On Saturation of Web Usage by Lay Internet users

Mario Christ[1]
Graduate School of Distributed Information Systems Berlin-Brandenburg
+49 (30) 2093-5662
christ@wiwi.hu-berlin.de

Ramayya Krishnan
Heinz School of Public Policy and Management
Carnegie Mellon University

Daniel Nagin
Heinz School of Public Policy and Management
Carnegie Mellon University

Oliver Günther
Institute for Information Systems
Humboldt University of Berlin

Robert Kraut
Human Computer Interaction Institute
Carnegie Mellon University

[1] in 2000/01 on leave at Heinz School, Carnegie Mellon University, partially funded by the Alexander-von-Humboldt Foundation, TransCoop program

Abstract

The growing presence of the Web in everyday life is inextricably connected to the exponential growth in number and variety of Web sites offering information, commerce and services. While the number of users making use of the Internet and the Web has also grown tremendously, little is known about the intensity of individual level web utilization (e.g., number of visits to Web sites) the Web and the trajectory of the change over time of such utilization. For example, we do not know whether the overall growth in Web usage is attributable to the increased numbers of users or to increased intensity of use of established users or both. This article reports the results of an analysis of eight months of longitudinal data on residential Web usage. This data was assembled as part of the HomeNet project at Carnegie Mellon University. Drawing upon recent advances in semi-parametric, group-based statistical modeling, we examine whether there are distinctive clusters of trajectories of Web usage. We find that Web users can be clustered into four groups with distinct trajectories of use. Each of these groups achieve saturation in their extent of Web usage as measured in the number of distinct Web sites they visit over time. We also develop demographic profiles of these different user groups. The results have important implications for Internet marketing strategy and public policy pertaining to the digital divide.

Keywords: User behavior, Internet and the World Wide Web, Statistical Analysis

1. Introduction

With the commercialization of the Internet, the Web has become a marketplace. Visits to given Web sites are considered an important measure of market share and success, and indeed, many Web sites have enjoyed a steady increase in the number of visits. Yet at the level of the individual user, little is known about the trajectory of change over time in the number of visits to Web sites.

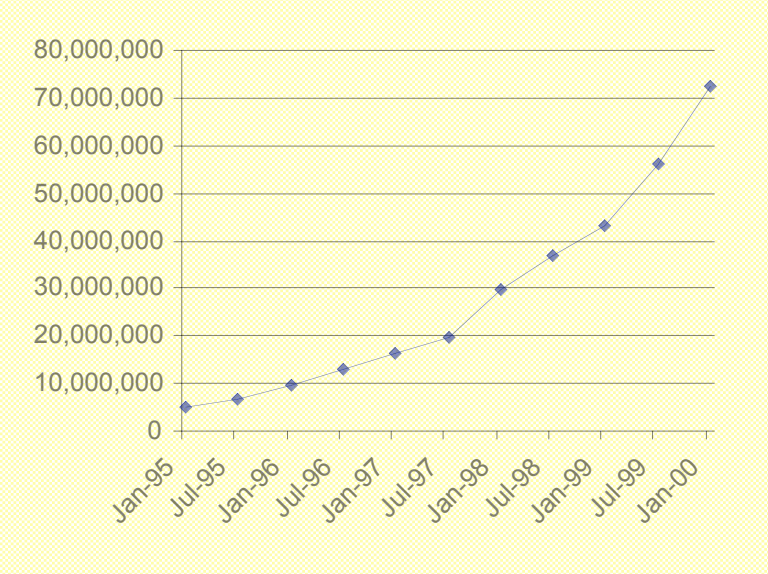

Figure 1: Number of hosts advertised in the DNS (Source: Internet Domain Survey, January 2000)

Figure 1 shows that over the period 1995 to 2000 there was an explosive growth in number of Web sites available to users. In this paper we examine how users responded to the exponential growth in Web site availability. We are specifically interested in exploring whether the increase in Web site visiting opportunities spurred an increase in the utilization rates of individual users. While there is much evidence of large increases in the number of users utilizing the Web,

aggregate utilization reflects a combination of two distinct usage components—number of users and the intensity of their use. The objective of this study is to better understand the unfolding utilization rates of individual users. Specifically, we report results of an analysis of eight months of longitudinal data on individual level Web usage that was extracted from data collected as part of the HomeNet Project [21]. We apply to these data a semi-parametric, group-based statistical method [29] designed to identify distinctive trajectories of individual Web usage. We focus on the analysis of the number of distinctive Web sites accessed per month as a measure of the user's interest in the World Wide Web. We thereby not only identify groups with different levels of usage, we also identify distinctive trajectories of the development of Web usage over time and provide demographic profiles of the identified user groups. The resulting trajectories are compared to the overall trend in the number of Web sites, which multiplied exponentially during the period of observation.

Our study advances research on Web usage in three important ways. With the exception of the recent work by Montgomery and Faloutsos [26], prior studies (e.g., [36], [24], and [4]) have relied on highly non-representative samples (e.g, individuals who worked or studied in computer science departments). Our study relies on a sample of households that is more closely representative of the general population. Second, our period of observation and the tenure of the individuals in the panel, eight months, is far longer than in prior studies. In our judgment an extended period of observation is required to calibrate credibly patterns of change in Web use. Eight months may still be too short a period of time to measure fully the development of Web usage behavior but it is clearly better than the much shorter study periods of prior studies. Third, as described below, the statistical methodology we employ allows us to address directly patterns of change in Web usage over time for individual users.

The paper is organized as follows. Section 2 discusses our source of data and the statistical method we apply to these data. Section 3 presents the results of our analysis. Limitations of this study and future work are discussed in Section 4. Section 5 concludes with a discussion of the implications of our results for Internet marketing strategy and public policy as it pertains to the digital divide.

2. HomeNet Data, Measurement of Web Use, and Statistical Method Used

2.1 The HomeNet project at Carnegie Mellon

Our study is based on individual use records from the HomeNet project. HomeNet is a field trial at Carnegie Mellon University whose aim was to understand usage of the Internet at home by lay users. Starting in 1995, it provided families in the Pittsburgh area with hardware and Internet connections and carefully documented their residential usage of on-line services such as electronic mail, computerized bulletin boards, chat groups, and the World Wide Web [21][37]. We used computer-generated use records from the HomeNet data set of Web sites visited for our purposes.

The data set we used consisted of 139 users performing 1,187,325 http requests between 11-6-1995 and 4-28-1997. Individuals start their Web use on different starting dates and exhibit different durations of Web usage. On average, Web usage behavior of users was observed for 311 days.

The available fields of the data set are as follows:

1. *Unique user ID*
2. *Time and date of the user action*
3. *URL accessed, which is comprised of*
    a) *Domain / Web site accessed (e.g., 'www.yahoo.com')*
    b) *Path on the Web server accessed (e.g., '/search/index.html')*.

Additionally, demographic data measuring age, race, sex, and family role (i.e., mother, father, son, daughter, and other) of all household members were assembled.

In conducting the analysis, we made the following normalizations and assumptions. First, the data set included many http requests that the user did not explicitly perform, such as requests for image files, which are automatically generated by the Web browser. Also, http invocations using the common gateway interface to external programs are often not explicitly requested by the user but instead are automatically loaded[1]. Including these download requests would incorrectly inflate Web usage. For example, 66.7% of the downloads were inline images, 5.13% of the requests were downloads using the common gateway interface (cgi), and 10.0% were downloads of other types, such as music or video files. Therefore, we included only 'page views' in our analysis, which are requests for documents (rather than to an inline image, movie, audio file, etc.). In this regard, we removed records from the database whose *path* fields did not have one of following suffixes: .htm, html, .jsp, and .asp. *Path* fields that point to a directory rather than to a

[1] For instance, an HTML page containing form elements such as those use to submit credit card numbers have references to cgi-bin programs. We count the visit to the HTML page but leave out the calls to the cgi-bin programs that are invoked automatically to process the entries submitted using the form.

file remained in the clickstream database, because Web servers automatically respond to these request by sending a standard document, such as 'index.html'. Removing all hits that were not page views reduced the size of the data set by 81.9% to 214,818 downloads.

Next, *domains,* which are identical except for the prefix, (e.g., 'www.yahoo.com' and 'yahoo.com') were treated as the same page. In this regard, domains, which have the suffix that indicates an explicitly requested port, such as 'yahoo.com:80' were truncated to 'yahoo.com'. Further, we assumed that each domain represents a single Web site.

Finally, because users began their use of the Web at different starting dates and because of different individual durations of Web use in the project, we had to deal with the issue of sparse or missing data at the end of each individual's monitored period of Web usage. Therefore, we analyzed the evolution of Web usage over an 8-month period of time beginning with each individual's starting date. Cutting off sparse data at the end of the period of observation reduces the size of the data set by 6.9%. After these normalizations, the data set consisted of 133,421 page views.

2.2 Measurement of Web Use

There are several conceptually reasonable alternatives to measure Web usage. Broadly, they may be classified into frequency-based measures and time-based measures. Time-based measures use the time spent by an individual at a given Web site as an indicator of the utility received from utilizing the content of the site. We did not pursue time-based measures because we were not confident that we could construct a measure that accurately reflected actual time

spent actively interacting with a Web site. Users may download a Web site but only actively attend to the Web site for a small fraction of the duration over which the site was displayed (i.e., leave the computer unattended for hours or even days). Thus, we focused on frequency-based measures.

Two frequency-based measures were analyzed—a count of the number of distinct Web sites visited by a given user per time period and total Web site visits per time period. The former measure does not count repeat visits to the same Web site whereas the latter measure does count such visits. Table 1 illustrates the calculation of these alternative measures of Web usage for a hypothetical user over a three-month period. In period 1, the user accesses a total of three sites. However, only two are distinctive because yahoo is visited twice. By this same counting logic, a total of four sites are visited in period 2 but only three are distinct. In month three, a total of two sites are visited but because they are the same, only one distinctive site is recorded.

Table 1: URL sets and number of distinctive Web sites of a fictitious user

Month	URLs accessed	#distinctive Web sites
1	www.yahoo.com www.yahoo.com www.amazon.com	2
2	www.yahoo.com www.amazon.com www.excite.com www.amazon.com	3
3	www.yahoo.com www.yahoo.com	1

We use the count of distinct Web site visited as our primary indicator of Web usage because of our interest in comparing the development of individual Web usage with the aggregate growth in Web site visiting opportunities. At the level of the individual the diversity of Web sites visited

provides an indication of individual-level willingness to search the exponentially expanding set of visiting opportunities. However, because this measure does not count repeat visits to a given Web site, it is also important to examine total Web site visits as an alternative utilization intensity measure. This permits an analysis of whether users that visit a few distinct Web sites in a time period are more intensive users of these sites than are users who visit a larger number of sites but use each of them less intensively. Further, the number of repeat visits to a site per user is equivalent to the pages downloaded per site. This is an important measure with relevance for advertising online using banner advertisements. These advertisements are served as part of a downloaded Web page and priced per thousand impressions of the advertisement (also known as cost per thousand impressions or CPM) (see http://www.iab.net/measuringsuccess/index.html for more on online advertising).

2.3 A Semi-parametric, Group-Based Approach for Analyzing Developmental Trajectories

Table 2 reports summary statistics on Web usage for 139 HomeNet users in our analysis. The mean number of distinct Web sites visited per month is 32.66. Users commonly make repeated visits to their favored Web site because the average number of page views per month is about 155 or 4.75 per distinct Web site visited. There is, however, much variation across users in utilization rates. The median number of distinct Web sites visited is only 10 sites per month, less than half the average. This implies a pronounced rightward skew in the population utilization rates, which is indeed reflected in the 90th percentile of distinct Web sites visited, 82.

Table 2: Summary information on Web usage

Overall number of page views	133,421
Average number of distinct sites visited / months	32.66
Average page views per month	155.15
Ratio of page views / site	4.75
10^{th} percentile of distinct sites visits	0
Median of distinct sites visits	10
90^{th} percentile of distinct sites visits	82
Users	139

Further there may also be large differences across individuals in the unfolding of their utilization rates over time. This brings us to the central goal of our analysis—identification of the developmental course of Web usage across distinctive subpopulations. To this end we apply a semi-parametric, group-based methodology [29] that was designed to identify distinctive trajectories of human development. In developmental psychology, a trajectory defines the developmental course of a behavior over age or time. Such trajectories might include groups of "increasers", "decreasers", and "no changers."

Using finite mixtures of suitably defined probability distributions, the group-based approach for modeling developmental trajectories is intended to provide a flexible and easily applied method for identifying distinctive clusters of individual trajectories within the population and for profiling the characteristics of individuals within the clusters. Technically, the group-based trajectory model is an example of a finite mixture model. Its parameters are estimated by maximum likelihood.

The fundamental concept of interest is the distribution of behavioral outcomes conditional on month of usage; that is, the distribution of behavioral trajectories denoted by $P(Y_i|month_i)$, where the random vector Y_i represents individual i's longitudinal sequence of behavioral outcomes (i.e. Web usage) and the vector $month_i$ represents i's month of Web usage when each of those measurements is recorded. The model assumes that this distribution arises from a finite mixture

of unknown order K. The likelihood for each individual i, conditional on the number of groups K, may be written as:

$$P(Y_i \mid Month_i) = \sum_{j=1}^{K} \pi_j \cdot P(Y_i \mid Month_i, j; \theta_j),$$

where π_j is the probability of membership in group j, and the conditional distribution of Y_i given membership in j is indexed by the unknown parameter vector θ_j. In most previous applications, θ_j is a vector of regression parameters determining the shape of the group-specific trajectory.

For given group j, conditional independence is assumed for the sequential realizations of the elements of Y_i, y_{it}, over the T periods of measurement. Thus, we may write

$$P(Y_i \mid Month_i, j; \theta_j) = \prod_{t=i}^{T} p(y_{it} \mid month_{it}, j; \theta_j),$$

where $p(.)$ is the distribution of y_{it} conditional on membership in group j and the month of Web usage of user i at time t.

One valuable feature of the model is that it is easily adapted to accommodate different forms of data by an appropriate distributional representation of $p(y_{it}|month_{it}, j; \theta_j)$. In this analysis the data is in the form of a count whereby y_{it} measures the number of distinct Web sites visited by individual i in period t. As is conventional in the analysis of count data, we assume that y_{it} follows the Poisson distribution. For the Poisson-based model it is assumed that, for each group j:

$$\log(\lambda_{it}^{j}) = \beta_0^j + \beta_1^j month_{it} + \beta_2^j month_{it}^2 + \beta_3^j month_{it}^3 \quad (1),$$

where λ_{jt} is the expected number of occurrences of the event of interest (e.g. visits to distinct Web sites) of subject i at time t given membership in group j.[2] The model's coefficients—β^j_0, β^j_1, β^j_2, and β^j_3—determine the shape of the trajectory and are subscripted by j to denote that the coefficients are not constrained to be the same across the K groups. See [29] for further details.

A key issue in the application of a group-based model is making a determination of how many groups define the best fitting model. One possible choice for testing the optimality of a specified number of groups is the likelihood ratio test. However, the null hypothesis (e.g. three components vs. more than three components) is on the boundary of the parameter space and hence the classical asymptotic results that underlie the likelihood ratio test do not hold [10].

Given these problems with the use of the likelihood-ratio test for model selection, we have followed the lead of [7] and use the Bayesian Information Criterion (BIC) as a basis for selecting the optimal model. For a given model, BIC is calculated as *log(L) - 0.5*log(n)*(d)*, where L is the value of the model's maximized likelihood, n is the sample size, and d is the number of parameters in the model. [19] and [20] argue that BIC can be used for comparison of both nested and non-nested models under fairly general circumstances. When prior information on the correct model is limited, they recommend selection of the model with the maximum BIC. In even more recent work, [20] demonstrates that BIC identifies the optimal number of groups in finite mixture models, a result specifically relevant for the mixture models demonstrated here.

[2] A log-linear relationship between λ_{jit} and *month* is assumed to ensure that the requirement that $\lambda_{jit} > 0$ is fulfilled in model estimation. Note also that the group-based specification accommodates population variation in λ. Such variation is the motivation for two important generalizations of the Poisson distribution, the negative binomial distribution and the zero-inflated Poisson distribution.

3. Results

3.1 Trajectories of Usage

Application of the group-based trajectory methodology to these data revealed that the best fitting model, based on the BIC, clustered users into four groups. Figure 2 depicts the actual and predicted trajectories of the four groups, which we label "very heavy users", "heavy users", "moderate users", and "light users". Because the utilization rates of the "very high users" groups is so much higher than the other three groups, figure 3 excludes the very heavy user group and depicts the predicted[3] and actual behavior of the other three groups. Table 3 shows the group percentages.

Table 3: Group percentages

light users	52.5%
moderate users	30.2 %
heavy users	13.7 %
very heavy users	3.6 %

[3] Predicted behavior is calculated as the expected value of each group's behavior and is computed based on model coefficient estimates. For this poisson-based model, this expectation equals the antilog of Equation 1. Actual behavior is computed as the mean behavior of all persons assigned to the various groups identified in estimation. As described in this section, the assignments are based on the posterior probability of group membership.

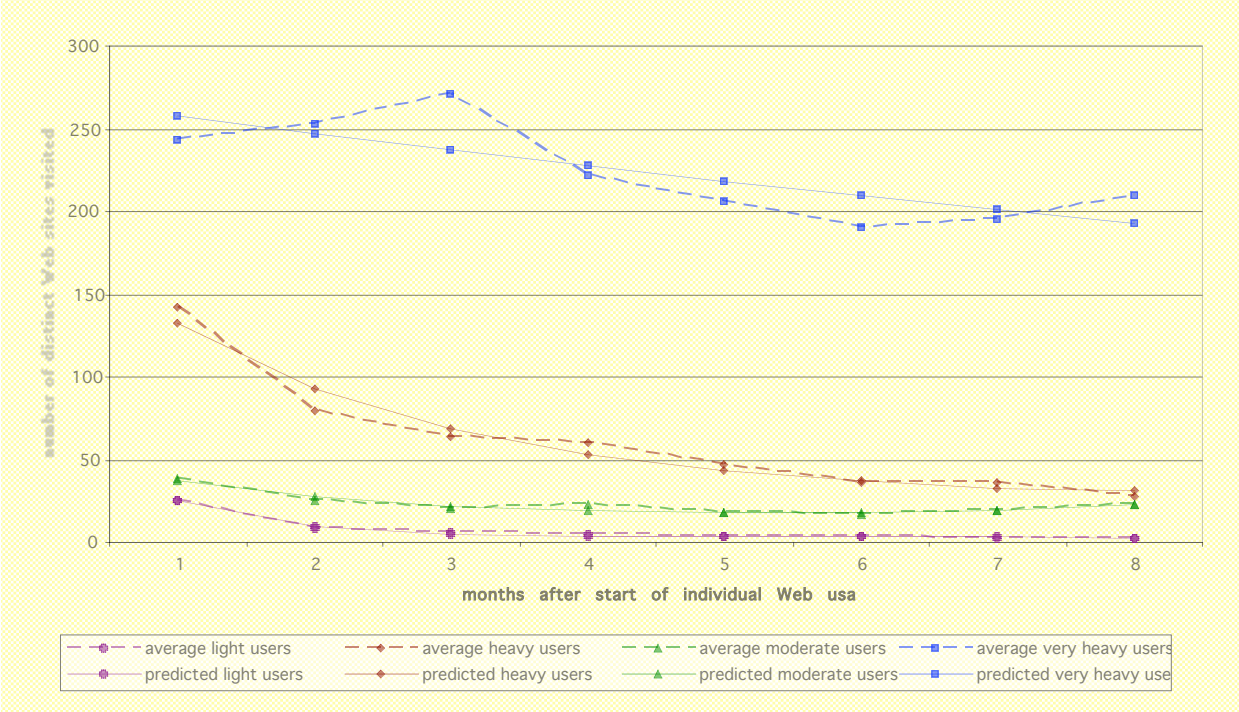

Figure 2: Residential use of the Web measured in number of distinctive Web sites accessed over time

Note that in contrast to the exponential growth in Web sites available as shown in figure 1, there is actually a decline in residential Web usage intensity as measured by number of distinctive Web sites accessed per month. But for a few initial visits to Web sites, the group of 'light users' is composed of individuals who make little use of the Web. This group is estimated to account for estimated 52.2% of the sampled population. The saturation level of 'light users' is only about 3 sites/month, indicating that this group did not find the Web particularly useful, following a short period of Web exploration.

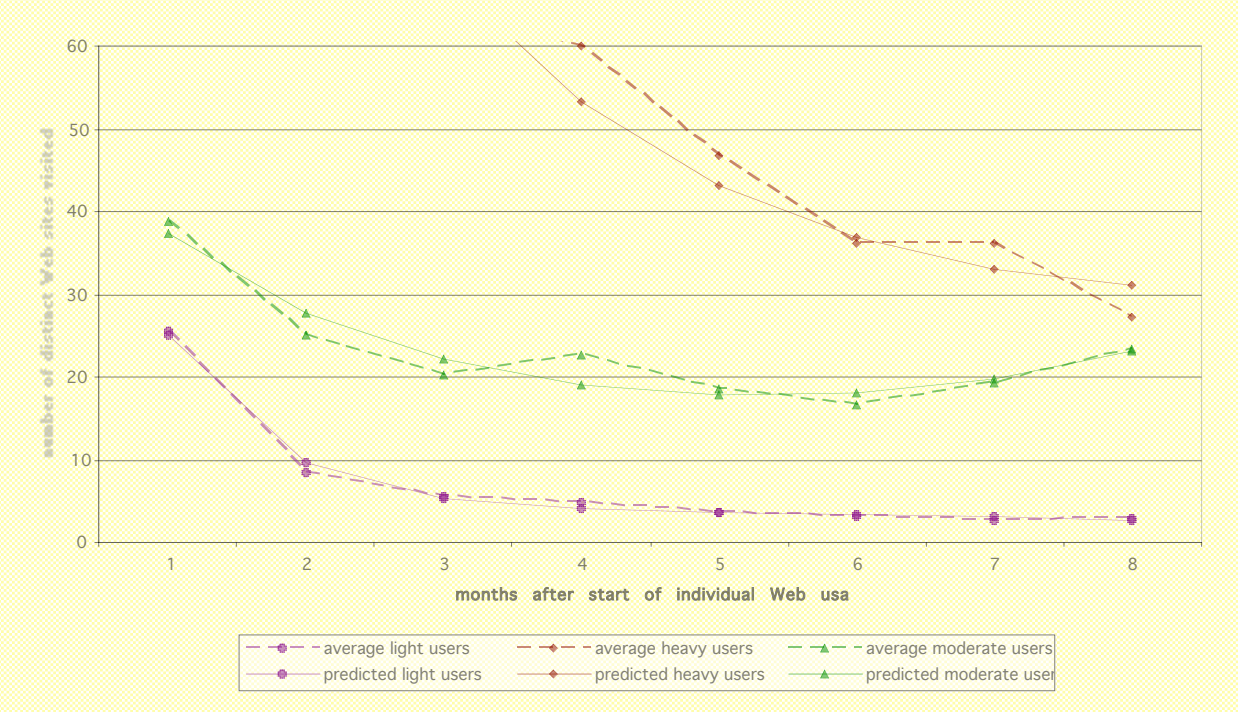

Figure 3: Number of distinctive Web sites visited over time; light users, moderate users, and heavy users only

The second group of individuals – moderate users – start Web usage at a higher level and follow a stable path in Web usage to a point of about 20 distinctive Web sites per month. This group is estimated to constitute 30.2% of the population.

The third group – heavy users who account for about 14% of the overall population – initiates Web usage at a high level of 140 distinctive Web sites per month. However, thereafter their utilization declines quickly to a saturation point of 33 distinctive Web sites per month, which is very close to the saturation point of the group of moderate users. Thus, while moderate users and heavy users differ in their initial Web usage, they converge to about the same utilization level in the long term. Finally, a 'very heavy user' group was identified and is estimated to make up 3.6% of the overall population in the HomeNet sample. This group consists of users who started

at a very high level, over 250 sites per month and who settle into a usage rate of about 200 sites per month.

In summary, all the groups appear to reach saturation in their extent of Web usage as measured by the average number of distinctive Web sites visited per month. For 52.5% of the population called 'light users', the saturation level is at a nominal level of usage of about 2 to 3 sites per month. For 'moderate users the saturation level is about 20 distinctive Web sites per month. After initial heavy utilization of the Web, 'heavy users' tend to visit about 33 distinctive Web sites per month. A small minority of 'very heavy' users has a saturation level that is about 200 distinctive Web sites per month.

We consider these trajectories 'learning curves' of Web usage. Our purpose was to test whether these learning curves tracked the rapid increase in number of Web sites available, and the commercialization of the Net that occurred during the period 1995-1997. They did not. While the sampled households initiated their Web usage in this period of dramatic change in the Internet, no group followed a trajectory of increasing usage. On the contrary, all the groups follow a downward path, indicating that, after a period of 'surfing around' and 'exploring' the Web, residential users seem to limit their Web usage. The increase in available Web sites and the commercialization of the Web with its intended effort to appeal to users did not lead to an increase of Web usage at the individual level.

3.2 Intensity of Web utilization

As shown in Table 3, the intensity of utilization as measured by numbers of page views is considerably larger than when measured in number of distinct Web sites. This indicates that individuals are making multiple visits to Web sites, which is desirable from the perspective of a Web site operator. Figure 4 depicts the distribution of page views by a trajectory group over time. As with visits to distinct Web sites, the number of page views is stable or slightly declining over time, depending on group membership.

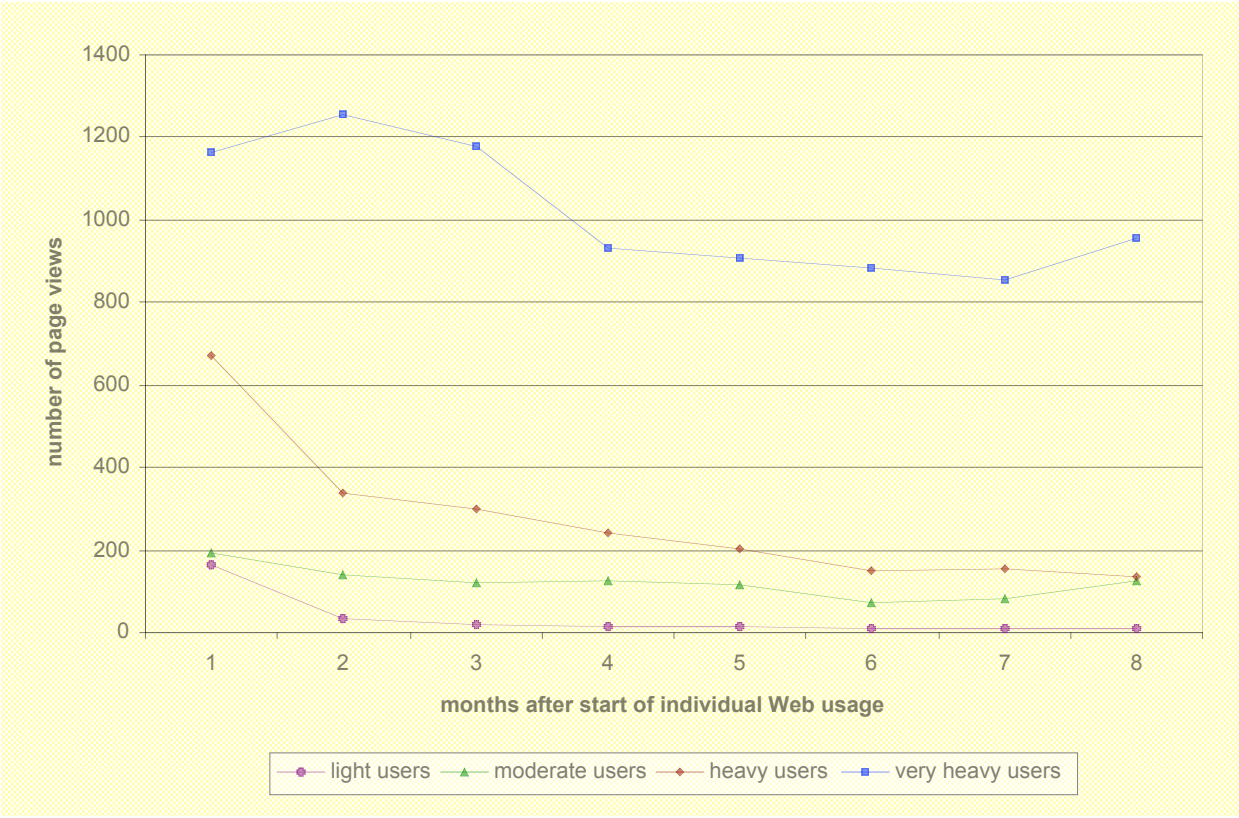

Figure 4: Distribution of monthly page views over time by trajectory group

The results confirm that there is no increase in Web pages viewed by individuals. However, this key result that all groups of users achieve saturation in their Web utilization (also measured as the number of Web viewings per user) is contrary to the key results of Montgomery and Faloutsos [26] who found that Web usage as measured by the number of Web viewings per user is increasing for all types of users.

Before we move on, it is important to explain this difference. The tenure of individuals in the panel is an important distinction between our work and that reported by Montgomery and Faloutsos. As noted in [27], the number of months an individual spent on their panel has a pronounced leftward skew. 38% of the users spent between 1-3 months on the panel and 22% of the users spent between 4-6 months on the panel. Thus 60% of the users spent less than 6 months on the panel. This implies that they have high churn in their panel. Data on Web usage includes both, data from novice and experienced users. In contrast our data set followed the browsing behavior of *each user* over the entire 8 month period over which data was collected and pertains to steady state browsing behavior of users. [4]

[4] While we believe panel tenure to be the most significant difference between our data, there are other differences that could potentially account for the different results. The Jupiter Media Metrix panel is a national panel whereas the HomeNet panel is limited to the Pittsburgh area. Further, the data they use covers the period July 1997 to December 1999 while our data is from June 1995 to April 1997. Another important difference relates to the manner in which data has been collected in the Homenet study and the JMM panel. The Jupiter Media Metrix (JMM) data is collected using a PC Meter resident on the machine being used by the panel member. In contrast, HomeNet data is collected using a proxy server. The meter permits fine-grained monitoring of user actions on the desktop. For example, the meter can record when a browser

In interpreting page views, it is important to keep in mind that individuals do not necessarily visit the same distinct Web sites from month to month. Indeed, there might be considerable churn in the specific Web sites visited from month to month. In this case, there would be limited overlap over time in the identities in the specific Web sites visited. Still while not a perfect measure of loyalty, the number of page views per site is an indicator of loyalty. The number of page views per site indicates how satisfied users are with a given site or how useful a site is to them. If users in fact view many pages on a given site, it is likely that this site interests them. However, **if** individual users have the same "capacity constraints" that determine the extent of their Web utilization as measured by total number of pages viewed in a time period, then one could develop a theory based on P, the number of page views per site and N, the number of sites visited in the time period. Heavy users, as measured by the number of distinctive Web site visited, have on average a larger value for N implying that their value for P should be lower than the value of P for light users who visit few distinctive sites but visit each site with great intensity.

---

window is activated and the length of time for which it is the active window. Proxy servers do not measure this sort of information. Thus, "Web viewings", the measure used by JMM is not the same as the measure of page views we use in this paper. A "web view" begins when a web page is accessed and the browser window used to access the page is active. If the browser window is de-activated, the web view ends even if no new web page has been accessed. In contrast a page view is defined by the user access to a web page and is not related to window activations. To summarize, apart from differences in the process of collecting data, measuring web use, and the nature of the panel, our results shed light on the steady state behavior of users while the work of Montgomery and Faloutsos focuses on the short term behavior of new users.

To test this theory, we investigate whether the intensity of usage as measured in page views / site varies by trajectory group.

Table 5 reports descriptive statistics for the last 50% of the observation period by which time users appear to have reached a steady state of Web usage. Table 5 reveals that there are no material differences between trajectory groups in terms of their utilization intensity as measured in page views per site. Further, the theory that users with fewer distinct Web site visits have a higher number of page views per site is not supported by the data when users achieve steady state behavior. Moreover, it seems that no direct relationship between distinct sites visited and page views per site can be identified. When users achieve a steady state, light users are the least loyal group. This is contrary to the hypothesis that some individuals are visiting a large number of distinct sites infrequently and other users are visiting relatively few sites with high frequency.

Table 5: Summary information on page downloads by trajectory group (last 4 months only)

	light users	moderate users	heavy users	very heavy users
Average number of distinct sites visited per month	3.2	18.5	34.1	183.8
Average number of page views per month	10.8	92.3	153.8	817.5
Ratio of average monthly page views per distinct Web site	2.95	5.12	4.60	4.42

We also examined whether the intensity rates were an exaggerated summary statistic of the typical level of usage intensity due to users having a favorite portal they visit very often. The four most dominant portals / search engines in the data set were Infoseek (0.25% of the records in the set), Yahoo (1.99%), Lycos (0.30%), and Excite (0.70%). Together, these sites account for 3.25% of the page views.

Of specific interest was whether trajectory groups use portals sites and search engines differently. For example, users with low or moderate utilization rates may visit more 'one-stop-surfing' sites such as portals but use these sites more intensively than other users. Indeed, the analysis revealed that the percentage of these sites in the set of very heavy users is only 1.29%, whereas it is 4.00%, 4.77%, and 6.30% for heavy users, moderate users, and light users respectively.

While this difference is not quite significant at the .0f level (Prob > F = 0.0898), we deleted all the page requests of Web servers on the four most popular portals/search engines in our data set: 'yahoo.com', 'excite.com', 'lycos.com', and 'infoseek.com'. Deleting these records reduced the size of the page view data set by 3.25%. Table 6 reports the results of this analysis.

Table 6: Comparison of page views per site by user groups with and without search engines (last 4 months only)

	light users	mo-derate users	heavy users	very heavy users
Average of individual monthly page views per distinct Web site (with portals)	2.95	5.12	4.60	4.42
Average of individual monthly page views per distinct Web site (without portals)	2.89	4.90	5.15	4.36

Except for the group of heavy users, we see a slight decrease in the ratio of page views per site. Judging from these results, the advent of Web portals had only a minor effect on individual Web usage.

In summary, our study reveals that the population of residential Web usage can be clustered into four groups with distinct trajectories of use, whose usage behavior is not increasing but

asymptotically saturating over time. Further, this finding does not seem to be dependent on the specific measure of number of distinct Web sites visited per month. An analysis of the distribution of page views over time leads to the same conclusion. Finally, the intensity with which people utilize their favorite Web site differs slightly across user groups. However, the average intensity rate did not seem to be skewed by portal sites.

3.3 Group profiles

In the next stage of the analysis, we examine the demographic profiles of the trajectory groups. Our purpose was to identify characteristics that distinguished individuals following these four distinctive trajectories. To perform this analysis individuals were assigned to the trajectory group that best conformed to their actual usage trajectory. This assignment was based on the posterior probability of group membership. Based on the model's estimated parameters, it is possible to compute conditional upon membership in a specific trajectory group, the probability of each individual's actual usage level over time, as measured by distinctive Web sites visited per month [29]. This probability is called the posterior probability of group membership. Individuals were assigned to the group with the largest such probability.

Table 4 shows the demographic differences across the groups with low and heavy Web usage and their statistical significance. Note that we aggregated the groups of light users and moderate users on the one hand, and the groups of heavy users and very heavy users on the other hand.

The summary statistics reveal that there is a difference in age across groups. Heavy users and very heavy users tend to be younger whereas light users and moderate users tend to be older. More significantly, there is a race effect and a gender effect. Individuals in the groups that use

the Web heavily tend to be male and white. Conversely, the groups that make little use of the Web ('light users and moderate users') were disproportionately comprised of females and minorities. These results conform with [21]. Surprisingly, there is no income effect. We discuss the implications of the observed race and gender difference from the perspective of the digital divide in Section 4.2.

Table 4: Overview of characteristics of users in the various groups

	all users	light users and moderate users	heavy and very heavy users	Prob > chi2, Prob > F
Percentage	100%	82.70%	17.30%	
Adult	74.30%	73.27%	78.89%	0.60
Female	51.40%	57.02%	26.81%	0.01
Minority	27.60%	31.35%	9.78%	0.05
Average age (years)	31.91	32.42	29.50	0.46
Household income ($1000/year)	54.41	54.77	53.72	0.74
Computer skill (5-point psychometric scale)	3.43	3.33	3.89	0.03
Connection time (hours weekly)	2.15	1.65	4.08	0.00
Mail usage (self-reported[5])	1.62	1.45	2.35	0.16
Phone usage (self-reported[5])	4.14	4.14	4.25	0.87

We also compared e-mail usage and connection time to other indicators of Internet usage. Not surprisingly, connection time increases as Web usage increases. However, as we previously noted this measure of connection time is suspect. We studied e-mail usage and its relationship to

[5] 0 = never, 1 = less than weekly, 2 = weekly, 3= few times/week, 4 =daily, 5=multiple times per day

web usage because these are substantively different Internet services. The Web enables information retrieval. E-mail permits communication. Perhaps users who do not use the Web intensively may be using their time in communication activities. This is not supported by the data. Email usage actually increases with web usage but the difference is not statistically significant across groups. We note, however, the Web-based email such as Hotmail or Yahoo! mail would be counted as a web service vs. e-mail. Understanding the relative utilization of different Internet services is a topic of future research.

Next, we tested if Internet usage is a substitute for telephone usage. For example, users might send email to friends instead of calling them. Also, users might retrieve information from the Web instead of calling somebody to get the needed information. However, this does not seem to be the case. Phone usage actually increases slightly as Web usage increases.

4. Implications for Electronic Commerce and Public Policy

4.1 Implications of saturation in distinct Web sites visited

The results of our study have important implications both for Business to Consumer electronic commerce and for public policy as it pertains to the digital divide. The Web can be thought of as a marketplace with sites competing to attract users to visit. The saturation levels for Web site visits in every trajectory group identified in our analysis can be interpreted to estimate the size of this market. For example, users visit on average about 33 distinct Web sites/month. If this were generalized to the online Web browsing population (let this number be N) at large, 33*N estimates the number of potential Web site visiting opportunities that Web sites will compete over each month. However, over the period of observation, the number of Web sites has grown

exponentially. According to [15] and [32], there were 72,398,092 sites and N=248,660,000 users online in January 2000. Thus, 72,398,092 sites are competing for these 33*248,660,000/month visiting opportunities. Over the past several years, N has continued to grow as the Internet has attracted new entrants. However, as the number of new entrants begins to decrease (this is already happening as can be seen from the estimates of online users at http://www.nua.com/surveys/how_many_online/n_america.html), the number of Web site visiting opportunities will reach a steady state and we expect competition among Web sites for these visiting opportunities to grow in intensity.

Discussions of Web site visiting opportunities are relevant to business models in use in business to consumer electronic commerce. To date portals such as Yahoo! have relied almost exclusively on advertising income generated from serving banner advertisements (so called page impressions). This business model is dependent on maximizing visits from individuals – both first time and repeat visitors- in each time period. Portals have implemented a variety of personalized services (e.g., myyahoo.com) to attract and retain visitors with varying degrees of success. Among the recent wave of dot com failures are several well funded portal sites. These include generic horizontal portals such as the Go portal (funded by Disney) and vertical portals such as dr.koop.com that failed to garner sufficient Web site visitors to sustain themselves – a situation further exacerbated by the decline in online advertising and online advertising rates.

In contrast to portal sites, e-retail sites have to convert visitors into buyers, manage churn rates which represent loss of customers to the competition and enhance repeat purchase rates. Given limited capacity for Web utilization, sites that can achieve high rates of repeat visits and purchases are likely at a clear advantage. Supporting this hypothesis is a recent article by Agarwal et al. (2001), which reports on key processes in business to consumer commerce states

that successful retail commerce companies need to achieve visitor conversion rates of 12 percent, customer churn rates below 20 percent, and repeat purchase rates of around 60 percent [2]. While our results do not shed light on the details of these conversion processes and how online companies should achieve these targets, limited capacity for Web site visiting opportunities among individuals is an important determiner of competition in this area.

This discussion highlights the need to understand the reasons underlying the capacity limits we observed. It is possible that the limited capacity for Web site visits is due to the current technical shortcomings on the Internet (e.g., ease of use of sites, difficulty in using search engines, ineffectiveness of banner advertisements). Breakthroughs in technology can potentially increase the capacity for Web utilization and in turn the size of the market. For example, recent surveys (see the article "Ads Click" at http://interactive.wsj.com/articles/SB10041153126863589 60.htm) demonstrate that ads returned in response to searches are effective in increasing clickthrough rates. Similarly, recent studies undertaken by MSN, Cnet and Doubleclick also demonstrate improved effectiveness of online marketing campaigns using reengineered advertising technology (http://www.iab.net/main/measuringsuccessfinal.pdf). However, it might well be the case that capacity limits on Web utilization are based on cognitive limits and cannot be mediated by technological breakthroughs [40]. Determining the reasons underlying the capacity constraints is an important topic for future research.

4.2 Policy implications

As the Internet has grown and become more widely used by government and organizations, concerns have been raised about the digital divide [9]. The digital divide refers to those members of society who are unable to benefit from the Internet due to their lack of access to it or their inability to make full use of it. Studies such as [14] have carefully examined the policy implications of the demographic patterns of Web usage. Issues such as the gender gap and the race gap have been discussed and numerous studies [6] predict that while the gender gap will likely close over time, the race gap will prevail [1]. In these discussions of the digital divide, the usual assumption has been that access to the Web will almost automatically trigger usage and thereby help close the digital divide. Indeed, a recent report in the Wall Street Journal titled "Closing the Gap" [11] discusses government subsidies that have been proposed as part of legislation such as Colorado's Information Technology Education Act for broadband access in rural and urban areas.

As discussed in Section 3.3. and shown in Table 4, there are race and gender differences in the trajectory groups. For example, the percentage of people who belong to a minority group is 27.60% in the overall sample, 31.35% for light users and moderate users, and 9.78% for heavy users and very heavy users. We observed similar differences in the utilization of the Web by gender. For example, 51.4% of the people in the HomeNet sample are female. This percentage decreases as usage increases (moderate users: 58.1% female, heavy users: 28.6% female, very heavy users: 20.0% female). These findings imply that increased utilization of the Web will require more than access. As noted, all users in the HomeNet panel received free computers with Internet connections and basic training in use of the technology. Informal reports indicate that customized training by gender or race may be needed in addition to access to enable different

segments of society to benefit fully from the Internet. [35] proposes gender-sensitive training to meet the diverse needs of the female Internet user community. Even though recent studies such as [41] show that the gender gap is closing in terms of time spent online, men & women use the Internet different in terms of services used [42]. In this regard, additional work is required to develop policies that will be more successful in promoting utilization of the Web.

5. Future work

This paper contributes to the literature by presenting the results of a long-term study with residential subjects. We encourage future work with respect to three major issues: age of data, length of period of observation, and representativeness of the sample.

The patterns of Web usage we found were based on usage data from 1995-1997. Technical advances, e.g., in the field human computer interaction in general or personalized recommender systems in particular can affect the intensity of Web usage. One of the main reasons for not using more recent data was to make sure that each user has a natural starting point of individual Web exposure. Further studies on people who did not use the Internet before are necessary to confirm the findings from 1998 onwards.

Our study is distinctive in its use of 8 months of continuous individual Web usage data. While recent work by Montgomery and Faloutsos [26] have used more recent data from the Jupiter Media Metrix Panel, the average tenure of users in their panel is much shorter and is biased towards behavior exhibited by new users in the short term. However, in order to gain insights in truly long-term changes in individual access behavior, the analysis of even longer samples of longitudinal data is desirable.

Observed development in browsing behavior might arise due to cultural and social peculiarities of the subject group. Also, a significant share of the population accesses the Internet at work. Therefore, future research is necessary in order to confirm the findings for all groups of users. A truly random nationally representative sample is necessary for this work. Further, we encourage conducting this study in various international settings.

Our results also emphasize that future research on the dynamics of usage will have to incorporate an analysis of churn in the Web. The objective is to measure loyalty of users in the different trajectory groups to the Web sites they visit more accurately than with average number of page views/site metric we used in this paper. By measuring the degree of loyalty of Web users to specific Web sites over time and analyzing whether a given level of Web usage intensity is directed to one site or many sites, one can answer the related question about the demographics of the loyal users on the Web and factors that determine loyalty. Recent work by Yoo and Donthu [36] develops measures of site quality and hospitality to users. Combining our data on browsing behavior with independent measures of site quality will help develop a theory to explain observed levels of stickiness of Web sites.

Acknowledgements

HomeNet is funded by grants from the National Science Foundation under Grants No. IRI-9408271, Apple Computer, AT&T, Bell Atlantic, Bellcore, Intel, Carnegie Mellon University's Information Networking Institute, Interval, the Markle Foundation, the NPD Group, the U.S. Postal Service, and US West. Farallon Computing and Netscape Communications contributed software

Development of the trajectory estimation method and software was supported by the National Science Foundation under Grant No. SBR-9513040 to the National Consortium on Violence and also by separate National Science Foundation grants SBR-9511412 and SES-9911370.

The author Mario Christ was supported by the German Research Society, Berlin-Brandenburg, Graduate School in Distributed Information Systems (DFG grant no. GRK~316). This research was also supported by the TransCoop program of the Alexander von Humboldt Foundation, Bonn, Germany.

The work of Ramayya Krishnan was funded in part by NSF grant CISE/IIS/KDI 9873005. The work of Robert Kraut was funded in part by NSF grant IIS-9980013.

References

[1] Abrams, A. "Diversity and the Internet. J*ournal of Commerce.* June 26. 1997.

[2] Agrawal, V., Arjona, L. D., Lemmens, R. E-performance: the path to rational exuberance. *McKinsey Quarterly*. 1/2001. 30-43.

[3] Cameron, A. C., & Trivedi, P. K "Econometric models based on count data; comparisons and applications of some estimators and tests". *Journal of Applied Econometrics, 1*, 1986, 29-53.

[4] Catledge, L., Pitkow, J. "Characterizing browsing strategies in the world wide Web". In Computer Systems and ISDN Systems. *Proceedings of the Third International World Wide Web Conferenc*e. 27. 1995.

[5] Christ, M., Krishnan, R., Nagin, D., Kraut, R., Günther, O. "Trajectories of individual Web usage: implications for electronic commerce". *Proc. 34th Hawaii International Conference on System Science (HICSS-34)*. 2001.

[6] CyberAtlas. "Demographics: Who's on the Net in the US?" [http://www.cyberatlas.com/demographics.html]. 1998.

[7] D'Unger, A., Land, K., McCall, P., & Nagin, D. "How many latent classes of deliquent/criminal careers? Results from mixed Poission regression analyses of the London, Philadelphia, and Racine cohorts studies". *American Journal of Sociology, 103*, 1998, 1593-1630.

[8] DoubleClick Frequency research findings, Juli 1996, http://www.doubleclick.net/us/resource-center/findings/banner-burnout.asp.

[9] Falling Through the Net: Toward Digital Inclusion. [http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html].

[10] Ghosh, J. K and Sen, P. K. (1985) On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer (Vol. II), L. M. LeCam and R. A. Olshen (eds.), Wadsworth, Monterey, 789-806.

[11] Grimes, A. "Closing the Gap", The Wall Street Journal, Oct 29, 2001.

[12] Hansell, S. Red Face for the Internet's Blue Chip. *The New York Times*. 3-11-2001.

[13] Hoffman, D. L., Novak, T. P. "Bridging the racial divide on the Internet". *Science.* April 17. 1998.

[14] Hoffman, D. L., William, D., Kalsbeek, D., Novak, T. P. "Internet and Web use in the United States: Baselines for Commercial Development". *Communications of the ACM. 39 (12),*1996, 36-46.

[15] Internet Domain Survey, January 2000, http://www.isc.org/ds/WWW-200001/report.html.

[16] Jones, B. L., & Rafaeli, S., "Time to Split, Virtually: 'Discourse Architecture' and 'Community Building' Create Vibrant Virtual Publics", in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, IEEE Press, 2000.

[17] Jones, B. L., & Rafaeli, S., "What Do Virtual " Tells" Tell ? Placing Cybersociety Research Into a Hierarchy of Social Explanation", in *Proceedings of the 33rd Hawaii International Conference on System Sciences,* IEEE Press, 2000.

[18] Jones, B.L., Nagin, D.S., & Roeder, K. (forthcoming). "A SAS procedure based on mixture models for estimating developmental trajectories", *Sociological Research Methods*.

[19] Kass, R.E., & Raftery, A.E. "Bayes factor". *Journal of the American Statistical Association, 190*, 1995, 773-795.

[20] Keribin, C. (1997). Consistent Estimation of the Order of Mixture Models. Working Paper. Laboratoire Analyse et Probabilite, Universite d'Evry-Val d'Essonne.

[21] Kraut, R. E., Scherlis, W, Mukhopadhyay, T., Manning, J., Kiesler, S. "HomeNet: A field trial of residential Internet services". *Communications of the ACM. 39 (12),*1996, 55-63.

[22] Land, K., & Nagin, D. "Micro-models of criminal careers: A synthesis of the criminal careers and life course approaches via semiparametric mixed poisson models with empirical applications". *Journal of Quantitative Criminology, 12*, 1996, 163-191.

[23] Mayo, E. The human problems of an industrial civilization. *Cambridge, MA: Harvard University Press*. 1933.

[24] McKenzie, B., Cockburn, A. "An empirical analysis of Web Page revisitation". *Proc. 34th Hawaii International Conference on System Science (HICSS-34)*. 2001.

[25] Miller, G. A., "The magical number seven, plus or minus two: Some limits on our capacity to process information". *Psychological Review* 63, 1956, 81-97.

[26] Montgomery, A. L., Faloutsos, C. "Trends and Patterns of WWW Browsing Behavior". GSIA Working Paper 2000-E20.

[27] Montgomery, A., "Tenure of panel members in Montgomery Faloutsos Study", Personal Communication, April, 2002. [12]

[28] Nagin, D. & Land, K. "Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poission model". *Criminology, 31*, 1993, 327-362.

[29] Nagin, D. “Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach”. *Psychological Methods*. Vol. 4, No. 2, 1999, 139-157.

[30] Nagin, D., & Tremblay, R.E. “Trajectories of boys’ physical aggression opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency”. *Child Development, 70,* 1999, 1181-1196.

[31] Nagin, D., Farrington, D. & Moffitt, T. “Life-course trajectories of different types of offenders”. *Criminology, 33,* 1995, 111-139.

[32] NUA Internet: How many online?. [http://www.nua.ie/surveys/how_many_online/world.html].

[33] Raftery, A. E. (1995). Bayesian model selection in social research. Sociological Methodology, 25, 111-164.

[34] Roeder, K., Lynch, K., & Nagin, D. “Modeling uncertainty in latent class membership: A case study from criminology”. *Journal of the American Statistical Association, 33*, 1999, 766-777.

[35] Shade, L.R. “Using A Gender-based Analysis in Developing a Canadian Access Strategy: Backgrounder Report“. Universal Access Project, http://www.fis.utoronto.ca/research/iprp/ua/gender/GenderBased.html.

[36] Tauscher, L., Greenberg, S. „How people revisit Web pages: empirical findings and implications for the design of history systems“. *International Journal of Human Computer Studies, Special Issue on World Wide Web Usability, 47*, 1997, 97-138

[37] The HomeNet project. [http://homenet.andrew.cmu.edu/progress].

[38] Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). Statistical analysis of finite mixture distributions. New York: Wiley.

[39] Yoo, B., Naveen, D. (2001), “Developing a Scale to Measure the Perceived Quality of an Internet Shopping Site (SITEQUAL)”, Quarterly Journal of Electronic Commerce, 2 (1), 31-36.

[40] Pool, I., Inose, H., Takasaki, N., & Hurwitz, R. (1984). Communication flows: A census in the United States and Japan. New York: North-Holland.

[41] Cummings, J. N., & Kraut, R. (2002). Domesticating computers and the Internet. The Information Society, 18(3), 1-18.

[42] Boneva, B., Kraut, R., & Frohlich, D. (2001). Using e-mail for personal relationships: The difference gender makes. American Behavioral Scientist Special Issue: The Internet in everyday life, 45(3), 530-549.