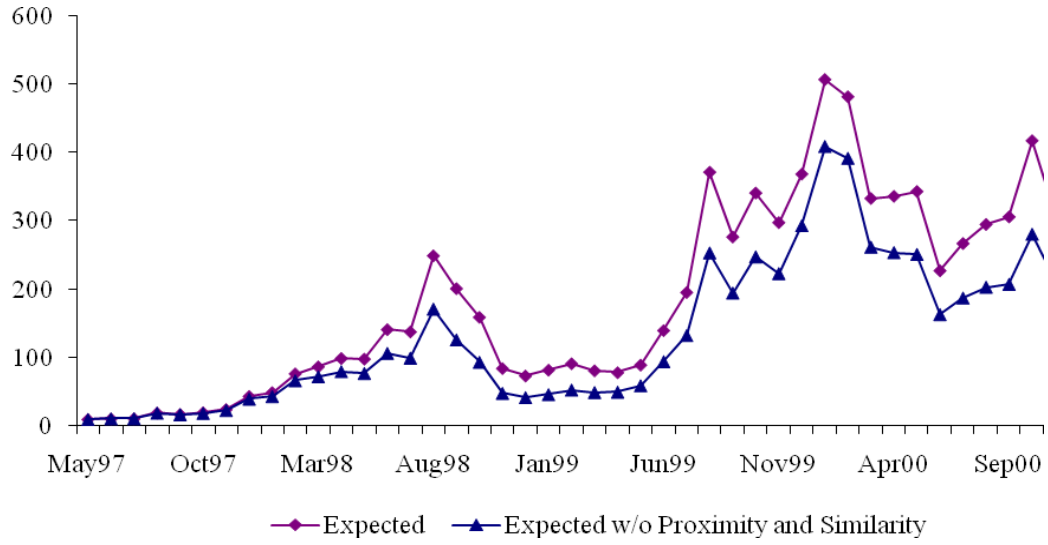


Figure 2.5 shows the marginal effects of imitation over *time*, by again comparing the expected number of buyers over time under the full model versus the expected number of buyers with imitation coefficients set to 0. The relative contribution of the imitation effects increases over time. This finding is intuitive as the larger the cumulative number of existing customers, the greater the potential for imitation of all types. This again underscores the importance of the installed base of new buyers for the ongoing acquisition of additional new buyers.

Figure 2.5: Expected Number of New Buyers in Each Month



Proximity and Similarity. Imitation effects for a focal zip code have three components: (1) the *within*-zip code effect of prior new buyers on the current period rate, (2) the *across*-zip code geographic proximity effect of prior new buyers in contiguous neighbors, and (3) the *across*-zip code demographic similarity effect. Since the first two components are based on short-range physical proximity and their relative magnitudes are relatively

stable over space and over time (the ratio of *within*- and *across*- proximity effects is about 0.5), we now combine them as one overall effect called "proximity" and compare it with the similarity effect.

Figure 2.6: Relative Magnitudes of the Proximity and Similarity Effects over Time

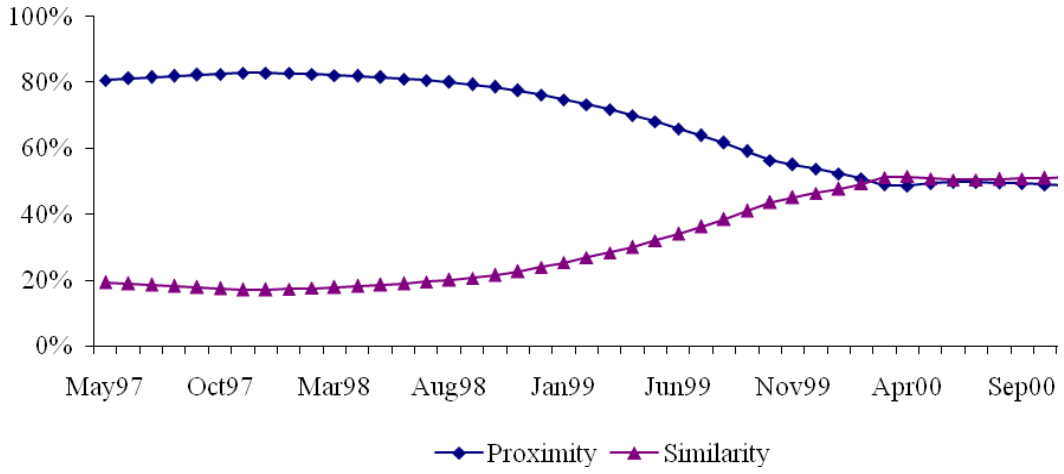


Figure 2.6 plots the relative magnitudes of the "proximity" and "similarity" effects over time. The proximity effect is relatively more important initially; however, from about thirty months out the similarity effect becomes just as important. This model-based insight complements the observed decreasing spatial concentration of new buyers implied by Figure 2.1. Initially, new buyers start to emerge in hot spot areas (such as Philadelphia and Pittsburgh) and areas that are geographically proximate areas to hot spots. Later on, new buyers increasingly emerge from new zip codes beyond the "core set" of zip codes that produce the early new buyers. The similarity effect plays a more significant role in explaining new buyers in laggard areas that are "similar" to previously successful areas. Despite the larger similarity effect in later time periods being aggregated over space, its

ultimate multiplicative effect in laggard areas does not generate as many new buyers *in total* as the proximity effect does early on in high popularity areas. The effect is nevertheless very important. This is because it helps drive orders from spatially-dispersed customers who are small in number individually, but collectively account for a significant percentage of total sales.

2.5.3 Market Seeding

Our findings suggest that the firm can influence the space-time demand trajectory through judicious market seeding (see also Godes and Mayzlin 2008). To explore this possibility we perform *hypothetical* simulations based on our model parameters and compare and contrast alternative seeding approaches. To perform this analysis, we assume that: (1) the firm knows all the imitation coefficients beforehand (perhaps from using an "analogous product" in an approach common for Bass imitation coefficients; see Lilien and Rangaswamy 2004), (2) the imitation coefficients are invariant to the firm's seeding actions, and (3) costs are equivalent across scenarios. Since validating these assumptions requires data that are beyond our sample, we must stress that the analyses presented here are purely conceptual and intended only to be treated as a springboard for future research.[11]

[11] The seeding experiment using the imitation parameter estimates is parallel to an oracle test in statistics and data mining which attempts to derive the best result given perfect knowledge of the parameters. If imitation estimates need to be predicted, the *proximity-and-similarity-based* strategy would not perform as well as it does here. Therefore, the *proximity-and-similarity-based* strategy in this paper should be interpreted as the best-case scenario.

With this caveat in mind, we explore the following "seeding" scenario. Suppose the firm considers seeding new buyers in month *t*. It then faces the decision of where these new buyers should be "planted" or allocated. Candidate zip codes are selected in accordance with the seeding policy and one buyer is added to each zip code in that month. We compare how many new buyers the alternative time *t* seeding strategies bring to Netgrocer.com from month *t* + 1 onwards. Following terminology in Libai, Muller, and Peres (2005) and in accordance with their study we compare and contrast the following four strategies (the first three draw on their work directly):

1) *Support-the-weak strategy*: The firm seeds new buyers in regions with the greatest remaining "market potential", i.e., current performance is relatively "weak" compared to what might be expected. A common heuristic is that the market potential is roughly proportional to population size so we pick candidate regions according to population yet to adopt at time *t*.

2) *Support-the-strong strategy*: The firm seeds in the historically (up to time *t*) best regions, i.e., those that have demonstrated "strong" performance to date.

3) *Uniform strategy*: The firm seeds new buyers randomly across regions regardless of market potential (based on population) or historical performance.

4) *Proximity-and-similarity-based strategy*: The firm seeds by choosing new zip codes that are the most responsive in month *t* when the combined impact of both effects is taken into account.

By December 1997, approximately eight months after the website was launched, 105 zip codes in Pennsylvania had at least one buyer. We implement our seeding experiment immediately thereafter; January 1998 is the first month available for seeding. For month *t*

we seed one new buyer into 50 regions selected by each strategy outlined above and simulate expected trajectories of incremental buyers that should result from this one-time seeding. As an illustration, the trajectory of incremental new buyers from the April 1998 seeding is shown in Figure 2.7 panel (a). In July 1998, for example, the 50 buyers seeded in April 1998 by the *support-the-weak* strategy have generated 3 new buyers.

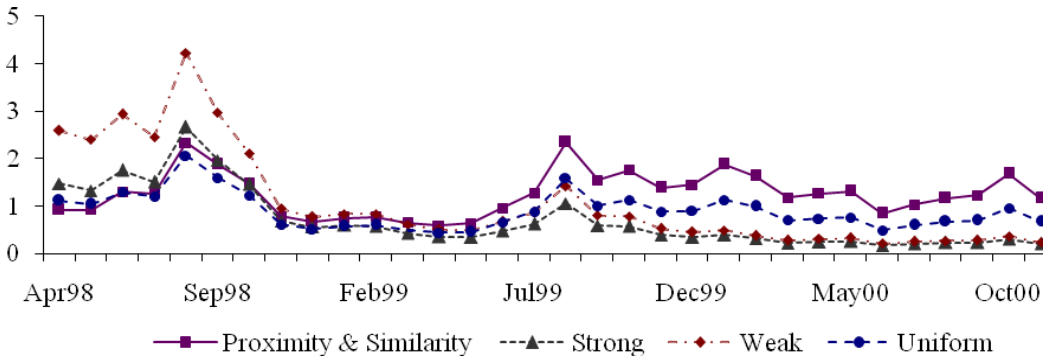
Among the strategies of Libai, Muller, and Peres (2005), the s*upport-the-weak* strategy shows the best performance early on (prior to January 1999), but later on it does not perform as well as it fails to target potential markets that are spatially-dispersed. With time the *proximity-and-similarity-based* strategy performs best as the similarity effect starts to impact new and distant areas. By adjusting the impact of proximity and similarity effects over time, the *proximity-and-similarity-based* strategy pinpoints the most promising areas for growth. This natural coordination makes this strategy consistently superior over time.

Figure 2.7 panel (b) shows the aggregate number of incremental buyers through January 2001 that result from three different one-time seeding months (January 1998, January 1999, and January 2000). "Jan 2000 Seeding", for example, shows that seeding 50 buyers in January 2000 using the *proximity-and-similarity-based* strategy yields 18 new buyers in total by January 2001. Our findings with respect to the three strategies studied by Libai, Muller, and Peres (2005) are consistent with theirs; spatially-dispersed efforts are generally superior to spatially-clustered efforts. When seeding is delayed, the *support-the-weak* strategy has less time to reap the benefit from proximity and its average performance deteriorates. The best overall outcome is induced by the *proximity-and-*

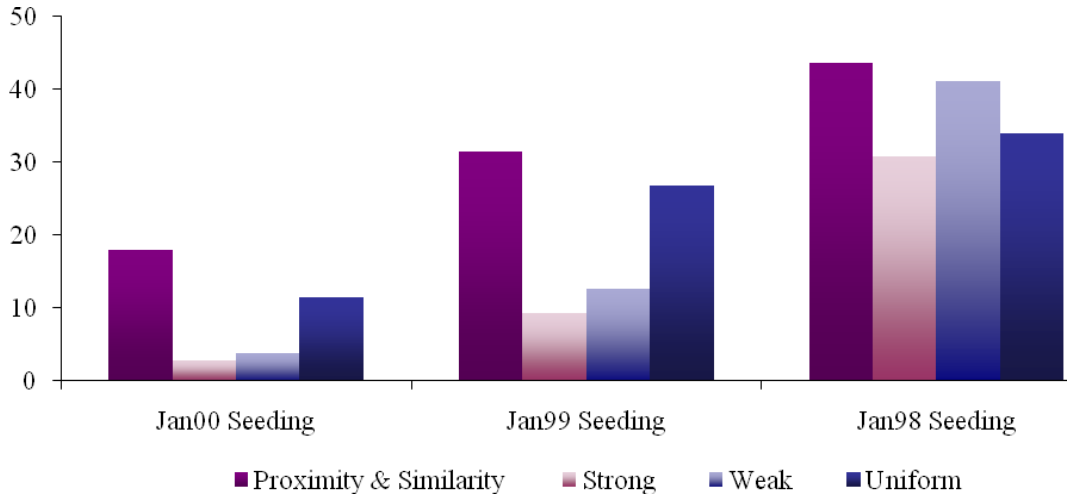
similarity-based approach and its superiority becomes more evident as the similarity effect gains momentum.

Figure 2.7: Hypothetical Seeding Experiments

(a) Temporal Trajectory of the Number of Incremental New Buyers from the One-Time Seeding in April 1998 through January 2001. (50 new buyers were seeded in April 1998)



(b) Aggregate Number of Incremental New Buyers Resulting from Three One-Time Seeding Months (in January 1998, January 1999, and January 2000) through January 2001. (50 new buyers were seeded in these seeding events)



Panels (a) and (b) of Figure 2.7 together provide insight into how to optimize seeding strategies over time. The *uniform* strategy is the best among the strategies of Libai, Muller, and Peres (2005), but seeding by *support-the-weak* very early on can outperform a *uniform* strategy continuously applied. This is because the model shows that very early on the proximity effect plays a significant role and the *support-the-weak* strategy (based on relatively under-performing areas with relatively large populations) can pick up zip codes with good potential for proximity effects. The *support-the-weak* strategy however fails to pick up spatially-dispersed markets and therefore its performance deteriorates fast with time. A switch from *support-the-weak* to *uniform* strategies might engender better performance. Unfortunately, it is "hard-to-impossible" (from a practical perspective), to predict when to switch strategies.

This implies that Internet retailers in their infancy should perhaps focus initially on populous metropolitan areas. However this strategy needs to be altered over time to incorporate the similarity effect as local concentration of demand declines. A spatially-expanded customer base is likely to be important to an Internet retailer's growth. Our *proximity-and-similarity-based* strategy is a good candidate to this end as it automatically balances the similarity effect against the proximity effect while avoiding the need to manually switch strategies. Moreover, the relative advantage of this strategy increases the later seeding is started (see Figure 2.7 (b)). Our finding highlights the insight that serving many small pools of somewhat *similar* buyers, who are spatially distant from each other, can be important to an Internet retailer as the relative contribution of these buyers to sales increases over time.

It is widely believed that a firm can offer an almost *unlimited product* assortment when the product stocking constraint is relaxed, and that small sales levels over a large number of products account for substantial aggregate sales, a phenomenon termed "The Long Tail" (Anderson 2006; Brynjolfsson, Hu, and Simester 2006). Our insight on the importance of the sales distribution over *obscure regions* (see Balasubramanian 1998) mirrors the importance of the sales distribution over *obscure products* in the Long Tail. Here the benefit comes primarily through the ability to sell in essentially *unlimited local markets*, rather than sell an *unlimited product assortment*. The Internet retailer with sufficient distribution capabilities, e.g., through use of a third party expert such as FedEx or UPS, is freed from the constraint of geography and can enjoy the benefit from serving sparse pockets of geographically-diverse demand.

2.6 Conclusion

The vastly expanded trading area of the Internet retailer is perhaps the starkest difference between it and a traditional retailer. As such, it is critical for the Internet retailer to understand how and why demand varies spatially. In this paper, we focus on the dynamic role of imitation based on geographic proximity and demographic similarity in generating new buyers over space and time. We find that in the initial phases of demand growth proximity effects are more prominent. New demand in a local area is influenced by the extent of prior demand not only in the same local area and but also in contiguous and "geographically close" regions. As time progresses the proximity effect diminishes in relative importance, but does not dissipate entirely. The similarity effect tends to increase

in relative importance over time and is particularly salient to demand generation in spatially-dispersed regions with relatively small absolute sales.

2.6.1 Limitations and Directions for Future Research.

Our study focuses on a description of the behavior of new buyers only, and does not explicitly measure the interactions among individuals. These limitations open several opportunities for future research, including the four areas described below.

- *Forecasting*: In this paper, we focus on building a descriptive model instead of a forecasting model. Moving from description to forecasting requires a different model formulation. In particular, one may want to utilize a Bayesian dynamic model (e.g., Bass et al. 2007; West and Harrison 1997) and assess its market seeding performance.
- *Incorporating social networks by demographic types*: We measure the demographic similarity by the extent of shared socio-demographic characteristics. One could allow for separate social networks by demographic types and examine which demographic network drives imitation (e.g., Conley and Topa 2002). One could also expand a model with demographically-correlated random effects in demographic space.
- *Incorporating WOM valence*: We have assumed, similar to Albuquerque, Bronnenberg, and Corbett (2007), that there is non-negative imitation, which could be driven in part by positive word-of-mouth from the earlier buyers. One interesting extension would be to allow for negative influence (e.g., Godes and Mayzlin 2004).

- *Incorporating marketing activities*: A unique aspect of our data is the absence of significant marketing efforts. We can thus assess the impact of imitation without controlling directly for potential marketing activities (e.g., advertising, promotions). If marketing activities are present, our model can be extended to control for them, perhaps using the method suggested in Bass et al. (2007). Moreover, one could build on the approach in Jank and Kannan (2005) who find significant spatial correlation in individual-level preference for PDF and print forms of books, and that this impacts price sensitivity at different geographical locations.

2.7 Appendix

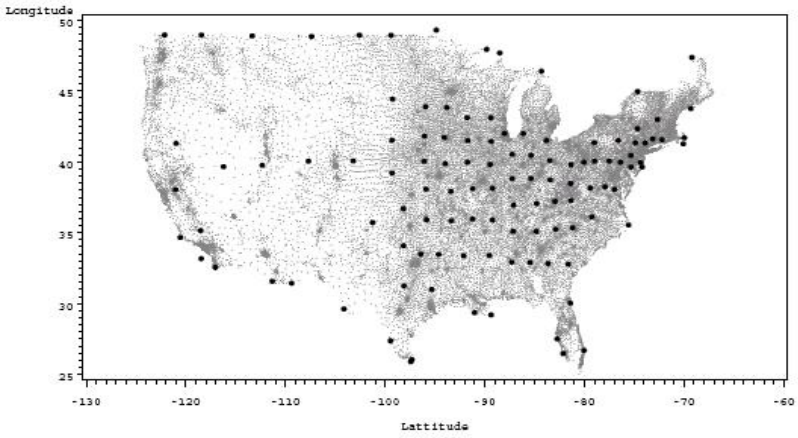
2.7.1 Low Rank Spatial Smoothing of the Broadband Access Variable

Our measure of the Internet access availability has a few known imperfections: there are some missing data for some zip codes, and in some cases services are under-reported. To improve the quality of this variable, we implement a low-rank thin plate spline smoother (Wand 2003) to correct for measurement errors. Here, we provide an outline for our implementation; readers are encouraged to see Wand (2003) for more details.

Step 1. Choose knots: We obtain "knots" based on centroids of a *k-d* tree (Molenberghs and Verbeke 2006). Starting with the entire set of zip codes, the *k-d* tree partitions the space until all partitions contain at most 300 regions. The region nearest to

the centroid of each partition is chosen as a knot which creates 117 knots in our application, as shown in the figure 2.8.

Figure 2.8: Knots Chosen for Spatial Smoothing of the Broadband Access Variable



Step 2. Bivariate radial smoothing: We then apply the low-rank thin plate spline smoothing with a radial basis function. ISPs in region *i* at time *t*, x_{it}, are specified to follow a negative binomial distribution with parameters r and κ_{it} (French, Kammann, and Wand. 2001; Molenberghs and Verbeke 2006). κ_{it} is then spatially-smoothed based on the Euclidean distances from the set of knots, $z_1, z_2, ..., z_K$, and a proper covariance function.

2.7.2 Alternative Measures for *G* and *D* Matrices

The matrix *G* can also be specified from information on the shared boundaries among zip codes. Two alternatives are: (1) the shared boundary approach, and (2) the contiguity approach. The shared boundary weighting matrix is

$$G_{ij} = \begin{cases} l_{ij} / l_i, & l_{ij} > 0 \\ 0, & otherwise \end{cases} \tag{2.11}$$

where l_{ij} is the length of zip code *i*'s boundary shared with zip code *j* and l_i is the total length of *i*'s boundary shared with all its contiguous zip codes, i.e., $l_i = \sum_j l_{ij}$. This weighting system is appropriate when two regions with a longer shared boundary might be expected to exert greater influence on each other. The shared boundary weighting matrix can be simplified to a case where two neighboring regions have equal influence on the focal region as long as they share boundaries with focal region, and this simpler form is called a contiguity weighting matrix,

$$G_{ij} = \begin{cases} 1, & l_{ij} > 0 \\ 0, & otherwise. \end{cases} \tag{2.12}$$

Two alternative measures for *D* are: (1) Inverse Exponential Mahalanobis Distance, and (2) "Affiliation" (Van Alstyne and Brynjolfsson 2005). The first measure, Inverse Exponential Mahalanobis Distance, is based on Mahalanobis distance as suggested by Van Alstyne and Brynjolfsson (2005). It measures scale-free *dissimilarity* between regions *i* and *j* and takes into account correlations in the data

$$d_{ij} = \sqrt{(v_i - v_j)' \Sigma^{-1} (v_i - v_j)}, \tag{2.13}$$

where v_i is a vector of socio-demographic characteristics of region i and Σ^{-1} is the corresponding covariance matrix. As equation (2.13) is a measure of dissimilarity, similarity can be specified as an inverse function of the exponentiated socio-demographic distance (Yang and Allenby 2003):

(2.14) $$D_{ij} = \exp(-d_{ij})$$

Affiliation is derived to be directly consistent with analytical work in Van Alstyne and Brynjolfsson (2005). Instead of using regional profile vectors directly, we define regional "vectors of types" in the following way. We compute the empirical distribution of each individual element of the fifteen elements of the profile vectors described in the paper. That is, we look across all 1,459 regions in the sample and compute the first quartile, median, and third quartile of the distribution of a particular characteristic. As a result, for each region and each characteristic, we can assign the region to one of four mutually exclusive and collectively exhaustive "types" along each element: "high" (top quartile and above) "moderate" (between median and top quartile), "low" (between bottom quartile and median), and "very low" (below bottom quartile). For example, imagine that the first quartile of the distribution of the ethnic subcategory "Black" is 10% (i.e., one quarter of the regions in the sample have a population which contains 10% or fewer Blacks). If the Black proportion of a region is 5%, then its type is defined as a region with a very low proportion of Black residents (compared to the overall population). If another region also has a small portion of Black residents, say 7%, then these two regions are assumed to be implicitly affiliated on the Black dimension. The extent of affiliation comparing two regions is as

(2.15) $$D_{ij} = \sum_{k} I(e_{ik}, e_{jk})$$

where e_{ik} is an element of the vector of socio-demographic types of region i, and $I(\cdot)$ is an indicator function which takes one if two elements are equal, and zero otherwise.

2.7.3 Justification of Poisson Distribution in Equation (2.3)

We denote the number of individiuals in zip code i as N_i, and the number of buyers as Y_i. Let y_{ij} ($j = 1, 2, \ldots N_i$) be an indicator variable which takes value 1 if the j-th individual in zip code i adopts, and 0 otherwise. In other words, we have $Y_i = \sum_{j=1}^{N_i} y_{ij}$. The Poisson distribution has been shown to be an adequate limiting distribution under the following three assumptions:

(i) The adoption probabilities are equal across individuals,

(ii) the adoption probabilities are low, and

(iii) adoption behaviors across individuals, *during the same time period*, are independent.

As discussed in Section 2.3, assumption (ii) holds in our dataset; assumption (i) and (iii), however, are fairly strong assumptions that may not hold in reality. In the following argument, adapted from Knorr-Held and Besag (1998) and Ross (1996), we show that under a reasonable relaxation of assumptions (i) and (iii), the Poisson distribution is still a valid approximation.

Heterogeneous adoption probabilities (*relaxing assumption* (*i*)). We begin by assuming that the *j*-th individual in the *i*-th zip code adopts with probability θ_{ij}, and θ_{ij} is beta-distributed across the different individuals, i.e., $\theta_{ij} \sim \text{Beta}(a_i, b_i)$.

We can now derive the marginal distribution of y_{ij} as follows. Since y_{ij} can take only the value 0 or 1, we consider the marginal probability that y_{ij} takes value 1:

$$P(y_{ij}=1) = \int_0^1 P(y_{ij}=1 \mid \theta_{ij}) \text{Beta}(\theta_{ij} \mid a_i, b_i) d\theta_{ij} \tag{2.16}$$

$$= \int_0^1 \theta_{ij} \text{Beta}(\theta_{ij} \mid a_i, b_i) d\theta_{ij} = \frac{a_i}{a_i + b_i}$$

By writing $p_i = \frac{a_i}{a_i + b_i}$, we can write the marginal distribution of y_{ij} as:

$$y_{ij} \sim \text{Bernoulli}(p_i) \tag{2.17}$$

Thus, the marginal distribution of Y_i is $\text{Binomial}(N_i, p_i)$. When p_i is small, we can use the classical Poisson approximation (Ross 1996) to obtain:

$$Y_i \sim \text{Poisson}(N_i p_i) = \text{Poisson}(\lambda_i) \text{ where } \lambda_i = N_i p_i. \tag{2.18}$$

Weakly-correlated adoption (*relaxing assumption* (*iii*)). To relax the assumption that *same-period* adoptions across individuals are independent, we need to consider the possibility of positive correlations across individual adoption behaviors. We assume that imitation behavior takes time to develop, and hence same-period imitation is weak; thus, individuals' adoption behavior during the same period is assumed to be at most weakly correlated. (Again, this is reasonable given the sparseness of our data; see Figure 2.2.).

Researchers have developed error bounds for the Poisson approximation when correlations across individuals are present. The bound, derived using the Stein-Chen method (Ross 1996), is as follows:

$$\left| P\{Y \in A\} - \sum_{i \in A} e^{-\lambda} \lambda^i / i! \right| \leq \min(1, 1/\lambda) \sum_{i=1}^{n} \lambda_i E[|Y - V_i|] \tag{2.19}$$

where V_i is such that $P\{V_i = k\} = P\left\{ \sum_{j \neq i} X_j = k \mid X_i = 1 \right\} \ \forall k$.

For a detailed derivation of (2.19), we encourage readers to refer to Ross (1996) or Barbour, Holst, and Janson (1992). Here, we note that the errors of the Poisson approximation are proportional to the quantity $E[|Y - V_i|]$, which is assumed to be arbitrarily small given our assumption that imitation takes time (thus same-period imitation is limited). Empirically, the validity of the Poisson approximation is also supported by the empirical evidence presented in Figure 2.2; the predicted distribution of adopters under our model closely resembles that of the actual empirical distribution.

2.7.4 Embedding a Polynomial Smoother within a Bayesian Model

In this appendix, we discuss how we embed a polynomial smoother within our Bayesian model by exploiting the parallel between the Gaussian random walk prior specification and polynomial smoothing. We begin by providing a brief introduction of smoothing techniques commonly used in Frequentist nonparametric statistics, and then explain how we embed such techniques into our model using Gaussian random walk priors.

Smoothing techniques. In non-parametric statistics, a smoother is often used to produce a smooth curve of *y* against *x*, given a scatterplot of (*x*,*y*) values (Simonoff 1986). The underlying model is of the form $y_i = f(x_i) + \varepsilon_i$, and interest is typically centered on estimating the function *f*(.). Since *y* is measured with error ε_i, smoothing helps the estimation of $f(x_i)$ by considering not only the observations at x_i, but also observations that have *x* values "close" to x_i. When estimating $f(x_i)$, these "neighboring" observations are down-weighted by their distance from x_i. For instance, a kernel smoother is of the form (Hastie, Tibshirani, and Friedman 2001):

$$\hat{y}_i = \sum_{j=1}^{n} y_j \, K\left(\frac{x_i - x_j}{b}\right) \Bigg/ \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b}\right) = \sum_{j=1}^{n} w(x_{i,} x_j) y_j \tag{2.20}$$

where $w(x_i, x_j)$ denotes the "weight" of the *j*-th observation on the estimation of y_i, which is governed by the distance of x_j to x_i. Different types of smoothers are defined based on the functional form of $w(x_i, x_j)$. In particular, for a *polynomial* smoother, the weights are defined to be proportional to $\rho^{|x_i - x_j|}$, $\rho < 1$.

Gaussian random walk prior. A Gaussian random walk prior, as defined in equation (2.7)-(2.10), allows us to embed a polynomial smoother within our Bayesian model. In the discussion below, we explore the parallel between a random walk prior and polynomial smoothing using a simplified set-up as follows ($t = 1,2,\ldots,T$):

$$y_t \mid \theta_t \sim N(\theta_t, 1) \tag{2.21}$$

$$\theta_t \mid \theta_{t-1} \sim N(\theta_{t-1}, \gamma^2) \tag{2.22}$$

$$\pi(\theta_0) \sim N(0, \sigma^2) \tag{2.23}$$

Equation (2.21) states that y_t is θ_t observed with error, in the same way that the adoption number y_{it} is a noisy observation based on the time-varying coefficients (and other controls) in equation (2.3). Equation (2.22) is the Gaussian random walk prior similar to that in equation (2.7)-(2.10). Equation (2.23) is a conjugate prior for the first period parameter; the variance term σ^2 can be set to a large number (e.g., 100^2) to obtain a diffuse prior.

We now explore the properties of the random walk prior in the simplified setting in equations (2.21)-(2.23). First, we show that the posterior mean estimate of θ_t (and hence $\hat{y}_t$) is a linear function of $\vec{y}$. We then proceed to show more concretely, using a numerical example, that the properties of these estimators mirror that of a polynomial smoother.

Since the conditional distribution for each $\theta_t \mid \theta_{t-1}$ is normal, it follows that their joint prior distribution is also normal (Ravishanker and Dey 2002). Thus, we can write

(2.24) $$\pi(\vec{\theta}) = MVN(\vec{\mu}_0, \Lambda_0)$$

where (after algebraic manipulations)

(2.25) $$\Lambda_0(i, j) = \sigma^2 + \gamma^2[\min(i, j) + 1] .$$

Clearly, given equation (2.23) and the structure of equation (2.22), the marginal expectation of $\vec{\theta}$ is a zero vector. Thus,

(2.26) $$\pi(\vec{\theta}) = MVN(0, \Lambda_0)$$

Equation (2.21) implies that

(2.27) $$\vec{y} \mid \vec{\theta} \sim N(\vec{\theta}, I)$$

From equation (2.26) and equation (2.27), we obtain (after some algebraic simplifications),

(2.28) $$E(\vec{\theta} \mid \vec{y}) = \left(\Lambda_0^{-1} + I\right)^{-1} \vec{y} = W\vec{y}$$

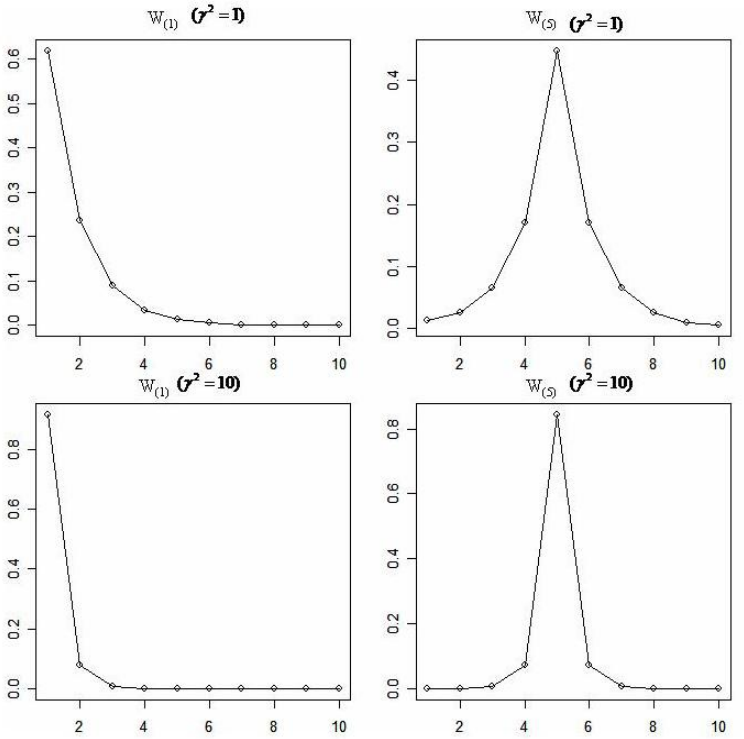
which is linear in $\vec{y}$, as desired.

To further explore the properties of the estimator in equation (2.28), we conduct a numerical experiment. In the numerical results below, we set $T = 10, \sigma = 100,$ and explore two values for γ^2, $\gamma^2 = 1$ (more smoothing) and $\gamma^2 = 10$ (less smoothing). Note that in the actual implementation, the random walk variances (i.e., $\sigma_\zeta^2, \sigma_W^2, \sigma_G^2, \sigma_D^2$) are all sampled along with other parameters, and hence the degree of smoothing is also governed by the data.

Figure 2.9 plots the 1st (to estimate $\hat{\theta}_1$) and 5th row (to estimate $\hat{\theta}_5$) of *W*, for both values of γ^2. The figure shows that the estimator induced by a random walk prior in equation (2.28) mirrors that of polynomial smoothing; for example, when estimating $\hat{\theta}_1$, a (polynomially) decreasing function is applied to the y_t's based on the distance between *t* and 1 (see the upper left panel of the figure). The same holds for the estimation of $\hat{\theta}_5$ (see the upper right panel). Further, by comparing the lower panels with the upper panels, we see that the amount of smoothing is controlled by the value of γ^2; a higher γ^2 leads to less smoothing (more weight on the observation at *t*).

Figure 2.9: Numerical Example of the Estimator in Equation (2.28)

Upper panel: $\gamma^2 = 1$; lower panel: $\gamma^2 = 10$



The Gaussian random walk prior offers several statistical advantages. First, it allows for smooth variation in the behavior of coefficients over time without a need to pre-define a parametric form (thus, data rather than model assumptions drive inferences about the temporal evolution of coefficients). Second, it links coefficients at different time points together, allowing our estimation procedure to "borrow strength" across all data observations and thereby yield more accurate estimates. Third, it is a special case of a Bayesian dynamic linear model (West and Harrison 1997), and behaves as a conjugate prior when we draw coefficients. Consequently, posterior samples are drawn very efficiently using the Gibbs sampler.

2.7.5 MCMC Procedure

As we discussed in the *Prior Specification and Smoothing* section, proper conjugate priors are specified for each of the parameters. That is, $\zeta_1, \beta_1^W, \beta_1^G, \beta_1^D \sim N(0, \sigma_1^2)$, $\vec{\tau} \sim N(0, \sigma_\tau^2 I)$, and the variance parameters are given $Inv-\chi^2(\kappa_0, \xi_0^2)$ conjugate priors. With these conjugate priors, the full conditional distribution for all parameters except λ_{it} are of standard forms. The Gibbs sampler (Casella and George 1992) is used to sample from them. Samples of λ_{it} are generated from a random walk Metropolis-Hastings algorithm (Hastings 1970).

In the discussion below, we outline the full conditional distributions for the other parameters. The following three expressions are used in this process.

(i) First, we consider a vector *y* of *n* i.i.d. observations from $N(\mu, \sigma^2)$ with known variance. Given the conjugate prior on μ ($\mu \sim N(\mu_0, \sigma_0^2)$), the conditional posterior distribution is:

(2.29) $$\mu \mid y \sim N\left(\frac{\left(\sigma_0^2\right)^{-1} \mu_0 + \left(\sigma^2 / n\right)^{-1} \bar{y}}{\left(\sigma_0^2\right)^{-1} + \left(\sigma^2 / n\right)^{-1}}, \frac{1}{\left(\sigma_0^2\right)^{-1} + \left(\sigma^2 / n\right)^{-1}} \right)$$

(ii) Second, we consider a linear model of $y_i \mid \vec{\beta}, \sigma^2, \vec{x} \sim N(\vec{x}'\vec{\beta}, \sigma^2)$ with known variance. Given the conjugate prior on $\vec{\beta}$ ($\vec{\beta} \sim N(\beta_0, \sigma_\beta^2 I)$), the conditional posterior distribution is:

$$\vec{\beta} \mid y \sim N\left(\left((\sigma^2)^{-1}X'X + (\sigma_\rho^2)^{-1}I\right)^{-1}\left((\sigma^2)^{-1}X'y + (\sigma_\rho^2)^{-1}\beta_0\right),\right.$$
$$\left.\left((\sigma^2)^{-1}X'X + (\sigma_\rho^2)^{-1}I\right)^{-1}\right) \tag{2.30}$$

(iii) Next, we consider a vector *y* of *n* i.i.d. observations from $N(\mu,\sigma^2)$ with known mean. Given a conjugate prior on σ^2 ($\sigma^2 \sim Inv-\chi^2(v_0, s_0^2)$), the conditional posterior distribution is:

$$\sigma^2 \mid y \sim Inv-\chi^2\left(v_0 + n, \frac{v_0 s_0^2 + n s^2}{v_0 + n}\right) \text{ where } s^2 = \frac{\sum (y_i - \mu_i)^2}{n} \tag{2.31}$$

We now outline how we sample each individual model parameter.

Regional random effects, γ_i.

Let $\phi_{it} = \log(\lambda_{it}) - \left(\log(n_{it}) + \zeta_t + \vec{x}_i'\vec{\tau} + \beta_t^W z_{it} + \beta_t^G G_{(i)}\vec{z}_t + \beta_t^D D_{(i)}\vec{z}_t\right)$. Then,

$\phi_{it} \sim N(\gamma_i, \sigma_\varepsilon^2)$. With the prior $\gamma_i \sim N(0, \sigma_\gamma^2)$, we apply equation (2.29).

Parameters of control variables, $\vec{\tau}$.

Let $\phi_{it} = \log(\lambda_{it}) - \left(\log(n_{it}) + \zeta_t + \gamma_i + \beta_t^W z_{it} + \beta_t^G G_{(i)}\vec{z}_t + \beta_t^D D_{(i)}\vec{z}_t\right)$. Then,

$\phi_{it} \sim N(\vec{x}_i'\vec{\tau}, \sigma_\varepsilon^2)$. With the prior $\vec{\tau} \sim N(0, \sigma_\tau^2 I)$, we apply equation (2.30).

Time-varying coefficients, $\zeta_t, \beta_t^W, \beta_t^G, \beta_t^D$. Let $\phi_{it} = \log(\lambda_{it}) - \left(\log(n_{it}) + \gamma_i + \vec{x}_i'\vec{\tau}\right)$.

Then, priors for $\vec{\zeta}$ are given below depending on time periods; those of $\beta_t^W, \beta_t^G, \beta_t^D$ are of the same form.

For *t* = 1, $\zeta_t \sim N\left(\frac{\left(\sigma_\zeta^2\right)^{-1}\zeta_2}{\left(\sigma_1^2\right)^{-1} + \left(\sigma_\zeta^2\right)^{-1}}, \frac{1}{\left(\sigma_1^2\right)^{-1} + \left(\sigma_\zeta^2\right)^{-1}}\right)$

For $1 < t < T$, $\zeta_t \sim N\left(\frac{\zeta_{t-1}+\zeta_{t+1}}{2}, \frac{\sigma_\zeta^2}{2}\right)$

For $t = T$, $\zeta_t \sim N\left(\frac{\zeta_{t-1}}{2}, \sigma_\zeta^2\right)$

With these priors, we apply equation (2.30) to obtain posterior distributions.

Variance parameters (e.g., σ_ε^2 *).* Let

$\phi_{it} = \log(\lambda_{it}) - \left(\log(n_{it}) + \gamma_i + \zeta_t + \vec{x}_i'\vec{\tau} + \beta_t^W z_{it} + \beta_t^G G_{(i)}\vec{z}_t + \beta_t^D D_{(i)}\vec{z}_t\right)$. Then,

$\phi_{it} \sim N(0, \sigma_\varepsilon^2)$. With the conjugate prior, $\sigma_\varepsilon^2 \sim Inv-\chi^2(\kappa_0, \xi_0^2)$, we can apply equation (2.31) to obtain posterior samples.

2.7.6 Effects of Control Variables ($\vec{\tau}$)

The posterior means of the coefficients of the control variables are shown in Table 2.3. Here we simply note a few interesting observations that may warrant future studies. In general, Netgrocer.com has a higher rate of new buyers in zip codes that have higher population growth, more urban housing units, and higher levels of educational attainment. While new buyers are gained more rapidly in urban areas (e.g., Philadelphia and Pittsburgh) this is driven mostly by the larger population size, and *not* a higher underlying adoption rate. Since the overall number of new buyers is relatively low, and there is a large disparity between population sizes in the highly urban areas and more rural ones, it turns out that the adoption rate, i.e., the number of buyers relative to the population, is negatively correlated with population density. The adoption rate is

negatively correlated with the density of general stores (e.g., Wal-Mart) and the presence of warehouse clubs, the two offline formats that compete directly with Netgrocer.com. It is positively related to the density of supermarkets, a complementary format (Netgrocer.com does not sell perishable products).

Table 2.3: Posterior Means of Control Variables and Variances

Variable	Description	Posterior Mean	Posterior Std Dev
Local Environment			
Population Density	Population density	-0.102	0.013
Population Growth	Annual population growth rate from 2000 to 2004	0.037	0.011
Home Value	% of homes valued at $250,000 or more	0.033	0.017
Urban Housing	% of houses with 50 units or more	0.138	0.010
Land Area	Area in square miles	-0.034	0.011
Household Characteristics			
Asian	% of Asians	-0.019	0.013
Black	% of Blacks	-0.155	0.016
White	% of Whites	-0.050	0.006
College	% with bachelors and/or graduate degree	0.360	0.019
Elderly	% aged 65 and above	-0.081	0.010
Wealthy	% of households earning $75,000+	-0.125	0.028
Access to Retail Services			
Density General	Density of general stores within the second order neighboring zip codes	-0.158	0.056
Density Supermarket	Density of supermarkets within the second order neighboring zip codes	0.256	0.056
Presence Warehouse	Presence of warehouse clubs within the second order neighboring zip codes	-0.042	0.016
Access to the Internet			
Broadband Access	Number of high-speed Internet service providers	0.026	0.004
Variances			
$\sigma_{\varepsilon}^2 \times 10$	Variance of errors	2.065	0.108
$\sigma_{\gamma}^2 \times 10$	Variance of regional random effects	1.159	0.070
$\sigma_{\zeta}^2 \times 10^2$	Variance of baseline adoption	9.361	1.928
$\sigma_{W}^2 \times 10^4$	Variance of within-region proximity effects	4.106	1.055
$\sigma_{G}^2 \times 10^4$	Variance of across-region proximity effects	7.069	2.864
$\sigma_{D}^2 \times 10^4$	Variance of across-region similarity effects	3.784	1.049

Notes: All the variables concerning the local environment, household characteristics, and access to retail services are cross-sectional and standardized, while the broadband access variable is time-varying and un-standardized.

Chapter 3

Social Influence from Existing to New Customers

3.1 Introduction

A traditional retailer serves a fixed trading area and is directly observable to customers in the local neighborhood. Conversely, an Internet retailer can tap into a potentially "limitless" pool of new customers however the retailer must first be "discovered" by these customers. Two main discovery processes are online keyword search (hereafter, "search") and offline word-of-mouth (hereafter, "WOM"). Prior studies suggest that the customers acquired in different ways might have not only different values to a firm (Lewis 2006; Villanueva, Yoo, and Hanssens 2008), but also different propensities to influence potential customers (Brown and Reingen 1987; Herr, Kardes, and Kim 1991; Richins and Bloch 1986). To the extent that two acquisition processes bring different qualities of customers to a firm, the *mix* of two types of customers in the installed base has a potentially critical influence on future new customer acquisitions. Consequently, it is vital that an Internet retailer understands how both types of existing customers exert possibly differential social influence over new customers.

Social influence arises from direct interactions or mere observations among "adjacent" or proximate individuals. When dealing with location-specific information (e.g., the number of new customers in a zip code or choices of customers living in a certain location), an *aggregate* measure of *prior* behavior of close neighbors is commonly used to proxy for social interactions (Bell and Song 2007; Manchanda, Xie, and Youn 2008; Yang and Allenby 2003). That is, prior research does not consider the acquisition status of the existing customers despite the potentially different propensities to influence new customers. I construct two separate measures to contrast the effect of social influence emanating from the existing search and WOM customers on new customers.

As retailers operate over time, prior research examines temporal parameter variation to uncover dynamic effects of a variable of interest on an outcome variable (Choi, Hui, and Bell 2009; Naik, Mantrala, and Sawyer 1998; Stremersch and Lemmens 2009). Unlike traditional retailers operating in fixed trading areas, Internet retailers acquire customers virtually from everywhere. Since local conditions dictate customer benefits from shopping at Internet retailers (Anderson et al. 2009; Forman, Ghose, and Goldfarb 2009), social influence might vary substantially not only over time but also by location. In this research, I allow the temporal parameter paths to vary over space. Thus, I expand on prior studies as follows. First, I account for (potentially) differential social influence from the two customer types in the installed base, and second, I model spatially-varying temporal parameter paths.

My goal is twofold. First, to show how the mix of customer types in the installed base drives the space-time trajectory of new customers. Second, to show what an Internet retailer can therefore do to expedite demand growth. I specify a functional mixed effects

model (e.g., Guo 2002; 2003) for the spatio-temporal demand evolution at Childcorp.com.[12] I focus on the social influence from the installed customers acquired by search (hereafter, "the installed search customers") and the installed customers acquired by WOM (hereafter, "the installed WOM customers"), and obtain temporal social influence parameter paths by county. To produce efficient parameter estimates, I model both the fixed (i.e., average) temporal path and county-specific random deviation paths from the fixed path as cubic smoothing splines in the same functional space. Estimation takes advantage of the state space representation of a smoothing spline (Wecker and Ansley 1983) and univariate Kalman filtering and smoothing (Guo 2002; Koopman and Durbin 2000) to alleviate the heavy computational demand. To the best of my knowledge, this implementation of the functional mixed effects model has yet to appear in marketing, and is a key methodological contribution of this paper.

I provide new insights into how demand evolves for an Internet retailer. First, customers acquired by WOM are on average of "more valuable" than those acquired by search. The fixed parameter paths for social influence from the installed WOM customers are generally larger than those for the installed search customers. Second, there is substantial variation in the temporal parameter paths over counties, and the spatial variation in social influence parameter paths is greater for the installed WOM customers than for the installed search customers. These findings suggest that locations still matter for new customer evolution as well as overall sales (see Forman, Ghose, and Goldfarb 2009) and that social influence from the installed WOM customers is especially sensitive to local markets. Lastly, not every market in my data favors WOM acquisitions despite

[12] For reasons of confidentiality I refer to this leading Internet retailer by the *nom de plume*, "Childcorp.com".

overall superiority of this mode. In summary, it is important that an Internet firm understands how to employ "locally-effective acquisition" to acquire the "right" customers.

The rest of the paper is organized as follows. Section 3.2 reviews relevant literature and Section 3.3 provides an overview of the data and describes key variables to proxy for the social influence from the installed customers. Section 3.4 specifies the spatio-temporal functional mixed effects model of online demand. Empirical findings and managerial implications for Internet retailers are given in Section 3.5. Section 3.6 concludes the paper.

3.2 Background and Literature Review

Internet retailers attract customers from two predominant sources—search and WOM—and the objective of this research is to evaluate how these two types in the installed base affect the acquisition of new customers, and the effects differ depending on location and time. Hence, I first review studies on customer acquisition and social influence, and then discuss modeling approaches for dynamic processes.

3.2.1 Customer Acquisition and Social Influence

Prior studies find that the mode by which a customer is acquired affects the customer's future behavior, and ultimately, their value to the firm. Acquisition mode could simply

reflect selection processes (e.g., certain customers find firms in different ways) or it could have an endogenous effect on future behavior (e.g., a customer acquired through WOM might be more satisfied with the firm and more inclined to influence potential customers). Lewis (2006), for example, finds that promotionally-acquired customers have lower repurchase rates and smaller lifetime values. Villanueva, Yoo, and Hanssens (2008) find that customers acquired via WOM add nearly twice as much value to an Internet firm as customers acquired via marketing efforts.

A large body of research documents that the installed customer base interacts with potential customers. Engel, Blackwell, and Kegerreis (1969), for example, find that sixty percent of the respondents named referrals from prior customers as the most influential source regarding the adoption of an automotive diagnostic center. That is, the installed customer base plays an important role in acquiring future new customers by serving as a "sales person" for the firm. These social interactions arise from direct interactions or mere observations among near neighbors, and thus, the observed *prior behavior* of close neighbors is used to construct measures that in turn influence the probability of later action by other proximate individuals. Bell and Song (2007) find that spatially-proximate prior customers influence Internet retailer trials by potential customers. In studying a new drug adoption Manchanda, Xie, and Youn (2008) find that the social contagion is greater for geographically close physicians. "Closeness" among neighbors can also be defined in ways other than physical proximity and recent studies introduce additional variables to measure different types of "closeness" in social networks. In studying individual automobile choice Yang and Allenby (2003) incorporate "demographic closeness" in their model of household behavior. Albuquerque, Bronnenberg and Corbett (2007) find

that ISO9000 diffusion across countries is driven by proximity and trade-based similarity, whereas ISO4000 diffusion is driven by proximity and cultural similarity.

While all these studies combine the measure to aggregate prior customers with one or more social network specifications they do not decompose the types of individuals (or firms) in the installed base into distinct subgroups. The *mix* of different types of customers in the installed base matters as different acquisition processes bring forth different "qualities" of customers (e.g., Lewis 2006; Villanueva, Yoo, and Hanssens 2008).[13] Thus, I construct two separate measures to contrast social influence from the installed WOM customers and social influence from the installed search customers in driving future new sales.

In general, there are several reasons that customers acquired by WOM are more likely to exert social influence than those acquired by search. First, customers acquired by WOM are more likely to be socially active and create subsequent referrals (Brown and Reingen 1987). Second, positive impressions formed during the initial WOM interactions (when *they* became customers) are likely to be reinforced through experience (Herr, Kardes, and Kim 1991). Prior impressions are persistent and resistant to change because impression-consistent information supports confidence, impression-inconsistent information is discounted, and ambiguous information is interpreted as consistent with the initial impression (Hoch and Deighton 1989; Lord, Ross, and Lepper 1979). Third, customers engaging in WOM are more likely to have enduring involvement with the firm

[13] Of course the types of *social relationship* play a role in social interactions and social influence. Brown and Reingen (1987) find that at the macro level, weak ties serve an important bridging function, allowing information to travel from one distinct subgroup of referral actors to another subgroup in the broader social system. At the micro level, strong and homophilous ties were more likely to be activated for the flow of referral information. Strong ties were also perceived as more influential than weak ties, and they were more likely to be utilized as sources of information for related goods.

while customers who search online are more likely to have been motivated situational interest. Customers with enduring involvement continue to be engaged in subsequent referrals (Richins and Bloch 1986).

If one studied firms operating from fixed locations, the effect of social influence on new customer acquisitions would be relatively stable over time within fixed trading areas and it would be relatively straightforward to measure. My focus, however, is on an Internet retailer who can acquire customers from multiple (and very different) locations, over a relatively long period of time. Thus, the conjectured superiority of customers acquired by WOM over those acquired by search in exerting social influence on potential customers might vary substantially by space and over time. That is, depending on local environment and time point, one type of already acquired customer (in the installed base) could be more socially active than the other type in influencing potential customers and driving new sales.

3.2.2 Modeling Dynamic Processes

Marketers have devoted considerable attention to substantive and methodological aspects of dynamic behavior in markets. Causal factors include changes in firms' marketing strategies, competitive market environments, consumers' tastes, social interactions, and so on. Temporal dynamics are specified in various ways (see Leeflang et al. 2009 for a comprehensive review). State space models (Kalman filters and DLMs) are well suited for this purpose as they have an observation equation for short-term effects and a time-varying state equation for long-term effects. Naik, Mantrala, and Sawyer (1998), for

example, use the Kalman filter procedure to accommodate inter-temporal dependencies in awareness buildup and decay via the use of conditional densities. Putsis (1998) studies temporal parameter variation in new product diffusion models, and finds that stochastic parameter specifications produce substantially better fits and are useful in the case of weak priors on the likely pattern of variation.

VAR (vector autoregressive) models are also used to examine the dynamic effects among multiple endogenous variables through impulse response functions. Dekimpe and Hanssens (1995), for example, show that not all advertising has strong trend-setting effects on sales and Yoo and Pauwels (2008) find stronger long-term response to price increases than to price decreases. Time varying parameters can be incorporated into VAR models using a moving window approach but this requires a judicious window choice.

Recent papers impose non-parametric or semi-parametric smoothing into temporal parameter paths. Stremersch and Lemmens (2009) model time-varying coefficients using penalized splines and Choi, Hui, and Bell (2009) embed a polynomial smoother within a Bayesian model using a random walk prior. In particular, Choi, Hui, and Bell (2009) demonstrate that the temporal dynamics in social influence explains demand evolution at an Internet retailer both in clustered hot spots and in more dispersed and somewhat "obscure" locations.

An Internet retailer is not bounded by a trading area, and customers in different locations see different net benefits from using one (Anderson et al. 2009; Forman, Ghose, and Goldfarb 2009). This implies that social influence can also vary over location. Indeed, dynamics are not only temporal, but most likely spatial, at least for an Internet retailer. While there is an abundance of empirical models that incorporate temporal dynamic

effects, no study in marketing (to the best of my knowledge) has addressed spatial and temporal dynamics of parameters together. In this study I model space-time dynamics in the effect of social influence from the installed to future new customers.

3.3 Data and Measures

I first introduce the raw data and describe how they are compiled and integrated from separate sources. I then discuss how I construct variables to capture social influence from existing customers to potential new customers.

3.3.1 New Customer Data

Childcorp.com sells a large selection of name brand diapers and was rated one of the fastest growing pure-play Internet retailers in the United States in 2007.[14] Shoppers obtain free shipping on orders over $49 (about 90% of the orders are shipped free of charge) and orders are shipped via UPS from company warehouses located in both the eastern and western United States. Management provided me with quarterly zip-level sales data from the start of the first quarter in 2005 after the site opened (Q1) through the end of the first quarter in 2008 (Q13).

I focus on zip codes within Metropolitan Statistical Areas (MSAs) where at least some customers adopted Childcorp.com during the first five quarters. Performing the

[14] For reasons of confidentiality I refer to this leading Internet retailer by the *nom de plume*, "Childcorp.com".

analysis on these data ensures that I avoid zip codes that are sparse in terms of both target population and purchase history, and keeps the estimation tractable. This results in a sample of 4,532 zip codes across all the contiguous states, with thirteen quarterly observations for each zip code. Table 3.1 presents descriptive statistics. Columns (1) and (2) in Table 3.1 show that the average number of new customers per zip code increases over quarters. In addition, so does the variability in the number of new customers. (I discuss the influence measures in columns (3) to (10) in Section 3.3.3.)

Table 3.1: Summary Statistics

(a) Summary statistics for the dependent variable and variables for social influence

Data Quarters	Calendar Quarters	New Buyers (y_{ijt})		Influence from the installed base acquired by word-of-mouth				Influence from the installed base acquired by search			
				Within-Zip ($x_{ijt,1}^{WOM}$)		Across-Zip ($G_{(ij)}\bar{x}_{t,2}^{WOM}$)		Within-Zip ($x_{ijt,3}^{SEARCH}$)		Across-Zip ($G_{(ij)}\bar{x}_{t,4}^{SEARCH}$)	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Q1	Q1, 2005	.237	.938	.000	.000	.000	.000	.000	.000	.000	.000
Q2	Q2, 2005	.735	1.427	.046	.364	.039	.179	.052	.282	.036	.126
Q3	Q3, 2005	.701	1.465	.214	.820	.150	.410	.238	.640	.162	.333
Q4	Q4, 2005	.941	2.282	.480	1.342	.326	.686	.434	.910	.290	.492
Q5	Q1, 2006	1.285	3.006	.749	2.230	.519	1.144	.774	1.366	.518	.774
Q6	Q2, 2006	1.315	3.347	1.103	3.276	.776	1.728	1.245	2.044	.829	1.194
Q7	Q3, 2006	2.488	4.327	1.533	4.538	1.098	2.430	1.619	2.544	1.091	1.523
Q8	Q4, 2006	2.256	4.776	2.207	6.040	1.594	3.293	2.121	3.085	1.449	1.916
Q9	Q1, 2007	3.127	5.982	2.856	7.712	2.068	4.185	2.591	3.680	1.788	2.313
Q10	Q2, 2007	4.668	6.182	3.790	9.935	2.751	5.439	3.128	4.293	2.169	2.721
Q11	Q3, 2007	4.228	6.235	4.832	11.992	3.508	6.588	4.197	4.856	2.884	3.103
Q12	Q4, 2007	4.820	7.275	5.982	14.161	4.353	7.910	4.774	5.335	3.300	3.483
Q13	Q1, 2008	7.946	9.168	7.088	16.209	5.169	9.205	5.338	5.826	3.699	3.877

Note: In the model, all variables are log-transformed.

(b) Summary statistics for zip-level covariates.

Variable	**Mean**	**SD**
Shipping Times		
Expected Days to Ship to Zip Code	2.144	0.958
Net Promoter Scores		
Net Promoter Scores	0.720	0.034
Percentage Promoters	0.787	0.026
Percentage Detractors	0.068	0.008
Sales Taxes		
Local Offline Conditional Sales Tax Rate (%)	5.637	3.098
Access to Offline Retailers		
Distance to the Nearest Supermarket	1.924	1.741
Distance to the Nearest Wal-Mart or Target	3.738	3.233
Distance to the Nearest Warehouse Club	6.800	7.621
Socio-Demographic Controls		
Number of Children $\leq$ 4 Years Old	1979.430	1295.840
Population Density	4.515	9.985
Population Growth Rate (2000-2004)	0.017	0.022
Percentage Population Aged 20 to 39 Years Old	0.297	0.069
Percentage Households with Working Female	0.575	0.075
Percentage of Whites	0.785	0.192
Percentage with College Education	0.620	0.149
Percentage Households Earning $75,000 and more	0.323	0.155
Percentage Homes Valued at $250,000 or more	0.242	0.259
Percentage Apartments with 50 or more units	0.060	0.095

3.3.2 Exogenous Variables

Acquisition Modes. During the registration process new customers are asked: "How did you hear about our website?" Multiple responses are prevented through the use of a drop-down list and all the answers are classified into mutually exclusive and collectively exhaustive categories. I focus on the two predominant categories, online search and offline WOM.[15] *Search* acquisition includes customers who came to Childcorp.com through keyword search from search engines or connections from sponsored price-comparison sites. *WOM* acquisition includes personal referrals from friends or acquaintances, and accidental referrals from unacquainted people in local regions. Thus, at each time period and location I can compute the number of customers acquired through search and through WOM.

Net Promoter Scores (*NPS*). Customers are asked their recommendation likelihood on a 10-point scale after they receive their first orders: customers with a rating of 9 and 10 (79% of respondents) are classified as "promoters" and customers with a rating of 6 or lower (7% of respondents) are classified as "detractors." Following Reichheld (2006), I compute zip-level Net Promoter scores by subtracting the proportion of detractors from the proportion of promoters. The zip-level variables are imperfect because of small-to-sparse responses in many zip codes. To improve the quality of the variables, I employ a low-rank thin plate spline smoother before incorporating them into the model (Choi, Hui,

[15] About 6% of responses indicate "online WOM" as the acquisition channel and the remaining answers were not amenable to any meaningful classification. The response rate is about 70%. The ordering behavior does not differ significantly between the non-respondent and respondent groups; hence, I believe the data are relatively free of non-response bias. Moreover, we have no reason to believe that individuals systematically distort their self-report.

and Bell 2009; Ruppert, Wand, and Carroll 2003; Stremersch and Lemmens 2009). Details are shown in Appendix in Section 3.7.1.

Shipping Times. Childcorp.com management provided information on expected shipping times (in days) from two warehouses (one on each coast) to each zip code, which I corroborated with data from UPS.com. Orders ship from whichever warehouse is closer to the zip code receiving the order, and shipping takes 1 to 4 days. Thus, I am able to control for new customer response to product convenience, as measured by shipping times.

Relative Online Prices—Sales Taxes. Childcorp.com charges the same price at every location however the effective "relative online price" varies by location in accordance with offline sales taxes. In locations with taxes on offline sales, Childcorp.com has an increased relative price advantage. State sales tax rates are straightforward to obtain, but in many regions taxes are assessed at the county or municipal level. Some states also have tax exemptions on baby diapers. The relevant local sales tax rates are therefore affected not only by state regulations, but also by local rules governing the sale of these products. In compiling the relevant zip level tax rates I started with publicly available information from the Department of Revenue in each state and undertook an exhaustive manual check of local tax rates.[16] Substantial variation in the zip code-level rates indicates that state-to-state variation alone is not sufficient to examine the tax effect. It underscores the importance of obtaining this disaggregate data.

[16] I made over 1,000 telephone calls to a random sample of major retailers across the United States including Wal-Mart, Walgreens, and CVS, and asked store employees to determine whether the focal products were tax exempt. I also requested that they verify their answer by physically scanning individual items.

Access to Offline Retailers. Local retail statistics in the 2007 Census of Retail Industries are obtained from ESRI (www.esri.com), using 8-digit NAICS (North American Industry Classification System) codes. While 6-digit NAICS codes are often used in research, greater accuracy is achieved with 8-digit NAICS codes. To reflect the local retail environment, I compute the distance from the zip code center to the actual location of the nearest supermarkets, discount stores (Wal-Mart and Target), and warehouse clubs as physical distance is taken as a parallel to transportation costs in spatial differentiation models (see e.g., Balasubramanian 1998; Bhatnagar and Ratchford 2004). Prices at Wal-Mart are comparable to those at Childcorp.com, so access to such stores and to warehouse clubs should make Childcorp.com less attractive to potential customers in the local area.

Socio-Demographic Controls. Local covariates which describe the socio-demographic environment are constructed from the 2000 US Census of People and Households. The zip code market size is measured by the density of population (population per square miles) and the annual growth rate in population from 2000 to 2004. The characteristics of people living there are represented by percentage variables (Bell and Song 2007; Dhar and Hoch 1997).

3.3.3 Measures to Capture Social Influence

My main interest centers on the spatio-temporal effects of social influence from search-acquired and WOM-acquired customers in the installed base. Hence, I allow both customer types in the installed base to affect the future evolution of new customers

differently. I define neighbors at the zip code level (e.g., Choi, Hui, and Bell 2009) and construct two zip-level measures of social influence for each customer type. This makes sense for several reasons. First, Childcorp.com takes customers from local offline retailers and a zip code is a relatively self-contained unit of customers and sellers (especially for a product like diapers).[17] Second, zip code sales are of particular managerial interest for potential targeting efforts. Third, individual-level neighbor covariate information is neither available nor practical to work with. Lastly, exogenous definition of zip-level "neighbors" helps avoid the well-known "reflection problem" (Manski 1993; 2000).

For each customer type in the installed base, I create two variables to reflect *within*-zip and *across*-zip social influence (see also Choi, Hui, and Bell 2009). The within-zip measure is defined as the log of the cumulative number of customers in region *p* prior to time *t*, x_{pt}. This serves as a proxy for the influence from *within*-zip proximate neighbors on new customers who may subsequently emerge in the same location. The definition of this variable and its construction and motivation for use are all straightforward; hence, no further elaboration is necessary.

The *across*-zip measure of social interaction is more complex. Following the standard approaches in the literature, I define neighborhood relationships through the use of weighting matrices (Anselin 1988; Bell and Song 2007; Yang and Allenby 2003). All pair-wise relations among *n* zip codes are summarized by a $n \times n$ weighting matrix, *G*, in which each nonnegative element, $G_{(p,q)}$, denotes the degree of geographic proximity of

[17] The most accessible local retail format for diapers is a supermarket and residential zip codes have on average four supermarkets. Also, zip codes are widely used as the unit of analysis in other studies of related phenomena, such as restaurant and bookstore variety (see Waldfogel 2007 for a review).

location *p* to location *q*.[18] The focal measure of across-region proximity is based on physical distances among zip codes. Following Choi, Hui, and Bell (2009) and Yang and Allenby (2003), geographic proximity is assumed to be an inverse function of the geographic distance (in miles)

$$G_{(p,q)} = \begin{cases} \exp(-d_{pq}), & p \neq q \\ 0 & , \; p = q \end{cases} \tag{3.1}$$

The distance-based measure helps control for the fact that different zip codes vary greatly in land area and number of contiguous neighbors. Two alternatives matrices are considered in Appendix in Section 3.7.2. The weighting matrix is symmetric by definition and row-normalized to take into account relative magnitudes of closeness among locations.

The row vector, $G_{(p)}$, represents the degree of geographic proximity of zip code *p* to the remaining zip codes. Post-multiplying this by the vector that captures the size of the installed customers in a neighborhood, $\vec{x}_t$, (i.e., the log of the cumulative number of customers in the remaining zip codes) produces a scalar value to summarize the installed customers across neighbor zip codes.

Given the nature of the data I am unable to disentangle—except in an ex post analysis of the marginal effect—whether prior customers are propagating positive or negative social influence. Since name-brand diapers and formula sold at Childcorp.com are sufficiently well known, shoppers' concern about ex ante quality is largely irrelevant.

[18] Weighting matrices to measure socio-demographic similarity can be included (see, e.g., Choi, Hui, and Bell 2009; Yang and Allenby 2003). Due to the prominent importance of the proximity effects in populous regions, in this paper, we focus on the proximity effect. Note that geographic proximity matrices are constructed using all the residential zip codes and are not subject to the issue of the "edge effect", i.e., border zip codes have fewer neighbors than the interior zip codes do and as a consequence, inferences can be biased.

Prices are also known. Hence, negative information about the experience with Childcorp.com is most likely to relate to delivery, which is handled by UPS. Therefore, positive social influence is likely; the more cumulative customers there are, the greater number of new customers that will emerge. However, I impose no restriction on the signs of the parameters of social influence, and negative parameters paths are possible.

In summary, I construct two separate variables to measure the social influence from the two customer types in the installed base. That is, I develop within-zip and across-zip measures to capture social influence from the installed search customers, and analogous measures for the installed WOM customers. Descriptive statistics are given in Table 3.1 columns (3) to (10). Both types of customers in the installed base increase over time, but the installed WOM customers grow faster than installed search customers, with more spatial variation.

3.4 Model

I first propose a spatio-temporal model of new customers and specify social influence due to *within*- and *across*-zip proximity to two customer types in the installed base. Next, I describe how the parameters for social influence are modeled to vary across time as well as space. Lastly, I outline the proposed estimation procedure.

3.4.1 A Spatio-Temporal Model of New Customers

I specify the spatio-temporal model for the emergence of new customers at the quarterly time interval. Denote y_{ijt} $(i=1,\cdots,n;\ j=1,\cdots,m_n;\ t=1,\cdots,T)$ as the number of new customers in log transformation in the j^{th} zip code in the i^{th} county at quarter t.[19] As discussed in the previous section, my interest centers on the spatio-temporal effects of social influence emanating from the two different types of customers in the installed base—those acquired through WOM and those acquired by search. Thus, I decompose y_{ijt} into the baseline effect, $BASE_{ijt}$, the social influence from the installed WOM customers, SI_{ijt}^{WOM}, the social influence from the installed search customers, SI_{ijt}^{SEARCH}, and measurement error, e_{ijt}.

$$y_{ijt} = BASE_{ijt} + SI_{ijt}^{WOM} + SI_{ijt}^{SEARCH} + e_{ijt},\quad e_{ijt} \sim N(0,\sigma_e^2) \tag{3.2}$$

where e_{ijt} is assumed to be independent and normally distributed with mean 0 and variance σ_e^2. I further model

$$BASE_{ijt} = \beta_{t,0} + \alpha_{it,0} + \vec{z}_{ijt}\vec{\delta} \tag{3.3}$$

where $\beta_{t,0}$ is the overall baseline path and $\alpha_{it,0}$ is the county-level random deviation from the overall baseline path. $\vec{z}_{ijt}$ is a vector of covariates that serve as controls for zip-level socio-demographic characteristics as shown in Table 3.1 (b), including the size of the target population who have yet to try the service at time t, and $\vec{\delta}$ is the vector of time-

[19] The model can be easily modified to accommodate the cases where observations are collected at different time points and some observations are missing. See Guo (2002) for details.

invariant parameters.

As discussed in Section 3.3.2, the social influence from the installed WOM customers consists of the within-zip influence due to the proximity to the installed WOM customers in the same market and the across-zip influence due to the proximity to the installed WOM customers in the local neighborhood. The model is given as

$$SI_{ijt}^{WOM} = \left(\beta_{t,1}^{WW} + \alpha_{it,1}^{WW}\right) x_{ijt,1}^{WOM} + \left(\beta_{t,2}^{AW} + \alpha_{it,2}^{AW}\right) G_{(ij)} \bar{x}_{t,2}^{WOM} \quad (3.4)$$

where $x_{ijt,1}^{WOM}$ is the size of the installed WOM customers in log-transformation (i.e., the log of the cumulative number of customers acquired by WOM acquisition) in the j^{th} zip code in the i^{th} county prior to time t. $\beta_{t,1}^{WW}$ and $\alpha_{it,1}^{WW}$ are the fixed (or overall) parameter path and the county-specific random deviation, respectively. Thus, $\left(\beta_{t,1}^{WW} + \alpha_{it,1}^{WW}\right) x_{ijt,1}^{WOM}$ captures the within-zip social influence by the installed WOM customers. The temporal parameter paths are obtained at the county level to accurately reflect the propensity of the installed WOM customers in driving future evolution of new customers. Recall from Section 3.3.3 that $G_{(ij)} \bar{x}_{t,2}^{WOM}$ summarizes the installed WOM customers across neighbor zip codes as a scalar value. $\beta_{t,2}^{AW}$ and $\alpha_{it,2}^{AW}$ are the fixed parameter path and the county-level random deviation, respectively, and $\left(\beta_{t,2}^{AW} + \alpha_{it,2}^{AW}\right) G_{(ij)} \bar{x}_{t,2}^{WOM}$ captures the across-zip code social influence from the installed WOM customers.

The model of the social influence from the installed search customers is also given in the same form as

$$SI_{ijt}^{SEARCH} = \left(\beta_{t,3}^{WS} + \alpha_{it,3}^{WS}\right) x_{ijt,3}^{SEARCH} + \left(\beta_{t,4}^{AS} + \alpha_{it,4}^{AS}\right) G_{(ij)} \bar{x}_{t,4}^{SEARCH} \quad (3.5)$$

where $x_{ijt,3}^{SEARCH}$ and $\bar{x}_{t,4}^{SEARCH}$ is the size of the installed search customers in a focal zip code

and in a neighborhood prior to time *t*. $\beta_{t,3}^{WS}$ and $\alpha_{it,3}^{WS}$, and $\beta_{t,4}^{AS}$ and $\alpha_{it,4}^{AW}$ are the corresponding fixed and the county-level random parameter paths for within- and across-zip social influence, respectively. Thus, $\left(\beta_{t,3}^{WS}+\alpha_{it,3}^{WS}\right)x_{ijt,3}^{SEARCH}$ and $\left(\beta_{t,4}^{AS}+\alpha_{it,4}^{AS}\right)G_{(ij)}\vec{x}_{t,4}^{SEARCH}$ represent the within-zip and across-zip social influence by the installed search customers.

3.4.2 Space-Time Varying Parameters

In order to make the average fixed parameter path and county-level random parameter paths have the same smoothness properties, $\beta_{t,k}$ and $\alpha_{it,k}$ ($k=0,\cdots,4$) are modeled in the same functional space as cubic smoothing splines (Guo 2002; 2003). Following Wahba (1978, 1983)'s Bayesian approach, they are modeled as

(3.6) $$\beta_{t,k}=B_{1k}+B_{2k}t+\lambda_{\beta k}^{-1/2}\int_0^t W_k(s)ds, \qquad k=0,\cdots,4$$

(3.7) $$\alpha_{t,k}=A_{1k}+A_{2k}t+\lambda_{\alpha k}^{-1/2}\int_0^t W_k(s)ds, \qquad k=0,\cdots,4$$

where B_{1k} and B_{2k} have diffuse priors ($[B_{1k},B_{2k}]^T\sim N(0,\tau I)$ with $\tau\to\infty$), $[A_{1k},A_{2k}]^T\sim N(0,\{\sigma_{\alpha 1k}^2,0;0,\sigma_{\alpha 2k}^2\})$, $W_k(s)$ and $W_l(s)$ are Weiner processes, and $\lambda_{\beta k}$ and $\lambda_{\alpha l}$ are smoothing parameters.[20] Each random path is modeled as realizations of Gaussian process with zero means. The smoothing parameter λ controls the trade-off between smoothness and bias: the larger λ is, the smoother the parameter path is. (Note that the absence of constraints on parameters paths allows for the possibility of negative paths.)

[20] I employ numerically diffuse priors by taking a large value for κ. For exact diffuse priors, see Ansley and Kohn (1990), Koopman (1997), and Koopman and Durbin (2002).

Conditional on the parameter estimates, the posterior estimate of $\hat{\beta}_{t,k} = E[\beta_{t,k} \mid Y]$ and $\hat{\alpha}_{it,k} = E[\alpha_{it,k} \mid Y_i]$ are obtained along with their posterior variance $V[\beta_{t,k} \mid Y]$ and $V[\alpha_{it,k} \mid Y_i]$, where Y is all the data and Y_i is the data from the i^{th} county. Since $\hat{\beta}_{t,k}$ and $\hat{\alpha}_{it,k}$ are cubic splines, the prediction curve, $\hat{y}_{ijt}$, is also a cubic smoothing spline. Moreover, conditional on the smoothing parameter, the estimate of the spline has a valid interpretation at any given t.

The differences between the fixed path, $\beta_{t,k}$, and the random paths, $\alpha_{it,k}$, are two-fold. First, $\beta_{t,k}$ is modeled as a single realization of a partially diffuse Gaussian process, while $\alpha_{it,k}$ is modeled as a random realization from the same Gaussian process with proper variances. Each random path shares the same degree of smoothness because they share the same correlation structure and smoothing parameter. Second, $\hat{\beta}_{t,k}$ is estimated by its posterior mean conditional on all the data, while $\hat{\alpha}_{it,k}$ is estimated as a posterior mean conditional on the i^{th} county profile.

3.4.3 Estimation

Estimation takes advantage of the state space representation of a smoothing spline (Wecker and Ansley 1983). The model in equation (3.2) is represented in the following state space form.

(3.8) $$\vec{y}_t = M_t \vec{\gamma}_t + e_t, \qquad \vec{e}_t \sim N(0, \sigma_e^2 I_{N\times N}), \qquad t = 1, \cdots, T$$

(3.9) $\vec{\gamma}_t = H_t \vec{\gamma}_{t-1} + \vec{\omega}_j, \quad \vec{\omega}_t \sim N(0, W_t)$

where $\vec{y}_t = \{y_{11t}, \cdots, y_{1m_1t}, \cdots, y_{n1t}, \cdots, y_{nm_nt}\}^T$ is the collection of observations at time *t* and $\vec{e}_t = \{e_{11t}, \cdots, e_{1m_1t}, \cdots, e_{n1t}, \cdots, e_{nm_nt}\}^T$ is the collection of measurement errors at time *t*. $N = \sum_{i=1}^{n} m_n$ is the total number of observations. $X_{it}^* = \{X_{i1t}^{*\,T}, \cdots, X_{im_it}^{*\,T}\}^T$ is the design matrix with X_{ijt} and the zero columns, $X_{ijt}^* = \{X_{ijt,1}, 0, \cdots, X_{ijt,p}, 0\}$, and $M_t = \left[\{X_{1t}^{*T}, \cdots, X_{nt}^{*T}\}^T, \text{diag}\{X_{1t}^*, \cdots, X_{nt}^*\}\right]$. $\gamma_t = \{\beta_t^{*T}, \alpha_t^{*T}\}^T$ is the collection of both fixed and random parameters with $\beta_t^* = \{\beta_{t,0}, \beta'_{t,0}, \cdots, \beta_{t,4}, \beta'_{t,4}\}^T$ and $\alpha_t^* = \{\alpha_{1t}^{*T}, \cdots, \alpha_{nt}^{*T}\}^T$ where $\alpha_{it}^* = \{\alpha_{it,0}, \alpha'_{it,0}, \cdots, \alpha_{it,4}, \alpha'_{it,4}\}^T$. The transition matrix from time *t*-1 to *t* by Δt is given by $H_t = \text{diag}\{H_{1t}, \cdots, H_{(5+5\times n)t}\}$ with $H_{gt} = \{1, \Delta t; 0, 1\}$ $g = 1, \cdots, (5 + 5 \times n)$ and the system error variance is given by $W_j = \text{diag}\{\lambda_1 W_{1j}, \cdots, \lambda_{(5+5\times n)} W_{(5+5\times n)j}\}$ with $W_{gt} = \left\{\Delta t^3/3, \Delta t^2/2; \Delta t^2/2, \Delta t\right\}$ where $\lambda_k = \lambda_{bk}$ if $k \le 5$, $\lambda_k = \lambda_{al}$ otherwise where *l* is the modulo of *k*-5 divided by 5. The initial values at *t*=0 are modeled as $\beta_0^* \sim N(0, \tau I)$ with $\tau \to \infty$, and $\alpha_{i0}^* \sim N(0, \bar{D})$, $\bar{D} = \text{diag}\{\sigma_1^2 D, \cdots, \sigma_5^2 D\}$ $i = 1, \cdots, n$.

Following Koopman and Durbin (2000) and Guo (2002), univariate Kalman filtering and smoothing are applied to a multivariate state space model to alleviate the heavy computational demand. Let θ be the vector with all the unknown parameters and $Y_t = \{\vec{y}_1, \cdots, \vec{y}_t\}$. The likelihood is

(3.10) $L(\theta \mid Y_T) = p(\vec{y}_T \mid Y_{T-1}, \theta)\, p(\vec{y}_{T-1} \mid Y_{T-2}, \theta) \cdots p(\vec{y}_2 \mid Y_1, \theta)\, p(\vec{y}_1 \mid \theta)$

where $p(\vec{y}_t \mid Y_{t-1}, \theta)$ is sequentially obtained using the Kalman filter. Conditional on the

estimates of the parameters, a smoothing algorithm runs sequentially backward to obtain the posterior estimate of $\hat{\beta}_{t,k} = E[\beta_{t,k} \mid Y]$ and $\hat{\alpha}_{it,l} = E[\alpha_{it,l} \mid Y_i]$ along with its posterior variance as $V[\beta_{t,k} \mid Y]$ and $V[\alpha_{it,l} \mid Y_i]$. Full estimation details are shown in Appendix in Section 3.7.3.

3.5 Empirical Findings

I first compare the proposed spatio-temporal model with reduced models and demonstrate its ability to describe both the spatial and temporal dimensions of the raw data. Next I present the parameter estimates of key variables, interpret them, and discuss social influence by acquisition type on the emergence of new customers. I also briefly discuss the control variables.

3.5.1 Model Fits

I compare the performance of the proposed model with reduced models that do not include some of the random variation components in the model parameters. In performing the comparisons I retain the fixed parameter paths ($\beta_{t,k}$) for all variables, but selectively turn on the random parameters ($\alpha_{it,k}$). (In the most restricted model, all random parameter paths are turned off, resulting in the conventional time-varying parameter model.) These reduced models are re-estimated and compared with the proposed model by using restricted maximum likelihood estimation (REML) (Guo 2002,

2003; Wahba 1978, 1983).[21] Table 3.2 shows that the log-likelihood for the proposed model is significantly better the reduced models—this indicates that the full model offers the better description of the actual data. Figure 3.1 panels (a) and (b) show both the actual number of customers (y_{ijt}) and the fitted number of customers according to the full model ($\hat{y}_{ijt}$) on the temporal and spatial dimensions, respectively, after aggregating over space and time. Figure 3.1 (a) shows that the model does an excellent job of tracking the growth over time in the number of new customers over the thirteen quarters of data. Figure 3.1 (b) ranks zip codes according to the cumulative number of new customers. It shows that the model does a very good job of explaining variation in performance by spatial location.

Table 3.2: Model Fit Comparison

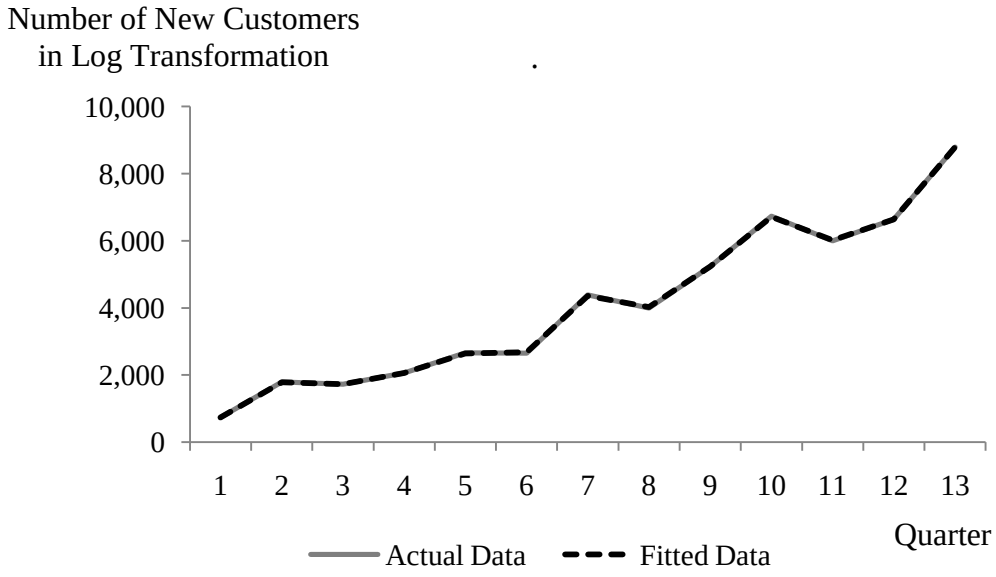
Model	**LL**[b]
Fixed parameter paths, $\beta_{t,k}$ *, for all variables*	
Random parameter paths, $\alpha_{it,k}$ *, for the following variables*[a]	
1. None (that is, time-varying parameter model)	11,299
2. Intercept	11,491
3. Intercept + Social influence from the installed WOM customers	11,746
4. Intercept + Social influence from the installed search customers	11,680
5. All variables	11,767

Note: $k = 0$ for the baseline; $k = 1, 2$ for within-zip and across-zip social influence from the installed WOM customers; $k = 3, 4$ for within-zip and across-zip social influence from the installed search customers.

[21] Log-likelihoods are computed as equation (3.17) in Appendix in Section 3.7.3 suggested by Koopman and Durbin (2000) and Guo (2002). Under the null hypothesis of, the asymptotic distribution of the likelihood ratio test statistic is a 50: 50 mixture of χ_3^2 and χ_2^2, where χ_ν^2 is the central Chi-square distribution with ν degrees of freedom.

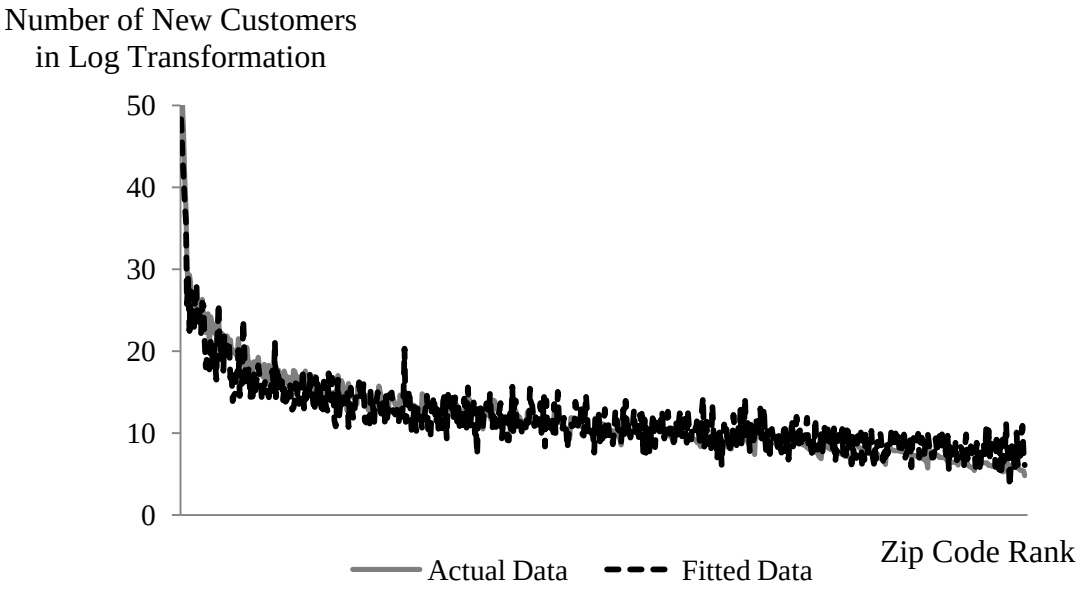
Figure 3.1: Model Fits over Space and Time

(a) Actual versus Fitted Number of New Customers in Log Transformation over Time



Note: Quarters on the *x*-axis are from the first quarter in 2005 to the first quarter in 2008.

(b) Fitted versus Actual Number of New Customers in Log Transformation over Space



Note: Zip code locations on the *x*-axis are given by the rank in terms of cumulative customers.

3.5.2 Parameter Estimates and Interpretation of Social Influence

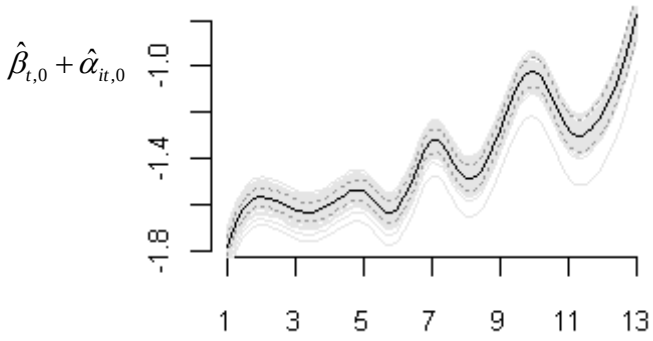
Overall Temporal Path. Figure 3.2 shows the trajectories of the spatio-temporal parameter estimates for the thirteen quarters of data. Recall that the estimate paths are meaningful at any time point, conditional on the smoothing parameter. Black solid and dotted lines denote posterior means and 95% confidence intervals of fixed parameter path ($\hat{\beta}_{t,k}$), respectively, and gray lines represent county-specific parameter paths ($\hat{\beta}_{t,k} + \hat{\alpha}_{it,k}$) after combining posterior means of fixed and random paths. I discuss the fixed parameter paths of interest and then move on to the county-specific parameter paths.

The temporal non-stationarity in the raw data in Table 3.1 (a) is closely followed by the baseline $\hat{\beta}_{t,0}$. The fixed parameter paths for social influence, $\hat{\beta}_{t,1}$ to $\hat{\beta}_{t,4}$, are (weakly) greater than zero in every period indicating positive social influence from existing customers. (Recall that I imposed no imposed restriction on the parameter paths and did not constrain them to be positive). Here, I note interesting observations on how the fixed parameter paths, $\hat{\beta}_{t,1}$ to $\hat{\beta}_{t,4}$ vary across the variables of social influence. First, the parameter paths for *within*-zip influence, $\hat{\beta}_{t,1}^{WW}$ and $\hat{\beta}_{t,3}^{WS}$, increase over time while the parameter paths for *across*-zip influence, $\hat{\beta}_{t,2}^{AW}$ and $\hat{\beta}_{t,4}^{AS}$ decrease. That is, within-zip social influence gets larger as more customers are acquired in a focal market, while the marginal increase from additional prior customers in neighboring zip codes gets smaller.

Figure 3.2: Posterior Means of Parameter Paths

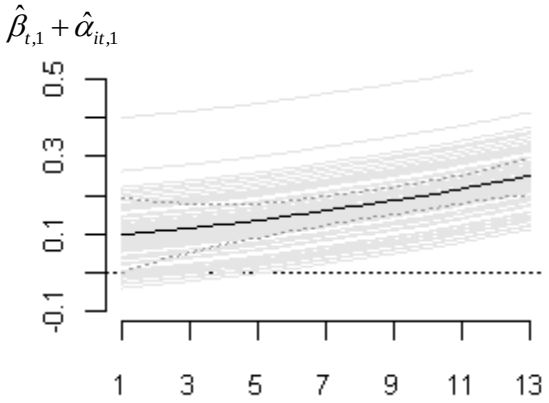
Black solid and dotted lines denote posterior means of the fixed paths ($\hat{\beta}_{t,k}$) and the corresponding 95% confidence intervals, respectively. Gray lines represent the posterior means of county-specific parameter paths ($\hat{\beta}_{t,k} + \hat{\alpha}_{it,k}$).

(a) Parameter paths for baseline

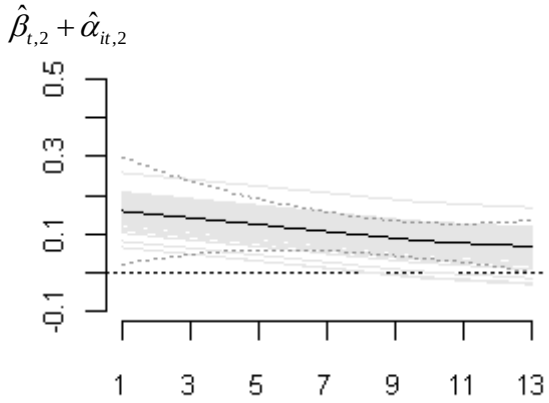


(b) Parameter paths for social influence from the installed customers acquired by WOM

Within-Zip Influence

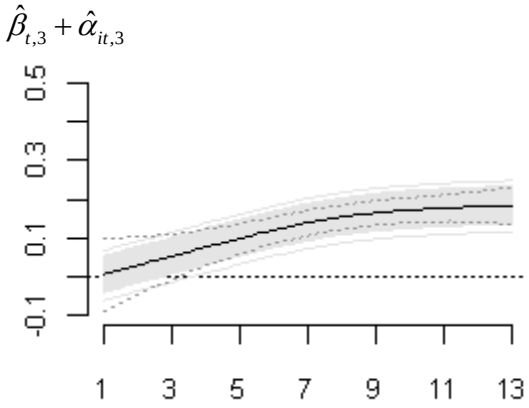


Across-Zip Influence

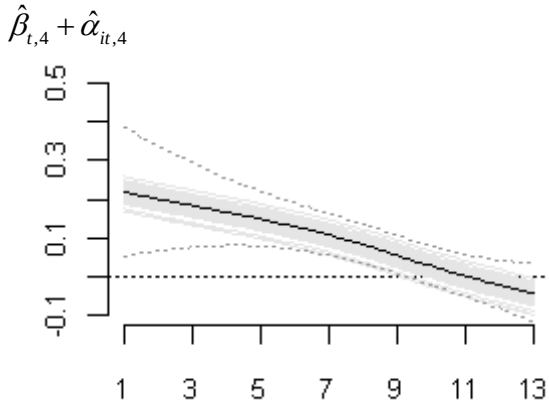


(c) Parameter paths for social influence from the installed customers acquired by search

Within-Zip Influence



Across-Zip Influence



Second, the parameter path for the *within*-zip influence from the installed WOM customers $\hat{\beta}_{t,1}^{WW}$ is larger than that for the installed search customers $\hat{\beta}_{t,3}^{WS}$, and $\hat{\beta}_{t,1}^{WW}$ continues to increase whereas $\hat{\beta}_{t,3}^{WS}$ levels off with time. Thus, customers acquired by WOM increasingly generate more new customers. Customers acquired by search also generate new customers, but to a lesser extent. Third, the parameter path for the *across*-zip influence from the installed WOM customers $\hat{\beta}_{t,2}^{AW}$ decreases steadily, and that for the installed search customers $\hat{\beta}_{t,4}^{AS}$ fades out quickly with time. The increasing within-zip parameter paths and decreasing across-zip parameter paths together suggest that social influence has a very "local" quality—most of the effect occurs within the *same* zip code. Moreover, search-acquired customers in neighboring zips have a more limited effect on future new customers in the focal zip code than WOM-acquired customers.

Table 3.3: Parameter Estimates

(a) Functional Parameter Estimates

Parameters	**Fixed Parameter Path**	**Random Parameter Path**		
	$\lambda_{\beta k}^{-1/2}$ Smoothing	$\lambda_{\alpha k}^{-1/2}$ Smoothing	$\sigma_{\alpha k,1}^{2}$ Variance	$\sigma_{\alpha k,2}^{2}$ Variance
Intercept	385.914	1.960×10^{-2}	7.224×10^{-2}	2.114×10^{-2}
Within-Zip WOM	6.247×10^{-2}	5.679×10^{-4}	1.040×10^{-1}	5.620×10^{-5}
Across-Zip WOM	8.051×10^{-2}	5.530×10^{-4}	6.176×10^{-2}	5.620×10^{-5}
Within-Zip Search	7.931×10^{-2}	5.466×10^{-4}	5.762×10^{-2}	5.620×10^{-5}
Across-Zip Search	1.884×10^{-1}	5.518×10^{-4}	4.715×10^{-2}	5.620×10^{-5}

Note: The null hypotheses of $\beta_{t,k}=0$ and $\alpha_{it,k}=0$ are rejected and all the fixed and random parameter paths are significant.

(b) Non-Functional Parameter Estimates

Variable	Estimate	Standard Error
Relative Convenience		
One-Day Shipping	0.034*	0.004
Two-Day Shipping	0.040*	0.003
Three-day Shipping	0.016*	0.003
Net Promoter Scores		
Net Promoter Scores	0.028+	0.017
Relative Online Prices		
Local Offline Conditional Sales Tax Rate (%)	0.0005+	0.0003
Access to Offline Retailers		
Distance to the Nearest Supermarket	0.009*	0.001
Distance to the Nearest Wal-Mart or Target	0.006*	0.000
Distance to the Nearest Warehouse Club	0.001	0.001
Socio-Demographic Controls		
Log (Number of Children $\leq$ 4 Years Old Yet to Adopt)	0.177*	0.001
Population Density	0.008*	0.000
Population Growth Rate (2000-2004)	0.029	0.040
Percentage Population Aged 20 to 39 Years Old	0.152*	0.016
Percentage Households with Working Female	0.120*	0.014
Percentage of Whites	0.137*	0.005
Percentage with College Education	0.311*	0.009
Percentage Households Earning $75,000 and more	0.111*	0.011
Percentage Homes Valued at $250,000 or more	0.265*	0.005
Percentage Apartments with 50 or more units	0.119*	0.013
Error Variance		
σ_e^2	0.237*	0.000

Note: $^{*}p < 0.05$, $^{+}p < 0.1$.

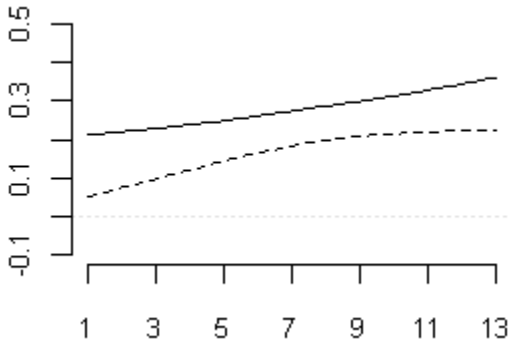
The second and third findings support the earlier conjecture for the different "qualities" of customers by acquisition process; the installed WOM customers are of better quality than the installed search customers in the sense that they exert greater social influence, i.e., they produce more new customers in the future. This finding is consistent with a large body of research in marketing (e.g., Villanueva, Yoo, and Hanssens 2008) that demonstrate that WOM is in general a superior acquisition mode for new customers. This seems to be true in both traditional and Internet settings.

Spatial Variation in County-Specific Temporal Paths. So far I have documented the overall temporal parameter paths in social influence from existing to new customers. It is also important to consider spatial variation. County-level spatial variation in parameter paths provides several new insights into how online demand evolves over time and over space according to the customer types in the installed base. Spatial variation in parameter paths is greater for the installed WOM customers than for the installed search customers. This is especially pronounced in the case of the parameter paths for *within*-zip influence from the installed WOM customers. Interestingly, the existence of this heterogeneity in the parameter paths suggests the *overall superiority* of the WOM acquisition process over the search acquisition can be reversed in some local markets.

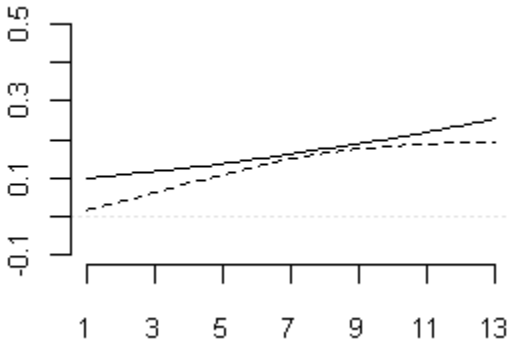
Figure 3.3: Parameter Paths in Four Counties in the State of California

Solid and dotted lines represent parameter paths for within-zip social influence from the installed WOM and search customers, respectively, in four selected counties in the state of California. Los Angeles County, Orange County and San Diego County are *contiguous* in the order by which they are listed. Sacramento County is *distantly* located from the three counties. For ease of exposition parameters for within-zip social influence are plotted.

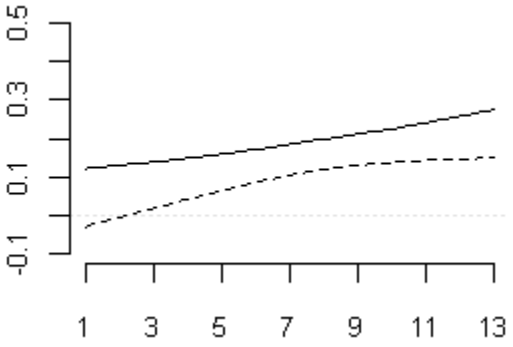
(a) Los Angeles County



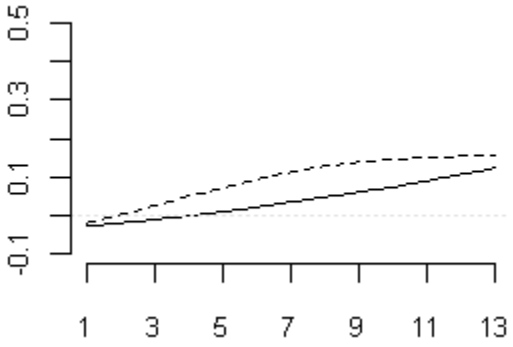
(b) Orange County



(c) San Diego County



(d) Sacramento County



To illustrate this heterogeneity across counties in the parameter paths, I pick three *contiguous* counties and one *distant* county in California and plot their parameter paths in Figure 3.3. For ease of exposition parameter paths for *within*-zip social influence are plotted. Solid and dotted lines denote the posterior means of parameter paths for within-zip influence from the installed WOM and search customers, respectively. The three

contiguous counties (Los Angeles, Orange, and San Diego) show different parameter paths in terms of not only the estimates' magnitudes but also the relative effectiveness. In Los Angeles County and San Diego County, the parameter path for *within*-zip influence from the installed WOM customers is larger than that from the installed search customers, while in Orange County, both are more or less the same. On the other hand, in Sacramento County, the parameter path for *within*-zip influence from the installed search customers is *larger* than that from the installed WOM customers. That is, in this county, it is potentially more profitable to the firm to acquire new customers via search since customers acquired in this way exert more social influence over potential customers in the future. In summary, the across-location discrepancy in which acquisition mode is locally-superior suggests that a firm can enhance business performance by locally adjusting its acquisition channels.

Further insights are obtained by decomposing the expected number of new customers per location and time period ($\hat{y}_{ijt}$) into three separate marginal effects. Specifically, the marginal effects come from $BASE_{ijt}$, SI_{ijt}^{WOM}, and SI_{ijt}^{SEARCH}, as denoted in equation (3.2). (As *within*-zip influences are dominant, SI_{ijt}^{WOM} and SI_{ijt}^{SEARCH} are not further decomposed into *within*-zip and *across*-zip influence). Figure 3.4 (a) and (b) show the marginal effects over time (after aggregating over space) and over space (after aggregating over time), respectively. The darkest area at the bottom of Figure 3.4 (a) indicates the baseline effect, $BASE_{ijt}$, i.e., the number of new customers acquired in a particular quarter due to location-specific factors. These include access to offline stores and zip code characteristics such as the size of the target population, etc. The middle light gray area

and top dark gray area represent the new customers accounted for by the social influence from the installed WOM and search customers, SI_{ijt}^{WOM} and SI_{ijt}^{SEARCH}, respectively.

Over time WOM-generated customers are increasingly important in generating new customers; search-generated customers are also important, but their relative contribution appears to have stabilized. Thus, Figure 3.4 (a) shows that social influence increasingly brings more new customers to Childcorp.com over time. This suggests that the success at Childcorp.com is attributable not only to the natural or organic growth but also to the active social interactions between already acquired customers and potential new customers.

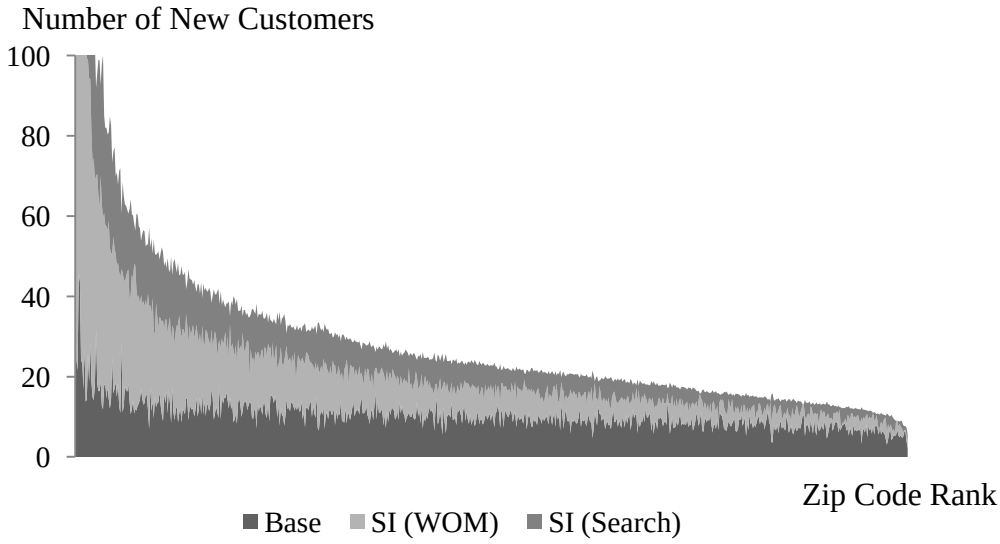
In Figure 3.4 (b) the location of zip codes on the *x*-axis is again determined by their performance in generating new customers. Since I focus on relatively homogenous zip codes within MSAs the baseline effect, i.e., organic contributions due to local market factors, does not vary much from the best to worst-performing zips. However, the implied social influence effects are not uniform. The best-performing markets benefit far more from social influence effects than the worst-performing markets do. Furthermore, the relative superiority of installed WOM customers over installed search customers is particularly prominent in the best-performing markets. Recall that parameter paths for the installed search customers are larger than those for the installed WOM customers in certain counties. However, as the installed WOM customers are often larger in absolute terms than the installed search customers in many local markets, the marginal effect from the installed WOM customers becomes larger after combining the parameter paths and the size of the installed base.

Figure 3.4: Marginal Effects of Social Interactions

The fitted number of new customers under the proposed model is decomposed into the baseline effect (denoted as Base), and social influence from the installed customers acquired by word-of-mouth and search, separately (denoted both as SI(WOM) and SI(Search)).

(a) Number of New Customers in Each Quarter.

(b) Number of New Customers over Space



Thus, it appears that customers acquired by WOM are on average of higher value to the firm than those acquired by search (see also Villanueva, Yoo, and Hanssens 2008). The fixed parameter paths for the installed WOM customers are larger than those for the installed search customers. However, the superiority of customers acquired via WOM fails to hold in every market. Some counties imply better profitability from acquiring new customers via search than via WOM. This suggests that the relative superiority of an acquisition mode needs to be addressed at the local level. (I elaborate on the implications stemming from the heterogeneity of parameter paths across counties in the following section.) The pattern of parameter estimates along with the superior fit of the proposed model demonstrates that it is critical to take into account the spatial variation in temporal dynamics in order to avoid model misspecification and biased inferences. Such an approach is necessary to fully understand demand evolution at an Internet retailer over space and time.

Change in the Mix of Customer Types. I examine the effect of the mix of two types of customers in the installed base on the firm's growth through social influence. The key question is: what if Childcorp.com had acquired one more customer of the type that a local market favors and one less of the other type? That is, I hold the total number of already acquired customers in time *t constant*, but slightly change the mix of two types of customers in every zip code. I then compute how many new customers the change in the mix bring to Childcorp.com from time *t* onward (through Q13), and average the differences in the numbers of customers over time and over all the locations.

Were the *mix of customer types* and not the *total number* of customers, changed slightly in the fourth quarter in 2005, 2006, and 2007, Childcorp.com would have

acquired 0.035, 0.055, and 0.094 more customers per zip code per quarter, respectively, by the end of data period. This suggests that the *aggregate* installed customer base alone cannot fully explain the firm's future growth. One must also consider the *decomposed* installed base by acquisition type. Also, Childcorp.com benefits more from acquiring "right" types of customers over time as the effect of social influence from the installed customer base increases.

Moreover, I examine the *number of zip codes* which favor search acquisition (i.e., favoring the case in which one search customer is added and one WOM customer is dropped over the opposing case where a WOM customer is added and a search customer is dropped). Not all the markets favor WOM acquisition. About 35% of local markets would be better off if they acquired customers via search.[22] That is, some local markets are better able to breed social influence from the installed search customers while the others benefit more from social influence from the installed WOM customers. Examining the decomposed installed base enables the firm to better allocate acquisition efforts over channels. Clearly a traditional modeling approach which looks at the customer base in aggregate and independent of acquisition mode cannot accomplish this.

[22] A logit analysis using zip-level socio-demographic characteristics in Table 3.1 (b) shows that in general WOM acquisitions are more likely in zip codes that have a higher value of population density and population growth rate, smaller percentage of households of child-bearing age, wealthy people, and urban housing units, and inconveniently located discount stores. In future research, I plan to further investigate these differences due to local characteristics.

3.5.3 Parameter Estimates and Interpretation of the Control Variables ($\vec{\tau}$).

Local markets differ in their average propensity to rely on Childcorp.com. Organic growth in new customers in a local market will reflect unique aspects of the local environment (such as distances to offline stores, local assortments, prices, etc). Measures of observed heterogeneity (i.e., the covariates listed in Table 3.1) serve the role of control variables, as the primary substantive interest is in the evolution of social influence just discussed. The signs and magnitudes of the coefficients have face validity, and here I simply note a few interesting observations that may warrant future studies.

More new customers emerge in areas with greater market potential, i.e., areas with a larger number of babies. Service satisfaction measured by higher Net Promoter Scores significantly improves online demand. A relative price advantage (i.e., higher offline taxes) and convenient access to products (i.e., faster shipping) also contribute to online demand growth. In general, Childcorp.com has greater demand in zip codes that have more population density, a greater percentage of households of child-bearing age, and more females in labor force. Furthermore, zip codes with a greater percentage of whites, college-educated individuals, higher incomes, higher home values, and more urban housing units show higher organic growth. Lastly, online demand is positively correlated with the relative inaccessibility of offline stores; the greater distances to offline supermarkets and discount stores the higher online demand.

3.6 Conclusion

An Internet retailer acquires new customers through search and WOM and both types of customers in the installed base influence potential customers. As the two acquisition modes can bring different "qualities" of customers to a firm, I construct two separate variables to measure social influence from the installed WOM customers and social influence from the installed search customers. Unlike traditional retailers operating within fixed trading areas, Internet retailers acquire customers over time from virtually everywhere. To reflect this, I specify a functional fixed effects model and I allow temporal paths to measure the effect of social influence on new customer acquisitions to vary over space.

I find that not all types of customers are created equal in terms of their ability to bring new customers to the firm. Customers acquired by word-of-mouth are of better quality to the firm than those acquired by search as the social influence fixed parameter paths for the installed WOM customers are larger than those for the installed search customers. Interestingly, substantial variation in the temporal parameter paths over counties suggests that not every market favors WOM acquisitions; the superiority of one channel over the other varies markedly over space. In the best performing zip codes the majority of new customer growth is driven by WOM-acquired customers. The worse performing zip codes produce almost as many customers *in absolute terms* from organic growth as the best performing ones do—they just produce very little from the installed base. Finally, I observe that social influence has a particularly local quality. The majority of the effect

occurs within a zip code—there is relatively little influence from customers in neighboring zip codes.

3.6.1 Future Research

There are at least three other avenues for future work. First, I focus on the zip-level local markets and one may want to model the underlying individual consumer choice process (see Jank and Kannan 2005). Second, I limit analyses to zip codes which have at least one customer in the first five quarters. One may want to increase the number of zip codes. As there will be much variation in adoption time, a two stage model can be employed, the first stage for adoption (e.g., Bell and Song 2007) and the second stage for growth since adoption by counting time since "birth", i.e., when the first new customers appear (e.g., Guo 2003). Third, I do not model the firm's marketing actions (e.g., advertising expenditures, price promotions) as they are not applicable in my data; instead, I focus on measuring how influential the installed customers are in contributing to future acquisition. One might want to develop a model to allocate acquisition and retention resources optimally for long-term profitability (e.g., Reinartz, Thomas, and Kumar 2005). I plan to address these issues in future research.

3.7 Appendix

3.7.1 Low Rank Spatial Smoothing of the NPS Loyalty Variable

The measures of Net Promoter related variable are volatile since many zip codes have small sales and even smaller responses. To improve the quality of this variable, I implement a low-rank thin plate spline smoother to spatially smooth it out (e.g., Choi, Hui, and Bell 2009; Wand 2003). In the discussion below, I provide an outline for the implementation; readers are encouraged to see Ruppert, Wand, and Carroll (2003) Chapter 13 and Wand (2003) for more details of the thin plate spline bivariate smoother including its theoretical justifications.

Step 1. Choose knots: I obtain "knots" based on centroids of a *k-d* tree (see also Molenberghs and Verbeke 2006 p. 379-384; Stremersch and Lemmens 2009). Starting with all the zip codes in MSAs, the *k-d* tree partitions the space until all partitions contain at most 20 locations. The region nearest to the centroid of each partition is chosen as a knot which creates 465 knots.

Step 2. Bivariate radial smoothing: I then apply the low-rank thin plate spline smoothing with a radial basis function. Bivariate smoothing is then based on the Euclidean distances from the set of knots, $z_1, z_2, ..., z_K$, and a proper covariance function.

3.7.2 Alternative Weighting Matrices

Two alternatives to define a proximity matrix are based on shared boundaries and contiguity information. The shared boundary weighting matrix is

$$G_{(p,q)} = \begin{cases} l_{pq} / l_p, & l_{pq} > 0 \\ 0, & otherwise \end{cases} \tag{3.11}$$

where l_{pq} is the length of zip code *p*'s boundary shared with zip code *q* and l_p is the total length of *p*'s boundary shared with all its contiguous zip codes, i.e., $l_p = \sum_q l_{pq}$. This weighting system is appropriate when two regions with a longer shared boundary might be expected to exert greater influence on each other. The shared boundary weighting matrix can be simplified to a case where two neighboring regions have equal influence on the focal region as long as they share boundaries with focal region, and this simpler form is called a contiguity weighting matrix,

$$G_{(p,q)} = \begin{cases} 1, & l_{pq} > 0 \\ 0, & otherwise. \end{cases} \tag{3.12}$$

3.7.3 Univariate Kalman Filtering and Smoothing to a Multivariate State Space Model

The dimensionality of fixed and random parameter paths introduces computational complexity to the estimation process. To alleviate this high computational demand, I

apply univariate Kalman filtering and smoothing to a multivariate state space model as suggested by Guo (2002) and Koopman and Durbin (2000).

I create a pseudo time series

$\{y_{111},\cdots,y_{1m_1 1},\cdots,y_{n11},\cdots,y_{nm_n 1},\cdots\cdots,y_{11T},\cdots,y_{1m_1 T},\cdots,y_{nm_n T}\}^T$ and re-define as

$\{y_{11},\cdots,y_{N1},\cdots\cdots y_{N1},\cdots,y_{NT}\}^T$ where $i=1,\cdots,N$. Kalman filter runs forward and the smoothing runs backward. The modified state space model is[23]

(3.13) $\quad y_{it}=M_{it}\vec{\gamma}_t+e_{it},\quad e_{it}\sim N(0,\sigma_e^2),\quad$ where M_{it} is the i^{th} row of M_t

(3.14) $$\vec{\gamma}_t=\begin{cases}H_t\vec{\gamma}_{t-1}+\vec{\omega}_t, & \vec{\omega}_t\sim N(0,W_t), \quad \text{if new time point}\\ \vec{\gamma}_t, & \text{otherwise}\end{cases}$$

Univariate Filtering. The object of filtering is to calculate the mean and error variance matrix of α_t given $\vec{y}_1,\cdots,\vec{y}_{t-1}$. Define $\vec{a}_{1t}=E[\vec{\gamma}_{1t}\,|\,Y_{t-1}]$ with $P_{1t}=Var[\vec{\gamma}_{1t}\,|\,Y_{t-1}]$ and $a_{i+1t}=E[\vec{\gamma}_{i+1t}\,|\,Y_{t-1},y_{1t},\cdots,y_{it}]$ with $P_{i+1t}=Var[\vec{\gamma}_{i+1t}\,|\,Y_{t-1},y_{1t},\cdots,y_{it}]$ for $i=1,\cdots,N$. Filtering equations are given by

(3.15) $\quad a_{i+1t}=a_{it}+K_{it}F_{it}^{-1}v_{it},\qquad P_{i+1t}=P_{it}-K_{it}F_{it}^{-1}K_{it}'$

where $v_{it}=y_{it}-M_{it}\vec{a}_{it}$, $F_{it}=M_{it}P_{it}M_{it}'+\sigma_e^2$, and $K_{it}=P_{it}M_{it}'$ for $i=1,\cdots,N$ and $t=1,\cdots,T$. The transition from time *t*-1 to time *t* is achieved by

(3.16) $\quad a_{1t}=H_t a_{N+1t-1}\qquad P_{1t}=H_t P_{N+1t-1}H_t'+W_t$

The values a_{1t} and P_{1t} are the same as the corresponding values from the standard

[23] When the variance of the vector of measurement errors, $Var[e_j]$, is not diagonal, the univariate representation of the multivariate state space model does not lead to an equivalent model because the correlations between the observation equations are lost. In this case, I can either put e_j into the state vector or use a singular value decomposition to make $Var[e_j]$ diagonal. See Koopman and Durbin (2002) for more detail.

Kalman filter.

Estimation. Let θ be the vector with unknown parameters. Outputs from the univariate filtering lets evaluate the log-likelihood via the prediction error decomposition for given θ.

(3.17) $$\log L(\theta) \propto -0.5\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\log F_{it} + v_{it}^2 F_{it}^{-1}\right)$$

Univariate Smoothing. The smoothed state vectors, conditional on the full set of observations, are evaluated by backward smoothing algorithm. The smoothing recursions are given by

(3.18) $$\hat{a}_{1t} = a_{1t} + P_{1t} r_{0t}, \qquad V_{1t} = P_{1t} - P_{1t} N_{0t} P_{1t}$$

where $r_{i-1t} = M_{it}' F_{it}^{-1} \upsilon_{it} + L_{it}' r_{it}$, $N_{i-1t} = M_{it}' F_{it}^{-1} M_{it} + L_{it}' N_{it} L_{it}$, $r_{Nt-1} = H_t' r_{0t}$, $N_{Nt-1} = H_t' N_{0t} H_t$, and $L_{it} = I - K_{it} M_{it} F_{it}^{-1}$ for $i = N, \cdots, 1$ and $t = T, \cdots, 1$. The initializations are $r_{NT} = 0$ and $N_{NT} = 0$. The transitions do not apply when *t*=1. The smoothed values $\hat{a}_{1t}$ and V_{1t} are the same as the corresponding values from the standard multivariate state space model.

Chapter 4

Preference Minorities and the Internet: Why Online Demand is Greater in Areas where Target Customers are in the Minority

4.1 Introduction

Local offline stores face trading area and space constraints, so the products they offer cater to the tastes of the local majority. As a result, agglomeration of individuals who share preferences improves their welfare as the local retail market brings forth products they want (Sinai and Waldfogel 2004). On the other hand, consumers whose preferences are dissimilar to the majority in trading area—*preference minorities*—are therefore likely to be under-served, or, perhaps, neglected by local retailers altogether. In this paper, we examine online demand from preference minorities; we explain why Internet retailers draw more sales from regions that contain them, holding the *absolute* number of target customers per region constant, and why members of the preference minority are less price-sensitive. Furthermore, the effect is exacerbated for niche products (relative to

popular products). We show that niche products in the tail of the Long Tail sales distribution (Anderson 2006) will draw a *greater proportion* of their total online demand from high preference minority regions. In summary, we explain why and how region-level category and brand demand emerges at Internet retailers through a process of virtual, rather than physical, agglomeration of customers. Specifically, we document how local market preference isolation explains variation in online consumer demand.

In formulating our hypotheses and empirical model we draw on recent developments in economics and economic geography. It is well known that larger markets deliver more product variety and increased consumption (Glaeser, Jolko, and Saiz 2001; Waldfogel 2003). In the online setting, the Internet has the potential to act like a "large market." A succinct explanation is given by Sinai and Waldfogel (2004, p. 3): "By agglomerating consumers into larger markets, the Internet allows locally isolated persons to benefit from product variety made available elsewhere." This has implications for how online demand might vary across locations by aggregating isolated diffuse preferences. Conceptually, there are two forms of isolation that might affect consumer demand online. The first is physical distance to offline alternatives. Recent research by Forman, Ghose, and Goldfarb (2009) shows that "isolation" in the form of increased distances to offline retailers leads to an increase in online demand for books. The second—and the focus of this paper—is preference isolation, holding physical distances to offline alternatives constant.

Profit-maximizing offline retailers allocate shelf space according to the Pareto or "80/20" rule (Chen et al. 1999; Reibstein and Farris 1995). Products are made available locally when they are wanted by a "sufficient" number of local neighbors so preference minorities with atypical needs are thereby implicitly harmed. This "push and pull" aspect

to distribution decisions is described by Farris, Olver, and De Kluyver (1989): "Retail buyers favor products that provide the greatest returns to the shelf space and the merchandizing resources allotted them" (p. 109). The goal of this research is to study how an Internet retail alternative assists local preference minorities, and what this implies for the Internet retailer's sales distribution over local markets. We propose and test four hypotheses on spatial variation in category and brand online demand.

The following example illustrates the main ideas. Imagine a local area in which the elderly are the majority of population. Since retailers have a fixed amount of space and their stocking rules take local preferences into account, young parents with newborns living in this area might not find a full assortment of baby diapers in the local market. That is, they assume the status of preference minorities. Local stores may still allocate *some* shelf space to baby products, but if they do, the brands and variety offered will be limited (e.g., perhaps restricted to the leading brand, Pampers).[24] The parents are not only unlikely to have local access to the full variety of Pampers, but also to other diaper brands as well. Therefore, the local market characteristic of a prevalent elderly population puts the young parents at a relative disadvantage when it comes to shopping locally for their newborns—something that is exacerbated when even more narrowly defined preferences are taken into account. Further suppose the newborn is sensitive to chlorine and the parents must use chlorine-free diapers, e.g., Seventh Generation. (Seventh Generation is a "niche product," i.e., a specialized product designed for a particular segment). As a relatively limited space is allocated to the *product category overall* it is even more difficult to purchase niche products locally.

[24] This is especially true for products such as diapers that are bulky and have high shelf space-to-profit ratios. We formalize this notion in the next section (see Figure 4.2).

Our data are from the leading online baby products retailer, Childcorp.com (see Section 4.3 for details).[25] As the largest U.S. online retailer carrying baby products, Childcorp.com provides an excellent setting for measuring differences across regions in online demand for diapers overall, and for specific brands. The diapers category has several features that make it well-suited for our study. First, per-capita consumption of diapers is relatively constant, and total consumption in a particular location is tied to the number of babies living there. Second, Childcorp.com carries leading national brands (Pampers, Huggies, and Luvs) and a leading niche brand (Seventh Generation) that is not available in all offline supermarkets. (We determine which offline stores in which locations carry this brand in order to control for region-level variation in access to popular and niche brands.) Third, the high shelf space-to-profit ratio for baby diapers limits product breadth and depth available in local markets more than would be the case for other products with lower shelf space-to-profit ratios (e.g., spices, vitamin pills, etc).

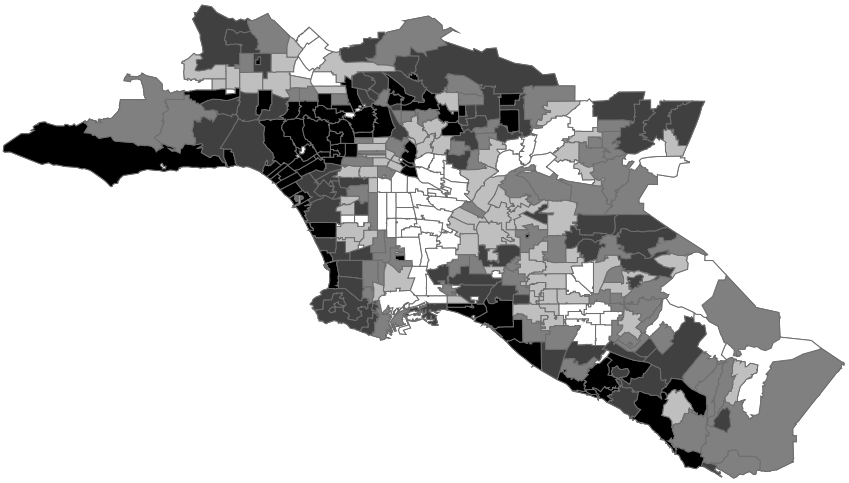
There is no absolute standard for defining "minor preferences"; hence, we define them by looking at the *relative size* of the target group in a local area. We construct a "preference minority index" (hereafter, the PM Index) using the following proportion defined at the local market level: [1 - (Target Population / Total Population)] (see Forman, Ghose, and Wiesenfeld 2008; Goolsbee and Klenow 2002; Sinai and Waldfogel 2004). The PM Index reflects a key assertion of our analysis: The amount of local product variety available *offline* to the target group depends on the *relative* size of the target group. Patterns in actual Childcorp.com data complement the illustrative example and help motivate our hypotheses (discussed shortly). Figure 4.1 (a) maps the extent to which

[25] For reasons of confidentiality I refer to this leading Internet retailer by the *nom de plume*, "Childcorp.com".

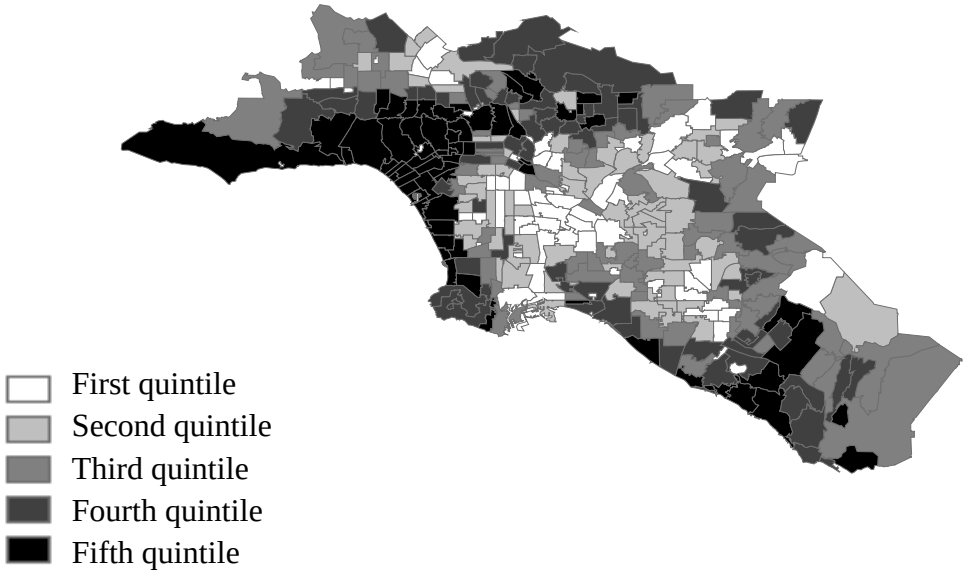
a target group in Los Angeles—"households with babies"—is a local preference minority. Figure 4.1 (b) maps the cumulative number of orders per target household placed at Childcorp.com. The shading illustrates a positive correlation. In zip codes where households with babies are in the minority, online sales per target household are higher.

Figure 4.1: Positive Correlation between Preference Minorities and Online Demand in Los Angeles County

(a) Zip-level PM Index



(b) Zip Level Orders per Target Household



We make four substantive contributions. First, we hypothesize and demonstrate that sales substitution, from offline retailers to online retailers, increases across local markets as their PM Indices increase, i.e., as the *relative size* of the target group decreases (H_1). Holding the characteristics of the local environment (and the total size of the target group) constant, online sales are higher in markets where the target group is more of a preference minority. On average, online sales in "high PM" markets (at the 90th percentile of the PM Index) are roughly 56% higher than in "low PM" markets (at the 10th percentile), *even though both these markets contain the same number of potential customers*. Second, we show that preference isolation also affects price sensitivity (H_2). Online demand in preference minority markets is less sensitive to relatively favorable online prices, given the difficulty in obtaining products. When the price advantage at Childcorp.com increases demand in low PM markets increases substantially (up 24%). Conversely, the demand increase in high PM markets is a more modest 8%.

Third, we find an interaction with the types of brands sold (H_3). By definition, total niche brand sales are lower than total popular brand sales; however, local online sales of "niche" brands respond more strongly to the presence of preference minorities than local online sales of "popular" brands do. Relative to "low PM" markets (at the 10th percentile of the PM Index), "high PM" markets (at the 90th percentile) have local *online* sales of popular brands that are about 50% higher, and local *online* sales of niche brands that are about 175% higher, *even though both markets contain the same number of potential customers*. Preference minorities turn online to alleviate the constraint of their limited local options and this online agglomeration intensifies for members of the preference minority who favor niche brands.

Fourth, we show that the difference in the sales sensitivity to the presence of preference minorities between popular and niche brands has an important implication for the Long Tail sales distribution (H_4). Popular brands and niche brands both have more offline-to-online substitution in local markets with a higher PM Index. However, niche brands with a lower overall sales rank (i.e., those in the "tail" of the Long Tail) draw a greater *proportion* of their total online demand from high PM regions.

The paper is organized as follows. The next section summarizes key ideas from the extant literature, introduces a conceptual framework, and describes the hypotheses. The subsequent section describes the data and measures, and validates some key empirical assumptions. Next, we describe the empirical model and report and interpret the findings. The paper concludes with a discussion of the implications for retailing theory and practice and for future research.

4.2 Background, Conceptual Framework, and Hypotheses

4.2.1 Market Size, Variety, and Preference Minorities

As noted in the Introduction, larger markets bring forth more variety and more consumption. Aggregation of like-minded consumers allows firms to serve their tastes and in turn allows individual consumers to find products locally that suit their needs. This is particularly important when the fixed cost of product provision is high. A classic example is the provision of media (e.g., newspapers and television programming); specialist products such as Spanish language television emerge only when there is

sufficient local density of customers demanding them (e.g., Waldfogel 2003). The fixed cost argument applies directly to retail settings as firms face hard constraints on shelf space. When products are unavailable locally because they are insufficiently attractive to overcome the fixed cost constraint of shelf-space allocation, the Internet has the potential to act like a "large market" by aggregating diffuse preferences across different geographical locations. This idea that "preference minorities" who are locally isolated might therefore turn to the Internet is rooted in recent work in economics and economic geography and can even be traced back to Central Place Theory (Christaller 1933).

4.2.2 Online-Offline Demand Substitution

Online retailers can offer several benefits to consumers including, lower prices (Brynjolfsson and Smith 2000; Chiou 2005; Goolsbee 2000), greater convenience (Balasubramanian, Konana, and Menon 2003; Cairncross 1997; Keeney 1999), and more variety (Brynjolfsson, Hu, and Raman 2008; Ghose, Smith, and Telang 2006). Among factors studied, price has received the most attention. Brynjolfsson and Smith (2000) and Goolsbee (2000) find that consumers shop online for lower prices and to avoid local sales tax, respectively. Anderson et al. (2009) find that when retailers open physical stores in a location—and acquire a nexus for tax purposes—Internet sales at that location suffer (since the firm now has to charge sales tax). Forman, Ghose, and Goldfarb (2009) find that when conventional booksellers enter specific offline locations Amazon.com sales at those locations decline. Increased convenience of offline alternatives mitigates the attractiveness of the online alternative. Thus the value proposition of Internet retailers to

consumers is determined by where those consumers live, which in turn directly reflects the options and constraints they face in the local offline market.

4.2.3 Local Preference Minorities and "Compromised Demand"

We focus on Internet retailing, yet the key ideas relate back to classic findings developed in the distribution channels literature prior to the introduction of the Internet. Farris, Olver, and De Kluyver (1989) and Reibstein and Farris (1995) note that not all consumers can find their first choice brands in all local stores. While market leader brands tend to be stocked in all stores in a local market, niche brands tend to be stocked only in local stores with considerable shelf space. A shopper looking for a niche brand, but shopping in a convenience store (for example), might be forced to buy the popular brand, as it is the only brand stocked in the category. In this case, the popular brand sales at the convenience store represent "compromised demand" (see Farris, Olver and De Kluyver 1989, p. 114, Figure 4).

Not all brands are distributed through all local stores because offline retailers are constrained by: (1) fixed retail space, i.e., shelf space and inventory space, and (2) the preferences of customers within the trading area. Due to high fixed costs of product provision in retail stores, rational shelf space allocation dictates that not all product categories, or brands within a category, "make the cut" (Farris, Olver, and De Kluyver 1989; Anderson 1979). A member of the preference minority in a local area (such our family with a newborn described in the Introduction) is like a customer who is forced to shop at a "convenience store" and who therefore faces limited or negligible assortment in

product categories of direct interest.

Figure 4.2 illustrates the idea by contrasting two product categories, diapers and vitamin pills, which differ substantially in their shelf space-to-profit ratios. In Markets A and B, retailers allocate shelf space to product categories in accordance with the different tastes of local customers. The allocation mechanism is simply a discrete probability distribution where the percentage of space allocated to a category is proportional to the size of the customer group that wants it. In Figure 4.2 (a), we show this with seven groups of declining size. The group "households with babies" is ranked third in Market A (17% of the total), but sixth in Market B (7%), while the group who wants vitamin pills is ranked seventh in both markets (3% in Market A and 2% in Market B).

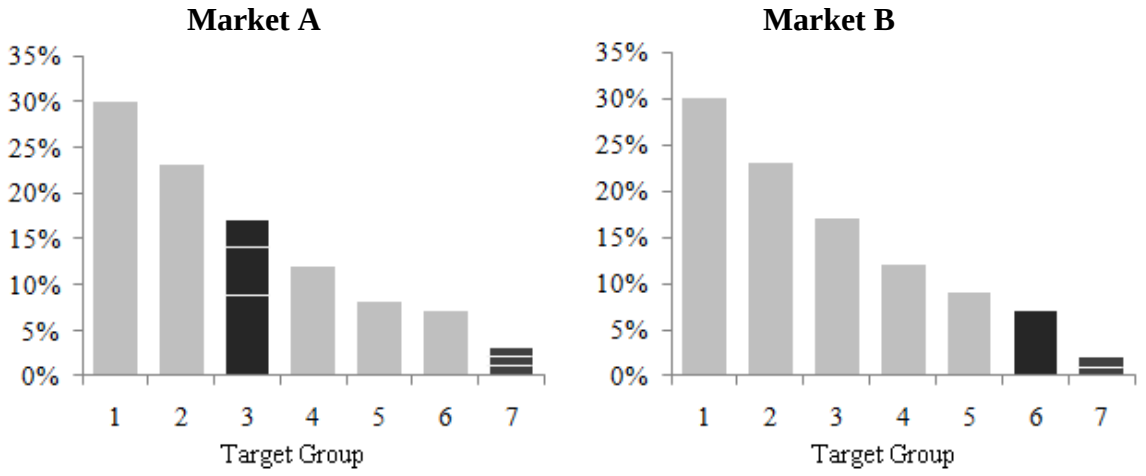
Space allocated to the category must then be further parceled out to different brands within the category. Stocking rules follow a step function whose steepness is determined by the shelf space-to-profit ratio of the category; Figure 4.2 (b) shows the stocking rules for diapers and vitamins. Since diapers are bulky, they will not be stocked locally unless 5% of the population has babies and every 5% increase in the baby population will bring one additional brand into the local market. Conversely, one additional brand of vitamin pills will be stocked with every 1% increase in the target population.

In the diaper category, three brands are available in Market A, whereas only the market leader brand is available in Market B. In the category for vitamin pills, although the target population is small, three brands are available in Market A and two brands are available in Market B. Therefore, the diaper category is expected to have more "comprised demand" than the category for vitamin pills. Moreover, Market B will have even more "comprised demand" for diapers than Market A, and consumers residing there

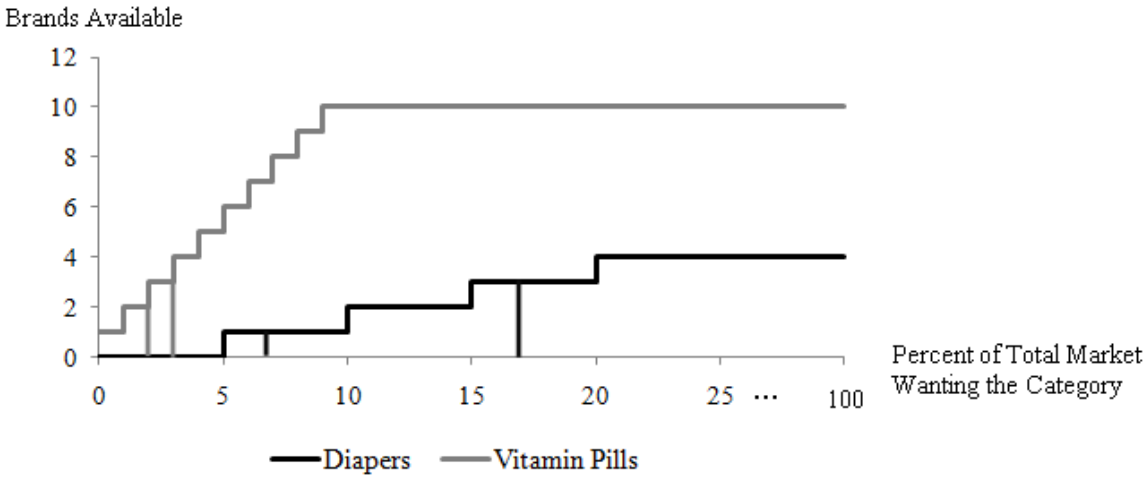
will be less satisfied with their local offline options for baby products (Fornell 1995).

Figure 4.2: Shelf Space Allocation for Categories and Brands in Local Markets

(a) Two Hypothetical Local Markets



(b) Hypothetical Stocking Rules for Diapers and Vitamins



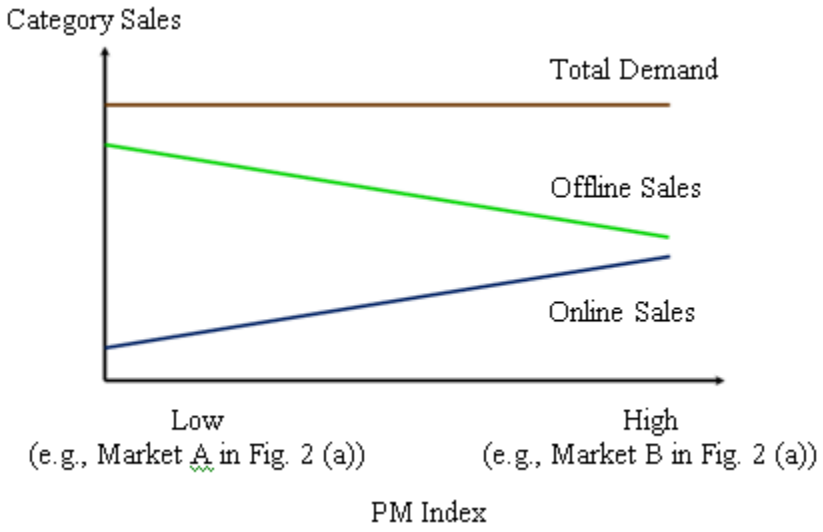
4.2.4 Hypotheses

Category Sales Online (H_1). The idea that individuals with heterogeneous preferences for public goods sort into different *physical* neighborhoods underlies much of the analysis in urban economics (see especially Tiebout 1956 and Dowding, John, and Biggs 1994 for a comprehensive review). This idea translates to private goods and online activities as follows. Local consumers with minority preferences living in geographically separate places may similarly *sort into online retailers* (i.e., *virtual* neighborhoods) to take advantage of essentially unlimited product assortment available online.

The interplay between local consumer demand and local retailers' stocking decisions affects the proportion of total local demand satisfied online versus offline. In Figure 4.3 the PM Index, i.e., 1-(Target Population)/(Total Population), increases across different local markets from left to right. This is because while the *total* population in a local market increases from left to right, the size of the target population stays the same. Imagine that a local target group (say households with babies) has relatively fixed per-capita consumption of a category (say diapers). That is, total consumption of the category in a local market has to be proportional to the size of a target group. In "high PM" markets (Market B in Figure 4.2) the focal group is small relative to other local groups (i.e., households without babies), so local retailers allocate limited space and attention to the category sought by the preference minority. Hence, customers are driven online. This idea is expressed in Figure 4.3 which shows that as the PM Index goes up, so does the online demand share.

H_1: Local Preference Minorities and Category Sales. Substitution from offline retailers to online retailers will be greater in markets that have a higher PM Index.

Figure 4.3: Local Preference Minorities and Category Sales



Note: The *x*-axis varies across local markets with the same-sized target population, but different total populations. The *y*-axis is total category demand from the target group. The PM Index is: [1-(Target Population)/(Total Population)]. Since the size of the target population is held constant, the total consumption by the target group is also constant over markets. When the size of the *total* population increases from left to right, then the target population becomes a smaller fraction of the total population from left to right, i.e., the target group's preferences become "more minor" from left to right. "High PM" markets have relatively limited local product availability; hence, consumers in these markets should be more likely to buy online. Conversely, in "Low PM" markets where the target population is a significant portion of the total population, offline alternatives will be relatively plentiful.

Category Price Sensitivity (H_2). Members of the preference minority will be less price-sensitive given the difficulty they have in obtaining what they want offline. While Childcorp.com charges identical prices in all markets, the *relative differential* between online and offline prices varies according to the presence or absence of local taxes on diapers. When the PM Index is low (i.e., there is a good amount of offline variety) and

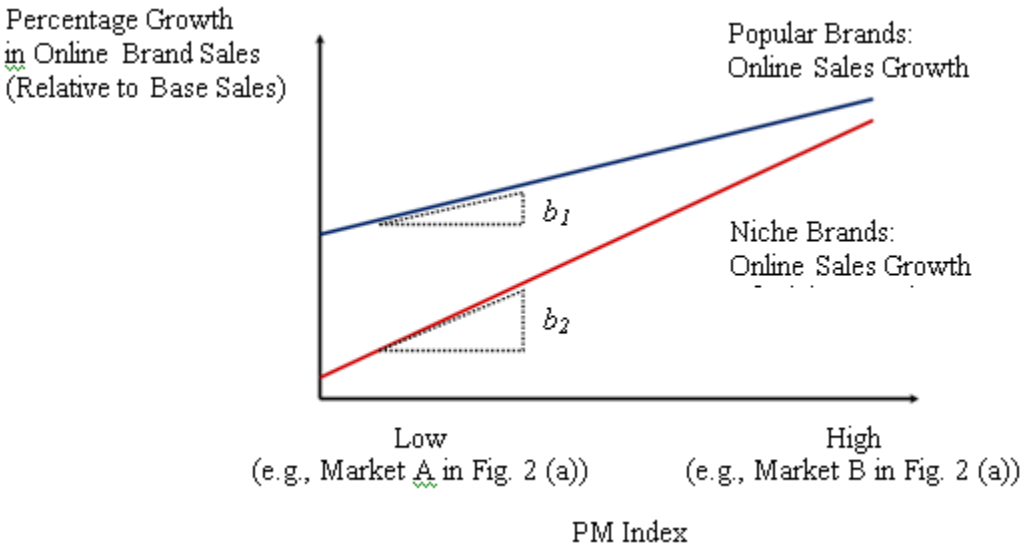
there are no offline taxes, more shopping should be done offline. Conversely, when the PM Index is high, shoppers should still be going online *even when* online prices are "relatively high" (i.e., there is no offline tax).

H_2: Local Preference Minorities and Price Sensitivity. Markets with a higher PM Index show diminished sensitivity to the *online price advantage*, i.e., consumers in these markets continue to shop online even when offline sales tax rates are low.

Brand Sales: Popular versus Niche (H_3). If a local retailer decides to stock a category of potential relevance to the preference minorities at all, she will most likely choose a "popular" brand such as a leading national brand (Farris, Olver, and De Kluyver 1989; see also Figure 4.2). Niche brands in preference minority markets are therefore subject to double jeopardy. By definition, fewer consumers prefer niche brands; hence, *even fewer* local retailers in the preference minority market will stock them. Conversely, brands with high sales and large market shares further increase market share and sales through a positive-feedback process (Reibstein and Farris 1995). Hence, category-level online-offline substitution in H_1 will intensify when local consumers in the preference minority do not favor "popular" brands.

H_3: Local Preference Minorities and Demand for Popular versus Niche Brands. Online-offline substitution for niche brands, relative to popular brands, will be *more* sensitive to changes in the PM Index. That is, as the PM Index increases across markets, online-offline substitution will be more pronounced for niche brands.

Figure 4.4: Local Preference Minorities and Demand for Popular versus Niche Brands



Note: The *x*-axis varies across local markets with the same sized target population, but different total populations. The PM Index is: [1-(Target Population)/(Total Population)]. Sales substitution from offline retailers to online retailers is intensified when consumers in the preference minority do not favor "popular" brands. An increase in the PM Index increases online sales of niche brands by a higher percentage: $b_2 > b_1$.

The Long Tail (H_4). Figure 4.5 summarizes the Long Tail. On the *x*-axis brands (within a category) are ranked from "best selling" to "worst selling". The darkest area in Figure 4.5 (a) shows 20% of popular brands accounting for 80% of category sales at offline retailers; the middle and medium gray area shows the remaining 80% of "less popular" brands. Online retailers can expand their inventory to include the light gray area, i.e., all those brands that would not meet the shelf space or customer preference constraints faced by local offline retailers. Figure 4.5 (b) shows the sales and profit implications. While offline retailers gain 20% of their sales from the 80% of brands in the "less popular" group, they might make little or no profit after taking inventory holding costs and turn into account. Conversely, online retailers have "infinite" shelf space,

negligible inventory holding costs, and are not subject to a locally-defined trading area. In aggregate, individual niche brand sales over a large number of niche brands can contribute significant profits (25% in the example).

Category-level online sales increase across markets as the PM Index increases (H_1), however online sales response to the PM Index is stronger for niche brands (H_3). The difference in the sales sensitivity to the presence of preference minorities between popular and niche brands generates a new insight regarding the online sales distribution across brands and across local markets. Specifically, we show that the rank ordering of a brand in the Long Tail has implications for the *proportional mix* of its sales across geographical markets that vary according to the PM Index.

H_4: The Long Tail and Brand Sales by Market Type. Niche brands with a lower overall sales rank (i.e., those in the "tail" of the Long Tail) draw a greater *proportion* of their total online demand from high PM regions, than popular brands do.

Proof:

Define online sales, y, of a popular brand at a particular location as $y = a_1(1+ b_1 x)$ where x is a PM Index value for a local market. Online sales are a_1 when $x = 0$, and b_1 is the sales growth rate. Similarly, $y = a_2(1+ b_2 x)$ for the niche brand. Without loss of generality, divide the space of all local markets into two groups: one group with a relatively low PM Index (0 to x_1) and the other with a relatively high PM Index (x_1 to x_2).[26] Aggregate sales of the popular brand in the two markets is determined by integrating out the relevant areas under the sales curve $y = a_1(1+ b_1 x)$ as follows.

$$\text{(H4.1)} \qquad A(\text{Low PM, Popular}) = \int_0^{x_1} a_1(1+b_1 x)dx = a_1(1+.5b_1 x_1)x_1$$

[26] It is straightforward to show that the general case of n partitions of local markets along the PM Index leads to an identical result, but requires additional integrals of the relevant sales areas.

(H4.2) $$B(\text{High PM, Popular}) = \int_{x_1}^{x_2} a_1(1+b_1x)dx = a_1(x_2 - x_1)\{1+.5b_1(x_2 + x_1)\}$$

Similarly, aggregate sales of the niche brand, in the low PM and high PM markets are

(H4.3) $$C(\text{Low PM, Niche}) = \int_{0}^{x_1} a_2(1+b_2x)dx = a_2(1+.5b_2x_1)x_1$$

(H4.4) $$D(\text{High PM, Niche}) = \int_{x_1}^{x_2} a_2(1+b_2x)dx = a_2(x_2 - x_1)\{1+.5b_2(x_2 + x_1)\}$$

H_1 and H_3 state that the niche brand, by definition, has lower overall sales, but that the niche brand's online sales respond more strongly to the PM Index. That is, $a_1 > a_2$, $a_1(1+ b_1 x_2) > a_2(1+ b_2 x_2)$, and $b_1 < b_2$. This yields the following relationship

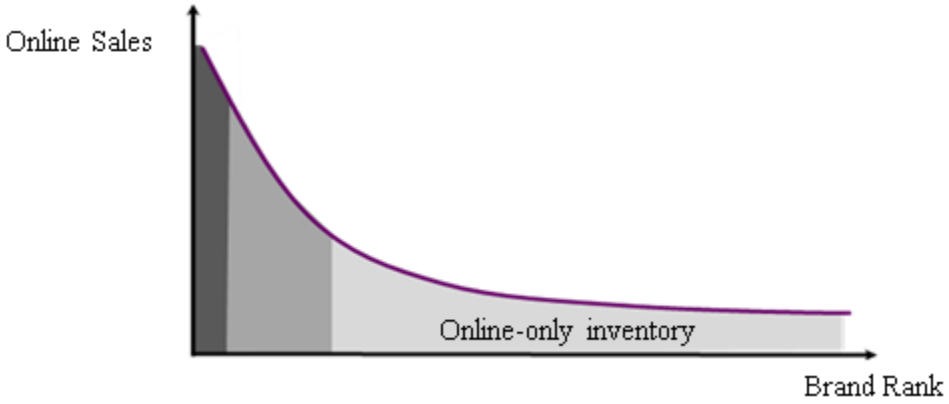
(H4.5) $$\frac{B(\text{High PM, Popular})}{A(\text{Low PM, Popular})} < \frac{D(\text{High PM, Niche})}{C(\text{Low PM, Niche})}$$ or

(H4.6) $$\frac{C(\text{Low PM, Niche})}{A(\text{Low PM, Popular})} < \frac{D(\text{High PM, Niche})}{B(\text{High PM, Popular})}$$

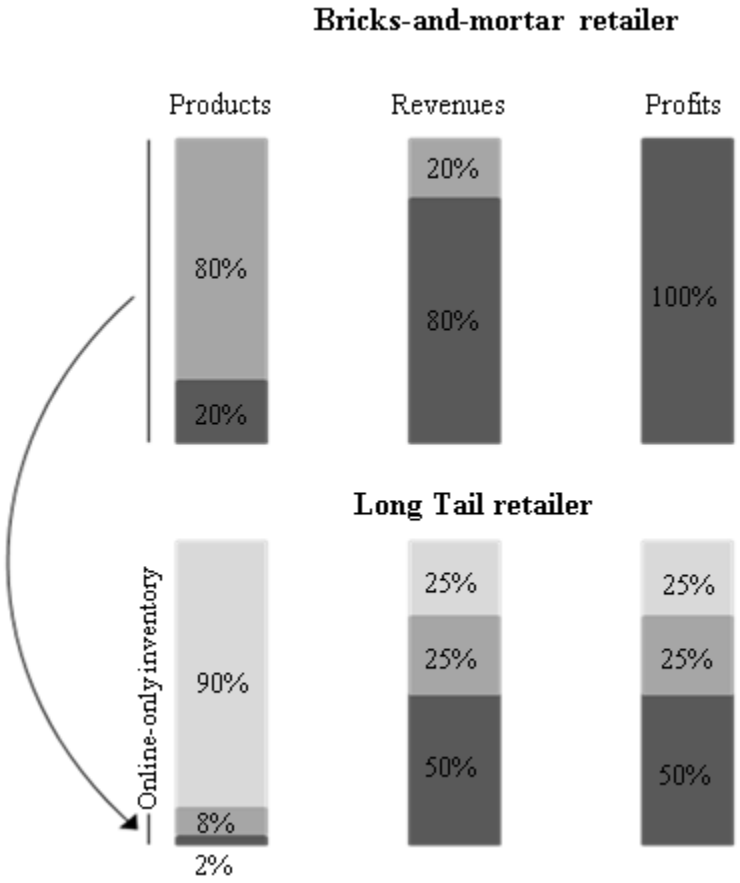
In words, the *sales ratio* of high PM-market sales to low PM-market sales for the niche brand, i.e., *D* / *C*, is greater than the same ratio for popular brands, i.e., *B* / *A*. Hence, looking *across markets*, niche brands get proportionally more online sales from high PM markets. Note also that equation (H4.6) implies a *within*-market comparison for online brand sales which we discuss subsequently. Since the *sales ratio* of the niche product to the popular product in a high PM-market, i.e., *D* / *B*, is greater than the same ratio in a low PM market, i.e., *C* / *A*, this implies a relatively more "even" distribution of online purchased assortments in high PM markets.

Figure 4.5: The Long Tail

(a) The Long Tail Sales Distribution Online



(b) Product Inventory, Revenues, and Profits for Offline versus Online Retailers



Note: Adapted from "The 80/20 Rule Revisited" at www.thelongtail.com.

4.3 Data and Measures

To test our hypotheses, we first choose a suitable category and define the unit of analysis for "local markets." Second, we profile and control for the local offline retail environment, and control for differences in geo-demographic characteristics across markets.

4.3.1 Product Category and Unit of Analysis

Product Category. Childcorp.com, the leading online retailer of diapers in the United States, provided: (1) zip-level cumulative numbers of buyers and orders from the firm's inception in January 2005 through March 2008, and (2) zip-level cumulative sales by brand between January 2007 and March 2008.[27] Three major national brands—Pampers, Huggies, Luvs—and one niche brand that is not available in all stores, Seventh Generation, are used in the analysis. (We identify exact locations of each store that carries Seventh Generation).[28] During the data period, Childcorp.com did no significant marketing activity that varied by location, i.e., the same assortments and prices were offered to all locations; thus, we can assess how preference minority status in a region affects online demand there, free of explicit marketing interventions.

[27] For reasons of confidentiality I refer to this leading Internet retailer by the *nom de plume*, "Childcorp.com".

[28] Seventh Generation limits distribution to bricks-and-mortar retailers that have an image of being "natural" or "organic" (e.g., Whole Foods). During the data period, Seventh Generation also decided to distribute household cleaning products through Target (but not Wal-Mart). We control for these store locations in the empirical analysis.

The diapers category is especially suitable for three reasons. First, name-brand diapers mentioned above are well known nationally and buyers can be relatively certain about product quality before placing orders (Lal and Sarvary 1999; Lynch and Ariely 2000). Second, diaper consumption is reasonably expected to be proportional to the total baby population, and to be constant at the per-capita level, i.e., the category is not "expandable" in the way that others (books, etc.) might be. Constant consumption across markets with the *same number of target customers* (see Figure 4.3) is a reasonable assumption. Third, a high shelf space-to-profit ratio for diapers limits offline product breadth and depth available in local markets (see Figure 4.2).

Unit of Analysis. Zip codes define local markets. This makes sense for two reasons. First, a zip code is a relatively self-contained group of buyers and sellers (especially for packaged goods such as diapers). The most accessible offline local retail format for diapers is the local supermarket and all zip codes that we examine have at least one supermarket.[29] Second, zip codes are widely used in other studies of related phenomena, such as restaurant and bookstore variety (see Waldfogel 2007 for a review). We focus our attention on zip codes that lie within Metropolitan Statistical Areas (MSAs).[30] Limiting the analysis to zip codes within MSAs is consistent with prior research (e.g., Forman, Ghose, and Goldfarb 2009; Sinai and Waldfogel 2004). It ensures that consumers have "reasonable" travel distances to offline alternatives and are not induced to shop online

[29] Residential zip codes have on average four supermarkets. There is roughly one discount store for every five zip codes, and one warehouse club for every fifteen zip codes. Supermarkets appear at approximately 2.5 miles intervals, discount stores every 8 miles, and warehouse clubs every 15 miles.

[30] The hypotheses are re-tested using *all* the residential zip codes in the United States, i.e., including those in non-MSAs, but after aggregation to the three-digit zip-code level. This aggregation leads to 877 geographical units, all of which have positive sales. Hence, we estimate a standard log-log linear model. Qualitatively identical results are obtained.

due to complete inaccessibility of local retail formats. Further, we avoid zip codes that are extremely "sparse" in terms of either focal population or total population.

We need the data to conform to the assumptions implied by Figures 4.3 and 4.4—the PM Index varies on the *x*-axis because only the denominator—total population—is changing. However, in the real data, *both* the number of households with babies (target group) *and* the total population (denominator of the PM Index) vary by location. Hence, we not only use the pooled data for analysis but also we define three separate bins of data (terciles) for analysis, based on the empirical distribution of the number of households with babies across all local markets. Each bin includes 2,979 residential zip codes, and zip codes within a bin have roughly equal size in terms of the target population but substantially different total populations i.e., within each tercile the PM Index varies in a manner consistent with Figures 4.3 and 4.4. We can also see how the results change (or do not change) when the size of the target population differs across bins. Summary statistics for the overall PM Index and the bin-specific values are in Table 4.1 (b). The use of bins also mitigates confounding effects by outside options other than disposable diapers, such as the availability of cloth diaper cleaning services in large target markets with more babies.

4.3.2 Local Environments

Offline Store Presence. Variables that capture the presence of local retailers are constructed from the 2007 US Census of Business and Industry using 8-digit NAICS (North American Industry Classification System) codes. While 6-digit NAICS codes are

often used in research, greater accuracy is achieved with our approach (6-digit codes can lead "candy stores" to be included with supermarkets, but 8-digit codes do not). We identify store locations of the major local retail competitors using NAICS 44511003 (retail grocery stores) and NAICS 45211204 (warehouse clubs). Wal-Mart and Target belong to the "discount department stores" classification and we obtain store locations directly. Since physical distance is taken as a parallel to transportation costs in spatial differentiation models (see e.g., Balasubramanian 1998; Bhatnagar and Ratchford 2004; Cheng and Nault 2007; Forman, Ghose, and Goldfarb 2009) distance from the focal zip code to the nearest store of each format is used in the model.

To further control for local depth and breadth of assortment of baby diapers we define a proxy variable using NAICS 44813001 and 44813002 (presence of stores selling baby accessories, e.g., Babies R Us). Since these stores serve broad neighborhoods, we set the indicator variable equal to 1 when a store is in the focal zip code or contiguous zip codes.

For Seventh Generation we have used the exhaustive national list of all local retailers in the United States that sell Seventh Generation products (see www.seventhgeneration.com) and confirmed availability of Seventh Generation diapers in each store. As above, we compute the distance from each zip code to the nearest store stocking Seventh Generation diapers. (Distances to the nearest retail formats that *do not* sell Seventh Generation diapers were re-computed in order to test H_3.) Thus, we have general measures of distance to stores that stock the popular brands and specific measures of distances to stores that stock the niche brand.

Relative Local Prices and Convenience. Childcorp.com offers *exactly the same* prices to every zip code in the United States however consumers in different zip codes clearly

face different *offline* prices. While we cannot gather data on offline prices for every zip code, we can nevertheless partially control for across zip code variation in the relative attractiveness of shopping online as follows. To help control for *relative* prices, we exploit the fact that Childcorp.com does not collect sales tax in locations where they have no retail nexus. In zip codes where offline stores collect sales tax, Childcorp.com has a greater *relative advantage* in price, compared to zip codes where they do not, which is helpful as our goal is to explain across zip code variation in the propensity to shop online. To compile the relevant zip level tax rates we started with publicly available information from the Department of Revenue in each state and undertook an exhaustive manual check of local tax rates.[31] We also account for online convenience (expected number of days to ship from Childcorp.com warehouses) using shipping time information from the UPS website (www.ups.com) which was confirmed by Childcorp.com management.

Geo-Demographic Controls. Geo-demographic covariates describe the local environment overall, and the characteristics of households who live there. They are derived from the 2000 US Census of People and Households.

4.3.3 The PM Index and Local Assortments

The PM Index. Empirically, there is no *absolute* determinant for "minority preferences". Drawing on published research (e.g., Chen et al. 1999), we assume that more local stores are available as population increases, but the physical size of stores need not be related to

[31] We made over 1,000 telephone calls to a random sample of major retailers across the United States including Wal-Mart, Walgreens, and CVS, and asked store employees to determine whether the focal products were tax exempt. We requested that they verify their answer by physically scanning individual items from the product category.

population size.[32] Hence, a greater *number* of supermarkets need not necessarily mean an increased amount of variety in a local market. An increase in total population limits the share of available retail space allocated for the target group (as the target group becomes proportionally smaller), and in turn, locally-available variety for the target group. Therefore, the PM Index is the following proportion defined at the local market level: [1 - (Households with Babies / Total Households)]. The measure is presumed to reflect the relative variety, from one market to another, of goods available locally for the baby group. (See Goolsbee and Klenow (2002), Sinai and Waldfogel (2004), and Forman, Ghose, and Wiesenfeld 2008 for similar approaches).

Relation to Local Assortments. A key assumption is that preference minorities, by definition, lack access to local variety and their presence is a good indication of the extent of local variety. Further arguments and data support this assumption. While we cannot get detailed assortment information for every zip code (for the same reason we cannot get offline prices), we can nevertheless use zip code variation in *online* sales to investigate our assumption. Childcorp.com offers identical assortments in every zip code. We can exploit the fact that while the *offered* assortments are identical across zip codes, the *purchased* assortments are not. The variation in purchased assortments could reflect heterogeneity in brand preferences, but it could also reflect heterogeneity in the amount

[32] Two assumptions are validated with our data. First, the total size in square footage of particular stores (e.g., Target, Whole Foods) tends to be driven more by chain level decisions, than population size per se. We examine store space using two variables describing (1) four ranges of square footage of local retailers and (2) eleven ranges of the number of employees working there. Among the 1,415 (224) local Target (Whole Foods) stores, for example, 99% (79%) belong to the highest range of being more than 40,000 square feet and 81% (69%) belong to the range of having 100-249 employees. Second, we examine the relation between the numbers of each retail format and population size using our using 8-digit NAICS codes at the MSA level. The number of households, for example, has the significant correlations of .97 with the number of supermarkets, .86 with the number of discount stores, and .96 with the number of warehouse stores.

of offline variety available. To measure the diversity of online variety purchased, we compute a category-level Herfindahl Index (HI) for each zip code—a smaller HI implies a greater level of purchased online variety. As expected, within a bin, this measure is negatively correlated with the PM Index. More variety is bought online—perhaps reflecting absence of variety offline—when the PM Index is higher. This negative correlation persists at the brand level, e.g., when we define the HI over 35 Pampers SKUs, and is also implied by H_4 (see equation H4.6).

4.3.4 Summary Statistics

Summary statistics for the model variables are presented in Table 4.1. *Within* each bin, the variation across markets in the size of the target group, i.e., the number of households with babies, is substantially smaller than the variation across markets in the total number of households (see Table 4.1 (b)); the coefficient of variation for the number of households with babies is about half of the coefficient of variation for total households. Interestingly, distances from each zip code to the nearest supermarket that *does not* sell Seventh Generation are roughly equal across three bins, but distances to the nearest store that *does* sell Seventh Generation decreases as the overall market size, reflected by the size of the total population, increases (see Table 4.1 (c)). Geo-demographic characteristics are largely similar across the three bins.

Table 4.1: Summary Statistics

(a) Dependent Variables

	All		Bin 1		Bin 2		Bin 3	
H_1:	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Number of Buyers	13.097	16.249	5.712	9.822	13.199	15.705	20.372	18.479
Number of Repeat Orders	26.940	58.256	12.767	44.815	29.193	61.990	38.845	63.119
H_2:	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Number of Pampers Packages	104.409	216.504	47.286	147.898	112.428	224.414	153.500	250.309
Number of Huggies Packages	28.653	63.337	13.667	47.390	29.836	61.729	42.418	74.615
Number of Luvs Packages	11.657	22.907	5.670	15.707	11.447	21.225	17.887	28.319
Number of SG Packages	22.936	64.449	10.519	34.748	26.710	74.618	31.590	73.787

(b) The PM Index

	All		Bin 1		Bin 2		Bin 3	
Preference Minority Variables	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Local Fraction of Households with Babies								
Number of Total Households	5620.390	3435.880	2355.030	1337.760	5223.820	2031.530	9278.650	3299.870
Number of Households with Babies	868.978	541.519	325.250	86.024	753.931	162.782	1527.090	321.786
PM Index = [1 - Fraction of Households with Babies]	.837	.054	.844	.054	.840	.058	.828	.049

(C) Independent Control Variables

	All		Bin 1		Bin 2		Bin 3	
Control Variable	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Online Price Advantage								
Percentage Local Sales Tax Rate	5.497	3.001	5.215	3.070	5.351	3.075	5.923	2.804
Offline Store Presence								
Local Presence of Stores Selling Baby Accessories	.374	.484	.241	.428	.385	.487	.495	.500
Distance to the Nearest Supermarket	2.347	2.304	2.715	2.708	2.148	1.929	2.178	2.160
Distance to the Nearest Discount Store	5.905	5.228	8.661	5.910	5.286	4.624	3.771	3.632
Distance to the Nearest Warehouse club	11.479	12.015	15.894	12.485	10.483	10.753	8.064	11.380
Distance to the Nearest Store Selling Seventh Generation	5.289	6.673	8.307	7.429	4.628	6.206	2.933	4.984
Distance to the Nearest Supermarket with No Seventh Generation	2.423	2.347	2.783	2.716	2.264	2.057	2.222	2.173
Relative Convenience								
One-Day Shipping (1=Yes, 0 = No)	.190	.392	.209	.407	.212	.409	.149	.356
Two-Day Shipping (1=Yes, 0 = No)	.199	.399	.228	.420	.199	.399	.170	.376
Three-Day Shipping (1=Yes, 0 = No)	.327	.469	.321	.467	.324	.468	.337	.473
2nd Warehouse Led to One-Day Shipping (1=Yes, 0 = No)	.030	.170	.019	.136	.028	.165	.043	.202
2nd Warehouse Led to Two-Day Shipping (1=Yes, 0 = No)	.104	.305	.082	.274	.093	.290	.138	.345
Geo-Demographic Controls								
Percentage of Population Aged 20 to 39 Years	.280	.063	.266	.065	.281	.068	.292	.051
Percentage with Bachelors and/or Graduate Degree	.529	.167	.491	.164	.544	.171	.552	.160
Percentage of Female Population in Labor Force	.556	.083	.552	.089	.555	.082	.563	.076
Percentage of Households Below the Poverty Line	.100	.078	.097	.073	.101	.083	.104	.078
Percentage of Blacks	.103	.176	.065	.130	.104	.182	.138	.199
Percentage of Apartment Buildings with 50 Units or More	.034	.068	.023	.078	.038	.069	.043	.052
Percentage of Homes Valued at $250,000 or More	.148	.210	.140	.203	.163	.227	.140	.198
Annual Population Growth Rate from 2000 to 2004	.014	.020	.014	.020	.014	.021	.014	.019
Population Density (thousands in square miles)	1.880	3.719	.955	3.002	1.923	3.402	2.760	4.391

Note: Discount stores and warehouse clubs did not stock Seventh Generation diapers at the time of data collection. All the retail formats selling Seventh Generation diapers are considered in the computations for the distance to nearest store selling Seventh Generation (used to test H_3). Similarly, the remaining supermarkets without Seventh Generation diapers are used to compute the "Distance to the Nearest Supermarket" in the test of H_3.

4.4 Empirical Analysis

4.4.1 Category Sales (H_1) and Price Sensitivity (H_2)

We test H_1 with two dependent variables: (1) the number of buyers per zip location, and (2) the number of repeat orders per zip location. The number of buyers (repeat orders) in zip code *z* in MSA *m* is Poisson distributed with rate parameter $\lambda_{z(m)}$. The Poisson is appropriate when the occurrence rate is low (Agresti 2002) which agrees with the properties of the data. Specifically, the number of buyers (repeat orders) in a zip code is very small compared with the number of households with babies, and thus, we include the number of households with babies as an offset variable with its parameter constrained to one (Agresti 2002; Knorr-Held and Besag 1998; Michener and Tighe 1992).

The Poisson rate, $\lambda_{z(m)}$, is modeled as a function of $PM_{z(m)}$ (the PM Index), $TAX_{z(m)}$ (local sales tax rates), the interaction of these two, and local market characteristics, $\vec{X}_{z(m)}$. MSA-level random effects help control for unobserved heterogeneity in the baseline rates. The error tem $\varepsilon_{z(m)}$ allows for over-dispersion and is *IID* Gamma distributed with shape and scale parameter equal to θ (Cameron and Trivedi 1986; Greene 2002).

(3.1) $$Y_{z(m)} \sim \text{Poisson}(\lambda_{z(m)})$$

(3.2) $$\log(\lambda_{z(m)}) = \beta_1 \cdot PM_{z(m)} + \beta_2 \cdot TAX_{z(m)} + \beta_3 \cdot PM_{z(m)} \cdot TAX_{z(m)} + \log(n_{z(m)}) + \alpha_0 + \alpha_m + \vec{\gamma}' \vec{X}_{z(m)} + \varepsilon_{z(m)}$$

$$\text{where } \alpha_m \sim N(0, \tau^2) \text{ and } \exp(\varepsilon_{z(m)}) \sim Gamma(\theta, \theta)$$

Category Sales. An increase in the PM Index means that the focal group's preferences are becoming more minor, so we expect $\beta > 0$ (H_1). Table 4.2 (a) shows the expected positive effect on the number of buyers and the number of repeat orders.[33] To understand the implied quantitative effects, suppose that two local markets are of equal size in terms of baby population, but differ in terms of *total* population, and therefore on the PM Index. Suppose one compares a "low PM market" (the 10th percentile market; PM Index = 0.79) and a "high PM market" (the 90th percentile; PM Index = 0.89). At the mean of the other covariates, this implies 6.67 (9.88) buyers and 10.28 (16.60) orders from the low PM (high PM) market. Combining trial and repeat orders, Childcorp.com sales are almost 56% higher in the high PM market compared to the low PM market *even though in both markets we hold the number of target consumers constant*.

Moving from the low to the high PM market *does not* change the size of the target population as both markets have an identical number of target customers. Instead, the increase in the total population makes the customers in the high PM market more isolated. As they are a smaller fraction of the total market, local retailers allocate less space to products (or variety) they want, which in turn drives them online.

[33] We estimate equations (3.1)-(3.2) with two additional variables for the number of households without babies and the reciprocal of the number of total households. Qualitatively identical results are obtained. The estimation results are in Table 4.4 (a) in the Appendix.

Table 4.2: Parameter Estimates at Category Level

(a) Parameter Estimates from the Pooled Data

	Buyers		**Repeat Orders**	
	Estimate	SE	Estimate	SE
Intercept	-10.253*	.277	-11.560*	.541
Preference Minority				
$PM_{z(m)}$ = [1 - Fraction of Households with Babies]	4.500*	.305	5.710*	.597
Online Price Advantage				
Percentage Local Sales Tax Rate	.114*	.038	.181*	.075
Preference Minority and Online Price Advantage				
$PM_{z(m)}$ × Percentage Local Sales Tax Rate	-.123*	.045	-.190*	.088
Offline Store Presence				
Local Presence of Stores Selling Baby Accessories	-.035*	.013	-.044*	.026
Distance to the Nearest Supermarket	-.011*	.003	-.021*	.006
Distance to the Nearest Discount Store	.018*	.002	.019*	.003
Distance to the Nearest Warehouse Club	.005*	.001	.007*	.001
Relative Convenience				
One-Day Shipping	.738*	.058	1.206*	.102
Two-Day Shipping	.362*	.045	.616*	.079
Three-Day Shipping	.209	.040	.341	.070
2nd Warehouse Led to One-Day Shipping	.194*	.073	.413*	.129
2nd Warehouse Led to Two-Day Shipping	.127+	.052	.375	.092
Geo-Demographic Controls				
Percentage of Population Aged 20 to 39 Years	2.471*	.136	2.826*	.266
Percentage with Bachelors and/or Graduate Degree	1.725*	.068	1.851*	.128
Percentage of Female Population in Labor Force	-.185*	.127	.213	.236
Percentage of Households Below the Poverty Line	-2.534	.178	-3.394*	.318
Percentage of Blacks	-.284*	.054	.011	.101
Percentage of Apartment Buildings	.726*	.103	.767*	.206
Percentage of Homes Valued at $250,000 or More	.848*	.047	1.738*	.095
Annual Population Growth Rate from 2000 to 2004	10.233*	.344	10.488*	.684
Population Density (thousands in square miles)	.018*	.002	.024*	.004
Variance				
θ	6.357*	.164	.819*	.014
τ^2	.038*	.005	.107*	.016
-2LL	51,996		65,649	

Note: * indicates significance at $p < .05$ and + indicates significance at $p < .10$.

(b) Parameter Estimates from Three Bins

	Bin 1		**Bin 2**		**Bin 3**	
	Buyers	**Repeat Orders**	**Buyers**	**Repeat Orders**	**Buyers**	**Repeat Orders**
$PM_{z(m)}$	4.635*	7.355*	4.274*	5.179*	4.349*	4.700
Local Sales Tax Rate	.178*	.326*	0.112+	.072*	0.092	.093
$PM_{z(m)}$ × Sales Tax	-.203*	-.354+	-0.125+	-.061*	-0.100	-.083
-2LL	13,894	17,361	17,775	22,738	20,240	25,147

Note: The estimates for the remaining variables are largely consistent with those from the model with the pooled data and are not shown for ease and clarity of exposition. * indicates significance at $p < .05$ and + indicates significance at $p < .10$

(c) Expected Demand (Total Number of Orders) by PM Indices

	Low *PM* Index	Medium *PM* Index	High *PM* Index	Demand Increase
All	16.945	21.171	26.471	56%
Bin 1	7.100	9.047	11.562	63%
Bin 2	21.046	26.392	33.131	57%
Bin 3	34.834	42.938	52.947	52%

(d) Expected Demand (Total Number of Orders) by PM Indices and Local Sales Taxes

	Low PM			High PM		
	Low Tax	High Tax	Demand Increase	Low Tax	High Tax	Demand Increase
All	14.686	18.214	24%	25.120	27.180	8%
Bin 1	6.174	8.262	34%	11.678	12.065	6%
Bin 2	17.551	20.730	18%	28.789	31.734	10%
Bin 3	29.052	34.892	19%	46.355	51.655	11%

To control for the effect of the absolute size of the target group, we estimate separate models for each of the three bins and report the results in Table 4.2 (b). The estimates for the PM Index are significant, economically meaningful, and provide support for H_1. We are able to notice, however, that as the target population increases across bins, the percentage changes in online demand from low to high PM markets *within* a bin become smaller. The average percentage sales increases when moving from a low to high PM market are 63%, 57%, and 52%, in the first, second, and third bins, respectively (see Table 4.2 (c)). This is consistent with the idea that more retail formats will exist to address the needs of target population in larger markets (see Christaller 1933). In summary, H_1 is strongly supported.

Price Sensitivity. As the local sales tax rate increases, the relative online price advantage increases, so we expect $\beta_2 > 0$. Thus, everything else held constant, we expect demand at Childcorp.com to be higher when it has a greater relative advantage in prices. From H_1, we know that an increase in the preference minority status leads to more online demand ($\beta_1 > 0$). H_2, however, concerns the interaction between the preference minority status and the relative price advantage of shopping online. Since preference minorities are motivated by lack of assortment, we expect consumers in high PM markets, relative to those in low PM markets, to be *less* responsive to an increase in the relative price advantage at Childcorp.com. Thus, we expect $\beta_3 < 0$. Table 4.2 (a) shows the parameter estimates with the expected signs from the model for the number of buyers and repeat orders.

We obtain the expected demand by varying both the PM index and the sales tax rate. (Sales tax rates range from zero to 8.25%.) The results are given in Table 4.2 (d). When

the price advantage at Childcorp.com increases, demand in low PM markets increases by 24%. An increase in the price advantage at Childcorp.com also helps in high PM markets, but to a far lesser extent—demand only goes up by 8%. Consumers in low PM markets (who have good offline options) can be motivated by price advantages to buy online. Conversely, those in high PM markets are more likely to buy online independently of relative online price advantages. Thus, we find strong support for H_2 as well.

Control Variables. Control variable estimates are largely plausible. The presence of "baby-oriented" stores has a negative effect on demand at Childcorp.com. Increased travel distances to discount stores and warehouse clubs has a positive effect (the less accessible these stores are the more likely shoppers buy online; see also Forman, Ghose, and Goldfarb 2009). Thus, it is important to see that our findings on the effect of "preference isolation" have been obtained in a model that also controls for "geographical isolation" (transportation costs) studied in other articles. Conversely, there is a negative effect for the distance to supermarket. Since most shoppers visit supermarkets anyway for perishable products they may buy larger baskets at less convenient supermarkets to amortize fixed travel costs (Tang, Bell, and Ho 2001), and therefore be *less* likely to buy individual categories online (e.g., Bell and Song 2007). In general, Childcorp.com performs better in zip codes that have higher percentages of the local population between 20 and 39 years old, college-educated individuals, working females, urban housing units, and homes valued at $250,000 or more, but have lower percentages of black households and households below the poverty line. Population growth and density also help and demand is higher in zip codes with expeditious delivery.

4.4.2 Popular versus Niche Products (H_3) and the Long Tail (H_4)

Popular Versus Niche Products. Customers in preference minority markets have difficulty finding diapers of all types and this will be exacerbated for those who seek the niche brand, Seventh Generation. (In our data preferences for the national brands and the niche brand are bi-modal and there is little switching between these brands, conditional upon shopping at Childcorp.com. Thus, we exclude the possibility that online buyers endogenously form the preference for variety.) To test H_3, we fit equations (4.1) and (4.2) at the brand level. The dependent variable is the number of brand *j* diaper "standard packages" purchased in each zip code *z* and MSA *m*. Each SKU has a different number of actual diapers; hence we standardize by frequently purchased package sizes. For example, the SKU "Pampers Swaddlers Jumbo Pack Size 2 – 80 counts" converts to 2 packages of "Pampers Swaddlers Super Mega Pack Size 2 – 40 counts." We convert all diapers purchased to "standard units" in the same way that scanner panel datasets with multiple sizes within a category are treated (e.g., Bucklin, Gupta, and Siddarth 1998).

Table 4.3: Parameter Estimates at Brand Level

(a) Parameter Estimates from the Pooled Data

	Popular Brands			Niche Brand
	Pampers	**Huggies**	**Luvs**	**Seventh Generation**
Intercept	-9.711*	-9.470*	-9.933*	-18.827*
Preference Minority				
$PM_{z(m)}$ = [1 - Fraction of Households with Babies]	5.319*	4.170*	4.350*	11.396*
Online Price Advantage				
Percentage Local Sales Tax Rate	.209*	.170	.216	.221
Preference Minority and Online Price Advantage				
$PM_{z(m)}$ × Percentage Local Sales Tax Rate	-.246*	-.180*	-.238	-.281
Offline Store Presence				
Local Presence of Stores Selling Baby Accessories	-.036	-.003	-.071	-.069
Distance to the Nearest SG Store	-.002	.000	-.006	.001
Distance to the Nearest Supermarket	-.008	-.023*	-.018	-.016
Distance to the Nearest Discount Store	.007+	.021*	.010	.038*
Distance to the Nearest Warehouse Club	.009*	.009*	.009*	.004*
Relative Convenience				
One-Day Shipping	1.012*	1.006*	.735*	1.469*
Two-Day Shipping	.636*	.488*	.494*	.512*
Three-Day Shipping	.294*	.298*	.411*	.424*
2nd Warehouse Led to One-Day Shipping	.166	.354*	.178	1.117*
2nd Warehouse Led to Two-Day Shipping	.194*	.359*	.135	1.049*
Geo-Demographic Controls				
Percentage of Population Aged 20 to 39 Years	2.495*	1.361*	1.551*	5.102*
Percentage with Bachelors and/or Graduate Degree	1.915*	1.726*	.671*	3.506*
Percentage of Female Population in Labor Force	.288	.111	1.409*	1.063+
Percentage of Households Below the Poverty Line	-3.447*	-3.324*	-2.831*	-2.523*
Percentage of Blacks	-.062	-.088	-.554*	.530*
Percentage of Apartment Buildings	1.127*	1.149*	.460*	.277
Percentage of Homes Valued at $250,000 or More	1.917*	1.494*	.017	1.311*
Annual Population Growth Rate from 2000 to 2004	11.575*	9.548*	6.419	12.642*
Population Density (thousands in square miles)	.020*	.026*	.030*	.021*
Variance				
θ	.717*	.428*	.211*	.186*
τ^2	.102*	.114*	.088*	.453*
-2LL	88,177	65,408	49,514	48,376

Note: * indicates significance at $p < .05$ and + indicates significance at $p < .10$. We estimate equations (3.1)-(3.2) with the parameter for the PM index held constant across the four brands. The full model with the free parameters is significantly better than the restricted model. (The full model has -2LL of 251,475 and the restricted model has -2LL of 251,916.) Moreover, pairwise comparisons of the parameters for the PM index show that the estimate for Seventh Generation is significantly larger than those for the other three brands ($p < .05$), which are not significantly different from each other.

(b) Parameter Estimates from Three Bins

		Popular Brands			Niche Brand
Bin 1	**Variable**	**Pampers**	**Huggies**	**Luvs**	**Seventh Generation**
	$PM_{z(m)}$	6.259*	4.634*	8.132*	15.767*
	Local Sales Tax Rate	.411*	.110	.554	.611
	$PM_{z(m)}$ × Sales Tax	-.490*	-.101	-.627	-.708
		Popular Brands			**Niche Brand**
Bin 2	**Variable**	**Pampers**	**Huggies**	**Luvs**	**Seventh Generation**
	$PM_{z(m)}$	4.914*	3.664*	2.468	9.559*
	Local Sales Tax Rate	.079	.097	-.201	-.052
	$PM_{z(m)}$ × Sales Tax	-.082	-.107	.240	.045
		Popular Brands			**Niche Brand**
Bin 3	**Variable**	**Pampers**	**Huggies**	**Luvs**	**Seventh Generation**
	$PM_{z(m)}$	3.681*	4.323*	.814	11.333*
	Local Sales Tax Rate	.005	.186	.015	.266
	$PM_{z(m)}$ × Sales Tax	.012	-.183	.020	-.356

Note: The parameter estimates for the control variables are largely consistent with those from the model with the pooled data and are not shown for ease of exposition. * indicates significance at $p < .05$ and $^{+}$ indicates significance at $p < .10$.

Table 4.3 (a) provides the estimates.[34] As the PM Index increases across markets, online demand responds more strongly for Seventh Generation than it does for the national brands, Pampers, Huggies, and Luvs. At the mean of the other covariates, this implies the following changes from low to high PM markets. Pampers' sales increase by 50% (from 40.11 to 60.26), Huggies by 39% (13.07 to 18.11), and Luvs by 37% (7.68 to 10.49).[35] The increase for Seventh Generation, albeit from a smaller sales base, is dramatically greater at 175% (from 4.71 to 12.96).

To control for the effect of the absolute size of the target group, we estimate three separate models in three bins and present the results in Table 4.3 (b). Seventh Generation sales increase by 245%, 173%, and 158%, in the first, second and third bins, respectively while national brand increases remain between 30% and 60%. As discussed earlier, the diminishing sales increase across three bins supports the idea that more retail formats will exist to address the needs of the target population in larger markets. Thus, H_3 is supported.

The Long Tail. Earlier we showed that H_3 leads to H_4—niche brands with a lower overall sales rank relative to popular brands, draw a greater *proportion* of their total online demand from high PM markets. We compute the empirical analog for H_4. In addition to low and high PM markets defined earlier, we define "medium PM" markets at the mean of the PM Index. Holding everything else equal, we compute the expected sales for each of the four brands in all three types of markets. Figure 4.6 shows the results.

Figure 4.6 (a) is a typical Long Tail plot (*x*-axis = brands, *y*-axis = sales) and Figure 4.6 (b) makes the sales aggregation over markets with different degrees of preference

[34] We estimate equations (3.1)-(3.2) with two additional variables for the number of households without babies and the reciprocal of the number of total households. Qualitatively identical results are obtained and the estimation results are in Table 4.4 (b) in the Appendix.

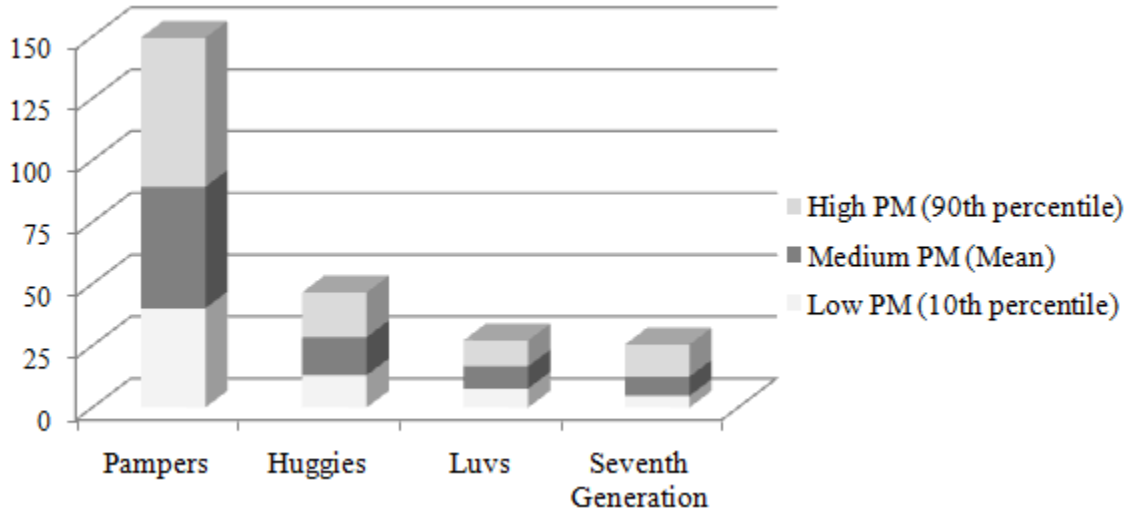
[35] For each brand we examine the *percentage* increase from different levels of base sales. This is because popular brands (by definition) will always have absolute sales that are higher than niche brand sales.

minority status explicit. All national brands draw proportionally more sales from markets that have a higher PM Index (the ratio of high PM market sales to low PM market sales is always strictly greater than one because the slope of the online sales line is positive—see equation H4.5). 40% of Pampers' total sales come from high PM markets and 27% from low PM markets. In fact, regardless of the differences in total sales of Pampers, Huggies, and Luvs, the sales share distribution across the three different types of local markets is remarkably similar. The sales distribution for the niche brand—Seventh Generation—shows a stark contrast. The ratio of sales from the high to low PM market is 51:18, or about three to one. Thus, H_4 is also supported.

Figure 4.6 (c) shows the online share distribution *across brands* within a given type of market. In the low PM market, Seventh Generation has a 7% share, and this nearly doubles to 13% in the high PM market. Earlier, we noted that in the raw data the PM Index is negatively correlated across zips with the Herfindahl Index, i.e., more online variety is purchased in high PM markets. The model estimates, which control for several zip-level differences, imply the same finding and line up with the theory implied by equation (H4.6) which shows that high PM *markets* have more heterogeneous online category demand than low PM markets do. This is noteworthy because it says that across-market diversity in online brand choices is potentially explained by the constraints that consumers face in their local markets, rather than preference heterogeneity alone.

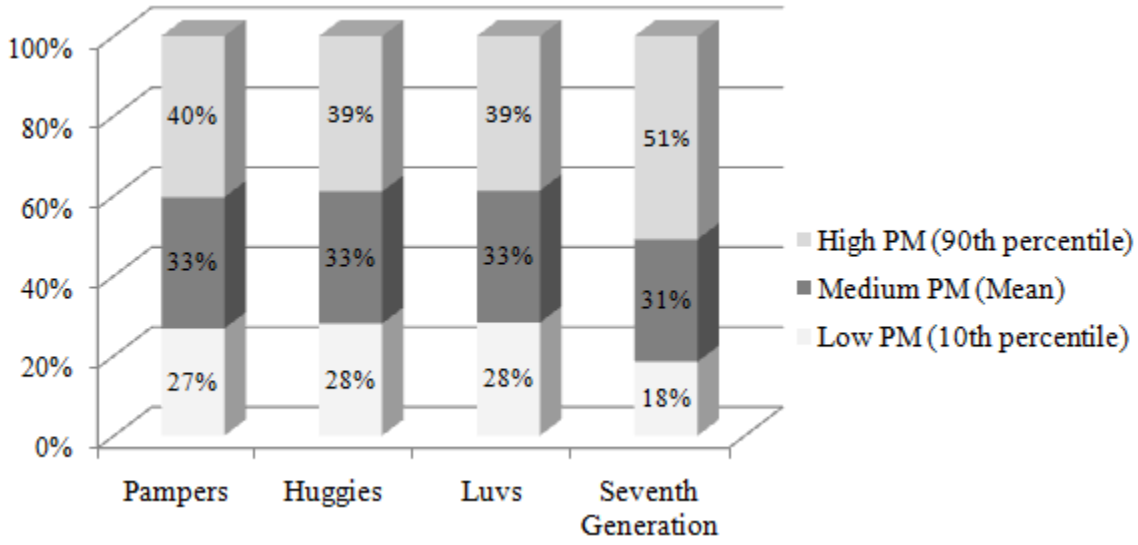
Figure 4.6: The Contribution of Local Markets to Total Brand Sales

(a) Long Tail Sales Distribution



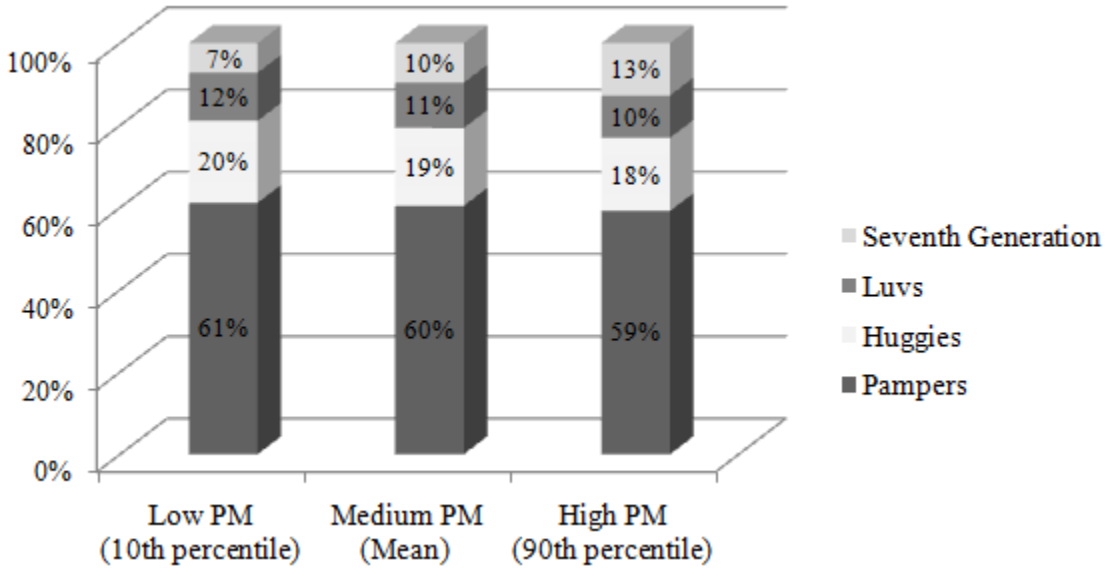
Note: We fix the PM Index at the 10th percentile, the mean, and at the 90th percentile. Holding everything else equal, we compute the expected sales for each brand in all three types of markets.

(b) Fraction of Total Brand Sales Drawn from Each Type of Local Market



Note: In accordance with H_3, all national brands draw proportionally more sales from markets with higher PM Indices, but the niche brand (Seventh Generation) draws *proportionally greater* demand from markets where the PM Index is high. The relative proportion of sales from the high to low PM market is 51:18, or about three to one.

(c) Brand Market Shares within Different Types of Local Market



Note: In accordance with equation (H4.6) the market share of the niche brand (Seventh Generation) increases from 7% in the low PM market to 13% in the high PM market. Thus, regions with a higher PM Index show more heterogeneous demand across brands, i.e., more online variety *purchased*, even through all three types of markets have the same number of target customers, and are offered the same variety.

4.5 Conclusion

Drawing on theory and empirical findings in economics and economic geography, we introduced the concept of local preference minorities as a driver of online-offline sales substitution in local markets. We tested four hypotheses and found each to be supported by the data. Category-level sales substitution to online retailers is greater in markets that have a higher PM Index (H_1) and high PM markets are less price sensitive (H_2). Online-offline substitution for niche brands, relative to popular brands, is *more* sensitive to changes in the PM Index (H_3). Finally, niche brands draw a greater *proportion* of their total sales from high PM markets (H_4).

4.5.1 Implications for Retailing Practice and Theory

The findings imply an interesting trade off. An Internet retailer could focus on local markets where the *absolute* number of potential customers is high. However, customers in these markets are likely to be well served by offline retailers. Marginal effects from the model suggest that overall category sales in high PM markets, relative to low PM markets, can be up to 56% higher, even though a naïve examination based on the absolute number of target customers would suggest both have about equal "potential." Moreover, Internet retailers selling niche brands face less competition from local retailers, and this is especially true when niche sales are made in high PM markets. Our findings imply a further reason that selling niche brands in high PM markets is an especially attractive proposition for an Internet retailer. Relative demand is not only strong in these locations, but also relatively insensitive to the difference between online and offline prices.

In the offline world, retailers attempt to improve the economics of stocking slower-moving SKUs (so that they can increase the variety they offer) by using distributors who stock in less-than-case pack-out quantities. Despite this, there may still be several categories (e.g., bulky or low value) or brands (e.g., niche) that need minimum facing and are difficult for offline retailers to justify. Online retailers can exploit this assortment gap, particularly in high PM markets. We demonstrate this for diapers, but other categories with similar properties should also benefit in the same way.

Finally, larger markets have more retail formats and more product variety, and make niche brands more accessible (Christaller 1933). This begs the question: Which local markets should the firm target when markets have *different* numbers of potential

consumers? Compare, for example, a high PM market with a small target population and a low PM market with a larger target population. Focusing on a low PM market in Bin 2 instead of a high PM market in Bin 1 roughly doubles the size of target population. Sales for national brands increase by 35-60%, but sales of Seventh Generation are flat. Hence, sales do not increase proportionally with the increase in the size of the target group, nor do they increase equally for national and niche brands.

In summary, our study provides further evidence that consumer benefits from the online marketplace are contextual and relative to offline alternatives (Anderson et al. 2009; Choi, Hui, and Bell 2009; Forman, Ghose, and Goldfarb 2009). In particular, we show how a specific form of consumer isolation—preference isolation—explains across-location variation in online demand. Internet retailers are ubiquitous but the net benefit to individual consumers still depends largely on where they live, and, *who* lives next to them. "Old economy" data can be used to find and target local markets with high potential. In other words, understanding local geography still matters a good deal for "borderless" retailing.

4.5.2 Limitations and Future Research

We use the diaper category and additional empirical support for the hypotheses could be pursued using other product categories. Second, we rely on zip-level sales data and do not assess the "preference minority status" of specific individuals (see Sinai and Waldfogel 2004). Rather, we characterize the preference minority status of a market segment within its local market. Third, we develop the preference minority arguments from a cross-

sectional perspective. Given appropriate data one could examine the evolution of preference minority status over time, and perhaps explore the dynamic nature of substitution between online and offline markets (e.g., Overby and Jap 2008).

There are at least two promising avenues for future research. First, Central Place Theory (CPT) (Christaller 1933; see also Shonkwiler and Harris 1996), a cornerstone of retailing thought that explains distances between cities of different sizes and the emergence of retail stores, could be reconfigured to address Internet retailing. According to CPT, larger towns have both more stores and more *variety* of stores. One might be able to develop a complementary theory for the distribution of *customers* acquired by Internet retailers, in contrast to the distribution of *stores* (given customers) implied by CPT. Second, the possibility of endogenous preference for variety might be examined. Preference minorities might go online for the reasons we suggest (H_1), but having got there, expand their brand preferences within a category. We intend to pursue these issues in future research.

4.6 Appendix

Table 4.4: Parameter Estimates from the Pooled Data

(a) Parameter Estimates at Category Level

	Buyers		**Repeat Orders**	
	Estimate	SE	Estimate	SE
Intercept	-10.527*	.281	-11.805*	.549
Preference Minority				
$PM_{z(m)}$ = [1 - Fraction of Households with Babies]	4.863*	.308	6.104*	.604
Online Price Advantage				
Percentage Local Sales Tax Rate	.110*	.038	.175*	.075
Preference Minority and Online Price Advantage				
$PM_{z(m)}$ × Percentage Local Sales Tax Rate	-.118*	.045	-.180*	.089
Offline Store Presence				
Local Presence of Stores Selling Baby Accessories	-.024+	.013	-.025	.026
Distance to the Nearest Supermarket	-.007+	.004	-.017*	.006
Distance to the Nearest Discount Store	.014*	.002	.017*	.003
Distance to the Nearest Warehouse Club	.004*	.001	.007*	.001
Relative Convenience				
One-Day Shipping	.729*	.058	1.204*	.101
Two-Day Shipping	.366*	.045	.590*	.077
Three-Day Shipping	.224*	.040	.347*	.069
2nd Warehouse Led to One-Day Shipping	.222*	.073	.439*	.128
2nd Warehouse Led to Two-Day Shipping	.142*	.052	.390*	.091
Geo-Demographic Controls				
Number of Households without Babies[1]	-.011*	.003	-.020*	.007
Reciprocal of Number of Total Households[1]	.133*	.052	-.006	.091
Percentage of Population Aged 20 to 39 Years	2.525*	.137	2.990*	.271
Percentage with Bachelors and/or Graduate Degree	1.784*	.068	1.946*	.130
Percentage of Female Population in Labor Force	-.243+	.128	.079	.238
Percentage of Households Below the Poverty Line	-2.551*	.178	-3.443*	.320
Percentage of Blacks	-.253*	.054	.013	.101
Percentage of Apartment Buildings	.691*	.104	.733*	.209
Percentage of Homes Valued at $250,000 or More	.800*	.048	1.645*	.095
Annual Population Growth Rate from 2000 to 2004	9.903*	.350	10.040*	.691
Population Density (thousands in square miles)	.019*	.002	.027*	.004
Variance				
θ	6.324*	.162	1.091*	.020
τ^2	.037*	.005	.102*	.015
-2LL	51,940		65,626	

Note: * indicates significance at $p < .05$ and + indicates significance at $p < .10$. [1] indicates that the variables are added to equations (3.1)-(3.2) along with the variables in Table 4.2 (a).

(b) Parameter Estimates at Brand Level

	Popular Brands			Niche Brand
	Pampers	**Huggies**	**Luvs**	**Seventh Generation**
Intercept	-10.058*	-9.773*	-9.968*	-18.895*
Preference Minority				
$PM_{z(m)}$ = [1 - Fraction of Households with Babies]	5.719*	4.577*	4.731*	11.598*
Online Price Advantage				
Percentage Local Sales Tax Rate	.197*	.166	.184	.206
Preference Minority and Online Price Advantage				
$PM_{z(m)}$ × Percentage Local Sales Tax Rate	-.225*	-.171	-.202	-.248
Offline Store Presence				
Local Presence of Stores Selling Baby Accessories	-.020	-.011	-.045	-.115+
Distance to the Nearest SG Store	-.004	-.002	-.006	-.002
Distance to the Nearest Supermarket	-.005	-.018+	-.013	-.011
Distance to the Nearest Discount Store	.006	.020*	.006	.041*
Distance to the Nearest Warehouse Club	.009*	.008*	.007*	-.004
Relative Convenience				
One-Day Shipping	.992*	.986*	.727*	1.504*
Two-Day Shipping	.623*	.451*	.498*	.510*
Three-Day Shipping	.295*	.310*	.432*	.411*
2nd Warehouse Led to One-Day Shipping	.212	.373*	.211	1.122*
2nd Warehouse Led to Two-Day Shipping	.192*	.377*	.169	1.089*
Geo-Demographic Controls				
Number of Households without Babies[1]	-.011	-.012	-.049*	-.021
Reciprocal of Number of Total Households[1]	.083	.058	-.262	-.308
Percentage of Population Aged 20 to 39 Years	2.643*	1.514*	1.786*	5.283*
Percentage with Bachelors and/or Graduate Degree	2.011*	1.803*	.855*	3.553*
Percentage of Female Population in Labor Force	.164	-.033	1.233*	.925+
Percentage of Households Below the Poverty Line	-3.529*	-3.422*	-2.830*	-2.621*
Percentage of Blacks	-.068	-.065	-.539*	.553*
Percentage of Apartment Buildings	1.087*	1.110*	.441	.283
Percentage of Homes Valued at $250,000 or More	1.826*	1.437*	-.120	1.244*
Annual Population Growth Rate from 2000 to 2004)	11.178*	9.172*	5.745*	12.254*
Population Density (thousands in square miles)	.021*	.027*	.033*	.021*
Variance				
θ	.725*	.427*	.211*	.186*
τ^2	.101*	.117*	.089*	.424*
-2LL	88,163	65,402	49,502	48,366

Note: * indicates significance at $p < .05$ and + indicates significance at $p < .10$. [1] indicates that the variables are added to equations (3.1)-(3.2) along with the variables in Table 4.2 (a).

Chapter 5

Conclusion

Through the three related essays in my dissertation, I intend to contribute to the literature both substantively and methodologically. In the first and second essays, I model Internet retail demand evolution from website inception, and compare and contrast different social influence effects. To provide a complete picture of these effects, I introduce two spatio-temporal models. In the third essay, I introduce the concept of "local preference minorities" and relate it to spatial variation in online demand and the Long Tail. The three essays together demonstrate that while Internet retailers are ubiquitous, the net benefit of using them still depends largely on *where* customers live and *who* lives next to them. Thus, understanding local geography still matters a good deal for "borderless" retailing.

New substantive and managerial insights generated in my dissertation along with limitations open several directions for future research. I briefly outline some interesting problems that warrant further study and have the potential to contribute to the understanding of the Internet retailing industry. First, the seeding experiments in the first and second essays are hypothetical as the proposed spatio-temporal models are descriptive models. One might want to develop a forecasting model and assess seeding performance, or evaluate geo-targeting performance more rigorously. Also, one might want to include the firm's marketing actions (e.g., advertising expenditures, price

promotions) as well. Lastly, I rely on zip-level aggregate sales data but one might want to assess social influence or the preference minority status at the individual level. I plan to pursue these issues in future research.

Bibliography

Agrawal, Ajay, Devesh Kapur, and John McHale (2008), "How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data," *Journal of Urban Economics*, forthcoming.

Agresti, Alan (2002), *Categorical Data Analysis*, Wiley: New York, NY.

Albuquerque, Paulo, Bart J. Bronnenberg, and Charles J. Corbett (2007), "A Spatiotemporal Analysis of the Global Diffusion of ISO 9000 and ISO 14000 Certification," *Management Science*, 53 (3), 451-468.

Anderson, Chris (2006), *The Long Tail: Why the Future of Business is Selling Less of More*, Hyperion: New York, NY.

Anderson, Eric T., Nathan M. Fong, Duncan I. Simester, and Catherine E. Tucker (2009), "How Sales Taxes Affect Customer and Firm Behavior: The Role of Search on the Internet," *Journal of Marketing Research*, forthcoming.

Anderson, Evan E. (1979), "An Analysis of Retail Display Space: Theory and Methods," *Journal of Business*, 52 (1), 103-118.

Angers, Jean-Francois and Mohan Delampady (1992), "Hierarchical Bayesian Curve Fitting and Smoothing," *The Canadian Journal of Statistics*, 20 (1), 35-49.

Anselin, Luc (1988), *Spatial Econometrics: Methods and Models*, Kluwer: Boston, MA.

Ansley, Craig F. and Robert Kohn (1990), "Filtering and Smoothing in State Space Models with Partially Diffuse Conditions," *Journal of Time Series Analysis*, 11, 275-316.

Avery, Jill, Thomas J. Steenburgh, John Deighton, and Mary Caravella (2008), "Adding Bricks to Clicks: The Effects of Store Openings on Sales through Direct Channels," Working Paper, Harvard Business School.

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

_____, Prabhudev Konana, and Nirup M. Menon (2003), "Customer Satisfaction in Virtual Environments: A Study of Online Investing," *Management Science*, 49 (7), 871-889.

Barbour, A. D., Lars Holst, and Svante Janson (1992), *Poisson Approximation*, Clarendon: Oxford, England.

Bass, Frank (1969), "A New Product Growth Model for Consumer Durables," *Management Science*, 15 (5), 215-227.

_____, Norris Bruce, Sumit Majundar, and B.P.S. Murthi (2007), "Wearout Effects of Different Advertising Themes: A Dynamic Bayesian Model of the Advertising-Sales Relationship," *Marketing Science*, 26 (2), 179-195

Bell, David R., and Sangyoung Song (2007), "Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing," *Quantitative Marketing and Economics*, 5 (4), 361-400.

Bhatnagar, Amit and Brian T. Ratchford (2004), "A Model of Retail Format Competition for Non-Durable Goods," *International Journal of Research in Marketing*, 21, 39-59.

Bronnenberg, Bart J., and Carl F. Mela (2004), "Market Roll-out and Retailer Adoption of New Brands," *Marketing Science*, 23 (4), 500–518.

Brown, Jacqueline Johnson and Peter H. Reingen (1987), "Social Ties and Word-of-Mouth Referral Behavior," *Journal of Consumer Research*, 14 (3), 350-362.

Brynjolfsson, Erik and Michael D. Smith (2000), "Frictionless Commerce? A Comparison of Internet and Conventional Retailers," *Management Science*, 46 (4), 563–585.

_____, Yu (Jeffrey) Hu, and Mohammad S. Rahman (2008), "Battle of the Retail Channels: How Product Selection and Geography Drive Cross-Channel Competition," Working Paper, Sloan School of Management, MIT.

Bucklin, Randolph E., Sunil Gupta, and S. Siddarth (1998), "Determining Segmentation in Sales Response across Consumer Purchase Behaviors," *Journal of Marketing Research*, 35 (2), 189-197.

Cairncross, Frances (1997), *The Death of Distance*, Cambridge, MA: Harvard University Press.

Cameron, A. Colin and Pravin K. Trivedi (1986), "Econometric Models Based on Count Data" Comparisons and Applications of Some Estimators and Tests," *Journal of Applied Econometrics*, 1, 29-54.

Casella, George and Edward I. George (1992), "Explaining the Gibbs sampler," *American Statistician*, 46 (3), 167-174.

Chen, Yuxin, James D. Hess, Ronald T. Wilcox, and Z. John Zhang (1999), "Accounting Profits Versus Marketing Profits: A Relevant Metric for Category Management," *Marketing Science*, 18 (3), 208-229.

Cheng, June and Barrie R. Nault (2007), "Internet Channel Entry: Retail Coverage and Entry Cost Advantage," *Information Technology and Management*, 8 (2), 111-132.

Chib, Siddhartha (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90 (432), 1313-1321.

_____ and Ivan Jeliazkov (2001), "Marginal Likelihood From the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96 (453), 270-281.

Chiou, Lesley (2005), "Empirical Analysis of Retail Competition: Spatial Differentiation at Wal-Mart, Amazon.com, and their Competitors," Working Paper, Occidental College.

Christaller, Walter (1933), *Die zentralen Orte in Suddeutschland*, Jena: Gustav Fischer. (Translated (in part) by Charlisle W. Baskin (1966), as *Central Places in Southern Germany*, Englewood Cliffs, NJ: Prentice Hall).

Claude, Besner (2002), "A Spatial Autoregressive Specification with a Comparable Sales Weighting Scheme," *Journal of Real Estate Research*, 24 (2), 193-211.

Conley, Timothy G. and Giorgio Topa (2002), "Socio-Economic Distance and Spatial Patterns in Unemployment," *Journal of Econometrics*, 17 (4), 303-327.

Cressie, Noel (1993), *Statistics for Spatial Data*, Wiley: New York, NY.

Dekimpe, Marnik G. and Dominique M. Hanssens (1995), "The Persistence of Marketing Effects on Sales", *Marketing Science*, 14 (1), 1-21.

Dhar, Sanjay K. and Stephen J. Hoch 1997 "Why Store Brand Penetration Varies by Retailer," *Marketing Science*, 16 (3), 208-227.

Dowding, Keith, Peter John, and Stephen Biggs (1994), "Tiebout: A Survey of the Empirical Literature," *Urban Studies*, 31 (4/5), 767-797.

Engel, James E., Roger D. Blackwell, and Robert J. Kegerreis (1969), "How Information is Used to Adopt an Innovation," *Journal of Advertising Research*, 9 (4), 3-8.

Farris, Paul, James Olver, and Cornelis De Kluyver (1989), "The Relationship Between Distribution and Market Share," *Marketing Science*, 8 (2), 107-128.

Fischer, Claude S. (1978), "Urban-to-Rural Diffusion of Opinions in Contemporary America," *American Journal of Sociology*, 84 (1), 151-159.

Forman, Chris, Anindya Ghose, and Batia 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.

_____, _____, and Avi 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.

Fornell, Claes (1995), "The Quality of Economic Output: Empirical Generalizations about its Distribution and Relationship to Market Share," *Marketing Science*, 14 (3), 203-211.

French, Jonathan L., Erin E. Kammann, and Matt P. Wand (2001), "Semiparametric Nonlinear Mixed-Effects Models and Their Applications: Comment," *Journal of the American Statistical Association*, 96 (456), 1285-1288.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2003), *Bayesian Data Analysis, 2nd Edition*, Chapman & Hall, New York.

Getis, Arthur, and J. Keith Ord (1992), "The Analysis of Spatial Association by Use of Distance Statistics," *Geographical Analysis*, 24, 189-206.

Ghose, Anindya, Michael D. Smith, Rahul Telang (2006), "Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact," *Information Systems Research*, 17 (1), 3-19.

Glaeser, Edward L., J Kolko, and Albert Saiz (2001), "Consumer City," *Journal of Economic Geography*, 1 (1), 27-50.

Godes, David and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, 23 (4), 545-560.

_____ and _____ (2008), "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," *Marketing Science*, forthcoming.

Goolsbee, Austan (2000), "In a World Without Borders: The Impact of Taxes on Internet Commerce," *Quarterly Journal of Economics*, 115 (2), 561-576

_____ and Peter J. Klenow (2002), "Evidence of Learning and Network Externalities in the Diffusion of Home Computers," *Journal of Law and Economics*, 45(2), 317-342.

Greene, William (2002), *Econometric Analysis*, Prentice Hall: Upper Saddle River, NJ.

Guo, Wensheng (2002), "Functional Mixed Effects Models," *Biometrics*, 58 (March), 121-128.

Guo, Wensheng (2003), "Dynamic State Space Models," *Journal of Time Series Analysis*, 24 (2), 149-158.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001), *The Elements of Statistical Learning*, Springer, New York.

Hastings, W. Keith (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57 (1), 97-109.

Hauser, John, Gerard J. Tellis, and Abbie Griffin (2006), "Research on Innovation: A Review and Agenda for Marketing Science," *Marketing Science*, 25 (6), 687-717.

Herr, Paul M., Frank R. Kardes, and John Kim (1991), "Effects of Word-of-Mouth and Product-Attribute Information on Persuation: An Accessibility-Diagnosticity Perspective," *Journal of Consumer Research*, 17 (4), 454-462.

Hoch, Stephen J. and John Deighton (1989), "Managing What Consumers Learn From Experience," *Journal of Marketing*, 53 (April).

Howard, Philip E., Lee Raine, and Steve Jones (2001), "Access, Civic Involvement, and Social Interaction," *American Behavioral Scientist*, 45 (3), 382-404.

Huff, David L. (1964), "Defining and Estimating a Trade Area," *Journal of Marketing*, 28 (3), 34-38.

Jank, Wolfgang, P. K. Kannan (2005), "Understanding Geographic Markets of Online Firms Using Spatial Models of Customer Choice," *Marketing Science*, 24 (4), 623-634.

Kalyanam, Kirthi and Thomas S. Shively (1998), "Estimating Irregular Pricing Effects: A Stochastic Spline Regression Approach," *Journal of Marketing Research*, 35 (1), 16-29.

Katz, James E., Ronald E. Rice, and Philip Aspden (2001), "The Internet, 1995-2000," *American Behavioral Scientist*, 45 (3), 405-419.

Keeney, Ralph L. (1999), "The Value of Internet Commerce to the Customer," *Management Science*, 45 (4), 533-542.

Knorr-Held, Leonhard and Julian Besag (1998), "Modeling Risk from a Disease in Time and Space," *Statistics in Medicine,* 17 (18), 2045-2060.

Koopman, Siem J. (1997), "Exact Initial Kalman Filtering and Smoothing for Nonstationary Time Series Models," *Journal of American Statistical Association*, 92 (440), 1630-1638.

Koopman, Siem J. and James Durbin (2000), "Fast Filtering and Smoothing for Multivariate State Space Models," *Journal of Time Series Analysis*, 21 (3), 281-296.

Lal, Rajiv and Miklos Sarvary (1999), "When and How Is the Internet Likely to Decrease Price Competition?" *Marketing Science,* 18 (4), 485-503.

Leeflang, Peter S.H., Tammo H.A. Bijmolt, Jenny van Doorn, Dominique M. Hanssens, Harald J. van Heerde, Peter C. Verhoef, and Jaap E. Wieringa (2009), "Creating Lift versus Building the Base: Current Trends in Marketing Dynamics," *International Journal of Research in Marketing*, 26, 13-20.

LeSage, J. P. and R. K. Pace (2005) "Matrix Exponential Spatial Specification," *Journal of Econometrics*, 140 (1), 190-214.

Lewis, Michael (2006), "Customer Acquisition Promotions and Customer Asset Value," *Journal of Marketing Research*, 43 (2), 195-203.

Libai, Barak, Eitan Muller, and Renana Peres (2005), "The Role of Seeding in Multi-Market Entry," *International Journal of Research in Marketing*, 22 (4), 375-393.

Lilien, Gary L. and Arvind Rangaswamy (2004), *Marketing Engineering, 2nd Edition*, Trafford: Victoria, B.C.

Lord, Charles G., Lee Ross, and Mark R. Lepper (1979), "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology*, 37 (11), 2098-2109.

Lynch, John G. and Dan Ariely (2000), "Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution," *Marketing Science*, 19 (1), 83-103.

Manchanda, Puneet, Ying Xie, and Nara Youn (2008), "The Role of Targeted Communication and Contagion in Product Adoption," *Marketing Science*, forthcoming.

Manski, Charles F. (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60 (3), 531-542.

_____ (2000), "Economic Analysis of Social Interactions," *Journal of Economic Perspectives,* 14 (3), 115-136.

Michener, Ron and Carla Tighe (1992), "A Poisson Regression Model of Highway Fatalities," *American Economic Review,* 82 (2), 452–456.

Molenberghs, Geert and Geert Verbeke (2006), *Models for Discrete Longitudinal Data*, Springer-Verlag: New York, NY.

Naik, Prasad A., Murali K. Mantrala, and Alan G. Sawyer (1998), "Planning Media Schedules in the Presence of Dynamic Advertising Quality", *Marketing Science,* 17 (3), 214-235.

Newton, Michael A. and Adrian E. Raftery (1994), "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap," *Journal of Royal Statistical Society Series B*, 56, 41-42.

Overby, Eric and Sandy Jap (2008), "Electronic and Physical Market Channels: A Multi-Year Investigation in a Market for Products of Uncertain Quality," Working Paper, Georgia Institute of Technology.

Pauwels, Koen and Scott A. Neslin (2008) "Building Bricks and Mortar: The Impact of Opening Physical Stores in a Multichannel Environment," MSI Working Paper 08-102, Cambridge, MA.

Putsis, William P. (1998), "Parameter Variation and New Product Diffusion," *Journal of Forecasting*, 17 (3/4), 231-257.

Rabe-Hesketh, Sophia and Anders Skrondal (2006), *Multilevel and Longitudinal Modeling Using Stata*, Statacorp LP: College Station, TX.

Ravishanker, Nalini, and Dipak K. Dey (2002), *A First Course in Linear Model Theory*, Chapman and Hall, Boca Raton, FL.

Reibstein, David J. and Paul W. Farris (1995), "Market Share and Distribution: A Generalization, a Speculation, and Some Implications," *Marketing Science*, 14 (3), 190-202.

Reichheld, Frederick F. (2006), The Ultimate Question: Driving Good Profits and True Growth. Boston: Harvard Business School Press.

Reinartz, Werner, Jacquelyn S. Thomas, and V. Kumar (2005), "Balancing Acquisition and Retention Resources to Maximize Customer Profitability," *Journal of Marketing*, 69 (1), 63-79.

Richins, Marsha L. and Peter H. Bloch (1986), "After the New Wears Off: The Temporal Context of Product Involvement," *Journal of Consumer Research*, 13 (2), 280-285.

Rosenblat, Tanya S. and Markus M. Mobius (2004), "Getting Closer or Drifting Apart?" *Quarterly Journal of Economics*, 119 (3), 971-1009.

Ross, Sheldon M. (1996), *Stochastic Processes, 2nd Edition*, Wiley, New York.

Rossi, Peter E. and Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22 (3), 304-328.

Ruppert, David, Matt P. Wand, and Ray J. Carroll (2003), Semiparametric Regression, Cambridge: Cambridge University Press.

Shonkwiler, J. Scott and Thomas R. Harris (1996), "Rural Retail Business Thresholds and Interdependencies," *Journal of Regional Science*, 36 (4), 617-630.

Simonoff, J. S. (1986), *Smoothing Methods in Statistics*, Springer, New York.

Sinai, Todd and Joel Waldfogel (2004), "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?" *Journal of Urban Economics*, 56 (1), 1-24.

Stremersch and Lemmens (2009), "Sales Growth of New Pharmaceuticals Across the Globe: the Role of Regulating Regimes," *Marketing Science*, forthcoming.

Tang, Christopher S., David R. Bell, and Teck-Hua Ho (2001), "Store Choice and Shopping Behavior: How Price Format Works," *California Management Review*, 43 (2), 56-74.

Tiebout, Charles M. (1956), "A Pure Theory of Local Expenditures," *Journal of Political Economy*, 64 (5), 416-424.

Van Alstyne, Marshall and Erik Brynjolfsson (2005), "Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities," *Management Science*, 51 (6), 851-868.

Van den Bulte, Christophe and Gary L. Lilien (1997), "Bias and Systematic Change in the Parameter Estimates of Macro-Level Diffusion Models," *Marketing Science*, 16 (4), 338–353.

_____ and Yogesh V. Joshi (2007), "New Product Diffusion with Influentials and Imitators," *Marketing Science*, 26 (3), 400-21.

Villanueva, Julian, Shijin Yoo, and Dominique M. Hanssens (2008). "The Impact of Marketing-Induced vs. Word-of-Mouth Customer Acquisition on Customer Equity," *Journal of Marketing Research*, 45 (1), 48-59.

Wahba, Grace (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society Series B*, 40 (3), 364-372.

Wahba, Grace (1983), "Bayesian Confidence Intervals for Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society (Series B)*, 45, 133-150.

Waldfogel, Joel (2003), "Preference Externalities," *RAND Journal of Economics*, 34 (1), 557-568.

_____ (2007), *The Tyranny of the Market: Why You Can't Always Get What You Want*, Harvard University Press: Cambridge, MA.

Wand, Matt P. (2003), "Smoothing and Mixed Models," *Computational Statistics*, 18, 223-249.

Wecker, William E. and Craig F. Ansley (1983), "The Signal Extraction Approach to Nonliear Regression and Spline Smoothing," *Journal of the American Statistical Association*, 78, 81-89.

Wedel, Michel and Jie Zhang (2004), "Analyzing Brand Competition Across Subcategories," *Journal of Marketing Research*, 41(4), 448-456.

West, Mike, and Jeff Harrison (1997), *Bayesian Forecasting and Dynamic Models*, Springer: New York, NY.

Wikle, Christopher K. and Mevin B. Hooten (2006), "Hierarchical Bayesian Spatio-Temporal Models for Population Spread," In Clark, J.S. and A. Gelfand (Eds.) *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*, Oxford University Press.

Yang, Sha and Greg M. Allenby (2003), "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research*, 40 (3), 282-294.

Yoo, Shijin and Koen Pauwels (2008), "Generalized Long-Term Price Effects: Are They Asymmetric, Nonmonotonic, and time Dependent?" MSI Working Paper 08-108, *Marketing Science Institute*, Cambridge: MA.