ORSA Journal on Computing
Vol. 7, No. 1, Winter 1995

0899-1499/95/0701-0089 $01.25
© 1995 INFORMS

Efficient Computation of Probabilities of Events Described by Order Statistics and Applications to Queue Inference

LEE K. JONES / *Institute for Visualization and Perception Research and Department of Mathematical Sciences, University of Massachusetts-Lowell, Lowell, Massachusetts 01854; Email: jones1@uml.edu*

RICHARD C. LARSON / *Operations Research Center and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; Email: rclarson@mit.edu*

(Received: March 1992; revised: February 1993, March 1994; accepted: May 1994)

This paper derives recursive algorithms for efficiently computing event probabilities related to order statistics and applies the results in a queue inferencing setting. Consider a set of N i.i.d. random variables in [0, 1]. When the experimental values of the random variables are arranged in ascending order from smallest to largest, one has the order statistics of the set of random variables. Both a forward and a backward recursive $O(N^3)$ algorithm are developed for computing the probability that the order statistics vector lies in a given N-rectangle. The new algorithms have applicability in inferring the statistical behavior of Poisson arrival queues, given only the start and stop times of service of all N customers served in a period of continuous congestion. The queue inference results extend the theory of the "Queue Inference Engine" (QIE), originally developed by Larson in 1990.[8] The methodology is extended to a third $O(N^3)$ algorithm, employing both forward and backward recursion, that computes the conditional average probability (averaged over all N customers in a congestion period) that the in-queue wait is less than t minutes, given the departure time data and assuming first come, first served service. To our knowledge, this result is the first $O(N^3)$ exact algorithm for computing points on the in-queue waiting time distribution function, conditioned on the start and stop time data. The paper concludes with an extension to the computation of certain correlations of in-queue waiting times. Illustrative computational results are included throughout.

Order statistics have applicability in many areas of science. If N students take a test and if each test score is modeled as an independent sample from a common "test score" distribution, then the minimum of the N scores has a distribution corresponding to the first order statistic, the maximum has a distribution corresponding to the largest order statistic, and in general the jth score from the lowest has a distribution of the jth order statistic of the set of N independent, identically distributed (*i.i.d.*) random variables (*r.v.*'s). Or, consider a homogeneous Poisson arrival process for which it is observed that N customers arrive to a bulk service queue over the pre-set interval $[0, T]$, where T is the time of bulk service. ("Bulk service" implies that all N waiting customers are served simultaneously, as one observes with elevators, buses, and pedestrian traffic crossing lights, for example.) Then the N unordered customer arrival times are independent, uniformly distributed over $[0, T]$ and the in-queue waiting time of the jth arriving customer in $[0, T]$ corresponds to $T - X_{(j)}$, where $X_{(j)}$ is the jth order statistic of the set of N uniform independent r.v.'s over $[0, T]$.

Throughout this paper we will focus the application of our results on queueing computations, although the recursive algorithms developed in §1 are of independent interest in any order statistics setting. The substance of the paper appears to lie at an intersection of the fields of statistics, operations research (particularly queueing) and computer science (particularly the design of algorithms).

Let $X_1, X_2, \ldots, X_{N(1)}$ be an i.i.d. sequence of random variables with values in [0, 1] where the sequence length $N(1)$ is an independent random integer. It is most natural to ask for a computationally efficient algorithm to calculate the probability of an order statistics vector lying in a given N-rectangle, i.e., to compute

$$\Gamma(\underline{s}, \underline{t}) \equiv Pr\{s_1 < X_{(1)} \leqslant t_1, s_2 < X_{(2)} \leqslant t_2, \ldots, s_N < X_{(N)} \leqslant t_N | N(1) = N\}, \qquad (1)$$

where $\underline{s} \equiv (s_1, s_2, \ldots, s_N)$, $\underline{t} \equiv (t_1, t_2, \ldots, t_N)$ and without loss of generality the sequences $\{s_i\}$ and $\{t_i\}$ are increasing. Apparently the question of efficient computation of event probabilities for the order statistics vector has not been previously treated in the literature. (See for example [1, 5].) Recently, in an application to queue inference,[8] an $O(N^3)$ algorithm was developed to compute the conditional cumulative probability of the vector of order statistics, $Pr\{X_{(1)} \leqslant t_1, X_{(2)} \leqslant t_2, \ldots, X_{(N)} \leqslant t_N | N(1) = N\}$, for the case of each X_i uniform. This is a special case of our computation, with $\{s_i\} = \{0\}$ and with uniform r.v.'s. We note that the method of computing $\Gamma(\underline{s}, \underline{t})$ by applying repeated differences to the cumulative distribution will require 2^N evaluations of the cumulative[8]; this is too slow for many applications. The algorithm presented here will efficiently

Subject classifications: Order statistics, queues, inference, computational probability.

calculate Eq. (1) for X_i having arbitrary cumulative distribution function (*c.d.f.*) $F(x)$. More generally, in this paper we develop two $O(N^3)$ algorithms to compute $\Gamma(\underset{\sim}{s}, \underset{\sim}{t})$ where the X_i have a given c.d.f. $F(x)$. New applications are shown for deducing queue statistics from transactional data.

We extend the logic of the two general algorithms to develop a special "*forward/backward 'kiss' (FBK)*" $O(N^3)$ algorithm that computes the conditional probability that a queued customer waited less than τ minutes in the *FCFS* (*F*irst *C*ome, *F*irst *S*erved) queue, given the observed departure times. To our knowledge, this is the first exact $O(N^3)$ algorithm for inferring from service time data the distribution of queue wait in a *FCFS* queueing system.

The paper concludes with an extension to the computation of certain correlations of in-queue waiting times. The method utilizes an $O(N^5)$ nested recursion to find the correlation coefficients of 0-1 indicator random variables that are equal to one only if the corresponding customers waited in queue less than τ time units, given the departure time data. This methodology may have applications in assessing queue inference approximations for very congested systems in which exact solutions require a prohibitively long time; it may also be applicable in estimating the equivalent number n of independent samples of queue delay contained in a data set having N customers, $n \leqslant N$.

1. The Primary Recursive Algorithms

In this section we derive the $O(N^3)$ algorithms for finding $\Gamma(\underset{\sim}{s}, \underset{\sim}{t})$. Assume $X_i \in (0,1]$ and $0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_N \leqslant 1, 0 \leqslant s_1 \leqslant s_2 \leqslant \cdots \leqslant s_N \leqslant 1$ and $s_i \leqslant t_i$ for $i = 1,2,\ldots,N$. Since $\{t_i\}$ and $\{s_i\}$ are each nondecreasing as sequences, we may merge the two sequences into $\{v_i\}_{i=1}^{2N}$, ordered according to magnitude, using only $O(N)$ operations. We now define a conditional probability indexed on k, the number of i.i.d. r.v.'s whose order statistics we are computing, and on i, which through v_i yields an upper bound to the allowed value of any of the order statistics. In the developed recursion k is progressed from 1 to N and i is increased from 1 to $2N$. More precisely, define

$$W_{ki} \equiv Pr\{s_1 < X_{(1)} \leqslant \min\{t_1, v_i\}, s_2 < X_{(2)} \leqslant \min\{t_2, v_i\}, \ldots, s_k < X_{(k)} \leqslant \min\{t_k, v_i\} | N(1) = k\},$$
$$k = 1,2,\ldots,N, i = 1,2,\ldots,2N. \quad (2)$$

For example, suppose $X_{(j)}$ is the time of the jth customer arrival to a queueing system; then W_{ki} is the conditional probability that each arrival j arrives *after* his earliest allowable arrival time s_j and *before* his entry-into-service time t_j or v_i, whichever is smaller, given exactly k arrivals in [0, 1]. If the last customer's earliest allowable arrival time s_k equals or exceeds v_i, then we have an impossible event and the conditional probability becomes $W_{ki} = 0$. Setting k and i equal to their respective maximum values, we easily verify that $W_{N,2N}$ is the desired probability given in (1), i.e., $W_{N,2N} = \Gamma(\underset{\sim}{s}, \underset{\sim}{t})$. Evaluating $W_{N,2N}$ requires recursive computation of entries of the matrix $\mathbf{W} = (W_{ki})$, starting with $k = 1$. Our recursive procedure requires as a boundary condition

$$W_{0i} = 1 \quad \text{for } i = 1,2,\ldots,2N, \quad (3)$$

which can be interpreted to be the probability that the event inequalities will be satisfied, given no random variables (hence no inequalities) in [0, 1].

We now give the *forward recursive algorithm* as

Theorem 1. *For $k = 1,2,\ldots,N$, $i = 1,2,\ldots,2N$, W_{ki} can be computed using the following recursion:*

$$W_{ki} = \left\{ \begin{array}{l} W_{k,i-1} + \\ \displaystyle\sum_{\substack{j=1 \\ \text{s.t. } v_i \leqslant t_{k-j+1}}}^{k} \binom{k}{j} W_{k-j,i-1}[F(v_i) - F(v_{i-1})]^j \\ \qquad \text{if } s_k < v_i \\ 0 \qquad \text{if } s_k \geqslant v_i \end{array} \right\} \quad (4)$$

Note. Eq. (4) essentially states that the conditional probability W_{ki} can be written as the sum of conditional probabilities of up to $k + 1$ disjoint events, the jth event corresponding to the first $k - j$ order statistics appropriately distributed over the inequality-constrained interval $(0, v_{i-1}]$ and the remaining j in the simple interval $(v_{i-1}, v_i]$, $j = 0,1,\ldots,k$. The *summation index constraint* $v_i \leqslant t_{k-j+1}$ assures that the summation index only produces events having the $k - j + 1$st order statistic less than or equal to its maximum allowed value, t_{k-j+1}. In queueing parlance, the summation index constraint assures that the earliest arriving customer in $(v_{i-1}, v_i]$, customer $k - j + 1$, arrives not later than her time of entry into service, t_{k-j+1}. The *left limit test condition* $s_k \geqslant v_i$ creates an impossible event, in queueing parlance requiring a customer to arrive before her earliest allowable arrival time.

Proof. The case $s_k \geqslant v_i$ yields an impossible event, implying $W_{ki} = 0$. Now consider $s_k < v_i$. We can write

$$\begin{aligned} W_{ki} &\equiv Pr\{s_1 < X_{(1)} \leqslant \min\{t_1, v_i\}, s_2 < X_{(2)} \leqslant \min\{t_2, v_i\}, \ldots, \\ &\quad \times s_k < X_{(k)} \leqslant \min\{t_k, v_i\} | N(1) = k\} \\ &= Pr\{s_1 < X_{(1)} \leqslant \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \\ &\quad \leqslant \min\{t_2, v_{i-1}\}, \ldots, s_k < X_{(k)} \\ &\quad \leqslant \min\{t_k, v_{i-1}\} | N(1) = k\} \\ &\quad + Pr\{s_1 < X_{(1)} \leqslant \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \\ &\quad \leqslant \min\{t_2, v_{i-1}\}, \ldots, s_{k-1} < X_{(k-1)} \\ &\quad \leqslant \min\{t_{k-1}, v_{i-1}\}, v_{i-1} < X_{(k)} \leqslant \min\{t_k, v_i\} | N(1) = k\} \\ &\quad + Pr\{s_1 < X_{(1)} \leqslant \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \\ &\quad \leqslant \min\{t_2, v_{i-1}\}, \ldots, s_{k-2} < X_{(k-2)} \\ &\quad \leqslant \min\{t_{k-2}, v_{i-1}\}, v_{i-1} < X_{(k-1)} \leqslant \min\{t_{k-1}, v_i\}, \\ &\qquad v_{i-1} < X_{(k)} \leqslant \min\{t_k, v_i\} | N(1) = k\} \\ &\quad + \cdots + Pr\{s_1 < X_{(1)} \leqslant \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \\ &\quad \leqslant \min\{t_2, v_{i-1}\}, \ldots, s_{k-j} < X_{(k-j)} \\ &\quad \leqslant \min\{t_{k-j}, v_{i-1}\}, v_{i-1} < X_{(k-j+1)} \\ &\quad \leqslant \min\{t_{k-j+1}, v_i\}, \ldots, \\ &\quad v_{i-1} < X_{(k)} \leqslant \min\{t_k, v_i\} | N(1) = k\} + \cdots \end{aligned}$$

The term explicitly displaying $X_{(k-j+1)}$ on the RHS can be nonzero only if $\min\{t_{k-j+1}, v_i\} = v_i$, implying that all order statistics constrained to be greater than v_{i-1} must lie in the simple interval $(v_{i-1}, v_i]$. By counting the number of ways that the original (unordered) r.v.'s can fall either into the simple interval $(v_{i-1}, v_i]$ or into the inequality-constrained interval $(0, v_{i-1}]$, and recognizing the appropriate probabilities that apply in each interval, we can write

$$\begin{aligned} W_{ki} = W_{k,i-1} &+ \binom{k}{1} W_{k-1,i-1}(F(v_i) - F(v_{i-1})) \\ &+ \binom{k}{2} W_{k-2,i-1}(F(v_i) - F(v_{i-1}))^2 \\ &+ \cdots + \binom{k}{j} W_{k-j,i-1}(F(v_i) - F(v_{i-1}))^j + \cdots \end{aligned}$$

where we include all j satisfying $v_i \leqslant t_{k-j+1}$ and $j \leqslant k$. ■

As a verification of the recursion we obtain as expected at the first iteration

$$W_{1i} = F(\min\{t_1, v_i\}) - F(s_1) \qquad i = 1, 2, \ldots, 2N.$$

The matrix $\mathbf{W} = (W_{ki})$ will have the maximum possible number of nonzero entries if the following strict inequalities hold: $0 < t_1 < t_2 < \cdots < t_N < 1$; $0 < s_1 < s_2 < \cdots < s_N < 1$; $s_i < t_i$; $F(x) - F(y) > 0$ for $x > y$, $x, y \in (0, 1]$. Then $\mathbf{W} = (W_{ki})$ can be partitioned into three regions:

(1) Always impossible events: $W_{ki} = 0$ for $k \geqslant i$;
(2) Sometimes impossible events: $W_{ki} \geqslant 0$ for $i - N < k < i$;
(3) Always possible events: $W_{ki} > 0$ for $k \leqslant i - N$;

Hence the maximum possible number of nonzero terms in row k is $2N - k$, and the minimum number is $N - k + 1$. The recursion to obtain W_{ki} requires computation and addition of up to $k + 1$ terms. Thus, row k of (W_{ki}) requires computation of up to $(2N - k)(k + 1)$ terms. The total number of terms required to compute (W_{ki}) is

$$\sum_{k=1}^{N} (2N - k)(k + 1) = \tfrac{2}{3}N^3 + 2N^2 - \tfrac{2}{3}N,$$

yielding an $O(N^3)$ procedure. For the special case $s_i = 0$, $i = 1, 2, \ldots, N$, we have the problem of [2, 3, 8], and all W_{ki} in columns $i = 1$ through N are equal to 0 and all W_{ki} in columns $i = N + 1$ through $2N$ are nonzero.

There is a comparable backward recursive algorithm that is used to compute

$$\begin{aligned} Y_{ki} \equiv Pr\{&\max\{s_{N-k+1}, v_i\} < X_{(1)} \leqslant t_{N-k+1}, \\ &\max\{s_{N-k+2}, v_i\} < X_{(2)} \leqslant t_{N-k+2}, \ldots, \\ &\max\{s_N, v_i\} < X_{(k)} \leqslant t_N \mid N(1) = k\}, \\ &i = 1, 2, \ldots, 2N,\ k = 1, 2, \ldots, N \end{aligned} \tag{5}$$

having boundary conditions $Y_{0i} = 1$ for $i = 1, 2, \ldots, 2N$. In a queueing context Y_{ki} is the conditional probability that each arrival j arrives after her earliest allowable arrival time s_{N-k+j} or v_i, whichever is larger, and before her entry-into-service time t_{N-k+j}, given exactly k arrivals in $[0, 1]$. If the first customer's entry-into-service time t_{N-k+1} is less than or equal to v_i, i.e., if $t_{N-k+1} \leqslant v_i$, then we have an impossible event and the conditional probability is $Y_{ki} = 0$. Here we want to compute $Y_{N,1} \equiv \Gamma(\underline{s}, \underline{t})$. This will require recursive computation of entries of the matrix $\mathbf{Y} = (Y_{ki})$, starting with $k = 1$. We now give the *backward recursive algorithm* as

Theorem 2. *For $k = 1, 2, \ldots, N$, $i = 2N, 2N - 1, \ldots, 1$, Y_{ki} can be computed using the following recursion:*

$$Y_{ki} = \left\{ \begin{array}{ll} Y_{k,i+1} + \displaystyle\sum_{\substack{j=1 \\ \text{s t. } v_i \geqslant s_{N-k+j}}}^{k} \binom{k}{j} Y_{k-j,i+1}[F(v_{i+1}) - F(v_i)]^j & \\ & \text{if } v_i < t_{N-k+1} \\ 0 & \text{if } v_i \geqslant t_{N-k+1} \end{array} \right\} \tag{6}$$

Proof. Omitted (similar to Theorem 1) ■

Intuitively, the forward algorithm processes disjoint events in the interval [0, 1] from left to right and the backward algorithm employs similar logic to process disjoint events in [0, 1] from right to left. The backward algorithm is also $O(N^3)$.

Example 1. The construction of the set $\{v_i\}$ from the sets $\{t_i\}$ and $\{s_i\}$ is shown in Figure 1 for an example with $N = 4$. In using Eq. (4) the first task is to decide which entries of the matrix $\underline{\mathbf{W}}$ are zero, i.e., probabilities equaling zero due to impossible events. This determination is made from testing the inequality $s_k \geqslant v_i$, as shown in Exhibit 1. In the matrix, a "Y" ("N") implies that the test yields "yes" ("no") and thus the corresponding entry is 0 (positive). Note that all entries in the "southwest" corner of the matrix, corresponding to $k \geqslant i$, are always "Y" regardless of the prob-

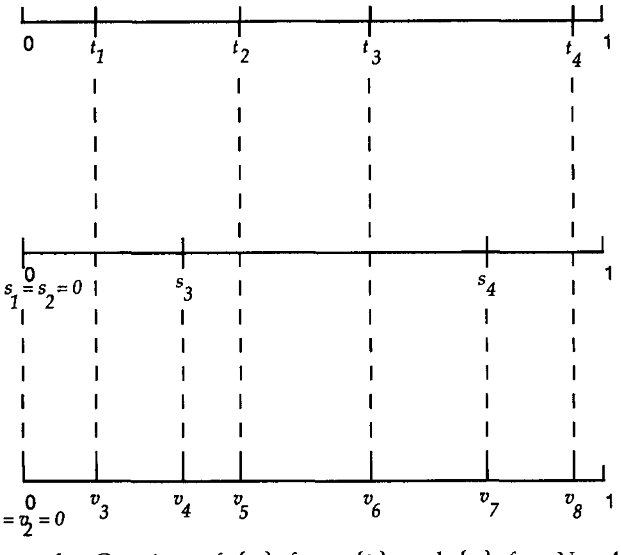

Figure 1. Creation of $\{v_i\}$ from $\{t_i\}$ and $\{s_i\}$ for $N = 4$ example.

lem data. (A dash ("-") as a matrix entry implies that the Y/N categorization is not relevant.)

To use Eq. (4) to compute a positive entry of the matrix, one must assure that the summation index constraint, $v_i \leqslant t_{k-j+1}$, is obeyed. This constraint guarantees that the number of r.v.'s allowed in the simple interval $(v_{i-1}, v_i]$ does not exceed the maximum permitted, given the inequalities in the inequality-constrained interval $(0, v_{i-1}]$. Defining m to be the maximum allowed value for the summation index j for a particular matrix entry, the respective values of m are shown in Exhibit 2.

Example 2. Now consider a simple numerical example with $N = 3$, $\underline{t} = (1/3, 2/3, 1)$, $\underline{s} = (1/6, 1/2, 5/6)$ and hence $\underline{v} = (1/6, 1/3, 1/2, 2/3, 5/6, 1)$. For simplicity assume uniform r.v.'s, implying $F(x) = x,\ 0 \leqslant x \leqslant 1$. The desired probability, $W_{3,6} = \Gamma(\underline{s}, \underline{t})$, corresponds to the event defined by exactly one of the r.v.'s assuming values in each of the intervals $(1/6, 1/3]$, $(1/2, 2/3]$ and $(5/6, 1]$, respectively. The intervals are disjoint, so by elementary arguments, $\Gamma(\underline{s}, \underline{t}) = 1/36$. Using the algorithm of Theorem 1, we find the matrix

$$\underline{\mathbf{W}} = \begin{array}{c|cccccc} \mathbf{k}\backslash & i=1 & 2 & 3 & 4 & 5 & 6 \\ \hline 0 & 1 & 1 & 1 & 1 & 1 & — \\ 1 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & — \\ 2 & 0 & 0 & 0 & 1/18 & 1/18 & — \\ 3 & 0 & 0 & 0 & 0 & 0 & 1/36 \end{array},$$

where $W_{3,6} = 1/36$, as expected. We illustrate the computation of

$$W_{2,4} = W_{2,3} + \binom{2}{1} W_{1,3}[2/3 - 1/2]^1 = 1/18;$$

the potential second term in the summation on the RHS, corresponding to $j = 2$, is not included due to the summation inequality constraint.

Now modify the example so that $\underline{t} = (1/3, 2/3, 1)$, $\underline{s} = (0, 0, 0)$ and hence $\underline{v} = (0, 0, 0, 1/3, 2/3, 1)$. The desired quantity $W_{3,6}$ is not so readily found by elementary arguments. By applying the algorithm of Theorem 1 we find

$$\underline{\mathbf{W}} = \begin{array}{c|cccccc} \mathbf{k}\backslash & i=1 & 2 & 3 & 4 & 5 & 6 \\ \hline 0 & 1 & 1 & 1 & 1 & 1 & — \\ 1 & 0 & 0 & 0 & 1/3 & 1/3 & — \\ 2 & 0 & 0 & 0 & 1/9 & 1/3 & — \\ 3 & 0 & 0 & 0 & 1/27 & 7/27 & 16/27 \end{array},$$

implying that the desired probability is $W_{3,6} = 16/27$.

Example 3. The Multinomial Distribution. The algorithm of Theorem 1 can be applied to the derivation of several well known probability laws. For example consider $F(x)$ and a vector $\underline{\tau} = \{\tau_i\}$ as shown in Figure 2, where we assume $\tau_{i+1} > \tau_i$, $i = 1, 2, 3, \ldots$. Suppose we are interested in

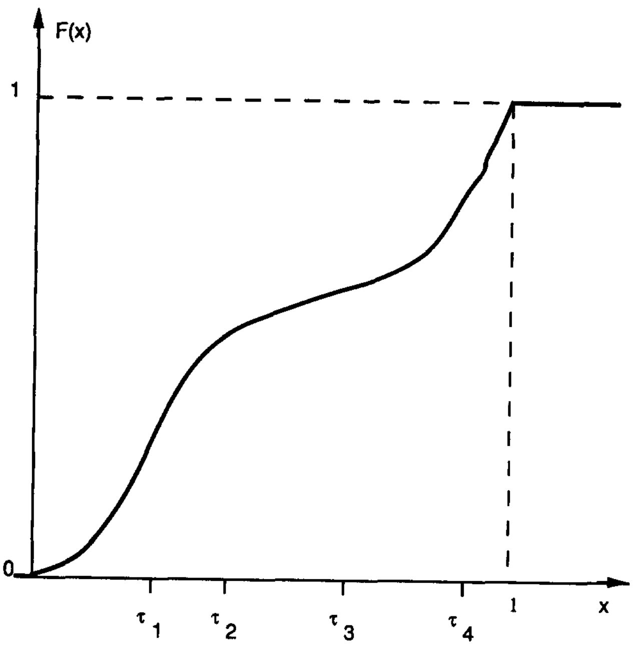

Figure 2. Graphical depiction of $F(x)$ for multinomial example.

$$P\Big\{ X_{(n_1)} \leqslant \tau_1, \tau_1 < X_{(n_1+1)}, X_{(n_1+n_2)} \leqslant \tau_2, \tau_2 < X_{(n_1+n_2+1)},$$
$$X_{(n_1+n_2+n_3)} \leqslant \tau_3, \tau_3 < X_{(n_1+n_2+n_3+1)}, X_{(n_1+n_2+n_3+n_4)}$$
$$\leqslant \tau_4 \,|\, N(1) = \sum_{i=1}^{4} n_i \Big\} \equiv \eta(\underline{n}, \underline{\tau}),$$

where $\underline{n} \equiv \{n_i\}$. The event whose probability we seek corresponds to a prespecified number of random variables falling into each of four disjoint intervals. If for $i = 1, 2, 3, 4$, we define $p_i = F(\tau_i) - F(\tau_{i-1})$, where $\tau_0 \equiv 0$, then we recognize this problem as an example of the multinomial distribution with

$$\eta(\underline{n}, \underline{\tau}) = \frac{N!}{n_1! n_2! n_3! n_4!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}, \text{ and } \sum_{i=1}^{4} n_i = N.$$

The multinomial is also derived by applying Theorem 1 as follows:

$$\underline{t} = \{t_1 = \tau_1, t_2 = \tau_1, \ldots, t_{n_1} = \tau_1, t_{n_1+1} = \tau_2, \ldots, t_{n_1+n_2}$$
$$= \tau_2, t_{n_1+n_2+1} = \tau_3, \ldots\}$$
$$\underline{s} = \{s_1 = 0, s_2 = 0, \ldots, s_{n_1} = 0, s_{n_1+1} = \tau_1, \ldots, s_{n_1+n_2}$$
$$= \tau_1, s_{n_1+n_2+1} = \tau_2, \ldots\}$$

and $\underline{v}$ is the ordered merging to $\underline{t}$ and $\underline{s}$. The vector $\underline{v}$ contains adjacent elements that are equal in value. Recognizing that for any i having $v_i = v_{i-1}$, Eq. (4) yields $W_{k_i} = W_{k,i-1}$, we can transform the vector $\underline{v}$ into a reduced

Exhibit 1. Test for Impossible Events in the $N = 4$ Example

"Negative length" right-most interval test (implying impossible event):

	$k \backslash$	$i = 1$	2	3	4	5	6	7	8
	0		—	—	—	—	—	—	—
?	1	Y	Y	N	N	N	N	N	—
$s_K \geqslant v_i$	2	Y	Y	N	N	N	N	N	—
	3	Y	Y	Y	Y	N	N	N	—
	4	Y	Y	Y	Y	Y	Y	Y	N

These entries always "Y"

Exhibit 2. Determining Maximum Allowable Value for Summation Index

Matrix $\underline{W}$: m = maximum number of r.v.'s allowed in $(v_{i-1}, v_i]$, given inequality contraints in $(0, v_{i-1}]$
= maximum value for summation index j in Eq. (4)

	k	$i = 1$	2	3	4	5	6	7	8
	0	1	1	1	1	1	1	1	—
$\underline{W} =$	1	0	0	$m = 1$	$m = 0$	$m = 0$	$m = 0$	$m = 0$	—
	2	0	0	$m = 2$	$m = 1$	$m = 1$	$m = 0$	$m = 0$	—
	3	0	0	0	0	$m = 2$	$m = 1$	$m = 0$	—
	4	0	0	0	0	0	0	0	$m = 1$

vector $\underline{v}' = \{0, \tau_1, \tau_2, \tau_3, \tau_4\}$. If we compute elements of the corresponding reduced matrix $\underline{W}' = (W'_{ki})_{k=1,\ldots,N, i=1,\ldots,5}$, then we are obtaining each of the *different* columns of the original matrix $\underline{W}$. We know that $W'_{ki} = 0$ if $s_k \geqslant v'_i$, implying $W'_{ki} = 0$ if $s_k \geqslant v_i$. Examining components of $\underline{s}$, we find that $s_k = 0$ for $k = 1, \ldots, n_1$; $s_k = \tau_1$ for $k = n_1 + 1, \ldots, n_1 + n_2$; $s_k = \tau_2$ for $k = n_1 + n_2 + 1, \ldots, n_1 + n_2 + n_3$; and $s_k = \tau_3$ for $k = n_1 + n_2 + n_3 + 1, \ldots, n_1 + n_2 + n_3 + n_4 = N$. Hence, we find the "southwest corner" elements of $\underline{W}'$ that are equal to zero: $W'_{k1} = 0, k = 1, \ldots, N$; $W'_{k2} = 0, k = n_1 + , \ldots, N$; $W'_{k3} = 0, k = n_1 + n_2 + 1, \ldots, N$; $W'_{k4} = 0, k = n_1 + n_2 + n_3 = 1, \ldots, N$.

The nonzero elements of $\underline{W}'$ are found by following a simple descending "staircase" through the matrix, starting at W'_{12} and ending at $W'_{N,5}$. The summation inequality constraint is $v_i \leqslant t_{k-j+1}$, implying for the reduced matrix $\underline{W}'$ the summation inequality constraint $v'_i \leqslant t_{k-j+1}$. For $i = 2$, we require $v'_2 = \tau_1 \leqslant t_{k-j+1}$, implying that the summation index is allowed values over its full range, $j = 1, \ldots, k$; using Eq. (4) with its "row 0" boundary condition this implies that for $k = 1, \ldots, n_1$, $W'_{k2} = F(\tau_1)^k$. For $i = 3$, we require $v'_3 = \tau_2 \leqslant t_{k-j+1}$, implying that for $k = n_1 + 1, \ldots, n_1 + n_2$, the summation index varies between $j = 1$ and $j = k - n_1$. Hence for $k = n_1 + 1, \ldots, n_1 + n_2$, we find using Eq. (4)

$$W'_{k3} = \binom{k}{k - n_1} F(\tau_1)^{n_1} [F(\tau_2) - F(\tau_1)]^{k-n_1}.$$

Following similar reasoning we find for $k = n_1 + n_2 + 1, \ldots, n_1 + n_2 + n_3$,

$$W'_{k3} = \binom{n_1 + n_2}{n_2} \binom{k}{k - n_1 - n_2} F(\tau_1)^{n_1} \times [F(\tau_2) - F(\tau_1)]^{n_2} [F(\tau_3) - F(\tau_2)]^{k-n_1-n_2}.$$

or, equivalently,

$$W'_{k3} = \frac{k!}{n_1! n_2! (k - n_1 - n_2)!} F(\tau_1)^{n_1} [F(\tau_2) - F(\tau_1)]^{n_2} \times [F(\tau_3) - F(\tau_2)]^{k-n_1-n_2}.$$

Finally, again using Eq. (4) and noting allowed values for the summation index, we obtain the desired multinomial distribution,

$$W'_{N,5} = \frac{N!}{n_1! n_2! n_3! (N - n_1 - n_2 - n_3)!} \times F(\tau_1)^{n_1} [F(\tau_2) - F(\tau_1)]^{n_2} [F(\tau_3) - F(\tau_2)]^{n_3} \times [F(\tau_4) - F(\tau_3)]^{N-n_1-n_2-n_3}.$$

2. Inferring the Performance of Poisson Arrival Queues

In [8] Larson uses events of order statistics to derive an algorithm, the "Queue Inference Engine (QIE)," to compute various estimates of performance measures of Poisson arrival queues from a set of transactional data. The transactional data are the moments of service initiation and service completion for each customer served. The "signature of a queue" from the transactional data is a back-to-back service completion and service initiation. That is, if at least one customer is delayed in queue, it is assumed that said customer will enter service virtually immediately following departure of a customer from service. If there are M servers and all are busy, any arriving customer must be delayed in queue.

Our period of analysis is a single congestion period, a continuous time interval when all M servers are busy and all arriving customers (assumed to be Poisson) must queue for service. A congestion period commences the moment that all M servers become busy and ends the moment that one of the servers completes service and then remains idle for a positive period of time. In our notation here, N is (1) the total number of customers who complete service during and before the end of a particular congestion period and, also, (2) the total number of new customers to arrive to the queueing system during the congestion period. By convention $N = 0$ for any congestion period having no queued customers. In a queue inferencing setting, t_i has two definitions: (1) it is the observed time of departure of the ith departing customer to leave the system during the congestion period; (2) it is also the observed time for the ith customer from the queue to enter service, not necessarily in a FCFS manner. The two sets of individuals comprising the set of arriving customers and the set of departing customers during a congestion period are never identical, and may be disjoint. The number of servers M does not enter into the analysis, nor do any distributional properties of the service times (e.g., there is no requirement for i.i.d. service times). We do assume that service times are independent of arrival times, else the assumptions regarding the distributional form of the order statistics may be incorrect. For any given congestion period, the QIE computations may occur any time after completion of the congestion period.

Using the fact that for a Poisson process the N unordered arrival times during any fixed time interval $(0, T]$ are i.i.d. and scaling the congestion period to $(0, 1]$, then in our notation $\Gamma(0, \underline{t})$ is the *a priori* probability that the (unobserved) arrival times $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ obey the inequalities $X_{(i)} \leqslant t_i$ for all $i = 1, 2, \ldots, N$, a condition that must hold for the congestion period to persist. That is, "$X_{(i)} \leqslant t_i$" simply says that the ith arriving queued customer must arrive (and enter the queue) before completion of service of the ith departing customer from service; otherwise there would be no customer to select from the queue at time t_i, negating the possibility of a back-to-back service completion and service initiation, a condition known to be true from the transactional data. If the Poisson arrival process is homogeneous then the unordered arrival times are i.i.d. *uniform* and the rate parameter of the process does not enter the analysis. If the arrival process is nonhomogeneous then the time-dependent arrival rate parameter $\lambda(t)$ must be known up to a positive multiplicative constant for use in computing the c.d.f. $F(x)$, i.e.,

$$F(x) = \frac{\int_0^x \lambda(t)\, dt}{\int_0^1 \lambda(y)\, dy}, \qquad 0 \leqslant x \leqslant 1.$$

For simplicity we assume that $F(x)$ is strictly monotone nondecreasing continuous, thereby reducing to zero the probability of simultaneous customer arrivals.

Queue inferencing, despite its youth, is a field with a growing number of researchers and papers. In [2] Bertsimas and Servi derived the first $O(N^3)$ algorithm for queue performance estimation based on N-dimensional integration. They also extend in various ways the original paper of Larson, for instance explicitly treating time varying Poisson arrivals, certain non-Poisson arrivals and bounding each moment of the customer waiting time in $O(N^3)$ and providing an exact (but not $O(N^3)$) expression for the waiting time c.d.f. of each customer. In [3] Daley and Servi demonstrate how the queue inference calculations may be performed in $O(N^3)$ time using ideas of nonhomogeneous Markov chains with taboo states. In [4] they extend these ideas to develop a "kernel" within the Markov framework which allows development of queue inferencing calculations within more complex environments involving, for instance, Markovian reneging. The net result of [2], [3], and [8] is the existence of three $O(N^3)$ algorithms for computing essentially the same queue statistics, each based on a different approach and each relying exclusively on transactional data. Gawlick[6] reports a successful application of these techniques within a communication network context. Recently in her Ph.D. thesis, published as an M.I.T. Operations Research Center Technical Report, Hall using order statistics arguments advanced queue inferencing in a number of directions[7]; among her results, she finds a stochastic dominance bound for queue inference approximations, an $O(N^3)$ algorithm for computing the conditional probability distribution of the waiting time of the jth customer to be delayed in queue (assuming FCFS queue discipline), and queue inferencing algorithms for complex situations in which for instance one knows that during a congestion period the queue did not exceed L_0 in length or in which one knows during the congestion period the precise times at which the queue exceeds L_0 in length. Following [7] and [8], our approach is based on order statistics.

To date, the QIE and related queue inferencing algorithms have been implemented to monitor queues and to help schedule servers in post offices, airports and banks. Additional applications have included telecommunications systems and automatic teller machines.

In the following subsections we illustrate the variety of performance charactertistics that can be computed simply using Theorem 1 (or equivalently Theorem 2). In the final two sections of the paper we expand the methodology to derive new computational algorithms focusing on (1) queue delay c.d.f. and (2) correlations of queue delays between customers in the same congestion period.

2.1. The Maximum Experienced Queue Delay

Assume we have a FCFS queue. Suppose we consider a congestion period having N customers with observed departure time vector $\underline{t}$, and we are interested in the maximum time that any of the N customers was delayed in queue, given $\underline{t}$. More precisely, we are interested in the c.d.f. of the maximum of N nonindependent r.v.'s, the in-queue waiting times of the N queued customers, given $\underline{t}$.

Define

$D(\tau|\underline{t}) \equiv$ conditional probability that none of the N
customers waited τ or more time units,
given the obseved departure time data.

Set $\underline{s} = \underline{t} - \tau$, i.e., $s_i = \max\{t_i - \tau, 0\}$ for all $i = 1, 2, \ldots, N$. Then $\Gamma(\underline{t} - \tau, \underline{t})$ is the *a priori* probability that the observed departure time inequalities will be obeyed *and* that *no* arrival waits τ or more time units in queue. Clearly,

$$D(\tau|\underline{t}) = \Gamma(\underline{t} - \tau, \underline{t})|\Gamma(0, \underline{t}). \qquad (7)$$

As an example, reconsider Example 2 with $N = 3$, $\underline{t} = (1/3, 2/3, 1)$, $\underline{s} = (1/6, 1/2, 5/6)$. Here $\Gamma(0, \underline{t})$ is the apriori probability that the Poison arrivals, assumed in this case to be homogeneous, will obey the departure time inequalities: $X_{(1)} \leqslant 1/3$, $X_{(2)} \leqslant 2/3$ and $X_{(3)} \leqslant 1$. This is the same example worked out by Larson in [8], and our result $\Gamma(0, \underline{t}) = 16/27$ agrees with his result. For this example we see that $\underline{s} = \underline{t} - 1/6$, and thus $\Gamma(\underline{t} - \tau, t)$ is the apriori probability that the arrivals will obey the departure time inequalities, and that no customer waits more than 1/6 time unit in queue. Hence, $D(\tau|\underline{t}) = (1/36)/(16/27) = 3/64 = 0.046875$ = conditional probability that none of the 3 customers waited in queue 1/6 or more time units, given the observed departure time data, $\underline{t}$.

2.2. Maximum Queue Length

The derived methodology is also readily applied to examination of the maximum length of the queue during any congestion period for which departure time data are known. Without any assumption regarding queue discipline, suppose we define $\underline{s} = \underline{s}^{*K}$ such that

$$s_i^{*K} = t_{(i-K)} \qquad \text{for all } i = 1, 2, \ldots, N;\ K = 1, 2, \ldots, N,$$

where a non-positive subscript on t implies a value of zero. These values for $\underline{s}$ imply that each arriving customer i has to arrive *after* the departure time of departing customer $i - K$ during the congestion period. Now we can compute the conditional probability that the queue length did not exceed K during the congestion period, given $\underline{t}$:

$$\begin{aligned} P(Q \leqslant K|\underline{t}) &= \Pr\{\text{queue length did not exceed } K \text{ during} \\ &\qquad \text{the congestion period} \\ &\qquad |\text{observed departure time data}\} \\ &= \Gamma(\underline{s}^{*K}, \underline{t})/\Gamma(0, \underline{t}). \end{aligned} \qquad (8)$$

For the departure time data of Example 2 with $N = 3$, $t = (1/3, 2/3, 1)$, we compute $P(Q \leqslant 1|\underline{t}) = 3/8$, $P(Q \leqslant 2|\underline{t}) = 15/16$ and $P(Q \leqslant 3|\underline{t}) = 1.0$.

2.3. Probability Distribution of Queue Length

Following the same arguments as in [8], we can utilize the $O(N^3)$ computational algorithm to determine for any queue discipline the probability distribution of queue length at departure epochs, and, by a balance of flow argument, this distribution is also the queue length distribution experienced by arriving customers.

2.4. The Cumulative Distribution of Queue Delay: An $O(N^4)$ Algorithm

The new recursive algorithms allow exact computation of points on the conditional in-queue waiting time distribution, given the observed departure data. To our knowledge, this computation was not feasible with any previous queue inference procedure.

Again assume we have a FCFS queue. Suppose we define

$$\begin{aligned} \beta_j(\tau|\underline{t}) \equiv \Pr\{&j\text{th customer to arrive during the congestion} \\ &\text{period waited less than } \tau \text{ time units}|\text{observed} \\ &\text{departure time data}\}, \end{aligned}$$

Then if we set $\underline{s} = \underline{s}^j$, defined so that

$$\begin{aligned} s_i^j &= 0 && i = 1, 2, \ldots, j - 1 \\ s_i^j &= \operatorname{Max}\{t_j - \tau, 0\} && i = j, j + 1, \ldots, N. \end{aligned}$$

we can write

$$\beta_j(\tau|\underline{t}) = \Gamma(\underline{s}^j, \underline{t})/\Gamma(0, \underline{t}). \qquad (9)$$

This result allows us to determine for any congestion period the probability that a *random* customer waited less than τ time units, given the observed departure data. We simply compute Eq. (9) once for each value of j and average the results. For Example 2 with $\tau = 1/6$ we find $\beta_1(1/6|\underline{t}) = 49/128$, $\beta_2(1/6|\underline{t}) = 30/128 = 15/64$ and $\beta_3(1/6|\underline{t}) = 36/128 = 9/32$, and thus the probability that a random queued customer waited less than 1/6 time unit is $(1/3)([49 + 30 + 36]/128) = 115/384 \cong 0.2995$. By applying Eq. (9) for differing values of τ, we can determine any number of points on the c.d.f. of queue delay, conditioned on the observed departure time data.

The problem with this approach is that for each value of τ an $O(N^3)$ algorithm must be performed for each of N customers, resulting in an $O(N^4)$ procedure. If a less accurate computation is permitted or if less computational work is required, one can select the customer j at random from the N available and apply Eq. (9) to the selected customer. But such a procedure will most likely increase the variance of the estimate of the cumulative delay distribution.

3. The Cumulative Distribution of Queue Delay: An $O(N^3)$ Algorithm

One can utilize the ideas of Theorems 1 and 2 to create an $O(N^3)$ algorithm to compute the average probability (averaged over all N customers in a congestion period) that the in-queue wait is less than τ minutes, given the departure time data. To create the required forward recursion, let $\{z_i\}$ be the ordered merging of $\{0, t_1, t_2, \ldots, t_N\}$ and $\{s_i\} = \{\max[0, t_i - \tau]\}$, and define

$$\begin{aligned} Z_{ki} \equiv \Pr\{&0 < X_{(1)} \leqslant \min\{t_1, z_i\}, 0 < X_{(2)} \leqslant \min\{t_2, z_i\}, \ldots, \\ &0 < X_{(k)} \leqslant \min\{t_k, z_i\}|N(1) = k\}, \\ &= 1, 2, \ldots, 2N + 1,\ k = 1, 2, \ldots, N. \end{aligned} \qquad (10)$$

Note that for each value of j the lower limit on $X_{(j)}$ in Eq. (10) is 0, not s_j as in Eq. (2). We have the standard boundary condition $Z_{0i} = 1$, $i = 1, 2, \ldots, 2N + 1$, and we note that since $z_1 = 0$, $Z_{k1} = 0$, $k = 1, 2, \ldots, N$. Since $\{z_i\}$ is a refinement of the appropriate $\{v_i\}$ sequence, by a similar argument to that of Theorem 1 the quantities Z_{ki} are computable in $O(N^3)$ time using the forward recursion,

$$Z_{ki} = Z_{k,i-1} + \sum_{\substack{j=1 \\ \text{s.t. } z_i \leqslant t_{k-j+1}}}^{k} \binom{k}{j} Z_{k-j,i-1}[F(z_i) - F(z_{i-1})]^j \tag{11}$$

Note that there is no test for impossible events, comparing s_k to v_i, since in this recursion the left hand side of each interval is 0, not s_k.

To obtain the required backward recursion we use the same definition of $\{z_i\}$ and define

$$\begin{aligned} R_{ki} \equiv \Pr\{z_i < X_{(1)} &\leqslant t_{N-k+1}, z_i < X_{(2)} \\ &\leqslant t_{N-k+2}, \ldots, z_i < X_{(k)} \leqslant t_N | N(1) = k\}, \\ &i = 1,2,\ldots,2N,2N+1,\ k = 1,2,\ldots,N. \end{aligned} \tag{12}$$

Note that for each value of j the lower limit on $X_{(j)}$ in Eq. (12) is z_i, not $\max\{s_{N-k+j}, z_i\}$ as in Eq. (5). Here we have the usual boundary condition $R_{0i} = 1$, $i = 1,2,\ldots,2N+1$, and we note $R_{k,2N+1} = 0$, $k = 1,2,\ldots,N$. By a similar argument as that in Theorem 2, the quantities R_{ki} are computable in $O(N^3)$ time using the backward recursion,

$$R_{ki} = \left\{ \begin{array}{ll} R_{k,i+1} + & \\ \displaystyle\sum_{j=1}^{k} \binom{k}{j} R_{k-j,i+1}[F(z_{i+1}) - F(z_i)]^j & \\ \quad \text{if } z_i < t_{N-k+1} & \\ 0 \quad \text{if } z_i \geqslant t_{N-k+1} & \end{array} \right\} \tag{13}$$

Note that there is no constraint on the summation index, comparing z_i to s_{N-k+j}, since in this recursion the lower limit on $X_{(j)}$ is simply z_i.

Upon completion of each of the two recursive algorithms, Eqs. (11) and (13), one has obtained the same final quantity, $Z_{N,2N+1} = R_{N,1} = \Gamma(0,\underline{t})$ = the *a priori* probability that the departure time inequalities are obeyed, which is the fundamental probability found in the original QIE paper.[8]

3.1. The Algorithm

Define the *average* probability that a customer is delayed in queue less than τ time units, given the departure time data, as

$$\beta(\tau|\underline{t}) \equiv (1/N) \sum_{j=1}^{N} \beta_j(\tau|\underline{t}). \tag{14}$$

The quantity $\beta(\tau|\underline{t})$ can be computed in $O(N^3)$ time using the *forward/backward "kiss" (FBK)* algorithm of

Theorem 3.

$$\begin{aligned} \beta(\tau|\underline{t}) = \left(\frac{1}{N}\right)\left(\frac{1}{\Gamma(0,\underline{t})}\right) \sum_{j=1}^{N} \sum_{k=1}^{j} \binom{N}{j-k} \\ \times Z_{j-k,i(s_j)} R_{N-j+k,i(s_j)}, \end{aligned} \tag{15}$$

where $i(s_j)$ is the index of the element of $\{z\}$ equaling s_j, i.e., $s_j = z_{i(s_j)}$.

Proof. It is sufficient to show that

$$\beta_j(\tau|\underline{t}) = \left(\frac{1}{\Gamma(0,\underline{t})}\right) \sum_{k=1}^{j} \binom{N}{j-k} Z_{j-k,i(s_j)} R_{N-j+k,i(s_j)}$$

or equivalently that

$$\Gamma(\underline{s}^j,\underline{t}) = \sum_{k=1}^{j} \binom{N}{j-k} Z_{j-k,i(s_j)} R_{N-j+k,i(s_j)}.$$

But we can write

$$\begin{aligned} \Gamma(\underline{s}^j,\underline{t}) = \Pr\{0 < X_{(1)} &\leqslant t_1, 0 < X_{(2)} \leqslant t_2, \ldots, \\ &\max\{0, t_j - \tau\} < X_{(j)} \leqslant t_j, \ldots, \max\{0, t_j - \tau\} \\ &< X_{(N)} \leqslant t_N | N(1) = N\}. \end{aligned}$$

Recall that $s_j = \max[0, t_j - \tau]$ and that $i(s_j)$ is the index of the element of $\{z\}$ equaling s_j, i.e., $s_j = z_{i(s_j)}$. By partitioning the arrivals into an *early* set and a *late* set, where the lth ordered entry in the early set arrives before $\min\{t_l, z_{i(s_j)}\}$, we can write

$$\begin{aligned} \Gamma(\underline{s}^j,\underline{t}) = \Pr\Big\{0 < X_{(1)} &\leqslant \min\{t_1, z_{i(s_j)}\}, 0 < X_{(2)} \\ &\leqslant \min\{t_2, z_{i(s_j)}\}, \ldots, 0 < X_{(j-1)} \\ &\leqslant \min\{t_{j-1}, z_{i(s_j)}\}, z_{i(s_j)} \\ &< X_{(j)} \leqslant t_j, z_{i(s_j)} < X_{(j+1)} \leqslant t_{j+1}, \ldots, z_{i(s_j)} \\ &< X_{(N)} \leqslant t_N | N(1) = N\Big\} \\ &\quad + \Pr\Big\{0 < X_{(1)} \leqslant \min\{t_1, z_{i(s_j)}\}, 0 \\ &< X_{(2)} \leqslant \min\{t_2, z_{i(s_j)}\}, \ldots, 0 < X_{(j-2)} \\ &\leqslant \min\{t_{j-2}, z_{i(s_j)}\}, z_{i(s_j)} \\ &< X_{(j-1)} \leqslant t_{j-1}, z_{i(s_j)} < X_{(j)} \leqslant t_j, \ldots, z_{i(s_j)} \\ &< X_{(N)} \leqslant t_N | N(1) = N\Big\} \\ &\quad + \cdots + \Pr\Big\{0 < X_{(1)} \leqslant \min\{t_1, z_{i(s_j)}\}, 0 \\ &< X_{(2)} \leqslant \min\{t_2, z_{i(s_j)}\}, \ldots, 0 < X_{(j-k)} \\ &\leqslant \min\{t_{j-k}, z_{i(s_j)}\}, z_{i(s_j)} < X_{(j-k+1)} \\ &\leqslant t_{j-k+1}, z_{i(s_j)} < X_{(j-k+2)} \\ &\leqslant t_{j-k+2}, \ldots, z_{i(s_j)} < X_{(N)} \\ &\leqslant t_N | N(1) = N\Big\} + \cdots, \end{aligned}$$

where the kth term on the RHS corresponds to $j - k$ arrivals in the early set and $N - j + k$ arrivals in the late set. But, recognizing that there are $\binom{N}{j-k}$ ways in which the N total arrivals may be partitioned into two subsets, we can write the generic kth term as the product of two conditional probabilities and $\binom{N}{j-k}$,

$$\Pr\Big\{0 < X_{(1)} \leqslant \min\{t_1, z_{i(s_j)}\}, 0 < X_{(2)} \leqslant \min\{t_2, z_{i(s_j)}\}, \ldots, 0 < X_{(j-k)} \leqslant \min\{t_{j-k}, z_{i(s_j)}\}, z_{i(s_j)} < X_{(j-k+1)} \leqslant t_{j-k+1}, z_{i(s_j)} < X_{(j-k+2)} \leqslant t_{j-k+2}, \ldots, z_{i(s_j)} < X_{(N)} \leqslant t_N | N(1) = N\Big\}$$

$$= \binom{N}{j-k} \times \Big[\Pr\Big\{0 < X_{(1)} \leqslant \min\{t_1, z_{i(s_j)}\}, 0 < X_{(2)} \leqslant \min\{t_2, z_{i(s_j)}\}, \ldots, 0 < X_{(j-k)} \leqslant \min\{t_{j-k}, z_{i(s_j)}\} | N(1) = j-k\Big\}\Big] \times \Big[\Pr\Big\{z_{i(s_j)} < X'_{(1)} \leqslant t_{j-k+1}, z_{i(s_j)} < X'_{(2)} \leqslant t_{j-k+2}, \ldots, z_{i(s_j)} < X'_{(N-j+k)} \leqslant t_N | N(1) = N-j+k\Big\}\Big],$$

where $\{X'_{(i)}\}$ is a set of $N - j + k$ order statistics over [0, 1], independent of the set $\{X_{(i)}\}$ over [0, 1]. But the last expression is simply $\binom{N}{j-k} Z_{j-k, i(s_j)} R_{N-j+k, i(sj)}$. ■

Remark. We call the algorithm *forward/backward "kiss"* (*FBK*) because it uses forward recursion to obtain Z_{ki}, then backward recursion to obtain R_{ki}, and the two algorithms "kiss" at the point $s_j = \max[0, t_j - \tau] = z_{i(s_j)}$.

3.2. Illustrative Computational Results

To illustrate the use of Theorem 3 we first apply it to the continuing $N = 3$ example that we have used throughout the paper. The results are shown in Exhibit 3 and are seen to be in agreement with the results found in §2.4 using the less efficient $O(N^4)$ algorithm. Note, as expected, $Z_{N,2N+1} = R_{N,1} = \Gamma(0, \underline{t}) = 16/27$, as found earlier in Example 2.

We also applied the algorithm in a set of Monte Carlo simulation runs modeling the well known M/M/1 (Poisson customer arrivals, i.i.d. negative exponential service times, single server) queue under alternative load factors (ratio of customer arrival rate to available customer service rate). As one illustrative example an M/M/1 queue was simulated with an average of 10 customers arriving per hour, available service rate of 20 customers per hour (i.e., mean service time of 1/20 hour or 3 minutes) for a total of 1000 + simulated hours. The average load factor was 0.5. The transactional data of each of the 4961 observed congestion periods were analyzed with Eq. (15) to estimate points on the steady state in-queue waiting time c.d.f. The random variable of interest is W_q, the in-queue delay experienced by a random customer. To compute estimates of the c.d.f. for W_q over many congestion periods, we must be careful to include also those customers experiencing zero queue delay, i.e., those who commence a congestion period by activating a previously idle server. Averaging in the appropriate way Eq. (15) yields the following c.d.f. estimates: $P\{W_q = 0\} = 0.4922$, $P\{W_q < 1 \text{ min.}\} = 0.5913$, $P\{W_q < 2 \text{ min.}\} = 0.6476$, $P\{W_q < 3 \text{ min.}\} = 0.6972$. From the theory of M/M/1 queues, the analytically obtained limiting results are $P\{W_q = 0\} = 0.5000$, $P\{W_q < 1 \text{ min.}\} = 0.5768$, $P\{W_q < 2 \text{ min.}\} = 0.6417$, $P\{W_q < 3 \text{ min.}\} = 0.6967$. [*Note*: We are using strict rather than nonstrict inequalities in the definition of the waiting time c.d.f., in order to be consistent with the strict LHS inequalities in the event definition of Eq. (2). Due to the continuous nature of the c.d.f., this convention is of no practical significance.]

Exhibit 3. Use of the $O(N^3)$ Algorithm with the Continuing $N = 3$ Example

$$\underline{t} = \{1/3, 2/3, 1\};\ \underline{s} = \{1/6, 1/2, 5/6\};\ \underline{z} = \{0, 1/6, 1/3, 1/2, 2/3, 5/6, 1\}$$

$$\mathbf{Z} = \begin{array}{c|ccccccc} \mathbf{k} = & i = 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & — \\ \mathbf{1} & 0 & 1/6 & 1/3 & 1/3 & 1/3 & 1/3 & — \\ \mathbf{2} & 0 & 1/36 & 1/9 & 2/9 & 1/3 & 1/3 & — \\ \mathbf{3} & 0 & 1/216 & 8/216 & 26/216 & 56/216 & 92/216 & 16/27 \end{array}$$

$$\mathbf{R} = \begin{array}{c|ccccccc} \mathbf{k} = & i = 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \mathbf{0} & — & 1 & 1 & 1 & 1 & 1 & 1 \\ \mathbf{1} & — & 5/6 & 2/3 & 1/2 & 1/3 & 1/6 & 0 \\ \mathbf{2} & — & 21/36 & 1/3 & 5/36 & 0 & 0 & 0 \\ \mathbf{3} & 16/27 & 49/216 & 0 & 0 & 0 & 0 & 0 \end{array}$$

$$\beta_1(1/6|\underline{t}) = \frac{1}{(16/27)}\binom{3}{0} Z_{0,2} R_{3,2} = 49/128$$

$$\beta_2(1/6|\underline{t}) = \frac{1}{(16/27)}\left\{\binom{3}{1} Z_{1,4} R_{2,4} + \binom{3}{0} Z_{0,4} R_{3,4}\right\} = 15/64 = 30/128$$

$$\beta_3(1/6|\underline{t}) = \frac{1}{(16/27)}\left\{\binom{3}{2} Z_{2,6} R_{1,6} + \binom{3}{1} Z_{1,6} R_{2,6} + \binom{3}{0} Z_{0,6} R_{3,6}\right\} = 9/32 = 36/128$$

$$\beta(\boldsymbol{\tau}|\underline{t}) \equiv (1/N) \sum_{j=1}^{N} \beta_j(\boldsymbol{\tau}|\underline{t}) = (1/3)\left\{\frac{49 + 30 + 36}{128}\right\} = \frac{115}{384} \cong 0.2995$$

3.3. Extensions

The distributional results above can be extended in a number of ways. For instance, suppose one wishes to obtain estimates of points on the c.d.f. of queue wait, given that in addition to the usual transactional data we also know that the queue did not exceed K in length during the congestion period. In that case, one redefines $\{s_i\} = \{\max[t_{i-K}, t_i - \tau)]\}$. (Again, we use the convention that a non-positive subscript on t implies a time of 0.) Then, when strictly positive, the earliest allowable arrival time for customer i is either (1) τ time units before her service commencement or (2) the service commencement time of the customer who was the Kth customer to arrive before customer i, whichever is greater. The new condition ensures that the queue does not exceed K in length during the congestion period. By employing this definition of $\{s_i\}$ one can obtain, using Eqs. (11), (13), and (15) with $O(N^3)$ computations, estimates of points on the queue wait distribution in a situation in which the queue length never exceeds K. In applying this result, use should be made of Eq. (8) to obtain the correct conditioning probability.

Similar extensions can be obtained using other conditioning events, with appropriate modifications to the definition of $\{s_i\}$.

4. Correlations

The recursive methods above can be applied to finding correlations between waiting times of two specified customers during a congestion period. Such results may be useful in exploring the validity of various approximations to exact recursive algorithms or in estimating the equivalent sample size of independent observatons of the queue, given the departure time data.

Using the notations of Sec. 3, we define

$$\begin{aligned} A_{l,j,k,i} \equiv \Pr\{ & s_j < X_{(1)} \leqslant \min\{t_l, z_i\}, s_j < X_{(2)} \\ & \leqslant \min\{t_{l+1}, z_i\}, \ldots, s_j < X_{(k)} \\ & \leqslant \min\{t_{l+k-1}, z_i\} \mid N(1) = k\}, \\ & \quad k = 1, \ldots, N - l + 1;\ j, l = 1, \ldots, N; \\ & \quad i = 1, \ldots, 2N + 1 \end{aligned} \tag{16}$$

We use as a boundary condition $A_{l,j,0,i} = 1$ for all i, j and l. By noting impossible events we see that for all i, j such that $\min\{t_l, z_i\} \leqslant s_j$, $A_{l,j,k,i} = 0$, implying that $A_{l,j,k,1} = 0$ for all j, k and l. For any fixed values of l and j one can compute in $O(N^3)$ time the quantities $A_{l,j,k,i}$ for the appropriate values of k and i using the forward recursion,

$$A_{l,j,k,i} = \left\{ \begin{array}{l} A_{l,j,k,i-1} + \\ \displaystyle\sum_{\substack{m=1 \\ \text{s.t. } z_i \leqslant t_{l+k-m}}} \\ \quad \binom{k}{m} A_{l,j,k-m,i-1}[F(z_i) - F(z_{i-1})]^m \\ 0 \qquad s_j \geqslant z_i \end{array} \right\} \tag{17}$$

Since l and j can each take on values $l, j = 1, 2, \ldots, N$ the entire four dimensional matrix $\mathbf{A} \equiv (A_{l,j,k,i})$ requires $O(N^5)$ computations.

We are now interested in the conditional joint probability that both arrival j_1 and arrival j_2 during the same congestion period waited less than τ time units, given the departure time data. For $j_2 > j_1$, define the unconditional joint probability

$$\begin{aligned} \Gamma(\underline{s}^{j_1, j_2}, \underline{t}) = \Pr\{ & 0 < X_{(1)} \leqslant t_1, 0 < X_{(2)} \\ & \leqslant t_2, \ldots, \max\{0, t_{j_1} - \tau\} < X_{(j_1)} \\ & \leqslant t_{j_1}, \ldots, \max(0, t_{j_1} - \tau) < X_{(j_2 - 1)} < t_{j_2 - 1}, \\ & \max\{0, t_{j_2} - \tau\} < X_{(j_2)} \\ & \leqslant t_{j_2}, \ldots, \max\{0, t_{j_2} - \tau\} < X_{(N)} \\ & \leqslant t_N \mid N(1) = N\} \end{aligned} \tag{18}$$

The quantity $\Gamma(\underline{s}^{j_1, j_2}, \underline{t})$ can be computed with a nested recursion that partitions [0, 1] into three subintervals, $[0, s_{j_1}]$, $(s_{j_1}, s_{j_2}]$, and $(s_{j_2}, 1]$. Intuitively, for the required event to occur, $s_{j_1} +$ is the earliest possible arrival time for arrival j_1, so that at most $j_1 - 1$ arrivals can occur in $[0, s_{j_1}]$. But there is also a minimum number of allowable arrivals in $[0, s_{j_1}]$; suppose that $a(s_{j_1})$ is the largest index j of t_j satisfying $t_j \leqslant s_{j_1}$; then at least $a(s_{j_1})$ arrivals must occur in $[0, s_{j_1}]$ or else the original arrival time inequalities implied in $\underline{t}$ would not be satisfied. For the second interval $(s_{j_1}, s_{j_2}]$ there are similar concerns. The earliest allowable arrival time for arrival j_2 is $s_{j_2} +$. And up to time s_{j_2} there must be at least $a(s_{j_2})$ arrivals, or else the arrival time inequalities of $\underline{t}$ would not be satisfied. Given the upper and lower bound constraints on the allowable number of arrivals in each of the first two intervals (either or both of which may be empty), distributing the unordered arrivals over the three intervals is rather straightforward. For a given number n of arrivals in $[0, s_{j_1}]$, where $n = a(s_{j_1}), \ldots, j_1 - 1$, the remaining $N - n$ arrivals are distributed over the other two intervals. Of those $N - n$ arrivals, for the desired event to occur, at least $N - j_2 + 1$ must be in $(s_{j_2}, 1]$ and the remainder (if any) are in $(s_{j_1}, s_{j_2}]$. Using essentially the same recursive logic as in Theorems 1–3, we obtain

Theorem 4.

$$\begin{aligned} \Gamma(\underline{s}^{j_1, j_2}, \underline{t}) = & \sum_{n = a(s_{j_1})}^{j_1 - 1} \sum_{m = \max[0, a(s_{j_2}) - n]}^{j_2 - 1 - n} \frac{N!}{n! m! (N - n - m)!} \\ & \times Z_{n, i(s_{j_1})} A_{n+1, j_1, m, i(s_{j_2})} R_{N-n-m, i(s_{j_2})} \end{aligned} \tag{19}$$

Proof. Omitted. ■

Consider that the matrices **Z** and **R** are computed in $O(N^3)$ time prior to insertion in Eq. (19), that the matrix **A** is computed in $O(N^5)$ time, and that Eq. (19) requires $O(N^2)$ computations for a fixed j_1 and j_2, the entire matrix of joint probabilities $\Gamma(\tau) = (\Gamma(\underline{s}^{j_1, j_2}, \underline{t}))_{j_1, j_2}$ can be obtained in $O(N^5)$ time.

Let the random variable $W_q(j)$ be the queue delay experienced by the jth customer to enter the system during a congestion period having a total of N customers ($j = 1, 2, \ldots, N$), given the usual departure time data. Consider as an application of Eq. (19) use of indicator random variables

$$X_j = \begin{Bmatrix} 1 & \text{if } W_q(j) < \tau \\ 0 & \text{Otherwise} \end{Bmatrix}, \text{ given the departure time data.}$$

Then using the property that an indicator r.v. can take on values of only 0 or 1, we have

$$E[X_j] = \Pr\{W_q(j) < \tau\} = E\left[X_j^2\right] = \beta_j(\tau|\underline{t}),$$

$$E[X_i X_j] = \Pr\{X_i = 1, X_j = 1\} = \Pr\{W_q(i) < \tau, W_q(j) < \tau\}$$

$$= \Gamma(\underline{s}^{i,j}, \underline{t})/\Gamma(0, \underline{t}),$$

$$\sigma_{X_j}^2 = E\left[X_j^2\right] - E[X_j]^2 = \beta_j(\tau|\underline{t})[(1 - \beta_j(\tau|\underline{t})]$$

With these quantities we can find for any customers i and $j (i \neq j)$ the correlation coefficient of X_i and X_j,

$$\rho_{ij} = \text{correlation coefficient}$$

$$= \frac{\text{COV}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{E[X_i X_j] - E[X_i]E[X_j]}{\sigma_{X_i}\sigma_{X_j}}$$

$$= \frac{[\Gamma(\underline{s}^{i,j}, \underline{t})/\Gamma(0, \underline{t})] - \beta_i(\tau|\underline{t})\beta_j(\tau|\underline{t})}{\sqrt{\beta_i(\tau|\underline{t})\{1 - \beta_i(\tau|\underline{t})\}\beta_j(\tau|\underline{t})\{1 - \beta_j(\tau|\underline{t})\}}}.$$

As usual, the correlation coefficient may vary between -1 and $+1$, with a value of 0 indicating that the random variables are uncorrelated.

Continuing the $N = 3$ example, the above computations are carried out in Exhibit 4. As we might expect intuitively, all three correlation coefficients are positive; this reflects the fact that a limited in-queue waiting time for one customer is statistically associated with a limited in-queue waiting time for other customers within the same congestion period. But, as we might expect, the degree of statisti-

Exhibit 4. Illustrative Computations with Joint Probabilities and Correlation Coefficients

To solve this problem we only need $\mathbf{A}_1 \equiv (A_{1,1,k,i})$, $\mathbf{A}_2 \equiv (A_{2,2,k,i})$. Using Eq. (17) we obtain

$$\mathbf{A}_1 = \begin{array}{c|ccccccc} \mathbf{k} = & i = \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} \\ \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & — \\ \mathbf{1} & 0 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & — \\ \mathbf{2} & 0 & 0 & 1/36 & 1/12 & 5/36 & 5/36 & — \\ \mathbf{3} & 0 & 0 & 1/216 & 7/216 & 19/216 & 34/216 & 49/216 \end{array}$$

$$\mathbf{A}_2 = \begin{array}{c|ccccccc} \mathbf{k} = & i = \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} \\ \mathbf{0} & & 1 & 1 & 1 & 1 & 1 & — \\ \mathbf{1} & 0 & 0 & 0 & 0 & 1/6 & 1/6 & — \\ \mathbf{2} & 0 & 0 & 0 & 0 & 1/36 & 1/12 & — \end{array}$$

$$\Gamma(\underline{s}^{1,2}, \underline{t}) = \sum_{n=0}^{0} \sum_{m=1}^{1} \frac{3!}{n!m!(3 - n - m)!} Z_{n,2} A_{n+1,1,m,4} R_{3-n-m,4}$$

$$= 3Z_{0,2} A_{1,1,1,4} R_{2,4} = 3(1)(1/6)(5/36) = 5/72$$

$$\Gamma(\underline{s}^{1,3}, \underline{t}) = \sum_{n=0}^{0} \sum_{m=2}^{2} \frac{3!}{n!m!(3 - n - m)!} Z_{n,2} A_{n+1,1,m,6} R_{3-n-m,6}$$

$$= 3Z_{0,2} A_{1,1,2,6} R_{1,6} = 3(1)(5/36)(1/6) = 5/72$$

$$\Gamma(\underline{s}^{2,3}, \underline{t}) = \sum_{n=1}^{1} \sum_{m=1}^{1} \frac{3!}{n!m!(3 - n - m)!} Z_{n,4} A_{n+1,2,m,6} R_{3-n-m,6}$$

$$= 6Z_{1,4} A_{2,2,1,6} R_{1,6} = 6(1/3)(1/6)(1/6) = 4/72$$

$$\rho_{ij} = \text{correlation coefficient} = \frac{[\Gamma(\underline{s}^{i,j}, \underline{t})/\Gamma(0, t)] - \beta_i(\boldsymbol{\tau}|\underline{t})\beta_j(\boldsymbol{\tau}|\underline{t})}{\sqrt{\beta_i(\boldsymbol{\tau}|\underline{t})\{1 - \beta_i(\boldsymbol{\tau}|\underline{t})\}\beta_j(\boldsymbol{\tau}|\underline{t})\{1 - \beta_j(\boldsymbol{\tau}|\underline{t})\}}}$$

$$\rho_{12} = \frac{[(5/72)/(16/27)] - (49/128)(30/128)}{\sqrt{(49/128)\{1 - 49/128\}(30/128)\{1 - 30/128\}}} = 0.13339$$

$$\rho_{13} = \frac{[(5/72)/(16/27)] - (49/128)(36/128)}{\sqrt{(49/128)\{1 - 49/128\}(36/128)\{1 - 36/128\}}} = 0.04357$$

$$\rho_{23} = \frac{[(4/72)/(16/27)] - (30/128)(36/128)}{\sqrt{(30/128)\{1 - 30/128\}(36/128)\{1 - 36/128\}}} = 0.14613$$

Exhibit 5. Matrix of Correlation Coefficients for an $N = 15$ Customer Congestion Period with $\tau = 0.03333$

$\underline{t}$ = (0.11638, 0.31580, 0.36149, 0.40434, 0.41049, 0.50302, 0.54814, 0.61105, 0.71665, 0.86028, 0.91845, 0.91988, 0.92351, 0.98407, 1.0)

Matrix of Correlation Coefficients:

-	0.02820	0.03011	0.02700	0.05097	0.01986	0.02109	0.01637	0.00799	0.00200	0.00134	0.00413	0.00802	0.00458	0.00570
-	—	0.18715	0.07616	0.08261	0.02037	0.01782	0.01184	0.00519	0.00122	0.00079	0.00237	0.00443	0.00243	0.00294
-	—	—	0.22103	0.18366	0.03875	0.03215	0.02060	0.00882	0.00206	0.00133	0.00395	0.00732	0.00398	0.00480
-	—	—	—	0.36959	0.06388	0.04857	0.02961	0.01230	0.00283	0.00181	0.00536	0.00987	0.00533	0.00639
-	—	—	—	—	0.15297	0.12006	0.07414	0.03102	0.00716	0.00459	0.01360	0.02508	0.01357	0.01627
-	—	—	—	—	—	0.25304	0.09937	0.03377	0.00716	0.00442	0.01267	0.02243	0.01167	0.01363
-	—	—	—	—	—	—	0.21594	0.06225	0.01251	0.00757	0.02128	0.03690	0.01883	0.02171
-	—	—	—	—	—	—	—	0.12341	0.02101	0.01204	0.03237	0.05331	0.02599	0.02912
-	—	—	—	—	—	—	—	—	0.04881	0.02254	0.05157	0.07162	0.03045	0.03144
-	—	—	—	—	—	—	—	—	—	0.10138	0.11372	0.08213	0.02376	0.01999
-	—	—	—	—	—	—	—	—	—	—	0.26677	0.10085	0.02696	0.01999
-	—	—	—	—	—	—	—	—	—	—	—	0.37176	0.09284	0.07101
-	—	—	—	—	—	—	—	—	—	—	—	—	0.21850	0.17518
-	—	—	—	—	—	—	—	—	—	—	—	—	—	0.35093
-	—	—	—	—	—	—	—	—	—	—	—	—	—	—

cal association is strongest for adjacent customers (i.e., customers 1 and 2 or customers 2 and 3) and weakest for the pair of nonadjacent customers (customers 1 and 3). Exhibit 5 displays the entire matrix of correlation coefficients for an $N = 15$ customer congestion period with $\tau = 0.03333$.

Acknowledgments

The work of the first author was supported by the Institute for Visualization and Perception Research. The work of the second author was supported by the National Science Foundation under Grant SES 9119962. Both authors thank Sue Hall for programming algorithms, obtaining computational results, and offering valuable suggestions and corrections based on careful readings of earlier drafts.

References

1. R.E. Barlow, D.J. Bartholomew, J.M. Bremmer, and H.D. Brunk, 1972. *Statistical Inference Under Order Restriction*, Wiley, NY.
2. D.J. Bertsimas and L.D. Servi, 1992. Deducing Queueing from Transactional Data: The Queue Inference Engine Revisited, *Operations Research 40*:S2, 217–228.
3. D.J. Daley and L.D. Servi, 1992. Exploiting Markov Chains to Infer Queue-Length from Transactional Data, *Journal of Applied Probability 29*, 713–732.
4. D.J. Daley and L.D. Servi, 1993. A Two-Point Markov Chain Boundary-Value Problem, *Advances in Applied Probability 25*, 607–630.
5. H.A. David, 1981. *Order Statistics*, Wiley, NY.
6. R. Gawlick, 1990. Estimating Disperse Network Queues: The Queue Inference Engine, *Computer Communication Review 20*, 111–118.
7. S.A. Hall, 1992. New Directions in Queue Inference for Management Implementation, Ph.D. Thesis in Operations Research, M.I.T., Cambridge, MA. Published as M.I.T. Operations Research Center Technical Report TR-200, June, 1992.
8. R.C. Larson, 1990. The Queue Inference Engine: Deducing Queue Statistics from Transactional Data, *Management Science 36*:5, 586–601, and 1991, The Queue Inference Engine: Addendum, *Management Science 37*:6, 1062.