Characterizing the Variability of Arrival Processes with Indices of Dispersion

Riccardo Gusella

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, California 94704

TR-90-051
September 26, 1990

ABSTRACT

We propose to characterize the burstiness of packet arrival processes with indices of dispersion for intervals and for counts. These indices, which are functions of the variance of intervals and counts, are relatively straightforward to estimate and convey much more information than simpler indices, such as the coefficient of variation, that are often used to describe burstiness quantitatively.

We define and evaluate the indices of dispersion for some of the simple analytical models that are frequently used to represent highly variable processes. We then estimate the indices for a number of measured point processes that were generated by workstations communicating to file servers over a local-area network.

We show that nonstationary components in the measured packet arrival data distort the shape of the indices and propose ways to handle nonstationary data. Finally, to show how to incorporate measures of variability into analytical models and to offer an example of how to model our measured packet arrival processes, we describe a fitting procedure based on the index of dispersion for counts for the Markov-modulated Poisson process.

I. INTRODUCTION

Since the first analyses of computer traffic in the mid- and late 1970s [10, 19], which showed that packet arrival processes are highly variable, researchers have frequently described computer communication patterns as ‘‘bursty’’. Yet few have bothered to define burstiness. Most seem to invoke the term bursty when confronted with processes whose interarrival time distributions show greater variability than Poisson processes. (Exceptions do exist; see for instance [13], in which a distribution with coefficient of variation 0.3 is termed bursty.) The vagueness surrounding the concept of burstiness stems both from its use to denote different types of variability in many disparate situations and from the difficulty of characterizing in meaningful ways the capricious nature of packet arrivals.

The variability of packet arrival processes is strikingly manifested in the following Figures 1 and 2, which represent respectively the times between subsequent arrivals and the times between every four arrivals for the messages sent by a single-user workstation to its file server over a local-area network. In each figure, the logarithm of the interarrival time is on the ordinate and the serial number on the abscissa, and artificial zero lines have been placed at the median values of the series. This graphical arrangement allows us to view short interarrival times, which would otherwise be obscured by long ones: the former are printed toward the bottom of the graph and the latter toward the top.

FIGURE 1. TIMES BETWEEN SUCCESSIVE ARRIVALS

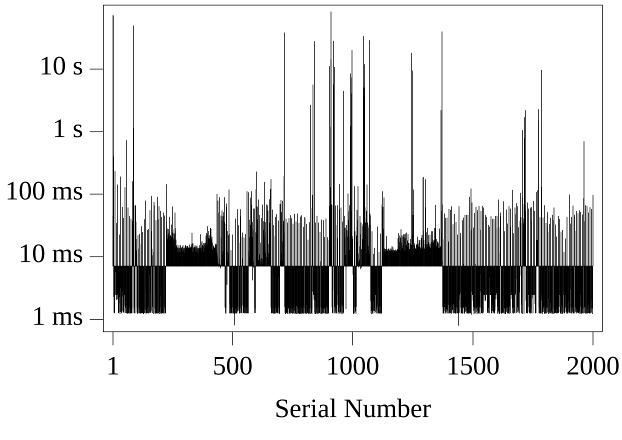
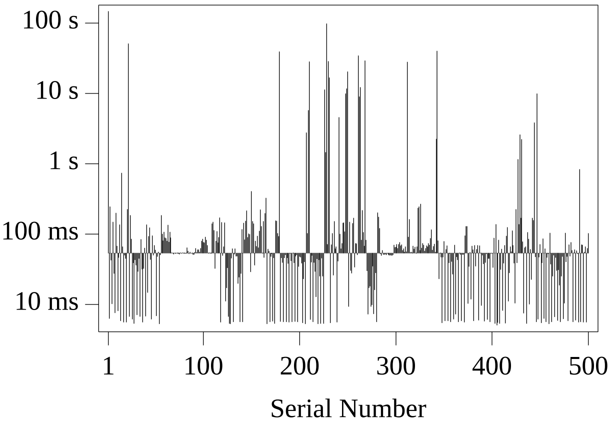


FIGURE 2. TIMES BETWEEN EVERY FOUR ARRIVALS



The variance of the interarrival times in Figure 2 is about six times larger than the variance of those in Figure 1. Clearly, the larger dispersion of values in the second figure stems from the clustering of small and large interarrival times in separate groups in the first figure. This bunching is caused by protocol features, such as fragmentation of large messages, that generate short interarrivals in batches. Similarly, a sequence of related remote procedure calls may generate a group of longer interarrivals because, for each request, the destination has to execute a procedure before returning the answer, which in turn will trigger the next request. Thus, the increase in variance is related to the temporal structure in the data and is not captured by simple burstiness indices such as the coefficient of variation, the peak-to-average ratio, an so on. In this paper we will use the term variability to refer explicitly to changes in the variance of the sum of consecutive interarrivals or in the variance of arrival counts over larger and larger time intervals.

Variability in packet arrivals has been connected to the queuing delays packets are subject to: the general rule is that more variability corresponds to longer delays. However, except in a few simple cases, the precise relationship between variability and queuing delays is difficult to represent analytically. Several attempts to resolve this issue have been made; in particular, Fendik and Whitt's approach [5], which uses a statistical index that models the variability of the arrival process but also captures the dependency between the interarrival and the service times, is worth mentioning.

In this paper we take a more narrow, focused approach: we characterize the variability of measured packet arrival processes with indices of dispersion functions and discuss the merits of these indices as well as the pitfalls of their indiscriminate use. Indices of dispersion have long been known in the statistics community as a powerful tool in the analysis of the second-order properties of point processes [1, 3], but, despite the flourishing in recent years of measurements and analyses of computer traffic data (for a survey see [16]), they have been rarely, if ever, applied to computer traffic measurements. Here, we demonstrate that indices of dispersion are valuable and valid tools for characterizing the variability of packet arrival processes. We also discuss how standard analytical models should be fitted to traffic measurements in order to take into account the variability of the data.

This paper is organized into three major parts. In Section II, we define the index of dispersion for intervals and the index of dispersion for counts, and review their basic properties. We then calculate one of these two indices for each of three classes of analytical models that are often used to represent bursty point processes: renewal models with hyperexponential interarrival times, batch Poisson processes, and Markov-modulated Poisson processes. Although the results are in most cases not new, the exercise serves the important purpose of clarifying the meaning and use of indices of dispersion. In Section III, we estimate the indices of dispersion for several measured packet arrival processes generated by single-user workstations communicating with file servers over a local-area network. We show how nonstationary data introduce difficulties, and suggest that semi-Markov models may model accurately both short- and long-term variability. Finally, in Section IV, as an example of how standard arrival models can incorporate the variability that we have analyzed, we develop a procedure to fit a Markov-modulated Poisson process to our arrival processes.

II. INDICES OF DISPERSION

A. The Index of Dispersion for Intervals

Let us consider first describing point processes in terms of the lengths of the intervals between subsequent arrivals. For packet-arrival processes, these intervals, which we will call *interarrival times*, are defined as the length of time between the beginning of the transmission of a given packet and the beginning of the transmission of the previous packet. (Cox and Miller [2, page 339] define the same quantity more esoterically as *backward recurrence-time*.) Notice that under this definition the transmission time of the previous packet is included in the interarrival time; thus, for networks that allow variable packet sizes, this definition introduces a source of dependency between intervals and packet lengths. (A constant packet size imposes a lower bound on interarrival times, but not the dependency.)

The variance of the sum of two random variables depends on the covariance between them, and, if they have common variance, is given by

$$\mathrm{var}(X_{i+1} + X_{i+2}) = 2\mathrm{var}(X) + 2\mathrm{cov}(X_{i+1}, X_{i+2}) ,$$

and, in general, for the sum of n variables we have

$$\mathrm{var}(X_{i+1} + \cdots + X_{i+n}) = n\ \mathrm{var}(X) + 2\sum_{j=1}^{n-1}\sum_{k=1}^{j}\mathrm{cov}(X_j, X_{j+k}) . \qquad (1)$$

We have indicated with $\mathrm{var}(X)$ the common variance of the X_i (we will also write $E(X)$ for the common mean), and thus have assumed implicitly that the processes under consideration are at least weakly stationary, i.e., that their first and second moments are time invariant, and that the autocovariance series depends only on the distance k, the *lag*, between samples: $\mathrm{cov}(X_i, X_{i+k}) = \mathrm{cov}(X_j, X_{j+k})$, for all i, j, and k.

It is the dependency on the autocovariance, or, equivalently, on the autocorrelation, that makes the variance of the sum of intervals useful in describing arrival processes. In fact, in situations like those of Figure 1 and Figure 2, in which interarrivals smaller than the mean as well as interarrivals larger than the mean are grouped together, the covariance will assume positive values.

We will use the variance above, normalized by the factor $n\,E^2(X)$, as a measure of the variability of packet arrival processes. The sequence of values

$$J_n = \frac{\mathrm{var}(X_{i+1} + \cdots + X_{i+n})}{n\,E^2(X)} , \qquad (2)$$

with $n = 1, 2, \cdots$, is called *index of dispersion for intervals* (IDI).

Notice that J_1 is $C_J^2 = \mathrm{var}(X)/E^2(X)$, the squared coefficient of variation for intervals. As a result of the normalization, for a Poisson process J_n has constant value 1 for all n; for a renewal process, whose interarrival times are identical and independently distributed (i.i.d.), J_n is also a constant in n of value C_J^2. Using equation (1) and the definition $\rho_n = \mathrm{cov}(X_i, X_{i+n}) / \mathrm{var}(X)$, we can express relation (2) in terms of the autocorrelation coefficients at lag n:

$$J_n = C_J^2 \left[1 + 2\sum_{j=1}^{n-1}(1 - \frac{j}{n})\rho_j \right] , \qquad (3)$$

which shows that point processes with positive correlation coefficients have monotonically increasing

IDI curves. Notice also that the limit of equation (3), when it exists, is proportional to the sum of all correlation coefficients (plus 1), that is,

$$\lim_{n \to \infty} J_n = C_J^2 \left(1 + 2 \sum_{j=1}^{\infty} \rho_j \right) . \tag{4}$$

The asymptote in the limit above depends on the sum (integral) of all the correlation coefficients. Since the interarrival times typically become statistically independent as the lag increases (the causation effects triggered by a packet transmission, such as additional queuing and increased disk activity, diminish as time increases), making their autocorrelation coefficients decrease to 0, for practical purposes the limit will be reached for a finite value of j. As noted above, packet-arrival processes normally have positive autocorrelation coefficients since both interarrivals shorter than the mean interarrival time and those longer than the mean interarrival time tend to occur in separate bursts. In packet-arrival processes, we would thus expect the IDI sequence to increase with n. Notice that, if the data are not stationary, we can still compute an estimate of J_n; however, equations (1), (3), and (4) are no longer generally valid.

B. The Index of Dispersion for Counts

We can also analyze point processes from the perspective of packet counts–the number of packets in an interval. We can define for packet counts a function similar to the index of dispersion for intervals. The *index of dispersion for counts* (IDC) is the variance of the number of arrivals in an interval of length t divided by the mean number of arrivals in t:

$$I_t$$