Social Contagion and Trial on the Internet: Evidence from Online Grocery Retailing

David R. Bell and Sangyoung Song

May 12, 2004

---

1Associate Professor and Doctoral Candidate respectively, The Wharton School, University of Pennsylvania. Address: 700 Huntsman Hall, 3730 Walnut Street, Philadelphia PA 19104. Corresponding author: David R. Bell. E-mail: davidb@wharton.upenn.edu, Tel: (215) 898-8253, Fax (215) 898-2534. We would like to thank seminar participants at Carnegie Mellon University, Dartmouth College, Massachusetts Institute of Technology, University of Auckland, and the University of Texas at Dallas for their many insights and comments. Peter Danaher, John Zhang and especially Christophe Van den Bulte provided detailed feedback and suggestions. We are also very grateful to Lisa Kent, CEO at netgrocer.com for supplying the data used in this study. Financial support was generously provided by WeBI, the Wharton eBusiness Initiative.
Social Contagion and Trial on the Internet: Evidence from Online Grocery Retailing

Abstract

For a traditional retailer, the size of the customer pool can evolve over time but is largely bounded in space. In contrast, an Internet retailer with the appropriate shipping infrastructure can draw customers from a wide-ranging geographical area (e.g., the entire United States). We examine the trial decision for customers shopping at an Internet grocery retailer (netgrocer.com). Drawing on literature in economics, marketing and sociology, we conjecture that exposure to the actions of spatially proximate others — either through direct social interaction or passive observation — influences the trial decision of those who have yet to experience the service.

This idea is tested in a discrete time hazard framework in which consumer trial decisions arise from utility-maximizing behavior. Moreover, our derivation allows use of region-level data to estimate a model consistent with individual utility maximization, even in the absence of individual level covariates. We find that the marginal impact of the so-called “neighborhood effect” is economically meaningful as it results in an approximately fifty percent increase in the baseline hazard of trial. The effect is robust to the inclusion of a broad set of covariates, region-level fixed effects, and time-dependent heterogeneity in the baseline rate. The model is calibrated on a unique dataset spanning: (1) 29,701 residential zip codes, (2) transactions from forty-five months since the inception of the service, and (3) zip-code specific contiguity data from Geographic Information System (GIS) analysis. Substantive implications for customer base evolution and Internet retailing are discussed.

Key Words: Discrete Time Hazard, Neighborhood Effects, Retailing, Social Contagion
1 Introduction

“... The choice of a store location has a profound effect on the entire business life of a retail operation. A bad choice may all but guarantee failure, a good choice, success.”

“Store Location: Little Things Mean A Lot” CBSC, Government of Canada.

For retailers “location, location, location” is a familiar mantra and a vast literature has emerged to substantiate the importance of this variable in determining success. While pricing and assortment are important as well location is thought to account for the majority of the variation in outlet choice in most retail settings.\(^1\) Moreover, the attractiveness of an outlet to a shopper declines exponentially the further the individual is from the store (e.g., Fotheringham 1988; Huff 1964). For the Internet retailer the physical location of the store relative to a particular set of potential customers is of almost no consequence (indeed, it seems logical to assert that the trading area of the Internet retailer is constrained only by the availability of shipment infrastructure for distributing orders). What may be critical is the location of existing customers relative to new potential customers.

The unique market context of the Internet retailer — geographically dispersed customers and competitors — raises important (and thus far unstudied) questions concerning the evolution of the customer franchise. One especially fundamental issue is the role of existing customers in recruiting or influencing new potential customers. The role of emulation in decision making has been the focus of considerable attention in recent theoretical and empirical research in the economics and sociology literatures (e.g., Burt 1987; Goolsbee and Klenow 2002; Tolnay, Deane and Beck 1996; Van den Bulte and Lilien 2001) and this is the central focus of our research. We are in a unique position to address this phenomenon by examining the space-time evolution of trial behavior from the inception of a new Internet retailing service. Before developing the modeling framework and its theoretical underpin-

\(^1\)For example, Progressive Grocer (April 1995) reports that location explains up to seventy percent of the variance in consumer choice of supermarket retailers.
nings, it is helpful to have a descriptive appreciation of the phenomenon. Figure 1 provides a summary of trial orders for netgrocer.com, specifically the total revenue earned and average order value by state, in panels (a) and (b) respectively. The empirical distribution of these two variables is broken into quintiles as indicated by the legend.2

[Figure 1 about here]

California, Texas, Florida and New York generate the greatest amount of revenue, while the average order values are higher in the interior western states of Nevada, Wyoming, Colorado and New Mexico. Likely explanations for the first observation include population size as these states are among the most populous, while the second may be a consequence of (more limited) access to retail services. The important observation is that the Netgrocer customer base spans the entire United States. The data summarized in Figure 1 are cumulated from the inception of netgrocer.com in May 1997 through January 2001. Orders were, and still are, shipped from a company warehouse in New Jersey and sent to customers via Federal Express.

For the remainder of the paper we focus on modeling the spread of initial orders over time and space, with special emphasis on the role of social contagion or “neighborhood effects” in speeding or inhibiting the process of trial as individuals learn from previous triers.3 Research from the economics and sociology literatures in particular motivates the role of social contagion in decision making (e.g., Bikhchandani, Hirshleifer and Welch 1998; Case 1991; Case, Hines and Rosen 1989; Singer and Spilerman 1983). We propose and

---

2For reasons of confidentiality we have excluded the dollar values from panel (a), however it is the case that all forty eight contiguous states generate revenue. The lower value for the bottom quintile is approximately $15,000 while the lower value of the top quintile is in the neighborhood of $500,000.

3We will use the terms “contagion” and “neighborhood effects” interchangeably throughout the paper and will provide precise empirical definitions in a subsequent section. The former term is preferred by sociologists and the latter by economists, but both have a common interpretation.
estimate a model in which trial decisions are conceptualized to result from utility-maximizing behavior (trial is observed when an individual-specific threshold for action is exceeded). The advantage of this conceptualization is that it allows one to consider a time-dependent process through a sequence of binary actions. Specifically, a discrete time hazard model estimated on time-dependent trial data is consistent with random utility maximization over binary outcomes (e.g., Van den Bulte and Lilien 2002).

**Contribution and Caveats**

This research demonstrates empirically the importance of neighborhood effects in the space-time pattern of trial for an Internet retailing service. The substantive message is in line with Goolsbee and Klenow (2002) who, after implementing the appropriate controls, find that individuals are more likely to buy home computers in areas where greater numbers of other individuals already own computers. We find similar neighborhood effects for trial decisions for an Internet retailing service, given a simple representation of social contagion derived from physical proximity. The effect is economically and statistically important and is robust to the inclusion of controls for region and time-specific fixed effects, region covariates and unobserved heterogeneity in the baseline hazard. In this research we make the following new contributions to the literature:

- First, we develop a framework for empirical analysis of a new phenomenon in retailing that is not yet well understood, namely the evolution of customer trials for an Internet retailer. In so doing, we provide insight into the consequences of having spatially dispersed customers and competitors.

- Second, we provide an analytical derivation to estimate an inherently individual-level decision process using region-level data. We exploit the relationship between random utility maximization and a discrete time hazard model (Van den Bulte and Lilien 2002) coupled with knowledge of the number of individuals in each region to accomplish this.
• Third, we generate several new substantive insights by showing: (1) neighborhood effects influence “private behavior” — trial of an Internet-based retailer — even in the absence of usage-based network effects, (2) neighborhood effects are economically significant and contribute to an approximately fifty percent increase in the baseline hazard, and (3) the space-time customer trial pattern for an Internet retailer is strongly related to local conditions (e.g., characteristics of the constituent population, physical and retail environments, etc.). These substantive findings have implications for segmentation strategies.

In summary, the primary contribution is to show that in an Internet retailing context the neighborhood or social contagion effect has a basis in theory, is present empirically, and is economically meaningful. We do not offer or claim a complete elaboration of all nuances of the effect, such as potential variation in magnitude across time and/or observational units. We are unable to fully articulate the exact nature of the mechanism as the data do not allow one to distinguish between contagion through direct social interaction and contagion through observation alone. We leave such pursuits for future research.

The remainder of the paper is organized as follows. The next section reviews related literature and offers a conceptual framework. The following section develops the statistical model in a random utility setting in which the neighborhood effect is both consistent with rational behavior by consumers and identified. Subsequent sections introduce the data and accompanying exploratory analysis, provide the estimation results and summarize the empirical findings. The paper concludes with implications for Internet retailing practice and ongoing research on social contagion in this setting.
2 Background and Motivation

This section briefly reviews findings from relevant research in marketing on consumer decision making and research in economics and sociology on neighborhood effects. It provides the conceptual motivation for a statistical model of consumer behavior introduced and developed in the next section.

Consumer Decision Making

A longstanding tradition in marketing posits a generic consumer decision making process that can be modeled as a series of discrete stages through which an individual is assumed to pass in an approximately linear fashion. Lilien, Kotler and Moorthy (1993, p. 26) describe a five stage process: need arousal, information search, evaluation, purchase and post purchase, and elements of this conceptualization can be traced back to earlier work in consumer behavior (e.g., Howard and Sheth 1969). Each stage of the process differs with respect to sources and use of information, time taken, and decision rules invoked and applied.

Given the potential complexity of the entire process and interactions among stages, researchers have typically addressed only parts of it in individual pieces of research, or have “aggregated” the process into a fewer number of steps. Work in this area includes evaluation models of consideration set composition with cost-benefit trade offs (e.g., Hauser and Wernerfelt 1990), and joint models of consideration set formation and choice (e.g., Roberts and Lattin 1991). Gensch (1987) focuses on use of non-compensatory decision rules in industrial buying while Van den Bulte and Lilien (2002) study the stages of awareness and evaluation. An novel feature of the latter work is the ability of the model to disentangle two separate stages when only single stage adoption data are observed.

The model we propose focuses exclusively on a single stage (the first purchase, or trial), yet it incorporates important elements suggested by prior literature. Trial outcomes arise because unobserved utility thresholds have been crossed, and new information is potentially
revealed to current non-triers as a consequence of trial by proximate others. We derive the trial utility threshold model in a discrete time setting and show why and how information about decisions made by others might affect the focal individual who has yet to try. The motivation for this conceptualization stems from the nature of the institutional setting and also from related literature in economics and sociology on the role of neighborhood effects in individual decision making.

**Neighborhood Effects**

Several theoretical and empirical studies have emerged in economics and sociology to deal with a phenomenon that may be broadly defined as neighborhood effects or social contagion. The underlying notion is that individuals may be dispelled to act (or defer action) based on what can be learned by observing the actions of others.\(^4\) In distilling the key ideas in this literature it is important to distinguish between theoretical reasons for the behavior and the mechanisms by which the phenomenon can be formalized and studied in empirical work. One line of reasoning is that in relatively closed communities individual knowledge (alone and through aggregation) may represent a public good. To capture this idea, Besley and Case (1993) propose an approach to developing dynamic structural models which accommodate the updating rules of individual agents when they are exposed to knowledge transmission by others. In examining technology adoption by farmers, Foster and Rosenweig (1995) show that aggregate-level imperfect knowledge about the management of new seeds was an impediment to adoption by individual farmers, however this barrier diminished with time. Case, Hines and Rosen (1993) offer strong empirical support for the finding that public spending in a particular region is strongly influenced by levels of expenditure in neighboring jurisdictions (for every one dollar spent by a contiguous neighbor an additional seventy cents is spent

\(^4\)Observational learning occurs when one decision maker observes the actions of others but can only make inferences about motivation for the observed action. This should be distinguished from word of mouth which arises from dialog (see Godes and Mayzlin 2004 for an analysis of word of mouth on the Internet).
by the focal region). Moreover, failure to account for such effects in estimation leads to an upward bias in other model parameters.

A distinct yet related literature deals with observational learning when the innovation itself (rather than simply the decision making process) exhibits network effects. In this case, the individual’s decision to adopt the innovation is influenced by the pool of current users and anticipated eventual penetration of the innovation. Examples here include studies by Hannan and McDowell (1984; 1987) and Saloner and Shepard (1995) on ATM technology adoption in the banking industry. Conversely, we focus exclusively on what could be termed an “informational network effect”. That is, information revealed by “relevant” neighbors is incorporated into the decisions of the focal individual solely because it facilitates (initial) decision making.⁵

In addition to these findings from the economics literature, a broad set of insights have emerged from the field of sociology. A large body of work has concentrated on defining measures of social connectedness, and in particular the extent of information transfer among and between groups of individuals (see Burt 1980 for a comprehensive treatment; also Greve, Strang and Tuma 1995; Strang and Tuma 1993). While much of the work in economics has assumed that social contagion (i.e., transmission of information, observation of actions of relevant others) is primarily grounded in geographical contiguity, sociologists have placed greater emphasis on richer definitions of affiliation. For example, Burt (1980) distinguishes structural equivalence between individuals and social cohesion. A number of related studies devote attention to describing appropriate forms of an affiliation matrix to capture the nature, strength and timing of interaction within groups. One example of this kind of work is provided in Chaves (1996) who shows that the diffusion of gender equality in churches is

⁵One could perhaps make the argument that individual trial decisions for Netgrocer are influenced by the size of the installed base because individuals perceive that a larger installed base increases the probability that the service continues as a going concern. In our model, any such effect is absorbed into the baseline hazard.
influenced by cultural boundaries and affiliations within the denominations studied.

The power of social contact in dissemination of information generally is shown in an interesting study by Oyen and De Fleur (1953). In a field experiment where leaflets were distributed by plane over four areas of Washington state, knowledge of the message content by individual discovery was found to decline dramatically with increased distance from the drop areas, while knowledge via social contact (i.e., learning the content from others) tended to increase within the circumscribed distance. Finally, it is also important to keep in mind that observational learning can induce both negative and positive dispositions with respect to the innovation. Tolnay, Deane and Beck (1996) study state-tolerated racist violence at the turn of the century and report that the number of lynchings occurring in a particular county decreased with the number of prior lynchings in contiguous neighbors. That is, contagion effects can be negative (i.e., slow the spread of the phenomenon) as well as positive.

3 Theory and Model

This section provides the utility theory and statistical model for individual trial decisions. We motivate the statistical approach by highlighting the link between individual utility maximization for time-dependent binary choices and a discrete time hazard model, and note that a model of this type can be formulated to estimate the parameters of an underlying continuous time process. With respect to trial, we claim that it may be rational for individuals to incorporate information on the behavior of others when making their own trial decisions. Particular attention is paid to the empirical representation of the neighborhood effect in the individuals’ utility functions in a way that allows it to be identifiable (Manski 1993) and consistent with rational behavior (Brock and Durlaf 2001).
**Individual Utility for Trial**

The analysis to follow centers on the timing of individual trial decisions, for individuals located in regions $z = 1, \ldots, Z$. Let $T_{iz}$ denote the uncensored time of occurrence of trial for individual $i$, and $\lambda(t)$ the hazard rate. The instantaneous probability that an event (i.e., trial) occurs for individual $i$ at time $t$ is

$$
\lambda(t) = \lim_{\Delta \to 0} \frac{P(t \leq T_{iz} \leq t + \Delta | T_{iz} \geq t)}{\Delta}.
$$

(1)

$\lambda(t)$ may also be interpreted as the expected number of events in a time interval of unit length. In order to introduce a behavioral underpinning for individual trial decisions, we instead work with the discrete time analog to equation (1)

$$
P_{iz}(t) = P(T_{iz} = t | T_{iz} \geq t, X_{iz}(t)).
$$

(2)

$X_{iz}(t)$ is the set of covariates that potentially influences the uncensored time of trial. Equation (2) is also the conditional probability that an event occurs at $t$, given that it has yet to occur and will be shown to depend upon parameters governing individual utility functions. Furthermore, it is important to note that use of a discrete time model need not result in a loss of information nor be subject to aggregation bias. Specifically, discrete time parameter estimates derived from the complementary log-log link function are also the estimates of an underlying continuous time proportional hazards model (Prentice and Gloeckler 1978).

In addition to the important substantive advantage of allowing a utility interpretation, the discrete time framework also confers methodological benefits associated with simplicity of estimation and the ability to incorporate time-dependent covariates directly into the hazard. To generate $P_{iz}(t)$, allow that every individual $i$ in region $z$ has the potential engage in trial in any period $t$, beginning at time period 1 (when the innovation first becomes available). In

---

6Heckman and Singer (1984) note that while the parameters of discrete time models are not invariant to the length of time intervals chosen, a discrete time model derived from a continuous time process does not suffer from this defect. Additional discussion of this point is given in Allison (2001, p. 216-219.)
each discrete time period $t$ for each individual $i$, $y_{iz}(t) \in \{0, 1\}$ is observed where $y_{iz}(t) = 1$ indicates that the individual experienced the trial in period $t$. The complete decision history of the individual is described by the time-indexed sequence $\{y_{iz}(t)\}$, $t = 1, \ldots, T_{iz} \leq T$ where $T_{iz}$ is the time period in which trial takes place for individual $i$. If trial never occurs, $\{y_{iz}(t)\}$ is a sequence of zeros of length $T$ (the end of the observation period).

Assume that individual $i$ at location $z$ has an unobserved utility value for trial at time $t$ given by

$$U_{iz}(t) = V_{iz}(t) - \epsilon_{iz}(t)$$  \hspace{1cm} (3)

where $V_{iz}(t)$ is a linear in parameters polynomial sum and $\epsilon_{iz}(t)$ is a stochastic disturbance term. In general, $V_{iz}(t)$ potentially depends on individual, region and time-dependent characteristics, and the relationship between the components of $V_{iz}(t)$ and $y_{iz}(t)$ is governed by the probability distribution of $\epsilon_{iz}(t)$.

**An Individual Model with Region Level Data**

Given individual-level covariate information it is straightforward to estimate the parameters of equation (3). As a practical empirical matter, we do not observe exogenous information at the level of the individual $i$ for the trial decision, making it impossible to specify $V_{iz}(t)$ in equation (3) at the desired level of aggregation directly. (We do observe the zip code region where the order came from.) Furthermore, the lack of individual-level information makes it impossible to characterize all members of the risk set, or individuals who have yet to experience trial, at each time period. Complete identification of the risk set at each period is necessary for proper specification of the discrete time model.

Fortunately, we show how these problems can be circumvented by utilizing data from a higher level of aggregation (region) in combination with knowledge of $n_z$ the number of individuals located in region $z$, and an appropriate choice of probability distribution for $\epsilon_{iz}(t)$. First, note that at the individual level trial occurs when the utility threshold for action is
crossed. Namely, \( y_{iz}(t) = 1 \) when \( U_{iz}(t) > \tau \) where the constant \( \tau \) can be normalized to zero without loss of generality. Let \( \epsilon_{iz}(t) \) be independently and identically distributed over individuals and time within region, with the following probability distribution

\[
f(\epsilon) = \frac{1}{\mu} \exp \left[ \frac{\epsilon - \eta}{\mu} \right] \exp \left\{ -e^{\frac{\epsilon - \eta}{\mu}} \right\}. \tag{4}
\]

The probability that individual \( i \) in region \( z \) experiences trial at time \( t \) is obtained from the cumulative distribution function \( F(\epsilon) \) as

\[
P(y_{iz}(t) = 1) = P(\epsilon_{iz}(t) \leq V_{iz}(t)) = 1 - \exp \left\{ - \exp \left[ \frac{V_{iz}(t) - \eta}{\mu} \right] \right\}. \tag{5}
\]

Rather than model this probability for every individual \( i \) in every region \( z \), we instead model the probability of the first trial in a region. Intuitively, the more individuals there are in a region, the greater the chance that at least one individual will experience trial by a particular date. In a model that combines data across regions, it is vital to take this into account. The probability that trial occurs in region \( z \) at time \( t \), given that trial has yet to occur there is equivalent to the probability that the utility of the maximal individual exceeds the threshold. Note that while this maximal individual cannot be described in terms of individual-level characteristics, s/he can be represented by a combination of region-specific characteristics and the implied individual-level stochastic component of utility. That is

\[
P(y_{z}(t) = 1) = P(\max_{i} \{ U_{iz}(t) \ i = 1,\ldots,n_z \} \geq 0)
= P(\max_{i} \{ V_{iz}(t) - \epsilon_{iz}(t) \} \geq 0)
= P( V_{z}(t) - \min_{i} \{ \epsilon_{iz}(t) \} \geq 0)
= P(\min_{i} \{ \epsilon_{iz}(t) \} \leq V_{z}(t)) \tag{6}
\]

Equation (6) reframes the event — trial in region \( z \) at time \( t \) — with respect to the distribution of the minimum of \( i = 1,\ldots,n_z \) random variables. In words, the probability that the
unobserved maximal individual’s utility exceeds zero is equivalent to the probability that the observed deterministic utility $V_z(t)$ for the representative individual from the region exceeds the minimum value of all $\epsilon_{iz}(t)$. The Gumbel or Type I Extreme Value distribution in equation (4) with location parameter $\eta$ and scale parameter $\mu$ has the useful property that the distribution of the minimum of $n_z$ independent random variables is also Gumbel

$$
\epsilon_{iz}(t) \sim G(\eta, \mu)
$$

$$
\epsilon_{iz}^{\text{min}}(t) = \min_i \{ \epsilon_{iz}, i = 1, \ldots, n_z \}
$$

$$
\sim G(\eta - \mu \ln(n_z), \mu).
$$

(7)

The probability that trial occurs in region $z$ given that it has not yet occurred is obtained from the distribution function as

$$
P(y_z(t) = 1) = F(\epsilon_{iz}^{\text{min}}(t)) = 1 - \exp \left\{ - \exp \left\{ \frac{V_z(t) - (\eta - \mu \ln(n_z))}{\mu} \right\} \right\}
$$

$$
= 1 - \exp \{ - \exp \{V_z(t) + \ln(n_z)\}\},
$$

(8)

where the second line follows from the standard normalizations for the location and scale parameters, $\eta$ and $\mu$ respectively. $\ln(n_z)$ is an “offset” factor that controls for the fact trial is more likely to be observed earlier in regions containing more individuals. In practical empirical terms $n_z$ is just the region population which can easily be obtained from the census. Furthermore, the inclusion of $\ln(n_z)$ in the probability expression is not arbitrary but arises from a specific model of individual behavior.

Hence, our solution to the absence of individual-level information for all potential triers is to: (1) retain an individual-level model of behavior but work at a higher level of aggregation (region) where we not only have extensive covariates but also can bound the risk set, and (2) make regions comparable by exploiting our assumption about the individual utility distribution. Our procedure is quite general and could be applied in any circumstance where the researcher does not have detailed information about individuals, but can count individuals and obtain covariate data for a higher level of aggregation.
Accounting for Region Level Heterogeneity

The derivation that led to the offset parameter $\ln(n_z)$ is important because it preserves an individual-level behavioral interpretation even though the model will be estimated using region level covariate information. In addition, it serves a key statistical purpose because it implies the need for region-level variation in the baseline hazard. To see this, assume that the region-level utility $V_z(t)$ (not including the offset) is equal to $\alpha_z + \beta X_z(t)$, where $X_z(t)$ contains region and time-varying covariates to be specified in the subsequent sections. We have

$$V_z(t) = \alpha_z + \beta X_z(t) + \ln(n_z)$$

$$= \alpha_0 + \beta X_z(t) + \ln(n_z) + (\alpha_z - \alpha_0)$$

$$= \alpha_0 + \beta X_z(t) + \phi \ln(n_z)$$

where $\phi = \frac{\ln(n_z) + (\alpha_z - \alpha_0)}{\ln(n_z)}$ (10)

Hence, when pooling data across regions imposing the theoretical constraint $\phi = 1$ required by equation (8) is equivalent to assuming that $\alpha_z = \alpha_0$ (in the absence of an additional random term in equation 9). In our setting this assumption is unlikely to be true empirically and we therefore allow a free parameter for the offset term and model the intercept as a random effect. The deterministic utility becomes

$$V_z(t) = \alpha_z + \beta X_z(t) + \phi \ln(n_z), \quad \alpha_z = \alpha_0 + \nu_z \quad \nu_z \sim N(0, \sigma^2).$$

(11)

In summary, the model exhibits the appealing property that the approach to accounting for heterogeneity falls naturally out of the analytical derivation, which in turn follows directly from an underlying behavioral model.\footnote{We also estimate $V_z(t) = \alpha_z + \beta X_z(t) + \ln(n_z), \quad \alpha_z = \alpha_0 + \eta_z \quad \eta_z \sim N(0, \sigma^2)$. Results are provided and discussed in the next section.}
Neighborhood Effects

Consider the possibility that in a region where no individual has yet tried, there is the potential for either direct communication with an individual from an adjacent region who has tried, or an ability to observe the behavior of these adjacent individuals. As an illustration, consider two adjacent regions, $z_1$ and $z_2$ and imagine trial occurs in region $z_1$ at $t - 1$. If individuals in $z_2$ gain knowledge of the event $\{y_{z_1}(t - 1) = 1\}$, this may lead to a change in the conditional probability of trial in $z_2$ where the conditioning is now on the prior event in $z_1$ such that

$$P(y_{z_2}(t) = 1|y_{z_1}(t - 1) = 1) \neq P(y_{z_2}(t) = 1|y_{z_1}(t - 1) = 0).$$

(12)

In the model, this notion can be reflected by a straightforward addition to the deterministic utility, $V_z(t)$

$$V'_z(t) = V_z(t) + \theta g(\{y_w(u)\}, w = 1, \ldots, W_z; \ u = 1, \ldots, t - 1),$$

(13)

where $w$ indexes the neighbors of focal region $z$ and $g(\cdot)$ is a function of past behavior by these neighbors. $\theta$ captures the effect of neighbor behavior on current utility of trial in the focal region. In the absence of interaction between proximate triers and current non-triers, the probability of the first trial occurring in the focal region depends only on region characteristics (which are constant over time) and the passage of time (e.g. as new information about the service becomes available), such that $\theta = 0$.

It is important to note that $g(\cdot)$ in equation (13) is not arbitrary and indeed one must impose conditions if the function is to be consistent with rational utility maximizing behavior on the part of individual decision makers (Brock and Durlaf 2001). In particular, $g(\cdot)$ specifies the relationship among regions considered to be part of a neighborhood or group, and assigns weights to the influence of the behavior of group members on the current utility of the focal

---

8As noted earlier, we do not distinguish between the two. Passive observation is facilitated by individuals observing deliveries to houses and apartment buildings (each box is clearly marked with “netgrocer.com”).
region. In this paper we focus on a model of contiguity relationships among regions and assign an equal weighting across neighbors in their influence on decision making in the focal region (see Anselin 1988). Similar methods for capturing spatial interaction among decision makers have been applied in the marketing literature. Bronnenberg and Mela (2004) find that manufacturers enter markets based on spatial proximity to markets already entered and Garber et al (2004) show the importance of spatial effects in determining new product success.

An example serves to clarify the structure of the relationships and the empirical specification of the neighborhood effect. Figure 2 illustrates a set of four regions \{z_1, z_2, z_3, z_4\} from which one can derive a corresponding first-order contiguity matrix \(C\). The elements of the contiguity matrix are binary indicators of contiguity, which in the equal weighting scheme are further normalized such that the rows of this matrix sum to one.

Each \((4 \times 1)\) row vector \(c_z\) corresponds to the influence relationship between potential neighbors and the focal region \(z\). The collective influence of neighbors on \(z\) is obtained by post-multiplication of \(c_z\) by \(Y_z(t-1)\) the column vector of neighbor behaviors summarized at time \(t-1\), resulting in a scalar variable bounded between zero and one.\[^9\] Under first-order contiguity and equal weighting of prior neighbor behaviors, equation (13) becomes

\[
V_z'(t) = V_z(t) + \theta c_z Y_z(t-1) \text{ where } \\
(14)
\]

\[
c_z Y_z(t-1) = \sum_j^{n_z} c_{zj} \cdot Y_{zj}(t-1).
\]

\[^9\]In this first-order contiguity model neighbor relationships do not change over time so that all time variation in the measure is induced by changes in neighbor status with respect to trial as captured in \(Y_z(t-1)\).
over the \( n_z \) neighbors in the set (if another region is not a neighbor the contiguity relationship is zero by definition).

Our substantive interest is in the sign of \( \theta \) and the implied marginal effect of the social contagion variable. It is therefore important to ensure that neighborhood effects can be correctly identified. A comprehensive treatment of this point ("the reflection problem") is given in Manski (1993) who distinguishes endogenous from contextual and correlated effects. In the context of this study a true endogenous neighborhood effect exists if, all else equal, the probability of trial for a focal region varies with a measure of the average probability of trial for the reference group.\(^\text{10}\) The first step in identifying the neighborhood effect is exogenous determination of reference groups — a condition met in this study through the use of geographical contiguity and lagged actions. Specification of the full set of covariates embedded in the deterministic component of utility is deferred to the next section.

4 Data and Preliminary Analysis

We examine evidence for neighborhood effects and estimate the model parameters using data drawn from three separate sources: (1) netgrocer.com transaction files, (2) the United States census, and (3) CACI (caci.com) retailing statistics. These separate datasets are linked via the common zip code variable and the unit of analysis is time to first trial in each zip code area. The composition of the datasets and some descriptive analysis of the basic consumer behavior processes being modeled are now described.\(^\text{11}\)

\(^{10}\)A contextual effect exists if probability of trial varies according to the socio-economic characteristics of the reference group. A correlated effect is present if individuals in the same location behave similarly as a result of selection or exposure to similar environmental stimuli (retail stores, etc.).

\(^{11}\)Given the size and unique nature of the data, this process is necessarily rather extensive. All summary analyses are available from the author upon request.
Raw Data

(1) Transaction Data. The website http://www.netgrocer.com was launched on May 7, 1997. Between that time and January 31, 2001 the site had a total of 382,478 transactions. These data are a complete and exhaustive list of all orders filled at the site during this time period (i.e., we observe the order process from its inception so there is no left censoring). The 382,478 orders were placed by 162,618 different customers and shipped to 19,418 unique zip codes. The average order value during this time period across all customers is $57.53 (standard deviation = $50.99) which is considerably larger than the average order of $26.26 (standard deviation = $29.18) for purchases in traditional grocery stores.

Netgrocer continues to evolve, however it is important to understand key aspects of the business model for the period covered by the data. During this time Netgrocer offered shipments of non-perishable grocery products throughout the United States. Netgrocer’s stated goal is to provide a service that supplements the services of traditional supermarkets. Ideally, Netgrocer customers continue to patronize local stores for perishable or fresh products and fill part or all of their non-perishable requirements at Netgrocer. Shipping infrastructure was provided by Federal Express and offered at a standard rate of $6.99 per order.

Each transaction at the site is described by four fields: (1) date, (2) customer identification number, (3) total dollar value, and (4) zip code where the order was shipped. Before merging these data with census and CACI information, some initial pruning is in order. Both

---

12While Netgrocer continues as a going concern, we do not have access to data after January 2001.
13This figure is based on 161,778 shopping trips taken by the 1,042 consumers contained in the Stanford Market Basket Database. These data cover June 1991-June 1993 so the inflation-adjusted average order value for the period of the Internet data is approximately $29.80.
14From time to time Netgrocer ran specials to entice new customers, or encourage customers to purchase larger orders. These specials included offers such as “free shipping.” We are unable to separately account for orders that might have resulted from such promotions. Furthermore, subsequent to the data collection window Netgrocer management introduced a non-linear shipping cost schedule. This schedule imposes shipping fees differentiated by order size and region of the country (larger orders to western states are more expensive).
the census data and a second collection of records provided by ESRI (esri.com) indicate that there are 29,701 residential zip codes in the United States and it is these zip codes that we focus on in this research.\textsuperscript{15} By the end of the data period, 17,910 of these zip codes had seen at least one order, while the remaining 11,791 had not: Netgrocer had achieved a zip code trial rate equal to sixty percent. After removing the data for non-residential zip codes 369,146 orders remained for subsequent analysis.

\textit{(2) Census Data}. The 2000 United States census provides information on several zip code characteristics. From this we created three categories of data useful as model covariates. While this process is necessarily a matter of judgment, it was performed with reference to prior literature on the compilation of socio-demographic information (e.g., Dhar and Hoch 1997). The purpose is to summarize the profiles of individuals living in each zip code and the conditions under which they live. The categories of covariate are:

1. \textit{Household Characteristics}: Ethnicity, Gender, Family Size.


For each variable there is one observation per residential zip code. Covariate values are typically expressed as percentages, for example, the variable “Elderly” is defined as “the percentage of the zip code population that is aged 65 and above” and “College” represents “the percentage of individuals with bachelors and/or graduate or professional degrees.” Defining the variables in this way induces greater variation across observational units and is consistent with the approach advocated and taken in Dhar and Hoch (1997). Table 1 lists the variables and associated descriptive statistics.

\textsuperscript{15}Netgrocer had also shipped orders to 1,508 non-residential zip codes which were predominantly Army Post Office (APO) addresses. These zip codes generated a total of 25,123 orders which we have excluded from the database. The average dollar value of these orders is $69.12. More detailed information is available from the authors upon request.
(3) Retail Competition Data. Headquartered in Arlington Virginia, CACI is a premier collector and disseminator of commercial information and is the source of retail competition data for this study. While nationwide distribution coverage gives Netgrocer access to a vast potential customer base, it also exposes them to literally thousands of retail competitors. Individuals in each zip code can continue to shop at local stores, hence it is important that the model include information on the availability of this outside or status quo option. The CACI database reports the number of outlets and the average sales volume ($000) for five classes of retailer (convenience stores, drug stores, general merchandisers, supermarkets and warehouse clubs) for each of the 29,701 residential zip codes. From this information we derive a measure of the maximum expected distance an individual within the zip code must travel to reach a particular class of store. Using information on the zip code land area (Table 1) and assuming it is approximately rectangular, we can compute the length of the hypotenuse. One measure of the expected maximum distance to a store is the length of this hypotenuse divided by the number of stores plus one.

Preliminary Analysis

Prior to examining the model results, it is useful to first gain a sense of space-time evolution of trial behavior and examine preliminary evidence for neighborhood effects.

Temporal Patterns. The transaction data are first organized into a data matrix with 29,701 rows (the number of zip codes) and 45 columns (the number of months from May
1997 through January 2001). This results in 180,634 unique observations of zip-month combinations in which orders were observed. Panel (a) of Figure 3 shows that the number of zip codes where trial occurred for the first time rose through August 1998 (where it first peaked) and subsequently declined before continuing to rise again. Panel (b) plots the empirical hazard, i.e., the proportion of zip codes where trial of Netgrocer occurred for the first time from among those where trial had not yet occurred at time \( t \) (the risk set). The modeling implication of both plots is that it will be important to control for time variation in the baseline hazard.

Spatial-Temporal Patterns. By the end of May 1997 (the first month of operation) trial had occurred in thirty-four distinct zip codes ranging from New Jersey to California. Figure 4, panels (a) through (d) show the cumulative space-time trial patterns at one year intervals. As noted previously, by January 31, 2001 at least one order had been placed from approximately sixty percent of all residential zip codes. Figure 4 indicates the pool of triers expanded rapidly throughout the country in a relatively short amount of time. These data clearly demonstrate the most dramatic difference between a traditional retailer (where new and existing customers arrive from a small circumscribed area close to the store) and an Internet retailer.

Closer (i.e., more disaggregate) visual inspection of the patterns of evolution raises the possibility that neighborhood effects play a role in this process. Trial at time \( t \) does not occur randomly in space — rather, locations of new trials appear more likely to be located
near contiguous neighborhoods who have experienced trial prior to \( t \). Figure 5 shows trial evolution in rolling three-month increments for two separate east and west coast snapshots of the United States. Visual inspection of these areas shows that as time moves along new trials are more likely to arise close to contiguous areas of prior trial.\(^{16}\)

---

**Neighbors.** A contiguous neighbor \( j \) is a zip code that shares an adjoining boundary with the focal zip code \( z \) and information on the presence of contiguous neighbors was obtained for each of the 29,701 zip codes. The set of all such neighbors for \( z \) is defined as \( \{N_j\}_z \ j = 1, \ldots n_z \) where \( n_z \) is the total number of neighbors as defined earlier. In the data, the average number of regions contained in the neighborhood set is 5.63 (standard deviation = 2.28). Almost all zip codes have at least one neighbor, however there are 136 “islands” who have no direct contiguous neighbors.

5 Empirical Findings

In reporting estimation results from the discrete time hazard model of trial, particular emphasis is given to demonstrating the neighborhood effect is properly identified and robust in the presence of controls for heterogeneity, selection and unobservables. Before exploring the model findings we give a brief summary of the estimation procedure as it involves a specific reorganization of the data matrix.

---

\(^{16}\)This general visual pattern is representative of other months and regions, however in the interests of brevity such figures are not shown but are available from the authors upon request.
Estimation

The discrete time model is implemented without loss of information as we choose the complementary log-log specification to mimic an underlying continuous time process. We exploit the fact that the complementary log-log function is the inverse of the Gumbel cumulative distribution function (see Allison 1982; Maritz and Munro 1976). Working from the earlier derivation and substituting from equation (8) we obtain

\[
\log \left[ -\log(1 - P_{iz}(t)) \right] = \log \left\{ \exp \left[ \frac{V_z(t) - (\eta + \mu \ln(n_z))}{\mu} \right] \right\} \\
= V_z(t) + \ln(n_z) \quad \text{(15)}
\]

In other words, the right hand side is equivalent to the deterministic component of utility. Exact specifications for \( V_z(t) \) rely on covariates introduced in Tables 1 and 2.\[^{17}\]

In order to estimate the discrete time model, the following reorganization of the data is necessary. Continuing with prior notation \( y_z(t) = 1 \) indicates that trial occurred in zip code \( z \) at time \( t \). Denote the time at which \( y_z(t) = 1 \) as \( T_z \). It follows that the number of observations zip code \( z \) contributes for estimation is therefore \( T_z \) with the dependent variable \( y_z(s) \) set equal to zero for all periods \( s < T_z \). For the 11,791 zip codes in which trial is never observed \( T_z = 45 \) — the number of months from May 1997 through January 2001 inclusive. The total number of stacked observations \( \sum_z T_z = 910,769 \). Estimation of the parameters of the discrete time hazard is accomplished via binary choice analysis on this stacked data set, using the complementary log-log link function. This follows from the factorization of hazard probabilities in the likelihood, specifically

\[
P(T_z = x) = P_z(x)[1 - P_z(x - 1)] \ldots [1 - P_z(1)] \quad \text{where}

P_z(t) = 1 - \exp \left\{ -\exp^{V_z(t)} \right\} \quad \text{follows from equations (8) and (15)}
\]

\[^{17}\]While the choice of a complementary log-log function follows from our theoretical derivation, for completeness we also estimated results using a logit specification. The substantive results are unchanged and available from the authors upon request.
and \( \prod_{z=1}^{Z} P(T_z = x) = \prod_{z=1}^{Z} \prod_{t=1}^{x} P_z(t)^{y_z(t)} \cdot [1 - P_z(t)]^{1-y_z(t)}. \)

Note that this factoring implies there is no time dependence within region provided each region has no more than one event (a condition satisfied by our study).

**Initial Evidence for Neighborhood Effects**

We first test for the presence of social contagion in trial with two models in which the neighborhood effect \((\theta)\) and the population offset \((\phi)\) are the sole covariates

- A lagged cumulative effect (LC) with individual elements of the column vector for neighbor behavior \(Y_z(t-1)\) equal to one if the relevant neighbor of \(z\) experienced trial at any point up to and including \(t-1\), and zero otherwise.
- A standard lagged effect (L) with individual elements of \(Y_z(t-1)\) equal to one if the relevant neighbor of \(z\) experienced trial at \(t-1\), and zero otherwise.

We settle on lagged formulations because contemporaneous representations violate the rationality conditions discussed previously, and the associated parameters are not theoretically estimable using maximum likelihood methods. Table 3 presents the parameter estimates for \(\theta\), model fits and Wald \(\chi^2\) statistics. Both models show significant neighborhood effects, however no attempt has been made to control for unobserved common traits or heterogeneity.

[ Table 3 about here ]

---

**Unobserved Common Traits.** If it is the case that individuals in contiguous regions share unobserved common traits that are positively correlated with the utility of trial, this will cause an upwards bias in any estimate of the neighborhood effect. One such example in the current context is the level of technological sophistication or Internet penetration. It is likely that this variable takes a similar value across contiguous neighbors. While one
could make analogous claims regarding other unobservables, it would be difficult to accept that individuals self select locations in which to live on the basis of their preferences for Netgrocer. Three approaches are taken in unison in order to mitigate the potential upwards bias accruing as a result of unobservables.

First, lagged cumulative trial (LC) is adopted as the appropriate empirical formulation of the neighborhood effect. Structuring the independent variable in this way ensures that neighboring regions that have the potential to exert influence are demonstrably different from the focal region (i.e., they have already experienced trial of the innovation). Isolating the focal region where trial has not been experienced from contiguous neighbors with prior trial helps mitigate any potential bias (a similar strategy is employed and advocated by Goolsbee and Klenow (2002) in their study of household-level diffusion of personal computers).

Second, a comprehensive set of covariates is developed from the variables presented in Tables 1 and 2. Collectively, these variables serve to control for observable differences across regions with respect to: (1) intrinsic characteristics of the population, (2) socio-economic status, (3) physical environment, and (4) access to competing retail services. This serves to control for any omitted variable bias that would otherwise amplify the effect attributed to contagion when these variables are excluded from the model. All these observables are measured at the level of the region, \( z \), and therefore directly related to the dependent variable, \( y_z(t) \).

Third, fixed effects are added to account for all other sources of variation that are attributable to differences across states, which are higher order regions. One could potentially incorporate region-specific fixed effects, however this is complicated by virtue of the number of fixed effects (there are 29,701 regions) and the unbalanced design of the data matrix. Regions who experience trial in the first period contribute only one observation making identification difficult.
Unobserved Heterogeneity. If there is unobserved heterogeneity in the baseline hazard with higher hazard regions experiencing trial earlier, this will cause an attenuation of the neighborhood effect, such that the parameter estimate will tend to zero. It is important to note however that while the coefficient may be attenuated the associated standard errors are not biased (Gail, Weiand and Piantadosi 1984). As noted by Allison (2001) there are two effective ways to deal with this problem, specifically through estimation of a non-parametric baseline hazard which changes over time, or via a random effect on the model intercept to allow for differences across observational units. Following equation (11) we already model a random effect on the intercept, so we augment this with time-specific fixed effects. Given the number of observations in the study, the addition of monthly time-specific fixed effects is not problematic.

More Evidence for Neighborhood Effects

Table 4 provides model fits and parameter estimates for \( \theta \) obtained after inclusion of the control procedures advocated above. Rows two and three report the benchmark fixed effects and non-parametric baseline hazard model fits. In all subsequent cases the neighborhood effect \( (\theta) \) enters the model and is formulated according to the Lagged Cumulative (LC) specification. Specifications 4 and 5 show that the neighborhood effect is robust to the separate inclusion of state fixed effects and a non-parametric baseline, however the estimates of \( \theta \) are lower than the value of 1.843 shown in Table 3 where the neighborhood effect and the population control variable \( \log(n_x) \) are the sole covariates.

[ Table 4 about here ]

Inclusion of a broad set of region characteristics from Table 1 and the retail environment variables from Table 2 again reduces the magnitude of \( \theta \), and produces a corre-
sponding improvements in model fit. The last row of Table 4 reports the estimate from
the final and most general model, and shows that $\theta$ is still highly statistically significant
($\theta = 0.4794$, Wald $\chi^2 = 200.5$). Collectively, these results imply that $\theta$ is capturing a real
behavioral phenomenon. To the extent that unobserved common traits, unobserved hetero-
geneity in the baseline hazard and observed variation across observational units have been
properly accounted for there is evidence of social contagion in trial for the Internet retailing
innovation studied here. In fact, inspection of the complete list of parameter estimates from
the final model (see Table 5) shows that the neighborhood effect is second only to “Percentage
of College Educated Households” (College) in its level of statistical significance. The
standard deviation of the random effect (equation 11) is not significantly different from zero
($\sigma = 0.0009$, LR test: $\rho = 0$ yields $\chi^2 = 0.000$, $p = 1.0$). This suggests a model with a free
population offset parameter ($\phi$), that incorporates a non-parametric baseline hazard coupled
with state-specific fixed effects fully accounts for sources of unobserved heterogeneity in the
base hazard rate.

**Alternative Models**

As a robustness check, we estimated three alternative but conceptually similar models. The
first is the model discussed earlier which imposes the constraint $\phi = 1$, but still allows for a
random effect on the intercept. The neighborhood effect remains ($\theta = 0.5829$, Wald $\chi^2 =
215.2$), but the standard deviation of the random effect is now highly significant as expected
($\sigma = 0.6247$, LR test: $\rho = 0$ returns $\chi^2 = 262.7$, $p < 0.001$). The model presented in Table
5 is however preferred on the basis of model fit ($-2 \log L = 142,797.4$ versus 142,861.0,
$\chi^2 = 63.6$, $p < 0.001$).

The remaining two models define $n_z$ as the number of households in the region, rather
than the number of individuals (one could reasonably argue that the decision making unit
for trial is in fact the household). When $\phi$ is a free parameter the neighborhood effect is
again significant ($\theta = 0.5195$ Wald $\chi^2 = 236.5$) and the random effect on the intercept is not ($\sigma = 0.0009$, LR test: $\rho = 0$ yields $\chi^2_1 = 0.000$, $p = 1.0$). With $\phi = 1$ we find $\theta = 0.6377$ (Wald $\chi^2 = 242.7$) and the random effect on the intercept is again highly significant ($\sigma = 0.6931$, LR test: $\rho = 0$ returns $\chi^2_1 = 305.0$, $p < 0.001$). The fits (-2 Log L) of these two alternative models are 143,418.6 and 143,817.0, respectively, which are inferior to the fit for the model presented in Table 5.\(^{18}\) In summary, the pattern of results support the conjecture that neighborhood effects are present, and that the individual is the appropriate decision-making unit. Moreover, the qualitative pattern of parameter estimates for other model covariates is consistent across all four specifications.

**The Effect of Region Covariates on Trial**

The estimates in Table 5 allow us to draw conclusions about the effect of region characteristics on time to trial (a positive coefficient indicates that the covariate *speeds up* the time to trial). Fourteen of the twenty-three parameters are significantly different from zero ($p < 0.01$) and the implied marginal effects are intuitively reasonable. Our discussion of the effects follows the variable classifications given in Tables 1 and 2.

[Table 5 about here]

\(^{(1)}\) *Household Characteristics.* Regions with greater percentages of minorities (Blacks and Hispanics) are slower to experience initial trial. This is consistent with other evidence for so-called “digital divide” in which these groups have both less access to the Internet and lower usage given access (see for example, U.S. Department of Commerce annual studies *Falling Through the Net — Defining the Digital Divide*). The percentage of solo person

\(^{18}\)A further model allowed for (monthly) time variation in $\theta$. The neighborhood effect declined (but remained positive and significant) after March 2000, however this model is again rejected on the basis of fit. The improvement (-2 Log L = 142,722.4) comes at the expense of forty-four additional parameters.
households is also important, but this interacts with gender: regions containing greater proportions of male-only households more likely to see earlier trial. Conversely, an increase in the proportion of large (greater than five person) households slows time to first trial. These larger families may be less willing to split their shopping baskets across formats (Netgrocer does not sell perishables).

(2) Household Economics. Regions with a higher percentage of wealthy and tertiary-educated individuals experience trial earlier. Furthermore, an increase in the number of young wealthy individuals (Generation X) shows an additional positive effect and a higher percentages of elderly individuals slows time to trial \((p < 0.05)\). Other variables held constant, working status of the either the male or female in the household shows no effect. Collectively these variables may capture individuals’ desire to try new innovations and/or economize on time spent shopping.

(3) Local environment. Region size in terms of land area per se is not important, however the number of households, population density and the extent of urbanization are all critical. Each of these factors has a significant positive effect on the time to first trial. These variables may operate as proxies for the potential for social interaction within a region and within and between household members in a given region. Home sizes and values and not significant given that these other important factors have been controlled for.

(4) Access to Retail Services. The parameter estimates for expected maximum distance to convenience stores and general merchandisers suggest that neither type of format affects initial trials. Increasing the expected maximum travel distance to drug stores and supermarkets decreases time to first trial. This result can be interpreted in light of the differences in format between these types of retailer and netgrocer.com. As Netgrocer offers neither perishable products nor a full complement of drug store items a household using Netgrocer and requiring these products would still need to visit a supermarket or drug store. The more inconvenient the supermarket the more it makes sense for a rational household to amor-
tize the full cost of the trip and do one-stop shopping at these outlets, thus obviating the need to purchase non-perishables at Netgrocer. The parameter estimate is consistent with the notion that Netgrocer complements traditional supermarkets: the easier it is to access a supermarket, the more sense it makes to split the shopping basket for the purchase of perishables (supermarket) and non-perishables (Internet). Conversely, Netgrocer appears to compete more directly with warehouse clubs. The greater the distance to a warehouse club, the more likely an individual household will try Netgrocer.

Taken individually and as a group, the parameter estimates paint an intuitive and plausible description of how region characteristics influence trial of an Internet retailing service. It is important to note that all these effects — and the focal neighborhood effect $\theta$ — are statistically significant after having controlled for the most important characteristic of all, the number of individuals $n_z$ that reside in the region (Wald $\chi^2 = 2354.2$). Moreover the inclusion of $\log(n_z)$ in the model is not arbitrary but derived from the underlying assumptions on the distribution of individual-level utility and allows us couch the analysis in terms of the first trial in each region, and to pool data across regions by implicitly putting them on the same “scale”.

6 Discussion and Conclusion

According to Forrester Research online retailers accumulated close to $100$ billion in sales during 2003 (up from $78$ billion in 2002) and this number is expected to grow in 2004. Our study therefore addresses a new, important and unresearched phenomenon: space-time evolution of trial decisions for an Internet retailer. The customer acquisition process is clearly different from that seen for offline retailers and while much has been learned about traditional retailing, little is known about how Internet retailers develop a customer franchise. In this paper, we develop a theoretical rationale for the presence of social contagion or neighborhood effects in individual trial decisions for Internet retailers, along with a statistical approach to
test for the effect in the absence of strictly individual-level data. Specifically, we exploit the relationship between utility maximization and a discrete time hazard model to examine the first trial in a particular region. Our choice of the distributional assumption on utility allows us to combine data across regions by rescaling the latent utility of this maximal individual to account for the number of individuals present in the region.

**Substantive Findings and Implications**

Other studies in the economics and sociology literatures have argued for and demonstrated the existence of neighborhood effects in a number of diverse contexts, however they have not been shown to operate on the Internet. Some researchers have even speculated that the Internet may contribute to individuals becoming more diffuse and solitary in their behavior (e.g., Townsend 2001). Conversely, our empirical findings are consistent with the proposition that social interaction stimulates trial of a new Internet service. Our goal is to demonstrate the existence of the effect — future work remains to be done on the exact nature of the mechanism. We have modeled social contagion through a measure of physical contiguity but have not been able to explicitly separate passive observation and direct inter-personal interaction.

The estimate for $\theta$ given in Table 5 implies that when all contiguous neighbors of a focal region have already experienced trial, this leads to a sixty-three percent increase in the log odds of trial for the focal region. With a baseline probability equal to about two percent and an imputed marginal effect of social contagion equal to 0.009, this translates to a fifty percent increase in the baseline for the average region. One implication is that an Internet retailer could benefit significantly from a judicious seeding of the trial process.

The demise of other forms of Internet-based supermarket retailing (e.g., Webvan) has been attributed to lack of customer density and the corresponding burden placed on the delivery infrastructure (Deighton 2001). Our research suggests that even when shipping is
handled by third party specialists such as Federal Express customer density is still important because of the role of neighborhood effects in stimulating new trials. Finally, we show that region characteristics are likely to be managerially useful segmentation variables: speed to trial is strongly influenced by education levels, population density, extent of urbanization, access to retail services and household composition.

Future Research

While we demonstrate the existence of the neighborhood effect and quantify the implications for customer behavior, several issues remain. One could:

- Broaden the construct of affiliation beyond geographical contiguity with equal influence by all neighbors. Bronnenberg and Mahajan (2001) propose different types of influence model in their study of interaction among retailers in the adoption of new brands. Graff and Ashton (1993) analyze the pattern of store openings by Walmart and find a reverse hierarchical process (i.e., from rural to urban areas). Such approaches may also be useful in our setting.

- Attempt to distinguish between effects that arise from direct word of mouth and those attributable to observational learning. The power of word of mouth in television viewing trends has been demonstrated in Godes and Mayzlin (2004).

- Incorporate information on purchase volumes (Chandrashekaran and Sinha 1995) or subsequent behaviors such as repeat (Ubran 1975). A preliminary examination of our data shows that initial orders for individuals who go on to repeat are approximately forty percent larger than those for individuals who try, but do not repeat.

We intend to visit these issues in future research.
References


Table 1: Region (Zip Code) Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(1) Household Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blacks</td>
<td>% of Blacks</td>
<td>0.0725</td>
<td>0.1563</td>
</tr>
<tr>
<td>Foreign</td>
<td>% of foreign born individuals (aged 18+)</td>
<td>0.0434</td>
<td>0.0790</td>
</tr>
<tr>
<td>Hispanics</td>
<td>% of Hispanics</td>
<td>0.0459</td>
<td>0.1141</td>
</tr>
<tr>
<td>Large family</td>
<td>% of families with five or more members</td>
<td>0.1515</td>
<td>0.0607</td>
</tr>
<tr>
<td>Solo female</td>
<td>% of single female households</td>
<td>0.0477</td>
<td>0.0245</td>
</tr>
<tr>
<td>Solo male</td>
<td>% of single male households</td>
<td>0.0356</td>
<td>0.0202</td>
</tr>
<tr>
<td><strong>(2) Household Economics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>% with bachelors and/or graduate or professional degree</td>
<td>0.0984</td>
<td>0.0785</td>
</tr>
<tr>
<td>Elderly</td>
<td>% aged 65 and above</td>
<td>0.1371</td>
<td>0.0586</td>
</tr>
<tr>
<td>Fulltime female</td>
<td>% of households with full time female worker</td>
<td>0.2545</td>
<td>0.0839</td>
</tr>
<tr>
<td>Fulltime male</td>
<td>% of households with full time male worker</td>
<td>0.4850</td>
<td>0.1197</td>
</tr>
<tr>
<td>Generation X</td>
<td>% of individuals 25-34 with incomes over $50,000</td>
<td>0.0102</td>
<td>0.0116</td>
</tr>
<tr>
<td>Wealthy</td>
<td>% of households earning $75,000+</td>
<td>0.0660</td>
<td>0.0833</td>
</tr>
<tr>
<td><strong>(3) Local Environment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Density</td>
<td>population density</td>
<td>1108.0700</td>
<td>4270.6200</td>
</tr>
<tr>
<td>Home value</td>
<td>% of homes valued at $250,000 or more</td>
<td>0.0232</td>
<td>0.0782</td>
</tr>
<tr>
<td>Households</td>
<td>number of households</td>
<td>3095.4000</td>
<td>4415.5400</td>
</tr>
<tr>
<td>Land area</td>
<td>area in square miles</td>
<td>110.2122</td>
<td>387.1567</td>
</tr>
<tr>
<td>Large house</td>
<td>% of homes with five bedrooms or more</td>
<td>0.0339</td>
<td>0.0324</td>
</tr>
<tr>
<td>Population</td>
<td>total population</td>
<td>8372.6100</td>
<td>11867.6000</td>
</tr>
<tr>
<td>Urban housing</td>
<td>% of urban housing units</td>
<td>0.1098</td>
<td>0.1393</td>
</tr>
</tbody>
</table>
Table 2: Access to Retail Services

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convenience</td>
<td>Number of convenience stores</td>
<td>2.383</td>
<td>3.726</td>
</tr>
<tr>
<td>Drugstore</td>
<td>Number of drug stores</td>
<td>1.474</td>
<td>2.447</td>
</tr>
<tr>
<td>General</td>
<td>Number of general stores</td>
<td>1.633</td>
<td>3.099</td>
</tr>
<tr>
<td>Supermarket</td>
<td>Number of supermarkets</td>
<td>5.303</td>
<td>7.629</td>
</tr>
<tr>
<td>Warehouse</td>
<td>Number of warehouse clubs</td>
<td>0.021</td>
<td>0.151</td>
</tr>
<tr>
<td>Distance to convenience</td>
<td>Expected maximum distance to a convenience store</td>
<td>3.158</td>
<td>4.010</td>
</tr>
<tr>
<td>Distance to drug</td>
<td>Expected maximum distance to a drug store</td>
<td>3.910</td>
<td>4.394</td>
</tr>
<tr>
<td>Distance to general</td>
<td>Expected maximum distance to a general store</td>
<td>3.944</td>
<td>4.214</td>
</tr>
<tr>
<td>Distance to supermarket</td>
<td>Expected maximum distance to a supermarket</td>
<td>2.154</td>
<td>3.005</td>
</tr>
<tr>
<td>Distance to warehouse</td>
<td>Expected maximum distance to a convenience store</td>
<td>5.533</td>
<td>4.881</td>
</tr>
</tbody>
</table>

Table 3: Initial Evidence for Neighborhood Effects in Trial

<table>
<thead>
<tr>
<th>Formulation</th>
<th>Estimate of θ</th>
<th>Wald χ²</th>
<th>-2 Log(L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept Only</td>
<td>—</td>
<td>—</td>
<td>176,199.8</td>
</tr>
<tr>
<td>Lagged Cumulative (LC)</td>
<td>1.843</td>
<td>7,316.2</td>
<td>154,322.6</td>
</tr>
<tr>
<td>Lagged (L)</td>
<td>2.047</td>
<td>626.0</td>
<td>160,268.1</td>
</tr>
</tbody>
</table>
Table 4: More Evidence for Neighborhood Effects in Trial

<table>
<thead>
<tr>
<th>Formulation</th>
<th>Estimate of $\theta$</th>
<th>Wald $\chi^2$</th>
<th>-2 Log(L)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Benchmark Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Intercept Only</td>
<td>—</td>
<td>—</td>
<td>176,199.8</td>
</tr>
<tr>
<td>2. State Fixed Effects</td>
<td>—</td>
<td>—</td>
<td>159,266.5</td>
</tr>
<tr>
<td>3. Non-Parametric Baseline Hazard</td>
<td>—</td>
<td>—</td>
<td>149,947.3</td>
</tr>
<tr>
<td><strong>Models with Neighborhood Effect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Lagged Cumulative (LC) $\theta$ + State Fixed Effects</td>
<td>1.779</td>
<td>6,467.4</td>
<td>153,458.5</td>
</tr>
<tr>
<td>5. LC $\theta$ + Non-Parametric Baseline Hazard</td>
<td>1.253</td>
<td>2,083.9</td>
<td>147,957.3</td>
</tr>
<tr>
<td><strong>Models with Neighborhood Effect and Covariates</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. LC $\theta$ + Non-Parametric Baseline Hazard + Region Characteristics</td>
<td>0.859</td>
<td>808.8</td>
<td>144,351.0</td>
</tr>
<tr>
<td>(Table 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. LC $\theta$ + Non-Parametric Baseline Hazard + Region Characteristics</td>
<td>0.873</td>
<td>782.3</td>
<td>144,239.4</td>
</tr>
<tr>
<td>(Table 1) + Retail Access (Table 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. LC $\theta$ + Non-Parametric Baseline Hazard + Region Characteristics</td>
<td>0.584</td>
<td>307.0</td>
<td>143,255.9</td>
</tr>
<tr>
<td>(Table 1) + Retail Access (Table 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ State-Level Means of Covariates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. LC $\theta$ + Non-Parametric Baseline Hazard + Region Characteristics</td>
<td>0.479</td>
<td>200.5</td>
<td>142,797.4</td>
</tr>
<tr>
<td>(Table 1) + Retail Access (Table 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ State Fixed Effects</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5: The Effect of Region Characteristics on Trial Behavior

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model intercept $^1$ ($\alpha_0$)</td>
<td>-9.4629</td>
<td>0.2071</td>
<td>2088.5</td>
<td>0.000</td>
</tr>
<tr>
<td>Population control $\log(n_z)$ ($\phi$)</td>
<td>0.7221</td>
<td>0.0149</td>
<td>2354.2</td>
<td>0.000</td>
</tr>
<tr>
<td>Neighborhood effect ($\theta$)</td>
<td>0.4794</td>
<td>0.0339</td>
<td>200.5</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Region Level Covariates

(1) Household Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blacks</td>
<td>-0.6635</td>
<td>0.0684</td>
<td>94.1</td>
<td>0.000</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.1013</td>
<td>0.1663</td>
<td>0.4</td>
<td>0.542</td>
</tr>
<tr>
<td>Hispanics</td>
<td>-0.5403</td>
<td>0.1264</td>
<td>18.2</td>
<td>0.000</td>
</tr>
<tr>
<td>Large family</td>
<td>-2.3178</td>
<td>0.2443</td>
<td>90.1</td>
<td>0.000</td>
</tr>
<tr>
<td>Solo female</td>
<td>-3.7293</td>
<td>0.6745</td>
<td>30.6</td>
<td>0.000</td>
</tr>
<tr>
<td>Solo male</td>
<td>5.8246</td>
<td>0.5233</td>
<td>123.9</td>
<td>0.000</td>
</tr>
</tbody>
</table>

(2) Household Economics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>2.9313</td>
<td>0.2010</td>
<td>212.6</td>
<td>0.000</td>
</tr>
<tr>
<td>Elderly</td>
<td>-0.5482</td>
<td>0.2363</td>
<td>5.4</td>
<td>0.020</td>
</tr>
<tr>
<td>Fulltime female</td>
<td>0.2669</td>
<td>0.1373</td>
<td>2.4</td>
<td>0.124</td>
</tr>
<tr>
<td>Fulltime male</td>
<td>-0.1092</td>
<td>0.1212</td>
<td>0.8</td>
<td>0.368</td>
</tr>
<tr>
<td>Generation X</td>
<td>7.3326</td>
<td>1.0420</td>
<td>49.6</td>
<td>0.000</td>
</tr>
<tr>
<td>Wealthy</td>
<td>0.8178</td>
<td>0.2375</td>
<td>11.8</td>
<td>0.001</td>
</tr>
</tbody>
</table>

(3) Local Environment

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>0.0000</td>
<td>0.0000</td>
<td>6.7</td>
<td>0.009</td>
</tr>
<tr>
<td>Home value</td>
<td>-0.3144</td>
<td>0.1583</td>
<td>4.0</td>
<td>0.047</td>
</tr>
<tr>
<td>Households</td>
<td>0.0000</td>
<td>0.0000</td>
<td>17.5</td>
<td>0.000</td>
</tr>
<tr>
<td>Land area</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.4</td>
<td>0.527</td>
</tr>
<tr>
<td>Large house</td>
<td>0.1708</td>
<td>0.3831</td>
<td>0.2</td>
<td>0.656</td>
</tr>
<tr>
<td>Urban housing</td>
<td>0.3920</td>
<td>0.1021</td>
<td>14.7</td>
<td>0.000</td>
</tr>
</tbody>
</table>

(4) Access to Retail Services

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to convenience</td>
<td>0.0055</td>
<td>0.0030</td>
<td>3.4</td>
<td>0.066</td>
</tr>
<tr>
<td>Distance to drug</td>
<td>-0.0085</td>
<td>0.0026</td>
<td>10.8</td>
<td>0.001</td>
</tr>
<tr>
<td>Distance to general</td>
<td>-0.0035</td>
<td>0.0023</td>
<td>2.3</td>
<td>0.126</td>
</tr>
<tr>
<td>Distance to supermarket</td>
<td>-0.0262</td>
<td>0.0046</td>
<td>32.9</td>
<td>0.000</td>
</tr>
<tr>
<td>Distance to warehouse club</td>
<td>0.0110</td>
<td>0.0022</td>
<td>26.2</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$^1$ Estimate of standard deviation of the random effect, $\sigma$ from equation (11) is equal to 0.0009. $\rho = \frac{\sigma^2}{\epsilon^2 + \sigma^2}$, LR test: $\rho = 0$ yields $\chi^2_1 = 0.000$, $p = 1.0$. 
Figure 1 (a) Total Trial Revenue By State

Figure 1 (b) Average Trial Order By State
Figure 2 First Order Contiguity Relationships

\[ Y_z(t-1) = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \]

\[
C = \begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
0 & 1 & 1 & 0
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
0 & 1/2 & 1/2 & 0 \\
1/3 & 0 & 1/3 & 1/3 \\
1/3 & 1/3 & 0 & 1/3 \\
0 & 1/2 & 1/2 & 0
\end{pmatrix}
\]
Figure 3 (a) Netgrocer Trials by zip code

Figure 3 (b) Empirical Hazard
Figure 4 Space-Time Evolution of Trial

(a) Trial as of December 1997

(b) Trial as of December 1998

(c) Trial as of December 1999

(d) Trial as of December 2000
Figure 5 Trial and Neighborhood Effects