Print Date: 2/24/2004

Modeling Online Browsing and Path Analysis Using Clickstream Data

Alan L. Montgomery, Shibo Li, Kannan Srinivasan, and John C. Liechty

November 2002
First Revision, September 2003
Second Revision, February 2004
Third Revision, February 2004

Alan L. Montgomery (e-mail: alan.montgomery@cmu.edu) is an Associate Professor at Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. Shibo Li (shibo_li@rbsmail.rutgers.edu) is an Assistant Professor of Marketing at Rutgers University, 228 Janice Levin Building, 94 Rockafeller Road, Piscataway, NJ 08854. Kannan Srinivasan (kannans@andrew.cmu.edu) is H.J. Heinz II Professor of Management, Marketing, and Information Systems and Director of the Center for E-Business Innovation at the Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. John C. Liechty (jcl12@psu.edu) is an Assistant Professor of Marketing and Statistics at the Pennsylvania State University, 710 M Business Administration Building, University Park, PA 16802. The corresponding author is Alan L. Montgomery. The authors wish to thank Comscore Media Metrix for their generous contribution of data without which this research would not have been possible. Additionally, we would like to thank Brett Gordon for his help with perl scripting, and Randy Bucklin, Ron Goettler, and Ajay Kalra for their comments.

Modeling Online Browsing and Path Analysis Using Clickstream Data

Abstract:

Clickstream data provides information about the sequence of pages or the path viewed by users as they navigate a web site. 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 web site. 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.

Keywords: Personalization, Multinomial Probit Model, Hierarchical Bayes Models, Hidden Markov Chain Models, Vector Autoregressive Models

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 records the navigation path that a user takes through the web site (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, MuCulloch, and Allenby 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 they browse through a web site.

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 web site, 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 which takes them 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 backing up and reviewing pages, until finally viewing 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 had no specific goal and is unlikely to

purchase, but was 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 web site. For example, which user is more likely to view another product page or leave the web site entirely within the next five clicks? Potentially this model could be used for web site design or setting marketing mix variables. For example knowing that a user is less likely to purchase the site could dynamically change the design of the site by adding links to helpful pages, while for those users likely to purchase the site could become more streamlined. 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 upon web browsing behavior using session level data. This aggregate data is 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. Since web sites must interact with users dynamically this sequencing data is 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, since we are modeling page-level movements through a web site and not collections of pages that correspond to tasks. This requires our model to be much more flexible since the paths we observe do not have nice, sequential properties as does Sismeiro and Bucklin.

We also contrast our work with that of Ansari and Mela (2003), who consider the personalization of e-mail messages—but whose work could potentially be applied in a clickstream environment. Again the basic difference is the type of data that we consider which dictates many modeling differences. Their data is derived from user clicks on hyperlinks to personalized e-mails. 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 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, how it is collected, and conduct an exploratory data analysis to motivate the model we introduce in §3. Our data is 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 computers users and tracks their usage at home (Coffey 1999). These panelists 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 their browser window. Since it 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., does the user select a hyperlink, a bookmark, or directly type in the URL to navigate to a page). Nor does the meter record the content of the page, only the URL.

2.1. Descriptive Analysis and Defining the Path

Our dataset consists of 1,160 users who visited barnesandnoble.com (or also 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, or 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 to 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 it is available to the authors; also it is more complete and has 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 their 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 their 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 in length from 2 to 239, with the median of 5 viewings. Most users have only one or two sessions that included activity at B&N; fewer than 25% of our users have 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. 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, slightly more than half are female, most are white, have some college education, and 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 this data from online vendors.

Variable	Mean	Std Dev	Min	Median	Max
Age	45.89	14.62	9	46	89
Age^2 (square of Age)	2326.48	1331.68	81	2209	7921
Male	.47	.50	0	0	1
White	.77	.42	0	1	1
Children under 18 in the house	.40	.49	0	0	1
Married	.29	.45	0	0	1
Some college education	.82	.39	0	1	1
High Income (>$50,000)	.32	.47	0	0	1
Medium Income ($25,000-$50,000)	.35	.48	0	0	1

Table 1. Demographic characteristics of 1,160 panelists, all of the means are proportions except age and age^2 which are continuous variates.

Potentially the clickstream is a very rich data source since 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 this 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 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 upon 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 to review order status. 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 since this data is 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 like 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. Finally, we have three measures of the cumulative number of pages viewed up to that point during the session: pages viewed at B&N, other sites, and other bookstores. Again B&N may not have access to these measures of external activity, but we include them to understand how helpful this data could be in predicting paths. Descriptive statistics for these variables are given in Table 2.

Variable	Mean	StdDev	Min	Med	Max
Presence of price information on page (Proportion)	.45	.50	0	0	1
Promotional image present (Proportion)	.83	.37	0	1	1
Presence of banner advertisement (Proportion)	.03	.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	.03	.18	0	0	1
Time Since Last Viewing (Seconds)	7.2	66.3	1	1	1193
Whether the Visit is on Weekend (Proportion)	.28	.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

Table 2. Descriptive Statistics for the 9,180 unique B&N pages requested.

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 them to other web sites. For example, a book publisher may wish to promote their 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, and a1books.com, etc.). 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 since pages may have already been viewed at other sites during the session but preceding the first B&N page viewed.

Group	User	Session
No purchase	1	HCCCCCCCCCPCCPCCCCCCCCCCCCCCCCCCCCCCCCCCCE
	2	IHHE
	3	IE
	4	IHICPPPCE
	5	IHHIIICIIE
Purchase	6	HIAAAAIAIIIICIIICICICCICICCIPPIPPIPPPIPIICCSIIIPPPPPIPIPSISISISSSOIIIIIHE
	7	HCCPPPCCPCCCCCCCCCPSCSCSPCCPCPCCCCCCCSAAAAAAAAAASSOIIIISASCCCE
	8	IIICICPCPPPCPCICICPCCCPCPPPIPSIIAASSSIIISOIIE
	9	IISIASSSSOIE
	10	IPPPPSASSSSSOIAAAHCCPCCCCCCE

Table 3. Listing of category of viewings for selected user sessions. (Types of pages: H=Home; A=Account; C=Category; P=Product; I=Information; S=Shopping Cart; O=Order; E=Exit.)

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 their session at a product page after considering an individual item. To illustrate our data we list the sessions of ten 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; their B&N session started at the home page, and then 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 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 their order status, and continued to the home page before exiting.

		Category of Previous Viewing							
Category of Current Viewing	Category	Home	Account	Category	Product	Inform.	ShopCart	Order	Exit
	Home	.23	.01	.01	.01	.10	.02	0	.16
	Account	.01	.69	.01	.01	.02	.15	0	.01
	Category	.17	.02	.60	.31	.15	.05	0	.16
	Product	.01	0	.20	.43	.10	.05	0	.05
	Information	.25	.06	.08	.12	.46	.15	.87	.61
	Shop. Cart	.01	.16	.01	.03	.02	.45	.13	.01
	Order	0	0	0	0	0	.10	0	0
	Exit	.32	.06	.09	.09	.14	.02	0	0
	Marginal	.06	.05	.32	.17	.23	.05	.01	.11
	Initial Prob.	.16	.02	.16	.06	.60	.01	0	0

Table 4. Sample transition matrix, marginal and initial probabilities for categories of viewings. (Notice that the columns sum to one, and there are a total of 14,512 observations.)

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 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 .43 to .69. The lack of transitions between exit pages is due to our definition of a session, in which we treat consecutive viewings at other web sites as a single exit observation. Finally, note that there is a 23% probability of viewing the home page in the next viewing even when you are already at the home page. Home to home page viewings usually arise when a user works with another program in 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 (back 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 web site may not offer links to all of 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 web site, 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 web site. For example, the many hypertext links from a category page to a product page make it likely the user may use one of these links to navigate the web site. However, navigation is not restricted to these links. Suppose a user requests additional information about shipping which appears as a popup window, the user may continue browsing through the web site but leave the popup window open. The user may re-select 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 they are 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 of these facets. The main problem being 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, McCulloch, and Allenby 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 non-flow 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 total of I users, C categories, Q_i sessions for user i, and T_{iq} viewings for the qth 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 amongst those that are available. We denote the user's selection as Y_{iqtc}, which yields the following observational equation:

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

Where $\mathbf{U}_{iqt} = [U_{iqt1}, U_{iqt2}, \cdots, U_{iqtC}]$ is a $C \times 1$ vector of latent utilities, ι_{iqtc} is an indicator variable that equals one when category c is available for user i during the tth viewing of session q, $\boldsymbol{\iota}_{iqt}$=[

$\iota_{iqt1}, \iota_{iqt2}, \ldots, \iota_{iqtC}$], the operator $\mathbf{U}<\boldsymbol{\iota}>$ denotes the set of elements from the vector $\mathbf{U}$ whose corresponding indicator operand ($\boldsymbol{\iota}$) equal to one. In our model $\boldsymbol{\iota}_{iqt}$ is a vector of ones except 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 unless the shopping cart was previously selected). 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 ($\mathbf{U}_{iqt}$) is modeled as:

$$\mathbf{U}_{iqt} = \boldsymbol{\Gamma}_{i,s<iqt>}\mathbf{X}_{iq,t-1} + \boldsymbol{\Phi}_{i,s<iqt>}\mathbf{U}_{iq,t-1} + \boldsymbol{\varepsilon}_{iqt,s<iqt>}, \qquad \boldsymbol{\varepsilon}_{iqt,s<iqt>} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{s<iqt>}), \tag{2}$$

where subscript $s<iqt>$ defines the user's state (s=1, …, 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 ($<iqt>$) 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), $\boldsymbol{\Gamma}_{is}$ is a $C \times L$ parameter matrix, $\boldsymbol{\Phi}_{is}$ is a $C \times C$ matrix of autoregressive coefficients, and the error vector $\boldsymbol{\varepsilon}_{iqts}$ 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 upon the information being viewed at time t-1 (*e.g.*, the number of links of the page, type of information display, time page viewed, etc.), since 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 since the user only views the covariates of the selected alternative. Hence our current framework relies upon 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). 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_{iqt8}=0), and from

equation (2) and hereafter set C=7 to refer to the remaining categories. Also, for identification we set the first element of the error covariance matrix to unity, $\left[\Sigma_s\right]_{1,1} = 1$.

3.2. Modeling State Transitions

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

$$\Pr\left[s < iqt >= s \middle| \boldsymbol{\nu}_i\right] = \nu_{is}, \text{ where } \boldsymbol{\nu}_i = \begin{bmatrix} \nu_{i1} & \nu_{i2} & \cdots & \nu_{iS} \end{bmatrix}'. \tag{3}$$

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

$$\Pr[w_{iqt} \mid \boldsymbol{\lambda}_i] = \lambda_{i,D_{iqt}} \exp\{-w_{iqt}\lambda_{i,D_{iqt}}\}, \text{ where } \boldsymbol{\lambda}_i = \begin{bmatrix} \lambda_{i1} & \lambda_{i2} & \cdots & \lambda_{iS} \end{bmatrix}'. \tag{4}$$

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:

$$\Pr\left[D_{iq,t+w_{iqt}} = s \middle| D_{iqt} = g, \boldsymbol{\nu}_i, \mathbf{P}_i\right] = P_{igs} \text{ if } t > 1, \text{ where } \mathbf{P}_i = \begin{bmatrix} 0 & P_{i12} & \cdots & P_{i1S} \\ P_{i21} & 0 & \cdots & P_{i2S} \\ \vdots & \vdots & \ddots & \vdots \\ P_{iS1} & P_{iS2} & \cdots & 0 \end{bmatrix}. \tag{5}$$

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 since same state transitions are captured through the waiting time. Finally, the initial state probability for the first viewing of a session is:

$$\Pr[D_{iqt} = s \mid \boldsymbol{\nu}_i] = v_{i,D_{iqt}} \text{ if } t = 1. \tag{6}$$

where we redefine the probability vector $\boldsymbol{\nu}_i$ as the initial starting probabilities.

---

[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 non-computer 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.

In order to identify the states we place a restriction on the means of the latent variable U_{iqtc}. Specifically, we introduce two metrics, W^a_{iqts} and W^b_{iqts}, to capture the user's tendency to browse (e.g., surf or a non-purchase orientation) or deliberate (e.g., focused navigation or a purchase orientation) during a session. W^a_{iqts} is the sum of the expected value of U_{iqtc} for the account, shopping cart, and order pages, while W^b_{iqts} 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^a_{iqts} and highest value of W^b_{iqts}) and state S is the state that exhibits the most deliberation orientation (highest value of W^a_{iqts} and lowest value of W^b_{iqts}). That is, $W^a_{iqtS} \geq W^a_{iqt(S-1)} \geq \cdots \geq W^a_{iqt1}$ and $W^b_{iqt1} \geq W^b_{iqt2} \geq \cdots \geq W^b_{iqtS}$. While we believe these metrics are useful in describing the likely cognitive state of a user, this is only a 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 non-zero 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, step 9). However, these two formulations are equivalent if the transitions occur at integer values. To illustrate this equivalence notice 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):

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 1-\exp\{-\lambda_1\} & \exp\{-\lambda_1\} \\ \exp\{-\lambda_2\} & 1-\exp\{-\lambda_2\} \end{bmatrix}. \tag{7}$$

This equivalence relies upon the "lack of memory" property of the exponential process (cf., Johnson, Kotz, and Balakrishnan 1994, Chapter 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 since the transitions may occur at non-integer values.

3.3. Specification of the Hyper Distributions

Following the usual hierarchical Bayesian framework (Rossi, McCulloch, and Allenby 1996) we incorporate heterogeneity across users by assuming that $\mathbf{\Gamma}_{is}$ has a random coefficient

specification. Specifically, the *l*th column of $\boldsymbol{\Gamma}_{is}$, which we denote as $\boldsymbol{\gamma}_{ils}$, a $C \times 1$ vector, follows a multivariate regression:

$$\boldsymbol{\gamma}_{ils} = \boldsymbol{\Pi}_{ls}\mathbf{R}_i + \boldsymbol{\varsigma}_{ils}, \quad \boldsymbol{\varsigma}_{ils} \sim MVN(\mathbf{0}, \boldsymbol{\Psi}_s), \tag{8}$$

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, since the first element is unity to incorporate an intercept), $\boldsymbol{\Pi}_{ls}$ is a $C \times K$ parameter matrix, and $\boldsymbol{\Psi}_s$ is a $C \times C$ covariance matrix.

We assume that the VAR coefficients are drawn from a hyper-distribution:

$$vec(\boldsymbol{\Phi}_{is}) \sim MVN(vec(\overline{\boldsymbol{\Phi}}_s), \boldsymbol{\Omega}_s), \tag{9}$$

where $\overline{\boldsymbol{\Phi}}_{is}$ is a $C \times C$ matrix of autoregressive coefficients and $\boldsymbol{\Omega}_s$ 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:

$$\mathbf{P}_{ij} \sim D(\boldsymbol{\tau}_j), \quad \nu_{is} \sim D(\boldsymbol{\alpha}_s), \quad \lambda_{is} \sim \Gamma(\hat{\lambda}_{sh}, \breve{\lambda}_{sc}), \tag{10}$$

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

3.4. Discussion of the Model

There are two dynamic elements that we include in our model of browsing behavior. First, we introduce persistent behavior through a vector-autoregressive (VAR) component of lagged latent utility values (Hamilton 1994, Chapter 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 their goals and decide to focus upon 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. 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:

$$\mathrm{E}[\mathbf{U}_{iqt} \mid \mathbf{U}_{iq,t-1}] = \boldsymbol{\Gamma}_{i1}\mathbf{X}_{iq,t-1} + \boldsymbol{\Gamma}_{i1}\mathbf{U}_{iq,t-1} + \frac{\exp\{-\lambda_{1i}\}}{\exp\{-\lambda_{1i}\} + \exp\{-\lambda_{2i}\}}\left\{(\boldsymbol{\Gamma}_{i2} - \boldsymbol{\Gamma}_{i1})\mathbf{X}_{iq,t-1} + (\boldsymbol{\Gamma}_{i2} - \boldsymbol{\Gamma}_{i1})\mathbf{U}_{iq,t-1}\right\}. \quad (11)$$

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 US real GNP. 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 an 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 1980). 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) and the hit rate. The marginal posterior distribution is computed by taking the mean across the Gibbs iterations weighted by the corresponding priors (Newton and Raftery 1994). 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 their data will only be used 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 their 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 their users based upon 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 only using 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 regards to changing the underlying browsing state, since the state may potentially change with every page viewing. For example, a user may begin their 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 that 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, since 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 being either purchase or non-purchase oriented. Secondly, notice 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 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 Raferty (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 45.7. We can also find similar pattern for our session and user-level models, indicating a two-state model is adequate.

State Time	State Process	Number of States	Log Marginal Density	In-Sample Hit Rate (%)	Out-of-Sample Hit Rate (%)
Page	Zero-Order	1	-9378.1	72.05	65.40
				(0.46)	(0.68)
		2	-9016.9	79.44	71.42
				(0.41)	(0.64)
		3	-9064.0	80.34	70.56
				(0.41)	(0.64)
	First-Order	1	-8545.4	83.23	79.95
				(0.38)	(0.57)
		2	-8428.3	89.71	83.15
				(0.31)	(0.53)
		3	-8474.0	89.97	81.14
				(0.31)	(0.56)
Session	Zero-Order	1	-9376.1	73.17	61.56
				(0.45)	(0.69)
		2	-9051.0	77.90	70.48
				(0.42)	(0.65)
		3	-9097.7	78.76	66.14
				(0.42)	(0.67)
	First-Order	1	-8573.5	83.05	73.57
				(0.38)	(0.63)
		2	-8464.9	88.44	81.48
				(0.33)	(0.55)
		3	-8487.0	88.73	78.42
				(0.33)	(0.59)
User	Zero-Order	1	-9411.1	64.38	61.50
				(0.49)	(0.69)
		2	-9124.2	70.04	64.12
				(0.47)	(0.68)
		3	-9193.8	70.85	63.99
				(0.46)	(0.68)

Table 5. Measures of fit for various dynamic probit models. The standard errors of the hit rates are provided in parentheses below the estimate.

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.

Model	Log Marginal Density	In-Sample Hit Rate (%)	Out-of-Sample Hit Rate (%)
Zero-Order Markov Model (1 State)	-20410.4	20.48	12.62
		(0.41)	(0.47)
Zero-Order Markov Model (2 States)	-19458.3	28.18	19.02
		(0.46)	(0.56)
First-Order Markov Model (1 State)	-16444.5	56.06	51.59
		(0.51)	(0.71)
First-Order Markov Model (2 States)	-16076.0	58.61	52.08
		(0.50)	(0.71)
Latent Class Model (1 State)	-17849.2	35.47	30.78
		(0.49)	(0.66)
Latent Class Model (2 States)	-17673.9	44.29	40.21
		(0.51)	(0.70)
Latent Class Model (3 States)	-17722.3	45.29	36.14
		(0.51)	(0.68)
Independent	-19086.4	33.23	30.35
		(0.48)	(0.66)
Only-Intercept	-19335.9	29.37	23.12
		(0.47)	(0.60)
VAR with Intercept	-13768.4	71.13	64.38
		(0.46)	(0.68)

Table 6. Measures of fit for other alternative model specifications. 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 estimate.

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:

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

Notice there is no multinomial probit or vector auto-regression 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:

$$\mathbf{U}_{iqt} = \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_{iqt}, \quad \boldsymbol{\varepsilon}_{iqt} \sim N(\mathbf{0}, \mathbf{I}). \quad (13)$$

Our Bayesian framework can yield results similar to the maximum likelihood estimator (MLE) when no individual level covariates are used (**R** is a vector of ones) and the covariance of the hyper-distribution is small ($\boldsymbol{\Psi} \rightarrow \mathbf{0}$). 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 non-purchasers (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 both zero-order Markov models have the worst fit and predictive ability amongst 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:

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

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 non-purchasers. 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 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 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 (**R** is a vector of ones) and the covariance of the hyper-distributions are small ($\boldsymbol{\Psi} \rightarrow \mathbf{0}$, $\boldsymbol{\Omega} \rightarrow \mathbf{0}$). Although for consistency sake we use the same prior settings as our other multinomial probit models; hence, our latent class model allows heterogeneity at a user level similar to the work of Allenby, Arora, and Ginter (1998), but the heterogeneity is shrunk towards 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 favors 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, $\mathbf{\Sigma}$=I) to judge the contribution of the correlated error structure (labeled as "Independent" model in Table 6). This model is estimated for a single state without VAR effects (S=1, $\mathbf{\Phi}_{is}$=**0**). Additionally, we estimate our multinomial probit model with only intercepts and no covariates (S=1, $\mathbf{X}_{iqt} = [1]$, $\mathbf{\Phi}_{is}$=**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 vector autoregression 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, Γ_{is}=**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 quite well against all the alternative models, demonstrating the 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 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 since it shows that not only is the frequency of viewing different content important, but that 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 state as a browsing-oriented and the second state as a deliberation-oriented, following our identification conditions. While purchases may occur in either state, they are more likely to occur in the deliberation-oriented state. 64% 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 make it impractical to discuss all parameters in this text (we refer the reader to the Technical Appendix D for a full report). However, in order to illustrate some of the findings from our model we consider the parameters associated with the home category.

The home category is common entry point for a web site. 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 since the user couldn't 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 visited another site. While users in a deliberation state 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 hyper-distribution is a linear model that relates the demographic characteristics of the user to parameters in $\boldsymbol{\Gamma}_{is}$. 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, Luce and Payne 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 when compared to a model with only intercepts. Similarly, a first-order Markov model has a 52% out-of-sample hit-rate compared with 13% 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), 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 simply interested in forecasting a single-step ahead, but we are potentially interested in predicting the entire path that a user may take. One way to measure the multi-step 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 under predict 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 tends to over predict. Only the two-state dynamic probit model does a good job of capturing the entire distribution. We present another multi-step ahead forecasting comparison in the Technical Appendix E which yields similar findings.

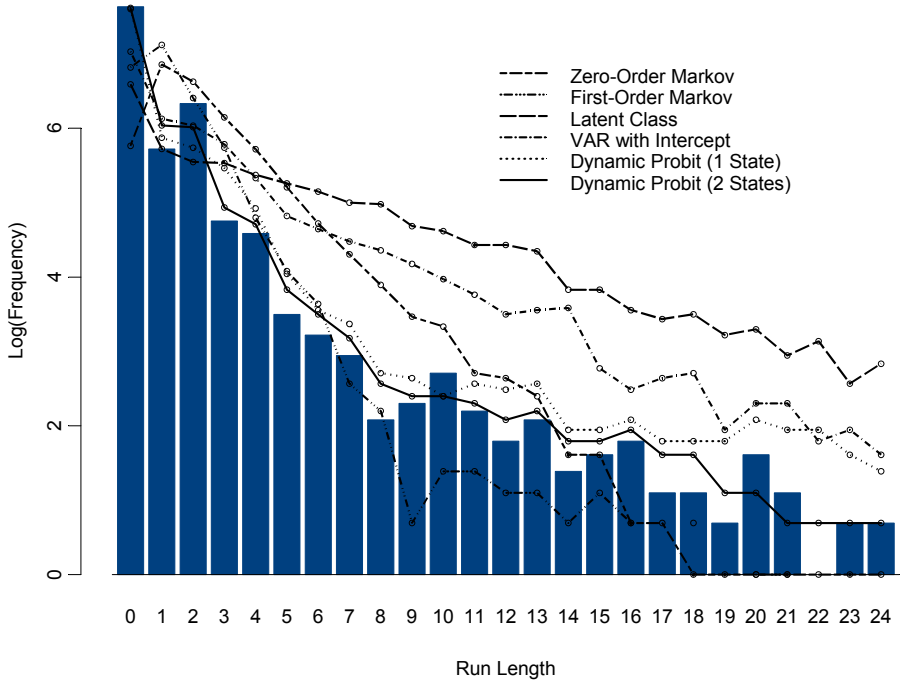

Figure 1. Frequency distribution of run-lengths between category viewings (e.g., run length of CC, C?C, and C??C, is 0, 1, and 2, respectively). The vertical axis is given in logged units to better illustrate the dispersion amongst the models.

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 since it provides a measure of how many visitors are turned into 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 1-2% (*New York Times* 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 (*Internet Week* 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 that the user will purchase sometime 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 do not include exit or order). Notice that we are only interested in forecasting orders before the end of the session, otherwise we would be forecasting the probability of a 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 web site within five viewings, etc.

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 his hidden Markov chain. Since 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. After five viewings we predict that there is better than an 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.

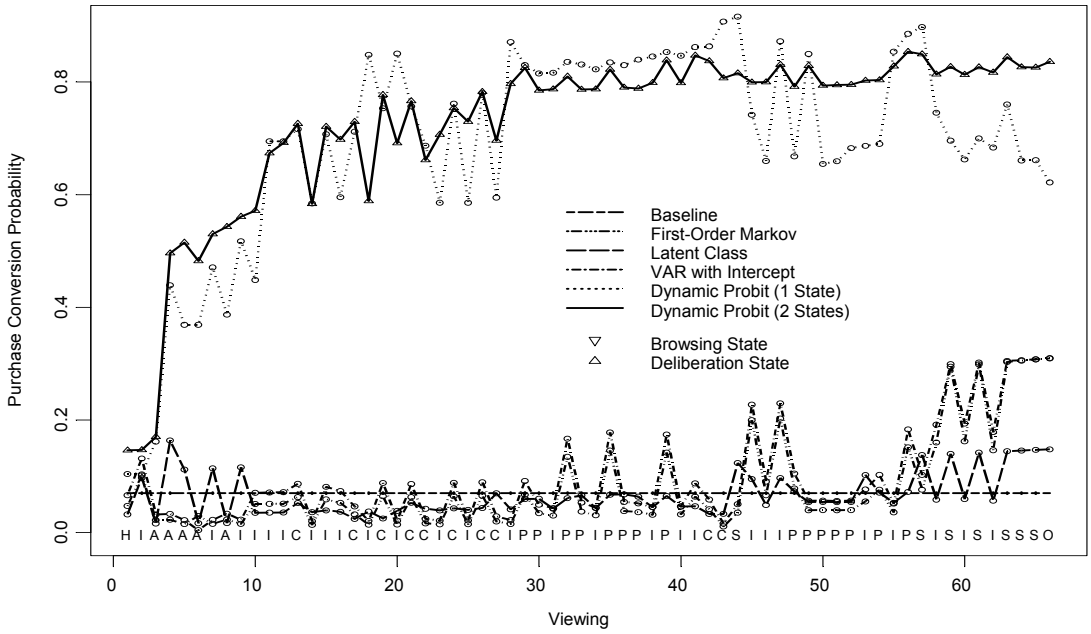

Figure 2. Probability of purchase sometime during the remainder of the user's session.

This graph also helps visually depict 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, like 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 only reaches 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 forecast purchase conversion, it only summarizes the sequential predictive performance for a single session. To assess the ability to predict user's conversion probabilities in a more systematic way, we compute the probability that a user will purchase sometime during their session based upon their 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 and 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 non-purchasers it is equal to the proportion of visitors who were incorrectly classified. (See Technical Appendix F for the forecasting performance on predicting purchase conversion of other selected models.)

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	1129	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	1212	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)

Table 7. Predicted purchase conversion probabilities (and standard errors in parentheses) given initial paths for the two-state dynamic multinomial probit model.

Ideally, a perfect model would yield 100% probability of purchase for purchasers and 0% for non-purchasers. 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 non-purchasers are close to the baseline at 6%, which means the odds-ratio of differentiating purchasers from non-purchasers is 2. As we observe the user's path we are better able to distinguish purchasers from non-purchasers, with the probability of correctly predicting a purchaser after six viewings goes to 42%; the odds-ratio goes to 14 after six viewings, which shows that path information can be quite valuable in differentiating purchasers from non-purchasers 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 non-purchasers 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 web site. 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 their 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 not), 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). For deliberation-oriented users the opposite customizations were made: 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% (.22) and 3.36% (.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% (.29) and -4.01% (.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 since users expect some consistency in the user interface of a web site, 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, which we hope to address 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 regards to understanding which users are likely to make a purchase or not. We can predict those users that 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 promotions that are important elements of this problem. Path analysis doesn't 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 one-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 web site. Bucklin and Sismeiro (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 web site 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 consumer's search and choice behavior (Johnson and Payne 1985, Bettman, Luce, and Payne 1998, Shafir, Simonson, and Tversky 1993, Svenson 1996). Incorporating structural models of consumer behavior (Payne, Bettman, and Johnson 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 (2001a, 2001b, 2002) 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, pg. 99) found that shoppers whose path through the store include 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 which 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 analysis at the viewing-by-viewing level to the problem of purchase conversion. We hope that 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 appears to be a powerful source 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.

References

Alba, Joseph, John Lynch, Barton Weitz, Chris Janiszewski, Richard Lutz, Alan Sawyer, and Stacy Wood (1997), "Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Markets", *Journal of Marketing*, 61 (3), 38-53.

Allenby, Greg M., Neeraj Arora, and James L. Ginter (1998), "On the Heterogeneity of Demand", *Journal of Marketing Research*, 35, 384-389.

Ansari, Asim and Carl F. Mela (2003), "E-Customization", *Journal of Marketing Research*, 40 (2), 131-145.

Bettman, James R., Mary F. Luce, and John W. Payne (1998), "Constructive Consumer Choice Process," *Journal of Consumer Research*, 25, 187-217.

Blattberg, R. and J. Deighton (1991), "Interactive Marketing: Exploiting the Age of Addressability", *Sloan Management Review*, Fall.

Bucklin, Randolph E. and Catarina Sismeiro (2003), "A Model of Web Site Browsing Behavior Estimated on Clickstream Data", *Journal of Marketing Research*, (August) 249-267.

Bucklin, Randolph E., James M. Lattin, Asim Ansari, Sunil Gupta, David Bell, Eloise Coupey, John D.C. Little, Carl Mela, Alan Montgomery, and Joel Steckel (2002), "Choice and the Internet: From Clickstream to Research Stream", *Marketing Letters*, 13 (3), 245-258.

Cadez, Igor, David Heckerman, Christopher Meek, and Padhraic Smyth (2000), "Visualization of Navigation Patterns on a Web Site Using Model Based Clustering", *Technical Report MSR-TR-00-18*, Microsoft Research.

Coffey, Steven (1999), "Media Metrix Methodology," Media Metrix Working Paper, http://www.mediametrix.com/Methodology/Convergence.html

Crosby, L.A. and J.R. Taylor (1981), "Effects of Consumer Information and Education on Cognition and Chocie", *Journal of Consumer Research*, 8(1): 43-56.

Haaijer, Rinus and Michel Wedel (2001), "Habit Persistence in Time Series Models of Discrete Choice," *Marketing Letters*, 12(1), 25-35.

Hamilton, James D. (1994), *Time Series Analysis*, Princeton Univ. Press: Princeton, New Jersey.

Heer, J. and Chi, E.H. (2002a), "Mining the Structure of User Activity using Cluster Stability", *SIAM International Conference on Data Mining, Workshop on Web Analytics*, Arlington, VA.

Heer, J. and Chi, E.H. (2002b), "Separating the Swarm: Categorization Methods for User Access Sessions on the Web", *Proceedings of the Human Factor in Computing Systems Conference*, Minneapolis, MN.

Heer, J. and Chi, E.H. (2001), "Identification of Web User Traffic Composition using Multi-Model Clustering and Information Scent", *Proceedings of the Workshop on Web Mining, SIAM Conference on Data Mining*, Chicago, IL 51-58.

Hoffman, Donna L. and Thomas P. Novak (1996), "Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations," *Journal of 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, Nov. 26, 2001.

Janiszewski, Chris (1998), "The Influence of Display Characteristics on Visual Exploratory Search Behavior," *Journal of Consumer Research*, 25 (3), 290-301.

Johnson, Eric J. and John W. Payne (1985), "Effort and Accuracy in Choice," *Management Science*, 31, 394-414.

Kass, Robert E. and Adrian E. Raftery (1995), "Bayes Factors," *Journal of The American Statistical Association*, 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", *Journal of the American Statistical Association*, 75, 369, 146-153.

Keenan, D.M. (1982), "A Time-Series Analysis of Binary Data", *Journal of the American Statistical Association*, 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, Volume 454, Berlin: Springer.

Lewin, Kurt (1943), "Defining the Field at A Given Time," *Psychological Review*, 50, 292-310.

Li, Shibo, John C. Liechty, and Alan L. Montgomery (2002), "Modeling Category Viewership of Web Users with Multivariate Count Models," Carnegie Mellon University, GSIA, *Working Paper*.

Liechty, J. C., and G. O. Roberts (2001), "MCMC Methods for Switching Diffusion Models," *Biometrika*, 88, 2, 229-315.

Marketing News (1988), "VideoCart Shopping Cart with Computer Screen Creates New Ad", *Marketing News*, Chicago, May 9, 1988, pp. 1-2.

Mena, Jesus (2001), "WebMining for Profit: E-Business Optimization", Butterworth-Heinemann.

Meyers-Levy, Joan (1989), "The Influence of Sex Roles on Judgement", *Journal of Consumer Research*, 14 (March), 522-530.

Moe, Wendy, W. (2003), "Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology*, 13 (1&2), 29-40.

Moe, Wendy, W. and Peter S. Fader (2004), "Dynamic Conversion Behavior at e-Commerce Sites", *Management Science*, forthcoming.

Moe, Wendy W., Hugh Chipman, Edward I. George, and Robert E. McCulloch (2002), "A Bayesian Treed Model of Online Purchasing Behavior Using In-Store Navigational Clickstream," *Working Paper*.

Montgomery, Alan L. (2001), "Applying Quantitative Marketing Techniques to the Internet", *Interfaces*, 30, 2, 90-108.

Newton, Michael A. and Adrian E. Raftery (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," Journal of the Royal Statistical Society, Series B (Methodological), 56 (1), 3-48.

New York Times (2000), "Easier-To-Use Sites Would Help E-tailers Close More Sales," Bob Tedeschi, June 12.

Paap, Richard and Philip Hans Franses (2000), "A Dynamic Multinomial Probit Model for Brand Choice with Different Long-Run and Short-Run Effects of Marketing-Mix Variables," *Journal of Applied Econometrics*, 15, 717-744.

Pal, Nirmal and Arvind Rangaswamy (2003), *The Power of One: Leveraging the Potential of Personalization Technologies*, AMACOM.

Park, Young-Hoon, and Peter S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, forthcoming.

Payne, John W., James R. Bettman, and Eric J. Johnson (1993), *The Adaptive Decision Maker*, Cambridge University Press.

Pirolli, P. and S. K. Card (1999), "Information foraging", *Psychological Review*, 106 (4), 643-675.

Pitkow, James (1997), "In Search of Reliable Usage Data on the WWW", *Proceedings of the Sixth International WWW Conference*.

Redish, Janice (2002), "Information-Rich Web Sites: Challenges and Opportunities", www.redish.net/cmu.pdf. Last accessed October 2002.

Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing", *Marketing Science*, 15 (4), 321-340.

Rossi, Peter E. and Robert E. McCulloch (2000), "Bayesian Analysis of the Multinomial Probit Model", in *Simulation-Based Inference in Econometrics*, edited by Roberto Mariano, Til Schuermann, and Melvyn J. Weeks. Cambridge University Press, 158-178.

Seetharaman, Seethu (2004), "Modeling Multiple Sources of State Dependence in Random Utility Models: A Distributed Lag Approach", *Marketing Science*, forthcoming.

Shafir, Eldar B., Itamar Simonson, and Amos Tversky (1993), "Reason-Based Choice," *Cognition*, 49, 11-36.

Shapiro, Carl and Hal R. Varian (1998), "Information Rules: A Strategic Guide to the Network Economy," Harvard Business School Press.

Shulman, Richard (1993), "New systems, old practices create a POS 'generation gap'", *Supermarket Business*, New York, 48 (11), 17.

Sismeiro, Catarina and Randolph E. Bucklin (2003), "Modeling Purchase Behavior at an E-Commerce Website: A Conditional Probability Approach", Anderson School at UCLA, Working Paper.

Svenson, Ola (1996), "Decision Making and the Search for Fundamental Psychological Regularities: What Can Be Learned from a Process Perspective?" *Organizational Behavior and Human Decision Processes*, 65(3), 252-267.

Teräsvirta, T. (1994), "Specification, estimation, and evaluation of smooth transition autoregressive models", *Journal of the American Statistical Association*, 89, 208-218.

Underhill, Paco (1998), *Why We Buy: The Science of Shopping*, Touchstone Books.

Zhang, Jie and Lakshman Krishnamurthi (2004), "Customizing Promotions in Online Stores", *Marketing Science*, forthcoming.