

RAHUL TELANG, PETER BOATWRIGHT, and TRIDAS MUKHOPADHYAY*

The authors extend the marketing literature on stochastic interpurchase-time models by allowing for purchase periodicities and unobserved heterogeneity in a proportional hazards mixture model. Their parsimonious framework builds on commonly used baseline hazard functions. They use the search-engine visits data to highlight the benefits of the proposed model.

A Mixture Model for Internet Search-Engine Visits

In the past decade or two, firms have increasingly relied on statistical and econometric models to facilitate marketing decisions. Because of the widespread availability of data in the consumer packaged goods industry, most of the developments in quantitative tools have been applied to the marketing of grocery products. The negative binomial distribution and, more recently, the proportional hazards model (PHM) have been used to study stochastic interpurchase times and store visits, particularly in demand estimation, sales forecasts, analysis of consumer segments, and the impact of product promotions (Ehrenberg 1972; Gupta 1991; Jain and Vilcassim 1991; Morrison and Schmittlein 1981, 1988; Seetharaman and Chintagunta 2002). A problematic restriction of the extant interpurchase-time models that inhibits their application to many firms is that they do not account for periodicity in purchases and visits. We extend the marketing literature by allowing for purchase periodicities through the use of a mixture-model framework. We use our proposed model to address the feasibility of personalization at popular Internet Web sites, which is one of many potential applications, and we investigate specific factors that affect frequency of Web search-engine use.

Because of people's propensity to use search engines as the entry point for Web browsing and search engines' central role in locating Web sites, search engines have become the arbiters of Web site information, and they possibly hold the greatest potential for marketing on the Web. Although it is tempting to use the traditional interpurchase models to fit search-engine visit data, they provide a poor fit because the data are highly periodic. Consider Panels A and B of Figure

1. Each panel shows a histogram of the same data, which demonstrates the intervisit times to the search engines in our data set. In each panel, time is on the horizontal axis, and the vertical axis gives counts of visits during the time periods. In Panel A, we aggregated the data to units of days, whereas Panel B depicts hourly data. The pattern in the aggregate data is one that can be fit by a well-known distribution such as the exponential. However, such simple distributions do not provide accurate fits to the hourly data because of the many modes. It is significant that this pattern of user behavior has been consistently observed in even traditional data, though extant applications have been able to rely on aggregate (unimodal) data (Dunn, Reader, and Wrigley 1983; Kahn and Schmittlein 1989). Moreover, researchers' not accounting for such regularities may lead to biased estimates of important covariates. A model that can parsimoniously incorporate regularly occurring peaks can also make better predictions at a disaggregate level. Such predictions can be effectively used for the purpose of personalization.¹

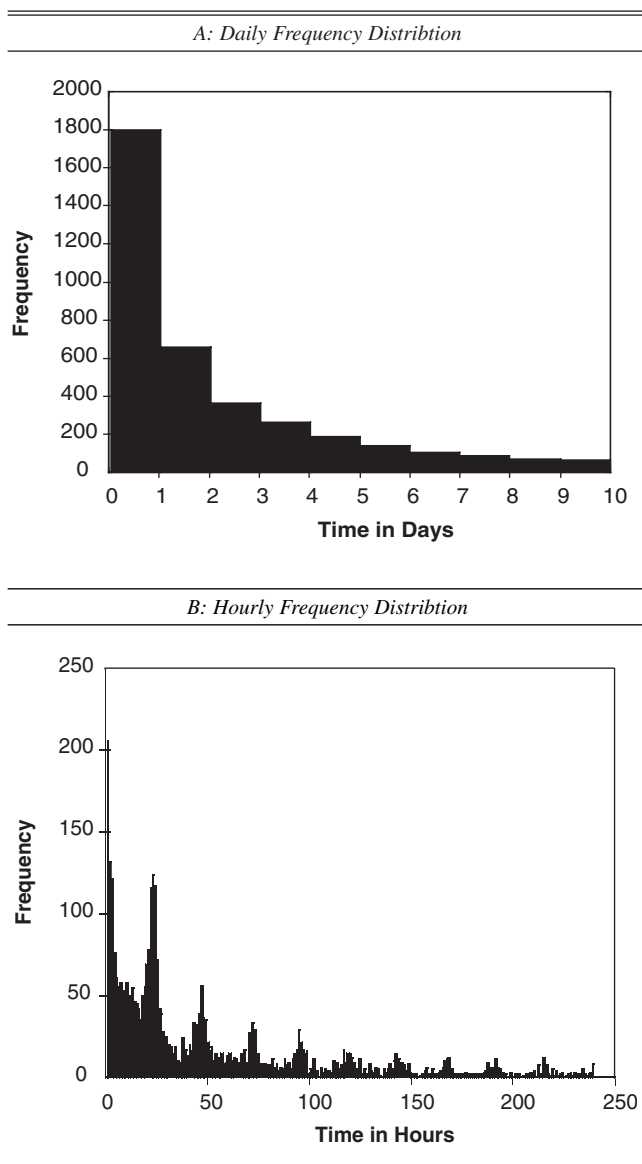
Because of the high volume of visits to search engines, most Web advertising dollars are spent at search-engine sites (Buchwalter, Ryan, and Martin 2001). However, as for other high-volume sites, it is difficult for search engines to provide personalization and targeting, a promising benefit of Internet marketing, because serving dynamic content imposes large costs on network infrastructure.² An aid to firms' personalization efforts is the more accurate prediction of a user's subsequent visit. For example, a manager of the IBM High-Volume Web Site Team suggested that it is useful for firms to know customers' time of visit. This knowledge enables firms either to create the personalized

*Rahul Telang is Assistant Professor of Information Systems, H. John Heinz III School of Public Policy and Management (e-mail: rtelang@andrew.cmu.edu), and Peter Boatwright is Assistant Professor of Marketing (e-mail: pbhb@andrew.cmu.edu) and Tridas Mukhopadhyay is Deloitte Consulting Professor of e-Business (e-mail: tridas@andrew.cmu.edu), Graduate School of Industrial Administration, Carnegie Mellon University.

¹Other applications of such a model would include demand forecasting, load balancing on the servers, user segmentation, analysis of the impact of product promotions, and inferred market structure (Grover and Rao 1989).

²Dynamic content is time sensitive; it often contains information such as weather forecasts, stock quotes, or news.

Figure 1
FREQUENCY DISTRIBUTIONS



pages beforehand and serve them or to shuffle the databases so that search time is reduced considerably (Chiu 2001).

Because content is dynamic, it is not possible to create personalized pages for all users all the time. To do so would require updates of pages in regular, frequent intervals, even though the updated pages may not be served. This process is a tremendous strain on server resources. However, if a model can predict the next user visit accurately, such personalized pages can be prebuilt for those time windows, thus considerably improving the efficiency of serving personalized content to individual users.³

In this article, we propose a mixture model that accounts for the regularities of the disaggregate data. At the same

time, our model does not lose the generality or predictive power of traditional models, but it captures the recurring peaks with only minimal additional parameters. Our structure also easily accommodates covariates and unobserved heterogeneity. In our application, we investigate factors that accelerate or delay user visits, including user experience and search-engine attributes.

The contributions of our article to the literature are three-fold. First, we provide a more complete model to predict visits to search-engine sites. Second, we offer a mixture model that not only incorporates the properties of the traditional models but also accounts for regularly occurring peaks and fits and predicts the disaggregate data more accurately. Third, we incorporate covariates that add precision to predictions, and we empirically test relevant hypotheses about search-engine use.

The rest of the article is organized as follows: We subsequently review the literature and then present our model. We follow up with a description and analysis of the data. We then show that our model outperforms the traditional model at both an aggregate and an individual level. Finally, we present concluding remarks and the managerial implications of our findings.

LITERATURE REVIEW

Because of the central role of search engines on the Internet, many studies have examined issues related to search engines and ShopBots. Bradlow and Schmittlein (2000) show that search-engine indexing is not extensive and that users need more than one engine to receive satisfactory results. Mukhopadhyay, Rajan, and Telang (2002) use this property to show that many engines can survive in the market. Moe and Fader (2002) capture the issue of intervisit time on online Web sites (e.g., Amazon.com, CDnow) by introducing nonstationarity into the mean rate of exponential distribution. Park and Fader (2002) model the users' visit patterns using a multivariate timing mixture model and by generalizing the univariate exponential gamma model. We contribute to this growing body of literature by modeling the recurring regularities in user behavior. Although our data are specific to search engines, our model is applicable in more general online and offline product settings.

The PHM has been used in the marketing literature to study the influence of marketing activities on household purchase decisions (Gupta 1991; Jain and Vilcassim 1991; Seetharaman and Chintagunta 2002). In the PHM, the hazard function is decomposed into two multiplicative components: (1) the baseline hazard function and (2) the covariate function. Many different baseline hazard functions have been used in the PHM framework. The most commonly used distributions are smooth (i.e., unimodal and monotonic on either side of the single mode) and thus cannot account for the additional modes caused by the regularity of usage. We develop a parsimonious continuous distribution model that can explicitly account for the regularities in a context of continuous data. To account for regularly occurring spikes, we mix a Laplace distribution (also called the double exponential distribution) with the baseline hazard. To incorporate user heterogeneity, we use a nonparametric discrete support point method (Heckman and Singer 1984; Jain and Vilcassim 1991) and covariates. Despite the mixture of two

³A current trend is to push even personalized content to the edge of the network to serve the pages quickly. However, to realize any cost savings, a Web server should push only the pages that have a greater chance of being hit (Datta et al. 2002), which emphasizes the need for a model that predicts visits.

distributions, the model is quite parsimonious and has only two extra parameters in its baseline hazard.

MODEL FORMULATION

We base our mathematical model on the simple concept that people tend to operate on schedules. For example, people tend to sleep at a regular period of the 24-hour day. A person using a search engine may interrupt a session to follow his or her schedule (e.g., go to sleep, go to work) and then perform a search the subsequent day at a time that his or her schedule permits. In contrast, a person may interrupt a session to eat dinner, for example, and perform the next search after only an hour or two. Thus, at the individual level, there is some probability that a person's subsequent log-in will occur relatively soon, and there is also the chance that a schedule will influence the visit. So, although the need to use search engines may occur at random, a user's schedule may cause visits to be more predictable. Although any parametric hazard function can model the randomness of the visits, we illustrate our general approach using four different baseline hazards and mixing each with the Laplace distribution to account for users' schedules. In Table 1, we list the hazard and survivor functions of the four distributions we test with our data. All four distributions allow for increasing, decreasing, or constant hazards. The expo-power distribution is especially flexible (Saha and Hilton 1997).

Mixture Model

To incorporate daily schedules into the model, we decompose the intervisit time into days and hours and mix in a distribution on the hours that accounts for schedules. We propose a unimodal distribution that reflects a preference for a particular time of day, given the day of the visit.

For the unimodal time-of-day distribution, we use the Laplace distribution (also known as the double exponential distribution). The thicker tails of the Laplace are a more accurate representation of the observed data than the more common Gaussian density would imply. The density of the Laplace distribution is

$$(1) \quad f_L(t) = .5 \phi^{-1} \exp\left(\frac{-(t - \theta)}{\phi}\right),$$

and the hazard function is

$$(2) \quad h_L(t) = \begin{cases} \frac{\exp[-(t - \theta)\phi^{-1}]}{2 - \exp[-(t - \theta)\phi^{-1}]} \times \frac{1}{\phi} & \text{for } \theta < t \text{ and} \\ \frac{1}{\phi} & \text{for } \theta \geq t. \end{cases}$$

The Laplace is a two-parameter distribution in which θ is the location parameter. Because the domain in our application is a single day, we truncate the Laplace distribution in one-day periods in which the peaks occur at multiples of 24 hours. In other words, we can specify the location parameter a priori (i.e., at multiples of 24 hours).⁴ The density function of the truncated Laplace is

$$(3) \quad f_{TL}(t) = \frac{.5\phi^{-1} \exp(-\phi^{-1}|t^*|)}{I(t)},$$

for $t^* = t - 24k$, $k = \text{round}[(t - 12)/24 + .5]$, and

$$(4) \quad I(t) = \begin{cases} \int_0^{12} \frac{e^{-\frac{|u|}{\beta}}}{2\beta} du & \text{for } t < 12 \\ \int_{-12}^{12} \frac{e^{-\frac{|u|}{\beta}}}{2\beta} du & \text{for } t \geq 12 \end{cases}.$$

Here, k represents days in integers, t is continuous time across days, and t^* is continuous time within a day.

In our conceptual framework, there is some probability p that a person's visit is unaffected by a schedule, and the probability $(1 - p)$ indicates that the visit follows a schedule. Thus, within a day (a period of 24 hours), the distribution of the data on the k th day is a mixture of the baseline density and the truncated Laplace distribution. In the following, we denote the baseline density as $f_B(t)$, where $f_B(t)$

⁴If it is preferred, the location parameter can be estimated from the data.

Table 1
BASELINE HAZARD AND SURVIVOR FUNCTIONS

Function	Hazard	Survivor
Weibull	$h_w(t) = \beta\alpha t^{\alpha-1}$	$S_w(t) = e^{-\beta t^\alpha}$
Log-logistic	$h_{LG}(t) = \frac{\beta\alpha(\beta t)^{\alpha-1}}{1 + (\beta t)^\alpha}$	$S_{LG}(t) = \frac{1}{1 + (\beta t)^\alpha}$
Expo-power	$h_E(t) = \beta\alpha t^{\alpha-1} e^{\gamma t^\alpha}$	$S_E(t) = e^{\gamma} [1 - e^{\gamma t^\alpha}]$
Conway-Maxwell-Poisson	$h_D(z) = \frac{\lambda^z}{(z!)^\nu} \frac{\sum_{k=0}^{\infty} \frac{\lambda^k}{(k!)^\nu} - \sum_{x=0}^z \frac{\lambda^x}{(x!)^\nu}}{\sum_{x=0}^z \frac{\lambda^x}{(x!)^\nu}}$	$S_D(z) = 1 - \left[\sum_{k=0}^{\infty} \frac{\lambda^k}{(k!)^\nu} \right]^{-1} \sum_{x=0}^z \frac{\lambda^x}{(x!)^\nu}$

Notes: For the Conway-Maxwell-Poisson distribution, z is a nonnegative integer and t is continuous.

represents any probability density, such as those in Table 1. The mixture density on the k th day following the previous Web site visit can be summarized as follows:

$$(5) f_k(t) = \begin{cases} p \frac{f_B(t)}{\int_0^{12} f_B(u)du} + (1-p)f_{TL}(t) & \text{for } t < 12(k=0) \\ p \frac{f_B(t)}{\int_{24k-12}^{24k+12} f_B(u)du} + (1-p)f_{TL}(t) & \text{for } t \geq 12(k > 0) \end{cases}$$

The likelihood in Equation 5 is conditional on day k (within a specific day). To define the likelihood over continuous time, we need probabilities both across days and within days. Because the baseline hazard that underlies the visit process is defined over continuous time, it can provide the probability of each day. We compute the probability distribution over the days (represented by the discrete random variable k) by integrating $f_B(t)$, or

$$(6) \Pr(K = k) = \begin{cases} \int_0^{12} f_B(u)du & \text{for } k = 0 \\ \int_{24k-12}^{24k+12} f_B(u)du & \text{for } k > 0 \end{cases}$$

Thus, the mixture density $f_M(t)$ is

$$(7) f_M(t) = \Pr(k)f_k(t) = \begin{cases} pf_B(t) + (1-p)f_{TL}(t) \int_0^{12} f_B(u)du & \text{for } t < 12 \\ pf_B(t) + (1-p)f_{TL}(t) \int_{24k-12}^{24k+12} f_B(u)du & \text{for } t \geq 12 \end{cases}$$

Unobserved Heterogeneity and Covariates

The previous model includes neither covariates nor unobserved heterogeneity. In the literature, covariates are typically incorporated in a multiplicative PHM (Gönül and Srinivasan 1993; Gupta 1991). In line with this approach, the hazard function with covariates is $h_1(t|x) = h_0(t)e^{x\lambda}$, where λ is a parameter vector to be estimated, and X is a vector of covariates. Note that $h_0(t)$ is the baseline hazard function of the mixture $h_0(t) = [f_M(t)/S_M(t)]$, where $f_M(t)$ is specified in Equation 7, and $S_M(t)$ is the corresponding survival function. The baseline hazard function for our mixture model can be specified as follows:

$$(8) h_0(t) = \{pf_B(t) + (1-p)[F_B(k^+) - F_B(k^-)]f_{TL}(t)\} / \{1 - pF_B(t) - (1-p)[F_B(k^-)] - (1-p)[F_B(k^+) - F_B(k^-)]F_{TL}(t)\},$$

where F_B and F_{TL} are cumulative densities of f_B and f_{TL} . $k^+ = 24k + 12$, and $k^- = 24k - 12$. We can easily derive the standard nonmixture hazard function from Equation 8 by setting p equal to 1.

Users show heterogeneity in their frequency of visits to search engines in our data. Although we can control for some of the differences by incorporating many observed covariates (including demographics), unobserved hetero-

geneity typically needs to be accounted for in the sample. If unobserved heterogeneity across users is not modeled, duration dependence may appear to characterize the sample data, even if it does not exist for any of the individual observations. Unobserved heterogeneity can be modeled either parametrically or nonparametrically. We adopt a nonparametric approach, as Heckman and Singer (1984) suggest. The new hazard function after incorporation of heterogeneity, conditional on support s , is

$$(9) h_s(t|x) = h_s(t)e^{x\lambda_s},$$

where h_s is the baseline hazard that corresponds to support s . We assume that s is fixed across spells for a given user and has a distribution $G(\cdot)$ across users. We assume that there is no parametric form for $G(\cdot)$, but we estimate it empirically from the data. This approach has been used extensively in both the economics and the marketing literature (Jain and Vilcassim 1991).

Note that there is a one-to-one relationship between the density and hazard function such that

$$(10) f_s(t|x) = h_s(t|x)e^{-\int_0^t h_s(u|x)du}$$

Therefore, we can write the likelihood function for the model with Equation 8:

$$(11) L = \prod_{n=1}^N \sum_{s=1}^R \left[q_s \prod_{j=1}^{J_n} f_s(t_j|x) \right] S_s(t_{last}|x),$$

where N is the total number of users in the sample, R is the number of support points, q_s is the membership probability of support s , J_n is the number of observations available for user n , S_s is the survivor function of a user for support s , and t_{last} is the elapsed time from the last observed log-in of user n to the date at which the data is right-censored.⁵

As for identifiability, our mixture borrows from the baseline density in such a way that its parameters are identified as long as we identify a PHM with the baseline density. Given Elbers and Ridder's (1982) three conditions, the first and third conditions are unaffected by our use of a mixture. The second condition requires that integrals of the hazard are finite for finite time points. In our mixture, time t is bounded above and below by the start and end of the day, the k^- and k^+ of Equation 8. Similarly, $\int_0^t h_0(u)du$ is bounded above and below by $\int_0^{k^-} h_0(u)du$ and $\int_0^{k^+} h_0(u)du$. In our mixture, the latter two integrals are integrals of the hazard of the baseline density f_B , not of the mixture of f_B and f_{TL} . Therefore, as long as the proportional hazard is identifiable for the baseline density (e.g., the Weibull), it is identified for our mixture of the baseline density and the truncated Laplace. Although theoretically the model is identifiable with few data points per user, its practical application requires longer time series for better precision, because we model regularities in user behavior.

DATA

The data we used for this study are from the HomeNet project (Kraut et al. 1999), which tracked Internet use in

⁵The data in this study are not left-censored because they contain the start date of the users in the study.

homes. For each user, the server recorded which search engine the user visited, the time at which the visit occurred, and the user's action on the site. The most important characteristic of the data set is that it was collected in an unobtrusive and natural setting: participants' homes. The data collected contain actual choices rather than elicited preferences.

The server maintained the detailed reports of the users' navigation patterns on the Internet. Using uniform-resource-locator information, we identified whether the user performed a search or visited some of the nonsearch features, such as e-mail, chat, and shopping. We captured data for 126 individual users for one year, from May 30, 1998, to May 30, 1999. Most participants had not used the Web before the study. We gathered demographic characteristics of individual users through a questionnaire.

In our analysis, we included the top eight search engines during the time of the study in our product class. The eight engines we included are Yahoo!, Excite, Lycos, AltaVista, Infoseek, Goto, Go, and HotBot. In our sample, more than 95% of users' visits to search engines were to one of these eight engines. We treat search engines as a product class rather than as individual brands, and thus we consider a visit to any of these engines a visit by the user to an engine.

Because of the precise nature of our data, we know the exact time of each visit. Because the goal is to predict when a user will return to a site after completing a session, we were required to define a search session. We analyzed the data under the assumption that searches separated by less than an hour are part of the same session.

Our final data sample consists of 126 users who made 4559 visits to engines during the year; we excluded users with only one visit to the engine during the period. We use the first nine months of data for calibration purposes, for 3398 data points, and retain the remaining three months of data as a holdout sample. The mean intervisit time in the calibration sample is 112 hours, with a standard deviation of 220. The median time is 34 hours. The average number of visits per user is approximately 27, with a standard deviation of 34. Of the 126 subjects, approximately 53% were women, almost 76% were white, and 59% were adults (older than age 21).

Covariates

Crockett (2000) reports that a lower percentage of online users relied on search engines in January 2000 than in July 1999 to find sites of interest. It is possible that with time, users tend to use the engines less frequently. The rationale for this hypothesis is that users become more experienced with the Internet. When users have become familiar with their preferred sites, they go directly to those sites without using search engines. Thus, experience with the Internet may be a useful covariate for prediction of frequency. Given that frequency of use is tied to the revenue of the engines, it is possible that experienced users' reduced usage casts doubt on the search engines' long-term profitability. Because most of the users in our sample were new to the World Wide Web, our data cover the time period during which the users learned how to use search engines. As a proxy for experience, we use log of count of visits before time t for each individual.⁶

⁶We also tested a discrete measure in which the dummy variable experience is equal to 0 for the first six months and equal to 1 for the remaining months; our results were analogous to the ones we report herein.

Search engines also spend resources to develop new features, such as chat, e-mail, and message boards. These are considered value-added features, and some authors have tried to measure the impact of them on prices and firms' market share (Brynjolfsson and Kemerer 1996). Telang, Mukhopadhyay, and Wilcox (2002) have studied the impact of these features on users' loyalty and search-engine choices, because these features are typically offered at no monetary cost to the users. We study the impact of feature use on frequency of visits; a primary aim of these features is to accelerate users' visits to the engines.

Our empirical measure of nonsearch features of the engine, f_{ij} , is a cumulative count of the number of times that individual i used such features before session j . The mean of f_{ij} is 8.94, and the standard deviation is 10.6. We use the square root of f_{ij} as a covariate to test whether more use of nonsearch features in previous periods would lead to more frequent visits to the engines in future periods.

In addition, search-engine features may counteract the potential negative effect of experience, because these nonsearch features offer even experienced users a reason to use search sites regularly. Therefore, although people who use engines mostly for search purposes may show a decline in their visit frequency over the period, users of nonsearch features such as e-mail and stock quotes may visit the engines consistently. Thus, we test the interaction of experience and feature use on the visit propensity.

Finally, we also introduced multiple demographic variables as covariates. As in previous studies (Rossi, McCulloch, and Allenby 1996), demographic variables are not a good predictor of frequency of visits to the engines, and thus we do not show them in our results.

RESULTS AND DISCUSSION

In Tables 2 and 3, we present the results of our estimation using four different baseline hazards. In Table 2, we present the results without the mixture model; in Table 3, we present the results with the mixture model. In each table, there are two columns for each baseline specification that correspond to two support points (two segments). The addition of another segment does not lead to any significant improvement in the likelihood. To account for heterogeneity properly, we estimated segment-specific slope parameters, not simply segment-specific intercepts.

In a comparison of Tables 2 and 3, our mixture model performs better in terms of fit in all four baseline specifications, which justifies the need to incorporate the Laplace distribution to account for the users' schedules. In addition, the mixture proportion p is significantly different from zero or one in all three specifications, which offers additional evidence in support of the mixture and shows that the users place significant value on their schedules.

Using the Conway–Maxwell–Poisson distribution, Boatwright, Borle, and Kadane (2003) estimate a mixture model for periodicities in a discrete context. To contrast our results with their model, we estimate that model but with two main differences. First, we use the Laplace for the mixture instead of a multinomial distribution. Note that the use of multinomial distribution would lead to estimation of more than 20 parameters for hours in a day for each segment. Second, in their application, they define the baseline density over weeks, a level of aggregation that

Table 2
CALIBRATION RESULTS

	<i>Weibull</i>		<i>Log-Logistics</i>		<i>Expo-Power</i>		<i>Geometric</i>	
β	.04 (2.93)	.02 (2.3)	.10 (2.35)	.006 (4.27)	.03 (2.12)	.010 (2.84)	.022 (4.72)	.003 (17.75)
α	.762 (4.2)	.714 (16.2)	.98 (3.41)	.85 (28.2)	.847 (3.96)	.765 (8.43)		
γ					-.006 (.36)	-.001 (.50)		
Experience	-1.07 (.9)	-.65 (.7)	.06 (.2)	-.46 (2.34)	-2.32 (.19)	1.08 (.23)	.05 (.01)	1.44 (29.5)
Weekend	.54 (.50)	1.25 (9.65)	.63 (1.16)	.68 (7.96)	.82 (2.63)	.84 (5.97)	1.08 (5.61)	.415 (12.3)
Features	.20 (.52)	.34 (5.76)	.25 (.44)	.59 (6.59)	.25 (.42)	.44 (3.96)	.154 (1.41)	.309 (4.52)
Interaction	.06 (.72)	.087 (2.71)	.47 (1.08)	.03 (.91)	.03 (.79)	.03 (.28)	.29 (.12)	.022 (4.72)
q (segment proportion)	38%	62%	29%	71%	37%	63%	27%	73%
Log-likelihood	-18,144		-18,221		-18,097		-18,454	
Bayesian information criterion	36,393		36,547		36,315		37,005	

Notes: t-values are in parentheses.

Table 3
MIXTURE-MODEL RESULTS

	<i>Weibull</i>		<i>Log-Logistics</i>		<i>Expo-Power</i>		<i>Geometric</i>	
β	.065 (2.1)	.02 (2.74)	.04 (2.35)	.006 (5.2)	.09 (2.12)	.016 (2.84)	.83 (50.1)	.0183 (83.2)
α	.718 (7.9)	.741 (17.9)	1.24 (6.3)	.99 (28.4)	.78 (4.3)	.79 (16.4)		
γ					-.010 (.43)	-.03 (2.27)		
ϕ	4.05 (2.25)	2.13 (9.4)	2.79 (2.5)	1.80 (10.0)	4.18 (2.01)	2.13 (9.18)	3.73 (9.2)	2.06 (6.2)
Experience	.29 (.33)	-.15 (.95)	-1.13 (1.8)	-.70 (2.87)	-.45 (.38)	-.81 (3.9)	.71 (3.6)	-1.70 (7.5)
Weekend	.69 (1.36)	1.15 (11.9)	.55 (1.6)	.65 (5.76)	.025 (.02)	1.12 (4.84)	.43 (1.5)	.45 (3.40)
Features	.037 (.11)	.47 (7.92)	.02 (.05)	.47 (6.11)	.09 (.25)	.278 (5.0)	.02 (.33)	.43 (24.1)
Interaction	-.07 (.17)	-.023 (.25)	-.45 (1.28)	.27 (2.11)	.23 (.36)	.34 (3.27)	-.74 (1.6)	-.78 (7.5)
P (mixture proportion)	.53 (14.2)	.39 (3.74)	.57 (23.2)	.40 (3.20)	.60 (5.43)	.538 (14.9)	.44 (3.83)	.52 (6.58)
q (segment proportion)	29%	71%	38%	62%	28%	72%	21%	79%
Log-likelihood	-17,659		-17,741		-17,632		-18,004	
Bayesian information criterion	35,456		35,620		35,418		36,129	

Notes: t-values are in parentheses.

would be analogous to days in our data. Because a baseline density defined over hours provides a better fit than that defined over days, we used hours. The parameter estimates of the Conway–Maxwell–Poisson distribution show that it is equivalent in this case to a geometric distribution, so we report results from a geometric distribution in Tables 2 and 3.⁷

In all four specifications, the parameter for the Laplace distribution (ϕ) is larger for the first segment. Because a smaller ϕ yields tighter variance of the Laplace distribution, Segment 2 users follow their schedules more consistently than do Segment 1 users. In Tables 2 and 3, the expo-power outperforms the other specifications. In what follows, we focus our discussion on the expo-power, though other results yield similar conclusions.

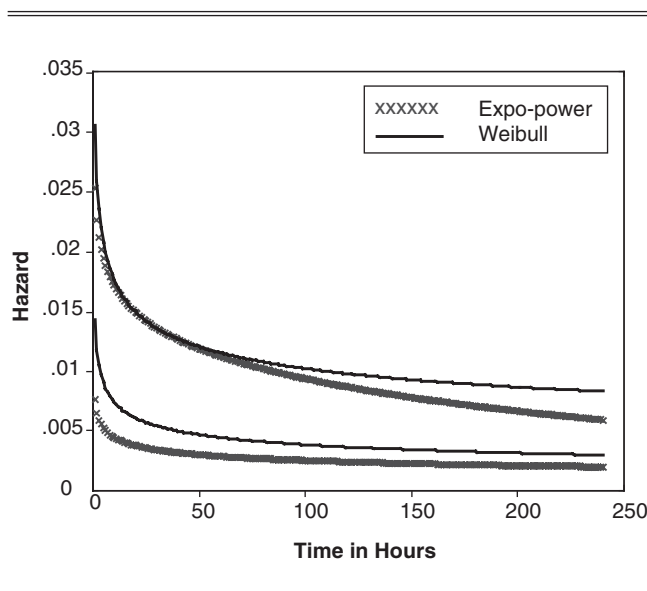
We plot the baseline hazards for the expo-power and Weibull distributions for both segments in Figure 2. As is seen in Figure 2, both hazards are monotonically decreasing, though the first segment exhibits higher propensity to return early in both models. As a basis for our model, we assume that multiple modes arise because of periodicities in users' schedules. It is also possible that the existence of multiple consumer segments leads to multiple modes.⁸ If the multiple modes are simply due to user heterogeneity, at least one of the hazards should not be monotonically decreasing. Because both segments of all baseline specifications are monotonically decreasing (see Figure 2), segmentation is not the underlying cause of modes in our data.

The hazard rates also suggest that Segment 1 users are less frequent visitors than are Segment 2 users. Notably, the covariates are not significant for Segment 1, but all covariate coefficient estimates are significant for Segment 2, and the effects are in the hypothesized directions. For Segment

⁷The Conway–Maxwell–Poisson distribution nests the Poisson, geometric, and Bernoulli distributions.

⁸We thank an anonymous reviewer for pointing out this explanation for the existence of multiple modes.

Figure 2
BASELINE HAZARD



2, the interpretation of the covariate parameters is as follows: The positive sign means that an increase in the corresponding regressor increases the hazard rate or that the intervisit time has decreased. The coefficient of the non-search features has a positive and significant coefficient. As people use more nonsearch features, they tend to visit the engines more frequently. This is consistent with our expectation. From the perspective of search-engine advertising, this result might be an important determinant of the viability of the nonsearch features.

In contrast with features, experience has a negative and significant coefficient. As we discussed previously, people probably become more experienced with Internet use over time, learning which sites are interesting to them and eventually relying less on search engines to navigate across the Web.

Finally, the interaction variable is positive and significant. For people who use the nonsearch features, the frequency of visits does not decrease. For people who use the nonsearch features frequently, the visit rate to engines increases, which contrasts with results of prior intuition. Therefore, nonsearch features help the engines in two ways. First, they lead to users' more frequent visits to the search engine. Second, the use of nonsearch features leads to strong and consistent visits to search engines across time.

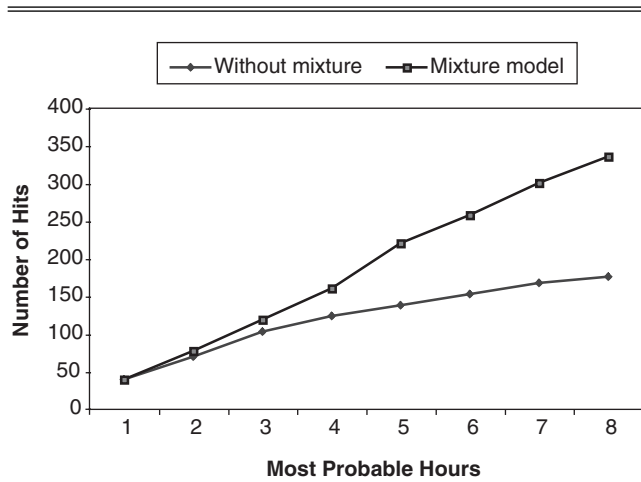
MODEL VALIDATION

We use the holdout sample to validate our results, comparing the mixture and nonmixture models with the expo-power baseline density. The holdout sample consists of the final three months of the data, which contain 1161 observations. Because not all panel members used search engines during this period, there are 105 users in the holdout sample. For the purpose of illustrating and testing the predictive power of our model, we present predictive hit rates of the two models.

To calculate hit rates, we use the models to identify the most probable hours when a user is likely to visit a search engine. If the user visits in those hours, we award the model a hit. The usefulness of this approach can be viewed from the perspective of a manager, who can use the model to decide for which time periods to prepare for a visit by a user. Because preparedness for visits involves costs, a firm needs to anticipate user visits in a time window of only a few hours. The decision of how many hours to include in the window can be viewed as a cost–benefit analysis. A hit generates some benefit $B(\cdot)$, but it costs $C(\cdot)$ to track the user within that hour. If the probability of a user visit in an hour is p , it may be worthwhile to track the user only if $B(\cdot)p - C(\cdot) > 0$. Because we do not know such costs, we present results for several different time windows.

For each user, conditional on segment membership and covariate values, we identify the hour (of any day, not within a particular day) that has the highest visit probability, the two hours with highest visit probability, and so on, up to the most likely eight hours. We calculate the hit rate for the various time windows and plot the results in Figure 3. We were not surprised that for the one- and two-hour windows, both models were quite similar. As is shown in Figure 1, the first two hours are the most likely hours for a visit, which the traditional model also predicts. Afterward, though, the traditional model completely misses the regularities. The

Figure 3
HIT RATE FOR THE HOLDOUT SAMPLE



mixture model outperforms the traditional (unmixed) model for three-hour windows onward. For an eight-hour window, for example, the mixture model has a hit rate of approximately 30%, compared with the traditional model, which has a hit rate of 15%. The traditional model offers little beyond a simple heuristic, a prediction that all visits will occur in the following few consecutive hours.

We also present predictions at an individual level. We calculate the hit rate for each individual over two-, five-, and eight-hour windows. The hit rates are shown in Figure 4, sorted in order of increasing probabilities as predicted by the traditional model. As is shown, our model outperforms the traditional model for all users for all time windows.⁹ Because the hit rate predicted by our model is quite high, the individual hit rates may enable managers to focus on select users as they do for many users.

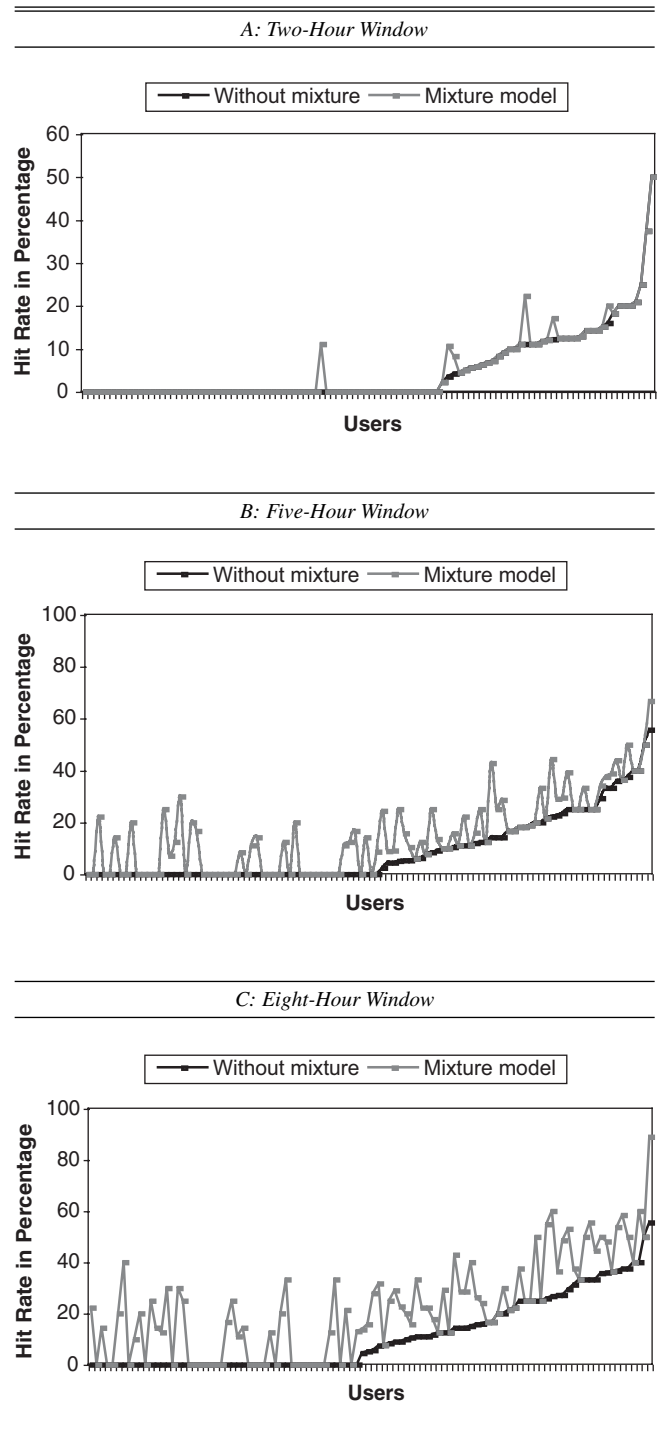
CONCLUSIONS

In the context of search-engine visits, we develop a mixture model that explicitly incorporates user schedules. We show that a model that accounts for periodicities fits the search-engine visit data well and can more precisely predict the timing of a subsequent visit.

A major benefit of our framework is that it is nested within the PHM family of models, which is used extensively in research and practice. Our mixture model is parsimonious, having only two additional parameters to be estimated. In addition to providing a better fit, our model provides better predictions of search-engine visits at a highly disaggregate level, such as the hourly data in our study. Predictive hit rates show that focusing on just a few hours can lead to large benefits to the firm. For example, if a firm focuses only on the best eight hours, it can correctly anticipate a user visit 30% of the time, whereas a traditional model achieves a hit rate of only 15%. Superior hit rates are achieved not only at the aggregate level but also at the individual level.

⁹Note that in Figure 4, the hit rate for approximately 50% of users is zero in a five-hour window with the traditional model, but the mixture model predicts a zero hit rate for only approximately 24% of users. Moreover, many users have no more than three or four observations in the hold-out sample, which leads to a lower hit rate.

Figure 4
INDIVIDUAL LEVEL HIT RATE FOR DIFFERENT TIME WINDOWS



We also incorporate covariates in our model. As does the popular press, we find that consumers tend to decrease their use of search engines over time. At the same time, the results indicate that the use of nonsearch features (e.g., e-mail, chat, news, stock quotes) increases the frequency of visits. Moreover, our results show that people who use the nonsearch features do not show a decline in search-engine use with time. Therefore, despite some doubts of long-term

profitability of search engines in light of decreasing usage, our results suggest that nonsearch features lead to regular and continued use.

Although we have tested our model with data on search-engine use, our model generalizes to many other on- and offline product categories. Habit formation has a long history in almost every field of user behavior. We observe the periodicity in usage of not only search engines but also many other portal sites, newspaper sites, ShopBots, e-mail sites such as Hotmail, and other e-tailers. Moreover, such regular interpurchase times are common even in household panel (grocery purchase) data, and researchers have noted the need to model such periodicities explicitly (Kahn and Morrison 1989). Therefore, our model will benefit analyses across various settings.

A limitation of our study is that we have not conditioned our analysis on the Internet log-in itself. Further research might address Web site design to measure how features of a Web site affect user behavior. Such research would almost certainly require experimentation rather than empirical data of the type used herein.

REFERENCES

- Boatwright, Peter, Sharad Borle, and Joseph B. Kadane (2003), "A Model of the Joint Distribution of Purchase Quantity and Timing," *Journal of the American Statistical Association*, 98, 564-72.
- Bradlow, Eric T. and David C. Schmittlein (2000), "The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science*, 19 (Winter), 43-62.
- Brynjolfsson, Erik and Chris F. Kemerer (1996), "Network Externalities in Microcomputer Software: An Econometric Analysis of the Spreadsheet Software," *Management Science*, 42 (12), 1627-47.
- Buchwalter, Charlie, Marc Ryan, and David Martin (2001), "The State of Online Advertising: Data Covering Fourth Quarter 2000," (February), (accessed March 29, 2002), [available at <http://www.adrelevance.com/intelligence/report20.pdf>].
- Chiu, Willy (2001), Personal e-mail communication with IBM High-Volume Web Site Team Manager, (April).
- Crockett, Roger O. (2000), "How to Bridge America's Digital Divide," *BusinessWeek*, (May 8), 56.
- Datta, Anindya, Kaushik Datta, Helen Thomas, Debra VanderMeer, and Suresha Krithi Ramamritham (2002), "Proxy-Based Acceleration of Dynamically Generated Content on the World Wide Web: An Approach and Implementation," paper presented at the Association for Computing Machinery Special Interest Group on Management of Data Conference, Madison, WI (June 3-6).
- Dunn, Richard, S. Reader, and Neil Wrigley (1983), "An Investigation of the Assumptions of the NBD Model as Applied to Purchasing at Individual Stores," *Applied Statistics*, 32 (3), 249-59.
- Ehrenberg, A.S.C. (1972), *Repeat Buying: Theory and Applications*. Amsterdam: North-Holland Publishing Company.
- Elbers, Chris and Geert Ridder (1982), "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model," *Review of Economic Studies*, 49, 403-411.
- Gönül, Füsün and Kannan Srinivasan (1993), "Consumer Purchase Behavior in a Frequently Bought Product Category: Estimation Issues and Managerial Insights from a Hazard Function Model with Heterogeneity," *Journal of the American Statistical Association*, 88 (424), 1219-27.
- Grover, Rajiv and Vithala R. Rao (1989), "Inferring Competitive Market Structure Based on a Model of Interpurchase Intervals," *International Journal of Research in Marketing*, 5 (1), 55-72.
- Gupta, Sunil (1991), "Stochastic Models of Interpurchase Time with Time-Dependent Covariates," *Journal of Marketing Research*, 28 (February), 1-15.
- Heckman, James J. and Burton Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52 (March), 271-320.
- Jain, Dipak C. and Naufel J. Vilcassim (1991), "Investigating Household Purchasing Decisions: A Conditional Hazard Function Approach," *Marketing Science*, 10 (1), 1-23.
- Kahn, Barbara E. and David C. Schmittlein (1989), "Shopping Trip Behavior: An Empirical Investigation," *Marketing Letters*, 1 (December), 55-70.
- Kraut, Robert E., Tridas Mukhopadhyay, Janusz Szczypula, Sara Kiesler, and William Scherlis (1999), "Information and Communication: Alternative Uses of the Internet in Households," *Information Systems Research*, 4 (10), 287-303.
- Moe, Wendy W. and Peter S. Fader (2002), "Capturing Evolving Visit Behavior in Clickstream Data" working paper, Wharton School, University of Pennsylvania.
- Morrison, Donald G. and David C. Schmittlein (1988), "Generalizing the NBD Model for Customer Purchase: What Are the Implications and Is It Worth the Effort?" *Journal of Business and Economic Statistics*, 6 (April), 145-66.
- Mukhopadhyay Tridas, Uday Rajan, and Rahul Telang (2002), "Competition Between Internet Search Engines," paper presented at the Institute of Electrical and Electronics Engineers Computer Society's Hawaii International Conference on System Sciences.
- Park, Young Hoon and Peter S. Fader (2002), "Modeling Browsing Behavior at Multiple Web Sites," working paper, Wharton School, University of Pennsylvania.
- Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321-40.
- Saha, Atanu and Lynette Hilton (1997), "Expo-Power: A Flexible Hazard Function for Duration Data Models," *Economics Letters*, 54 (July), 227-33.
- Seetharaman, P.B. and Pradeep K. Chintagunta (2002), "The Proportional Hazard Model for Purchase Timing: A Comparison of Alternative Specifications," *Journal of Business and Economic Statistics*, 21 (3), 1-15.
- Telang, Rahul, Tridas Mukhopadhyay, and Ronald Wilcox (2001), "An Empirical Analysis of Internet Search Engine Choice," working paper, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University.