Click Here for Internet Insight:
Advances in Clickstream Data Analysis in Marketing

Randolph E. Bucklin
Catarina Sismeiro*

March 2008
(Forthcoming at Journal of Interactive Marketing)

* Randolph E. Bucklin is Peter W. Mullin Professor, UCLA Anderson School, 110 Westwood Plaza, Los Angeles, CA 90095. Catarina Sismeiro is a Lecturer at Imperial College London, Tanaka Business School. Email correspondence to rbucklin@anderson.ucla.edu. The authors thank the editor and two anonymous reviewers for comments.
Click Here for Internet Insight:
Advances in Clickstream Data Analysis in Marketing

Abstract

Clickstream data are defined as the electronic record of Internet usage collected by Web servers or third-party services. The authors discuss the nature of clickstream data, noting key strengths and limitations of these data for research in marketing. The paper reviews major developments from the analysis of these data, covering advances in understanding (1) browsing and site usage behavior on the Internet, (2) the Internet’s role and efficacy as a new medium for advertising and persuasion, and (3) shopping behavior on the Internet (i.e., electronic commerce). The authors outline opportunities for new research and highlight several emerging areas likely to grow in future importance. Inherent limitations of clickstream data for understanding and predicting the behavior of Internet users or researching marketing phenomena are also discussed.

Keywords: Internet, Marketing, E-commerce, World Wide Web, Clickstream Data
1. Introduction

The pervasiveness of the Internet, and its increasing and widespread influence as an information source, marketplace, and setting for social contact, have sparked growing interest in studying what people do online and how their behavior can be predicted and influenced. A thorough understanding of online behavior has also become a necessity for the success of websites and other online businesses given the complex environment in which they compete. In addition, regulators and advocacy groups, among others, have also demonstrated a keen interest in the analysis of online user activity, either for monitoring purposes or for protecting users from the potentially harmful actions of others.

Unlike traditional media and commercial settings, the Internet allows fast, easy, and unobtrusive collection of detailed information on individual activities. This record of an Internet user’s actions online has come to be known as clickstream data. The clickstream data available to an analyst can differ in detail and coverage of user activity, a feature that provides both opportunities and challenges. And though the data are not without pitfalls, they are collected in the users’ own environment with no artificial interruptions. This makes clickstream data a rich resource for researchers and practitioners who seek to better understand the behavior and choices made by individuals. Because clickstream data provide a detailed new window on behavior, insights developed from clickstream analysis need not be confined entirely to online settings.

The purpose of this paper is to provide an overview of key developments and research advances relevant to marketing that have been based on clickstream data over the past decade. We also explain advantages and challenges of using and studying these data and discuss future research opportunities. In presenting our review, we recognize that there is a substantial literature outside the marketing field that is also based on clickstream data analysis (e.g., computer science, sociology, etc.); generally, this material will be beyond the scope of this review. Within marketing, we also focus on clickstream
data collected in the natural course of website operations; thus, research based on
electronic records from laboratory or field experiments will not be extensively reviewed.

The balance of the paper is organized as follows. We begin by discussing the
different types of clickstream datasets, their characteristics, and the issues that arise when
using one type of clickstream dataset versus another. We then review key research
findings in marketing that have been based on clickstream data analysis, organizing these
into three themes: (1) usage of the Internet, (2) advertising on the Internet, and (3)
shopping on the Internet (e-commerce). We end with a discussion of emerging areas of
analysis, promising research opportunities, and challenges in harnessing the power of
these data.

2. Clickstream Data
Clickstream data are defined as the electronic record of a user’s activity on the Internet.
Thus, the data trace the path a visitor takes while navigating the Web. This path reflects
choices, often very large in number, made by the user both within and across websites.
For example, a clickstream dataset might include a record of every website visited and
every page seen, the time the user spent on each site or page, and the order the sites and
pages were visited. A key unit of observation in clickstream data is the page view – the
recording of a user’s exposure to a given website page. Technically, the assembly of a
“page view” from the user’s perspective can involve numerous “hits” to the Web server.
These reflect the downloading of various page elements before they are assembled in the
user’s Internet browser window. In many cases, clickstream data is automatically
aggregated from hits to page views but in other cases (e.g., raw server log files), the
analyst may need to perform this step.

A clickstream dataset could also include information on newsgroups the user
participated in, the banner advertisements the user clicked on, the sequence of bids
performed in online auctions, and the products and services that were purchased by the
user online. Because there is yet no unifying definition of precisely what will be
contained in a clickstream dataset, we briefly discuss the various sources of this
information, the types of datasets in common use, and some of the characteristics that
make clickstream data distinctive.
2.1 Clickstream Data Sources

Raw clickstream data can be captured in multiple ways. Server log files maintained by, or on behalf of, a website can record all the requests and information transferred between the server and the user’s computer. Because these data are collected by a single website, they are known as “site-centric.” Site-centric clickstreams can provide very detailed records of what visitors do when navigating and interacting with a given site. A key drawback, however, is that these data lack information regarding the activities of site users on other websites. They may also lack user-specific information such as demographic profiles. Server log files can record information on the visitor’s cookie ID and IP address, allowing the identification of unique users and return visits. (Though the use of cookies is hardly infallible, early research by Dréze and Zufryden (1998) established that their use does not pose a significant problem in practice). To give an analogy from the offline world of conventional retailing, site-centric data are akin to data captured from a store’s register receipts or shopper loyalty cards: they reveal the actions of users at the site (or store), but not what they do elsewhere.

Providers of syndicated Internet panel data, such as ComScore and Nielsen Net Ratings, represent an alternative data source. Internet panel data capture the Universal Resource Locators (URLs) of all pages requested during Web usage by transmitting this information from the user’s computer to the panel-data supplier. Clickstream data also can be collected by a user’s Internet service provider (ISP) or by Java Applets or Javascript code downloaded to the users’ computers. In that case, when users perform any activity online, the ISP, the Java Applets or the Javascript code can record it, thereby creating another source of panel clickstream data. All of these sources are known as user-centric, because they record activities for a panel of users across multiple (potentially competing) websites. A useful offline analogy is UPC scanner panel data for consumer products.

User-centric panel data offer the advantage of combining online behavior tracing across sites with information from each user. This has potential modeling and managerial advantages, as pointed out by Padmanabhan et al. (2001). On the other hand, user-centric data suffers from various limitations, many of which are related to potential sampling problems. Despite the very large number of panelists maintained by syndicated data
suppliers or tracked by ISPs (e.g., more than one million users), sample size issues can arise when analyzing activities at a single website. This is because the number of relevant observations can become small if the site has low traffic volumes or, in the case of e-commerce analysis, has a very low visit-to-purchase conversion rate. The representativeness of the panel sample and the churn that occurs within panels also create potential sampling-related issues. Finally, surfing activity is typically tracked at the machine level. This can create challenges in representing activity at the individual level when a panelist uses multiple computers or multiple individuals use a single machine. Though site-centric data offer the advantage of complete coverage of user activity at a site, such data can also suffer from the machine/user matching problem unless the site forces users to register or login.

In addition to sampling, another difference between user-centric and site-centric data is the level of detail captured about each page view. For example, in user-centric data, the URL corresponding to a page request is often truncated. This can make it difficult to match information recorded in the user-centric panel to specific page content or activities taking place on a website (e.g., a purchase transaction or the addition of items to a shopping cart). In sum, researchers and practitioners need to balance the strengths and weaknesses of each type of clickstream data when weighing how to proceed.

Clickstream data can also be collected in an experimental setting, recording the actions of subjects in the laboratory or in the field. In this paper we focus our discussion on clickstream data that has been collected from the natural operations of company servers (which can also be recorded by syndicated third parties). For research on laboratory or experimental clickstream data, the interested reader is referred to articles discussing pros and cons of clickstream data use in experiments (e.g., Bimbaum 1999; McGraw, Tew, and Williams 2000; Jonhson and Mick 2001) or reporting results from such experiments (e.g., Mandel and Johnson 2002, Moe 2006a).

Regardless of the way it is captured, a clickstream dataset tracks the activities of online users and records the virtual trail each user leaves behind while surfing the Web. In general, we can expect the raw data to include an identification of the computer or individual (e.g., cookie ID, IP address, or username), the type of browser used, the pages
requested or banners clicked, a time stamp for each activity, the previous URL visited, and other page specific variables (e.g., keywords) transmitted between the server and the user computer.

2.2 New Data for Old Problems

Clickstream data now permits the study of well-known phenomena in ways that were hitherto infeasible, too difficult, or costly. These advances are not specific to the online world, but have application in broader contexts (e.g., Johnson and Mick 2001). As an example, consider the catalogue and telesales businesses and compare these with a standard commercial website. The traditional businesses provided some of the advantages of current e-commerce websites (e.g., removing the requirement of physical proximity) and faced issues that e-commerce businesses now also confront (e.g., the need to process significant amounts of data, efficiently test the best stimuli, customize commercial messages and target specific audiences). The detailed data on what consumers do online (e.g., observing how consumers study a category or browse a catalog in its Web version) opens research opportunities into such topics as the formation of consideration sets (e.g., Wu and Rangaswamy 2003) and the modeling of decision making processes and stages (Moe 2006a). As a result, the availability of high-quality clickstream data has allowed research that can potentially aid traditional businesses, including bricks-and-mortar retail, catalogue, and telesales.

Clickstream data also provide a means for improving the measurement of audience size and characteristics in media. Because of the detailed electronic trace provided by clickstream records, potentially more accurate measures of reach (unique visitors) and exposure frequency (repeat visits) are available for Internet media versus many forms of traditional media. Interestingly, user-centric and site-centric data sometimes produce different metrics in this regard (e.g., Story 2007). The resulting discrepancies in visitor counts and page views are often attributed to the different sampling properties of the two data sources. Resolving these differences is, in itself, a potentially interesting topic for clickstream research.
2.3 New Data for New Problems

While “new data” for the research of “old” problems is clearly of interest, we believe that clickstream data has had (and will have) its greatest impact on “new” problems brought about by the Internet. Many of these new problems involve new types of targeting opportunities. For example, websites can respond quickly to the actions of individual consumers by providing information, in real time, likely to be the most relevant for that consumer. Websites can post banner ads that are dynamically created to include the current best travel deals from the location of a specific user, after quickly checking the latest data. Websites can limit the number of specific banner exposures each individual sees, and businesses can also change the content of their website pages given the path individuals have previously taken.

It is clickstream data that provide the detailed information on consumer interactions required for sophisticated targeting actions. For example, online product recommendations based on collaborative filtering are possible because clickstream data is unobtrusively and immediately collected and processed by websites while users navigate and search for specific products. Other types of customization that rely on individual parameters and their estimation (e.g., Ansari, Essegaier, and Kohli 2000) are also possible only with the detailed clickstream data collected by websites. For more details on personalization, both offline and online, please refer to AN01 (2008) and AN15 (2008).

While clickstream data do not provide every detail that researchers and practitioners might seek, they provide far more than the scanner panel data used in the development and testing of choice models from the early 1980s (Bucklin et al. 2002). This enhanced detail also goes hand in hand with increased complexity. While scanner data provided information on the purchase decisions made by panelists, clickstream data also track the path that the user follows to purchase. This information enables researchers to study more aspects of search and purchase behavior, but it also makes the size of the datasets much larger and challenges researchers to find ways to organize user activity into sensible units of analysis. It is possible that this may slow the pace of some advances, but it is also clear that the opportunities are rich and varied. We now turn to a
review of some of the key advances that clickstream data has made possible over the past 10 years.

3. Clickstream Research
The detailed records of Web usage behavior have allowed researchers working with clickstream data to pursue a wide variety of topics. We group these into three broad research themes. First, we discuss clickstream research into how people navigate the new medium. This includes studies of browsing behavior, website choice, and the extent and nature of search across websites. Second, we discuss how clickstream is shedding light on online advertising methods. This includes studies on banner advertising, paid search, and email. Lastly, we examine what clickstream research has shown about how people shop online and how online purchases can be predicted. This includes studies on purchase conversion, task completion before purchase, consideration sets, and online auctions.

3.1 Browsing and Navigation of Websites
Clickstream data enable researchers to study how users browse or navigate websites, how they respond to site design and structure, and how they move from site to site. We first look at studies focusing on behavior within a given website visit. We then discuss research on usage across websites.

Within-Site Browsing, Learning, and Site Design
One of the earliest clickstream analyses published is the study by Huberman et al. (1998) on site browsing behavior. The authors propose that an inverse Gaussian represents the distribution of page requests across visitors to a website. This is derived from the simple assumption that users will request an additional page view from a given site so long as the value of viewing the new page exceeds the cost of doing so. The authors show that the proposed distribution fits the behavior observed in the clickstream across several different websites and also could be useful in predicting the number of page requests across visitors. The study highlights the idea that a cost-benefit perspective may be useful for understanding and predicting individual behavior on the Internet. On the other hand, Huberman et al. did not incorporate the effect of any covariates.
(marketing or otherwise) and did not allow for user navigation behavior to change over time due to, for example, possible learning effects.

The idea that Web users might learn how to use a given site more efficiently with experience was explored in detail by Johnson et al. (2003). Their major finding was that visitors spend less time per session the more they visit the same website. This was formalized by applying the power of law of practice model to clickstream site visit and duration data obtained from Media Metrix (now part of ComScore). Johnson et al. reported that as the number of visits to a website doubled, the time spent on the site during a visit dropped by an average of 19 percent. This suggests that users become more efficient in their time allocation as they return to a website, probably as a result of learning how to navigate the site and becoming familiar with its content.

The total time a user spends in a site session is a function of both the number of page views and the durations of those page views. Thus, one question that the Johnson et al. study raises is whether the learning or efficiency gains they identified come from fewer pages being viewed or less time per page or a combination of the two. In their study of browsing behavior using site-centric data from an automotive website, Bucklin and Sismeiro (2003) showed that users changed their behavior in a manner consistent with learning effects. In particular, the study found that repeat visits by the same user led to fewer page views but had no effect on page view duration. These results corroborate the findings reported by Johnson et al. (2003) but show, at least for the website under study, that the reduction in session duration is the result of fewer page views and not less time spent viewing each page. Bucklin and Sismeiro also found evidence for within-site “lock-in” (i.e., as users requested more page views within a session, the site became progressively “stickier”) and that user behavior was consistent with rational considerations involving time constraints and cost-benefit trade-offs (corroborating Huberman et al. 1998).

An e-commerce site, seeking to speed users from visit to purchase transaction, might understandably benefit from fewer page views and/or shorter visit durations. On the other hand, media and entertainment sites that rely on advertising support may have different goals. Management at such sites may seek to maximize the number of page views users see so as to increase the number of advertising impressions served. If sites
spread content across page views to do this, then page view counts and page view
durations could be misleading indicators of usage. In these cases, it may be more
appropriate to rely on visit duration times to gauge site usage.

The factors affecting website visit duration were investigated in a study by
Danaher et al. (2006). The authors analyzed clickstream data from a NetRatings user-
centric dataset covering the top 50 websites visited by panelists in New Zealand. The
study found that variability in visit duration is driven mostly by the situation and, only to
a smaller degree, by the traits of the individual or the fundamental aspects of the website
itself (including text, graphics, advertising content, and functionality). These findings
provide an interesting counterpoint to prior work which had suggested that behavior
follows regular patterns and that design issues might play an important role.

The foregoing studies were not specifically set up to examine website design.
One reason for this is that clickstream data lack detailed information on the visual
components of the site. To link the clickstream records of page views to design and
content elements, researchers need to separately collect this information. Sismeiro and
Bucklin (2004), for example, augmented their clickstream dataset with ratings from two
independent judges who classified each website page on a set of design characteristics
expected to influence the visitor’s experience with the site (e.g., clutter, dynamic content,
presence of figures, number of links, and page size). While augmenting clickstream data
is one approach, another is to conduct controlled experiments that manipulate site design
features and content (e.g., Mandel and Johnson 2002, Yoo and Kim 2005, Lam, Chau,
and Wong 2007).

Patterns of Within-Site Browsing

In addition to visit duration, page requests and page view duration, it is also
possible to examine detailed browsing records to identify different patterns of site usage
across visitors. Using site-centric data from an online retailer, Moe (2003) classified Web
pages by type (e.g., home page, types of product information, purchase, etc.). She then
developed a series of user-specific site usage metrics (e.g., number of page views,
average time per page) and conducted a cluster analysis. She found that visitors can be
effectively clustered into one of four browsing strategies: directed buying,
search/deliberation, hedonic browsers, and knowledge-building. Of the meaningful user
visits (i.e., sessions representing more than a perfunctory visit), the clusters accounted for 15, 14, 66, and 5 percent of the site visits, respectively. Each of these strategies was associated with different navigational patterns by users on the site and different propensities to purchase from the site. For example, users in the directed buying cluster were most likely to make a purchase while users in the knowledge-building group were least likely.

Browsing behavior patterns were also the focus of a modeling study conducted by Montgomery et al. (2004). Examining user-centric clickstream data from Media Metrix, the authors focused on activity at the Barnes and Noble website. Like Moe (2003), they classified pages on the site into several different page-type categories (e.g., home page, information, product, shopping cart, order etc.) but then went on to study the transitions users made from page-type to page-type as they navigated the site. Modeling these transition choices, the authors found that navigational behavior was best explained by allowing for two modes of site usage: deliberation and browsing. Their model also allowed users to switch between these two modes during a single session. The empirical results established that the dynamic model represented browsing behavior at the site better than a comparable static model, indicating that at least some users change their goals while in the midst of a site visit.

**Browsing and Searching Across Websites**

In addition to within-site browsing behavior, a number of studies have employed user-centric clickstream data to investigate browsing and search across multiple websites. One of the advantages of user-centric data is that visits to multiple websites are recorded for each user. This opens the possibility that what users do at one website might help predict behavior at competing or complementary websites. This idea was explored by Park and Fader (2004) who studied Media Metrix data for two competing sites in each of two product categories, books and CD’s. They develop a stochastic timing model of cross-site visit behavior which accounts for the potential correlations in visit behavior between the competing sites. Their findings demonstrate that information in the visit pattern to another site improves the prediction of future visit behavior at a given site. In particular, the authors show how this can be used to forecast when a user might make their first visit to a website, given that they are already visiting competing sites.
Though cross-site visitation across competing sites is clearly evident, one of the important substantive findings from clickstream research is that the overall extent of cross-site search is low. Indeed, clickstream-based studies have found Internet users to have substantial site loyalty and high switching costs. In one key study, Johnson et al. (2004) analyze clickstream panel data from more than 10,000 Internet households who shopped for commodity-like products (books, compact discs, and air travel services). The authors report that shoppers search across very few websites in a given shopping month, with about 70% of the CD and book shoppers loyal to just one website. Similar results were also found for the travel category. This finding is also supported by Smith and Brynjolfsson’s (2001) study of an Internet price-comparison service or “shopbot.” These authors found that the three most heavily branded online book retailers were able to hold a significant price advantage over others in obtaining clickthrough to their sites from the shopbot.

Several additional studies also give further evidence of online users’ high switching costs. Goldfarb (2006a) investigated Internet portal choice with user-centric clickstream data. He finds that users exhibit switching costs with respect to portal choices, and that the loyalty generated as a result of these costs drive a large fraction of portal visits and generate considerable revenues. In a follow-up study, Goldfarb (2006b) looked at cross-site usage in the first days following an Internet denial of service attack. For websites not affected by the attack, lock-in (i.e., the tendency to remain with a site unless forced to go elsewhere) accounted for most of the gains in their traffic. However, lock-in had little long-term impact on the performance of these sites as, over time, users returned to their steady-state patterns of search and browsing. Chen and Hitt (2003) looked at clickstream data from online brokerage firms and, like Goldfarb, also found significant switching costs. These costs also varied substantially across individuals, by as much as a factor of two. Interestingly, most of the variation in estimated individual-level switching costs could not be explained by customer demographics. Instead, system usage measures, together with firm characteristics such as product line breadth and quality, were strongly associated with reduced switching.

The site loyalty levels documented in cross-site clickstream research indicate that “lock-in” (e.g., due to learning and other cognitive switching costs) is indeed an
important factor in everyday Internet usage. This suggests that firms may further enhance customer retention in the online channel by introducing new “frictions” even as traditional ones (e.g., the costs involved in conducting offline search) have been removed. Examples include frequent-purchaser programs, use of user profiles for personalization, “clickthrough” rewards, affiliate programs, and the provision of online product recommendation systems. Others have suggested that online retention is influenced indirectly through engaging website design (Novak et al. 2000) and strong branding (Smith and Brynjolfsson 2001), which might serve as a proxy for retailer credibility in non-contractible aspects of the product and service bundle, such as shipping reliability.

3.2 Advertising on the Internet

Much of the Internet is sustained by revenues from the sale of various forms of advertising. Indeed, Internet advertising has now become an important part of the marketing mix of most companies, with spending in the U.S. growing from $4.6 billion in 1999 to $16.9 billion in 2006 (PricewaterhouseCoopers 2006). Clickstream data offer the capability to track exposure to Internet advertising along with the user’s subsequent actions (e.g., a clickthrough or a purchase), offering the prospect of a tight link between exposure and the user’s behavioral response.

In terms of spending, the two major categories of online advertising are display advertising (so-called “banner” ads) and paid search advertising (the text ads served as sponsored links by search engines such as Google and Yahoo!). So far, existing clickstream research has focused primarily on banner-type advertising, but work is now underway in paid search. Clickstream-based research has also been conducted on email as an advertising medium. More recently, online word-of-mouth and recommendation systems have become important topics, but clickstream-based studies of these vehicles are yet to emerge in the literature. We now discuss the variety of advances in clickstream research which have occurred in the area of Internet advertising.

Banner Ads

When an Internet user is exposed to a banner-type advertisement, one measurable response is whether or not the user clicks on the ad. Such a clickthrough typically takes the user to another site, usually where more detailed information about the product or service can be found. Chatterjee et al. (2003) model banner ad clickthrough using an
individual-level binary logit (click/no-click formulation). Using site-centric clickstream data, they model the probability that the user clicks on the banner, given that the ad has not yet been clicked by that user during the current session. They find significant heterogeneity in (unobservable) click proneness across consumers and find also that repeated banner exposures increase the clickthrough rate when consumers are less click-prone. They also find that consumers are equally likely to click on banner ads placed early or late in the navigation path.

Though clickthrough rates have been a popular measure of banner ad effectiveness, they are not without problems. Issues include fraud and the dramatic decline in clickthrough rates observed since the early days of the commercial Web (e.g., DoubleClick 2003). Despite this trend, recent growth in spending on banner ads suggests that advertisers believe in the value of the advertising format. Part of this may come from other significant outcome variables (some not measurable in clickstream data) that may be affected by exposure to banner advertising. For example Dréze and Hussherr (2003) suggest banners could be processed at a pre-attentive level and that brand awareness and recall could provide a better measure of banner effectiveness than clickthrough. In a study based partly on clickstream data, Ilfeld and Winer (2002) corroborate this notion. They report that online advertising is effective in attracting site visitors and that it also impacts site awareness and brand equity.

If exposure measures to banner advertising become the desired metric, traditional advertising touchstones such as reach and frequency are likely to be of practical interest on the Internet. This raises the question whether advertisers will be able to assess and predict the reach and frequency of impression-based ads on the Internet. Danaher (2007) developed a stochastic modeling approach designed to represent the reach and frequency of page view exposure across websites. The approach was applied to user-centric data from Comscore and produced a variety of insights regarding how page views from different websites can be decomposed into reach versus frequency. We note also that frequency and reach are of such importance to online advertisers that today ad-servers (like the one provided by DoubleClick) allow advertisers to control who sees the ads (via demographic and behavioral targeting) and the number of exposures each user will receive of each banner.
In addition to clickthrough and page view exposures, another approach to assessing banner advertising is to look at the possible effects it may have on e-commerce purchase transactions. Manchanda et al. (2006) use individual-level data on user exposure to the banner ads sponsored by an e-commerce website matched to purchase transactions. Using a hazard modeling approach, the authors investigate purchase timing as a function of banner ad exposure. They find that banner advertising accelerates the timing of purchases made by existing customers. More specifically, the number of ad exposures, the number of websites where the exposure occurred, and the number of pages on which a consumer is exposed to the banner advertising all had a positive effect on the repeat purchasing of existing customers.

In addition to inducing clickthrough (albeit at low levels) and purchase acceleration, exposure to banner ads may also affect the subsequent browsing behavior of consumers within a website. Rutz and Bucklin (2007a) study how exposure to banner advertising at an automotive website changes the ensuing page view choices made by visitors. Using individual-level site-centric data, they model the page-specific choices for product-related information that users make following banner ad exposure. They find that banner ads served in the current session have segmented response effects on subsequent browsing activity across users. In one segment, the effect is positive (ad exposure leads to page requests for more information about the brand), in a second segment the effect is negative, and in a third segments the banner ads have no effect.

**Paid Search**

Paid search is a service offered by Internet search engines in which the advertiser selects specific keywords and creates a text ad that will appear when a user searches for these keywords. It has been the main driver of Internet advertising growth over recent years (growing between 26 and 174% yearly from 2003 to 2007; eMarketer 2007). This growth is probably due to the unique opportunities that paid search and search engine marketing offer. Search engine users are actively looking to satisfy a goal and by providing information that might help to fulfill that goal, paid search qualifies as direct as well as ‘just-in-time’ marketing. Additionally, search engine traffic originates from a voluntary, user-driven activity rather then from a forced exposure.
Despite the significant (and recent) importance of paid search, and the success of firms such as Google and Yahoo, there is as yet no published academic research in marketing studying paid search using clickstream data. Economists have become interested in the properties of the online auctions for text ad placement for search terms (or so-called key words) where sponsors bid against each other to obtain higher positions in the display of search results to the user (e.g., Ostrovsky, Edelman, and Schwarz 2007), but this work is theoretical and analytical. There is also an extensive practitioner literature on search engine marketing, consisting primarily of shorter “white papers” (see, for example, www.sempo.org).

One example of early academic work in marketing is a study by Rutz and Bucklin (2007b). Using daily data on paid search activity for a lodging industry company in the U.S., the authors investigate the role of branded versus generic search terms. A branded search term includes a company brand name (e.g., “Hilton Las Vegas”), whereas a generic search term does not (e.g., “Las Vegas Hotels”). The cost per click pertaining to generic search terms runs substantially higher than for branded terms, yet generic terms are associated with a much lower rate of user conversion from click to purchase (in this case, a lodging reservation) than branded terms. In their paper, Rutz and Bucklin develop a model of advertising dynamics in which the effects of generic search are allowed to “spill over” into effects on subsequent branded search. Using daily data from the Google and Yahoo search engines, the study finds that the spill over effect is large enough to justify the premium placed on generic search terms. In a follow-up study, Rutz and Bucklin (2007c) model the performance of individual search terms as opposed to groups of these terms.

Email

Another important advertising tool on the Internet is email. Companies send emails to registered users or simply to potential customers for whom an email address is available. These outbound communications may contain text, editorial content, and graphics designed to promote a site, attract visitors, and sell products or services. Typically a link or set of links is contained within the email, enabling the user to move directly to the site (or purchasing pages). Though emails are inexpensive to send on a
variable cost basis, the wrong content can make email ineffective and, worse, lead to its classification by the user a “junk email.”

Given the differences across users in potential interests, an important question is whether or not customizing outbound emails might be effective in increasing clickthrough rates. This question is comprehensively addressed by Ansari and Mela (2003). They develop a two-phase approach to customizing email communications using individual-level clickstream data. The first phase specifies and estimates a model of clickthrough probability as a function of the content and design features of the email. The second phase uses the probability model parameter estimates as input into an optimization module which, in turn, recommends an email configuration customized for the recipient and time. The authors report that a 62% improvement in the expected number of clicks can be produced. While this study focused on email, the combined response and optimization modeling approach developed by Ansari and Mela could also be applied to other problems involving customization of online content via clickstream data analysis.

One aspect of outbound email management not directly addressed by Ansari and Mela (2003) is the frequency with which the firm should send out emails to its users. The very low marginal costs of email communication need to be balanced against the potential for wear-out—or opting out—from receiving additional emails from the site. While no published paper specifically addresses this problem, ongoing work first reported by Bonfrer and Dréze (2006) provides a first look at capturing these issues. For additional discussion of issues in online media and advertising, please see AN09 (2008).

3.3 Online Shopping and E-Commerce

One of the most active areas of clickstream research has been the understanding and modeling the online purchase behavior of visitors to e-commerce websites. Much of the modeling work has focused on whether or not a site visitor will complete a transaction and the factors which might predict this. Several different approaches to this problem have emerged, with each using clickstream data in various ways. Besides purchase conversion, clickstream data have also played an important role in understanding online auctions as well as the role of shopbots.
Purchase conversion has been modeled with clickstream data using several approaches. One such approach to predicting online purchase, given a site visit, has used stochastic models. This method has been described and tested in articles by Moe and Fader (2004a, 2004b). They develop stochastic models of evolving visit behavior to e-commerce sites (e.g., CDNow and Amazon) using Media Metrix panel data from 1998. Among their modeling results, they show that users who visit a retail site more frequently have a greater propensity to buy, and that the evolution in individual-level visit frequency is informative of which customers are more likely to buy online. The flexible models proposed by Moe and Fader predict purchase as a function of prior visit behavior. This parsimony means that implementing the models requires only minimal data to be extracted from the clickstream (i.e., the occurrence of visits and purchases for users over time).

A limitation of the stochastic modeling approach is that it is not set up to examine the actions of users while browsing a site (e.g., page view requests) and the possibility that those actions could influence or be associated with purchase outcomes. Doing this requires the analyst to delve deeper into the clickstream, developing measures of within-site browsing behavior and relating them to purchase conversion. One example of this approach is the model proposed by Sismeiro and Bucklin (2004). The model breaks down the purchase process into a series of tasks which users must complete in order to buy. Studying site-centric data from an Internet car retailer, the authors modeled three user tasks: (1) completion of product configuration, (2) input of personal information, given (1), and (3) order confirmation with credit card provision, given (1) and (2). (More generally, these steps follow the familiar sequence of e-commerce tasks of placing an item in the “shopping cart,” entering shipping information, and placing the order with a credit card.) The clickstream data showed that only about two percent of site visitors completed an order transaction. When decomposed by task, the purchase process looked as follows: 30 percent of visitors to the site completed task (1), 20 percent of those users then went on to complete task (2), and 34 percent of those completing tasks (1) and (2) went on to complete task (3).
One of the interesting empirical findings was a set of significant sign reversals for many of the covariate parameters from one task to another. For example, exiting and returning to the site was not predictive of task one completion (product configuration), but it was positively related to task two (personal information) and negatively related to task three (order and payment). The results also showed that the number of site visits, per se, was not predictive of purchase. This finding differs from some of the Moe and Fader results, where they documented that more site visits lead to greater purchase likelihood. The discrepancy may be due to the differences in shopping behavior for books and CDs versus new cars. For example, knowledge building and hedonic browsing may be higher for new cars than books, so that many users often return to the site multiple times without ever purchasing.

The Montgomery et al. (2004) modeling approach, discussed earlier in Section 3.1, also can be used to predict purchase conversion. The authors show how the model can provide an improving dynamic forecast of purchase likelihood as the user proceeds through page views on the site. For the Barnes and Noble clickstream data in their study, the mean purchase rate was 7 percent. Using their model, after one page view 13% of the visitors who end up as purchasers were correctly classified, almost double the baseline rate. After three page views this grows to 23% and after six page views it is 41%. Thus, the study highlights how the information contained in the clickstream records of browsing behavior can be used to predict subsequent online purchase.

**Consideration Sets**

While clickstream datasets record purchase, they also record user activities undertaken prior to purchase, such as the use of decision aids. This means that clickstream data can provide a window into the purchase process as online shoppers form consideration sets and then make final selections. Using data from the online grocer Peapod, Wu and Rangaswamy (2003) model consideration and choice for liquid detergent. In their data, shopper use of two online decision aids (a personal list and a sorting capability) was available along with prices, promotions, and product characteristics. Estimating their model, the authors found that the shoppers formed consideration sets in different ways. In a two-segment analysis, they report that one
group, “seekers,” actively used online search tools while a second group, “non-seekers,” relied more heavily on personal lists shoppers kept on the Peapod system.

In addition to the use of decision aids, clickstream data can also provide information on the products viewed by the user. Taking advantage of this, Moe (2006a) modeled product viewing and choice for two product categories (weight loss aids and meal replacements) at an online retailer of nutritional products. (Product viewing during a site visit does not necessarily indicate consideration, of course, but the decision to view a product is likely to be closely related to it.) In Moe’s model, the probability of viewing a given product option in the category is modeled as a first stage and the purchase decision as a second stage. She finds that the two-stage model provides better predictive validity than a single-stage approach. Moe’s model is also useful for examining the role of different product attributes in the viewing phase versus the purchase phase. For example, she finds that fewer attributes are used in the first stage than in the final stage. In particular, price and product size tend to be used in only a single phase while attributes for product ingredients tend to be given weight by shoppers in both phases. Applying the model more generally could prove useful for site design and targeting of online interventions such as promotions or product displays.

*Auctions and Alternative Pricing Mechanisms on the Internet*

Lower transaction costs offered by the Internet along with new possibilities to interact with customers online have enabled new pricing mechanisms to flourish on the Internet. These mechanisms let buyers actively participate in price discovery and offer sellers a chance to increase sales by means of price discrimination and the attraction of new customer segments. Some of these pricing mechanisms, such as various online auction formats stem from the offline world but have grown quickly on the Web. Other mechanisms, such as “name-your-own-price” systems, were previously unknown.

Because of the unobtrusive way it is collected, clickstream data have made possible a wealth of empirical research on these pricing mechanisms. A good example is the work on so-called “sniping” or late bidding in auctions (bids that arrive during the last seconds of an Internet auction). Late bidding is now an accepted empirical regularity for most Internet auctions. Ockenfels and Roth (2006) find in a sample of 240 antique
auctions on eBay, that 89 auctions had bids in the last minute and 29 in the last 10 seconds. Wilcox (2000) also reports similar findings.

Access to a significant volume of clickstream data from real online auctions has allowed many theories of late bidding to be tested (e.g., Bajari and Hortaçsu 2003; Roth and Ockenfels 2002). These theories include tacit collusion, the response of experienced bidders to naïve bidders, and multiplicity of auction listings for the same product. For example, Zeithammer (2006) provides initial evidence that forward-looking bidders who participate in a sequence of auctions for substitutes products reduce their bids in anticipation of future auctions. Thus, the author demonstrates that in the online auction marketplace, useful information about other (future) purchase opportunities is available, and this information enters current observed demand, effectively blending elements of simultaneous choice among several purchase opportunities into the underlying sequential search. Though definitive conclusions have yet to emerge, clickstream data has enabled significant advance in this area (see also Bajari and Hortaçsu 2004 for a review).

As we discuss in Section 2 above, part of the potential contribution from clickstream data is to provide new data for old problems. With respect to auctions, a good example is information asymmetry. Asymmetries can lead to the “winner’s curse” phenomenon, among other problems. Using clickstream data from online auctions, researchers have found that bidder behavior is consistent with fear of the winner’s curse. This often leads to depressed auction prices and, in particular, occurs even in auctions where information asymmetries should not play a major role (e.g., Bajari and Hortaçsu 2003).

One important source of information asymmetry on online auctions is buyer and seller anonymity. For example, eBay does not require users to report their actual names or addresses; only eBay IDs are revealed. There are also very few repeat transactions between buyers and sellers, which prevents learning via direct experience. Resnick and Zeckhauser (2001), using a large dataset from eBay, report that fewer than 20% of transactions are between repeated buyer-seller pairs within a five-month period. To ensure honest behavior, online auction sites rely on voluntary feedback mechanisms, in which buyers and sellers alike can post reviews of each others’ performance.
Because buyer and seller reviews are available along with bidding and transaction activity, empirical study of these feedback mechanisms in online auctions has been possible. Bajari and Hortaçsu (2004) and Resnick et al. (2003) review these studies and conclude that an “established” eBay seller (i.e., someone with hundreds or thousands of largely positive or neutral feedback comments) could enjoy a 10-12% price premium when compared to a seller with no track record. Zhang (2006) reports that negative selling comments are weighted more heavily than positive ones; it is also important to distinguish between buying versus selling feedback for each individual because each type of reputation can yield a different impact. Zhang also finds that buyers prefer to bid on auctions by more trusted sellers and that when they do so they also tend to bid above the starting prices or reserve prices.

Finally, in two recent articles Park and Bradlow (2005) and Bradlow and Park (2007) have applied sophisticated modeling to notebook auction data that explicitly captures the critical features of bidding behavior established in the literature. The authors provide valuable tools for managers at auction sites to conduct their customer relationship management efforts and to evaluate the quality of the listed auction items and of potential bidders. In addition, the authors find that larger bid and time increments significantly influence the bidding behavior of the remaining bidders and that the number of latent bidders varies considerably across auctions.

While auctions have captured most of the empirical research on new pricing mechanisms, several studies have been conducted with data collected by “name-your-own-price” websites. Findings suggest that consumers do not follow rational decision making processes (Spann and Tellis 2006) and that frictional costs are substantial (Hann and Terwiesch 2003; Spann, Skiera and Schäfers 2004). In general, these results tend to corroborate the spirit of findings from fixed price e-commerce websites and studies on multiple site search.

Shopbots

A website purchase decision might be modeled as a binary outcome conditional on the visit to an e-commerce site. However, users may also consider online purchases across multiple sites. Researchers have studied part of this choice process by examining the choice decisions that users make at “shopbots” – 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. (2004) 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. They demonstrate the validity of their model using observed prices at online bookstores over a six-month period.

4. The Future for Clickstream Analysis: Opportunities and Challenges

Unlike traditional media, the Internet is both digital and interactive. Because it is digital, new business concepts have been made possible (e.g., the digital download of entertainment) and new opportunities for data collection and analysis are available. Because it is interactive, companies have an expanded range of marketing opportunities and contact tools, including individual-level targeting and customization, chat rooms, networks, and blogging. Indeed, an established trend since the inception of the Internet is the growth of dialogs and interactions (e.g., Shankar and Malthouse 2007). These are not only between the firm and the consumer, but also among consumers. MySpace.com and Facebook.com provide two well-known examples of conversational platforms, and their success reveals the power of online interactions. In this section we discuss some of the untapped research opportunities in the analysis of clickstream data. We also note some of the challenges that using these data still pose to researchers.

4.1 Opportunities

The analysis of clickstream data should provide major opportunities for advancement in several fast-growing new areas of the Internet. These include interactivity in the form of word-of-mouth (WOM) and social networking, managing multiple channels, and
recommendation systems. Clickstream research is also likely to lead to further methods developments in marketing as investigators work to overcome various challenges in using the data to understand Web usage, advertising effects, and e-commerce.

Interactivity, WOM and Social Networking

Though interactivity is one of the basic features of the Internet (individuals can communicate and interact with each other, with companies, and with websites at a level previously unseen), this feature has also revealed itself as one the most difficult to study, especially using clickstream data collected in the user’s own environment. A good example is the study of online word-of-mouth (WOM) communications—communications originated from customers or users towards others in which the firm does not directly control the content, timing, or nature of the message.

The Internet has created not only a new forum for WOM it has also revitalized the interest in this phenomenon. The Internet has lowered the costs of passing WOM and expanded its reach. Blogs, chat rooms, social networks and other forms of user generated content are growing quickly and user access to them is simple and easy. Search engines also facilitate rapid collection of thoughts and opinions from people and contexts to which one would otherwise not be exposed. Finally, commercial websites now commonly allow consumers to post feedback, ratings, and comments about almost anything related to their business, institutionalizing WOM inside their own environment.

The challenges in using clickstream data to study WOM can be significant. It is often not easy to track WOM online and maybe even harder to connect it to consumer transactions or other behavior. Nevertheless, some initial studies are reporting progress. For example, Chevalier and Mayzlin (2006) study the impact of online book reviews on the relative online sales of a given book at Amazon versus Barnes and Noble. They show that changes in the number of book reviews and the nature of those reviews are likely to be associated with online sales. Interestingly, their study is not actually based on clickstream data per se, but relies instead on a database developed by the authors from publicly available information reported on the two websites. Managers operating internally, however, should be able to use their own clickstream data to conduct similar investigations and could readily track the product review metrics proposed by Chevalier and Mayzlin.
Another investigation of online WOM is being pursued in a paper by Trusov et al. (2007a). Using clickstream data from an Internet social network site, the authors track the number of outbound email invitations sent by users to invite others to join the network. This is paired with data on the number of new members who sign up. Though the two data streams cannot be linked at the individual-level, an aggregate time series analysis reveals that the outbound WOM significantly affects the number of people who subsequently sign up to join the network. Thus, if researchers can track online WOM and connect it to sales or other outcome variables, clickstream-based research may be able to provide new insights into the effectiveness of this form of marketing communications.

Once users have joined a social networking site, the firm’s interest turns to generating revenue from the activity of users on the site (e.g., by the sale of banner ads on the site’s pages). Understanding what drives higher or lower usage of a social networking site at the member level is therefore a potentially important area of research in this domain. As in the case of WOM, clickstream-based analyses should be able to provide researchers and practitioners with insight into the question. Initial study on this topic has been carried out by Trusov et al. (2007b), who model the site usage behavior of a sample of members at a social networking site. The study shows how firms can identify members who are disproportionately influential in bringing others to the site, thus opening the door to possible targeting and customization opportunities. More discussion of online social networks is also provided in AN05 (2008).

Managing Multiple Channels

Firms are increasingly selling through multiple channels and contact points. Companies use websites to drive people to their retail stores and to also to sell online. Retailers that previously sold primarily through prime-location bricks-and-mortar stores now have websites and discount outlets. In addition, consumers may use the Web to gather information, but use other channels to make purchases. Being able to connect the actions of users across the various channels a firm operates, and to study how to best leverage the strengths of each one, would represent a significant advantage for firms. (For a review of the literature in this area, please see AN07 2008.) So far, there is very little work on this topic using clickstream-based data. One exception is the study by Ansari et al. (2008), which studies the migration of one firm’s catalogue customers to its new e-
commerce site and the role of marketing activity, including outbound emails, in influencing this migration.

**Recommendation Systems**

While creating opportunities for consumers, the rapid growth of the Internet and proliferation of websites also creates new difficulties. In particular, how can consumers find the relevant information, website, or product online when millions of similar items co-exist on the Web? Indeed, these difficulties are part of the reasons why price dispersion and economic frictions that were thought to be irrelevant on the Internet (e.g., Brynjolfsson and Smith 2000) have turned out to be so persistent. To help users, a multitude of intelligent agents and recommendation systems have been developed (e.g., Ansari et al. 2000, Iacobucci et al. 2000, Adomavicius and Tuzhilin 2005). The use of these systems is a very important topic in industry (e.g., in 2007 Netflix publicly offered a one million dollar prize for improving its movie recommendation system) and clickstream-based data is used internally by companies to implement these systems. Thus far, academic work in marketing has focused on experimental findings on decision making (e.g., Häubl and Trifts 2000), the analysis of ratings (Ansari et al. 2000, Ying et al. 2006) and purchase histories (Bodapati 2008). Clearly, this should be a vital topic for future clickstream research.

**Methods Development**

Due to the challenges and opportunities in clickstream analysis, researchers have responded with methodological advances which have had an impact not only on Internet-related work but on research in marketing more generally. One example of this is the study by Goldfarb and Lu (2006). The authors show how individual-level regression analysis -- not applying Bayesian shrinkage methods or random effects approaches -- can provide a better understanding of online behavior than other methods (in this case, the drivers of Internet portal choice). A second example is the work of Telang, Boatwright, and Mukhopadhyay (2004). They enhance existing stochastic inter-purchase time models by allowing for purchase periodicities and unobserved heterogeneity in a proportional hazards mixture model. The authors used search engine visits data to highlight the benefits of the proposed model. A third example is how Sismeiro and Bucklin (2004) show that rare binary events can be predicted better by decomposing what users do online
into sequential tasks. By modeling a chain of conditional binary probabilities instead of a single event, predictive power was increased and insightful findings extracted. Additional examples of methods advances from clickstream research include Goldfarb (2006b), Jank and Kannan (2006), Park and Fader (2004), Moe and Fader (2004a,b), Montgomery et al. (2004), and Gatarski (2002). We believe that methods development will remain a promising area as researchers continue to grapple with the challenges inherent in using clickstream data.

4.2 Challenges

Despite the potential clickstream data analysis holds for research and practice in marketing, these data are almost surely being underutilized. In the fields of computer science and computer mediated interactions, the number of published research studies based on clickstream data is astounding when compared to the number in marketing. Managerially, a principal use of clickstream data is to track site traffic across websites, the average length of time that a user spends at a given website, and the demographics of website visitors. These audience metrics, while interesting, do not come close to fully exploiting the wealth of information that is available. On the other hand, it is important to recognize that there are technical factors which pose challenges in collecting and analyzing clickstream data. Finding ways to address these challenges could help expand the use of clickstream data analysis in marketing.

The Volume of Data

Working with clickstream data confronts researchers with the need to store and process very large amounts of information. Some early studies relied on site-centric raw log file data, which required intensive pre-processing before a usable dataset could be obtained. Fortunately, more recent papers have been able to rely on clickstream data collected directly by data warehouse and database systems that can dramatically cut pre-processing work. Despite these advances, researchers must still develop models and estimation methods than can cope with the size of the datasets and the dimension of the problems.

Scalable algorithms have been an important area of research in computer science but not a focus in marketing. With tens of thousands of users being monitored in the clickstream (even millions in some cases), the idea of estimating individual-level
parameters can be a daunting task. In fact, the individual-level studies using clickstream data in marketing have tended to rely on small samples relative to the actual number of users tracked (e.g., Moe 2006a studies 150 buyers; Bucklin and Sismeiro 2003 analyze 5,000 site visitors). In so doing, however, researchers could be losing the attractive sampling properties and insights that easy access to online behavior of very large numbers of Internet users can provide.

Depth and Quality of Clickstream Data

Though companies can quickly collect a very large volume of data tracking their individual visitors, the data collected often lack key pieces of information needed for marketing analysis. For example, sales data on margins might need to be integrated with clickstream data for measuring design effectiveness and return on investment. Call center data should be paired with clickstream data to understand customer behavior across multiple channels, to allow for proper sales attribution (e.g., when customers shop on the Web and purchase over the phone), and to permit service representatives to provide customized interactions. Demographic data from market research companies could be integrated with other profile and behavioral data to help segment consumers for marketing campaigns.

As a result, some studies in marketing have combined clickstream data with other data sources. For example, Ilfeld and Winer (2002) use information from a variety of sources to estimate the model on data for 88 Internet companies. Media Metrix panel data provided website visits, past website visits, and page views. Advertising was measured by tabulating the spending for each company on both online and offline media (radio, television, cable, outdoor, newspapers and magazines). The number of links from other websites to the focal site was determined from the Internet search engine, Google.

Another issue is that server-based tracking can be inaccurate. Servers can fail to account for Web-page requests that are cached by the user’s computer or by proxy servers, might not clearly identify users (e.g., due to the use of different computers with different cookies), and can lose part of the information in their communication with the client computer. (Lost “data packets” are not uncommon in the Internet and pose challenges, for example, for the tracking of banner clickthrough and other online performance metrics.) Finally, what servers register may not be what is seen by users and
information might need to be recomposed into page views, which could require specific algorithms.

The foregoing problems make site-centric clickstream research complex and challenging for marketers. A partial solution may be to rely more on user-centric clickstream panels. As noted above, however, these datasets can have issues of their own, including low sample sizes for unpopular sites or few counts for certain key within-site activities (such as a purchase transaction), even at popular sites. Because both types of clickstream data record detailed records of user behavior online, privacy concerns are, of course, a potential issue. If laws, regulations, or even industry norms limit access to clickstream data, progress could be seriously impeded. (Recently, Jerath, Fader, and Hardie (2007) proposed an approach using aggregated data to circumvent at least some privacy concerns.) Clearly, more research is needed to understand the significance of these problems and how to resolve them. Please see AN04 (2008) for a discussion of research on information privacy.

Automated Analysis Systems

Empirical research is traditionally a very labor intensive task in marketing, economics and statistics. Part of the answer to the clickstream data challenges might include the use of more automated systems able to handle the raw data and to automatically restore missing information or build a better image of what users do online (e.g., Ting, Kimble, and Kudenko 2005; Cooey, Mobasher, and Srivastava 1999). To take one example in the study of site design issues, it could be useful to develop new models using automatic systems to classify Web pages and characterize the visual stimuli people are exposed to while browsing. The computer science community has already developed similar systems for the automatic processing of images and used some of these systems to improve targeting of mobile phone commercial messages (e.g., Sismeiro et al. 2007 and Battiato et al. 2007). Another fruitful area for development of automated systems is in the analysis needed to administer real-time marketing activities. These could include, for example, real-time site customization and bidding in keyword search auctions.

5. Conclusion

Many academic fields including marketing share a keen interest in studying the Internet and e-commerce. In order to do so empirically, detailed data chronicling what transpires
in this domain is required. Fortunately, this is available to both researchers and practitioners in what has come to be known as clickstream data, the electronic records of what users do on the Internet and to what they are exposed. The purpose of this paper has been to provide an introduction to the use of clickstream data for addressing problems of interest to marketing.

Clickstream data can be collected and compiled in a variety of ways. For example, individual websites can maintain detailed records of the interactions between their site and their visitors. Such “site-centric” datasets can provide very detailed looks at what happens within a given website, but because they are collected by individual firms, there is little in the way of data about what those visitors do elsewhere. Alternatively, clickstream data also can be collected by tracking the Internet activity of participating panelists. Such “user-centric” data offers the advantage of covering all of a users Internet site visits, including those to competing sites, but can pose challenges with respect to sample sizes in some instances and the ability to precisely track the user’s actions within a given website of interest. In either case, the detailed records provided open a wealth of opportunities to advance the understanding of online behavior and to improve website performance.

Over the past ten years, significant advances directly relevant to marketing have been made in the analysis of clickstream data. These advances can be grouped into three broad categories: (1) website usage and navigation, (2) advertising on the Internet, and (3) online shopping and e-commerce.

In the first category, usage and navigation, clickstream studies have shown how users alter their browsing as they navigate sites and make return visits. The behavior patterns uncovered are consistent with learning effects as well as attention to cost-benefit trade-offs between time and information value. Researchers have also found that users conduct surprisingly low levels of cross-site search, despite the seemingly low costs of doing so. This behavior helps explain the ability of companies to build sustainable businesses on the Web as well as to command price premiums.

In the second category, advertising, clickstream data has enabled researchers to understand and model, in sophisticated fashion, the consumer response to banner advertisements and email solicitations. These studies have shown not only that these
advertising vehicles can be effective, but how marketers can begin to target and customize them to enhance their efficiency. Clickstream research is also well underway in the areas of paid search and online word-of-mouth, but is not yet well developed.

The third category, online shopping and e-commerce, has seen models developed which successfully predict the purchase conversion behavior of visitors to e-commerce websites. A variety of approaches have been employed to do so, including stochastic models and sophisticated forms of binary choice models. These models enable site managers not only to forecast and target, but to better understand the impact that site design and structure elements might have on purchase conversion. An extensive literature has also developed using clickstream data to study the empirical properties of online auctions, including late bidding or “sniping” and various sources of information asymmetry.

While we have noted that clickstream research is emerging in the areas of paid search and online WOM, there is clearly vast potential for study on these topics along with very strong current interest from industry practitioners. Future clickstream research into the management of multiple channels and recommendation systems should also be able to yield new insights and methods. Because clickstream data capture so many different elements of marketing activity on the Internet along with such a wide range of consumer decision-making, researchers should also be able to build integrated models of different behaviors.

When compared to computer science, the study of clickstream data in marketing looks relatively light. Several challenges inherent in the use of clickstream data could explain at least part of this. First, the size and complexity of the datasets is often daunting for both researchers and practitioners. Second, to effectively use the information, clickstream data usually needs to be augmented and matched with other sources within the firm. As researchers continue to demonstrate the value of clickstream-based research and produce new findings, we are optimistic that these barriers will erode over time. From a marketing perspective, clickstream analysis is probably best characterized as being in the early growth phase of its life cycle. It has been successfully introduced to the field, awareness and use are spreading quickly, but there is no question that maturity lies, if you will, many clicks into the future.
References


