# Summary of WWW Characterizations

James E. Pitkow
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto CA 94304 USA
pitkow@parc.xerox.com

## Abstract

To date there have been a number of efforts that attempt to characterize various aspects of the World Wide Web. This paper presents a summary of these efforts, highlighting regularities and insights that have been discovered across the variety of access points available for instrumentation. Characterizations that are derived from client, proxy, and server instrumentation are reviewed as well as efforts to characterize the entire structure of the WWW. Given the dynamic nature of the Web, it may be surprising for some readers to find that many properties of the Web follow regular and predictable patterns that have not changed in form over the Web's lifetime. Understanding these aspects as well as those that vary is critical to designing a better Web, and as a direct consequence, creating a more enjoyable user experience.

**Keywords:** WWW Characterizations, Statistics, Metrics, Analysis, and Modeling

## 1. Introduction

The decentralized nature of the Web makes measuring and gathering representative characterizations of Web usage difficult. As reviewed by [Pitkow 1997], various infrastructure, privacy, and protocol level issues can make even the simplest of metrics, like determining how many people visit a Web page, difficult to obtain. This difficulty is expressed as a common thread in the literature that attempts to analyze various aspects of the Web. Furthermore, it reduces the ability of any one study to generalize results to the entire WWW, as one can not be certain that the data used is an unbiased sample of WWW usage. To compensate for this shortcoming, several studies try to include as many and as diverse data sets as possible to reduce the possibility that a significant aspect of WWW usage was not examined. Table **1** summarizes several invariants discovered across the client, proxy, server, and Web studies reviewed—a remarkable phenomenon given the dynamic nature of the Web. Still, until a method exists of drawing truly random and representative usage samples from the Web, the strongest conclusions that can be drawn are of the form "There seems to be an overwhelming amount of evidence that suggests…". It is from this perspective that the author encourages the reader to frame the following summary of WWW characterization literature.

This paper summarizes the published literature of prior Web characterization efforts. Research that analyses client, proxy and server traffic will be examined followed by a review of attempts to characterize the entire World Wide Web. The literature is reviewed in historical order within each category. This is followed by a discussion of the regularities that have been identified across studies and categories. The bibliography contains hyperlinks to nearly all of the papers mentioned in the paper.

This paper does not focus on the body of literature that deals with caching algorithms, pre-fetching strategies, or other applications that leverage WWW characterizations. While the summary is meant to be complete, unintentional errors and omissions undoubtedly are present—apologies are extended to those efforts and researchers in advance. Finally, this paper should not be used as a substitute to reading the actual literature, but will hopefully serve to help provide a springboard for future research.

## 2. Client

Characterizations of user behavior captured at the browser are the most informative yet the rarest of characterizations. The scarcity of browser level analysis is primarily due to the difficulty in instrumenting WWW browsers and/or ensuring that the data collection environment is not biased towards a particular user segment. Despite this, method such as routing requests through proxies, monitoring operating system level events, and privacy flaws in various browser implementations enable browser level characterizations to be performed today. Still, without access to source code[1] or sufficient APIs to instrument browsers, none of the above techniques permit all user interface events to be captured in context, limiting the scope and accuracy of the results.

[1] While the Open Source program offered by Netscape appears to provide a solution, key aspects of browser technology like Java are currently missing from the code release which severely limits the user experience, and hence the generalizability of the results.

| Regularity | Sources | Metric |
|---|---|---|
| Requested file popularity | [Glassman 1994;<br>Cunha *et al.* 1995;<br>Almeida *et al.* 1996] | Zipf Distribution often with a slope of -1 indicating Zipf's Law |
| Reoccurrence Rate | [Tauscher 1996;<br>Douglis *et al.* 1997;<br>Lorenzetti *et al.* 1996;<br>Cao and Irani 1997] | Roughly 50% of all files are requested more than once by the same client, with the probability of re-referencing within *t* minutes being proportional to *log(t)* |
| File sizes (requested and from the entire Web) | [Cunha *et al.* 1995;<br>Bray 1996; Woodruff *et al.* 1996;<br>Arlitt and Williamson 1996] | Heavy tailed (Pareto) with average HTML size of 4-6 KB and median of 2 KB, images have an average size of 14 KB |
| Traffic properties | [Sedayao 1994;<br>Cunha *et al.* 1995;<br>Mogul 1995;<br>Arlitt and Williamson 1996] | Small images account for the majority of the traffic and document size is inversely related to request frequency |
| Self-similarity of HTTP traffic | [Crovella and Bestavros 1995;<br>Gribble and Brewer 1997;<br>Abdulla 1998] | Bursty, self similar traffic between the micro second and minute time range |
| Periodic nature of HTTP traffic | [Bolot and Hoschka 1996;<br>Abdulla *et al.* 1997c;<br>Gribble and Brewer 1997] | Periodic traffic patterns able to be modeled by time series analysis at the hour to weekly time range |
| Site popularity | [Arlitt and Williamson 1996;<br>Abdulla *et al.* 1997a] | Roughly 25% of the servers account for over 85% of the traffic |
| Life span of documents | [Worrell 1994;<br>Gwertzman and Seltzer 1996;<br>Douglis *et al.* 1998] | Around 50 days, with HTML files being modified and deleted more frequently than images and other media. |
| Occurrence rate of broken links while surfing | [WCG 1997-Xerox PARC, Virginia Tech] | Between 5-8% of all requested files |
| Occurrence rate of redirects | [WCG 1997-Xerox PARC, Virginia Tech] | Between 13-19% of all requested files |
| Number of page requests per site | [Huberman *et al.* 1998;<br>Catledge and Pitkow 1995;<br>Cunha *et al.* 1995] | Heavy tailed (Inverse Gaussian) distribution with typical mean of 3, standard deviation of 9, and mode of 1 page request per site |
| Reading time per page | [Catledge and Pitkow 1995;<br>Cunha *et al.* 1995] | Heavy tailed distribution with an average 30 seconds, median of 7 seconds, and standard deviation of 100 seconds |

**Table 1. Summary of some of the WWW characterization regularities discovered to date.**

The first study to characterize WWW client behavior was performed by [Catledge and Pitkow 1995] during the summer of 1994. The study was conducted for a three-week period using a fully instrumented version of Xmosaic at the Georgia Institute of Technology.  The 107 person study provided a complete breakdown of user interface activity, including the verification of the overwhelming use of navigational features (90+% of all user interface events), especially hyperlink following (52%) and use of the "Back" button (41%). Only 2% of all requested URLs were typed in by users via the "Open URL" dialog box. The mean time between interface events was 9.3 minutes, with sessions being delineated by inactivity intervals of 25 minutes (1 1/2 standard deviations) or greater. A new classification method that categorizes user modes as searching, browsing, or serendipitous based up the length of common navigation subsequences within sites was motivated. Additionally, the "hub and spoke" navigation pattern, where people use a central page to explore from and return to, was observed in many of the traces.

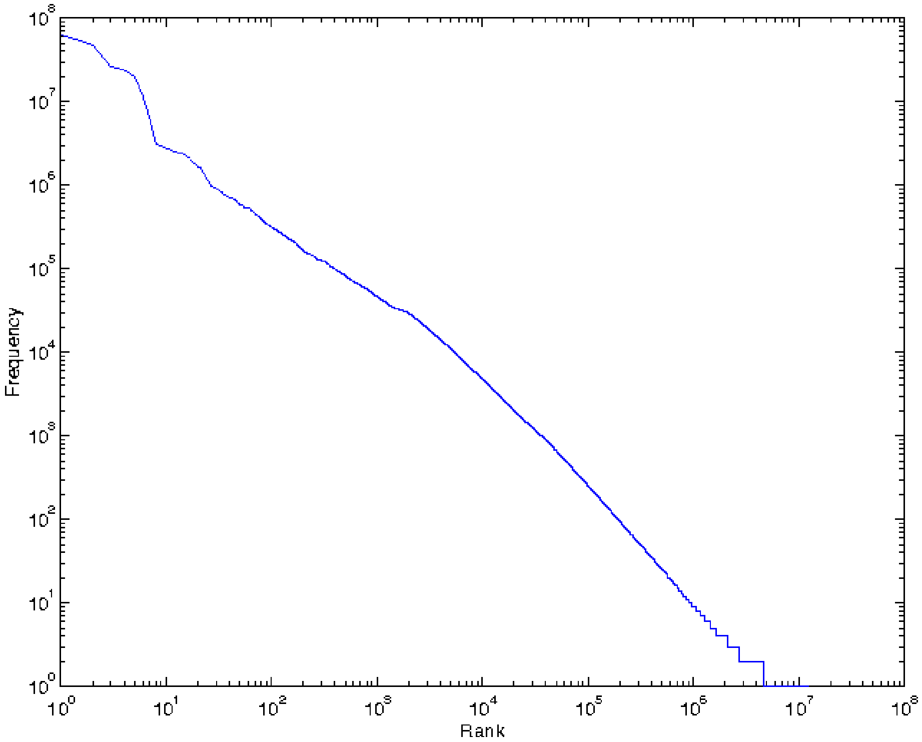

**Figure 1. The Zipf distribution of number of page hits versus rank for five days of AOL December 1997 data.**

Shortly thereafter during the fall of 1994, the Boston University Oceans Group conducted a much larger study—over 600 users during a six-month period—using an instrumented version of Xmosaic that focused solely on navigation events. This data set has been widely used by a number of research efforts on Web characterization. Many interesting heavy tail distributions were documented in [Cunha *et al.* 1995] and later in [Crovella *et al.* 1997] and [Cunha 1997] including: transmission times (Pareto), document sizes (Pareto), document size versus number of requests (Pareto), and a reconfirmation of the applicability of Zipf's distribution for document popularity noted initially by [Glassman 1994] (See Figure 1). Additionally, over 75% of the requested URLs were to off campus servers and 96% of the URLs requested by all users were requested more than once. Even back then, Yahoo! (resided at http://akebono.stanford.edu) was the most popular WWW site. A comparison of file sizes to UNIX file sizes which showed that more small and large files exist on UNIX file systems than on the Web. Implications of these findings on caching algorithms were also presented. An application level specific caching analysis using the same data can be found in [Bestavros *et al.* 1995].

The same data set was used to demonstrate and explain the self-similar nature of WWW traffic for time ranges between 1 second and 100 seconds [Crovella and Bestavros 1995; Crovella and Bestavros 1996]. A clear day/night cycle of network demand is noted at the 16.6 minute/bin level. The number of bytes per unit time was used as the primary metric to gauge burstiness. Their explanation of self-similarity utilizes the Pareto distribution of file sizes, transmission times (Pareto), reading/quiet times (heavy tailed), and the inverse relation between the number of times a document is requested and the size of the document. From this, [Crovella and Bestavros 1995; Crovella and Bestavros 1996] conclude that the self-similar nature of Web traffic is more a basic property of information storage and processing systems than a side effect of network protocols or user preferences. Interestingly, traffic samples from only the busiest periods were used to reveal self-similar behavior. The authors clearly note that many of the less-busy hours in the logs do not show self-similar characteristics, but suggest that this is a consequence of low aggregate traffic.

During the tail end of 1995, [Tauscher 1996] conducted a study of 28 university students who used a fully instrumented version of Xmosaic for a six-week period. An extensive analysis was performed on the user interface events and compared to [Catledge and Pitkow 1995]. The study confirmed that over 90% of all user interface events were navigational, with link following and the "Back" button accounting for 50% and 43% respectively. Only 9% of the requested URLs were typed in. The rate new URLs are visited per user as a function of all requests was measured to be 42%, with the rate URLs are revisited being 58%. The number of references, or distance, between revisits followed a lognormal distribution (the highest reoccurrence rate was 18% for a distance of two clicks) with a jagged form due to the parity induced by the "Back" button. An analysis of longest repeating subsequences was also presented that went into significantly more

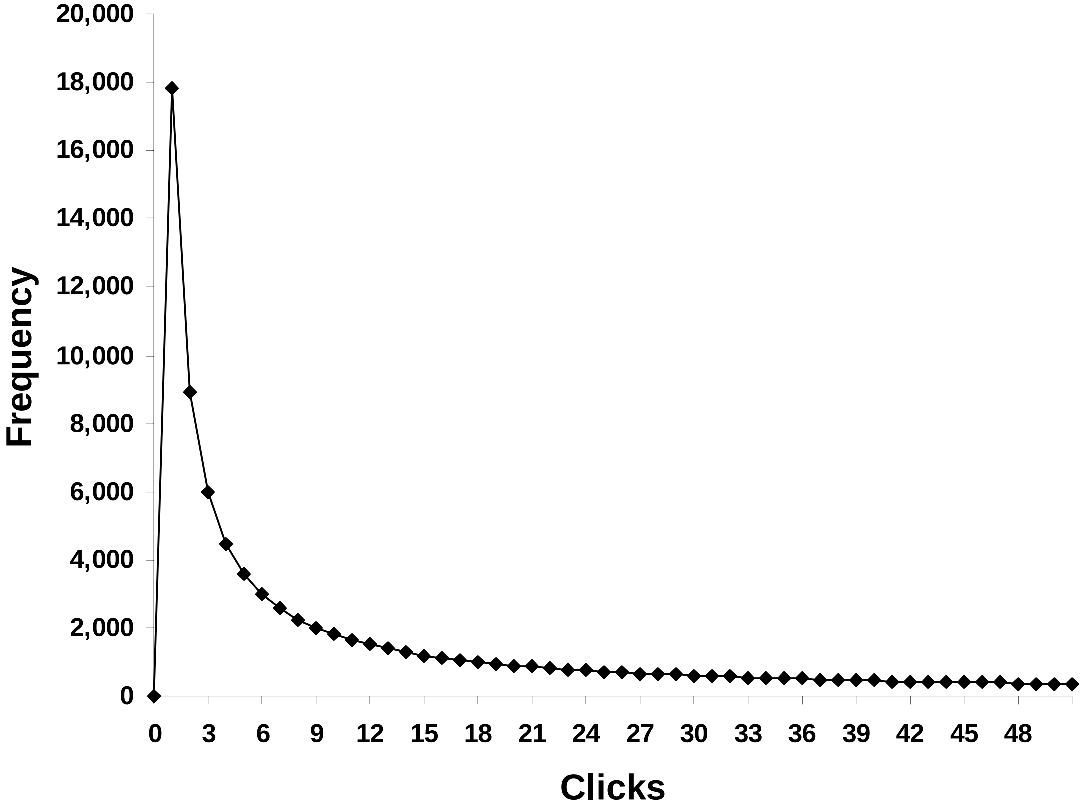

**Figure 2. The number of clicks per user at the Xerox WWW Site during May 1998. The curve follows an Inverse Gaussian Distribution, which has a heavy right tail.**

depth than [Catledge and Pitkow 1995]. The results were used to measure the effectiveness of the algorithms used to manage bookmarks/favorites. A concise summary of the findings can be found in [Tauscher and Greenberg 1996].

Huberman *et al.* 1998] show that the number of clicks users make within sites exhibits strong regularities. The model assumes that users make a sequence of decisions to proceed to another page, continuing as long as the value of the current page exceeds some threshold, which yields the probability distribution for the number of pages a user visits within a site. By using the client traces collected from [Catledge and Pitkow 1995; Cunha *et al.* 1995], and five days of AOL proxy traces during December of 1997, the distribution of all user clicks per site was observed to be inverse Gaussian. The inverse Gaussian distribution of clicks as also found for all users of the Xerox Web Site (See Figure 2), thus bridging the gap between client, proxy and server observed behavior. [Huberman *et al.* 1998] point out that with heavy tailed distributions like the inverse Gaussian, the average case is not the typical case, which for all of the traces was one click. That is, while the average number of clicks per site from the Georgia Tech trace was reported as 8.32 clicks the typical user only requests one page per site and then leaves. The authors note the limitations this regularity places upon protocol constructs like "keep-alive" and pre-fetching strategies.

## 3. Proxy and Gateways

Another way to characterize user behavior is by analyzing proxy traffic [Luotonen and Altis 1994], where the proxy resides on the client machine or in the network. While the availability of proxy traces is greater than client traces, much of the proxy trace research focuses on improving caching algorithms and therefore does not typically specify the shapes and parameters of key distributions like file size, inter-request times, etc. This makes comparing the results and extracting useful characterizations from these works difficult. It is hoped that future proxy trace research will include a section that describes the distributions and characterizations of various parameters.

The first characterization of WWW proxy traffic was performed by [Glassman 1994] using the Digital Equipment Company (DEC) proxy-cache. The proxy was established in January of 1994 and served nearly 300,000 pages to 600 different users, averaging 4,000 requests per day from 100 users. Several interesting observations were drawn. First, document popularity followed a Zipf distribution. Second, one third of the requested pages were not in the cache (cache misses), another third were in the cache (cache hit), and the final third were in the cache but no longer valid (cache invalid). Third, smaller files were requested with greater frequency than larger files. Finally, the rate pages change was uniformly distributed over the sample. As a comparison, higher hit rates were reported by [Smith 1994] for the HENSA proxy in the United Kingdom and lower hit rates by [O'Callaghan 1995] for a proxy in Australia during September 1994 through February 1995 and by [Nabeshima 1997] for the national cache in Japan. Another DEC proxy trace was collected by [Mogul 1996]. That trace has been used by many caching algorithm specific papers.

Also using a corporate Intranet, [Sedayao 1994] characterized the distribution of mime types, mime type traffic, domains accessed, and temporal nature of WWW traffic generated by roughly 800 people at Intel during spring 1994. [Sedayao 1994] found that images are the most requested items and account for the most traffic. GIF files averaged 19,413 bytes and HTML files 5402 bytes, with both distributions exhibiting large variances. In contrast to the early client studies, the most popular domains accessed via the Intel network were equally split between educational and commercial domains at 30% a piece.

With the intent of studying cache invalidation algorithms, [Worrell 1994] conducted an experiment that gathered URLs from a variety of sources including Harvest Caches, publicly readable browser history files, the top Harvest Web pages, and by a modified version of Xmosaic. As a side effect of this research, [Worrell 1994] measured the lifecycles of Web objects. HTML files were found more likely to be modified than images (mean lifetime for HTML files was 75 days versus 107 days for images). The distribution of averaged estimated lifetimes revealed a heavy tail, with the majority of URLs exhibiting small lifetimes (under 50 days). The time between modifications varied significantly, with the variance increasing as a function of time since last modification. A few years would pass before the issue of Web page lifecycles was reexamined.

One of the more prolific research groups studying proxies is the Network Resource Group at Virginia Tech. Using data from three student populations at Virginia Tech during the spring of 1995, [Abrams *et al.* 1995] report a maximum hit rate between the range of 30-50%. This range compares well with the corporate hit rate published by [Glassman 1994] a year earlier. Although distributions are not provided, the paper contains the location of the collected data files form which the distributions could be derived. The traces were expanded in [Williams *et al.* 1996] to include proxies at various access points in the Virginia Tech network. Interestingly, they show that the number of requests per server also follows a Zipf distribution. The percentage of CGI requests was observed not to exceed 0.5% for all traces in the study. Traces from the same sources, albeit from a later time (fall of 1996), were used in [Wooster and Abrams 1997] (see [Wooster 1996] for in depth analysis) to compare caching algorithms as well. [Williams *et al.* 1996] and [Wooster and Abrams 1997] contain a media type breakdown for each data set.

[Gwertzman and Seltzer 1996] used data from a proxy at Microsoft, a Boston University Web server, and several Harvard College Web servers during 1996 to study cache coherency. Confirming the earlier work of [Bestavros 1995] and [Bestavros 1996], file popularity was found to vary inversely with frequency of change, with the average life span of HTML pages being 50 days and GIF files being 85 days, which confirmed [Worrell 1994]. The study also found that images compose over 65% of the total number of requests, with CGI material accounting for 9%, which is notably higher than [Williams *et al.* 1996].

Using a combination of Fourier and statistical analysis techniques, [Abdulla *et al.* 1997c] used time series analysis to model and forecast Web proxy traffic, extending the periodicity and predictability that was modeled by [Bolot and Hoschka 1996] for Web server traffic. Additionally, [Scheuermann *et al.* 1997] reconfirmed the Zipf distribution of file size by popularity using a two-day trace from students at Northwestern University during November 1996. The observation made by [Glassman 1994] and [Cunha *et al.* 1995] that small files are much more frequently accessed was also validated.

[Gribble and Brewer 1997] performed an interesting analysis of dial-in users with connections less than 28.8 kb/sec at the University of California at Berkeley for 45 days starting October 1996 and later during April 1997. They confirm the findings of [Crovella and Bestavros 1995; Crovella and Bestavros 1996] and [Abdulla *et al.* 1997c] that while bursty behavior occurs on small time scales (microsecond and seconds) large time scales (hours, days, weeks) show very predictable diurnal cycles. An important distinction though was raised by [Gribble and Brewer

1997] on the cause of the bursty behavior. While the authors observed bursty behavior, they found no support for self-similar traffic. Specifically, the distribution of GIF image inter-arrival times along with other file types was found to be exponential, not heavy-tailed as would be predicted by self-similar traffic. It is worthwhile to note that the studies used different metrics to determine self-similarity, with [Crovella and Bestavros 1995] using bytes per unit time whereas [Gribble and Brewer 1997] used inter-arrival times.

[Gribble and Brewer 1997] report that cache-specific HTTP headers sent by servers have not increased (54% of the servers issued "last-modified" headers in November 1996 and April 1997). The second most used cache-specific header was the "expires" header, which accounted for 5% of all headers, cache-specific and non cache-specific. On the client side, 22% of the clients issued "if-modified-since" headers, followed by 7% "no-cache". The authors also observe that locality of reference increases with the size of the client population and that 58% of the bytes and 67% of the files transferred are images.

[Abdulla *et al.* 1997a] completed an extensive analysis of proxy logs from ten sources. The sources include the two 1996 DEC proxy traces, the undergraduate and graduate 1994-1995 client BU traces, the three 1996 departmental proxy Virginia Tech proxy traces, a fall 1995 trace from the Korean national proxy, a few minute trace from AOL during December 1996, and a 1996 proxy trace from Auburn high school. Using [Arlitt and Williamson 1996] as a guide, [Abdulla *et al.* 1997a] identified nine invariants. The regularities found in common across the diverse set of proxy logs include a) the median file size is approximately 2KB, b) the mean file size is less than 27 KB, c) 90-98% of the accessed files are either HTML, image, or CGI-bin files, d) images account for the most byte traffic, e) less than 5% of the servers are accessed only once, f) less than 12% of the accesses are to unique servers, g) 25% of the servers are responsible for 80-95% of all accesses, f) 90% of the bytes are from 25% of the servers, and g) 88-99% of the requested files were retrieved successfully. [Abdulla 1998] extends the set of invariants by noting that the Hurst parameter, which is used to measure the degree of traffic self-similarity, varies between 0.59 and 0.94 across all proxy traces. Given that in order for traffic to be considered self-similar, the Hurst value should be between 0.5 and 1.0, [Abdulla 1998] concludes that self-similarity is also an invariant, though the sources may be different than noted in [Crovella and Bestavros 1996].

[Abdulla *et al.* 1997b] presents one of the only studies to systematically study WWW queries across a diverse set of proxies. Studying queries from proxy traces can be difficult, as several proxy caches like Squid remove CGI parameters due to privacy reasons. The percentage of all requests that were queries varied considerably between traces, with a low of 4% (October 1995 Korea national cache), a high of 12% (November 1996 Virginia Tech Library), and a mode of 9% (November 1996 AOL). Also showing considerable variability is number of users accessing search engines (43% of all users in the Korean trace versus 60% Virginia Tech). The number of words per query revealed a mode of two words, with the distribution being right skewed across traces. Nearly all (99%) of the queries did not use any Boolean or other advanced operators. An analysis that characterized user session in terms of browsing, searching, and requesting the next set of search results was also preformed.

[Douglis *et al.* 1997] contains an exhaustive study of the dynamics of the Web drawing on a DEC proxy-level trace during December 1996 and from an AT&T packet-level trace during November 1996. As with other traces, images account for the majority of accesses and bytes transferred. Plotting of the time between accesses to the same URL across all users indicates two notable peaks at one minute and one day, with a mean inter-arrival time of 25.4 hours and a median of 1.9 hours. Nearly half (49%) of all requests reoccurred during the AT&T trace, similar to [Tauscher 1996]. A bit surprising, 16% of the resources that were accessed more than once changed with each access, and 15% of all requested resources reflected a modified resource. These results suggest that using CGI-bin style heuristics, e.g., "cgi", "cgi-bin", htbin" in the URL, to determine the amount of dynamically generated content on the Web may significantly underestimate the true value. Contrary to [Bestavros 1996] and [Gwertzman and Seltzer 1996], more popular files were found to change more frequently. By calculating checksums on the full-body responses, [Douglis *et al.* 1997] found that 18% of AT&T trace requests that resulted in a new instance were unchanged from at least one other instance of a different URL (e.g., mirror sites, session ids embedded in URLs, etc.).

[Duska *et al.* 1997] presents an analysis of locality and sharing characteristics of seven internationally diverse proxy servers during 1996 and 1997 including the DEC proxy trace [Mogul 1996]. "Sharing" occurs when more than one user requests the same URL. The authors point out that clients from both small and large populations request roughly the same proportion of non-shared objects, even though there is more sharing in large populations (clarifying the observation of [Gribble and Brewer 1997] that reference locality increases with population size). Interestingly, while a very large portion of accesses are to shared objects (71% for DEC), only a small portion of objects are shared (23% for DEC). [Duska *et al.* 1997] also found that while about half of each trace's sharing only occurs in that trace and involves only a few clients, the other half of the trace is more general, overlapping between other traces and involving many hosts. The implications of the bimodal nature of sharing on caching are also presented.

While the focus of [Cao and Irani 1997] is on cost-aware proxy caching strategies, the authors do note that the probability of a document being re-referenced within *t* minutes is proportional to *log(t)*. This property was initially observed by [Lorenzetti *et al.* 1996], where 20% of all re-accesses occurred within 15 minutes. In one of the only studies to measure the prevalence of cookies, [Cáceres *et al.* 1998] found that 30% of all requests the AT&T Worldnet modem bank contained cookies, which had a significant effect on cache performance.

## 4. Server

While nearly every Webmaster performs some form of Web site traffic analysis, the following summary focuses only on scientific attempts to characterize site usage. As reviewed in [Crovella *et al.* 1997], the Web, especially on the server side, is filled with many heavy tailed distributions and regularities, e.g., file size, transmission time, document popularity. While gaining access to a server log of some sort or another seems fairly easy, identifying and acquiring a representative server sample frame remains a difficult, unsolved problem. In addition, most studies fail to take adequately address the issue of separating out robots from ordinary users, which may bias the results. Future research will hopefully refine our existing understanding of Web site usage as well as incorporate more diverse and generalizable set of log files.

Within the educational domain, several studies were performed during 1994. [Braun and Claffy 1994] analyzed the cluster of servers at NCSA during August 2-3 1994. Although the focus of [Braun and Claffy 1994] was on the geographical nature of WWW requests with respect to distributed caching, the paper contains several time and size dependent distributions. Traces were taken from NCSA's server, one of the busiest WWW servers at the time, receiving over 500 requests/minute.

At the same conference, [Pitkow and Recker 1994] published the results of applying a model of human memory to the January through March 1994 traffic to the Georgia Institute of Technology WWW server. Their findings suggested that recency (the time since last access) is a better predictor than frequency of future access using a one-day granularity. The paper also documented a server side cache-hit rate of nearly 80% using a LRU policy with a one-day refresh period.

One of the most comprehensive analyses of a WWW server during 1994 was performed by [Mogul 1995] using the November 9, 1994 Californian Congressional election server set up by DEC. The site received over 1,500 requests/minute during peak periods. [Mogul 1995] modified the server to record connection duration, number of disk I/Os, and CPU usage in addition to the normally logged fields. Numerous important findings were made including: a) no correlation existed between file size and connection time for files under 30K, b) the majority of traffic was consumed by small images, c) the cumulative distribution of requestors as a function of the number of requests is log-log, and d) the inter-arrival time of requests did not appear to follow a pure Poisson process. This latter point lends support for the self-similar nature of Web server traffic as put forth in [Leland *et al.* 1993] for Ethernet traffic and [Crovella and Bestavros 1995] for client generated HTTP traffic.

Another early attempt to model server behavior was done by [Burchard 1995]. File size, external link references, and session lengths are shown to generally fit lognormal distributions while the time between requests from the same host were modeled as an exponential distribution.

In an extensive analytical study, [Arlitt and Williamson 1996] identified ten invariants using six different data sets that spanned the course of one week to one year. The data sets included the departmental servers at the University of Waterloo and the University of Calgary, a campus server at the University of Saskatchewan, the Kennedy Space Center (NASA) server, a commercial ISP server in Baltimore, Maryland, and NCSA's server. The first invariant discovered was that 88% of all requests were transfers of content, with another 8% being requests where the client was checking to see if a resource had been modified. Second, images and HTML files account for 90-100% of the files transferred, with images typically being requested far more than HTML files. This is consistent with the client side observations of [Sedayao 1994] and [Cunha *et al.* 1995]. Third, the mean file transfer size was below 21K for all requested objects, confirming the observations by [Braun and Claffy 1994]. One third of the files and bytes accessed are accessed only once.

The fifth invariant presented was the Pareto distribution of file sizes (previous noted on the client side by [Crovella and Bestavros 1995]). The file inter-reference times were found to be independent and exponentially distributed a finding that also confirms [Burchard 1995]. Another invariant is that 10% of the documents account for 90% of all requests and bytes transferred. Temporal locality, as measured by least recently used stack distance, was found in all log files as well. The ninth invariant discovered was that remote sites account for over 70% of the requested files and over 60% of the requested bytes (similar to the client side findings of [Cunha *et al.* 1995]). The final invariant uncovered was that 10% of the thousands of domains accessing the servers account for over 75% of the usage (similar to the proxy trace findings of [Abdulla *et al.* 1997a]).

Shortly thereafter, [Almeida *et al.* 1996] performed analysis on access logs from NCSA, SDSC, EPA, and BU during the fall of 1996. Zipf's Law was shown to apply to popularity for documents served by Web sites as well as for sequences of requests from clients ([Glassman 1994] had shown that a Zipf distribution provided a good fit for document popularity at proxies). Despite this, the Zipf-based model was not able to synthetically generate representative workloads. [Almeida *et al.* 1996] argue that this is due to the failure of the Zipf-based model to capture spatial and temporal locality of reference, a property that was documented on the client side by [Tauscher 1996]. Their study provided further support for the limited utility for LAN based caching due to the lack of shared common interests as originally predicted by [Glassman 1994] and reiterated by [Abrams *et al.* 1995].

Survival analysis of files within the GVU Center at Georgia Institute of Technology during a 226-day period starting in 1996 was performed by [Pitkow and Pirolli 1997]. The origin of accesses was used to stratify the data, where files were classified as either being primarily accessed from users within Georgia Tech (internally driven), outside of Georgia Tech (externally driven), or accessed equally from both communities (mutually driven). Internally driven files were more likely to be deleted, than mutually and externally driven files. Confirming [Worrell 1994] and [Gwertzman and Seltzer 1996], HTML files were found to have shorter lifecycles than non-HTML files. They also found that the more frequently a file has been modified, the greater the chances of file being eventually deleted.

In his undergraduate thesis, [Manley 1997] examined the access log files from ten different sites across a variety of domains. The sites were composed of the following servers: Harvard Engineering, Harvard Arts and Sciences, Rice University Engineering, an adult content, a professional organization, a government agency, a free software company, a Web site designer, and a traditional business. All logs were from 1996 or early 1997 and typically spanned over a year. Unlike previous analyses, [Manly 1997] focused on the changes in access patterns and a Web site's content to categorize sites. The primary characteristics that categorize a site are the growth over time of the following metrics: a) the number of Web users, b) whether the site experiences a redesign, c) the number of documents on the site, d) the number of documents visited per user, e) the number of search engine hits, and f) whether the site charges for access. The results challenge the common belief that CGI is becoming increasingly more important and that the main cause of latency is heavily loaded servers. A more concise summary of the central factors effecting the growth of Web sites can be found in [Manley and Seltzer 1997].

With the intent of advancing the field of server benchmarking [Manley *et al.* 1997] developed a model of WWW server traffic. The model incorporates the site's page set, a collection of user profiles, and the inter-arrival rate between users. In order to construct a page-based model, files requested within a two-second inter-request interval were considered a page. Users profiles were created by using IP numbers to identify users, and modeling the number of requests per session, session length (5-minute time outs), and the time between requests. Inter-request times were derived by dividing the log into sub-logs and using the average arrival time for each sub-log. The ability of the model to accurately reproduce server traffic was measured against five factors including file size, file type, number of requests per user, number of requests per method, and server responses. This approach was able to successfully reproduce traffic as well as predict traffic.

[Barford and Crovella 1998] have also developed a tool to create realistic workload called SURGE (Scalable URL Reference Generator). SURGE draws upon a number of statistical regularities in WWW server usage to generate traffic, including: file sizes - body of the distribution (lognormal), file sizes - tail of the distribution (Pareto), file popularity (Zipf), temporal locality (lognormal), request sizes (Pareto), time between requests for embedded images (Weibull), reading times (Pareto) and the number of embedded references per page (Pareto). From these empirically derived distributions, SURGE creates a mock Web site that essentially contains pages with embedded images that match the statistical properties measured from real Web sites. SURGE also creates a set of clients that each issue a set of requests that mimic the behavior of real clients. [Barford and Crovella 1998] show that SURGE more realistically benchmarks server performance than other commercially available tools. Given the analytical strength of SURGE, it was used by the HTTP-NG project as the basis for the representative testbed.

## 5. WWW

One of the earliest uses of citation analysis on the Web, if not the earliest published account, was performed by [Mauldin 1994] using the Lycos search engine. Of the 335,000 documents retrieved at the time, the Genome Database was the most cited, followed by a now extinct educational server, an RFC FTP site, CERN, and NCSA. For this analysis, citations from within host site were not excluded, hence the probable cause for the Genome site being the most cited.

[Woodruff *et al.* 1996] used the Inktomi collection of 2.6 million documents as of November 1995 to characterize the Web. Many interesting descriptive statistics were provided although no attempts were made to model the distributions. The statistics presented include: a) the mean size for HTML documents was 4.4 KB with a median size of 2 KB and a maximum size of 1.6 MB, b) the average number of HTML tags per page was 71 with an average of 11 unique tags, c) the most cited sites were Xerox PARC, Yahoo, and The Cool Site of the Day (though as with [Mauldin 1994], self referencing citations were not excluded).

Appearing at the same conference was a paper by [Bray 1996], which used the Open Text collection of 1.5 million documents as of November 1995. [Bray 1996] reports several similar findings, notably: a) the mean file size is between 6.5 KB, with a median size of 2 KB, b) slightly over 50% of all HTML pages contain at least one embedded image, with 15% of the pages containing exactly one image, c) 75% of all pages contains at least one hyperlink, with 15% of the pages containing no hyperlinks, d) over 80% of the sites are linked to by between 1 and 10 other sites, e) 80% of the sites contain no hyperlinks to other sites (which means a few sites are doing most of the navigation), f) NCSA, MIT, Yahoo, CMU, and Netscape respectively were the most linked to sites by other sites, and g) 44% of the files did not have a mime type extension, 36% were HTML, and 3% were GIF, which reverses the findings of media type requests and bytes transferred noted at the client and server level.

[Bharat and Broder 1998] and [Lawrence and Giles 1998] independently arrived at a method for computing the total size of the Web as a consequence of determining the amount of the Web covered by search engines. While differences exist in the methodologies, the basic premise is the same: pose a series of queries to search engines and measure the number of pages returned in common between the search engines. Essentially, this amount to finding the intersection, or overlap, between the results. The most striking findings are that the intersection between search engines is less than 2% of the total Web and that with no engine indexes more than about one third of the Web. While the researchers differ on the total size of the Web (between 200 million and 320 million as of November 1997), both agree that Web search engines index only a fraction of the total number of documents on the Web.

## 6. Discussion

**Table 1** summarizes several invariants discovered across the client, proxy, server, and Web studies reviewed in this paper. A brief discussion of the more interesting and unresolved areas follows.

The Zipf distribution of file popularity has been observed at the client [Cunha *et al.* 1995], proxy [Glassman 1994], and server [Almeida *et al.* 1996] levels. In more cases than not, the slope of the distribution is -1, which expresses Zipf's Law in its original form. The heavy tailed nature of the Zipf distribution means that only a few files account for the majority of the distribution. Findings from [Tauscher 1996; Douglis *et al.* 1997; Lorenzetti *et al.* 1996; Cao and Irani 1997] paint a common picture that roughly half of all files are requested more than once, proportional to the log of the last access time; however, the relationship between file popularity and re-occurrence rate remains a topic for future research. This property of a few accounting for a lot is also found in numerous other places like number of access per server, number of requests per user, requests by bytes transferred [Arlitt and Williamson 1996; Abdulla 1998].

The basic nature of Web files appears to be well understood. Efforts that treat HTML files and image files differently, whether from the client, proxy, server, of the Web itself all result in heavy tailed (Pareto) distributions [Cunha *et al.* 1995; Arlitt and Williamson 1996; Bray 1996; Woodruff *et al.* 1996]. The fact that small images dominate traffic also seems to be well supported [Sedayao 1994; Mogul 1995; Cunha *et al.* 1995; Arlitt and Williamson 1996]. Another strong regularity is that documents on the Web change frequently as noted by [Worrell 1994; Gwertzman and Seltzer 1996; Douglis *et al.* 1997].

Moving over to work that characterizes how users experience the Web, the work on SURGE by [Barford and Crovella 1998] incorporates a basic model of surfing using reading times, file popularity, and a complex model of Web page characteristics. The ability of this model to mimic real traffic is a testimonial to the types of applications that can be built once fundamental characterizations have been discovered. By using a variety of data sets, [Huberman *et al.* 1998] demonstrated that the number of clicks of users in a Web site could be modeled by the inverse Gaussian distribution.

Despite the regularities that have been discovered, several areas remain either unexplained or unexplored. For example, while numerous researchers have noted the self-similar nature of Web traffic, the access points and metrics used by each group differ as well as the explanations [Crovella and Bestavros 1995; Gribble and Brewer 1997; Abdulla 1998]. Even to this day, very little work has explored the exact nature of user paths through Web sites. That is, how common and predictive are user paths? The work of [Padmanabhan and Mogul 1996; Cunha and Jaccoud 1997] are exceptions, though both of these efforts concentrate on pre-fetching by assuming either a Markov or random walk model of surfing, neither of which have been demonstrated to be necessarily applicable to the Web. Although [Abdulla *et al.* 1997b] presents a characterization of how people use Web information retrieval systems, more work that utilizes real data and attempts to understand the intention of users is indeed warranted.

## 7. Conclusion

One of the biggest obstacles still facing Web characterization is that of instrumentation. Gathering representative traces from the client, proxy, and server access points remains difficult. The majority of characterizations to date focus on US education settings, resulting in a scarcity of client characterizations from commercial Internet Service Providers, corporate environments, and international users. These sectors represent significant portions of today's WWW traffic and may have very different usage patterns. Another shortcoming of the research is that many of the studies reviewed are outdated. While one could argue that Web usage behavior has not changed significantly in the past year or more, empirical evidence is necessary to construct representative characterizations.

Still, despite limitations of client monitoring and reliable usage measurement issues, the amount of high quality research being done that characterizes the Web continues to increase and will hopefully yield more characterizations. It is possible that better instrumentation at the server, network, proxy, and client access points would enable stronger conclusions to be made, moving this area of Web characterization from correlation to causation.

## 8. References

Abdulla, G., Fox, E., and Abrams, M. (1997). Shared user behavior on the World Wide Web. In *Proceedings of WebNet97*, Toronto, Canada. http://www.cs.vt.edu/~chitra/docs/97webnet/

Abdulla, G., Liu, B., Saad, R., Fox, E. A. . (1997). Characterizing WWW Queries. Technical Report TR-97-04, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA. http://csgrad.cs.vt.edu/~abdulla/ckim/WWWquery.ps

Abdulla, G., Nayfeh, A., and Fox, E. (1997). A realistic model of request arrival rate to caching proxies. Submitted for publication. http://www.cs.vt.edu/~chitra/docs/abdulla-nayfeh-fox/paper.pdf

Abdulla, G. (1998). Analysis and modeling of World Wide Web traffic. Doctoral Thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA. http://www.cs.vt.edu/~chitra/docs/nrgpub/gdiss.pdf

Abrams, M., Standridge, C. R., Abdulla, G., Williams, S., and Fox, E. (1995). Caching proxies: limitations and potentials. *The World Wide Web Journal,* 1(1). http://www.w3.org/Journal/1/abrams.155/paper/155.html

Almeida, V., Bestavros, A., Crovella, M., and de Oliveira, A., (1996). Characterizing reference locality in the WWW. In *Proceedings of PDIS'96: The IEEE Conference on Parallel and Distributed Information Systems*, Miami Beach, FL. http://www.cs.bu.edu/~best/res/papers/pdis96.ps

Arlitt, M. and Williamson, C. (1996). Web server workload characterization: the search for invariants. In *Proceedings of the ACM SIGMETRICS Conference*, Philadelphia, PA.

Barford, P. and Crovella, M. (1997). Generating representative Web workloads for network and server performance evaluation. In *Proceedings of ACM SIGMETRICS Conference*. http://www.cs.bu.edu/techreports/97-006-surge.ps.Z

Bestavros, A. (1995). Demand-based resource allocation to reduce traffic and balance load in distributed information systems. In *Proceedings of SPDP '95: The 7th IEEE Symposium on Parallel and Distributed Processing*, San Antonio, TX. http://www.cs.bu.edu/faculty/best/res/papers/spdp95.ps

Bestavros, A., Carter, R., and Crovella M. (1995). Application-level document caching in the Internet. In *Proceedings of the Second International Workshop on Services in Distributed and Networked Environments (SDNE '95)*. Whistler, Canada. http://www.cs.bu.edu/faculty/best/res/papers/sdne95.ps

Bestavros, A. (1996). Speculative Data Dissemination and Service to Reduce Server Load, Network Traffic and Service Time for Distributed Information Systems. In *Proceedings of ICDE'96: The 1996 International Conference on Data Engineering*, New Orleans, LA. http://www.cs.bu.edu/faculty/best/res/papers/icde96.ps

Bharat, K. and Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of the 7th International World Wide Web Conference*, Elsevier Science, Brisbane, Australia. http://decweb.ethz.ch/WWW7/1937/com1937.htm

Bolot, J., and Hoschka, P., (1996). Performance engineering of the World Wide Web: Application to dimensioning and cache design. In *Proceedings of the Fifth International WWW Conference*, Paris, France. http://www5conf.inria.fr/fich_html/papers/P44/Overview.html

Braun, H. and Claffy, K. (1994). Web traffic characterization: an assessment of the impact of caching documents from NCSA's Web server. In *Proceedings of the Second International WWW Conference*, Chicago, IL. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/claffy/main.html

Bray, T. (1996). Measuring the Web. *The World Wide Web Journal* 1(3). http://www5conf.inria.fr/fich_html/papers/P9/Overview.html

Burchard, P. (1995). Statistical properties of the WWW. http://www.cs.princeton.edu/~burchard/www/stats/

Cáceres, R., Douglis, F., Feldmann, A., Glass, G., and Rabinovich, M. (1998). Web proxy caching: the devil is in the details. In *Proceedings of the ACM SIGMETRICS Workshop on Internet Server Performance*. http://www.research.att.com/~ramon/papers/wisp98.ps.gz

Catledge, L. D. and J. E. Pitkow (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems* 26(6): 1065-1073. http://www.igd.fhg.de/www/www95/papers/80/userpatterns/UserPatterns.Paper4.formatted.html

Crovella, M., and Bestavros, A. (1995). Explaining World Wide Web traffic self-similarity. Technical Report BUCS-TR-95-015, Department of Computer Science, Boston University, Boston, MA. http://www.cs.bu.edu/techreports/95-015-explaining-web-self-similarity.ps.Z

Crovella, M. and Bestavros, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In *Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems*. Philadelphia, PA. http://www.cs.bu.edu/~best/res/papers/sigmetrics96.ps

Crovella, M., Taqqu, M., and Bestavros, A. (1997). Heavy-Tailed probability distributions in the World Wide Web. In *Applications of Heavy-Tailed Probability Distributions*, Adler, Feldman, and Taqqu Ed., Birkhauser, Boston, MA.

Cunha, C., R., Bestavros, A., and Crovella M. (1995). Characteristics of WWW client-based traces. Department of Computer Science, Boston University. Boston, MA, http://www.cs.bu.edu/techreports/95-010-www-client-traces.ps.Z

Cunha, C. (1997). Trace analysis and its application to performance enhancements of distributed information systems. Doctoral Thesis, Department of Computer Science, Boston University, Boston, MA. http://www.cs.bu.edu/students/grads/carro/thesis.ps.Z

Cunha, C. and Joccoud, C. F. B. (1997). Determining WWW user's next access and its application to pre-fetching. In *Proceedings of the International Symposium on Computers and Communication*, Alexandria, Egypt.

Douglis, F., Feldmann, A., Krishnamurthy, B., and Mogul, J. (1997). Rate of change and other metrics: a live study of the World Wide Web. In *Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems*, Monterey, CA. http://www.usenix.org/publications/library/proceedings/usits97/douglis_rate.html

Duska, B. M., Marwood, D., and Feeley, M. J. (1997). The Measured Access Characteristics of World-Wide-Web Client Proxy Caches. In *Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems*, Monterey, CA. http://www.usenix.org/publications/library/proceedings/usits97/duska.html

Glassman, S. (1994). A caching relay for the World Wide Web. *Computer Networks and ISDN Systems* 27(2). http://www1.cern.ch/PapersWWW94/steveg.ps

Gribble, S. D, and Brewer, E. A. (1997). System design issues for Internet middleware services: eeductions from a large client trace. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, Monterey, CA. http://HTTP.CS.Berkeley.EDU/~gribble/papers/sys_trace.ps.gz

Gwertzman, J., and Seltzer, M. (1996). World Wide Web cache consistency. In *Proceedings of the 1996 Usenix Technical Conference*, Boston, MA, Harvard College. http://www.eecs.harvard.edu/~vino/web/usenix.196/

Huberman, B, Pirolli, P., Pitkow, J., and Lukose, R. (1998) Strong regularities in WWW surfing. *Science*, Volume 280. http://www.sciencemag.org/cgi/content/abstract/280/5360/95

Lawrence, S., and Giles, C. L., (1998). Searching the World Wide Web. *Science*, Volume 280. http://www.sciencemag.org/cgi/content/abstract/280/5360/98

Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1993). On the self-similar nature of Ethernet traffic, In *Proceedings of ACM SIGCOMM '93*. San Francisco, CA.

Lorenzetti, P., Rizzo, L., and Vicisanno, L. (1996). Replacement policies for a proxy cache. Technical report LR-960731, DEIT, University of Pisa. Italy. http://www.iet.unipi.it/~luigi/caching.ps  Rewritten version available as Technical Report RN-98-13 by Rizzo, L. and Vicasano, L. (1998). http://www.iet.unipi.it/~luigi/lrv98.ps.gz

Luotonen, A. and Altis, K. (1994). World-Wide Web proxies. *Computer Networks and ISDN Systems* 27(2). http://www1.cern.ch/PapersWWW94/luotonen.ps

Manley, S. An analysis of issues facing World Wide Web servers. Bachelor of Arts. Department of Computer Science, Harvard College, Cambridge, MA. http://www.eecs.harvard.edu/~vino/web/manley_thesis.ps.gz

Manley, S., and Seltzer., M., (1997). Web facts and fantasy. In *Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems*, Monterey, CA. http://www.eecs.harvard.edu/~vino/web/sits.97.html

Manley, S., Courage., M., Seltzer., M (1997) A Self-Scaling and Self-Configuring Benchmark for Web Servers, unpublished document. http://www.eecs.harvard.edu/~margo/papers/hbench-web.ps

Mauldin, M., and Leavitt, J. (1994) Web agent related research at the Center for Machine Translation. *Meeting of the ACM Special Interest Group on Networked Information Discovery and Retrieval*, McLean, VA. http://fuzine.mt.cs.cmu.edu/mlm/signidr94.html

Mogul, J. (1995). Network behavior of a busy Web server and its clients. Digital Western Research Laboratory, CA. ftp://gatekeeper.dec.com/pub/DEC/WRL/research-reports/WRL-TR-95.5.ps

Mogul, J. (1996). Digital's Web proxy traces. Online reference. ftp://ftp.digital.com/pub/DEC/traces/proxy/webtraces.v1.2.html

Nabeshima, M. (1997). The Japan cache project: an experiment on domain cache. In *Proceedings of the Sixth International WWW Conference*, Santa Clara, CA. http://www6.nttlabs.com/HyperNews/get/PAPER21.html

O'Callaghan, D. (1995). A central caching proxy server for WWW users at the University of Melbourne. In *Proceedings of AusWeb95, the First Australian WWW Conference*, University of Melbourne, Australia. http://www.its.unimelb.edu.au/papers/AW12-02/

Padmanabhan, V. N.,  and Mogul, J. C. (1996). "Using predictive pre-fetching to improve World Wide Web latency," *Computer Communications Review*, 26. July 1996.

Pitkow, J., and Recker, M. (1994). A simple yet robust caching algorithm based on document access patterns. In *Proceedings of the Second International WWW Conference*. Chicago, IL. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/pitkow/caching.html

Pitkow, J, and Pirolli, P. (1997). Life, death and lawfulness on the electronic frontier. In *Proceedings of Human Factors in Computing Systems (CHI 97)*. Atlanta, GA. http://www.acm.org/sigchi/chi97/proceedings/paper/jp-www.htm

Pitkow, J. (1997). In search of reliable usage data. In *Proceedings of the Sixth International WWW Conference,* Santa Clara, CA. http://www6.nttlabs.com/HyperNews/get/PAPER126.html

Scheuermann , P., Shim, J., and Vingralek, R. (1997). A case for delay-conscious caching of Web documents. In *Proceedings of the Sixth International WWW Conference*, Santa Clara, CA. http://www6.nttlabs.com/HyperNews/get/PAPER20.html

Sedayao, J. (1994). "Mosaic will kill me network!" Studying network traffic patterns of Mosaic use. In *Proceedings of the Second International WWW Conference*. Chicago, IL. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/sedayao/mos_traf_paper.html

Smith, N. (1994). What can archives offer the World-Wide Web. In *Proceedings of the First International WWW Conference*, Geneva, Switzerland. http://www1.cern.ch/PapersWWW94/ngs.ps

Tauscher, L. (1996). Evaluating history mechanisms: an empirical study of reuse patterns in World Wide Web navigation. MS Thesis, Department of Computer Science. University of Calgary, Alberta, Canada. http://www.cpsc.ucalgary.ca/Redirect/grouplab/papers/96-Tauscher.Thesis/thesis.html

Tauscher, L. and Greenberg, S. (1996). How people revisit Web pages: Empirical findings and implications for the design of history systems. *International Journal of Human Computer Studies* 47(1). http://www.cpsc.ucalgary.ca/Redirect/grouplab/papers/97-HowUsersRepeat.IJHCS/RevisitArticle.ps.zip

The W3C HTTP-NG Web Characterization Group: Boston University Ocean's Group, Harvard College's Vino Group, INRIA, Microsoft, Netscape, Virginia Tech's Network Resource Group, and Xerox PARC. (1997). Work in progress.

Williams, S., Abrams, M., Standridge, C., Abdulla, G., and Fox, E. (1996). Removal policies in network caches for World-Wide Web documents. In *Proceedings of ACM SIGCOMM 96*, Stanford, CA. http://ei.cs.vt.edu/~succeed/96sigcomm/

Woodruff, A., Aoki, P., Brewer, E., Gauthier, P., and Rowe, L. (1996). An investigation of documents from the World Wide Web. *The World Wide Web Journal* 1(3). http://www5conf.inria.fr/fich_html/papers/P7/Overview.html

Wooster, R. (1996). Optimizing response time, rather than hit rates, of WWW proxy caches. MS Thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA. http://scholar.lib.vt.edu/theses/materials/public/etd-34131420119653540/etd-title.html

Wooster, R., and Abrams, M. (1997). Proxy caching that estimates page load delays. In *Proceedings of the Sixth International WWW Conference*, Santa Clara, CA. http://www6.nttlabs.com/HyperNews/get/PAPER250.html

Worrell, K. (1994) Invalidation in large scale network object caches. MS Thesis, Department of Computer Science, University of Colorado, Boulder, CO. ftp://ftp.cs.colorado.edu/pub/techreports/schwartz/WorrellThesis.ps.Z