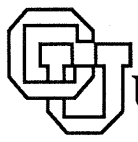
A Measurement Study of Changes in Service-Level Reachability in the Global TCP/IP Internet: Goals, Experimental Design, Implementation, and Policy Considerations

Michael F. Schwartz

CU-CS-551-91 October 1991



University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

A Measurement Study of Changes in Service-Level Reachability in the Global TCP/IP Internet: Goals, Experimental Design, Implementation, and Policy Considerations

Michael F. Schwartz

CU-CS-551-91 October 1991

Department of Computer Science
Campus Box 430
University of Colorado
Boulder, Colorado 80309-0430
(303) 492-3902
Internet: schwartz@cs.colorado.edu

Abstract

In this report we discuss plans to carry out a longitudinal measurement study of changes in service-level reachability in the global TCP/IP Internet. We overview our motivation, experimental design, considerations of network and remote site load, mechanisms used to control the measurement collection process, and efforts to inform administrators at sites measured by this study, along with concomitant privacy and security issues. A list of references and information on how to contact the Principal Investigator are included.

Introduction

The global TCP/IP Internet interconnects millions of individuals at thousands of institutions worldwide, offering the potential for significant collaboration through network services and electronic information exchange. At the same time, such powerful connectivity offers many avenues for security violations, as evidenced by a number of well publicized events over the past few years. In response, many sites have imposed mechanisms to limit their exposure to security intrusions, ranging from disabling certain inter-site services, to using external gateways that only allow electronic mail delivery, to gateways that limit remote interactions via access control lists, to disconnection from the Internet. While these measures are preferable to the damage that could occur from security violations, taken to their logical extreme they could eventually reduce the Internet to little more than a means of supporting certain pre-approved point-to-point data transfers. Such diminished functionality could hinder or prevent the deployment of important new types of network services, impeding both research and commercial advancement.

To understand the evolution of this situation, we have designed a study to measure changes in Internet service-level reachability over a period of one year. The study considers upper layer service reachability instead of basic IP connectivity because the former indicates the willingness of organizations to participate in inter-organizational computing, which will be an important component of future wide area distributed applications. The data we gather will be useful for both Internet research and engineering planning activities. They will also be of general interest, as it represents direct measurements of the evolution of a global electronic society.

Clearly, a study of this nature raises a number of potential concerns regarding privacy, security, and network/remote site load. In this note we overview our experimental design, considerations of network and remote site load, mechanisms used to control the measurement collection process, and efforts to inform administrators at sites measured by this study, along with concomitant privacy and security issues.

A point we wish to stress from the outset is that this is not a study of network security. The experiments do not attempt to probe the security mechanisms of any machine on the network. The study is concerned solely with the evolution of network connectivity and service reachability.

Experimental Design

The experiment consists of a set of runs of a program over the span of one to two days each month, repeated monthly for a period of one year (November 1, 1991-November 1, 1992). Each program run attempts to connect to 13 different TCP ports at each of a 12,865 Internet domains worldwide, recording the failure/success status of each attempt. The program attempts no data transfers in either direction. If a connection is successful, it is closed immediately. (Note in particular that this means that the security mechanism behind individual network services is not tested.) The machines on which connections are attempted are selected at random from a large list of machines in the Internet, constrained such that at most 1 to 3 machines is contacted at any particular domain. The ports at which connections are attempted are:

Port Number	Service
13	daytime
15	netstat
21	FTP
23	telnet
25	SMTP
53	Domain Naming System
79	finger
111	Sun portmap
513	rlogin
514	rsh
540	UUCP
543	klogin
544	krcmd, kshell

This list was chosen to span a representative set of services that can be expected to be found on any machine in a domain (so that probing random machines is meaningful). The one exception is port 53, for which the machines to probe are selected from information obtained from the Domain Naming System. Only TCP ports are used, since they allow one to determine if a server is running in an application-independent fashion.

Network and Remote Site Load

The measurement program has been designed to be quite careful to avoid generating unnecessary Internet packets, and to avoid congesting the Internet with too much concurrent activity. Once it has successfully connected to a particular port in a domain, the program never tries that port on any machine in that domain again, for the duration of the current measurement run (i.e., the current month). Once it has recorded 3 connection refusals at any machines in that domain at a port, it does not try that port at that domain again during the current measurement run. If it experiences 3 timeouts on any machine in a domain, it gives up on the domain for that measurement run, possibly to be retried at another time. In the worst case there will be 3 connection failures for each port at 3 different machines, which amounts to 37 connection requests per domain (3 for each of the 12 ports other than the Domain Naming System, and one for the Domain Naming System). However, the average is much less than this.

To quantify the actual Internet load, we now present some measurements from test runs performed in August 1991. In total, 50,549 Domain Naming System lookups were performed, and 73,760 connections were attempted. This measurement run completed in approximately 10 hours, never initiating more than 20 network operations (name lookups or connection attempts) concurrently. The total NSFNET backbone load from all traffic sources that month was approximately 5 billion packets. Therefore, our measurement study increased Internet load by at less .5% on the day that it ran. Because the Internet contains several other backbones beyond NSFNET, the proportionate increase in total Internet traffic was significantly less than .5%.

The cost to a remote site being measured is effectively zero. From the above measurements, on average we attempted 5.7 connections per remote domain. The cost of a connection open/close sequence is quite small, particularly when compared to the cost of the many electronic mail transmissions that any site experiences on a given day.

Control Over Measurement Collection Process

The measurement program evolved from an earlier set of experiments used to measure the reach of an experimental Internet white pages tool called netfind, and has been tested extensively over a period of 2 years. During this time it has been used in a number of experiments of increasing scale. The program uses several redundant checks and other mechanisms to ensure that careful control is maintained over the network operations that are performed. In

addition, we monitor the progress and network loading of the measurements during the measurement runs, observing the log of connection requests in progress as well as physical and transport level network status (which indicate the amount of concurrent network activity in progress). Finally, because the measurements are controlled from a single centralized location, it is quite easy to stop the measurements at any time.

Informing Site Administrators; Privacy and Security Issues

When we ran our initial test runs of this study, we attempted to inform site administrators at each study site about this study, by posting a message on the USENET newsgroup "alt.security" and by sending individual electronic mail messages to each site. We also informed the Computer Emergency Response Team (CERT) at CMU of the study. As a practical matter, informing all sites turned out to be infeasible. Part of the problem was that no channels exist to allow such information to be easily disseminated. Approximately half of the messages we sent out were returned as undeliverable. Moreover, the network traffic and remote site administrative load caused by the study announcement messages far outstripped the network and administrative load required by the study itself. Some sites felt that the announcement was an unnecessary imposition of their time.

In addition to these practical problems, a broad announcement of this study could affect the measurements it attempts to gather. Some sites would likely react to the announcement by changing the reachability of their services. Asking for explicit permission from sites would yield even worse methodological problems, as this would have provided a self-selected study group consisting of sites that are less likely to disconnect from the Internet.

In contrast with our attempts to announce the study, running the study without announcing it caused only a handful of site administrators to notice the traffic and inquire about it to either the CERT or to one of the responsible network contacts at the University of Colorado. The significantly lower remote site administrator and network overhead of running the study unannounced, coupled with the practical and methodological problems of announcing the study, lead us to the decision to run the study without further broad announcements. This decision is in line with the Internet Activities Board's recent policy statement regarding appropriate Internet measurement activity, since the study itself performs only legitimate network connections, and does not impose undue burden on a remote machine or network.

Clearly, the data collected by this study is somewhat sensitive to privacy and security concerns, in the sense that it might be used as a "road map" of accessible network services. We will treat the raw data as private information, publishing measurements only in global statistical terms, divorced from the actual sites that make up the underlying data points. We previously carried out a study with much larger privacy implications than the current study (concerning organizational patterns in the global electronic mail community), and successfully masked the data to protect individual privacy.

For Further Information

Information about the general research program within which these experiments fit is available by anonymous FTP from latour.cs.colorado.edu, in pub/RD.Papers. This directory contains a "README" file that describes the overall research project (which focuses on resource discovery), and includes a bibliography. Particularly relevant are:

- [Schwartz 1991c], a project overview;
- [Schwartz 1991a], about an earlier, simpler version of the current study;
- [Schwartz & Tsirigotis 1991b], about the netfind white pages tool;
- [Schwartz & Tsirigotis 1991a], which considers a number of the techniques used in this experiment, including those for controlling the progress of the measurements;

and

- [Schwartz & Wood 1991], about an earlier study we carried out that raises significant potential privacy questions, for which we carefully masked the underlying data, presenting the results without sacrificing individual privacy.

Also:

- [Cerf 1991], IAB guidelines for Internet measurement activity.

We are currently preparing a proposal to the National Science Foundation about the current study. Once the results are complete, we will publish them in a conference or journal, as well as by anonymous FTP.

Communication With Principal Investigator

If you would like to have your site removed from this study, or you would like to be added to the list of people who receive results from this study, or you would like to communicate with the Principal Investigator for some other reason, please send electronic mail to schwartz@cs.colorado.edu.

References

[Cerf 1991] V. G. Cerf, editor. Guidelines for Internet Measurement Activities. Request For Comments 1262, Internet Activities Board, October 1991.

[Schwartz 1988]
M. F. Schwartz. Autonomy vs. Interdependence in the Networked Resource Discovery Project. Position paper, ACM SIGOPS European Workshop, Cambridge, England, September 1988.

[Schwartz 1989]
M. F. Schwartz. The Networked Resource Discovery Project. *Proceedings of the IFIP XI World Congress*, pp. 827-832, San Francisco, California, August 1989.

[Schwartz 1990]
M. F. Schwartz. A Scalable, Non-Hierarchical Resource Discovery Mechanism Based on Probabilistic Protocols. Technical Report CU-CS-474-90, Department of Computer Science, University of Colorado, Boulder, Colorado, June 1990. Submitted for publication.

[Schwartz & Wood 1991]
M. F. Schwartz and D. C. M. Wood. A Measurement Study of Organizational Properties in the Global Electronic Mail Community. Technical Report CU-CS-482-90, Department of Computer Science, University of Colorado, Boulder, Colorado, August 1990; Revised July 1991. Submitted for publication.

[Schwartz & Tsirigotis 1991a]
M. F. Schwartz and P. G. Tsirigotis. Techniques for Supporting Wide Area Distributed Applications. Technical Report CU-CS-519-91, Department of Computer Science, University of Colorado, Boulder, Colorado, February 1991; Revised August 1991. Submitted for publication.

[Schwartz & Tsirigotis 1991b]
M. F. Schwartz and P. G. Tsirigotis. Experience with a Semantically Cognizant Internet White Pages Directory Tool. *Journal of Internetworking: Research and Experience*, 2(1), pp. 23-50, March 1991.

[Schwartz et al. 1991a]
M. F. Schwartz, D. H. Goldstein, R. K. Neves and D. C. M. Wood. An Architecture for Discovering and Visualizing Characteristics of Large Internets. Technical Report CU-CS-520-91, Department of Computer Science, University of Colorado, Boulder, Colorado, February 1991. Submitted for publication.

[Schwartz et al. 1991b]
M. F. Schwartz, D. R. Hardy, W. K. Heinzman and G. Hirschowitz. Supporting Resource Discovery Among Public Internet Archives Using a Spectrum of Information Quality. *Proceedings of the Eleventh IEEE International Conference on Distributed Computing*

Systems, pp. 82-89, Arlington, Texas, May 1991.

[Schwartz 1991a]
M. F. Schwartz. The Great Disconnection? Technical Report CU-CS-521-91, Department of Computer Science, University of Colorado, Boulder, Colorado, February 1991.

[Schwartz 1991b]
M. F. Schwartz. The Role of Resource Discovery in Support of a National Software Exchange. Position paper, RIACS National Software Exchange Workshop, March 1991.

[Schwartz 1991c]
M. F. Schwartz. Resource Discovery and Related Research at the University of Colorado. *ConneXions - The Interoperability Report*, pp. 12-20, Interop, Inc., May 1991.