Quant Market Econ (2007) 5:361–400
DOI 10.1007/s11129-007-9025-5

Neighborhood effects and trial on the Internet: Evidence from online grocery retailing

David R. Bell · Sangyoung Song

Received: 21 March 2006 / Accepted: 8 March 2007 /
Published online: 23 August 2007

Abstract For traditional retailers the customer pool is largely bounded in space, whereas an Internet retailer can obtain customers from a wide geographical area. We examine customer trials at Netgrocer.com, and drawing on studies in marketing and economics conjecture that exposure spatially to proximate others (through direct social interaction or observation), can influence decisions of those who have yet to try. Trials arise from utility-maximizing behavior and the model is estimated as a discrete time hazard. The data span: (1) 29,701 residential zip codes, (2) 45 months of transactions since inception, and (3) zip code contiguity relationships. The estimated neighborhood effect is significantly positive and economically meaningful.

Keywords Discrete time hazard · Neighborhood effect · Random utility · Retailing

JEL Classification C25 · M30

D. R. Bell (✉)
The Wharton School, University of Pennsylvania,
700 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, USA
e-mail: davidb@wharton.upenn.edu

S. Song
The Zicklin School of Business, Baruch College,
1 Bernard Baruch Way, New York, NY 10010, USA
e-mail: sangyoung_song@baruch.cuny.edu

> "... 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 substantiates its importance. While pricing and assortment are critical as well location accounts for the most variation in outlet choice in many retail settings.[1] For an e-tailer, physical location of the store relative to potential customers is no consequence. Indeed, the trading area of the e-retailer is constrained only by the availability of shipment infrastructure for distributing orders. The location of existing *customers* relative to potential customers and *their* interactions, may however be critical.

An e-tailer's unique market context—geographically dispersed customers and competitors—raises important (and thus far unstudied) questions about the evolution of the customer base. The role of existing customers in recruiting or influencing new potential customers is especially fundamental. Emulation in decision making has been studied in theoretical and empirical research in economics and sociology (e.g., Burt 1980; Goolsbee and Klenow 2002; Tolnay et al. 1996; Van den Bulte and Lilien 2001) and is the focus of our research. We are uniquely positioned to address this issue in an e-tail setting through the space-time evolution of trials from the inception of a new Internet grocery retailer. A descriptive characterization of the data motivates the modeling framework and underlying theory. Figure 1 summarizes trial orders for Netgrocer with total revenue earned and average order value by state shown in panels (a) and (b), respectively.

The empirical distribution of these two variables is broken into quintiles.[2] 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. Population size is a likely explanation for the first observation, while the second may result from greater travel distances to retail services. The important fact is that the customer base spans the entire United States. The data in Fig. 1 are cumulative

[1] For example, *Progressive Grocer* (April 1995) reports that location explains up to 70% of the variance in consumer choice of supermarket retailers. 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).

[2] For reasons of confidentiality we have excluded the dollar values from panel (a), however all 48 contiguous states generate revenue.

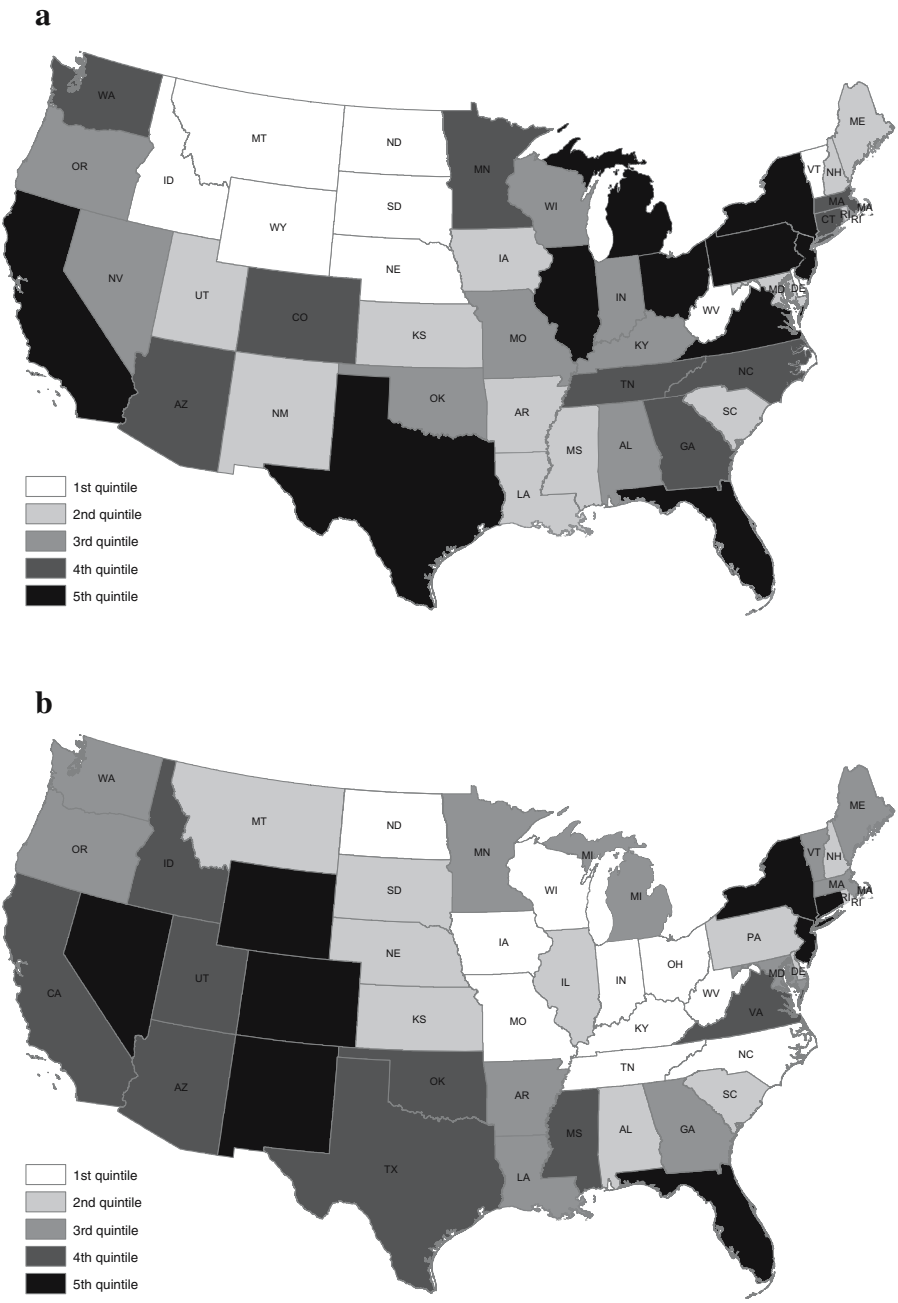


Fig. 1 **a** Total trial revenue by state **b** Average trial by state

from the inception of Netgrocer in May 1997 through January 2001. Orders were, and still are, shipped via Federal Express from a company warehouse in New Jersey.

For the remainder of the paper we focus on the spread of initial orders over time and space, with special emphasis on social influence or "neighborhood effects" in speeding up or inhibiting the process.[3] Studies in economics and sociology (e.g., Bikhchandani et al. 1998; Case 1991; Case et al. 1989; Singer and Spilerman 1983) motivate our representation of neighborhood effects, however marketing researchers are also beginning to analyze spatial aspects of diffusion processes (see for example, Bronnenberg and Mela 2004; Garber et al. 2004). We propose and estimate a model in which trial decisions result from utility-maximizing behavior. Trial is observed when an individual-specific threshold for action is exceeded. The advantage of this conceptualization is that a time-dependent process can be examined through a sequence of binary actions.[4]

Contribution and caveats

We demonstrate empirically the importance of neighborhood effects in generating trial at an Internet grocery retailer. The substantive message is in line with Goolsbee and Klenow (2002) who 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, given simple representations of influence derived from physical proximity. The estimated effect is economically and statistically important and is robust to controls for region and time-specific fixed effects, region covariates, unobserved heterogeneity in the baseline hazard, and alternative specifications for access to the Internet. We contribute the following:

- First, we develop a framework for empirical analysis of a new phenomenon in retailing, namely the evolution of customer trials for an e-tailer. In so doing, we provide insight into the consequences of spatially dispersed customers and competitors.
- Second, we offer an analytical derivation to estimate parameters of an inherently individual-level decision process using region-level data. The relationship between random utility maximization and a discrete time hazard model coupled with knowledge of the number of individuals in each region accomplishes this. Our approach avoids individual-level covariates, unrealistic assumptions about right censoring, and problems in exogenously defining neighbor relationships.
- Third, we find that neighborhood effects influence the "private behavior" of e-tailer trial, and are economically important. Moreover, the space-time

[3]The former term is preferred by sociologists and the latter by economists, but have complementary interpretations: Neighborhood effects emanate from the influence of well-defined exogenous groups on a focal group, whereas social influence refers to the broader behavioral process. We focus on an empirically grounded neighborhood effect without speculation as to the exact mechanism.

[4]Specifically, a discrete time hazard model estimated on time-dependent trial data is consistent with random utility maximization over binary outcomes.

trial pattern for an e-tailer is strongly related to local conditions (e.g., population characteristics, physical environment, etc.).

The estimated neighborhood effect is grounded in theory and economically meaningful. Our analysis is facilitated through a new variant of the discrete time model which allows individual level behavior to feed naturally into a representation based on region level data. Insights into the behavioral process are obtained, however we do not claim a complete elaboration of all nuances of the effect, nor do we fully articulate the exact nature of the mechanism as the data prohibit a distinction between influence through direct social interaction and through observation alone. Such pursuits are left to future research.

The paper is organized as follows. The next section briefly reviews related literature. The following section develops the statistical model in a random utility setting in which the neighborhood effect is both correctly specified and identified. Subsequent sections introduce the data and accompanying exploratory analysis, provide the estimation results and summarize the empirical findings. The paper concludes with some implications for e-tailing practice and ongoing research on social influence and neighborhood effects.

1 Background and motivation

Yang and Allenby (2003, p. 282) point out that "Quantitative models of consumer purchase behavior often do not recognize that preferences and choices are interdependent" and propose and estimate a model of automobile purchase that incorporates demographic and geographic proximity among choosers. They find strong evidence for interdependence; moreover, geographic proximity is somewhat more important than demographic similarity in explaining auto choice. Other recent work in marketing (e.g., Bronnenberg and Mela 2004; Godes and Mayzlin 2004) investigates interdependence in retailer brand adoption decisions and consumer opinions on television programming, respectively. Collectively, these studies motivate and substantiate our interest in interdependence. They also inform the specification of the empirical models (points of similarity and departure are discussed subsequently). While the empirical study of interdependence is relatively new to marketing, it has a longer history in economics and sociology—findings relevant to the current study are reviewed next.

1.1 Neighborhood effects: selected evidence

Economic analysis of neighborhood effects shows that in relatively closed communities individual knowledge can be aggregated to create a public good. 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. Besley and Case (1993) describe models which accommodate the updating rules of individual agents when they are exposed

to knowledge transmission by others. Case et al. (1993) find 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 by the focal region). Moreover, failure to account for this in estimation leads to an upwards bias in other model parameters.

Sociologists have also contributed a number of insights. Many studies focus on social connectedness and the extent of information transfer among and between groups of individuals (see Burt 1980 for a comprehensive treatment; also Greve et al. 1995; Strang and Tuma 1993). While social contagion can be grounded primarily in geographical contiguity, sociologists have also examined the structure of interpersonal affiliation. Burt (1980) discusses social cohesion and related studies describe how an affiliation matrix can be constructed to capture the nature, strength and timing of interaction within groups. Chaves (1996) for example, shows that the diffusion of gender equality in churches is influenced by cultural boundaries and affiliations within the denominations studied.

The power of social contact in information dissemination is demonstrated in a clever study by Oyen and De Fleur (1953). In a field experiment 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 important to keep in mind that observational learning can induce both negative and positive dispositions with respect to the innovation. Tolnay et al. (1996) study state-tolerated racist violence in the US 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.

1.2 Neighborhood effects: conceptualization and measurement

A longstanding tradition in marketing posits a generic consumer decision making process in which an individual passes through discrete stages in an approximately linear fashion. Lilien et al. (1993, p. 26) describe a five stage process: need arousal, information search, evaluation, purchase and post purchase. Each stage differs with respect to sources and use of information, time taken, and decision rules invoked and applied.[5]

Our model focuses exclusively on a single stage (trial) and incorporates important elements suggested by prior literature. Trials 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. To properly investigate how the trial behavior of spatially proximate neighbors

[5]This conceptualization can be traced back to early work by Howard and Sheth (1969).

affects current non-triers we require exogenous definitions of groups and neighborhood relationships.[6] In our application neighborhood relationships are known at the level of the region (zip code) but *not* at the level of the individual. We know the exact spatial proximity of different zip codes, but nothing about relative locations of individuals residing in the same zip code. Moreover, no individual-level covariate information is available for either triers or non-triers. Thus, the pattern of social influence, or neighborhood effect, will be specified empirically as a region-to-region phenomenon. We do not however ignore the potential for influence that occurs among individuals who share a specific region, rather we separate out the *within* and *across* region possibilities for transmission of information and emulation of triers by non-triers. The motivation for this conceptualization stems from the institutional setting, the data, and from the studies referenced above. The next section presents a model based on random utility maximization that incorporates all these elements in an integrated way.

In summary, theories in economics and sociology motivate why new triers of an Internet service could be influenced by existing users who are spatially proximate. In addition, recent empirical work in marketing highlights the presence of neighborhood effects in a variety of contexts (brand adoption by retailers, auto purchases by consumers, television viewership). Our study extends the set of contexts to include the Internet. Moreover, we exploit the relationship between random utility models and discrete time hazard models to offer a method that allows examination of neighborhood effects in the absence of strictly individual level data. Neighborhood effects are modeled as a direct covariate (Bronnenberg and Mela 2004; Goolsbee and Klenow 2002), rather than through an auto-regressive error structure (Yang and Allenby 2003).

2 Empirical model

We motivate the statistical model by highlighting the link between individual utility maximization for time-dependent binary choices and a discrete time hazard model. Moreover, we show our discrete time formulation estimates the parameters of an underlying continuous time proportional hazards model. The hazard model imposes three important requirements on the data: (1) exact knowledge of the risk set (those observational units yet to experience trial) at each time period, (2) covariate information for all risk set members, and (3) exact knowledge of neighbor relationships in order to identify the neighborhood effect. A full individual level model would therefore require very detailed covariate data on 300 million individuals and information on where they live in relation to each other. This requirement is clearly impossible to meet. We therefore derive a model based on region level data which satisfies (1)–(3) above, yet is consistent with individual-level decision making.

[6]For a complete treatment of this issue see Anselin (1988) and Manski (1993).

2.1 Individual utility for trial

Consider trial decisions for individuals located in regions $z = 1, \ldots, Z$ and let T_{iz} denote the uncensored time of occurrence of trial for individual i. To allow a behavioral underpinning (utility maximization) for individual trial decisions, we work with the discrete time hazard

$$P_{iz}(t) = P(T_{iz} = t | T_{iz} \geq t, X_{iz}(t)). \tag{1}$$

$X_{iz}(t)$ are covariates that potentially influence the uncensored time of trial. Equation 1 is the conditional probability that an event occurs at t, given that it has yet to occur, and can result from a model of random utility maximization over binary choices. Furthermore, it is important to note that 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).[7]

In addition to the substantive advantage of a utility interpretation offered by the discrete time approach, methodological benefits are simplicity of estimation and the ability to incorporate time-varying covariates. Allow individual i in region z the potential for in trial in any period t, beginning at period 1 (when the innovation first becomes available). The observed variable $y_{iz}(t) \in \{0, 1\}$ indicates that the individual experienced trial ($y_{iz}(t) = 1$) or not ($y_{iz}(t) = 0$). The complete decision history 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).

To see the link to random utility maximization, assume individual i at location z has a latent utility for trial at time t

$$U_{iz}(t) = V_{iz}(t) - \epsilon_{iz}(t), \tag{2}$$

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

2.2 An individual model with region level data

With complete individual-level covariate information one could estimate the parameters of Eq. 2. As noted previously, the data requirements would be

[7] Please see Appendix. Parameters of discrete time models are usually not invariant to the length of time intervals chosen (Heckman and Singer 1984a; Ryu 1995); the discrete time model with complementary log-log link function is the exception (see also Allison (2001 p. 216-219.). Ter Hofstede and Wedel (1999) document aggregation biases in discrete time models. Ryu (1995) shows that even for a standard discrete time model a time interval to average event time ratio of 1/16 is generally sufficient to mitigate bias.

enormous. It is therefore impossible to specify $V_{iz}(t)$ in Eq. 2 at the individual level of aggregation directly. We now derive a region level model which allows identification of the risk set, and covariates for set members. The region level model also enables us to specify neighborhood relationships exogenously. Finally, right censoring is less problematic as it is reasonable to assume that given enough time, each *zip code* will see at least one trial.[8]

At the individual level trial occurs when the utility threshold is crossed. Namely, $y_{iz}(t) = 1$ when $U_{iz}(t) > \tau$ where τ can be normalized to zero without loss of generality. Let $\epsilon_{iz}(t)$ be independently and identically distributed over individuals and time within region

$$f(\epsilon) = \frac{1}{\mu} \exp\left[\frac{\epsilon - \eta}{\mu}\right] \exp\left\{-e^{\frac{\epsilon-\eta}{\mu}}\right\}. \tag{3}$$

The probability that individual i in region z experiences trial at time t is obtained as

$$\begin{aligned} P(y_{iz}(t) = 1) &= P(\epsilon_{iz}(t) \leq V_{iz}(t)) = F_{\epsilon}(V_{iz}(t)) \\ &= 1 - \exp\left\{-\exp\left\{\frac{V_{iz}(t) - \eta}{\mu}\right\}\right\}. \end{aligned} \tag{4}$$

For reasons given above, we do not model this probability but instead model the probability of the *first* trial in a region. 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

$$\begin{aligned} P(y_z(t) = 1) &= P(\max_i \{ U_{iz}(t) \;\; i = 1, \ldots, n_z \} \geq 0) \\ &= P(\max_i \{ V_{iz}(t) - \epsilon_{iz}(t) \} \geq 0) \\ &= P(V_z(t) - \min_i \{ \epsilon_{iz}(t) \} \geq 0) \\ &\quad \text{since we have } \; V_{iz}(t) = V_z(t) \;\; \forall i \\ &= P(\min_i \{ \epsilon_{iz}(t) \} \leq V_z(t)). \end{aligned} \tag{5}$$

Equation 5 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

[8] It is much less reasonable to assume that given enough time each *individual* will eventually try Netgrocer.

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 $\epsilon_{iz}(t)$. It is worthwhile to reflect on the statistical and behavioral appeal and consequences of this IID assumption. The IID assumption implies that there is no *within* region contagion for the *first* trial which is not unreasonable.[9]

The Gumbel distribution in Eq. 3 with location parameter η and scale parameter μ has the useful property that the distribution of the minimum of n_z independent random variables is also Gumbel

$$\begin{aligned} \epsilon_{iz}(t) &\sim G(\eta, \mu) \\ \epsilon_z^{min}(t) &= \min_i \{ \epsilon_{iz}, i = 1, \ldots, n_z \} \\ &\sim G(\eta - \mu \ln(n_z), \mu). \end{aligned} \tag{6}$$

Setting $\eta = 0$ and $\mu = 1$ as standard normalizations, the probability that trial occurs in region z given that it has not yet occurred is

$$\begin{aligned} P(y_z(t) = 1) &= F_\epsilon^{min}(V_z(t)) \\ &= 1 - \exp\left\{ - \exp\left\{ \frac{V_z(t) - (\eta - \mu \ln(n_z))}{\mu} \right\} \right\} \\ &= 1 - \exp\{ - \exp\{ V_z(t) + \ln(n_z) \} \}. \end{aligned} \tag{7}$$

Intuitively, the more individuals there are in a region, the greater the chance that *at least one* will experience trial by a particular date. When combining data across regions, it is vital to take this into account. $\ln(n_z)$ is therefore an "offset" factor controlling for the fact trial is more likely to be observed earlier in regions containing more individuals. In practical empirical terms n_z is simply the region population and easily obtained from the census. The inclusion of $\ln(n_z)$ in the probability expression is not arbitrary but arises from a specific model of individual behavior. As shown in the Appendix, Eq. 7 is also a complementary log-log link function and therefore estimates an underlying continuous time proportional hazards model.

Equation 7 does not yet include a neighborhood effect covariate, which will be specified subsequently. Arrival at a region level specification where neighbor relations are exogenously known makes it possible to investigate neighborhood effects (or region-to-region influence) in trial. At the same time,

[9] Conceptually, this says that first individual to try in a region was a "local innovator" and not influenced by others in the region. This is behaviorally plausible because none of the other same-region individuals had in fact tried. At the same time, the model will however allow for the *first* trier in region z to be influenced by *prior* triers in region j, if region j is a neighbor of z. Details follow shortly.

the model will allow the *first* trier in region z to be influenced by *prior* triers in region j, if region j is a contiguous neighbor of region z.

2.3 Accounting for region level heterogeneity

The derivation above preserves an individual-level behavioral interpretation even though the model will be estimated using region level data. It also serves a statistical purpose because it implies 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 shortly. We have

$$\begin{aligned} V_z(t) &= \alpha_z + \beta X_z(t) + \ln(n_z) \qquad (8)\\ &= \alpha_0 + \beta X_z(t) + \underbrace{\ln(n_z) + (\alpha_z - \alpha_0)}\\ &= \alpha_0 + \beta X_z(t) + \phi \ln(n_z)\\ \text{where } \phi &= \frac{\ln(n_z) + (\alpha_z - \alpha_0)}{\ln(n_z)} \qquad (9) \end{aligned}$$

Hence, when pooling data across regions imposing the theoretical constraint $\phi = 1$ in Eq. 7 is equivalent to assuming that $\alpha_z = \alpha_0$ (in the absence of an additional random term in Eq. 8). This is unlikely to be true empirically so we allow a free parameter for the offset term and model the intercept as a random effect (see also Appendix for details)

$$V_z(t) = \alpha_z + \beta X_z(t) + \phi \ln(n_z), \quad \alpha_z = \alpha_0 + \nu_z \quad \nu_z \sim N(0, \sigma^2). \qquad (10)$$

The model exhibits the appealing property that the control for heterogeneity falls naturally out of the derivation, which in turn follows directly from an underlying behavioral model.[10]

2.4 Neighborhood effects

Imagine that in a region where no individual has yet tried, there is potential for either direct communication with, or passive observation of, individuals from an adjacent region where trial has occurred.[11] 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

[10] 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 discussed in the next section.

[11] As noted earlier, we do not distinguish between the two. Passive observation is facilitated by individuals observing deliveries (each box is clearly marked with "netgrocer.com"). Unfortunately, we cannot address Internet-based communication directly as we have no way to track it.

now on the prior event in z_1 such that $P(y_{z_2}(t) = 1|y_{z_1}(t-1) = 1) \neq P(y_{z_2}(t) = 1|y_{z_1}(t-1) = 0)$. This notion is reflected in the deterministic utility

$$V'_z(t) = V_z(t) + \theta[w_z Y_z(t-1)], \tag{11}$$

where w_z is a row vector whose elements capture the relationship between region z and its neighbors. It has dimension $1 \times N_z$ where N_z equals the number of neighbors in the neighborhood set, including z itself. $Y_z(t-1)$ is an $N_z \times 1$ column vector of the lagged trial behavior of individuals who reside in zip codes contained in the neighborhood set.

An example clarifies these relationships. Figure 2 shows a set of four regions $\{z_1, z_2, z_3, z_4\}$ and the corresponding first-order contiguity matrix C. Elements of C are binary indicators of contiguity and as noted by Anselin (1988, p. 21) " ... the weight matrix should bear a direct relation to the theoretical conceptualization of the structure of dependence ... ". In general, the row vector w_z corresponds to an appropriate weight or influence relationship between neighbors and the focal region z. The collective influence of prior neighborhood set activity on z is obtained by post-multiplication of w_z by $Y_z(t-1)$. Again, the researcher faces various choices in describing $\pi_z(t-1)$, the elements of $Y_z(t-1)$. For example, $\pi_z(t-1)$ could indicate the number of prior triers in region z or even be a simple binary indicator of the presence of prior trial.

In the empirical analysis we define $\pi_z(t-1)$ as the number of previous triers in z which means that the *first* trier in z is therefore potentially influenced not

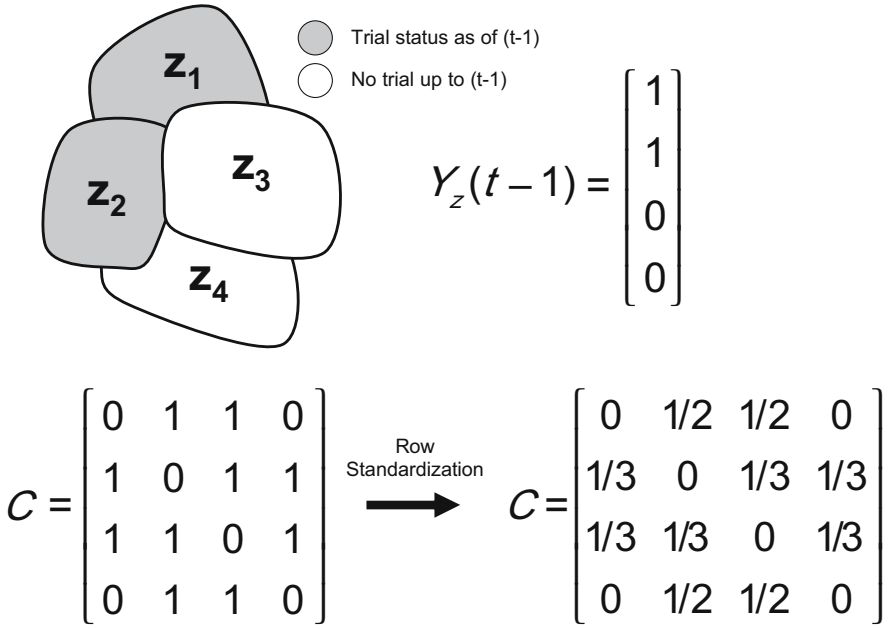


Fig. 2 First order contiguity relationships

only by the first trier in j, but also implicitly by all *subsequent* triers in j who try between time periods s and $t-1$.[12]

2.5 Summary of model properties

Consequences of the model assumptions are as follows.

- *Within Regions* Individuals within a focal region z have IID utilities, leading to an analytical representation of the unobserved *first* trier, or local innovator. Regions are equalized by controlling for the fact that the first trial is likely to occur earlier in regions with more people, all else equal.
- *Across Regions* Influence flows across exogenously defined neighborhood groups. The *first* trier in a region yet to experience trial is affected by the cumulative weight of *all* previous triers in adjacent regions.
- *Rationality* The model is consistent with individual utility maximization. Specification of $w_z Y_z(t-1)$ in accordance with exogenous groupings and lagged behavior satisfies the conditions for rational influence (see Brock and Durlaf 2001) and precludes potential reflection and identification problems discussed in Manski (1993).[13]

Covariates in $V_z(t)$ are described next along with additional procedures to control for heterogeneity. We also address endogeneity and the possibility that spatial-temporal patterns of trial are affected by time and region-varying marketing efforts of Netgrocer.

3 Data and preliminary analysis

Data are drawn from three sources: (1) Netgrocer transaction files, (2) the United States census, and (3) CACI (caci.com) retailing statistics. These datasets are linked via the common zip code variable. Descriptive analyses are provided to illustrate basic properties of the data (all summary analyses are available upon request).

[12] In order to check that results are robust to alternative formulations, we estimate a wide variety of alternative models that use different standard definitions for these two constructs. The results are reported in a subsequent section. We thank an anonymous reviewer for suggesting that the *number* of individuals (or a function thereof) as the measure for $\pi_z(t-1)$ gives the most behaviorally meaningful interpretation. Our final specification relies on $\ln[w_z Y_z(t-1)+1]$, and a previous working paper reports (qualitatively similar) results from a model that uses equal influence weights and binary indicators of lagged neighbor behavior.

[13] Manski (1993) 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. 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.).

3.1 Internet access

An important rival hypothesis for any neighborhood effect finding is that the space-time pattern of evolution for Netgrocer simply mirrors the diffusion of the Internet. We therefore need to appropriately control for space time variation in access to the Internet. Our preferred model specification will use the entire region-level dataset (29,701 zip codes) and utilize a proxy variable called "Broadband Access Providers" as described below. In settling on this specification we do however consider three separate complementary approaches to controlling for access to the Internet. First, we note that a semi-annual reporting requirement for the Federal Communications Commission (FCC), FCC Form 477, has been used since 1999 to determine the extent of local telecommunications competition and deployment of broadband services at the zip code level. From a linear interpolation based on the four different data points (Dec 1999, Jun 2000, Dec 2000, Jun 2001), we created a "Broadband Access Providers" variable that varies over time (month) and space (zip code). This variable is a raw count of the number of providers and is available for all residential zip codes used in our analysis.

Second, we utilize the fact that the number of households with an Internet connection is available as supplementary data to the Current Population Survey (CPS), beginning in 1997. Internet usage data were collected in the form of supplementary questions in the CPS in October 1997, December 1998, August 2000, September 2001, and October 2003. CPS data are available at various geographical units including State, CMSA, County, and MSA, MSA being the smallest. The 173 MSAs in the CPS cover 8,185 zip codes and for these zip codes we again use linear interpolation to create the variable "CPS Internet Penetration" which varies over time and region and include this variable in models estimated on the reduced dataset.

Third, we modify the definition of the potential users. Our chosen specification still assumes that the unit of analysis is the region (zip code), yet for each region we try to estimate the proportion of people with Internet access who reside there and judge this group to be those who are able to try Netgrocer.[14] To proxy this, we take the 8,185 zip codes identified above and interact the variable CPS Internet Penetration with the total population of the zip code. In order to assess which of the three variables performs best as a proxy for Internet access, we estimate three separate formulations on the dataset with 8,185 zip codes (on which all three measures can be computed). The empirical results are given subsequently.

3.2 Raw data

(1) *Transaction Data and Institutional Setting* It is important to understand the Netgrocer business model during the data period. Netgrocer offered shipments of non-perishable grocery items nationwide. The stated goal

[14] We thank an anonymous reviewer for explicitly suggesting this approach.

was to provide a service supplementary to that of traditional supermarkets. Customers could shop at local stores for perishable products and fill part or all of their non-perishable requirements at Netgrocer. Shipping was provided by Federal Express at a standard rate of $6.99 per order.[15] Netgrocer.com launched on May 7, 1997 and by January 31, 2001 had accumulated 382,478 transactions (Netgrocer is a going concern but we do not examine data after January 2001). The 382,478 orders were placed by 162,618 different customers and shipped to 19,418 unique zip codes. The process is observed from inception so there is no left-censoring. The average order value of $57.53 (std. dev. $50.99) is larger than that at traditional grocery stores $26.26 (std. dev. $29.18).[16]

Each transaction is described by: (1) date, (2) customer identification number, (3) total dollar value, and (4) zip code where the order was shipped. Some pruning is needed before these data are merged with other information. The census data and records provided by ESRI (esri.com) show 29,701 residential zip codes for the United States and we focus on these.[17] By January 2001 17,910 of these zip codes had seen at least one order, while the remaining 11,791 had not: Netgrocer had achieved trial in sixty percent of the residential zip codes.

(2) *Census Data* From the 2000 census we created three categories of covariates. While this process is necessarily a matter of judgment, it was performed with reference to prior literature on the compilation of sociodemographic information (e.g., Dhar and Hoch 1997). Zip code profiles are summarized by

a. *Household Characteristics:* Ethnicity, Gender, Family Size.
b. *Household Economics:* Age, Education, Employment Status, Income.
c. *Local Environment:* Home Value, Land Area, Population, Urbanization.

Table 1 lists all variables and associated descriptive statistics for both the full sample of 29,701 zip codes and the reduced sample of 8,185 zip codes to be used with the CPS data.

Each variable contributes one observation per zip and is often expressed as a percentage. "Elderly" is defined as "the percentage of the zip code

[15]Netgrocer periodically ran specials to entice new customers, or encourage existing customers to buy larger orders. We are unable to separately account for orders that might have resulted from such promotions. Subsequent to the data collection period management introduced a non-linear shipping fee schedule. Fees are differentiated by order size and region of the country (larger orders to western states are more expensive).

[16]Based on 161,778 shopping trips taken by the 1,042 consumers in the Stanford Market Basket Database. Those data cover June 1991–June 1993. The inflation-adjusted average order value for the period of the Netgrocer data is approximately $30.

[17]Netgrocer shipped 25,132 orders to 1,508 non-residential zip codes which were predominantly Army Post Office (APO) addresses. The average dollar value of these orders is $69.12 and we have excluded them from our analysis. Detailed information is available upon request.

Table 1 Region (zip code) characteristics and access to retail services

Variable	Description	Full data (n = 29,701)		CPS subset (n = 8,185)	
		Mean	Std dev	Mean	Std dev
(1) Household characteristics					
Blacks	% of Blacks	0.0725	0.1563	0.0880	0.1665
Foregn	% of foreign born individuals (aged 18+)	0.0434	0.0790	0.0411	0.0661
Hispanics	% of Hispanics	0.0459	0.1141	0.0514	0.1294
Large family	% of families with five or more members	0.1515	0.0607	0.1473	0.0589
Solo female	% of single female households	0.0477	0.0245	0.0481	0.0277
Solo male	% of single male households	0.0356	0.0202	0.0346	0.0204
(2) Household economics					
College	% with bachelors and/or grad/prof degree	0.0984	0.0785	0.1069	0.0776
Elderly	% aged 65 and above	0.1371	0.0586	0.1268	0.0616
Fulltime female	% of households with f-t female worker	0.2545	0.0839	0.2799	0.0780
Fulltime male	% of households with f-t male worker	0.4850	0.1197	0.5033	0.1120
Generation X	% of individuals 25-34, incomes > $50k	0.0102	0.0116	0.0118	0.0101
Wealthy	% of households earning $75k+	0.0660	0.0833	0.0692	0.0681

(3) Local environment					
Density	Population density	1,108.0700	4,270.6200	999.7380	1,773.7715
Home value	% of homes valued at $250k or more	0.0232	0.0782	0.0154	0.0477
Households	Number of households	3,095.4000	4,415.5400	4,313.3381	4,526.0870
Land area	Area in square miles	110.2122	387.1567	60.1757	185.3236
Large house	% of homes with five bedrooms or more	0.0339	0.0324	0.0279	0.0264
Population	Total population	8,372.6100	11,867.6000	11,519.0160	11,916.3060
Urban housing	% of urban housing units	0.1098	0.1393	0.1614	0.1426
(4) Access to retail services					
Distance to convenience	Expected max. distance to a convenience store	6.3172	8.0202	3.8641	4.8240
Distance to drug	Expected max. distance to a drug store	7.8207	8.7894	5.3370	5.7469
Distance to general	Expected max. distance to a general store	7.8893	8.4296	5.5629	5.6300
Distance to supermarket	Expected max. distance to a supermarket	4.3086	6.0105	2.4736	3.2666
Distance to warehouse	Expected max. distance to a warehouse store	11.0665	9.7632	8.7685	6.4391
(5) Access to internet					
Broadband access providers	No. of high-speed ISPs	0.3776	0.6585	0.5446	0.7473
CPS Internet penetration	MSA measure from CPS			0.2474	0.1422

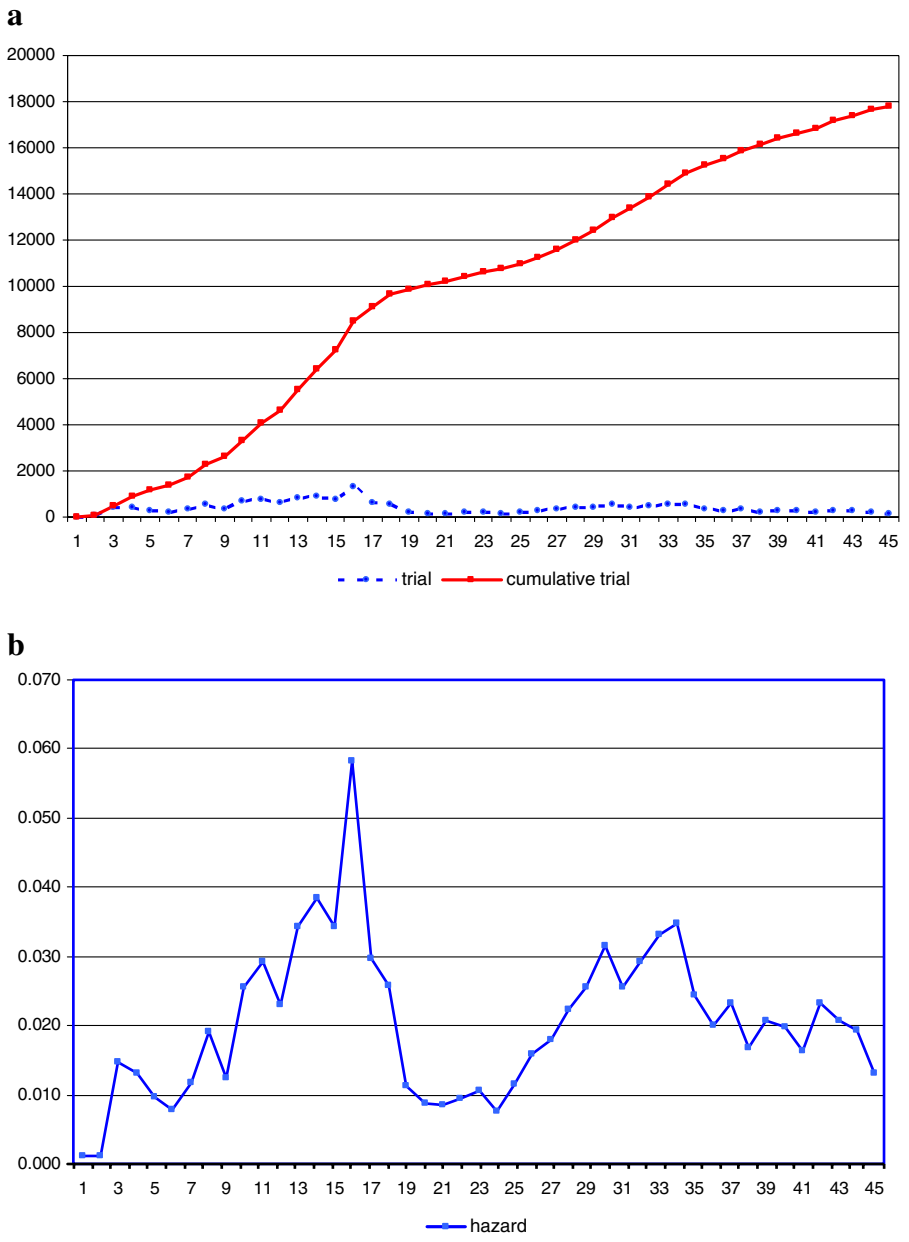
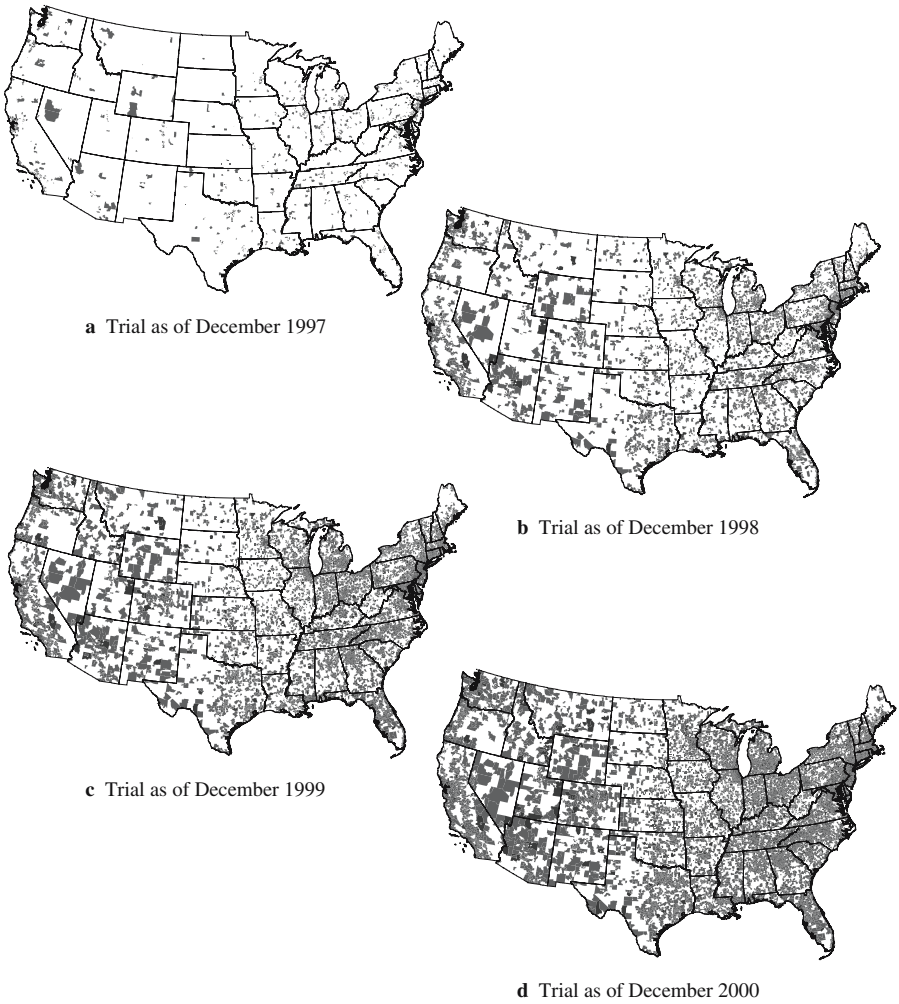


Fig. 3 **a** Netgrocer trials by zip code **b** Empirical hazard

population that is aged 65 and above," "College" represents "the percentage of individuals with bachelors and/or graduate or professional degrees," and so forth. Defining variables this way induces greater variation across observational units and is consistent with Dhar and Hoch (1997).



a Trial as of December 1997

b Trial as of December 1998

c Trial as of December 1999

d Trial as of December 2000

Fig. 4 Space–time evolution of trial

This more "extreme" representation of the zip code characteristics (as opposed to using say average income, etc.) fits nicely with the idea that the first individual to try is a local innovator.

(3) *Retail Competition Data.* Nationwide distribution gives Netgrocer access to a vast potential customer base and also exposes them to thousands of competitors. As individuals in each zip code can still shop at local stores it is important to include information on the availability of this outside or status quo option. CACI report the number of outlets and average sales volume by zip code for five classes of retailer: convenience stores, drug stores, general merchandisers, supermarkets and warehouse clubs. We compute a measure of the maximum expected distance an individual within the zip code must travel to reach each type of store. Using zip

Fig. 5 Trial and neighborhood effects

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.

We do not have information on any marketing efforts undertaken by Netgrocer but we know these were: (1) sporadic, (2) small-budget, and (3) often focused on emails to existing customers. Explicit covariates therefore cannot be used to account for them, however a combination of time and region fixed effects, coupled with random effects in the baseline across zip codes are used to absorb, rather than explain, their impact. Details follow shortly.

3.3 Preliminary analysis

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

Spatial–Temporal Patterns By the end of May 1997 trial had occurred in thirty-four distinct zip codes ranging from New Jersey to California. Figure 4 shows cumulative space-time trial patterns at one year intervals with the pool of triers expanding rapidly throughout the United States. These data reveal the most dramatic difference between a traditional retailer (where customers are contained within a relatively small area) and an e-tailer. More disaggregate visual inspection of the patterns raises the possibility that neighborhood effects play a role. Trial at time t does not occur randomly in space. Rather, 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 3-month increments for two separate east and west coast snapshots. As time moves along new trials are more likely to arise close to contiguous areas of prior trial.[18]

Neighbors A contiguous neighbor j is a zip code that shares an adjoining boundary with the focal zip code z. Neighbor connectivity data was obtained for all 29,701 zip codes. The average number of regions in a neighborhood set is 5.63 (std. dev. = 2.28). Most zip codes have at least one neighbor, however there are 136 "islands" who have no direct contiguous US neighbors.

4 Empirical findings

Special emphasis is given to demonstrating the neighborhood effect θ is properly identified and robust to controls for heterogeneity, selection, and unobservables. The importance of measures of observed heterogeneity should not however be overlooked. Our data set contains a far greater number of covariates than is typical in models of spatial effects. Bronnenberg and Mela (2004) for example, note the importance of random effects in their model to account for the influence of omitted variables such as market demographics.

[18]This visual pattern is representative of other months and regions. In the interests of brevity other figures are not shown but are available from the authors upon request.

4.1 Estimation

The discrete time model with a complementary log-log specification mimics an underlying continuous time process (see Appendix). 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 1967) used earlier to derive a regional level model from the individual behavior. Working from that derivation and substituting from equation (7)

$$\begin{aligned}\log\left[-\log(1-P_z(t))\right] &= \log\left[-\log\left[\exp\left\{-\exp(V_z(t)+\ln(n_z))\right\}\right]\right] \\ &= V_z(t)+\ln(n_z) \qquad (12)\end{aligned}$$

The right hand side is therefore equivalent to the deterministic utility for the first trier from the region. Specifications for $V_z(t)$ utilize covariates given in Table 1.

Recall $y_z(t) = 1$ indicates the first trial occurred in zip code z at time t. For non-censored observations, let T_z reflect the time at which $y_z(t) = 1$. It follows that the number of observations zip z contributes for estimation is T_z with the dependent variable $y_z(s)$ equal to zero for all periods s with $s < T_z$. For the 11,791 zip codes where trial is never observed and the data are censored, $T_z = 45$ (May 1997 through January 2001, inclusive). The total number of stacked observations in the full dataset is $\sum_z T_z = 910,769$. For the reduced dataset created to examine Internet access using variables constructed from the CPS, there are 8,185 zip codes which generate 211,032 observations. In all instances, parameters are estimated via binary choice analysis with a complementary log-log link function on these stacked data.

4.2 Initial evidence for neighborhood effects

In our first models the neighborhood effect and the population offset are the sole covariates. The neighborhood effect is formulated in two different ways

- As a lagged cumulative effect (LC) with elements of the column vector for neighbor behavior $Y_z(t-1)$ containing the counts of all previous trials occurring up to and including time $t-1$ in the neighbors of z, and
- As a standard lagged effect (L) with elements of $Y_z(t-1)$ containing only the counts of trials that occurred at $t-1$.

Contemporaneous representations violate the rationality conditions discussed previously, and the associated parameters are not theoretically estimable using maximum likelihood methods. Table 2 shows estimates for θ, model fits and Wald χ^2 statistics. θ is positive and significant for both formulations. The superior fit of the LC model occurs because it captures the influence on focal region z of not only the *first* trial of a neighbor j, but also *all* trials that have occurred in j between the initial trial at time $s \leq t-1$ and time $t-1$.

Unobserved Common Traits If contiguous regions share unobserved common traits that are positively correlated with the utility of trial, the

Table 2 Initial evidence for neighborhood effects on trials

Formulation	Estimate of θ	Wald χ^2	-2 Log(L)
Intercept only	–	–	176,199.8
Lagged Cumulative (LC)	0.614	8,491.6	168,606.5
Lagged (L)	1.245	7,675.5	170,446.1

neighborhood effect will be biased upwards. One candidate is unobserved technological sophistication or Internet access. At the same time, it is difficult to accept that individuals self select locations in which to live on the basis of preferences for Netgrocer. Three procedures help mitigate potential upwards bias from these unobservables.

First, lagged cumulative trial (LC) is adopted as the empirical formulation of the neighborhood effect. This 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 occurred from contiguous neighbors where it has is similar to the strategy used by Goolsbee and Klenow (2002).

Second, a comprehensive set of covariates are added to the model (see Table 1). These variables control for observable differences across regions in: (1) intrinsic characteristics of the population, (2) socio-economic status, (3) physical environment, and (4) access to competing retail services. This helps eliminate omitted variable bias that would otherwise amplify the effect attributed to neighborhood effects. All these observables are measured at the level of the region, z, and therefore directly related to the dependent variable, $y_z(t)$.

Third, we proxy for region and time-specific access to the Internet using the zip code specific Broadband Access Providers variable described previously.

Unobserved Heterogeneity Over Space and Time Region fixed effects control for other sources of variation that are attributable to unobserved cross-sectional differences. While zip code specific fixed effects are possible in theory, implementation is difficult due to the number required and the unbalanced design of the data matrix. In initial models we therefore use state fixed effects (states are higher order regions) and later settle on two-digit zip fixed effects. These effects are in addition to the random effect on the baseline hazard already shown in Eq. 10.

Figure 3 implies variation in the baseline over time occurs in addition to any variation over region. Negative duration dependence with higher hazard regions experiencing trial earlier would cause an attenuation of the neighborhood effect but would not bias the standard error (Gail et al. 1984). Allison (2001) suggests a non-parametric time-dependent baseline hazard modeled with time-specific fixed effects to accommodate this. Our final and most general model specification is given in Eq. 21 in the Appendix. It includes all control procedures just discussed, namely a rich set of covariates, two-digit zip fixed effects, non-parametric baseline, and normal mixing.

4.3 More evidence for neighborhood effects

Table 3 provides model fits and estimates for θ obtained after implementation of the control procedures outlined above. Rows 2 and 3 report the benchmark fixed effects and non-parametric baseline hazard model fits. Where used the neighborhood effect (θ) enters according to the Lagged Cumulative (LC) specification. Rows 4 and 5 show that the effect is robust to the separate inclusion of two-digit zip fixed effects and a non-parametric baseline. Including region characteristics reduces the magnitude of θ and improves fit. The last row shows θ for Model 10, the best fitting and most general model. It is still positive and highly significant ($\theta = 0.170$, Wald $\chi^2 = 173.2$). This model is especially instructive as Internet access is proxied for and fixed effects are at the two-digit zip (rather than state) level.

Collectively, these results provide some assurance θ captures a behavioral process and does not simply mimic access to the Internet or unobserved

Table 3 More evidence for neighborhood effects on trials

Formulation	Estimate of θ	Wald χ^2	−2 Log(L)
Benchmark models			
1. Intercept only	–	–	176,199.8
2. State fixed effects	–	–	159,266.5
3. Non-parametric baseline hazard	–	–	149,947.3
Models with neighborhood effect			
4. Lagged Cumulative (LC) θ + Two-digit zip code fixed effects	0.555	5,635.5	152,962.0
5. LC θ + Non-parametric baseline hazard	0.409	1,483.0	148,045.7
Models w/ neighborhood effect and covariates			
6. LC θ + Non-parametric baseline hazard + Region characteristics	0.272	484.9	144,359.9
7. LC θ + Non-parametric baseline hazard + Region characteristics + Retail access	0.278	490.2	144,220.9
8. LC θ + Non-parametric baseline hazard + Region characteristics + Retail access + Two-digit zip code fixed effects	0.144	159.3	142,682.0
9. LC θ + Non-parametric baseline hazard + Region characteristics + Retail access + Two-digit zip code fixed effects + Broadband access	0.131	127.7	142,638.8
10. LC θ + Non-parametric baseline hazard + Region characteristics + Retail access + Two-digit zip code fixed effects + Broadband access + Random effect on intercept	0.170	173.2	142,510.0

marketing effort. The increase in θ from Model 9 to Model 10 when the random effect is added is consistent with the presence of some negative duration dependence. Inspection of the full list of parameter estimates from the final model shows that θ is second only to "Percentage of College Educated Households" (College) in its level of statistical significance and that other covariates have plausible signs and significance levels (see Table 4). This finding was tested extensively in a number of alternative models and found to be consistent (details follow).

Table 4 The effect of region characteristics on trials (29,701 zip codes)

Variable	Coefficient	Std Err	Wald χ^2	p-value
Model intercept[a] (α_0)	−9.039	0.263	1182.0	< 0.001
Population control $\log(n_z)$ (ϕ)	0.727	0.018	1700.7	< 0.001
Broadband access	0.129	0.014	88.4	< 0.001
Neighborhood effect (θ)	0.170	0.013	173.2	< 0.001
Region level covariates				
(1) Household characteristics				
Blacks	−0.746	0.082	83.5	< 0.001
Foreign	−0.049	0.214	0.1	0.820
Hispanics	−0.622	0.159	15.2	< 0.001
Large family	−3.053	0.287	113.0	< 0.001
Solo female	−2.637	0.783	11.4	0.001
Solo male	5.693	0.649	77.1	< 0.001
(2) Household economics				
College	3.791	0.244	241.5	< 0.001
Elderly	−1.544	0.283	29.7	< 0.001
Fulltime female	−0.328	0.199	2.7	0.099
Fulltime male	−0.156	0.138	1.3	0.260
Generation X	7.635	1.283	35.4	< 0.001
Wealthy	0.110	0.281	0.2	0.696
(3) Local environment				
Density	0.000	0.000	0.4	0.515
Home value	−0.546	0.198	7.6	0.006
Households	0.000	0.000	27.4	< 0.001
Land area	0.000	0.000	0.8	0.382
Large house	0.833	0.427	3.8	0.051
Urban housing	0.428	0.119	13.0	< 0.001
(4) Access to retail services				
Distance to convenience	0.005	0.003	1.8	0.174
Distance to drug	−0.010	0.003	12.0	0.001
Distance to general	−0.003	0.003	1.7	0.191
Distance to supermarket	−0.027	0.005	30.1	< 0.001
Distance to warehouse club	0.011	0.002	18.1	< 0.001

[a]Estimate of σ (Eq. 10) = 0.533;

LR test: $\rho = \frac{\sigma^2}{1+\sigma^2} = 0$ yields $\chi^2_1 = 128.8$, $p < 0.001$.

4.4 Alternative models and specifications of the neighborhood effect

Several alternative models varied the definition of the risk set, restrictions on the population offset, weights of the contiguity matrix (w_z), and the elements of the lagged neighborhood actions vector ($Y_z(t-1)$). These models were estimated in order to examine the robustness of the basic neighborhood effect to different treatments (full results available upon request).

Formulations, Variables, Residuals A model with the constraint $\phi = 1$ with a random effect on the intercept and gives results essentially identical to those in Table 3. Other specifications define n_z as the number of households in the region, not individuals. Again, the qualitative results are unchanged. A further model uses cumulative adoptions at time t in addition to the neighborhood effect, and the effect remains. We tested the empirical veracity of the logarithmic form for population which follows from our derivation. A model with linear through fourth order polynomial terms for population falls short on the basis of fit, supporting our theory-driven choice. The Heckman–Singer (1984b) non-parametric approach to heterogeneity produced results essentially identical to for our model with Normal mixing.

We checked for evidence of spatial autocorrelation in the residuals of our final model, computing Moran's I statistic using neighborhood contiguity matrices as weights for all zip codes, for all 45 months. Ten months show significant positive autocorrelation, two show significant negative autocorrelation, and the mean spatially-weighted residual is 0.005. The residuals are small in absolute terms and decline in value with time (only two of the last 20 months show significant positive values). As a point of comparison, we computed the same test values for a model that is identical, but does *not* include θ. Here the pattern of autocorrelation is identical and the correlation between residuals for models with and without θ is 0.98. Thus, we conclude that while some limited evidence for spatial autocorrelation exists, it is clear that the neighborhood effect θ is not simply picking this up. As a final check, we re-estimated the model but included neighborhood averages of all the demographic regressors and Broadband access measure as covariates. In this case the qualitative results were unchanged and the point estimate of θ was 0.197 (Wald $\chi^2 = 173.3$); the number of instances of significant Moran's I values declined from 12 to 9. An alternative approach would be to use non-parametric methods to obtain robust standard errors (see Conley 1999; Conley and Molinari 2007).

Neighborhood Effect Covariate and Relationship to Other Studies We investigated alternative representations of the neighborhood effect through modifications of $w_z Y_z(t-1)$. Following the spirit of Bronnenberg and Mela (2004) who constructed w_z using relative category volume at neighboring retailers we use relative population size: $w_z = POP_z / \sum_j^{N_z} POP_j$. We also amended $Y_z(t-1)$ with the proportion of population trying, and simple binary indicators of the presence or absence of prior trials. All such variations lead to θ values which

are positive and significantly different from zero, which again gives us some confidence in the basic empirical finding.

Summary The pattern of results assessed through many alternative specifications is consistent with the presence of neighborhood effects—every model specification produced qualitatively identical results. Our final model (Model 10) employs a rich specification of observed heterogeneity, two-digit zip fixed effects, a random effect on the baseline, and non-parametric time-dependence. In this model, as in all others, θ remains positive and significant, while the other coefficients have plausible signs and levels of significance.

4.5 Substantive findings on the effect of region covariates on trial

Table 4 shows how region characteristics affect time to trial (a positive coefficient means that the covariate *speeds up* the time to trial). Fourteen of twenty-three parameters are significantly different from zero ($p < 0.01$) and the implied marginal effects are intuitive. The magnitude and level of significance of θ is unaffected by the presence of Broadband Access and this variable itself is highly significant and correctly signed. Discussion of the remaining variables follows the classification in Table 1.

(1) *Household Characteristics* Regions with greater percentages of minorities experience trial later, consistent with evidence for "digital divide" in which these groups have 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*). Percentage of solo person households is also important, but interacts with gender: Regions with greater proportions of male-only households see trial earlier. Conversely, an increase in the proportion of large (greater than five person) households slows time to first trial. Larger families may prefer one stop shopping and Netgrocer does not sell perishables.

(2) *Household Economics* Higher percentages of tertiary-educated individuals leads to earlier trial. An increase in the number of young wealthy individuals (Generation X) shows an additional positive effect, whereas a higher percentage of elderly individuals slows time to trial. Other variables held constant, working status household members shows no effect.

(3) *Local environment* The number of households, size of the housing unit and the extent of urbanization have a positive effect on the time to first trial. The latter two effects are weak (population is already controlled for), but the collective impact is likely a proxy for the potential for social interaction within a region, and within and between household members.

(4) *Access to Retail Services* Estimates for convenience stores and general merchandisers suggest travel distance to either has no effect on Netgrocer trial. An increase in expected travel distance to drug stores and supermarkets *decreases* time to first trial. While this may seen counterintuitive, it can be reconciled in light of the format differences. Netgrocer offers

neither perishable products nor a full complement of drug store items. A household using Netgrocer would still need to visit a supermarket or drug store. If the supermarket (for example) is relatively far away, then a rational household might amortize the fixed cost of a trip by doing do one-stop shopping, thus eliminating any need to purchase non-perishables at Netgrocer. A household with better access to traditional stores might be more willing to split the shopping basket for perishables (supermarket) and non-perishables (Internet). Conversely, Netgrocer seems to compete more directly with warehouse clubs. The less convenient the warehouse club, the more likely shoppers are to try Netgrocer.

4.6 Internet access, censoring, and time variation in θ

Our substantive conclusions have been based on a model that assumes: (1) the variable Broadband Access Providers is a suitable proxy for Internet access, (2) all 29,701 zip codes can be used in estimation (i.e., this is a reasonable definition of the risk set), and (3) θ is a constant, conditional on the other covariates. We provide results that suggest not only are these assumptions empirically reasonable, but also that departure from them does not lead the neighborhood effect to break down or disappear. Table 5 provides estimates from three different approaches to proxying for Internet access discussed earlier. Note that all three models use only the 8,185 zip codes in the reduced dataset because the CPS Internet penetration variables do not cover the entire US. One other argument in favor of using the reduced MSA-level dataset is the following. It is difficult to construct a good measure of "proximity with regard to social contacts" across regions that vary greatly in terms of population density; hence, the MSA-only sample which uses primarily high density zip codes may constitute a superior sample.

Model 1 shows that the Broadband Access Providers covariate and the neighborhood effect are still positive and significant in the reduced dataset. Model 2 shows that while the neighborhood effect remains significant when Internet access is proxied for with the CPS Internet Penetration variable derived from the CPS, the variable itself is not significant. Model 3 suggests that there is perhaps a better way to use the CPS data. Here, CPS Internet Penetration (CPSPenet) is interacted with the zip code population to in effect adjust the region-specific "risk set" to only those who are estimated as having Internet access.[19] The Broadband Access Providers variable is still included as a control (given the results from Model 1 in Table 5). Again, the neighborhood effect remains significant.

One could also look for evidence of spillover or neighborhood effects in zip codes that have no high speed access. A neighborhood effect estimate

[19]We control for over time variation in zip code level access to the Internet; one could also use such data to model changes in the "at risk" population. While only relatively crude controls for the size of the risk set are available, the substantive results from a variety of formulations produce qualitatively identical results.

Table 5 The effect of region characteristics on trials (8,185 zip codes)

Variable	Model 1: broadband access				Model 2: CPS internet penetration				Model 3: at-risk population			
	Coef	Std err	Wald χ^2	p-value	Coef	Std err	Wald χ^2	p-value	Coef	Std err	Wald χ^2	p-value
Neighborhood effect (θ)	0.143	0.023	38.2	< 0.001	0.152	0.022	45.9	< 0.001	0.150	0.023	41.9	< 0.001
Population control $\log(n_z)$ (ϕ)	0.721	0.036	391.7	< 0.001	0.744	0.035	443.9	< 0.001				
Population control $\log(n_z * CPSpenet)$ (ϕ')									0.627	0.033	368.5	< 0.001
Broadband access	0.122	0.024	25.1	< 0.001					0.136	0.024	31.0	< 0.001
CPS Internet penetration					−0.155	0.411	0.1	0.707				
Model intercept[a] (α_0)	−8.656	0.688	158.3	< 0.001	−8.626	0.706	149.3	< 0.001	−7.739	0.673	132.2	< 0.001
Region level covariates												
(1) Household characteristics												
Blacks	−0.631	0.150	17.8	< 0.001	−0.601	0.144	17.4	< 0.001	−0.611	0.149	16.8	< 0.001
Foreign	1.642	0.553	8.8	0.003	1.581	0.535	8.7	0.003	1.631	0.549	8.8	0.003
Hispanics	−1.482	0.336	19.4	< 0.001	−1.422	0.325	19.1	< 0.001	−1.444	0.335	18.6	< 0.001
Large family	−3.331	0.632	27.8	< 0.001	−3.284	0.612	28.8	< 0.001	−3.193	0.627	25.9	< 0.001
Solo female	−0.647	1.537	0.2	0.674	−0.723	1.492	0.2	0.628	−0.514	1.525	0.1	0.736
Solo male	7.324	1.547	22.4	< 0.001	7.807	1.488	27.5	< 0.001	6.595	1.534	18.5	< 0.001
(2) Household economics												
College	2.991	0.475	39.7	< 0.001	2.839	0.457	38.6	< 0.001	2.955	0.472	39.2	< 0.001
Elderly	−2.217	0.536	17.1	< 0.001	−2.139	0.519	17.0	< 0.001	−2.137	0.532	16.1	< 0.001
Fulltime female	0.160	0.397	0.2	0.688	0.227	0.386	0.3	0.557	0.216	0.394	0.3	0.584
Fulltime male	−0.109	0.288	0.1	0.706	−0.097	0.280	0.1	0.729	−0.065	0.286	0.1	0.820
Generation X	11.458	2.685	18.2	< 0.001	11.551	2.594	19.8	< 0.001	11.019	2.662	17.1	< 0.001
Wealthy	0.115	0.565	0.0	0.839	0.202	0.546	0.1	0.712	0.185	0.561	0.1	0.741

Table 5 Continued

(3) Local environment												
Density	0.000	0.000	4.9	0.028	0.000	0.000	6.0	0.014	0.000	0.000	3.4	0.065
Home value	−0.248	0.552	0.2	0.653	−0.211	0.532	0.2	0.692	−0.315	0.551	0.3	0.567
Households	0.000	0.000	8.2	0.004	0.000	0.000	6.6	0.010	0.000	0.000	22.3	< 0.001
Land area	0.000	0.000	0.1	0.748	0.000	0.000	0.4	0.509	0.000	0.000	0.0	0.924
Large house	1.920	0.941	4.2	0.041	1.982	0.912	4.7	0.030	1.764	0.934	3.6	0.059
Urban housing	0.130	0.213	0.4	0.542	0.188	0.206	0.8	0.361	0.344	0.211	2.7	0.103
(4) Access to retail services												
Distance to convenience	−0.009	0.008	1.2	0.275	−0.010	0.008	1.6	0.212	−0.008	0.008	1.1	0.306
Distance to drug	−0.001	0.007	0.0	0.871	−0.002	0.007	0.1	0.745	−0.002	0.007	0.1	0.799
Distance to general	−0.011	0.006	3.7	0.056	−0.011	0.005	4.4	0.035	−0.010	0.006	3.1	0.080
Distance to supermarket	−0.018	0.011	2.5	0.112	−0.017	0.011	2.2	0.138	−0.024	0.011	4.6	0.033
Distance to warehouse club	0.014	0.006	5.6	0.018	0.013	0.006	4.8	0.028	0.020	0.006	11.3	0.001
	[a]Estimate of σ (Eq. 10) = 0.470; LR test: $\rho = \frac{\sigma^2}{1+\sigma^2} = 0$ yields $\chi_1^2 = 28.65, p < 0.001$.				[a]Estimate of σ (Eq. 10) = 0.383; LR test: $\rho = \frac{\sigma^2}{1+\sigma^2} = 0$ yields $\chi_1^2 = 16.91, p < 0.001$.				[a]Estimate of σ (Eq. 10) = 0.468; LR test: $\rho = \frac{\sigma^2}{1+\sigma^2} = 0$ yields $\chi_1^2 = 28.47, p < 0.001$.			

obtained from such a sample is not entirely free of a potential broadband access confound because a customer who ships an order to a region with no broadbrand access could have done so by placing the order from a broadband access zip code (say at work). Nevertheless, it is comforting to see that even in this case, a similar estimate of θ is obtained (see Table 6).

As noted during the introduction to the model, one advantage of our region level approach is it dramatically reduces the size of the risk set while at the same time making a more reasonable assumption about who should be contained in it. That is, it is more reasonable to conclude that all 29,701 zip codes could eventually see one trial of Netgrocer (and are therefore "at risk") than it would be to assume that all 300 million individuals in the US will eventually try. A more extreme approach would be to estimate the model only using those zip codes that eventually see trial within the observed forty-five periods of the data.[20] That is, all 11,791 right censored zip codes that each contribute 45 observations with the dependent variable equal to zero are dropped. As shown in Table 6, while the magnitude of the neighborhood effect is somewhat attenuated (θ drops from 0.170 in the full sample to 0.136 here), it remains significant. Of secondary interest (given our discussion above) is the observation that the coefficient on Broadbrand Access Providers also falls from 0.129 to 0.093.

Finally, one might expect that the magnitude of the neighborhood effect could differ over time. To investigate this possibility, we reestimate our full sample model with two digit zip fixed effects and a non-parametric baseline and allow θ to vary at 5-month discrete intervals.[21] The results for this model are shown in Table 7. The estimate of θ is significantly higher over the first 10 months and by the last 5 months of the observation period has reached a value of 0.072, which is considerably smaller than the constant value full sample estimate of 0.170. This pattern is certainly consistent with the notion that as time passes along, potential customers may have been more likely to have tried competing or services, or that later triers are less susceptible to neighborhood effects. We do not explore the behavioral underpinnings of this empirical finding here, but note that time variation in θ could be consistent with standard theories of diffusion.

Summary The estimates paint an intuitive and plausible picture of how region characteristics influence trial. All significant effects—and the focal neighborhood effect θ—remain so after the inclusion of many controls, the most important of which is n_z, the number of individuals residing in the region. The inclusion of $\log(n_z)$ is not arbitrary but derived from the assumptions on the distribution of individual-level utility and allows us couch the analysis

[20] We are indebted to an anonymous reviewer for suggesting this check.

[21] We thank an anonymous reviewer for prompting this analysis. Our choice of 5 month intervals is somewhat arbitrary, however popular press articles and some leading market research companies suggest that new product related "buzz" can last for up to 20 weeks. See Bzzagent.com.

Table 6 The effect of region characteristics on trials (alternative models)

Variable	Zip areas with no broadband access				Right-censored zip areas dropped			
	Coefficient	Std err	Wald χ^2	p-value	Coefficient	Std err	Wald χ^2	p-value
Model intercept[a] (α_0)	−9.534	0.513	345.2	< 0.001	5.202	76.808	0.0	0.946
Population control $\log(n_z)$ (ϕ)	0.834	0.027	970.9	< 0.001	0.272	0.021	165.3	< 0.001
Broadband access	–	–	–	–	0.093	0.017	31.4	< 0.001
Neighborhood effect (θ)	0.178	0.023	58.1	< 0.001	0.136	0.017	67.5	< 0.001
Region-level covariates								
(1) Household characteristics								
Blacks	−0.421	0.161	6.8	0.009	−0.624	0.107	34.0	< 0.001
Foreign	0.848	0.454	3.5	0.062	0.453	0.271	2.8	0.095
Hispanics	−0.325	0.320	1.0	0.310	−0.845	0.209	16.3	< 0.001
Large family	−3.593	0.455	62.4	< 0.001	−2.168	0.376	33.2	< 0.001
Solo female	−1.916	1.521	1.6	0.208	−2.665	1.092	5.9	0.015
Solo male	5.557	1.575	12.5	< 0.001	5.745	1.016	32.0	< 0.001
(2) Household economics								
College	4.036	0.421	92.0	< 0.001	4.187	0.319	171.8	< 0.001
Elderly	−1.806	0.495	13.3	< 0.001	−2.076	0.362	32.9	< 0.001
Fulltime female	0.259	0.317	0.7	0.415	−1.293	0.261	24.5	< 0.001
Fulltime male	0.335	0.208	2.6	0.108	−0.424	0.181	5.5	0.019
Generation X	5.262	2.321	5.2	0.023	11.872	1.674	50.3	< 0.001
Wealthy	0.650	0.514	1.6	0.206	−0.495	0.354	1.9	0.163

(3) Local environment								
Density	0.000	0.000	0.6	0.421	−0.000	0.000	8.4	0.004
Home value	0.210	0.411	0.3	0.609	0.148	0.248	0.4	0.551
Households	0.000	0.000	29.7	< 0.001	0.000	0.000	388.0	< 0.001
Land area	0.000	0.000	7.1	0.007	0.000	0.000	0.5	0.462
Large house	1.209	0.583	4.3	0.038	0.554	0.531	1.1	0.297
Urban housing	1.367	0.213	41.2	< 0.001	0.586	0.151	15.0	< 0.001
(4) Access to retail services								
Distance to convenience	0.000	0.005	0.0	0.952	0.002	0.005	0.1	0.713
Distance to drug	−0.015	0.005	11.5	0.001	−0.004	0.004	1.0	0.306
Distance to general	−0.009	0.004	5.8	0.016	−0.004	0.003	1.7	0.186
Distance to supermarket	−0.024	0.006	14.5	< 0.001	−0.006	0.006	0.9	0.342
Distance to warehouse club	0.011	0.004	7.1	0.008	0.002	0.003	0.4	0.531
	[a]Estimate of σ (Eq. 10) = 0.575; LR test: $\rho = \frac{\sigma^2}{1+\sigma^2} = 0$ yields $\chi^2_1 = 161.1$, $p < 0.001$.				[a]Estimate of σ (Eq. 10) = 0.954; LR test: $\rho = \frac{\sigma^2}{1+\sigma^2} = 0$ yields $\chi^2_1 = 412.5$, $p < 0.001$.			

Table 7 Time variation in θ

Variable	Coefficient	Std err	Wald χ^2	p-value
Model intercept[a] (α_0)	−8.767	0.282	968.5	< 0.001
Population control $\log(n_z)$ (ϕ)	0.693	0.020	1256.7	< 0.001
Broadband access	0.078	0.013	37.5	< 0.001
Neighborhood effect (θ)	0.408	0.083	24.2	< 0.001
NE ∗ Month 6–10	−0.126	0.088	2.0	0.152
NE ∗ Month 11–15	−0.271	0.084	10.4	0.001
NE ∗ Month 16–20	−0.285	0.085	11.3	0.001
NE ∗ Month 21–25	−0.328	0.089	13.7	< 0.001
NE ∗ Month 26–30	−0.220	0.086	6.6	0.010
NE ∗ Month 31–35	−0.277	0.086	10.4	0.001
NE ∗ Month 36–40	−0.368	0.087	17.9	< 0.001
NE ∗ Month 41–45	−0.326	0.088	13.8	< 0.001
Region level covariates				
(1) Household characteristics				
Blacks	−0.696	0.077	81.9	< 0.001
Foreign	−0.064	0.210	0.1	0.762
Hispanics	−0.588	0.144	16.7	< 0.001
Large family	−2.654	0.289	84.1	< 0.001
Solo female	−2.848	0.851	11.2	0.001
Solo male	5.315	0.728	53.3	< 0.001
(2) Household economics				
College	3.272	0.243	180.9	< 0.001
Elderly	−1.224	0.300	16.6	< 0.001
Fulltime female	−0.133	0.192	0.5	0.487
Fulltime male	−0.099	0.138	0.5	0.472
Generation X	5.730	1.342	18.2	< 0.001
Wealthy	0.101	0.272	0.1	0.712
(3) Local environment				
Density	0.000	0.000	0.3	0.604
Home value	−0.466	0.194	5.8	0.016
Households	0.000	0.000	12.0	0.001
Land area	0.000	0.000	1.1	0.294
Large house	0.726	0.421	3.0	0.084
Urban housing	0.331	0.111	8.9	0.003
(4) Access to retail services				
Distance to convenience	0.005	0.003	2.3	0.131
Distance to drug	−0.009	0.003	10.8	0.001
Distance to general	−0.002	0.002	1.1	0.296
Distance to supermarket	−0.026	0.006	22.8	< 0.001
Distance to warehouse club	0.008	0.002	10.0	0.002

in terms of the first trial in each region, and to pool data across regions by implicitly putting them on the same "scale."

5 Discussion and conclusion

According to the 2007 Statistical Abstract compiled by the US Census Bureau, online retail sales were forecast at $95.3 billion for 2006, which represents

a nearly 20 percent increase over 2005. Our study therefore addresses a new and important process: Space-time evolution of trials for an e-tailer. Consumer behaviors in traditional retail formats have been studied extensively, but relatively little is known about how e-tailers acquire customers. We develop a theoretical rationale for neighborhood effects on individual trial decisions, along with a statistical approach to test for them in the absence of strictly individual-level data. We exploit the relationship between utility maximization and a discrete time hazard model to examine the first trial in a region. Our choice of distributional assumption on utility allows us to combine data across regions by rescaling the latent utility of the unobserved maximal individual, to account for the number of individuals present in the region. Neighborhood effects are then examined at the region level where neighbor relations are properly defined, covariates are available, and the assumption of eventual trial is much more reasonable.

5.1 Substantive findings and implications

Researchers have speculated that the Internet could cause individuals to become more diffuse and solitary in their behavior (e.g., Townsend 2001; see also Van Alstyne and Brynjolfsson 1996, 2005). Conversely, our empirical findings are consistent with the idea that social interaction grounded in physical proximity stimulates trial of a new Internet service. We explored neighborhood effects in this context through a variety of measures of first-order physical contiguity, under a number of different assumptions. In each instance, the estimate was positively signed and significant.

The estimate of θ given in Table 4 implies that an additional prior trial by an individual residing in a neighborhood adjacent to the focal region translates to a roughly 19 percent proportional increase in the baseline hazard for the average region. (For the proportional hazards model this effect is simply $\exp(\theta) = 0.185$). To understand the implications of the model for the probability of an individual trial at Netgrocer, consider the following example. Imagine a zip code with the following characteristics: 8,700 individuals inhabitants (the average value) and 48,000 adjacent neighbors (recall that the average zip code has approximately 5.6 neighboring zip codes). The point estimates imply that the marginal effect of going from zero to 20,000 neighbors who have tried netgrocer increases the point estimate probability of an individual trying netgrocer from 1/12,000 to 1/2,000. The corresponding point estimate of the probability of trial within this zip code now increases from approximately 2.7% to 14.0%.

The demise of some 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; Tanskanen et al. 2002). Our research suggests that even when shipping is handled by third party specialists, customer density is still important because neighborhood effects help stimulate new trials. We also show that region characteristics are managerially useful segmentation variables as speed to trial is strongly

influenced by education levels, population density, extent of urbanization, access to retail services, and household composition.

5.2 Future research

We view this research as a first step and several issues remain. One could:

- Broaden the affiliation concept beyond that rooted in first-order geographical contiguity. This could involve new variations on the retailer interaction models discussed in Bronnenberg and Mahajan (2001). Graff and Ashton (1993) find a reverse hierarchical process (from rural to urban areas) in the pattern of store openings by Walmart. Ter Hofstede et al. (2002) show that customer similarities may reach across geographical boundaries. All these approaches could shed light on how e-tailers gain customers.
- Distinguish between effects that arise from direct word of mouth, observational learning, and electronic communication. Godes and Mayzlin (2004) show the power of electronic word of mouth in predicting television viewing habits.
- Study purchase volumes. Chandrashearan and Sinha (1995) or repeat behaviors (Urban 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. Moreover, the neighborhood effect disappears for the repeat decision. Individuals, having tried, appear to rely on their personal cost-benefit assessment in deciding to repeat.

We intend to visit these issues in future research.

Appendix

First, we show that the discrete time hazard with a complementary log-log link function mirrors an underlying continuous time proportional hazards model.[22] Second, we show how to incorporate heterogeneity into the model using Normal mixing (estimates from this model are given in Table 4).

A Complementary log-log link function

Beginning with the continuous time hazard

$$\lambda_z(t) = \lim_{h \to 0} \frac{P(t < T_z \leq t + \Delta | T_z \geq t)}{\Delta}, \tag{13}$$

we obtain the standard parameterized proportional hazards form

$$\lambda_z(t) = \lambda_0(t) \exp(x_z(t)' \beta). \tag{14}$$

[22] We are grateful to anonymous reviewers for providing references that led to this Appendix.

Here $\lambda_0(t)$ is the baseline hazard at time t, which is unknown, $x_z(t)$ is a vector of time-dependent explanatory variables for region z, and β is a vector of unknown parameters. The probability that a spell lasts until time $t+1$ given that it has lasted until t is easily written as a function of the hazard

$$P(T_z \geq t+1 | T_z \geq t) = \exp\left[-\int_t^{t+1} \lambda_z(u)du\right]$$
$$= \exp\left[-\exp(x_z(t)'\beta) \cdot \int_t^{t+1} \lambda_0(u)du\right] \quad (15)$$

given that $x_z(t)$ is constant between t and $t+1$. Equation 15 can be written as

$$P(T_z \geq t+1 | T_z \geq t) = \exp\left[-\exp((x_z(t)'\beta) + \gamma(t))\right] \quad (16)$$

where

$$\gamma(t) = \log\left[\int_t^{t+1} \lambda_0(u)du\right]. \quad (17)$$

This last expression shows that $\gamma(t)$ is the logarithm of the integrated continuous time hazard. It then follows that the analog of Eq. 7 in the paper is simply

$$P(t \leq T_z < t+1 | T_z \geq t) = 1 - \exp\left[-\exp((x_z(t)'\beta) + \gamma(t))\right] \quad (18)$$

which is the claimed complementary log-log link function used for all models in the paper with time variation in the baseline hazard. The interested reader is referred to Seetharaman and Chintagunta (2001, equations 14 and 23) and Meyer (1990) for additional details. A second and final important point about this modeling choice concerns the distinction between the observation interval—which we define exogenously as 1 month—and what one might called the consumer decision interval (which is unobserved by the analyst). Krishnan and Seetharaman (2002, see equations 8 and 9) provide a very nice result which links the decision interval and the exogenous observation interval. They show that the complementary log-log link function results as a limiting case when decisions are made instantaneously. Absent other information, instantaneous decision making at the zip code level is a reasonable approximation for Internet shopping behavior.

B Heterogeneity

While the discrete time model just shown incorporates time variation in the baseline via $\gamma(t)$, one may also wish to allow for heterogeneity over observational units. In our case, we also need to allow for variation in the baseline over regions $z = 1, 2, \ldots, Z$. Our most general model accomplishes this and the estimates are given in Table 4. Technical details for arriving at this specification are given below.

First, as noted in the paper, the data are organized into "sequential binary response" form (Prentice and Gloeckler 1978; Han and Hausman 1990). The panel is arranged such that region z contributes T_z observations where $1 \leq$

$T_z \leq 45$. Let t index observations for zip code z such that $t = 1, 2, \ldots, T_z$. We assume proportional hazards and introduce a positive-valued random variable or mixture ν

$$\begin{aligned}\lambda_z(t, \nu_z) &= \lambda_0(t)\nu_z \exp(x_z(t)'\beta) \\ &= \lambda_0(t) \exp(x_z(t)'\beta + u_z), \end{aligned} \tag{19}$$

where $\lambda_0(t)$ is the baseline hazard, $x_z(t)$ is the same vector of observable covariates as above, and $u \equiv \log \nu$ has density $f_u(u)$. The likelihood $L_z(\beta, \gamma)$ for each region with observed covariates $x_z(t)$ in this "mixed proportional hazards" model is

$$\begin{aligned} L_z(\beta, \gamma) = \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_z} P_z(t, u_z)^{y_z(t)} [1 - P_z(t, u_z)]^{1-y_z(t)} \right] \\ \times f_u(u_z) du_z, \end{aligned} \tag{20}$$

$$\text{where } P_z(t, u_z) = 1 - \exp(-\exp(x_z(t)'\beta + \gamma(t) + u_z)) \tag{21}$$

Because of the proportional hazards assumption, the covariates affect the hazard via the complementary log-log link. In estimation $x_z(t)$ also includes the state or two-digit zip code fixed effects. As shown in the first section of this Appendix, the $\gamma(t)$ are interpreted as the log of a non-parametric piecewise linear baseline hazard.

We model unobserved heterogeneity over z using Normal mixing. As noted earlier we also experimented with non-parametric mixing (Heckman and Singer 1984b), however the impact of the covariates and the shape of the baseline hazard are very similar, but the fit slightly worse. Thus, results given in Table 4 are based on Normal mixing.

Acknowledgements We thank seminar participants at Baruch College, Carnegie Mellon University, Dartmouth College, HEC Paris, Humboldt University, Massachusetts Institute of Technology, Singapore Management University, Stanford University, Tilburg University, University of Auckland, UC San Diego, University of Houston, University of Michigan, University of Southern California, University of Texas at Dallas and the 2004 Invitational Choice Symposium for insights and comments, and Netgrocer CEO Lisa Kent for providing data and support. Detailed comments were provided by Eric Bradlow, Peter Danaher, Steve Hoch, Ka Kok Lee, Phillip Leslie, Gary Russell, Seenu Srinivasan, Christophe Van den Bulte and John Zhang. Financial support for Song was generously provided by WeBI, the Wharton School eBusiness Initiative and the Marketing Science Institute through the Alden G. Clayton Doctoral Dissertation Proposal Award. We are very grateful to the Editor and two anonymous *QME* reviewers for several excellent suggestions that improved the paper. Any remaining errors are, of course, our own.

References

Allison, P. D. (1982). Discrete time methods for the analysis of event histories. In S. Leinhardt (Ed.), *Sociological Methodology 1982* (pp. 61–98) . San Francsico, CA: Jossey-Bass.

Allison, P. D. (2001). *Survival analysis using the SAS system: A practical guide*. Cary, NC: SAS Institute.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Norwell, MA: Kluwer.

Besley, T., & Case, A. (1993). Modeling technology adoption in developing countries. *American Economic Review 83*(2), 396–402.
Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads and informational cascades. *Journal of Economic Perspectives*, *12*(3), 151–170.
Brock, W. A., & Durlaf, S. N. (2001). Discrete choice with social interactions. *Review of Economic Studies*, *68*, 235–260.
Bronnenberg, B., & Mahajan, V. (2001). Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Science, 20*(3), 284–299.
Bronnenberg, B. J., & Mela, C. F. (2004). Market roll-out and retailer adoption for new brands. *Marketing Science*, *23*(4), 500–518.
Burt, R. S. (1980). Models of network structure. *Annual Review of Sociology*, *6*, 79–141.
Case, A. C. (1991). Spatial patterns in household demand. *Econometrica*, *59*(4), 953–965.
Case, A. C., Hines, J. R. Jr., & Rosen, H. S. (1989). Copycatting: Fiscal policies of states and their neighbors. NBER Working Paper No. 3032.
Case, A. C., Hines, J. R. Jr., & Rosen, H. S. (1993). Budget spillovers and fiscal policy interdependence: Evidence from the states. *Journal of Public Economics*, *52*(3), 285–307.
Chandrashekaran, M., & Sinha, R. K. (1995). Isolating the determinants of innovativeness: A split-population tobit (SPOT) duration model of timing and volume of first and repeat purchase. *Journal of Marketing Research*, *32*(3), 444–456.
Chaves, M. (1996). Ordaining women: The diffusion of an organizational innovation. *American Journal of Sociology*, *101*(4), 840–873.
Conley, T. G. (1999). GMM estimation with cross-sectional dependence. *Journal of Econometrics*, *92*, 1–45.
Conley, T. G., & Molinari, F. (2007). Spatial correlation robust inference with errors in location or distance. *Journal of Econometrics*, doi:10.1016/j.jeconom.2006.09.003. (forthcoming)
Deighton, J. (2001). *WebVan*. Boston, MA: Harvard Business School Press.
Dhar, S., & Hoch, S. J. (1997). Why store brand penetration varies by retailer. *Marketing Science*, *16*(3), 208–227.
Foster, A. D., & Rosenweig, M. R. (1995). Learning by doing and learning from others: Human capital and technological change in agriculture. *Journal of Political Economy*, *103*(6), 1176–1209.
Fotheringham, S. A. (1988). Consumer store choice and choice set definition. *Marketing Science*, *7*(3), 299–310.
Gail, M. H., Weiand, S., & Piantadosi, S. (1984). Biased estimates of treatment effects in randomized experiments with non-linear regressions and omitted covariates. *Biometrika*, *71*(3), 431–444.
Garber, T., Goldenberg, J., Libai, B., & Muller, E. (2004). From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science*, *23*(4), 419–428.
Godes, D., & Mayzlin, D. (2004). Using online conversations to study word of mouth communication. *Marketing Science*, *23*(4), 530–534.
Goolsbee, A., & Klenow, P. J. (2002). Evidence on learning and network externalities in the diffusion of home computers. *Journal of Law and Economics*, *45*(2), 317–342.
Greve, H. R., Strang, D., & Tuma, N. B. (1995). Specification and estimation of heterogeneous diffusion models. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 377-420). Cambridge, MA: Blackwell.
Graff, T. O., Ashton D. (1993). Spatial diffusion of Wal-Mart: Contagious and reverse hierarchical elements. *Professional Geographer*, *46*(1), 19–29.
Han, A., & Hausman, J. A. (1990). Flexible parametric estimation of duration and competing risk models: Summary. *Journal of Applied Econometrics*, *5*(1), 1–28.
Heckman, J., Singer, B. (1984a). Econometric duration analysis. *Journal of Econometrics*, *24*(1,2) 63–132.
Heckman, J., Singer, B. (1984b). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*, 271–318.
Howard, J. A., & Sheth, J. N. (1969). *The theory of buyer behavior*. New York, NY: Wiley.
Huff, D. (1964). Redefining and estimating a trading area. *Journal of Marketing*, *28*(3), 34–38.

Krishnan, T. V. & Seetharaman, P. B. (2002). A flexible class of purchase incidence models. *Review of Marketing Science*, *1*(3), Working paper 4.
Lilien, G. L., Kotler, P., & Moorthy, S. (1993). *Marketing models*. Englewood Cliffs, NJ: Prentice Hall.
Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, *60*(3), 531–542.
Maritz, J. S., & Munro, A. H. (1967). On the use of the generalised extreme-value distribution in estimating extreme percentiles. *Biometrics*, *23*(1), 79–103.
Meyer, B. D. (1990). Unemployment insurance and unemployment spells. *Econometrica*, *58*, 757–782.
Oyen, O., & De Fleur, M. L. (1953). The spatial diffusion of an airborne leaflet message. *American Journal of Sociology*, *59*(2), 144–149.
Prentice, R., & Gloeckler, L. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, *34*(1), 57–67.
Ryu, K. (1994). Group duration analysis of the proportional hazard model: Minimum chi-squared estimators and specification tests. *Journal of the American Statistical Association*, *89*(428), 1386–1397.
Ryu, K. (1995). Analysis of a continuous-time proportional hazard model using discrete duration data. *Econometric Reviews*, *14*(3), 299–313.
Seetharaman, P. B. & Chintagunta, P. K. (2003). The proportional hazard model for purchase timing: A comparison of alternative specifications. *Journal of Business and Economic Statistics*, *21*(3), 368–382.
Singer, B., & Spilerman, S. (1983). The representation of social processes by Markov models. *American Journal of Sociology*, *82*(1), 1–54.
Strang, D., & Tuma, N. B. (1993). Spatial and temporal heterogeneity in diffusion. *American Journal of Sociology*, *99*(3), 614–639.
Ter Hofstede, F., & Wedel, M. (1999). Time aggregation effects on the baseline of continuous-time and discrete-time hazard models. *Economics Letters*, *63*(5), 145–150.
Ter Hofstede, F., Wedel, M., & Steenkamp, J.-B. E. M. (2002). Identifying spatial segments in international markets. *Marketing Science*, *21*(2), 160–177.
Tanskanen, K., Yrjola, H., & Holstrom, J. (2002). The way to profitable Internet grocery retailing—six lessons learned. *International Journal of Retail and Distribution Management*, *30*(4), 169–178.
Tolnay, S. E., Deane, G., & Beck, E. M. (1996). Vicarious violence: Spatial effects on southern lynchings, 1890-1919. *American Journal of Sociology*, *102*(3), 788–815.
Townsend, A. M. (2001). Network cities and the global structure of the Internet. *The American Behavioral Scientist, 44*(10), 1697–1716.
Urban, G. L. (1975). Perceptor—A model for product positioning. *Management Science*, *21*(8), 858–870.
Van den Bulte, C., & Lilien, G. L. (2001). *Medical Innovation* revisited: Social contagion versus marketing effort. *American Journal of Sociology*, *106*(5), 1409–1453.
Van Alstyne, M., & Brynjolfsson, E. (1996). Could the Internet balkanize science? *Science*, *274*(5292), 1479–1480.
Van Alstyne, M., & Brynjolfsson, E. (2005). Global village or cyber-balkans? Modeling and measuring the integration of electronic communities. *Management Science*, *51*(6), 851–868.
Yang, S., & Allenby, G. M. (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research 40*(3), 282–294.