

## Visualizing Very Large Internet Traffic Databases

Don Sun and William S. Cleveland  
AT&T Laboratory

Data collection on Internet wires can result in large databases (VLDBs) because packets come across Internet wires continuously at high speed. To exploit a VLDB we need to do more than simply compute summary statistics; we need to study the data in detail and in its full complexity. To help achieve this we developed S-Net, a traffic measurement and analysis system that begins with packet header collection on network wires, and ends with data analysis on a cluster of Linux PCs running S/S-PLUS, a language and system for organizing, visualizing, and analyzing data.

Visualization tools have been vital in our traffic analyses. One issue is screen real estate. Because the databases are very large and the structure, complex, we must accept the notion that single displays need to cover tens and perhaps hundreds of pages with many panels on each page. Data visualization is often limited to a display of a set of data that can be placed all at once in our visual field. So it can be shocking at first to contemplate looking at so many pages. But using the structure of trellis display, a visualization system available in S/S-PLUS, it is easy to generate many pages. And using a document viewer, it is possible to learn a great deal about a VLDB from these multipage, multipanel displays.

The intensity of analysis and the extensive visualization have led to a number of new developments including the following: (1) Traffic variables on an Internet wire exhibit a pervasive nonstationarity. As the TCP connection rate changes, marginal distributions and long range dependence change. The cause of the nonstationarity is superposition: the intermingling of sequences of connections between different source-destination pairs, and the intermingling of sequences of packets from different connections. (2) The burstiness of Internet traffic was established in pioneering work in the early 1990s, which demonstrated that packet arrival times are not Poisson, and packet and byte counts in fixed-length intervals are long-range dependent. Our study of Internet traffic has shown that these results are one end of a continuum of traffic characteristics. At the other end are Poisson behavior and independence. (3) PackMime, an IP traffic model, generates synthetic Internet packet traffic that mimics the behavior of live traffic on a TCP/IP network. Validation of PackMime has been achieved through a set of statistical tools for analyzing packet processes that are applied to live traffic and to synthetic traffic from PackMime.

For more information, see [cm.bell-labs.com/stat/InternetTraffic](http://cm.bell-labs.com/stat/InternetTraffic).