Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples*

András Horváth, Miklós Telek

Dept. of Telecommunications, Budapest University of Technology and Economics,
{horvath,telek}@webspn.hit.bme.hu

Abstract. In order to support the effective use of telecommunication infrastructure, the "random" behavior of traffic sources has been studied since the early days of telephony. Strange new features, like fractal like behavior and heavy tailed distributions were observed in high speed packet switched data networks in the early '90s. Since that time a fertile research aims to find proper models to describe these strange traffic features and to establish a robust method to design, dimension and operate such networks.
In this paper we give an overview of methods that, on the one hand, allow us to capture important traffic properties like slow decay rate, Hurst parameter, scaling factor, etc., and, on the other hand, makes possible the quantitative analysis of the studied systems using the effective analysis approach called matrix geometric method.
The presentation of this analysis approach is associated with a discussion on the properties and limits of Markovian fitting of the typical non-Markovian behavior present in telecommunication networks.

1 Introduction

In the late 80's, traffic measurement of high speed communication networks indicated unexpectedly high variability and burstiness over several time scales, which indicated the need of new modeling approaches capable to capture the observed traffic features. The first promising approach, the fractal modeling of high speed data traffic [28], resulted in a big bum in traffic theory. Since that time a series of traffic models were proposed to describe real traffic behavior: fractional Gaussian noises [30, 37], traditional [7] and fractional ARIMA processes [18], fractals and multifractals [49, 13], etc.

A significant positive consequence of the new traffic engineering wave is that the importance of traffic measurement and the proper statistical analysis of measured datasets became widely accepted and measured datasets of a wide range of real network configurations became publicly available [52].

In spite of the intensive research activity, there are still open problems associated with these new traffic models:

* This work is supported by the OTKA-T34972 grant of the Hungarian Research Found.

- None of the traffic models is evidently verified by the physical behavior of the networks. The proposed models allow us to represent some of the features of data traffic, but some other features are not captured. Which are the important traffic features?
- The traffic features of measured data are checked via statistical tests and the traffic features of the models are checked using analysis and simulation methods. Are these tests correct enough? Is there enough data available for reliable tests?
- The majority the proposed traffic models has important asymptotic properties, but all tests are based on finite datasets. Shall we draw consequence on the asymptotic properties based on finite datasets? And vice-versa, shall we draw consequence from the asymptotic model behavior on the performance of finite systems.
- Having finite datasets the asymptotic properties extracted from tests performed on different time scales often differ. Which is the dominant time scale to consider?

The above listed questions refer to the correctness of traffic models. There is an even more important issue which determines the utility of a traffic model, which is computability. The majority of the mentioned traffic models are not accompanied with effective analysis tools which would allow us to use them in practical traffic engineering.

In this paper we discuss the application of Markovian models for traffic engineering. The most evident advantage of this modeling approach with respect to the above mentioned ones is that it is supported with a set of effective analysis techniques called matrix geometric methods [34, 35, 27, 29]. The other features of Markovian models with respect to the answers of the above listed questions are subjects to discussion. By the nature of Markovian models, non-exponential asymptotic behavior cannot be captured, and hence, they are not suitable for that purpose. Instead, recent research results show that Markovian models are able to approximate arbitrary non-Markovian behavior for an arbitrary wide range of scales.

The paper summarizes a traffic engineering procedure composed by the following steps:

- statistical analysis of measured traffic data,
- Markovian approximation of traffic processes,
- analysis of performance parameters based on the Markovian model.

All steps of this procedure are supported with a number of numerical example and the results are verified against simulation and alternative analysis methods.

The paper is organized as follows. Section 2 discusses some relevant characteristics of traffic processes and describe models that exhibit these features. Statistical tests for identifying these characteristics in datasets are described in Section 3. A short introduction to Markovian models is given in 4. An overview of the existing fitting methods with connected application examples is given in 5. The survey is concluded in 6.

2 Traffic models and their properties

The traffic process at a given point of a telecommunication network is characterized by the data packet arrival instances (or equivalently by the interarrival times) and the associated data packet sizes. Any of these two processes can be composed by dependent or independent samples. In case of identically distributed independent samples the process modeling simplifies to capturing a distribution, while in case of dependent samples the whole stochastic process (with its intrinsic dependency structure) has to be captured as well.

2.1 Heavy tailed distributions

One of the important new observations of the intensive traffic measurement of high speed telecommunication networks is the presence of heavy tailed distributions. Marginal distributions of specific traffic processes, file size distribution on HTTP servers, etc, were found to be "heavy tailed". The random variable Y, with cumulative distribution function (cdf) $F_Y(x)$, is said to be heavy tailed if

$$1 - F_Y(x) = x^{-\alpha} L(x),$$

where $L(x)$ is slowly varying as $x \to \infty$, i.e., $lim_{x\to\infty} L(ax)/L(x) = 1$ for $a > 0$. (There are several different naming conventions applied in this field. Heavy tailed distributions are called regularly varying or power tail distributions also.) Typical member of this distribution class is the Pareto family.

There is an important qualitative property of the moments of heavy tailed distributions. If Y is heavy tailed with parameter α then its first $n < \alpha$ moments $E(Y^n)$ are finite and its all higher moments are infinite.

There are other classes of distributions whose tail decay slower than the exponential. The random variable Y, with distribution $F_Y(x)$, is said to be long tailed if

$$\lim_{x\to\infty} e^{\gamma x}(1 - F_Y(x)) = \infty, \qquad \forall \gamma > 0$$

The Weibull family ($F(x) = 1 - e^{-(t/a)^c}$) with $c < 1$ is long tailed, even if all moments of the Weibull distributed random variables are finite. The heavy tailed distributions form a subclass of the long tailed class.

A characteristic property of the heavy tailed class is the asymptotic relation of the distribution of the sum of n samples, $S_n = Y_1 + \ldots + Y_n$, and the maximum of n samples, $M_n = \max_{1\le i\le n} Y_i$:

$$Pr(S_n > x) \sim Pr(M_n > x) \tag{1}$$

where the notation $g(x) \sim f(x)$ denotes $lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$. In words, the sum of heavy tailed random variables is dominated by a *single* large sample and the rest of the samples are negligible small compare to the dominant one for large values of x. The probability that S_n is dominated by more than one "large" samples or it is obtained as the sum of number of small samples is negligible for

"large" values of S_n. This interpretation gives an intuitive explanation for a set of complex results about the waiting time of queuing models with heavy tailed service time distribution [6].

2.2 Processes with long range dependence

The definition of long range dependence of traffic arrival processes is as follows. Let us divide the time access into equidistant intervals of length Δ. The number of arrivals in the ith interval is denoted by X_i. $\mathcal{X} = \{X_i, i = 0, 1, \ldots\}$ is a stochastic process whose aggregated process is defined as follows:

$$\mathcal{X}^{(m)} = \{X_i^{(m)}\} = \left\{ \frac{X_1 + \ldots + X_m}{m}, \ldots, \frac{X_{mk+1} + \ldots + X_{(m+1)k}}{m}, \ldots \right\}$$

The autocorrelation function of $\mathcal{X}^{(m)}$ is:

$$r^{(m)}(k) = \frac{E\{(X_n^{(m)} - E(X^{(m)})) \cdot (X_{n+k}^{(m)} - E(X^{(m)}))\}}{E\{(X_n^{(m)} - E(X^{(m)}))^2\}}$$

The process $\mathcal{X}$ exhibits long-range dependence (LRD) of index β if its autocorrelation function can be realized as

$$r(k) \sim A(k)k^{-\beta}, \qquad k \to \infty$$

where A(k) is a slowly varying function.

Self-similar processes Using the above definition of the aggregated process, $\mathcal{X}$ is

a) exactly self-similar if $\mathcal{X} \stackrel{d}{=} m^{1-H}\mathcal{X}^{(m)}$, i.e., if $\mathcal{X}$ and $\mathcal{X}^{(m)}$ are identical within a scale factor in finite dimensional distribution sense.
b) exactly second-order self-similar if $r^{(m)}(k) = r(k), \quad \forall m\ ,\ k \geq 0$
c) asymptotically second-order self-similar if $r^{(m)}(k) \to r(k), \quad (k, m \to \infty)$

where H is the Hurst parameter, also referred to as the self-similarity parameter.

For exactly self-similar processes the scaling behavior, which is characterized by the Hurst parameter (H), can be checked based on any of the absolute moments of the aggregated process:

$$log(E(|X^{(m)}|^q)) = log(E(|m^{H-1}X|^q)) = q(H-1)log(m) + log(E(|X|^q)). \quad (2)$$

According to (2), in case of a self-similar process, plotting $log(E(|X^{(m)}|^q))$ against $log(m)$ for fixed q results in a straight line. The slope of the line is $q(H-1)$. Based on the above observations the test is performed as follows. Having a series of length N, the moments may be estimated as

$$E(|X^{(m)}|^q) = \frac{1}{\lfloor N/m \rfloor} \sum_{i=1}^{\lfloor N/m \rfloor} |X_i^{(m)}|^q,$$

where $\lfloor x \rfloor$ denotes the largest integer number smaller or equal to x. To test for self-similarity $log(E(|X^{(m)}|^q))$ is plotted against $log(m)$ and a straight line is fitted to the curve. If the straight line shows good correspondence with the curve, then the process is self-similar and its Hurst-parameter may be calculated by the slope of the straight line. This approach assumes that the scaling behavior of all absolute moments, q, are the same and it is captured by the Hurst-parameter. If it is the case we talk about mono-fractal behavior. The variance-time plot, which is used widespread to gain evidence of self-similarity, is the special case with $q = 2$. It depicts the behavior of the 2nd moments for the centered data.

It is worth to point out that self-similarity and stationarity imply that either $E(X) = 0$, or $E(X) = \pm\infty$, or $H = 1$. But $H = 1$ implies as well that $X_i = X_j$, $\forall i, j$ almost surely. As a consequence, to test for statistical self-similarity makes sense only having zero-mean data, i.e., the data has to be centered before the analysis.

Multi-fractal processes Statistical tests of self-similarity try to gain evidence through examining the behavior of the absolute moments $E(|X^{(m)}|^q)$. In case of monofractal processes the scaling behavior of all absolute moments is characterized by a single number, the Hurst parameter. Multifractal processes might exhibit different scaling for different absolute moments. Multifractal analysis looks at the behavior of $E(|X^{(m)}|^q)$ for different values q and results in a *spectrum* that illustrates the behavior of the absolute moments. This analysis procedure is detailed in Section 3.3.

Fractional Gaussian noise By now we provided the definition of the large class of self-similar stochastic processes, but we did not provide any specific member of this class. The two simplest self-similar processes that are often used in validation of self-similar modeling assumptions are the fractional Gaussian noise and the ARIMA process.

Fractional Gaussian noise, $X_i, i \geq 1$, is the increment process of fractional Brownian motion, $B(t), t \in \mathbb{R}^+$:

$$X_i = B(i+1) - B(i),$$

Fractional Brownian motion with Hurst parameter H $(0.5 < H < 1)$ is characterized by the following properties: *i)* $B(t)$ has stationary increment, *ii)* $E(B(t)) = 0$, *iii)* $E(B^2(t)) = t^{2H}$ (assuming the time unit is such that $E(B^2(1)) = 1$), *iv)* $B(t)$ has continuous path, *v)* $B(t)$ is a Gaussian process, i.e., all of its finite dimensional distributions are Gaussian. The covariance of fractional Brownian motion is $E(B(t) \cdot B(s)) = 1/2(s^{2H} + t^{2H} - |s-t|^{2H})$, and hence, the auto-covariance function of fractional Gaussian noise $\gamma(h) = E(X_i X_{i+h}) \sim H(2H-1)h^{2H-2}$ is positive and exhibits long-range dependence.

ARIMA process An other simple self-similar process is the fractional ARIMA(0,d,0) process. It is defined as:

$$X_i = \sum_{j=0}^{\infty} c_j \epsilon_{i-j}$$

where ϵ_i are i.i.d. standard normal random variables and the c_j coefficients implement moving average with parameter d according to $c_j = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)}$. For large values of j the coefficients $c_j \sim \frac{j^{d-1}}{\Gamma(d)}$. The asymptotic behavior of the auto-covariance function is

$$\gamma(h) = E(X_i X_{i+h}) \sim C_d h^{2d-1}$$

with coefficient $C_d = \pi^{-1}\Gamma(1-2d)\sin(\pi d)$. For $0 < d < 1/2$ the auto-covariance function has the same polynomial decay as the auto-covariance function of fractional Gaussian noise with $H = d + 1/2$.

The better choice among these two processes depends on the applied analysis method. The fractional Gaussian noise is better in exhibiting asymptotic properties based on finite number of samples, while the generation of fractional ARIMA process samples is easier since it is based on an explicit expression.

3 Statistical analysis of measured traffic datasets

3.1 Estimation of the heavy tail index

In this section we discuss methods for identifying the heavy tail index of datasets. Application of the methods is illustrated on the dataset EPA_HTTP which can be downloaded from [52] and contains a day of HTTP logs with about 40000 entries. The experimental complementary cumulative distribution function (ccdf) of the length of the requests is depicted in Figure 1.

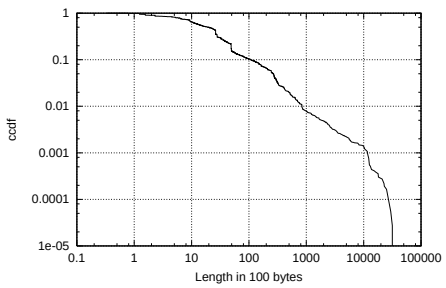

Fig. 1. Experimental ccdf of the length of requests arriving to the server

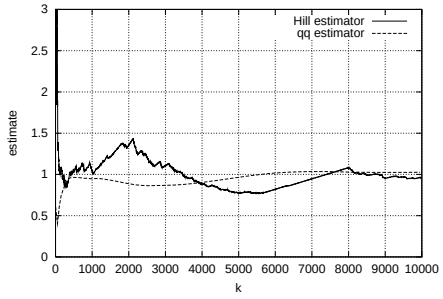

Fig. 2. The Hill- and the dynamic qq-plot for the EPA trace

Hill estimator A possible approach to estimate the index of the tail behavior α is the Hill estimator [20]. This estimator provides the index as a function of the k largest elements of the dataset and is defined as

$$\alpha_{n,k} = \left(\frac{1}{k} \sum_{i=0}^{k-1} \left(\log X_{(n-i)} - \log X_{(n-k)} \right) \right)^{-1} \tag{3}$$

where $X_{(1)} \leq ... \leq X_{(n)}$ denotes the order statistics of the dataset. In practice, the estimator given in (3) is plotted against k and if the plot stabilizes to a constant value this provides an estimate of the index. The Hill-plot (together with the dynamic qq-plot that will be described later) for the EPA trace is depicted in Figure 2.

The idea behind the procedure and theoretical properties of the estimator are discussed in [39]. Applicability of the Hill estimator is reduced by the fact that

- its properties (e.g. confidence intervals) are known to hold only under conditions that often cannot be validated in practice [39],
- the point at which the power-law tail begins must be determined and this can be difficult because often the datasets do not show clear border between the power-law tail and the non-power-low body of the distributions.

By slight modifications in the way the Hill plot is displayed, the uncertainty of the estimation procedure can be somewhat reduced, see [39, 40].

Quantile-quantile regression plot The above described Hill estimator performs well if the underlying distribution is close to Pareto. With the quantile-quantile plot (qq-plot), which is a visual tool for assessing the presence of heavy tails in distributions, one can check this. The qq-plot is commonly used in various forms, see for example [8, 41]. Hereinafter, among the various forms, we follow the one presented in [25].

Having the order statistics $X_{(1)} \leq ... \leq X_{(n)}$ plot

$$\left\{ \left(-\log \left(1 - \frac{j}{k+1} \right), \log X_{(j)} \right), n-k+1 \leq j \leq n \right\} \tag{4}$$

for a fixed value of k. (As one can see only the k upper order statistics is considered in the plot, the other part of the sample is neglected.) The plot, if the data is close to Pareto, should be a straight line with slope $1/\alpha$. By determining the slope of the straight line fitted to the points by least squares, we obtain the so-called qq-estimator [25].

The qq-estimator can be visualized in two different ways. The *dynamic* qq-plot, depicted in Figure 2, plots the estimate of α as the function of k (this plot is similar to the Hill-plot). The *static* qq-plot, given in Figure 3, depicts (4) for a fixed value of k and shows its least square fit. As for the Hill-plot, when applying the qq-estimator, the point at which the tail begins has to be determined.

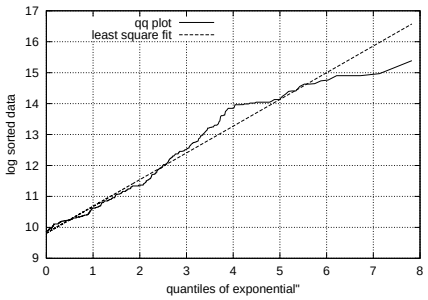

Fig. 3. Static qq-plot for the EPA trace

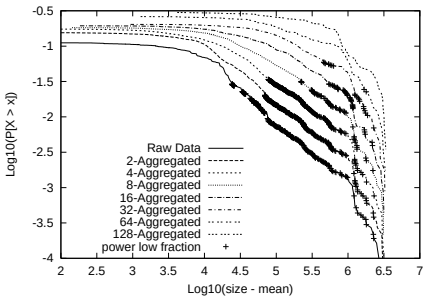

Fig. 4. Complementary distribution function for different aggregation levels for the EPA trace

Estimation based on the scaling properties Another method is proposed in [9] which, in contrast to the Hill- and qq-estimator, does not require to determine where the tail begins. The procedure is based on the scaling properties of sums of heavy tailed distribution. The estimator, which is implemented in the tool *aest*, determines the heavy tail index by exploring the complementary distribution function of the dataset at different aggregation levels. For the EPA trace, the index estimated by *aest* is 0.97. In order to aid further investigation, the tool produces a plot of the complementary distribution function of the dataset at different aggregation levels indicating the segments where heavy tailed behavior is present. This plot for the considered dataset is depicted in Figure 4.

3.2 Tests for long range dependency

Recently, it has been agreed [28, 36, 37] that when one studies the long-range dependence of a traffic trace the most significant parameter to be estimated is the degree of self-similarity, usually given by the so-called Hurst-parameter. The aim of the statistical approach, based on the theory of self-similarity, is to find the Hurst-parameter.

In this section methods for estimating the long-range dependence of datasets are recalled. Beside the procedures described here, several other can be found in the literature. See [3] for an exhaustive discussion on this subject.

It is important to note that the introduced statistical tests of self-similarity, based on a finite number of samples, provides an approximate value of H only for the considered range of scales. Nothing can be said about the higher scales and the asymptotic behavior based on these tests.

Throughout the section, we illustrate the application of the estimators on the first trace of the well-known Bellcore dataset set that contains local-area network (LAN) traffic collected in 1989 on an Ethernet at the Bellcore Morristown Research and Engineering facility. It may be downloaded from the WEB site collecting traffic traces [52]. The trace was first analyzed in [16].

Variance-time plot One of the tests for *pseudo* self-similarity is the variance-time plot. It is based on the fact that for self-similar time series $\{X_1, X_2, \ldots\}$

$$\mathrm{Var}(X^{(m)}) \sim m^{-\beta} \quad , \quad \text{as} \quad m \to \infty, \quad 0 < \beta < 1.$$

The variance-time plot depicts $Log(\mathrm{Var}(X^{(m)}))$ versus $Log(m)$. For *pseudo* self-similar time series, the slope of the variance-time plot $-\beta$ is greater than -1. The Hurst parameter can be calculated as $H = 1 - (\beta/2)$. A traffic process is said to be *pseudo* self-similar when the empirical Hurst parameter is between 0.5 and 1.

The variance-time plot for the analyzed Bellcore trace is depicted in Figure 5. The Hurst-parameter given by the variance-time plot is 0.83.

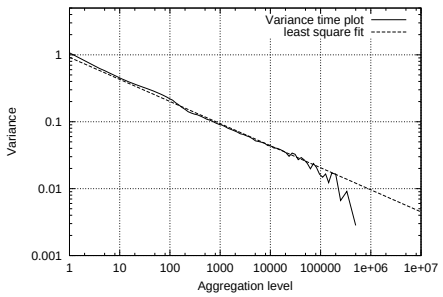

Fig. 5. Variance-time plot and its least square fit for the Bellcore trace

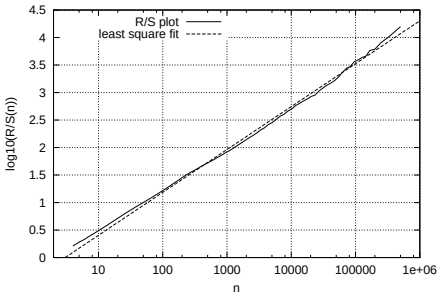

Fig. 6. R/S plot and its least square fit for the Bellcore trace

R/S plot The R/S method is one of the oldest tests for self-similarity, it is discussed in detail in [31]. For interarrival time series, $\mathcal{Z} = \{Z_i, i \geq 1\}$, with partial sum $Y_n = \sum_{i=1}^{n} Z_i$, and sample variance

$$S^2(n) = \frac{1}{n} \sum_{i=1}^{n} {Z_i}^2 - \frac{1}{n^2} \cdot Y_n^2,$$

the R/S statistic, or the rescaled adjusted range, is given by:

$$R/S(n) = \frac{1}{S(n)} \left[\max_{0 \leq k \leq n} \left(Y(k) - \frac{k}{n} Y(n) \right) - \min_{0 \leq k \leq n} \left(Y(k) - \frac{k}{n} Y(n) \right) \right].$$

$R/S(n)$ is the scaled difference between the fastest and the slowest arrival period considering n arrivals. For stationary LRD processes $R/S(n) \approx (n/2)^H$. To determine the Hurst parameter based on the R/S statistic the dataset is divided into blocks, $\log[R/S(n)]$ is plotted versus $\log n$ and a straight line is fitted on the points. The slope of the fitted line is the estimated Hurst parameter.

The R/S plot for the analyzed Bellcore trace is depicted in Figure 6. The Hurst-parameter determined based on the R/S plot is 0.78.

Whittle estimator The Whittle estimator is based on the maximum likelihood principle assuming that the process under analysis is Gaussian. The estimator, unlike the previous ones, provides the estimate through a non-graphical method. This estimation takes more time to perform but it has the advantage of providing confidence intervals as well. For details see [17, 3]. For the Bellcore trace, the estimated value of the Hurst parameter is 0.82 and its 95% confidence interval is $[0.79, 0.84]$.

3.3 Multifractal framework

In this section we introduce two techniques to analyze multifractal processes.

Legendre spectrum Considering a continuous-time process $\mathcal{Y} = \{Y(t), t > 0\}$ the scaling of the absolute moments of the increments is observed through the *partition function*

$$T(q) = \lim_{n\to\infty} \frac{1}{-n} \log_2 E\left[\sum_{k=0}^{2^n-1} |Y((k+1)2^{-n}) - Y(k2^{-n})|^q\right]. \tag{5}$$

Then, a multifractal spectrum, the so-called *Legendre spectrum* is given as the *Legendre transform* of (5)

$$f_L(\alpha) = T^*(\alpha) = \inf_q (q\alpha - T(q))$$

Since $T(q)$ is always concave, the Legendre spectrum $f_L(\alpha)$ may be found by simple calculations using that

$$T^*(\alpha) = q\alpha - T(q), \text{ and } (T^*)'(\alpha) = q \text{ at } \alpha = T'(q). \tag{6}$$

Let us mention here that there are also other kinds of fractal spectrum defined in the fractal world (see for example [42]). The Legendre spectrum is the most attractive one from numerical point of view, and even though in some cases it is less informative than, for example, the large deviation spectrum, it provides enough information in the cases considered herein.

In case of a discrete-time process $\mathcal{X}$ we assume that we are given the increments of a continuous-time process. This way, assuming that the sequence we examine consists of $N = 2^L$ numbers, the sum in (5) becomes

$$S_n(q) = \sum_{k=0}^{N/2^n-1} |X_k^{(2^n)}|^q, \ 0 \le n \le L, \tag{7}$$

where the expectation is ignored. Ignoring the expectation is accurate for small n, i.e., for the finer resolution levels. In order to estimate $T(q)$, we plot $\log_2(S_n(q))$ against $(L - n)$, $n = 0, 1, ..., L$, then $T(q)$ is found by the slope of the linear line fitted to the curve. If the linear line shows good correspondence with the

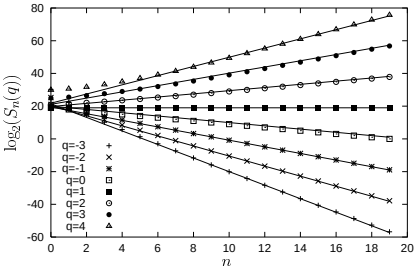

Fig. 7. Scaling of log-moments with linear fits for the interarrival times of the Bellcore *pAug* trace

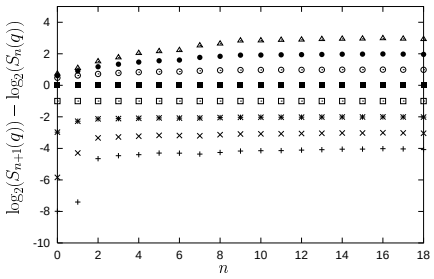

Fig. 8. Increments of log-moments for the interarrival times of the Bellcore *pAug* trace

curve, i.e., if $\log_2(S_n(q))$ scales linearly with $\log(n)$, then the sequence $\mathcal{X}$ can be considered a multifractal process.

Figure 7, 9, 8 and 10 illustrate the above described procedure to obtain the Legendre spectrum of the famous Bellcore *pAug* traffic trace (the trace may be found at [52]). Figure 7 depicts the scaling behavior of the log moments calculated through (7). With q in the range $[-3, 4]$, excluding the finest resolution levels $n = 0, 1$ the moments show good linear scaling. For values of q outside the range $[-3, 4]$ the curves deviate more and more from linearity. As, for example, in [43] one may look at non-integer values of q as well, but, in general, it does not provide notably more information on the process. To better visualize the deviation from linearity Figure 8 depicts the increments of the log-moment curves of Figure 7. Completely horizontal lines would represent linear log-moment curves.

The partition function $T(q)$ is depicted in Figure 9. The three slightly different curves differ only in the considered range of the log-moments curves, since different ranges result in different linear fitting. The lower bound of the linear fitting is set to 3, 5 and 7, while the upper bound is 18 in each cases. (In the rest of this paper the fitting range is 5 - 18 and there are 100 moments evaluated in the range $q \in [-5, +5]$.) Since the partition function varies only a little (its derivative is in the range $[0.8, 1.15]$), it is not as informative as its Legendre transform is (Figure 10). According to (6) the Legendre spectrum is as wide as wide the range of derivatives of the partition function is, i.e., the more the partition function deviates from linearity the wider the Legendre spectrum is. The Legendre transform significantly amplifies the scaling information, but it is also sensitive to the considered range of the log-moments curves.

See [43] for basic principles of interpreting the spectrum. We mention here only that a curve like the one depicted in Figure 10 reveals a *rich multifractal spectrum*. On the contrary, as it was shown in [51], the fractional Brownian motion (fBm) has a trivial spectrum. The partition function of the fBm is a straight line which indicates that its spectrum consists of one point, i.e., the behavior of its log-moments is identical for any q.

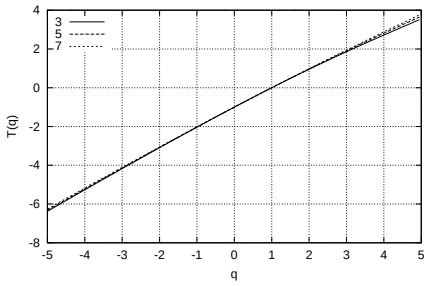

Fig. 9. Partition function estimated through the linear fits shown in Figure 7

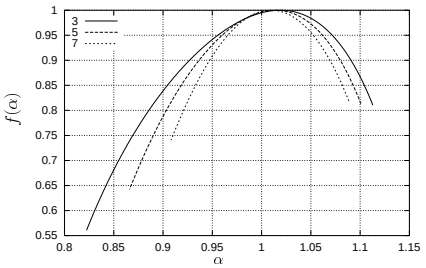

Fig. 10. The Legendre transform of the partition function (Figure 9) results in the Legendre spectrum

Haar wavelet Another way to carry out multiscale analysis is the Haar wavelet transform. The choice of using the *unnormalized* version of the Haar wavelet transform is motivated by the fact that it suits more the analysis of the Markovian point process introduced further on.

The multiscale behavior of the finite sequence $X_i, 1 \leq i \leq 2^L$ will be represented by the quantities $c_{j,k}, d_{j,k}, j = 0, \ldots, L$ and $k = 1, \ldots, 2^L/2^j$. The finest resolution is described by $c_{0,k}, 1 \leq k \leq 2^L$ which gives the finite sequence itself, i.e., $c_{0,k} = X_k$. Then the multiscale analysis based on the unnormalized Haar wavelet transform is carried out by iterating

$$c_{j,k} = c_{j-1,2k-1} + c_{j-1,2k}, \tag{8}$$

$$d_{j,k} = c_{j-1,2k-1} - c_{j-1,2k}, \tag{9}$$

for $j = 1, \ldots, L$ and $k = 1, \ldots, 2^L/2^j$. The quantities $c_{j,k}, d_{j,k}$ are the so-called scaling and wavelet coefficients of the sequence, respectively, at scale j and position k. At each scale the coefficients are represented by the vectors $c_j = [c_{j,k}]$ and $d_j = [d_{j,k}]$ with $k = 1, \ldots, 2^L/2^j$. For what concerns c_j, the higher j the lower the resolution level at which we have information on the sequence. The information that we lost as a result of the step from c_{j-1} to c_j, is conveyed by the sequence of wavelet coefficients d_j. It is easy to see that c_{j-1} can be perfectly reconstructed from c_j and d_j. As a consequence the whole $X_i, 1 \leq i \leq 2^L$ sequence can be constructed (in a top to bottom manner) based on a normalizing constant, $c_L = c_{L,1} = \sum_{i=1}^{2^L} X_i$, and the $d_j, j = 1, \ldots, L$ vectors.

By taking the expectation of the square of (8) and (9)

$$E[c_{j,k}^2] = E[c_{j-1,2k-1}^2] + 2E[c_{j-1,2k-1}c_{j-1,2k}] + E[c_{j-1,2k}^2], \tag{10}$$

$$E[d_{j,k}^2] = E[c_{j-1,2k-1}^2] - 2E[c_{j-1,2k-1}c_{j-1,2k}] + E[c_{j-1,2k}^2], \tag{11}$$

Let us assume that the series we analyze are stationary; then, by summing (10) and (11) and rearranging the equation, we have

$$E[c_{j-1}^2] = \frac{1}{4}\left(E[d_j^2] + E[c_j^2]\right). \tag{12}$$

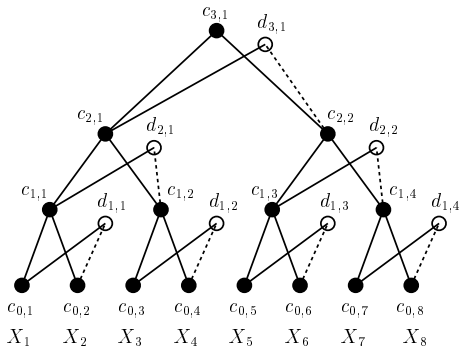

Fig. 11. Haar wavelet transform

Similarly, by consecutive application of (12) from one scale to another, the $E[d_j^2], j = 1, \ldots, L$ series completely characterize the variance decay of the $X_i, 1 \leq i \leq 2^L$ sequence apart of a normalizing constant ($c_L = c_{L,1} = \sum_{i=1}^{2^L} X_i$). This fact allows us to realize a series with a given variance decay if it is possible to control the 2nd moment of the scaling coefficient with the chosen synthesis procedure. In Section 5 we will briefly discuss a method that attempts to capture the multifractal scaling behavior via the series $E[d_j^2], j = 1, \ldots, L$.

4 Markovian modeling tools

Markovian modeling tools are stochastic processes whose stochastic behavior depends only on the state of a "background" Markov chain. The research and application of these modeling tools through the last 20 years resulted in a widely accepted standard notation. The two classes of Markovian processes considered in this paper are Phase type distributions and Markovian arrival processes. Here, we concentrate our attention mainly on continuous time Markovian models, but it is also possible to apply Markovian models in discrete time [33, 5, 27].

4.1 Phase type distribution

$Z(t)$ is a continuous time Markov chain with n transient state and one absorbing state. Its initial probability distribution is $\hat{\alpha}$ and generator matrix is $\hat{\mathbf{B}}$. The time to reach the absorbing state, T, phase type distributed with representation $\alpha, \mathbf{B}$, where α is the sub-vector of $\hat{\alpha}$ and $\mathbf{B}$ is the sub-matrix of $\hat{\mathbf{B}}$ associated with the transient states. The cumulative distribution function (cdf), the probability density function (pdf), and the moments of this distribution are:

$$F_T(t) = 1 - \alpha e^{\mathbf{B}t} h, \quad f_T(t) = \alpha \mathbf{B} e^{\mathbf{B}t} h, \quad E[X^i] = i!\, \alpha\, (-\mathbf{B})^{-i} h,$$

where h is the column vector of ones. The number of unknown in the $\alpha, \mathbf{B}$ representation of a PH distribution is $\mathcal{O}(n^2)$.

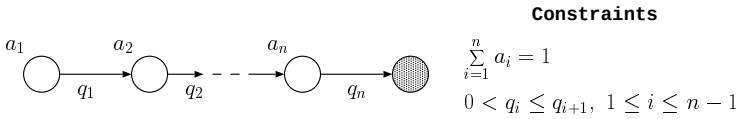

Fig. 12. Canonical form for Acyclic continuous-time PH distributions

When $Z(t)$ is an acyclic Markov chain the associated PH distribution is referred to as Acyclic PH (APH) distribution. The popularity of APH distributions (specially in PH fitting) lies in the fact that all APH distributions can be uniquely transformed into a canonical from (Figure 12) which has only $\mathcal{O}(n)$ parameters [10] and the flexibility of the PH and the APH class of the same order is very close. E.g., the 2nd order PH and APH classes exhibit the same moments bounds [50].

4.2 Markovian arrival process

Let $Z(t)$ be an irreducible Markov chain with finite state space of size m and generator $\mathbf{Q}$. An arrival process is associated with this Markov chain in the following way:

- while the Markov chain stays in state i arrival occurs at rate λ_i,
- when the Markov chain undergoes a state transition from i to j arrival occurs with probability p_{ij}.

The standard description of MAPs is given with matrices $\mathbf{D_0}$ and $\mathbf{D_1}$ of size $(m \times m)$, where $\mathbf{D_0}$ contains the transition rates of the Markov chain which are not accompanied with arrivals and $\mathbf{D_1}$ contains the transition rates which are accompanied with arrivals, i.e.:

- $\mathbf{D_0}_{ij} = (1 - p_{ij})\mathbf{Q}_{ij}$, for $i \neq j$ and $\mathbf{D_0}_{ii} = \mathbf{Q}_{ii} - \lambda_i$;
- $\mathbf{D_1}_{ij} = p_{ij}\mathbf{Q}_{ij}$ for, $i \neq j$ and $\mathbf{D_1}_{ii} = \lambda_i$.

Many familiar arrival processes represent special cases of MAPs:

- the Poisson process (MAP with a single state),
- interrupted Poisson process: a two-state MAP in which arrivals occur only in one of the states and state jumps do not cause arrival,
- Markov modulated Poisson process: state jumps do not give rise to arrivals.

The class of MAPs is closed for superposition and Markovian splitting.

5 Fitting Markovian models to datasets

Fitting a Markovian model to a measured dataset is to find a Markovian model which exhibits a stochastic behavior as close to the one of the measured dataset as possible. In practice, the order of approximate Markov models should kept

low, both, for having few model parameters to evaluate and for obtaining computable models. The presence of slow decay behavior (heavy tail or long range correlation) in measured datasets makes the fitting more difficult. Typically a huge number of samples needed to obtain a fairly reliable view on the stochastic behavior over a range of several orders of magnitude, and, of course, the asymptotic behavior can not be checked based on finite datasets. A class of fitting methods approximates the asymptotic behavior based on the reliably known ranges (e.g., based on 10^6 i.i.d. samples the cdf. can be approximated up to the $1 - F(x) \sim 10^{-4} - 10^{-5}$ limit). The asymptotic methods are based on the assumption that the dominant parameters (e.g., tail decay, correlation decay) of the known ranges remain unchanged in the unknown region up to the asymptotic limit.

Unfortunately, Markovian models can not exhibit any complex asymptotic behavior. In the asymptotic region Markovian models have exponential tail decay or autocorrelation. Due to this dominant property Markovian models were not considered for fitting datasets with slow decaying features for a long time. Recently, in spite of the exponential asymptotic decay behavior, Markovian models with slow decay behavior for several orders of magnitude were introduced. These results broaden the attention from asymptotically slow decay models to models with slow decay in given predefined range. The main focus of this paper is on the use of Markovian models with slow decay behavior in applied traffic engineering.

A finite dataset provides only a limited information about the stochastic properties of traffic processes. Especially, the long range and the asymptotic behavior cannot be extracted from finite dataset. To overcome the lack of these important model properties the set of information provided by the dataset is often accompanied by engineering assumptions in practice. One of the most commonly applied traffic engineering assumptions is that the decay trends of a known region continuous to infinity.

The use of engineering assumptions has a significant role in model fitting as well. With this respect there are two major classes of fitting methods:

- fitting based on al the samples,
- fitting based on information extracted from the samples,

Naturally, there are methods which combines these two approaches.

The fitting methods based on extracted information find their roots in traffic engineering assumptions. It is a common goal in traffic engineering to find a simple (characterized by few parameters), but robust (widely applicable) traffic model which is based on few representative traffic parameters of network traffic. The traffic models discussed in Section 2 are completely characterized by very few parameters. E.g., the tail behavior of a power tail distribution is characterized by the heavy tail index α, fractional Gaussian noise is characterized by parameter H and the variance over a natural time unit. Assuming that there is representative information of the dataset, it is worth to complete the model fitting based on this compact description of the traffic properties instead of using all the very large dataset. Unfortunately, a commonly accepted, accurate and compact traffic

characterization is not available up to now. This way, when the fitting is based on extracted information, the goodness of fitting strongly depend on the descriptive power of the selected characteristics to be fitted.

In this section we introduce a selected set of fitting methods from both classes. The fitting methods that are based on extracted information are composed by two mains steps: the statistical analysis of the dataset to extract representative properties and the fitting itself based on these properties. The first step of this procedure is based on the methods presented in the previous section, and only the second step is considered here.

5.1 PH fitting

General PH fitting methods minimizes a *distance measure* between the experimental distribution and the approximate PH one. The most commonly applied distance measure is the relative entropy: $\int_0^\infty f(t) \log\left(\frac{f(t)}{\hat{f}(t)}\right) dt$ where $f(t)$ and $\hat{f}(t)$ denote the pdf of the distribution to be fitted and that of the fitting distribution, respectively. The number of parameters to minimize in this procedure depends on the order of the approximate PH model. The required order of PH models can be approximated based on the dataset [48], but usually small models are preferred in practice for computational convenience. It is a common feature of the relative entropy and other distance measures that the distance is a non-linear function of the PH parameters.

General PH fitting methods might perform poorly in fitting slow decaying tail behavior [22]. As an alternative, heuristic fitting procedures can be applied that focus on capturing the tail decay behavior. In case of heuristic fitting methods, the goal is not to minimize a properly defined distance measure, but to construct a PH distribution which fulfills a set of heuristic requirements.

According to the above classification of fitting procedures general fitting methods commonly belong to the *fitting based on samples* class and heuristic fitting methods to the *fitting to extracted model properties* class.

The literature of general PH fitting methods is quite large. A set of methods with a comparison of their fitting properties are presented in [26]. Here we consider only those methods which were applied for fitting slowly decaying behavior in [11] and [22]. Among the heuristic methods we discuss the one proposed in [14] and its extension in [22].

EM method The *expectation maximization* (EM) method was proposed to apply for PH fitting in [2]. It is a statistical method which performs an iterative optimization over the space of the PH parameters to minimize the relative entropy. It differs from other relative entropy minimizing methods in the way it searches for the minimum of the non-linear distance measure. Based on the fact that hyper-exponential distributions can capture slow decay behavior ([14]), a specialized version of the EM algorithm, which fits the dataset with

hyper-exponential distributions, is applied for fitting measured traffic datasets in [11].

Starting from an initial guess $\alpha^{(0)}$, $\lambda^{(0)}$ and denoting the pdf of the hyper-exponential distribution with initial probability vector α and intensity vector λ by $\hat{f}(t|\alpha,\lambda)$, the iterative procedure calculates consecutive hyper-exponential distributions based on the samples $t_1,\ldots,t_N$ as:

$$\alpha_i^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N}\frac{\alpha_i^{(k)}\hat{f}(t_n|e_i,\lambda^{(k)})}{\hat{f}(t_n|\alpha^{(k)},\lambda^{(k)})}, \quad \alpha_i^{(k+1)} = \frac{\frac{1}{N}\sum_{n=1}^{N}\frac{\alpha_i^{(k)}\hat{f}(t_n|e_i,\lambda^{(k)})}{\hat{f}(t_n|\alpha^{(k)},\lambda^{(k)})}}{\frac{1}{N}\sum_{n=1}^{N}t_n\frac{\alpha_i^{(k)}\hat{f}(t_n|e_i,\lambda^{(k)})}{\hat{f}(t_n|\alpha^{(k)},\lambda^{(k)})}}$$

where e_i is the vector of zeros with a one at the ith position.

The computational complexity of this simplified method using hyper-exponential distributions is much less than the one for the whole PH class. Nevertheless, a reliable view on the (slow decaying) tail behavior requires very large number of samples. The complexity of the simplified fitting method is still proportional to the size of the dataset, hence the applicability of this approach is limited by computational complexity ($\sim 10^7$ samples were reported in [11]). On the other hand, due to the strict structure of hyper-exponential distributions (e.g., there is no fork in the structure), less iterations are required to reach a reasonable accuracy ($5-10$ iterations were found to be sufficient in [11]).

This simplified EM fitting method is a potential choice for model fitting when we have a large dataset, but we do not have or do not want to apply any engineering assumption on the properties of the dataset.

Tail fitting based on the ccdf The method proposed by Feldmann and Whitt [14] is a recursive fitting procedure that results in a hyper-exponential distribution whose cumulative distribution function (ccdf) at a given set of points is "very close" to the ccdf of the original distribution. This method was successfully applied to fit Pareto and Weibull distributions.

Combined fitting method In [22] a PH fitting method is proposed that handles the fitting of the body and the fitting of the tail in a separate manner. This is done by combining the method proposed by Feldmann and Whitt [14] and a general method to.

The limitation of this combined method comes from the limitation of the method of Feldmann and Whitt. Their method is applicable only for fitting distributions with monotone decreasing density function. Hence the proposed combined method is applicable when the tail of the distribution is with monotone decreasing density. In the case of the combined method, this restriction is quite loose since the border of the main part and the tail of the distribution is arbitrary, hence the restriction of applicability is to have a positive number C such that the density of the distribution is monotone decreasing above C.

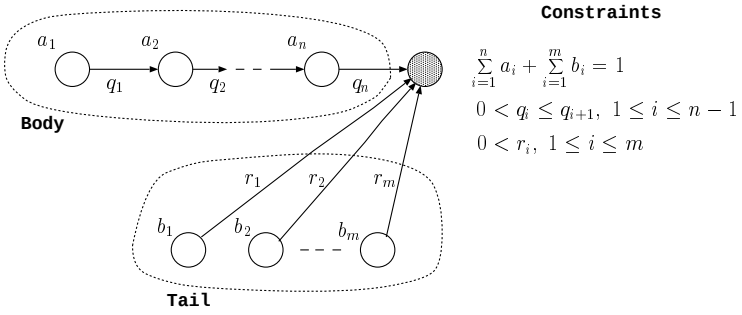

Fig. 13. Structure of approximate Phase type distribution

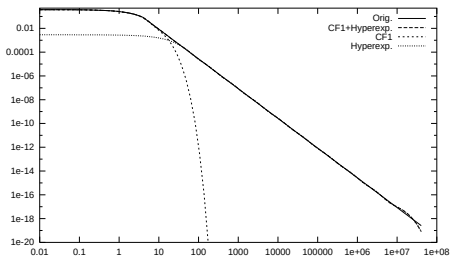

Fig. 14. Different parts of the pdf are approximated by different parts of the PH structure

The result of this fitting algorithm is a Phase type distribution of order $n+m$, where n is the number of phases used for fitting the body and m is the number of phases used for fitting the tail. The structure of this Phase type distribution is depicted in Figure 13 where we have marked the phases used to fit the body and those to fit the tail. The parameters $\beta_1, \ldots, \beta_m, \mu_1, \ldots, \mu_m$ are computed by considering the tail while the parameters $\alpha_1, \ldots, \alpha_m, \lambda_1, \ldots, \lambda_2$ are determined considering the main part of the distribution.

To illustrate the combined fitting method, we consider the following Pareto-like distributions [45]:

$$\text{Pareto I:} \quad f(t) = \begin{cases} \alpha B^{-1} e^{-\frac{\alpha}{B}\ t} & \text{for } t \leq B \\ \alpha B^{\alpha} e^{-\alpha} t^{-(\alpha+1)} & \text{for } t > B \end{cases}$$

$$\text{Pareto II:} \qquad f(t) = \frac{b^{\alpha} e^{-b/t}}{\Gamma(\alpha)} x^{-(\alpha+1)}$$

For both ditributions α is the heavy tail index.

Figure 14 pictures how different parts of the PH structure (Figure 13) contributes to the pdf when fitting distribution Pareto I with parameters $\alpha = 1.5, B = 4$. In this case 8 phases are used to fit the body and 10 to fit the tail.

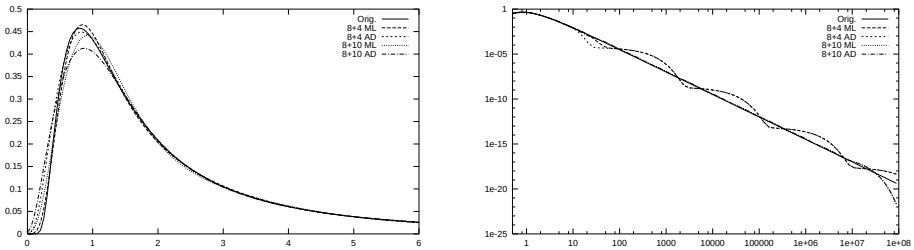

Fig. 15. Pareto II distribution and its PH approximation with the combined method

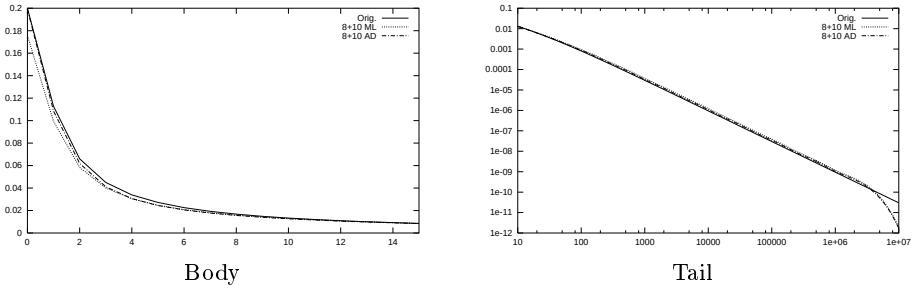

Fig. 16. Queue length distribution of an M/G/1 queue and its approximate M/PH/1 queue

Figure 15 illustrates the fitting of distribution Pareto II with parameters $\alpha = 1.5, b = 2$. In the legend of the figure ML indicates that the relative entropy measure was applied to fit the main part (corresponding to the maximum likelihood principle), while AD stands for area difference of the pdf. Still in the legend, X+Y means that X phases was used to fit the body, while Y to fit the tail. Figures 16 shows the effect of Phase type fitting on the M/G/1 queue behaviour with Pareto II service (utilization is 0.8). Exact result of the M/G/1 queue was computed with the method of [45].

At this point we take detour to discrete-time models. Discrete-time counterpart of the fitting method, i.e. when discrete-time PH distributions are applied, is given in [24]. We apply discrete PH distributions to fit the EPA trace. The ccdf of the body and the tail of the resulting discrete PH distribution are shown in Figure 17 and 18. In Figure 18 we depicted the polynomial fit of the tail behaviour as well.

5.2 MAP fitting based on samples

Similarly to the case of PH fitting, MAP fitting methods can be classified as general and heuristic ones. General methods utilize directly the data samples, and hence they do not require any additional engineering knowledge. Our numerical experiences show that MAP fitting is a far more difficult task than PH fitting.

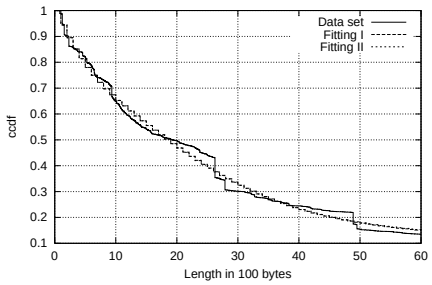

Fig. 17. Body of the approximating distributions

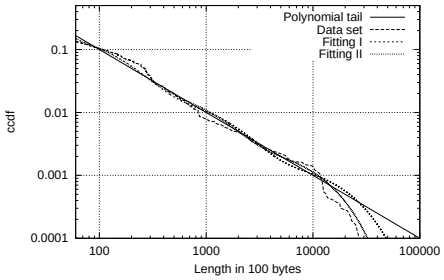

Fig. 18. Tail of the approximating distributions

A simple explanation is that fitting a process is more difficult than fittign a distribution. Capturing slow decaying behaviour with general MAP fitting seems impossible.

Anyhow, there are numerical methods available for fitting low order MAPs directly to datasets. In [32, 15, 46] a fitting method based on maximum likelihood estimate is presented, and in [47] the EM method is used for maximizing the likelihood estimate.

Simple numerical tests (like taking a MAP, drawing samples from it, and fitting a MAP of the same order to these samples) often fail for MAPs of higher order (≥ 3) and the accuracy of the method does not necessarily improve with increasing number of samples.

5.3 Heuristic MAP fitting

An alternative to general MAP fitting is to extract a set of (hopefully) dominant properties of the traffic process from the dataset and to create a MAP (of particular structure) that exhibits the same properties. This kind of heuristic methods fail to satisfy the above mentioned "self test" by their nature, but if the selected set of parameters are really dominant with respect to the goal of the analysis we can achieve "sufficient" fitting. [19] proposed to fit the following parameters: mean arrival rate, variance to mean ratio of arrivals in $(0, t)$, and its asymptotic limit. After the notion of long range dependence in traffic processes the Hurst parameter was added to this list. The following subsections introduces heuristic fitting methods with various properties to capture and various fitting MAP structures.

MAP structures approximating long range dependent behaviour An intuitive way to provide long range dependent behaviour for several time scales with Markovian models is to compose a combined model from small pieces each of which represents the model behaviour at a selected range of the time scales. One of the first models of this kind was proposed in [44]. The same approach was applied for traffic fitting in [38], but recently this approach is criticized

in [12]. Renewal processes with heavy tailed interarrival times also exhibit self-similar properties. Using this fact the approximate heavy tailed PH distributions can be used to create a MAP with PH renewal process. In [1] superposition of 2 state MMPPs are used for approximating 2nd order self-similarity. The proposed procedure fits the mean arrival rate, the 1-lag correlation, the Hurst parameter and the required range of fitting.

Fitting based on separate handling of long- and short-range dependent behavior In [21] a procedure is given to construct a MAP such a way that some parameters of the traffic generated by the model match predefined values. The following parameters are set:

- The fundamental arrival rate describes the expected number of arrivals in a time unit.
- In order to describe the burstiness of the arrival stream, the index of dispersion for counts $I(t) = \mathrm{Var}(N_t)/E(N_t)$ is set for two different values of time: $I(t_1)$ and $I(t_2)$. The choice of these two time points significantly affects the goodness of fitting.
- A higher order descriptor, the third centralized moment of the number of arrivals in the interval $(0, t_3)$, $M(t_3) = E[(N_{t_3} - E(N_{t_3}))^3]$ is set.
- The degree of *pseudo* self-similarity is defined by the Hurst parameter H. The Hurst parameter is realized in terms of the variance-time behavior of the resulting traffic, i.e., the straight line fitted by regression to the variance-time curve in a predefined interval has slope $2(H-1)$.

The MAP resulting from the procedure is the superposition of a PH arrival process and a two-state MMPP. In the following we sketch how to construct a PH arrival process with *pseudo* self-similar behavior and describe the superposition of this PH arrival process with a two-state MMPP. Detailed description of the procedure is given in [21].

Let us consider an arrival process whose interarrival times are independent random variables with heavy tail probability density function (pdf) of Pareto type

$$f(x) = \frac{c \cdot a^c}{(x+a)^{c+1}}, \;\; x \geq 0. \tag{13}$$

The process X_n $(n > 0)$ representing the number of arrivals in the nth time-slot is asymptotically second-order self-similar with Hurst parameter $H = (3-c)/2$ ([49]).

Using the method of Feldman and Whitt [14] one may build an arrival process whose interarrival times are independent, identically distributed PH random variables with pdf approximating (13). To check *pseudo* self-similarity of this PH renewal processes Figure 19 plots $Var(X^{(m)})$ of PH arrival processes whose interarrival time is a 6 phase PH approximation of the pdf given in (13) for different values of c. As it can be observed $Var(X^{(m)})$ is close through several orders of magnitude to the straight line corresponding to the self-similar case with slope $2(H-1)$. The aggregation level where $Var(X^{(m)})$ drops compared to

the straight line may be increased by changing the parameters of the PH fitting algorithm.

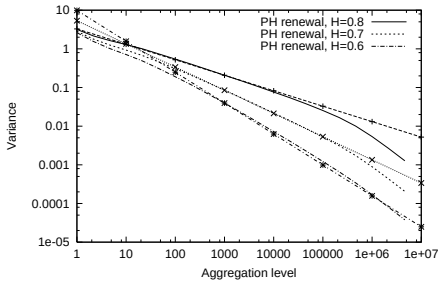

Fig. 19. Variance-time plot of *pseudo* self-similar arrival processes with i.i.d. PH interarrival

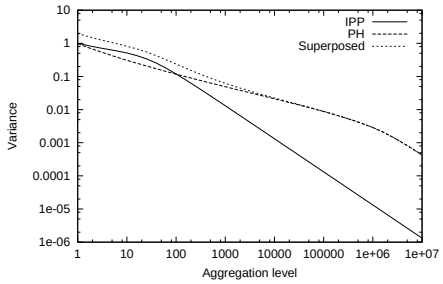

Fig. 20. Superposition of the PH arrival process with an MMPP

The parameters of the two-state MMPP with which the PH arrival process is superposed are calculated in two steps:

1. At first we calculate the parameters of an Interrupted Poisson Process (IPP). The IPP is a two-state MMPP that has one of its two arrival rates equal to 0. The calculated parameters of the IPP are such that the superposition of the PH arrival process and the IPP results in a traffic source with the desired first and second order parameters $E(N_1)$, $I(t_1)$ and $I(t_2)$.
2. In the second step, based on the IPP we find a two-state MMPP that has the same first and second order properties as the IPP has (recalling results from [4]), and with which the superposition results in the desired third centralized moment.

If the MMPP is "less long-range dependent" than the PH arrival process, the *pseudo* self-similarity of the superposed traffic model will be dominated by the PH arrival process. This fact is depicted in Figure 20. It can be observed that if the Hurst parameter is estimated based on the variance-time plot the Hurst parameter of the superposed model is only slightly smaller than the Hurst parameter of the PH arrival process. In numbers, the Hurst parameter of the PH arrival process is 0.8 while it is 0.78 for the superposed model (based on the slope in the interval $(10, 10^6)$). This behavior is utilized in the fitting method to approximate the short and long range behavior in a separate manner.

We illustrate the procedure by fitting the Bellcore trace. Variance-time plots of the traffic generated by the MAPs resulted from the fitting are depicted in Figure 21. The curve signed by (x_1, x_2) belongs to the fitting when the first (second) time point of fitting the IDC value, t_1 (t_2), is x_1 (x_2) times the expected interarrival time. R/S plots for both the real traffic trace and the traffic generated by the approximating MAPs are given in Figure 22. The fitting of the traces were tested by a $\bullet$/D/1 queue, as well. The results are depicted in Figure 23.

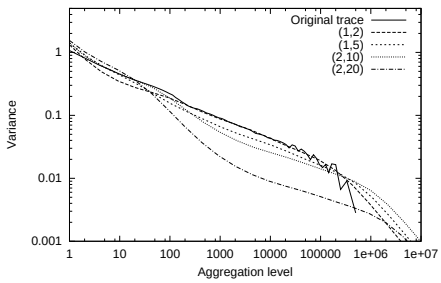

Fig. 21. Variance-time plots of MAPs with different time points of IDC matching

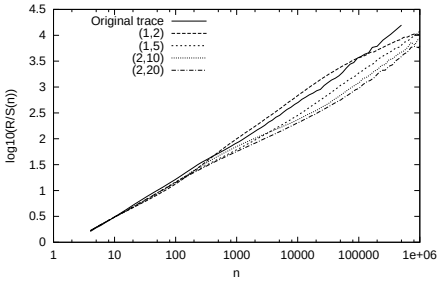

Fig. 22. R/S plots of MAPs with different time points of IDC matching

The $\bullet$/D/1 queue was analyzed by simulation with different levels of utilization of the server. As one may observe the lower t_1 and t_2 the longer the queue length distribution follows the original one.

The fitting method provides a MAP whose some parameters are the same as those of the original traffic process (or very close). Still, the queue length distribution does not show a good match. This means that the chosen parameters do not capture all the important characteristics of the traffic trace.

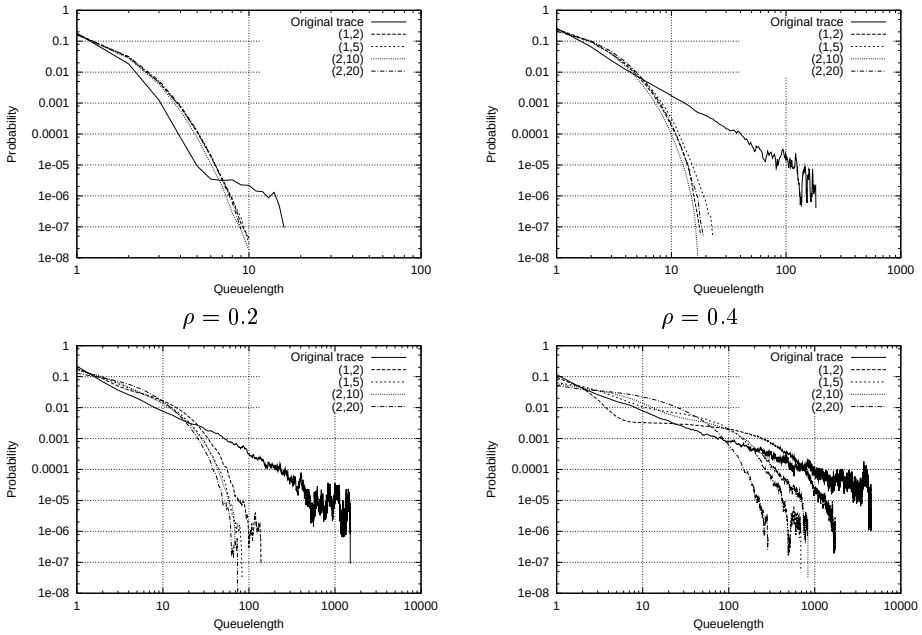

Fig. 23. Queue-length distribution

MMPP Exhibiting Multifractal Behavior In [23] a special MMPP structure is proposed to exhibit multifractal behavior. The background CTMC of the MMPP has a symmetric n-dimensional cube structure and the arrival intensities are set according to the variation of the arrival process at the different time scales. The special choice of the structure is motivated by the generation of the Haar wavelet transform. Basically the Haar wavelet transform evaluates the variation of the dataset at different aggregation levels (time scales), and similarly, the proposed MMPP structure provide different variation of the arrival rate at different time scales.

The composition of the proposed MMPP structure is similar to the generation of the Haar wavelet transform (a procedure for traffic trace generation based on this transform is introduced in [43]). Without loss of generality, we assume that the time unit is such that the long term arrival intensity is one. A MMPP of one state with arrival rate 1 represents the arrival process at the largest (considered) time scale.

At the next time scale, $1/\lambda$, an MMPP of two states with generator

$$\begin{array}{|cc|}\hline -\lambda & \lambda \\ \lambda & -\lambda \\ \hline \end{array}$$

and with arrival rates $1-a_1$ and $1+a_1$ ($-1 \leq a_1 \leq 1$) represents the variation of the arrival process. This composition leaves the long term average arrival rate unchanged.

In the rest of the composition we perform the same step. We introduce a new dimension and generate the n-dimensional cube such that the behavior at the already set time scales remains unchanged. E.g., considering also the $1/\gamma\lambda$ time scale an MMPP of four states with generator

$$\begin{array}{|cc|cc|}\hline \bullet & \lambda & \gamma\lambda & \\ \lambda & \bullet & & \gamma\lambda \\ \hline \gamma\lambda & & \bullet & \lambda \\ & \gamma\lambda & \lambda & \bullet \\ \hline \end{array}$$

and with arrival rates $(1-a_1)(1-a_2)$, $(1+a_1)(1-a_2)$, $(1-a_1)(1+a_2)$ and $(1+a_1)(1+a_2)$ ($-1 \leq a_1, a_2 \leq 1$) represents the variation of the arrival process. With this MMPP, parameter a_1 (a_2) determines the variance of the arrival process at the $1/\lambda$ ($1/\gamma\lambda$) time scale. If γ is large enough ($>\sim 30$) the process behavior at the $1/\lambda$ time scale is independent of a_2. The proposed model is also applicable with a small γ. In this case, the only difference is that the model parameters and the process behavior of different time scales are dependent.

A level n MMPP of the proposed structure is composed by 2^n states and it has $n+2$ parameters. Parameters γ and λ defines the considered time scales, and parameters $a_1, a_2, \ldots, a_n$ determines the variance of the arrival process at the n considered time scales. It can be seen that the ratio of the largest and the smallest considered time scales is γ^n. Having a fixed n (i.e., a fixed cardinality of the MMPP), any large ratio of the largest and the smallest considered time scales can be captured by using a sufficiently large γ.

A simple numerical procedure can be applied to fit a MMPP of the given structure to a measured dataset. This heuristic approach is composed by "engineering considerations" based on the properties of the measured dataset and a parameter fitting method.

First, we fix the value of n. According to our experience a "visible" multiscaling behavior can be obtained from $n = 3 \sim 4$. The computational complexity of the fitting procedure grows exponentially with the dimension of the MMPP. The response time with $n = 6$ (MMPP of 64 states) is still acceptable (in the order of minutes).

Similarly to [43], we set γ and the λ based on the inspection of the dataset. Practically, we define the largest, T_M, and the smallest, T_m, considered time scales and calculate γ and λ from

$$T_M = \frac{1}{\lambda}\ ; \qquad T_m = \frac{1}{\gamma^n \lambda}.$$

The extreme values of T_M and T_m can be set based on simple practical considerations. For example when the measured dataset is composed by N arrival instances, T_M can be chosen to be less than the mean time of $N/4$ arrivals, and T_m can be chosen to be greater than the mean time of 4 arrivals. A similar approach was applied in [43]. These boundary values can be refined based on a detailed statistical test of the dataset. E.g., if the scaling behavior disappears beyond a given time scale, T_M can be set to that value.

Having γ and λ, we apply a downhill simplex method to find the optimal values of the variability parameters $a_1, a_2, \ldots, a_n$. The goal function that our parameter fitting method minimizes is the sum of the relative errors of the second moment of Haar wavelet coefficients up to a predefined time scale S:

$$\min_{a_1,\ldots,a_n} \sum_{j=1}^{S} \frac{|E(d_j^2) - E(\hat{d}_j^2)|}{E(d_j^2)}\ .$$

The goal function can be calculated analytically as it is described in [23].

Application of the fitting procedure is illustrated on the Bellcore trace. We applied the fitting method with $n = 5$ and several different predefined setting of γ, λ. We found that the goodness of the fitting is not very sensitive to the predefined parameters around a reasonable region. The best "looking" fit is obtained when T_m is the mean time of 16 arrivals and $\gamma = 8$. In this case T_M is the mean time of $16{*}8^5 = 2^{19}$ arrivals which corresponds to the coarsest time scale we can analyze in the case of the Bellcore trace. The simplex method minimizing the sum of the relative error of the second moments of the Haar wavelet coefficients over $S = 12$ time scales resulted in: $a_1 = 0.144, a_2 = 0.184, a_3 = 0.184, a_4 = 0.306, a_5 = 0.687$. The result of fitting the second moment of the Haar wavelet transform at different aggregation levels is plotted in Figure 24. At small time scales the fitting seems to be perfect, while at larger time scales the error enlarges. The slope of the curves are almost equal in the depicted range.

The multiscaling behavior of the obtained MAP and of the original dataset are illustrated via the log-moment curves in Figure 25. In the figure, the symbols

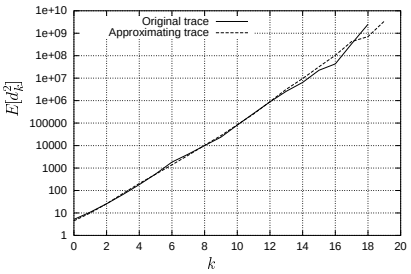

Fig. 24. The second moment of the Haar wavelet transform at different aggregation levels

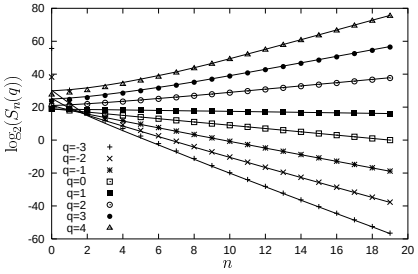

Fig. 25. Scaling of log-moments of the original trace and the fitting MMPP

represent the log-moment curves of the fitting MAP and the solid lines indicate the corresponding log-moment curves of the Bellcore trace. In the range of $n \in (3, 19)$ the log-moment curves of the fitting MAP are very close to the ones of the original trace. The log-moment curves of the approximate MAP are also very close to linear in the considered range.

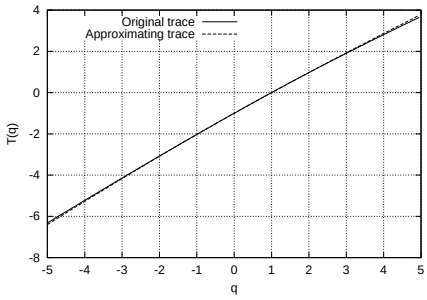

Fig. 26. Partition function estimated through the linear fits shown in Figure 25

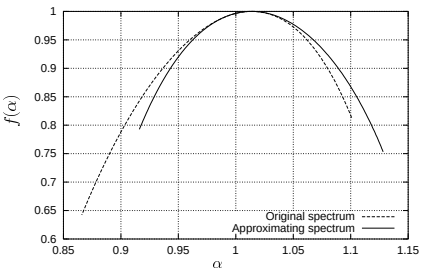

Fig. 27. The Legendre transform of the original dataset and the one of the approximate MMPP

The partition functions of the fitting MAP and of the original trace are depicted in Figure 26. As it was mentioned earlier, the visual appearance of the partition function is not very informative about the multifractal scaling behavior. Figure 27 depicts the Legendre transform of the partition functions of the original dataset and the approximating MAP. The visual appearance of the Legendre transform significantly amplifies the differences of the partition functions. In Figure 27, it can be seen that both processes exhibit multifractal behavior but the original dataset has a bit richer multifractal spectrum.

We also compared the queuing behavior of the original dataset with that of the approximate MAP assuming deterministic service time and different levels of utilization, ρ. Figure 28 depicts the queue length distribution resulting from the

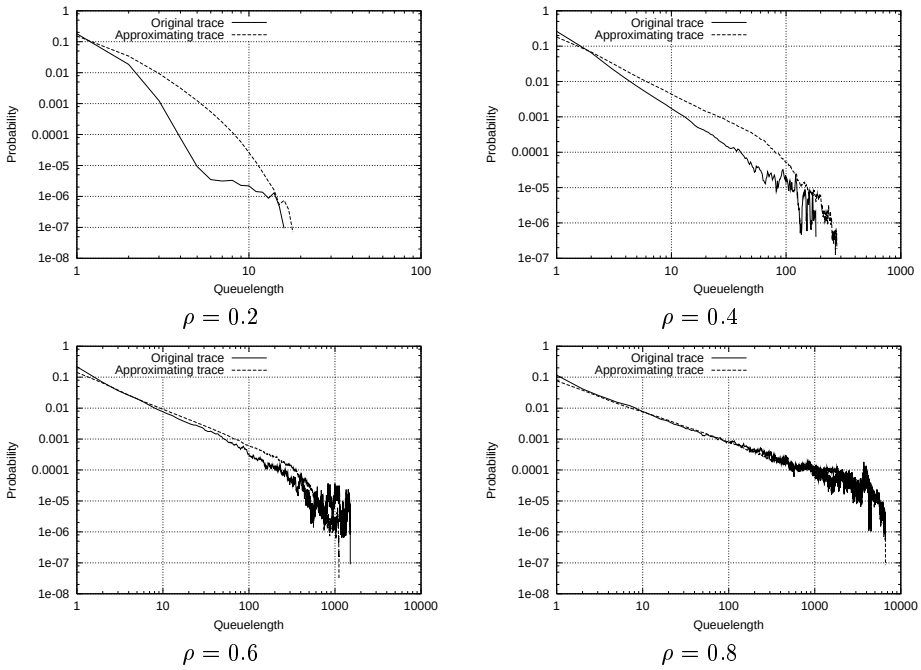

Fig. 28. Queue-length distribution

original and the approximate arrival processes. The queue length distribution curves show a quite close fit. The probability of an empty queue, which is not displayed in the figures, is the same for the MAP as for the original trace since the MAP has the same average arrival intensity as the original trace. The fit is better with a higher queue utilization, which might mean that different scaling behaviors play a dominant rule at different utilizations, and the ones that are dominant at high utilization are better approximated by the proposed MAP.

6 Conclusions

This paper collects a set of methods which can be used in practice for measurement based traffic engineering. The history of traffic theory of high speed communication networks is summarized together with a short introduction to the mathematical foundation of the applied concepts. The common statistical methods for the analysis of data traces and the practical problems of their application is discussed.

The use of Markovian methods is motivated by the fact that an effective analysis technique, the matrix geometric method, is available for the evaluation of Markovian queuing systems. To obtain the Markovian approximation of measured traffic data a variety of heuristic fitting methods are applied. The properties and abilities of these methods are also discussed.

The presented numerical examples provide insight to the qualitative understanding of the strange traffic properties of high speed networks.

References

1. A. T. Andersen and B. F. Nielsen. A markovian approach for modeling packet traffic with long-range dependence. *IEEE Journal on Selected Areas in Communications*, 16(5):719–732, 1998.
2. S. Asmussen and O. Nerman. Fitting Phase-type distributions via the EM algorithm. In *Proceedings: "Symposium i Advent Statistik"*, pages 335–346, Copenhagen, 1991.
3. J. Beran. *Statistics for long-memory processes.* Chapman and Hall, New York, 1994.
4. A. W. Berger. On the index of dispersion for counts for user demand modeling. In *ITU*, Madrid, Spain, June 1994. Study Group 2, Question 17/2.
5. A. Bobbio, A. Horváth, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. submitted to Performance Evaluation, 2000.
6. S. C. Borst, O. J. Boxma, and R. Nunez-Queija. Heavy tails: The effect of the service discipline. In *Tools 2002*, pages 1–30, London, England, April 2002. Springer, LNCS 2324.
7. G. E. P. Box, G. M Jenkins, and C. Reinsel. *Time Series Analysis: Forecasting and Control.* Prentice Hall, Englewood Cliff, N.J., third edition, 1994.
8. E. Castillo. *Extreme Value Theory in Engineering.* Academic Press, San Diego, California, 1988.
9. M. E. Crovella and M. S. Taqqu. Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, 1(1):55–79, 1999.
10. A. Cumani. On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics and Reliability*, 22:583–602, 1982.
11. R. El Abdouni Khayari, R. Sadre, and B. Haverkort. Fitting world-wide web request traces with the EM-algorithm. In *Proc. of SPIE*, volume 4523, pages 211–220, Denver, USA, 2001.
12. R. El Abdouni Khayari, R. Sadre, and B. Haverkort. A valiadation of the pseudo self-similar traffic model. In *Proc. of IPDS*, Washington D.C., USA, 2002.
13. A. Feldman, A. C. Gilbert, and W. Willinger. Data networks as cascades: Investigating the multifractal nature of internet WAN traffic. *Computer communication review*, 28/4:42–55, 1998.
14. A. Feldman and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31:245–279, 1998.
15. W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.
16. H. J. Fowler and W. E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE JSAC*, 9(7):1139–1149, 1991.
17. R. Fox and M. S. Taqqu. Large sample properties of parameter estimates for strongly dependent stationary time series. *The Annals of Statistics*, 14:517–532, 1986.
18. C. W. J. Granger and R. Joyeux. An introduction to long-memory time series and fractional differencing. *Journal of Time Series Analysis*, 1:15–30, 1980.

19. H. Heffes and D. M. Lucantoni. A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, 1986.
20. B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174, 1975.
21. A. Horváth, G. I. Rózsa, and M. Telek. A map fitting method to approximate real traffic behaviour. In *8th IFIP Workshop on Performance Modelling and Evaluation of ATM & IP Networks*, pages 32/1–12, Ilkley, England, July 2000.
22. A. Horváth and M. Telek. Approximating heavy tailed behavior with phase type distributions. In *3rd International Conference on Matrix-Analytic Methods in Stochastic models*, Leuven, Belgium, 2000.
23. A Horváth and M. Telek. A markovian point process exhibiting multifractal behaviour and its application to traffic modeling. In *Proc. of MAM4*, Adelaide, Australia, 2002.
24. A. Horváth and M. Telek. Phfit: A general phase-type fitting tool. In *Proc. of 12th Performance TOOLS*, volume 2324 of *Lecture Notes in Computer Science*, pages 82–91, Imperial College, London, April 2002.
25. M. Kratz and S. Resnick. The qq–estimator and heavy tails. *Stochastic Models*, 12:699–724, 1996.
26. A. Lang and J. L. Arthur. Parameter approximation for phase-type distributions. In S. R. Chakravarty and A. S. Alfa, editors, *Matrix-analytic methods in stochastic models*, Lecture notes in pure and applied mathematics, pages 151–206. Marcel Dekker, Inc., 1996.
27. G. Latouche and V. Ramaswami. *Introduction to Matrix-Analytic Methods in Stochastic Modeling*. Series on statistics and applied probability. ASA-SIAM, 1999.
28. W. E. Leland, M. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions in Networking*, 2:1–15, 1994.
29. D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Commun. Statist.-Stochastic Models*, 7(1):1–46, 1991.
30. B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1969.
31. B. B. Mandelbrot and M. S. Taqqu. Robust R/S analysis of long-run serial correlation. In *Proceedings of the 42nd Session of the International Statistical Institute*, volume 48, Book 2, pages 69–104, Manila, 1979. Bulletin of the I.S.I.
32. K.S. Meier. A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operations Research*, 29:370–377, 1987.
33. M. Neuts. Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206. University of Louvain, 1975.
34. M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, 1981.
35. M.F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications*. Marcel Dekker, 1989.
36. I. Norros. A storage model with self-similar imput. *Queueing Systems*, 16:387–396, 1994.
37. I. Norros. On the use of fractional brownian motion in the theorem of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13:953–962, 1995.
38. A. Ost and B. Haverkort. Modeling and evaluation of pseudo self-similar traffic with infinite-state stochastic petri nets. In *Proc. of the Workshop on Formal Methods in Telecommunications*, pages 120–136, Zaragoza, Spain, 1999.

39. S. Resnick. Heavy tail modeling and teletraffic data. *The Annals of Statistics*, 25:1805–1869, 1997.
40. S. Resnick and C. Starica. Smoothing the hill estimator. *Advances in Applied Probability*, 29:271–293, 1997.
41. J. Rice. *Mathematical Statistics and Data Analysis*. Brooks/Cole Publishing, Pacific Grove, California, 1988.
42. R. H. Riedi. An introduction to multifractals. Technical report, Rice University, 1997. Available at `http://www.ece.rice.edu/~riedi`.
43. R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Transactions on Information Theory*, 45:992–1018, April 1999.
44. S. Robert and J.-Y. Le Boudec. New models for pseudo self-similar traffic. *Performance Evaluation*, 30:1997, 57-68.
45. M. Roughan, D. Veitch, and M. Rumsewicz. Numerical inversion of probability generating functions of power-law tail queues. tech. report, 1997.
46. T. Rydén. Parameter estimation for Markov Modulated Poisson Processes. *Stochastic Models*, 10(4):795–829, 1994.
47. T. Ryden. An EM algorithm for estimation in Markov modulated Poisson processes. *Computational statist. and data analysis*, 21:431–447, 1996.
48. T. Rydén. Estimating the order of continuous phase-type distributions and markov-modulated poisson processes. *Stochastic Models*, 13:417–433, 1997.
49. B. Ryu and S. B. Lowen. Point process models for self-similar network traffic, with applications. *Stochastic models*, 14, 1998.
50. M. Telek and A. Heindl. Moment bounds for acyclic discrete and continuous phase-type distributions of second order. In *in proc. of Eighteenth Annual UK Performance Engineering Workshop (UKPEW)*, Glasgow, UK, 2002.
51. J. Lévy Véhel and R. H. Riedi. Fractional brownian motion and data traffic modeling: The other end of the spectrum. In C. Tricot J. Lévy Véhel, E. Lutton, editor, *Fractals in Engineering*, pages 185–202. Springer, 1997.
52. The internet traffic archive. http://ita.ee.lbl.gov/index.html.