LOG-NORMAL DURATIONS CAN GIVE LONG RANGE DEPENDENCE

JAN HANNIG
Colorado State University

J.S. MARRON
Cornell University and
University of North Carolina

GENNADY SAMORODNITSKY
Cornell University

F.D. SMITH
University of North Carolina

Size distributions for internet connections are fit using a novel visualization. While no standard distribution is exactly right, both heavy tail Pareto and light tail log-normal distributions appear sensible in the tails. As noted by Downey (2000), goodness of fit of the log-normal raises interesting questions about the widely accepted view of internet traffic, that only heavy tailed duration distributions lead to long range dependence. Some non-standard mathematical analysis reveals that both tail distributions are actually consistent with long range dependence, because with appropriate choice of parameters a system with log-normal durations can have correlation consistent with long range dependence over a wide range of lags.

1. Introduction

A number of studies of internet traffic suggest that internet flows (transfers of data from one computer to another one) often have heavy tailed duration distributions, and that the aggregated traffic (e.g., the collection of all data flowing through a particular point on the internet) exhibits long range dependence, see, e.g., Garrett and Willinger (1994) and Paxson and Floyd (1995). An elegant mathematical theory, see, e.g., Mandelbrot (1969), Cox (1984), Taqqu and Levy (1986) and Heath, Resnick and Samorodnitsky (1998), provides a convincing connection between these phenomena.

A graphical illustration of this behavior is given in Figure 1, where IP (Internet Packet) flows are represented as horizontal lines. The heights of the lines are random, which allows simple visual separation. Details of the data are given below, but a striking feature is that the lengths of the lines include many very short flows, and also some very long flows.

The data shown in Figure 1 were gathered from packet headers, during approximately a 40 minute period on a Sunday morning in 2000, at the main internet link of the University of North Carolina, Chapel Hill. This time period was chosen as being "off peak," having relatively light traffic. An IP "flow" is defined here as the time period between the first and last packets transferred between a given pair of IP sending and receiving addresses. For more details on the data collection and processing methods, see Smith, Hernandez, Jeffay and Ott (2001).