Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds

QUANTILE ESTIMATION: A MINIMALIST APPROACH

Yury Bakshi

AT&T Laboratories
200 S Laurel Ave
Middletown, NJ 07751, U.S.A.

David A. Hoeflin

AT&T Laboratories
200 S Laurel Ave
Middletown, NJ 07751, U.S.A.

ABSTRACT

Managing telecommunication networks involves collecting and analyzing large amounts of statistical data. The standard approach to estimating quantiles involves capturing all the relevant data (what may require significant storage/processing capacities), and performing an off-line analysis (what may delay management actions). It is often essential to estimate quantiles as the data are collected, and to take management actions promptly. Towards this goal, we present a minimalist approach to sequentially estimating constant/changing over time quantiles. We follow prior work and devise a fixed-point algorithm, which does not estimate the unknown probability density function at the quantile. Instead, our algorithm uses the log-odds transformation of the observed fractions, and the exponentially smoothed estimates of the standard deviation to update the quantile estimate. For large data streams, this algorithm can significantly reduce the amount of collected data and the complexity of data analysis.

1 INTRODUCTION

Managing telecommunication networks involves collecting and analyzing large amounts of statistical data, e.g. call durations, transmitted/received bytes per minute, loads on a set of interfaces, etc. The standard approach to estimating quantiles involves capturing all the relevant data (that may require significant storage and processing capacity), and performing an off-line analysis (that may be too slow if the immediate action is required). It is often essential to estimate and track specific (and often changing over time) quantiles along with the associated mean and variance as the data are collected so that management action can be taken promptly. Frequently, the analyst tries to reasonably track, but not necessarily exactly estimate, the targeted quantile using the least amount of resources as possible. Towards this goal, we present a minimalist approach to sequentially estimating and tracking a constant or a changing over time quantile, typically defined as percentage target, e.g., 95^{th} or 99^{th} percentile.

Various algorithms for sequential quantile estimation were proposed in earlier papers. Most of them are intended for estimating static quantiles (i.e., quantiles that do not change over time) (Greenwald and Khanna 2001; Manku and Rajagopalan 1998; Lee, McNickle and Pawlikowski 1998). Several suggested algorithms estimate static quantiles together with static histogram from the collected data, e.g., (Chen 2002; Chen and Kelton 2001). The algorithm for tracking dynamic quantiles (i.e., quantiles that change over time) is described in (Chen, Lambert and Pinheiro 2000).

The algorithm we propose estimates static and tracks dynamic quantiles. The gain factor is defined differently for static and dynamic quantiles (refer to sections 3 and 4 for more details). In our approach, we follow prior work (Tierney 1983; Chen, Lambert and Pinheiro 2000), and devise a fixed-point algorithm, Log Odds Ratio Algorithm (LORA), that uses the observed fractions of measurements exceeding the most recent quantile estimate to adjust the estimate value. Unlike in prior work, we do not try to estimate the unknown probability density function at the quantile. Instead of Newton-Raphson approach, the proposed algorithm uses the log-odds transformation of the observed fractions, and the exponentially smoothed estimates of the standard deviation to update the quantile estimate. These data are generally available when monitoring stochastic processes. The algorithm stores only the most recent estimates of standard deviation and quantile. For large data streams, this can significantly reduce the amount of data collected and the complexity of data analysis and storage.

First, we examine the trivial case with no sampling and with known cumulative distribution. For this case, we define the sufficient conditions for LORA convergence to the true quantile value, when the initial quantile estimate is chosen sufficiently close to the true quantile value. We show that the convergence conditions are met for several common distributions. Second, we examine the case of a fixed but unknown distribution. Here we also address the standard approaches of estimating quantiles using the Tierney's Stochastic Approximation (TSA) and Moving Average (MA) algorithms. We compare how LORA, TSA and

MA algorithms estimate 95th percentile for the steady state (stationary) case with observation samples drawn from a fixed distribution. Third, we enhance LORA for tracking quantiles in non-stationary cases, when the underlying distribution is subject to location and scale changes. We compare the enhanced LORA and Exponentially Weighted Stochastic Algorithm, EWSA (Chen, Lambert and Pinheiro 2000) by tracking 95th percentiles for non-stationary Normal and Shifted Exponential distributions. Finally, we summarize the results and discuss future work.

2 CASE OF THE KNOWN DISTRIBUTION

The trivial case assumes that the cumulative distribution F and the probability density function *f* are known. Let *s* denote the standard deviation of *F*, and T_p denote the true pth quantile value, i.e. $F(T_p) = p$, and $q = 1- p$.

The proposed algorithm, Log Odds Ratio Algorithm (LORA), is defined as

$$\hat{T}_{new} = h(\hat{T}_{old}) = \hat{T}_{old} + s \times q \times \ln\left(\frac{(1-F(\hat{T}_{old}))}{F(\hat{T}_{old})} \times \frac{p}{q}\right).$$

We note several important properties of LORA:

1. $h(T_p) = T_p$, that is, T_p is a fixed point of h,
2. For $T > T_p$, $h(T) < T$, and for $T < T_p$, $h(T) > T$,
3. $\frac{\partial h(T)}{\partial T} = 1 - sq\left(\frac{f(T)}{(1-F(T))F(T)}\right)$,
4. At T_p, $\left|\frac{\partial h(T_p)}{\partial T}\right| < 1 \Leftrightarrow 0 < sf(T_p) < 2p$.

When condition 4 is satisfied, an interval *I* exists such that T_p belongs to *I*, and for any initial T_{init}, chosen within *I*, the algorithm convergences to the true quantile value T_p.

For several commonly used distributions, condition 4 is satisfied for the typically sought T_p, i.e., for larger *p*. For example:

1. In an Exponential distribution with mean μ, condition 4 is equivalent to requiring $0 < q < 2p$ that is satisfied for larger p regardless of the value of μ.
2. In a Normal distribution with mean μ and standard deviation σ, condition 4 is equivalent to $0 < \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{T_p-\mu}{\sigma}\right)^2} < 2p$. It is generally satisfied for larger p regardless of the values of μ and σ.
3. In a Weibull distribution with shape parameter *c* and scale parameter 1, condition 4 is equivalent to:

$$\left(\Gamma\left(\frac{2}{c}+1\right) - \Gamma\left(\frac{1}{c}+1\right)^2\right) c T_p^{\,c-1} e^{-T_p{}^c}.$$

When $c<0.3$, this expression is larger than 2 and convergence is not expected. When c>0.35, the expression is smaller than one and the proposed algorithm converges.

Clearly, any algorithm can converge slowly. It takes 66 iterations to achieve 1% relative error when estimating 95th percentile for the Exp(1) distribution estimating the 95th percentile if the starting value T_{init} is 2. Convergence can be hastened by adding a gain factor, i.e.,

$$\hat{T}_{new} = h(\hat{T}_{old}) = \hat{T}_{old} + s \times q \times \ln\left(\frac{(1-F(\hat{T}_{old}))}{F(\hat{T}_{old})} \times \frac{p}{q}\right) \times gain \cdot$$

If gain factor of 10 is used, the algorithm gets to 1% relative error after only six iterations. As we show later, the choice of the gain factor/function is of prime importance when tracking a stochastic process.

3 STEADY STATE CASE: SAMPLES FROM A FIXED UNKNOWN DISTRIBUTION

Let $x_{i1}, x_{i2} \ldots x_{iM}$ denote the i^{th} observation sample of size *M* from an unknown distribution *F*. Let n_{T_i} denote the number of observations in the i^{th} sample that are larger than a given threshold, T_i. Let $\hat{s}_i$ denote the most recent estimate of the standard deviation, obtained, for example, via a simple Exponential Weighting (EW), i.e.

$$\hat{s}_i = \omega \times \hat{s}_{i-1} + (1-\omega) \times s_i,$$

where $\hat{s}_{i-1}$ is a previous EW estimate of the standard deviation and s_i is a standard deviation of the i^{th} observation sample. We assume that $\omega = 0.95$ for the reminder of the paper.

LORA: We construct the sample estimates for known *p* and *q* (where $p = F(T_p)$ and $q = 1-p$) as $\hat{p}_i = (M - n_{T_i} + 0.5)/(M+1)$ and $\hat{q}_i = 1 - \hat{p}_i$. The update procedure for the quantile estimate is

$$\hat{T}_i = \hat{T}_{i-1} + \hat{s}_i \times q \times \ln\left(\frac{\hat{q}_i}{\hat{p}_i} \times \frac{p}{q}\right) \times gain.$$

For larger (smaller) gains, the algorithm will converge quicker (slower), but the variation in the steady state will be larger (smaller). In a fixed steady state case, it is desirable to reduce the contribution of later observations. This can be easily accomplished by defining the gain as a de-

creasing positive function of the sample index i, for example, as $1/\sqrt{i}$. Then, the quantile update procedure becomes

$$\hat{T}_i = \hat{T}_{i-1} + \hat{s}_i \times q \times \ln\left(\frac{\hat{q}_i}{\hat{p}_i} \times \frac{p}{q}\right) \times \frac{1}{\sqrt{i}}.$$

The gain factor is defined differently for non-stationary case, which is discussed later. Next, we describe two common algorithms of estimating quantiles and, then, compare these algorithms with LORA.

MA: One common technique uses the Moving Average (MA) to estimate the p[th] quantile (unadjusted to account for bias). The update procedure is given by

$$\hat{T}_i = \hat{T}_{i-1} \times \left(1 - \frac{1}{i}\right) + \frac{t_i}{i},$$

where t_i is a p^{th} quantile of the i^{th} observation sample. In this method, the top *[Mq]+1*, where *q=1-p*, elements in the i[th] observation sample must be determined to calculate t_i . Depending on the sample size and value of *q*, this could require significant storage and processing resources.

TSA: Another common technique uses the modified Newton-Raphson algorithm with decreasing weights assigned to later observations. This algorithm is called Tierney's Stochastic Approximation algorithm (TSA) (Tierney 1983). The updating procedure is

$$\hat{T}_i = \hat{T}_{i-1} + \frac{(p - \tilde{p}_i)}{\max\left(\hat{f}_{i-1}, \frac{\hat{f}_0}{\sqrt{i-1}}\right)} \times \frac{1}{i},$$

where $\tilde{p}_i = \frac{n_{T_i}}{M}$, $\hat{f}_0$ is an initial estimate, and $\hat{f}_{i-1}$ is a prior estimate of the density function at quantile, i.e., of $f(T_p)$. The estimate $\hat{f}_i$ is updated via

$$\hat{f}_i = \left(1 - \frac{1}{i}\right) \times \hat{f}_{i-1} + \frac{m_{T_i}}{2 \times M} \times \frac{1}{\sqrt{i}},$$

where m_{T_i} is the number of the observations in the i^{th} sample that are located within the distance of $1/\sqrt{i}$ or less from T_{i-1}.

We used three described algorithms, MA, TSA and LORA, to estimate the 95[th] percentile of the normal distribution with mean of 50 and standard deviation of 12.5. Typical results are shown below in Figures 1 and 2. The initial estimate of the quantile, $\hat{T}_0$, is set to the sum of the true quantile value, T_p, and normally distributed variable with zero mean and unit variance. The initial value of density function, $\hat{f}_0$, is set to it's true value of $f(T_p)$. The sample size *M* is set to 120 (first figure), and to 40 (second figure). Black straight lines show location of the true quantile value T_p. Grey straight lines show locations of $T_p \pm 1$. LORA estimates are shown in red, TSA estimates are shown in blue, and MA estimates are shown in purple.

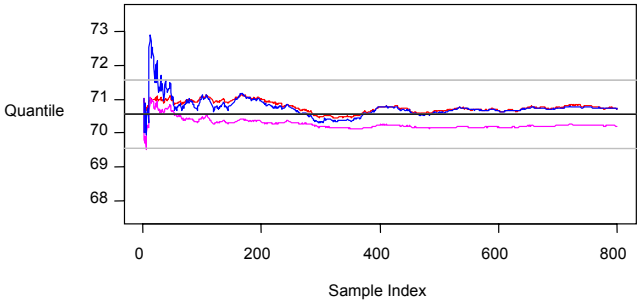

Figure 1: *N(50,12.5), M=120, To~N(Tp,1), p=0.95*

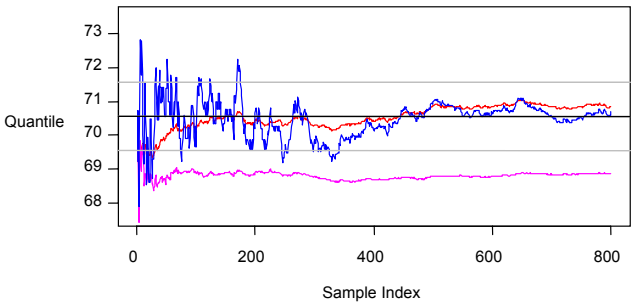

Figure 2: *N(50,12.5), M=40, To~N(Tp,1), p=0.95*

We observed the following for the chosen sample sizes (M=40 and M=120):

- LORA and TSA have significantly smaller estimating bias than MA.
- LORA and MA settle down faster than TSA.
- Visibly, LORA performs no worse (or maybe even better) than either MA or TSA

4 NON-STATIONARY CASE: SAMPLES FROM AN UNKNOWN CHANGING DISTRIBUTION

To track changes in the underlying distributions, the weights for new observations should not shrink to zero. While there are many possibilities for such weight adjustment, we chose to use exponential smoothing.

The enhanced update procedure for LORA is

$$\hat{T}_i = \hat{T}_{i-1} + \hat{s}_i \times q \times \ln\left(\frac{\hat{q}_i}{\hat{p}_i} \times \frac{p}{q}\right) \times gain \times (1-\beta),$$

$$\hat{p}_i = (M - n_{T_i} + 0.5)/(M+1),\ \hat{q}_i = 1 - \hat{p}_i,$$

$$\hat{s}_i = \omega \times \hat{s}_{i-1} + (1-\omega) \times s_i$$

where n_{T_i} = number of observations x_{ij} in the i[th] sample $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{iM})$ such that $x_{ij} > T_{i-1}$; β and ω are the smoothing constants. We observed that gain=10, β = 0.95 and ω=0.95 yields good results. The initial estimate of the quantile, $\hat{T}_0$, is set to the sum of the true quantile value, T_p, and normally distributed variable with zero mean and unit variance, i.e.,

$$T_0 = T_p + N(0,1),\ s_0 = stdev(x_{01}, x_{02}, ..., x_{0M}).$$

EWSA: We compared LORA to the modified Newton-Raphson algorithm (NR) that uses fixed weight for smoothing and a fixed interval length for estimating the probability density function at the quantile. This algorithm, Exponentially Weighted Stochastic Approximation (EWSA), was developed by Chen, Lambert and Pinheiro (2000); we used an estimate of standard deviation instead of an inter-quantile range recommended by the authors. Our numerical simulations indicate that using standard deviation leads to as good or better quantile tracking in cases where data contamination is not a concern. The update procedure is

$$\hat{T}_i = \hat{T}_{i-1} + \frac{(p - \tilde{p}_i)}{\hat{f}_{i-1}} \times (1-\alpha),$$

$$\tilde{p}_i = 1 - \frac{n_{T_i}}{M},$$

$$\hat{f}_i = \alpha \times \hat{f}_{i-1} + \frac{m_{T_i}}{2 \times M \times c_{i-1}} \times (1-\alpha),$$

$$c_i = \frac{\hat{s}_i}{M} \times \sum_{i=M+1}^{2\times M} \frac{1}{\sqrt{i}} \text{ for } i \geq 1,$$

$$\hat{s}_i = \omega \times \hat{s}_{i-1} + (1-\omega) \times s_i,$$

where n_{T_i} is the number of observations x_{ij} in the i^{th} sample $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{iM})$, such that $x_{ij} > T_{i-1}$. m_{T_i} is the number of observations x_{ij} in the i^{th} sample $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{iM})$, such that $\left|x_{ij} - \hat{T}_{i-1}\right| \leq c_{i-1}$, with initial values set as:

$$T_0 = T_p + N(0,1),$$

$$c_0 = \frac{s_0}{M} \times \sum_{i=1}^{M} \frac{1}{\sqrt{i}},$$

$$f_0 = \frac{1}{2 \times c_0 \times M} \times m_0,$$

$$s_0 = stdev(x_{01}, x_{02}, ..., x_{0M}),$$

m_0 = max(1, number of x_{0j} in the initial observation sample $\vec{x}_0 = (x_{01}, x_{02}, ..., x_{0M})$ such that $\left|x_{0j} - \hat{T}_0\right| \leq c_0$), where α, ω are the smoothing constants.

5 NUMERICAL SIMULATION STUDY

We compared how LORA and EWSA track quantiles for the steady-state and non-stationary cases. The set of changes that we often encounter in field are jumps in the location, scale or both. While the results depend on the choices for smoothing constants α, β and ω, we permanently fixed these to α = 0.95, β = 0.95, and ω=0.95 (we observed that α=0.95, β = 0.95 and ω=0.95 usually yields good results). We implemented update procedures and chose initial values for LORA and EWSA as described in the previous sections. In all simulations, 95[th] percentile was estimated.

5.1 Steady-State Simulation Results

The measured distribution did not change during the steady-state simulation runs. We worked with four exponential distributions with different combinations of two means of 30 and 40, and two shifts of 20 and 30. We also worked with four normal distributions with different combinations of two means of 30 and 40, and two standard deviations of 7.5 and 15. In each simulation we chose the sample size of 40, 80 or 120, and performed 100 simulation runs (a simulation case) with 200 observation samples in each run. After each simulation run we calculated RMSE, and after all 100 runs, in each simulation case, we calculated an average of 100 RMSEs that are presented below.As the results indicate the LORA's performance is quite similar to the performance of the somewhat more complex EWSA, and can serve as the EWSA alternative in the cases when algorithm simplicity is essential. The simu-

lation results are presented below in Tables 1 and 2, and in Figures 3 and 4.

Table 1: Exp. Distribution-SteadyState

Experiment: Exponential Steady-State			
Case	Mean	Shift	Sample Size
1	30	20	40
2	30	20	80
3	30	20	120
4	40	20	40
5	40	20	80
6	40	20	120
7	30	30	40
8	30	30	80
9	30	30	120
10	40	30	40
11	40	30	80
12	40	30	120

Table 2: Normal Distribution - Steady State

Experiment: Normal Steady-State			
Case	Mean	Sigma	Sample Size
1	30	7.5	40
2	30	7.5	80
3	30	7.5	120
4	30	15	40
5	30	15	80
6	30	15	120
7	40	7.5	40
8	40	7.5	80
9	40	7.5	120
10	40	15	40
11	40	15	80
12	40	15	120

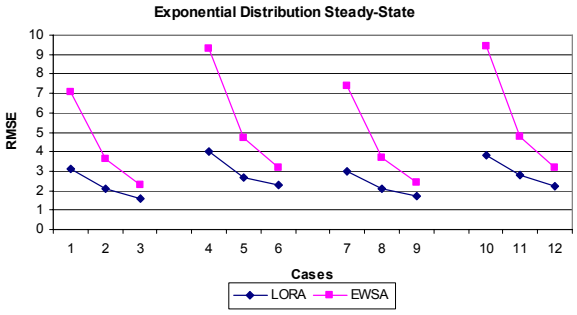

Figure 3: Exponential Distribution - Steady State

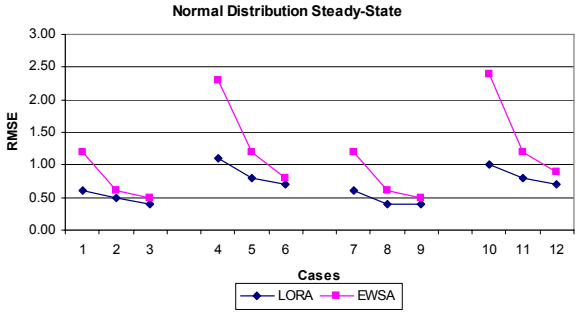

Figure 4: Normal Distribution - Steady State

5.2 Non-Stationary Simulation Results

During non-stationary simulation runs, the distribution changed between the 100 and 101 observation samples. We analyzed the same sets of exponential and normal distributions and used the same sample sizes as in steady-state cases. The distribution changes that we modeled included increases in means and/or shifts for exponential distributions, and increases of means and/or standard deviations for normal distributions. The simulation results are presented below in Tables 3 and 4, and in Figures 5 and 6. Once again, the results indicate that LORA is a competitive alternative to EWSA.

Table 3: Non-Stationary Exponential Distribution

Experiment: Exponential Cases			
Case	Change in Mean	Change in Shift	Sample Size
1	10	0	40
2	10	0	80
3	10	0	120
4	0	10	40
5	0	10	80
6	0	10	120
7	10	10	40
8	10	10	80
9	10	10	120

Table 4: Non-Stationary Normal Distribution

Experiment: Normal Cases			
Case	Change in Mean	Change in Sigma	Sample Size
1	10	0	40
2	10	0	80
3	10	0	120
4	0	7.5	40
5	0	7.5	80
6	0	7.5	120
7	10	7.5	40
8	10	7.5	80
9	10	7.5	120

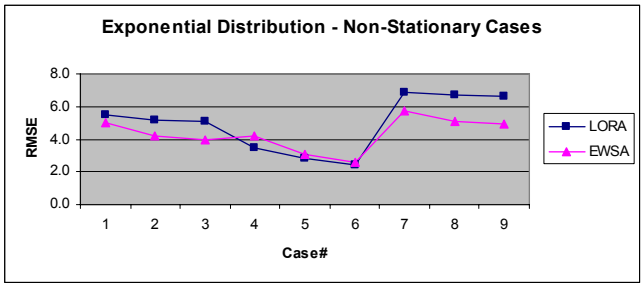

Figure 5: Non-Stationary Exponential Distribution

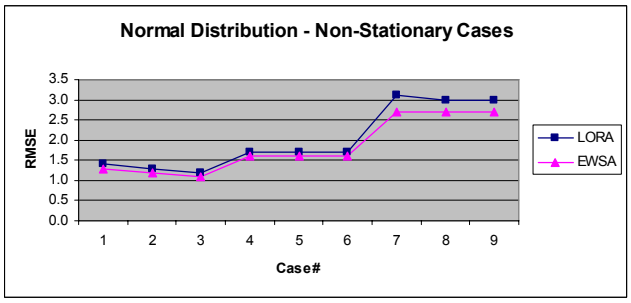

Figure 6: Non-Stationary Normal Distribution

6 CONCLUSION AND FUTURE WORK

We introduced a simple fixed-point algorithm, LORA, for estimating quantiles of unknown fixed (steady state) and changing (non-stationary) distributions. Based on completed numerical simulation study, we concluded that, in spite of very modest memory requirements and a simple sequential update procedure, LORA is quite competitive to the algorithms commonly used for estimating a quantile of a stochastic process: TSA and MA for processes with fixed distributions, and EWSA for processes with changing distributions.

We plan to enhance LORA to be able to use a set of inter-dependent observations for estimating a process quantile. While evaluating 95th percentile of queue length for simulated M/M/1 queue, we observed that LORA performs well at the lower server utilization (when the inter-dependency of queue length observations is weak), but for server utilization exceeding 80% (and highly inter-dependent queue length observations) the convergence to the true quantile value is rather problematic. Another planned enhancement is adjusting weights assigned to new observations based on perceived distance between estimated and true quantile values.

We also plan to quantify LORA's performance while tracking quantiles of non-stationary distributions that change frequently (e.g., every 20-50 samples) and significantly (e.g., from exponential distribution to normal one). Another area of interest is tuning algorithm parameters (e.g., sample size, gain factor, smoothing constants) for improved performance.

REFERENCES

Chen, F., D. Lambert, and J. C. Pinheiro. 2000. Incremental Quantile Estimation for Massive Tracking. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 516-522.

Chen, J.E., and D. W. Kelton. 2001. Quantile and Histogram Estimation. *Proceedings of the Winter Simulation Conference*, 451-459. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Chen, J.E. 2002. Two-Phase Quantile Estimation. *Proceedings of the 2002 Winter Simulation Conference*, 447-455. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Greenwald, M., and S. Khanna, S. 2001. Space-Efficient Online Computation of Quantile Summaries. *Proceedings of the ACM SIGMOD Conference*, 47-57.

Lee, J.R., D. McNickle, and K. Pawlikowski. 1998. Sequential Estimation of Quantiles. *Technical Report TR-COSC 05/98*. Department of Computer Science, University of Canterbury, Christchurch, New Zealand.

Manku, G.S., S. Rajagopalan, and B. G. Lindsay. 1998. Approximate Medians and other Quantiles in One Pass and with Limited Memory. In *Proceedings of the ACM International Conference on Management of Data*, 426-435.

Tierney, L. 1983. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*. Vol. 4, Issue 4, 706-711.

AUTHOR BIOGRAPHIES

YURY BAKSHI is a member of the staff at AT&T Laboratories. His current research focuses on VoIP network modeling, and performance tuning of UNIX/LINUX systems. Yury holds 9 U.S. and foreign patents in the areas of telecommunications and chromatography. He has advanced degrees in Electrical Engineering and Systems Engineering.

DAVID A. HOEFLIN After receiving a Ph.D. in Mathematics from Iowa State University in 1984, David joined AT&T Bell Labs to do performance and reliability analysis and remained in AT&T Labs to do more of the same. He is now a Technical Manager in the Network Design and Performance Analysis department of AT&T Labs.