Results from the Third WWW User Survey

James E. Pitkow & Colleen M. Kehoe

<URL:http://www.cc.gatech.edu/gvu/user_surveys/>

Graphics, Visualization, & Usability (GVU) Center
Georgia Institute of Technology, Atlanta GA 30332-0280
Email {pitkow, colleen}@cc.gatech.edu

Abstract

The tremendous success of the World Wide Web has led to an ever-increasing user base. Intuitively, one would expect this base to change over time as more people from different segments of the population become Web users and advocates. What exactly have these changes been? How do the original Web users differ from the new users from major online service providers like Prodigy? What trends exist and what picture do they paint for the future of the Web user population? This paper, drawing on results from three User Surveys spanning over a year and a half, attempts to answer these and other questions about who is using the Web and why. Additionally, a review of the methodology, questionnaires, and new architectural enhancements is presented. Although the surveys lack the scientific rigor of controlled and accepted methods of surveying, we discuss analyses that help us understand the limitations and process of this new type of surveying. Finally, new quantitative analysis techniques are presented based upon post-hoc log file analysis, yielding guidelines for Web-based survey design.

Keywords

World Wide Web, surveys, demographics, log file analysis, design guidelines.

Introduction

Even with its limited, but expanding degree of interactivity, the Web poses unique opportunities for distributed surveying across loosely-coupled heterogenous environments like the Internet. Yet, these same opportunities to pioneer a new terrain require conservative interpretation of collected data due to the absence of time-tested validation and correction metrics which exist for other surveying techniques.

Part of the initial impetus behind the surveys was to experiment with the Web to determine its viability as a powerful surveying medium. This hypothesis was based upon the easy-to-use, point-and-click interface Web browsers provide. Supporting evidence that the Web is indeed an effective medium is twofold. First, the response rates for the surveys (1,300 respondents for the First Survey, 3,500 for the Second, and 13,000 for the Third) are orders of magnitude above those reported for Usenet news-based surveys and non-specialized emailings. Second, a Usenet news pilot study conducted during the fall of 1994 found a two to one preference for responding to survey announcements via the Web versus email [Aloa 94]. The User Surveys continue to provide fertile testing ground for this hypothesis.

Yet, the advantages of Web-based surveying are not limited to response rates. Foremost, the use of adaptive surveying decreases the number and complexity of questions asked of each user, as only pertinent questions and choices are presented. Because each questionnaire dynamically *adapts* based on the user's input, the database of potential questions can be large while the number of questions given to a particular user remains relatively small. Additionally, the submission, storage, collation, and analysis processes all occur in an electronic medium, limiting human effort to developing processing programs and ensuring the integrity of the collected data[1]. This removes any errors that typically occur in surveying techniques that rely upon human encoding of the collected data. Despite these advantages, we observe that the Web's degree of interaction does not produce the ideal surveying environment**--**one where adaptation occurs instantly on the client.

GVU's First WWW User Survey was conducted during January 1994 and was the first publicly accessible Web-based survey. The initial idea behind the survey was to begin to characterize WWW users as well as demonstrate the Web as a powerful surveying tool. The survey was per-

1. Theoretically, since users are presented with a fixed set of choices, the data received by the server ought to be free of errors. In practice however, we typically observe that several browsers mangle the returned values due to internal programming errors. For this reason, we highly encourage activities that establish test suites for WWW browsers to identify and correct FORM submission-based problems before the browsers are publicly released.

ceived successful as over 1,500 users world wide participated. The response rate was limited, however, due to the lack of non-UNIX clients that correctly processed FORMs.

The Second Survey was advertised and made available to the Web user population for 38 days during October and November 1994. During this period, over 18,000 total responses to the questionnaires from over 4,000 users were received. This survey provided the first cross-platform analysis of Web users as FORMs capable browsers were readily available. New to the surveys was the addition of adaptive questioning to the survey software and the incorporation of the Consumer Sections as pre-tests developed by the University of Michigan's Hermes Project.

Walk-Through of the Survey Interaction

In order to convey the sense of interaction present while completing the surveys, a quick walk-through follows (see [Pitkow 94] for more details on survey execution and architecture). Essentially, the respondents are led through a series of *"question-answer-adapt/re-ask"* cycles. Upon selection of a questionnaire from the Main Launching Page that provides access to all the questionnaires, the surveying software generates the default set of questions from the question database. No adaptation occurs during this stage. The user then responds to the questions displayed by their WWW browser by selecting options presented via radio buttons, pulldown menus, scrolling lists, and check boxes. The surveys intentionally avoid the use of open-ended text entry, as this increases the complexity of response processing.

Once the user completes the set of questions, they click on the "Submit Responses" button of the page. This returns the responses to the survey server. Upon receipt, the survey software inspects each response which results in the one of the following three scenarios:

1. The response triggers an adaptive question based upon the value of the returned response. The corresponding follow-up question is extracted from the database and added to the list of questions returned to the user for the next iteration.
2. The software determines that a question has been asked but not yet answered. In this case, the question is added to the list of questions returned for completion.
3. The response is an acceptable reply to a non-adaptive question. The response is noted and no follow up action occurs.

After all the responses have been inspected, the list of adapted and unanswered questions is returned to user, and another iteration occurs. This cycle continues until all questions have been asked and have been responded to completely. When this happens, the software records that the user has completed the questionnaire and writes the results to disk. The user is returned to the Main Launching Page that lists all the questionnaires that have yet to be completed.

Since the software keeps track of who has filled out which questionnaires, multiple submissions are easily detected. When this occurs, the user is presented with the option to overwrite their previous responses or to preserve them. No method currently exists for a user to inspect the submitted responses above those facilities offered by the browser, e.g. use of the "Back" button.

The integration of adaptive questions into the surveys serves several purposes. Most importantly, it reduces the number and complexity of questions presented to each user. For example, an interesting marketing question is "Where are you located?". Clearly, the space required to list all countries would easily fill several screens; this is undesirable and inefficient. However, staging the question in two parts, one that asks for the primary geographical region of the user and the other that provides a list of countries in that region, reduces the amount of space required to pose the question as well as the cognitive load necessary for the user to correctly answer the question. This method also enables the acquisition of detailed responses, which facilitates a more in-depth understanding of the user population.

Architectural Enhancements

The Third Survey included a trial implementation of *longitudinal tracking* for survey participants. Longitudinal tracking is a method for studying a specific group of users over several surveys. This allows us to investigate how these users' answers change over time and to ask more questions than a one-time survey allows. Since the questionnaires are designed for new as well as returning survey participants, many questions are duplicated from previous surveys. However, when a former survey participant returns to take the current survey, duplicated questions are already filled in with their previous answers. These answers can then be reviewed and changed if necessary. We expect that this implementation of longitudinal tracking will encourage users to participate in more than one survey and will enable us to collect an enriched set of data.

Before answering any of the questionnaires, each user is asked to enter an ID to be used for tracking. Users are cautioned against using an existing password as their ID to avoid potential security hazards. Users are then assigned an internal, unique identifier which is a combination of their ID and a part of their IP address, supplied by their browser. (In the released datasets, these identifiers are replaced by generic identifiers of the form "idxx" to preserve the participants' anonymity.) Finally, users are asked to "Hotlist" the page whose URL contains their identifier and to use this entry whenever accessing the surveys.

When a user returns to take the next survey, the survey software tries to determine the user's unique identifier. During this process, the option to "Choose a New ID" is always present, so that users can choose not to participate in longitudinal tracking if they prefer. If the user returns through their Hotlist entry, their identifier can be immediately extracted from the URL. If not, the user can enter their ID by hand. If the user cannot remember their ID (or has changed to an IP addresses outside their previous domain and class), they can enter the machine name from which they answered the last survey. They are then presented with a list of valid IDs for that machine from which to choose. If the user still cannot find their old ID, they are asked to simply enter a new ID.

Once a user's identifier has been found, to confirm their identity, they are asked for their age and geographic location during the last survey. Note that this is not an attempt at true, reliable authentication; it is designed to minimize errors in identification and to discourage blatant mis-identification attempts. If the user's answers match those given for the last survey, the user is marked in the survey database as "verified" for the remainder of the current survey. If the answers do not match, the user may enter a different ID and try again, or simply choose a new ID and continue with the survey.

This implementation of longitudinal tracking will be fully deployed in the Fourth Survey.

Survey Questions

As with the Second Survey, the questionnaires were separated into four main categories: General Demographics, Web and Internet Usage, Authoring & Information Providers, and the Consumer Section. Since very little is known about the new and expanding market segment of Web Service Providers (companies that offer Web-based services like page design, server space, etc.), we included a pre-test questionnaire for this category. The use of high level categories enabled users to quickly finish sections and select only those areas that are applicable. We note that one long survey containing all questions may discourage potential respondents.

The number of questions in the General Demographics category was doubled since the last survey to 21. Presuming that most people would fill in this portion of the survey, but maybe not others, we included some of the top-level questions from other categories. Thus, users were asked the usual demographic questions regarding age, gender, geographical location, occupation, income, race, level of education, marital status, impairments, etc., as well as questions regarding frequency of Web browser use, primary computing platform, the nature of their primary Internet access provider, etc. For sensitive questions, we provided a "Rather Not Say!" option. Standard to all questionnaires was the inclusion of a text-entry comment box at bottom of the page soliciting users' free-form input.

Of interest is not only who is using the Web, but how they are using it. The second category addressed this topic by posing 28 questions directed toward user's behavior and motivations. Respondents were queried about their frequency and periodicity of Web use, preferences for different types of Web sites and pages, regularity of accesses to different information sources, etc. Questions directed toward users' primary reasons for using the Web were also asked.

Another area of interest surrounds the creation and publishing of HTML documents and their maintenance. The Authoring and Providers section (13 questions) initially identified users who have published information and those that also have maintained HTTPd servers. For authors, questions that determine the learnability of HTML, the sources consulted during learning, as well as understanding some of the advanced features like CGI were posed. Additional questions were asked regarding the number of documents they have authored and converted and the types of pages they create. For Webmasters, information is gathered about which server they operate, which port it listens to, whether proxy and mirroring services are provided, and policies for advertising.

In cooperation with the Hermes Project at the University of Michigan, (and in line with our open policy of incorporating other research agendas into the surveys), the Consumer Section that was pre-tested during the Second Survey was fully deployed. These questions were directed towards understanding consumer purchasing behaviors, attitudes towards online commerce and security as well as plans for future purchases. The questions were specifically designed to allow for comparisons of Web commerce to more traditional practices, such as catalogue shopping and ordering via telephone.

Limitations of the Results & Methodology

Highly distributed, heterogeneous, electronic surveying is a new field, especially with respect to the Web. Our adaptive WWW based surveying techniques are pioneering and as such, require conservative interpretation of collected data due to the absence of tested validation metrics. These metrics depend upon data collected via accepted methods. To date, we know of no such study has been published and the datasets made available to perform these analyses, though several such studies are underway.

Basically, the survey suffers two problems: self-selection and sampling. When people decide to participate in a survey, they select themselves. This decision may reflect some systematic selecting principle (or judgment) that affects the

collected data. Just about all surveys suffer from self-selection problems. For example, when a potential respondent hangs up on a telephone-based surveyor, self-selection has occurred. Likewise, when a potential respondent does not send back a direct mail survey, self-selection has occurred.

The other issue is sampling. There are essentially two types of sampling: random and non-probabilistic. Random selection is intended to minimize bias and make the sample as typical of the population as possible. To accomplish this, steps need to be taken to get respondents in a random manner, e.g., drawing numbers out of a hat. Our survey uses a form of non-probabilistic sampling which relies on users who see announcements of the survey to participate. Since respondents are gathered in this manner, segments of the entire Web users population may not be aware of the surveys and therefore may not participate. As a result, all segments of the user population may not be represented in our sample. This reduces the ability of the gathered data to generalize to the entire user population.

Since the Web does not have a broadcast mechanism (yet), we used the following diverse mediums to attract respondents:

- a special link on the Prodigy Web access page
- links on high exposure WWW pages, e.g. links for the duration of the survey on NCSA's 'What's New', Hotwired, Lycos, etc.
- announcements on WWW & Internet related Usenet newsgroups, e.g. `comp.infosystems.www.*`, `comp.internet.net-happenings`, etc.--two postings at equal intervals
- unsolicited write-ups in numerous computer and Internet related trade magazines, and daily newspapers
- `www-surveying` mailing list announcement

One could argue that this diversified exposure minimizes any systematic effect introduced via the sampling method. We tend to agree, but have taken steps to further explore this issue.

Specifically, we designed the Third Survey to enable us to determine how the respondents found out about the survey. This allows us to group respondents accordingly and look for significant differences between these user populations. For all users, 50% found out about the survey via other WWW pages, with 20.3% finding out via "Other" sources, and 17.9% finding out via Usenet newsgroup announcements. WWW-based listserver/mailing lists, e.g. `www-announce`, etc. accounted for only 6% of all respondents finding out about the survey and thus are not tremendously lucrative means of attracting attention.

"Remembered from last survey" was the least effective method cited (0.4%). This indicates that reliance on former survey participant's memories is not a very robust means of accomplishing longitudinal user tracking. While very few users found out about the Third Surveys via the `www-surveying` mailing announcement (1.1%) compared to other methods, we note that the 142 users who did respond accounted for one fourth of the survey mailing list at the time. Thus, specialized mailing lists seem to be a fairly effective way to announce the beginning of a survey.

In order to determine if the way people found out about the survey systematically biases the sample, we stratified users into groups based upon how they enter the survey. Statistical analysis was performed to determine if these subsamples differed. There were no significant differences between the ways women and men found out about the surveys for the following categories: remembering to take the survey, other Web pages, the newspaper, other sources, and listserver announcements. There were differences found for finding out via friends, magazines, Usenet newsgroups, and the `www-surveying` mailing list. Given the low effectiveness of all but other Web pages and Usenet news announcements, we conclude that these differences lead to nominal effects.

Thus, the surveys do not appear to suffer critically from sampling biases with respect to gender[2]. If a segment of the Web user population were excluded, statistically we'd expect to find similar response distribution for women and men. Still, the data we're about to present is only a snapshot of users who chose to respond--we do not make the claim that the data is representative of the entire Web population.

Execution Environment

The survey ran from April 10th through May 10th, 1995. The server used for the survey operates NCSA's HTTP version 1.3 and ran on a dedicated Sun OS 4.1.3 Sparc 2 installed with a four 75 MHz co-processor HyperSparc. The machine had three gigabytes of disk and 128 megabytes of RAM. The server resided on the College of Computing's internal CDDI ring via a CDDI jumper. This internal ring connects to Georgia Tech's internal and subsequently external FDDI rings, which has a T3 connection to SuraNET. The Survey Modules are written in Perl 4.36 and were not compiled. No notable disruptions or denial of service occurred during the sampling period.

Results

Overall, there were a total of 26,468 responses to all questionnaires combined (38,602 including the Consumer Sections). These responses were submitted from 13,982 unique users.[3] This represents the largest response rate to any Web-based survey known to date. It also represents the most

2. Despite this, we remain unconvinced that the survey's sampling methodology is optimal and welcome suggestions and comments on this subject.

comprehensive online survey of Web users, asking over 138 questions across all questionnaires. Below we present some of the more interesting findings and trends, since presentation of all the results is not possible. Interested readers should consult the online version <URL:http://www.cc.gatech.edu/gvu/user_surveys> to access all results, which include over 200 graphs and detailed interpretations for each question. For the Consumer Surveys, please see the pages maintained by the Hermes Team at <URL:.http://www.umich.edu/~sgupta/hermes/survey3>.

Statistical Inferences

All analyzes were performed using Splus version 3.1 for Unix. Tests for significant interactions amongst variables were performed using the classical chi-squared for independence of categorical data, with significance being determined at $p <= 0.01$ level. Tests for differences between stratified samples was performed using a two-sided alternative for the Wilcox rank sum statistic. Since all tests included $N > 49$, the normal approximation was used, which was replaced by the Lehmann approximation in the event of ties. Significance was determined at the $p <= 0.01$ and confirmed by checking that Z was either < -2.58 or > 2.85.

General Demographics

Analysis of the data for the Third Survey resulted in many interesting findings. Overall, we observed substantial shifts between the demographics of the users who filled out the first two surveys and the third. The users in the Third Survey represent less and less the "technology developers/pioneers" of the First Survey (primarily young, computer-savvy users) and more of what we refer to as the "early adopters/seekers of new technology." These adopters are not typically provided access to the Web through work or school, and as a result, actively seek out local or major Internet access providers, like Prodigy.

Why all this mentioning of Prodigy?

Due to an arrangement with Prodigy (the first major online service to enable Web access), a link to the surveys was placed on Prodigy's Web entry page for ten days during the surveying period. This provided us with the ability to compare Prodigy's users to users in general--the first comparison of these two populations that we know of. Additionally, we stratified the respondents by location (Europe & US) and gender (Women & Men) and performed statistical tests on all questions for differences between groups. All analyses showed differences between groups except where noted, which is not surprising given the large number of data points.

What is the average age?

One category that has changed considerably over time is age. The mean age for the Third Survey is 35.0 (median 35.0), up almost four years from the Second Survey. Also, only 30.4% were between the ages of 21 and 30, compared to 56% of the respondents for the First Survey. We observe no statistically significant differences across gender for age (average age for women is 35.2 years old vs. 35.2 for men).

What is the gender ratio & how has it changed over time?

As for gender, 15.5% of the users are female, 82.0% male and 2.5% chose to "Rather not say!" Compared to the Second Survey, women represent a 6% increase and men a 8% decrease. Compared to the First Survey in January of 1994, this represents a 10% increase for women and 12% decrease for men. This trend is quite linear ($R^2 = 0.98$) and suggests an even male/female ratio could be achieved during the first quarter of 1997. In summary, there exists a trend for the Web towards older users and towards more balanced gender ratios. This progression is clearly away from the young technically savvy male population of a year and half ago.

Also, we observe higher female ratios in the US, with 17.1% of the users being female, 80.3% male and 2.6% chose to "Rather not say!" For Prodigy, the ratios were even more in favor of women, with 19.1% female and 78.8% male. This 1 to 4 female to male ratio more accurately reflects the proportions outside the Web and suggests that as more major online services join the Web and Internet, more balanced female/male ratios are likely to occur. The US and Prodigy ratios also indicate that the US is integrating women more quickly into the user population than other parts of the world.

What is the average and median income?

The overall median income is between US$50,000 and US$60,000, with an estimated average household income of US$69,000. European respondents continue to lag in income, with an average income of US$53,500. Prodigy users' income is the highest of all sampled groups, with a median income in the range of US$60,000 and US$75,000 and an estimated average income of US$80,000.

What about location, marital status, race, & occupation?

For classification by major geographical location, 80.6% of the respondents are from the US, 9.8% from Europe, and 5.8% from Canada and Mexico, with all other major geographical locations represented to a lesser degree. Steps toward replicating the survey on other continents and providing some multilingual support might alter these differences. Overall, 50.3% of the users are married, and 40.0% are single. The percentage of users who report being

3. All collected datasets are publicly available online via the URL listed in the title section and <URL:ftp://ftp.cc.gatech.edu/gvu/www/survey/survey-04-1995/datasets>.

divorced is 5.7%. Occupation-wise, computer-related fields (31.4%) and education-related fields (including students) (23.7%) still represent the majority of respondents, though their dominance over other occupations has been declining. Professional (21.9%), management (12.2%), and "other" occupations (10.8%) fill out the other categories. 82.3% of the respondents are white, with none of the other groups reporting over 5% of the responses. To characterize the sampled population, we find that the respondents are typically white, married, and North American, with computer or educational occupations.

How willing are users to pay for access to Web sites?

Overall, 22.6% of the respondents stated outright that they would not pay fees to access material from WWW sites. This is the same ratio observed in the Second and First Surveys. Additionally, there were no statistically significant differences found between the Prodigy and non-Prodigy response distributions for this question (despite the fact Prodigy users already pay in a direct sense for accessing Web sites). This implies that as the Web continues to increase its user base, we expect to find a 20% negative response to paying for access to Web sites.

What is the primary computing platform?

The distribution of primary computing platforms across all sampled populations more closely resembles computer marketing reports than previous surveys: 52.0% Windows, 26.2% Macintosh, & 8.8% Unix. These platforms account for 87% of all platforms reported.

WWW Usage and Preferences

While our survey does not answer the question, "How many Web users are there?" it does provide insight into potentially more interesting areas such as why people use Web and in what manner. Thus, regardless of overall size, we can gain an understanding of users behind the explosive revolution of the Web.

How often do people use their Web browser?

In general, people spend a considerable amount of time on the Web, with 41% of the users report using their browser between 6 and 10 hours/week and 21% between 11 and 20 hours/week, an increase of 5% and 6%, respectively, since last October. Over 72% responded that they use their Web browser at least once a day. These findings are very encouraging for services like electronic news that attempt to provide daily content--the audience is tuned-in and present.

Why do people use their Web browser?

The most common use of browsers is simply for browsing (82.6%) followed by entertainment (56.6%) and work (50.9%). The category with the least number of responses is shopping (10.5%) (respondents were allowed to choose more than one answer). More users from Europe primarily use their browsers for academic research than do users in the US (45.1% vs. 32.6%). Thus while "surfing" still constitutes the primary reason for using the Web, more serious endeavors like work and research are emerging. These findings support the claim that the Web is not just for fun and games.

What do people do with their Web browser and with what regularity?

The following questions are scored on a 1 (never) to 9 (regularly) scale. The most popular activity for using Web browsers is to replace other interfaces for accessing information (6.7) such as those for FTP, & Gopher. Other categories include accessing reference information (6.2), electronic news (5.7) and product information (5.1). Thus, we find support for the notion that Web browsers are becoming the default interface to the Internet. The least-frequently cited activity for using Web browser is shopping (2.9), which may very well be due to the lack of merchandise on the Web and ubiquitous, secure payment schemes. Interestingly, the response distributions are quite similar to those from the Second Survey, indicating a stable characteristic.

How likely are people to archive Web documents?

In general, users print and save documents with approximately the same regularity (3.9 for print and 4.5 for save). These numbers are right around the "Sometimes" option (4.5), which indicates that not many documents are pulled off the Web. Interestingly, this finding is supported by the research done by Catledge & Pitkow on Web browsing strategies, which also observed low archiving rates based upon monitoring actual user's browsing behavior [Catledge 94].

How fast are people's connection to the Internet?

The most common connection speed is 14 Kb/sec (43.8%) followed by 10 Mb/sec (13.1%). This uneven distribution is a result of the Prodigy users, 73.2% of which have 14.4 connections, and those users which have high speed connections provided via work or school.

Authoring and Providers

How easy was it for people to learn HTML?

Good news, HTML, the markup language used for writing Web documents, is easy to learn. Most users (82.0%) spent between 1 and 6 hours learning HTML. Many users learned

HTML in only 1 to 3 hours (55.2%). CGI was rated the most difficult (5.0) to learn followed by FORMs (4.0), ISMAP (3.9), and HTML overall (2.5). Interestingly, none of these averages are near the maximum difficulty rating of 9.0. Nearly 25% of the users sampled have authored HTML.

How do users learn about HTML?

Online documentation was consulted by 88.4% of users in learning HTML. The next two most popular sources, books and friends, were consulted by only 29.2% and 25.2% of users, respectively (respondents were allowed to choose more than one answer). Hence, use of the Web as a learning medium or to disseminate reference materials corresponds to the behavior of many Web users and is thus recommended for such purposes.

How much does advertising on the Web typically cost?

When queried about charging for advertising on their site, the vast majority of Webmasters replied that the question was "Not Applicable" (70.6%) or that they "Don't Allow Ads" (24.0%) for a total of 94.6%. For those that do allow ads, the largest percentage (3.3%) charge under $50 per week. Only 0.4% charge over $501 per week. Thus, the Web provides an inexpensive advertising medium for most sites.

What about HTTPd servers?

As far as HTTPd servers, the most popular server is NCSA's (38.6%) followed by MacHTTP (20.8%) and CERN's (18.5%). In Europe, however, the most popular server is CERN's (34.9%). Only a small percentage of sites operate a proxy server (12.6%) and most HTTP servers do not mirror other sites (91.5%). The most common server connection speed is 10 Mb/sec (32.3%). The next most common are 1 Mb/sec with 18.0% and 56 Kb/sec with 14.1%, indicating ample throughput to the Internet for over half of the HTTPd servers. This suggests that the lag often experienced by users is primarily a result of their connection speed or the load experienced by the server. Roughly 11% of the users population sampled is composed of Webmasters.

Web Service Providers

What types of services are being offered?

Over half of all Web Service Provider companies sampled (633 total) provide page design (79.0%). Other services are also offered, in the following proportions: Internet/Web consulting (72.8%), other types of services (67.8%), disk space (59.4%), Internet/Web marketing advice (56.2%), CGI scripting (54.7%), and traffic analysis of page accesses (52.1%). Additionally, the providers were equally likely to provide Domain Name Service (DNS) Registration (46.1%) as to not provide DNS services (46.9%).

How many customers and employees do they have?

Nearly half the providers report having between 1 and 10 customers (42.6%), with 9.3% reporting having no customers and 23.2% reporting having over 100 customers. US providers are more likely to have a larger customer base (23.7% US vs. 15.8% European with over 100 customers). The majority of providers have under 10 employees (67.3%), with 16.9% having between 11 and 50, 3.3% having between 51 and 100, and 12.5% having over 100 employees.

How long have they been in business?

Over half of the providers have been in business over 10 months (53.7%). Between January and March 1995, 17.5% of the providers surveyed went online. The startup rate for Web Service Provider companies is fairly consistent (around 10% per month). Thus, most of the Web Services Providers sampled appear to be smaller, recently established companies, with moderate client-bases.

Population Analysis

The response rate to the three surveys has risen dramatically from 1,300 to 3,500 to 13,000 users. The growth is linear under log transform, with the regression equation (R^2 = 0.987) being:

$$f(\text{response rate}) = 5.96 + e^{1.51X}$$

Given this limited model, it becomes possible to predict the response rate for future surveys. The log transform model predicts 38,000 users for the Fourth Survey. This estimate represents an upper bound, with the lower and middle bounds predicted as 18,000 users based upon a linear model and 30,000 users based upon a 2^{nd} degree polynomial curve fitting the equation:

$$f(\text{response rate}) = 3650\ X^2 - 8750X + 6400$$

Log File & Path Analysis

Web-based surveying is a new and exciting medium for collecting data. However, very little is known about how users take Web-based surveys and what parameters effect survey completion. Towards this end, we employed several existing log file analysis techniques and defined new ones to begin to determine and quantify the parameters at play. The next section presents our findings and one of the new techniques.

During the surveying period, 279,770 files totalling over 1.3 gigabytes were transmitted (average 6,824 files/day and 3.3 megabytes/day). Of these requests, slightly over 0.1% resulted in errors, with the most frequent error being "Code 404 Not Found Requests". This indicates that nearly all users who attempted to participate in the surveys were able to successfully access the pages. This round of surveys was

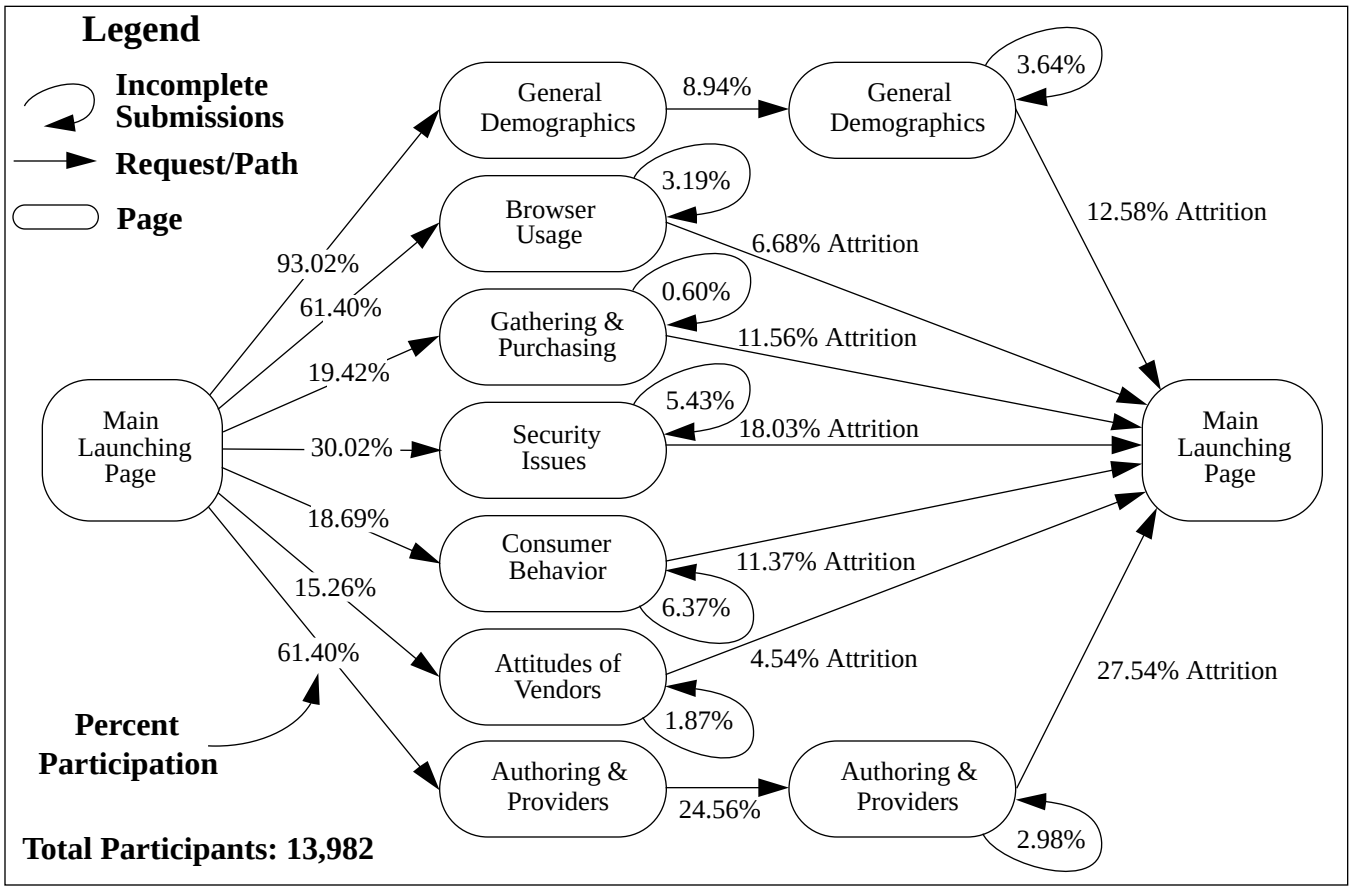

Figure One: The above diagram represents the paths taken by respondents for each questionnaire. Access to each questionnaire was provided via the "Main Launching Page." The loop-backs result from users who did not complete all questions on the questionnaire. The adaptive questionnaires are displayed as two consecutive nodes. Attrition for each traversal is shown as the percent of users who did not proceed along that path. For a more complete discussion of the diagram see the below explanation.

the smoothest to date, with accesses to the Help and FAQ (Frequently Asked Questions) pages each accounting for under one percent of the total file requests. Over 71% of the browsers had image loading turned on, as measured by the ratio of the number of accesses to the Entry pages verses the number of accesses to the image (13 KB interlaced GIF) embedded within these pages.

These descriptive statistics are not very informative. Additionally, exploration into previous hypertext research into the analysis of event driven log files did not reveal many useful techniques. The research typically takes a users path and converts this into a state matrix for subsequent clustering analysis. These analyses usually involve a small number of paths and lose the important sequential nature of hypertext traversals. Given that we wish to explore the paths of over 13,000 users and not have to individually inspect each instance, we developed several new methods of analysis.

One method introduces the notion of *attrition* and *attrition curves*. Attrition can best be thought of in terms of the paths taken by users through an information space. These paths are determined by the underlying structure of hyperlinks, that is, which pages are connected to what. We know that certain users will visit a page and not continue traversing the hyperlinks contained in that page. Others, however, will proceed to traverse the presented links, thus continuing down a path. *Attrition* can be understood as a measure of users who stop traversing verses the users who continue to traverse the hyperlinks from a given page. Attrition is calculated across a group of users. *Attrition curves* are defined as the plot of attrition ratios for all pages along a certain path.

In order to compute attrition, we gathered the paths taken by all users and applied software that tabulates the occurrences of *k-substrings* in an *n-string* for all *k* between 1 and 50. The actual paths taken by users are collated and compose the n-string. Our software exploits the fact the set of k-substrings within the n-string may be a subset of the information space if not all possible paths were traversed. In practice, we observe this property to be true, which greatly reduces the complexity of the computation.

For example, suppose a user takes the path {*a*, *b*, *c*, *a*, *b*}, where *a*, *b* represents the user traversing the hyperlink link contained in *a* to *b*. The tabulation of the 2-substring of the 6-string would be {*ab*, 2} {*bc*, 1}{*ca*, 1}. Stated in terms of paths, we note that the user traversed the subpath from *a* to *b*

Questionnaires (in order of position)	Ques-tions	Participation (%)	Attrition (%)	Net Loss (People)	Avg. Reading Time (sec)
General Demographics	21	93.02	12.58	1636	44.15
Browser Usage	28	61.40	6.68	130	37.03
Author & Providers	13	24.00	27.54	924	46.13
Security Issues	21	30.02	10.03	421	36.43
Consumer Behavior	16	18.69	11.37	297	35.91
Attitudes of Vendors	28	15.26	4.54	97	28.41
Gathering & Purchasing	41	19.42	11.56	314	26.44

Table One: Summary of exploration into the effect of questionnaire ordering on other attributes.

twice and the subpaths from *b* to *c* and *c* to *a* once. The calculation of k-substrings was computed for all paths taken by all users (48,243 total paths) for the entire survey information space. This set of k-substrings provides the input for calculating attrition, which we define next.

Let G = (V, E) be the directed graph with vertices V and edges E. Let P = {N} be the set of all paths taken by all users through G with N being a subset of V and $p(u, v)$ defining the path from vertex *u* to *v*. Attrition for $p(u, v)$ is thus the sum of accesses to *u* minus the sum of $p(u, v)$ traversals divided by sum of accesses to *u*. That is:

$$Attrition(u, v) = \frac{\sum(u) - \sum(u, v)}{\sum(u)}$$

Now, let T be the defined as the n-string composed of all vertices along $p(u, z)$ where n equals the length of $p(u, z)$ and I equal the set of vertices from *u* to *z*. The attrition curve for a given vertex *u* to vertex *z* is defined as the attrition plots for all pairs (u, i) where *i* is an element of I.

Figure One shows the results of the attrition analysis for the main body of the survey. Given the sensitive nature of some of the questions on the General Demographics questionnaire, it is not surprising to see such a high attrition rate. Loosing 8 our of every 100 users may indeed be enough of a loss to warrant the removal of these questions in future surveys. Plus we see that over a quarter of the users went to the information providers page and did not continue. This is most likely due to the fact that they were neither HTML authors or Webmasters

Loop-backs occur when a user fails to complete the entire set of questions, which results for our software enforcing question completion. The attrition rates for loop-backs range from 0.60% (Gathering & Purchasing) to 5.43% (Security Issues). Interestingly, the Security Issues questionnaire managed to cause problems with some Web browsers, which were unable to successfully submit the results event though all questions had been completed. The Gathering & Purchasing questionnaire did not enforce question completion as all answers were optional check boxes. Thus, the results of these analyses make sense and help quantify the effects of attrition on user behavior.

Other, more conventional analysis were also performed to gain a better understanding the effects of questionnaire ordering on the Main Launching Page. That is, what effect does the ordering of possible questionnaires have on which questionnaire users participate? For starters, we determined each user's reading time per page. This was then averaged across all users for all pages. Table One displays the results of this analysis along with the relationship between position, participation, attrition, and the number of questions each questionnaire contained. Correlation analysis of these factors reveals that there is a significant relation between position and reading time (Spearman's r = -0.89) which is modeled as a linear fit ($R^2 = 0.73$) of the form:

$$f(\text{reading time}) = -2.88 * \text{position} + 4.78$$

This means that questionnaires placed first on the Main Launching Page were read for longer periods of time. Intuitively this makes sense as users get tired during the surveys presented later. Also, we observe correlation between position and participation (r = -0.83), which is modeled as a power fit ($R^2 = 0.94$) of the form:

$$f(\text{participation}) = 9.63 * \text{position}^{-0.95}$$

Thus, the positioning of questionnaires has a significant effect on how many people take the surveys. The extent of this relationship can been seen from Table 1 as nearly 60% of the users who took the General Demographics section did not take any of the last five questionnaires.

As one would then expect, reading time and participation are also correlated (r = 0.78), yielding an interpretation that the higher the chance of users taking a survey, the greater the chance they are to spend time reading it. In addition, the number of questions and the time spent reading each ques-

tionnaire are correlated (r = -0.67). There are no significant interactions between all the other variables as tested by pairwise analysis. Important to note is that attrition is not correlated to participation. This further confirms that the effects of self-selection on the collected data are nominal.

Empirically Supported Guidelines

- Place the most important questionnaires at the top of the page.
- Place the questionnaires with the most questions near the top of the page.
- Understand the trade-off between gathering sensitive information and attrition.
- Enforcing question completion does not drastically increase attrition.

Conclusions

Clearly, today's Web is not the same Web of January 1994. The infusion of National and Global Information Infrastructure focus combined with easily acquired interfaces to the Web has left its trail across the surveys. The surveyed Web user populations have rapidly flowed from the originators of the technology to the initial users in the educational and research settings to the users provided connectivity at work and school to those who actively seek out Web connectivity. The WWW User Surveys are able to keep pace with the fluidity by identifying and quantifying real changes in the adaptation of what may very well be the most important revolution since Guttenberg.

The use of the Web as a surveying tool has also provided the means for research into a number of areas beyond the collected demographic data. Part of our research efforts are currently being spent creating a surveying environment in Java which is closer to the ideal surveying environment. It will allow instantaneous survey adaptation as opposed to the *"question-answer-adapt/re-ask"* cycle currently used. We are also exploring the relationship between user characteristics and navigational behavior. New log file analysis techniques as well as the development of our log file analysis software will also continue. As always, we welcome and encourage the participation of other research agendas and thank the Web community for their participation and for providing us with this opportunity.

References

Alao, F. (1994). Pilot Study of Network Surveying Techniques. (manuscript not published).

Catledge, L. and Pitkow, J. (1995). Characterizing Browsing Strategies in the World-Wide Web *Journal of Computer Networks and ISDN systems*, Vol. 27, no. 6.

Pitkow, J. and Recker, M. (1994). Results from the First World-Wide Web Survey. *Journal of Computer Networks and ISDN systems*, Vol. 27, no. 2.

Pitkow, J. and Recker, M. (1995). Using the Web as a Survey Tool: Results from the Second World-Wide Web User Survey. *Journal of Computer Networks and ISDN systems*, Vol. 27, no. 6.

Acknowledgments

Georgia Tech's Graphics, Visualization, & Usability (GVU) Center operates the surveys as a pubic service as part of its commitment towards the Web and Internet communities.

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Thanks to all members of the GVU, its director Dr. Jim Foley, and staff for their support and help. Special thanks extend to Kipp Jones, Dan Forsyth, Dave Leonard, and Randy Captenter and the entire Computer Network Services staff for their technical support and generous donation of equipment. Additional thanks go to Laurie Hodges for guidance. Finally, James would like to thank Dr. Jorge Vanegas for implicit funding throughout the surveys.

Author Information

JAMES PITKOW received his B.A. in Computer Science Applications in Psychology from the University of Colorado Boulder in 1993. He is a Graphics, Visualization, & Usability (GVU) Center graduate student in the College of Computing at Georgia Institute of Technology. His research interests include event analysis, user modeling, adaptive interfaces, and usability.

COLLEEN KEHOE received her B.S. in Computer Science from Stevens Institute of Technology in Hoboken, NJ in 1994. She is currently a Ph.D. student in the Graphics, Visualization, and Usability Center of the College of Computing at the Georgia Institute of Technology. Her current interests include educational technology, visualization, cognitive science and Web-related technologies.