۱-برداشت خودتان را متن زیر بنویسید

۲- بخش فرمول ۲-۲۲ و مبحث خود متشابهی را به دقت بیان کنید

## 2.3 MODELING NETWORK TRAFFIC

In this section we consider the problem of describing and modeling the distributions of various statistics of network traffic. In the previous section, we looked at a particular kind of inference, determining the traffic loads between source and destination pairs. Here we are more interested in characterizing network traffic in general.

The previous work assumed that the traffic intensity was distributed as a Poisson random variable. As we will see, this assumption breaks down as we investigate network traffic in more detail.

Network traffic is quite complex. For example, consider Web traffic. When you type in a URL to your browser, many things that you are mostly unaware of happen behind the scenes. We have seen that the TCP handshaking is used to initialize the connection. Data pass back and forth until the session is closed. However, a Web session usually consists of more than just one session. Even if you only go to a single page, you will probably initiate several sessions. Each image you download is a separate session, as are other files that may be loaded by the page. After a few seconds of reviewing the page, you then select a new URL and the process starts over. In addition to these explicit sessions, there are generally DNS lookups that occur to obtain the IP addresses of the pages and images.

Thus, there are several levels to the network data. Within each TCP session there are the individual packets, whose statistics are determined primarily by the network and hardware considerations. The number of sessions spawned by the main session is different for each Web page and thus has another distribution associated with it. Finally, there is the user input, which determines the time between new sessions.

I will consider several models for network traffic. These are by no means a complete listing of the work done on this problem. I will provide references to other work, so interested readers can learn more.

In the simplest case, we consider the initiation times of TCP user sessions. A basic probability course would tell us that arrival times are Poisson, so this is a good starting place. Note that it is also reasonable to believe that the rate is slowly varying (diurnal), and in fact several groups (Paxson and Floyd [1995], Nuzman et al. [2000]) have found that (some) user session initiation times are well modeled by a Poisson distribution with a slowly time-varying rate. This model breaks down, however, if one considers other factors such as the data transfer times for FTP traffic or activity that spawns new activity, such as news and Web transfers.

It is shown in Paxson and Floyd [1995] that telnet and FTP session arrival times are well-modeled by the Poisson process (Equation (2.9)). Recall that a Poisson process is one where the counts within fixed time intervals are distributed as Poisson. Thus, when talking about continuous random variables as being "Poisson," we mean they come from a Poisson process. In the same paper, the claim is made that email, Web and news are not Poisson.

Let us look at some data. Figure 2.3 shows data collected over a two-hour period. The interarrival times between connection requests to a mail server are plotted against arrival times. Notice the burstiness of the data (gaps in the plot). The data are quite correlated (with a correlation of 0.33 in this case), as can be

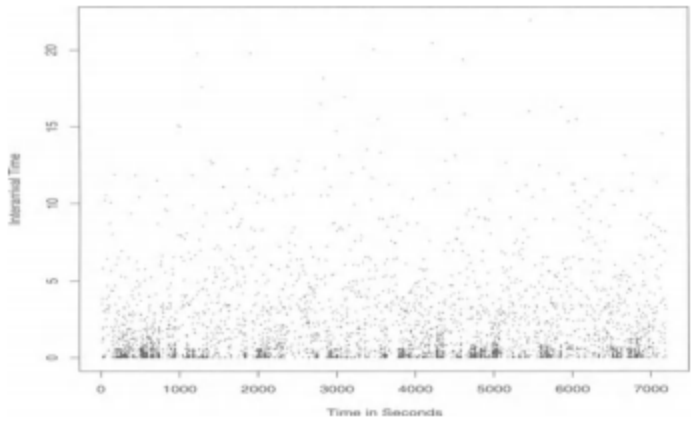


***Fig. 2.3*** Scatter plot of the interarrival times of connections to a mail server. The data were collected over a two hour time period.

seen in Figure 2.4. This tells us that the data are not independent - that the time between packets now is to an extent related to the time between packets in the recent past.

Figure 2.5 shows another view of these data. Here, the interarrival times between connection requests are depicted as a histogram with a gamma distribution

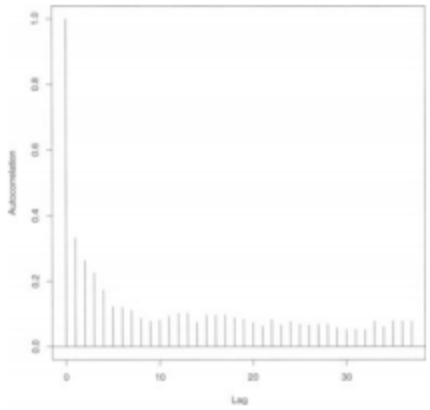


***Fig. 2.4*** Autocorrelation of the data in Figure 2.3.

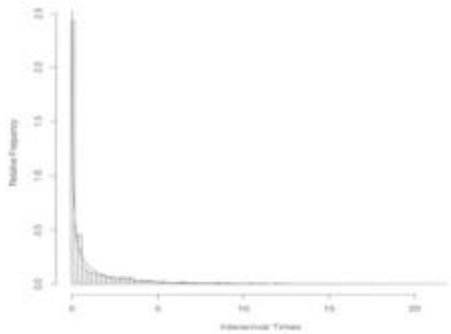

***Fig. 2.5*** Histogram of the interarrival times of connections to a mail server. The data were collected over a two hour time period. The times are in seconds. A gamma distribution fit to the data is shown as a curve overlaid on the histogram.

fit to the data,

$$\gamma(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}. \tag{2.21}$$

The parameters of the gamma distribution fit are $\alpha = 0.278$ and $\beta = 4.241$. There are 6104 observations in these data. Figure 2.6 depicts the same graph zoomed in to the range from 0 to 5 seconds. As can be seen, the fit is not too bad, although there seems to be something interesting happening at around 1/2 second.

Paxson [1994] calculates statistics for several different quantities, such as number of bytes per session and duration of session for several different applications

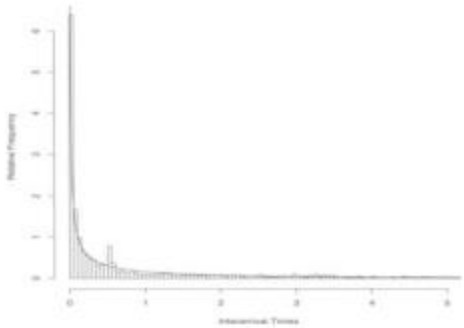

***Fig. 2.6*** Histogram of the interarrival times of connections to a mail server (Figure 2.5) zoomed in to the range from 0 to 5 seconds. A gamma distribution fit to the data is shown as a curve overlaid on the histogram.

(telnet, email, news, and FTP). He fits a number of distributions to the data, performing tests of fit to determine the best fit. The distributions investigated are:

$$\begin{aligned} \text{Extreme: } f(x) &= \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} e^{-e^{-(x-\alpha)/\beta}}; \\ \text{Pareto: } f(x) &= \frac{\alpha k^{\alpha}}{x^{\alpha-1}}; \\ \text{Exponential: } f(x) &= \frac{1}{\beta} e^{-x/\beta}. \end{aligned}$$

In addition, he considers the lognormal and log-extreme distributions. A random variable is said to have a log-f distribution if $\log(x)$ has distribution $f$. In this case, the logarithms are taken base 2.

Using a series of tests based on the $\chi^2$ test, a "best fit" distribution is found for various quantities. Some of these are reproduced in Table 2.2. These data represent the analysis of 3 million connections. In the table, "originator bytes" refers to the number of bytes per packet from the originator of the session and similarly for "responder bytes." In FTP, data transfers often occur in "bursts," sessions that occur less than 4 seconds apart. These can be caused by multiple gets or puts, in which several files are transferred in rapid succession. The number of bytes in these burst sessions is referred to as the "burst bytes" in the table.

The conclusion of the Paxson [1994] paper is that although the models are not perfect, they do a good job of approximating the empirical distributions of the network data investigated. Further, different quantities are distributed according to different families of distributions, and none of the quantities considered was well-modeled as a Poisson process.

As we saw in our small example (Figures 2.3–2.6), network data are complicated, even if we restrict our investigation to relatively simple problems such as session interarrival times. Several authors have investigated these issues and found interesting structure. Leland et al. [1994] were among the first to comment on the self-similar nature of network traffic. They first illustrate the self-similarity graphically by noting that the traffic looks similar at different scales. Our little 6000 point data set is not large enough to provide evidence nearly as convincing as in Leland et al. [1994]; however we can illustrate it on a small scale (Figure 2.7).

***Table 2.2*** Models selected in Paxson [1994] for various traffic quantities.

| Protocol | Variable | Model |
|---|---|---|
| Telnet | Originator bytes | log-extreme |
| | Responder bytes | log-normal |
| | Duration in seconds | log-normal |
| NNTP | Originator bytes | log-normal |
| SMTP | Originator bytes | log-normal |
| FTP | Connection bytes | log-normal |
| | Session bytes | log-normal |
| | Burst bytes | Pareto |

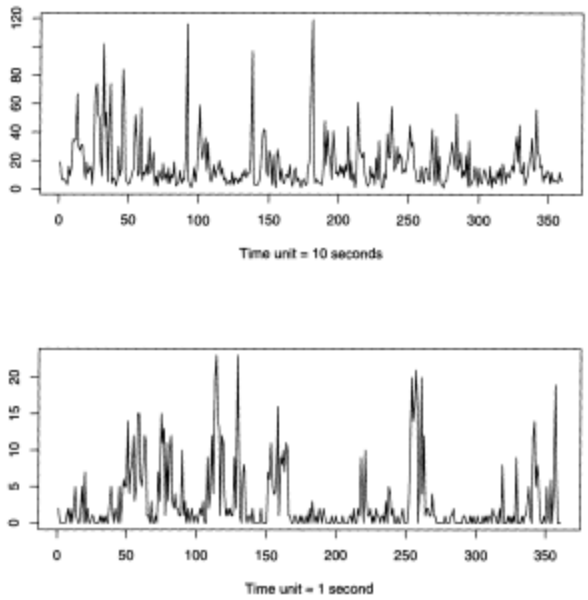


***Fig. 2.7*** Two views of the data from Figure 2.3. The number of packets per 10 seconds is depicted in the top graph, while the number of packets per second is depicted on the bottom.

This shows some evidence of self-similarity at two scales. Leland et al. [1994] show similar plots across five scales, ranging from a time unit of 100 seconds down to one of 1/100 of a second. This is very subjective and to strengthen the argument Leland et al. [1994] show similar plots for a synthetic data set based on a compound Poisson model fit to the data, for which self-similarity is not evident.

The concept of self-similarity can be made precise. Recall that the autocorrelation function of a process $X = X_1, X_2, \ldots$ is defined to be

$$r(k) = \text{cov}(X_t X_{t+k}). \tag{2.22}$$

Assume that $X$ has a finite variance $\sigma^2$ and that $r(k) \sim k^{-\beta} L(t)$ as $k \to \infty$ and $L$ is asymptotically constant. Processes with these properties are called self-similar. Define the aggregate time series $X^{(m)}$ to be the average of $X$ taken over nonoverlapping blocks of size $m$: $X_k^{(m)} = 1/m(X_{km-m+1} + \cdots + X_{km})$. Denote the autocorrelation function of this process by $r^{(m)}(k)$. We say that $X$ is second-order self-similar if $r^{(m)}(k) = r(k)$. Similarly, we call the process asymptotically second-order self-similar if the autocorrelation function of the aggregate time series converges to $r(k)$ as $m$ goes to infinity. The parameter $\beta$ gives a measure of the self-similarity, usually defined to be the Hurst exponent $H = 1 - \beta/2$.

Several papers describe the self-similar nature of network traffic. The interested reader is encouraged to investigate Feldmann et al. [1998], which analyzes data collected over several years, Crovella and Bestavros [1997], which looks at self-similarity for Web traffic, and Willinger et al. [1997] and Yang et al. [1999], who describe and apply the so-called "On/Off" model and variants to network data.