Efficient fitting of long-tailed data sets into hyperexponential distributions

Alma Riska Vesselin Diev Evgenia Smirni
Department of Computer Science
College of William and Mary
Williamsburg, VA 23187-8795, USA

e-mail {riska,vdiev,esmirni}@cs.wm.edu

Abstract— **We propose a new technique for fitting long-tailed data sets into hyperexponential distributions. The approach partitions the data set in a divide and conquer fashion and uses the Expectation-Maximization (EM) algorithm to fit the data of each partition into a hyperexponential distribution. The fitting results of all partitions are combined to generate the fitting for the entire data set. The new method is accurate and efficient and allows one to apply existing analytic tools to analyze the behavior of queueing systems that operate under workloads that exhibit long-tail behavior, such as queues in Internet-related systems.**

I. Introduction

In the process of designing, capacity planning, and resource allocation of network-related systems, engineers are faced with the complexity of the workload experienced by such systems as well as with the complexity of models that can be used for their analysis. Internet-related workloads are characterized by high variability, long-range dependence, or both [1], [10], [14]. These characteristics can be captured well by models such as the Markovian Arrival Process (MAP) and Batch Markovian Process (BMAP) [4], [6], [14] and an entire analytic methodology exist for their analysis [11], [8]. A special and widely used class of MAPs is the phase-type distribution.

Phase-type distributions are characterized by the number of exponential phases and the interaction between them [8]. The latter determines special cases of phase-type distributions such as the Erlang, Coxian, and hyperexponential distributions. Phase-type distributions can capture the long-tail behavior in data sets, that is often observed in Internet-related workloads, making them an attractive modeling tool. The fitting techniques available for phase-type distributions are either based on moment matching or on maximum likelihood estimators (MLEs). Moment matching techniques are computationally efficient, but apply to somewhat more restrictive models [7]. Among the techniques that are based on MLEs, we distinguish a numerical optimization method to fit long-tailed distribution functions into Coxian distributions [6], the Feldmann-Whitt algorithm for fitting distribution functions such as Weibull and Pareto into hyperexponential distributions [5], and the Expectation-Maximization (EM) algorithm for fitting both data and distributions into general phase-type distributions [3].

Accuracy and *efficiency* are equally important in our approach for capturing the long-tail behavior of data sets, because we are driven by the need for *on-line* performance modeling tools. From this perspective, moment matching methods [7] that apply to simple models fail to accurately capture the long-tail behavior, while numerical optimization-based fitting techniques [6] are too complex despite their good fitting accuracy. Fitting methods that deal with distribution functions instead of data sets [5], [6], despite the computational efficiency that some of them provide [5], introduce additional errors in the fitting procedure since one additional step is required to first fit the data set into a distribution function such as Pareto or Weibull. The EM algorithm [3], one of the most popular MLE-based approaches, fits either data sets or distribution functions into general phase-type distributions making EM superior to other fitting methods. However, while the accuracy of the EM algorithm increases with the number of phases, so does its computational complexity. Furthermore, the EM algorithm might fail to accurately capture the tail of the distribution, because it searches for the “global” optimal solution [3].

In this paper we propose a new technique for fitting data sets into hyperexponential distributions. The new technique strives for accuracy and efficiency by applying the popular EM algorithm in a *divide-and-conquer* fashion. The divide-and-conquer approach to the EM algorithm ensures that the new method benefits from the strengths of the EM algorithm and at the same time reduces the effects of its known weaknesses.

The rest of the paper is organized as follows. In Section II we describe the proposed fitting method. In Section III we present experimental results obtained using the proposed method on three different data sets. Section IV discusses the accuracy of the new approach from the queueing systems perspective. Section V reports on the efficiency of the proposed method. Concluding remarks and future directions are outlined in Section VI.

This work has been supported by National Science Foundation under grands EIA-9974992, CCR-0098278, and ACI-0090221.

II. Fitting Method

In this paper we present a new methodology for fitting data sets that exhibit high variability into hyperexponential distributions. The hyperexponential distribution is characterized by the number of exponential phases ph and the mean λ_i and probability p_i associated with each phase. The proposed methodology applies the EM algorithm in a divide and conquer fashion over the initial data set. We call our approach D&C-EM (Divide-and-Conquer-EM).

The high level idea of the proposed method (see Figure 1), is based on the observation that for data sets that exhibit long-tail behavior, it may be beneficial to partition their entire range of values, so as to ensure that each partition exhibits significantly reduced variability in comparison to the variability of the entire data set. The data set of each partition is then fitted into a hyperexponential distribution using the EM algorithm [12] and the final fit for the entire data set is generated by combining together the fitting results for all partitions.

The divide and conquer approach increases the accuracy of the EM algorithm because the portion of the data set belonging to the tail of its continuous data histogram (CDH) [9] could possibly fit in one or more partitions, reducing the possibility that the EM algorithm does not capture it correctly while searching for the global optimal solution. The approach is efficient because each partition has less data entries and less variability than the entire data set, facilitating an accurate fit in a few phases only.

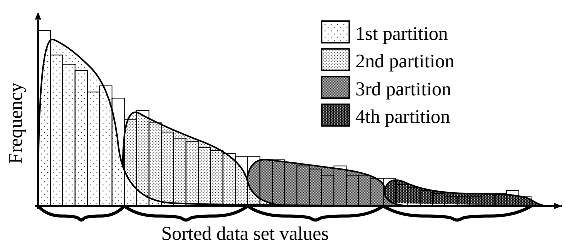

Fig. 1. Splitting of the continuous data histogram (CDH).

Since we are interested to split the CDH such that each partition has reduced variability, we use the coefficient of variation (CV) to determine the partition boundaries. For each partition we accumulate bins until the accumulated coefficient of variation for that partition, CV_{Acc}, is larger than a threshold, CV_{max}. The value of CV_{max} determines the number of partitions for a given data set. We select CV_{max} to be between 1.2 and 1.5, i.e., slightly higher than the CV of the exponential distribution, in order to fit each partition into a hyperexponential distribution with few phases only using EM. The EM algorithm requires as input only the number of phases, ph, and the actual data and in our experiments we have used $ph = 4$.

We generate the final result by combining together the weight of each partition to the entire CDH with its respective fitted hyperexponential distribution. Figure 2 summarizes the D&C-EM algorithm.

1. Build CDH from the data set.
2. while (there are still CDH bins to be considered)
  a include data of current bin into current partition i
  b update CV_{Acc}
  c if ($CV_{Acc} > CV_{max}$)
    Use EM to fit partition i into ph phases
    Obtain p^i_j and λ^i_j, $1 \leq j \leq ph$
    Compute weight w_i for partition i
    $CV_{Acc} = 0$.
3. if ($CV_{last\ partition} < CV_{max}$)
  a Merge last two partitions and perform step **2c**
4. Generate final result
  for i from 1 to # of partitions
    for j from 1 to ph
      $p^i_j = p^i_j \cdot w_i$

Fig. 2. The D&C-EM fitting algorithm.

III. Experimental results

In this section we present results obtained by fitting three different data sets into hyperexponential distributions using the D&C-EM algorithm. We first describe the characteristics of the selected data sets.

A. Workload

We have selected three highly variable data sets to test our approach. The first data set (indicated as "trace 1") is a trace from the 1998 World Soccer Cup Web site. It contains the sizes of the files requested by clients from this Web site in the course of an entire day. The other two traces are synthetically generated from analytic models that closely approximate Web server traffic [1]. Trace 2 is generated from a Lognormal distribution with shape parameter 1.85 and scale parameter 7.0. Trace 3 is generated from a Weibull distribution with shape parameter 0.25 and scale parameter 9.2. The statistical characteristics of these traces are shown in Table I. The number of entries and the number of unique entries for each trace are significant for the performance of the D&C-EM since the running time of the EM algorithm depends on these parameters [12]. Observe that the real trace has less unique entries than the synthetically generated ones.

Trace	Entries	Unique	Mean	CV
1	16045065	12122	4407.80	7.28
2	25000	25000	6358.22	5.87
3	25000	22969	227.26	7.36

TABLE I
Statistical characteristics of the data sets.

B. Fitting results

The size of the data sets precludes using goodness-of-fit tests such as the Kolmogorov-Smirnov and χ^2 tests [9]. Therefore, we evaluate the accuracy of D&C-EM by checking the matching of the first and second moments and by plotting PDFs, CDFs, and CCDFs (Complimentary Cumulative Distribution or Survival function).

Table II illustrates the means and the CVs of the original data sets, plus various hyperexponential fittings using EM and D&C-EM. "EM x ph" means that the EM algorithm is used to fit the entire data set into a hyperexponential with x phases. Observe that the D&C-EM models match the mean of the traces with maximal error of 4 percent and the CV with a maximal error of 20 percent (trace 2). The relatively large error for trace 2 is related to the fact that the hyperexponential distribution is a better fit for data sets with monotone CDHs [5]. The EM algorithm alone could not generate results for traces 3 within reasonable amount of computation time (less than a week). The results of Table II show that the divide and conquer approach captures better the characteristics of the data sets, when compared to the EM algorithm. D&C-EM fits match better even the higher moments of the three data sets that we examine here. We observed a third moment maximum error of 40% for the D&C-EM fits and a maximum error of 90% for the EM fits.

	Data	EM 4 ph	EM 8 ph	D&C-EM
Trace 1				
Mean	4407.80	4355.39	4291.12	4405.22
CV	7.28	3.47	3.45	6.84
Trace 2				
Mean	6358.22	6241.61	6233.73	6343.02
CV	5.87	3.72	4.12	4.73
Trace 3				
Mean	227.26	n/a	n/a	218.41
CV	7.36	n/a	n/a	6.79

TABLE II

Statistical evaluation of the fittings.

Figure 3 plots the PDH, CDF, and CCDF for each trace, the D&C-EM model and the best of the EM models. PDF plots for each of the traces are shown in the first column of graphs in Figure 3 (note the logscale of the x-axis). The PDF of trace 1 is heavily jagged, characteristic of real trace data, which makes matching the PDF more challenging. D&C-EM offers accurate fits for all traces. The fits for trace 2, both D&C-EM and EM ones, do not match well the original PDF at the start. This happens because trace 2 does not have monotone PDF while the hyperexponential distribution has a complete monotone PDF [5].

The CDF plots for all traces (middle column of graphs in Figures 3) illustrate that the D&C-EM provides a good match for the body of the distribution of all traces. In order to investigate the accuracy of the fitting for the distribution tail, we also present the CCDF plots (third column of graphs in Figure 3). Observe that even for the tail of the distribution, which is directly connected with the observed variability of the data sets, D&C-EM generates models that very closely match the data set characteristics.

IV. Queueing System Analysis

Because we want to provide a methodology that allows for analytic modeling of Internet-related systems, we also examine the accuracy of the D&C-EM from a queueing perspective. We consider an $M/H_r/1$ server queue with exponentially distributed interarrival times and service times drawn from a hyperexponential distribution. The hyperexponential model for the service process is generated from the test data sets using either D&C-EM or the EM algorithm. We opt for exponential interarrival times in order to concentrate on the effects of the service process on queueing behavior.

To analyze the $M/H_r/1$ queue that resulted from our fittings, we used the matrix-geometric method that is implemented in the MAMSolver tool [13]. The tool provides the entire stationary probability distribution for the queueing system under study, the average queue length, and the queue length distribution. In our analysis we focus on both the average queue length and the queue length distribution. The queue length distribution is an important metric because it can guide system design and at the same time is a strong indicator of model accuracy.

To examine the accuracy of the fitting algorithm, we compare the above performance measures that are analytically derived using the $M/H_r/1$ queue with those obtained from trace-driven simulations. The simulation model consists of a single server queue with the same arrival process as in the $M/H_r/1$ model and service times from the data sets of Table I. Results are presented in Figure 4. For all data sets, we plot the average queue length as function of the arrival rate (first column of graphs in Figure 4), the body of the queue length distribution (middle column in Figure 4) and the tail of the queue length distribution (last column in Figure 4)[1]. The queue length distributions in Figure 4 correspond to 80% system utilization levels.

We observe that the models obtained from the D&C-EM fitting generate queueing system results that are quite close to the ones obtained from simulation. The $M/H_r/1$ queueing system captures accurately the performance metrics of interest. Consistent with discussions in [5], we note that for traces 1 and 3, whose CDH is completely monotone, the hyperexponential distribution is a

[1] The queue length tail distribution is obtained by representing the queue length distribution in log-log scale plots.

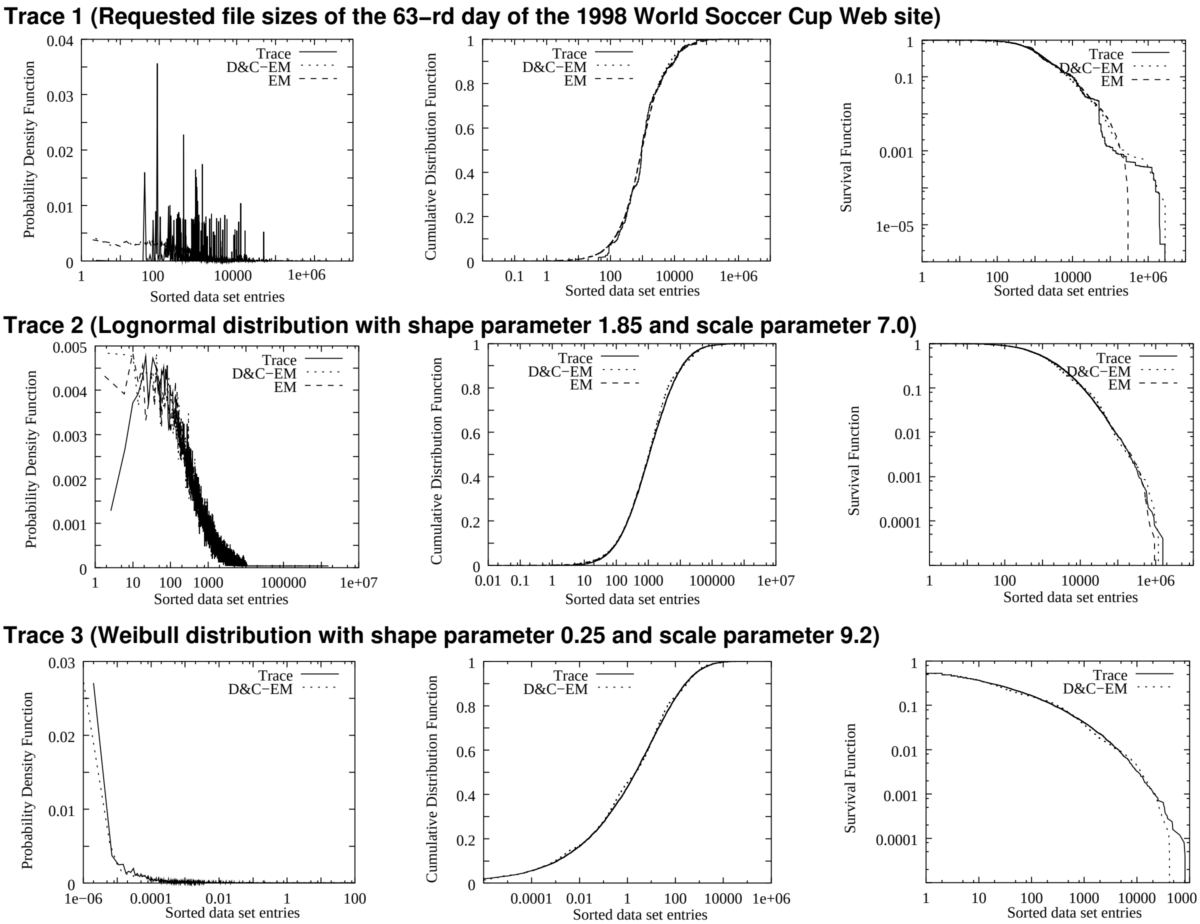

Fig. 3. PDF, CDF, and CCDF of the real data and the fitted models.

better approximation, than for trace 2 whose CDH is not completely monotone.

V. Computational efficiency of D&C-EM

In this section, we report on the computational efficiency of D&C-EM for fitting data sets into hyperexponential distributions. In Figure 5 we illustrate the computation time needed to obtain fittings for the data sets indicated in Table I using the D&C-EM and EM algorithms. The experiments were conducted in a Linux machine with Pentium III 800MHz processor and 1GB of memory. D&C-EM is much faster than EM, even for fittings that consist of few phases only. The efficiency of D&C-EM increases as the number of unique entries in the data sets increases. Note that for trace 3, we were not able to obtain results even for 4-phase models within a timeframe of a week, so the timing results for this trace comes from D&C-EM only.

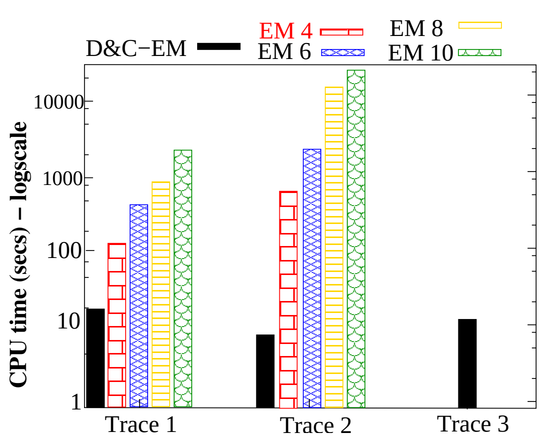

Fig. 5. CPU time for running EM and D&C-EM on all the traces.

VI. Conclusions and Future Work

In this paper we propose a new method to fit data sets exhibiting long-tail behavior into hyperexponential distributions. The proposed method uses the EM algorithm in a divide and conquer fashion, which improves the computational efficiency and the accuracy of the fits. We demonstrate the goodness of fit of the new method from both a statistical perspective and a queueing system perspective. Our results show that the proposed approach is an efficient and accurate way for analyzing performance of Internet-related systems that are characterized by work-

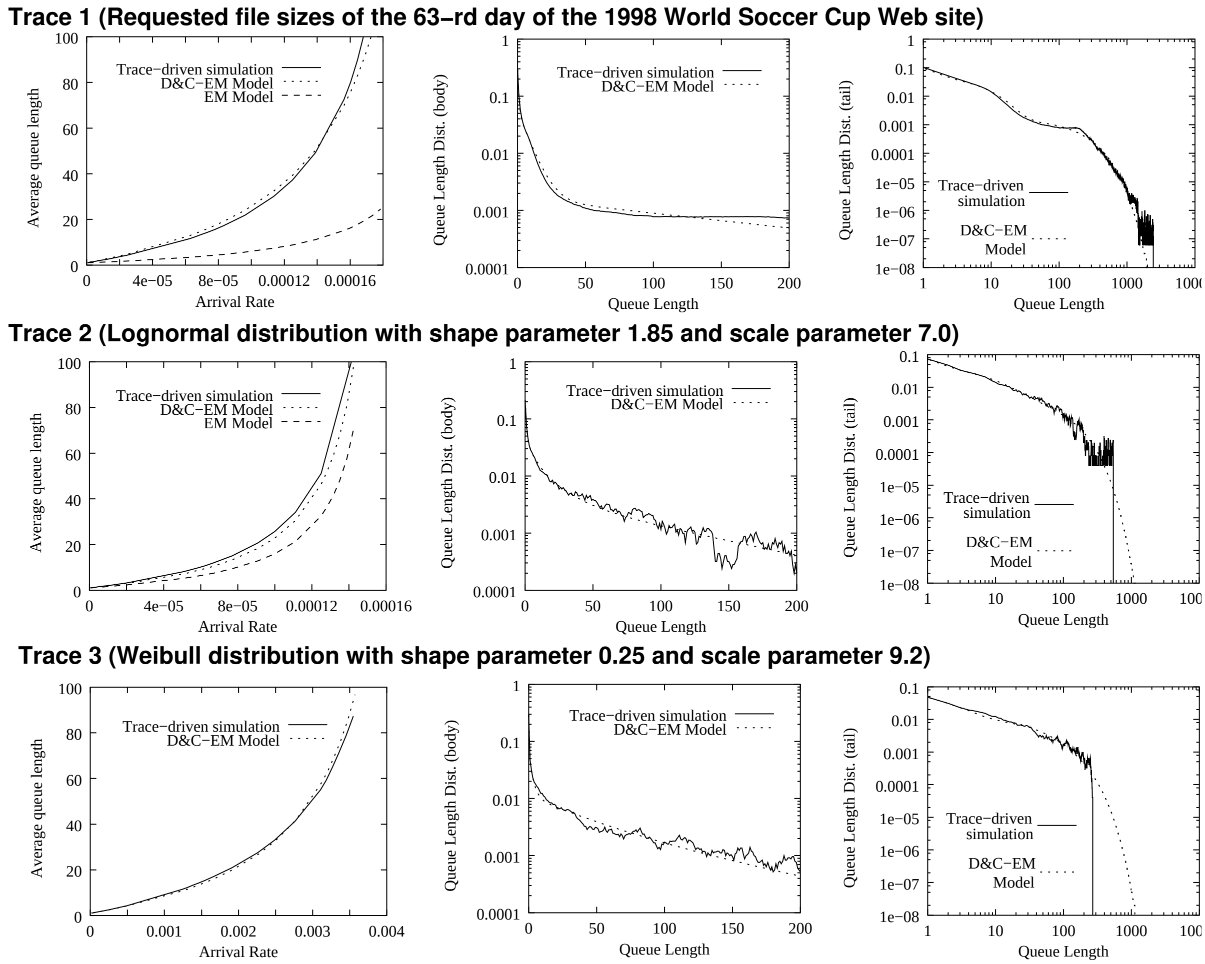

Fig. 4. Average queue length, queue length distribution and tail queue length distribution from simulation and model analysis.

loads of high variability. Currently we are working to improve D&C-EM such that it can be used to fit data sets into more general phase-type distributions. We plan to incorporate D&C-EM in on-line performance monitoring and prediction tools for such systems.

References

[1] M. Arlitt and T. Jin. Workload characterization of the 1998 World Cup Web Site. HP Labs Technical Report, Hewlett Packard, Palo Alto, CA, Sept. 1999.

[2] M. Arlitt and C. L. Williamson. Internet Web servers: Workload characterization and performance implications. *IEEE/ACM Transactions on Networking*, Vol. 5(5), pages 631-645, October 1997.

[3] S. Asmussen, O. Nerman, and M. Olson. Fitting phase type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, Vol. 23, pages 419-441, 1996.

[4] S. Asmussen and G. Koole. Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability*, Vol. 30, pages 365–441, 1993.

[5] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Perf. Eval.*, 31(8):963–976, Aug. 1998.

[6] A. Horvath and M. Telek. Approximating heavy tailed behavior with phase type distribution. In *Advances in Algorithmic Methods for Stochastic Models*, G. Latouche and P. Taylor (eds.), pages 191–214, Notable Publications, 2000.

[7] M. A. Johnson and M. R. Taffe. Matching moments to phase distributions: mixtures of Erlang distribution of common order. *Stochastic Models*, 5:711–743, 1989.

[8] G. Latouche and V. Ramaswami. *Introduction to Matrix Geometric Methods in Stochastic Modeling.* ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia PA, 1999.

[9] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis, Third Edition.* McGraw-Hill, 2000.

[10] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking 2*, pages 1–15, 1994.

[11] M. F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications.* Marcel Dekker, New York, NY, 1989.

[12] M. Olsson. The EMpht-programme. Technical report, Department of Mathematics, Chalmers University of Technology, June 1998. http://www.math.lth.se/matstat/staff/asmus/pspapers.html.

[13] A. Riska and E. Smirni. MAMSolver: a Matrix-analytic methods tools, In T. Field et al. (eds.), *TOOLS 2002, LNCS 2324*, pages 205–211, Springer-Verlag, 2002.

[14] A. Riska, M. Squillante, S. Yu, Z. Liu and L. Zhang. Matrix-analytic analysis of a MAP/PH/1 queue fitted to Web server data, In *Matrix-Analytic Methods; Theory and Applications*, G. Latouche and P. Taylor (eds.), pages 333–356, World Scientific, 2002.