WORLD WIDE WEB SITE VISITOR STUDIES TECHNIQUES USING SERVER LOG FILE DATA

By

Randy Michael Russell

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology and Special Education

1998

ABSTRACT

WORLD WIDE WEB SITE VISITOR STUDIES TECHNIQUES USING SERVER LOG FILE DATA

By

Randy Michael Russell

The World Wide Web has grown at a phenomenal rate. Much effort has been devoted to creating Web sites, including ones intended for educational use. Efforts to study the effectiveness of such materials have not, however, kept pace with site development efforts. Educators need tools to evaluate the effectiveness and influence of Web sites. Site developers need techniques to apply to formative evaluations of sites still under construction. Such techniques must allow researchers to produce results quickly, since the findings of many traditional approaches to educational research could be rendered obsolete prior to dissemination due to the rapid pace of evolution of the Web. Such methods of gathering formative feedback should also be straightforward enough to appeal to the many site developers who do not view themselves primarily as educational researchers.

The present study built upon methods used in museum visitor studies. Museum visitor studies researchers often use the time visitors spend viewing displays as a proxy indicator of the amount such visitors likely learned from those displays. Similarly, educational researchers have found correlations between students' "time on task" and learning outcomes. It would be useful to be able to measure "time on page" or "site visit durations" for visitors to Web sites. Such data could form the basis for determining whether correlations between Web site viewing times and learning exist.

This study used file request records stored in a Web server's log file as a

source of data for studying site visitor behaviors and trends. Such data is automatically recorded for all file requests by the Web server software, and is thus very simple to collect. These data were analyzed and displayed using inexpensive and easy-to-use server log analysis software, standard spreadsheet and graphing programs, and common database filtering and sorting techniques. Reports showing long term trends in page view and visitor counts for an entire site were created. Distributions of page views by time, site sections, network addresses, and other categories for a selected "typical" week were examined. Finally, detailed records of visit "paths" through the site and of visit durations for a smaller group of site visitors during that case study week were analyzed.

Server log data was found to be inadequate for accurately monitoring visit durations, largely because of gaps in the data record caused by caching of pages by visitors' browsers. Attempts to test correlations between "time on page" and learning outcomes should seek other means to monitor visit durations. Many of the methods employed in this study are, however, suitable for establishing broad-brush overviews of site usage trends, and supply useful data with minimal resource expenditures. The basic research techniques used here are scalable; evaluators can dig deeper into the data to uncover greater detail in a flexible, adaptable way. These methods can produce results in a short time, which is more suitable to the rapidly evolving Web than many traditional approaches to educational research. The methods used in this study are simple enough to be adopted by developers who are not primarily researchers. They provide information which developers can use to fine-tune ongoing site development, and lead to insights which might not be evident without such a formal approach to the study of a site's impact.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

INTRODUCTION

The explosive growth of the World Wide Web has generated a proliferation of new, widely accessible sources of information. Many Web sites are intended, in a broad sense, to fulfill educational purposes. The rapid growth of the Web has outpaced the rate of development of methods for studying the effectiveness of Web sites as tools to support teaching and learning. The primary purpose of this study is to explore methods for studying visitor behaviors of Web site users that are relevant to educational goals by analyzing visitors to a specific Web site.

The rise of the Web as a new publishing medium has provided many site developers with a public podium. Many developers of Web sites intended for educational purposes are not primarily educators by vocation; fewer still consider educational research their central concern with regard to the creation of their Web sites. A major emphasis of this study is discovery of research methods which can be used to improve educationally oriented sites, but which are sufficiently palatable to site developers who are not primarily researchers to be frequently applied in real world settings. This study also seeks to identify research methods which are sufficiently efficient to support studies with rapid turnaround times, since many traditional approaches to educational research produce results in a time frame unsuitable for the rapidly evolving Web, where the pace of technological progress can render information obsolete in a matter of months.

The Web is often used as a setting for informal educational endeavors. In this regard it is very similar to museums and zoos. This study draws heavily,

therefore, on the findings of museum visitor studies researchers for inspiration in formulating research methodologies. Because "Web visitor studies" are in their infancy, this study focuses on broad exploration of possible approaches to research in this arena, as opposed to trying to test well established principles of proper Web design to support specific types of learning.

Statement of Research Problem

The World Wide Web is a new phenomenon that is evolving rapidly. Methods for studying the impact of Web-based materials are in their infancy. Most Web site design is based on the intuitions of the site's authors, with little or no grounding in established principles of effective design based on research findings. Formative evaluation efforts to support improvement of sites under development are often limited or non-existent.

Many traditional methods for studying the educational impacts of materials could be applied to Web-based resources. For example, developers could directly observe learners using a site, could poll users via online surveys, or could administer tests to site visitors to assess learning associate with use of a site. Two major factors inhibit the widespread use of such techniques. Traditional educational research methods often require investigators to invest a large amount of effort into studies. Such studies often do not produce usable results for months or years after the study begins. Many Web site developers are primarily responsible for site development, and view site evaluation efforts as secondary tasks which can be conducted only if time and resource availability permit. They may be unwilling to commit to evaluation efforts which they view as overly complex, or to those which divert too many resources away from site development. Even if

developers are committed to evaluation, the incredible pace of change of the Web demands that research results be achieved in time frames which are much shorter than many educational research methods are designed to accommodate. Studies that take a year to complete may yield interesting results, but such results might be obsolete in terms of their utility in informing ongoing site modification efforts.

Many of the people involved in development of Web sites intended for educational use are not primarily educators by trade. Fewer still view themselves mainly as educational researchers. Site developers generally desire to make their sites as valuable as possible and are interested in feedback from users of their sites. They may not be willing, however, to invest a lot of effort into research techniques that they deem overly complex or obscure. Research methods which yield results of immediately apparent worth to such individuals could encourage them to value the research process and to gradually invest more effort into increasingly sophisticated studies.

Much of the research into techniques for studying Web site traffic and visitors so far has been conducted from the perspective of advertisers trying to measure the commercial impact of Web sites and their influence on users' buying habits. Such studies are of some value to educators, since much of the focus is on technological issues and the likelihood that a site visitor will recall information about a product is akin to certain types of learning of factual data. However, educators are interested in types of learning beyond simple recall of facts, and the final measure of successful educational efforts is usually not how much consumer spending on a certain product increases. Although the advertising oriented research techniques have some relevance to educational researchers, studies conducted from an educator's perspective would likely yield further insights.

Purpose of Study

This study is designed to help Web site developers choose site evaluation techniques which fulfill the specific formative evaluation needs of their unique sites. Techniques employed in this study are described in terms of the types of insights they provide, the amounts of effort or expertise required to use such techniques, and the limitations of those techniques in terms of reliability or availability of various types of information. One result of this study is a description of a suite of research techniques from which site developers can choose elements to assemble into an evaluation program suited to the analysis of their site. The techniques described in this report emphasize efficiency and scalability. The research techniques included herein can be used to produce results quickly and without a tremendous investment of effort. Evaluators can begin a study by using a small subset of the techniques described here, and can later expand the scope of the study by including more of these methods if they find such evaluations useful.

The approach used in this study takes advantage of automated data collection technologies supported by most Web server software packages. Those server log file generating technologies enable site evaluators to collect large amounts of detailed data about numerous site visitors with relatively little effort. Most of the data analysis techniques employed in this study also support analysis of records for many site visitors. The approaches to both data collection and data analysis used here are not dependent on overly specific data formats or processing software packages, thus insuring their widespread and ongoing availability to investigators and insulating them somewhat from idiosyncrasies of specific computer platforms and from changes in Web technologies over time. Other research techniques, such as visitor tracking

using cookie-based technologies, enable greater reliability and detail in tracking data as compared to the server log based approach used here, but at the expense of portability across computer platforms and server software packages and at a greater risk of obsolescence over time.

Four specific research questions are addressed in this study. The research questions posed by this study are:

1. What types of information relevant to educators is it possible to deduce about visitors and visitor behaviors, and with what degree of certainty?
2. What sorts of skills or tools, and what amounts of labor investments, are required to obtain those various types of information and degrees of certainty about such findings?
3. Is it possible to accurately measure a quantity such as "time on page" or "time on site," which could be tested for its correlation with learning outcomes in a fashion similar to the use by some researchers of quantities such as "time on task"?
4. What sorts of data should educational Web site developers collect and in what ways should they process those data to efficiently gain useful insights into how their own sights are being used?

The methods chapter of this report describes these questions and background information related to them in greater detail.

Limitations of Scope of Study

The data sources used in this study are Web server log files. These logs, which include records of all file requests received by Web server software, are generated automatically by the server software as a normal part of its operations. The great advantage of this source of data is that its collection requires very little effort on the part of a researcher. A second advantage is

that numerous server log analysis software packages which are easy to use are widely available. The log analysis programs produce reports that summarize server activity in variety of formats which can be customized by investigators. Sophisticated summaries of visitor behaviors can be produced for large numbers of visitors with relative ease thanks to these automated data collection and data processing programs.

This study uses data about the Digital Learning Center for Microbial Ecology (DLC-ME) Web site to provide illustrative examples of research techniques and the types of results they generate. This study is not about the specific results revealed concerning the DLC-ME site. The DLC-ME results are intended to show the types of information that site developers can expect to learn about their own sites by using the methods employed in this study. Specific trends concerning the DLC-ME site should not be construed as findings that should serve as general principles for other site developers. If this study had been intended to reveal principles of effective site design, the results from several diverse sites and from a variety of user populations would have had to be included. The DLC-ME examples included here serve merely to provide concrete examples of visitor studies techniques and the types of results they generate.

Server data logs have limitations and are by no means the only data source evaluators can use to study the behaviors and learning of visitors to Web sites. Many research techniques which are well documented elsewhere in the literature could be used to study Web site use. For example, researchers might wish to conduct user surveys, possibly online, to ask visitors about their experiences with and feelings about a site. There are also other potential Web usage analysis techniques which have advantages over the use of server log data. Such techniques require either greater expertise on the part of

evaluators, more time to conduct, greater expense, or some combination of these factors; they were ignored in this study in favor of the simpler and less labor intensive opportunities afforded by server log data analysis. I chose to focus my study of techniques for this new medium on the simplest methods unique to Web studies that held promise of generating useful results in a timely fashion.

Other investigators may wish to apply established methods to this new medium, or to document the value of the use of more sophisticated data collection or processing techniques. Researchers may wish to query Web users with surveys or interviews or to directly observe or videotape the actions of users. Investigator might collect and systematically code and analyze e-mail messages submitted by site visitors. Researchers could gather detailed data about visitors' clickstreams by installing software directly on users' computers, by using advanced server-side tracking technologies such as cookies (explained in detail later in this report), or by requiring visitors to enter a user ID code each time they logged onto a site. Although such techniques can yield data beyond that attainable via server logs, some have a disadvantage in that they are more intrusive upon research subjects, a concern often mentioned by museum visitor studies researchers when evaluating the suitability of research methods in informal education settings.

Chapter Two

LITERATURE REVIEW

Largely due to the youth of the World Wide Web, there has as yet been little published in academic literature about monitoring the use of Web sites intended to support education. However, several areas of research with longer histories provide insights into many of the problematic aspects of studying educationally-oriented Web sites. Likewise, progress in monitoring the use of Web sites in general, or from the perspective of promotion of commercial enterprises, has proceeded rapidly, and can shed light on some issues germane to the study of educationally oriented sites.

This review begins with consideration of literature about the relationship between time and learning. The amount of time during which learners are exposed to different educational materials or treatments is a critical control variable that researchers must account for in order to make fair comparisons between the outcomes caused by various treatments. A great deal of educational research has been dedicated to the study of the relationship between the amount of time students spend engaged in learning activities (generally referred to as "time on task") and educational outcomes. A major goal of this study is to determine whether time measures related to Web site usage, which can tentatively be labeled "time on page," can be reasonably established and to document the methods required to establish such metrics

Educational Web site use often involves informal educational environments and heterogeneous populations of site "visitors." Such use shares many characteristics with the educational roles played by museums and zoos. Many educationally-oriented Web sites are, in fact, "virtual"

representations of some physical zoo or museum. The second section of this literature review describes some of the research methods employed by and some of the findings of practitioners of the field of museum visitor studies.

People involved in commercial marketing of products share some of the goals of educators. They wish to convey some sort of message to other people and to have those people remember that message over time. Although educators often seek to impart more sophisticated forms of knowledge to students or to have students create their own understandings (as opposed to having them merely transferred from teachers), some forms of learning involve techniques similar to those employed by advertisers. The third section of this review describes the terminology used in the field of traditional (such as print, television, and radio) advertising media and some of the research that has been done in that field about how people remember information.

The rapid commercialization of the Web has created huge demand by commercial sponsors for accountability of online advertisers; sponsors wish to know whether their advertisements are being seen and whether they are being remembered by and are having an influence on consumers. This demand for accountability has driven rapid progress in efforts to develop terminology to explain and to develop software to monitor Web site visitation patterns. The fourth section of this review describes the progress that advertisers have made in these areas and examines the possibility of adopting or adapting some of those tools to support educational research causes.

Creators of Web sites presenting information about a wide variety of topics have gradually gained sophistication in the methods they employ for tracking use of their sites and in the language they use to describe such tracking. Some

of those methods and terms overlap the concepts used by advertisers, but others are more general purpose in nature. The fourth section of this literature review describes such visitor tracking concepts, especially those which are applicable to educationally-oriented sites in particular. Although some of these concepts are oriented toward support of Web sites from a technical perspective (such as estimating server load in number of bytes of data transmitted per hour at peak times), others are more relevant to monitoring activities of concern to educators (such as which page in a site was most frequently visited).

As the Web and its users have matured, most users have increasingly turned from browsing (following hypertext links from site to site) to searching as a major means for seeking information of importance to them on the Web. The fifth section describes the structure and function of directory and search sites. Most major search sites use similar techniques for creating catalogs of Web sites and pages and for reporting the results returned to users in response to queries. Many search sites employ autonomous software agents, called "Webcrawler robots" or "spiders," to build up their databases of site listings. An understanding of the way search sites and robots work can aid site developers and researchers in determining how visitors found their sites and which pages are most likely to serve as "entry portals" to a site from the search and directory sites.

Most people realize that the Internet as a whole and the World Wide Web in particular have grown at remarkable rates in recent years. The final section of this review presents data describing the rates of growth of the Web and of the Internet. Although precise definitions of concepts such as "connected to the Internet" and "Web site" are elusive, these data provide a useful backdrop against which usage rate trends over time of specific sites can be compared.

Time and Learning

Bloom (1974) concisely stated the basic relationship between time and learning, saying that "All learning, whether done in school or elsewhere, requires time" (p. 682). In order to learn anything from a Web site, visitors must spend time examining and interacting with the pages contained within it. This section describes research concerned with the relationships between time and learning. The first segment explains the role that time measures can play as research variables. The second subsection describes some of the correlations that have been found between time and learning. Finally, the third segment details some of the implications of measuring time for investigations of learning associated with educationally oriented Web sites.

Time as a research variable. In order to make comparisons between the effects of educational treatments, the amount of time during which learners are exposed to those treatments must be measured. Berliner and Fisher (1985) stated that "Unless duration is taken seriously in designing treatments and interpreting data from treatment comparisons in educational experiments, the potential for reaching faulty conclusions about the effects of educational treatments is quite high" (p. 345). Filby, Marliave, and Fisher (1977) asserted that keeping records of time allocated to instruction and the amount of time students spend engaged with learning activities or educational materials is a necessary requisite for researchers attempting to produce an adequate description of an instructional treatment. Fisher, Filby, and Marliave (1977) noted the importance of accounting for time on task when evaluating the success of educational initiatives. Good and Brophy (1995), in a critique of certain computer-based learning studies, observed:

> Several qualifications on these positive findings should be noted, however. First, there was no control for the amount of instructional

> time in about half of the studies, so that much of the reported achievement advantage to computer-assisted instruction may be due to greater opportunity to learn the material than to use of the computer. (p. 173)

In new, emerging fields of educational research, learning how to measure aspects of time relevant to experimental treatments should be a high priority.

The complexity of educational research often makes comparisons between methods employed, variables measured, and results obtained by various studies extremely difficult. Bloom (1974) claimed that:

> For the educational researcher, there are many attractive features in the use of time as a variable. Time can be measured with as much precision as the researcher desires. The measures of time have many properties that are almost impossible to secure in our conventional measures of academic achievement: equality of units, an absolute zero, and clear and unambiguous comparisons of individuals. Furthermore, time as a variable can be put into economic and resource costs for the individual learner, for groups of learners, and for the school and communities. (p. 684)

He further noted that, as a research variable, time makes available various scales, from seconds to years, for varying research tasks.

Which aspects of educational activities that can be measured in terms of time have researchers used in their studies? Carroll (1963) described a "model of school learning" that incorporates five factors—aptitude, ability to understand instruction, perseverance, opportunity, and the quality of instruction. He expressed three of these factors purely in terms of time. Aptitude is the amount of time, all other things being equal, for a given student to complete a specific learning task. Perseverance is the amount of time a student is willing to engage actively in learning. Opportunity is the amount of time the student is permitted to spend learning. In describing quality of instruction, Carroll stated that "the learner must be put into adequate sensory contact with the material to be learned (for example, one

must insure that the learner will adequately see or hear the materials of instruction)" (p. 726). He also noted that quality of instruction "applies not only to the performance of a teacher but also to the characteristics of textbooks, workbooks, films, teaching-machine programs, etc." (p. 726).

Most researchers interested in time and learning have identified two key elements which can be measured in terms of time. The first is allocated time, or the amount of time available for learning. The second is alternately referred to as engaged time or time on task. In Brophy's (1979) review of the findings of process-product research, he stated it revealed that "Students' opportunity to learn materials is a major determinant of their learning. This is indexed both by the time scheduled for instruction (allotted time) and the time actually engaged in learning activities (engaged time)" (p. 735). Anderson (1976) measured elapsed time and time-on-task and tested their relationship to achievement. Rich and McNelis (1987) distinguished between allocated time and time-on-task in their study of the use of student time in elementary schools. Filby, Marliave, and Fisher (1977) compared differences between "the amount of time devoted to instruction" and "student engaged time" across various classrooms.

Brophy and Good (1986) reviewed and summarized the findings of process-product research studies. They identified "student engaged time," "academic learning time," and time allocated by teachers to academic activities as significant correlates with student achievement. Fisher, Filby, and Marliave (1977) measured teacher allocations of instructional time and observed measures of student engaged time. Bloom (1974) found that "some students were spending three times as much time in active learning as others during the same period of elapsed time" (p. 686). Good's (1983) review of process-product research used the terms allocated time, engaged time, on-task,

and academic learning time. Time spent by students engaged with appropriate materials in a particular content area is the main measure of achievement in the Academic Learning Time (ALT) research program (Shulman, 1990). Shulman described the central constructs of process-product research as "teacher effectiveness, direct instruction, active learning, time-on-task" (p. 20). Gagné (1985) described the importance of measuring the time students spend in actual learning, or "time on task."

Relationship between time and learning. As noted earlier, Bloom (1974) concisely summarized the relationship between time and learning, stating that "All learning, whether done in school or elsewhere, requires time" (p. 682). The section immediately preceding this one describes how most researchers divide time affiliated with measuring learning into two categories: time allocated for learning or instruction, and student engaged time or time on task. In order to measure engaged time, researchers have turned to observable, overt behaviors that seem to indicate when students are paying attention to or expressing interest in learning activities or educational materials. This section describes some of the ways researchers have viewed attention and interest and the correlations they have found between allocated time, engaged time, and learning.

Keller (1983) developed an instructional design model focused on the importance of motivation, which included "interest" as one of the four basic categories of motivational conditions. Keller's definition of interest stated "*Interest* refers to whether the learner's curiosity is aroused, and whether this arousal is sustained appropriately over time" (p. 395). Keller, introducing his section on instructional strategies to promote interest, wrote "Practically every theory of learning includes some assumption about interest. A student has to at least be paying attention to a stimulus for learning to occur" (p. 398).

In a similar vein, Gage and Berliner (1979) described the necessity of attention for learning: "We have already noted, when discussing the information-processing model of learning, that without attention there can be no learning" (p. 334). Good (1983) reviewed the findings of process-product research, noting that studies had shown large variations between schools, classrooms, and students in rates of attentiveness, with rates within a single classroom varying by as much as 40 percent. Berliner and Fisher (1985) claimed that "Certainly, it is evident to all that attention, time-on-task, or some similar term is a necessary and possibly even a sufficient condition for some kinds of learning" (p. 338). Edminston and Rhoades (1959) measured rates of attention among high school students and compared them with standardized test scores.

Numerous educational researchers have reported results confirming correlations between allocated time, time on task, or both with learning outcomes. Bloom (1974) reported that he and his colleagues at the University of Chicago "have found that these indices of the amount of time the student is spending directly on the learning (either overt or covert) are highly predictive of the learning achievement of the student. The correlations when corrected for reliability account for about three fifths of the achievement variation of students" (p. 686). Fisher, Filby, and Marliave (1977) measured teacher allocations of instructional time and observed measures of student engaged time, and found a correlation between such time measures and achievement in reading and mathematics classes. Brophy and Good (1986) reviewed and summarized the findings of process-product research studies. Their list of teacher behaviors that maximize student achievement includes several measures of the quantity of content material covered or the amount of time allocated to learning. They described studies that indicated "student

engaged time," "academic learning time," and time allocated by teachers to academic activities are correlated with student achievement. They also described "opportunity to learn/content covered" as a major influence of achievement, stating: "Amount learned is related to opportunity to learn, whether measured in terms of pages of curriculum covered or percentage of test items taught through lecture or recitation" (p. 360).

Seifert and Beck (1984), in a study of high school algebra classes, found that "Achievement gain is positively correlated with minutes spent on task. This finding appears to indicate that the more time students spend on-task the more they will learn" (p. 9). Likewise, Anderson (1976) found the amount of time pupils spend on-task to be highly predictive of student learning. In another review of the findings of process-product research, Good (1983) observed that "Most recent studies of time and learning involve engaged time, reflecting the opinion of many persons that an indisputable relationship has been established between engaged time and amount of learning" (p. 130).

Gagné (1985) summarized his view of the relationship between time and learning as follows:

> The amount of time devoted to learning may be expected to affect the amount of learning. As a number of empirical studies have shown, the time students spend in actual learning ("time on task") is a particularly potent variable in the determination of what is learned, as indicated by student proficiency in school subjects. (p. 256)

Edminston and Rhoades (1959) found a positive correlation between attention and standardized test scores for high school students.

Berliner (1992) claimed that "time needed to learn is the crucial variable around which schools should be organized" (p. 9) and that "rate of learning is a better predictor of future learning than is intelligence" (p. 9). He strongly

advocated the importance to educational research of measuring time invested in learning. He described the relationship between learning time and achievement, stating "Teachers can find ways to give some students more time, thereby increasing their learning" (p. 9). In Good's (1979) review of the findings of process-product research, he noted that various studies reported that "the *time allocated* to instruction in a content area and the degree of student *engaged time* in reading and math is positively associated with students' learning gains in those subjects" (p. 56).

Implications of time for educational Web site research. Various measures of time, particularly allocated time and engaged time, have been shown to correlate with learning outcomes. Many researchers have stated that some sort of time measures are important, and possibly the most fundamental, elements to include in research projects concerned with learning. Time measures have characteristics that make them attractive variables to include in educational research studies. Especially in new, emergent areas of research such as Web-based learning, time measures should be a high priority ingredient to include.

A fundamental question for research about learning associated with Web-based resources, therefore, is: "Can reasonable, accurate measures of time affiliated with Web site usage be developed?" If so, what form should such measures take, how reliable are they, what are their limitations, and what methods are required to record them? Since many researchers distinguish between allocated and engaged time, it would be useful, if possible, to establish such distinctions with regards to Web usage as well. However, Web site use is often an informal, self-directed and self-selected educational activity. As such, it may prove difficult to separate indicators of interest, indicators of attention, allocation of time, and time on task from each other.

Finally, I wish to emphasize an intuitively obvious corollary to the notion that "the amount learned corresponds to the amount of time learners spend with educational materials." If learners spend no time with material associated with a particular topic, they are unlikely to learn anything about that topic. Thus, Web pages, or entire sites, that users never visit will fail to have an impact on the learning of those people. This notion is the reason that researchers who study visitors to museums, which is the topic of the next section, collect data concerning both the amount of time visitors spend viewing specific exhibits and which exhibits visitors skip over altogether.

Museum Visitor Studies

The study of visitors to museums and zoos is similar in many regards to the study of "visitors" to educationally oriented Web sites. Both museums and Web sites are typically informal educational environments, as compared to the formal educational environments of schools. Museum, zoo, and Web site visitor populations are heterogeneous in terms of age, educational background, and visit goals. Unlike Web visitor studies, which are still in their infancy, museum visitor studies have had time to develop and mature, and thus may offer insights into appropriate research methods and types of data to collect to support Web visitor studies research. This section describes some of the methods of museum visitor studies that seem likely to inform the study of Web site visitors.

This review of museum visitor studies and their relationship to Web visitor studies is composed of six subsections. The first section describes the elements of museum and Web visitor studies which are similar. The next section lists and describes research methods used in museum visitor studies.

The third section is an overview of the types of data collected by museum visitor studies researchers. The following section describes some findings from museum visitor studies relevant to Web visitor studies. The fifth subsection details some of the relationships between time, attention, interest, and learning that have been gleaned from museum visitor studies. Finally, the last subsection explains the aspects of museum visitor studies research that have been applied to this dissertation.

Comparison between museum and Web site visitor studies. The term "museum visitor studies" refers to the study of visitors to zoos, aquariums, greenhouses and gardens, and similar attractions, as well as various types of museums. Museums come in a variety of flavors, including art, science, history, children's, technology, and hands-on museums. The parallels between traditional zoos and the online, virtual Microbe Zoo Web site are obvious. However, several themes common to the study of visitors to zoos, the various types of museums, and many educationally oriented Web sites bind these attractions together in terms of research methodologies.

Like many Web sites, museums are typically informal educational environments. There is no explicit curriculum, tests and grading are largely absent, and attendance is voluntary. The "students" visiting these institutions form a heterogeneous population in terms of age, gender, educational background, interests, and purpose of visit. Affective, as well as cognitive, dimensions of learning play a large role in the educational impact of museum visits (Falk, 1983; Greene, 1988; Morrissey, 1991). Although visitors are frequently unable to recall much in the way of factual information from their trip to a museum or zoo, fascination with creatures and exhibits viewed often inspires interest, study, and learning at a later date.

Practical issues often assert a strong influence on the conduct of research

in museums and zoos. The people conducting studies are rarely full-time researchers, but commonly wear two or more hats at different times. Zookeepers must first tend the animals, while visitor studies are a peripheral sidelight. Museum curators catalog items and design and build exhibits; visitor studies are often an afterthought. Likewise, Web site developers are responsible for keeping the server running, coding pages, designing sites, and creating artwork. Site traffic evaluation is often a low-priority task that is set aside until other duties are fulfilled. In each environment, levels of staff availability to conduct research often dictates the scope of studies. Likewise, the research expertise of available staff members, whose primary training is usually not in research methodologies, can have a large impact on the nature of the research techniques employed (Haeseler, 1989).

Designers of museum research programs must also take great care to avoid intrusive studies (Beer, 1987; Falk, 1983; Haeseler, 1989). Museum visitors are decidedly not a "captive audience," and visitors who are keenly aware that they are being watched are likely to alter their behaviors or object to the observation.

Research methods employed in museum visitor studies. Two aspects of the methods employed during museum visitor studies research projects are detailed below. The first section describes some issues concerning the subjects selected for study. The second segment describes several data collection methods used by museum researchers, including written surveys, oral interviews, tests, and visual observation.

Many visitor studies choose groups of people, not individuals, as the basic social unit to investigate (Diamond, 1986; Greene, 1988; Morrissey, 1991). Factors such as group size and composition, and behaviors such as interactions between group members, play significant roles in research

designs. The kinds of information gathered are described later in the "Types of data collected during museum visitor studies" section. Web browsing, by contrast, is far more often an individual activity.

Techniques used for selection of research subjects seek to balance randomization requirements with practical issues. Subjects are often selected as they enter the museum; for example, every tenth group through the door or the third individual or group to enter after noon might be chosen as prospective subjects. Koran, Foster, and Koran (1989) studied undergraduate students who were specifically sent to the museum and instructed to view certain exhibits. Falk (1983) studied London schoolchildren who had been told to learn about cells in an exhibit adjacent to the one of interest; they were chosen partly because of the ease with which pre- and posttests could be administered through their schools. Morrissey (1991) observed all groups that entered an exhibit during a limited, predetermined time. Beer (1987) collected data on all visitors, except those who were part of tour groups or were not English-speakers, over an extended period of time. Diamond (1986) tracked the first group of an appropriate composition through the doors after 1:00 p.m. on the days the study was being conducted.

Museum visitor studies researchers employ numerous data collection methods to observe the behaviors of and gain insights into the thoughts of visitors. Written surveys, generally administered to visitors as they exit the museum, supply data about aspects of the visitors' experiences and opinions of their visits (Falk, 1983; Pierce, 1989). Surveys are used to inquire about factual information, such as which exhibits visitors viewed, how long they stayed at the museum, and what they remember about the displays they perused. Surveys can also help researchers understand affective aspects of visitors' experiences, such as which exhibits they found most and least

interesting (Koran, Foster, & Koran, 1989). Because surveys involve self-reported data from the research subjects, care is need to distinguish between what people say they did and how they actually behaved. For example, reports by visitors of their estimates of the amount of time spent in specific display areas is often inaccurate.

Researchers frequently use written tests of knowledge to determine what, if anything, visitors learned in the course of their museum visit (Patterson & Bitgood, 1988). Often, due to practical limitations of staffing or desire not to intrude upon visitors, only posttests administered at the end of the visit are employed (Koran, Foster, & Koran, 1989; Pierce, 1989). In other cases, visitors are given both pre- and posttests (Falk, 1983), which provide a truer picture of the change in knowledge, or learning, that occurred during a visit.

Oral interviews are another method used to encourage visitors to report upon their visit experiences (Beer, 1987; Diamond, 1986; Falk, 1983; Greene, 1988). Such interviews can encompass the same scope of topics as written surveys or tests, including attitudes, learning, and self-reports of behaviors. Oral interviews are often considered less intrusive or threatening than written surveys or tests, and thus may be employed for practical reasons to elicit more user response.

Visual observation (or "tracking") of museum visitors is commonly used to directly note the actual behaviors of research subjects. In some cases, a researcher actively follows subjects, sometimes overtly but more often covertly, during their visit and records behaviors of interest. In other instances, video cameras deployed within the museum are used to covertly observe behaviors. The cameras may be specifically placed for use by the researcher (Morrissey, 1991), or may be preinstalled security surveillance cameras that can also be used for visitor studies (Falk, 1983). Studies that use

cameras can record all behaviors from a given perspective for later analysis. Tracking by human observers generally requires use of some sort of coding or behavioral rating scales (Diamond, 1986; Falk, 1983) to convert observational data to quantitative variables for analysis.

Where and when does "tracking" of visitors commence and where does it halt? Tracking may begin when visitors enter the museum (Diamond, 1986; Haeseler, 1989), when they enter a section of the museum (Haeseler, 1989), as they approach a specific exhibit (Koran, Foster, & Koran, 1989), or when they wander into an area which includes an exhibit of interest to the researcher (Morrissey, 1991). Visitors may be tracked only if they enter the facility through a specific door (Diamond, 1986). Tracking which begins upon entry into the museum generally ends when the visitors exit the building (Haeseler, 1989). Haeseler (1989) also noted that head counts of entrance and exit volumes over fixed time intervals are occasionally used to estimate the lengths of visits for large visitor volumes.

Time spent in the museum or viewing specific exhibits is one of the most common visitor behaviors noted by museum studies researchers (Beer, 1987; Diamond, 1986; Haeseler, 1989; Koran, Foster, & Koran, 1989). Haeseler (1989) described four techniques frequently employed to measure such time; they include rough estimates by the management of the facility, estimates by visitors gleaned from exit interviews, recording entrance and exit times of individual visitors, and covert tracking of visitors during their visits. Data about how visitors spend their time, as ascertained by observation during tracking, is a crucial ingredient in many studies (Greene, 1988; Koran, Foster, & Koran, 1989; Morrissey, 1991).

Types of data collected during museum visitor studies. This section describes some of the types of data museum researchers collect during visitor studies. The first segment lists demographic information about museum visitors that researchers gather. The second subsection describes aspects of visitor behaviors noted by museum studies researchers. The next segment explains some of the ways researchers attempt to gauge the attitudes of visitors. The final portion briefly describes the impact of architecture, exhibit design, and various other traits of the museum and the displays that affect visitors' experience.

Visitors to museums and zoos make up a heterogeneous population. Visitor studies researchers therefore consider collection of demographic information about this subject population critical to research (Patterson & Bitgood, 1988; Pierce, 1989). Because they are both important and easily ascertained, the age (Diamond, 1986; Patterson & Bitgood, 1988) and gender (Diamond, 1986; Morrissey, 1991) of visitors are some of the most commonly gathered bits of demographic data. Other important, but less easily determined, data include educational level, socioeconomic status, and ethnicity. Some research projects necessitate knowledge of visitors' previous museum experience (Diamond, 1986), whether this is a subject's first visit to this particular museum (Pierce, 1989), or how a visitor found out about the museum involved in the study (Pierce, 1989).

Many museum visitors arrive in groups. Therefore, whether an individual is part of a group is an important piece of research data (Diamond, 1986; Morrissey, 1991). If a subject is a member of a group, her or his relationship to other group members may be significant (Diamond, 1986). The size and composition of groups also can impact the behavior of its members (Morrissey, 1991). Significant aspects of group composition include: presence

of children, presence of adults, presence of adults and children, and presence of adults but no children. Three specific, distinct group types commonly found in museums and zoos are family groups, school groups, and groups that are part of organized tours. Family groups make up the most numerous types of visitors to museums and zoos (Patterson & Bitgood, 1988). School groups represent the second most numerous visitor population (Patterson & Bitgood, 1988).

The second major category of data collected by museum visitor studies researchers concerns the behavior of visitors during their visit. Greene (1988) studied the interaction between individuals within a group. Whether members of a group or alone, the duration of their museum visit and the amount of time spent in certain galleries or in the proximity of specific exhibits are crucial aspects of visitors' behavior recorded by numerous researchers (Beer, 1987; Diamond, 1986; Falk, 1983; Greene, 1988; Haeseler, 1989; Koran, Foster, & Koran, 1989; Morrissey, 1991; Patterson & Bitgood, 1988; Pierce, 1989). Researchers have also observed whether visitors stopped at or skipped over certain displays (Beer, 1987), and whether visitors read labels, touched manipulables, or listened to audio clips associated with individual exhibits (Beer, 1987; Greene, 1988; Pierce, 1989). Pierce (1989) studied label-reading behaviors extensively, noting the percentage of visitors who stopped and read labels, duration of label reading, total time in the exhibition hall, time spent viewing exhibit objects, and performance on a quiz about label contents.

Museums and zoos can exert strong influences on the attitudes of visitors towards the topics and creatures represented in such establishments. Museum visitor studies researchers attempt to observe behaviors that indicate the interests and opinions of visitors. Indicators of interest include time spent

observing certain displays (Patterson & Bitgood, 1988) or responses rating levels of interest expressed via a written survey (Koran, Foster, & Koran, 1989). Pierce (1989) sought to understand visitors opinions of their museum experience by asking them which exhibits they liked best and least, what they would come to see again, and what needed to be improved.

Museum visitor studies researchers also realize that characteristics of the museum buildings, zoo grounds, and the exhibits and displays found therein can play a large role in how visitors react (Patterson & Bitgood, 1988). The scope of the variables in this realm which can influence visitors' experiences and learning is large. Factors as gross as the overall architectural layout of museum buildings to as minute as the choice of font size on label placards can affect the way visitors react to the items housed in zoos and museums.

Selected findings of museum visitor studies research. This section describes some results of museum visitor studies relevant to my dissertation research. The first segment describes relationships between time, attention, interest, and learning. The second part describes factors that influence how visitors allocate their time when visiting a museum or zoo.

Morrissey (1991) stated that "In the museum field, time has been consistently correlated with both cognition and affective outcomes" (p. 110). Falk (1983) noted the relationship between pre- and posttest scores ("change scores") of schoolchildren and the time they spent viewing the exhibit which was the focus of his study. He found that "Raw time scores showed a significant amount of correlation with change score ($r = 0.597$, $p < 0.0001$)" (p. 272). In their study of visitors to a natural history museum, Koran, Foster, and Koran (1989) discovered that "Multiple regression analysis also indicated that there was a significant relationship between attention and the score on the criterion measure, with greater attention yielding greater learning

($p<.05$)" (p. 242). They quantified attention in terms of time, further stating that "the factor that was most important to learning was the length of attention" (p. 243). Based on these and similar findings, many museum studies researchers have developed a rule of thumb that claims that time spent observing, or paying attention to, an exhibit can serve as rough indicator of how much a visitor is likely to learn from that exhibit.

Attention, being a mental state, is not directly measurable. Museum visitor studies researchers infer "attention" from the behavior of visitors. The amount of time visitors are observed to spend viewing an exhibit is frequently used as a quantifiable measure of attention (Koran, Foster, & Koran, 1989). Whether a visitor stopped to view or entirely skipped over a certain exhibit is another basic indicator used to gage whether the exhibit captured the visitor's attention (Beer, 1987). Beer also considered signs of attention to include whether a visitor read an exhibit's label, touched manipulables, or listened to audio clips.

Attention is an immediate, short-term mental state. Visitor studies researchers also attempt to gage attention's longer lasting sibling, interest. Zoo and museum visits often evoke affective as well as cognitive reactions from visitors, leading researchers to desire means to measure the levels of interest in topics that visitors have or develop (Falk, 1983; Greene, 1988). Koran, Foster, and Koran (1989) used responses to a written survey, which had visitors rate exhibits on a Likert scale ranging from "dull" to "very interesting," as a means to measure interest.

Behaviors that indicate attention are proxy measures representing "likelihood of learning." Koran, Foster, & Koran (1989) used written posttests as a more direct measure of visitors' knowledge at the end of their museum stay. Falk (1983) used both written pretests and posttests to determine the

change in visitors' knowledge, or learning. Patterson & Bitgood (1988) asked children to describe the aspects of their museum visits that they remembered, and found that "In recall tests, children generally mentioned the exhibit where they had spent the most time" (p. 44). Time, attention, interest, and learning—and methods for measuring them—are intimately interwoven in museum visitors studies research.

Since time spent by visitors in museums is frequently correlated with the amount they learn, factors which influence the duration of visitors' stays may affect how much they learn. Haeseler (1989) described several factors that influence the amount of time visitors spend at museums and zoos. They include the facility's setting (whether indoors or outside), the attraction of exhibit contents, visitor services (such as gift shops and snack bars), visitor fatigue, seasonality, crowding, demographics of visitors (especially the child/adult mix of groups), and time budgets (such as plans to visit other attractions in the same day). Aspects of attraction content that strongly influence time expenditure include the extent or size of the attraction (Haeseler, 1989; Patterson & Bitgood, 1988; Pierce, 1989), dynamic versus static exhibits (Haeseler, 1989; Patterson & Bitgood, 1988; Pierce, 1989), and physical versus sedentary activities (Haeseler, 1989).

Besides examining factors that affect the duration of an entire visit to a museum or zoo, visitor studies researchers also assess factors that influence the time spent with individual exhibits or displays. Factors that play a role in time spent with an exhibit include the presence of moving parts or specimens, the size of an exhibit, and the location of an exhibit within a gallery (Patterson & Bitgood, 1988; Pierce, 1989). Pierce (1989) intensively investigated the impact that labeling had on time spent with exhibits, finding that the size of the text, the amount of text, and the location of the label—both

with respect to the exhibit and with respect to the natural field of view of visitors—influenced viewing time behaviors. Not surprisingly, Pierce (1989) found that visitors who read labels spent more time viewing exhibit objects than those who did not. However, Greene (1988) noted that "only one in four zoogoers will read an informative sign" (p. 51).

Greene (1988) claimed that "Zoogoers look at exhibits for about 90 seconds" (p.51). In her study, Diamond (1986) found that "The average science museum visit lasted slightly over two hours" (p. 143). In each of these cases, the duration reported represents an average of all research subjects and gives no indication of the variation about that average. Using a slightly different approach to describe the amassed behavior of numerous subjects, Diamond (1986) reported that "57% of the exhibit visits lasted less than one minute" (p. 144). Data about the amount of time spent by visitors in museums and viewing exhibits is a major element of many visitor studies projects. Numerous factors influence the duration of visits and time spent with exhibits, and there are a variety of means for summarizing time-based data across large groups of research subjects.

Implications of museum visitor studies for Web visitor studies. This section describes some of the research methods, questions, and findings from museum visitor studies research which are appropriate to apply to Web visitor studies research. Some aspects of museum visitor studies are not practically applicable to Web visitor studies; these aspects are noted as well.

Time spent visiting a museum or zoo, or viewing specific exhibits plays an important role in museum visitor studies and is often correlated with learning. It is important, therefore, to attempt to measure the amount of time Web users spend visiting an entire site and various portions of it. This goal raises important questions. Is it possible to measure the amount of time Web

users spend visiting a site? If so, what methods can be used to do so, what technical or other constraints influence or restrict such measurement, and what form will the data obtained take? Museum researchers study the duration of entire visits to museums and the duration of interactions with individual galleries or exhibits. Duration of a user's visit to an entire Web site, to sections of a site, and to individual pages within a site are natural analogs to museum, gallery, and exhibit visit durations. Measurement of these aspects of Web site visits, if possible, should be a priority.

Tracking of visitors through museums is a major visitors studies research tool. Can visitors to a Web site be similarly "tracked," noting which pages they viewed and in what order, how long they tarried at each page, and what they did at each such location? What are the technical hurdles or limitations to such tracking? Is it possible to determine which pages, like exhibits, were skipped over and which were viewed? Can the entry point, analogous to the external doors of a museum, via which visitors arrived at a Web site be determined? Is it possible to determine when a visitor departs, and through which "door"?

Museum researchers use surveys, interviews, and tests to study visitors. Such instruments can also be used to study Web site visitors. To some extent, surveys and tests can be administered remotely using electronic renditions of the tests or surveys. Proper development of surveys and tests is the subject of entire disciplines of study. Although use of such instruments could prove valuable to Web visitor studies researchers, that topic is beyond the scope of this research project. My study will focus on the use of techniques analogous to tracking, leaving investigation of techniques for directly questioning site users to other researchers.

Which site visitors should research studies focus on? One approach is to

study a relatively small number of visitors intensively, using surveys, tests, tracking, or some combination of techniques. Another approach is to gather less data per visitor but to gather such data for a much larger subject population, using techniques such as "head counts" of visitors arriving through the doors during certain intervals. Another issue of data collection scope involves duration of study, such as whether visitors being tracked should be studied for the duration of their visit to the entire site or only in the vicinity of specific attractions or pages. Studies could focus on all visitors over a predetermined time, on randomly selected visitors from the overall population, or on selected visitors matching a demographic profile of particular interest. The focus of many museum studies on groups of visitors, as opposed to individuals, may distinguish museum studies from Web studies. Web use is generally a solitary activity, and it is unclear to what extent it is possible to remotely note the presence of multiple users sharing a social Web browsing session.

Visitor demographics are another important element of museum studies. Is it possible to determine demographic information about Web users, and, if so, which information and by what means? Some of the key demographics which museum researchers commonly note include age, gender, educational level, socioeconomic status, and ethnicity. Some of these data, though easily determined at a glance in museums, may be difficult or impossible to ascertain in the case of Web visitors. Some researchers have inquired of visitors whether this was their first visit to a particular museum, or how they found out about the museum or zoo. It might likewise be useful to know whether Web visitors were "first-timers" or repeat visitors, and how they discovered or were led to the Web site.

Finally, it is important to note that pragmatic issues strongly influence

research in museums. Availability of staff, and in some instances people with certain skills or expertise, often exerts a tremendous influence on the scope of museum visitor studies projects and sometimes dictates the data collection and analysis techniques which are used or avoided. Avoiding intrusive studies that interfere with visitors' museum experiences is another important concern of museum researchers. Web site studies researchers should also keep an eye out for the influence such practical issues may exert on the design and implementation of Web visitor studies research projects.

Advertising Media Metrics

The World Wide Web has opened a new frontier for commercial advertising. Advertisers have vast financial resources available, as compared to educators, to apply towards getting their messages across to consumers. The information transfer goals of advertising are, in some respects, similar to some educational endeavors, especially those involving direct instruction or the learning of factual information. Advertisers are accountable to their sponsors for ascertaining the effectiveness of advertising campaigns. The combination of financial backing and need for accountability has pushed Web-based advertising to the forefront of efforts to monitor the use of Web pages and sites. Most software that has been developed for monitoring Web use is primarily intended for measuring the effectiveness of advertising. Likewise, much of the terminology and many of the techniques applied to monitoring Web usage have their roots firmly planted in the advertising world. This section reports on the approaches used by advertisers to understand Web usage. The relevance of this issue to educators is whether such techniques can be applied to the study of learning associated with Web

sites, which aspects of advertisement monitoring are applicable to learning assessment, and what modifications to advertisers' approaches are needed to support educational research.

The first portion of this section describes studies of remembering that advertisers apply towards determining how many times people need to be exposed to advertisements in order to recall them. The second segment describes metrics used to measure advertising campaign effectiveness in traditional media, such as newspapers, radio, and television. The final section describes measures of advertising effectiveness that have been adapted to or developed expressly for measurement of Web-based advertisements.

Studies of remembering. A core concern of advertisers is for consumers to remember advertisements long enough to influence purchasing behavior. Factors that influence such remembering include how many times a person was exposed to an advertisement, and how long after such exposures the memory needs to persist. Ebbinghaus (1855/1964) is generally credited with conducting the first scientific study of the relationship between time and learning (Good & Brophy, 1995; Surmanek, 1993). Ebbinghaus investigated his ability to remember invented nonsense syllables. His data concerning the relationship between the amount of time spent memorizing syllables and the amount he could recall led to the invention of the concept of a "learning curve." Ebbinghaus also explored the rate at which the memorized syllables were forgotten over time. His work, and derivatives of it, are used by advertising researchers to estimate the number of exposures to a commercial message, and the desired time frame for those exposures, required to assure that consumers will remember that message.

Zielske (1959) measured the rates of remembering and forgetting of print media advertisements among consumers, using two different advertising rate

schemes. Both groups of subjects were exposed to a total of 13 advertisements. Group I subjects were exposed to the advertisements once per week for 13 consecutive weeks. Group II subjects were exposed to advertisements once every four weeks, spread out over a year-long period. Members of Group I had the highest peak recall rate of any single week, showed a rapid increase in recall rate early in the advertising campaign, displayed a rapid dropoff in recall rate after the advertisements stopped, and had a lower level of recall at the end of the one-year study period. Subjects in Group II displayed a gradually increasing recall rate that rose in a "sawtooth" pattern, and had a higher recall rate at the end of the year.

Zielske and Henry (1980) examined the rates of remembering and forgetting of television advertisements by consumers, using a variety of advertisement rate patterns. They tested various exposure rates, ranging from all advertisements presented in a "burst" over a short time period to very gradual, periodic exposure over a long time frame. Their results for televisions advertisements were similar to those of Zielske's (1959) earlier study of print advertising. They found only minor variations between recall patterns of print and television advertisements. Advertisers are used to dealing with advertising campaigns that span several media types, and prefer metrics that translate well across multiple media types. Since these studies indicate that advertisement recall rate patterns are similar between print and television advertisements, advertisers are inclined to believe that metrics of advertisement effectiveness can span media types. They are prone, therefore, to attempt to apply measures used with traditional media to Web-based advertising strategies. Surmanek (1993) noted that most advertisers assume a "3+ exposure" rate as a rough rule of thumb; consumers must view an advertisement, irrespective of media type, at least three times in order to

have a substantial likelihood of remembering it in a way that would influence purchasing behavior.

Traditional advertising media. Traditional advertising media include newspapers, magazines, billboards, direct mail, television, and radio. Surmanek (1993) presented definitions of the terms used by advertisers to assess the scope and success of advertising campaigns, including impressions, reach, frequency, duplication, ratings, share, HUT, PUR and PUT. His definitions, along with the pages in his book upon which the definitions are stated, are summarized in Table 1. Surmanek noted that reach takes duplication into account, while impressions does not:

> You've noted that GRPs and impressions are indicators of *gross* delivery, without regard for duplication. Neither indicates how many *different* people will be exposed to a medium; **reach** does. Reach is the number of different individuals (or homes) exposed to a media schedule within a given period of time. (p. 106)

Another concept used, though not exclusively, by advertisers, is that of an "index" that indicates upward or downward trends in values over time. The Consumer Price Index is an example of such use of this term. Surmanek defines index, in this context, as "A number indicating change in magnitude relative to the magnitude of some other number (the base) taken as representing 100. A 110 index indicates a 10 percent positive change in magnitude; a 90 index a 10 percent negative change" (p. 84). A common use of such an index is to indicate the weekly change in television program viewership rates.

Advertisers acknowledge that they are able to count only the advertisements that are delivered to consumers via mass broadcast media, not the ones that consumers actually see or listen to. Surmanek (1993) explained this distinction as follows:

Table 1 - Definitions of Traditional Advertising Terms

Term	Definition (page number in Surmanek, 1993)
Impressions	The gross sum of all media exposures (numbers of people or homes) without regard to duplication. (p. 81)
Reach	The number or percentage of a population group exposed to a media schedule within a given period of time. (p. 106)
Frequency	The number of times people (or homes) are exposed to an advertising message, an advertising campaign, or to a specific media vehicle. Also, the period of issuance of a publication, e.g., daily or monthly. (p. 125)
Duplication	The number or percentage of a medium's audience, or of those reached with a media schedule, who are exposed to more than one media vehicle or to more than one advertising message. (p. 332)
Rating	The percentage of a given population group consuming a medium at a particular moment. Generally used for broadcast media, but can be used for any medium. One rating point equals one percent. (p. 51)
Share	"Share of audience" is the percentage of HUT (or PUT, PUR, PVT) tuned to a particular program or station. "Share of market" is the percentage of total category volume (dollars, units, etc.) accounted for by a brand. "Share of voice" is the percentage of advertising impressions generated by all brands in a category accounted for by a particular brand, but often also refers to share of media spending. (p. 65)
HUT	The percentage of Homes Using (tuned in to) TV at a particular time. (p. 60)
PUR	The percentage of People Using Radio at a particular time. (p. 60)
PUT	The percentage of People Using TV at a particular time. Identical to *PVT*, People Viewing TV. (p. 60)

> A rating, therefore, is only an indicator of the percentage of a group of individuals that have the *opportunity* to be exposed to the advertising. The percentage of the people who will actually see or hear the commercials can vary substantially, ranging from zero (although this is highly improbable) to 100 percent of the viewing/listening audience. (p. 53)

Ratings for television programs are typically calculated for each 15-minute segment of broadcast time; ratings for commercials within such programming blocks are inferred to be the same as for the overall quarter-hour broadcast segment. This approach is obviously imprecise, for people often leave the room to raid the refrigerator during commercial breaks. Factors such as the creative effectiveness of a particular commercial and the relative position of a commercial within a commercial break can dramatically influence the likelihood that a given advertisement will be watched.

Bayne (1997) noted that advertisers often inform consumers of phone numbers or mailing addresses via which they can obtain further information about an advertised product. She further explained that such information can be used to track advertisement effectiveness by assigning different addresses or phone numbers to different advertisement media types, geographic advertisement distribution regions, or other distinct channels. Bayne further noted that use of these techniques can be extended to electronic media: "To track the effectiveness of your traditional marketing communications programs, assign different e-mail addresses or different Web page addresses to each activity" (p. 349).

Web-based advertising terminology. Advertising and other commercial enterprises have expanded rapidly on the World Wide Web. In order to communicate about the relative effectiveness of Web-based advertising campaigns, advertisers have invented a preliminary set of terminology. This new vocabulary draws upon traditional advertising terms, incorporates early

conventions of Web terms that are not specific to commercial endeavors, and includes terms that have been invented solely to describe events affiliated with advertising on the Web. This section describes some of the terminology that has been developed so far to explain Web-based advertising, emphasizing terms that are used reasonably consistently by different groups.

"Hits" is probably the most fundamental term applied to the measurement of Web site activity. A hit is a request received by a server for a transfer of a single file of any type ("CASIE guiding principles," 1997; "Microsoft Site Server," 1997). Files might include HTML pages, images such as JPEGs or GIFs, Java applets, or any of a number of other file formats. Since Web pages typically include several elements, especially graphics, a request for a single page can generate numerous hits. This fact, along with the widely variable number of elements that comprise different pages, drastically detracts from the usefulness of hits as a measure of Web activity. However, largely because of the ease with which hit counts can be tabulated, hits are probably the most widely reported Web traffic metric. Some groups distinguish between requests received by the server and files successfully transferred in reporting hits ("Glossary of NetCount terms," 1997), which accounts for events such as server errors that prevent file transfers when heavy traffic overloads the file server.

A "visitor" or "user" is an individual person who visits a Web site ("CASIE guiding principles," 1997; "Glossary of NetCount terms," 1997; Lee, 1996a; "Microsoft Site Server," 1997). This definition does not explain how the identity of a unique individual should be determined, which can be a difficult issue to resolve. Some groups also refer to "identified users," defined as visitors about whom demographic data is known ("CASIE guiding principles," 1997; "Glossary of NetCount terms," 1997; Lee, 1996a). A "visit"

generally refers to a series of consecutive requests made by a single user at one Web site ("CASIE guiding principles," 1997; "Glossary of NetCount terms," 1997; "I/PRO: FAQ," 1997; "Microsoft Site Server," 1997). The end of a visit can come when the user leaves the Web site, or when a predetermined "timeout" period (typically 30 minutes) elapses between requests. A "session" is a series of transactions by a single user that spans multiple Web sites ("Glossary of NetCount terms," 1997).

In Web parlance, a "banner" or "advertisement" is typically a clickable advertisement that links to the advertisement's sponsor's Web page ("Glossary of NetCount terms," 1997). Banners are often but not always located at the top of a Web page. Banner advertisements are usually some sort graphic, but can be an animation, Java applet, or other element. Advertisers are concerned with whether consumers actually see their advertisements, and use the term "ad views" to refer to number of times that an advertisement banner has been downloaded and presumably viewed by visitors (Cooper, 1996; "I/PRO: FAQ," 1997). Two terms borrowed from traditional advertising lingo, "impressions" (Andrews, 1997a; "Glossary of NetCount terms," 1997; Lee, 1997a) and "exposures" ("I/PRO: FAQ," 1997), are essentially synonymous with ad views in the manner they are used in the language of Web-based advertising.

Once consumers have seen, or at least had the possibility of being exposed to, a banner advertisement, advertisers wish to know whether they react to that advertisement. An "ad click" or "clickthrough" describes a situation in which a visitor clicks on an advertisement banner ("I/PRO: FAQ," 1997). Some advertisers reserve "clickthrough" for advertisement clicks that successfully deliver the visitor to the advertiser's Web site ("Glossary of NetCount terms," 1997), thus factoring in failed transfers due to network

errors or busy servers. "Clickthrough rate," "clickthrough ratio," and "ad click rate" all describe the percentage of ad views that result in clickthroughs or ad clicks (Andrews, 1997a; Bayne, 1997; "Glossary of NetCount terms," 1997; "I/PRO: FAQ," 1997).

"Clickstream" describes the path, in terms of pages requested and clicks registered, that a visitor follows while viewing a Web site (Bayne, 1997). Clickstream information may include data about how much time a user spent on each page or where they went upon leaving the site. Whether included in clickstream analyses or not, advertisers sometimes gather other time-based information, such as the "average time on page" ("Glossary of NetCount terms," 1997).

Web Visitor and Traffic Tracking Technologies

The realm of Web visitor and traffic tracking, like the World Wide Web itself, is very young and is evolving rapidly. This section begins with a description of the terminology and metrics currently in vogue. Next, it describes the technologies, standards, and conventions presently being used to measure and compare Web site traffic. The third segment describes the software that has so far been developed to log and analyze site traffic. The last part of this section describes some of the factors that interfere with traffic tracking efforts, and mentions some of the emerging and projected future technologies that may alter traffic measurement procedures and reshape the World Wide Web and other educational uses of the Internet.

The World Wide Web represents an incredibly vast collection of information resources. Methods for finding appropriate, relevant information amongst the innumerable sites and pages are important skills for

Web users to acquire. Similarly, efforts to determine significant patterns of Web site use as a result of monitoring visitation patterns of users of even a moderately popular site can generate huge quantities of data. Collection of certain types of data can be a fairly simple, straightforward process. Making sense of such collections of data is a much more difficult undertaking.

Computer scientists have begun to apply computational technologies to the problem of transforming large quantities of data into useful, comprehensible information sources. Computer scientists refer to techniques designed to extract useful information from large collections of data, typically stored in some sort of database, as "data mining" (Glymour, Madigan, Pregibon, & Smyth, 1996). Fayyad, Piatetsky-Shapiro, and Smyth (1996) described the situation as follows:

> As we march into the age of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining. (p. 27)

Data mining techniques have been applied to problems in fields such as business (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996), science (Fayyad, Haussler, & Stolorz, 1996), medicine, and government (Fayyad, Piatetsky-Shapiro et al, 1996). Their use in education has so far been quite limited. Efforts to make sense, from an educational perspective, of the usage and navigation patterns of visitors to Web sites intended to support learning are, in part, an application of data mining techniques to the field of education.

Current terminology and metrics. This section describes the types of measures that are being used to gage Web site traffic, and the terms so far adopted to refer to those measures. An earlier section of this report described terminology that advertisers use to monitor Web traffic; this section is concerned with the broader audience of all people and groups involved with Web traffic monitoring. Terms that are used by that broader community, as well as by advertisers, are described here only briefly; the reader is directed to that earlier section for more detailed explanations of those terms.

"Hits" are a measure of the number of files of any type requested for transfer by a Web server (Buchanan & Lukaszewski, 1997; Stout, 1997). Some authors distinguish between files requested and files actually transferred, which accounts for events such as server errors or network transfer errors (Shaffer, 1996). Because of the widely variable number of elements that can appear on a single Web page, the number of hits recorded when a visitor requests a page is extremely page-content dependent (Cooper, 1996; Shaffer).

"Page views" refers to the number of HTML documents, or pages, transferred to users by a Web server (Buchanan & Lukaszewski, 1997; Cooper, 1996; "New media companies," 1996). Page views are sometimes shortened to "views" (Stout, 1997). Lee (1996a) noted that sites that employ frames complicate the issue of defining a page, since an item that appears on a visitor's screen as a single "page" actually contains multiple HTML "pages."

"Visitors" are individual people who view portions of a Web site (Buchanan & Lukaszewski, 1997). Lee (1996b) notes that some software packages also detect "repeat visitors" who return to a site for multiple visits. "Visits," "unique visits," "sessions," or "user sessions" are a series of sequential hits by a single visitor (Buchanan & Lukaszewski; Lee, 1996a; Stout, 1997). Visits duration is defined in terms of the user's entry into and exit from

the site, or by a presumed "exit" when 30 minutes or more passes between successive hits. Buchanan and Lukaszewski also distinguish between hits by actual human visitors and hits registered by the Webcrawler "robots" or "spiders" that stock search engine databases, referring to activity generated by such software entities as "spider visits."

Buchanan and Lukaszewski (1997) use the term "referring links" to describe external sites and pages that users "came from" upon "arrival" at the site being studied. They also define terms for noting the apparent network or geographic affiliations of site visitors. "Originating domain" describes the domain name, or network address, of a visitor. "Originating country" describes the top-level domain, whether a country or other top-level domain (such as those represented by the codes ".edu" or ".com"), of a visitor. Buchanan and Lukaszewski also use "platform type" to distinguish between the computer types or operating systems of site visitors.

Current technologies and standards. Web server log files, which record information about all file request transactions received by the server software, store large quantities of data of potential use for tracking the browsing habits of site visitors. This section begins with a description of common log file formats and the data fields that such logs record. The next segment describes "cookies," a technology developed by Netscape that enables more accurate identification of individual site visitors. Finally, the third subsection briefly describes approaches to gathering data about visitors that yield even greater levels of refinement and certainty, but which require more cooperation and effort on the part of Web site users.

Most Web servers generate a log file that records all file requests received by the server software (Stein, 1997b). The server log file is effectively a database representing the server's activity, though the file is often initially

created as a simple text file. Each record, or line if in the form of a text file, in this database corresponds to a single file request, and generally represents a single "hit" on the server (Wiederspan & Shotton, 1996). The fields within each record store information such as the URL of the requested file, the date and time of the request, the number of bytes of data transmitted, and so on. Several server log analysis software packages have been developed to process the raw log files and generate summary reports describing server activity.

The most widely supported server log file format is the NCSA (National Center for Supercomputing Applications) Common Log File format (Stein, 1996), which includes fields for seven basic types of information. The NCSA Combined (or Extended) Log File format adds two more fields to the seven in the basic Common Log File format (Stout, 1997). As is the case with most Web-related "standards," there are several other log file formats in common use that are generated by various server software packages. Microsoft has developed several log file formats for use with its server packages (D. Brown, 1997), while MacHTTP and WebStar logs ("WebSTAR technical reference," 1995) are common products of Macintosh-based servers. Fortunately, the data fields most useful, from an educator's perspective, for monitoring site traffic are reasonably uniformly implemented across this assortment of log file formats.

The NCSA Common Log File Format (D. Brown, 1997; Stein, 1996; Stout, 1997), as well as other log formats, includes a field called "host," "remote host," or "hostname" that lists the network address (as a DNS hostname or an IP address) of the computer from which a file request was received. Server logs also include a field containing the date and time the request was received. A third important datum recorded in the log is the URL of the requested file. A "status" or "result" field notes whether the file request

resulted in a successful transmission, generated an error, was redirected, and so on. Another field records, in bytes, the size of the file transmitted.

The NCSA Combined (or Extended) Log File format (Stout, 1997), as well as some other log formats, adds "user agent" and "referrer" fields to log file records. The user agent field lists the Web browser software (and possibly the version number) the visitor is using, the operating system of the visitor's computer, and the general type of computer the visitor is using. This information is often used by Web site developers to create custom pages that have different appearances on different browser and computer platforms, or that take advantage of non-standard features supported by certain platforms. The referrer field records the URL of the visitor's Web "location" immediately prior to the current file request. Such information can inform Webmaster's of common points of entry, such as major directory or search sites, that lead visitors to their site.

The data recorded by server log files is adequate for inferring site hit counts, but does not supply, in many instances, sufficient information for accurate tracking of visitors or visits. For technical reasons that allow the Web to support many small file transfers from scattered servers, the HTTP protocol upon which the Web is based is a "stateless" system (Stein, 1997a). Each hit, or individual file request, appears to Web servers as separate, new network connection, so that relationships between individual requests cannot be explicitly determined. In essence, this statelessness makes it impossible to reliably track a series of file requests from a single computer. Although the hostname field in server log files can, in many instances, aid identification of visitor's computers, circumstances exist in which such information does not uniquely identify a single computer. To overcome this limitation, Netscape introduced, with the release of version 1.1 of its Navigator Web browser

software, a new technology called "cookies" ("Persistent client state HTTP cookies," 1997).

Cookie technology uses a visitor's Web browser to create a small database, in the form of a text file, on the visitor's hard drive (Clark, 1997). When a visitor first logs on to a particular Web site, the site can send a cookie to the visitor's browser, which writes information into the local cookie file. When the visitor returns to the same Web site, the server can use the browser to read information from the visitor's cookie file (Floyd, 1997). Cookies can be used to assign a unique identification number to site visitors' browsers, thus enabling more accurate tracking of the activity of individual users over time and across visits (Waring, 1997). Cookies also support features such as online "shopping carts" at commercial sites and the searching of complex databases that require the server to have a "memory" of the search's progress as a user refines it during a series of steps.

Other approaches can produce more detailed or more reliable information about users and their Web browsing habits, but typically require greater effort and cooperation on the part of visitors. Surveys can be filled out, submitted, and administered online. Some site developers require users to establish ID numbers and passwords, set up via identification forms, to gain access to their sites (Cooper, 1996; "New media companies," 1996). Such information, submitted at the beginning of a visit, aids sites in tracking repeat visitors from one visit to the next. Some groups, such as the company PC Meter, avoid "site-centric" approaches to monitoring visitor activity. Instead, PC Meter has installed tracking software in the households of 10,000 Web users who they claim are a demographically balanced sample of the population of United States PC owners ("First year of PC Meter," 1997). This approach enables definitive identification of visitors' computers and tracking across multiple

Web sites.

Current logging and analysis software. One approach to Web site visitor tracking involves use of a log analysis software package to process the Web server's log file. Such packages read through the log files, count and group individual entries based on criteria selected by the user, and produce textual or graphical summaries of site activity over a given time period. These summaries typically report the number of bytes of data transmitted, of interest for monitoring server load levels, or the count of hits on files, of interest for analyzing visitor traffic patterns. Most packages allow users to specify that only certain file types be included in hit count summaries, so that reports configured to count only Web pages (HTML files) can effectively convert hit counts into page view counts. Such page view count summaries form a basis for visitor traffic analyses enabled by server log processing software.

Common log analyses include summaries based on time and date, on the network addresses of visitors, and on the specific pages visited by users. Typical time-based tallies include hit counts by individual day (date), by hour of each day over the course of the summary period, or by days of the week over the summary period. Summaries based on visitors' network addresses may show hits for each separate address, or may cluster hits into groups representing entire network domains (such as those ending with ".edu"). Finally, reports can show total hit counts for all files within an entire Web site, or can indicate hits on individual pages within the site.

Titles of some of the commonly available server log analysis programs include AccessWatch, ServerStat, Site Stat, Statbot, WebReporter, WebTrends, Wusage, and wwwstat (Lee, 1996c, 1996e; Patten, 1997). Six of these programs are either freeware or are currently priced below $100. Many of the more sophisticated visitor tracking software packages, to be described next, cost

thousands to tens of thousands of dollars, and thus may be priced beyond the means of many educators and institutions. Log analyzers are generally fairly easy to learn how to use, and produce similar sorts of reports.

Many commercial sites use visitor tracking software that is more sophisticated and more expensive than basic log analysis packages. Such programs are usually tightly integrated with particular Web server software programs, often logging information about client-server transactions into advanced databases. They frequently employ cookies to identify and track individual users throughout and across visits. Some programs enable observation of site traffic levels within seconds of "real-time" server activity. Examples of this type of tracking software include Andromedia's ARIA and EveryWare's Bolero (Lee, 1996c, 1996d; Pearlstein, 1997; Seiter, 1997).

Some packages go beyond mere real-time tracking of visitor behaviors. Programs such as Accipiter's AdManager, NetGravity's AdServer, and WebThreads use data about visitors and visitors' actions to present users with pages tailored to their supposed interests (Cooper, 1996; Murphy, 1996). Such packages are primarily intended for advertising purposes, and are meant to present banner advertisements for certain products to those users most likely to purchase those products. Software such as this, if capable of supporting educational initiatives, might be used to present learners with Web pages appropriate to their interests, levels of understanding, and prior online history browsing a given topic.

Finally, besides covertly tracking visitor behaviors, software can also support direct questioning of site visitors. Power Knowledge Software's PowerTab package helps site managers formulate and implement online surveys (Cortinas, 1997b). PowerTab tallies survey results, and then uses an expert systems approach to help it's users select and run tests of statistical

significance on those results.

Factors that foil tracking and emerging and future technologies. Several aspects of the way the World Wide Web and Web browsers operate have the potential to render invalid assumptions about the implications of data collected concerning visitor traffic levels. Also, the Web, the way people use the Web, and the population of people who use the Web are constantly evolving, thus complicating the task of translating the significance of findings about Web visitor patterns into recommendations for further studies and into advice for Web site designers and developers. Likewise, the Web, though a great driver towards popularizing use of the Internet, is not the end of the road for Internet-based technologies with likely widespread appeal and influence. This section describes some of the factors that can confound measurement of Web visitor behaviors and traffic levels, and briefly explores some of the emerging technologies that may partially or wholly supplant the Web, decreasing the relevance of monitoring Web visitor traffic.

Site developers who use hits as a measure of site activity can be seriously misled in their estimations of traffic levels by visitors who turn of the "auto-load images" feature of their browsers (Shaffer, 1996). For instance, a page containing ten images would account for eleven hits when requested by a user who downloaded the graphics, but only a single hit when requested by a user with a low-speed connection who was viewing the site in "text-only" mode. Similarly, caching can create a mismatch between the number of files, including HTML documents, viewed by a visitor and the number of requests for documents that a Web server receives (Shaffer). If a user had recently visited a page, the visitor's browser could retrieve files from the browser's cache, stored locally on the visitor's hard drive, instead of sending a request to the Web server for those files. Such cached page views may not create new

records of hits on the server's log file. Likewise, some service providers and network gateways cache frequently requested documents, thus intercepting requests for files before they reach the server and preventing their appearance in the server's log.

Savvy Web users concerned about privacy issues can use redirection technologies to hide their identities from the owners of sites that they visit. By visiting a site such as "the Anonymizer" (www.anonymizer.com) at the beginning of a browsing session, users can have all of their file requests redirected through a separate Internet address or proxy server (Schwartz, 1997; "the Anonymizer," 1997). The proxy server's address appears in the logs of Web sites the user visits, shielding the visitor's identity from the site's owners.

Recently developed tools which download groups of Web pages, or even entire sites, can also confound tracking efforts. Products such as WebWhacker and Web Buddy (Duncan, 1997) can be set to automatically download pages and sites while their users are otherwise occupied. Users can view sites later, from locally stored copies, without waiting for slow page downloads caused by busy networks or low bandwidth connections. This creates two problems for site visitor trackers: some pages may be retrieved but never viewed, while other pages may be viewed numerous times after only a single request is recorded on the server's log. Use of such "offline browsing" technologies is likely to increase, since the most recent versions of both Netscape's and Microsoft's Web browsers have these features built in to the software (Andrews, 1997a, 1997d).

Some emerging technologies and trends are beginning to influence the way people use the World Wide Web, while other new technologies may divert Internet users away from the Web. Such trends and technologies seem

likely to alter educational uses of the Internet, and to thereby influence the importance of and methods for tracking use of educationally oriented Internet resources. Web users have gradually moved away from "surfing" or "browsing" the Web, following hypertext links from site to site, towards searching for specific information with the aid of directory or search sites (Lemay, 1997). This trend influences the way visitors find their way to sites, and can increase the number and variety of site entry points. Many Web sites are no longer collections of pre-built pages, but instead consist of dynamically created documents assembled from chunks of information drawn from databases in response to user queries (Andrews, 1997b; Manes, 1997). Page view counts lose their relevance when pages are more of a continuum of collections of information instead of discrete files. Streaming audio and video technologies are likewise calling into question the appropriateness of using discrete files as a measure of the quantity of content quantity viewed by visitors.

Although Netscape's cookie technology has aided site managers in their attempts to identify unique visitors, it has also raised the ire of many Web users who view its use as an invasion of privacy (Clark, 1997; Floyd, 1997). Some software developers have begun to distribute products that prevent browsers from setting cookies (Cortinas, 1997a), thus negating the tracking advantages gained by sites employing cookies and introducing a bias into the visitor statistics generated by such sites. An industry consortium has begun work on a standard that could supersede cookies (Lee, 1997b), which would give site managers similar tracking capabilities as cookies but which might alleviate fears of privacy advocates by requiring explicit permission from users before releasing identity information to Web sites.

Other emerging technologies may capture some of the Web's "market

share" of visitor interest, bringing new challenges to administrators interested in visitor tracking. Virtual Reality Modeling Language (VRML) certainly has much of the sexy, graphical appeal that has played a role in the popularity of the World Wide Web. "Push" technologies that deliver information to users, instead of waiting for users to come seeking information, seem a blend of traditional broadcast media with the Internet-based Web (E. Brown, 1997). Since push technology requires users to subscribe to services, it is easier for site administrators to uniquely identify and track such subscribers. However, since users only partially control which content is delivered to them, push developers share the disadvantages of traditional broadcast media vendors of not being able to distinguish between content users are actually interested in and content that is delivered but ignored.

Directories, Search Engines and Robots

The World Wide Web is vast and is constantly in flux. Web users need tools to help them locate the information they seek. Web users often turn to directory and search sites to help them find relevant resources (Lemay, 1997). Most search sites, in turn, rely on automated systems called "Webcrawler robots" or "spiders" to create and update their databases of Web sites. Details of how search and directory sites list and display data, and how robots feed data to search sites, have a large impact on trends in visitor traffic to Web sites. Visitor "head counts" and which pages serve as "entry portals" to sites are especially influenced by the way sites appear to visitors on directories and search sites. This section describes robots, search engines and sites, and directory sites.

Directory sites. Directory sites list Web sites based on subject categories devised by people, as opposed to automatic techniques that sort sites into categories based on keywords embedded within documents. Since sites are selected by and placed in categories by humans, development of such sites is a labor intensive undertaking. Some directories are general listings spanning a broad range of categories. Yahoo! (www.yahoo.com), which logged a billion page views in the third quarter of 1996 (Andrews, 1997c), is probably the most widely known and used general directory site. Some directories cover only specific topic areas. For example, science education directories, biology or microbiology directories, and microscopy directories might all list a site such as the Microbe Zoo. Some directories employ people to seek out sites which are appropriate to add to their listings. Many directories allow visitors to suggest sites for inclusion in their listings.

The submission procedure used by Yahoo! is typical. Visitors choose a Yahoo! category which seems appropriate for the site they wish to add ("Yahoo! - How to Suggest Your Site," 1997). An online submission form requests the name of the site, it's URL, and a brief (25 words or fewer) description of the site. Optionally, visitors can also suggest creation of a new Yahoo! category if they think the submitted site does not fit into the existing categories, or they can suggest multiple listing categories by which the site should be cross-referenced. The submission form also prompts the user to enter contact information (name and E-mail address), which purportedly is used for verification if the submitter later wishes to change listing information. The form also requests information about the geographical location of the site, since some of the Yahoo! categories are only regionally relevant. Also, the form can accept information about dates, in case a site is only accessible or relevant for a limited time frame. Once submitted by a

visitor, the site information is examined, processed, and placed in appropriate categories by a Yahoo! "Surfer," or employee ("Advanced Help on Suggesting Sites to Yahoo!," 1997).

Directory sites tend to be less comprehensive in their listings, due to the need to have people enter and update information, than sites which use automation to build and maintain their catalogs (Hamit, 1996). However, directories are sometimes more helpful because they are more concise than automated search sites, and may place information in categories that are more sensible to users than search sites that use algorithms to classify sites based on keywords.

Search sites and search engines. Search sites use Web pages as an interface to databases that list other Web sites. Visitors type in keywords relevant to the information they are seeking, click a "search" button, and then view listings of pages containing their search terms retrieved from the search site's database. Common search sites include AltaVista (www.altavista.digital.com), Excite (www.excite.com), HotBot (www.hotbot.com), Infoseek (www.infoseek.com), Lycos (www.lycos.com), Magellan (www.mckinley.com), and WebCrawler (webcrawler.com). Most search sites have also gradually adopted a directory-style interface to their listings, thus giving visitors the option to access information in whichever mode is most appropriate for their task. Search sites generally use some type of Webcrawler robot agent software, described in the following subsection, to populate their databases with Web page listings.

Web users have gradually shifted their exploration techniques towards searching for specific information, as opposed to the earlier dominant mode of browsing by following links from site to site (Andrews, 1997b). Searches begin with visitors typing in words or phrases likely to occur on pages that

have the information they seek. The search engine examines it's database of Web sites and pages, locates entries containing the relevant terms, applies some sort of algorithm to rank the likely relevance of the results to the visitor's search goals, and then reports those results in order of likely relevance. Early search sites indexed only the titles of Web pages, or the titles plus the first several words at the top of pages. As faster, more sophisticated computers and search engines have evolved, most search sites have progressed to indexing page titles plus the full text of Web pages (Hamit, 1996).

Some search sites index additional information that site developers can embed within pages to aid visitors' searches. Keyword and description <META> tags, introduced with the HTML 2.0 specification (Lemay, 1997), allow Web page developers to embed information about pages within HTML documents in a way that supports cataloging by search sites but that remains invisible to site visitors viewing pages. The keyword tag allows site developers to include words that a visitor might search for, but that are not explicitly included in the text of a page. For example, a site that includes "bacterium" in the text of a page might help search sites index that page by including a keyword tag containing terms such as "microbiology," "microbe," "microbes," and "bacteria." Note that plural forms of words that appear on pages (or the singular form if the plural appears in text), especially those that change the form of the term as opposed to adding an "s" to the end of the word, are good candidates for <META> keyword lists. Synonyms of important terms are also good choices for keyword lists. Most search engines that recognize <META> tags assign their contents a relatively high score when creating relevancy rankings, so the tags can also be used to emphasize important terms that also appear in text.

HTML supports labeling of image files via the "ALT" tag. When users have the "auto-load images" feature turned off in their browser, typically to speed page loading over slow connections, the text of the ALT tag appears in place of an image. The ALT tag text usually describes the image, helping visitors decide whether to choose to download that file. Some search sites catalog ALT tag text, thus providing an indexing method for pictures that are not explicitly described in the text of a page ("Submit It!," 1997). Other media elements, such as sound files and video clips, can also have descriptive ALT tags attached to their HTML code.

Some search sites also index comments included within the HTML coding of a Web page file. Though invisible to users, comments can include words that the page developer wishes search sites to incorporate into the page's record in their databases. This approach can fill a role similar to that played by <META> keyword tags in the cases of search sites which do not recognize the <META> tags.

Most search sites include page titles, file URL's, and brief page descriptions or summaries in their listings of Web pages that result from visitors' searches. Many search sites also report a relevancy rating, usually expressed as a percent, for each "found" page. The rating represents the degree to which the listed page matches the user's search criteria, according to the rating algorithm the search engine employs. Some results listings include the size of each found file (in bytes or kilobytes), and the most recent modification date of each page or the date on which the page's record was last updated in the search site's database.

The page titles in search result listings are drawn from the HTML document's title tag (not from the first line of text displayed on the page). The same title appears in users' bookmark lists when they bookmark sites for later

reference. Sites that recognize <META> tags use the description tag text as the page summary description reported in search results lists. Most sites that do not recognize <META> tags place the first several words on a Web page into the page description listing. Typically, the first 20 to 25 words, or the first 200 characters, are listed in such descriptions. In many cases, such "descriptions" are not very good representations of page contents, since they may include navigational listings or similar functional, but not descriptive, elements of text from a page. Since some sites include HTML comments in the text that they index, placing a descriptive comment on a page before the actual visible page text can improve search site listings.

Several factors influence the listing order or priority, or relevance ratings, of pages returned by searches. Most sites place the top priority on matches between users' search terms and words which appear in page titles. Sites that recognize <META> tags place high relevance priority on terms in <META> keyword lists. Words that appear early in the text of a page, as opposed to those appearing further down the page, are often assigned greater relevance. Matched search terms which appear numerous times on a page, as compared to words mentioned once or twice, usually generate higher rankings. Some site developers, in recognition of this fact, place numerous repetitions of key terms on some pages for the sole purpose of gaining greater visibility in search results listings. This practice, which is sometimes also used with <META> keyword tags, is called "spamming." Some search sites have altered their ranking algorithms to detect spamming and to assign lower ratings to, or to not report at all, pages that seem to be using spamming techniques (Lemay, 1997; "Submit It!," 1997).

Since search sites receive information about users' interests, in the form of search terms entered by visitors, they can also serve as visitor behavior and

interest monitoring tools (Andrews, 1997c). Some search sites use cookies to allow visitors to select preferences that are maintained from one session to another, and to customize results and relevancy rankings based on the cookie-based "memories" of a visitor's previous requests. Some server log formats store information about search terms ("WebSTAR technical reference," 1995), enabling analysis of those records based on a type of direct input from site visitors.

Webcrawler robots. Search sites rely on automated software agents, called "Webcrawler robots" or "spiders," to populate their databases with Web site and page listings. Search site administrators supply robots with some initial starting point, or points, from which to begin their site cataloging processes. These starting points are often home pages of sites submitted by visitors to the search sites in a manner similar to the submission procedures used by directory sites such as "Yahoo!". The robot first logs all of the text in the starting point's page into the search site's database, and then proceeds to pages that the seed page links to (Burner, 1997). This process is followed recursively, so that the robot gradually explores and records all pages linked to from the seed page, all pages linked to from the pages the starting point links to, and so forth (Koster, 1995). By using a number of starting point seed pages, and relying on the extensively interlinked nature of the Web, robots are able to eventually traverse and record the majority of publicly accessible Web pages. Robots generally are unable, however, to catalog pages hidden behind corporate firewalls or protected by passwords.

Minimally, robots record the titles and URLs of the pages they "visit." Many robots also read, and log into their parent search site's database, the full text of visited pages. Some robots also record data in <META> keyword and description tags, in the HTML ALT tags which describe graphics images

embedded within pages, and in HTML comment lines in documents. Tracking down and recording such vast quantities of data takes time, so that robot "sweeps" covering the entire Web tend to take several weeks to complete (Hamit, 1996). Robot "visits" to Web sites are, at first glance, indistinguishable from accesses by human visitors, and are recorded similarly in server log files.

For various reasons, some site administrators do not want their pages listed by search sites and therefore do not want cataloging robots to visit certain pages or entire sites (Duncan, 1997). To support this desire, many robot developers have agreed to create robots that comply with a voluntary standard called the "Robot Exclusion Protocol" (Koster, 1997). Sites can inform protocol-compliant robots to avoid indexing their pages by placing an appropriately formatted text file, named "robots.txt," in their root directory. Control of some compliant robots can also be exerted at the individual Web page level, by using <META> robots tags which tell robots whether or not to index the page on which the tag appears, whether or not to follow links from that page in their indexing search, and combinations of the two option choices. Compliance with the Robot Exclusion Protocol is entirely voluntary, so there is no guarantee that all robot implementations will honor it.

Sites that dynamically generate Web pages on the fly, such as those assembled from information drawn from databases in response to user input, cannot readily be cataloged by robots (Carl, 1995). Sites based on database and dynamic page creation technologies present virtually infinite numbers of possible pages, thus foiling search sites which rely on robots. Ironically, search sites themselves are usually implemented in just such a manner. Sites based on dynamically created pages are a growing phenomenon, sometimes referred to as the "invisible Web" (Andrews, 1997b; Manes, 1997), and present

an increasing challenge to search sites that use robots to create their catalogs.

Growth of the Internet and the World Wide Web over Time

The Internet and the World Wide Web have both grown at a phenomenal pace. Studies of visitation trends to specific Web sites occur in the context of this rapid growth of the overall Web. Precise, consensus definitions of "Web site" and "connected to the Internet" are elusive, making it difficult to accurately measure Web and Internet growth rates. Likewise, measures of the number of computers connected to networks do not necessarily indicate the number of people who use those networks, or the amount of time those people spend online. However, some general trends in network growth rates are measurable, and provide a rough sense of how quickly such growth is proceeding. This section describes and presents measures of the rate of growth of the World Wide Web and of the Internet as a whole.

Network Wizards, a California-based company, conducts a survey every six months to estimate the "size" of the Internet (Lottor, 1997). Their data about the estimated number of "hosts" connected to the Internet are widely cited as evidence of the Internet's rapid growth in recent years. Host counts, which roughly represent the number of computers "connected to" the Internet, serve as an approximate indicator of the number of people who have access to the Internet. Table 2 lists the number of hosts on the Internet, as reported by the Network Wizards' survey, during the past seven years. Disclaimers presented with the survey note that the Network Wizards "consider the numbers presented in the domain survey to be fairly good estimates of the *minimum* size of the Internet." If we consider host count a fair representation of the "size" of the Internet, we find that the Internet is

Table 2 - Number of Internet Hosts over Time

Date	Host count (thousands)
January 1991	376
July 1991	535
January 1992	727
July 1992	992
January 1993	1,313
July 1993	1,776
January 1994	2,217
July 1994	3,212
January 1995	4,852
July 1995	6,642
January 1996	9,472
July 1996	12,881
January 1997	16,146
July 1997	19,540

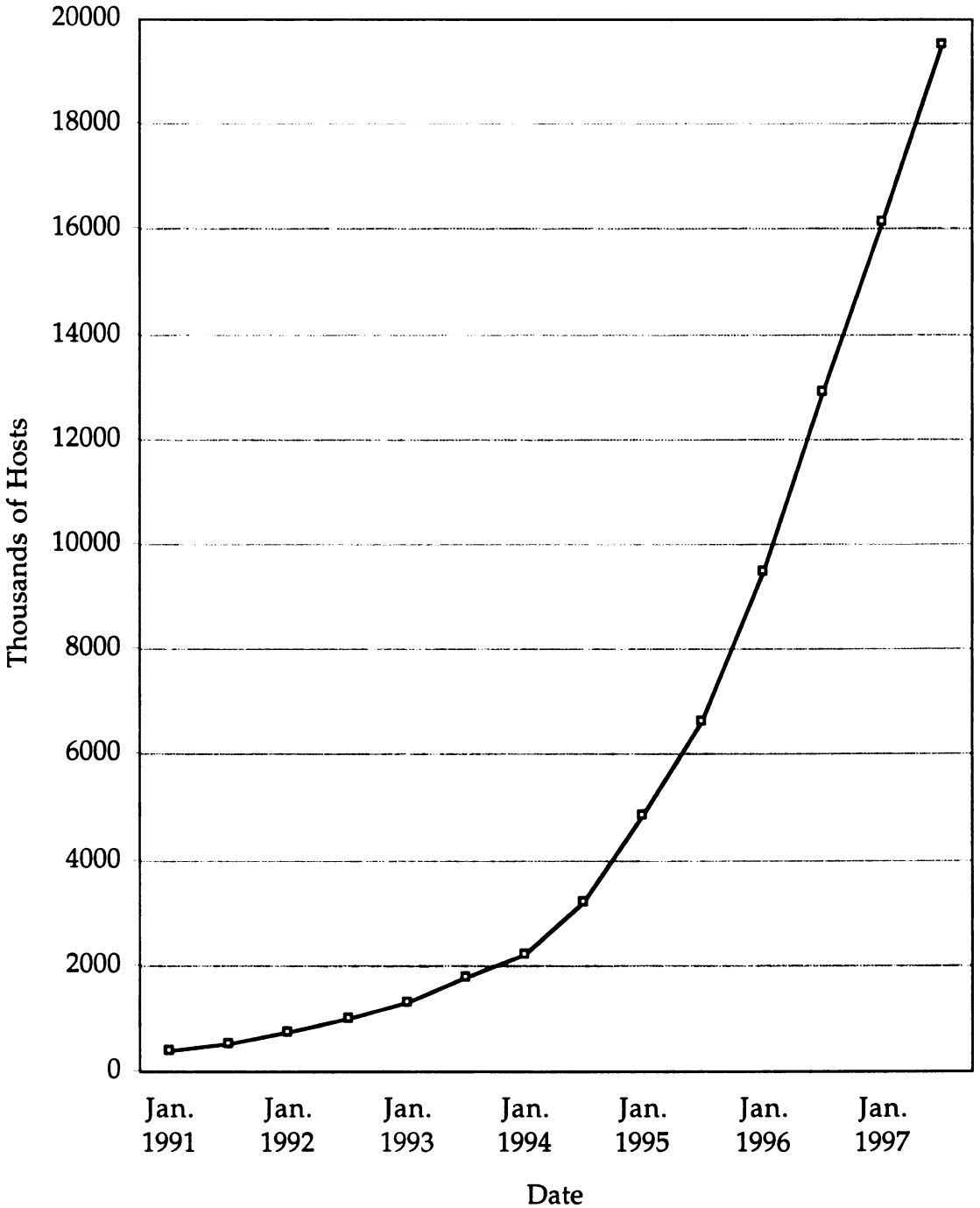

Figure 1 - Number of Internet Hosts over Time

typically doubling in size ever 12 to 18 months, as it has been doing since the early 1990's. Figure 1 shows Internet growth data in graphical form.

The World Wide Web portion of the Internet has undergone especially rapid growth during the last few years. Table 3 lists Gray's (1997) data about the number of Web sites available online, at six-month intervals, between 1993 and 1997. Definitions of what constitutes a Web "site" vary. Gray defines a site as "All documents with urls beginning with a unique hostname. That is, http://www.mit.edu/people/mkgray/ and http://www.mit.edu/madlibs are part of the same site, but a document http://web.mit.edu/ is a separate site." By this measure, the Web is more than doubling in size every six (or fewer) months, and has been doing so since its inception. Figure 2 is a graphical representation of the growth of the Web.

Gray (1997) also calculated the ratio of hosts to Web servers. This proportion has been declining over time, since the Web has been growing more rapidly than the Internet as a whole. Table 4 lists the ratio of hosts to Web servers, at roughly six-month intervals, since mid-1993. These data were mostly provided by Gray; the value for January, 1997, is a combination of Gray's data with values from the latest Domain Survey by the Network Wizards (Lottor, 1997). Based on the decline of this "hosts to Web sites" ratio, the "competition" between Web sites for viewers' limited time and attention may be increasing as the Web's growth rate outpaces the growth of Internet. Figure 3 shows the decline over time of this ratio. Only values from December 1993 onward are displayed on this graph, since inclusion of the June 1993 value would necessitate a logarithmic scale for the vertical axis.

Table 3 - Number of World Wide Web Sites over Time

Date	Number of Web sites	
June 1993	130	
December 1993	623	
June 1994	2,738	
December 1994	10,022	
June 1995	23,500	
January 1996	100,000	
June 1996	230,000	(estimated)
January 1997	650,000	(estimated)

Data from Matthew Gray of the Massachusetts Institute of Technology

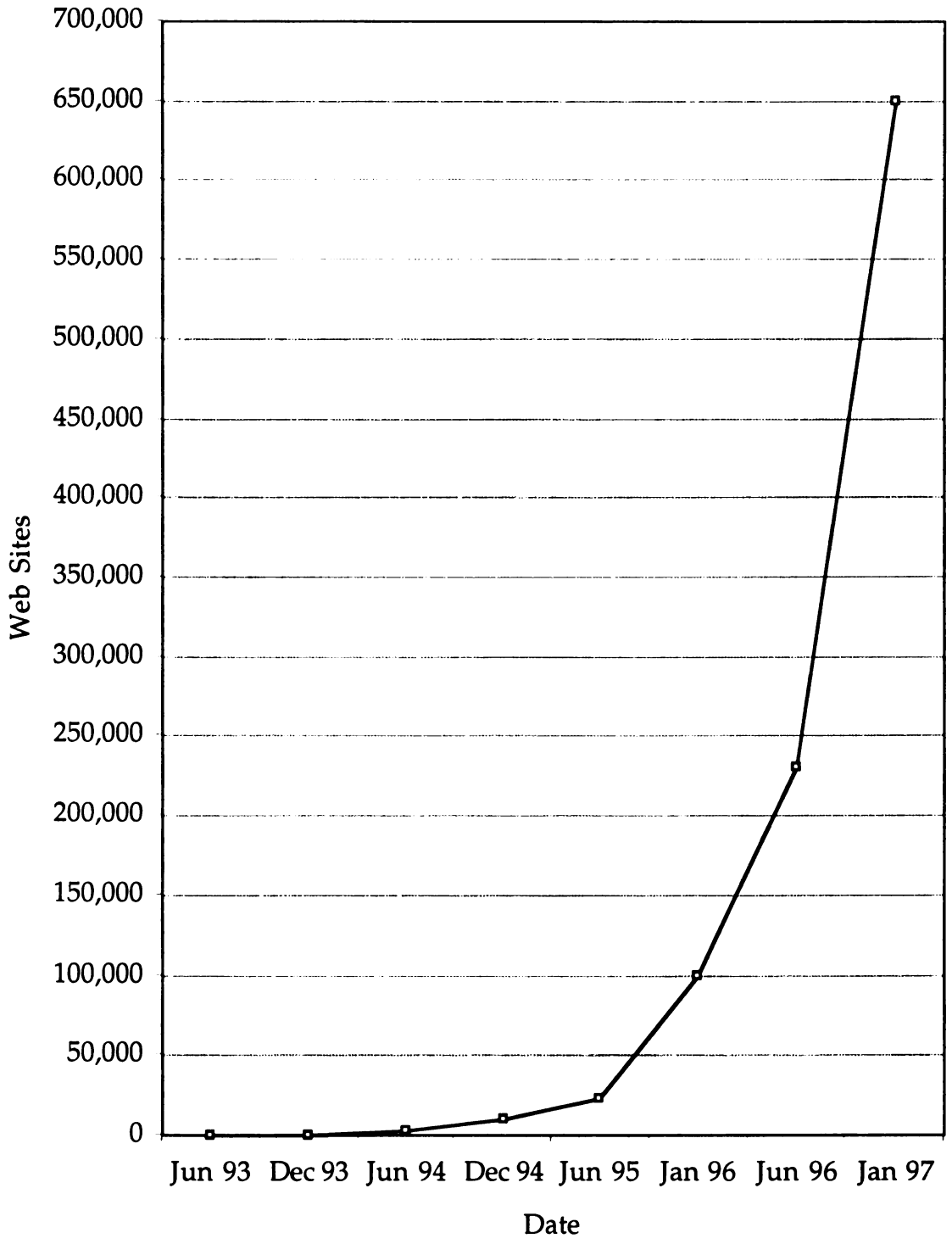

Figure 2 - Number of World-Wide Web Sites over Time

Table 4 - Hosts per Web Site ratio over Time

Date	Host per Web site
June 1993	13,000
December 1993	3,475
June 1994	1,095
December 1994	451
June 1995	270
January 1996	94
June 1996	41
January 1997	25

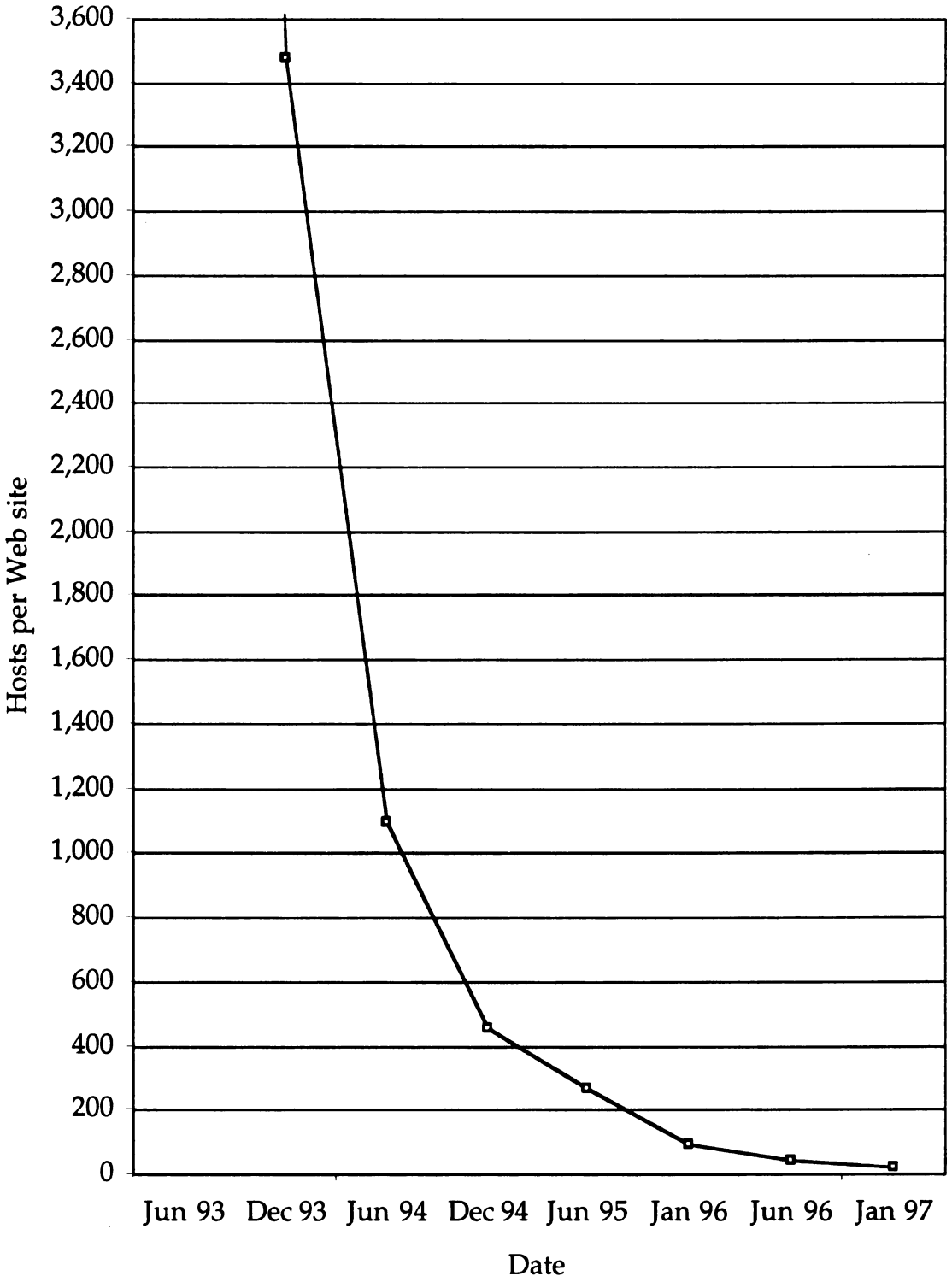

Figure 3 - Hosts per Web Site ratio over Time

Smith (1997) estimated that by May of 1997 there were about 80 million HTML pages and approximately one million Web sites on the publicly accessible Web. He noted that the increase in dynamically generated pages is rapidly making total Web page counts a meaningless figure.

Chapter Three

METHODS

This chapter opens with a description of the intended audience of this report, the applicability of the methods used in this study to formative and summative evaluations, the exploratory as opposed to definitive nature of this investigation, and the illustrative role played by the specific Web site studied throughout this research project. The first major section of this chapter states the major research questions which serve as the focus of this study. The next section describes the research subjects whose behaviors this study records and in some instances attempts to explain. The final two segments of this chapter explain the data collection and data analysis procedures used in this research project.

The primary target audience of this report consists of designers and developers of Web sites intended for educational use. This document endeavors to aid such developers by providing them with insights into methods they can apply to the sites they are developing to better understand how those sites are being used, and by whom. This report describes several analyses that site developers can apply to their sites, helps developers understand the amounts and types of efforts required to apply each analysis, and informs developers about the sorts of information that each of those analyses reveals.

The analysis techniques investigated via this study are mainly intended to be used as formative evaluation tools. My hope is that developers will apply these analyses while the site is still being created, and will use the results of the analyses to make better informed judgments about how to proceed with

site development. Although my main emphasis throughout this report is on use of these techniques for formative evaluation, they could certainly be applied by site evaluators to a summative evaluation as well. Given the ongoing "This site under construction!" nature of many Web sites, the distinction between formative and summative evaluation is often quite fuzzy.

The World Wide Web is still a relatively new phenomenon, and Web visitor studies are in their infancy. This investigation is intended to be exploratory in nature, and is thus not designed to reveal definitive prescriptions for how sites should be developed. The target audiences and intended purposes of sites, design features of Web pages, and technologies used within Web sites vary widely, so "one size fits all" prescriptions for successful site design techniques or site evaluation methodologies are inappropriate goals. This study seeks to provide site designers, developers, and evaluators a grab bag of site evaluation tools and techniques which they can consider applying to the specific needs appropriate to their sites and evaluation goals. Many of the avenues of investigation described in this study will reveal dead-ends, which later investigators will likely avoid. Other techniques described here may merit further, more detailed scrutiny by later researchers. The goal of this study is to conduct a broad exploratory overview investigation, and to thus begin to separate the wheat from the chaff in the realm of Web visitor study techniques. The results of this investigation are not intended to be the final word in site analysis techniques, but rather to give site evaluators a bit of a head start into the investigations of their own sites.

The analysis techniques I have developed in the course of my dissertation research were applied to the DLC-ME Web site. This study is an investigation of the visitor studies techniques as applied generally to Web sites; it's focus is

not on the specific findings about the use of the DLC-ME site itself. The application of these techniques to the DLC-ME site is intended merely to provide a concrete example to use as a reference. The DLC-ME is not intended to play the role of an especially typical or representative example of a site, but is simply a convenient case with which I am familiar and for which I have ready access to an extensive amount of server log data. My hope is that the techniques described in this report are widely applicable to many Web sites. Determination of whether such generalization is valid, however, would require application of these techniques to data from other sites. That work is beyond the scope of this study.

Research Questions

The information recorded in the log file of a Web server serves as the source of data upon which this study is based. The research questions described in this section are framed in the context of that data source; other data sources would likely lead to different results in terms of the answers to the research questions which follow. Data from server log files has limitations in how much it unambiguously tells an evaluator about the identities and behaviors of Web site visitors. However, such data is readily available from most Web-hosting computers, can be automatically collected for large numbers of visitors, can be processed via computer in several potentially useful forms by inexpensive software, and is generally significantly less labor intensive to collect and process for large numbers of subjects than several other types of data. Such "other types of data" include online surveys, installation of software which records user actions onto the user's computer, and technologies such as cookies and user ID's which can more readily identify specific visitors. Although such techniques can enable

closer scrutiny of visitors, the additional information generally comes at the cost of greater intrusion upon research subjects; greater burdens of time, money, or skills requirements upon the researchers; or restrictions on the number of subjects who can feasibly be studied.

The four major research question addressed, in the context of Web server log data, by this study are:

1. What types of information relevant to educators is it possible to deduce about visitors and visitor behaviors, and with what degree of certainty?
2. What sorts of skills or tools, and what amounts of labor investments, are required to obtain those various types of information and degrees of certainty about such findings?
3. Is it possible to accurately measure a quantity such as "time on page" or "time on site," which could be tested for its correlation with learning outcomes in a fashion similar to the use by some researchers of quantities such as "time on task"?
4. What sorts of data should educational Web site developers collect and in what ways should they process those data to efficiently gain useful insights into how their own sights are being used?

The following paragraphs describe, in further detail, the issues associated with each of these major questions which are addressed by this report.

Two types of data sought via this study are information about the identity of visitors and information about the behaviors of visitors. What do the network addresses of visitors tell us about who they are in terms of the organizations they are affiliated with? Where are they coming from, in terms of both geographical and cyberspace locations? Where do they go upon arrival at the site being studied, in terms of pages and site sections visited? When do users visit the site, in terms of calendar dates, days of the week, and times of

day? What sorts of paths do visitors follow through the landscape of hypermedia-based information presented by a site? To what extent must evaluators make assumptions in interpreting data about visitors and behaviors, and how certain or tenuous are such assumptions?

This study is primarily concerned with formative evaluation of site usage patterns. Since formative evaluation is just part of the total development effort required to produce a site, analysis techniques must be evaluated in light of the resources available for their execution. What software tools are required to conduct or support various analyses? How much do such tools cost, how much effort will be required of the users of such tools to learn their use, and will those tools become obsolete over time and thus require analysts to acquire and learn how to use replacements? What skill levels in terms of understanding of analytical techniques or technical software sophistication will be required of the personnel who will conduct analyses? How much of the project personnel's time will be required to conduct various analyses? Do more labor- or skill-intensive analyses yield results with greater credibility, and how much and what type of certainty is gained by expenditures of how much additional resources? What steps must site developers take to prepare a site for evaluation, in terms of keeping track of the site's state at various times and in terms of ethically alerting visitors to their status as research subjects?

Many educational researchers have noted that measures of the amount of time students spend actively engaged in learning activities, typically referred to as "time on task," is often correlated with various learning outcomes. Some researchers have used measures of time on task as proxy indicators of student learning when direct measurement of learning outcomes was not feasible. Is it possible to reliably measure a similar metric describing the amount of time visitors to a Web site spend on specific pages, on specific

sections of a site, or on their entire visit to a site? If such quantities are measurable, they could be used to test whether a correlation between them and learning outcomes exists. If such a correlation were found, such "time on page" or "time on site" measures could conceivably be used as proxy measures of learning. In a similar vein, museum visitor studies researchers often use the time spent viewing an exhibit as a learning indicator. Since page view counts per Web site visit may be even more practical to measure than visit durations in terms of time, it would also be interesting to determine whether a correlation between page views per visit and visit duration in terms of elapsed time exists.

Finally, a major pragmatic goal of this study is to provide Web site developers with advice about conducting research about their own sites. I will attempt to postulate some guidelines, and to describe a set of tools and techniques, developers might use to study their sites. The site evaluation goals and other circumstances of particular evaluation efforts for diverse Web sites vary too much to expect that a single prescription for site analysis is viable. Instead, I will attempt to provide site developers insights into the types of inquiries they could conduct, the sorts of information that those inquiries would likely yield, and the resource costs that they would have to invest to conduct such studies. Hopefully, such guidance will aid developers in formulating evaluation plans appropriate to their needs and levels of resource availability, and will encourage developers who might not otherwise consider conducting such evaluations to apply at least some analyses beyond their own intuitions to the study of their sites.

Subjects

The research subjects "selected" for inclusion in this study were simply all "visitors" to the DLC-ME during the course of the study's time frame. This approach reflects two goals of this study: the likely desire of other site developers conducting a formative evaluation about their site to learn about as many of the actual visitors to their site as possible, and the automation advantages in terms of both data collection and data processing inherent to use of server log files as a data source which enable analysis of data about large numbers of visitors. Instead of selecting representative samples of subjects and extrapolating their behavior patterns onto larger populations, most of this study focuses on studying the entire user population.

The Microbe Zoo portion of the DLC-ME Web site is a form of online, virtual zoo. Museum visitor studies researchers consider institutions such as gardens, aquaria, and zoos sufficiently similar to museums to fall with in the realm of museum studies research. A common approach to the study of museum visitors is to track some sample of the people who come in through the doors. Resource limitations usually require museum researchers to select some subset of visitors to study. The subjects studied in the project described by this report were essentially all of the "visitors" who "came in through the doors" of the DLC-ME Web site.

Museum studies researchers typically limit the scope of their data collection efforts by selecting some subset of their total visitor population to track. One means to limit data collection efforts is to randomly select some visitors to study; selection criteria such as "track every fifth visitor through the doors" are common. Another approach to limiting the scope of data collection efforts is to track all visitors, but for a very limited time period. A

similar technique is used in a portion of this study. Although data collection for this study spanned most of the site's history, one segment of the analysis focuses on detailed analysis of data from a specific week. That week was chosen because it seemed typical or representative in several ways, and thus results of that week's analysis could reasonably be expected to be similar to those which might be obtained from analyses of data from many other weeks. Although this selection of a restricted population to study is similar to the techniques used by museum researchers previously mentioned, the approach is somewhat different in that the paring down of the subject population was done at the data analysis, not the data collection, stage. Should detailed study of other specific time periods become a priority later, the data to support such study would be available.

The definition of a "visitor," as used in this study, requires some clarification. Although "visitors" frequently correspond to individual human beings, the data collected via server log files is actually about individual network addresses of computers. There are enough cases in which such network addresses do not correspond to unique individuals to require researchers to keep the nature of the data source used here firmly in mind. Computers are often in libraries or labs where they are shared by several or many people. Network addresses used by Internet Service Providers are assigned dynamically when users dial in; thus a single address could be assigned to many different users at different times, and a single user could be assigned different addresses when she or he logs on at different times. Some users have access to different computers, so a single person might appear as multiple addresses; for instance, when browsing the Web from home or from work. Finally, some "visitors" to Web sites are not even people. Webcrawler robots are actually pieces of software, and can generate large numbers of hits

on Web sites in the process of building up catalogs for search sites. These complications in identifying visitors who are actually unique human beings must be kept in mind during any analysis of Web site traffic which relies on network addresses of computers as an identifier of "visitors."

A significant goal of this study is to discover the limitations to Web visitor studies based on the server log file as a source of data. Automated collection of data and the ease with which some types of data analyses can be conducted for large numbers of subjects using the processing power of computers are major strengths of this approach. The choice of this approach for this study does not, however, belittle the need for other approaches that demand different methods for selecting research subjects. For example, the study of people who are aware of a site's existence, but choose not to "visit" it, would likely provide interesting insights. Identifying such people in order to study the reasons for their choices would be a major undertaking; one that is beyond the scope of this study. The emphasis of this study is on efficiency in data collection and on preliminary exploration of the potential benefits and limitations of using the server log file to study large numbers of visitors to a site.

Data Collection

The sources of data used in this study were the Web server log files from the WebStar Web server software running on a Macintosh computer in the Communication Technology Laboratory at Michigan State University. Those files are routinely generated and saved as a normal part of operation and monitoring of the performance of the Web server. Each hit on the Web server generates a one-line record in the log file. Each record contains eight

fields of data items separated by tabs: date, time, result, hostname, URL, bytes sent, referrer and transfer time. The records are saved as text files. The server administrator periodically archives record files as the files grow large, replacing the active file with a new, empty file to fill with data. Files are typically replaced when they grow to five to ten megabytes in size. Early in the site's history, files of that size typically included several weeks worth of visitor data. More recently, as traffic levels have increased, log files of that size have spanned only a few days' time.

The server log files required a small amount of formatting before they were in a form suitable for analysis. Each time a new file was started, the server placed a header line at the top of the file. Replacement of a log file required the server administrator to temporarily suspend logging; sometimes this caused a record to be broken mid-line. To construct an unbroken log file spanning a given time period, typically one week in this study, two or more raw log files spanning shorter periods might have to be merged, especially later in the site's history when log files covered shorter times. "Clean" log files have all opening header lines and all broken trailing lines removed. Note that some data was lost during switchover from an old to a fresh, new log file. This lost data typically represented just a minute or two of server activity.

The Comm Tech Lab's Web server host's several Web sites associated with the lab's programs. Hits on the server are recorded chronologically on the server log file, so that accesses to files on all of the Web site's hosted by the lab's server are thoroughly intermingled. In order to study hits on a single site, the records for that site must be separated from the others. The URL field in each record contains the information necessary to make that separation. In this study, the log file records were processed using a database and a server log

analysis program. Each of those pieces of software is capable of separating files, based on the data in the URL field, in order to facilitate analysis of a single site hosted by a server that is home to other sites.

To alert visitors to the DLC-ME site that their actions were being monitored, I placed an announcement describing that fact at the bottom of each page on the site. The announcement reads:

> Please note! Use of this World-Wide Web site is being monitored for educational research purposes. Data concerning use of this site by visitors such as yourself may be included in published research reports. If you are not comfortable with this, you may wish to end your visit to this Web site.

This notification was employed at the suggestion of Michigan State University's human subjects research committee. It is not clear that such a notification is strictly required. It seemed to me, in my discussion of this issue with the University's representatives, that this was a new issue for them at the time I was inquiring about it. It seemed to me prudent to follow their suggestions, since it did not seem to create any great hardship or negatively impact the site's design in a significant way. It also seemed fair to me; as a user of Web sites, I would like to be informed that I was "being watched," so I could decide whether that was of concern to me. Web sites may be sufficiently "public" forums that such notifications are not, legally, required. I have not received a single complaint from a user of the site regarding their status as a research subject. I did receive a request for clarification of precisely what "being monitored" meant in this case from a visitor who also was the administrator of a site considering placing a link to the DLC-ME site on her site. She was primarily looking out for the interests of the users of her site; she found the explanation of how I was planning on using my data completely unobjectionable.

Data Analysis

Data analysis procedures used in this study can be grouped into four categories. First, weekly page view and visitor counts for the entire DLC-ME site were determined and graphed throughout the history of the site. Based on trends in traffic levels, a "typical" week was selected for closer scrutiny. Page view counts during the selected case study week, arranged into several groupings such as "page views by hour of the day" and "page views by top level domain of visitors' addresses," were summarized and in many instances graphed. The third analysis type focused on detailed descriptions, covering issues such as the paths (sequences of pages visited) visitors followed on their visit to the site and the duration of visitors' stays, of site visits by smaller subsets of the entire visitor population during the case study week. The last analysis is a rich description of various aspects of an unusual and illustrative event in the site's history, dubbed the "Microbes on Mars incident."

Computer software was used throughout this study for data analysis. Computer-assisted analysis procedures used in this study can also be divided into four major categories. A Web server log analysis program called ServerStat, which is produced by Kitchen Sink Software, was used extensively to produce elaborate reports of site activity from raw server log record files. Spreadsheet software was used to collect data and to support exploratory investigation of potentially interesting combinations of data, such as dividing weekly page view counts by weekly visitor counts to determine weekly average page views per visitor values. Graphing capabilities within the spreadsheet program enabled rapid construction of visual displays of data trends, thus bringing the pattern recognition capacities of human vision into

use for data analysis. Finally, a database program, Claris' FileMaker Pro, was used to filter and sort raw server log records to support detailed analysis of behaviors of individual site visitors.

Reports from ServerStat log analysis program. Kitchen Sink Software's ServerStat log analysis program is typical of the genre. Use of software of this nature is within reach of most educators; it is inexpensive, not difficult to learn how to operate, and sufficiently common and in widespread use to insure continued availability of comparable tools for a reasonable time period. Many log analyzers are available for free or at low cost. ServerStat comes in a "Lite" version which is free and fully capable of support the analyses used in this study. We used the commercial version of ServerStat, which cost $100, and supported some time-saving batch processing and automation features absent in the Lite version. Use of such software requires a person with a level of savvy in using computers comparable to that required for use of a database or spreadsheet program. There are many analyzer programs on the market which produce reports similar to ServerStat's, so ongoing projects reliant on such software need not be concerned about the continued availability of a particular product to support long-term research efforts.

ServerStat's preferences allow a user to generate reports covering only specific files in certain directories, or with certain trailing extensions, or both. In this study, that feature allowed me to focus solely on the DLC-ME Web site, as opposed to the other sites hosted by the Comm Tech Lab's server which were logged to the same file. It also allowed me to filter out hits on elements such as graphic files (which end with the ".gif" and ".jpg" extensions), thus concentrating the focus of reports onto page views (files ending with ".html").

Each log analysis report covered one week. The top of each report included

a summary of total site page views for that week as well as a tally of the total number of unique network addresses which accessed the site during each week. For the purpose of this study, unique address counts have been equated with "visitor" counts. Complications in relating such "visitor" counts to tallies of human visitors to the site are described elsewhere in this report. Weekly reports were generated for a period spanning most of the DLC-ME site's history, and cover 88 weeks or more than one and one-half years, from November 1995 to July 1997.

Each log analysis report also provides breakdowns of file transfers (which equate to page views in this case, since only HTML files were counted) by various categories. The categories generated for and used in this study include: transfers by day of the year, transfers by hour of the day, transfers by day of the week, transfers by client domain, transfers by client reversed subdomain, and transfers by file section. These summaries were primarily used for the in-depth description of the "typical" week chose for the detailed case study, but also provided some of the data used in describing the "Microbes on Mars incident."

Spreadsheet and graphing. The spreadsheet section of ClarisWorks was used to accumulate data from the weekly log analysis reports. A spreadsheet provided a simple template in which to assemble tables of data. It also supported exploration of data trends involving combinations of basic values. Weekly average page views per visitor, arrived at by dividing the page view count column values by the visitor count column values, proved to be an interesting derived quantity. Other derived values, such as the site page views to site home page page views ratio, were less informative and were not included in this final report. Assembly of the data in a spreadsheet made it easy to try "what if" combinations to see which were valuable, and which

were less interesting.

ClarisWorks' spreadsheet also has a built-in graphing feature, which enabled further data trend exploration through visualizations. Several trends were most easily spotted when portrayed graphically. Integration of the graphing features with the main data depository supported data exploration efforts; I tried portraying many of the data sets used in this study in a variety of graphical forms before settling on formats found in the results section of this report. Bar charts, scatter plots, pie charts and line graphs all seemed useful in some instances for some data types; the final versions of graphs presented in the results chapter frequently represent a series of experiments in how best to convey information about specific data sets.

Use of database for filtering and sorting. Reports from log analysis programs provide a means for creating massed summaries of data trends for many page views and visitors but are ill-suited to detailed tracking of individual visitors. Raw log files contain data about individual visitors and visits, but in a form that is very difficult to interpret. In this study, I imported log files into a database analysis program in order to study detailed behaviors of fairly small numbers of visitors. I used Claris' FileMaker Pro database software, since I am familiar with it, but any of a number of commercially available products could be used. The database's searching and sorting features allowed me to examine the paths, or sequences of pages visited, which visitors followed during their site visits. Those features also enabled me to investigate the amount of time visitors spent on each page visited, and the duration of entire visits to the DLC-ME site.

Server log files, which are arranged as a series of records in subsequent lines containing tab-delimited data fields, are initially stored as simple text files. A small amount of formatting was required to prepare these files for

import into a database. Each file has a descriptive header line, which contains no visitor data, and which therefore was removed. When log files are swapped, replacing an empty new one with an old full one, the last record in the file may be broken when logging is briefly interrupted. Using a text editor, I looked at the end of each log file and cut out the first part of any broken final records. In many cases, individual log files span less than one week's time. In such cases, I opened the files and simply pasted each successive file onto the end of the previous one, again using a text editor, in order to assemble a single file covering a one-week long period. At this stage, the file was ready for import, as a tab-delimited text file, into the database program.

After the data was imported into the database, I did some "data cleaning" to eliminate records irrelevant to my needs. I searched for all records that ended with ".html" in the URL field, and removed all of those which did not meet that criterion. The "cleaned" data thus contained only page views, instead of all hits; this action primarily removed image files in JPEG and GIF formats. Another search for records starting with "http://commtechlab.msu.edu/dlc-me/" identified requests for files in the DLC-ME Web site, as opposed to the other sites hosted by the Comm Tech Lab's server. Again, I removed all records which did not match my search criterion. The database records at his point represented page views of DLC-ME site pages, and were ready for analysis.

I selected two groups of visitors, based on page view counts per visitor, to study in depth using the database records. Preliminary data analysis using ServerStat reports provided criteria for selecting which two visitor groups to study in this fashion. Details of that preliminary analysis, and the rationale behind the selection, are described in the results chapter of this report. The two visitor groups selected were those with seven and with thirteen page

views during the course of the case study week. I collected the network addresses of all such visitors from the "transfers by client reversed subdomain" section of the ServerStat report. I used the network addresses thus obtained as input values for a search of the database's hostname field to locate all log entries associated with each relevant "visitor." Next, I sorted the found records by date and time to create a sequential listing of all file requests for a given "visitor." Sequential lists of page views created in this manner for each of the visitors studied formed the basis of the remainder of my inquiries.

The date and time fields supplied the information necessary to determine whether the page views for a single visitor were all from a single site visit, or from multiple visits on different days or at widely separated times. Those fields also enabled calculation of the amount of time between requests for pages, and thus provided a basis for estimation of visit durations and for "time on page" calculations. The URL field in each of the records allowed me to determine the path through the site, in terms of new pages visited, of each visitor. I noted URL of the first page viewed in a visit sequence, which is of interest since it serves as the "front door" or site entry point of that visit. I also noted the value contained in the referrer field for each first page of a visit record, since it often contains information about the location on the Web of a visitor immediately prior to arrival at the DLC-ME site.

Finally, I compared the referrer field value of each sequential record in a given visit with the URL of the immediately prior record. These two values should match if the log is a complete record of the sequence of page views in a given visit. I suspected that that might not always be the case, because of caching of pages by browsers and other causes. This referrer to prior URL matching helped me understand how complete, and thus to a large extent how useful and reliable, the record of page views in the server log file is.

Chapter Four

RESULTS AND DISCUSSION

The first section of this chapter presents the long-term trends, spanning most of the online history of the DLC-ME site, of page view and visitor counts for the entire site. That data was used to select a typical week upon which to focus some more detailed investigations. The second section of the results chapter presents analyses of data describing that typical week. The first segment of the typical week's analysis describes "where" visitors came from, in terms of their network addresses or in terms of the "referrer" sites and pages that visitors were viewing immediately before visiting a page at the DLC-ME site. The second segment summarizes "where" in the DLC-ME site visitors went during their visits, in terms of specific Web pages or sections of the site. The third and final segment of the typical week's analysis summarizes the "when" of high and low visitor traffic levels during the course of one week, in terms of hours of the day and in terms of day of the week.

The third major section of the results chapter covers detailed investigation of visitor behaviors of small numbers of visitors during the typical week which are grouped into two specific categories based on visit durations in terms of page views. That section also presents data on the viability of measuring visitors' visit durations in terms of time, in hopes of establishing a "time on page" metric similar to the "time on task" measure employed by many educational researchers. The fourth and final section of this chapter presents data and commentary about an informative anomalous event, dubbed the "Microbes on Mars incident," which was in many ways the polar

opposite of the typical week intensively analyzed in much of the rest of the chapter.

Page View and Visitor Count Trends Over Time

Figure 4 shows the total page view count per week for the entire DLC-ME site over time. The figure spans the period from November 1995, when we first began saving the Web server's log files, through June 1997. The dominant feature displayed by the graph is the growth of site traffic, from an initial level of one thousand page views per week to a sustained level of more than six thousand page views per week. This figure also shows that dramatic short-term fluctuations in traffic levels are common. The overall traffic growth trend was interrupted by slumps during the summer months (in the northern hemisphere). These seasonal slumps correspond to the summer vacation periods when most schools are not in session. Page view counts picked up rapidly at the end of summer, when students returned to classrooms. Dramatic dips in page view counts also occurred during late December, around the time of Christmas, New Years, and other year-end holidays and the corresponding school vacation periods.

Figure 4 shows large fluctuations in page view counts during the spring and early summer of 1997. An especially large increase occurred in late May. I have not yet attempted a detailed analysis of the causes of these trends in the data. The drop-off in page view counts in June of 1997 is similar to the summertime slump of 1996, and is probably associated with the end of school and the beginning of summer vacations.

Figure 5 shows total "visitor" counts, on a weekly basis, for the entire DLC-ME Web site from November 1995 through June 1997, the same time span

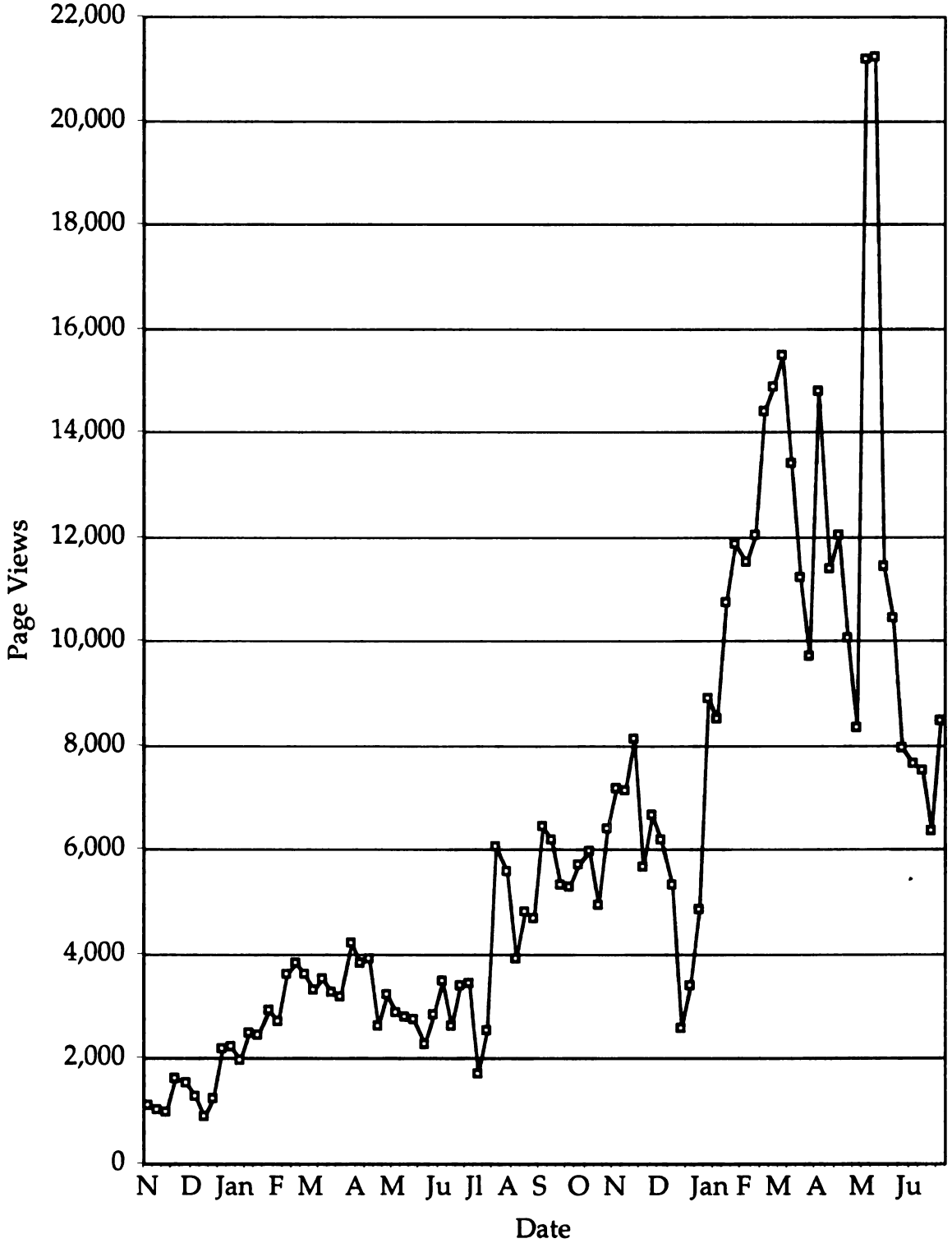

Figure 4 - Weekly Page View Count for Entire Site (11/10/95 - 7/11/97)

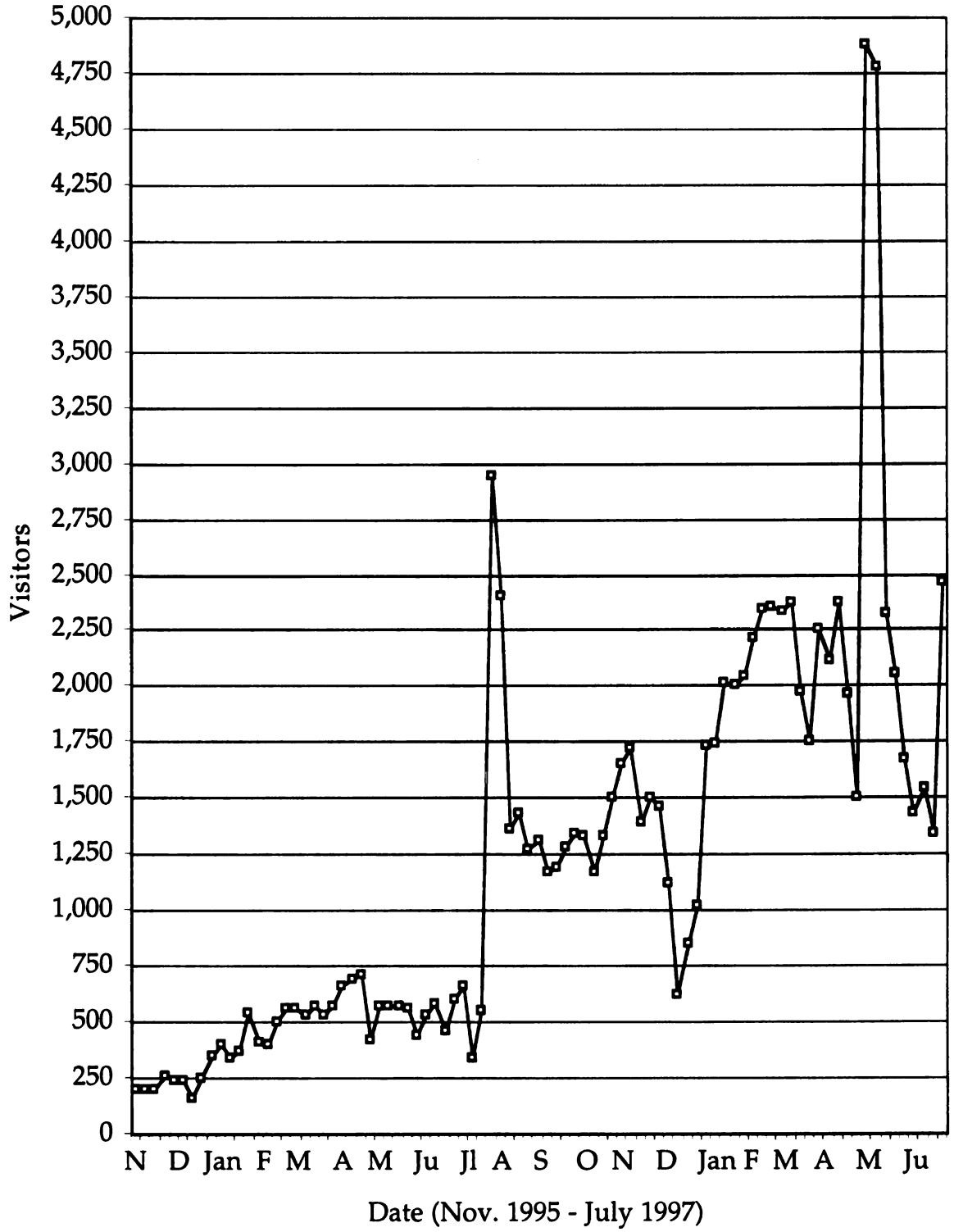

Figure 5 - Weekly Visitor Count for Entire Site (11/10/95 - 7/11/97)

covered in Figure 4. A "visitor," in this context, is actually a unique network address, which in many cases does not identically correspond to an individual person. The trends in visitor counts displayed in Figure 5 are similar in many respects to those evident in page view counts as shown in Figure 4. Site traffic has grown over time, from an initial count of around 250 visitors per week to a sustained level of well over 1,000 visitors per week. Weekly visitor counts sometimes showed substantial fluctuations from week to week, although these fluctuations were generally smaller than those found in page view counts. Summertime and year-end dips in visitor counts were also similar to reductions in page view counts at those times. Likewise, visitor counts increased in early autumn, at the beginning of the new school year, as was the case with page view counts. Erratic fluctuations in visitor counts during the spring of 1997, and the large peak in late May of 1997, are also similar to the trends in page view counts displayed in Figure 4. The increase in visitors during August 1996 is much more prominent than the surge in page view counts at that time.

Figure 6 is essentially the union of Figure 4 with Figure 5, and shows both page view and visitor counts for the DLC-ME site over the November 1995 through June 1997 period. Page view and visitor counts are fortuitously of such a magnitude to permit their simultaneous display on a single graph with one scale for the vertical axis, without the two trends overlapping and thus confusing a viewer. This presentation readily permits comparison of the two trends over time, allowing a viewer to easily spot similarities and differences. The vertical scale, which accommodates the largest values for page view counts, compresses the range of the visitor count data, thus making it difficult to see the smaller fluctuations in visitor counts. A display such as Figure 6 would be most valuable to a site evaluator when

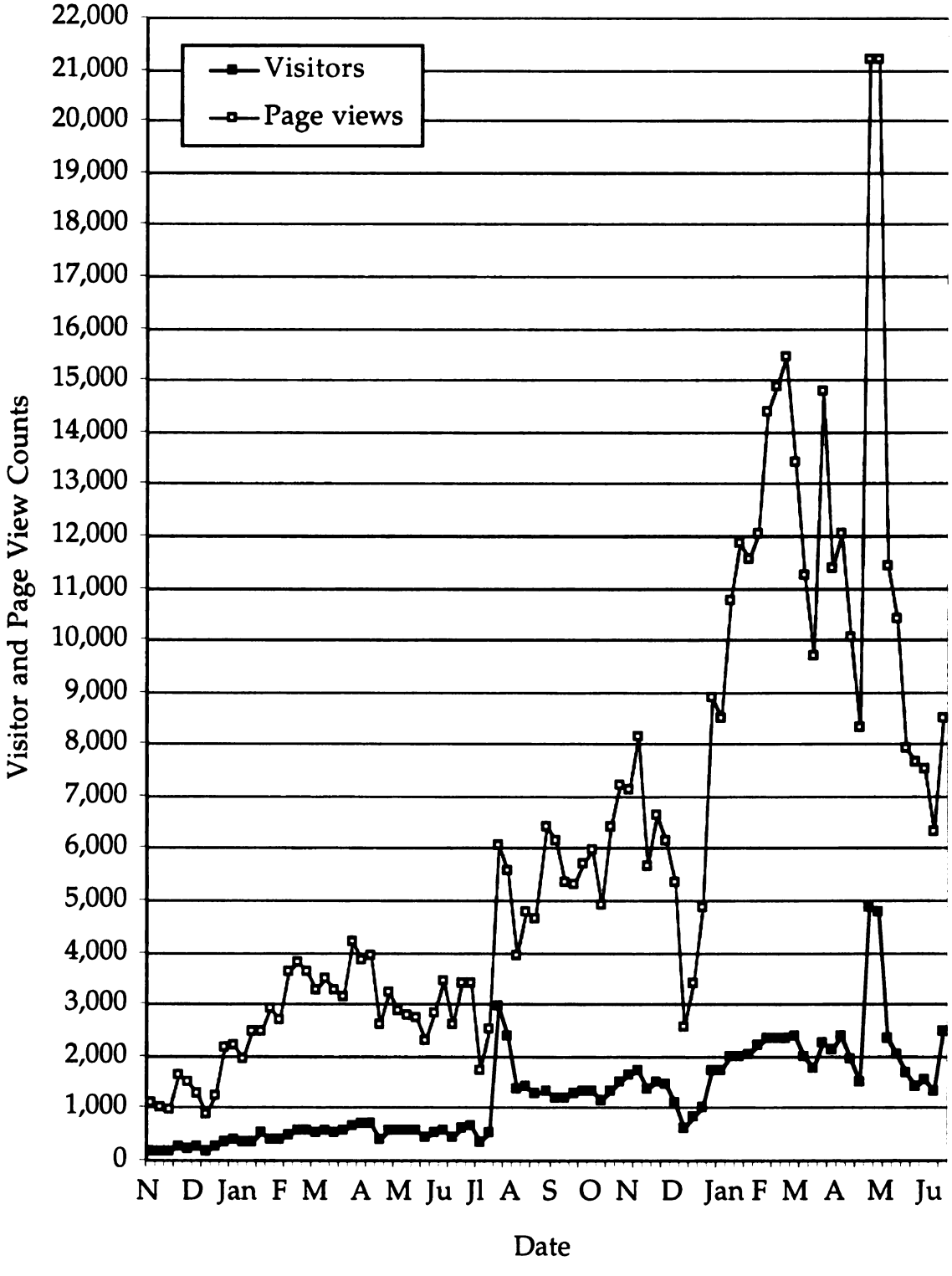

Figure 6 - Weekly Page Views and Visitors for Entire Site

comparisons between the two trends over time is of paramount importance. However, the individually tailored vertical axis scale of Figure 5 permits viewing of finer details. Site evaluators need to balance the extra time required to create multiple representations of the data against their needs and the intended emphases of their studies in deciding whether to produce such multiple renditions of the same data. In my study, these displays of trends in overall site traffic over the history of the site proved invaluable in identifying specific periods upon which to focus more detailed examinations.

The disparity between the size of the increase in page view and visitor counts during August 1996, as well as curiosity about how many pages the "average" site visitor viewed, inspired me to calculate the page views per visitor ratio for each week. Since I was recording and graphing site traffic data in a spreadsheet program, it was easy to create a new column listing the page views divided by the number of visitors. Figure 7 presents these data as a graph, again covering the history of the DLC-ME site from November 1995 through June 1997 on a weekly basis. The most striking feature shown by this graph is the remarkable consistency of average page views per visitor throughout the site's history, which falls almost entirely within the four to seven pages per visitor range. I expected that there would be a strong upward or downward trend in this value as the site's traffic levels rose dramatically throughout the site's history. Instead, it has remained remarkably constant. The main exception to this consistent behavior was during the "Microbes on Mars incident" in August 1996, which was a noteworthy aberration that is discussed in greater detail later in this report. Average page views per visitor is a rough measure of the average amount of content seen by visitors, and thus is a quantity of potentially substantial educational interest. It is somewhat analogous to the amount of time visitors spend viewing exhibits

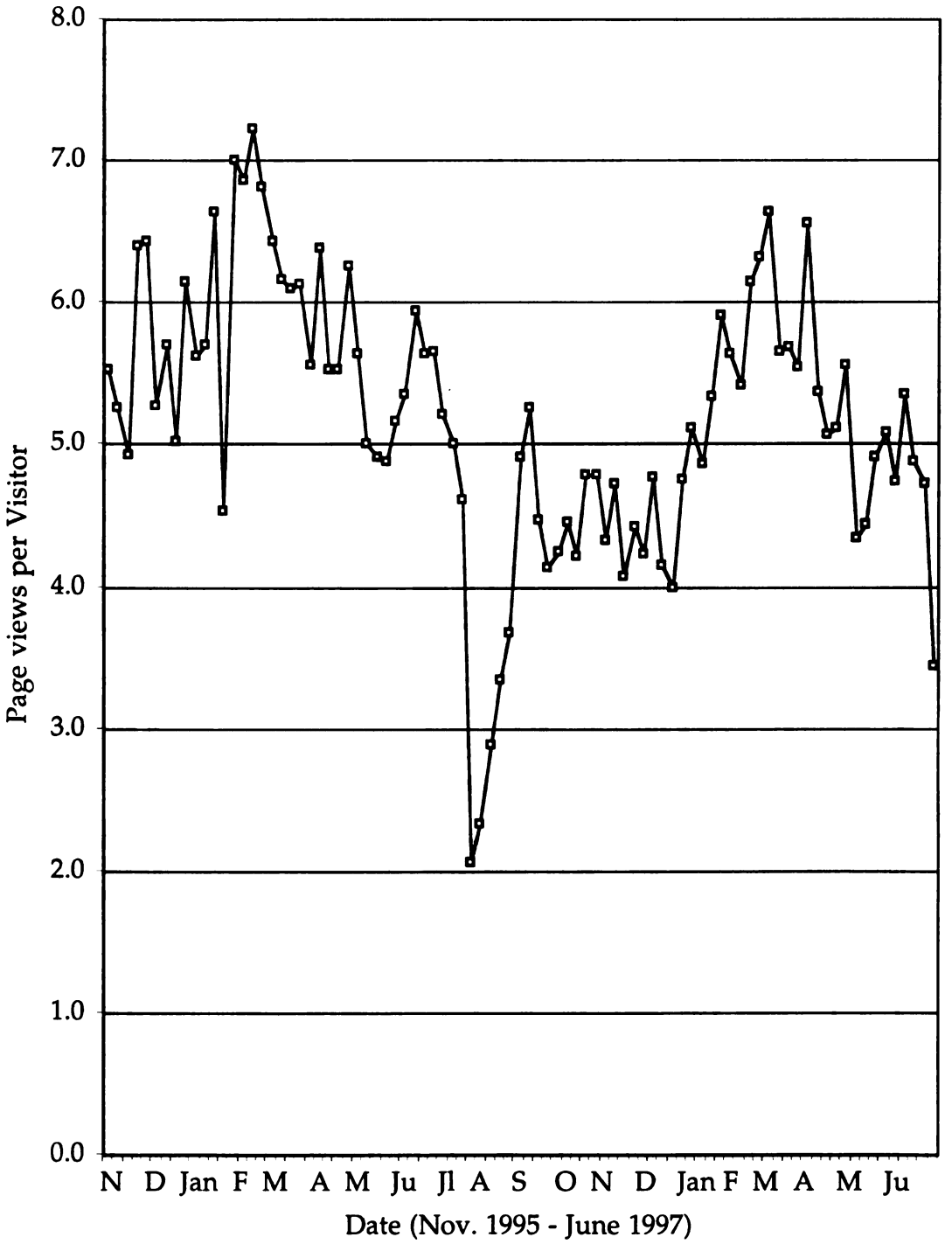

Figure 7 - Weekly Page Views per Visitor Ratio (11/10/95 - 7/11/97)

in a museum, a quantity often measured by museum visitor studies researchers because of it's frequent correlation with learning. Graphs similar to Figure 7 may prove a valuable tool for Web visitor studies researchers, since this figure clearly shows the unusual nature of the period around August 1996, indicating that a phenomena worthy of closer scrutiny was occurring.

Case Study of a Typical Week

The next two major sections of this chapter focus on a specific case-study week and detailed investigation of patterns of site traffic during that one-week period. The long-term trend analyses helped me identify a "typical" week to study in greater detail. I chose a week that was not too early in the site's history, for which the traffic levels were fairly high so that a wide range of visitor behaviors would likely be represented and analyses dependent on larger sample sizes would have greater validity. I chose a week during the school year, so that school-related site usage would likely be represented. The chosen week was not too close to either the beginning or the end of the school year, or to the December break time, so that fluctuations associated with those times would be absent. I avoided times when sudden fluctuations in page view counts were evident, such as during August 1996 and May 1997. Since both visitor and page view counts have grown steadily throughout the history of the site, I chose a week which fell within a steady growth trend period. The case study week I selected is the week of February 1-7, 1997.

This case study week was selected specifically because it was likely to represent a typical week, and therefore results gleaned from such study seem likely to apply to other typical weeks. Atypical weeks are also potentially

interesting specifically because of their unusual nature; the "Microbes on Mars incident" discussed later in this chapter is an example of the study of an atypical period during the history of this Web site. Decisions about whether typical or atypical periods during a site's history should be examined must be made by site developers studying their own sites, and will depend on their research goals. In either case, long-trends in page view and visitor counts can aid identification of typical and atypical periods in the site's history.

Figure 8 shows the distribution of page views per visitor during the week of February first through the seventh of 1997. The term "visitor" once again refers to a single network address, not an individual person. Most visitors, in this case 840 of them, visited only a single page before leaving the site. The next most common behavior was to visit two pages in the DLC-ME site during this case study week; 266 visitors did so. Note that visitors who visited more than one page may not have viewed multiple pages in succession during a single visit, but may have looked at one page on one day and another a few days later during the same week. The largest number of page views per visitor shown in Figure 8 is 20; 4 visitors logged that many page views during the case study week. The most pages viewed by a single visitor during that week was 953, which made it impractical to show the range of all possible page view counts per visitor in Figure 8. Most visitor counts for page views per visitor values above twenty were small; eight visitors logged 22 page views, eight others logged 24 page views, six visitors logged 28 page views, and all other visitor counts for a single page views per visitor value were four or less. All page views per visitor values above 44 logged at most one visitor per value.

The dominant message conveyed by Figure 8 is that the vast majority of visitors looked at a very small number of pages. This finding is reminiscent

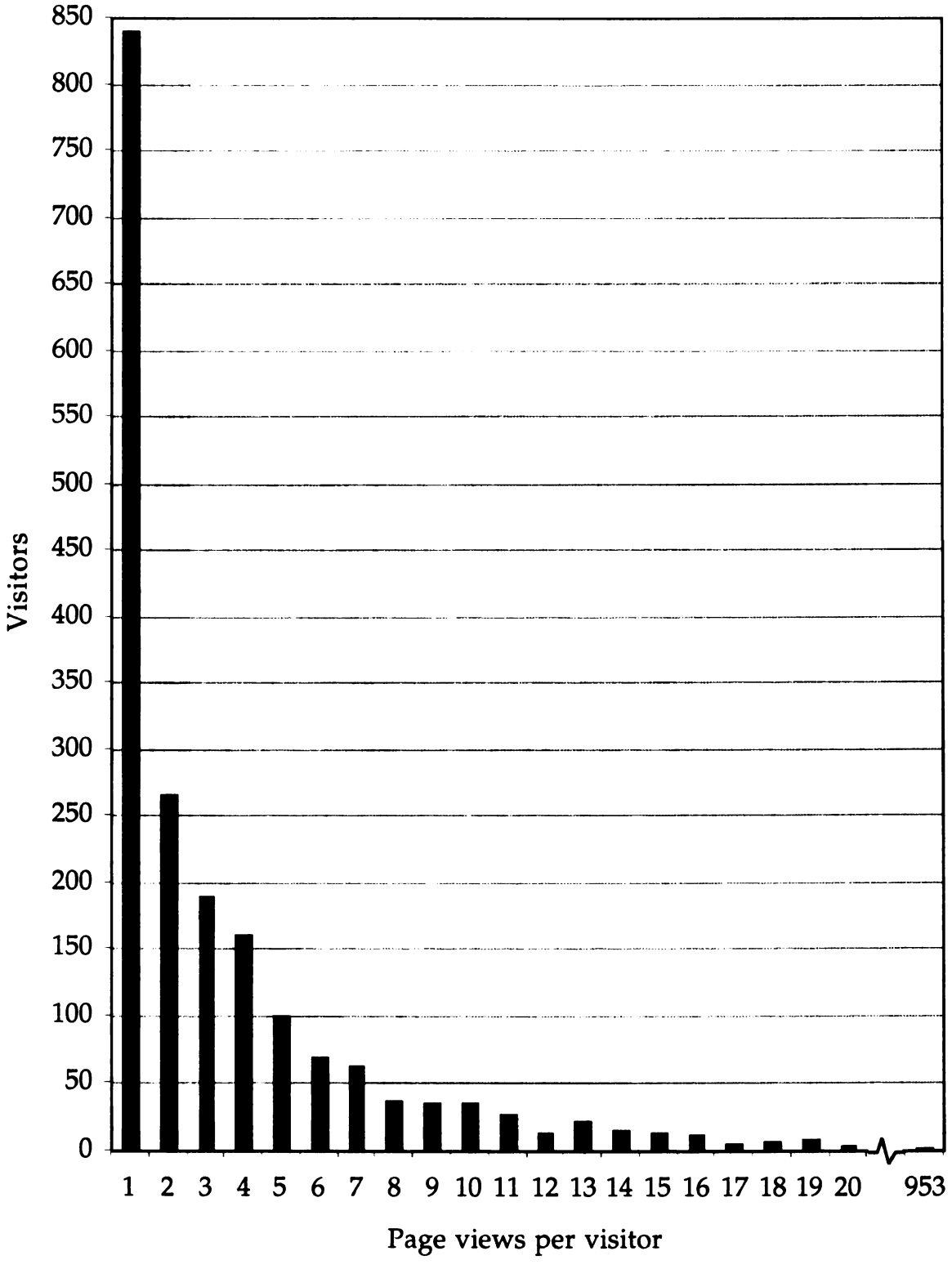

Figure 8 - Page Views per Visitor Distribution (case study week)

of the discovery by museum visitor studies researchers that most visitors spend mere seconds viewing most exhibits. The mean number of page views per visitor during the case study week is about 5.9, while the median page views per visitor is 2. There were 2,004 visitors during the case study week, and 11,849 total page views.

Where visitors came from. Figure 9 shows the page view counts during the case study week of the individual network addresses which accounted for the largest number of page views. The computer with the network address "css6.cl.msu.edu" was by far the site's most persistent visitor, logging 953 page views during the course of the week. The top three addresses accounted for approximately 15 percent of all page views for the week. An apparently related group of computers, whose IP addresses all begin with "198.146.15", also registered a large number of page views. Most of these high page view count addresses probably represent webcrawler robots, proxies at Internet service providers which assign temporary addresses to dial-up users, or computers which are shared by several users in labs, classrooms, or similar environments. The lowest page view count address displayed in Figure 9, "35.8.111.115", is the address of my workplace computer on campus. It's inclusion in this figure underscores the need for site evaluators to account for site accesses by site developers, which may be very plentiful, when assessing the significance of site traffic levels. We had initially filtered out page views registered by the DLC-ME development groups, using an option in our server log analysis program, but overlooked changing that filter list when I moved to a new work location and a new network address on campus.

Figure 10 shows the distribution of page views during the case study week by top level domains of the network addresses of visitors. Top level domains are the most general classification of network addresses, and include the six

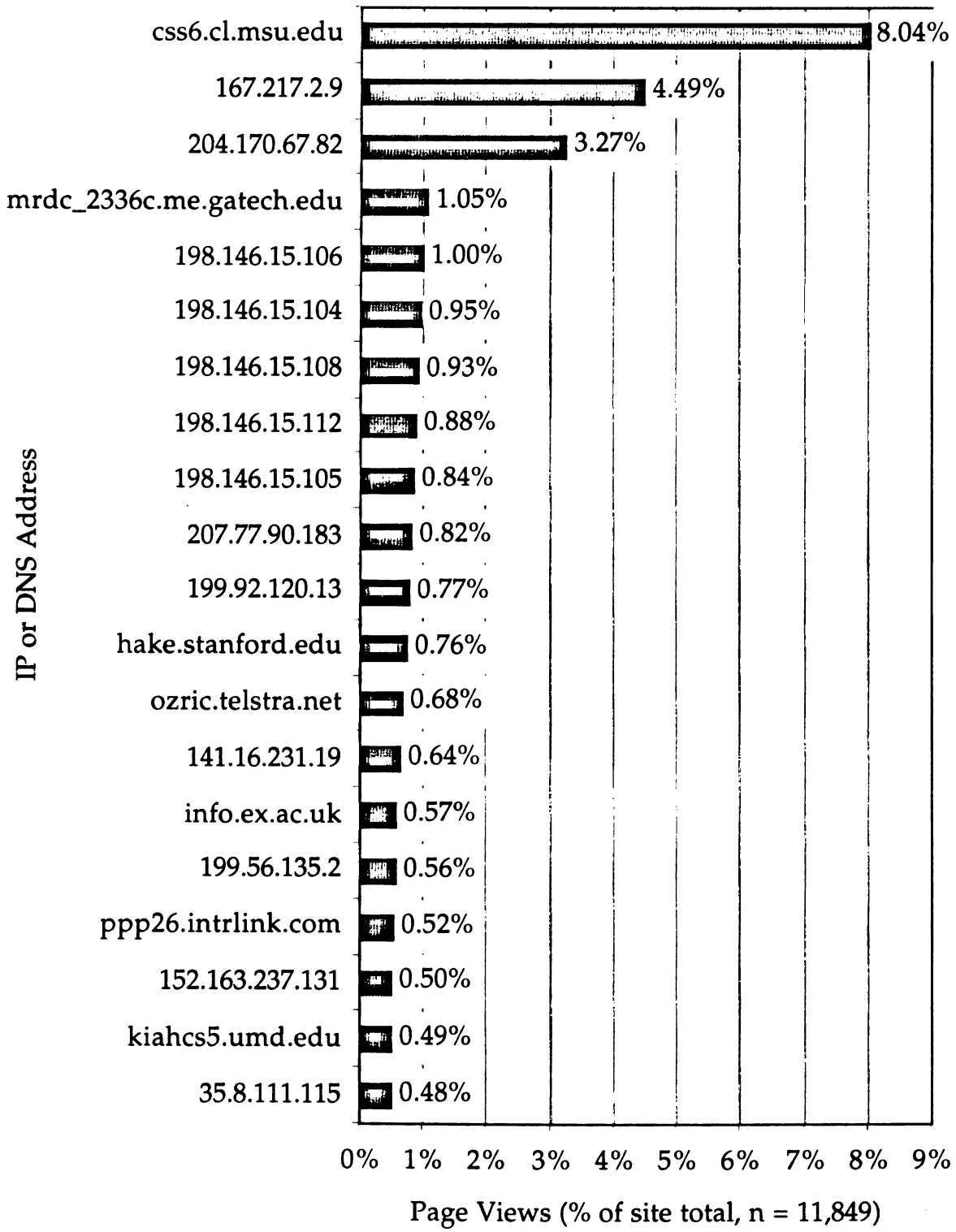

Figure 9 - Page Views of Top 20 Visitor Addresses (case study week)

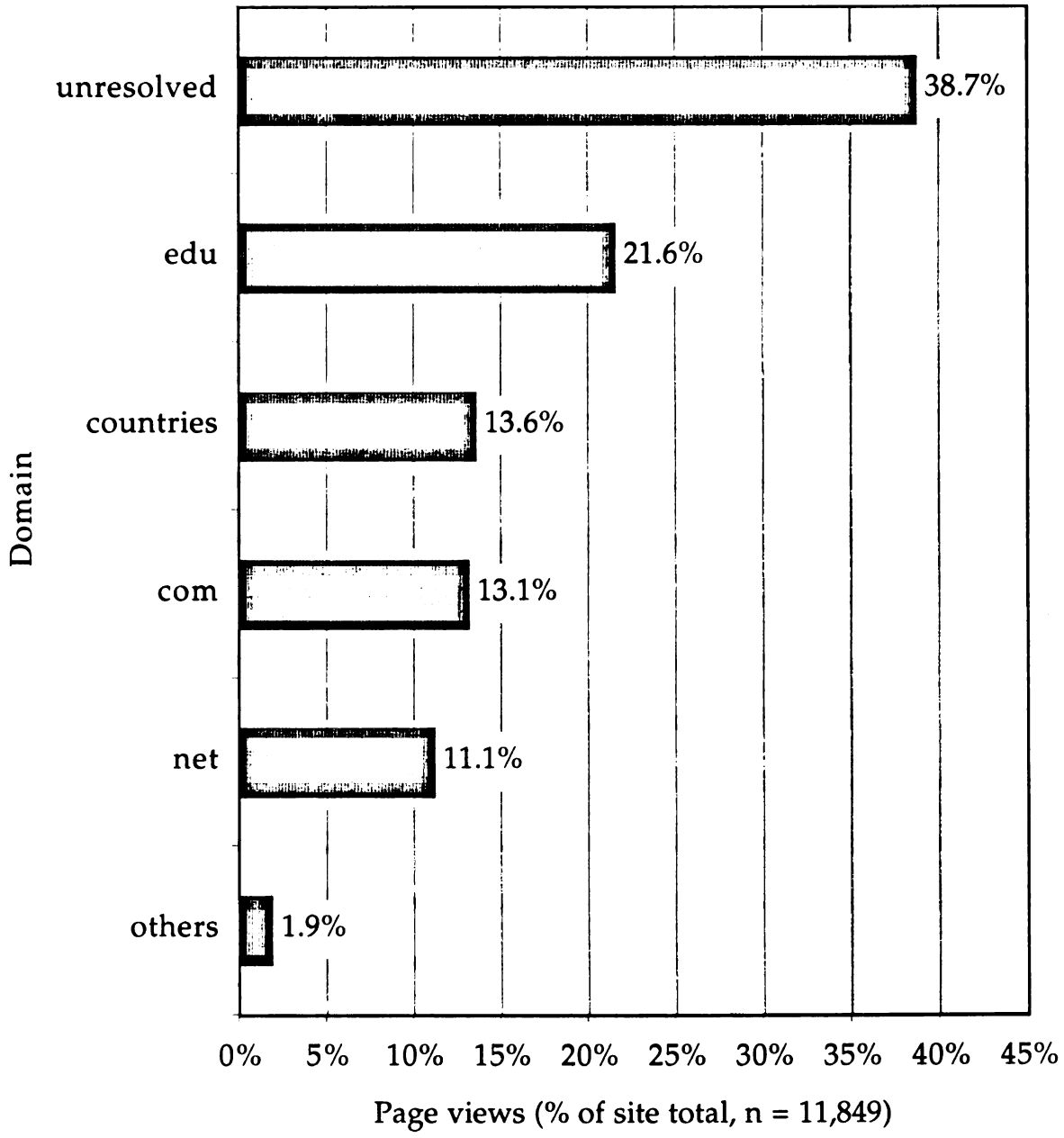

Figure 10 - Page Views by Top Level Domains

main U.S. top level domains ("com", "edu", "gov", "mil", "net" and "org") as well as the more numerous international country codes (such as "uk" for United Kingdom, "ca" for Canada, "de" for Germany, and so on). Many network addresses cannot be resolved in terms of their Domain Name System (DNS) identities, of which the top level domains are an element, and are logged by the Web server in terms of their less informative, numeric Internet Protocol (IP) addresses. Such unresolved addresses account for the largest single category in terms of page views in Figure 10, representing 39 percent of the count total. The largest identifiable source of site traffic came from the educational (edu) top level domain, which accounted for 22 percent of page views. The various country top level domains, as a group, accounted for the next largest portion of identifiable addresses. Figure 11 shows the distribution of page views per individual country top level domains which make up the "countries" category in Figure 10. The commercial (com) and network (net) top level domains, respectively, accounted for most of the remainder of page views, as shown in Figure 10. The "others" category, which consists of the organization (org), government (gov), and military (mil) top level domains, generated a small fraction of the total page views for the DLC-ME site during the case study week.

Figure 11 shows the distribution of page views by country top level domains. The number of page views for each country code is displayed as a percentage of all international top level domains. There were 1,608 page views by addresses with international top level domains, which is about 14 percent of the total of 11,849 page views for the entire site from all addresses, during the case study week. Country top level domains that registered more than twenty page views are shown in Figure 11. Predominantly English-speaking countries account for four of the top six page view generating

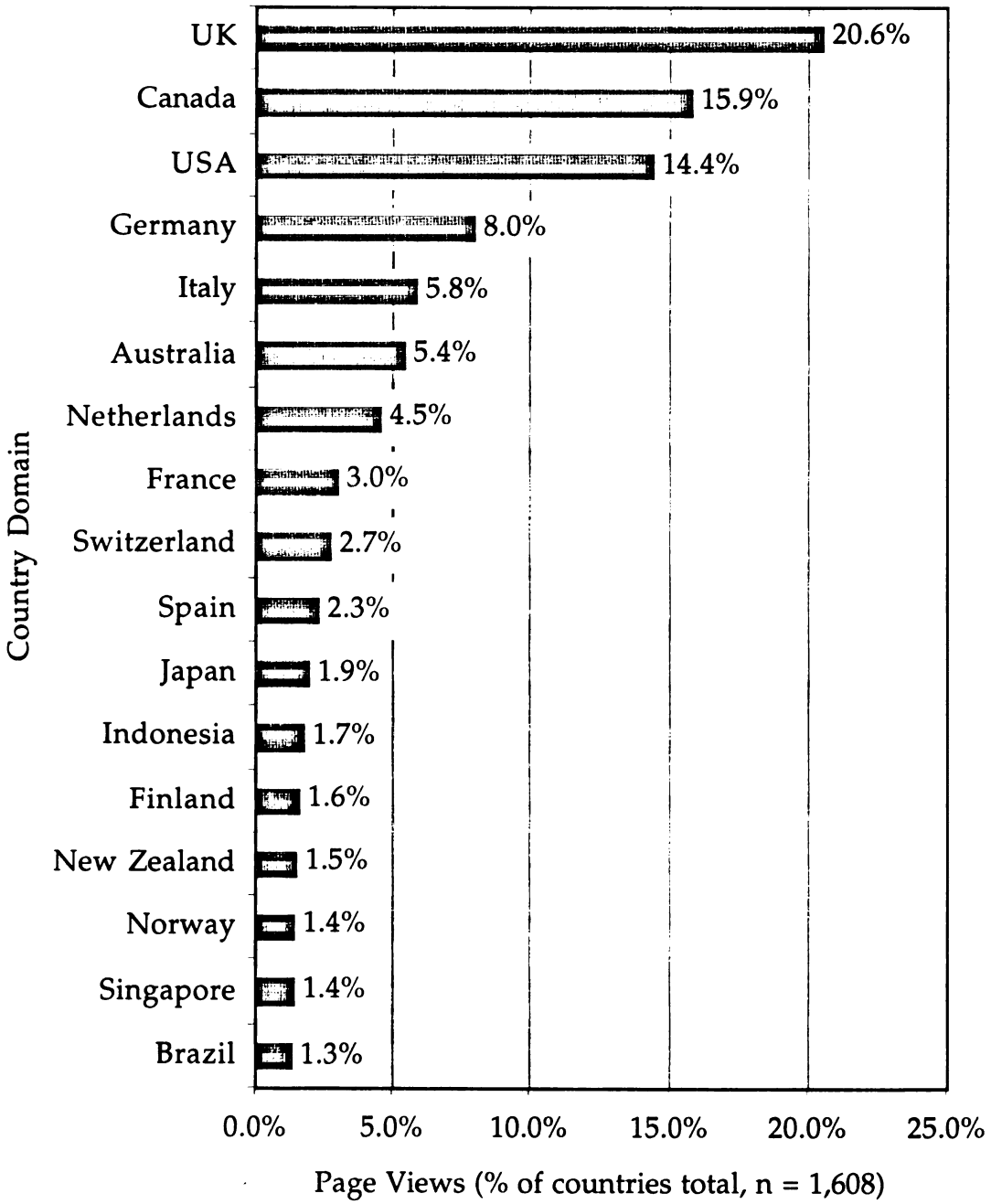

Figure 11 - Page Views by Country Domains

countries. Eight of the top nine page view generating countries are in North America or Europe. "Developing" nations possessing limited computer technology and telecommunications infrastructure resources are noticeably under represented in or absent from Figure 11; Brazil is the sole Latin American country shown, and no African nations broke the twenty page view level.

Table 5 lists lists several significant network sub-domains that accounted for large numbers of page views. For instance, fifty different network addresses ending with "aol.com" tallied a total of 397 page views during the case study week. The server log analysis program which I used for these analyses lists page views by individual network addresses in alphabetical order, with the domain and sub-domain orders reversed. In other words, the address "commtechlab.msu.edu" is listed in the report as "edu.msu.commtechlab". This approach facilitates recognition of related addresses via visual scanning of the log analysis program's report, since all addresses ending with "aol.com" (for instance) appear in a cluster sequentially in the report. In this case, there were fifty consecutive entries beginning with "com.aol", which were easy to spot while scrolling through the report file. Each of the entries in Table 5 was identified by this visual scanning method. All of the groups in Table 5 apparently represent site accesses by subscribers to major Internet service providers (ISPs).

The technique used to identify clusters of network addresses, as was done in constructing Table 5, could help site evaluators determine various groupings of related addresses. Which groupings site evaluators choose to identify will depend on the goals of their evaluation efforts. Clusters of related addresses may represent visitor constituencies of interest to site evaluators and developers. For instance, developers of a Web site intended

Table 5 - Major Sub-domains

Sub-domain	Page views	Distinct network addresses
aol.com	397 page views	50 different addresses (ending with ipt.aol.com or proxy.aol.com)
compuserve.com	96 page views	23 different addresses
netcom.com	85 page views	21 different addresses
prodigy.com	61 page views	6 different addresses

Table 6 - Top Referrers

Referrer	Page view count	Page views (% of total)
Internal (other dlc-me pages)	7113	60 %
None	2934	25 %
External	1802	15 %

External referrer	Page view count	% of External Referrers
yahoo	603	33.5 %
altavista	397	22.0 %
lycos (several different?)	186	10.3 %
excite	109	6.0 %
webcrawler	93	5.2 %
infoseek	66	3.7 %
comet (Cells Alive)	63	3.5 %
metacrawler (several different?)	38	2.1 %
hotbot	28	1.6 %
asmusa	25	1.4 %

primarily for use by Michigan State University students might wish to identify addresses ending with "msu.edu", and to determine the number of page views logged by all visitors with such addresses.

Large user constituencies identified via clusters of network addresses might help site developers understand the needs of their users, and tailor their sites to those needs. For instance, for a while America Online users had a non-standard Web browser client program, which required site developers to consider and test special Web page design constraints in order to insure AOL users would see the same pages as users of other browsers. The emergence of television-based Web browsing, via services such as WebTV, which also places constraints on page design techniques, may encourage site developers to investigate the sources of traffic to their sites.

Table 6 provides summaries of some of the referrer information embedded within server log analysis reports. Server log records often include data about the address of the Web page a visitor was "at" immediately before accessing a page on the server's site. This information can help site evaluators determine visitors' paths through their Web site, and can also help them discover which external sites led visitors to their site. Table 6 shows that 60 percent of DLC-ME site page views resulted from "jumps" from other pages within the DLC-ME site. Fifteen percent of site page views resulted from links to the DLC-ME site from other Web sites. About twenty-five percent of server log entries contained no referrer data. I have not explicitly investigated circumstances under which referrer data is not included in log entries, and thus can only speculate upon the causes of such omissions. I suspect that site accesses via URLs typed in by users, accesses to pages via previously set bookmarks, and accesses via menu driven lists of recently visited pages may not generate referrer data in log entries. I suspect that clicking on links

associated with text or images does generate referrer data.

The lower half of Table 6 lists some of the external referrers that "led" visitors to the DLC-ME site. The directory and search site "Yahoo!" and the search site "AltaVista" served as the gateway to the DLC-ME site in the case of more than half of the identifiable external referrers. Other search sites, including Lycos, Excite, Webcrawler, Infoseek, and HotBot, figured prominently in the external referrer list. The "Yahoo!" category shown in Table 6 is an amalgam of various pages throughout the "Yahoo!" site, including directory pages in several categories, "Yahoo!" search pages (which are powered by a copy of the AltaVista search engine), and the "Yahoo!" kid's, page which is called "Yahooligans." The Lycos category in Table 6 also represents various referrer pages, in this case apparently at different Web sites. Lycos has sold and licensed its search engine and webcrawler robot to other site administrators; the Lycos entry in Table 6 apparently represents several such licensee sites, not just the main Lycos search site. Two entries in Table 6, "comet" and "asmusa," are content-oriented sites covering topics related to the focus of the DLC-ME. The "comet" referrer address belongs to a well-know microbiology site titled "Cells Alive"; the "asmusa" address belongs to the American Society for Microbiology. Both sites have lists of related Web sites which include links to the DLC-ME, which was readily verified by visiting the referrer addresses using my Web browser. Site administrators might wish to use referrer information to discover "where" their visitors are "coming from," and to check whether their site is adequately represented on the major search and directory sites.

Where visitors went. The next two figures illustrate the distribution of site traffic throughout the sections and pages of the DLC-ME site. Figure 12 shows the distribution of page views across the major sections of the Web site. The

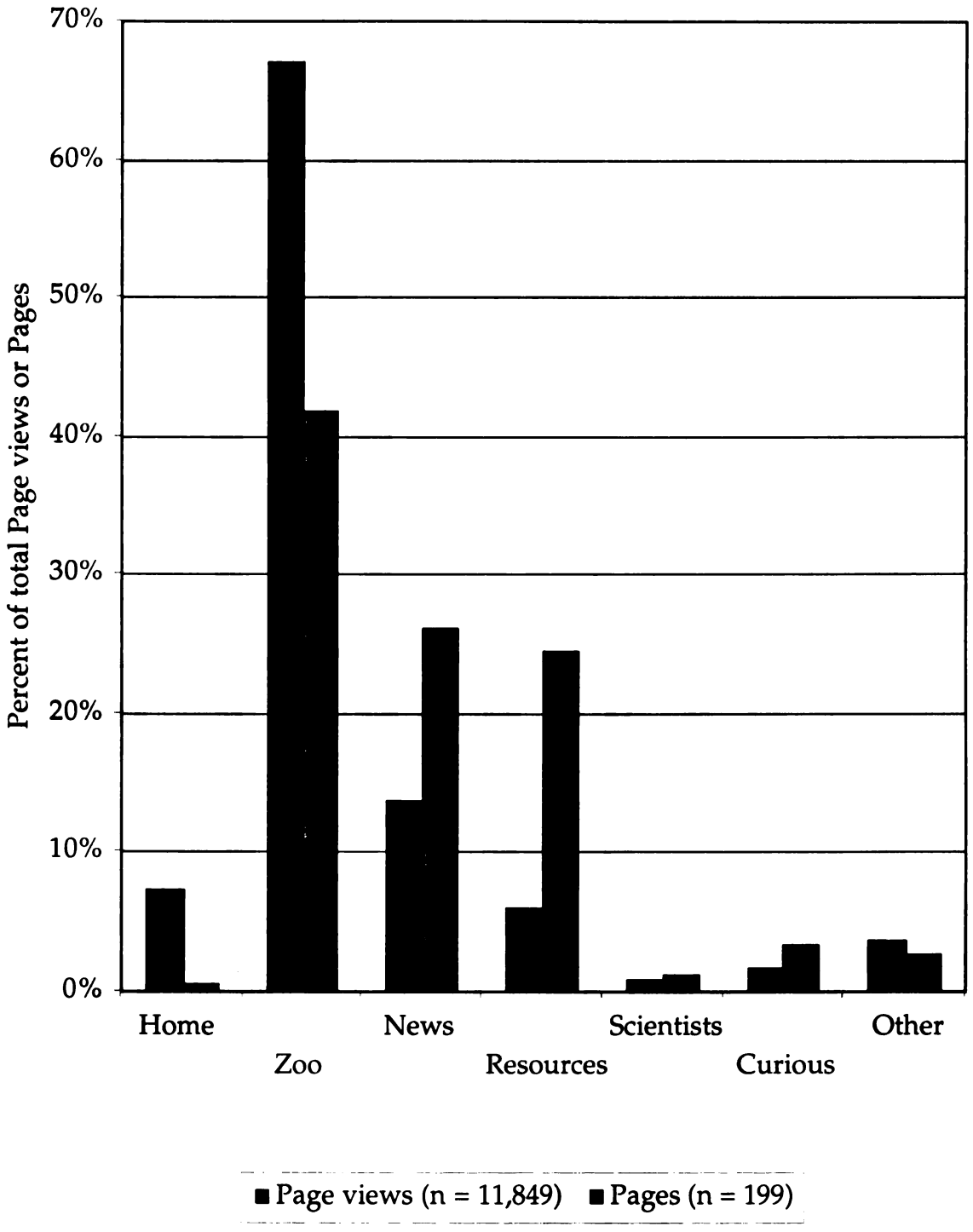

Figure 12 - Page View Distribution by Site Sections (case study week)

main site sections are: the Microbe Zoo, Microbes in the News, Microbial Ecology Resources, Meet the Scientists, and The Curious Microbe. Figure 12 also includes traffic level data for the DLC-ME site's home page and for an "other" category that encompasses pages that are not part of any of the major site sections. Each category has two types of data associated with it: the portion of site traffic which "visited" that section in terms of percentage of site page views during the case study week, and the portion of site pages in that section as a percentage of all pages in the DLC-ME site. The entire DLC-ME Web site consists of 199 pages, and logged 11,849 page views during the case study week.

The Microbe Zoo is the largest section of the site in terms of pages; it's 84 pages are about 42 percent of the site's total. The Microbe Zoo section also logged the most page views, accounting for over 67 percent of site traffic. The second "largest" section of the site, in terms of both pages and page views logged is the Microbes in the News section. The site's single home page logged more page views than any of the remaining site sections. The comparisons of page views and pages, as shown in Figure 12, supports evaluation of "relative efficiency" of pages in terms of the amount of traffic they attract or support. Not surprisingly, the home page is very efficient; its seven percent "share" of page views is much greater than its half percent of total site pages development "investment." Likewise, the Microbe Zoo section is an "efficient investment" in terms of the return in page views as compared to the number of pages developed. In those terms, the less than six percent share of page views logged by the Microbial Ecology Resources section as compared to its nearly 25 percent proportion of site pages makes it an "inefficient" section. Such evaluations are potentially useful, but must take other factors into account. Not all pages are equal, since some are more complex and require considerably more development effort investment than others. The location

of links to site sections on menu pages, such as a site's home page, can dramatically influence the number of visitors who explore or ignore a given section of the site.

Figure 13 shows all pages that logged at least one percent or more of the total page views for the entire site, and the percentage of site page views logged by each page. Not surprisingly, the Microbe Zoo section's home page and the DLC-ME site's home page were visited the most. Many of the other high traffic pages are part of the Microbe Zoo section, the most heavily used part of this site. The "Microbe of the Week" page, which is linked to directly from the site's home page and was one of the supposed "Martian microbes" discovered by NASA scientists in a meteorite from Antarctica, was also very "popular." Displays such as Figure 13 could be valuable formative evaluation tools for site developers, since they could help in determining whether pages intended to be heavily trafficked were indeed frequently visited. They could also reveal surprisingly popular topics or presentation techniques, or placement of links that are especially appealing. The prospect of finding life on Mars is apparently intriguing to many people, as evidenced by the high ranking in Figure 13 of the "Martian Microbe of the Week" page and a Microbes in the News article relating to the search for microorganisms on the red planet. A Microbe Zoo specimen page, "Spirogyra in Pond," apparently attracted site traffic with its animation of spirogyra chloroplasts, placing it within the top dozen most visited pages and possibly illustrating the impact of interesting media elements.

Busy times and slow times. The next two graphs show aspects of the temporal distribution of page views during the case study week. Figure 14 displays the distribution of page views by day of the week during the case study week. The lowest traffic levels occurred during the weekend, with

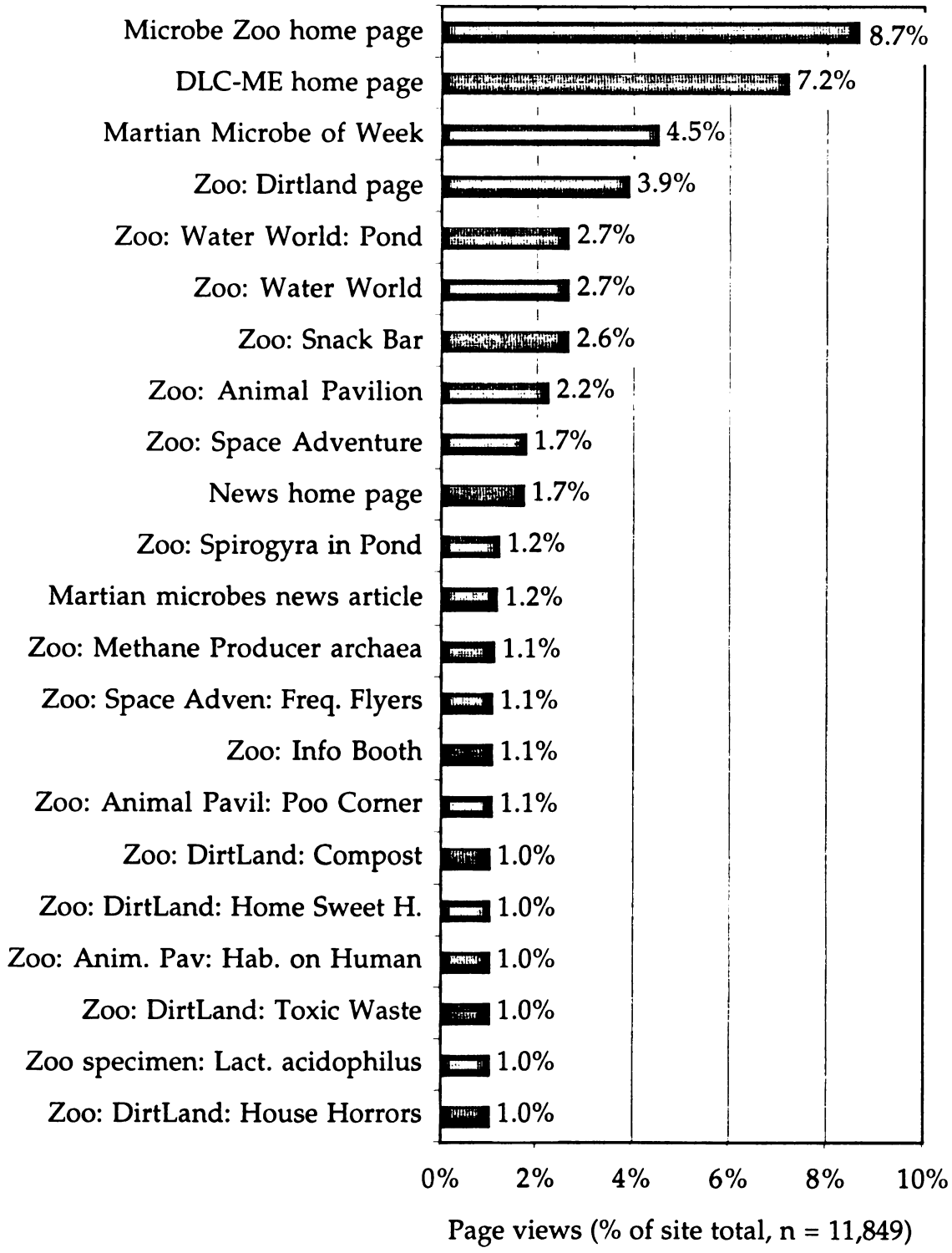

Figure 13 - Top Page View Pages (case study week)

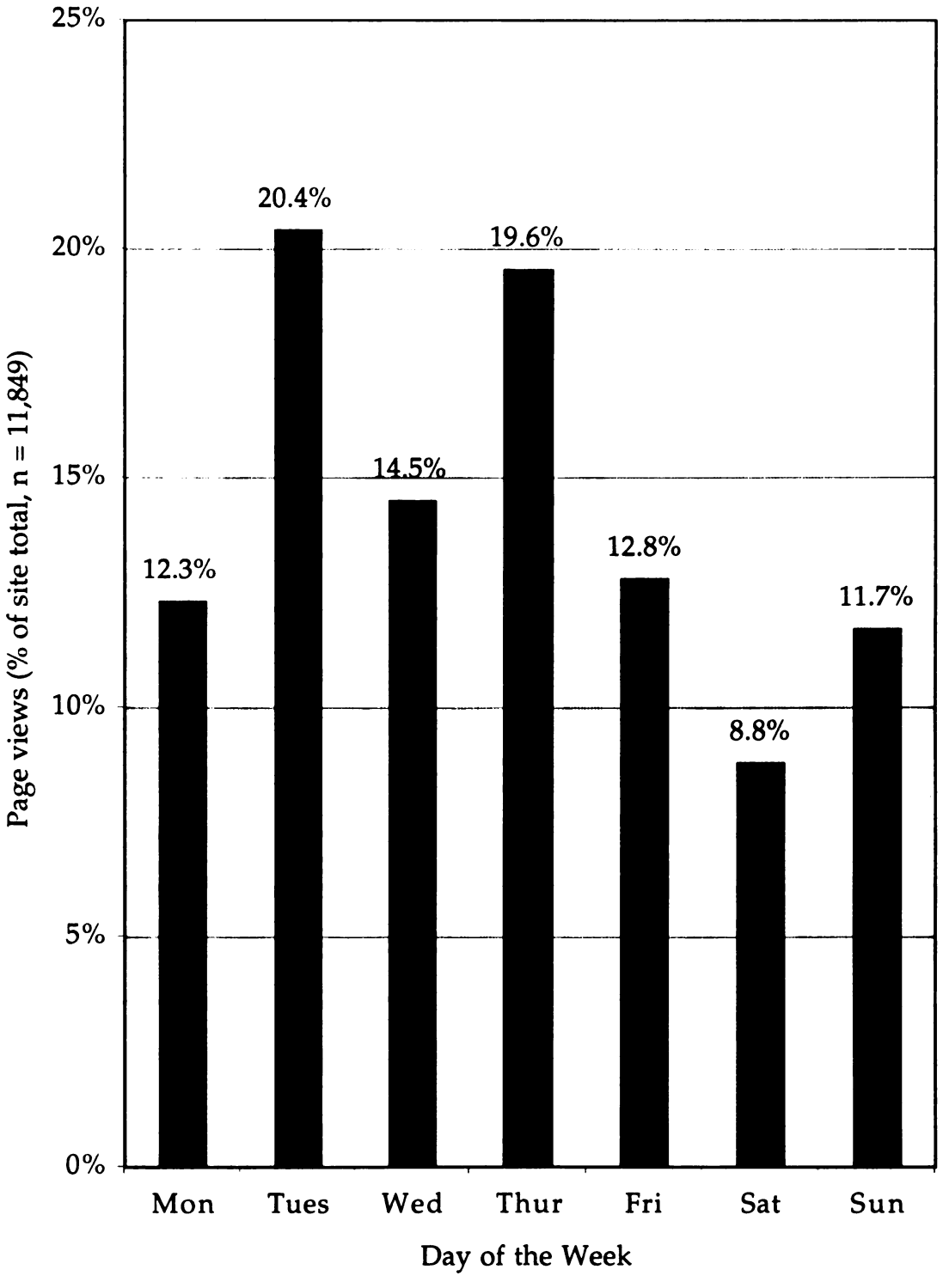

Figure 14 - Page Views by Day of the Week (case study week)

Saturday being slowest. Traffic levels peaked in the middle of the working and school week, with Tuesday leading the pack with slightly over 20 percent of all page views for the week. Such information could help site administrators schedule events, such as taking the site offline for maintenance or presentation of special live chat sessions, which could be set for low and high traffic periods, respectively.

Figure 15 shows the distribution of page views by hours of the day during the case study week. Values shown in the figure span the entire week, representing sums of page views on all seven days at a given time of day. Highest traffic levels fall between 8 AM and 10 PM, peaking in the early to mid afternoon. Lowest levels, not surprisingly, fall between 3 and 7 AM. All times are local relative to the Web site server, which is in the Eastern Time Zone of the Western Hemisphere. Site traffic from visitors in the western U.S. and Canada, from visitors in Europe, and from the populous regions of eastern Asia could influence traffic timing patterns in a graph such as this, emphasizing the truly World Wide nature of the Web and the difficulties in separating time zone and time of day issues in traffic monitoring analysis. As is the case with Figure 14, a graph such as Figure 15 can aid site administrators in deciding when to schedule special events in order to avoid or capitalize upon especially high traffic periods for a given site.

Individual Visitor Tracking

Previous sections of this report described data about site traffic trends of large numbers of visitors considered en masse. It is not feasible to consider, and to report on, all of the details of visitors' recorded behaviors for large numbers of visitors at once. It is possible, however, to more closely examine

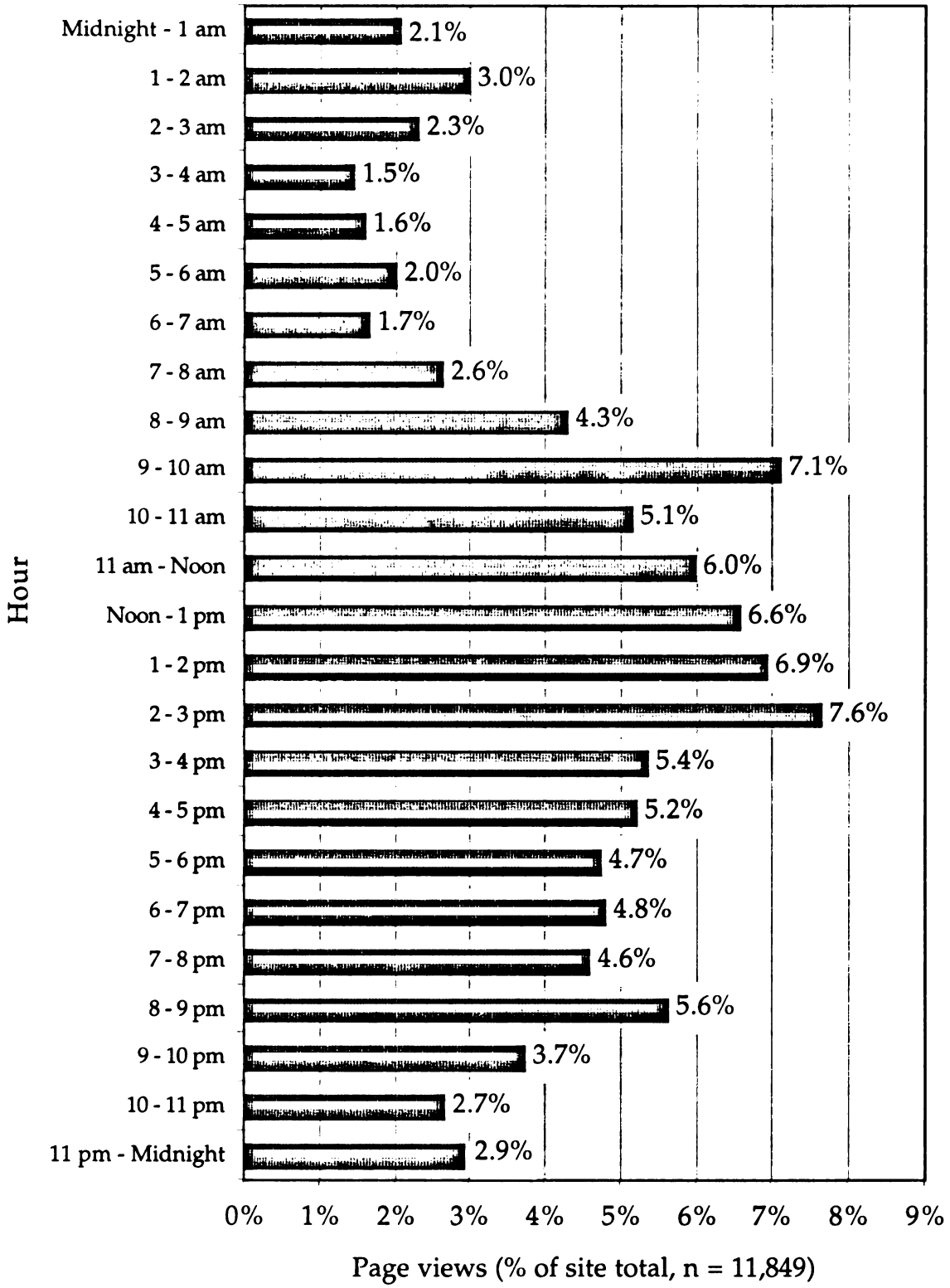

Figure 15 - Page Views by Hour of the Day (case study week)

behaviors of smaller numbers of visitors. In this section, a subset of the visitors who looked at the DLC-ME Web site during the case study week are subjected to more careful scrutiny. This investigation is roughly equivalent to the individual visitor tracking techniques used by museum visitor studies researchers. This section describes the extent to which behaviors can be determined or inferred, limitations to these visitor tracking techniques, and methods and data required for detailed visitor tracking.

Selection of visitors for intensive tracking. I chose two distinct subsets of the 2,004 unique network addresses which visited the DLC-ME Web site during the case study week to examine in closer detail. I selected addresses that logged seven page views during the week as the first subset, and addresses that logged thirteen page views as the second subset. I decided that visits which consisted of too few page views could not illustrate complex behaviors, and that I should therefore rule out such addresses. I also determined that addresses which logged too many page views could include cases likely to complicate analysis, such as multiple repeat visits by the same visitor during the week or visits by multiple users with a shared network addresses, and therefore removed them from consideration. I wanted to choose a well-bounded group of addresses, and decided that all addresses with a given number of page views would define such a grouping. I wanted to choose a page view count with a large enough number of addresses to represent a variety of behaviors, and upon which statistical tests could be reliably used.

I used Figure 8, the distribution of visitor counts by page views per visitor, to select a case to study. The seven page view visitors category logged 63 visitors, nearly as many as the six page view group (which had 69) and several more than the eight page view category (which had 38). A seven page visit

could include fairly complex behaviors, but is not such a large number that it would be an unlikely single visit tally. One goal of my tracking analyses is to measure the duration, in minutes, of visits. As explained in a later section, the time spent on the last page viewed by a given visitor is not measurable, so duration data for a seven page visit includes only the first six pages viewed. I chose to examine thirteen page view addresses as my second study group, partially because the twelve pages for which time data can be know is exactly twice the count for which such data is available for seven page view visitors. If the average "time on page" for visitors remains constant across visit durations, measurable visit durations for thirteen page view visits would be exactly twice that for seven page view visits. This choice of groups for study readily supports exploration of such visit duration conjectures.

Data used for intensive tracking. Four types of data recorded in the Web server's log file are relevant to the visitor tracking techniques I employed. These four data types are contained in five fields in the log file records. The four germane data types are the time a file request was logged, the URL of the requested file, the hostname of the client requesting a file, and the referrer address URL.

Data about the time a file request was logged is recorded in two fields: time of day and date. Both are in terms of the time zone of the Web server, not of the client making the request. Data about the URL of the requested file provides a record of the Web page which was "visited." The hostname field records the network address of the computer being used by the visitor, either as a numerical IP address or as a more descriptive alphanumeric DNS address. The referrer field provides data about where on the Web a visitor "was coming from," or which site and page the visitor was "at" immediately before coming to the requested page. This field is often blank, providing no

referring page data. Sometimes the referrer field includes data beyond which page a visitor "arrived from," such as the search keywords the visitor used if he or she reached the site via a search engine, or "anchor" tags which describe specific locations within a Web page that the visitor was viewing.

Defining visits and visit duration. An ideal definition of a "visit" to a site would describe a series of uninterrupted page views during which the visitor's attention was continuously engaged. In practice, indirect observation of visitor behavior via server log records does not allow us to distinguish between visitors who spend a long time reading a page and those who were interrupted by a phone call or went off to get a cup of coffee. In this study, I have used a commonly employed "rule of thumb" criterion for deciding what constitutes a visit (Buchanan & Lukaszewski, 1997; Lee, 1996a). A visit is a series of page requests by a single network address without a pause, or timeout, lasting 30 minutes or more between successive requests.

Time and date information is recorded in a server log file when a page is requested. In effect, this information tells one when a visitor arrives at a page, but not when she departs. A sequence of page views at a single site allows one to infer the amount of time spent on a page. The arrival time at the second page visited can be equated with the departure time from the first page visited. When a visitor leaves the site being studied, by going to another site, quitting from her Web browser, or so on, there is no record in the log file of the departure time. There is no information in the log file, therefore, of the duration of a visitor's "stay" on the last page viewed during a visit. This limitation influences attempts to measure the duration of a visit in terms of time (as opposed to in terms of pages viewed) and attempts to gage a visitor's average "time on page." This influence causes the greatest uncertainty for visits consisting of few page views, and lesser uncertainty for longer visits.

The duration, in terms of time, of a seven page view visit can only be measured for the first six pages; the duration of a thirteen page view visit measures only the time on the first twelve pages.

Seven page view visitors. The report generated by the server log analysis program listed 63 network addresses that logged seven page views each during the case study week. To facilitate examination of the details of usage patterns by these 63 visitors, I imported the server log file, which is in the form of a tab-delimited text file, into a database program. I "cleaned" the database by removing all hits that were not page views by searching for and deleting all hits that did not end with ".html". I also removed all records that were requests for files that are not part of the DLC-ME site, since the server which hosts the DLC-ME site also hosts other sites run by the same lab. I then used each of the network addresses for seven page view visitors, gleaned from the log analysis report, to search the database for the pertinent records. This approach enabled me to examine the full server log record for each page view for each of the seven page view visitors in detail.

Of the 63 seven page view "visitors," 51 (81 percent) were records of single visits. The remaining twelve sets of records represented multiple visits, usually on separate days, but in some cases on the same day at different times.

I noted the file request times for the first page viewed by each of the single visit visitors, and for the last page viewed. The difference between those times is the visit duration, recalling that only the total time spent on the first six of the seven pages viewed during the visit is measurable. The mean visit duration for the 51 single visit visitors was eight minutes six seconds. For the six pages for which time data can be known of a seven page visit, that equates to an average "time on page" of one minute twenty-one seconds. However, the variation among visit durations for those 51 visitors was very large. The

standard deviation of visit durations was six minutes seventeen seconds, a substantial fraction of the slightly over eight minutes mean visit duration. Table 7 shows a summary of these values for both the seven page view visits and the thirteen page view visits, which are described later.

Figure 16 shows the distribution of visit durations for the 51 single visit seven page view visitors. Visits lasting up to 60 seconds are grouped together in the one minute visits category, visits lasting 61 to 120 seconds are in the two minute category, and so on. Recall that these visit duration values cover only the measurable first six pages visited. The majority of the visits lasted ten minutes or less. The two long visits, at 27 and 28 minutes, seem likely outlier candidates which one might wish to exclude from statistical analyses. Presumably, those visitors were not continuously looking at just six pages for such long times, but were apparently engaged in other activities in the midst of their visits to the DLC-ME site. We do not, however, have any direct evidence to support such a claim. The nine visitors who spent between 12 and 21 minutes at the DLC-ME site present us with a more difficult interpretation problem. It is far from clear which, if any of those visitors should be dismissed as outliers. If some were removed from consideration for statistical analyses, the decision as to where the cut-off point should be placed would not be straightforward.

The first page a visitor views during a visit can be thought of as a front door or entryway to a site, which is an especially apt analogy when using museum visitor studies as a model for Web visitor studies. Though most sites are designed with a single front door in the form of a home page, the increasing use of search engines generates a large number of alternative entrances in terms of the ways many visitors arrive at a site. There were 22 different entry pages that the 51 single visit seven page view visitors to the

Table 7 - Summary Data for 7 and 13 Page View Visitors

Visit duration (page views)	7	13
"Visitors" (network addresses)	63	22
Single visit addresses (count)	51	17
Single visit addresses (percent)	81%	77%
Mean visit duration	8 min. 6 sec.	14 min. 1 sec.
Visit duration std. deviation	6 min. 17 sec.	11 min. 53 sec.
Mean time on page	1 min. 21 sec.	1 min. 10 sec.

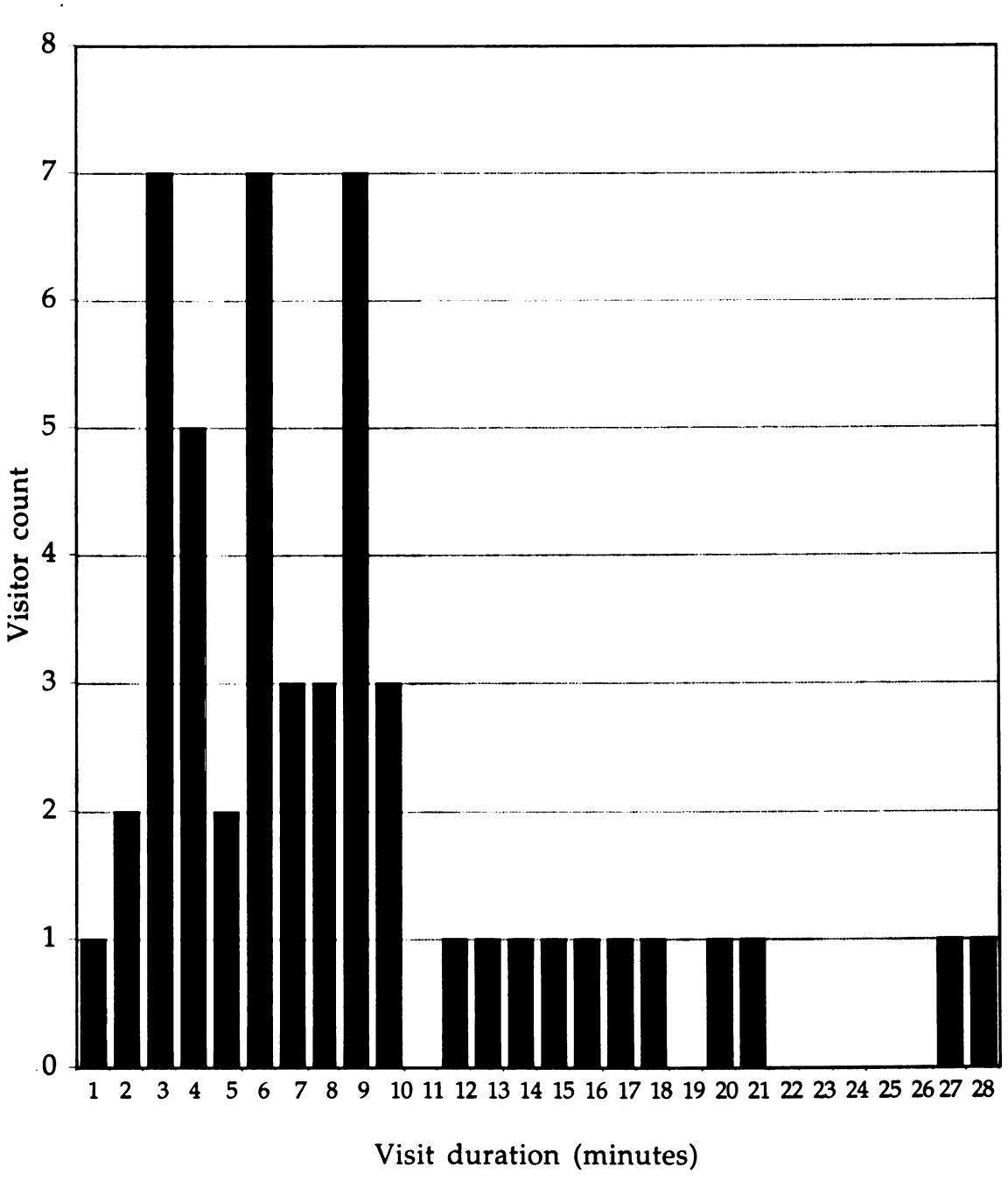

Figure 16 - Visit Duration Distribution (7 page view visitors)

DLC-ME site first arrived at. Table 8 lists the breakdown by site sections and by pages of these visitors' site entry points. Such data, combined with the extremely limited number of pages which most visitors view, can provide developers insights into what a user's experience of a site is like. The visitor's experience may be dramatically different than what one might imagine if one assumed all visitors arrived via the site's home page.

The referrer data field in the server log records provides information about where on the Web visitors were before coming to the site being evaluated. Referrer data for the site entry pages enables determination of which external sites and pages "led" visitors to the DLC-ME site. Table 9 provides a summary of entry point external referrer data for the single visit seven page view visitors. Nine of the visitors' records had no referrer data in the entry point listing, so we cannot tell what led them to the DLC-ME site. Over forty percent of the visitors "arrived" via search or directory sites; "Yahoo!" was the most frequently used gateway.

The referrer field data for directory sites tells us not only which site visitors arrived from, but sometimes also provides information about the topic category hierarchy within the directory which the visitor found a DLC-ME listing. For example, the full listing of one of the "Yahoo!" referrer entries was "www.yahoo.com/text/Science/Biology/Education/K_12/". Many of the individual search and directory site listings which are grouped together in Table 9 represent various referrer addresses. In some cases, especially those search engines which have licensed their technologies to other groups, these referrer addresses may represent different unrelated sites. Grouping of referrer addresses into clusters that represent pages which are indeed part of a single search site or are otherwise logically affiliated is straightforward in some cases, more tentative in others.

Table 8 - Site Entry Points for 7 Page View Visitors

Entry section	Visitors (count)	Visitors (percent)
Microbe Zoo	24	47%
DLC-ME home page	12	24%
Microbes in the News	9	18%
Microbial Ecology Resources	3	6%
others	3	6%

Entry page	Visitors (count)	Visitors (percent)
DLC-ME home page	12	24%
Microbe Zoo home page	10	20%
Pond (in Water World in Microbe Zoo)	4	8%
Frequent Flyers (in Space Adventure in Zoo)	3	6%
Spirogyra (Zoo specimen in Pond, animated)	3	6%
What is the DLC-ME?	2	4%
Ancient bacterium from amber (News article)	2	4%
other pages (1 visitor each)	15	29%
Total visitors	51	100%

Table 9 - External Referrers (entry points for 7 page view visitors)

External referrer site	Visitor count
no data	9
Search/directory sites	(22 total)
Yahoo!	8
Lycos	6
Webcrawler	3
AltaVista	2
Infoseek	2
Metacrawler	1
Cells Alive	5
Comm Tech Lab home page	2
Center for Microbial Ecology site	1
www.gene.com	2
other	10

Table 10 - Search Keywords Included in Search Site Referrers

Word or term	Number of occurrences
microbe(s)/microbial	5
bacteria	4
mold	3
spirogyra	3
fungus/fung	2
viruses	2
algae, protists, rotifer, spores	1 each

Other external referrers listed in Table 9 include Web sites concerned with subject matter similar to that covered by the DLC-ME, sites with organizational affiliations to the DLC-ME project, and an assortment of sites that defy ready classification. "Cells Alive" and "www.gene.com" are concerned with microbiology-related topics, as is the DLC-ME. The Comm Tech Lab is the multimedia research and development lab at Michigan State University where most of DLC-ME project development was done; the Center for Microbial Ecology, also at M.S.U., was the content expertise affiliate organization behind the DLC-ME. Sites, such as these four, often have pages with listings of "links to related sites" which frequently "steer" users to specific Web "locales." Site developers and evaluators might wish to know how visitors found their way to a site, in order to make informed decisions about site dissemination or publicity planning.

Referrer field data from search sites sometimes contains information about the keyword terms a user was searching for. For example, the following referrer field entry from an AltaVista search:

> www.altavista.digital.com/cgi-bin/query?pg=aq&what=web&fmt=
> .&q=fungus+AND+Mold&r=&d0=&d1=

indicates the user was searching for information relating to the terms "fungus" and "mold." Table 10 lists identifiable keywords or keyword fragments which were included in referrer strings from search sites used by seven page view visitors. Such data could help site developers understand what types of information visitors were seeking at a site, and what terminology those visitors were using to refer to concepts in which they were interested.

The next few pages of this report describe detailed examination of the entire visit records of a few seven page view visitors. These cases illustrate

some of the visitor behaviors that can be inferred from such data, and also show some of the limitations to such analyses and potential pitfalls inherent in cursory examination of the data. For each of these cases I have assembled a table which presents each of the seven page view records which comprise the visit. Each record includes the date and time of the file requests, the URL of each requested page, and the referrer field data associated with each request.

Table 11 presents the data affiliated with the first of these detailed seven page view visit analyses. This visit, which occurred on February 4th, lasted precisely six minutes, excluding the unknown time spent on the last page visited. The third column of Table 11, labeled "URL," lists the sequence of pages visited. This visitor viewed seven different pages in the "Microbes in the News" section of the DLC-ME site, identified by the ":news" directory identifier character string portion of each URL listing. This visitor viewed five separate news article summary pages: ns295dis1, ns994tim1, ns395dis3, ns000nyt1, and ns395sn3. The visitor also viewed two Microbes in the News subsection menu pages: ncdangerous and ncstrange.

Referrer data, in column four of Table 11, indicates the Web page the visitor was viewing immediately before "going to" each page (URL) in this listing. This visitor "arrived at" the first page of this visit from the United Kingdom "Yahoo!" directory site; specifically the "Archaea" subsection of the "Genetics" subsection of the "Molecular Biology" subsection of the "Biology" subsection of the "Science" section of that directory. The referrer for the second page visited is the same as the URL of the first page visited, indicating this visitor "went from" the page "ns295dis1" to the page "ncdangerous." This pattern of the current page's referrer matching the previous page visited (ncdangerous) is repeated in the case of the third page viewed during this visit.

Table 11 - Detailed Record of a Seven Page View Visit (number 1)

Date	Time	URL	Referrer
2/4/97	22:34:33	:CTLProjects:dlc-me: news:ns295dis1.html	www.yahoo.co.uk/Science/ Biology/Molecular_Biology/ Genetics/Archaea/
2/4/97	22:35:58	:CTLProjects:dlc-me: news:ncdangerous.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ns295dis1.html
2/4/97	22:36:23	:CTLProjects:dlc-me: news:ns994tim1.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ncdangerous.html
2/4/97	22:39:00	:CTLProjects:dlc-me: news:ncstrange.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ns295dis1.html
2/4/97	22:39:13	:CTLProjects:dlc-me: news:ns395dis3.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ncstrange.html
2/4/97	22:39:48	:CTLProjects:dlc-me: news:ns000nyt1.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ncstrange.html
2/4/97	22:40:33	:CTLProjects:dlc-me: news:ns395sn3.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ncstrange.html

The fourth page this visitor viewed, however, breaks this straightforward trend. The referrer (ns295dis1) for the fourth page visited (ncstrange) does not match the third page visited (ns994tim1). If we had only looked at the sequence of URLs for this visit, we would have probably concluded that the sequence of pages this visitor viewed was:

ns295dis1 —> ncdangerous —> ns994tim1 —> ncstrange —> etc.

However, including the referrer field data in this analysis allows us to realize that the page visit sequence could be better described as:

ns295dis1 —> ncdangerous —> ns994tim1 —> ?

? —> ns295dis1 —> ncstrange —> etc.

The simplest explanation of this event is that the visitor returned to one of the Microbes in the News pages previously visited before proceeding to the "ncstrange" page. If the previously visited page was still stored in the visitor's browser's cache, the page would be retrieved locally from the cache, and no file request would be sent to the Web server. The visitor may have simply returned to "ns295dis1" from "ns994tim1," possibly using the browser's "go recent" menu feature, before proceeding to "ncstrange." The referrer data for the "ncstrange" page request record tells us which page the visitor came to the "ncstrange" page from. We do not, however, know whether the visitor went to other pages, either previously cached DLC-ME pages or pages at other sites, between viewing "ns994tim1" and returning to "ns295dis1." The 2 minute 37 second delay between the request for "ns994tim1" and the request for "ncstrange" seems to indicate that any such "side excursion" was not very lengthy, but provides no definitive clues about where the visitor went during the unrecorded portion of this visit.

The referrer data fields for the last two records of this visit also indicate interruptions in the server log's record of the sequence of pages included in

this visit. These three gaps indicate that what initially appeared to be a seven page view visit actually consisted of a sequence of ten or more page views. Both the increased duration of this visit, in terms of pages visited, and the uncertainty regarding the number of pages in the sequence call into question the validity of using the visit duration data, in terms of time, affiliated with this visit. Since the number of actual pages in the sequence is unknown, the "average time per page" figure extracted from that value and the visit duration time data is inaccurate. Visitor page view sequences that involve return visits to previously viewed pages also raise the issue of whether analyses should distinguish between page visits and page revisits when interpreting user behaviors and applicable statistics. It seems likely that visitors might spend a longer time viewing a page the first time they see it that when they return to it, especially if that page serves as a menu for accessing other pages. Revisited pages might also appear on users' screens much more rapidly, if cached, than new pages, especially if the users' network connections were slow. "Time on page," as measured using server log records, includes both time spent viewing and time spent waiting for file transfers.

Table 12 shows the page view records for another hostname that logged seven page views during the case study week. Examination of the date and time fields reveals that this record includes four distinct visits. The hostname associated with these page views, "www-aj2.proxy.aol.com", is apparently one of the proxy servers assigned to dial-up users of America Online. This record, therefore, may represent as many as four distinct users. Each of the four visits in Table 12 begins with a different external referrer, which may be further evidence that these visit records were generated by distinct users, since they came to the DLC-ME site from different Web "locations." Repeat visits by a

Table 12 - Detailed Record of a Seven Page View Visit (number 2)

Date	Time	URL	Referrer
2/4/97	16:32:00	:CTLProjects:dlc-me: resources:rv_5.html	altavista.digital.com/ cgi-bin/query
2/5/97	01:51:19	:CTLProjects:dlc-me: zoo:index.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ index.html
2/5/97	01:54:59	:CTLProjects:dlc-me: zoo:zamain.html	commtechlab.msu.edu/ CTLProjects/dlc-me/zoo/
2/5/97	01:57:51	:CTLProjects:dlc-me: zoo:zsmain.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ zoo/zamain.html
2/6/97	16:39:30	:CTLProjects:dlc-me: zoo:index.html	sln.fi.edu/qa96/ spotlight12/spotlight12.html
2/6/97	16:40:09	:CTLProjects:dlc-me: zoo:index.html	commtechlab.msu.edu/ CTLProjects/dlc-me/zoo/
2/6/97	23:55:54	:CTLprojects:dlc-me: zoo:zwpmain.html	webcrawler.com/cgi-bin/ WebQuery

single visitor would probably share a common entry point. This case shows that page view counts which are grouped by hostnames in reports generated by log analysis programs need not represent single visits or even individual users. Site evaluators must look beyond such reports to the details of server log records to ferret out such information.

Table 13 shows a type of visitor behavior that a site evaluator might not intuitively expect based on preconceived notions of how users normally browse the Web. The seven page views listed in Table 13 fit the "no 30 minute timeouts" rule for defining a single visit, and were all generated by a single hostname. The times between file requests for this "visit" exhibit two very distinct patterns; some were very short, while others were quite long. The apparent time on page values, in order, were: 28 seconds, 31 seconds, over 13 minutes, 33 seconds, almost 6 minutes, and 6 seconds. The 13 minute stay on the third page visited strains the credibility of using a 30 minute timeout as the litmus test for classifying single visits, especially in light of the brief durations of the other "time on page" values for this visit. The 13 minute and the six second time on page values clearly indicate unusual visitor behaviors. In the former case, it seems likely that the user was in some way interrupted from, or chose to take a break from, her or his Web browsing activities. Interpretation of the extremely short page visit value requires examination of the URL column of Table 13 and knowledge of the nature of a particular page on the DLC-ME Web site.

The URL column of Table 13 reveals that this visit consisted of seven visits to a single page, "zwpspiro," in the Microbe Zoo section of the DLC-ME site. That page has a short animation showing the internal structure of a Spirogyra which plays once when the page is first loaded. This visitor apparently returned to the same page seven times in order to view that

Table 13 - Detailed Record of a Seven Page View Visit (number 3)

Date	Time	URL	Referrer
2/3/97	10:33:43	:CTLProjects:dlc-me: zoo:zwpspiro.html	www-csi.lycos.com/cgi-bin/ pursuit?cat=lycos&query= spirogyra&x=41&y=5
2/3/97	10:34:11	:CTLProjects:dlc-me: zoo:zwpspiro.html	www-csi.lycos.com/cgi-bin/ pursuit?cat=lycos&query= spirogyra&x=41&y=5
2/3/97	10:34:42	:CTLProjects:dlc-me: zoo:zwpspiro.html	www-csi.lycos.com/cgi-bin/ pursuit?cat=lycos&query= spirogyra&x=41&y=5
2/3/97	10:47:45	:CTLProjects:dlc-me: zoo:zwpspiro.html	www-csi.lycos.com/cgi-bin/ pursuit?cat=lycos&query= spirogyra&x=26&y=1
2/3/97	10:48:18	:CTLProjects:dlc-me: zoo:zwpspiro.html	www-csi.lycos.com/cgi-bin/ pursuit?cat=lycos&query= spirogyra&x=26&y=1
2/3/97	10:54:04	:CTLProjects:dlc-me: zoo:zwpspiro.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ zoo/zwpspiro.html
2/3/97	10:54:10	:CTLProjects:dlc-me: zoo:zwpspiro.html	www-csi.lycos.com/cgi-bin/ pursuit?cat=lycos&query= spirogyra&x=26&y=1

animation repeatedly. In the case of each of the first three page views, the referrer for each is the same character string indicating the visitor searched for the term "spirogyra" at the Lycos search site. Apparently, this visitor used her or his "go back" button on the browser client to return to Lycos and then reload the spirogyra page twice after first encountering it. The referrer string is slightly different between the first three page views and the fourth, fifth and seventh page views—the "&x=41&y=5" character string in the former was replaced by "&x=26&y=1" in the latter. This change corresponds to the long thirteen minute gap in the page viewing sequence. My guess is that this visitor viewed the animation thrice, perhaps went off to browse other Web sites, and then did a fresh search on Lycos for "spirogyra" to return for more viewings of this animation.

The spirogyra animation page has instructions on it, below the animation, informing visitors that they can use the "reload" button on their Web browser to reload the Web page and thus replay the opening animation. It seems that this visitor did so on the sixth page view in this sequence, where the URL and the referrer field both contain the address of the spirogyra page. Why didn't this visitor use the reload button in the other cases? Perhaps she or he didn't read the text, or was unfamiliar with browser operation and unaware of the reload button. Most of my description of possible user behavior regarding this case is admittedly speculative, but is also reasonably plausible. This case illustrates extremely atypical user browsing behaviors, the degree to which careful scrutiny of log records and knowledge of the details of a site can enable formulation of possible explanations of such unusual behaviors, and the lack of certainty that such explanations entail.

Table 14 shows data which is more in line with common views about a typical visit to a Web site. In this case, all of the referrers match the URLs of

Table 14 - Detailed Record of a Seven Page View Visit (number 4)

Date	Time	URL	Referrer
2/3/97	00:25:04	:CTLProjects:dlc-me: news:ns595ap1.html	www.search.com/AltaVista/ 1,57,0,00.html?mode= simple&query=bacteria &what=web&format=2
2/3/97	00:26:38	:ctlprojects:dlc-me: index.html	commtechlab.msu.edu/ CTLProjects/dlc-me/ news/ns595ap1.html
2/3/97	00:26:55	:CTLProjects:dlc-me: zoo:index.html	commtechlab.msu.edu/ ctlprojects/dlc-me/
2/3/97	00:28:17	:CTLProjects:dlc-me: zoo:zdmain.html	commtechlab.msu.edu/ CTLProjects/dlc-me/zoo/
2/3/97	00:29:40	:CTLProjects:dlc-me: zoo:zdtmain.html	commtechlab.msu.edu/ CTLProjects/dlc-me/zoo/ zdmain.html
2/3/97	00:31:53	:CTLProjects:dlc-me: zoo:zamain.html	commtechlab.msu.edu/ CTLProjects/dlc-me/zoo/ zdtmain.html
2/3/97	00:32:55	:CTLProjects:dlc-me: zoo:zapmain.html	commtechlab.msu.edu/ CTLProjects/dlc-me/zoo/ zamain.html

the preceding page view, apparently indicating that there are no "holes" in the data record. The "time on page" values seem reasonable, ranging from about thirty seconds to around two and one half minutes. This visitor used a "meta" search site (www.search.com) to access the AltaVista search engine, using "bacteria" as a search keyword. She or he discovered a news article page (ns595ap1.html) in the Microbes in the News section of the DLC-ME site, proceeded to the site's home page, and then went on to explore some of the Microbe Zoo pages. Table 14 displays, as best we can determine from a server log file record of the event, the full set of pathway data for a typical seven page view visit to the DLC-ME site.

Thirteen page view visitors. The report generated by the server log analysis program listed 22 network addresses that logged thirteen page views each during the case study week. Of those 22 thirteen page view "visitors," 17 (77 percent) were records of single visits. The mean visit duration for the 17 single visit visitors was fourteen minutes one second. Recall that this visit duration spans only the first twelve pages of the visit, as the departure time from the last page visited is unknown. For the twelve pages for which time data is known for these visits, the average "time on page" was one minute ten seconds. As was the case for the seven page view visits, the variation among visit durations for these 17 visitors was large in comparison to the mean visit durations. The standard variance of visit durations was eleven minutes fifty-three seconds. Table 7 shows a summary of these values.

Figure 17 shows the distribution of visit durations for 16 of the 17 single visit thirteen page view visitors. As was the case in Figure 16, visits lasting up to 60 seconds are grouped together in the one minute visits category, visits lasting 61 to 120 seconds are in the two minute category, and so on. One extreme outlier was left out of Figure 17; one of the visits lasted nearly 55

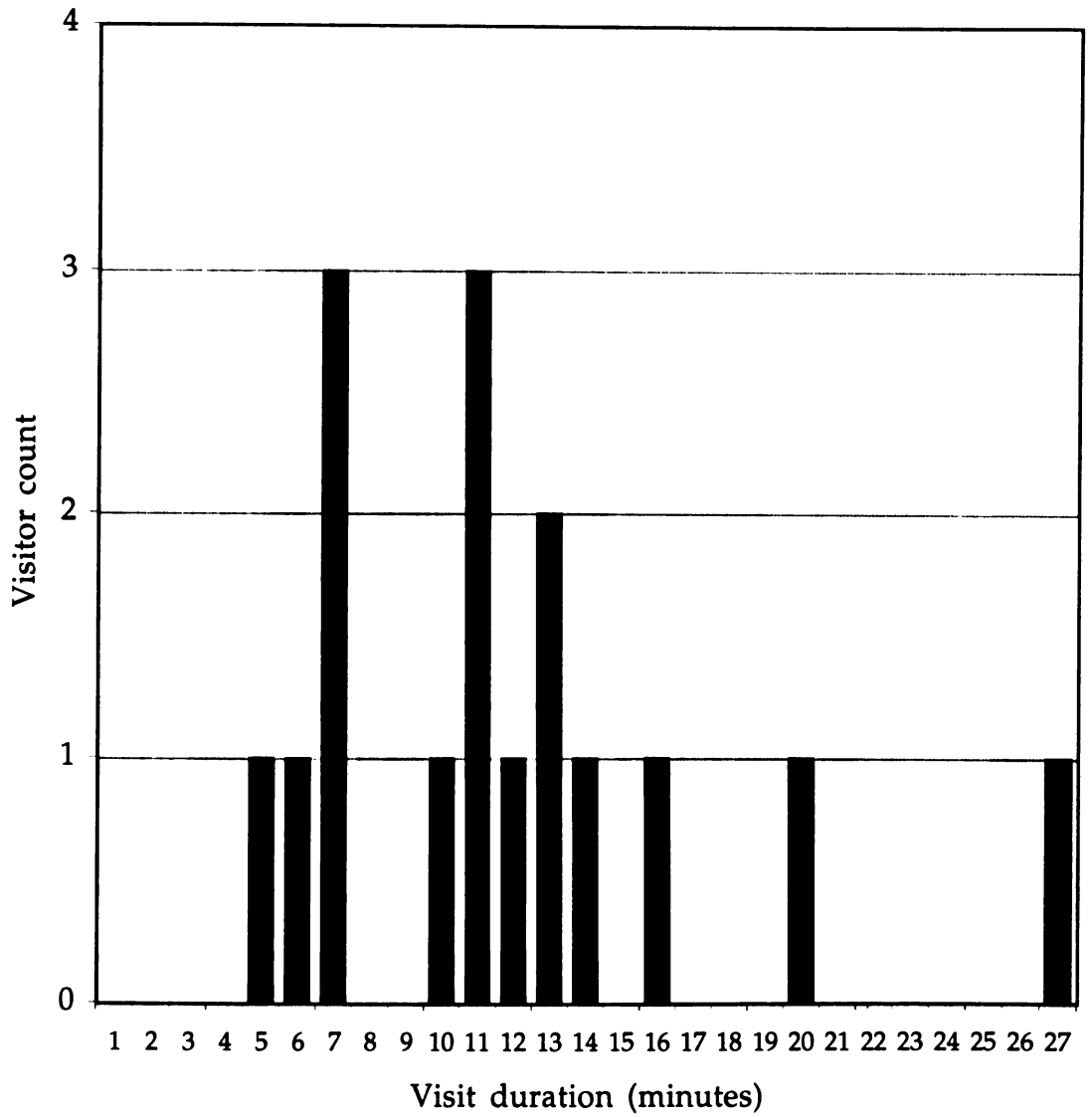

Figure 17 - Visit Duration Distribution (13 Page View Visitors)

minutes without violating the 30 minute timeout criterion for considering it a single visit.

Comparing seven and thirteen page view visitors. The sample sizes of the groups of seven page view and thirteen page view visitors examined in the preceding sections are quite small, so any bold pronouncements about the significance of patterns found in these data sets would be inappropriate. The detailed examination of some of the "seven page view" visits also made apparent the wide range of visitor behaviors which can appear in server log records, and showed that some such visits were not actually simple sequences of seven page views at all. Many of the seven and thirteen page view visits described above are probably actually records of visits of other page view durations "in disguise." A statistically valid analysis of actual seven and thirteen page view visit trends would need to start with a much larger pool of records, given how many records would have to be thrown out and given the wide range of visit behavior types that would likely be represented. However, a couple of interesting patterns worth noting are apparent in the current limited data sets.

Table 7 shows a summary of the major traits of the two data sets examined in this report. The percentage of network addresses which were records of single site visits were quite similar for the seven page view (81 percent) and the thirteen page view (77 percent) visits. The mean time on page for each of the two visit duration groups were also remarkably similar. However, in each case the standard deviation of visit durations was quite large, casting doubt upon the reliability of projecting trends in mean visit durations onto large visit data populations. The standard deviation of durations of the seven page view visits was nearly 78 percent as large as the mean of visit durations; the standard deviation of durations for the thirteen page view visits was almost

85 percent of the mean of visit durations.

The Microbes on Mars Incident

This section describes an anomalous surge in DLC-ME site traffic which began during August 1996, and some comments about the likely causes of that traffic increase. That event, which I have dubbed the "Microbes on Mars incident," illustrates the effect a major news story can have on a Web site which covers related topics. Investigation of that incident also sheds light on the relationship between search sites and content-based Web sites, and on how the details of such an event may be gradually discovered via increasingly focused investigation of the server log record. This section is presented as a roughly chronological narrative, in an attempt to give the reader a flavor of the mysteries and revelations the "Microbes on Mars incident" offered.

Discovery of an anomaly. Figure 18 shows the page view counts for the entire DLC-ME site from the November 1995 through July 1996. Site traffic levels had grown steadily up until April 1996. That growth trend had dropped off during the late spring and summer of 1996. We assumed that student absence from schools during the summer recess was the main cause of the curtailed level of site traffic. A move from indoor computer use to outdoor activities as the weather turned fairer in late spring, or other changes in student activities associated with the end of the school year, may have been responsible for the start of the page view count decline in late April and May. During the summer of 1996, we expected that site traffic would remain at suppressed levels for the remainder of the summer, and would likely pick up when school resumed in the fall.

The summer of 1996 brought some personnel changes to the staff of the

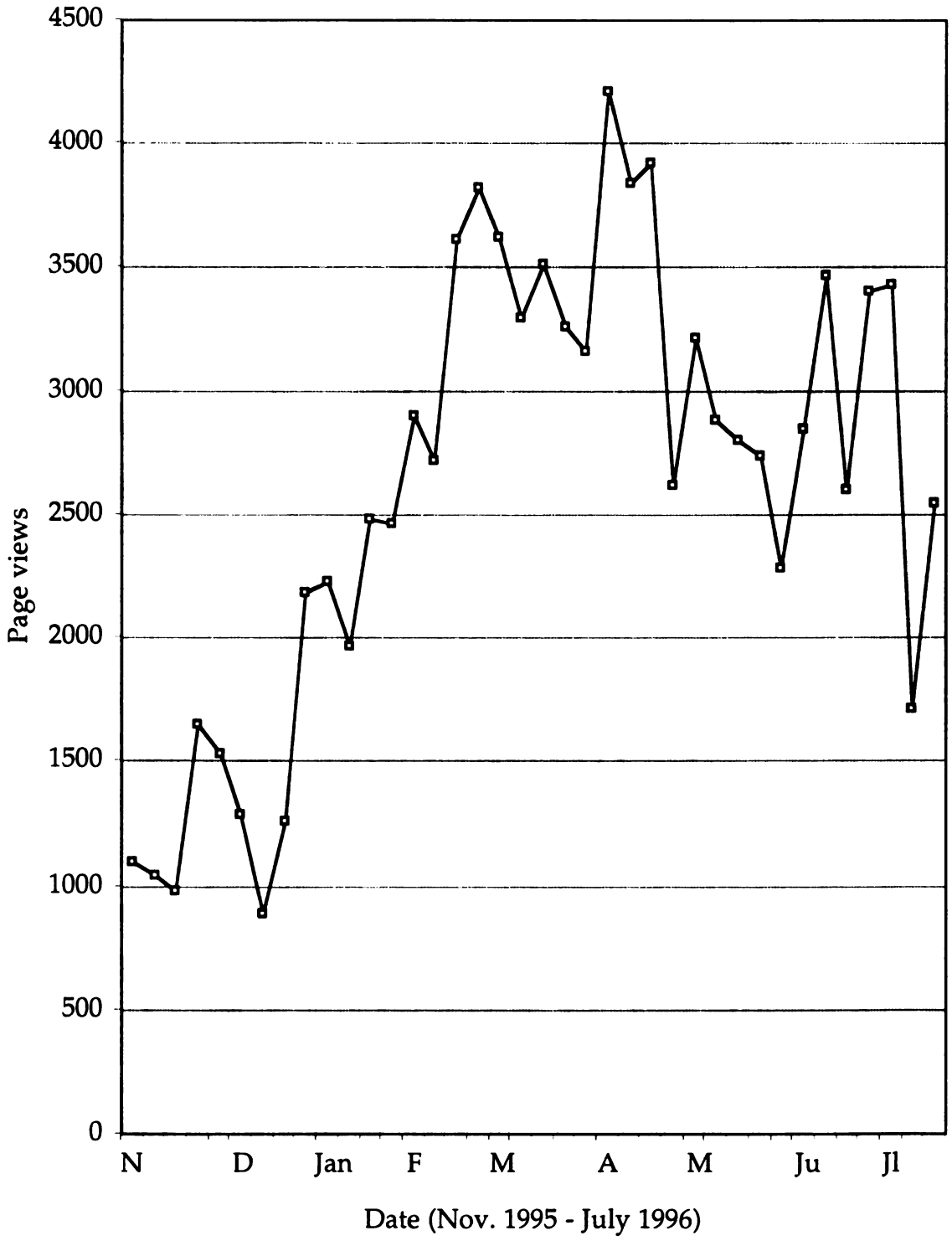

Figure 18 - Site Page Views through July 1996

lab which supports the server upon which the DLC-ME site is located. The person who had been the lab's Webmaster since the inception of the DLC-ME site moved on to a new job, and a new person assumed the Webmaster duties in the lab. Those duties included processing raw server log files, using the ServerStat log analysis software, and producing the site analysis reports. Those weekly reports were typically produced about once every one to two months using a batch processing feature of ServerStat. Thus there was typically a one or two month lag between site activity and the production of reports about that activity. Reports covering the latter half of the summer of 1996 were produced in early September. Initially, the lab's new Webmaster experienced some difficulties accurately configuring the batch processing ServerStat runs, and the first couple of report batches produced had to be re-run to correct some errors.

As I began to examine the reports from such an amended batch, I was relieved to discover that the errors had apparently been ironed out and the report generation process appeared to now be running smoothly at the hands of the new Webmaster. The figures for late June and throughout July fit the trend of decreased traffic levels that had begun in the latter part of spring. The site page view tally for the week ending August 9th, however, was surprisingly large. The page view count for that week was 6069; more than double the previous week's total of 2543 page views, and substantially greater than the largest previous weekly tally of 4212 page views from mid-April. My initial suspicion was that an error had occurred in setting the dates for ServerStat report generation for that week. Some earlier report errors resulted in reports covering periods longer than one week, which generated site page view totals exceeding common weekly totals. Closer inspection revealed that the dates on the report were accurate. Examination of the page views by file

sections portion of the report revealed that the report covered only DLC-ME pages, and had not accidentally been set to include pages from other sites hosted on the same Web server.

Examination of the next few weeks reports, covering the remainder of August, revealed a continuing dramatically increased page view count trend, especially as compared to traffic levels during the summer slump. Figure 19 shows page view counts for the entire DLC-ME site through August 1996. Total site visitor counts, in terms of unique network addresses, was the other major site traffic metric which I had been routinely monitoring. The surge in site traffic, in terms of visitors, during August was even more dramatic than the page view count increase. Figure 20 shows the site visitor count trend over the history of the site up through August 1996.

Suddenly popular page. Page view and visitor counts had suddenly and unexpectedly soared in early August 1996. Besides the increase in the sheer quantity of traffic, there were apparently changes in the types of visits people were making in terms of the number of pages viewed. As shown in Figure 7, page views per visitor had suddenly dropped from typical values of five to seven page views per visitor to around two page views per visitor. It seemed that there had been a sudden influx of visitors who were viewing just a small number of pages per visit; possibly viewing just a single page! This realization led me to examine the records of page view counts for individual pages, to see if there were specific pages these new visitors were being drawn to.

I first checked the page view counts for the site's two main menu pages, the DLC-ME site home page and the Microbe Zoo home page, for an increase in traffic levels. For the week ending August 9th, the DLC-ME home page registered 335 page views and the Microbe Zoo home page registered 291 page views. Neither value was significantly larger than the counts for the

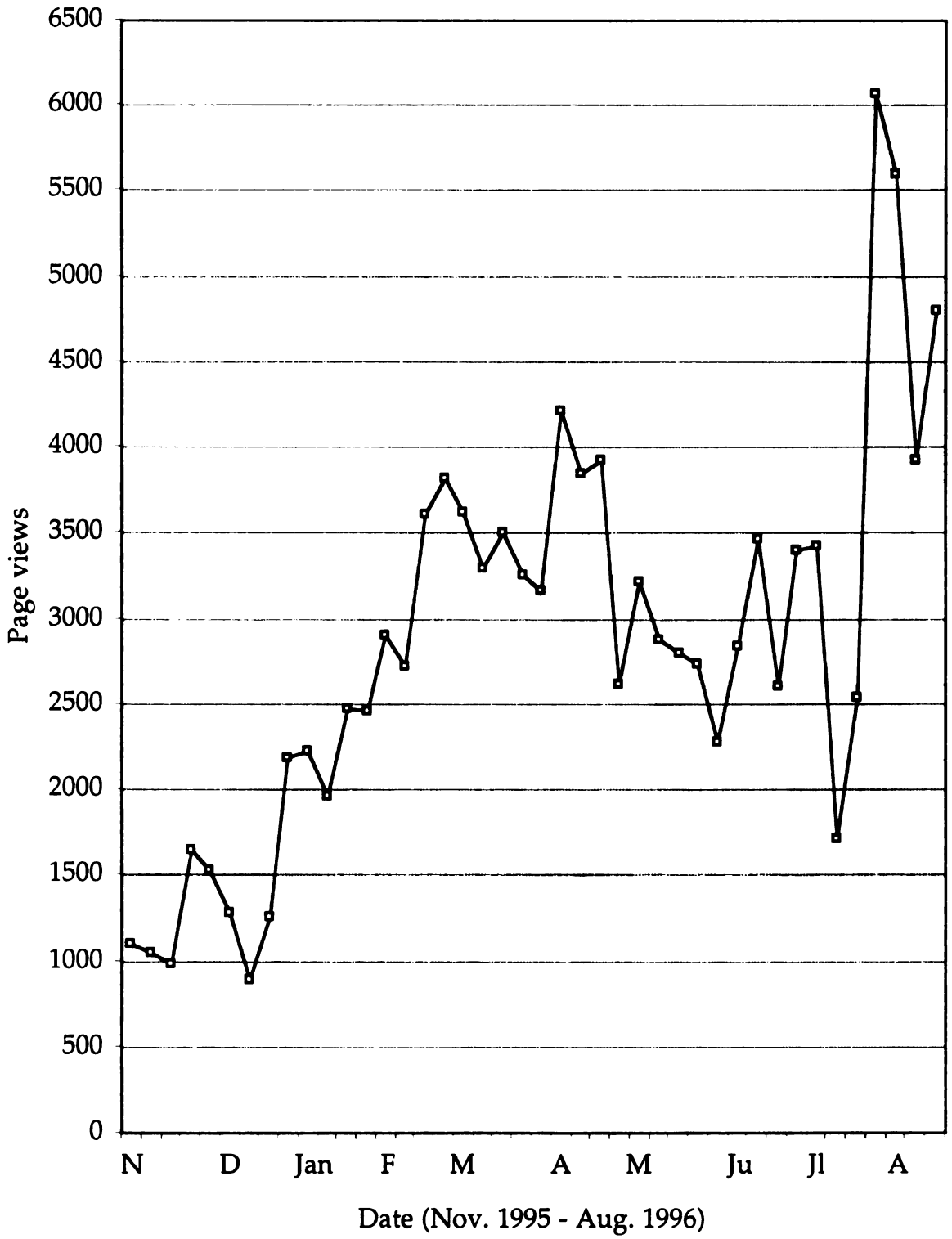

Figure 19 - Site Page Views through August 1996

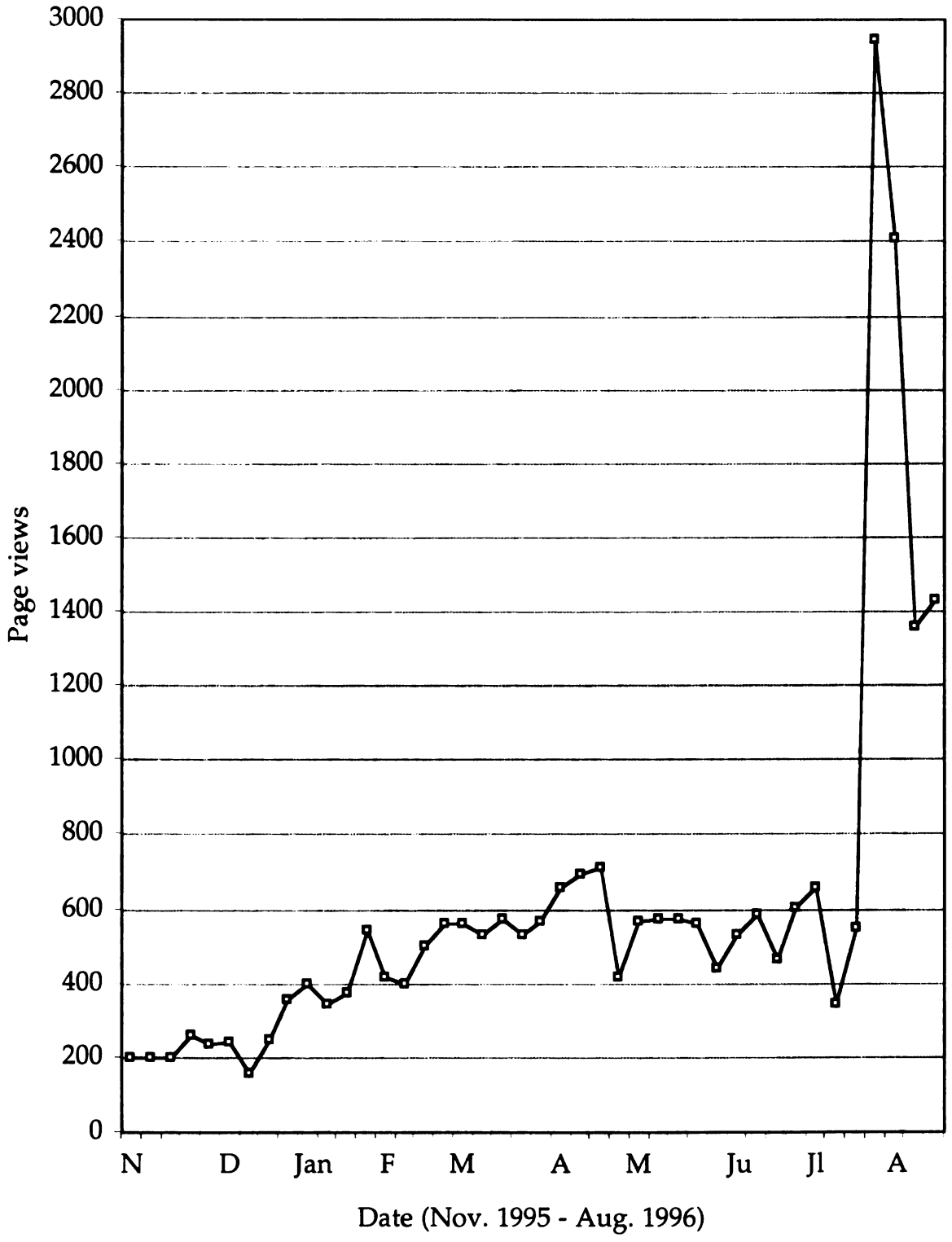

Figure 20 - Site Weekly Visitor Counts through August 1996

preceding weeks, and both fell well short of the highest counts for the busiest previous weeks, which were 612 page views for the site home page and 432 page views for the Zoo home page. The sudden influx of visitors was not, apparently, coming into the DLC-ME site through the "main front doors." A quick scan of other major section heading pages, such as the Microbes in the News home page and the main Zoo section menu pages, revealed that the increased traffic was not spread evenly throughout the site. Having determined that the DLC-ME site had not generally become suddenly more popular, I began searching for specific pages which were attracting large numbers of visitors.

The DLC-ME site is composed of approximately 200 Web pages. The listing of page views by individual pages, which is included in the ServerStat reports, is thus too large to view in a single glance, but is short enough to allow careful examination in a fairly short time. I visually scanned the page views by pages section of the report, seeking pages beyond the site's core menu pages which might account for sudden increase in traffic. Besides the site's core menu pages, few of the site's pages register more than 100 page views in a given week. It was easy, therefore, to scan the page view listings for any page view counts that were larger than double-digit values. I quickly discovered that a fairly obscure single page in the Microbe Zoo section had recorded a startling 1945 page views! The file name of that Web page, "zslmain.html", allowed me to immediately identify it as a subsection menu page in the Space Adventure section of the Microbe Zoo portion of the DLC-ME Web site.

The interest in "zslmain" had arisen very suddenly. In the two weeks preceding the week ending August 9th, the page had logged seven and eight page views, respectively. The peak page view count for "zslmain" for any

week during the five weeks prior to the sudden increase was 22. The only hypertext link to "zslmain," which is one of three subsection menu pages under the Space Adventure menu page in the Microbe Zoo, is from the Space Adventure page. Although traffic to the Space Adventure page had risen to 91 page views during the week ending August 9th from 28 page views during the preceding week, that rise was clearly far too small to account for the increased traffic arriving at "zslmain." The huge influx of new visitors was apparently not arriving at "zslmain" via the DLC-ME's and Microbe Zoo's hierarchies of menu pages.

"Microbes" and "Mars" keywords for Web searches. The Web page which the file "zslmain" underlies is titled "Microbes on Mars?", and addresses speculation by scientists about the possible existence of microbial life on other planets. On August 6, 1996, NASA scientists announced the discovery of organic compounds in a meteorite that originated from Mars and was discovered in Antarctica, and speculated that the sample might be evidence for the existence of microbial life on Mars (Wilford, 1996). The story was widely publicized in the popular press, and undoubtedly led many people to seek out information about the possible existence of microbial life on Mars from many sources, including the World Wide Web. When I realized this was the probable source of the sudden surge in DLC-ME traffic, I was able to further focus my investigation on the effects of the story upon site visitor behaviors.

When I reexamined the pages of the DLC-ME site in the context of public interest in the possibility of microbial life on Mars, I discovered that four of the site's pages made mention of Mars. Two pages, the Space Adventure section menu page (zsmain) and the "Microbes on Mars?" subsection page (zslmain), are in the Microbe Zoo. Two pages in the Microbes in the News

section, ns995ap1 and ns1095ap4, describe news articles that mention Mars. Both of those pages contain the word "Mars" at least three times, and include the exact phrase "life on Mars" within their text. I decided that all of these pages might have attracted increased visitor interest, as the "Microbes on Mars?" page in the Zoo had, and examined the ServerStat reports to see what the page view counts for each of those pages had been. Table 15 shows the page view counts for each of those Mars-related pages, starting a few weeks before the NASA announcement and running up until just before the year-end holidays traffic drop-off. One of the news article pages, ns995ap1, showed an huge increase in visitor traffic similar to the surge in visits to the "Microbes on Mars?" zoo page. The other news article, ns1095ap4, like the Space Adventure page in the zoo, showed a definite but not particularly dramatic rise in traffic volume.

The "Microbes on Mars incident" illustrates several interesting aspects of Web visitor behaviors, approaches to site evaluation, the impact of major news stories on the Web, and the way search sites are assembled and used. I conducted a search via AltaVista, using "microbes" and "Mars" as keywords, in late 1996. AltaVista reported finding about 200 pages matching both words, with the "best matches," according to AltaVista's criteria, listed first. The DLC-ME "Microbes on Mars?" page topped the list, and the Microbe Zoo's Space Adventure page was the third item. I suspect that people seeking information about the NASA Martian meteorite announcement were using various search sites and similar combinations of keywords. I further suspect that many of those search results probably listed some of the Mars-related DLC-ME pages in prominent positions near the top of results reports, and that many Web users were thus led to those DLC-ME pages, causing the sudden rise in site traffic levels. Search sites, and the Webcrawler robots they employ,

Table 15 - Page View Counts for Mars-related Pages Over Time

Start date	End date	zslmain	zsmain	ns995ap1	ns1095ap4
6/29/96	7/5/96	17	29	4	9
7/6/96	7/12/96	12	33	5	7
7/13/96	7/19/96	22	49	6	9
7/20/96	7/26/96	7	15	3	4
7/27/96	8/2/96	8	28	6	7
8/3/96	8/9/96	1945	91	987	76
8/10/96	8/16/96	1254	86	973	57
8/17/96	8/23/96	573	39	271	25
8/24/96	8/30/96	326	48	226	26
8/31/96	9/6/96	268	36	176	22
9/7/96	9/13/96	240	71	152	21
9/14/96	9/20/96	129	61	100	25
9/21/96	9/27/96	149	60	150	18
9/28/96	10/4/96	162	53	143	16
10/5/96	10/11/96	126	65	146	18
10/12/96	10/18/96	188	81	147	38
10/19/96	10/25/96	116	61	106	10
10/26/96	11/1/96	159	102	128	15
11/2/96	11/8/96	172	99	227	23
11/9/96	11/15/96	153	76	296	18
11/16/96	11/22/96	199	134	198	13
11/23/96	11/29/96	90	62	142	10
11/30/96	12/6/96	66	98	196	19
12/7/96	12/13/96	120	79	181	18

typically take days to weeks to add newly created Web pages into their databases. Sudden public interest in a certain topic, channeled through search sites, would be directed towards pages which had been online well in advance of the breaking news story. New pages created to address a suddenly popular issue would not appear in search site listings until days or weeks later. We had systematically submitted the main pages of the DLC-ME site to the major search sites early in our site's history. The search sites' robots had had time to track down and catalog all of the pages in the DLC-ME, since they were linked to from the submitted pages. The DLC-ME had serendipitously staked out its turf on the major search sites with regards to the breaking "Microbes on Mars" story well in advance of the NASA announcement.

Search sites use various algorithms to rank search result pages in terms of supposed relevance to the keywords or phrases entered by a user. Some of the criteria used in ranking results include the number of times a keyword appears in the text of a page, whether a keyword appears in the page's title, how many keywords (if the user included more than one) submitted by the user appear on the page, and how near the start of the page's text a keyword appears. Pages that generate high rankings according to such criteria for certain keywords would be displayed prominently on search site results pages when a user conducts a search using those keywords, and would likely direct a lot of traffic to the Web site which those pages were a part of. Table 16 lists the four Mars-related pages in the DLC-ME site, and shows how these pages might fare in relevancy criterion rankings on search sites for searches using terms associated with the NASA Martian meteorite announcement.

Many Web site designers develop their sites based on the implicit assumption that most site visitors will enter the site via the "front door," the site's home page, and will continue their browsing of the site from that

Table 16 - Mars-related Pages Search Criteria Ranking Relevancy

Page name: Space Adventure menu (zsmain)

Mars in page title?: no

First keyword mention: "Microbes on Mars" in first line of text, words 3-5

Keywords (repetitions): microbes (11), life (6), Mars (4), "life on Mars" (1)

Page name: Microbes on Mars? (zslmain)

Mars in page title?: yes (title includes phrase "Microbes on Mars?")

First keyword mention: Martian (first word), NASA (third word), and Mars (fifth word) in first line of text; "Microbes on Mars?" is the second line of text

Keywords (repetitions): Mars (28), life (21), Martian (6), NASA (5), microbes (3)

Page name: "Robot seeks Martian microbes' cousins" news article (ns995ap1)

Mars in page title?: yes (title includes phrase "Life on Mars")

First keyword mention: "Martian microbes'" in first line of text, words 3-4

Keywords (repetitions): Mars (4), NASA (3), "life on Mars" (2), microbes (2), Martian (1), microfossils (1)

Page name: "SLiME may exist on Mars" news article (ns1095ap4)

Mars in page title?: yes

First keyword mention: Mars is the fifth word in the first line of text

Keywords (repetitions): Mars (3), "life on Mars" (1), microbial (1)

starting point. What are the most likely sequences of pages a visitor to the DLC-ME site would follow, based on such an assumed browsing behavior pattern, to reach one of the site's Mars-related content pages? A visitor "arriving" at the DLC-ME home page could follow a link from there to the Microbe Zoo home page, then proceed to the Space Adventure section of the zoo, and then move on to the "Microbes on Mars?" subsection page. Since the Microbe Zoo subsection of the DLC-ME site has been widely publicized, a visitor might also first arrive at the Microbe Zoo home page, and then follow the latter two links in the aforementioned sequence to reach the "Microbes on Mars?" page. Paths leading to the Mars-related news articles might start at the DLC-ME home page, link to the Microbes in the News home page, and then proceed to either of the Mars-related news article pages. If visitors had been entering through the site's "main front door" pages during the "Microbes on Mars incident," page view counts during that time frame for the DLC-ME home page, Microbe Zoo home page, and Microbes in the News home page should have increased dramatically, as did the page view counts of the Mars-related content pages. Table 17 shows that the page view counts of the "front door" pages rose only slightly. Apparently visitors, arriving via search sites, came directly in through "side doors" to the pages containing the content they were searching for, largely bypassing the site's main pages. This realization could have a powerful impact on how site designers construct their sites, and what assumptions they should make about which pages visitors are likely to see.

Figure 21 shows the long-term trend of weekly page view counts for the "Microbes on Mars?" page. Page view counts declined quickly from the initial peak over the course of the next several weeks, implying that intense public interest in the freshly announced prospect of life on Mars had largely waned

Table 17 - Mars-related and Site Gateway Pages Page View Trends

	Page views by dates (weeks)						
Page	7/6-7/12	7/13-7/19	7/20-7/26	7/27-8/2	8/3-8/9	8/10-8/16	8/17-8/23
DLC-ME home	292	313	148	217	335	336	270
Zoo home	282	290	134	225	291	224	198
Space Adventure	33	49	15	28	91	86	39
Microbes on Mars?	12	22	7	8	1945	1254	573
News home	88	91	63	72	103	111	108
ns995ap1	5	6	3	6	987	973	271
ns1095ap4	7	9	4	7	76	57	25

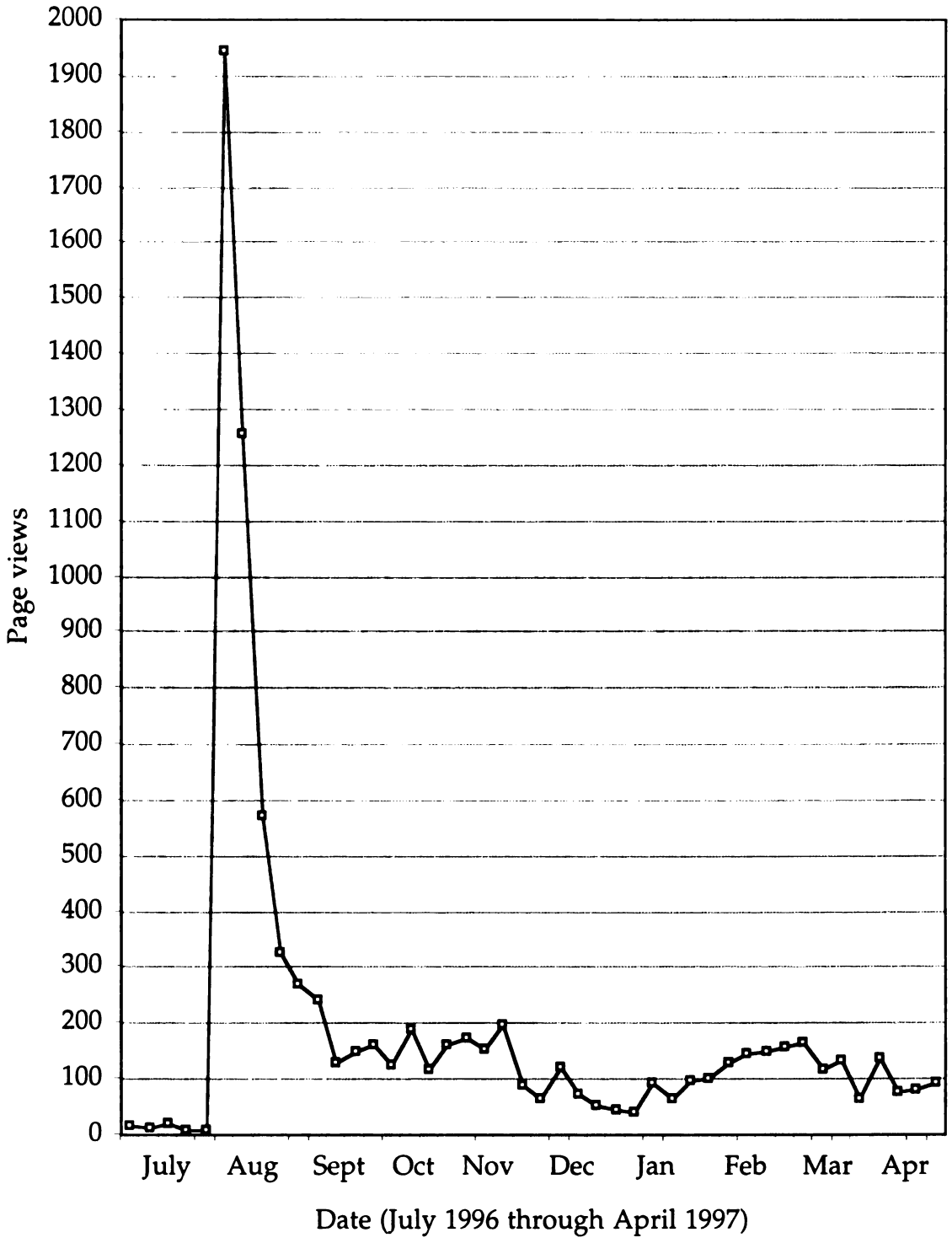

Figure 21 - Page Views for "Microbes on Mars?" Page Over Time

within a month or so after the NASA announcement. However, the weekly page view counts had not fallen back to the levels found before the announcement, even after eight months had elapsed since the news story first broke. The continued elevated level of page views for this page may reflect ongoing or recurring interest in the life on Mars issue, the general increase in traffic levels for the entire DLC-ME site (as shown in Figure 4), the continuing growth in numbers of World Wide Web users, or some combination of these factors.

Summary

In this chapter I have discussed a variety of analyses, visual displays of data from these analyses, and possible interpretations of these analyses. These analyses have included examination of Web visitors over a year and a half period, intensive analyses of a one week period, analysis of paths for 7-page view visitors, and examination of an anomalous event. The next chapter presents conclusions that may be drawn from these analyses.

Chapter Five

CONCLUSIONS

This chapter describes the information relevant to educators which can be discovered about Web site visitors and their behaviors during visits. This study uses a specific data source, the page view records from a Web server log file; all findings reported here are relevant only to use of that specific type of data. This chapter also explains some of the limitations of findings derived from the server log file, in terms of both what types of information cannot be determined and what skills and resources must be applied to an evaluation effort in order to discover various types of information. In some cases where the limitations imposed by the source of data used herein are especially problematic, and where other approaches provide clear advantages, I make brief mention of alternative means for conducting important evaluations. Finally, I describe some recommendations to Web site evaluators conducting formative evaluations for possible approaches to their task. Each of the four major sections of this chapter corresponds to one of the primary research questions posed earlier in the methods chapter of this report.

What Can Be Learned About Visitors and Visits

Analysis of server log data can help evaluators spot major trends in site traffic levels. Researchers can learn about identities of site visitors, the locations (in both geographic and cyberspace terms) visitors are coming from, which pages and sections in a site visitors are going to, and times and dates when traffic levels tend to be high and low. This approach is much better suited to broad characterizations of trends for large numbers of visitors than

for detailed examination of behaviors of specific visitors. Factors such as page caching make the record of page views incomplete, and reliance on network addresses to identify visitors makes identification of individual human users unreliable. Visitor behaviors can sometimes be deduced in great detail from log records, but such analysis is spotty, for researchers cannot control for which subjects such detailed useful data is available. Significant and previously unnoticed trends can be discovered using server log data, however; the "Microbes on Mars incident" and the realization that many DLC-ME site visitors "entered" the site at numerous points other than the site's main home pages are two important insights revealed in the course of this study.

Visitor identities. The network addresses of visitors often provide a great deal of information from which one can infer, but not be certain about, aspects of a visitor's identity. Country codes and state abbreviations in domain names imply geographic locations. Domain names also often provide clues about the institutional affiliations, such as ties to universities or school districts, of visitors. Commercial domain names of major Internet service providers also can imply facets of visitors' locations, connection bandwidths, and browser software. However, many numerical IP addresses cannot be resolved into descriptive domain names, thus preventing researchers from being able to read clues found in such addresses.

Attempts to uniquely identify visitors are confounded by several factors. Computers in labs and other public locations may be shared by multiple users. A single user may have access to more than one computer; at work and at home, for instance. Dial-up services often dynamically assign addresses of proxy servers to users, so a single user may have a different address from one visit to the next, and a single address can represent multiple users at different

times. Some "visitors" are not actually people, but Webcrawler robots scouring the Web for pages to log into search site databases. Finally, individuals have begun to use "personal robots" in the form of programs such as WebWhacker, to automatically download many pages in bulk. Such downloaded pages seem to a server's log to have been viewed once apiece, when in truth they may be viewed many times or not at all by human eyeballs.

The largest single problem with visitor identification is that it relies on the assumption that a network address is equivalent to an individual person. This limitation is inherent to, but not unique to, use of server log data as a source of information. Problems associated with this means of identification can be partially, but not completely, overcome by use of passwords and user IDs or cookie-based technologies.

Where did visitors go? Server log data does enable evaluators to discern which pages and site sections received the heaviest traffic, and which ones were infrequently visited. Site developers presumably intend certain sections of their sites to be focal points, and other pages to be of peripheral importance, so such data can aid them in understanding whether their sites are being used as intended by their design. Such data can also help designers clarify their goals with regards to traffic distribution, since most site developers probably do not set precise goals regarding optimal usage patterns. These data may encourage developers to ask themselves what the ratio of menu page versus content page page views should be, or what percentage of their visitors should visit the site's home page, if visitors are using their pages as desired by developers.

Page caching, primarily by browsers, is a major impediment to accurate visitor tracking using server log data. A significant portion of page views may

be page revisits which a browser services by reloading a local copy of files, thus avoiding sending page requests to the Web server. Since such caching is done for previously visited pages, it likely introduces a bias to page view counts by undercounting the most commonly visited pages and overrepresenting the proportion of visits to less "popular" pages. Webcrawler robots and automatic site downloading software such as WebWhacker also produce deceptive page views, since they generate file requests that do not correspond to actual user viewings of pages. In the case of technologies such as WebWhacker, the bias introduced may go either direction; a retrieved page may never be viewed by a user, or may be viewed many times while producing but a single file request.

Tracking paths of individual visitors is a "hit or miss" proposition. In some cases, such tracking can reveal highly detailed and accurate views of visitor behaviors. In other cases, gaps in the data, generally created by browser caching, prevent reliable analysis. If an evaluator wants to determine tracking details about at least some visitors, and is not too concerned about which visitors she or he studies, that researcher may be able to obtain some important insights into visitor behaviors. However, attempts to study specific, preselected visitors will likely require that many of the subjects' data be disregarded because it is incomplete. Furthermore, the eliminated data could introduce a bias to the remaining results, since visitors with certain behavioral patterns (return visits to previously viewed pages) will be the ones whose data is most commonly disregarded. The existence of referrer data in server log records is a bright spot in this story. Knowledge about a referring page often clarifies slightly ambiguous visitor path records, thus aiding researchers' decisions about which data is accurate and which contains too many gaps and must be thrown out.

Where did visitors come from? Three aspects of "where" visitors "come from" can, in some cases at least, be determined via server log records. Network addresses, when expressed as DNS entries, may contain information about the geographical location of a visitor (in country code top level domains, for instance) or about the institutional affiliation of a visitor (such as "msu.edu" for Michigan State University. Referrer data may indicate "where" on the Web a visitor "was" immediately before visiting the site being studied. As with most server log derived data, such "location" is not always available. Many numerical IP addresses cannot be resolved into DNS entries, so nothing can be discerned about locations of visitors with such network addresses. Referrer fields are sometimes blank, again depriving researchers of a potentially useful source of data.

Location data can help site developers better understand and accommodate their actual user audience. For instance, a site designer who discovered that her or his site is heavily used by visitors from Spanish-speaking countries might consider translating the site's text into Spanish. External referrer data can help developers determine whether their site is represented to the extent they desire on search sites and other sites with related topics, and to take steps to amend that situation if desired. Referrer field data, when it reveals that a referrer was external to the site being studied, also enables evaluators to determine which pages serve as "entry points" or "front doors" to their site, and to adjust the site to work better in light of such knowledge.

When did users visit? The server log is also a source of data about when high and low visitor traffic levels occurred. Site evaluators can track page view and visitor count trends over long periods, such as weeks, months, or years. Such tracking can reveal seasonal trends, such as slumps around

holidays or during schools' summer recesses. Researchers can also study traffic fluctuations during shorter time periods, such as days of a week or hours of a day. Knowing when high and low traffic periods occur can help site developers plan the timing of special events, such as conducting site maintenance on slow days or at off-peak times, or hosting online seminars or other special features at times that usually attract large numbers of visitors.

Time measures are sometimes confounded by the truly "World Wide" aspects of the Web and the distribution of its user population. Seasonal trends are opposite in the Earth's northern and southern hemispheres. File request times recorded in the server log are in terms of the time zone where the server is located, and thus do not reflect the time of day for visitors from distant parts of the globe. A site with greater appeal to distant visitors than to local ones might find that traffic levels are highest in the middle of the night local time. Since server log records may contain country codes in some visitors' DNS addresses, it is theoretically possible to at least partially resolve some visitors' visit times to their approximate locale time frame. In practice, this would be very difficult and time consuming to accomplish with accuracy for large numbers of visitors.

Skills, Tools, and Labor Investments Required for Evaluations

Automated data collection is the great advantage as a research methodology which use of server log data brings to Web visitor studies. Since a computer readily records the actions of thousands of visitors as a routine part of its operations, researchers can gather large amounts of detailed data with ease. The existence of numerous log analysis programs which are inexpensive or free, are not difficult to learn how to use, and which can easily

generate valuable reports about usage trends is the second great advantage of this approach to Web visitor studies. More detailed, focused analyses can be conducted with the support of common software packages such as databases, spreadsheets, and graphing utilities; whether such in depth analyses are justified in terms of value of information revealed as compared to amount of effort invested will likely vary depending on the goals of particular studies.

Long term trends in site total visitor and page view counts are simple to measure and graph, and can give site evaluators a powerful tool for selecting times which merit closer scrutiny. Similarly, evaluators can use log analysis reports to track page view counts over time of a small number of notable pages, such as the site's home page or recently introduced features, with little expenditure of effort. Tracking traffic trends for larger numbers of pages becomes more of a chore and is a data visualization problem. Superimposed weekly page view counts for more than a couple of pages could be very cluttered, whereas multiple graphs would be difficult to compare to see whether trends at certain times spanned the different pages or were localized to specific pages. Log analysis reports also produce "snapshots" of site activity spanning specific periods with ease, offering considerable detail of page view counts, common address domains of visitors, high traffic times of day, and so on. If evaluators know what time period they are interested in, analysis software can readily produce detailed summaries of many aspects of site activity spanning that time.

There are a few types of data which are frequently absent from at least some server log records which dilute the value of such records as a data source. Network addresses in the form of numeric IP addresses cannot always be resolved to DNS entries. Such unresolved do indicate unique "visitors" in the least precise meaning of that term, but provide essentially no other useful

information about visitor identities. Missing file requests, primarily caused by browser caching, eliminates much data about a user's actual "clickstream" or path through a site. Similarly, "non-eyeball" hits created by webcrawler robots and similar software skews page view counts away from the quantities researchers actually wish to measure. Visit duration and "time on page" measures are affected by the problem of uncertainty regarding the departure time from the last page viewed, which is especially acute for short visits involving small numbers of page views.

Close scrutiny of visit durations and paths followed by individual visitors requires more time and somewhat greater expertise than does generation of broad overview reports about many visitors created by log analysis software. Many visitor records must be thrown out because of missing data elements, which increases the amount of effort required per acceptable record analyzed. Careful path analysis requires a much greater investment of effort than does generation of summary reports, but yields much richer views of visitor behaviors. Although many records must be ignored because of missing data, the amount of data available due to automatic data collection may be immense enough to largely offset this limitation. Removal of records due to missing data requires analysts, however, to carefully monitor the introduction of biases which may result from trends in which records are not suitable for use.

How do other options for studying Web site visitors compare to use of server log data? Some sites require visitors to supply user IDs and passwords when they log onto the site, thus making visitor identification more reliable, especially from one visit to the next. Other sites use cookies and similar technologies to assign "tokens" to visitors, accomplishing a similar goal supporting more reliable identification, though in a way that can be viewed as

"less intrusive" to visitors or as more covert. Some researchers install software on the computers of users which monitors their clickstreams directly, thus avoiding the uncertainties associated with caching and with the lag introduced by data transmission over networks between user actions and reception of such signals at a server. Taking that concept a step further, some researchers also visually observe user actions (often in conjunction with recording clicks), either using human observers or by videotaping behaviors for later analysis. Some researchers employ surveys, which can readily be administered remotely over the Web, and which overcome limitations of merely observing outward behaviors and can inquire about users thoughts and feelings.

Each of the research methodologies has drawbacks as well as advantages as compared to server log based research. For example, use of cookies requires a Web server that supports that technology, implying that an evaluator has control over the choice of server software for a site, which may not be the case especially if the site is hosted on a server with other sites with varying needs. Installation of clickstream monitoring software on users' computers introduces greater reliability in recording all results, but is far more intrusive on research subjects and likely limits the range of subjects studied. Some analysis packages, especially ones employing cookie supported recording schemes, can automatically produce very detailed descriptions of the precise paths of users, but are often very expensive (several cost well over $10,000) and require much greater expertise on the part of the operator than do simple log analysis packages.

Web site visitor studies using different methodologies appear to require evaluators to choose priorities based on tradeoffs between seven basic factors. Ideally, one would like an evaluation effort to be inexpensive, to provide

information about many visitors, to provide in-depth information about visitors, to provide accurate and complete information, to require minimal technical expertise in use of analysis software on the part of evaluators, to require minimal time investments on the part of data analysts, and to produce results quickly. Server log based analysis can produce overviews of site activity which sacrifice some accuracy and completeness and are short on in-depth analyses of individual visitors, but which are relatively strong with regards to the remaining five criteria. To accomplish in-depth analyses of individual visitor behaviors based on server log records, some combination of the number of visitors analyzed, the amount of analysis effort invested, and the amount of time required to produce results must be sacrificed. Such in-depth analysis may also suffer from inaccuracy or incompleteness of records, depending on which subjects and records are studied.

Measuring Visit Durations and "Time on Page"

In some cases it is possible to make reasonably accurate estimates for visit durations and average time on page values based on server log data. Cases for which such estimates are reliable are ones for which no page view records are missing from a user's clickstream record (due to browser caching, for instance). Since departure time from the last page visited cannot be derived from log data, visit duration data for page view sequences containing larger numbers of page views will likely be more reliable and accurate than for visits lasting just a few page views.

Any system which measure time on page from the server's perspective can only record the time when a request is received, not when it is made by a user's actions. Variation in data transmission rates over the networks might

render such measurement techniques inaccurate compared to a researcher's need to measure the actual rate of activity by visitor.

Measurement schemes which record only users' clicks within the context of Web browser software cannot determine whether users were pausing from activity at times, or were doing other actions within their computer's environment, such as typing notes on a word processor. A system which measures all user clicks and keystrokes on that person's computer partially overcomes such limitations, but is likewise unable to distinguish between pauses and actions such as verbal discussions about page contents with a companion.

Server log data is not, in my opinion, up to the task of measuring users behaviors with the degree of accuracy required to firmly establish a relationship between time on page and learning. Initial studies to discover whether such a relationship exists in at least some cases should use a combination of automated recording on a user's computer of all keystrokes and mouse clicks in conjunction with visual observation of users. Visual observation, by a person or recorded using a video camera, would enable researchers to note whether a pause between clicks signified that a user was apparently reading a page, talking to a companion, away from the computer altogether, and so on. If research using visual observation and clickstream data showed a correlation between time on page and learning outcomes, it might be worthwhile to see how reliably a similar correlation could be measured using clickstream data alone, since existence of such a correlation would simplify research.

Correlations between time on page and learning would have to be tested on a variety of subjects using pages covering various topics and with sites with an assortment of page designs before measurement of visit durations

could generally be used as a proxy indicator of learning. Because of the number of variable factors involved, it seems unlikely that such a general relationship would be a precise measurement tool. However, as museum visitors studies researchers' use of similar measures of the amount of time visitors spend at exhibits shows, even such rough estimates of learning have utility on certain occasions. Some estimate of how much visitors are learning, imprecise and unspecific as it may be in such cases, can be better than no estimate at all. Sometimes rough measures of learning by large numbers of visitors that are compiled via relatively labor unintesive means are a good choice, just as other situations require careful study that reveals more precise details but necessitate larger labor investments and limit the number of subjects which can be studied.

If careful studies of visit durations and learning reveal a correlation between those quantities, it would be prudent to examine whether learning is correlated to the number of pages viewed as well as to the amount of time spent viewing such pages. Such a correlation would likely be weaker than a time and learning correlation, but would provide an even more readily measurable means by which to estimate learning. Such a relationship would further expand the set of instances in which at least some measure of learning could be estimated, and would benefit researchers by enabling them to study more visitors or by covering the same number of visitors with a smaller labor investment.

Server log data would be only marginally suitable for use in estimating learning if a correlation between page view counts and learning outcomes were discovered. The main limitation of use of log data for such a purpose stems from the absence of some page view records from the log, caused primarily by browser caching. This study contains one slight bit of data that

may hint at a link between page view counts and visit durations. As reported in the results chapter, visit durations for the thirteen page view visitors studied were about twice as long as durations for seven page view visitors. These results should be viewed with extreme caution, however, since the variances in visit durations were large, the number of subjects studied was small, and the selection of subjects for whom complete clickstream records were available may have introduced a bias into the results. This data does not prove anything, but may help later researchers understand some of the obstacles they will have to overcome to test for relationships between page view counts and learning and between visit durations and learning.

Guidelines for Site Evaluators

In light of my experiences studying visitors to the DLC-ME Web site, I devote the final section of this report to recommendations for visitor studies methods developers of other educationally oriented Web might wish to apply to their site evaluation efforts. Which of these techniques developers choose to employ should depend on the goals of their evaluations, the levels of expertise and types of software available to them, and the amount of effort they wish to apply towards evaluation efforts. The approach I suggest is scalable; some very simple measures can be used as the entire evaluation, can form the basis for decisions about which direction to head with more focused inquiries, or can temporarily be used as the entire evaluation which could later be expanded in various directions.

A good starting point for any investigation is to count total page views and total visitors to an entire site on a regular basis over fixed time periods. Such counts roughly indicate overall site traffic, and the "Microbes on Mars

incident" clearly indicates that unexpected events can dramatically alter traffic levels in ways that developers may wish to be aware of. Two likely candidates for an appropriate "fixed time period" over which such counts should be tallied are once per week or once per month. Weekly tallies provide equal-length time spans which support simple comparisons, routinely include similar common high and low traffic trend times (such as weekends), and provide sufficiently frequent sampling to allow developers to respond fairly quickly to events. Monthly tallies ease the evaluators labor investment burden, provide units that human observers are familiar with and can readily compare with the time frames of other phenomena (such as summer recess), and may be more suitable for following long term trends without creating visual clutter in graphs associated with too many data points.

Once site page view and visitor counts have been measured, dividing page views by visitors is a simple matter and provides another useful value, average page views per visitor. If page view visitor counts are logged into a spreadsheet, this derived quantity can be automatically calculated. Spreadsheets with graphing features built in can also aid trend evaluation efforts by enabling the creation of visual representations of page view, visitor, and page views per visitor trends.

Evaluators may be especially curious about page view counts for specific pages within their sites. A site's home page, main section menu pages, and newly introduced features' pages are common candidates for such scrutiny. If the number of such pages thus analyzed is small, both the data analysis and recording efforts and the techniques required to effectively make sense of results generated are readily manageable. If the number of such pages becomes large, more work is required to track them and methods for clearly presenting the results become tricky. In some cases evaluators may not have

access to server log files, may not be able to generate reports using log analysis programs, or may be unable to use server logs as a data source for some other reason. Evaluators interested in tracking page view counts for individual pages have a readily available alternative in such cases. Many freeware page view counters have been implemented and are available from numerous sources on the Web. Evaluators could install such counters on the pages they are keenly interested in. The values of such counters would have to be routinely checked and recorded on a periodic basis, but share the advantage as a data source possessed by server logs of automatic data generation. Also, users of such counters should avoid "hit" counters in favor of page view counters, which offer data that is more directly useful to evaluators.

Several types of more detailed analyses can be produced easily with the aid of log analysis software. Which ones evaluators should choose depends on the goals of their evaluations. Evaluators can assemble page view counts for sections of a site composed of multiple pages, which provide information that has some features of whole site counts and some of counts for individual pages. If evaluators are interested in the identities of visitors, they might wish to create reports of page view counts listed by network addresses in terms of domains or sub-domains. Developers wishing to know which other Web sites led visitors to their site might generate reports of page views broken down by referrers, and examine only the listings for external referrers. Developers wishing to schedule special events at specific times, such as server downtime for maintenance or live online seminars, can generate reports of server traffic levels by time of day or by day of the week to ascertain when their site's high and low traffic times typically occur.

Site evaluators may wish to use an adjustable, layered approach to site evaluation. For example, the routine evaluation effort might be to simply

record page view and visitor counts for the whole site on a weekly basis. If a week or series of weeks with unusual trends, such as high or low counts or discrepancies between visitor and page view counts, was spotted, a closer examination of that period might be called for. Likewise, evaluators tracking page view counts for sections of a site might decide to look at the counts for specific pages within a section if an unusual trend in section page view count levels arose. This layered approach allows evaluators to minimize effort invested in study of the site most of the time, but to dig more deeply into the details contained within existing datasets when simple efforts point out abnormal trends. This technique allows investigators to decide on a case by case basis how much effort to invest and how much detail is desired, to the level of tracking the clickstreams of individual site visitors during single visits.

Site developers can take some steps prior to or during the development phase of a site that can assist evaluation efforts. Developers can decide whether to include notification to site visitors that their actions may be studied as part of a research program, what the wording of such notification should be, and how such a notification will be integrated with page designs. If such notification is included, it is simpler to add it to all pages as they are created than appending it to them as an afterthought, and it is better to integrate the placement of such notification into page designs from the start than to "shoehorn" it in at the end.

Site designers may wish to create directory structures for Web file elements and file naming schemes that simplify the research process. Log analysis programs such as ServerStat can easily be told to produce a report covering all files within a given directory. If directory structures correspond to distinct sections of a site in terms of content arrangement, reports about those

sections will be easy to produce separately from reports covering other segments or the site as a whole. In the case of the DLC-ME site, the Microbe Zoo files are all in one directory, the Microbes in the News files are in another, and so on. Site developers might also wish to choose file names with analysis reports in mind. Reports listing page views by page titles typically can list the file names in alphabetical order. Carefully chosen names can group logically related pages together in reports, making it easier for human analysts to grasp trends and compare values for related pages. In the case of the DLC-ME site, all Microbe Zoo page file names begin with the letter "z", all Microbe News page file names begin with "n", Microbial Ecology Resources pages start with "r", and so on. Another way site developers might wish to distinguish between page types is by contrasting menu and content pages, possibly starting file names of the former with "m" and of the latter with "c", for instance.

Site developers may wish to keep a log of the page development process to support later analyses of visitors' reactions to the posting of new pages, alterations of old ones, or changes in link structures within a site. Many Web sites are constantly evolving entities, and without meticulous records evaluators may have a very difficult time determining the state of a site at any point in the past when they attempt to study visitors' reactions to certain features. Researchers might want to know how soon after a page was posted online visitors began to take notice of it. They might want to know how changes to a page influenced visitor behaviors relative to that page and others it links to. New links to previously existing pages could dramatically alter the number of visitors to such pages. The current status of a site often reveals little about its status at some prior time. Log record data is collected automatically, but analysis might be deferred to a much later date. An accurate record of the ongoing alterations made to a site's structure could

greatly facilitate later analyses.

Finally, this study has explored a variety of analyses and data displays based upon generally available data logs of Web visits. Although these analyses are not without considerable ambiguity, nevertheless they can inform Web designers as to the patterns of use of the site and suggest ways of redesigning the site to better meet their goals for the site. In view of the rapidly evolving nature of the Web, designers would be wise to pay attention to such measures in order to provide continuous formative feedback to the Web design process.

LIST OF REFERENCES

REFERENCES

Advanced help on suggesting sites to Yahoo! (1997). Retrieved November 11, 1997 from the World Wide Web: http://www.yahoo.com/docs/info/appropriate.html

Anderson, L. W. (1976). An empirical investigation of individual differences in time to learn. Journal of Educational Psychology, 68, 226-233.

Andrews, W. (1997a). Browser robots raise issues for advertisers. Web Week, 3(11). Retrieved October 22, 1997 from the World Wide Web: http://www.webweek.com/97Apr21/markcomm/robots.html

Andrews, W. (1997b). Challenge for spiders: Searching invisible web. Web Week, 3(3), 48-51.

Andrews, W. (1997c). Damn the push media, engines will focus on targeting this year. Web Week, 3(1), 46-47.

Andrews, W. (1997d). Offline browsing could foil agents. Web Week, 3(11). Retrieved October 22, 1997 from the World Wide Web: http://http://www.webweek.com/97Apr21/markcomm/offline.html

the Anonymizer. (1997). Retrieved January 8, 1997 from the World Wide Web: http://www.anonymizer.com

Bayne, K. M. (1997). The Internet marketing plan. New York: John Wiley and Sons.

Beer, V. (1987). Great expectations: Do museums know what visitors are doing? Curator, 30, 206-215.

Berliner, D. C. (1992). Redesigning classroom activities for the future. Educational Technology, 32(10), 7-13.

Berliner, D. C., & Fisher, C. W. (1985). One more time. In C. W. Fisher & D. C. Berliner (Eds.), Perspectives on instructional time (pp. 333-347). New York: Longman.

Bloom, B. S. (1974). Time and learning. American Psychologist, 29, 682-688.

Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. Communications of the ACM, 39 (11), 42-48.

Brophy, J. (1979). Teacher behavior and its effects. Journal of Educational Psychology, 71, 733-750.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 328-375). New York: Macmillan.

Brown, D. (1997). Description of Windows log entries. Retrieved October 30, 1997 from the World Wide Web: http://netsell.com/dbrown/loganalysis.html

Brown, E. (1997, Feb. 13). The Web gets pushy. NewMedia. Retrieved October 30, 1997 from the World Wide Web: http://www.newmedia.com/Today/97/02/13/Web_Pushy.html

Buchanan, R. W., Jr., & Lukaszewski, C. (1997). Measuring the impact of your Web site. New York: John Wiley and Sons.

Burner, M. (1997). Crawling towards eternity. Web Techniques, 2(5), 37-40.

CASIE guiding principles of interactive media audience measurement. (1997, April 2). Retrieved October 26, 1997 from the World Wide Web: http://www.commercepark.com/AAAA/bc/casie/guide.html

Carroll, J. B. (1963). A model of school learning. Teachers College Record, 64, 723-733.

Carl, J. (1995). Protocol gives sites way to keep out the 'bots. Web Week, 1(7). Retrieved October 21, 1997 from the World Wide Web: http://www.webweek.com/95Nov/news/nobots.html

Clark, S. (1997). The Webmaster's bakery. Web Developer, 3(2), 86-90.

Cooper, L. F. (1996). A new generation of Web metrics. Information Week, 63. (Issue no. 607).

Cortinas, M. (1997a). Luckman, 1.0 Technologies eat cookies. MacWeek, 11(34), 30.

Cortinas, M. (1997b). Power Knowledge surveys customers. MacWeek, 11(23), 27-28.

Diamond, J. (1986). The Behavior of Family Groups in Science Museums. Curator, 29, 139-154.

Duncan, G. (1997). Send in the robot. MacWorld, 14(1), 153-156.

Ebbinghaus, H. (1964). Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1855)

Edminston, R. W. & Rhoades, B. J. (1959). Predicting achievement. Journal of Educational Research, 52, 177-180.

Falk, John H. (1983). Time and behavior as predictors of learning. Science Education, 67, 267-276.

Fayyad, U., Haussler, D., & Stolorz, P. (1996). Mining scientific data. Communications of the ACM, 39 (11), 51-57.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39 (11), 27-34.

Filby, N., Marliave, R., & Fisher, C. (1977, April). Allocated and engaged time in different content areas of second- and fifth-grade reading and mathematics curriculum. Paper presented at the meeting of the American Educational Research Association, New York, NY.

First year of PC Meter® Web measurement data reveals: If you promote it (and offer value), they'll come. (1997, March 10). Retrieved October 25, 1997 from the World Wide Web: http://www.pcmeter.com/pcmpr23.htm

Fisher, C., Filby, N., & Marliave, R. (1977, April). Instructional time and student achievement in second-grade reading and mathematics. Paper presented at the meeting of the American Educational Research Association, New York, NY.

Floyd, M. (1997). The way the cookie crumbles. Web Techniques, 2(2), 5.

Gage, N. L., & Berliner, D. C. (1979). Educational psychology. Chicago: Rand McNally.

Gagné, R. M. (1985). The conditions of learning and theory of instruction. Orlando, FL: Holt, Rinehart and Winston.

Glossary of NetCount terms. (1997). Retrieved October 27, 1997 from the World Wide Web: http://www.netcount.com/glossary.html

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1996). Statistical inference and data mining. Communications of the ACM, 39 (11), 35-41.

Good, T. L. (1979). Teacher effectiveness in the elementary school. Journal of Teacher Education, 30, 52-60.

Good, T. L. (1983). Classroom research: A decade of progress. Educational Psychologist, 18, 127-144.

Good, T. L., & Brophy, J. (1995). Contemporary educational psychology (5th ed.). White Plains, NY: Longman.

Gray, M. (1997). Web growth summary. Retrieved October 14, 1997 from the World Wide Web: http://www.mit.edu/people/mkgray/net/web-growth-summary.html

Greene, M. (1988). Stalking the average North American zoogoer. Museum News, 67(1), 50-51.

Haeseler, J. K. (1989). Length of Visitor Stay. In Bitgood, S. (Ed.), Visitor studies: Theory, research, and practice (Vol. 2) (pp. 252-259). Anniston/Oxford, Alabama: Jacksonville State University.

Hamit, F. (1996). Inktomi search engine, Hotbot, delivers power, unparalled speed. Silicon Graphics World, 6(12), 18-19.

I/PRO: FAQ. (1997). Retrieved October 26, 1997 from the World Wide Web: http://www.ipro.com/faq.html

Keller, J. M. (1983). Motivational design of instruction. In C. M. Reigeluth (Ed.), Instructional-design theories and models: An overview of their current status (pp. 383-434). Hillsdale, NJ: Erlbaum.

Koran, J. J. Jr., Foster, J. S., & Koran, M. L. (1989). The relationship among interest, attention and learning in a natural history museum. In Bitgood, S. (Ed.), Visitor studies: Theory, research, and practice (Vol. 2) (pp. 239-244). Anniston/Oxford, Alabama: Jacksonville State University.

Koster, M. (1995). Robots in the Web: threat or treat? ConneXions, 9(4). Retrieved October 21, 1997 from the World Wide Web: http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html

Koster, M. (1997). The Web Robots pages. Retrieved October 21, 1997 from the World Wide Web: http://info.webcrawler.com/mak/projects/robots/robots.html

Lee, L. (1996a, Nov. 6). Beyond the hits: Mining Web sites for traffic data. NewMedia. Retrieved October 30, 1997 from the World Wide Web: http://www.hyperstand.com/Today/96/11/06/Web_Traffic_Data.html

Lee, L. (1996b, April 25). Market Focus 2. NewMedia. Retrieved October 30, 1997 from the World Wide Web: http://www.newmedia.com/Today/96/04/25/Market_Focus_2.html

Lee, L. (1996c, March 5). More ways to analyze Web traffic. NewMedia. Retrieved November 10, 1997 from the World Wide Web: http://www.newmedia.com/Today/96/03/05/Analyze_Web_Traffic.html

Lee, L. (1996d). Site analysis tools get serious. NewMedia. Retrieved October 30, 1997 from the World Wide Web: http://www.newmedia.com/Today/96/11/06/analysis.html

Lee, L. (1996e, July 22). Updated Open Market software. NewMedia. Retrieved November 10, 1997 from the World Wide Web: http://newmedia.com/Today/96/07/22/Open_Market_Software.html

Lee, L. (1997a). ClickWise serves up ads intelligently. NewMedia, 7(2), 32.

Lee, L. (1997b). Open Profiling Standard emerges. NewMedia, 7(9), 34.

Lemay, L. (1997). Lost and found. Web Techniques, 2(3), 18-22.

Lottor, M. (1997). Internet domain survey, July 1997. Retrieved November 15, 1997 from the World Wide Web: http://www.nw.com/zone/WWW/report.html

Manes, S. (1997, Feb. 11). Missing links on search engines. The Nando Times. Retrieved October 22, 1997 from the World Wide Web: http://www.nando.net/newsroom/ntn/info/021197/info26_9831.html

Microsoft Site Server, Enterprise Edition Usage Analyst FAQ. (1997). Retrieved October 27, 1997 from the World Wide Web: http://www.backoffice.microsoft.com/products/features/UsageAnalyst/FAQ UaEnterprise.asp

Morrissey, K. (1991). Visitor behavior and interactive video. Curator, 34, 109-118.

Murphy, K. (1996). Tracking tools follow visitors' footsteps. Web Week, 2(11). Retrieved October 30, 1997 from the World Wide Web: http://www.webweek.com/96Aug05/software/tracking.html

New media companies want better traffic measurement tools. (1996, October 28). NewMedia Week, 5. Retrieved October 29, 1997 from the World Wide Web: http://www.andromedia.com/who/inmedia/inmediadex.html

Patten, D. (1997). Preparing your Web server for a siege. NetscapeWorld, 2(3). Retrieved November 10, 1997 from the World Wide Web: http://www.netscapeworld.com/netscapeworld/nw-03-1997/nw-03-siege.html

Patterson, D., & Bitgood, S. (1988). Some evolving principles of visitor behavior. In Bitgood, S., Roper, J. T., Jr., & Benefield, A. (Eds.), Visitor studies: Theory, research, and practice (pp. 40-50). Anniston/Oxford, Alabama: Jacksonville State University.

Pearlstein, J. (1997). Bolero revision links to Oracle. MacWeek, 11(16), 14-16.

Persistent client state HTTP cookies. (1997). Retrieved October 30, 1997 from the World Wide Web: http://home.netscape.com/newsref/std/cookie_spec.html

Pierce, M. (1989). Four years of visitor evaluation at the Anniston Museum of Natural History. In Bitgood, S. (Ed.), Visitor studies: Theory, research, and practice (Vol. 2) (pp. 180-191). Anniston/Oxford, Alabama: Jacksonville State University.

Rich, H. L., & McNelis, M. J. (1987). A study of academic time-on-task in the elementary school. Educational Research Quarterly, 12(1), 37-46.

Schwartz, R. L. (1997). How to be virtually anonymous. Web Techniques, 2(2), 30-33.

Seifert, E. H., & Beck, J. J., Jr. (1984). Relationship between task time and learning gains in secondary schools. Journal of Educational Research, 78, 5-10.

Seiter, C. (1997). Bolero: Web-site logging for pros. MacWorld, 14(4), 68.

Shaffer, R. A. (1996, Oct. 7). The new math. ComputerLetter, 12(13). Retrieved October 30, 1997 from the World Wide Web: http://www.andromedia.com/who/inmedia/inmediadex.html

Shulman, L. S. (1990). Paradigms and programs. New York: Macmillan.

Smith, Z. (1997). The truth about the Web. Web Techniques, 2(5), 38.

Stein, L. (1996, April). Logs, logs, and more logs! Web Techniques. Retrieved June 10, 1997 from the World Wide Web: http://www.webtechniques.com/features/april96logs.html

Stein, L. D. (1997a). Cookies, sweet and sour. Web Techniques, 2(2), 8-11.

Stein, L. D. (1997b, September 3). The World Wide Web security FAQ: Server logs and privacy. Retrieved October 30, 1997 from the World Wide Web: http://www-genome.wi.mit.edu/WWW/faqs/wwwsf6.html

Stout, R. (1997). Web site stats: Tracking hits and analyzing traffic. Berkeley, CA: Osborne/McGraw-Hill.

Submit It!: Tips for announcing Web sites to search engines and directories. (1997). Retrieved October 21, 1997 from the World Wide Web: http://www.submit-it.com/subopt.htm

Surmanek, J. (1993). Introduction to advertising media. Lincolnwood, IL: NTC Publishing Group.

Waring, B. (1997). The 1997 hyper awards. NewMedia, 7(1), 46-54.

WebSTAR technical reference [Computer software manual]. (1995). Berkeley, CA: StarNine Technologies.

Wiederspan, J., & Shotton, C. (1996). Planning and managing Web sites on the Macintosh. New York: Addison-Wesley.

Wilford, J. N. (1996, August 7). Clues in meteorite seem to show signs of life on Mars long ago. The New York Times, pp. A1, A10.

Yahoo! - How to suggest your site. (1997). Retrieved November 11, 1997 from the World Wide Web: http://www.yahoo.com/docs/info/include.html

Zielske, H. A. (1959). The remembering and forgetting of advertising. The Journal of Marketing, 23, 239-243.

Zielske, H. A., & Henry, W. A. (1980). Remembering and forgetting television ads. Journal of Advertising Research, 20(2), 7-13.