Analyzing Website Choice Using Clickstream Data

By: Avi Goldfarb*

June 5, 2002

Joseph L. Rotman School of Management
University of Toronto

* This research was supported by the Social Science Research Council through the predissertation fellowship of the Program in Applied Economics and by a Plurimus Corporation Research Fellowship. I would like to thank Plurimus Corporation for providing me with the clickstream data and J. Walter Thompson Company for providing me with advertising data. I would also like to thank Shane Greenstein, Charles Manski, and Robert Porter for helpful comments. All remaining errors are my own.
Correspondence to: avigoldfarb@rogers.com.

Abstract: This paper estimates demand for Internet portals using a clickstream data panel of 2654 users. It shows that familiar econometric methodologies used to study grocery store scanner data can be applied to analyze advertising-supported Internet markets using clickstream data. In particular, it applies the methodology of Guadagni and Little (1983) to better understand households' Internet portal choices. The methodology has reasonable out of sample predictive power and can be used to simulate changes in company strategy. (JEL classification numbers: M31, C25. Keywords: Internet, search engine, clickstream, website, multinomial logit)

1. Introduction

The growth of the Internet has provided economists, marketers, and statisticians with a potentially rich and informative data source. Since everything on the Internet is necessarily digital, all activity can be easily recorded and stored in a database for future examination. This data has found disparate uses, from advertisement targeting to law enforcement. One prevalent but relatively under-used example of such data is clickstream data. This data consists of each website visited by a panel of users and the order in which they arrive at the websites. It is often accompanied by the time of arrival at and departure from the website as well as the degree of activity at the website and the demographic characteristics of the users. Examples of companies that collect clickstream data based on broad panels are Netratings Inc., MediaMetrix Inc., and Plurimus Corp. This paper uses data from Plurimus Corp. to analyze user choice of Internet portals. It will show that commonly used econometric models for examining grocery scanner data can be applied to clickstream data in advertising-based online markets.

Following Hargittai (2000, pp. 233), I define an Internet portal as "any site that classifies content and primarily presents itself as a one-stop point-of-entry to content on the Web." Portals, such as Yahoo, Altavista, and MSN have search engine capabilities, but they also have other features. These may include email, news, and a link-based directory to the web separate from the search service. There are few, if any, pure search engines remaining. I narrow Hargittai's definition further. I am interested in the portal as a starting point and not as a destination, and I therefore look at the use of portal main pages, directory pages, and search pages, but not at email and shopping pages.

The methodology used here closely mimics that of Guadagni and Little's (1983) paper that estimates a multinomial logit model with scanner data to examine consumer coffee purchases. It shows that the model has reasonably good out-of-sample predictive ability. Furthermore, informative simulations can be conducted on the effects on market share of changing a variable. For example, it can derive an estimate of the impact on number of visits of increasing advertising by one dollar.

Developing a framework to study consumer choices of free (advertising-supported) websites is an essential step to better understanding user behavior on the Internet. According to the data set used in this study, more than two-thirds of all

consumer Internet traffic is at advertising-supported sites. With the exceptions of Amazon and EBay, the top twenty sites in terms of unique visitors are all advertising-supported. The literature on this important aspect of the Internet is sparse. Three studies that focus on advertising-supported websites are Adar and Huberman (1999), Gandal (2001), and Goldfarb (2001). Adar and Huberman (1999) show that portals can discriminate between users as those looking for certain topics are willing to spend more time. This means that search engines can capture more consumer surplus (in the form of advertising revenue) by forcing consumers that are willing to spend more time to view more pages and advertisements. Gandal (2001) examines market share at an aggregate level to try to examine the portal market. He finds that early entrants have an advantage and that certain features matter more than others. Goldfarb (2001) examines concentration levels in advertising-supported Internet markets.

Lynch and Ariely (2000) is one of few Internet studies that looks at choice-specific data. They construct a simulated environment for the purchase of wine and examine purchase choice. Like Lynch and Ariely's study, this paper takes advantage of the choice-specific data. Unlike their study, I look at the choice of free web sites using actual user clickstreams.

The main data for this study was supplied by Plurimus Corporation. It is a clickstream data set consisting of every website visited by 2654 users from December 27, 1999 to March 31, 2000. It also contains data on the time of arrival at and departure from each site. In total, the data set contains 3,228,595 website visits, of which 859,587 (2622 people) are to Internet portals. Using this data, I construct measures of past search success, past time spent searching, whether a site is an individual's starting page, whether an individual has an email account at the site, and the number of pages viewed at each site. A considerable section of this paper is dedicated to explaining the construction of these variables from the raw data. I link the Plurimus data to monthly advertising spending data from J. Walter Thompson Company and media mentions data found through the Lexis-Nexis Academic Universe.

The next section of the paper provides a brief history of Internet portals. Section three describes the application of the methodology used by Guadagni and Little to the present problem, and section four explains the data set. Section five presents the results,

tests the model's predictive ability, and examines market response to changes in the control variables. The paper concludes by summarizing the key results and proposing several potential areas for future research.

2. The Internet Portal Market

Portals operate in a peculiar environment. The compete in two distinct markets. They compete in quality for users, but users provide no direct revenue. The revenue comes from advertising. Hotwired magazine pioneered this business model in October 1994, when the first banner advertisement (for Zima alcoholic beverages) appeared on their website. The advertising market is largely competitive: portals compete with thousands of other websites, as well as television, radio, newspapers, magazines, and billboards.

In their first generation, Internet portals were search engines. They maintained large databases of websites and allowed users to search them. In exchange, users viewed banner advertisements. By late 1996, it became obvious that there were too many undifferentiated search engines competing for the same users and the same banner advertisements. OpenText, once a large search engine, closed in the middle of 1997. Webcrawler and Magellen were taken over by Excite and neglected. The more successful search engines began to provide proprietary content in an attempt to differentiate themselves from their rivals. As search engines began offering email accounts, stock quotes, and news services, they became known as 'portals' because they were gateways to the Internet. Over the next few years, these portals bought other content companies and used that content to generate traffic and revenue. Yahoo expanded aggressively, buying broadcast.com, Geocities, and dozens of other smaller companies. Lycos bought Gamesville and quote.com. Excite bought Bluemountainarts (then the largest e-card company) and classifieds2000 (then the largest provider of online classified listings). Search engines were differentiated by the proprietary content they provided following a search.

The richer content generated more banner advertising, but many looked to new revenue streams. 'Partnerships' were the first of these new revenue streams to succeed. In July 1997, Amazon became the 'preferred book merchant' to Excite and Yahoo.

Preferred sponsor programs soon arrived at AOL and Lycos. Unlike banner advertisements, sponsors' names appeared in response to certain keyword searches. Furthermore, their logos were placed prominently in the middle of the website. Goto.com (now Overture) took the idea of sponsorships one step further. Founded in 1998, Goto's Internet database consists only of paid sponsors. Advertisers bid on keywords. The advertiser who pays most for a given keyword is listed first in the search results, the advertiser who paid second-most is listed second, and so on. Today, Goto continues to thrive and many other portals, including looksmart, about.com, and iwon.com, use its technology. The results in Goto's database are supplemented by another search engine's results if there are not enough paid listings.

There are several other revenue streams that portals use. Pop-up advertisements have become widespread. Many portals charge businesses a fee to be included in their directory manually. For example, a business can pay Yahoo to get listed immediately, or wait and hope that Yahoo's directory editors stumble across the website at some point in the future.

Today, the Internet portal market is stabilizing. The last major entries occured in 1999 with Google and Iwon.com. Since then, exit has been much more common. NBCi, Go.com, and Excite have all folded. Yet, quality improvements in search technology and in content continue. Google recently added 'pdf' files to its search capabilities, and MSN recently entered a partnership with ESPN. One finding in this paper is that both search efficacy and usable content are important to driving users to portals.

3. Using the Multinomial Logit With Clickstream Data

Internet users choose which website to visit just as they make several other economic choices: given the alternatives available and the information they have about those alternatives, they choose the alternative that will give them the highest utility. In terms of grocery products such as coffee (studied by Guadagni and Little), this means that households buy the product that has the best attributes for the lowest price. In terms of portals, this means that households will use the portal that will allow them to maximize the probability of finding what they seek and minimizing the time spent. Conceptually, I assume households are exogenously given a "goal" when they go online. They go to the

portal that they expect will help them achieve that goal in the least time with the most accuracy.

In the multinomial logit model, the expected utility of the portal is based on past history, several website characteristics (that may vary over time), outside influences such as advertising and media mentions, and an idiosyncratic error term. Formally, household *i* visits website *j* on choice occasion *t* when

$$Eu_{ijt} \geq Eu_{ikt} \tag{1}$$

for all $k \neq j$. Here Eu_{ijt} is defined by

$$Eu_{ijt} = X_{ijt}\beta_{ijt} + \varepsilon_{ijt} \tag{2}$$

X_{ijt} may include variables that change over any or all of *i*, *j*, and *t*. β may vary over *i*, *j*, or *t*, implying household heterogeneity, brand heterogeneity, time (choice occasion) heterogeneity or any combination of the three. In this chapter, X_{ijt} will never vary over just *t*, just *i*, just *t* and *i*, or just *t* and *j*. It will vary over just *j* in the form of portal-specific dummy variables. β will be assumed constant. There are I households, J websites, and T_i choice occasions for each household.

It is expected utility to the user, not to the observer, that is of interest. It is assumed that the user knows ε_{ijt}. The expectation is taken over relevant variables that the user may not know the value of before visiting the website. For example, the user does not know how long she will spend on the website. She does, however, have an expectation of how long it will take based on her past experience at that website.

In order to get the multinomial logit form, ε_{ijt} is assumed to be independently distributed random variables with a type II extreme value distribution. Given the above assumptions, the probability of household *i* choosing brand *j* at choice occasion *t* can be expressed as:

$$P_{it}(j \mid X_{ijt}, \beta_{ijt}) = \frac{\exp(X_{ijt}\beta_{ijt})}{\sum_{k=1}^{J} \exp(X_{ikt}\beta_{ikt})} \tag{3}$$

The model, as expressed above is a combination of Theil's (1969) multinomial logit and McFadden's (1974) conditional logit. It is commonly referred to as a mixed logit or as a

multinomial logit. Since this paper assumes β is fixed, the model here is a conditional logit. The log likelihood function is as follows:

$$\sum_{i=1}^{I}\sum_{t=1}^{T_i}\sum_{j=1}^{J} d_{ijt} \ln P_{it}(j \mid X_{ijt}, \beta) \qquad (4)$$

where d_{ijt} is equal to one if alternative j is chosen by individual i at time t, and is equal to zero otherwise.

A significant potential problem with this framework is that it implies an assumption of independence of irrelevant alternatives (IIA). If a household is offered a new alternative that is almost identical to one of the current alternatives, say k, then this new alternative should be expected to only draw buyers from k; however, under IIA, the new alternative draws buyers from all the other alternatives. IIA is not a major factor in the questions addressed in this study. Furthermore, it complicates the econometric analysis considerably.

In this model, the researcher observes the choice by each household on each choice occasion. Let $y_{ijt}=1$ if household i chooses website j on choice occasion t and let $y_{ijt}=0$ otherwise. The researcher also observes the characteristics of each website at that choice occasion for that household X_{ijt}.

4. Data

4.1 Raw data sources and description

The main data set consists of 3,228,595 website visits by 2654 households from December 27, 1999 to March 31 2000. Also included in the initial data set was the time of arrival at and departure from a website, the beginning and end of each online session, and the number of pages visited at that site. This data, collected by Plurimus Corporation, was used to construct a data set of 859,587 portal choices by 2622 households. This study uses only 2008 of these households and keeps the others to test the model out of sample. Furthermore, it only looks at the eight most frequently used portals comprising eighty percent of all portal visits. Therefore the final data set consists of 519,705 portal choices by 2005 households.

Plurimus has an anonymizing technology that allows them to collect information about users without needing the users' permission. Plurimus avoids significant privacy

concerns because the users are anonymous and the data cannot be traced to any actual person. They are regularly audited by PriceWaterhouseCoopers in order to ensure they exceed the privacy requirements of the FCC guidelines. Unlike volunteer panel data, behavioral records from anonymized users are not biased by the wish to be seen in a socially desirable light. Moreover, there is no selection bias into the sample itself, yielding a sample from a broader spectrum of socioeconomic status than is typically available from panel studies.

This data, however, has five limitations that need to be considered when extending the results of this study to the entire Internet. First, the geographic distribution of the sample is considerably biased. New York, Chicago, and Los Angeles are under-represented. Roughly half the sample comes from the Pittsburgh area. Another quarter is from North Carolina and another eighth is from Tampa. This problem is not as severe as it may first appear because portals are a national product.

The second limitation is that it does not collect data on America Online (AOL) users. Since AOL subscribers make up roughly 50% of all American home Internet users, this could bias the results. AOL, however, provides a different product from the other Internet service providers. AOL users are encouraged to stay within the gated AOL community and they generally do not venture out onto the rest of the Internet. Moreover, preliminary surveys commissioned by Plurimus show that when AOL users do leave the gated AOL community, they have similar habits to other web users. This data limitation will, however, put a downward bias on visits to the AOL portal.

Third, the data contains information on few users at work. Online habits at work are likely different from those at home; however according to a study by Nie and Erbring (2000), 64.3% of Internet users use the Internet primarily at home; just 16.8% use it primarily at work. Few data sets, however, contain reliable at-work panel data.

With the exception of AOL, Plurimus' market share numbers for Internet portals are well within the range of the other companies. The correlation coefficients for monthly market shares from January to March 2000 are 0.90 for Plurimus and MediaMetrix and 0.78 for Plurimus and PC Data Online. Since the numbers are generally quite close, the above issues with the data may not be important for understanding portal choice by users who are not AOL subscribers.

The fourth limitation is that the data is collected at the household level rather than at the individual level. If two people in a given household have considerably different habits this will show up as one person with widely varying habits. While this makes it difficult to assess the extent of learning over time, it is a standard problem in consumer panels.

Fifth, it does not contain information on households from the first time they go online. Therefore initial conditions are potentially a problem. Although the observations may not be independently and identically distributed, this problem may be partially alleviated by the law of large numbers due to the number of observations per household in the data set. More than 79% of the households in the final data set make 30 or more choices. The mean household makes 259 portal choices and the median household makes 120 portal choices.

Together, these five data limitations mean that results should be extended to different geographic distributions, AOL users, and at-work users with caution. Furthermore, the fourth and fifth limitations mean that understanding learning behavior is not possible.

I join this clickstream data set with two other data sets. The first is an advertising data set provided by J. Walter Thompson Company. This data set consists of all advertising spending by each of the portals used in this study on a monthly basis. The spending is determined by a thorough sampling of television, radio, newspaper, magazine, outdoor, and Internet advertising by each of the portals. The number of advertisements is then multiplied by the average cost of advertising in each medium (at the program level in television and the issue level in magazines). Since this data is not individual-specific, it will likely underestimate the impact of advertising. The methodology used in this paper, however, can easily be adapted to individual-specific advertising data.

I also constructed a data set of ‘media mentions’ for each of the relevant companies. If a company is mentioned on network television news (ABC, CBS, or NBC), in the Wall Street Journal, in the New York Times, or in USA Today on a given day or the day before then the media mentions variable is equal to one. Otherwise it is equal to zero. Unfortunately, I do not know which individuals were actually watching or

reading which media. It is likely, however, that mentions in these media are highly correlated with mentions in other media such as local newspapers.

In the data set, several dozen portals are observed to be chosen. For computational feasibility, I limit the number of portals to the eight with the most visits (in order): Yahoo, Microsoft Network (MSN), Netscape, Excite, AOL, Altavista, Iwon, and Lycos. These eight make up eighty percent of all visits and all portals with more than 2.5% of total visits. There was a natural break after Lycos because the ninth most visited portal, MyWay, is a site that is the default of several Internet Service Providers and is rarely chosen as anything but a start-up page. Go.com is not included because, although it is commonly ranked in the top five portals, a large percentage of those visits are to destination websites such as ESPN.com, Disney.com, and MrShowbiz.com. The Go.com portal page itself ranks tenth in total visits. Qualitative results, however, do not change with the addition of more portals.

4.2 Data set Construction

I used the above information to construct several variables from the raw clickstream data. Table 1 shows a sample of ten lines of raw data. Using only this information, I constructed the following variables: email, goal of search, start page, view length at the portal, links, repeated search, whether a portal was the first visited in the search process, and Guadagni & Little's weighted loyalty variable. I will describe the derivation of each in turn.

TABLE 1 ABOUT HERE

A household was considered to have an email account at a site if the household used the email feature at that site more than that at any other portal. I know that a household used email at a given site because the 'host' in the data would reveal this. For example, 'com.yahoo.mail' is Yahoo's email provider and 'com.hotmail' is MSN's email provider. No household used more than one email account a large number of times, so I did not allow for households to have more than one portal as an email provider. Many households did not use a portal email provider. This *same email* variable is potentially

endogenous when individual heterogeneity is not taken into account because users will set up an email account at their favorite portal. As such it can be used as a proxy for some individual heterogeneity. Furthermore, if the goal is to predict future choices or to simulate changes, then this endogeneity is not relevant. It was the initial decision to use the email that was endogenous; once that account is set up, each choice of portal is based on the existence of the email account.

I cannot identify search failure exactly, but I proxy it with a *repeated search* variable. If a household visited two portal sites in a row, and there was less than five minutes between visits, then the first search is likely a failure. Furthermore, if the household conducts a search and then searches again for the same goal[1] (at the same site or at a different one) within five minutes of the first search then the *repeated search* variable is equal to one. While five minutes is arbitrary, extending the time to ten minutes or shortening it to three did not change the number of repeats much. As with time spent, it is whether previous searches at a portal were repeated that matters. Also as with time spent, more complicated functions of past repeated searches do not yield qualitatively different results. I call this variable *last search repeated*.

A portal is considered to be a household's *start page* if at least 50% of all online sessions begin with that page. An online session is considered to end if a user does not do any activity for thirty minutes. While imperfect, this method determines a starting page for almost all of the households. Like, *same email*, *start page* is potentially endogenous. People often change their start page to their favorite website. Again like *same email,* this can proxy individual heterogeneity and the endogeneity is not relevant if the goal is to predict future choices or to simulate changes. 28% of households have their start page at a portal. This is likely lower than the general population due to the lack of AOL users.

The view length spent at a portal is the time of departure minus the time of arrival (in seconds). Recall that it is time spent during *previous* visits that is important for whether a household returns to that portal.

---

[1] The goals were divided into roughly one hundred overlapping categories including news, music, email, shopping for computers, automotive information and travel. I did not include goal of search in the final analysis because including it did not satisfy the Akaike information criterion or the Bayesian information criterion for goodness of fit.

The number of pages viewed at a portal may reflect the depth of search. While individuals likely want to minimize time spent generally, search depth may be an important control factor. As with view length, it is number of pages viewed during previous visits that is important for whether a household returns to that portal. This study only reports results from a one period lag on *last view length* and *last number of pages*. More complicated functions of past time spent and previous number of pages viewed do not yield qualitatively different results.

Links were determined by visiting each portal and recording which websites were directly linked to the main page. I recorded links in early April for each of the portals. While it is possible that several of the links changed, there were no relevant changes in partnerships over that time. If the site that an individual visited following a portal visit was linked to a portal, the *link* variable takes on a value of one. Otherwise, it equals zero. Note that the link variable can equal one even if the household did not visit that portal. For example, a household could search for financial information on Yahoo and the search may turn up information on MSNmoneycentral. The *link* variable serves as a proxy for portal features. Instead of listing whether a portal has features, this variable proxies whether people actually use these features. In other words, if people use a link, it means they are using a feature at that site, rather than the search capabilities.

If a portal was the first visited in the search process, then $firsttry_{ijt}=1$. If an individual has already searched, then $firsttry_{ijt}=0$.

This paper mimics Guadagni and Little's methodology for constructing their 'loyalty' variable almost exactly. In their paper, loyalty is considered to be a weighted average of past purchases of the brand, treated as dummy variables. Let $portsame_{ijt}=1$ if household *i* bought brand *j* as its previous purchase and zero otherwise.

$$loyalty_{ijt} \equiv \alpha loyalty_{ijt-1} + (1-\alpha) portsame_{ijt} \qquad (5)$$

Rather than estimate α by maximum likelihood which would significantly complicate the computational problem they calibrate α based on dummies for lags of length one to ten. In the present study, the value for alpha that minimizes the sum of the difference between the actual dummy coefficients and the loyalty function above was 0.7782. This loyalty variable can be a result of either individual preferences for a given portal or from some kind of lock-in. I do not separate these two effects, but the variable is an important

predictor of portal choice. In a recent study, Abramson, Andrews, Currim, and Jones (2000) find this to be the best loyalty measure they tried. Defining loyalty as *portsame* rather than *GL Loyalty* does not change the qualitative results.

In this study, I define the $portsame_{ijt}$ variable to depend on the previous portal visited of any kind, not just the previous of the eight portals used in this study. Therefore, if a household visits Yahoo then About.com and then Yahoo again, $portsame_{ijt}$ on the second visit to Yahoo is equal to zero, even though only two observations are included in the data set. This means that a household is not considered brand loyal if it went to a rival portal's website, even if that rival portal is not in the sample. If I only include the sample, the coefficient on the loyalty variable increases slightly but its significance falls slightly. The initial conditions problem frequently encountered in this literature does not apply here due to the large number of observations per household.

How much time a household's previous visit to a portal took and whether that search was repeated are only observed when the household has visited that portal previously in the data set. Since not every household visits every portal, these variables are missing for a large number of observations. I therefore created a dummy variable for missing data. I also interact one minus the missing data variable with the view length of previous search and the repeated search variables. This overcomes the significant potential bias of assuming a value for the missing data or of ignoring it entirely. The missing data dummy has no economic interpretation.

TABLE 2 ABOUT HERE

Table 2 contains descriptive statistics of the final data set. Yahoo has over twice the market share of its closest competitor, MSN. This table suggests that Yahoo's success may be largely a function of the features of its website. Searches are repeated much less often on Yahoo than elsewhere, it is the start page for the largest number of users, and it is the email provider for the second largest number of users. Furthermore, Yahoo advertises heavily, is frequently mentioned in the media, has frequently used links, and does not take long to search. Lycos searches, on the other hand, are repeated

frequently, Altavista's links are rarely used, and only Yahoo and MSN have a large number of email users.

5. Results

5.1 Coefficients

Table 3 presents the main results of the paper. Model (1) presents the base model. Here, the potentially endogenous variables of *same mail*, *link*, and *start page* are not included. The variables all have the expected signs, although *last view length* is barely significant: *loyalty*, *advertising*, and *media mentions* are all correlated with a higher probability of search. *Last view length* and *last search failed* are all correlated with a lower probability of search. The positive sign on *last view length squared* suggests that the effect of *last view length* is concave. There was no expectation on the sign of *missing data*. The coefficient on advertising likely underestimates the actual effect of advertising as the data is aggregated over the month rather than actual advertising viewed by the user.

TABLE 3 ABOUT HERE

Model (2) adds *same email* and *link* with the expected results. Taking these into account makes *last view length* significant. Model (3) adds *last number of pages* and *first try*. *Last number of pages* is found to have an increasing and concave relationship with choice probability. This is consistent with the assumption that pages viewed proxy depth of search. In this regression, *last view length* is significant at the 99% confidence level. Thus, controlling for depth, households prefer to spend less time at a portal. *First try* reveals that Netscape and MSN are preferred as first pages in a search than as later pages. This makes sense as they are the pages that appear when using the search function in the Netscape Navigator and Microsoft Internet Explorer browsers. They are also often default start pages, but the results do not change in models (4) through (6) which control for the start page.

Model (4) adds the *start page* variable to model (2). The coefficient on this variable is very large compared to the other dummy variables and the likelihood improves

more for this variable than for any others; however, the coefficient is not significantly different from zero as it has an extremely high standard error.

Model (5) is the same as model (4) except that is adds the interaction variable of media mentions and loyalty. Of particular interest here is the increase in the significance of *media mentions*. This suggests that being mentioned in the media has a larger effect for households that are less loyal to the brand.

Model (6) is the 'kitchen sink' regression in that it includes all of the variables in the study. The coefficients and their significance are similar to models (1) through (5).

Another interesting aspect of all of the models is that there is a clear brand preference for Yahoo over the others. Models (1) through (3) have negative coefficients for all brand dummies (Yahoo is the base). Models (4) through (6) also have negative dummies for Yahoo. While others may seem preferred on the first try, adding the coefficients together leaves a negative number meaning that Yahoo is generally preferred even on the first try.

The Akaike information criterion revealed that *last view length squared*, *last number of pages squared*, and *media mentions*loyalty* should be included. Other variables such as *advertising squared* and *advertising*loyalty* did not satisfy the Akaike information criterion. Note that including *start page* increases the likelihood a great deal, even though the effect is statistically insignificant. Any variables included in this study that satisfy the Akaike information criterion also satisfy the Bayesian information criterion.

5.2 Market Response to Variable Changes

Table 4 explores the market responses to variable changes in model (2) assuming no competitive response. This table presents elasticities in the form of changes in number of total visits over the three month period, assuming that there are a total of 43.3 million online households, Plurimus' estimate for the month of February 2000. Taking the results at face value, if MSN users' searches were repeated just 1% less often, MSN would get almost 1.7 million more site visits. If each site visit is worth five cents (about the revenue received from the five advertisements seen over typical two page views at a typical search engine), then it would be worth it for MSN to implement this change as

long as it cost less than eighty-five thousand dollars over three months. Links, search time, and search failure (proxied by repeated search) matter, suggesting that usable content and search efficacy are important factors in driving users to portals.

TABLE 4 ABOUT HERE

The advertising results are perhaps the most interesting. An increase in advertising by one dollar would bring 3.7 more visits to Altavista but 15 more to Yahoo. Therefore, Altavista should increase its advertising if each new site visit brings in twenty-seven cents of revenue and Yahoo should increase its advertising if each new site visit brings in just 6.7 cents of revenue. More effective links and more media mentions are other ways companies can drive traffic to their websites.

Caution should be used in interpreting these results because of the lack of IIA and because the functional form of the error term is important to deriving these results. While the numbers themselves may not be completely accurate, it is likely that an extra dollar of advertising by Yahoo has a larger effect than an extra dollar of advertising by Altavista. The current exercise should be viewed as an approximation that demonstrates potential marginal gains from the variables.

Another way to simulate policy changes by the firms is to change the underlying data and reestimate the market shares given the known coefficients. This method underestimates changes because it does not count dynamic effects. It does, however, provide a lower bound for the impact. Again using model (2), I undertook this exercise for several variables. If MSN advertised as much as AOL, then MSN would gain 7,843,659 more visits assuming 43.3 million households. If, on the other hand, Iwon advertised as much as AOL then it would only gain 1,617,622 visits. If Lycos searches were successful as often as Yahoo searches, Lycos traffic would rise by 14,561,538 or four percent. If Altavista had the same links as MSN then it would get 56,005,914 more visitors or ten percent. The exact quantities of these predictions should be interpreted with caution. The general trends, however, are informative.

6. Conclusion

This study has provided a preliminary look at estimating demand for advertising-supported Internet websites based on clickstream data. The methodology provides a reasonable fit to the actual patterns in the data. It has good predictive power and is informative about the potential impact of various policy changes.

With respect policy implications, the study provides a framework for understanding policy effects. The simulations in section 5.3 show the impact of potential policy changes on market shares. While they do not take into account supply side reactions or individual heterogeneity, they do give better estimates of policy effects than currently exist. More detailed policy analysis can also be explored in this framework. For example, a portal could simulate a link to a commonly used website, say americangreetings.com. It could then determine the effect of this link on market share. The actual increase in share resulting from this change would be no more than the simulated level. It may be less because it may be that people who go to a given portal are also the kind of people who like the links it has. Thus the effects of the new link may be less than predicted. Because it does not account for individual heterogeneity, this model does not provide an effective framework for examining the effects of major industry changes such as bankruptcies, nor does it provide a way to look at the welfare impact of improved technology.

The main purpose of this study was to show that demand for free online services can be estimated using methodologies that are common in both the economics and the marketing literature. The coefficients on the variables in the study have the expected signs and the predictive ability of the model, though not perfect, captured the major trends. Furthermore, I present informative simulations about the effects on share of changing variable values. Clickstream data will be an important tool in understanding online demand. This study has shown that the standard econometric methods that have previously been applied to grocery scanner data can successfully be applied to clickstream data. By bringing more econometric sophistication to this analysis, economists and marketers can gain a better understanding of online user behavior.

References

Abramson, Charles, Rick L. Andrews, Imran S. Currim, and Morgan Jones, 2000. Parameter Bias from Unobserved Effects in the Multinomial Logit Model of Consumer Choice. Journal of Marketing Research 37, 410-426.

Adar, Eyton, and Bernardo Huberman, 1999. The Economics of Surfing. Working Paper No. 42, Center for eBusiness at MIT.

Gandal, Neil, 2001. The Dynamics of Competition in the Internet Search Engine Market. International Journal of Industrial Organization 19, 1103-1117.

Goldfarb, Avi, 2001. Concentration in Advertising-Supported Online Markets: An Empirical Approach. Unpublished Working Paper, Northwestern University.

Guadagni, Peter M. and John D. C. Little, 1983. A Logit Model of Brand Choice Calibrated on Scanner Data. Marketing Science 2, 203-38.

Hargittai, Eszter, 2000. Open Portals or Closed Gates? Channeling Content on the World Wide Web. Poetics 27, 233-254.

Lynch, John G. and Dan Ariely, 2000. Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution. Marketing Science 19, 83-103.

McFadden, Daniel, 1974. Conditional Logit Analysis of Qualitative Behavior, in Zarembka, P. (Ed.), Frontiers of Econometrics. The Academic Press, Inc., New York, pp. 105-142.

Nie, Norman H., and Lutz Erbring, 2000. Internet and Society. Unpublished Working Paper, Stanford Institute for the Quantitative Study of Society.

Theil, H., 1969. A Multinomial Extension of the Linear Logit Model. International Economic Review 10, 251-259.

TABLE 1: Clickstream Data Sample

USER	HOST	START TIME	END TIME	BYTES FROM	BYTES TO	# PAGES VIEWED AT HOST
1	com.yahoo	14MAR00:08:42:55	14MAR00:08:45:28	196593	34484	3
1	com.allrecipes	14MAR00:08:45:28	14MAR00:08:50:59	65825	656	12
1	com.ivillage	14MAR00:08:55:00	14MAR00:09:09:48	541337	72005	53
1	com.allrecipes	18MAR00:12:27:10	18MAR00:12:34:46	75403	4454	5
1	com.allrecipes	21MAR00:12:31:01	21MAR00:12:36:51	75873	658	2
1	com.excite	28MAR00:13:13:59	28MAR00:13:15:22	105884	4006	4
1	com.adobe	28MAR00:13:15:06	28MAR00:13:19:39	70732	11988	9
1	gov.nara	28MAR00:13:19:38	28MAR00:13:21:57	1259	2340	1
1	gov.nara	28MAR00:13:34:09	28MAR00:13:38:00	60155	9074	13
1	com.allrecipes	30MAR00:16:44:18	30MAR00:16:52:05	86186	1857	4

TABLE 2: Summary statistics

Portal	Percent share of visits to top 8 portals	Average time spent at site (in seconds)	Percentage of times search repeated	Percentage of households with portal as start page	Percentage of households with same email	Percentage of visits using a link	Percentage of days with media mentions	Average monthly advertising spending (thousands of dollars)
Yahoo	42.0	96.67	7.03	9.76	19.92	3.20	58.33	2361.5
MSN	20.9	116.72	12.10	7.17	32.97	4.41	6.35	277.4
Netscape	13.5	114.0	13.33	5.38	4.38	3.62	13.54	198.7
Excite	6.5	93.21	11.28	1.29	2.39	2.57	15.63	397.7
AOL	5.5	93.89	11.11	0.75	4.48	2.78	82.29	7263.4
Altavista	5.0	109.7	14.41	0.30	0.40	0.17	5.21	1161.0
Iwon	3.6	152.0	14.81	0.30	1.59	0.69	1.04	0
Lycos	3.0	96.21	31.55	0.20	4.63	1.82	16.67	1570.2
Mean over all observations	N/A	105.45	15.30	2.41	11.32	1.87	33.92	1772.5
Standard deviation over all observations	N/A	171.59	36.01	15.34	31.72	13.53	47.34	2389.6

TABLE 3 – Model coefficients (with standard errors in parentheses)

Variable	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
GL Loyalty	1.35***	1.31***	1.32***	1.21***	1.27***	1.27***
	(0.00235)	(0.00245)	(0.00247)	(0.00261)	(0.00368)	(0.00368)
Missing Data	-2.35***	-2.32***	-2.28***	-2.26***	-2.24***	-2.21***
	(0.0126)	(0.0127)	(0.0129)	(0.0129)	(0.0130)	(0.013165)
Last view time at that site	-1.90E-05^	-2.20E-05*	-0.000120***	-2.60E-05*	-2.60E-05*	-0.000110***
	(1.34E-05)	(1.35E-05)	(1.59E-05)	(1.41E-05)	(1.41E-05)	(1.67E-05)
Last view time squared	2.08E-09**	2.31E-09**	6.69E-09***	2.43E-09**	2.49E-09**	5.95E-09***
	(9.89E-10)	(9.87E-10)	(9.90E-10)	(9.87E-10)	(9.93E-10)	(9.97E-10)
Last search failed	-0.476***	-0.440***	-0.425***	-0.451***	-0.452***	-0.451***
	(0.00608)	(0.00618)	(0.00620)	(0.00645)	(0.00646)	(0.00646)
Advertising ($ 000)	5.89E-06*	6.08E-06**	6.17E-06**	4.59E-06^	5.30E-06*	5.53E-06*
	(3.01E-06)	(3.07E-06)	(3.09E-06)	(3.17E-06)	(3.16E-06)	(3.16E-06)
Media Mentions	0.0137**	0.0136**	0.0124*	0.0109^	0.129***	0.128***
	(0.00667)	(0.00680)	(0.00683)	(0.00712)	(0.00857)	(0.00857)
Media Mentions*loyalty					-0.144***	-0.143***
					(0.00590)	(0.00590)
Same email		0.166***	0.174***	0.174***	0.181***	0.181***
		(0.00511)	(0.00513)	(0.00544)	(0.00544)	(0.00544)
Link		1.98***	2.02***	2.05***	2.06***	2.05***
		(0.0109)	(0.0110)	(0.0113)	(0.0113)	(0.0113)
Last number pages viewed at that site			0.0103***			0.00875***
			(0.000710)			(0.000726)
Last number of pages squared			-6.70E-05***			-5.10E-05***
			(9.19E-06)			(8.38E-06)
Start page				34.12	41.11	36.11
				(146.12)	(247.87)	(203.90)
Altavista	-0.530***	-0.494***	-0.287***	-0.248***	-0.246***	-0.258***
	(0.0103)	(0.0105)	(0.0141)	(0.0142)	(0.0142)	(0.0142)
AOL	-0.571***	-0.700***	-0.764***	-0.726***	-0.769***	-0.779***
	(0.0169)	(0.0173)	(0.0202)	(0.0205)	(0.0205)	(0.0205)
Excite	-0.479***	-0.612***	-0.548***	-0.540***	-0.543***	-0.553***
	(0.00971)	(0.0101)	(0.0145)	(0.0147)	(0.0147)	(0.0147)
Iwon	-0.415***	-0.430***	-0.662***	-0.633***	-0.639***	-0.662***
	(0.0135)	(0.0138)	(0.0204)	(0.0205)	(0.0206)	(0.0207)
Lycos	-0.686***	-0.808***	-0.489***	-0.494***	-0.499***	-0.496***
	(0.0105)	(0.0108)	(0.0147)	(0.0149)	(0.0148)	(0.0148)
MSN	-0.0270***	-0.174***	-0.592***	-0.654***	-0.674***	-0.670***
	(0.00953)	(0.00971)	(0.0128)	(0.0133)	(0.0133)	(0.0133)
Netscape	-0.157***	-0.261***	-0.695***	-0.779***	-0.791***	-0.798***
	(0.0101)	(0.0104)	(0.0144)	(0.0150)	(0.0150)	(0.0151)
First Try (Altavista)			-0.393***	-0.345***	-0.353***	-0.353***
			(0.0169)	(0.0170)	(0.0171)	(0.0171)
First Try (AOL)			0.0924***	0.135***	0.137***	0.139***
			(0.0167)	(0.0169)	(0.0168)	(0.0168)
First Try (Excite)			-0.126***	-0.153***	-0.165***	-0.168***
			(0.0171)	(0.0176)	(0.0176)	(0.0177)
First Try (Iwon)			0.321***	0.361***	0.357***	0.361***
			(0.0219)	(0.0221)	(0.0223)	(0.0223)
First Try (Lycos)			-0.580***	-0.468***	-0.474***	-0.475***
			(0.0195)	(0.0197)	(0.0196)	(0.0196)
First Try (MSN)			0.631***	0.632***	0.632***	0.633***
			(0.0123)	(0.0129)	(0.0129)	(0.0129)
First Try (Netscape)			0.646***	0.668***	0.665***	0.667***
			(0.0144)	(0.0154)	(0.0154)	(0.0154)
Log likelihood	-442,856	-425,651	-421,531	-386,956	-386,659	-386,581

*** significant at a 1% level in a two-tailed test
** significant at a 5% level in a two-tailed test
* significant at a 10% level in a two-tailed test
^ significant at a 10% level in a one-tailed test

TABLE 4: Increase in number of site visits over sample period due to small changes in variable*

	Increase advertising by one dollar	One more media mention	Searches take one second less on average	Searches repeated 1% less often	Links used 1% more often
Altavista	3.70	13,137	6761	352,195	175,599
AOL	3.52	1,296,860	8088	329,865	3,154,163
Excite	12.60	164,221	8551	399,711	3,532,417
Iwon	0.0160	7819	2385	116,384	532,908
Lycos	1.91	18,951	4134	310,342	1,143,310
MSN	11.65	630,456	30,042	1,676,960	16,626,927
Netscape	8.42	3,344,629	17,572	873,186	9,856,314
Yahoo	15.02	1,780,132	48,501	1,368,822	160,735

*Assumes 43.3 Million total online households. This is Plurimus' estimate of the total number of online households in February 2000