On generating self-similar traffic using pseudo-Pareto distribution

It has been shown in the literature that self-similar or long-range-dependant (LRD) network traffic can be generated by multiplexing several sources of Pareto-distributed ON and OFF periods. In a context of a packet-switched network the ON periods correspond to packet train – packets transmitted back to back, or separated only by a relatively small preamble (as defined in IEEE standard 802.3, for example). OFF periods are the periods of silence between packet trains.

Multiple sources contributing to resulting synthetic traffic trace may be thought of as individual flows (connections). It is reasonable to assume that packet sizes within a connection remain constant. Different connections, however, will have packets of different sizes.

To generate a Pareto-distributed sequence of ON periods, one can generate a Pareto-distributed sequence of packet train sizes. The minimum train size is 1, which corresponds to a single packet transmitted.

Pareto distribution has the following probability density function:

$$P(x) = \frac{ab^{a}}{x^{a+1}} \quad , \quad x \geq b \qquad (1)$$

Where α is a shape parameter (tail index), and b is minimum value of x. When $a \leq 2$, the variance of the distribution is infinite. When $a \leq 1$, the mean value is infinite as well. For self-similar traffic, α should be between 1 and 2. The lower the value of α, the higher the probability of an extremely large x. Figure 1 shows the density functions for various values of α.

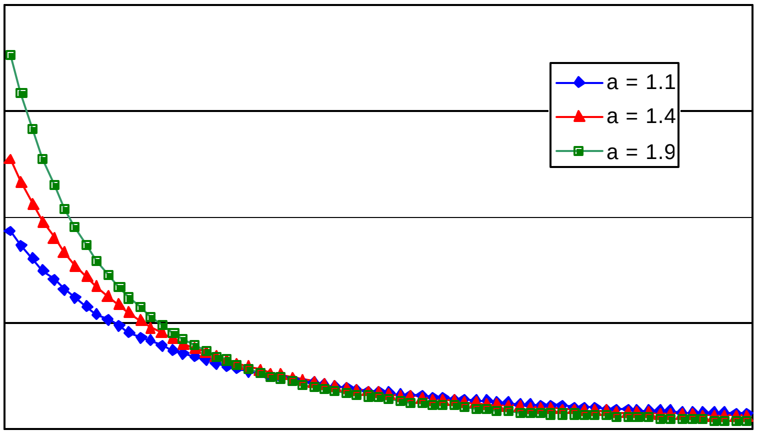

Figure 1. Probability density functions for Pareto distribution with a **= 1.1, 1.4, 1.9**

Mean value of a Pareto distribution equal is

$$E(x) = \frac{ab}{a-1}. \tag{2}$$

The formula to generate a Pareto distribution is

$$X_{PARETO} = \frac{b}{U^{1/a}} \tag{3}$$

where U is a uniformly distributed value in the range (0, 1]

Very often it is desirable to generate a synthetic traffic of a predefined load. Obviously, the resulting load *L* is just a sum of loads L_i generated by each individual source *i*. Given N sources,

$$L = \sum_{i=1}^{N} L_i \tag{4}$$

Thus, it is important to be able to get a good estimation of the load generated by one source. The load generated by one source is mean size of a packet train divided over mean size of packet train and mean size of inter-train gap, or putting it differently, it is a mean size of ON period over mean size of ON and OFF periods.

$$L_i = \frac{\overline{ON_i}}{\overline{ON_i} + \overline{OFF_i}} \tag{5}$$

Formula (2) gives the mean value of a true Pareto distribution. However computers using equation (3) generate a pseudo-Pareto distribution. One of problems comes from the fact that computers cannot generate arbitrarily large value. However, any true Pareto distribution of sufficiently large length will have values that exceed the range generated by computers. Thus, what we have is a truncated-value distribution.

Let's denote S to be the smallest non-zero value that uniform random generator may produce. Then, the generated Pareto-distributed values will not exceed q:

$$q = \frac{b}{S^{1/a}} \tag{6}$$

Then, the mean value of a Pareto distribution can be calculated as shown below:

$$\begin{aligned} E(x) &= \int_b^q x\ f(x)\ dx = \int_b^q x \frac{ab^a}{x^{a+1}} dx = ab^a \int_b^q \frac{dx}{x^a} \\ &= ab^a \left. \frac{x^{1-a}}{1-a} \right|_b^q = \frac{ab}{a-1}\left[1 - \left(\frac{b}{q}\right)^{a-1}\right] \end{aligned} \tag{7}$$

Substituting (6) into (7) we get

$$E(x) = \frac{ab}{a-1}\left[1 - S^{\frac{a-1}{a}}\right] \tag{8}$$

Equation (8) gives the mean value of a truncated-value Pareto distribution. Now, if we are given load Li and the packet size k for a given source, we can find the minimum value of the OFF period.

First, lets find the mean value of OFF period. From equation (2) we get

$$\overline{OFF_i} = \overline{ON_i}\left(\frac{1-L_i}{L_i}\right) \tag{9}$$

Let's denote M_{OFF} and M_{ON} to be the minimum ON and OFF periods respectively. We mentioned above that the minimum packet train size is just one packet, i.e., $M_{ON} = 1$

Then,

$$\frac{M_{OFF} a_{OFF}}{a_{OFF} - 1}\left[1 - S^{\frac{a_{OFF}-1}{a_{OFF}}}\right] = k \frac{M_{ON} a_{ON}}{a_{ON} - 1}\left[1 - S^{\frac{a_{ON}-1}{a_{ON}}}\right]\left(\frac{1-L_i}{L_i}\right) \tag{10}$$

where α_{ON} is the shape parameter for the ON periods, and α_{OFF} is the shape parameter for the OFF periods.

Denoting $T_{ON} = \frac{a_{ON} - 1}{a_{ON}}$ and $T_{OFF} = \frac{a_{OFF} - 1}{a_{OFF}}$, we get

$$M_{OFF} = k\frac{T_{OFF}}{T_{ON}} \times \frac{1-S^{T_{ON}}}{1-S^{T_{OFF}}} \times \left(\frac{1}{L_i} - 1\right) \qquad (11)$$

Thus, given values for k, L_i, α_{ON}, and α_{OFF} , the formula (10) gives us the value for M_{OFF} that would result in link load closer to L_i.

However, if we generate traffic using the above formulas, we will notice the mean values for ON and OFF periods in the generated series still slightly off.

The problem appears to be in the way computers generate Pareto-distributed values (formula 3). While Pareto distribution assumes continuous sample space, computers generate discrete values with uniform probability. The Pareto-like distribution is achieved by having higher density of samples toward lower end of the scale. The Figure 2 illustrates this idea. It shows probability distribution function for the pseudo-Pareto distribution. Note that every value has exactly the same probability of being chosen. Also note that there are no values between 12 and 16, or 16 and 26.

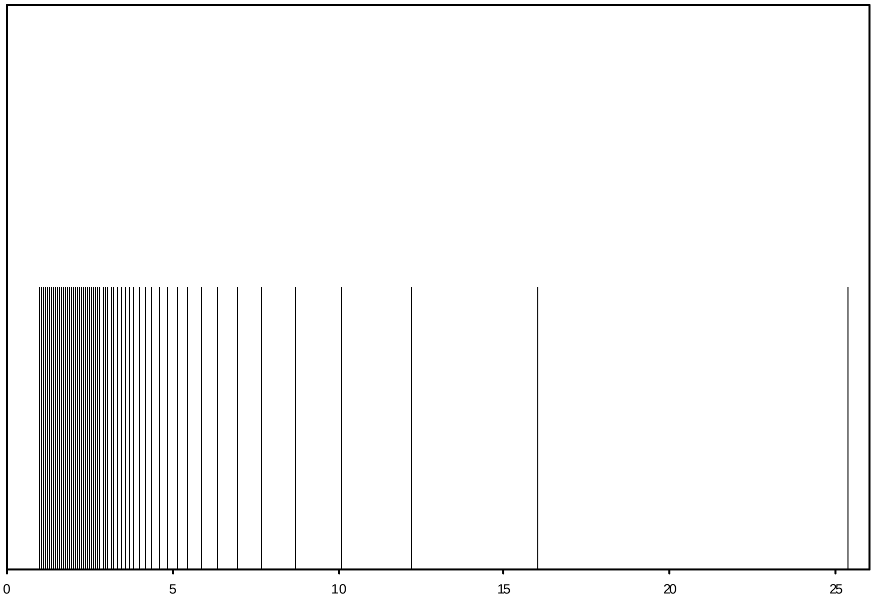

Figure 2. Probability density function for the pseudo-Pareto distribution

If we build distribution function by aggregating samples over some window size, we will get a plot somewhat close to the one shown in Figure 1. But still, no matter how large our window is, at the tail end the distance between two neighboring points will exceed the window size. That means that some windows will contain zero samples, even if number of samples approach infinity. Of course, that introduces an error to the mean size of ON and OFF periods.

To correct for this error, we found that the calculated values $\overline{ON}$ and $\overline{OFF}$ should be multiplied by coefficient C

$$C = (1.19a - 1.166)^{-0.027} \qquad (12)$$

Thus, formula (11) becomes

$$M_{OFF} = k \times \frac{C_{ON}}{C_{OFF}} \times \frac{T_{OFF}}{T_{ON}} \times \frac{1 - S^{T_{ON}}}{1 - S^{T_{OFF}}} \times \left(\frac{1}{L_i} - 1 \right) \qquad (13)$$

On a final remark, if we choose α_{ON}, and α_{OFF} to be the same, the equation (13) will reduce to

$$M_{OFF} = k \times \left(\frac{1}{L_i} - 1 \right) \qquad (14)$$

That, however, may limit the usefulness of the traffic generator. It is very reasonable to assume that in real traffic, probability of having extremely large OFF period is higher then the probability of having extremely large ON period. That means that the shape parameter α_{ON} should be larger than α_{OFF} . The above heuristic coefficient results in generated load being very close to the specified load with all combinations of α_{ON}, and α_{OFF} .