SLAC–PUB–7873
April 1999

DISCUSSANT REMARKS ON SESSION: STATISTICAL ASPECTS OF MEASURING THE INTERNET

R. Les Cottrell
Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94309

Abstract

These remarks will briefly summarize what we learn from the talks in this session, and add some more areas in Internet Measurement that may provide challenges for statisticians. It will also point out some reasons why statisticians may be interested in working in this area.

Published in Computing Science and Statistics
Volume 30, Proceedings of the 30th Symposium on the Interface
Minneapolis, MN, May 13 – 16, 1998

Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94309
Work supported by Department of Energy contract DE–AC03–76SF00515.

Discussant Remarks on session: Statistical Aspects of Measuring the Internet

R. Les. Cottrell
SLAC,
Stanford University
Stanford, California, 94309

Abstract

These remarks will briefly summarize what we learn from the talks in this session, and add some more areas in Internet Measurement that may provide challenges for statisticians. It will also point out some reasons why statisticians may be interested in working in this area.

1 Summary of Talks

The talks are very complementary and may be divided into tackling three important aspects of Internet Monitoring, as seen by a philosopher, an artist, and an analyst.

In the first talk, Vern Paxson acting as the philosopher, introduces us to the exciting challenges of Internet monitoring. In particular he identifies the size and tremendous growth (both of which were also well illustrated in many of the transparencies in the following talk by John Quarterman), the diversity and the complexity of the Internet. From there he goes on to identify the need for new statistical approaches, since the older Poisson related methods, used for the connection oriented phone system, do not work for the connectionless oriented Internet. These approaches have to be parsimonious in their approach and require the development of statistically robust invariants. He also shows the tremendous interest in the area.

Following Vern, John Quarterman, acting as the artist, shows some beautiful ways of visualizing information on the Internet going back many years. This information shows the growth (and in some cases such as BITNET the decline) of many areas, and the extent of the Internet, using some innovative Web accessible graphical plots. Such ways of showing data are extremely important given the widespread interest among people who have little knowledge of statistics and need the information to be presented in an easy to comprehend visual manner.

Finally, Darryl Downing, Valerii Fedorov, Yehuda Vardi and Walter Willinger are all acting as the analysts (or more accurately as statisticians) and apply a variety of statistical techniques to model, to analyse and to plan measurements on networks.

Darryl Downing and Valerii Fedorov are concerned with analytical visualization of time series based on such simple statistics as histograms or covariograms, for instance. He developes a matrix of plots that allow, at a visual level, one to identify the presence of systematic trends, compoundness or double stochasticity of the underlying random behavior, etc. For instance, double stochasticity and temporal trends may show the identical histograms, and only matching them with the corresponding covariograms leads to an unique inference.

Optimal sampling for network monitoring is the main topic of the talk presented by Valerii Fedorov (results are obtained together with S. Batsell and D. Flanagan). The emphasis is on statistical methods of optimal monitoring that use the very modest set of assumptions (only the existence of the mean and variance of measured variables is assumed). The relationship between some heuristic approaches in optimal monitoring design and the first order numerical algorithms of convex design theory are explored. Together they lead to the very practical methods which can work with large networks.

The talk presented by Yehuda Vardi is a part of his journal paper on estimation of source-destination traffic intensities from link data and thus are not included in the Proceedings; see Vardi (1996). The main results are derived under Poisson assumptions and two types of traffic-routing regimens: deterministic (a fixed known path) and Markovian (a random path generated by a known Markov chain).

Walter Willinger shows how the application of the latest techniques such as self-similarity, wavelets, multifractals and texture plots can make comprehensible a mathematical description of how Internet sessions work. This description is shown not only to work over a wide

range from microseconds to hundreds of seconds, but is also able to identify the historical impact of the Web on the behavior of Internet sessions. The impact of self-similarity and heavy tails has serious implications to network designers since it makes it difficult to estimate queue lengths and requires new approaches to queue management in network devices such as routers.

2 Further Challenges for Internet Measurements

Most users, even the network managers of the site networks connected to the Internet, have no ability to directly monitor Internet traffic or components at points within the Internet itself. Yet they need to know what is going on in order to: help identify and track problems, to audit the Internet Service Provider (ISP) performance, and to compare services to make price-performance tradeoffs and planning. Thus there is a great need for what is referred to as Internet End-to-end Performance Monitoring to provide the end user with the above capabilities. Typically such measurements are made with existing tools such as Ping and Traceroute (cf. Cottrell (1998a)) and require no prior agreements with the ISPs.

Ping provides an almost universally available ability to inject a packet into the Internet and have it echoed back from a remote node, thus providing measures of round trip time and packet loss. Traceroute provides the ability for intermediate nodes, such as routers, along the path from the measuring node to the remote node, to provide packet loss and round trip response time between the monitoring node and the intermediate node.

2.1 Deducing the internals of the Internet from external measurements

Ping and traceroute measurements made from multiple monitoring nodes to multiple remote nodes may be analyzed to deduce information about the internals of the Internet. Such information might include topology information, bandwidth measures for individual links, and identification of common congestion points. Deducing this information takes careful statistical analysis to extract the required information and could be a fruitful field for a statistician. In some ways this is analogous to deducing the structure of the proton by bombarding it with electrons and looking at how the electrons are scattered.

2.2 Aggregating End-to-end measurement data

There are several organizations and companies that are making fairly extensive end-to-end measurements on the Internet, and the numbers are growing (cf. Cottrell (1998b)). These typically involve a few to several tens of monitoring sites, each of which may monitor many tens to hundreds of remote sites. The measurements are often publicly available and may contain many gigabytes going back several years, and could be a fruitful starting point for statistical analyses.

Given the large number of possible monitor-remote site pairs it is apparent that it is critical to aggregate these pairs into affinity groups that are related to the information the user wants to deduce. Such groups might include remote sites connected to a given ISP, or pairs whose connections pass through a particular Internet eXchange Point (IXP), or remote sites that are in particular geographical regions, or pairs which are in a community of interest such as a set of sites collaborating in a High Energy or Nuclear Physics (HENP) experiment. The task of evaluating useful aggregations given the information required, and selecting the monitoring and remote sites could usefully be helped by statistical expertise.

2.3 Statistical design of experiments to optimize placement

In order to improve the measurements, some groups are starting to deploy special PC-based dedicated measurement stations (see, for example, Adams et al. (1998) and the Surveyor team (1998)) at critical Internet points. Typically these stations will monitor each other across the Internet. It is very important to decide where these stations should be placed to optimize the usefulness of the information with respect to the time to accumulate it, and the amount of storage needed. Such design would identify the number and locations of the stations. There is a wealth of existing publicly available monitoring information that can be used to act as data to the design process.

2.4 Modeling

One needs models of the Internet that can accurately reflect the behavior based on a few relatively simple metrics that can be readily measured. These models are needed to assist in automatic problem identification, diagnosis and repair. They are also needed to aid in prediction, for example knowing the packet loss, and round trip time, can one estimate the bulk throughput to be expected, and how robust/reliable are the estimates. These

results are needed to provide user expectations, to be able to provide just in time upgrades to the network, and to aid in network management. An idea of the current importance attached to the need for such work is illustrated by the front-page headline in the Network World edition of May 11, 1998 that heralded "Network Management (barely) Passing Grade" and made the statement "overarching, age-old problems of automation ...".

3 So why should a statistician care?

I have given you some of the reasons why I, as a network manager, Internet user, and leader of some Internet measurement efforts, wish to bridge the gap between statistics and Internet monitoring and build on each other's expertise. Now let me suggest why you as a statistician might also be interested.

The first reason is for academic curiosity, the satisfaction and challenge of applying your skills to a new intensely interesting area.

The second reason is more aimed at the leaders who need funding and may be interested in the financial food chain of the world's fastest growing source of revenue. The immense interest is causing the creation of many new Internet measurement startups such as Keynote (see www.keynote.com) and Inverse (see www.inverse.net), to be added to the more mature monitoring companies such as MIDS (www.mids.org) and the Automobile Industry effort; see, for instance, Moskowitz (1996). In addition to this there is considerable government interests, several user groups have been set up in the last couple of years to monitor the Internet including groups such as CAIDA, the XIWT/IPWT, the and the ICFA-NTF. In addition to this the ISPs themselves are beginning to make monitoring data available. More information on such companies, groups etc. can be found in Cottrell (1998b). This interest in turn means that there is venture capital available as well as sponsorship and grants.

For the younger folks who are possibly too altruistic to be interested in money, there is still the attraction of being able to light up at a gathering to people you are attracted to, with a commonly interesting topic like the Internet or the Web, rather than being limited to talking about non-linear estimation, or Markov chains etc.

Finally, and I hope this also bears some attraction, there is the need to provide the world with cost-effective access to the information and services it so desperately needs, whether it be email, the Web, eCommerce, Internet voice, video-conferencing, reference information etc.

4 Acknowledgments

I would like to thank Vern Paxson and Val Federov for many useful discussions and for help in reviewing this paper. John Quarterman and Walter Willinger also kindly provided me with early versions of their presentations and helped me to understand them.

References

A. Adams, J. Mahdavi, M. Mathis and V. Paxson (1998), "Creating a Scalable Architecture for Internet Measurements," *http://www.psc.edu/ mahdavi/nimi_paper/NIMI.html.*

R. Les Cottrell (1998a), "Traceroute Servers for HENP and ESnet," *http://www.slac.stanford.edu/ comp/net/ wanmon/traceroute-srv.html.*

R. Les Cottrell (1998b), "Internet Monitoring Sites," *http://www.slac.stanford.edu/comp/net/wanmon/ netmon.html.*

R. Moskowitz (1996), "Position Statement on AIAG for the NSF Workshop on INternet Statistics Measurements and Analysis," *http://www.nlanr.net/ ISMA/Positions/ moskpwitz.html.*

The Surveyor Team (1998), "About Surveyor," *http://www.advanced.org/csg-ippm/.*

Y. Vardi (1996), "Network Tomography: Estimating Source-Destination Traffic Intensities From Link Data," *JASA*, **91**, 365-377.