

Modeling Online Browsing and Path Analysis Using Clickstream Data

Alan L. Montgomery

Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213,
alan.montgomery@cmu.edu

Shibo Li

Rutgers University, 228 Janice Levin Building, 94 Rockafeller Road, Piscataway, New Jersey 08854,
shibo_li@rbsmail.rutgers.edu

Kannan Srinivasan

Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213,
kannans@andrew.cmu.edu

John C. Liechty

Pennsylvania State University, 710 M Business Administration Building,
University Park, Pennsylvania 16802, jcl12@psu.edu

Clickstream data provide information about the sequence of pages or the path viewed by users as they navigate a website. We show how path information can be categorized and modeled using a dynamic multinomial probit model of Web browsing. We estimate this model using data from a major online bookseller. Our results show that the memory component of the model is crucial in accurately predicting a path. In comparison, traditional multinomial probit and first-order Markov models predict paths poorly. These results suggest that paths may reflect a user's goals, which could be helpful in predicting future movements at a website. One potential application of our model is to predict purchase conversion. We find that after only six viewings purchasers can be predicted with more than 40% accuracy, which is much better than the benchmark 7% purchase conversion prediction rate made without path information. This technique could be used to personalize Web designs and product offerings based upon a user's path.

Key words: personalization; multinomial probit model; hierarchical Bayes models; hidden Markov chain models; vector autoregressive models

History: This paper was received November 18, 2002, and was with the authors 7 months for 3 revisions; processed by Peter Fader.

1. Introduction

One of the original promises of the Web was that online stores would be able to fully realize the potential of interactive marketing (Blattberg and Deighton 1991, Hoffman and Novak 1996, Alba et al. 1997) through personalization (Pal and Rangaswamy 2003, Ansari and Mela 2003). Currently, online stores target visitors (Mena 2001) using many types of information, such as demographic characteristics, purchase history (if any), and how the visitor arrives at the online store (i.e., did the user find the site through a bookmark, search engine, or link on an email promotion). Another potentially rich—but underutilized—source of information is clickstream data, which record the navigation path that a user takes through the website (Montgomery 2001). Unfortunately, marketers have lacked a methodology for analyzing path information (Bucklin et al. 2002). Our paper proposes a new model that draws upon past work in choice modeling (Rossi et al. 1996, Paap and Franses 2000, Haaijer

and Wedel 2001) to extract information from the path. In particular, we develop a statistical model that analyzes the page-by-page viewings of a visitor as he browses through a website.

Path data may contain information about a user's goals, knowledge, and interests. The path brings a new facet to predicting consumer behavior that analysts working with scanner data have not considered. Specifically, the path encodes the sequence of events leading up to a purchase, as opposed to looking at the purchase occasion alone. To illustrate this point consider a user who visits the Barnes and Noble website, barnesandnoble.com (B&N). Suppose the user starts at the home page and executes a search for "information rules," selects the first item in the search list, and is directed to a product page with detailed information about the book *Information Rules* by Shapiro and Varian (1998). Alternatively, another user arrives at the home page, goes to the business category, surfs through a score of book descriptions, repeatedly back-

ing up and reviewing pages, and finally views the same *Information Rules* product page.

Which user is more likely to purchase a book: the first or second? Intuition would suggest that the directed search and the lack of information review (e.g., selecting the back button) by the first user indicates an experienced user with a distinct purchase goal. The meandering path of the second user suggests a user who has no specific goal and is unlikely to purchase, but is simply surfing or foraging for information (Pirolli and Card 1999). It would appear that a user's path can inform about a user's goals and potentially predict future actions.

Our proposed statistical model can make probabilistic assessments about future paths, including whether the user will make a purchase. Our results show that the first user is more likely to purchase. Moreover, our model can be applied generally to predict any path through the website. For example, which user is more likely to view another product page or leave the website entirely within the next five clicks? Potentially this model could be used for website design or setting marketing mix variables. For example, the site design could be dynamically changed by adding links to helpful pages if it is known that a user is less likely to purchase, while the site could become more streamlined if it is known that a user is likely to purchase. A simulation study using our model suggests that purchase conversion rates could be improved using the prediction of the model, which could substantially increase operating profits.

From a marketing perspective, there has been recent interest in mining Web data to predict purchase conversion (Moe and Fader 2004, Moe et al. 2002, Park and Fader 2004). These studies have focused on web-browsing behavior using session-level data. These aggregate data are quite different from the page-level clickstream data we consider. One criticism of aggregate clickstream data is that sequential information is lost, while in our click-by-click level analysis it is retained. Because websites must interact with users dynamically, these sequencing data are crucial.

Sismeiro and Bucklin (2003) do consider some sequencing information. Specifically, they model the completion of tasks that correspond with groups of Web pages. However, our work is much more detailed because we are modeling page-level movements through a website and not through collections of pages that correspond to tasks. This requires our model to be much more flexible because the paths we observe do not have nice, sequential properties, as do Sismeiro and Bucklin's model.

We also contrast our work to that of Ansari and Mela (2003), who consider the personalization of email messages—but whose work could potentially be applied in a clickstream environment. Again, the

basic difference is the type of data we consider; the data dictate many modeling differences. Their data are derived from user clicks on hyperlinks to personalized emails. These emails may be separated by many days, hence modeling the dependence between choices is not crucial. Their choice model assumes independence both within a page and across time. In contrast, our goal is to focus on the sequence of the choices made, which tend to occur within seconds of one another. Hence we find it critical to introduce correlation across choices, as well as to introduce time series elements to capture the timing of the choices.

2. Clickstream Data

Given that clickstream data may be unfamiliar to many readers, we first explain our data, describe how they are collected, and conduct an exploratory data analysis to motivate the model we introduce in §3. Our data are derived from a panel of Web users maintained by Jupiter Media Metrix, which is now known as Comscore Media Metrix (CMM). CMM randomly recruits a representative sample of personal computer (PC) users and tracks these users' usage at home (Coffey 1999). These users agree to install a computer program (or PC meter) that runs in the background and monitors computer usage. It records any URL viewed by the user in his browser window. Because the meter records the actual pages viewed in the browser window, it avoids the caching problems commonly found by recording page requests at an Internet Service Provider (ISP) or a Web server. However, the meter does not distinguish how the user navigates between pages (e.g., whether the user selects a hyperlink, a bookmark, or directly types in the URL to navigate to a page). In addition, the meter does not record the content of the page, but records only the URL.

2.1. Descriptive Analysis and Defining the Path

Our dataset consists of 1,160 users who visited Barnes and Noble online by going to barnesandnoble.com, books.com, or bn.com between April 1, 2002, and April 30, 2002. (We abbreviate references to barnesandnoble.com as B&N.) This dataset represents all users in the full CMM panel who visited B&N for April 2002, which is almost 6% of the full panel. We selected B&N for our analysis because it is a popular online bookstore and has a relatively clean and stable site structure compared with other online stores. Although we use clickstream data collected by CMM, our methodology could be applied directly to clickstream data collected from B&N's Web servers. Again, our reason for using CMM clickstream data is that they are available to the authors, and that they

are more complete and have a cleaner format than Web-server logs (Pitkow 1997).

First, we define the following terms to describe Web browsing: page request, page viewing, and session. A *page request* refers to a user's requesting a URL through her browser program. In turn this page request will appear as a hit in the server's log file. A *page viewing* refers to the actual rendering of a page request in the user's browser window. A user may hit the back button in his browser window to review a page, which will generate another page viewing but not a page request. (Instead, the browser program will render the page from a previously stored or cached copy.) Often, pages are viewed multiple times, so page viewings generally exceed page requests. Finally, a *session* is defined as a period of sustained Web browsing or a sequence of page viewings. If a user has not viewed any pages for 20 minutes, we assume that the viewing session has ended and that the next page viewing marks the beginning of a new session. Sessions include all of a user's page viewings, both at B&N and other sites.

Our 1,160 users requested 9,180 unique URLs or pages at B&N on 14,512 viewing occasions over the course of 1,659 sessions. The average B&N page was viewed 1.5 times. The average number of B&N pages viewed during a session was 8.75. The number of B&N viewings during a session ranged from 2 to 239, with the median being 5 viewings. Most users had only one or two sessions that included activity at B&N; fewer than 25% of our users had more than two sessions. Out of these 1,659 sessions, 114 of these sessions had a purchase (two sessions had two purchases), which yields a purchase conversion rate of 7%. (This rate is higher than the industry average, either due to B&N's success or the fact that our estimate is not contaminated by automated traffic from search engines and robots, as is commonly the case.)

The descriptive statistics for the demographic information about our user sample is given in Table 1.

Table 1 Demographic Characteristics of 1,160 Panelists

Variable	Mean	Std dev	Min	Median	Max
Age	45.89	14.62	9	46	89
Age ² (square of age)	2,326.48	1,331.68	81	2,209	7,921
Male	0.47	0.50	0	0	1
White	0.77	0.42	0	1	1
Children under 18 in the house	0.40	0.49	0	0	1
Married	0.29	0.45	0	0	1
Some college education	0.82	0.39	0	1	1
High income (>\$50,000)	0.32	0.47	0	0	1
Medium income (\$25,000–\$50,000)	0.35	0.48	0	0	1

Note. All of the means are proportions except age and age², which are continuous variates.

All of our demographic variables, except age, are coded as dummy variables. Notice that the average user is 46 years old with a range from 9 to 89 years old; slightly more than half are female; most are white; most have some college education; and most have higher-than-average incomes. While it is unlikely that B&N would have such detailed information, we include this information to assess its predictive power; in the future it is possible that online retailers could purchase these data from online vendors.

Potentially, the clickstream is a very rich data source because the full text and HTML content of each URL is known (or can be recaptured). Practically, however, without some structure it is difficult to analyze these free-format and textual data. We choose to do so by focusing on the category that corresponds with each page viewed. Every page is classified into one of seven categories: Home, Account, Category, Product, Information, Shopping Cart, Order, and Enter/Exit pages. (See Technical Report Appendix C at <http://mktsci.pubs.informs.org> for our text matching algorithm to categorize pages and an example session.)

Redish (2002) proposed this categorization scheme as a common taxonomy across e-commerce sites based on a task analysis of what users want to do on an e-commerce sites from a human-computer interaction standpoint. Moe et al. (2002) also employed a similar classification scheme. The home page is a common starting point for new tasks. Account pages are used for logins, address changes, and order status reviews. Category pages present lists of items, categories, or search results. Product pages contain detailed product information, item description, price information, availability, and product reviews. Shopping cart pages are used to add or delete products and enter purchase information. Order pages are confirmation pages that denote an order has been placed. The enter/exit category is used to denote a non-B&N page and denotes either the beginning or end of a B&N session.

We augment our data by writing a perl script that queries B&N to reconstruct the page content viewed, because these data are not collected by CMM. The text of the page was parsed and scanned for information about the presence of price information, promotion images, banner ads, and the numbers and types of hypertext links on the page. (Some variables, such as the number of links to the shopping cart or pictures on a page, are omitted due to multicollinearity.) Additionally, we include a variable that measures whether or not a purchase was made at B&N during the user's last session and whether or not the visit occurred during a weekend. To capture timing information we compute the time between page viewings in seconds.

Table 2 Descriptive Statistics for the 9,180 Unique B&N Pages

Variable	Mean	Std dev	Min	Med	Max
Presence of price information on page (proportion)	0.45	0.50	0	0	1
Promotional image present (proportion)	0.83	0.37	0	1	1
Presence of banner advertisement (proportion)	0.03	0.16	0	0	1
Number of links to a home page	2.4	1.0	0	3	5
Number of links to a product page	10.1	18.1	0	0	110
Number of links to an account page	2.0	1.1	0	2	9
Number of links to an information page	28.8	33.9	0	17	303
Whether made a B&N purchase during last session	0.03	0.18	0	0	1
Time since last viewing (seconds)	7.2	66.3	1	1	1,193
Whether visit is on weekend (proportion)	0.28	0.45	0	0	1
Cum. no. of viewings at B&N during session (visit depth)	8.8	16.4	1	5	238
Cum. no. of viewings at other sites during session	44.3	84.6	0	17	891
Cum. no. of viewings at other bookstores during session	4.3	17.3	0	0	174

Finally, we have three measures of the cumulative number of pages viewed up to that point during the session: pages viewed at B&N, at other sites, and at other bookstores. Again, B&N may not have access to these measures of external activity, but we include them to understand how helpful these data could be in predicting paths. Descriptive statistics for these variables are given in Table 2.

Notice that in Table 2 we find that 45% of the pages viewed in our data have price information, while 83% of the pages have promotion information (e.g., free shipping or discounts). Only about 3% of the pages have banner ads provided by Double Click Inc. These banner ads only redirect a user within the B&N site

and do not take her to other websites. For example, a book publisher may wish to promote his book with a link to a corresponding B&N product page. We find many hypertext links to category pages, product pages, and information pages, while there are few links to the home, shopping cart, or account pages (although these links tend to be prominently displayed at the top of the page). The average time duration between page viewings is 7.2 seconds, although this average is highly influenced by many repeat viewings that last for only a second. Notice that during an average viewing users have cumulatively viewed 44.3 pages at other sites during their session, and 4.3 pages at competing online bookstores (such as amazon.com, borders.com, booksamillion.com, albooks.com, and so on). The cumulative variables are reset to zero whenever a session starts. Notice that the cumulative variables may not be zero when the user starts at B&N because pages may have already been viewed at other sites during the session, before the first B&N page is viewed.

2.2. Describing Page Transitions with a Markov Model

We can compactly represent paths using the first initial of our categories as an abbreviation. For example, the string HCPE would denote a user who starts at a home page to search for a book, moves to a category page to review the results, and concludes her session at a product page after considering an individual item. To illustrate our data we list the sessions of 10 selected users in Table 3. Notice the first five users do not make a purchase, while the second five do (notice the O or order page in the path). To illustrate these paths consider the first user. This user has a total of 44 viewings; his B&N session started at the home page, and then he viewed many category pages with only a couple of interruptions to product pages. Finally, the user ended the session without purchasing. Next consider user 6. This user started by visiting the home page, looked at an information page, and then moved to an account page. These actions suggest the session

Table 3 Listing of Category of Viewings for Selected User Sessions

Group	User	Session
No purchase	1	HCCCCCCCCPCCPCCCCCCCCCCCCCCCCCCCCCCCCCCE
	2	IHHE
	3	IE
	4	IHICPPPC
	5	IHHIICIE
Purchase	6	HIAAAIAIHHIICICICICICICIPPIPIPIPIIICSIIPPPPIPIPSISISISSOIIHIE
	7	HCCPPPCPCCCCCCCCPSCSCSPCCPCPCCCCCSAAAAAASSOIIHISASCCE
	8	IICICPCPPPCPCICICPCCPCPPPIPSIIAASSIIISOIE
	9	IISASSSOIE
	10	IPPPSASSSSOIAAAHCCPCCCCCE

Note. Types of pages: H = Home; A = Account; C = Category; P = Product; I = Information; S = Shopping Cart; O = Order; E = Exit.

Table 4 Sample Transition Matrix, Marginal and Initial Probabilities for Categories of Viewings

Category	Category of previous viewing							
	Home	Account	Category	Product	Inform.	ShopCart	Order	Exit
Category of current viewing								
Home	0.23	0.01	0.01	0.01	0.10	0.02	0	0.16
Account	0.01	0.69	0.01	0.01	0.02	0.15	0	0.01
Category	0.17	0.02	0.60	0.31	0.15	0.05	0	0.16
Product	0.01	0	0.20	0.43	0.10	0.05	0	0.05
Information	0.25	0.06	0.08	0.12	0.46	0.15	0.87	0.61
Shopping cart	0.01	0.16	0.01	0.03	0.02	0.45	0.13	0.01
Order	0	0	0	0	0	0.10	0	0
Exit	0.32	0.06	0.09	0.09	0.14	0.02	0	0
Marginal	0.06	0.05	0.32	0.17	0.23	0.05	0.01	0.11
Initial prob.	0.16	0.02	0.16	0.06	0.60	0.01	0	0

Note. Notice that the columns sum to one, and there are a total of 14,512 observations.

is more purchase directed. This is confirmed by the user's frequent searches for products, some of which are added to the shopping cart later on. Finally, this user made a purchase, checked her order status, and continued to the home page before exiting.

To better summarize the transitions between categories we report the sample transition matrix in Table 4 along with the marginal and initial probabilities of viewing a category. The transition matrix provides an estimate of a first-order Markov model (cf. Cadez et al. 2000). Additionally, in the last row of Table 4 we report the probability that this was the category of the first B&N page viewed during a session. Clearly, users do not randomly enter the site, but tend to start either at an information (60%), home (16%), or category page (16%). Notice that there is high degree of persistence for viewing the same type of Web page again; the diagonal elements for all types of Web pages except home, order, and exit range from 0.43 to 0.69. The lack of transitions between exit pages is due to our definition of a session, in which we treat consecutive viewings at other websites as a single exit observation. Finally, note that there is a 23% probability of viewing the home page in the next viewing even when the user is already at the home page. Home page to home page viewings usually arise when a user works with another program between viewings (for example sends an email message) or visits a non-B&N page in between home page viewings.

The transition matrix contains information about the sequence of pages viewed. We presume that users clicking on hypertext links generate most of the sequences. However, pages may be selected through the browser's interface (backward and forward buttons, bookmarks, or history). Hence, it is quite possible that two pages that are not linked together will appear as consecutive viewings in our dataset due to the use of the browser interface. This is especially important in understanding that even though the

website may not offer links to all the other categories, it is possible that the user could navigate to the page using another control. Navigation is complicated by the fact that a user could have multiple windows (Web browsers) open in different areas of the website. Unfortunately, in our dataset we see all viewings as a continuous stream and do not know with which window a viewing is associated or even if there are multiple windows open.

Many of the patterns in the transition matrix are due to the construction of the website. For example, the many hypertext links from a category page to a product page make it likely that the user may use one of these links to navigate the website. However, navigation is not restricted to these links. Suppose a user requests additional information about shipping that appears as a popup window; the user may continue browsing through the website but leave the popup window open. The user may reselect the popup window later in the session even though there is no link to the popup on the current Web page. There is one exception to this stochastic approach that we make for the order page: The only way for a user to get to an order page is from the shopping cart page; even if multiple windows are open, the previous page must be a shopping cart page. Hence, we assume that a user cannot select the order page unless he is currently on a shopping cart page.

3. A Dynamic Multinomial Probit Model of Web Browsing

Our exploratory analysis from the previous section prescribes two important elements needed in a formal statistical model: a categorical choice model of Web page movement and memory to capture dependence in the sequences chosen. Additionally, past research suggests that we need to account for other possible features of the data, such as consumer heterogeneity, the use of user characteristics and demographics to

explain some of this heterogeneity, dynamic behavior, the use of covariates to explain transitions between pages, and general error covariance patterns to capture unexplained transitions. The simple first-order Markov model proposed in the previous section cannot accommodate all these facets. The main problem is that the first-order Markov model has a one-period memory, i.e., the present viewing given the last one is independent of the previous path.

Modeling categorical data can be accommodated using a multinomial choice model, such as a multinomial probit model. The probit model makes it easy to incorporate covariates that may explain Web navigation choices. However, the traditional probit model has no memory. To overcome this problem we propose a dynamic multinomial probit model. Specifically, we introduce a vector autoregressive component to the model that can capture dynamics in choice (Paap and Franses 2000, Haaijer and Wedel 2001). Additionally, we incorporate a correlated error structure in the model to capture unexplained patterns (i.e., those that cannot be attributed to our covariates), which also overcomes the IIA property of multinomial logit models. We frame our model in the context of a hierarchical Bayesian model to allow for heterogeneity across users (Rossi and McCulloch 2000, Rossi et al. 1996). As a final point, we note that consumer behavior research has found that users may have goal-directed or exploratory search (Moe 2003, Janiszewski 1998) or flow and nonflow states (Hoffman and Novak 1996). Hence, we incorporate a mixture process where the model parameters for an individual can switch during the course of a session to reflect the possibility that browsing behavior may be quite different and change suddenly, depending upon a user's current goals or state of mind. This dynamic mixture process has not been previously considered in a choice context. Overall, our contribution is a conjunction of these techniques applied to a substantively new problem.

3.1. Model Specification

Formally, we assume that user i has latent utility U_{iqtc} associated with viewing a page in category c on viewing occasion t of session q , where there are totals of I users, C categories, Q_i sessions for user i , and T_{iq} viewings for the q th session of user i . In our dataset, $I = 1,160$, $C = 8$, Q_i ranges from 1 to 17, and T_{iq} ranges from 2 to 239. The consumer selects the category (or more precisely, the page associated with the category, as summarized in Table 3) that has the highest utility from among those that are available. We denote the user's selection as Y_{iqtc} , which yields the following observational equation:

$$Y_{iqtc} = \begin{cases} 1 & \text{if } \mathbf{u}_{iqtc} = 1 \text{ and } U_{iqtc} = \max(\mathbf{U}_{iq} \langle \mathbf{u}_{iq} \rangle) \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathbf{U}_{iq} = [U_{iq1}, U_{iq2}, \dots, U_{iqC}]$ is a $C \times 1$ vector of latent utilities, \mathbf{u}_{iqtc} is an indicator variable that equals one when category c is available for user i during the t th viewing of session q , $\mathbf{u}_{iq} = [\mathbf{u}_{iq1}, \mathbf{u}_{iq2}, \dots, \mathbf{u}_{iqC}]$, and the operator $\mathbf{U} \langle \mathbf{u} \rangle$ denotes the set of elements from the vector \mathbf{U} whose corresponding indicator operand (\mathbf{u}) equal to one. In our model \mathbf{u}_{iq} is a vector of ones except for the element that corresponds with the order page, which only equals one when the previous page is a shopping cart. The purpose of allowing only a subset of pages to be selected is to eliminate impossible transitions (i.e., moving to the order page without selecting the shopping cart on the preceding page). We note that our technique could easily be modified to include other exceptions, but as noted in §2 this is the only one present in our dataset.

The latent utility vector (\mathbf{U}_{iq}) is modeled as

$$\mathbf{U}_{iq} = \mathbf{\Gamma}_{i, s(iqt)} \mathbf{X}_{iq, t-1} + \mathbf{\Phi}_{i, s(iqt)} \mathbf{U}_{iq, t-1} + \mathbf{\epsilon}_{iq, s(iqt)}, \\ \mathbf{\epsilon}_{iq, s(iqt)} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}_{s(iqt)}), \quad (2)$$

where subscript $s(iqt)$ defines the user's state ($s = 1, \dots, S$) for user i during session q at viewing t . For notational simplicity we may drop the dependence of state on the user, session, and viewing ($\langle iqt \rangle$) and assume that this dependence is understood. Also, $\mathbf{X}_{iq, t-1}$ is a vector of the $L - 1$ covariates given in Table 4 for user i during session q at time t ($L = 14$ as the first element of \mathbf{X} is unity to incorporate an intercept), $\mathbf{\Gamma}_{is}$ is a $C \times L$ parameter matrix, $\mathbf{\Phi}_{is}$ is a $C \times C$ matrix of autoregressive coefficients, and the error vector $\mathbf{\epsilon}_{iqts}$ follows a multivariate normal distribution. We have included only the first lagged vector of utility which is supported in our empirical analysis, although future researchers may want to consider more general lag structures.

Notice that the $\mathbf{X}_{iq, t-1}$ covariates associated with the utility at time t (see Equation (2)) correspond to the page contents of the previous viewing at time $t - 1$. We assume that a user decides which category to view at time t largely based on the information being viewed at time $t - 1$ (e.g., the number of links of the page, type of information display, time page viewed, and so on), because the page at time t has not yet been displayed. A potential direction for future research is to replace the observed covariates with predicted values, because the user views only the covariates of the selected alternative. Hence our current framework relies on an assumption of the user's rational expectation of these covariates. Additionally, the initial utility values for the first viewing of each session (\mathbf{U}_{iq0}) are not observed but are inferred from the observed data (see Technical Appendix at <http://mktsci.pubs.informs.org>). To ensure identification of our model, we follow the usual practice of setting utility of a base category to zero, which in our

case is the enter/exit category ($c = 8, U_{igt8} = 0$), and from Equation (2), and hereafter we set $C = 7$ to refer to the remaining categories. Also, for identification we set the first element of the error covariance matrix to unity, $[\Sigma_s]_{1,1} = 1$.

3.2. Modeling State Transitions

We consider both a zero and first-order hidden Markov process to model the state variate ($s\langle iqt \rangle$). The zero-order Markov process states that there is a vector \mathbf{v}_i —which is independent of time and path—that defines the probability of user i being in state s :

$$\Pr[s\langle iqt \rangle = s \mid \mathbf{v}_i] = v_{is},$$

$$\text{where } \mathbf{v}_i = [v_{i1} \ v_{i2} \ \cdots \ v_{iS}]'. \quad (3)$$

We formulate the first-order Markov process by assuming that there is a hidden, continuous-time Markov chain, D_{igt} , which indicates the state. Note that $s\langle iqt \rangle$ is only defined at integer time values, while D_{igt} is continuous and equal to $s\langle iqt \rangle$ at integer values.¹ The waiting time between transitions (w_{igt}) in our continuous-time domain follows an exponential distribution:

$$\Pr[w_{igt} \mid \boldsymbol{\lambda}_i] = \lambda_{i, D_{igt}} \exp\{-w_{igt} \lambda_{i, D_{igt}}\},$$

$$\text{where } \boldsymbol{\lambda}_i = [\lambda_{i1} \ \lambda_{i2} \ \cdots \ \lambda_{iS}]', \quad (4)$$

where λ_{is} is an intensity parameter for state s , and the expected waiting time till the next is the inverse of this parameter ($1/\lambda_{is}$). Given that a transition has occurred, the transition matrix (\mathbf{P}_i) that defines our first-order Markov process is

$$\Pr[D_{igt, t+w_{igt}} = s \mid D_{igt} = g, \mathbf{v}_i, \mathbf{P}_i] = P_{igs} \quad \text{if } t > 1,$$

$$\text{where } \mathbf{P}_i = \begin{bmatrix} 0 & P_{i12} & \cdots & P_{i1S} \\ P_{i21} & 0 & \cdots & P_{i2S} \\ \vdots & \vdots & \ddots & \vdots \\ P_{iS1} & P_{iS2} & \cdots & 0 \end{bmatrix}, \quad (5)$$

where P_{igs} denotes the conditional distribution for user i to switch to state s given the previous state was g ; hence the rows sum to one. Notice that the diagonal elements are zero because same-state transitions are captured through the waiting time. Finally,

¹ We define time between viewings as a standard time unit and not as the time of day. The elapsed clock time between viewings is irregular and may include viewings at other sites or noncomputer activities (e.g., getting a cup of coffee). The disadvantage of this approach is a loss of information. However, we do include elapsed time as a covariate but find that it is a poor predictor of browsing, which supports our standardization of time.

the initial state probability for the first viewing of a session is

$$\Pr[D_{igt} = s \mid \mathbf{v}_i] = v_{i, D_{igt}} \quad \text{if } t = 1, \quad (6)$$

where we redefine the probability vector \mathbf{v}_i as the initial starting probabilities.

To identify the states we place a restriction on the means of the latent variable U_{igt} . Specifically, we introduce two metrics, W_{igts}^a and W_{igts}^b , to capture the user's tendency to browse (e.g., surf or participate in a nonpurchase orientation) or deliberate (e.g., have a focused navigation or a purchase orientation) during a session. W_{igts}^a is the sum of the expected value of U_{igt} for the account, shopping cart, and order pages, while W_{igts}^b is the sum of the expected value of home, category, product, and information pages. We assume state 1 corresponds with the highest browsing orientation (lowest value of W_{igts}^a and highest value of W_{igts}^b) and state S is the state that exhibits the most deliberation orientation (highest value of W_{igts}^a and lowest value of W_{igts}^b). That is, $W_{igts}^a \geq W_{igt(S-1)}^a \geq \cdots \geq W_{igt1}^a$ and $W_{igt1}^b \geq W_{igt2}^b \geq \cdots \geq W_{igtS}^b$. While we believe these metrics are useful in describing the likely cognitive state of a user, this is merely conjecture on our part and can be thought of as convenient labels to refer to each state.

An alternative to this waiting time formulation in continuous time would be to assume a discrete-time Markov chain with a transition matrix that had a nonzero diagonal. The advantage of the waiting time formulation is that it is more amenable to the Reversible Jump Algorithm that we use to estimate this model (see Technical Appendix B at <http://mktsci.pubs.informs.org>, step 9). However, these two formulations are equivalent if the transitions occur at integer values. To illustrate this equivalence, we show that the transition matrix for a two-state Markov model can be parameterized in terms of the waiting times (the i subscript is suppressed for clarity):

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \exp\{-\lambda_1\} & \exp\{-\lambda_1\} \\ \exp\{-\lambda_2\} & 1 - \exp\{-\lambda_2\} \end{bmatrix}. \quad (7)$$

This equivalence relies on the lack of memory property of the exponential process (cf., Johnson et al. 1994, Ch. 19), in which the conditional distribution of an exponential process given that an event has not occurred is equivalent to the original marginal distribution. The primary difference between these formulations is that our model is more general because the transitions may occur at noninteger values.

3.3. Specification of the Hyper Distributions

Following the usual hierarchical Bayesian framework (Rossi et al. 1996), we incorporate heterogeneity across users by assuming that Γ_{is} has a random coefficient specification. Specifically, the l th column of Γ_{is} , which we denote as γ_{ils} , a $C \times 1$ vector, follows a multivariate regression

$$\gamma_{ils} = \Pi_{is} \mathbf{R}_i + \mathbf{s}_{ils}, \quad \mathbf{s}_{ils} \sim MVN(\mathbf{0}, \Psi_s), \quad (8)$$

where \mathbf{R}_i is the $K \times 1$ vector of demographic measures, plus an intercept, listed in Table 1 for user i ($K = 10$, because the first element is unity to incorporate an intercept), Π_{is} is a $C \times K$ parameter matrix, and Ψ_s is a $C \times C$ covariance matrix.

We assume that the VAR coefficients are drawn from a hyperdistribution:

$$\text{vec}(\Phi_{is}) \sim MVN(\text{vec}(\bar{\Phi}_s), \Omega_s), \quad (9)$$

where $\bar{\Phi}_s$ is a $C \times C$ matrix of autoregressive coefficients and Ω_s is a $C^2 \times C^2$ covariance matrix. Finally, we assume that the row vectors of the transition matrix and the vector of initial probabilities follow a Dirichlet distribution, and the waiting times follow a gamma distribution:

$$\mathbf{P}_{ij} \sim D(\tau_j), \quad \nu_{is} \sim D(\alpha_s), \quad \lambda_{is} \sim \Gamma(\hat{\lambda}_{sh}, \check{\lambda}_{sc}), \quad (10)$$

where \mathbf{P}_{ij} denotes the j th row of the matrix \mathbf{P}_i , and $\hat{\lambda}_{sh}$ and $\check{\lambda}_{sc}$ denote the shape and scale parameters, respectively.

3.4. Discussion of the Model

We include two dynamic elements in our model of browsing behavior. First, we introduce persistent behavior through a vector-autoregressive (VAR) component of lagged latent utility values (Hamilton 1994, Ch. 11). The idea is that a higher-than-average affinity for a type of page may persist for many viewings. For example, a user may view many product pages consecutively. This dictates the need for incorporating a memory or time series element in the model. Secondly, we allow for abrupt changes in browsing behavior by incorporating a time-varying mixture model. A hidden Markov chain governs the transitions between the states of this mixture model. For example, during a session a user may change her goals and decide to focus on making a purchase, or alternatively decide to just look around instead of making a purchase.

While it might seem redundant to have two time series components in the model, the purpose of each is quite different from the other. The VAR model is meant to capture smoothly trending behavior reflective of a certain type of browsing, while the mixture process with the Markov transitions is meant

to capture abrupt changes in browsing styles. Additionally, the Markov process may be able to capture longer-term dynamics and help lessen the dimensionality of the autoregressive process, leading to a more parsimonious model. To illustrate, consider the conditional mean of the latent utility of a two-state process ($S = 2$), which can be derived from the Markov property of the hidden Markov chain under the assumption of stationarity:

$$\begin{aligned} E[\mathbf{U}_{iqt} | \mathbf{U}_{iq, t-1}] \\ = \Gamma_{i1} \mathbf{X}_{iq, t-1} + \Gamma_{i1} \mathbf{U}_{iq, t-1} + \frac{\exp\{-\lambda_{1i}\}}{\exp\{-\lambda_{1i}\} + \exp\{-\lambda_{2i}\}} \\ \cdot \{(\Gamma_{i2} - \Gamma_{i1}) \mathbf{X}_{iq, t-1} + (\Gamma_{i2} - \Gamma_{i1}) \mathbf{U}_{iq, t-1}\}. \end{aligned} \quad (11)$$

Notice that the state parameters (λ_{1i} , λ_{2i}) control the mixing of the VAR parameters.

The Markov switching-autoregressive models that we employ for our latent process were first proposed in a univariate model by Hamilton (1989) to describe the U.S. business cycle by modeling U.S. real gross national product. Subsequently there has been a great deal of interest in the use of this model for studying macroeconomic processes. Krolzig (1997) studies the multivariate form of the model that we employ. The primary difference with this past work is that we assume our latent process follows this structure and not the observed process. These models all employ discrete jumps between states. Another approach is the smooth-threshold autoregressive model by Teräsvirta (1994), which may offer an interesting direction for future research.

A univariate version of our dynamic probit model has been considered in the binary time-series literature (Kedem 1980a), specifically in the context of a discretized autoregressive-moving average (DARMA) (Kedem 1980b, Keenan 1982). DARMA models are quite flexible in capturing time-series trends in binary processes. Keenan (1982) showed that any stationary time-series process could be modeled as the discretization of a latent continuous process. This approach contrasts with the usual Markov modeling strategy, which models the observed process directly. While DARMA models can well approximate Markov models, the conditional transition probabilities of DARMA processes lose their Markov property (Keenan 1982). In other words, knowledge of the previous state is not sufficient for forecasting the next state. Our use of a VAR model should result in a good approximation to the Markov model described in our exploratory analysis. Recently, there have been several applications of VAR models in marketing choice models to capture brand inertia and loyalty effects (Paap and Franses 2000, Haaijer and Wedel 2001, Seetharaman 2004). These VAR models can be thought of as a multivariate generalization of a univariate DARMA process.

4. Estimation Results

In this section we present the empirical results from estimating the model presented in §3 using the data discussed in §2. We start by considering the fit and predictive performance of various formulations of our model as well as other potential benchmark models, and then continue with a discussion of some specific features and properties of our best model. To provide an overall comparison of the various model specifications we compute the marginal posterior distribution (or the marginal density) following Newton and Raftery's (1994) method and the hit rate. The hit rate refers to the percentage of viewings whose categories are correctly predicted. For example, random guessing should yield odds of 1 in 8 of a correct guess, or a 12.5% hit rate.

As another measure of model adequacy, we compute out-of-sample predictive performance. Each user's sessions are divided into two parts; the earlier sessions are used for estimation and the later ones are used for prediction. If a user has one session, then his data will be used only for estimation. Fractions of a session in the estimation and holdout sample are rounded up and down respectively (e.g., a user with three sessions would have her first two in the estimation sample and the last in the holdout sample). The construction of the validation sample is meant to closely approximate the type of information that B&N would have available for its users, based on past information. There are 1,160 users in the estimation sample and 268 users in the holdout sample, with 9,589 observations and 4,923 observations, respectively. The disparity in the number of users is due to the large number of users with only one session. For the users in the holdout sample, we predict their parameters and the states of the hidden Markov model using only the information from the estimation sample.

4.1. A Predictive Analysis with State Changes at the Page, Session, and User Level

Our model permits a great deal of flexibility with regard to changing the underlying browsing state, because the state can potentially change with every page viewing. For example, a user may begin his session in a state where purchase is unlikely, but then switch later in the session to a state where purchase is likely. However, allowing state changes at the viewing level may result in too much variability. Hence we consider two additional formulations of our model that restricts state changes to the session or user level. Restricting viewings within a session to share a common state reduces the potential of a user switching states in the midst of a session. Restricting all of a user's viewings to a single state permits heterogeneity across users (although not within a user), which can

capture departures from the normality assumption in our hierarchy.

We estimate our model using the assumption that states are allowed to change at the page and session level with both a zero and first-order Markov model, and another set of models where the state is constant for all of a user's viewings (only a zero-order Markov process is estimated, because there is no pre-user information necessary for a first-order Markov process). The question of how many states should be included is an empirical one; hence, we estimate our model for one, two, and three states ($S = 1, 2$, or 3) to allow the data to inform about this parameter. This yields a total of 15 models from which to investigate the amount of within-user heterogeneity. We report the fit and out-of-sample prediction validation in Table 5.

First, notice transitions defined at the page level outperform models estimated at a session or user level. This supports the notion that users are likely to change states in the midst of a session. Hence it is inaccurate to describe a user or an entire user session simply

Table 5 Measures of Fit for Various Dynamic Probit Models

State time	State process	Number of states	Log marginal density	In-sample hit rate (%)	Out-of-sample hit rate (%)
Page	Zero-order	1	−9,378.1	72.05 (0.46)	65.40 (0.68)
		2	−9,016.9	79.44 (0.41)	71.42 (0.64)
		3	−9,064.0	80.34 (0.41)	70.56 (0.64)
	First-order	1	−8,545.4	83.23 (0.38)	79.95 (0.57)
		2	−8,428.3	89.71 (0.31)	83.15 (0.53)
		3	−8,474.0	89.97 (0.31)	81.14 (0.56)
	Zero-order	1	−9,376.1	73.17 (0.45)	61.56 (0.69)
		2	−9,051.0	77.90 (0.42)	70.48 (0.65)
		3	−9,097.7	78.76 (0.42)	66.14 (0.67)
Session	First-order	1	−8,573.5	83.05 (0.38)	73.57 (0.63)
		2	−8,464.9	88.44 (0.33)	81.48 (0.55)
		3	−8,487.0	88.73 (0.33)	78.42 (0.59)
	Zero-order	1	−9,411.1	64.38 (0.49)	61.50 (0.69)
		2	−9,124.2	70.04 (0.47)	64.12 (0.68)
		3	−9,193.8	70.85 (0.46)	63.99 (0.68)
User	Zero-order	1	−9,411.1	64.38 (0.49)	61.50 (0.69)
		2	−9,124.2	70.04 (0.47)	64.12 (0.68)
		3	−9,193.8	70.85 (0.46)	63.99 (0.68)

Note. The standard errors of the hit rates are provided in parentheses below the estimate.

being either purchase or nonpurchase oriented. Second, notice that when the hidden states are governed by a first-order Markov model the fit is superior to a zero-order process. This suggests that the goals, to the extent that they are reflected in a state, show some persistence. It also suggests that the VAR process cannot fully capture a user's behavior, perhaps due to abrupt changes in a consumer's goal, e.g., a user changes from a browsing orientation to a purchase orientation. We also compute Bayes factors following Kass and Raftery (1995) for three of our page-level proposed models: a one-state, two-state, and three-state version of the hidden Markov model. The two-state model is favored over the one-state model by odds of 117:1. Also, the two-state model is favored over the three-state model by odds of 46:1. We can also find similar pattern for our session and user-level models, indicating a two-state model is adequate.

4.2. Predictive Comparisons with Alternative and Nested-Model Specifications

The model presented in §3 is quite general and nests many common models as special cases, such as the multinomial probit and latent-class model. To better understand these cases we show the relationship with these nested models. Additionally, we propose some alternative benchmark models to help evaluate the fit and predictive ability of our models, which are reported in Table 6 and discussed below.

Zero-Order Markov Model. Perhaps the simplest model assumes there is a fixed probability for each user to select a category, which can be described as a multinomial distribution or a zero-order Markov model. Specifically, the probability that user i chooses category c during session q at viewing t is independent of other viewings:

$$\Pr(Z_{iqt} = c) = \delta_c, \quad \text{where } Z_{iqt} = c \text{ iff } Y_{iqt} = 1. \quad (12)$$

Notice there is no multinomial probit or VAR component. To estimate this in a Bayesian specification we employ a diffuse prior on δ_c . This structure can be duplicated in our full model with a single state ($S = 1$) when the only covariate is an intercept and the errors are independent, standard normal variates:

$$U_{iqt} = \gamma_i + \epsilon_{iqt}, \quad \epsilon_{iqt} \sim N(0, I). \quad (13)$$

Our Bayesian framework can yield results similar to the maximum likelihood estimator (MLE) when no individual level covariates are used (\mathbf{R} is a vector of ones) and the covariance of the hyperdistribution is small ($\Psi \rightarrow 0$). Essentially, the individual-level parameters are shrunk to a common, pooled value.

This model is estimated for both a pooled sample (one state) and a split sample of purchasers versus

nonpurchasers (two state). Notice that these states are determined exogenously in contrast with the hidden Markov process proposed in the previous section. The estimation results are reported in Table 6 and show that zero-order Markov models have both the worst fit and predictive ability among all the models considered.

First-Order, Markov Model. The first-order Markov model assumes that the category of the current viewing can be predicted with knowledge of only the past category. Specifically, the probability that user i chooses category c during session q at viewing t will be

$$\Pr(Z_{iqt} = c \mid Z_{iq, t-1} = d) = \delta_{cd}. \quad (14)$$

To estimate this in a Bayesian specification we employ a diffuse prior on δ_{cd} . Although this model is not nested within our framework, the VAR process may provide a good approximation to such a model. As with the zero-order Markov model we estimate a two-state model with each state made up of purchasers or nonpurchasers. The results in Table 6 show an order-of-magnitude increase in fit over the zero-order model, illustrating the importance of memory in path analysis, but the results are inferior to our dynamic multinomial probit model.

Latent-Class Model. A popular model in marketing is the latent-class model (Kamakura and Russell 1989), which occurs as a special case of our model. If the state transitions in our model are restricted to a

Table 6 Measures of Fit for Other Alternative Model Specifications

Model	Log marginal density	In-sample hit rate (%)	Out-of-sample hit rate (%)
Zero-order Markov model (1 State)	-20,410.4	20.48 (0.41)	12.62 (0.47)
Zero-order Markov model (2 States)	-19,458.3	28.18 (0.46)	19.02 (0.56)
First-order Markov model (1 State)	-16,444.5	56.06 (0.51)	51.59 (0.71)
First-order Markov model (2 States)	-16,076.0	58.61 (0.50)	52.08 (0.71)
Latent-class model (1 State)	-17,849.2	35.47 (0.49)	30.78 (0.66)
Latent-class model (2 States)	-17,673.9	44.29 (0.51)	40.21 (0.70)
Latent-class model (3 States)	-17,722.3	45.29 (0.51)	36.14 (0.68)
Independent	-19,086.4	33.23 (0.48)	30.35 (0.66)
Only intercept	-19,335.9	29.37 (0.47)	23.12 (0.60)
VAR with intercept	-13,768.4	71.13 (0.46)	64.38 (0.68)

Notes. The zero- and first-order Markov models directly model the observed process and do not include VAR or hidden-Markov processes to govern state transitions as with the other models. The standard errors of the hit rates are provided in parentheses below the estimates.

single state for each user, then the s subscript in §3 can be interpreted as an index of the class of the mixing distribution. Hence, the traditional latent-class model in a multinomial probit framework occurs when the state transition process is characterized by a zero-order Markov model, as defined in §3.2, and when each user is restricted to a single state for all sessions. Our Bayesian framework yields estimates similar to the MLE of the traditional latent-class model when no individual-level covariates are used (\mathbf{R} is a vector of ones) and the covariance of the hyperdistributions are small ($\Psi \rightarrow \mathbf{0}$, $\Omega \rightarrow \mathbf{0}$), although for consistency's sake we use the same prior settings as our other multinomial probit models. Our latent-class model thus allows heterogeneity at a user level similar to the work of Allenby et al. (1998), but the heterogeneity is shrunk toward the aggregate parameter vectors for the user's assigned state. We estimate latent-class models with one, two, and three states, and find the data favor the two-state model using both the log marginal density and out-of-sample hit rate. The predictive accuracy falls between the zero- and first-order Markov model, suggesting that the covariates are not able to fully capture persistence in latent utility.

Multinomial Probit Model (Independent, Intercepts Only, and VAR). We estimate an independent probit model with the page and session covariates ($S = 1$, $\Sigma = I$) to judge the contribution of the correlated error structure (labeled as the Independent model in Table 6). This model is estimated for a single state without VAR effects ($S = 1$, $\Phi_{is} = \mathbf{0}$). Additionally, we estimate our multinomial probit model with only intercepts and no covariates

$$(S = 1, \mathbf{X}_{iqt} = [1], \Phi_{is} = \mathbf{0}),$$

to assess the effect of the covariates (labeled as Only Intercept in Table 6). These two models benchmark the performance of the popular multinomial probit model without memory. Finally, to help judge the contribution of our covariates in our full model we estimate a discretized VAR model without covariates (labeled as VAR with intercept). This is identical to our full model without the page and session covariates but with only a single state ($S = 1$, $\Gamma_{is} = \mathbf{0}$). Notice from Table 6 that the correlated intercept-only multinomial probit model does a poor job, since it ignores the contribution of covariates with marketing mix variables, Web page content characteristics, and Web user's demographic variables. The independent multinomial probit model also does poorly because of the unexplained dependency across Web page categories. The VAR with intercept does well against all the alternative models, demonstrating that much of the improvement of our dynamic multinomial probit model comes from its VAR component.

Discussion. Models with memory, such as the dynamic multinomial probit models (from Table 5), the first-order Markov models (from Table 6), and the VAR model (from Table 6) perform an order of magnitude better in predicting than comparable memory-less models, such as the independent, only-intercept, and latent-class models. This clearly demonstrates the importance of memory in predicting paths. This is an important finding because it shows that not only is the frequency of viewing different content important, but there is also a good deal of information contained within the sequence of viewings. This affirms the central thesis of this paper, that path analysis is informative. Additionally, the VAR model by itself outperforms the first-order Markov model, which shows that while a first-order Markov model may be a good first-order approximation, browsing behavior is better represented by a richer memory model.

4.3. Parameter Estimates

We focus our discussion of parameter estimates on our best model, which in terms of posterior odds and out-of-sample prediction is a two-state dynamic, multinomial probit model with transitions at the page level. This model is strongly favored using both the posterior odds and the out-of-sample hit rate. First, we consider the parameters that govern our state transitions. We label the first as a browsing-oriented state and the second as a deliberation-oriented state, following our identification conditions. While purchases may occur in either state, they are more likely to occur in the deliberation-oriented state. Sixty four percent of users start in a browsing-oriented state. Users tend to stay in a browsing-oriented state for about three viewings, while they tend to stay in a deliberation-oriented state for about four viewings.

Second, we consider the relationship between our covariates and the selection of the home category. The large number of parameters makes it impractical to discuss all parameters in this text (we refer the reader to the Technical Appendix D at <http://mktsci.pubs.informs.org> for a full report). However, to illustrate some of the findings from our model we consider the parameters associated with the home category.

The home category is a common entry point for a website. Intuitively its use signals the beginning of a session (16% of sessions begin at this page) or even within a session the beginning of a new goal (for example, a previous path was terminated because the user could not find the right book and is starting over at the home page). Our results show that the home page is more likely to be viewed in the browsing state than the deliberation state, as indicated by a higher intercept value. The effect of each variable depends

quite a bit upon the state of the user. Users who are browsing are more likely to visit a home page if they have previously purchased, are viewing the page during the weekend, or have visited another site, while users who are deliberating are much less likely to visit the home page as a result of these effects.

The presence of price information and advertisements tends to lessen the chance of viewing the home page for both types of viewers. Users in a browsing-oriented state are more likely to visit the home page if they have bought in a previous session, have a long session, visit during the weekend, or visit other sites during their B&N session. The differences between the states illustrate the importance of allowing heterogeneity both across and within users.

Demographic characteristics are also predictive of browsing tendencies. Nested within the hyperdistribution is a linear model that relates the demographic characteristics of the user to parameters in Γ_{is} . Again, there are a large number of effects, and for illustration purposes we consider only the effect of the demographics upon the Price Present coefficient for selecting the home page. We find that higher-income males with children are more likely to use the home page when they are in the browsing state. In contrast, the demographics of users in a deliberation-oriented state are much less helpful in prediction. It is possible that gender (Meyers-Levy 1989), age (Bettman et al. 1998), and education (Crosby and Taylor 1981) are indicative of cognitive strategies, but demographic variables can be correlated with many other characteristics that were not measured in this study.

4.4. Capturing Memory

At the heart of path analysis is the ability to use summaries of past movements or memory to predict future movements. The VAR process was important in improving the predictive ability of the model. For instance, the out-of-sample hit rate jumped from 23% to 64% when a VAR process was included, compared with a model with only intercepts. Similarly, a first-order Markov model has a 52% out-of-sample hit rate compared with a 13% out-of-sample hit rate for a zero-order Markov model. Clearly, memory plays a crucial role in the predictive ability of models for clickstream data. The memory effects in our model are captured by the first-order Markov model that governs state transitions, the VAR parameters, and the time-varying covariates. Instead of discussing the parameter estimates further (which are reported in our Technical Appendix D at <http://mktsci.pubs.informs.org>), we focus on illustrating the dynamic performance of our model.

The fit and hit rate provided in Tables 5 and 6 are measures of one-step-ahead forecasts. However, we are not interested in simply forecasting a single-step

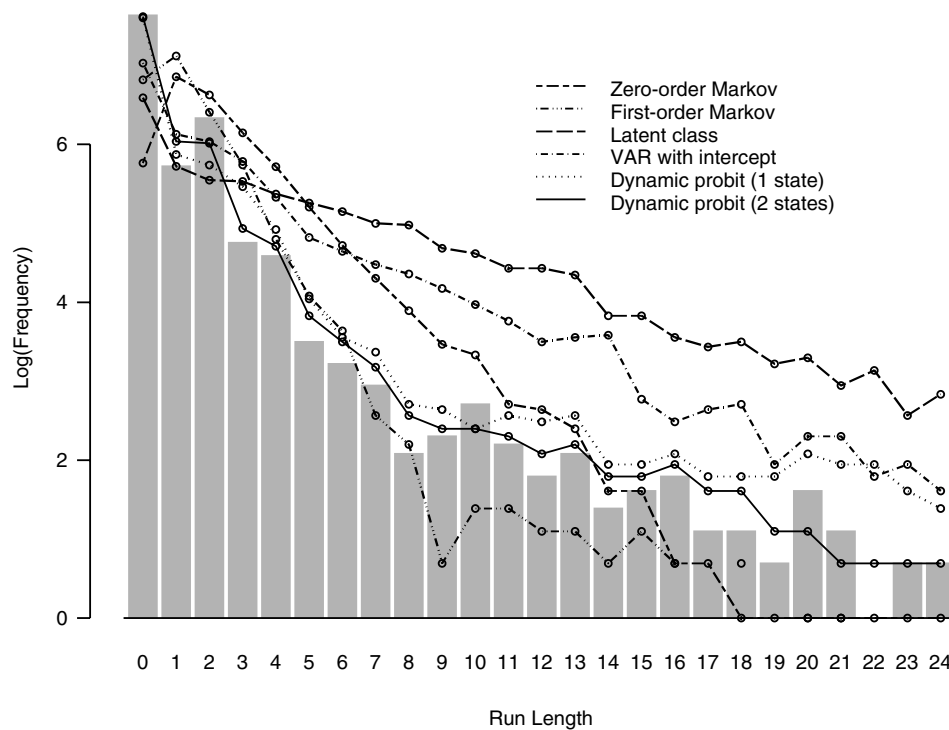
ahead—we are potentially interested in predicting the entire path that a user may take. One way to measure the multistep-ahead accuracy of our model is the ability to predict the run length of a path, where we define a run length as the number of intervening viewings between two events of interest, say two category viewings. For example, the run length of CC, C?C, and C??C would be 0, 1, and 2, respectively (where ? represents any category other than exit). Figure 1 illustrates the frequency distribution of run lengths for our actual data (using the estimation sample) as well as the predicted run lengths for various models. Notice that all the models, except for the dynamic multinomial probit models, substantially underpredict the count of runs with zero length. Also, the zero-order Markov model tends to underpredict the length of the remaining runs, while the first-order Markov and latent-class models tend to overpredict. Only the two-state dynamic probit model does a good job of capturing the entire distribution. We present another multistep ahead forecasting comparison in the Technical Appendix E at <http://mktsci.pubs.informs.org> which yields similar findings.

5. Predicting Purchase Conversion

Purchase conversion refers to the percentage of Web visitors who make a purchase during a visit to an online retailer. It is a key metric of the success of an e-commerce site because it provides a measure of how many visitors become customers. Despite the rapid growth of e-commerce, online purchase conversion rates have remained low. Online retailers such as Amazon.com, Macys.com, JCPenney.com, and MarthaStewart.com have purchase conversion rates that range between 1% to 2% (Tedeschi 2000). E-commerce managers are interested in understanding what influences purchase conversion and how to improve their conversion rates by dynamically adapting to customers' preferences (Kemp 2001).

In this section we consider how our model can be used to predict the purchase conversion as a user browses through B&N. We wish to forecast the probability that, given a sequence of pages viewed, the user will purchase during his session. For example, if the user has visited the category (C) and shopping cart (S) pages, we wish to know the probability the user will order (O) on the next page, or have the sequence CSO. Additionally, we need to consider the chance that the user will order two viewings out (i.e., CS?O, where ? stands for any page other than exit, since exit terminates a session), three viewings out (i.e., CS??O), or any path that will lead to a purchase during this session (i.e., CS*O*E, where * stands for any sequence of pages that does not include exit or order). Notice that

Figure 1 Frequency Distribution of Run Lengths Between Category Viewings



Note. For example, run length of CC, C?C, and C??C, is 0, 1, and 2, respectively. The vertical axis is given in logged units to better illustrate the dispersion among the models.

we are interested only in forecasting orders before the end of the session; otherwise, we would be forecasting the probability of ordering in the current session or in any future session, and not just the current session. Additionally, we note that although we focus on purchase conversion the same technique could be used to forecast any metric of interest, such as the probability the user will return to the home page, exit the website within five viewings, and so on.

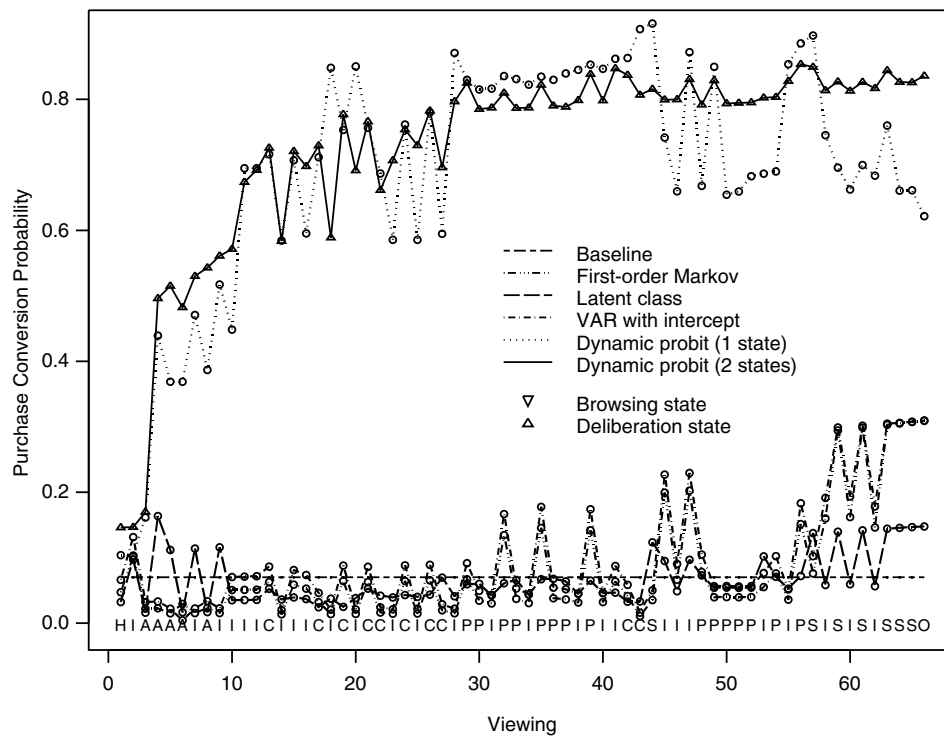
To construct these forecasts we use a simulation method. For each sweep of our MCMC estimation algorithm we simulate the latent category utilities starting with the specified forecasting origin and continue until the session is predicted to end (i.e., until the E category is encountered). Next, we calculate the purchase conversion probability as the percentage of sequences that include an order (O). The individual-level waiting time and hidden states are generated from each user's corresponding estimates of her hidden Markov chain. Because the covariates of these simulated pages are not known, we use the expected value for the corresponding category (see Table 2; i.e., expected time to the next viewing is 7.2 seconds).

To illustrate these purchase conversion forecasts we consider the session produced by User 6 as described in Table 3. We plot the predicted conversion probabilities as a function of the amount of the path that has been observed for various models in Figure 2.

The category abbreviations of each viewing are given along the horizontal axis. For comparison, we plot the baseline conversion rate of a 7% probability that a visitor will make a purchase during a session. This user starts at a home page but then immediately moves to a series of viewings at information and account pages. We predict that after five viewings there is a better-than-even chance of this user making a purchase. Notice that while the user's purchase probability continues to rise to around 80%, the rate of increase slows down significantly after 30 viewings. Our model predicts that this user is in a deliberation-oriented state throughout this session.

This graph also helps illustrate why the two-state model is better than a one-state model. The two-state model is better able to capture the notion that some viewings of a user in a deliberation-oriented state, such as viewing a category or information page, are not shifting the user from their overall purchase goal, but instead simply supporting the user's goal. In contrast, the one-state model is not able to make this distinction and results in predictions that are more susceptible to apparent browsing-oriented behavior. The other models do a much poorer job of tracking conversion; even the intercept-only VAR model still reaches only about a 35% prediction of purchase by the final observation.

While this graph illustrates how our model could be applied on a viewing-by-viewing basis to fore-

Figure 2 Probability of Purchase at Some Time During the Remainder of the User's Session

cast purchase conversion, it summarizes the sequential predictive performance for only a single session. To assess the ability to predict a user's conversion probabilities in a more systematic way, we compute the probability that a user will purchase sometime during his session based upon his initial viewings, and report these results in Table 7. We split our sample into those sessions where a purchase occurred and those with no purchase; the cells of Table 7 report the predicted purchase conversion rate. For those who purchase this is equivalent to the proportion of visitors who were correctly classified, while for nonpur-

chasers it is equal to the proportion of visitors who were incorrectly classified. (See Technical Appendix F at <http://mktsci.pubs.informs.org> for the forecasting performance on predicting purchase conversion of other selected models.)

Ideally, a perfect model would yield 100% probability of purchase for purchasers and 0% for nonpurchasers. Recall that the average purchase conversion rate is 7%. Notice that based on only knowing the initial viewing we are able to predict purchases with 13% accuracy, while nonpurchasers are close to the baseline at 6%, which means the odds ratio of differ-

Table 7 Predicted Purchase Conversion Probabilities (and Standard Errors in Parentheses) Given Initial Paths for the Two-State Dynamic Multinomial Probit Model

Sample	Session type	Number of sessions	Forecast origin/number of viewings during session					
			1	2	3	4	5	6
Estimation	Purchase	83	13.3% (0.48)	16.3% (0.52)	23.4% (0.60)	30.9% (0.65)	34.4% (0.67)	41.5% (0.69)
	No purchase	1,129	6.1% (0.33)	5.4% (0.32)	4.6% (0.30)	3.7% (0.27)	3.4% (0.26)	3.1% (0.25)
	All	1,212	6.6% (0.35)	6.1% (0.34)	5.9% (0.33)	5.6% (0.33)	5.5% (0.32)	5.7% (0.33)
Holdout	Purchase	31	10.4% (0.97)	12.8% (1.06)	15.2% (1.14)	18.0% (1.21)	19.1% (1.24)	21.2% (1.29)
	No purchase	416	6.9% (0.80)	5.5% (0.72)	5.1% (0.70)	4.2% (0.63)	3.5% (0.58)	3.2% (0.56)
	All	447	7.2% (0.82)	5.9% (0.75)	5.8% (0.74)	5.1% (0.70)	4.6% (0.66)	4.4% (0.65)

entiating purchasers from nonpurchasers is 2. As we observe the user's path we are better able to distinguish purchasers from nonpurchasers, with the probability of correctly predicting a purchaser after six viewings rising to 42%; the odds ratio rises to 14 after six viewings, which shows that path information can be quite valuable in differentiating purchasers from nonpurchasers, even with a limited amount of path information.

The forecast probabilities for the estimation and holdout samples are reported separately in Table 7. The parameter estimates of the estimation sample use all information when estimating the user's parameters (such as inferring state), while those in the holdout sample use only past information to estimate the parameters. However, in either case we do not use any future information when forecasting but only condition on past values up to the forecasting origin. Notice that there is a magnitude of decline in out of sample predictions, going from 41.5% to 21.2%. However, our odds ratio of differentiating purchasers from nonpurchasers is still good, about 7.

In summary, our analysis shows that path analysis can be quite helpful in predicting purchase conversion, even early in a session. The next step would be to consider how managers could use these predictions to improve their conversion rates and profitability by dynamically customizing the website. While a full customization is beyond the scope of this paper, it is helpful to understand how the predictive ability of the model would translate into decision making. Suppose B&N were to classify each user after her 5th viewing as either deliberation oriented or browsing oriented, and then—based upon this classification—customize the subsequently requested pages with the objective of encouraging purchase conversion. Specifically, we make the following changes for users who are browsing oriented: Delete price information (if any), add promotion image (if there is none), delete banner ads, reduce the number of links to home pages by half, and double the number of links to product, account, and information pages. These choices were made by examining the expected response to these covariates (see §4.3 and Technical Appendix E at <http://mktsci.pubs.informs.org>). For deliberation-oriented users B&N made the opposite customizations: Add price information, delete promotion image, delete banner ads, double the number of links to a home and product page, and reduce the number of account and information page links by half.

Applying these rules to our holdout sample we find that the conversion rate would increase an additional 2.46% (0.22) and 3.36% (0.26) for the browsing- and deliberation-oriented users, respectively. (Standard errors are given in parentheses.) In summary, conversion rates would increase from

around 7% to more than 9%. Given that the gross profit margin for B&N is around 25% and currently their operating profit margin is negative, these gains in conversion rate would substantively impact B&N's profitability. We also point out that reversing these rules (i.e., applying the customizations for browsing-oriented users to deliberation-oriented users and vice versa) could substantially reduce purchase conversion. Specifically, purchase conversion rates would drop by -4.25% (0.29) and -4.01% (0.28) for the browsing-oriented and deliberation-oriented users, respectively. Hence, overall Web design changes may not improve conversion.

Obviously, these rules could and should be customized at an individual level. Our purpose in following this more simplistic approach is to avoid potential computational problems in optimizing designs. Additionally, optimizing the Web design may not be straightforward because users expect some consistency in the user interface of a website, and abrupt changes may hurt a user's navigation ability (e.g., orphaned pages with no links to the home). Finally, one could consider jointly optimizing marketing mix policies (price, promotion, and assortment) at the same time as optimizing the Web design (see Zhang and Krishnamurthi 2004 for a discussion of promotional customization in an online store). These issues represent limitations of our research; we hope to address these limitations in future research. These results suggest that a fully dynamic, customized approach to Web design could be very profitable.

6. Conclusions

Our primary purpose has been to show that the sequence of Web viewings is informative in predicting a user's path. In our dataset we found that models that incorporated sequence or path information doubled the hit rates over those that did not. Additionally, we have shown that our model has reasonable predictive power with regard to understanding which users are likely to make a purchase and which are not. We can predict those users who are likely to purchase with 42% accuracy with as few as six viewings. This contrasts with a baseline prediction of a 7% conversion rate. Furthermore, dynamically changing Web design could increase conversion rate from 7% to beyond 9%, which indicates personalization would be quite profitable.

Certainly there are many aspects of Web design, advertising, and promotion that are important elements of this problem. Path analysis does not negate or supplant the need for studying these other aspects of the problem. Our narrow focus on path analysis is simply to understand the contribution of sequencing information. Our study also has many limitations.

First, we have studied paths at only one online retailer for a 1-month period. The usual caveat holds, that these results may not be representative of other retailers and other time periods. Second, we have only focused on the path taken through a website. Bucklin and Sismeiro's (2003) results suggest that incorporating timing information about viewings could also be helpful. We leave the integration of timing information for future research. Third, we have abstracted the website by categorizing pages, which has resulted in the loss of graphical and textual information concerning the page content. Additionally, other categorization schemes could alter the predictive accuracy of our model.

Finally, our approach is largely statistical in nature, although it has been motivated by behavioral research. We speculate that underlying the improved predictive ability of purchase behavior is the fact that navigation paths reflect a user's goals. Goals are defined as cognitive states that people desire to attain (Lewin 1943) and have been shown to drive consumers' search and choice behavior (Johnson and Payne 1985, Bettman et al. 1998, Shafir et al. 1993, Svenson 1996). Incorporating structural models of consumer behavior (Payne et al. 1993) could result in even better use of pathing information. Recently, cognitive psychologists have applied ecological models of food-gathering behavior in the context of information search (Pirolli and Card 1999). Heer and Chi (2001, 2002a,b) have used the idea of a hunter following a scent in modeling Web-browsing activity. We would hope that future research could better incorporate behavioral models of consumer search and goals into structural models of Web navigation beyond our reduced-form approach.

We also believe that path analysis has implications outside of online shopping. For example, Underhill (1999, p. 99) found that shoppers whose path through the store includes a visit to a dressing room were more likely to purchase. While collecting path information in traditional brick and mortar stores with human observers is prohibitively expensive and intrusive, there are some technologies that may make the collection and analysis of path information economically viable. For example, IRI's VideOcart installed a radio-tracking device to collect path information and interact with a shopper (*Marketing News* 1988). Unfortunately, researchers were not able to make use of the tracking information (Shulman 1993). The advent of GPS-enabled devices such as cell phones and mobile computers seem to raise these issues anew. Hence we believe the use of path analysis is not limited to online environments.

Research on clickstream analysis, both in academics and business, is only beginning (Bucklin et al. 2002). To the best of our knowledge our study is the first within the marketing literature to apply path analy-

sis at the viewing-by-viewing level to the problem of purchase conversion. We hope it will generate more widespread interest in path analysis and the analysis of clickstream data. Our interest in this problem is not because e-commerce is new, but because path data appear to be powerful sources of information with which to infer consumer goals and predict behavior. Economists have not had access to such detailed information search information, which provides both theoretical and empirical challenges for future research.

Acknowledgments

The authors thank Comscore Media Metrix for their generous contribution of data without which this research would not have been possible. Additionally, they thank Brett Gordon for his help with perl scripting and Randy Bucklin, Ron Goettler, and Ajay Kalra for their comments.

References

- Alba, Joseph, John Lynch, Barton Weitz, Chris Janiszewski, Richard Lutz, Alan Sawyer, Stacy Wood. 1997. Interactive home shopping: Consumer, retailer, and manufacturer incentives to participate in electronic markets. *J. Marketing* 61(3) 38–53.
- Allenby, Greg M., Neeraj Arora, James L. Ginter. 1998. On the heterogeneity of demand. *J. Marketing Res.* 35 384–389.
- Ansari, Asim, Carl F. Mela. 2003. E-customization. *J. Marketing Res.* 40(2) 131–145.
- Bettman, James R., Mary F. Luce, John W. Payne. 1998. Constructive consumer choice process. *J. Consumer Res.* 25 187–217.
- Blattberg, R., J. Deighton. 1991. Interactive marketing: Exploiting the age of addressability. *Sloan Management Rev.* 33(1) 5–14.
- Bucklin, Randolph E., Catarina Sismeiro. 2003. A model of Web site browsing behavior estimated on clickstream data. *J. Marketing Res.* 40(3) 249–267.
- Bucklin, Randolph E., James M. Lattin, Asim Ansari, Sunil Gupta, David Bell, Eloise Coupey, John D. C. Little, Carl Mela, Alan Montgomery, Joel Steckel. 2002. Choice and the Internet: From clickstream to research stream. *Marketing Lett.* 13(3) 245–258.
- Cadez, Igor, David Heckerman, Christopher Meek, Padhraic Smyth. 2000. Visualization of navigation patterns on a Web site using model based clustering. *Proc. Sixth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, Boston, MA, 280–284.
- Coffey, Steven. 1999. Media metrix methodology. Working paper, Media Metrix <http://www.mediametrix.com/Methodology/Convergence.html>.
- Crosby, L. A., J. R. Taylor. 1981. Effects of consumer information and education on cognition and choice. *J. Consumer Res.* 8(1) 43–56.
- Haaijer, Rinus, Michel Wedel. 2001. Habit persistence in time series models of discrete choice. *Marketing Lett.* 12(1) 25–35.
- Hamilton, James D. 1994. *Time Series Analysis*. Princeton University Press, Princeton, NJ.
- Hamilton, James D. 1989. A new approach to the economic-analysis of nonstationary time-series and the business-cycle. *Econometrica* 57(2) 357–384.
- Heer, J., E. H. Chi. 2001. Identification of Web user traffic composition using multi-model clustering and information scent. *Proc. Workshop Web Mining. SIAM Conf. Data Mining*, Chicago, IL, 51–58.
- Heer, J., E. H. Chi. 2002a. Mining the structure of user activity using cluster stability. *SIAM Internat. Conf. Data Mining, Workshop on Web Analytics*. Arlington, VA.

- Heer, J., E. H. Chi. 2002b. Separating the swarm: Categorization methods for user access sessions on the Web. *Proc. Human Factors Comput. Systems Conf.*, Minneapolis, MN.
- Hoffman, Donna L., Thomas P. Novak. 1996. Marketing in hypermedia computer-mediated environments: Conceptual foundations. *J. Marketing* **60**(3) 50–68.
- Internet Week. 2001. Macy's doubles conversion rate—Overhauled search, content and customer service help online unit turn site visitors into buyers. Ted Kemp (November 28). <http://www.internetweek.com>.
- Janiszewski, Chris. 1998. The influence of display characteristics on visual exploratory search behavior. *J. Consumer Res.* **25**(3) 290–301.
- Johnson, Eric J., John W. Payne. 1985. Effort and accuracy in choice. *Management Sci.* **31** 394–414.
- Johnson, Norman L., Samuel Kotz, B. Balakrishnan. 1994. *Continuous Univariate Distributions*, Vol. 1, ed. 2. John Wiley & Sons, New York.
- Kamakura, Wagner A., Gary Russell. 1989. A probabilistic choice model for market-segmentation and elasticity structure. *J. Marketing Res.* **26**(4) 379–390.
- Kass, Robert E., Adrian E. Raftery. 1995. Bayes factors. *J. Amer. Statist. Assoc.* **90**(430) 773–795.
- Kedem, Benjamin. 1980a. *Binary Time Series*. Marcel Dekker, New York.
- Kedem, Benjamin. 1980b. Estimation of the parameters in stationary autoregressive processes after hard limiting. *J. Amer. Statist. Assoc.* **75**(369) 146–153.
- Keenan, D. M. 1982. A time-series analysis of binary data. *J. Amer. Statist. Assoc.* **77**(380) 816–821.
- Krolzig, H.-M. 1997. *Markov-Switching Vector Autoregressions, Modelling, Statistical Inference and Application to Business Cycle Analysis. Lecture Notes in Economics and Mathematical Systems*, Vol. 454. Springer, Berlin, Germany.
- Lewin, Kurt. 1943. Defining the field at a given time. *Psych. Rev.* **50** 292–310.
- Marketing News. 1988. VideoCart shopping cart with computer screen creates new ad. Chicago, IL, May 9, Vol. 22, No. 10, 1–2.
- Mena, Jesus. 2001. *WebMining for profit: E-business optimization*. Elsevier Digital Press.
- Meyers-Levy, Joan. 1989. The influence of sex roles on judgement. *J. Consumer Res.* **14**(March) 522–530.
- Moe, Wendy W. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *J. Consumer Psych.* **13**(1&2) 29–40.
- Moe, Wendy W., Peter S. Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Management Sci.* **50**(3) 326–335.
- Moe, Wendy W., Hugh Chipman, Edward I. George, Robert E. McCulloch. 2004. A Bayesian treed model of online purchasing behavior using in-store navigational clickstream. Working paper, Robert H. Smith School of Business, University of Maryland, College Park, MD.
- Montgomery, Alan L. 2001. Applying quantitative marketing techniques to the Internet. *Interfaces* **30**(2) 90–108.
- Newton, Michael A., Adrian E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B (Methodological)* **56**(1) 3–48.
- Tedeschi, Bob. 2000. Easier-to-use sites would help e-tailers close more sales. *New York Times* (June 12) C14.
- Paap, Richard, Philip Hans Franses. 2000. A dynamic multinomial probit model for brand choice with different long-run and short-run effects of marketing-mix variables. *J. Appl. Econometrics* **15** 717–744.
- Pal, Nirmal, Arvind Rangaswamy. 2003. *The Power of One: Leveraging the Potential of Personalization Technologies*. Trafford Publishing, Victoria, British Columbia, Canada.
- Park, Young-Hoon, Peter S. Fader. 2004. Modeling browsing behavior at multiple Web sites. *Marketing Sci.* **23**(2) 280–303.
- Payne, John W., James R. Bettman, Eric J. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge University Press, New York.
- Pirolli, P., S. K. Card. 1999. Information foraging. *Psych. Rev.* **106**(4) 643–675.
- Pitkow, James. 1997. In search of reliable usage data on the WWW. *Comput. Networks IDN Systems* **29**(8-13) 1343–1355.
- Redish, Janice. 2002. Information-rich Web sites: Challenges and opportunities. www.redish.net/cmu.pdf. Retrieved October 2002.
- Rossi, Peter E., Robert E. McCulloch, Greg M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Sci.* **15**(4) 321–340.
- Rossi, Peter E., Robert E. McCulloch. 2000. Bayesian analysis of the multinomial probit model. Roberto Mariano, Til Schuermann, Melvyn J. Weeks, eds. *Simulation-Based Inference in Econometrics*. Cambridge University Press, New York, 158–178.
- Seetharaman, Seethu. 2004. Modeling multiple sources of state dependence in random utility models: A distributed lag approach. *Marketing Sci.* **23**(2) 263–271.
- Shafir, Eldar B., Itamar Simonson, Amos Tversky. 1993. Reason-based choice. *Cognition* **49** 11–36.
- Shapiro, Carl, Hal R. Varian. 1998. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Cambridge, MA.
- Shulman, Richard. 1993. New systems, old practices create a POS “generation gap.” *Supermarket Bus.* **48**(11) 17.
- Sismeiro, Catarina, Randolph E. Bucklin. 2003. Modeling purchase behavior at an e-commerce website: A conditional probability approach. *J. Marketing Res.* **41**(3) 306–323.
- Svenson, Ola. 1996. Decision making and the search for fundamental psychological regularities: What can be learned from a process perspective? *Organ. Behavior Human Decision Processes* **65**(3) 252–267.
- Teräsvirta, T. 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. *J. Amer. Statist. Assoc.* **89** 208–218.
- Underhill, Paco. 1999. *Why We Buy: The Science of Shopping*. Simon & Schuster, New York.
- Zhang, Jie, Lakshman Krishnamurthi. 2004. Customizing promotions in online stores. *Marketing Sci.* **23**(4) 561–578.