CAPTURING EVOLVING VISIT BEHAVIOR IN CLICKSTREAM DATA

Wendy W. Moe
Peter S. Fader

ABSTRACT

Many online sites, both retailers and content providers, routinely monitor visitor traffic as a useful measure of their overall success. However, simple summaries such as the total number of visits per month provide little insight about individual-level site-visit patterns, especially in a changing environment such as the Internet. This article develops an individual-level model for evolving visiting behavior based on Internet clickstream data. We capture cross-sectional variation in site-visit behavior as well as changes over time as visitors gain experience with the site. In addition, we examine the relationship between visiting frequency and purchasing propensity at an e-commerce site. We find evidence supporting the notion that people who visit a retail site more frequently have a greater propensity to buy. We also show that changes (i.e., evolution) in an individual's visit frequency over time provides further information regarding which customer segments are likely to have higher purchasing conversion rates.

JOURNAL OF INTERACTIVE MARKETING
VOLUME 18 / NUMBER 1 / WINTER 2004
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/dir.10074

WENDY W. MOE is assistant professor of marketing at the University of Texas at Austin; e-mail: wendy.moe@mccombs.utexas.edu

PETER S. FADER is the Frances and Pei-Yuan Chia professor of marketing at the Wharton School of the University of Pennsylvania; e-mail: faderp@wharton.upenn.edu

We thank the Wharton e-Business Initiative and Media Metrix, Inc. for generously providing the data used in this study. We also acknowledge Bruce Hardie and Eric Bradlow for their useful comments and suggestions. This article stems from Wendy Moe's dissertation work, and she extends special thanks to the other members of her dissertation committee (Barbara Kahn, Don Morrison, & David Schmittlein) and the Marketing Science Institute for supporting this research through the 1999 Alden Clayton Dissertation Award.

TABLE 1
Summary of Visit Data Over Time

	Amazon		*CDNOW*	
	Months 1–4	*Months 5–8*	*Months 1–4*	*Months 5–8*
Total number of visits	5,238	6,025	1,729	1,890
Number of unique visitors	2,645	2,756	988	920
Visits/visitor	1.98	2.19	1.75	2.05

INTRODUCTION

Ever since the Internet first emerged as a viable medium for commercial and informational purposes, analysts have closely tracked visitor traffic as a principal yardstick to gauge the success of online retail and content sites. In the early days of e-commerce, simple measures of *unique visitors* served as a proxy for site performance. Many sites were content to "buy eyeballs" under the assumption that these visitors would stay with the site long enough to justify the costs of attracting them there in the first place. But most retailers subsequently learned (the hard way, in many cases) that this is not a useful indicator; instead, measures of visitor retention and loyalty have proven to be more closely linked to the health of their businesses. Accordingly, measures such as *visits per visitor* have gained more prominence and are now highlighted in the e-commerce "scorecards" that firms such as comScore and Nielsen/NetRatings routinely publish. Whether a site is attempting to sell products or is primarily interested in attracting and retaining a set of regular readers (e.g., for a content site), this type of measure is widely accepted as the among the most diagnostic indicators of site performance.

As a specific example, Table 1 provides an aggregate summary of the dynamics in visit patterns for two prominent online retailers, Amazon and CDNOW. For both sites, not only do we see growth in the total number of visits over time but there also appears to be an increase in the number of visits per visitor. On the surface, these aggregate measures seem like great news for the site managers. However, these numbers may be misleading as the customer base is changing with the influx of new visitors (perhaps with relatively high visit rates) and the exit of more experienced users, potentially masking the true visit dynamics that exist.

As a consequence, any attempt to summarize intervisit times directly from observed data may be unable to provide accurate estimates of the true, underlying rates of repeat visiting. The only way to overcome these selection biases while still ensuring the representativeness of the entire set of visitors is to use a well-specified individual-level model (with suitable assumptions about heterogeneity and nonstationarity) to obtain valid inferences about differences in visit patterns across people and over time.

Therefore, the primary objective of this article is to develop a probabilistic model that carefully sorts out all of these issues. An important (and unique) aspect of this model is the manner in which we allow for *evolving behavior* in visitor traffic. Traditional stochastic models of purchasing behavior assume that purchase rates are unchanging over time (e.g., Morrison & Schmittlein, 1988). When these models are tested in stable and mature markets, such an assumption may indeed hold. But many new markets go through a state of flux for quite some time (Bronnenberg, Mahajan, & Vanhonacker, 2000). In other words, an individual's visit behavior often changes as she continually adapts to a new environment. The model presented here will relax the usual assumption of stationarity.

From our evolving visit model, we can estimate how likely (and when) a given customer

will return to the site as he or she gains experience there. Do intervisit times tend to speed up or slow down over a person's history, and how do these changes vary across people? Answers to these questions will give us the ability to forecast future visits to better anticipate and manage Web site traffic. We will show that our evolving model of visiting behavior forecasts future traffic patterns significantly better than an equivalent static model. Additionally, the evolutionary component of the model will offer useful diagnostics that will help shed light on other aspects of online visit behavior. For example, do frequent customers necessarily comprise the most valuable segment for targeting purposes?

Though Internet clickstream data is rich with behavioral information such as duration of visits, number of page views, characteristics of items viewed, and so on, we examine only the timing and frequency of site visits, as understanding visitor traffic at this level is an important managerial issue in itself. However, despite the limited data that we use, we find that simple visiting rates (and trends in these rates) are strong indicators of an individual's buying propensity at an e-commerce site. As we better describe customers in terms of their visiting behavior, we relate visiting frequency to purchasing propensity. Previous studies suggest that people who shop frequently may be more likely to make a purchase on any given shopping occasion (Bellinger, Robertson, & Hirschman, 1978; Janiszewski, 1998; Jarboe & McDaniel, 1987; Roy, 1994). As a result, frequent visitors are often the preferred target segment. Our clickstream analysis strongly confirms this hypothesis, and then extends it by showing that *changes* (i.e., evolution) in an individual's visit frequency over time provide even better information regarding which customers (and customer segments) are more likely to buy. Rather than simply targeting all frequent visitors, our results suggest that a more refined segmentation approach that incorporates how much an individual's behavior is changing can more efficiently identify profitable customers for targeting purposes.

In the next two sections, we develop the model and address some of the key estimation issues that arise from the model. We then describe the clickstream data that we will be using. In Section 5, we present the results of the model as applied to Amazon and CDNOW, and validate the model by demonstrating its forecasting ability over a 4-month holdout period. In Section 6, we compare our model to other benchmark models. Finally, in Section 7, we illustrate how purchasing behavior varies across visitors as a function of their latent visit rates as well as changes in these rates over time.

MODEL DEVELOPMENT

To understand the overall pattern of site visits, let us imagine that each customer tends to return to a site in accordance with a latent visit rate inherent to that individual. *When* that individual will visit the site next is driven largely by this rate of visit. Additionally, since customers are heterogeneous, this rate of visit varies from person to person. Some people may visit the site fairly frequently while others may not. But in addition to varying rates of visit across individuals, behavior also may change over time for a given individual. As customers mature, perhaps as a result of increased knowledge and experience, their behavior may evolve thereby changing their rates of visit over time. The model presented in this section will attempt to capture, describe, and measure the degree of this behavioral evolution. We will refer to this nonstationary model as the evolving visit (EV) model.

To capture the processes described previously, our model has three main components: (a) a timing process governing an individual's rate of visiting, (b) a heterogeneity distribution that accommodates differences across people, and (c) an evolutionary process that allows a given individual's underlying visit rate to change from one visit to the next.

As an appropriately robust starting point, repeat visit behavior can be modeled as an exponential-gamma (EG) timing process. That is, each individual's intervisit time is assumed to be exponentially distributed governed by a rate, λ_i. Furthermore, these individual rates of visit vary across the population. This heterogeneity can be captured by a gamma distribution with shape

parameter, r, and scale parameter, α. These distributions are given by the following two densities:

$$f(t_{ij};\ \lambda_i) = \lambda_i e^{-\lambda_i(t_{ij} - t_{i(j-1)})} \quad \text{and}$$

$$g(\lambda_i;\ r,\ \alpha) = \frac{\lambda_i^{r-1}\alpha^r e^{-\alpha\lambda_i}}{\Gamma(r)} \quad (1)$$

where λ_i is individual i's latent rate of visit, t_{ij} is the day when the jth repeat visit occurred, and t_{i0} is the day of their first observed visit. For a single visit occasion, this leads to the following familiar exponential-gamma mixture model:

$$f(t_{ij};\ r,\ \alpha) = \int_0^\infty f(t_{ij};\ \lambda_i) \cdot g(\lambda_i;\ r,\ \alpha)\, d\lambda$$

$$= \frac{r}{\alpha}\left(\frac{\alpha}{\alpha + (t_{ij} - t_{i(j-1)})}\right)^{r+1} \quad (2)$$

While the EG may be an excellent benchmark model, even in contexts in which the exponential and/or gamma assumptions may be violated (Morrison & Schmittlein, 1988), it does not adequately capture systematic changes in individual-level behavior over time. To account for nonstationarity, a novel extension of this basic model is described next.

Although studies have suggested that store visit behavior may evolve as a function of experience (Johnson & Russo, 1984; Park, Iyer, & Smith 1989), they have been inconclusive in identifying the direction of this evolution. Therefore, our model is a descriptive one that captures and characterizes any evolution that may exist—without taking a one-sided stand on this issue. We develop a flexible model that will accommodate varying magnitudes and directions of the behavioral change and offer a method to characterize the nature of this evolution.

Researchers in marketing have used several different mechanisms to introduce time-varying effects into the traditional stochastic modeling framework. For instance, Sabavala and Morrison (1981) incorporated nonstationarity by introducing a renewal process into a probability mixture model in accordance with the "dynamic inference" framework first set out by Howard (1965). Sabavala and Morrison applied this model to explain patterns of advertising media exposure over time.

While renewal models provide a statistical mechanism to introduce nonstationarity into a timing process, they do not capture the incremental, visit-to-visit changes in behavior that we have described. Instead, renewal models operate under the assumption that each visitor probabilistically discards the old rate parameter and draws an entirely new one from the original heterogeneity distribution, independent of previous values. This process allows for drastic changes in an individual's behavior while maintaining the same heterogeneity distribution for the population as a whole. While this may be a powerful and effective way to improve the fit of the model by accommodating changes over time, it does not in any way describe or measure the degree of such changes either at the population level or at the individual level. Furthermore, in new and developing market environments such as the Internet, it may be incorrect to assume that the overall heterogeneity distribution is not changing over time. The EV model that we propose allows for the population heterogeneity distribution to change as the customers that comprise the population gradually reevaluate their preferences and update their behavior.

Specifically, our behavioral assumption is that customers' underlying rates of visiting are continually and incrementally changing from one visit to the next. A simple way to specify this updating process is as follows:

$$\lambda_{i(j+1)} = \lambda_{ij} \cdot c \quad (3)$$

where λ_{ij} is the rate associated with individual i's jth repeat visit and c is a multiplier that will update this rate from one visit to the next. If the updating multiplier, c, equals 1, visiting rates are considered unchanging, and the stationary EG would remain in effect. But if c is greater than 1, visitors are visiting more frequently as

they gain experience; if c is less than 1, they are visiting less frequently as they gain experience.

However, using a constant multiplier to update the individual λs would be a very restrictive (and highly unrealistic) way of modeling evolutionary behavior in a heterogeneous environment. A more general approach is to replace the scalar multiplier, c, with a random variable c_{ij} to acknowledge that these updates can vary over time and across people. Each individual visit will lead to an update that may increase, decrease, or retain the previous rate of visit depending on the particular sequence of draws of these (stochastic) updating multipliers.

To generalize Equation (3) in this manner, we assume that these probabilistic multipliers, c_{ij}, arise from a gamma distribution, common across individuals and visits, with shape parameter s and scale parameter β[1]:

$$h(c_{ij};\ s,\ \beta) = \frac{c_{ij}^{s-1}\beta^{s}e^{-\beta c_{ij}}}{\Gamma(s)} \qquad (4)$$

This gamma distribution essentially describes the nature of the behavioral evolution faced by a given site. The updated $\lambda_{i(j+1)}$ then becomes a product of two independent gamma-distributed random variables: the previous rate, λ_{ij}, and the multiplier, c_{ij}. This model simultaneously captures cross-sectional heterogeneity and evolving visiting behavior with two parameters (r and α) describing the initial heterogeneity in visiting rates and another two parameters (s and β) describing the updating process.

An interesting characteristic of the updating distribution is that it allows for customer attrition since the gamma distribution can yield a draw of c_{ij} extremely close to zero. When this occurs, the individual's rate of visit falls dramatically and essentially reflects attrition. Such attrition may be very common for Web sites and has been the centerpiece of other types of models in this general methodological area (Fader, Hardie, & Lee, 2003; Schmittlein, Morrison, & Colombo, 1987). The fact that we can accommodate attrition in such a simple, natural manner is an appealing aspect of the proposed modeling approach. In fact, capturing attrition in this manner may be more realistic than using a separate model component. That is, traditional attrition models presume that many customers permanently drop out at some point and never return to the site. The EV model allows a less stringent dropout process where the "inactive" customer merely has a very small probability of returning, but that probability still does exist. Most datasets are not long enough or precise enough to be able to sort out "true" attrition from positive (but near-zero) visit rates.[2]

We also test a model that allows for an interrelationship between an individual's rate of visit and his or her updating process. That is, more frequent visitors may undergo a different updating process than those less frequent visitors. To incorporate this potential correlation between visiting rates and updates, we allow the rate of visit to effectively shift the updating distribution. In other words, we calculate the shape parameter, s, of the updating distribution as a function of the individual's rate of visit, λ.

$$s_{ij} = a + b \cdot \lambda_{ij} \qquad (5)$$

MODEL ESTIMATION

The likelihood function for a stationary EG timing model can be written as follows (Appendix A provides a more detailed discussion of how this likelihood is derived.):

$$L = \prod_{i=1}^{N}\prod_{j=1}^{J_i}\left(\frac{r+j+1}{\alpha + t_{i(j-1)} - t_{i0}}\right)$$

[1] An informal look at the ratio of intervisit times for a given level of repeat to the intervisit times of the last visit cycle suggests that the gamma distribution is an adequate descriptor of visit-to-visit changes.

[2] In addition to our proposed EV model, we also estimated a visit frequency model that included an explicit attrition component much like that specified by Eskin (1973) where the probability of being an active visitor after the jth visit, π, is determined by $\pi = \phi[1-\exp(-\theta j)]$. We find that a visit model with attrition provides a significantly poorer fit than the EV model proposed in this article.

$$\times \left(\frac{\alpha + t_{i(j-1)} - t_{i0}}{\alpha + t_{ij} - t_{i0}} \right)^{r+j} \cdot S(T - t_{ij})$$

$$\text{where} \quad S(T - t_{iJ_i}) = \left(\frac{\alpha + t_{iJ_i} - t_{i0}}{\alpha + T - t_{i0}} \right)^{r+J_i} \quad (6)$$

When we introduce the nonstationary updating distribution, the multipliers (c_{ij}) change the value of λ_i from visit to visit. We need to capture two forms of updating after each visit: one due to the usual Bayesian updating process and the other due to the effects of the stochastic evolution process. Therefore, the distribution of visiting rates at each repeat visit level is the product of two gamma distributed random variables—one associated with the updating multiplier and one capturing the previous visiting rate (with Bayesian updating). For the case of Panelist i making her jth repeat visit at time t_{ij}:

$$G(\lambda_{i(j+1)} | \text{arrival at } t_{ij}) = \text{gamma}(r_{ij} + 1, \alpha_{ij} + t_{ij} - t_{i(j-1)}) \cdot \text{gamma}(s, \beta) \quad (7)$$

One issue with this approach is that the product of two gamma random variables does not lend itself to a tractable analytic solution. However, there is an established result (e.g., Kendall & Stuart, 1977, p. 248) suggesting that the product of two gamma distributed random variables can be approximated by yet another gamma distribution, which can be easily obtained using the parameters of the two distributions. As shown in Appendix B, this approximation[3], used in conjunction with Bayesian updating, allows us to recover the updated gamma parameters that determine the rate of visit, λ_{ij}, for Panelist i's jth repeat visit as follows:

$$r_{i(j+1)} = \frac{\lfloor r_{ij} + 1 \rfloor \cdot s}{[r_{ij} + 2] \cdot (s + 1) - [r_{ij} + 1] \cdot s} \quad (8)$$

$$\alpha_{i(j+1)} = \frac{\lfloor \alpha_{ij} + t_{ij} - t_{i(j-1)} \rfloor \cdot \beta}{[r_{ij} + 2] \cdot (s + 1) - [r_{ij} + 1] \cdot s} \quad (9)$$

After incorporating the evolutionary process into our model, the likelihood function to be maximized is:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J_i} \left(\frac{r_{ij}}{\alpha_{ij}} \right) \left(\frac{\alpha_{ij}}{\alpha_{ij} + t_{ij} - t_{i(j-1)}} \right)^{r_{ij}+1} \cdot S(T - t_{iJ_i}) \quad (10)$$

where $r(i, j)$ and $\alpha(i, j)$ are defined in Equations (8) and (9) while r_{i1} and α_{i1} are estimated and represent the initial values of r and α. The survival function, $S(T - t_{ij})$, is defined as:

$$S(T - t_{ij}) = \left(\frac{\alpha_{i(J_i+1)}}{\alpha_{i(J_i+1)} + T - t_{iJ_i}} \right)^{r_{i(J_i+1)}} \quad (11)$$

For the special case in which behavior is not evolving and the nonstationary updating distribution degenerates to a spike at 1.0 (i.e., $s = \beta = M$, where M approaches infinity), then this equation collapses down exactly to the ordinary (stationary) EG model.

DATA

We apply the models described in the previous section to clickstream data collected by Media Metrix, Inc. in 1998. Media Metrix, whose panel operations were subsequently acquired by comScore Networks, maintained a panel of approximately 10,000 households whose Internet behavior was recorded, pageview by pageview, over time. Participating households installed customized software on their personal computers which recorded the date, time, and duration of each and every page being viewed when they surfed the Internet.

[3] We performed numerous simulations to verify the accuracy of using our approximation. In each simulation, we first generated 1,000 random draws from a gamma distribution with randomly determined shape and scale parameters to represent initial λ values. Then, a matrix of updating multipliers also was simulated for a series of five updates (i.e., five future repeat visits). A Kolmogorov-Smirnov test indicated that, for every one of the simulations, the distribution of values resulting from the moment-matching approximation was not significantly different from those resulting from the direct multiplication of the simulated random variables. Therefore, we are confident that the moment-matching approximation accurately captures the gamma distributed updating process we use in our model.

TABLE 2
Model Results for Amazon

	r	α	s*	β	b	$-LL$	k	$CAIC$
(i) Stationary model	0.48	42.96				34,347	2	68,711
	(0.01)	(0.02)						
(ii) Evolving visits	0.32	16.86	2.30	2.30		33,658	4	67,330
	(0.01)	(0.36)	(0.05)	(0.08)				
(iii) EV with correlation	0.32	16.78	2.28	2.30	0.15	33,648	5	67,337
	(0.01)	(1.32)	(0.04)	(0.02)	(0.04)			

* For Model (iii), this is the a parameter shown in Equation (4).

Despite the age of this dataset, we find it to be very appropriate for two primary reasons: First, it captures a period of great change in Internet usage habits, so it gives us a good glimpse at an unusually important time of evolving behavior. Second, it reflects a time when both focal sites (Amazon & CDNOW) were both "pure-play" retailers, selling only one type of product (books and music CDs, respectively), so we get a clean view of panelist behavior without worrying about mixing in behavioral patterns across a variety of product categories. We do not claim that the specific behaviors we observe here will remain the same for other product categories and time periods, but we have every reason to believe that the same model (albeit with different parameter estimates) will continue to perform well in these other settings.

For our purposes, we are interested in the dates of the visits each panelist makes to a given site. Any session in which the Web user views a URL with a particular domain name (Amazon or CDNOW) is considered a visit to that site. To consolidate the data just a bit, we aggregated visits to the daily level. For example, a visitor may leave a site briefly and return later that day. However, this second visit is unlikely to be considered a repeat visit, but rather an extension of the first visit. Therefore, if a given panelist were to visit a particular site multiple times in a single calendar day (a pattern that rarely happens in our dataset), we would encode that behavior as just one visit for the day when the session began. In some cases, a session may begin before midnight and conclude after midnight on the next calendar day. These sessions would be considered visits that occurred on the day the session began (i.e., when the first page was viewed). Since we are interested in the timing and frequency of repeat visits to a site, our dataset describes each panelist as a sequence of days when visits were made. All panelists who have visited the site of interest at least once during the observation period were included in this dataset. We use data from March 1 to October 31, 1998. During this period, Amazon attracted 4,379 unique visitors to its site during this 8-month period totaling 11,301 visits; CDNOW had 1,670 visitors making 3,619 visits (see Table 1 for more detailed summaries of the data).

MODEL RESULTS

In Table 2, we contrast the parameter estimates and fit statistics for the stationary EG model with those from the EV model (with and without the correlation component). When the static, two-parameter model is applied to the 8 months of Amazon data, we find that the mean rate of visit ($E[\lambda] = r/\alpha$) is 0.011. In other words, the expected intervisit time for an average visitor is 89.5 days, which seems high but is reasonably consistent with the summary statistics mentioned earlier. In contrast, the simple (uncorrelated) EV model has an average intervisit time of only 52.7 days. The main reason for this substantial difference is the fact that the stationary model assumes that all visitors remain active throughout the data period. That is, when an individual drops out and fails to return

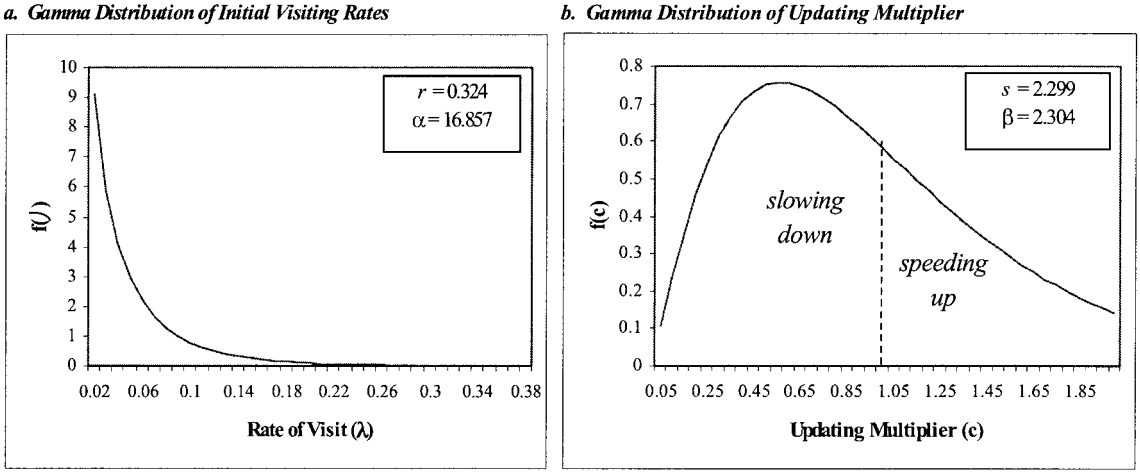

FIGURE 1
Evolving Visit Model Distributions for Amazon Data

within the remainder of the observation period, the stationary model assumes that he or she has an extremely long intervisit time instead of being written off as a customer who is no longer active.

Figure 1 illustrates the gamma distributions of the customers' initial rates of visit as well as the distribution of updating multipliers. Coupled with Bayesian updating after observing an individual's behavior, these distributions are what allow us to estimate when the visitor will return to the site.

According to the EV model, the mean update for any given visit (s/β) is very close to 1 (0.998) suggesting, perhaps, that it is a fairly stationary process. However, a closer look at the distribution (see Figure 1b) shows that there is significant variance about this mean. Though the mean update is close to 1, the distribution is quite skewed. With a median value of c_{ij} = 0.858, visitors tend to return to the site at slower rates from visit to visit. The implications of these results are in stark contrast to the measures summarized in Table 1 that implied increased visiting frequency over time. The distribution presented in Figure 1b seems to suggest that Amazon was gradually losing favor with its active customers. More visitors tend to be less satisfied with each subsequent visit and, as a result, gradually reduce their visit frequency.

Finally, a separate observation from Table 2 is that there is little appreciable correlation between an individual's rate of visit and the updates. This is shown by the relatively small improvement in fit that occurs when such correlation is taken into account (Model iii).

Validation

In this section, we validate the EV model by examining the accuracy of longitudinal forecasts. Because the EV model relies on an approximation to specify and estimate the model, we need to perform simulations to generate data for tracking/forecasting purposes. This is a straightforward and computationally efficient task. For each iteration of the simulation, we create a simulated panel that matches the actual panel in terms of its size and the distribution of its initial visit times. We then generate a sequence of repeat visits using the parameter estimates from the model. This requires us to maintain a time-varying vector of λs for each panelist, which starts with random draws from the initial (r, α) gamma distribution, and then gets updated using the (s, β) gamma distribution after each simulated exponential arrival occurs. We continue this process until every simulated panelist gets past the tracking/forecasting horizon of interest to us. It is then a simple matter to count the number of visits on a week-by-week basis for each iteration of the simulation. We then average across 1,000 itera-

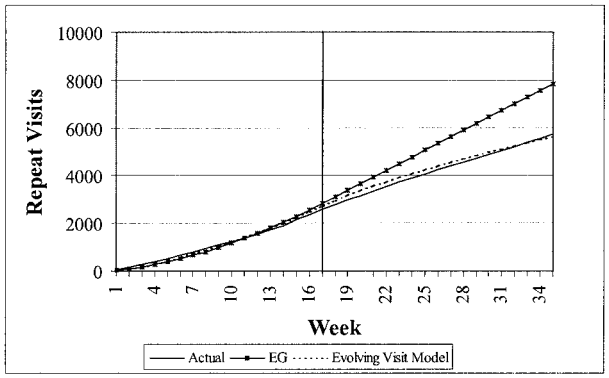

FIGURE 2
Forecasts of Repeat Visits

tions to generate the tracking and forecasting plots.

Before creating the forecasts, we reestimate both models (stationary and evolving EG) using only the first half (i.e., 4 months) of the dataset. To generate the forecasts for the EV model, we use the simulation procedure described previously. For the stationary EG model, the expected number of repeat visits per week can be calculated directly as follows:

$$E[\text{repeat visits}_w] = N_w z\left(\frac{r}{\alpha}\right) \qquad (12)$$

where N_w is the number of eligible repeat visitors in Week w and z is the time period of interest, i.e., 7 days in this case. Figure 2 shows cumulative forecasts as well as actual visits for the Amazon site.

Both models seem to track the data quite well over the initial 4-month calibration period. However, as we enter the forecasting period, the stationary EG model begins to diverge, ultimately overpredicting by 37% for Amazon at the end of the 8-month period. This model overestimates the number of visits per week because it does not recognize that visitors are returning less frequently over time. The EV model, however, forecasts quite accurately, well within 5% of the actual sales line throughout the forecast period.

Results for CDNOW

The same set of models and analyses also were applied to CDNOW data (results in Table 3). We see a remarkably similar set of patterns as in the case of Amazon. In moving from the static EG model to the EV specification, we see significantly shorter intervisit times, and the mean update is close to 1.0 (0.991). But with a median of 0.837, customer visit frequency is more likely to decrease than increase after each visit—once again contradicting the summary statistics from Table 1. Again, no appreciable correlation between visit frequency and the updating multiplier seems to exist.

Finally, our forecast validation offers encouraging results, with projected visits only 2% above the actual number at the end of the 8-month period compared to a 40% overforecast for the stationary model. While we are very encouraged by these strong initial results, we also are surprised at the degree of similarity seen for these two sites. We certainly do not want to suggest that the specific patterns cap-

TABLE 3
Model Results for CDNOW

	r	α	s*	β	b	$-LL$	k	$CAIC$
(i) Stationary model	0.26	28.31				9,460	2	18,934
	(<0.01)	(0.04)						
(ii) Evolving visits	0.16	8.89	2.08	2.10		9,121	4	18,271
	(0.01)	(0.96)	(0.21)	(0.18)				
(iii) EV with correlation	0.16	8.62	1.98	2.08	0.31	9,119	5	18,274
	(0.01)	(0.68)	(0.17)	(0.19)	(0.13)			

* For Model (iii), this is the a parameter shown in Equation (4).

tured here will generalize to all online sites, but this should be ample motivation for future studies to find and describe a broader range of online visiting behavior.

OTHER BENCHMARK MODELS

In addition to using variations of the exponential timing process as benchmarks, we also tested alternative specifications. The exponential distribution assumes a memoryless process. However, in many situations, there is some regularity in intervisit times. Therefore, we also estimate the proposed EV model replacing the exponential distribution with an Erlang-2 distribution, which allows for more regularity. Again, we incorporate gamma heterogeneity in the rate parameter, λ, and estimate two versions of the model—one with and one without the updating process described. In both cases (i.e., with and without updating), the EV model that uses an underlying exponential timing process outperforms its Erlang-2 counterpart in terms of likelihood (see Table 4 for log-likelihoods resulting from each of the benchmark models presented in this section). This suggests that the timing of online visits at Amazon and CDNOW are fairly random and do not exhibit a high degree of regularity.

Another possibility is that, as the online consumer population matures, the site visit process evolves toward the more regular pattern represented by the Erlang-2 distribution. That is, current behavior may be best modeled using an exponential process, but over time there may be a shift toward a Erlang-2 process. To test this possibility, we divide the data set into two 4-month periods and estimate both an exponential model and an Erlang-2 model (both with gamma heterogeneity) on the early half first and then the later half of the data. We see in both halves of the data that the exponential-based model outperforms the Erlang-2 based model, suggesting that the evolutionary process in the data is not one that can be described as a shift toward regularity.

Finally, as an alternative to the variety of continuous-time models that we have considered so far, we also examine a discrete-time duration model to describe the visit patterns. Instead of asking "How long will it be until this customer next visits the focal site?" we can ask "Will the next visit occur sometime today?" In other words, we use each day as the unit of observation (and record whether a visit occurs) instead of modeling the length of each intervisit gap. Seetharaman and Chintagunta (2003) provided an excellent review of timing models and a thorough comparison of various model specifications. For example, they found that a log-logistic baseline hazard specification, where $F(t) = (\gamma t)^{\alpha}/[1 + (\gamma t)^{\alpha}]$ provides one of the best fits when modeling the timing of offline grocery store visits. In our application, the log-logistic duration model performs less well than the proposed EV model in terms of log-likelihood for both sites. Again, we conclude that the timing of online site visits is best described by a continuous-time exponential process that evolves through the proposed updating process as the visitor makes repeated visits.

VISIT FREQUENCY AND EVOLUTION: ASSOCIATIONS WITH PURCHASING BEHAVIOR

There is considerable evidence (theoretical and empirical) implying that more frequent visitors also are more likely buyers at any given visit occasion (see Bellinger et al., 1978; Janiszewski, 1998; Jarboe & McDaniel, 1987; Roy, 1994). In this section, we explore this relationship between customers' visiting patterns and their purchasing propensities. We then extend the framework to incorporate (and separate out) the effects of evolving behavior on purchasing.

As an initial test of the traditional frequency–propensity hypothesis, we first calculate each panelist's expected rate of visit, λ_i, given the EV model's estimated parameters and the panelists' observed behavior during the 8-month observation period. We then calculate each repeat visitor's mean rate of visit at the end of the observation period as $r(i, J_i)/\alpha(i, J_i)$. Across the 2,098 repeat visitors to Amazon, the median expected-visit rate at the end of our time period was 0.0349, or an intervisit time of 29 days.

Additionally, we calculate each panelist's purchasing propensity by dividing the number of visits during which a purchase occurred by the total number of visits made by that individual. The average conversion rate across the repeat visitors was 0.139; that is, almost 14% of the visits made by these customers was accompanied by a purchase. However, conversion rates differ for frequent visitors, whom we define as customers with visiting rates greater than or equal to the median (n = 1,062), versus infrequent visitors, whom we define as customers with visiting rates less than the median (n = 1,036). Frequent visitors have significantly higher conversion rates, averaging 16.6% compared to an average across the infrequent visitors of 11.1%, $t = 6.04$, $p < 0.001$.[4] These results confirm the hypothesis that frequent visitors tend to be more valuable customers since they are relatively more likely buyers, both on a percentage and an absolute basis.

However, the main objective of this article is to capture—and capitalize upon—nonstationarity in individual's visiting behavior. Though the overall visit rates provide some information about the attractiveness of the visitor as a buyer, these rates change over time, and the nature of this change may have implications for the panelist's buying propensity. Therefore, in addition to segmenting panelists into frequent and infrequent visitors, we also characterize and segment panelists based on the extent of the behavioral evolution they have undergone during the observation period.

To determine the extent of updating a panelist has undergone, we need to calculate a baseline rate of visit that would best capture the behavior if no evolution had taken place. Therefore, we calculate each individual's latent rate of visit given the observed behavior and the model results absent of any updating distribution (i.e., the value of γ associated with a stationary EG model). The extent of updating for each panelist is the difference between the rate of visit as given by the nonstationary model and this baseline rate.[5]

TABLE 4
Log-Likelihoods for Benchmark Models

	Amazon	*CDNOW*
Erlang-2 without updating (8 mo.)	−37,839.1	−10,564.1
Erlang-2 with updating (8 mo.)	−33,988.1	−9,172.1
Exponential (Months 1–4)	−12,571.3	−3,796.1
Exponential (Months 5–8)	−21,991.5	−6,015.4
Erlang-2 (Months 1–4)	−13,540.9	−3,796.1
Erlang-2 (Months 5–8)	−24,452.5	−6,812.0
Log-logistic	−34,224.9	−9,498.6

The median update for repeat Amazon visitors is 0.000. A median split along this dimension divides visitors into those who became more frequent visitors over time versus those who became less frequent visitors. We also see a difference in conversion rates (CR) along this dimension: Those who increased their rate of visit were more likely to buy (n = 1,056, CR = 15.1%) than those who decreased their rate of visit (n = 1,042, CR = 12.7%). Once again, this difference is highly significant, $t = 2.68$, $p = 0.007$, suggesting that the degree of evolution is indeed related to purchase propensity.

After seeing these two strong effects, a natural question is whether each one is still present when both are taken into account simultaneously. Table 5 examines the issue by dividing repeat visitors along both dimensions into four cells, using the same median splits as before. It is interesting to note that the number of visitors in each cell is quite balanced, indicating that there is not a dominant association between frequency and updating.

An ANOVA on these data confirms that both

[4] To account for the nonnormality of these proportions, we utilize a standard arc-sine transformation of the conversion rates for all of the statistical tests discussed in this section.

[5] There is no significant difference in the relative position of each household in terms of its extent of evolution when the change in visiting rates is measured as an absolute difference versus a percentage change.

TABLE 5
Amazon's Conversion Rates

Median = 0.0349	*Decreasing Frequency*	*Increasing Frequency*
Infrequent visitors	CELL 1 CR = 10.9% (n = 526)	CELL 2 CR = 11.3% (n = 510)
Frequent visitors	CELL 3 CR = 14.6% (n = 516)	CELL 4 CR = 18.6% (n = 546)

main effects remain highly significant, $F(1, 2094) = 35.765$, $p < 0.001$, for high versus low frequency, and $F(1, 2094) = 6.473$, $p = 0.011$, for increasing versus decreasing frequency. Furthermore, a strong interaction, $F(1, 2094) = 5.035$, $p = 0.025$, emerged as well, and its presence is easily seen in Table 5. For infrequent visitors (top row), there is no meaningful difference in conversion rates regardless of the nature of the household's updates over time. But for frequent visitors, the purchase-to-visit rate is considerably higher for those who have experienced increasing frequency. The households in the lower right cell are particularly conspicuous, with a conversion rate nearly 40% higher than the rest of the panel. This is clearly a very attractive group of repeat buyers.[6]

Table 6 presents the same analysis for the 581 households that made at least one repeat visit to CDNOW. The patterns are remarkably similar to those seen for Amazon, with the exception of smaller sample sizes and lower conversion rates. The ANOVA model reveals significant main effects for frequency, $F(1, 577) = 4.044$, $p = 0.045$, and updating, $F(1, 577) = 8.810$, $p = 0.003$, with a very strong interaction, $F(1, 577) = 6.405$, $p = 0.012$, once again highlighting the unique nature of those households that have accelerated their visiting behavior to a relatively high rate over the course of the 8-month data-collection period. The conversion rate for the households in this cell is over 60% higher than that of the three cells combined. While this translates to only 3 percentage points on an absolute basis, this represents a very significant improvement in an industry that is just becoming aware of the critical importance of this single statistic as the most useful indicator of an online retailer's performance and future prospects (Gurley, 2000).

Taken together, the analyses for these two leading online retailers suggest not only that frequent visitors are more likely buyers but also that a more refined segmentation of visitors that incorporates *changes* in visiting behavior can identify an even more valuable segment of customers to target. This is a new and important result, worthy of management attention and further research.

DISCUSSION AND CONCLUSIONS

The detailed, disaggregate clickstream data available online make it possible for us to study not just store visiting behavior more carefully but also the evolution of behavior at a Web site. Use of the EV model reveals that individual-level behavior patterns appear to contradict the perspective that one would obtain from examining the aggregate data alone. Specifically, the aggregate data seem to indicate an acceleration of visiting behavior at each of two leading e-commerce sites, yet our model parameters suggest that the typical visitor is experiencing a gradual slowdown in visiting rate over time. The

TABLE 6
CDNOW's Conversion Rates

Median = 0.0431	*Decreasing Frequency*	*Increasing Frequency*
Infrequent visitors	CELL 1 CR = 3.8% (n = 129)	CELL 2 CR = 5.7% (n = 161)
Frequent visitors	CELL 3 CR = 4.0% (n = 160)	CELL 4 CR = 7.6% (n = 131)

[6] In addition to this ANOVA conducted on the two dichotomous variables discussed here, we also examined equivalent regression models on the household-level data. The results are quite similar across the two datasets.

difference here is that an increasing number of new visitors is coming to each site over time, masking the slowdown that may be occurring for many experienced visitors. This effect could have dramatic implications for managers who neglect to examine their data at a sufficiently fine level of disaggregation.

Beyond the intuitive appeal of the model specification and its estimated parameters, we also show that it has excellent validity from an out-of-sample forecasting perspective. For both retail sites, the model tracks future visiting patterns extremely well, remaining within 5% of the actual data over the entire duration of a 4-month holdout period. While this model was not constructed with forecasting in mind as a principal objective, this result certainly speaks well about its overall versatility.

Perhaps the most dramatic demonstration of the model's validity and usefulness is its ability to delineate highly significant differences in purchasing behavior across visitors. There is a significant amount of literature suggesting that customers who visit a particular store frequently also tend to buy something during a relatively high proportion of those visits. We provide strong confirming evidence of this hypothesis. But the evolutionary nature of our model allows us to test an equally compelling, complementary hypothesis: People who experience increases in their visiting rates over time are more likely to purchase something at any given visit than those who are slowing down.

Both sites provide solid support for this new hypothesis, but also exhibit a powerful interaction that combines both effects. Specifically, panelists who combine high frequency with an upwards evolutionary trend in visiting behavior have dramatically higher conversion rates than all other panelists. As noted earlier, measuring and managing conversion rates have become crucial to e-commerce executives, so this is an important finding that merits additional investigation in future research.

LIMITATIONS AND FUTURE RESEARCH

Since this article is among the first attempts to carefully examine online visiting behavior using clickstream data, we have deliberately kept the model as clear and simple as possible to highlight the chief phenomena that we have observed in these datasets. However, one limitation is the fact that the data do not reveal when each customer first started visiting each site. As a result, the model is able to provide only a description of customer visiting rates and how they are changing during the data period being examined. In time, as all potential customers have adopted and become accustomed to the online environment, perhaps no evolution will be detected. However, the EV model presented here will allow site managers to monitor trends until that time comes and also know *when* that time has arrived.

Additionally, the results of our analysis have highlighted a potential relationship between store visit behavior and purchasing conversion. In this article, we have modeled store visiting as a process independent from purchasing while treating conversion rates as a purely static summary measure. Moe and Fader (2004) more carefully examined and developed a model of dynamic conversion behavior. In contrast to the present article, they took the observed pattern of visits as given, and model whether a purchase takes place at each visit. A natural next step is to combine these two different stochastic models to obtain a comprehensive, fully integrated view of online visit and purchase behavior. Research that unifies visiting and buying behavior in a single integrated model—while still allowing for a careful understanding of each subprocess—will offer great benefits in building a more complete picture of online shopping patterns. The models presented in this article as well as that of Moe and Fader (in press) are early steps in that direction.

REFERENCES

Bellinger, D.N., Robertson, D.H., & Hirschman, E.C. (1978). Impulse Buying Varies by Product. Journal of Advertising Research, 18(December), 15–18.

Bronnenberg, B.J., Mahajan, V., & Vanhonacker, W.R. (2000). The Emergence of Market Structure in New Repeat-Purchase Categories: The Interplay of Mar-

ket Share and Retailer Distribution. Journal of Marketing Research, 37 (February), 16–31.

Eskin, G.J. (1973). Dynamic Forecasts of New Product Demand Using a Depth of Repeat Model. Journal of Marketing Research, 10(May), 115–129.

Fader, P.S., Hardie, B.G.S., & Lee, K.L. (2003). 'Counting Your Customers' The Easy Way: An Alternative to the Pareto/NBD Model. Wharton Marketing Department.

Gurley, J.W. (2000). The One Internet Metric That Really Matters. Fortune, 141(5), 392.

Howard, R.A.(1965). Dynamic Inference. Operations Research, 13(2), 712–733.

Janiszewski, C. (1998). The Influence of Display Characteristics on Visual Exploratory Search Behavior. Journal of Consumer Research, 25(3), 290–301.

Jarboe, G.R., & McDaniel, C.D. (1987). A Profile of Browsers in Regional Shopping Malls. Journal of the Academy of Marketing Science, 15 (Spring), 46–53.

Johnson, E., & Russo, J.E. (1984). Product Familiarity and Learning New Information. Journal of Consumer Research, 11(June), 542–550.

Kendall, M.G., & Stuart, A. (1977). The Advanced Theory of Statistics (3rd ed., Vol. 2). New York: Hafner.

Media Metrix, Inc. (1998).

Moe, W.W., & Fader, P.S. (2004). Dynamic Conversion Behavior at E-commerce Sites. Management Science.

Morrison, D.G., & Schmittlein, D.C. (1988). Generalizing the NBD Model for Customer Purchases: What are the Implications and Is It Worth the Effort? Journal of Business & Economic Statistics, 6(2), 145–159.

Park, C.W., Iyer, E.S., & Smith, D.C. (1989). The Effects of Situational Factors on In-Store Grocery Shopping Behavior: The Role of Store Environment and Time Available for Shopping. Journal of Consumer Research, 15(4), 422–433.

Roy, A. (1994). Correlates of Mall Visit Frequency. Journal of Retailing, 70(2), 139–161.

Sabavala, D.J., & Morrison, D.G. (1981). A Nonstationary Model of Binary Choice Applied to Media Exposure. Management Science, 27(6), 637–657.

Schmittlein, D.C., Morrison, D.G., & Colombo, R. (1987). Counting Your Customers: Who Are They and What Will They Do Next? Management Science, 33, 1–24.

Seetharaman, P.B., & Chintagunta, P. (2003). A Proportion Hazard Model for Purchase Timing: A Comparison of Alternative Specifications. Journal of Business and Economic Statistics, 21(3), 1–15.

APPENDIX A: LIKELIHOOD FUNCTION FOR THE STATIONARY EG MODEL

When estimating the ordinary (stationary) EG model, there are two ways of obtaining the likelihood function for a given individual. The usual approach is to specify the individual-level likelihood function, conditional on that person's (unobserved) value of γ_i. This likelihood is the product of J_i exponential timing terms, where J_i is the number of repeat visits made by Panelist i, plus an additional term to account for the right censoring that occurs between that customer's last arrival and the end of the observed calibration period (at Time T):

$$L_i|\lambda_i = \lambda_i e^{-\lambda_i(t_{i1}-t_{i0})} \cdot \lambda_i e^{-\lambda_i(t_{i2}-t_{i1})} \cdot \ldots \cdot \lambda_i e^{-\lambda_i(t_{ij_i}-t_{i(J_i-1)})} \cdot e^{-\lambda_i(T-t_{ij_i})} \quad \text{(a1)}$$

To get the unconditional likelihood, we then integrate across all possible values of γ, using the gamma distribution as a weighting function:

$$L_i|r, \alpha = \int_0^\infty L_i|\lambda_i \cdot \text{gamma}(\lambda_i; r, \alpha) d\lambda_i \quad \text{(a2)}$$

where gamma(γ_i; r, α) denotes the gamma distribution as shown in (a1). This yields the usual EG likelihood, which can be multiplied across the n panelists to get the overall likelihood for parameter estimation purposes:

$$L = \prod_{i=1}^{N} \frac{\Gamma(r + J_i)}{\Gamma(r)} \left(\frac{\alpha}{\alpha + T - t_{i0}}\right)^r \left(\frac{1}{\alpha + T - t_{i0}}\right)^{J_i} \quad \text{(a3)}$$

An alternative path that leads to the same result is to perform the gamma integration separately for each of the $J_i + 1$ exponential terms, and then multiply them together at the end. This involves the use of Bayes Theorem to refine our "guess" about each individual's value of γ_i after each arrival occurs. Specifically, it is easy to show that if someone's first repeat visit occurs at Time t_{ij}, then the heterogeneity distribution governing γ is distributed gamma($r + 1$, $\alpha + t_{i1}$

$- t_{i0})$, instead of gamma(r, α) as initially shown in Equation (a1). More formally,

$$g(\lambda_{i2}|\text{arrival at } t_{i1}) = \text{gamma}(r + 1, \alpha + t_{i1} - t_{i0}) \quad \text{(a4)}$$

The gamma distribution governing the rate of visit for subsequent arrivals follows:

$$g(\lambda_{i(j+1)}|\text{arrival at } t_{ij}) = \text{gamma}(r + j, \alpha + t_{ij} - t_{i0}) \quad \text{(a5)}$$

Using this logic, we can re-express the likelihood as the product of separate EG terms

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J_i} \left(\frac{r + j + 1}{\alpha + t_{i(j-1)} - t_{i0}} \right) \times \left(\frac{\alpha + t_{i(j-1)} - t_{i0}}{\alpha + t_{ij} - t_{i0}} \right)^{r+j} \cdot S(T - t_{ij_i})$$

$$\text{where} \quad S(T - t_{ij_i}) = \left(\frac{\alpha + t_{ij_i} - t_{i0}}{\alpha + T - t_{i0}} \right)^{r+J_i} \quad \text{(a6)}$$

which collapses into the same expression as (a3).

APPENDIX B: MOMENT-MATCHING APPROXIMATION OF THE PRODUCT OF TWO GAMMA DISTRIBUTIONS

If x and y are two gamma distributed random variables,

$x \sim$ Gamma (r, α)

$y \sim$ Gamma (s, β)

then the product, $z = xy$, can be assumed to be a gamma distributed random variable

$z \sim$ *Gamma* (R, A)

with shape and scale parameters, R and A, such that the first two raw moments of the z distribution is the product of the moments of the x and y distributions.

$$m_1^x = \frac{r}{\alpha} \qquad m_2^x = \frac{r(r+1)}{\alpha^2}$$

$$m_1^y = \frac{s}{\beta} \qquad m_2^y = \frac{s(s+1)}{\beta^2}$$

$$m_1^z = m_1^x \cdot m_1^y = \frac{rs}{\alpha\beta} \quad m_2^z = m_2^x \cdot m_2^y = \frac{r(r+1)s(s+1)}{\alpha^2\beta^2} \quad \text{(b1)}$$

Since the first moment of the z distribution, m_1^z, is R/A and the second moment, m_2^z, is $R(R + 1)/A^2$, we can solve for R and A with the following two equations:

$$\frac{R}{A} = \frac{rs}{\alpha\beta} \quad \text{(b2)}$$

$$\frac{R(R+1)}{A^2} = \frac{r(r+1)s(s+1)}{\alpha^2\beta^2} \quad \text{(b3)}$$

Therefore, the gamma distribution describing the product of two independently distributed gamma random variables has shape and scale parameters that can be calculated from the parameters of the multiplying distributions.

$$R = \frac{rs}{(r+1)(s+1) - rs}$$

$$A = \frac{\alpha\beta}{(r+1)(s+1) - rs} \quad \text{(b4)}$$

with Bayesian updating after observing one arrival at Time t. . .

$$R = \frac{(r+1)s}{(r+2)(s+1) - (r+1)s}$$

$$A = \frac{(\alpha + t)\beta}{(r+2)(s+1) - (r+1)s} \quad \text{(b5)}$$