DOCUMENT RESUME

ED 448 704 IR 020 467

AUTHOR	Chen, Hsiang; Tan, Zixiang
TITLE	Toward a Standardized Internet Measurement.
PUB DATE	1999-10-00
NOTE	7p.; In: WebNet 99 World Conference on the WWW and Internet Proceedings (Honolulu, Hawaii, October 24-30, 1999); see IR 020 454.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Data Collection; *Evaluation Criteria; *Internet; *Measurement Techniques; Research Methodology; Standards

ABSTRACT

This paper investigates measurement issues related to elements of the Internet and calls for a standardized measuring scheme to resolve the problem of the measurement. The dilemmas in measuring the elements of the Internet are identified, and previous studies are reviewed. Elements of the Internet are categorized into population, usage, protocol flow, hardware, software, traffic, and visitors. The paper proposes the following four criteria in measuring the elements of the Internet as the guidelines for Internet research: (1) validity and reliability check; (2) continuation of data collection; (3) a period of data collection limited to days, not months; and (4) a methodology that accommodates multiple users' needs and answers multiple questions. (Contains 11 references.) (Author/MES)

Toward a Standardized Internet Measurement

Hsiang Chen, hchen04@mailbox.syr.edu
School of Information Studies, Syracuse University, Syracuse, New York 13244 U.S.A.

Zixiang (Alex) Tan, ztan@syr.edu
School of Information Studies, Syracuse University, Syracuse, New York 13244 U.S.A.

Abstract: This paper investigates measurement issues related to elements of the Internet and calls for a standardized measuring scheme to resolve the problem of the measurement. In this paper, the dilemmas in measuring the elements of the Internet are identified and previous studies are reviewed. Elements of the Internet are categorized into population, usage, protocol flow, hardware, software, traffic, and visitors. At last, this paper proposes four criteria in measuring the elements of the Internet as the guidelines for Internet research.

1. Introduction

This paper investigates the dilemmas in measuring the elements of the Internet and reviews current attempts in monitoring the growth of the Internet. By examining these important issues, this paper aims to provide guidelines and criteria in measuring the entity of the Internet for further research.

The original ARPANET evolved into the present day Internet which itself has changed much in the last two decades since it came into existence. During the late 1980s, the population of Internet users and network constituents expanded internationally and began to include commercial facilities [Cerf 1998]. Today the Internet is made up of private networking facilities in educational and research institutions, businesses, and government organizations across the globe. It was developed in the era of time-sharing, but has survived into the era of personal computers, client-server architecture, peer-to-peer computing, and network computers [Leiner et al. 1998].

From a human perspective, no single person or single organization has contributed totally or controlled the growth and the development of the Internet. From a technological perspective, from 1969 to 1975, when the Internet (called APRANET) was still in a research and development period, drawing maps and topology diagrams, calculating traffic and performance statistics, and measuring the size and diffusion of the net were feasible. After 1975, when the Internet was turned over to the Department of Defense, tracking the Internet became a headache for the National Science Foundation [Press 1997]. Today, as more and more border-crossing backbones have been further developed and intermesh, tracking the development of the Internet and measuring related growth elements has become harder and harder.

Since it came into existence, the Internet has grown beyond its initial purposes and includes both a broad user community and increased commercial activity. Over the years, the Internet has become a medium with world-wide broadcasting capability; a mechanism for information dissemination; and a medium for collaboration and interaction between individuals and their computers by overcoming obstacles of distance [Leiner et al. 1998]. From a marketing perspective, the Internet represents one of the most successful examples of the benefits of sustained investment and commitment to businesses; and for academic researchers, the diffusion of the Internet represents an interesting phenomenon needing further diagnosis with scientific explorations [Morris & Ogan 1996].

Depending on how the Internet is viewed and what units of analysis are defined, studying elements of the Internet can generate different interpretations and different meanings. There is an urgent need to understand the growth of the Internet. In one sense, the Internet is rapidly growing in its number of users, its volume of protocol traffic, its complexity of topologies, its impacts on human beings' lifestyles, its value in economic activities, and its coverage. In another sense, the global diffusion of the Internet, the fostered knowledge organized by the Internet, the changes in users' behavior, and the impacts to human beings' global perspectives are intriguing issues in academic fields. More explorations of these phenomena would increase our understanding of the implications of this global technology and how it can potentially affect the lives of human beings.

As the Internet continues to evolve, the need to increase our understanding of the elements of the Internet becomes more urgent. Especially, we need a feasible and acceptable standard of Internet measurement which can allow us to monitor, track, and compare the size of the Internet, the growth rate of the Internet, the usage of the Internet, and the attributes of the Internet. This paper first investigates dilemmas to which researchers encounter when measuring the elements of the Internet. Further, related studies in measuring the elements of the Internet through are reviewed. This paper then provides a few useful criteria in resolving dilemmas of the measurement of the Internet.

2. Dilemmas in Measuring the Internet Elements

Since the Internet is rapidly growing in different dimensions, measuring the Internet from different perspectives becomes more important. However, researchers do not know exactly what to measure and how to measure this multi-dimension issue. Previous studies, such as [Novak & Hoffman 1997, Hoffman & Novak 1996] suggested that lacking the standards for what to measure and how to measure the element of the Internet would limit industries' further participation in the Internet activities. Today, different units of analysis were used when researchers attempted to measure the Internet [December 1996]. Under the circumstance of lacking guidelines and criteria, the problems of inconsistency arose and the results were controversy, and contradicted to each other in many areas.

Depending on how the Internet is viewed and what units of analysis are defined, measuring the elements of the Internet may generate different interpretations. For example, from the new technology viewpoint, the Internet can be measured as its penetration rate, adoptions of innovations, and its evolvement during diffusion process. When taking the viewpoint as mass medium, the measurement of the Internet may then focus on one-to-one, many-to-one, or one-to-many communication schemes, as well as send, recipients and message contents. If the Internet is studied as a CMC (computer mediated communication) multimedia environment, researchers may tend to explore hits, visits, uses and effects of the Internet for its commercial purposes. If the view of information dissemination is taken, it is then possible that researchers would study issues involving information flow across border, content censorship and regulation, or topics related to search engines. If a more behavior-orientated approach is taken, researchers would focus on demographics. From the geographical view, the measurement of Internet then becomes which university, town, city, state or country has the highest penetration rate. From the point of view of economics, researches may look at the relationships between economic development and the uses of the Internet on the linkages between telecommunication and trade flows in the economy.

3. Reviews of Internet Measurement

In order to establish baseline data for standardized procedures of the Internet measurement, the authors reviewed previous empirical studies researched on the Internet measurement issues. When reviewing such a rich data over time, we have limited ourselves on three questions: What were measured? Why were they measured? and How have they been measured? We sought the answers during our review process and later realized that the answers and the questions are intertwined. In reality, how the elements of the Internet were measured and what were measured actually depends upon how the final statistical numbers were used. Different research procedures and methodologies serve different purposes.

In the reviewing process, we made no attempt to review all previous studies. Rather, we tried to extend dimension of the Internet measurement as diverse as possible. Therefore, in this paper we did not list all similar studies in the same dimension. We have created two criteria when constructing this list. First, they must be empirical studies. Second, they must be able to make comparisons with other. To meet the first criterion, a study should provide its detailed description of methodology. To meet the second criterion, a study should present its results by numerical statements. From our preliminary literature review from traditional press format as well as the Internet, we were able to categorize the measurement of the Internet into the following categories:

3. 1. Population

The most common measurement of the Internet is about the Internet population. Virtually, there is no way to determine how many users are on the net without making guesses. The approach of counting human heads usually includes two sub-categories: number of users in a social system and attributes of demographics. The former refers to those studies investigating "how many" research questions, such as "How many Internet (adult or kid) users in United States or in the world" (e.g., http://www.nua.ie/surveys, http://www.cyberdialogue.com/free_data/index.html, and http://www.commerce.net/news/press/fact0699.html), "How many Web users in California State?" (http://www.commerce.net/news/press/030398_1.html) "What's the most wired big city in the country?" (http://www.zdnet.com/yil/content/mag/9803/wired.html) or "Which city in the United States owns the highest penetration rate of the Internet?" (http://www.commerce.net/news/press/030398_1.html) The later refers to those studies investigating "Who are they" questions, such as "What is the age range for those Internet users in June 1998?", "What are their educational background?" and "Are male users still more than female users?" (e.g., (http://www.gvu.gatech.edu/user_surveys/survey-1998-10/) The methodologies used to elicit the population number and attributes were through a telephone survey or through a self-administered Web survey. In the telephone survey, usually random sampling technique and statistical inference were used. In the Web survey, they usually suffered from problems of convenience sample.

3.2. Use

In addition to the study of Internet population, the usage of the Internet is also one of researchers' concerns. This approach attempts to answer the "how often" question. In order to investigate the "frequency" question, researchers usually conduct a telephone survey or a Web based survey. Usually, respondents were asked to identify the time of their last access to the Internet, hours of using the Internet per week, number of years of using the Internet, their Internet connection speed and so on (e.g., http://www.gvu.gatech.edu/user_surveys/survey-1998-10/). Using the methodology of telephone interview or Web based survey to answer "how often" questions usually generated distorted data. For example, because respondents were asked to describe their past experiences with memory, the results may not represent the truth. Respondents tend to increase the number of their Internet uses when asked, and in some cases, respondents are confused with the meaning of "uses."

3.3 Protocol flow

The Internet is comprised of several application layer of TCP/IP, such as HTTP, FTP, SMTP, POP and NNTP. Some of protocols, which have been very active and popular on the Internet, are now fading away (Gopher and Hypernet) while other new protocols (HTTP and POP) provide the majority traffic. One of purposes in measuring the application layer protocols is to observe their variations and to predict the trends of future developments, in addition to the performance of data flow in a timely manner [Monk & Claffy 1997]. In order to predict and observe the trend, research focusing on the protocol measurement should not limit their studies to one shot, and should extend their studies over time. Unlike the measurement of the Internet traffic (discussed later in this section), measuring the Internet protocol focuses on the comparisons among different protocols, their growth and decay over time, and the patterns of traffic flow. They look at the traffic of packets, bytes and flow in a specific time interval and monitor how Internet protocols flow over backbones [Apisdorf et al. 1997]. The traffic flow may include flow type (which protocol is observed), source/destinations of traffic (efficiency of traffic flow) and distributions of packet sizes and duration (effectiveness of networking) [Monk & Claffy 1997]. Today, however, it is practically almost impossible and cost prohibitive to detect the variation of protocols on the whole Internet because there are more than 30 backbone in the United States and more around the world. Detecting the bit flow over backbones in different time slots would be tedious with a considerable expense, if not possible.

3.4. Hardware

Previous attempts in measuring the size and the growth of the hardware focus on the Internet hosts and the Internet domains. See Robert H'obbes' Zakon's Internet Timeline v3.3 (http://www.isoc.org/guest/zakon/Internet/History/HIT.html) for the full description of these data and Netsizer for a commercial tool serving this purpose (http://www.netsizer.com). Understanding the number of domain or hosts worldwide or in a specific country may potentially increase our knowledge of the diffusion process of the Internet. Since the adoption rate of hardware may relate to culture, social and economic issues, analyzing these data can also contribute to our understandings of the Internet diffusion and societal factors. The basic procedure to collect data of hardware number is to ping the IP addresses through the Internet. By adding the replies after pinging, it is then possible to find out a total number of hosts and domains on the Internet. There are a few potential limitations of using this approach to measure the Internet. First, when some parts of the Internet choose to limit access to themselves to various degrees, the data collected would be distorted. To solve the above problems, researchers in Network Wizards (http://www.nw.com) created another new survey methodology in 1998 (http://www.nw.com/zone/WWW/new-survey.html). In their survey, in stead of ping every 4.3 billion IP address on the Internet, they tried 879,212 delegations, or just 223,319,848 possible hosts. The potential problem of this newer methodology is that the degree of the precision decreases. Second, in order to collect data over time, researchers need to conduct this procedure frequently. This would increase the unnecessary traffic of the Internet. Third, as the Internet becomes bigger and bigger, it then turns out practically impossible to ping over the Internet in a short time frame. Fourth, the definition of a host has changed in recent years due to virtual hosting. Today, a single machine can act like multiple systems with multiple domain names and IP addresses. Fifth, since an increasing number of domain names are registered in the USA, instead of in their own countries, the measurement of domain names becomes less meaningful.

3.5. Software

The war between Microsoft and Netscape is not about browsers but about the standard. Measuring the Internet software can reveal the fact which software dominates the Internet by studying the ration of their market sharing. Software related to the Internet use includes two parts: server and clients. As Web technology becomes prosperous, studies of Internet software basically focus on Web server applications and Web client applications. The methodology used to collect data of Web server applications and Web client applications are different. To detect which Web server application is running in a Web site, researchers can simply send a request for server's header by using HTTP. Researchers can then parse server's initial response to identify the server type. In June 1999, the Netcraft Web Server Survey (http://www.netcraft.com/survey) received 6,177,453 responses from their systematical poll. In their report, Apache takes 56.19% of the Internet Web server market while Microsoft IIS, 22.34%. To find out which Web browsers a Web user is using, researchers can use general telephone survey (http://www.psrinc.com/browser.htm) or just analyze the log file automatically stored in a Web server' directory. When a link between a Web browser and a Web server is established, a browser sends its browser type as "User-Agent" to Web server for recognition and communication purpose. Therefore, if a researcher can access log files from some popular Web sites, he should be able to tell the ratio to which each Web browser takes. By studying 2,000,000 Internet session data points over a time period of 21 months, Positive Support Review reported in June 1999 that Microsoft's Internet Explorer's market share has remained consistent in recent months, reaching as high as 66.61% (http://www.psrinc.com/PressRelease/PR_19990623.htm).

3.6. Traffic

Traditionally, issues of traffic are researchers' most concerns in the field of telecommunications because traffic data represent the precision and intensity of activities in a distributed communication environment. By analyzing traffic data, researchers can reveal a clear image of data flow in terms of amounts, directions, and growth. Changes in traffic volume can provide information related to "carrier productivity, tariff levels, market entry and the basis for settlements between interconnecting carriers, both domestic and foreign" [Staple & Mullins, 1989]. Besides, traffic data involves some complicated issues, such as information flow and content regulation. Crossed-border information basically is not a problem when the information per se is neutral and acceptable by people on the other side of border. It starts to cause troubles when the information is considered biased or unhealthy. When a country wants to keep its traditional values, like Singapore, from pollution of the Internet, content regulation becomes a necessary step to shut out border across information. Traffic can be measured by its total amount of bit passing through backbones in a time slot or by the amount of data passing through borders. Measuring the traffic flow on the Internet requires a higher degree of cooperation and involvement by service providers as well as multiple-nation cooperation. In order to make the data sensible, traffic measurement should be conducted over time and make comparisons periodically. In late 1995, Munzner et al. [Munzner et al. 1996] created a visually depicting Internet traffic components, which displayed how information on the Internet was routed over national borders.

3.7. Visitor

Measuring the number of visitors to a Web site or a Web page has its industry value because this summarized number is considered highly correlated to the exposure and interactivity of advertisements on the Web. Several different reporting schemes have been created, such as reach, frequency, duration time, and exposure. Unfortunately, there is no consensus on the definitions of these terms. See [Novak & Hoffman 1997] for detailed discussions of these terms. Those measurement units include hit, request, visits, user, organization, request duration, visit duration, Ad view, Ad click, Ad yield and Geography (see the example in United Expressline Web site: http://www.unxpres.com/usage/summary.html). Basically, to measure the effectiveness of a Web site or a set of Web pages includes counting and summarizing the visitor transactions on a Web site. Data from these counting processes may summarize who visited (visitor identities), when they visited (visitors' accessing time), where they were from (referred Web pages or entry point), what they visited (linkages between Web pages), how long they have stayed (elapse time), what they have done (transactions and interactivities), and where they exit (exit point). These summarized reports may tell managers a possible way to adjust the content and organization of his Web site, also how to charge clients of a Web site's advertisement. Besides, by applying this concept to a Internet Web site, it is then possible to compare the hottest Web sites on the Internet, usually measured by visits & hits.

4. Analysis

The Internet itself, especially with the view of marketing, is characterized by uncertainty. As suggested by [Hoffman & Novak 1997], one way to decrease this feeling of uncertainty is to build up an open methodological standard. However, we tend to argue that today's Internet is like frontier of the big West in the 18th century, where orders and procedures were yet constructed. Further, as the Internet keeps evolving with an accelerated rate, newly contrived procedures and orders may soon be collapsed due to their inability to adapt to the newly evolved world. To decrease the uncertainty and to increase our understanding of the elements of the Internet, a practical and acceptable measuring scheme should be established and a standard of measurement should also be built up.

However, after reviewing those attempts in measuring the elements of the Internet through diverse dimensions, the authors tend to believe that building a standardized, universally acceptable scheme for the measurement of the Internet is not feasible yet. Since the methodology aims to resolve research questions and should be consistent with research purposes, different research purposes would generate different methodologies. A standardized measuring scheme that aims to resolve all research questions related to elements of the Internet could actually be difficult to achiee. Nevertheless, the authors tend to believe building up some general guidelines which may be applied to all diverse research issues is still necessary and possible.

After reviewing those studies related to the measurement of the elements of the Internet, the following guidelines are proposed. These guidelines, which may be applied to different environment, attempt reveal the essential characteristics in measuring the Internet.

First, many studies, which investigated the elements of the Internet, are under the threats of validity and reliability. Due to the uncertainty characteristic built into the Internet, any researchers who want to study the elements of the Internet should first attempt to solve the problems of validity and reliability. Further, validity and reliability check should be the first criterion built into research designs and should be reported in detail

Second, the continuation of the study should be taken care. Since the Internet is evolving, growing up and expanding, the process of data collection should be extended to a longer session. The continuation of the data collection is necessary otherwise the results concluded from today's data may not be effectively in predicting tomorrow's situations. In most cases, multiple data collections over time are required to make meaningful and practical conclusions.

Third, human beings' Internet behaviors are constantly changing. Therefore, the period of data collection should be limited to days, not months. The Internet is like an arena, where different forces keep fighting and wrestling. Consequently, human beings' behaviors are open to change while the external environment constantly moves to different directions. The characteristic of the fast Internet evolution may influence the pattern of data, which later may invalidate analysis. For other academic and scientific research, the researching settings are relatively stable. Even though in some situations the research context may change, these variations are usually observable. It is very different on the Internet.

Fourth, if a methodology can accommodate to serve multiple users' needs and answer multiple questions through different dimensions at the same time, this methodology would be more desirable and useful [Staple and Mullins 1989]. This criterion is important because it can not only limit the resources invested on the Internet but also limit the Internet traffic.

5. Conclusion

In summary, in this paper we have discussed the needs of studying elements of the Internet and dilemmas of measuring the elements of the Internet. We also reviewed previous studies that attempted to measure the elements of the Internet. From the knowledge we learned from the reviewing process, we proposed four criteria, which should be followed in conducting research in measuring the elements of the Internet.

6. Reference

Apisdorf, J., Claffy, K., & Thompson, K.. (1997). OC3MON: Flexible, Affordable, High-Performance Statistics Collection. Proceedings of INET 97. http://www.isoc.org/inet97/proceedings/F1/F1_2.HTM.

Cerf, Vint. (1998). Brief History of the Internet. http://www.isoc.org/internet-history/cerf.html.

December, John. (1996). Units of Analysis of Internet Communication. Journal of Communication 46(1), Winter. 14-38.

Hoffman, D.L., W.D. Kalsbeek and T.P. Novak (1996), "Internet and Web Use in the United States: Baselines for Commercial Development," Special Section on "Internet in the Home," Communications of the ACM, 39 (December), 36-46.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G. & Wolff, S. (1998). A Brief History of the Internet. http://www.isoc.org/internet-history/brief.html.

Monk, Tracie & Claffy, K. (1997). Internet Data Acquisiton and Analysis: Status and Next Step. http://www.isoc.org/inet97/proceedings/F1/F1_3.HTM.

Morris, Merrill and Ogan, Christine. (1996). The Internet as Mass Medium. Journal of Communication. 46(1), Winter. 39-50.

Munzner, T., Hoffman, E., Claffy, K. & Fenner, B. (1996). Visualizing the Global Topology of the Mbone. Proceedings of the 1996 IEEE Symposium on Information Visualization, pp. 85-92, October 28-29 1996, San Francisco, CA, 1996.

Novak, T.P. and D.L. Hoffman (1997), "New Metrics for New Media: Toward the Development of Web Measurement Standards," World Wide Web Journal, Winter, 2(1), 213-246.

Press, Larry. (1997) Tracking the global diffusion of the Internet. Communications of the ACM, Nov 1997 v40 n11 p11(7).

Staple, Gregory C. & Mullins, Mark. (1989). Telecom Traffic Statistics -- MiTT Matter: Improving economic Forecasting and Regulatory Policy. Telecommunications Policy, June (1989): 105-127.