PhFit: A General Phase-Type Fitting Tool*

András Horváth and Miklós Telek

Dept. of Telecommunications, Technical University of Budapest, 1521 Budapest, Hungary, {horvath,telek}@webspn.hit.bme.hu

Abstract. PhFit, a new Phase-type fitting tool is presented in this paper. The presented tool is novel from several aspects. It allows for approximating distributions or set of samples not only by continuous but by discrete Phase-type distributions as well. The implemented algorithms separate the fitting of the *body* and the *tail* part of the distribution which results in satisfactory fitting also for heavy-tail distributions. Moreover, PhFit allows the user to choose the distance measure according to which the fitting is performed.

1 Introduction

In the recent decades, a lot of research was carried out to handle stochastic models in which durations are Phase-type (PH) distributed. Queuing models with PH distributed interarrival times and/or PH distributed service times are considered in numerous papers. In addition, PH distributions are introduced into different modeling formalisms, e.g. Petri Nets (PN).

In order to exploit advances in handling models with PH distributed activity durations, one needs algorithms to determine the parameters of the applied PH distribution. The two main classes of fitting methods differ in the kind of information they utilize. The methods that use *incomplete* information to perform the fitting results in a PH distribution whose selected parameters match those parameters of the original distribution. These methods typically use the moments of the original distribution. Nevertheless, it is possible to use other parameters as well like, for example, the procedure proposed in [3] that use the value of the complementary cumulative distribution function (ccdf) at a given set of points. The methods that use the *complete* distribution to perform fitting, i.e. the fitting is aimed at minimizing the *distance* between the original and the approximating distribution according to a given way of measuring the distance between the two distributions. MLAPH and EMPHT implement the maximum-likelihood principle which results in minimizing the cross entropy. The procedure presented in [4] gives the possibility to apply an arbitrary distance measure (a measure of the difference between the original and the approximating PH distribution).

Naturally, it is possible to combine the above concepts in a single fitting procedure. Given a set of samples, MEDA provides a PH distribution whose

* This work was partially supported by Hungarian Scientific Research Fund (OTKA) under Grant No. T-34972.

first three moments are identical to the empirical moments given by the samples, and, moreover, it minimizes the area difference between the empirical and the approximating cumulative distribution function. A detailed comparison of all the above mentioned statistical fitting procedures can be found in [5].

PhFit performs the fitting using the combined algorithm described in [4]. The body of the distribution is fitted according to a distance measure chosen by the user, while the *tail* part is fitted applying the heuristic method of [3]. In this paper, in addition to the description of the tool, the method described in [4] is reformulated for fitting discrete PH distributions as well.

The paper is organized as follows. Section 2 gives a short overview of both continuous and discrete PH distributions and their properties. Section 3 describes the algorithm implemented in the tool to approximate distributions. Application of the tool is illustrated in Section 4 and the paper is concluded in Section 5.

2 Phase-type distributions

A DPH distribution is the distribution of the time to absorption in a DTMC with n transient states and one absorbing state. The one-step transition probability matrix of the corresponding DTMC can be partitioned as

$$\widehat{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{b} \\ \boldsymbol{0} & 1 \end{bmatrix} \tag{1}$$

where $\boldsymbol{B} = [b_{ij}]$ is the $(n \times n)$ matrix collecting the transition probabilities among the transient states, $\boldsymbol{b} = [b_{i,n+1}]^T, 1 \leq i \leq n$ is a vector of length n that groups the probabilities from any state to the absorbing one, and $\boldsymbol{0}$ is the zero vector. The initial probability vector $\widehat{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}, \alpha_{n+1}]$ is of length $(n+1)$ with $\sum_{j=1}^{n} \alpha_j = 1 - \alpha_{n+1}$. In the tool described hereinafter, we consider only the class of DPH distributions for which $\alpha_{n+1} = 0$.

Similarly, a CPH distribution is the distribution of the time to absorption in a CTMC with n transient states and one absorbing state. The infinitesimal generator of the CTMC $\widehat{\mathbf{Q}}$ can be partitioned in the following way:

$$\widehat{\boldsymbol{Q}} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{q} \\ \boldsymbol{0} & 0 \end{bmatrix} \tag{2}$$

where $\boldsymbol{Q}$ is a $(n \times n)$ matrix that describes the transient behavior of the *CTMC* and $\boldsymbol{q}$ is a column vector of length n that groups the transition rates to the absorbing state. Let $\hat{\boldsymbol{\alpha}} = [\,\boldsymbol{\alpha},\ \alpha_{n+1}\,]$ be the $(n+1)$ initial probability (row) vector with $\sum_{i=1}^{n} \alpha_i = 1 - \alpha_{n+1}$. The tuple $(\boldsymbol{\alpha}, \boldsymbol{Q})$ is called the representation of the CPH, and n the order.

Some properties of continuous and discrete PH distributions are summarized in Table 1, where $\boldsymbol{e}$ denotes the vector with all entries equal 1.

Property	Discrete PH	Continuous PH
probability density (mass) function	$f_i = \boldsymbol{\alpha}\boldsymbol{B}^{i-1}\boldsymbol{b}$	$f(x) = \boldsymbol{\alpha}\exp(\boldsymbol{Q}x)\boldsymbol{q}$
	$f_i \geq 0$	$f(x) > 0,\ x > 0$
cumulative distribution function	$F_i = 1 - \boldsymbol{\alpha}\boldsymbol{B}^{i}\boldsymbol{e}$	$F(x) = 1 - \boldsymbol{\alpha}\exp(\boldsymbol{Q}x)\boldsymbol{e}$
min. coefficient of variation ($n > 0$)	≥ 0	$\geq 1/n$
max. coefficient of variation ($n > 1$)	∞	∞
support	finite or infinite	infinite
tail decay	geometric	exponential

Table 1. Properties of PH distributions

3 Fitting according to distribution functions

3.1 Fitting method for arbitrary distance measure

The aim of PH fitting is to find a PH distribution with such parameters that according to a given measure the PH distribution is close to the distribution that we want to approximate. The parameters to be found are the initial probability vector and the transition matrix. In case of an n-phase PH distribution, the number of free parameters in the initial probability vector is $n-1$ when $F(0) = 0$ and n when $F(0) \geq 0$, while it is n^2 in the transition matrix. In order to decrease the number of free parameters, only cyclic Phase-type (APH) distributions are considered[1].

Since any ACPH or ADPH distribution of order n can be transformed into the canonical form (CF1) in Figure 1 [2, 1] the approximating PH distribution is looked for in this form, described by the vectors $\boldsymbol{a}$ and $\boldsymbol{q}$. This way the number of free parameters is decreased to $2n - 1$.

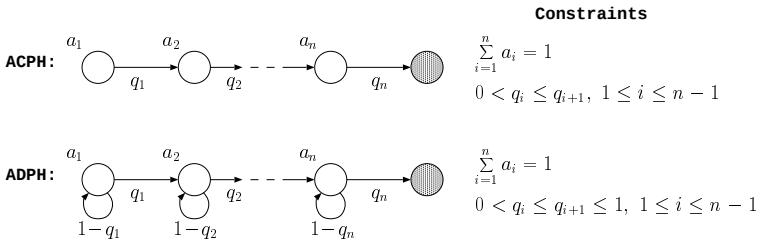

Fig. 1. Canonical form for ACPH and ADPH

The fitting is performed according to a predefined distance measure. Table 2 recites common measures that are implemented in PhFit, where $f(\cdot)$ ($f.$) denotes the probability density function, pdf, (probability mass function, pmf) of the original and $\hat{f}(\cdot)$ ($\hat{f}.$) of the approximating CPH (DPH) distribution.

The PH fitting algorithm is designed to find the vectors $\boldsymbol{a}$ and $\boldsymbol{q}$ with which the distance between the approximating PH distribution and the orig-

[1] The procedure described in this section would be able to fit any Phase type structure by relaxing some constraints.

Distance Measure	Continuous case	Discrete case
Relative entropy	$\int_0^\infty f(t)\ log\left(\frac{f(t)}{\hat{f}(t)}\right) dt$	$\sum_{i=1}^\infty f_i\ log\left(\frac{f_i}{\hat{f}_i}\right)$
pdf area difference	$\int_0^\infty \mid \hat{f}(t) - f(t) \mid dt$	$\sum_{i=1}^\infty \mid \hat{f}_i - f_i \mid$
cdf area difference	$\int_0^\infty (\hat{F}(t) - F(t))^2\, dt$	

Table 2. Distance measures

inal distribution -denoted by $\mathcal{D}(f_{PH}(\boldsymbol{a},\boldsymbol{q}), f)$- is minimal. The algorithm implemented in the PhFit is the following. The initial point $(\boldsymbol{a}^{(0)},\boldsymbol{q}^{(0)})$ is the *best* of 1000 random generated pairs of vectors (satisfying the constraints of the applied canonical form) with proper mean. The *best* means the one that gives the least distance according to the applied distance measure. The distance measure $\mathcal{D}(f_{PH}(\boldsymbol{a}^{(0)},\boldsymbol{q}^{(0)}), f)$ is evaluated numerically.

Starting from the initial guess, the non-linear optimization problem is solved by an iterative linearization method. In each step the partial derivatives with respect to the parameters $(\boldsymbol{a}^{(i)},\boldsymbol{q}^{(i)})$ are numerically computed:

$$\frac{\partial\mathcal{D}(f_{PH}(\boldsymbol{a}^{(i)},\boldsymbol{q}^{(i)}), f)}{\partial a_j^{(i)}}, \qquad \frac{\partial\mathcal{D}(f_{PH}(\boldsymbol{a}^{(i)},\boldsymbol{q}^{(i)}), f)}{\partial q_j^{(i)}}, \qquad j = 1, ..., n.$$

Then, the simplex method is applied to determine the direction in which the distance measure decreases optimally. The constraints of the linear programming is given by probabilistic constraints (the initial probabilities have to sum up to one and, in the discrete case, $0 \leq q_j^{(i+1)} \leq 1, 1 \leq q \leq n$), by the restriction on the structure of the PH distribution (entries of $\boldsymbol{q}^{(i+1)}$ are increasing) and by confining the change of parameters (since the derivatives are valid only in a small area around $(\boldsymbol{a}^{(i)},\boldsymbol{q}^{(i)})$). A search is performed in the direction indicated by the linear programming. The next point of the iteration $(\boldsymbol{a}^{(i+1)},\boldsymbol{q}^{(i+1)})$ is chosen to be the border of the linearized area (defined by the allowed maximum change in the parameters) in the optimal direction if the goal function is decreasing in that direction all the way to the border of the area. The next point is set to the (first) minimum if the goal function has a minimum in the optimal direction inside the linearized area. The iteration is stopped if the relative difference of the parameters in consecutive iteration steps are less than a predefined limit (10^{-5}), or if the number of iterations reaches the predefined limit (1000). The allowed relative change of the parameters greater than 10^{-3} is less than Δ, where Δ starts from 0.1 and is multiplied by 0.995 in each step.

3.2 Fitting heavy-tailed distributions

Recent experiences have shown that the method described in Section 3.1 does not provide satisfactory goodness of fitting for distributions with heavy tail [4]. In [3], Feldman and Whitt suggest a simple, heuristic method to fit heavy-tail behavior with mixture of exponential distributions. Their method is computationally cheap, provides good fitting of the tail behavior but it has the disadvantage of being restricted to decreasing probability density functions. In [4] the method described in Section 3.1 and the heuristic method of [3] are combined

into a powerful general PH fitting algorithm that is capable of capturing both heavy-tails and not only decreasing probability functions.

Hereinafter we reformulate the method of Feldman and Whitt for discrete distributions, and then, in Section 3.3 describe in short the combined method introduced in [4].

The resulting fitting distribution is a mixture of geometric distributions, and, as a consequence, belongs to the family of DPH distributions. Figure 2 depicts the mixture of m geometric distributions as a DPH distribution and gives the related constraints. The probability mass function and the cumulative distribution function are given by

$$f_n = \sum_{i=1}^{m} b_i (1-r_i)^{n-1} r_i, \text{ and } F_n = 1 - \sum_{i=1}^{m} b_i (1-r_i)^n.$$

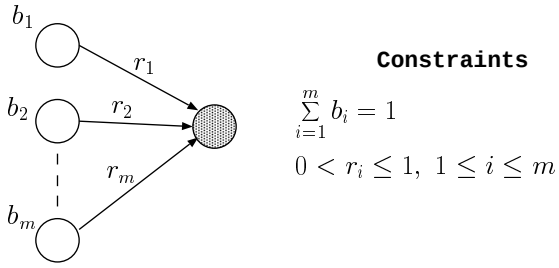

Fig. 2. Mixture of geometric distributions

The proposed heuristic method determines the parameters of the mixture of m geometric distributions ($\boldsymbol{b}$ and $\boldsymbol{r}$) in a recursive manner such that the ccdf of the resulting distribution is "very close" to that of the original distribution at $2m-1$ points. The $2m-1$ points are the following:

$$0 < t_c = t_m < t_{m-1} < bt_{m-1} < \ldots < t_1 = t_d < bt_1,$$

with

$$t_i = t_d \delta^{-i+1}, \quad i \in \{1, 2, \ldots, m\}, \quad \text{where } \delta = \sqrt[m-1]{\frac{t_d}{t_c}}, \text{ and } b < \delta.$$

In order to avoid cumbersome notation, we assume that the introduced $2m-1$ points are integers. If this is not the case, instead of a point t its integer part $\lfloor t \rfloor$ has to be used.

The above $2m-1$ points are defined by m, t_c, t_d and b. The points t_c and t_d define the lower and the upper bound of the range in which the distribution is considered. In the present section t_c will be set always to 1. The upper bound t_d

can be defined by deciding how far the tail should be considered. The constant b, which defines the distance of adjacent points, is defined by z as $b = \delta^z$ with $0 < z < 1$. Further considerations for the choice of the values m, t_c, t_d and b can be found in [3] for the continuous version of the algorithm. These considerations are valid for the discrete case as well.

First, the algorithm provides such b_1 and r_1 that the ccdf of the approximating distribution matches the ccdf of the original distribution F_t^c at t_1 and bt_1. This is done by assuming that the other components of the mixture will not impact the ccdf significantly at t_1 and bt_1, i.e. we assume that $r_i, i = 2, \ldots, m$ will be significantly larger than r_1, and so that

$$\sum_{i=1}^{m} b_i(1-r_i)^t \sim b_1(1-r_1)^t \quad \text{for } t \geq t_1. \tag{3}$$

The parameters b_1 and r_1 are obtained by arranging the two equations

$$b_1(1-r_1)^{t_1} = F_{t_1}^c, \quad \text{and} \quad b_1(1-r_1)^{bt_1} = F_{bt_1}^c,$$

as

$$r_1 = 1 - \sqrt[t_1(1-b)]{\frac{F_{t_1}^c}{F_{bt_1}^c}}, \quad \text{and} \quad b_1 = F_{t_1}^c(1-r_1)^{-t_1}.$$

In the ith step, $2 \leq i \leq m-1$, the algorithm determines b_i and r_i such way that the ccdf of the mixture of geometric distributions matches F_t^c at t_i and bt_i. As when determining b_i and r_i, this is done by assuming that $r_j, j = i+1, \ldots, m$ will be significantly smaller than r_i, so that

$$\sum_{j=1}^{m} b_j(1-r_j)^t \sim \sum_{j=1}^{i} b_j(1-r_j)^t, \quad \text{for } t \geq t_i. \tag{4}$$

The parameters b_i and r_i are obtained by the equalities

$$b_i(1-r_i)^{t_i} = F_{t_i}^c - \sum_{j=1}^{i-1} b_j(1-r_j)^{t_i}, \tag{5}$$

$$b_i(1-r_i)^{bt_i} = F_{bt_i}^c - \sum_{j=1}^{i-1} b_j(1-r_j)^{bt_i}, \tag{6}$$

where the contribution of the components with index lower than i is considered by the extraction on the right hand side of the equations. Using the notation

$$F_{i,t}^c = F_t^c - \sum_{j=1}^{i-1} b_j(1-r_j)^t,$$

rearranging (5) and (6), b_i and r_i are obtained as

$$r_i = 1 - \sqrt[t_i(1-b)]{\frac{F^c_{i,t_i}}{F^c_{i,bt_i}}}, \qquad b_i = F^c_{i,t_i}\ (1-r_i)^{-t_i}. \tag{7}$$

In the mth step, since b_m is already determined by the constraint

$$\sum_{i=1}^{m} b_i = 1, \text{ as } b_m = 1 - \sum_{i=1}^{m-1} b_i,$$

we cannot have both (5) and (6) for $i = m$. So that, instead of using both the points t_m and bt_m, we determine r_m using

$$b_i(1-r_i)^{t_i} = F^c_{m,t_i}, \text{ as } r_i = 1 - \sqrt[t_i]{\frac{F^c_{m,t_i}}{b_i}}.$$

We have guarantee neither that properties (3) and (4), which are assumed throughout the procedure, hold, nor that the sum of the initial probabilities associated with the hyper-geometric part of the Phase type structure ($\sum_{i=1}^{m} b_i$) is lower than 1 [3]. However, these can be easily checked after the procedure is complete, and in general it is not a problem to define the set of points in such way that these properties hold. For example, if the sum is greater than 1 it may help to increase t_d or m. It is discussed in details in [3] how the $2m-1$ points may be chosen efficiently.

In order to illustrate the application of the procedure, let us define the following heavy-tail discrete distribution function with Pareto-like tail:

$$f_n = \frac{1}{\mathcal{Z}(a+1)} n^{-(1+a)}, \tag{8}$$

where $\mathcal{Z}(\cdot)$ denotes the Riemann zeta-function and a determines the decay of the tail.

Figure 3 and 4 depicts the probability mass function (pmf) of approximations of the above defined distribution for different values of a with either 6 or 8 phases.

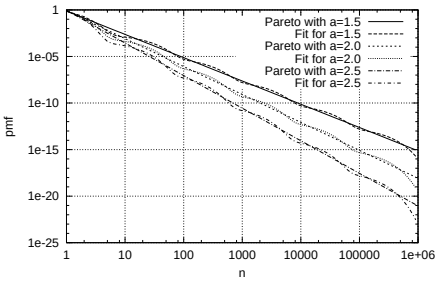

Fig. 3. Fitting different discrete Pareto distributions with 6 phases

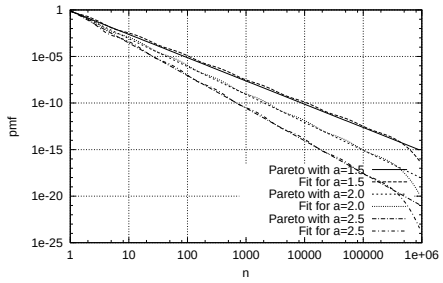

Fig. 4. Fitting different discrete Pareto distributions with 8 phases

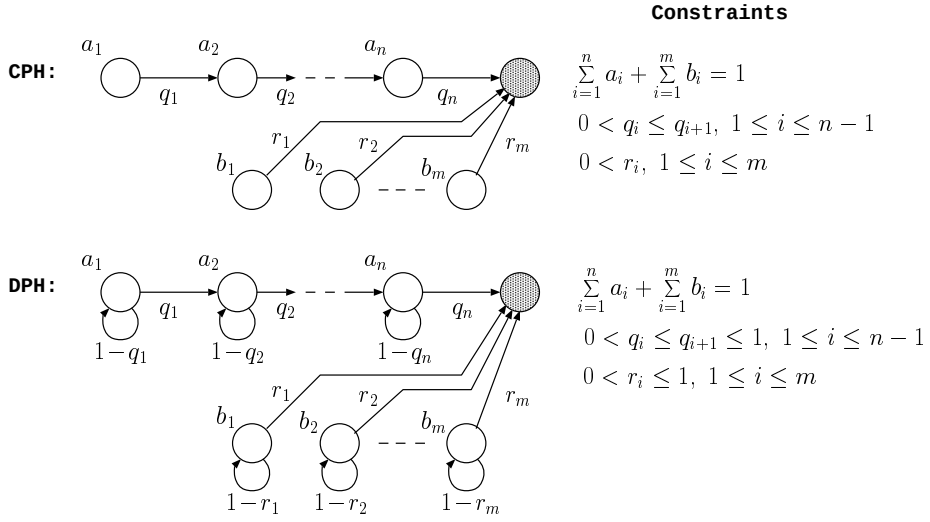

Fig. 5. Structured by CF1 and mixtures

3.3 Combined fitting method

As we mentioned before, the heuristic method presented in Section 3.2 provides distributions with decreasing probability mass function. In order to relax this limitation, the heuristic method is combined with the one described in Section 3.1. Figure 5 depicts the structure and constraints of the resulting PH distribution both for continuous and discrete case.

To this end, we modify the step of the algorithm in which b_m and r_m are defined. As the parameters $b_i, r_i, 2 \leq i \leq m-1$, also the parameters b_m and r_m are defined based on (7). Now the requirement for $\sum_{i=1}^{m} b_i$ is to be less than 1.

In case of fitting with continuous PH distributions the continuous counterpart of the algorithm presented in Section 3.2 is used to determine $\boldsymbol{b}$ and $\boldsymbol{r}$ (see [4]).

Having $\boldsymbol{b}$ and $\boldsymbol{r}$ we use the algorithm described in Section 3.1 to fit the body of the distribution with two differences:

1. Having the hyper-geometric (or hyper-exponential) part the constraint on the initial probabilities modifies to $\sum_{i=1}^{n} a_i = 1 - \sum_{i=1}^{m} b_i$.
2. The structure of the approximate PH distribution differs from the one used before (Figure 1). The parameters associated with the additional m phases ($\boldsymbol{b}$ and $\boldsymbol{r}$) are fixed during this stage of the fitting process.

4 Application of the tool

In this section, we illustrate how the tool can be applied to fit a set of data, in particular, we describe an example with heavy-tail behavior. The fittings are performed with discrete PH distributions.

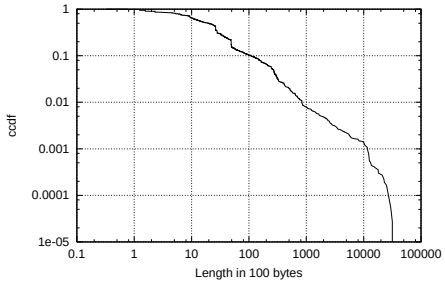

Fig. 6. Experimental ccdf of the length of requests arriving to the server

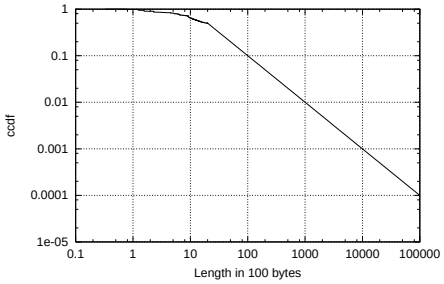

Fig. 7. The ccdf applied during the fitting

Parameter	Fitting I	Fitting II
Number of phases to fit the body	4	
Measure to fit the body	cdf area difference	
Upper bound of fitting the body	100	
Number of phases to fit the tail	3	
Lower limit of fitting the tail (t_3)	100	
Upper limit of fitting the tail (t_1)	10^4	6×10^3
Distance of adjacent points (z)	0.4	

Table 3. Parameters used to perform fitting

The chosen data set, which is called EPA_HTTP and can be downloaded from [6], contains a day of HTTP logs with about 40000 entries from a WWW server. The experimental ccdf of the length of the requests is depicted in Figure 6.

The linearity of the ccdf in the interval [20:10000] (Figure 6) suggests that the tail behavior can be approximated accurately by polynomial (Pareto-like) tail. In order to determine the parameter of the tail, regression is applied to the log-log curve of the ccdf in the interval [20:10000]. Having the parameter, the tail of the experimental ccdf is replaced with a polynomial tail in the interval $[20 : \infty]$ as depicted in Figure 7. The value of the ccdf at 20 is 0.503, so that the tail has this value at 20, and then it has the decay given by the regression.

Having defined the shape and parameter of the tail, the parameters of the fitting itself have to be chosen. Hereinafter two fittings are presented. Table 3 collects the parameters of these two fittings. For both cases, the stepsize of the DPH distribution corresponds to 100 bytes.

Figure 8 and 9 depicts the body and the tail of the ccdf of two possible approximations of the data set. As it can be observed, the only difference between the two fittings is the upper limit of the tail fitting. This difference is clearly reflected by the tails of the two PH distributions. One could easily find such upper limit for the fitting of the tail that the tail of the ccdf of the approximating distribution breaks down where the tail of the experimental ccdf breaks down.

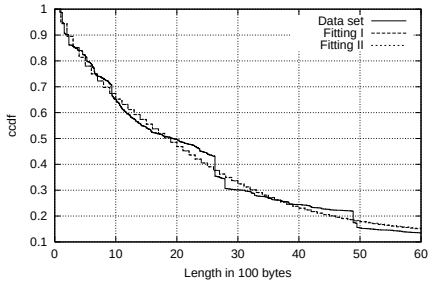

Fig. 8. Body of the approximating distributions

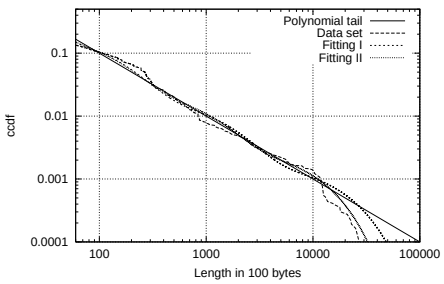

Fig. 9. Tail of the approximating distributions

However, in general, it is not easy to justify that the end of the tail of the random variable under observation is really there where it is suggested to be by its samples. The break we see in the experimental ccdf can easily be caused by the low number of observations as well. This is why -apart from the difficulty of identifying it automatically- we allow the user to choose the upper limit of the tail fitting.

5 Conclusion

This paper presented a new Phase-type fitting tool, PhFit. PhFit offers advanced functionality compared to existing tools from several points of view: i) PhFit allows fitting with both discrete and continuous PH distributions; ii) PhFit offers special treatment to fit slowly decaying tail behaviour; iii) PhFit allows to set the measure according to which the fitting is performed.

References

1. A. Bobbio, A. Horváth, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. Tech. Rep. of TU. Budapest, 2000.
2. A. Cumani. On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics and Rel.*, 583–602, 1982.
3. A. Feldman and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Perf. Eval.*, 31:245–279, 1998.
4. A. Horváth and M. Telek. Approximating heavy tailed behavior with phase type distributions. In *MAM3*, Leuven, Belgium, 2000.
5. A. Lang and J. L. Arthur. Parameter approximation for phase-type distributions. *MAM1*, pp 151–206. Marcel Dekker, Inc., 1996.
6. EPA-HTTP Trace, The Internet Traffic Archive. At `http://ita.ee.lbl.gov`. Collected by Laura Bottomley (`laurab@ee.duke.edu`) of Duke University.