Using Clickstream Data to Predict WWW Usage

by

Alan L. Montgomery
Carnegie Mellon University
Graduate School of Industrial Administration
5000 Forbes Ave.
Pittsburgh, PA 15213-3890

e-mail: alan.montgomery@cmu.edu

August 1999

Outline:

Abstract:

   The purpose of this study is to consider the determinants of World Wide Web (WWW) usage using clickstream data. Clickstream data is a natural byproduct of a user accessing WWW pages, and refers to the sequence of pages visited and the time these pages were viewed. A key component of this study is a large scale empirical analysis of clickstream data from a representative sample of PC owning households in the U.S. This sample comes from Media Metrix's PC Meter using a nationwide panel of over 5,000 households during 1997 and 1998. A primary finding of this research is that demographics are not a primary driver of web usage. While it is true that access to the WWW is influenced by demographic characteristics, the best way to predict current usage is to use information about past usage, and not use demographics. Additionally, we show there is a great deal of diversity in the usage of individual sites. These findings imply that if managers really want to learn how to fulfill the promise of interactivity on the WWW they need to learn how to capture and use their customers' clickstreams.

1 Introduction

The amount of time households spend browsing the World Wide Web (WWW) has received little attention. We know that overall growth has occurred in WWW usage, but there are many basic and unanswered questions about usage. Does individual usage exhibit positive growth trends similar to overall usage? What influences the amount of time a household will spend online? Are demographics predictive of usage? If not, are there other variables that can predict how much time a household will spend online? Are patterns in a user's monthly WWW viewership similar to those of individual web sites? The marketing mix elements that account for the increased household ownership of personal computers (PC), such as price, income, and technical innovations in the computer industry, have been thoroughly studied in principle and we can anticipate how they will impact computer ownership. In contrast, the factors of WWW usage have not been examined.

The purpose of this study is to answer these and other questions about WWW usage using a large-scale empirical analysis of clickstream data from a representative panel of PC owning individuals in the U.S.. Clickstream data refers to the sequence of WWW URLS or pages visited and the time these pages were viewed. This type of data is new to academic marketing research, and this research has been made possible by a clickstream dataset collected through Media Metrix's PC Meter. The dataset employed in this paper is comprised of a nationwide panel of over 48,000 individuals during June 1997 through November 1998. This data is discussed in further detail in section 2. Prior research about the WWW has provided valuable conceptual frameworks (Hoffman and Novak 1996, Pant and Hsu 1996), but empirical analyses have been lacking. The empirical studies that have been conducted on the WWW have been based upon self-reports of usage from surveys (Clemente 1998, Hoffman et al. 1996, Pitkow and Kehoe 1996, O'Reilly 1995). An important advantage of clickstream data is that it represents actual WWW usage data and does not rely upon self-reports of usage.

In this paper we use the term usage to refer to the amount of time users spend actively viewing web pages. Usage is an important measure of web activity for many reasons. First, web usage is used as an indicator of the success of a web site. Web managers closely monitor site usage through programs that can analyze their web log files (Niccolai 1997), as well as subscribe to services that track competing sites (Green 1998). Second, usage is a measure of exposure, which is a frequently employed measure in advertising

research. Many sites generate a large portion of their revenues from advertising, which means web usage can be a direct measure of profitability. Finally, usage may also provide additional information about household decisions to buy a new or additional PCs, i.e., one would expect a household will be more likely to buy a computer if they anticipate a high degree of use.

The outline of this paper is as follows. In section 2 we outline the possible sources of clickstream data and briefly discuss the merits and disadvantages of each. Section 3 presents a model from which we can analyze individual usage of the WWW. In section 4 we examine the application of this model to understanding the patterns of monthly WWW usage. We conclude in section 5 with a discussion of our findings. One primary finding of this research is that demographics are not a primary driver of web usage. While it is true that access to the WWW is influenced by demographic characteristics, the best way to predict current usage is to use information about past usage, and not use demographic characteristics. This implies that if managers really want to learn how to fulfill the promise of interactivity on the WWW they need to learn how to capture and use their customers' clickstreams. This represents both a mandate for managers to learn how to collect this new data set, but also requires further study in academics on how to extract summary measures from this new and massive data source.

2 Clickstream Data

A primary source for clickstream data is to monitor a user's browser window and record when the browser window is in focus and what URL is currently being viewed (Catledge and Pitkow 1995). This is the technique that is employed by Media Metrix, the source of clickstream data for our paper. A key advantage of the WWW for marketing purposes is that clickstream data is a natural byproduct of a user accessing WWW pages, and can be generated automatically without requiring interaction with the user that could alter his browsing behavior. However, this does not mean that collecting clickstream data is not without cost. The amount of data that can be collected can quickly make clickstream data cumbersome to manage and costly to analyze. Additionally, the cost of creating random samples of nationally representative WWW users and maintaining a panel of users quite costly.

Besides collecting clickstream data at the source there are two other potential sources for clickstream data which become apparent if we consider the sequence of events that occur when a user opens a URL in

a browser window on their PC. The browser issues a request to the server on which the URL is located. Along with the page request, the browser may provide other information about a user's PC, such as a user's domain address, operating system, browser type and version, and previous page request[1]. However, the browser request itself does not go directly to the server on which the requested page resides. Instead, the request is passed along through a network of computers until it reaches the desired host. Most frequently the initial computer in this chain is an Internet Service Provider (ISP), for example AOL. The ISP does not always pass the request on, but may instead satisfy the request using a locally stored cache, as a way of lessening traffic on the Internet. As the ISP processes these requests, it may also record them to measure the user's clickstream. If the request is passed onto the server, this server may also record the computer that generated the request. This provides two additional sources for collecting clickstream data.

To illustrate the clickstream consider an actual user session is listed in Table 1. Here the user started at their ISP, *voicenet.com*, and then proceeded to *weather.com* and *astrology.com*. Notice that the user frequently views the same page multiple times and may pause between pages. Also notice, that in this case none of the pages have links, which means that the user has navigated to these URLS by relying upon their bookmarks and history. This data also illustrates why it is important for the data to be collected at a user's machine and not from the host site. Notice that only the italicized URLS might occur in the logs of the host server, which may lead to biased assessments of usage if only host server logs are used. The detailed level of information that is collected shows that the clickstream can be costly to capture and analyze, certainly it is not cost free.

The clickstream data used in this paper comes from the PC Meter created by Media Metrix. The PC Meter is a Microsoft Windows or Apple Macintosh program that records information about who is using the computer, what applications are running or URLS are being viewed, and how long they are in use (PC Meter 1997). This program runs in the background, and is virtually transparent to the user. The PC Meter program is distributed on a monthly basis by Media Metrix to a nationwide panel of about 10,000 house-holds[2]. At the end of the month households return a disk that contains their monthly usage. These house-

---

1. To see what information is conveyed by a seemingly "anonymous" connection refer to http://privacy.net/anonymizer/.
2. The data from this study comes exclusively from Media Metrix's at home panel. There is also an at work panel. The current size of the at home panel is now in excess of 20,000 panelists.

holds are recruited largely through direct mail and represent a random sample of U.S. PC owning house-holds. To encourage households to participate they are given periodic gifts and prizes as incentives, although the nominal values of these gifts tend to be small in magnitude as with other panels. In addition, to the web usage data collected by the PC Meter, extensive demographic information about the household is collected through conventional surveys. In summary, the PC Meter represents a hybrid approach between online collection techniques and conventional mail panels.

In this research we measure monthly WWW usage by aggregating the total time actively browsing[3] as recorded in the clickstream. The time period for the sample of panelists used in this research are for a sixteen month span starting in July 1997 and ending in October 1998. This sample is representative of the population of WWW users from home during this time. There was an average of 18,950 panelists per month of whom 10,350 used the web on average[4]. Media Metrix is able to project this sample to the national level to forecast the total number of home WWW users, for example during October 1998 they estimate that there were 33.8 million users. If an individual connected at least once during the month to the WWW then that user spent an average of 3.6 hours actively viewing web pages per month. Active monthly WWW usage varies from less than one minute to a maximum of 210 hours, with a standard deviation of 7.1 hours. Clearly there is great deal of diversity in web usage. An alternative measurement of usage is the number of viewings or hits, which average 542.4 URLS per month during the same period. We concentrate upon active time viewing web pages since it better captures usage than viewings, but the correlation between the two measures is .86 so both measures are highly related.

3 Modeling Monthly WWW Usage

To better illustrate the clickstream data that is modeled in this research, several usage histories for six selected individuals are plotted in Figure 1. This monthly WWW usage data possess several pronounced features. First, the scale of usage differences greatly across users, indicating a log transformation would improve the comparability of usage across users by placing usage on a relative scale. Secondly, the observed

---

3. We use active viewership to mean that the WWW browser containing the page being viewed is the currently selected application. Also, if the user has not touched their keyboard for 3 minutes then application is no longer considered active.
4. This yields a total of over 48,000 unique individuals in this panel for this time frame. The median number of months in the panel was 7 months.

history for an individual may be very short. As illustrated in Figure 1, user 2 actually used the WWW for only one month out of the four that she participated in the study. In total 37% of users are observed for one to three months, less than 30% are observed for more than twelve months. The short histories and high variability of usage for each individual makes it difficult to assess whether time trends are present in individual usage.

Finally, the data features many censored and missing observations. There are many users that exhibit no usage during a month, i.e., the observation is censored at zero. In total 40% of observations were observed as zero. Censoring occurs when a user participated in the panel but chose not to access the WWW, perhaps either due to lack of interest or lack of time. Obviously light users are more prone to censoring, but even heavy users, i.e., those with more than 3.2 hours, had 16% of their observations observed as zero. Notice that user 6 exhibits high usage during the last two periods, but during the preceding months exhibited little or no usage. A further facet of the data is that 14% of observations are missing. For example, user 1 was continuously present in the panel for 13 months, missed a month, and then resumed the following month. There is no information why this user did not participate in the panel.

The lack of causal information about why usage occurs during a specific month, means that we have to rely upon previous patterns in past usage to predict current usage. These types of autoregressive models have been shown to be quite versatile. The use of time series effects of models with binary information has been explored previously (Kedem 1980, Keenan 1982). A strength of the data that is employed in this study is that a great deal of demographic characteristics about individuals is known. This demographic information can be used to explain the variation in usage across individuals. Table 2 lists the demographic variables employed in this study along with their frequencies for the 16 month period of the data for this study. The demographic variables are coded as categorical data for this study.

The model that we employ to predict monthly usage (u_{it}) for the ith panelist during month t is a Type II Tobit Regression model (Nelson 1977, Amemiya 1985, pp. 385-389):

$$u_{it} = \begin{cases} u_{it}^{*} & if\ z_{it}^{*} > 0 \\ \\ 0 & otherwise \end{cases} \tag{1}$$

Where u_{it}^{*} is considered to be observed when the latent process z_{it}^{*} is positive, but unobserved when z_{it}^{*} is

negative. A Type I Tobit would equate the u_{it}^* and z_{it}^* processes. The reason that a Type I Tobit is not used is that this type of model would imply a smooth distribution of usage from 0 to large amounts. However, our observed data contradicts this property, since even heavy users occasionally stop using the WWW. A Type II Tobit model permits different patterns to exist for the question of how much versus whether to use the WWW, and does not require a smooth distribution of usage around 0.

The corresponding models for the u_{it}^* and z_{it}^* processes are:

$$\ln(u_{it}^*) = \phi_{i0} + \phi_{i1} \ln(u_{i,t-1}^*) + \epsilon_{it}, \quad \epsilon_{it} \sim N(0,\sigma_i^2) \tag{2}$$

$$z_{it}^* = \delta_{i0} + \delta_{i1}\ln(u_{i,t-1}^*) + \alpha_{it}, \quad \alpha_{it} \sim N(0,1) \tag{3}$$

Notice that these models could be extended to incorporate exogenous variables, like time spent on other leisure and work activities, if this data was available. We assume that u_{it}^* and z_{it}^* are independent and set the variance of z_{it}^* to unity to guarantee identification of this model.

The expected value of usage given the usage during the previous month is:

$$E[\, u_{it} \,|\, u_{i,t-1},\phi_i,\delta_i,\sigma_i \,] = \Phi\left(-\delta_{i0} - \delta_{i1}\ln(u_{t-1}^*)\right) \exp\left\{\phi_{i0} + \phi_{i1}\ln(u_{i,t-1}^*) + \frac{1}{2}\sigma_i^2\right\}$$

where Φ is the CDF of the standard normal distribution. The elasticity of usage with respect to last month's usage is:

$$\frac{\partial E[\, u_{it} \,|\, u_{i,t-1} \,]}{\partial y_{i,t-1}} \frac{u_{i,t-1}}{E[\, u_{it} \,|\, u_{i,t-1} \,]} = \phi_{1i} + \delta_{1i}\frac{\phi\left(-\delta_{0i}-\delta_{1i}\ln(u_{i,t-1}^*)\right)}{\Phi(-\delta_{0i}-\delta_{1i}\ln(u_{i,t-1}^*))}$$

where ϕ is the PDF of the standard normal distribution. Notice the elasticity of usage is comprised of two components. The first is due to an increase in the expectation of u_{it}^*, while the other results from the contribution of the truncation term associated with z_{it}^*. When $-\delta_{0i}-\delta_{1i}\ln(u_{it}^*)$ is negative, as would happen with an individual with low usage, the ratio of ϕ/Φ is large, while positive values of $-\delta_{0i}-\delta_{1i}\ln(u_{it}^*)$ drive the ratio of ϕ/Φ towards 0.

A good deal of demographic information exists that may help explain the variation in usage across individuals, therefore a second stage model using the parameters from (2) and (3) is employed:

$$\varphi_h = \theta' d_h + v_h, \ v_h \sim N(0, \Psi)$$
$$\delta_h = \omega' d_h + w_h, \ w_h \sim N(0, \Pi)$$

The benefit of structuring the model in this hierarchy is that the strength of the cross-sectional variation in user coefficients can be used to improve the parameter estimates of individual panelists. These types of hierarchical Bayesian model schemes have been successfully employed in other marketing problems involving panel data with limited observations in the context of scanner data (Allenby and Rossi 1993). This model is estimated using the Gibbs sampler, and is described in the Appendix.

4 Predictors of Monthly WWW Usage

We apply the model developed in the previous section to the dataset described in Section 2. The average parameter estimates across individuals for (6) are given in Table 3. The φ_1 and φ_0 parameters captures the question of how much the user will use the WWW, and the δ_1 and δ_0 parameters captures the user's propensity to access the web. Consider how the model captures the usage of the average user. The expected value of usage for the average user who accessed the WWW is .67 hours, $E[u_t | z_{ti}^* > 0] = \exp\{\varphi_0/(1-\varphi_1) + \frac{1}{2}\sigma^2/(1-\varphi_1^2)\} = 2400$ seconds. The probability the average user accesses the WWW is .648,

$$^{*} > 0] = 1 - \Phi\left(-(\delta_0 + \delta_1 \varphi_0/(1-\varphi_1)) / \sqrt{1 + \delta_1^2 \sigma^2/(1-\varphi} \right. $$ See Appendix B for a proof of these relationships.

Notice that in both cases the autoregressive relationship is weak. If distinct individual level trends been present we would have seen values of φ_1 and δ_1 close to 1, indicating more persistence and memory in usage patterns. An inference that can be drawn from this analysis is that any trends in aggregate WWW usage patterns are not coming from trends of individual users, but trends in the number of people accessing the WWW. In other words, the current increasing trend in aggregate WWW usage is coming from an increasing number of new users accessing the WWW. However, there is a fairly high correlation across users between φ_0 and φ_1 of -.55 and between δ_0 and δ_1 of -.87, which indicates the function of the intercept and autoregressive parameters are inversely related. Hence if heterogeneity in the intercepts across users was omitted from the model, much of the variation of the intercepts would be projected back onto the autoregressive parameters.

The variation of the individual level parameters illustrates the amount of heterogeneity across individuals. A natural question is whether this variation be explained through demographic and characteristics that describe the user? These issues are addressed directly in the second stage of the model as described by equations (6) and (7). The parameters from the first stage model for each household in (2) and (3) are regressed upon demographic characteristics for each user. The estimates for these relationships between the demographics and first stage model parameters are given in Table 4. We summarize the effects of each of these variables below.

Age: The effect of the age and gender are frequently cited as predictors of usage. The usual stereotype of the WWW user is a young male user. While these users are over represented in the WWW population, their usage patterns do not differ as dramatically from other groups. In Table 5a we can observe that younger users do have higher base levels of usage, as indicated by their effects on ϕ_0. However, these effects are not substantially different from zero except for the 35-39 age group in which we can attribute an increase in ϕ_0 of .41. Since the dependent variable is on a logarithmic scale we can think of this as being approximately representing a 41% increase in predicted usage over users in the under 25 group. Older users tend to show more persistence as demonstrated by the increasing values of ϕ_1. Hence, while younger users have a high propensity to spend time online, if we relate current usage to past usage we find that older individuals show more persistence in their usage levels.

A different pattern emerges when studying the effects of age on probability of accessing the WWW. Users in the under 25 age group are most likely to access the WWW, but individuals in their 30's and 40's have higher probabilities of not accessing the WWW. This indicates that factors that influence time spent browsing may not be the same as those that influence the decision to access the WWW.

Gender: Along with age, gender is one of the best predictors of usage. As expected we do notice a baseline usage of males both in terms of their base levels of usage, as well as their persistence in using the WWW. Additionally, males have a higher likelihood of accessing the WWW as exemplified by the low coefficients associated with the the δ_0 parameter.

Race: We find no effect of race upon usage of given that a user has a personal computer that the race is predictive of how much the user will access the WWW. This is some inconclusive evidence that would support the notion that blacks are less likely to use the WWW but again given the amount of measurement

error, we cannot access a statistically significant impact to this relationship. This finding conflicts with Hoffman and Novak (1998) who find an impact of race. A possible explanation of this discrepancy may that their data was from 1996—a year earlier than the data considered in this analysis, and any race based differences have diminished.

Household characteristics: Besides information about the user we also have information about the other individuals within the user's household. For example, the effects of children, household size, gender of the head of household, number of vehicles, and marriage status are all measured. For the most part these variables are not predictive of usage. The one set of variables that does have some effect on usage is the presence of children, more children results in lower usage, with the absence of older children having the least impact on an individual's usage.

Householder characteristics: Additionally we can consider the effects of the head of household on an individuals usage. Here we find that households with younger householders tend to show more persistence in whether they access the WWW. For example, if we know that a household is headed by an individual who is less than 40 we know that past usage is more predictive than households headed by older individuals. However, householder education and income is not predictive with one notable exception. If the house-holder opted not to report their income we notice a sizable increase in usage.

Location: Finally, we find the area and size of city in which an individual lives does not impact usage, except for users located in rural areas do show higher usage levels over those in urban areas.

Table 5 lists the amount of variation that is explained by these demographic variables. First notice that on average for the four different types of coefficients the demographics explain 20% of the variation across users. The demographics are most predictive of the constant for determining the probability whether a user will access the WWW (δ_0) accounting for 38% of total variation. The demographics are least predictive of the constant level in predicting usage (ϕ_0) accounting for only 6% of total variation. To better assess which demographics are most helpful in predicting usage we break the demographics into four different sets of predictors. We find that the age, gender, and race variables tend to account for the most variation, with household location accounting for the least. In general, we note that in general demographics are poor to fair predictors of usage. This corresponds well with findings of household purchase behavior that show demographics are poor predictors of purchase.

5 Conclusions and Future Research Directions

There has been much discussion about the World Wide Web (WWW) and its future for business both in academia and the general press. A key interest in the WWW is due to its ability to fulfill the potential of interactivity (Blattberg and Deighton 1991). A requirement of interactivity is the ability to monitor and respond to an individual customer's actions. Since the WWW satisfies these prerequisites it forms an excellent foundation for building interactive systems. However, the promise of clickstream data to implement interactive marketing systems is far from realized (Deighton 1996). This research has shown that the information stored within the clickstream has much greater predictive ability than demographics.

The collection of clickstream data is not without substantial cost. Clickstream data can generate huge amounts of data. The collection, processing, and analysis of this data can represent significant costs. While the incremental costs are small for operating web sites, they are quite substantial for panels such as the PC Meter. However, the clickstream generated by server logs do not contain the full information set that can be collected by at the user's site. Information about viewing, and how the user reached a site may be critical to better improving the design of a web site. Alternately identifying individual users using cookies also incurs costs in terms of processing and tracking the information, as well as potential user backlash since violations of stated privacy policies may negatively impact traffic.

This research has provided the first academic study of a representative sample of WWW users. As such a primary contribution is an analysis of a new dataset that is bound to become critical for implementing interactive marketing systems. We have presented a set of findings that we believe helps shed light on WWW usage. Namely, demographic characteristics are not predictive of clickstream behavior. Trends in usage do not appear to be present at the individual user level. If accurate predictions about web browsing are desired then information about the past clickstream is necessary. The finding that observed behavior is more predictive of future behavior than demographics is a common one in marketing research. Additionally, the prediction of individual level browsing behavior is possible, and actually quite good, due to the persistence in user's web browsing habits.

Many marketing researchers are likely to make a comparison between single source scanner data and clickstream data. Clearly clickstream data has possibilities that have not been possible or even considered

with supermarket scanner data. The promise of clickstream data is to allow collection of a complete purchase history, the consideration set, as well as a recording of the information viewed used to make a decision or selection was made. Additionally clickstream data allows the possibility of dynamically interacting with a user. However, the datasets collected in their present form are far from realizing this potential. A primary difficultly in the present study is the lack of measures about the goals an individual has during a web session, as well as a lack of measures about purchases or decisions that were made due to this session.

Perhaps a better analogy between the clickstream dataset explored in this paper is with telephone dialing or television viewership. Consider the data stream that would be generated from people making telephone calls. Specifically, we would have information about the phone number called, the time the call was made, and the duration of the call. The phone number can be cross-referenced in a phone directory to determine who was called. This information can be used to measure social networks and possible information flows. An analogous measure in the clickstream is the URL. However, there is a huge difference in the quantity of the information However, the vast majority of research about phone dialing has focused on not on what information can be inferred from these phone calls, but simply information describing the frequency and duration of phone calls.

Appendix A

To complete the specification of the model in a Bayesian form, priors are placed upon the following parameters: $\boldsymbol{\phi}_h, \sigma_h, \delta_h, \theta_h, \boldsymbol{\Psi}, \omega_h, \boldsymbol{\Pi}$. These priors are specified in natural conjugate form. Specifically, the following priors are used:

$$\frac{\nu_\sigma \lambda_\sigma}{\sigma_h^2} \sim \chi_{\nu_\sigma}^2 \,,\; \theta \sim N(\bar{\theta}, V_\theta)\,,\; \boldsymbol{\Psi}^{-1} \sim W(\nu_\phi, V_\phi^- 1)\,,$$

$$\omega \sim N(\bar{\omega}, V_\omega)\,,\; \boldsymbol{\Pi}^{-1} \sim W(\nu_\Pi, V_\Pi^- 1)\,,\; \ln(u_{h0}^*) \sim N(\ln(\bar{u}_0), V_{u_0})$$

where the parameters of these priors are specified with diffuse information relative to the data.

Appendix B

If we assume that the autoregressive equation in (2) which describes usage is stationary, i.e., $|\phi_1|<1$, then we can show that its mean and variance are:

$$E[\ln(u_t^*)] = \phi_0 + \phi_1 E[\ln(u_{t-1}^*)] \Rightarrow E[\ln(u_t^*)] = \frac{\phi_0}{(1-\phi_1)}$$

$$Var[\ln(u_t^*)] = \phi_1^2 Var[\ln(u_{t-1}^*)] + \sigma^2 \Rightarrow Var[\ln(u_t^*)] = \frac{\sigma^2}{(1-\phi_1^2)}$$

Since, $\ln(u_t^*)$ is a sum of normally distributed variates, we can deduce its marginal distribution:

$$\ln(u_t^*) \sim N\left(\frac{\phi_1}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2} \right)$$

In a similar manner we can deduce that the marginal distribution of z_t^* from equation (3) which describes the probability of access to the WWW is:

$$z_t^* \sim N\left(\delta_0 + \delta_1\frac{\phi_1}{1-\phi_1}, 1 + \delta_1^2\frac{\sigma^2}{1-\phi_1^2} \right)$$

References

Becket, Dave J. (1997), "30% accessible—a survey of the UK Wide Web", *Computer Networks and ISDN Systems*, 29, 1367-1375.

Blattberg, Robert C. and John Deighton (1991), "Interactive Marketing: Exploiting the Age of Addressability", *Sloan Management Review*, 33(1), 5-14.

Catledge, Lara D. and James E. Pitkow (1995), "Characterizing browsing strategies in the World-Wide Web", *Computer Networks and ISDN Systems*, 27, 1065-1073.

Chen, Bo, and Huifang Wang (1997), "A human-centered approach for designing World-Wide Web browsers", *Behavior Research Methods, Instruments, & Computers*, 29 (2), 172-179.

Deighton, John (1996), "The Future of Interactive Marketing", *Harvard Business Review*, Nov-Dec, 151-162.

Green, Heather (1998), "The New Ratings Game", *Business Week*, 27 Apr 1998, 73.

Hoffman, Donna L. and Thomas P. Novak (1996), "Marketing in hypermedia computer-mediated environments: Conceptual foundations", *Journal of Marketing*, 60(3), 50-68.

Hoffman, Donna L., William D. Kalsbeek, and Thomas P. Novak (1996), "Internet and Web Use in the U.S.", *Communications of the ACM*, 39 (12), 36-46.

Niccolai, James (1997), "Web-analysis tools can measure more than a site's success", *Infoworld*, 7 Apr 1997, 19 (14), 62.

Pant, Somendra, and Cheng Hsu (1996), "Business on the Web: Strategies and Economics", *Communications of the ACM*, 39 (12), 36-46.

PC Meter (1997), "Quarterly Update: Subscriber Presentation".

Pitkow, James (1997), "In search of reliable usage data on the WWW", *Computer Networks and ISDN Systems*, 29, 1343-1355.

Pitkow, James E. and Colleen M. Kehoe (1997), "Emerging Trends in the WWW User Population", *Communications of the ACM*, 39 (6), 106-108.

Sakagami, Hidekazu and Tomonari Kamba (1997), "Learning personal preferences on online newspaper articles from user behaviors", *Computer Networks and ISDN Systems*, 29, 1447-1455.

Table 1. Sample of clickstream data, italicized URLS denote those hits that would be likely to occur in the logs of the host site.

Date and Time of Access	Seconds spent Viewing	URL
18JUL97:18:55:57	47	*http://www.voicenet.com/*
18JUL97:18:56:44	37	*http://www.weather.com/weather/us/cities/HI_Lahaina.html*
18JUL97:18:57:25	105	*http://www.weather.com/weather/us/cities/MI_Traverse_City.html*
18JUL97:19:03:00	7	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:03:56	2	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:03:58	6	http://www.weather.com/weather/us/cities/HI_Lahaina.html
18JUL97:19:04:58	2	http://www.weather.com/weather/us/cities/HI_Lahaina.html
18JUL97:19:05:00	1	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:15:24	39	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:17:00	7	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:17:07	13	*http://www.realastrology.com/*
18JUL97:19:17:20	44	http://www.realastrology.com/libra.html

Table 2. Listing of Demographic Variables

Age
Under 25
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
>64

Gender
Male
Female

Children
Children Present
No Children under 6
No Children under 6-12
No Children under 13-17

Householder
Age: <40

Table 3. Summary of Parameter Estimates Across Households.

Variable	Mean	Standard Deviation	Minimum	Maximum
Parameters associated with u_{it}^{*}:				
ϕ_1	.113	.144	-.719	.959
ϕ_0	6.763	1.490	-.403	12.560
σ	.803	.153	.568	2.034
Parameters associated with z_{it}^{*}:				
δ_1	.151	.278	-1.388	.726
δ_0	-.763	.961	-3.628	8.562

Table 4a. Effects of individual demographics on predicted parameters.

Variable	ϕ_0 Mean	ϕ_0 StdErr	ϕ_1 Mean	ϕ_1 StdErr	δ_0 Mean	δ_0 StdErr	δ_1 Mean	δ_1 StdErr
Intercept	8.57	-1.91	0.02	-0.21	8.38	-4.04	-1.33	-0.42
Age: Under 25	.	(.)	.	(.)	.	(.)	.	(.)
Age: 25-29	0.01	-0.22	0.1	-0.02	0	-0.27	0.03	-0.04
Age: 30-34	0.23	-0.19	0.07	-0.02	-0.47	-0.21	0.14	-0.03
Age: 35-39	0.41	-0.17	0.07	-0.02	-0.25	-0.21	0.14	-0.03
Age: 40-44	0.21	-0.16	0.09	-0.02	-0.23	-0.19	0.13	-0.03
Age: 45-49	0.22	-0.15	0.09	-0.02	-0.41	-0.20	0.18	-0.03
Age: 50-54	0.02	-0.18	0.12	-0.02	-0.44	-0.20	0.2	-0.03
Age: 55-59	-0.06	-0.20	0.14	-0.02	-0.33	-0.26	0.22	-0.04
Age: 60-64	-0.06	-0.23	0.13	-0.03	-0.47	-0.27	0.24	-0.04
Age: Over 64	0.08	-0.19	0.09	-0.02	-0.19	-0.22	0.19	-0.03
Male	0.17	-0.09	0.04	-0.01	-0.42	-0.11	0.08	-0.02
Female	.	(.)	.	(.)	.	(.)	.	(.)
Race: White	-0.07	-0.23	0.01	-0.03	0.36	-0.32	-0.06	-0.05
Race: Black	-0.45	-0.32	0.05	-0.04	-0.21	-0.42	0.03	-0.06
Race: Oriental	0.22	-0.39	0.02	-0.05	1.05	-0.48	-0.14	-0.07
Race: Missing	-0.31	-0.50	0.02	-0.06	0.06	-0.65	-0.01	-0.09
Race: Unknown	.	(.)	.	(.)	.	(.)	.	(.)

Table 4b. Effects of household demographics on predicted parameters.

Variable	ϕ_0 Mean	ϕ_0 StdErr	ϕ_1 Mean	ϕ_1 StdErr	δ_0 Mean	δ_0 StdErr	δ_1 Mean	δ_1 StdErr
Children present	0.48	-0.20	-0.06	-0.02	0.21	-0.25	-0.03	-0.04
No children under 6	0.31	-0.16	-0.02	-0.02	0.08	-0.18	0	-0.03
No children 6-12	0.3	-0.14	-0.02	-0.02	-0.1	-0.17	0.02	-0.03
No children 13-17	0.01	-0.16	-0.02	-0.02	-0.25	-0.18	0.02	-0.03
Household size: 1	0.12	-0.24	0.01	-0.03	0.3	-0.30	-0.01	-0.04
Household size: 2	0.21	-0.17	-0.02	-0.02	0.11	-0.20	-0.02	-0.03
Household size: 3	0.06	-0.14	0.01	-0.02	-0.24	-0.14	0.04	-0.02
Household size: 4 or more	.	(.)	.	(.)	.	(.)	.	(.)
Female Head Employed	0.14	-0.11	-0.02	-0.01	0.19	-0.14	-0.02	-0.02
Male Head Present	-0.2	-0.33	0.03	-0.04	-0.72	-0.39	0.11	-0.06
Female Head Present	-0.37	-0.28	0.04	-0.03	-0.67	-0.34	0.11	-0.05
Number of vehicles: none	.	(.)	.	(.)	.	(.)	.	(.)
Number of vehicles: 1	0.52	-0.24	-0.05	-0.03	0.72	-0.32	-0.14	-0.04
Number of vehicles: 2	0.09	-0.13	-0.01	-0.01	0.35	-0.14	-0.07	-0.02
Number of vehicles: 3+	0.2	-0.16	-0.02	-0.02	0.05	-0.19	-0.03	-0.03
Married	-0.43	-0.64	0.02	-0.07	0.42	-0.72	-0.12	-0.10
Widowed	-0.59	-0.63	0.03	-0.07	-0.56	-0.74	0	-0.10
Divorced/Separate	-0.7	-0.58	0.05	-0.06	-0.31	-0.61	-0.02	-0.08
Single	-0.56	-0.60	0.07	-0.07	-0.49	-0.62	0.03	-0.09
Marriage status unknown	.	(.)	.	(.)	.	(.)	.	(.)

Table 4c. Effects of householder demographics and location on predicted parameters.

Variable	ϕ_0 Mean	ϕ_0 StdErr	ϕ_1 Mean	ϕ_1 StdErr	δ_0 Mean	δ_0 StdErr	δ_1 Mean	δ_1 StdErr
HH age: <40	0.14	-0.15	0.01	-0.02	-0.12	-0.18	0.08	-0.03
HH age: 40-49	0.16	-0.13	0	-0.02	-0.08	-0.15	0.05	-0.02
HH age: 50+	.	(.)	.	(.)	.	(.)	.	(.)
HH educ: HS/less	-0.2	-0.13	0.01	-0.02	-0.34	-0.16	0.04	-0.02
HH educ: Some college	-0.15	-0.12	0.02	-0.01	-0.12	-0.13	0.03	-0.02
HH educ: College	-0.04	-0.11	0	-0.01	-0.09	-0.12	0.03	-0.02
HH educ: Grad school	.	(.)	.	(.)	.	(.)	.	(.)
HH income: < 25k	-1.86	-2.01	-0.01	-0.22	-8.34	-4.19	1.28	-0.44
HH income: 25-40k	-1.85	-2.02	-0.02	-0.22	-8.07	-4.18	1.24	-0.44
HH income: 40-70k	-2	-2.01	-0.02	-0.21	-7.86	-4.18	1.21	-0.44
HH income: 70k+	-2.15	-2.01	-0.01	-0.21	-7.95	-4.18	1.24	-0.44
HH income: missing	.	(.)	.	(.)	.	(.)	.	(.)
Non-MSA market	0.39	-0.13	-0.01	-0.01	0	-0.17	0.01	-0.02
MSA size: 50k-250k	-0.13	-0.15	0.02	-0.02	0.12	-0.18	0	-0.03
MSA size: 250k-1m	0.15	-0.10	-0.01	-0.01	-0.17	-0.12	0.03	-0.02
MSA size: 1m or more	.	(.)	.	(.)	.	(.)	.	(.)

Table 5. Amount of variation in parameter estimates across households attributable to demographic predictors.

Predictor Set	Dependent Variable				
	ϕ_0	ϕ_1	δ_0	δ_1	Average
All Demographics	6.3%	17.6%	37.6%	19.0%	20.1%
Age, Gender, and Race	1.7%	13.8%	10.7%	10.4%	9.2%
Children, Household Size, Vehicles, Marriage Status	2.2%	4.5%	14.4%	4.2%	6.3%
Householder age, Education, and Income	1.6%	0.1%	13.1%	2.3%	4.3%
Household location and city size	1.1%	0.3%	0.8%	0.2%	0.6%

Table 6. Proportion of variation explained by various predictors.

Predictor Set	Average
Demographics	.3%
Last months usage	66.8%
Last months usage + knowledge of whether user access the WWW or not	73.3%

Figure 1. Total WWW Usage for selected individuals.