CHOICE AND THE INTERNET:
FROM CLICKSTREAM TO RESEARCH STREAM

Randolph E. Bucklin
Anderson Graduate School of Management, UCLA

James M. Lattin
Graduate School of Business, Stanford University

Asim Ansari
Columbia Business School, Columbia University

David Bell
Wharton School, University of Pennsylvania

Eloise Coupey
Pamplin College of Business, Virginia Tech

Sunil Gupta
Columbia Business School, Columbia University

John D.C. Little
Sloan School of Management, MIT

Carl Mela
Fuqua School of Business, Duke University

Alan Montgomery
Graduate School of Industrial Administration, Carnegie Mellon University

Joel Steckel
Leonard N. Stern School of Business, New York University

Last Revised February 10, 2002

Acknowledgements: We thank Tulin Erdem and Russ Winer for organizing the U.C. Berkeley 5th Invitational Choice Symposium, during which time the authors discussed the ideas presented in this paper. We also thank two anonymous reviewers and Russ Winer for their comments and suggestions.

CHOICE AND THE INTERNET:
FROM CLICKSTREAM TO RESEARCH STREAM

ABSTRACT

The authors discuss research progress and future opportunities for modeling consumer choice on the Internet using clickstream data (the electronic records of Internet usage recorded by company web servers and syndicated data services). The authors compare the nature of Internet choice (as captured by clickstream data) with supermarket choice (as captured by UPC scanner panel data), highlighting the differences relevant to choice modelers. Though the application of choice models to clickstream data is relatively new, the authors review existing early work and provide a two-by-two categorization of the applications studied to date (delineating search versus purchase on the one hand and within-site versus across-site choices on the other). The paper offers directions for further research in these areas and discusses additional opportunities afforded by clickstream information, including personalization, data mining, automation, and customer valuation. Notwithstanding the numerous challenges associated with clickstream data research, the authors conclude that the detailed nature of the information tracked about Internet usage and e-commerce transactions presents an enormous opportunity for empirical modelers to enhance the understanding and prediction of choice behavior.

Key words: Internet, Choice Models, Clickstream Data

1. Introduction

The detailed electronic information that is tracked and stored about Internet usage and e-commerce transactions (known generally as "clickstream data") presents empirical researchers with a significant opportunity to advance the understanding and prediction of consumer choice behavior. Because, for example, it tracks so much of what takes place prior to a purchase, a clickstream data set might provide more detailed information regarding the choice process followed by consumers than the records contained in UPC scanner panel data sets. The Internet also opens up new types of choice decisions for study that are not present in the supermarket/ scanner data context. These include choosing page links, choosing web sites, and choosing products culled or recommended by electronic agents. On the Internet, choice is not only ubiquitous, it is also electronically recorded in great detail.

Although the Internet has had about six years of consumer use, relatively few papers have been produced (either published or in working paper form) in which choice models have been applied to clickstream data. Part of the problem is attributable to delays in obtaining access to clickstream data (which firms often view as highly proprietary), and to the difficulties associated with cleaning, filtering, and processing the data. Because work in this area is relatively new, there are also many unresolved questions about what problems are worthwhile to study and how best to go about building models to address them.

The purpose of this paper is to accelerate the pace of research in this area. We begin by benchmarking clickstream research against scanner panel research, asking the question, “What, if anything, is different?” We then go on to discuss four areas of choice model application in which researchers have begun to make progress: (1) within-site navigation choices, (2) e-commerce purchase decisions, (3) choice among web sites and (4) purchase choices involving electronic agents, particularly shopbots. We outline new areas of choice-related research to explore, and we discuss the challenges associated with choice modeling on clickstream data sources.

2. Clickstream Data: What is Different?

The term “clickstream” denotes the path a visitor takes through one or more web sites. The pathway reflects a series of choices made both within a web site (e.g., which pages to visit, how long to stay, whether or not to make an online purchase) and across web sites (e.g., which sites to visit). For the most part, raw clickstream data are captured in one of two ways. First,

server log files maintained on behalf of a web site owner can record all the requests and information transferred between the visitor's computer and the server during a web site visit. Server log files can record information on the cookie id of the visitor, so that it is possible to identify unique users and return visits.[1] Second, panel data (e.g., supplied by ComScore, NetRatings, MediaMetrix) captures the URL's of all pages requested during web usage by transmitting that information from the user's computer to the panel-data supplier. Third, clickstream data can also be collected by a user's Internet service provider (ISP). When users request web pages the ISP which may record these requests thereby creating another source of clickstream data.[2]

Panel (or ISP) data sources may also be able to associate detailed demographic information with each individual panelist. But, on the other hand, they lack some of the detailed information about the interaction between user and server collected in web server log files. Both sources typically include information on the IP address of the visitor, the type of browser used, a time stamp, and the previous URL visited. Although much information of potential interest is not captured by these sources (see Section 5 below), clickstream data provide far more detail than the scanner panel data that have been used in the development and testing of choice models in marketing since the early 1980's.

2.1 Is Internet choice behavior different?

Just the presence of a new data source is not necessarily sufficient to call for an entirely new program of research. The question we need to ask is whether or not Internet choice behavior is any different from the kind of choice behavior that marketers have already extensively studied in the context of the supermarket. If it is not, then we already have a suitable research paradigm and a wide range of methods available (e.g., Bucklin and Gupta 1999). In many respects, there do appear to be substantial (and substantive) differences in the choice behavior we observe during the course of Internet usage.

Table 1 shows several points of contrast between Internet choice behavior and the brand/category/store choice behavior that has been the focus of most choice modeling activity in marketing. For example, brand choice behavior in the supermarket context is typically modeled

[1] While this does not guarantee that every site visit can be identified, Dreze and Zufryden (1998) have shown that this limitation does not pose a problem in practice.

[2] Clickstream data can also be collected in a laboratory setting to record the actions of experiment subjects. In this paper we limit our discussion to clickstream data on actual (i.e., non-laboratory) web usage.

as a single discrete choice event characterized by a static choice context and a fully informed consumer. Given the practical difficulties associated with personalizing a supermarket shopping experience, this modeling approach fits the choice environment well.[3]

Internet Choice	**Supermarket Choice**
▪ Intent unclear at outset o Browse? Search? Buy? ▪ Active/Interactive o Visitor participates in creating choice context ▪ Addressable o Choice context is personalizable ▪ Dynamic o Marketers can intervene at low cost	▪ Store visit reveals purchase intent ▪ Passive ▪ Fixed o Choice context common ▪ Static

Table 1. *Internet choice behavior versus supermarket choice behavior.*

In contrast to the relatively static supermarket environment captured by scanner panel data, Internet choice behavior is dynamic. Internet choice behavior may be described as an evolving series of interrelated choices, where both consumer and marketer can play a role in shaping the context of subsequent choice events depending upon the outcome of earlier encounters. The website visitor (who may or may not be interested in purchase but who is a consumer of the information provided) makes a request by clicking on a link (or providing information) that leads to a new page containing new information and new choices (in the form of more links). At each request by the visitor (conditional on what is known about the visitor to that point), the marketer has an opportunity to respond in a myriad of ways.

Modelers in the clickstream environment must therefore grapple with a larger volume of consumer choices being recorded in the data, different types of choices, and the potential to personalize the choice environment on a dynamic basis. Considering the added complexity that the intent of the Internet visitor (e.g., browsing? search? purchase?) may not be discernible to the

[3] An exception might be the targeted couponing system pioneered by Catalina Marketing. However, the extent of the personalization is confined to the coupons made available for redemption at a subsequent store visit.

marketer (or, for that matter, to the visitor), it seems clear that new models based on new assumptions will be necessary for understanding this behavior and predicting it well.

3. Clickstream Research on Choice Behavior

Though there is a great deal being written about the Internet in marketing, there is relatively little work that applies choice models to clickstream data. Nevertheless, there are a number of promising early papers that merit discussion. We classify this existing work using the following two dimensions: (1) the objective of the individual (i.e., search or browsing versus purchase) and (2) the realm of the choice decision (i.e., within website versus across websites). Using this organizing framework, we categorize the existing research in Internet choice as falling into one of four (potentially overlapping) clusters: (1) site navigation, (2) e-commerce and recommendation, (3) site choice, and (4) shopbots. We now briefly discuss the existing research in each of these four areas.

	Within website	Across websites
Search	Site Navigation	Site Choice
Purchase	E-Commerce and Recommendation Systems	Shopbots

Table 2. *Framework for organizing Internet choice applications.*

3.1 Site Navigation

Site navigation choices involve the number of pages to view, the time to spend viewing a page (or the site as a whole), the decision to stay or exit the site on a given page, and the choice of which links to follow and/or which pages to view. All of these decisions are conditional upon a site visit taking place (i.e., predicated on the user having already arrived at the site in question). These decisions may reflect the attractiveness or "stickiness" of the web site and may also affect the ability of the site owner to raise revenue through banner advertising impressions or other exposure-related vehicles (e.g., pop-unders, pop-ups).

Perhaps the earliest published application of choice theory to site navigation decisions is Huberman et al (1998). The authors assumed that the utility of an additional page view is equal to the utility of the current page plus a normally distributed error. From this assumption, they derived that the distribution of the total number of page requests made by users at a site will

follow an inverse Gaussian. The model provides a good fit to cross-sectional data collected from company clickstreams, but it is not predictive of the actions of individual users, nor does it incorporate any covariates.

In addition to the total number of pages viewed, another outcome of navigation choice decisions is the total amount of time spent at a web site. Johnson, Bellman and Lohse (2000) model this duration as a function of the number of repeat visits made by a user to a given site. Drawing from cognitive science, they propose a power law of practice, in which session duration is expected to fall with repeat visitation according to a power function. The model is mathematically analogous to the learning curve or experience curve. Using MediaMetrix panel data, they test their model on the session durations of individual users and find that users do spend less time per session the more they visit the site. This finding suggests that learning may play a role in influencing site navigation choice decisions.

Bucklin and Sismeiro (2001) propose a model of web site browsing behavior that incorporates both page views and duration time. They use a binary probit to model the user's decision to stay or exit the site at each page view, and they use a proportional hazard model to predict the user's decision of how long to view each page. Based on data from server log files, the authors find that a visitor's propensity to stay with a site changes dynamically as a function of the depth of site visit and the number of repeat visits to the site. More specifically, repeat visits by a user lead to fewer pages viewed per session, but to no change in average page view duration. This suggests that the learning effects noted above are due to better navigation through the site, as opposed to faster processing of information from each page.

When deciding to remain on the site by requesting an additional page, the website visitor may have the choice of clicking on one of many available links. The location and number of links may influence the user's decision to click-through, and which link to select. To our knowledge, however, no research in marketing has yet specifically addressed this problem. Ansari and Mela (2000), in developing an approach to customize e-mails for the purpose of attracting visitors to a site, examine the effect of link order and placement on the probability of clicking. However, these are links within an e-mail and not links on a web page within a site.

3.2 *E-Commerce and Recommendation Systems*

The second application area also involves within-site behavior, but instead of search and browsing it focuses explicitly on purchase. In e-commerce, there is much concern among

practitioners with low rates of visit-to-purchase conversion. To address this problem, we need to understand the factors that influence the purchase decision and build models that make it possible for firms to intervene to influence purchase behavior. Moe and Fader (2001) develop a probability model of visit-to-purchase conversion that accounts for heterogeneity. Using the pattern of previous visits and purchases recorded in MediaMetrix data for Amazon and CDNow, the model predicts those visits most likely to convert into purchases. The authors find that repeat visits are positively related to purchase.

In recommendation systems, the goal is to provide users with guidance about what products to select, using the preference structure of the individual user (which may be evaluation ratings or revealed choice) and the preferences and choices of other site visitors. Typically, these recommendation systems are based on some form of collaborative filtering (Breese, Heckerman and Kadie 1998), which predicts an individual's evaluation (or choice) as a weighted sum of the evaluations (or choices) of other site visitors. Ansari et al. (2000) suggest that it may be possible to improve upon collaborative filtering by using a hierarchical Bayesian approach to calibrate a model that accounts for unobserved heterogeneity across individuals and across the alternatives being evaluated.

One source of low visit-to-purchase conversion on the Internet may be that many visitors do not have a purchase objective in mind when coming to the site. An important first step for choice models may be to use data from the clickstream to infer the goal of the individual Internet user (i.e., is this person simply browsing for information, or intent on making a purchase?). Researchers have had some success in using Bayesian methods to forecast the characteristics of individuals based on the sites they have visited (Montgomery 2001a);[4] it may also be possible to conduct this type of "user profiling" to separate casual browsers from those intent on purchase. Moe (2001) also provides a contribution with a typology of store visits that vary in terms of conversion rate. Using clickstream data from an online store, she finds that the in-store browsing patterns of users and the general content of pages viewed cluster together with different conversion rates. Thus, marketers may be able to begin to discern which visitors are more likely to be the best prospects for online purchasing.

[4] For example, DoubleClick is able to use the site visit history of an individual to predict with confidence the gender of the user, which is helpful in selecting targeted advertising material for future web site visits.

Another question is whether we can model the probability of an online purchase, given visit, by following the progress of the user through a session. If navigation-related choices are predictive of purchase, then it should be possible to improve the prediction of purchase likelihood in real time or "on the fly." Montgomery (2001b) proposes a dynamic, ordered probit model of site exit, browsing, and purchase that is designed to take into account the effect of covariates on a user's decision to exit, remain on the site browsing, or purchase.

Sismeiro and Bucklin (2001), using clickstream data from an online reseller of automobiles, propose a probit model to predict the probability of purchase given visit as a function of what the user does on the site, repeat visitation, and other covariates. They also incorporate the interim completion of so-called "nominal user tasks" prior to purchase (e.g., filling the shopping cart, configuring a car). This approach provides a way of decomposing the source of purchase likelihood to improve prediction power and provide better diagnostics about the effectiveness of the site. They find that what users do and are exposed to on the site is strongly predictive of purchase probability. Surprisingly, repeat visitation is not diagnostic of buying propensity in their data, nor does the site's offering of sophisticated decision aids produce increased conversion rates.

3.3 *Site Choice*

Turning to across-site choice behavior, two papers based on clickstream data have proposed models bearing on the user's decision to visit one site versus another. Johnson et al (2000), using MediaMetrix panel data, studied across-site search behavior for three product categories. They found that there was surprisingly little search across web sites in each of the three categories, but that strict loyalty to a single site was not commonly observed. The authors concluded that users "have some propensity to search, but rarely exercise it." Using clickstream panel data from Foveon, Goldfarb (2001) applies the structure of the Guadagni and Little (1983) brand choice model to predict user choice of web portals. He constructs a site loyalty measure analogous to the brand loyalty measure first proposed by Guadagni and Little and widely used in scanner panel research. Like brand loyalty in scanner data, he finds that site loyalty in clickstream data is strongly predictive of site choice. Goldfarb's study also includes a series of other covariates predictive of site choice that should be of interest to researchers, including use of links, email, and the user's previous experience with each site. Also related to site choice are

ad click-through (Chickering and Heckerman 2000) and the use of search engines (e.g., Bradlow and Schmittlein 2000).

3.4 *ShopBots*

While a purchase decision might be modeled as a binary outcome conditional on the visit to an e-commerce site, users may also consider online purchases across multiple sites. Researchers have begun to study part of this choice process by examining the choice decisions that users make at "ShopBots" – shorthand for shopping robots. In response to queries, a shopbot presents users with a set of alternative products and prices from competing online merchants. Users are then free to select among the alternatives or exit the shopbot site. Brynjolfsson and Smith (2000) study the choice decisions users make at shopbot sites using a multinomial logit choice model in which the choice set consists of the items presented to the user by the shopbot. They report that both price and brand name of the online merchant are important determinants of the choice decisions made.

Despite the appeal of shopbots, consumer usage rates remain low. Montgomery, et al. (2001) adopt the perspective of the shopbot operator and seek to identify factors under the operator's control that might improve the appeal of using shopbots (e.g., waiting time, merchants searched, and extent of offerings presented). The authors show how the design of a shopbot can be improved by modeling consumer utility for shopbot purchases, demonstrating the validity of their model using observed prices at online bookstores over a six-month period.

4. Research Opportunities in Clickstream Choice Modeling

The availability of clickstream data affords choice modelers a wide range of opportunities for studying Internet choice behavior and understanding how it may be influenced. In addition to encouraging work ongoing in the four application areas discussed above, we outline several additional research opportunities that may be worthwhile for choice modelers to pursue.

4.1 Improved Focus and Opportunity

UPC scanner data was once described as the "electron microscope of marketing." It was a device capable of measuring things that had never been captured before, and allowed marketers to develop theories and build models of phenomena that prior to the scanner were effectively unobservable. Now, with data from the clickstream, it is possible that we have a measurement device of even greater resolution. Consider, for example, the choice of multiple items from a single product category. With scanner panel data, we knew only which items ended up in the

basket at the end of the shopping trip. But with clickstream data from an online grocer such as Netgrocer (Bell 2001), it is possible to know not only the order in which the items from the category are chosen, but also what other items are in the basket at the time of choice (and where in the store the customer has visited before coming to this particular category). Due to the inflexibility of the physical supermarket environment, it is impossible to do any experimentation to assess customer response and test theory; in the virtual shopping environment it is possible to deliver a coupon (or otherwise promote a particular product) at an exactly specified stage of the shopping trip.

Clickstream data from online retailers like Netgrocer could also help researchers understand more about consumer search and the formation of consideration sets. The detailed trace of products examined but not purchased could provide information on consideration and enable researchers to test formal models of search. Thus, research on consideration and choice could see a resurgence in marketing as investigators are able to move beyond the limitations of the purchase-only information contained in scanner panel records. Presumably, this sort of opportunity should also extend to researchers who study the formation of market baskets by showing which categories might or might not enter the final basket as part of the search and consideration process at the category level.

4.2 Customization

The opportunity to customize a web site for a given user, adjust content "on the fly," and use the web for targeted marketing actions opens up a wide range of modeling challenges. First, the customization actions require the development and calibration of models for user response to website and Internet stimuli (e.g., to promotional e-mails). Because the opportunity to customize exists at the individual level, it is natural to turn to choice models for this purpose. Second, customization also requires that the decisions of what to do be optimized over the heterogeneity in response. Thus, researchers confront the challenge – and opportunity – of combining the ability of choice models to parsimoniously represent response heterogeneity with optimization algorithms. We expect a plethora of research opportunities to arise as modelers explore not only customization of recommendation systems (Ansari, et al. 2000) and emails (Ansari and Mela 2000), but also the customization of web site pages, link choices, promotional interventions, and prices and product assortments. Indeed, customization could ultimately extend to portal design and to search engine functionality.

4.3 *Integrating Data Mining and Choice Models*

As indicated above, new clickstream data sources are vast in size and potentially quite complex in terms of their level of detail. Many of the traditional statistical methods developed and used in marketing may not be scalable; that is, they may not be well-suited to handling databases with millions of customers and hundreds of millions of transactions. Part of the problem is that the traditional statistical methods were developed to deal with the consequences of data scarcity (i.e., small, clean samples whose sampling properties -- usually IID -- are well understood).

In situations characterized by an abundance of data, a practice known as data mining has emerged. Data mining is an inductive, exploratory analysis aimed at uncovering unsuspected relationships in the data (Berry and Linoff 1997; see also Cooper and Giuffrida 2000). The tasks (typically classification, profiling, and prediction) are accomplished by a range of different tools (e.g., basket analysis, CART, neural networks) that are efficient enough to be used with data sets of enormous size, obviating the need for working with only a subset of the data. Data mining is also characterized by a different philosophy: its focus is more on predictive results (e.g., inferences about future co-incidence of items in a basket) than about model parameter estimates and statistical significance (Steckel 2001). As marketing researchers turn their attention to large sets of clickstream data, it may be counterproductive to rely primarily on standard statistical methods. Emphasizing scalable models and predictive results may enable us to observe a richer set of behavioral phenomena in clickstream data.

4.4 *Automation*

Every click by an Internet user is a potential opportunity for personalization, a chance for a website owner to change the choice context or to provide the user with specially selected information. Such a high degree of targeted content would be unthinkable without some form of automation. Fortunately, in the context of Internet choice, the conditions are ideal for some form of automated system; almost everything on the Web must be programmed, decision rules and models are easily deployed (e.g., using tools like Broadvision), and huge amounts of data are automatically collected with which to calibrate these models. The question remains: what should website owner X do when visitor Y arrives at the site on Monday morning?

Little (2001) proposes a five-point framework for marketing automation, focused specifically on the needs of online retailers such as Amazon.com or Circuit City. The idea is to

deploy real-time decision rules (e.g., pricing and promotion decision, page design on the fly), to calibrate them using historical data (not only clickstream data from visitors and customers, but also information from comparison engines and web crawlers), and to update them through some form of adaptive experimentation. A key requirement of the system is that it provide constant feedback to website management in the form of quality control, trend monitoring, and early warning on market changes. Such a feedback mechanism might be similar in form to the expert systems used to monitor changes in the supermarket environment or to those needed to implement automation in marketing decision-making (e.g., Bucklin, Lehmann, and Little 1998).

One of the principle challenges to effective automation is getting a handle on the dizzying array of marketing decision variables that come into play at each possible choice juncture. As the visitor arrives at the home page of a typical Internet retailer, he/she is usually presented with a search function, a category index, special offers, promoted products, top sellers, and a wide range of sponsored advertising; each promoted product may by accompanied by information such as a picture, a price, endorsements, advertising copy, product reviews, and a link to a sales advisor (or other contact information). Understanding how the user responds to these mix elements is critical in designing systems to automate how they are set.

4.5 *Lifetime Value Analysis*

Increasingly, companies are turning to accounting systems that recognize the importance of the lifetime value (LTV) of the customer to the firm (Berger and Nasr 1998). Long-term value depends upon the ability of the firm to retain the customer (strengthening the relationship by engaging in frequent interaction) and to make the relationship profitable (through cross-selling of complementary products and services). With clickstream data to track the pattern of request and response between the customer and the firm, it should be possible to identify patterns that increase the success or cost-effectiveness of cross-selling and to diagnose early-warning signs of potential customer defection. Clickstream data give us the opportunity to better understand and influence customer retention, which is a primary driver of customer LTV (Gupta, Lehman and Stuart 2001).

5. Challenges for Choice Modelers Using Clickstream Data

While the research opportunities associated with clickstream data are manifold, the challenges of dealing with them are also numerous. The decisions about what to record in server log files were almost certainly made based on what it was *possible* to capture, rather than what

would be most *useful* for choice modelers. Though we now have an extraordinary amount of detailed information, there is also a great deal missing from clickstream data. One of the major limitations of server log files is that they are limited to capturing the interaction at a single site. This is analogous to having scanner panel data from a single store in a market served by many stores; it is impossible to understand the extent to which a website visitor is meeting his/her needs by visiting other sites on the Internet.

This limitation is partly addressed by Internet panel data (e.g., Media Metrix). These data have the advantage of tracking the navigation behavior of a sample of individuals across all websites, capturing each URL visited (as well as information about the time spent in each domain). While this provides us with insight into website loyalty and site switching behavior, the data are surprisingly sparse. Even with several hundred thousand panelists, it is difficult to achieve reasonable coverage of the Internet choice behaviors of over 300 million Internet users across perhaps a billion distinct web pages worldwide. To compound the problem, Internet panel data are not as rich as server log files, which provide a record of all the information requested from the server for a given page view. Internet panel data record only the URL of the site visited (and the URL is often truncated), which makes it difficult (if not impossible) to reconstruct what the user actually saw when visiting the web page or to understand the interaction that takes place between pages of a website.

Both sources of clickstream data (within and across site) are beset by other problems as well. For example, how do we know whether or not an individual has achieved his/her objective from the session? Unlike a supermarket, where the objective of the shopping trip is generally the same for everyone and where the success of merchandising activity can be inferred from observed purchase behavior, it is not always clear what an Internet user hopes to accomplish when visiting a web site. Without some independent form of corroboration, it is difficult to determine whether a short visit is indicative of finding exactly the right information or of a wrong turn in the search.

Early in the life cycle of scanner panel data, researchers and practitioners realized that the data did not capture all of the important elements influencing consumer choice behavior, so they invested in building comprehensive "single source" datasets. We may have reached the same stage with clickstream data. We believe that current clickstream data sources may actually capture a smaller proportion of information relevant to overall Internet choice behavior than the

early scanner data sets did with respect to supermarket choice behavior. Even if we are able to address the single source issue, we still may not find as much information at the individual level in clickstream data. Unlike scanner data, where over the course of a year the average household makes 100 trips to the store and multiple purchases per product category, clickstream data may reveal that most users make only one or two visits to a particular site.

The nature and limitations of existing forms of clickstream data may require choice modelers to develop methodological innovations. To begin with, potentially sparse information about a given user can create a need for models to be more parsimonious with respect to the way in which they capture heterogeneity (e.g., random effects models calibrated using hierarchical Bayesian methods). Because only a portion of Internet usage and choice behavior may be captured in the data, the choice decisions that are modeled may increasingly have to be conditional ones. For example, in studying the role of shopbots, the user's choice of partner link can be modeled *given* that a link is selected *and* a visit to a shoptbot takes place. Similarly, in studying e-commerce, the purchase decision can be modeled as a series of conditional choices (visit to the site, search on the site, items placed in shopping cart, order submitted). Indeed, it was the conditional approach used by Guadagni and Little (1983) – brand choice *given* category purchase and store visit -- that launched the ensuing stream of choice modeling work on scanner panel data. As with any conditional model, researchers should be mindful of the risk of selectivity bias and that in narrowing their focus other important choice decisions may go neglected.

We recommend that researchers encourage panel data providers or companies from whom they obtain server log files to capture the information needed to identify purchases and/or the achievement of other site-related objectives. To model the effects that web page content has on choice behavior, the clickstream data on browsing and purchase also needs to be tied to the actual page files and web page content viewed by the users involved. This is an important need if researchers are to take exploit the multi-site coverage offered by panel (or ISP) data versus the single-site coverage provided by server log files. Sufficient time periods also need to be covered in order to capture repeat visits and provide multiple visit observations for individual users. While it may take a few years to assemble the clickstream analog of "single source" data, we believe researchers can still make a great deal of progress in the meantime by working cleverly with existing types of data.

6. Conclusion

In this paper, we discussed four areas of early application of choice models to clickstream data: within-site navigation, e-commerce and recommendation, site choice, and shopbot use. We believe that researchers will find much of significance to pursue in each area. We also identified several additional areas in which choice modeling can play an important role. These include taking advantage of the finer resolution provided by clickstream data versus scanner data (e.g., NetGrocer), customization, data mining, automation, and customer lifetime value analysis.

Clickstream data offer a wide range of research opportunities for choice modeling. To structure a long-term research agenda on clickstream data and Internet choice, we can organize research into three domains, substantive, conceptual, and methodological (Brinberg and McGrath 1985). The substantive domain contains the applied problems of interest, the conceptual domain contains the ideas or theories that can be used to explain the phenomenon, and the methodological domain contains the measures and designs used in a research project (Coupey 1999, p. 197). Our discussion has emphasized the substantive domain of application areas, as well as some of the modeling methodologies and measurement challenges that we believe researchers will confront. As we accumulate greater empirical evidence about choice on the Internet, the new data provided by clickstreams should also enable us to develop new theories and concepts about consumer choice, thereby completing the research triangle.

The advent of the Internet and the detailed tracking information provided by clickstream data provides a vast and exciting array of research opportunities for empirical choice modelers. Scanner panel data opened up a major research stream in choice modeling that is still delivering contributions to knowledge almost two decades since it first began. Such data, however, was limited to the relatively stable choice context of the supermarket and consumer packaged goods. Clickstream data, on the other hand, will give researchers a window into the choice process prior to purchase, enable us to examine many types of products and services beyond packaged goods, and permit inquiry into choice behavior in a dynamic environment where marketers can quickly personalize the stimuli to which their customers are exposed. Indeed, the nature of cilckstream data itself is likely to continue to evolve rapidly. New devices and applications (e.g., hand held computers, PDA's, Real Audio, Flash Media, and peer-to-peer protocols such as Gnutella) will provide information on choice behavior well beyond the confines of WWW surfing and pageviews, and purchases. It promises to be an exciting time.

REFERENCES

Ansari, Asim, Skander Essegaier and Rajeev Kohli. (2000). "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (3), August, 363-375.

Ansari, Asim and Carl Mela. (2000). "E-Customization," *Working Paper*, Columbia University.

Bell, David R. (2001). "Customer Base Evolution at NetGrocer.com," Paper Presented at the UC Berkeley Fifth Invitational Choice Symposium, Monterey, CA. *www.andrew.cmu.edu/user/alm3/presentations/choicesymposium2001/bell.pdf*

Berger, Paul and Nada Nasr. (1998). "Customer Lifetime Value: Marketing Models and Applications." *Journal of Interactive Marketing*, 12 (1), 17-30.

Berry, Michael and Gordon Linoff. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support.* New York: John Wiley & Sons.

Bradlow, Eric and David Schmittlein. (2000). "The Little Search Engines That Could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science*, 19 (1), 43-62.

Breese, John, David Heckerman, and Carl Kadie. (1998). "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Technical Report MSR-TR-98-12*, Microsoft Research.

Brinberg, D. and J.E. McGrath. (1985). *Validity and the Research Process.* Newbury Park, CA: Sage Publications, Inc.

Brynjolfsson, Erik and Michael D. Smith. (2000). "The Great Equalizer? The Role of Shopbots in Electronic Markets." Working Paper, MIT Sloan School of Management, Cambridge, MA.

Bucklin, Randolph E. and Sunil Gupta. (1999). "Commercial Use of UPC Scanner Data: Industry and Academic Perspectives," *Marketing Science*, 18 (3), 247-273.

Bucklin, Randolph E., Donald R. Lehmann, and John D.C. Little. (1998). "From Decision Support to Decision Automation: A 2020 Vision," *Marketing Letters*, 9 (3), 235-246.

Bucklin, Randolph E. and Catarina Sismeiro. (2001). "A Model of Web Site Browsing Behavior Estimated on Clickstream Data," *Working Paper*, Anderson School at UCLA.

Chickering, David and David Heckerman. (2000). "Targeted Advertising with Inventory Management," *ACM Special Interest Group on E-Commerce (EC00)*, Minneapolis, MN, 145-149, October.

Cooper, Lee G., and Giovanni Giuffrida. (2000). "Turning Datamining into a Management Science Tool," *Management Science, 46 (2).*

Coupey, Eloise. (1999). "Advertising in an Interactive Environment: A Research Agenda," in David Schumann and Esther Thorson, editors, *Advertising and the World Wide Web*, Erlbaum: Mahwah, NJ, pp. 197-215.

Dreze, Xavier and Fred S. Zufryden. (1998). "Is Internet Advertising Ready for PrimeTime?" *Journal of Advertising Research*, 38 (3).

Goldfarb, Avi. (2001). "Analyzing Website Choice Using Clickstream Data." *Working Paper*, Northwestern University, Department of Economics.

Guadagni, Peter and John D.C. Little. (1983). "A Logit Model of Brand Choice Calibrated on Scanner Panel Data," *Marketing Science*, 2 (3), 203-238.

Gupta, Sunil, Donald Lehmann, and Jennifer Stuart. (2001). "Valuing Customers," *Working Paper*, Graduate School of Business, Columbia University.

Huberman, Bernardo A. Peter Priolli, James E. Pitkow and Rajan M. Lukose. (1998). "Strong Regularities in World Wide Web Surfing," *Science*, 280 (3), April, 95-97.

Johnson, Eric J. Steven Bellman, Gerald L. Lohse. (2000). "What Makes a Web Site 'Sticky'? Cognitive Lock In and the Power Law of Practice," *Working Paper*, Graduate School of Business, Columbia University.

Johnson, Eric. J., Wendy Moe, Peter Fader, Steven Bellman, and Gerald Lohse. (2000). "On the Depth and Dynamics of Online Search Behavior," *Working Paper*, Graduate School of Business, Columbia University.

Little, John D.C. (2001). "Marketing Automation on the Internet," Presentation at the UC Berkeley Fifth Invitational Choice Symposium, Monterey, CA. *www.andrew.cmu.edu/user/alm3/presentations/choicesymposium2001/lattin.pdf*

Moe, Wendy. (2001). "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstreams," *Journal of Consumer Psychology*, forthcoming.

Moe, Wendy W. and Peter S. Fader. (2001). "Which Visits Lead to Purchases? Dynamic Conversion Behavior at e-Commerce Sites." *Working Paper*, Wharton School.

Montgomery, Alan L. (2001a). "Applying Quantitative Marketing Techniques to the Internet," *Interfaces*, 31 (2).

Montgomery, Alan L. (2001b). "Modeling Purchase and Browsing Behavior Using Clickstream Data," Presentation at the UC Berkeley Fifth Invitational Choice Symposium, Monterey, CA. *www.andrew.cmu/edu/user/alm3/presentations/choicesymposium2001/montgomery.pdf*

Montgomery, Alan L., et al. (2001). "Designing a Better Shopbot," *GSIA Working Paper*, Carnegie Mellon University.

Sismeiro, Catarina and Randolph E. Bucklin. (2001). "Modeling Purchase Behavior at an E-Commerce Web Site: A Task Completion Approach," *Working Paper*, Anderson School at UCLA.

Steckel, Joel. (2001). "What Can We Learn from Data Mining?" Presentation at the UC Berkeley Fifth Invitational Choice Symposium, Monterey, CA. *www.andrew.cmu.edu/user/alm3/presentations/choicesymposium2001/steckel.pdf*