INTERPRETING INTERNET CLICKSTREAM DATA

By

Avi Goldfarb*

Northwestern University

a-goldfarb@northwestern.edu

First Version: September 26, 2000
This Version: March 26, 2002

Abstract

In this paper, I use a survey of 57 individuals to inform future analysis of clickstream data. Respondents performed four search tasks and answered several questions about their Internet habits. I use their responses to determine how to interpret the raw clickstream data in other papers. This paper was written as a chapter of my doctoral dissertation and is intended as a supplement for other papers, rather than as a stand-alone work. In particular, it is relevant to "Analyzing Website Choice Using Clickstream Data" and "Using Household-Specific Regressions to Estimate True State Dependence at Internet Portals"

The growth of the Internet has provided economists, marketers, and statisticians with a mountain of new data to analyze. One prevalent but relatively underused example of such data is clickstream data. This data format consists of each website visited by a panel of users and the order in which they arrive at these 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 are Netratings Inc., MediaMetrix Inc., and Plurimus Corporation. This paper provides a

* This research was supported by the Social Science Research Council through the predissertation fellowship of the Program in Applied Economics. I would like to thank Shane Greenstein, Charles Manski, and Robert Porter for helpful suggestions. Correspondence to: a-goldfarb@northwestern.edu

guide to several issues that arise when analyzing clickstream data, with a focus on data about Internet portals.

In both academic and business studies, clickstream panels are generally used to provide static analysis: the market shares of the various websites, the average length of time that a user spends at a given website, the demographics of a website's users, etc. These statistics, while interesting, do not fully exploit the wealth of information provided by panel data. There are, however, several difficulties that need to be overcome in order to analyze the data. Plurimus Corporation's data, for example, lists arrival and departure times for each website for each household in their sample. Interpreting the reasons people visit these websites and the experiences that they have from the raw data is difficult. This paper was done with a specific purpose in mind, although much of it applies more broadly. In Goldfarb (2002a) and Goldfarb (2002b), I build a model of individual Internet portal choice based on a data set provided by Plurimus Corporation. For each member of their panel, Plurimus records every website visited and the time of arrival at and departure from that website. Table 1 provides a sample of the raw data. This paper seeks to answer seven questions that relate to building models from the raw data. The first four apply to all clickstream data analysis, and the last three are Internet portal specific. The questions are:

1. Are website choices independent of each other?

2. Can individual preferences be aggregated, or does each individual's demand for a website have to be determined separately? In other words, can it be assumed that people have the same coefficients on the variables included in the study?

3. Which variables can proxy experience?

4. Do women and men have considerably different habits?

5. Which variables are relevant to an analysis of Internet portal competition?

6. Is faster search more desirable?

7. How can the researcher determine whether an individual's search fails?

Table 1.1
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	28MAR00:13:15:26	28MAR00:13:19:39	70732	11988	9
1	gov.nara	28MAR00:13:19:39	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

In order to answer these questions, I conducted an email survey of fifty-seven individuals. Each individual was asked to complete four search tasks and to answer questions about the searches. The survey (a copy of which has been included in the Appendix) also included several questions about Internet usage. I have considerably more information about the actual search choices of these individuals than is contained in a typical clickstream data set. This paper will show how to use this information to better interpret raw clickstream data.

Using surveys to inform data interpretation is relatively rare in economics. Helper (2000) asserts that economists should use more surveys and field research in order to better understand data. She emphasizes that this type of research "allows exploration of areas with little preexisting data or theory." Clickstream data definitely falls into this category. Manski (2000) recommends questionnaires to elicit agent's preferences and expectations directly. Jaffe, Trajtenberg, and Fogarty (2000) use surveys to determine whether patent citations are a good proxy variable for communication. They use a survey to determine how to interpret a data set. In this paper, I also use the rich information provided in a survey to determine how to use clickstream data to proxy other variables such as a successful search.

Using surveys to inform econometric modeling is a more common methodology, often used in experimental economics. Fischer and Nagin (1981) use surveys of parking preferences to compare individual utility estimation and panel estimation. They conclude that panels are better predictors of behavior.

The next section describes the survey methodology. Each of the seven questions is then answered in turn. Finally, I provide a brief conclusion.

2 Methodology

The survey was sent to each participant as an email attachment in Microsoft Word format. In the accompanying email, I explained that I am a doctoral student in economics studying Internet habits. Respondents came from two groups. The first group, henceforth referred to as the 'spammed' group, consists of the thirty-four respondents to unsolicited email. I sent two waves of five hundred unsolicited emails to addresses available in Yahoo's white pages directory. The addresses in this directory are either registered by the owner of the account or they are purchased from data services. Most addresses are from Hotmail and YahooMail, although some university accounts, other websites, and independent service provider accounts are represented. The first wave was sent on June 5, 2000. For each letter in the alphabet, except X, I started at the top and sent emails to every third American until I had sent twenty emails. The second wave, sent on June 29, 2000, was chosen similarly except that I started at the bottom of each letter. In the second wave, I included X and excluded Q. For the few cases where there were not enough addresses for that letter in the directory, I added addresses from the more common letters: A, B, and M.

The second group of respondents consisted of twenty-three 'friends of friends'. After receiving a response rate of roughly three percent, I decided to augment my numbers by asking several friends to forward the survey to their mailboxes. When there is sufficient data, I present results in this paper for the thirty-four 'spammed' respondents and for the fifty-seven in the total sample.

Clearly, this is a biased sample. With a 3.4% response rate to the random sample, it is likely that this group has some characteristics that non-respondents did not. For

example, it is likely that those who responded are more experienced Internet users, and were comfortable with the online survey format. They may also be more likely to have a respect for graduate research since they are responded to a survey with no reward promised (this may explain the extraordinary number with graduate degrees). Furthermore, not every Internet user is listed on Yahoo. Users who are listed may also be skewed toward certain characteristics. The 'friends of friends' group is biased for the obvious reason of being indirectly connected to me. Table 1 shows descriptive statistics of the demographic characteristics of the total data set and of the spammed respondents. The sample is younger and better educated than the general Internet population.

Table 1
Respondent demographics

	#	%		if spam, #	if spam, %
Spam	34	59.65%		34	100.00%
Male (Part 3 Ques. 1)	30	52.63%		13	38.24%
Student (Part 3 Ques. 4)	15	26.32%		4	11.76%
Modem (Part 3 Ques. 5)	19	33.33%		11	32.35%
Age (Part 3 Ques. 2)					
<18	1	1.75%		1	2.94%
18-25	20	35.09%		5	14.71%
26-35	25	43.86%		19	55.88%
36-50	7	12.28%		5	14.71%
51-65	4	7.02%		4	11.76%
>65	0	0.00%		0	0.00%
Education (Part 3 Ques. 3)					
Some high school	1	1.75%		1	2.94%
High school diploma	0	0.00%		0	0.00%
Some college	5	8.77%		1	2.94%
College degree	21	36.84%		14	41.18%
Some graduate school	8	14.04%		4	11.76%
Graduate degree	22	38.60%		14	41.18%
TOTAL	57			34	

These biases, however, do not imply that this research will be useless. The survey results are informative about individual surfing habits. It is common practice in psychology and in experimental economics to draw candidates from undergraduate classes, and then to use this information to inform theory. This study is no different. It uses information about the habits of a biased sample to explore results that should not depend on this bias. I am interested in modeling individual habits, not in determining collective truths. By observing a biased sample of people, I can follow the search process more closely than I can with a broader sample. I will then apply these results to the relatively unbiased clickstream data. I am therefore assuming that the bias does not significantly affect the answers to the seven questions I ask.

The survey itself asks respondents to search for driving directions, medical information, an MP3, and something of their own choosing. Respondents then answered several questions about the searches (see the Appendix). The search tasks were chosen to be diverse and to reflect common search activities.

2 Results

In this section, I will begin with a basic model of individual behavior. I will then add features to the model as a result of the answers to the seven questions asked in the introduction. The basic model is as follows: An individual, *i*, will choose website *j* at time *t*, if and only if the utility gained from that website is greater than the utility gained from any other choice at that time. Formally, if $U_{ijt}(X_i,W_{ijt},Z_{jt}|S_{it}) \geq U_{ikt}(X_i,W_{ikt},Z_{kt}|S_{it})$ for all $k \neq j$ then individual *i* chooses website *j* at time *t*, where $U_{ijt}()$ is the utility function, X_i are individual *i*'s personal time and product invariant characteristics, W_{ijt} are individual *i*'s time and product variant characteristics for product *j* at time *t*, Z_{jt} are product *j*'s characteristics at time *t*, and S_{it} is the state of the world faced by individual *i* at time *t*. The purpose of the seven questions is to determine which variables make up Xi, W_{ikt}, Z_{kt}, and S_{it} and how to measure them given the data available.

2.1 Are website choices independent of each other?

The immediate answer to this question is no, they are not independent. Past choices of websites by an individual are highly correlated with current choices. In this survey, people who used a given search engine during the first task were more likely to use it on subsequent tasks than were the other respondents. The more interesting question is how to treat two search engine visits during the same online session. In other

words, is the choice of search engine different if an individual has already performed an unsuccessful search?

To further understand the issues, consider a large firm that hires people for several jobs (just as an individual performs several searches over time). Does rejecting an applicant for a given job affect the probability of the next applicant to get hired? The answer will depend on several factors such as the number of applicants, whether both individuals applied for the same job, and the urgency of filling the positions. While overall hires may be correlated due to an overall company culture or to learning over time, decisions to hire for a given job may or may not be further correlated. When looking at website choice, the choices may be correlated over time due to individual preferences or learning over time; however, having visited a website in that category that day may or may not be further correlated with the next choice.

Out of 151 different tasks attempted using search, twenty nine of them (19%, 22% of spammed) used more than one search engine. Therefore using more than one engine for a given search is frequent enough to merit consideration in the model. Furthermore, in response to Part 2 Question 10C, 68% of respondents (76% of spammed) say they use search engines other than their favorite because a search already failed on their favorite. In other words, 68% of respondents assert that they are more likely to use a search engine that is not their favorite when a search has already failed at the engine they prefer.

This suggests that searches within a given session are correlated. Therefore the number of searches already in that session should be included in the analysis whenever possible. I include this variable in Goldfarb (2002a).

2.2 Can individual preferences be aggregated, or does each individual's demand for a website have to be determined separately?

Fischer and Nagin (1981) use survey data to explore whether taste parameters vary across individuals. If taste parameters do vary across individuals, then the standard panel data methods are not applicable. Either each individual's utility function will have to be estimated separately or a distribution must be assumed for the coefficients. This survey does not formally test whether taste parameters vary; it does however explore whether different people claim different factors to be more important in choosing a website.

In the survey, Part 2 Question 9 asks which search engine respondents use most often and why. The answers were surprisingly consistent. Thirty-five respondents (21 spammed) cited previous experiences, eighteen claimed it was due to habit (11 spammed), and four (2 spammed) said it was because their email accounts were on that search engine. There is still enough variation, however, to suggest that some information will be lost due to aggregation.

The survey presents no overwhelming evidence one way or the other about the assumption of constant taste parameters across individuals. Both possibilities should be considered when analyzing the data. In Goldfarb (2002a), I use constant taste parameters. In Goldfarb (2002b), I allow the taste parameters to vary.

2.3 Which variables can proxy experience?

Most clickstream data providers, including Plurimus, do not collect user information from the first day a user goes online. In addition to initial conditions, another potentially important piece of data that is lost is user experience online. Potential

proxies of this experience, however, do exist in the data. These proxies are hours online per week and variety of websites visited. Using the survey data, I examine whether hours online per week and variety of websites visited (through the number of tasks previously attempted) are correlated with years of experience and the users' perception of their own experience.

The potential measures of experience that can be easily derived from clickstream data are the number of tasks that people had done before (Part 1 Question 2) and the number of hours spent online per week (Part 2 Question 2). These two variables potentially proxy both the number of years online (Part 2 Question 1) or the respondents' own assessment of their experience (Part 2 Question 3) which are unobservable to the researcher that uses clickstream data. Tables 2 and 3 show correlations between the various potential experience proxies. In these tables, the *used before* variable is the number (from zero to three) of tasks that the respondent had tried before; *years* takes a value of 1 if the person has before on the Internet for less than a year, 2 if the person has been online from one to two years, 3 if the person has been online from two to five years, and 4 if the person has been online more than five years; *hours/week* takes a value of 1 if the person is online less than one hour per week, 2 if one to three hours, 3 if three to ten hours, and 4 if more than ten hours; *own opinion* takes a value of 1 if the person claims to be not very experienced, 2 if the person claims to be somewhat experienced, and 3 if the person claims to be very experienced.

Table 2
Correlation coefficients between experience measures for all respondents

	Use before	Years	Hours	Own Opinion
Used before	**1**			
Years	**0.161**	**1**		
Hours	**0.255***	**0.389*****	**1**	
Own Opinion	**0.312****	**0.323****	**0.439*****	**1**

*** 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

Table 3
Correlation coefficients between experience measures for spammed respondents only

	Use before	Years	Hours	Own Opinion
Used before	**1**			
Years	**0.122**	**1**		
Hours	**0.291***	**0.471*****	**1**	
Own Opinion	**0.409****	**0.249^**	**0.444*****	**1**

*** 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

For both all users and the spammed user subset, hours online per week are significantly correlated with both years of experience and perceived experience at the one percent level (where $\rho \approx 0.4$ in all cases). The variety of search task previously undertaken is only correlated with perceived experience at the five percent level and is not significantly correlated with years online, even at the ten percent level in a one-tailed test. Therefore, hours online is a good experience proxy as it is correlated with both types of experience measured here. Variety of websites visited, on the other hand, is a poor experience proxy. I use hours online as a proxy for experience in Goldfarb (2002b).

2.4 Do women and men have considerably different habits?

This question is particularly important for Plurimus' data because they do not have the demographic characteristics of their users. They do, however, have data about each household's census block. Therefore they have good approximations of income, household size, whether the home is rented or owned, and education, but they have no information about the gender of the users. Therefore, if men and women have different search habits, statistical analysis should rely on individual-specific time-invariant factors. For example, using individual fixed effects, rather than random effects, would eliminate the influence of demographic variables. Conditional logit instead of multinomial logit would serve the same purpose in a different setting.

Table 4
Survey results by gender

	men	women
% of tasks completed	87.50%	79.60%
Time	22.85	19.4
Favorite search engine is same as start page	23.33%	7.41%
Online goals (Part 2 Ques. 4)		
Several	40.00%	40.74%
One + email	36.67%	51.85%
One	13.33%	7.41%
None	10.00%	0.00%
Search methods (Part 1 Ques. 3)		
Bookmark	11.36%	9.32%
Typed in address	27.27%	23.73%
Search engine keyword	56.82%	52.54%
Search engine link	4.55%	10.17%
Other	0.00%	4.24%
How get to engine (Part 1 Ques. 4)		
Bookmark	33.00%	21.11%
Typed in address	55.00%	66.67%
Link	12.00%	12.22%
# engines used to complete each task (Part 1 Ques. 4)		
None	35.00%	32.41%
One	53.33%	53.70%
Two	5.83%	12.04%
Three	5.00%	1.85%
Four	0.83%	0.00%
N	30	27

Table 4 shows survey results by gender. There are few significant differences between men and women. Men take slightly longer on average to search, but they have a higher success rate. Men appear to be much more likely to personalize their web browser to help search. They are more likely to use bookmarks to arrive at search engines and they are more likely to have their browsers start at their favorite search engine. No women go online just to surf while ten percent of the men do. Overall however, these differences are not large. The lack of gender differences suggests that the choice of

random or fixed effects and the choice of conditional or multinomial logit should be determined by other aspects of the data set involved.

2.5 Which variables are relevant to an analysis of Internet portal competition?

This question seeks to determine the components of X_i, W_{ijt}, Z_{jt}, and S_{it}. Age should be included as it is highly correlated with the search time. I found no large correlations between education and habits. The above section suggests gender does not need to be included. By all measures of experience, more experienced individuals use a wider variety of search engines. Therefore, experience is an important factor. Most panel data sets, including Plurimus', do not, however, have experience data. Hours online per week should be included as a variable for analysis.

Bookmarks do not seem to be an important factor in search engine choice. In response to Part 2 Question 8, respondents listed a total of 258 bookmarks. Only thirteen were search engines, and the twelve individuals with these thirteen bookmarks were only slightly more likely to use the bookmarked websites. Start-up pages were, however, much more important. Sixteen percent (18% of spammed) have their most often used search engine as their start page. This suggests that start pages play a significant role in search engine choice and should therefore be included.

Another important variable that the survey suggests should be included is the goal of search. In response to Part 2 Question 10C, fifty-eight percent of respondents (59% of spammed) say that they use different search engines because "Different search engines are better suited to different tasks". In response to this question, other respondents said links and location-specific information matter. If information on these is available, it should be included. Unlike the search goal, they do not appear to be essential to

estimation. Surprisingly, aside from email, specific search engine features were not cited as important factors in choice. Goldfarb (2002a) and Goldfarb (2002b) apply this advice.

2.6 Is faster search more desirable?

If, as suggested in section 2.5, past experiences at a website are important for future choices, then it is necessary to try to determine proxies for measuring the quality of past experiences. This section will focus on the length of time past searches have taken for that individual at that website, while the next section will focus on defining whether a given search is successful.

In this section, I explore whether faster search is desirable. It may not be. It is possible that people prefer Internet portals that allow for more depth in search than those that find a given website more quickly. In order to answer this question, I examine whether more experienced users take less time to search than less experienced users. There are two fundamental assumptions that allow me to conclude that if more experienced users take less time then faster search is better. The first is that more experienced users are better at using the Internet. Practice has allowed them to improve their skills and do what they wish to do. Therefore if we observe them searching more quickly, it is because they want to spend less time searching rather than searching in more detail. The second assumption is that more experienced users are not inherently faster searchers than others, controlling for their learning from experience. I believe both of these assumptions are reasonable.

I use four different measures of experience and two measures of time to determine this correlation. As described above, the measures of experience are the number of tasks that people had done before (Part 1 Question 2), the number of years online (Part 2

Question 1), the number of hours spent online per week (Part 2 Question 2), and the respondents' own assessment of their experience (Part 2 Question 3). The time measures are the time spent and the time spent per successful search. Fifty three respondents recorded the time spent and, since one respondent had no successful searches, there are fifty-two respondents for analysis of time spent per successful search.

Table 5 shows correlations between the various measures of experience and time. I present regression results in Table 6. As expected, both time and time per success are negatively correlated with the experience variables. This suggests that less time spent searching is better.

Table 5
Correlation coefficients between time and experience

	Time	Time per success	Time (spammed)	Time per success (spammed)
Years	-0.229*	-0.274**	-0.143	-0.200
Own Opinion	-0.338**	-0.380***	-0.382**	-0.555***
Hours	-0.175	-0.212^	-0.323*	-0.372**
Used Before	-0.232*	-0.328**	-0.268^	-0.293^

*** 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 6 shows the results for regressing time on dummies for the various experience variables. Using time per success rather than time yields similar qualitative results. Only the results on having used an MP3 before and on considering oneself to be 'very experienced' are significantly negative; although most other results are negative with coefficients ranking in the expected order and no result is significantly positive. While not as clear as Tables 5, Table 6 also suggests that less time spent searching is better. Therefore a weighted measure of the past time each individual spent searching at

each website should be included to give a measure of expected quality. This is done in Goldfarb (2002a) and Goldfarb (2002b).

Table 6
Regressing time on experience (standard errors in parentheses)

	Model (1)	Model (2)	Model (3)	Model (4)
Used Map Before	1.40 (3.18)			
Used Health Before	-0.283 (3.49)			
Used MP3 Before	-7.83*** (-3.06)			
2-5 Years Experience		6.43 (7.01)		
>5 Years Experience		-1.37 (7.25)		
1-4 Hours Online/Week			-1.44 (5.09)	
4-10 Hours Online/Week			-1.47 (4.52)	
>10 Hours Online/Week			-3.90 (5.05)	
Self-identify as Very Experienced				-8.79*** (2.61)
<25 Years Old		-8.65*** (3.17)	-7.91** (3.67)	-8.46*** (3.09)
Spam	-4.22 (3.28)	-8.56*** (3.20)	-8.50** (3.61)	-8.44*** (2.99)
Modem	2.16 (3.02)			
Constant	24.64*** (3.21)	25.06*** (7.41)	30.56*** (5.15)	32.05*** (2.89)
df	47	48	47	49
R^2	0.171	0.267	0.165	0.312

*** 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

2.7 How can the researcher determine whether an individual's search fails?

Ideally each person would only conduct one task during each online session. Therefore if the researcher observes the individual go to a search engine and then to a website without searching again, then it would be reasonable to assume the search was

successful. In this scenario, if the researcher observes the individual search again after going to the website then the search would appear to have been a failure. Responses to Part 2 Question 4 reject this ideal situation. Just 10.5% of respondents (11.8% of spammed) only do one thing when online. Another 43.9% (41.2% of spammed) say they usually use email plus one other thing. More than 45% of the respondents either perform several tasks or have no specific task in mind, considerably complicating the definition of a failed search.

The group with no specific task in mind makes up only five percent of respondents (6% of spammed). Defining how they search and the reasons for it are beyond the scope of this survey. Much more important is controlling for the more than forty percent of total respondents (and spammed respondents) who do several tasks when they go online. One way to do this is to compare the goals of searches that occur during a given session. If the goals are the same, it is more likely that they are part of the same search task. Also, the elapsed time between searches and the number of websites seen between the visits to search engines may be relevant.

Therefore, if people search twice for the same thing in a short period of time, it seems reasonable to assume that the first search was a failure and the second a success. This relies on one further assumption: that people do not go to the destination website from the search engine by typing in the name of the website. They only use links from the search engine page. Only 5.8% of 155 searches (4.7% of 85 for spammed) were followed by the use of a non-search website that was not the final destination. This means that using the above method, over ninety-four percent of searches labeled as

successful would in fact have been successful. While this is not perfect, it seems to be a reasonable measure.

This is only one of several possible ways a failed search could be observed. If a person goes directly from one search engine to another then the visit to the first website is likely a failure. Furthermore, if a person does not go to a new website after visiting a search engine, then the visit may have been a failure. In creating the variables, it is important to distinguish between the possible types of failure because some may be more reliable proxies than others.

In Goldfarb (2002a) and Goldfarb (2002b), I use the following method to proxy search failure (in Goldfarb (2002a), I also experiment with another measure). I cannot identify search failure exactly, but I proxy it with a *repeated search* variable. If a household visited two portal websites 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 (at the same website 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 website were repeated that matters. Also as with time spent, more complicated functions of past repeated searches do not yield qualitatively different results. In Goldfarb (2002b), I only identify a search as repeated if it occurred in a previous session. This avoids confusion over the use of a browser's back button. I call this variable *last search repeated*.

3 Conclusion

This chapter has used a survey to examine some of the assumptions needed to analyze clickstream Internet panel data. By observing people's search habits directly and by asking several questions, this chapter has determined which assumptions can be made and which are questionable. It has also suggested several variables to include in analysis and how to structure some of the econometric analyses when studying Internet portals.

References

Fischer, Gregory W., and Daniel Nagin. 1981. "Random versus Fixed Coefficient Quantal Choice Models." In *Structural Analysis of Discrete Data with Econometric Applications*, ed. Charles F. Manski and Daniel McFadden, 273-304. Cambridge, MA: The MIT Press.

Helper, Susan. 2000. "Economists and Field Research: "You Can Observe a Lot Just by Watching." *American Economic Review Papers and Proceedings* 90: 228-32.

Goldfarb, Avi. 2002a. "Analyzing Website Choice Using Clickstream Data." Northwestern University. Mimeographed.

Goldfarb, Avi. 2002b. "Using Household-Specific Regressions to Estimate True State Dependence at Internet Portals." Northwestern University. Mimeographed.

Jaffe, Adam B., Manuel Trajtenberg, and Michael S. Fogarty. 2000. "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors." *American Economic Review Papers and Proceedings* 90: 215-18.

Manski, Charles. 2000. "Economic Analysis of Social Interactions." *Journal of Economic Perspectives* 14: 115-36.

APPENDIX : QUESTIONNAIRE

Thank you for giving me a few moments of your time. I am conducting this survey as part of my Ph.D. dissertation on Internet surfing habits.

This survey consists of three parts. Part 1 asks you to complete some simple tasks and then to answer questions about them. Part 2 asks questions about your Web usage. Part 3 asks for some background information.

PART 1: ONLINE TASKS

Complete each of the following four tasks. Please tell me when you start and finish each task. After completing each task, answer the questions relating to it.

The tasks are written at the bottom of each page.

START HERE

What time is it? _____

TASK ONE:

Find driving directions from Chicago to New York.

Task One questions:

Note that "search engines" include the following websites (in alphabetical order):

about.com, altavista, aol.com, ask.com/askJeeves, Britannica.com, Canada.com, dogpile, excite, go.com, google, goto, go2net, hotbot, infoseek, iWon, looksmart, Lycos, mamma.com, metacrawler, msn.com, netscape.com, Northern Light, search.com, snap.com, sympatico, webcrawler, and Yahoo!

1) If you found a site that gives the directions, what is the name of the site?

☐ Site name: _____ ☐ I didn't find one

2) Have you ever searched for directions online before?

☐ Yes ☐ No

3) How did you find this site? If you tried more than one method, please say the order in which you tried them (by marking 1, 2, 3, …, in the space provided).

___ a) With a bookmark on your computer direct to a travel or map site

___ b) By entering the web address of an online travel or map site that you know offhand

___ c) By looking up a keyword in a search engine

___ d) Other use of a search engine site. Please specify use _____

___ e) Other. Please specify _____

4) If you used a search engine, please state in order which search engine(s) you used and mark how you arrived at each one. Otherwise please go to the next task

	Name	Bookmark	Typed in address	Link from other page
Example	*Quicksearch*		*Yes*	
1)				
2)				
3)				
4)				

TASK TWO:

Find the definition of 20/20 vision.

Task Two questions:

1) If you found a site that gives the definition, what is the name of the site?

☐ Site name: _____ ☐ I didn't find one

2) Have you ever searched for medical information online before?

☐ Yes ☐ No

3) How did you find this site? If you tried more than one method, please say the order in which you tried them (by marking 1, 2, 3, …, in the space provided).

___ a) With a bookmark on your computer direct to an online health or information site

___ b) By entering the web address of an online health or information site that you know offhand

___ c) By looking up a keyword in a search engine

___ d) Other use of a search engine site. Please specify use _____

___ e) Other. Please specify _____

4) If you used a search engine, please state in order which search engine(s) you used and mark how you arrived at each one. Otherwise please go to the next task

	Name	Bookmark	Typed in address	Link from other page
Example	*Quicksearch*		*Yes*	
1)				
2)				
3)				
4)				

TASK THREE:

Find a recording of Hey Jude by The Beatles (i.e. an MP3).

Task Three questions:

1) If you found a site that has the recording, what is the name of the site?

☐ Site name: _____ ☐ I didn't find one

2) Have you ever searched for an online recording before?

☐ Yes ☐ No

3) How did you find this site? If you tried more than one method, please say the order in which you tried them (by marking 1, 2, 3, …, in the space provided).

___ a) With a bookmark on your computer direct to an online music site

___ b) By entering the web address of an online music site that you know offhand

___ c) By looking up a keyword in a search engine

___ d) Other use of a search engine site. Please specify use _____

___ e) Other. Please specify _____

4) If you used a search engine, please state in order which search engine(s) you used and mark how you arrived at each one. Otherwise please go to the next task

	Name	Bookmark	Typed in address	Link from other page
Example	Quicksearch		Yes	
1)				
2)				
3)				
4)				

TASK FOUR:

What do you use search engines for most often? Find something that you regularly look for on search engines.

Please specify what you will look for _____

Task Four questions:

1) If you found a site that gives the information you want, what is the name of the site?

☐ Site name: _____ ☐ I didn't find one

2) How did you find this site? If you tried more than one method, please say the order in which you tried them (by marking 1, 2, 3, …, in the space provided).

___ a) With a bookmark on your computer

___ b) By entering a web address that you know offhand

___ c) By looking up a keyword in a search engine

___ d) Other use of a search engine site. Please specify use _____

___ e) Other. Please specify _____

3) If you used a search engine, please state in order which search engine(s) you used and mark how you arrived at each one. Otherwise please go to the next task

	Name	Bookmark	Typed in address	Link from other page
Example	Quicksearch		Yes	
1)				
2)				
3)				
4)				

4) What time is it? _____

PART 2: WEB HABITS SURVEY

Please answer the following questions by putting a check beside the appropriate answer.

1. When did you start using the World Wide Web?

- ☐ In the last year
- ☐ From one to two years ago
- ☐ From two to five years ago
- ☐ More than five years ago

2. How many hours do you use the Web per week?

- ☐ Less than one hour
- ☐ 1-3 hours
- ☐ 4-10 hours
- ☐ More than 10 hours

3. Do you consider yourself?

- ☐ Very experienced with the Web
- ☐ Somewhat experienced with the Web
- ☐ Not at all experienced with the Web

4. Which of the following best describes your Web habits?

When I go online and do not *just* check my email:

- ☐ I usually have several things to do
- ☐ I usually check my email and do one other thing
- ☐ I usually have only one thing to do
- ☐ I do not generally have a specific task in mind

5. At what page does your browser start? (i.e. what page do you see first when you go online?) _____

6. Which of the following best describes why you generally go to a website for the first time?

- ☐ I go in order to do a specific task such as purchase a product or find information
- ☐ I go in order to see what the site is like so that I can keep it in mind for future reference

7. If you have only been to a website one time, how long do you think it would take for you to forget what it is for?

☐ Less than one week

☐ 1-4 weeks

☐ 4-12 weeks

☐ More than 12 weeks

8. Which pages do you have bookmarked on your computer?

a) _____

b) _____

c) _____

d) _____

e) _____

f) _____

9. What search engine do you use most often? _____

Why?_____

_____

10A) Do you ever use other search engines?

☐ Yes ☐ No

10B) If you do use other search engines, Which ones?

a) _____

b) _____

c) _____

d) _____

10C) If you do use other search engines, why do you use them? (mark all that apply)

☐ A search already failed on the search engine I used most

☐ Different search engines are better suited to different tasks

☐ They're all the same, so I use whatever comes to mind

☐ Other (Please specify) _____

PART 3: BACKGROUND INFORMATION

Please answer the following questions by putting a check beside the appropriate answer.

1. Gender:

☐ Male ☐ Female

2. Age:

☐ < 18 ☐ 18-25 ☐ 26-35 ☐ 36-50 ☐ 51-65 ☐ > 65

3. Education (please specify the highest level achieved)

☐ Some high school ☐ High school diploma

☐ Some College ☐ College Degree

☐ Some Graduate School ☐ Graduate Degree

4. Are you a full-time student?

☐ Yes ☐ No

5. In doing this survey, were you connected to the Internet by a telephone modem or by a faster connection?

☐ Telephone Modem ☐ Faster Connection

THANK YOU VERY MUCH FOR YOUR TIME. Please hand your questionnaire to me.