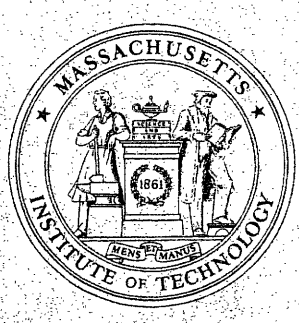
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Deducing Queue Statistics From Transactional Data

by

Richard C. Larson

OR 167-87 August 1987
Revised May 1988

Deducing Queue Statistics From Transactional Data

by

Richard C. Larson
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

August 1987
Revised May 1988

Abstract

The transactional data of a queueing system are the recorded times of service commencement and service completion for each customer served. With increasing use of computers to aid or even perform service one often has machine readable transactional data, but virtually no information about the queue itself. In this paper we propose a way to deduce the queueing behavior of Poisson arrival queueing systems from only the transactional data and the Poisson assumption. For each congestion period in which queues may form, the key quantities obtained are mean wait in queue, time-dependent mean number in queue, and probability distribution of the number in queue observed by a randomly arriving customer. The methodology builds on arguments of order statistics and usually requires a computer to evaluate a recursive function. The paper concludes with a proposed procedure for estimating the extent of balking and/or reneging present in a queueing system.

I. Introduction

Consider a Poisson arrival queueing system for which we have transactional data. That is, we know the time of service commencement and time of service completion for each customer who has been served by the system. Whenever there is a queue of customers waiting for service, we assume that following a customer's departure from service the next customer to enter service from the queue does so virtually immediately following said departure. Given this assumption the transactional data, when rank ordered, allow us to identify "congestion" periods that may involve the back-to-back sequential service of two or more customers. Such congestion periods are periods during which arriving customers must wait in queue prior to service.

Our objective is to derive the queue statistics, including mean time spent waiting in queue, and the time-dependent mean number in queue from the transactional data. In other words, we wish to deduce queue behavior without observing the queue, but by drawing inferences from the transactional data and from the Poisson arrival assumption. There are many potential applications, including analysis of customers queueing at automatic teller machines (ATM's), automobile traffic delayed at signalized intersections, and individuals queued awaiting access to a limited number of communications channels.

Our approach focuses on a single congestion period. Since the completion (or commencement) of a congestion period constitutes a renewal point in any Poisson arrival queue, once we have obtained the results for one congestion period we have in essence solved the entire problem. As will become clear, our approach exploits arguments drawn from the field of "order statistics." (cf. Barlow et. al. [1972] and David [1981]). We will find that for most of our results we do not need to know the

arrival rate parameter of the Poisson process. In all of our work, the server or servers can be completely general; for instance, successive service times need not be i.i.d.

II. Examples

Example 1: Automatic Teller Machines

Consider a facility housing k automatic teller machines (ATM's) fed by a single queue. The system is said to be operating within a congestion period whenever all k ATM's are simultaneously busy, requiring any new arrivals to wait in queue. A congestion peiod commences (terminates) whenever the number of busy ATM's jumps from k-1 to k (k to k-1, respectively). A customer service time is the time (s)he "occupies" the space directly in front of the ATM. For many ATM systems this time is closely approximated by the magnitude of the difference in times between the customer's ATM card insertion and the machine's card ejection. These transaction times may be routinely recorded in a master data file. When the data for all k ATM"s are merged and time-ordered, they constitute (to close approximation) the customer transaction times required to determine queue statistics developed herein. These queue statistics may in turn be used by bank managers to monitor the use of ATM sites, providing an accurate means to identify those sites requiring additional (fewer) ATM's.

Example 2: "Invisible" Queues in Communications Systems

Many finite capacity communications systems have during periods of congestion invisible queues of customers outside the system, continuously trying to gain access to it.

One example is a k-channel land mobile radio system. Whenever all k channels are simultaneously in use, potential users in the field (in vehicles) having a message to transmit continuously monitor channel use and attempt to acquire a channel as soon as any one of the current k communications is completed. If at a given time

there are n such potential users awaiting a channel, they constitute a spatially dispersed invisible queue, a queue in which one of the waiting customers enters service very shortly after another customer completes service. Service discipline is not necessarily first-come, first-served. Within the context of this paper the customer transaction times are the moments of gaining channel access (service initiation) and message termination (service completion). These times can be routinely monitored and recorded by electronic sensors measuring energy in the various broadcast channels.

Another communication system example is a telephone system having system capacity j , capacity measured by the maximum number of customers allowed in service and in queue. This system is "congested" whenever j customers are in the system and subsequent potential customers ("callers") are lost (they get a "busy" signal). If all such lost customers continuously and repeatedly call back until they successfully enter the system, then the real time population m of such lost customers constitutes an invisible queue. Within the context of this paper, initiation of "service" occurs the moment a caller successfully enters the system and "termination" of service occurs the moment the telephone conversation is completed; hence the "service time" of this paper represents the sum of queueing delay and telephone conversation time in the telephone system.

Example 3: Traffic Queued at Intersections

Imagine a street intersection in which one of the streets entering the intersection is equipped with a pressure-sensitive cable placed across the street. Whenever a vehicle passes over the cable, its presence is detected and recorded. Suppose that vehicles traveling along that street toward the intersection arrive in the vicinity of the intersection according to a Poisson process. As the vehicles stop at the

intersection, perhaps due to a stop sign or a traffic light, a queue may form. This queue is depleted as vehicles pass over the cable and enter the intersection.

Within the context of this paper, the service initiation time for each vehicle is the time that the vehicle's front axle passes over the cable. The service completion time is the time the rear axle passes over the cable plus some reasonable constant (perhaps dependent on vehicular speed - which can be estimated) to allow for space between vehicles. A congestion period exists whenever the cable is registering vehicular movement and, if the intersection is signalized, whenever the light is "red" for vehicles attempting to pass over the cable and enter the intersection. Note that with a signalized intersection (1) successive moveups in vehicular queue position are not i.i.d., and (2) congestion periods can be caused by exogenous events (a "red light") as well as by simple queueing congestion.

The methods of this paper allow a traffic engineer to deduce the queueing behavior of vehicles at the intersection simply from the cable-recorded information, without ever observing the queue.

Example 4: Queueing Networks

A not so obvious application is in communication networks. At any given node of a communications network one has in general a complex queueing system in which arrivals are not Poisson (and not even regenerative) and the service process is complicated, typically not following i.i.d. or other "nice" assumptions. However, the cause of analytical tractability would be served if the (complex) arrival process could be approximated to be Poisson. Using transactional data (from the real system), one could estimate queue behavior at the node using the methods herein and compare to

observed queue behavior; if the two are "similar," then the Poisson arrival assumption is probably a reasonable approximation for modeling purposes.

III. Preliminaries

Suppose we consider a homogeneous Poisson process with rate parameter $\lambda > 0$. Over a fixed time interval [0,T] we are told that precisely N Poisson events occur. The N ordered arrival times are $0 \leqq X_1 \leqq X_2 \leqq \ldots \leqq X_N \leqq T$ (by implication $X_{N+1} > T$). The N unordered arrival times are $U_1, U_2, \ldots, U_N$, $0 \leqq U_i \leqq T$ $(i = 1, 2, \ldots, N)$. From the theory of order statistics, it is well-known that the U_i's are independent, uniformly distributed over [0,T]. If we now let N(t) be the number of arrivals over [0, t], $0 \leqq t \leqq T$, without further conditioning information the following are well-known for N(t):

$$E\left[N(t)\right] = (t/T)N \qquad \textbf{(a)}$$

$$\text{VAR}\left[N(t)\right] \equiv \sigma^2_{N(t)} = N\left(\frac{t}{T}\right)\left(\frac{T-t}{T}\right) \qquad \textbf{(b)} \qquad \textbf{(1)}$$

$$\Pr\left\{N(t) = k\right\} = \binom{N}{K}\left(\frac{t}{T}\right)^k\left(\frac{T-t}{T}\right)^{N-k} \qquad \textbf{(c)}$$

In a queueing environment, N(t) could represent the number of customers in queue at time t, assuming bulk service of all waiting customers at time T, such as occurs as signalized pedestrian crosswalks.

In more general queueing environments, customers usually leave one-at-a-time. Their service completion times within a congestion period impose a set of inequalities on the arrival times of other customers who waited in queue. It is this set of inequalities that produces precise conditioning information within the general

context of order statistics, conditioning information that we use to deduce queue behavior.

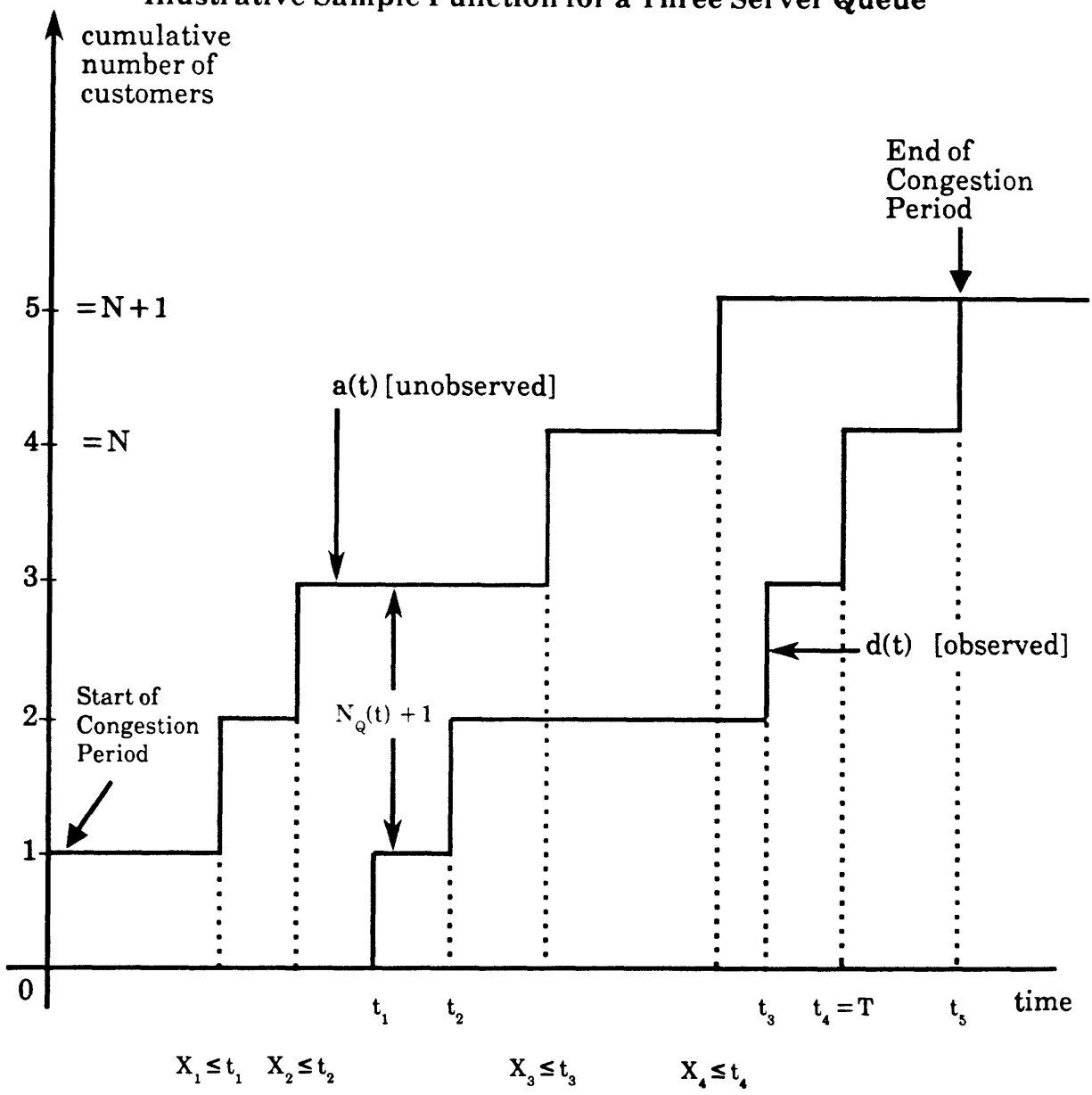
To illustrate key ideas and introduce notation, consider the sample function for a three server queue shown in Figure 1. In the example the congestion period commences at $t=0$ upon arrival of a customer who changes the remaining idle server's status from idle to busy. From transactional data the queue exhibits both service departures and service commencements at times t_1, t_2, t_3 and t_4, indicating that (1) all three servers were continuously busy during this time; (2) a queue existed at least at times t_1-, t_2-, t_3-, and t_4-; and (3) that the total number of customers queued was $N=4$. At time t_5 the transactional data indicate a service completion but no service commencement, thus ending the congestion period and thereby creating an idle server. From the transactional data, the cumulative number of departures through time t, d(t), is an observed function whereas the cumulative number of arrivals a(t) is not. From the conditioning information we know that the first arrival during the congestion period occurred at $X_1 \leqq t_1$, and that subsequent arrivals obey the inequalities $X_2 \leqq t_2$, $X_3 \leqq t_3$, $X_4 \leqq t_4 = T$. (Note that the end point of the conditional arrival interval for queued customers is $T = t_4$, not t_5). During the congestion period the number of customers in queue is $N_Q(t) = a(t) - d(t) - 1$. (For values of t equal to service completion times, i.e., $t = t_j$, one must be careful whether one is considering t_j+ or t_j-, as the former substracts from the queue the customer who enters service at time t, whereas the latter does not). The number of customers in the system (in service and in queue) is $N(t) = N_Q(t) + 3$.

The same concepts apply in more general queueing systems, including those with state-dependent service rates, shortest-job-first queue discipline, etc. The key idea

is to locate those service completion times which are accompanied by (nearly) simultaneous service commencement times.

Figure 1

Illustrative Sample Function for a Three Server Queue



a(t) = cumulative number of arrivals from commencement of congestion period

d(t) = cumulative number of departures from commencement of congestion period

$N_Q(t)$ = number of customers in queue

t_i = departure time of i^{th} customer served

X_i = arrival time of i^{th} customer to enter queue (i = 1, 2, 3, 4)

IV. Main Results

In this section we show how to deduce from transactional data (1) mean number of customers in queue τ time units after commencement of a congestion period; (2) time average queue length; (3) mean delay in queue; and (4) incidence probabilities. All of the results follow simply once we can determine, using order statistics, the a priori probability that the arrivals during a congestion period obey the time orderings imposed by the observed departure times.

1. Computing the Fundamental A Priori Probability

Recalling that X_i is the i^{th} arrival time during a congestion period and that t_i is the i^{th} departure time during the congestion period ($i = 1, 2, ...,, N$), define the a priori k-rank ordering probability,

$$\Psi_k(t_1, t_2, ..., t_k \mid N(T) = k) = \Pr\{X_1 \leqq t_1, X_2 \leqq t_2, ..., X_k \leqq t_k | N(T) = k\},$$

with $\Psi_0 \equiv 1$.

For $N = 1$ we have,

$$\Psi_1(t_1|N(T) = 1) = \Pr\{X_1 \leqq t_1 | \text{ precisely one Poisson arrival in } [0,T]\}$$

or

$$\Psi_1(t_1|N(T) = 1) = t_1/T. \tag{2}$$

We now find that $\Psi_k(\cdot)$ can be computed from $\Psi_0(\cdot), \Psi_1(\cdot), ..., \Psi_{k-1}(\cdot)$ by the recursion in

Lemma 1.

$$\Psi_k\left(t_1, t_2, ..., t_k | N(T) = k \right) = \tag{3}$$

$$\sum_{j=1}^{k} \binom{k}{k-j+1} \left(\frac{t_1}{T}\right)^{k-j+1} \left(\frac{T-t_1}{T}\right)^{j-1} \Psi_{j-1}\left(t_{k-(j-2)} - t_1, ..., t_{k-1} - t_1, t_k - t_1 | N(T-t_1) = j-1 \right)$$

Proof: (Induction) Equation 2 demonstrates that (3) holds for $k = 1$. Suppose (3) holds for k; we prove it holds for $k+1$.

Define the vector of k unordered arrival times $\underline{U}_k = (U_1, U_2, ..., U_k)$ and $\ell_i(\underline{U}_k) \equiv$ i^{th} largest of $U_1, U_2, ..., U_k$. For instance $\ell_1(\underline{U}_k) = MAX\{U_1, U_2, ..., U_k\}$ and $\ell_k(\underline{U}_k) = MIN\{U_1, U_2, ..., U_k\}$.

The argument proceeds as follows:

$$\Psi_{k+1}\left(t_1, t_2, ..., t_k, t_{k+1} | N(T) = k+1 \right) =$$

$$\Pr\left\{ X_1 \le t_1, X_2 \le t_2, ..., X_k \le t_k, X_{k+1} \le t_{k+1} | N(T) = k+1 \right\}$$

$$= \Pr\left\{ X_{k+1} \le t_1 \right\} + \Pr\left\{ X_k \le t_1 \text{ and } t_1 < X_{k+1} \le t_{k+1} \right\}$$

$$+ \Pr\left\{ X_{k-1} \le t_1 \text{ and } t_1 < X_k \le t_k, t_1 < X_{k+1} \le t_{k+1} \right\} + ...$$

$$+ \Pr\left\{ X_1 \le t_1, \text{and } t_1 < X_2 \le t_2, ..., t_1 < X_k \le t_k, t_1 < X_{k+1} \le t_{k+1} \right\}$$

$$= \Pr\left\{ \ell_1\left(\underline{U}_{k+1}\right) \le t_1 \right\} + \Pr\left\{ \ell_2\left(\underline{U}_{k+1}\right) \le t_1 \text{ and } t_1 < \ell_1\left(\underline{U}_{k+1}\right) \le t_{k+1} \right\}$$

$$+ \Pr\left\{ \ell_3\left(\underline{U}_{k+1}\right) \le t_1 \text{ and } t_1 < \ell_2\left(\underline{U}_{k+1}\right) \le t_k, t_1 < \ell_1\left(\underline{U}_{k+1}\right) \le t_{k+1} \right\}$$

$$+ \ldots + \Pr\left\{\ell_{k+1}\left(\underline{U}_{k+1}\right) \le t_1 \text{ and } t_1 < \ell_k\left(\underline{U}_{k+1}\right) \le t_2, \ldots,\right.$$

$$\left. t_1 < \ell_2\left(\underline{U}_{k+1}\right) \le t_k,\ t_1 < \ell_1\left(\underline{U}_{k+1}\right) \le t_{k+1}\right\}$$

$$= \left(\frac{t_1}{T}\right)^{k+1} + \left(k+1\right)\left(\frac{t_1}{T}\right)^{k}\left(\frac{T-t_1}{T}\right)\Psi_1\left(t_{k+1} - t_1 | N(T-t_1) = 1\right)$$

$$+ \binom{k+1}{2}\left(\frac{t_1}{T}\right)^{k-1}\left(\frac{T-t_1}{T}\right)^{2}\Psi_2\left(t_k - t_1, t_{k+1} - t_1 | N(T-t_1) = 2\right) + \ldots$$

$$+ \binom{k+1}{k}\left(\frac{t_1}{T}\right)\left(\frac{T-t_1}{T}\right)^{k}\Psi_k\left(t_2 - t_1, \ldots, t_k - t_1, t_{k+1} - t_1 | N(T-t_1) = k\right) .$$

■

Consider a queue congestion period starting at $t = 0$ during which N customers arrive. Observed departure epochs followed immediately by a service commencement are $t_1, t_2, \ldots, t_N = T$, where $0 \leqq t_i \leqq t_{i+1} \leqq T, i = 1, 2, \ldots, N - 1$. The a priori probability that the Poisson generated order statstics will obey the observed orderings in the data, given N arrivals in [0, T], is,

$$P\{X_1 \le t_1, X_2 \le t_2, \ldots, X_N \le t_N | N(T) = N\} \equiv P\{E(\underline{t}) | N(T) = N\} = \tag{4}$$

$$\Psi_N(t_1, t_2, \ldots, t_N | N(T) = N),$$

where $\underline{t} \equiv (t_1, t_2, \ldots, t_N)$ and $E(\underline{t}) = \text{Event}\{X_1 \leqq t_1, X_2 \leqq t_2, \ldots, X_N \leqq t_N\}$. Unfortunately computing (4) from the recursion in Lemma 1 requires $0(2^N)$ computations.

In Lemma 1, we may consider the "left-hand" interval $[0, t_1]$ as containing the "arrival time" of one or more tagged customers, with the arrival times of the remaining customers appropriately distributed over $[t_1, T]]$ and subsumed in the rank ordering probability. An alternative approach is to place the tagged arrival time(s) of the recursion in a corresponding "right-hand" interval, with the remaining arrival times dispersed appropriately from 0 to the boundary of that interval. The advantage of this approach is that it reduces the computational work from $0(2^N)$ to $0(N^3)$ by utilizing efficiently previously computed quantities.

Define

$$a_{ki}(\underline{t}) \equiv P\{X_1 \le t_1, X_2 \le t_2, \ldots, X_i \le t_i, \ldots, X_k \le t_i \mid \text{k arrivals in } [0, t_N]\} \qquad \text{for } k \ge i$$

This is the conditional probability that the first i arrival times obey the departure time inequalities and that the next k-i arrival times are also less than t_i, given that there are precisely k arrivals in $[0, t_N]$. In this notation the key quantity of interest is

$$P\{E(\underline{t}) \mid N(T) = N\} = a_{NN}(\underline{t}).$$

To calculate $a_{ki}(\underline{t})$ first note that

$$a_{k1}(\underline{t}) = \left(t_1 / t_N\right)^k \tag{5}$$

The fundamental recursion is given by

Lemma 2.

$$a_{ki}(\underline{t}) = \sum_{j=0}^{k-i+1} \binom{k}{j} a_{(k-j)(i-1)}(\underline{t}) \left(\frac{t_i - t_{i-1}}{t_N}\right)^j, \; k \ge i \tag{6}$$

Proof: The proof is similar in nature to that of Lemma 1 and will not be reported here.

To compute Eq. (6) iteratively one is filling out a lower triangular matrix $\underline{A}(\underline{t}) \equiv (a_{ki}(\underline{t}))$, including terms on the diagonal. One first uses Eq. (5) to compute all N entries of the first column of $\underline{A}(\underline{t})$. Then to compute the k^{th} entry ($k \geqq 2$) in the second column, one adds k terms, the j^{th} involving a multiplication with entry $(k-j+1)$ in the first column. In this way, one sweeps through the matrix column by column, starting in column one. The number of separate terms that have to be computed to complete the matrix is equal to

$$\sum_{i=1}^{N} \frac{i(i+1)}{2} = \frac{1}{6} N^3 + \frac{1}{2} N^2 + \frac{1}{3} N,$$

hence yielding an $0(N^3)$ procedure for evaluating $P\{E(\underline{t}) \mid N(T) = N\}$.

2. Computing Arrival Time Cumulative Probabilities

Consider now the "arrival time cumulative probabilities,"

$$\beta_{ki}(\underline{t}) \equiv \Pr\left\{X_k \leq t_i \mid E(\underline{t}), N(T) = N\right\}.$$

In words, $\beta_{ki}(\underline{t})$ is the conditional probability that the k^{th} arrival in $[0, t_N = T]$ occurs before t_i, given that all N arrivals obey the inequalities imposed by the observed service completion times. For instance, $\beta_{42}(\underline{t})$ is the probability that the fourth ordered arrival time in [0,T] is less than or equal to t_2, given by the conditioning event $E(\underline{t})$ that it must be less than or equal to t_4 (and, of course, given $E(\underline{t})$). Clearly

$$\beta_{ki}(\underline{t}) = 1 \text{ for all } k = 1, 2, \ldots, i.$$

There are two alternative methods for computing the matrix $\underline{\beta} \equiv (\beta_{ki}(\underline{t}))$, depending on whether one uses Lemma 1 or Lemma 2 for the fundamental recursions. In the context of Lemma 1, for $k > i$, we compute the arrival time cumulative probabilities as follows:

$$\beta_{ki}(\underline{t}) = \Pr\left\{X_k \le t_i | E(\underline{t})\right\} = \frac{\Pr\{X_1 \le t_1, \ldots, X_i \le t_i, X_{i+1} \le t_i, \ldots, X_k \le t_i, X_{k+1} \le t_{k+1}, \ldots\}}{\Psi_N(t_1, t_2, \ldots, t_N | N(T) = N)}$$

or

$$\beta_{ki}(\underline{t}) = \frac{\Psi_N(t_1, t_2, \ldots, t_i, t_i, \ldots, t_i, t_{k+1}, \ldots, t_N | N(T) = N)}{\Psi_N(t_1, t_2, \ldots, t_N | N(T) = N)} \tag{7}$$

With Lemma 2 the notation for determining $\underline{\beta} = (\beta_{ki}(\underline{t}))$ is somewhat more complex, but the computational effort for large N is considerably less. First, it should be clear that the bottom row of $\underline{\beta}$ is obtained by a simple division,

$$\beta_{Ni}(\underline{t}) = \frac{a_{Ni}(\underline{t})}{a_{NN}(\underline{t})} \tag{8}$$

For the general term, write

$$\beta_{ki}(\underline{t}) = P\{X_k \le t_i \,|\, E(\underline{t}), N(T)\}$$

$$= \frac{P\{X_1 \le t_1, \ldots, X_i \le t_i, \ldots, X_k \le t_i, X_{k+1} \le t_{k+1}, \ldots, X_N \le t_N \,|\, N(T)\}}{P\{E(\underline{t}) \,|\, N(T)\}} \tag{9}$$

Recognizing the denominator as $a_{NN}(\underline{t})$, we can write

$$\beta_{ki}(\underline{t}) = \frac{1}{a_{NN}(\underline{t})} \Big\{ P\{X_1 \le t_1, \ldots, X_i \le t_i, \ldots, X_k \le t_i, X_{k+1} \le t_i, X_{k+2} \le t_{k+2}, \ldots, \,|\, N(T)\}$$

$$+ P\{X_1 \le t_1, \ldots, X_i \le t_i, \ldots, X_k \le t_i, t_i \le X_{k+1} \le t_{k+1}$$

$$t_i < X_{k+2} \le t_{k+2}, \ldots, t_i < X_N \le t_N \,|\, N(T)\} \Big\}$$

The first probability in the brackets, when divided by $a_{NN}(\underline{t})$, is seen to be $\beta_{(k+1)i}(\underline{t})$, thereby giving rise to a recursion. To compute the second term, consider the ( ) ways of assigning k of the N unordered arrival times to the "left-hand" interval $[0, t_i]$ and the remaining (N-k) to the "right-hand" interval (t_i, t_N). Those assigned to the left would have to obey the first k in equalities in the second probability term above, while those assigned to the right would have to obey the final N-k inequalities. Invoking independence of the unordered arrival times, we can now write the recursion

$$\beta_{ki}(\underline{t}) = \beta_{(k+1)i}(\underline{t}) + \binom{N}{k} a_{ki}(\underline{t})\,\Gamma_{ki}(\underline{t}) / a_{NN}(\underline{t}) \tag{10}$$

where

$$\Gamma_{ki}(\underline{t}) \equiv P\{t_i < \ell_{N-k}(\underline{U}_{N-k}) \le t_{k+1}, \ldots, t_i < \ell_1(\underline{U}_{N-k}) \le t_N \mid N-k \text{ arrivals in } [0, t_N]\} \tag{11}$$

If we define the time-shifted vector $\underline{t}' \equiv (t_j')$,

$$t_j' = \begin{cases} t_{k+j} - t_i & \text{for } j = 1, 2, \ldots, N-k \\ t_N & j = N-k+1, \ldots, N \end{cases}$$

and invoke the uniformity property of the unordered arrival times, (11) can be computed using the algorithm for computing $a_{ki}(\cdot)$ as follows,

$$\Gamma_{ki}(\underline{t}) = a_{(N-k)(N-k)}(\underline{t}') \tag{12}$$

Hence, computation of $\beta_{ki}(\underline{t})$ using this technique requires $0((N-k)^3)$ new computations (i.e., to evaluate Eq. (12)). The worst case performance of the entire algorithm, i.e., to compute the entire matrix $\underline{\beta} = (\beta_{ki}(\underline{t}))$, is $0(N^5)$, although the

occurrence of numerous "near-zero" probability events for large N results in much better performance in practice.

3. The Mean Cumulative Number of Arrivals at Time t

We now wish to compute

$\bar{N}_a(t) \equiv$ the expected cumulative number of arrivals to the system up to and including time t, given E(t).

This is the quantity analogous to E[N(t)] displayed in Equation (1) (a) for unconditioned order statistics. To avoid counting ambiguities we assume in Lemma 3 a strict ordering of the t_i's: $0 < t_1 < t_2 < \ldots < t_N$. The generalization to nonstrict inequalities is straightforward and will not be stated here.

Lemma 3

(i) $$\bar{N}_a(t_j) = \sum_{k=1}^{N} \beta_{kj}(\underline{t}) \qquad \text{for all } j = 1, 2, \ldots, N \tag{13}$$

(ii) Define $t_0 \equiv 0$. For $t_{j-1} < t \leq t_j$, $j = 1, 2, \ldots, N$,

$$\bar{N}_a(t) = \frac{t_j - t}{t_j - t_{j-1}} \bar{N}_a(t_{j-1}) + \frac{t - t_{j-1}}{t_j - t_{j-1}} \bar{N}_a(t_j) \tag{14}$$

Remark: (i) states that the expected cumulative number of arrivals up to and including time t_j is equal to a simple sum of arrival time cumulative probabilities. (ii) states that $\bar{N}_a(t)$ grows linearly during any time interval between two successive conditioning times t_{j-1} and t_j.

Proof:

$$\text{(i)}\quad \overline{N}_a(t_j) = \sum_{k=1}^{N} k \;\Pr\left\{\text{exactly k arrivals in } [0, t_j] \,|\, E(\underline{t}), N(T) = N\right\}$$

$$= \sum_{k=1}^{N} k \left(\Pr\{\text{at least k arrivals in } [0, t_j] \,|\, E(\underline{t}), N(T) = N\}\right.$$

$$\left. - \Pr\{\text{at least k+1 arrivals in } [0, t_j] \,|\, E(\underline{t}), N(T) = N\}\right)$$

$$= \sum_{k-1}^{N} k \left(\beta_{kj}(\underline{t}) - \beta_{(k+1)j}(\underline{t})\right),$$

where

$$\beta_{(N+1)j}(\underline{t}) = 0 \qquad \text{for all } j = 1, 2, \ldots, N.$$

But the last expression is

$$\overline{N}_a(t_j) = 1 \cdot \left(\beta_{1j}(\underline{t}) - \beta_{2j}(\underline{t})\right) + 2 \cdot \left(\beta_{2j}(\underline{t}) - \beta_{3j}(\underline{t})\right) + 3 \cdot \left(\beta_{3j}(\underline{t}) - \beta_{4j}(\underline{t})\right) + \ldots$$

which simplifies to Equation (13).

(ii) Suppose $N_a(t_{j-1}) = \ell$ and $N_a(t_j) = \ell + m$, $m \geqq 0$. Then over $(t_{j-1}, t_j]$ we have m random variables that are conditionally independent, uniformly distributed, the m "unordered arrival times" over $(t_{j-1}, t_j]$, where the expected value of the cumulative number of arrivals through time t, $t_{j-1} < t \leqq t_j$, grows linearly with t (with zero growth, of course, for the case $m = 0$). Thus,

$$\overline{N}_a(t \,|\, N_a(t_{j-1}) = \ell \text{ and } N_a(t_j) = \ell + m) = \ell + \frac{m}{t_j - t_{j-1}}\,(t - t_{j-1}).$$

Unconditioning first on $N_a(t_{j-1})$,

$$\overline{N}_a(t \,|\, N_a(t_j) - N_a(t_{j-1}) = m) = \overline{N}_a(t_{j-1}) + \frac{m}{t_j - t_{j-1}}\,(t - t_{j-1}).$$

Then unconditioning on $N_a(t_j)$,

$$\bar{N}_a(t) = \bar{N}_a(t_{j-1}) + \frac{\left[\bar{N}_a(t_j) - \bar{N}(t_{j-1})\right](t - t_{j-1})}{t_j - t_{j-1}}$$

which simplifies to Equation (14).

As a final interesting property regarding $\bar{N}_a(t)$, we have

Lemma 4: For $t \geqq 0$, $\bar{N}_a(t)$ is a concave function of t.

Proof See Appendix I.

Note that none of the results of this section depend on the value of the Poisson rate parameter λ.

4. Numerical Example

To illustrate the mechanics, we solve using Lemma 1 a simple N = 3 example with $t_1 = 1/3$, $t_2 = 2/3$ and $t_3 = T = 1$. These data correspond to a queueing system for which (1) a congestion period commences at time t = 0; (2) departures followed immediately by service initiations occur at t_1, t_2, and t_3; and (3) the departure occurring sometime later at time t_4 is not followed immediately by a service initiation, thereby signaling the end of the congestion period. Hence, a queue existed at least at times t_1-, t_2-, and t_3-.

First we compute from (3) and (4)

$$P\left\{E(\underline{t}) \mid N(T) = N\right\} = P\left\{E\left(\frac{1}{3}, \frac{2}{3}, 1\right) \mid N(1) = 3\right\} = \Psi_3\left(\frac{1}{3}, \frac{2}{3}, 1 \mid N(1) = 3\right)$$

$$= \left(\frac{1}{3}\right)^3 + \binom{3}{2}\left(\frac{1}{3}\right)^2 \frac{2}{3}\, \Psi_1\left(1 - \frac{1}{3} \mid N\left(\frac{2}{3}\right) = 1\right) + \binom{3}{1}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2 \Psi_2\left(\frac{1}{3}, \frac{2}{3} \mid N\left(\frac{2}{3}\right) = 2\right)$$

Clearly

$$\Psi_1\left(1 - \frac{1}{3} \mid N\left(1 - \frac{1}{3}\right) = 1\right) = 1$$

$$\Psi_2\left(\frac{1}{3}, \frac{2}{3} \mid N\left(\frac{2}{3}\right) = 2\right) = \left(\frac{1/3}{2/3}\right)^2 + 2\left(\frac{1/3}{2/3}\right)\frac{1/3}{2/3}\,\frac{1/3}{1/3}$$

Combining results, we obtain

$$\Psi_3\left(\frac{1}{3}, \frac{2}{3}, 1 \mid N(1) = 3\right) = \frac{16}{27}$$

This is the a priori probability that the arrival times, given 3 arrivals over [0, 1], obey the departure time inequalities. Now we wish to obtain the matrix of arrival time cumulative probabilities,

$$\underline{\beta}(\underline{t}) = \begin{bmatrix} 1 & 1 & 1 \\ \beta_{21}(\underline{t}) & 1 & 1 \\ \beta_{31}(\underline{t}) & \beta_{32}(\underline{t}) & 1 \end{bmatrix}$$

We illustrate by computing the most complicated entry,

$$\beta_{32}(\underline{t}) = \Pr\left\{X_3 \le t_2 \mid E(\underline{t}), N(T) = 3\right\} = \frac{\Psi_3(1/3, 2/3, 2/3 \mid N(1)=3)}{16/27}$$

$$= \frac{27}{16}\left[\left(\frac{1}{3}\right)^3 + \binom{3}{2}\left(\frac{1}{3}\right)^2\frac{2}{3}\cdot\ \frac{1}{2} + \binom{3}{1}\frac{1}{3}\left(\frac{2}{3}\right)^2\Psi_2\left(\frac{1}{3}, \frac{1}{3} \mid N\left(\frac{2}{3}\right)=2\right)\right].$$

But

$$\Psi_2 = \left(\frac{1}{3}, \frac{1}{3} \mid N\left(\frac{2}{3}\right)=2\right) = \left(\frac{1}{2}\right)^2, \quad \text{thus} \quad \beta_{32}(t) = \frac{7}{16}.$$

The complete matrix, together with the column sums representing mean cumulative number of arrivals, is given by

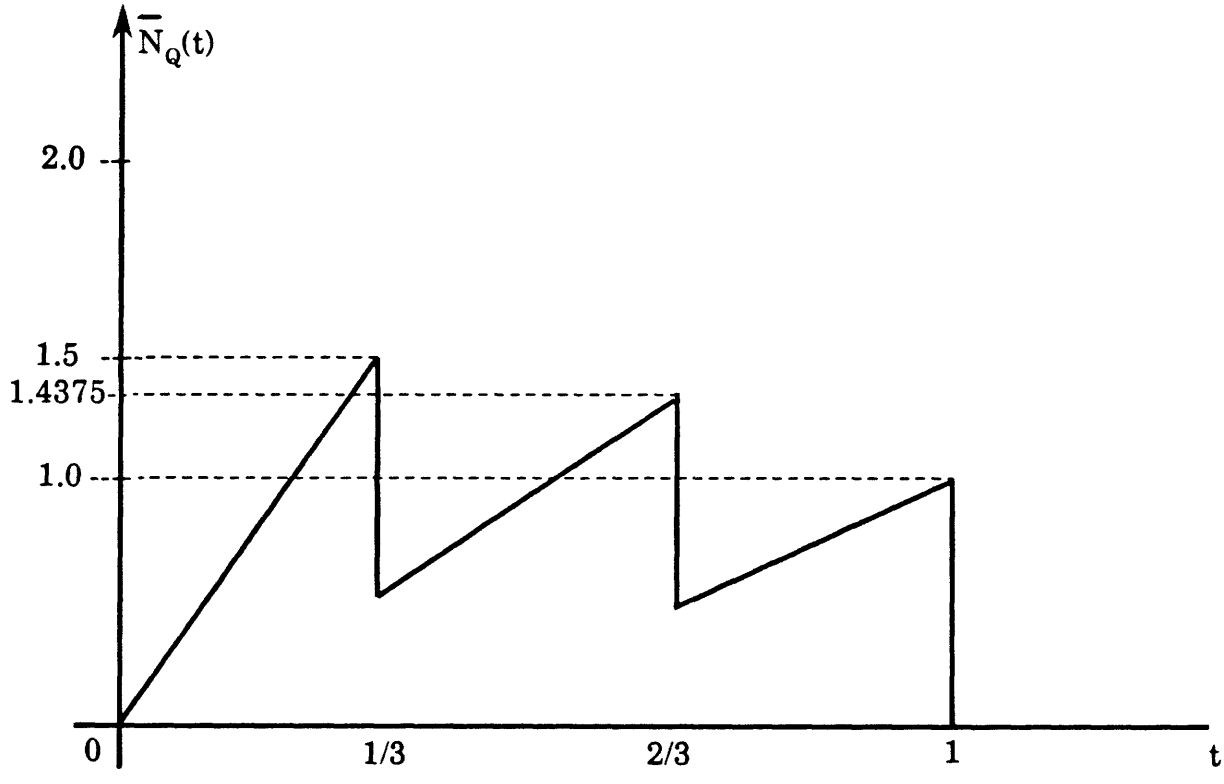
$$\underline{\beta}(\underline{t}) = \begin{bmatrix} 1 & 1 & 1 \\ 7/16 & 1 & 1 \\ 1/16 & 7/16 & 1 \end{bmatrix}$$

$$\overline{N}_a: \quad 1.5 \qquad 39/16 \qquad 3$$

Finally, using (6) and (7) the mean queue length as a function of time is displayed in Figure 2.

Figure 2

Mean Queue Length as a Function of Time for N = 3 Numerical Example



5. Expected Queue Length

Letting $\bar{N}_Q$ represent the time average queue length over a congestion period of length T, we have

$$\bar{N}_Q = \frac{1}{T} E\left[\int_0^T N_Q(t)\, dt \right] = \frac{1}{T} \int_0^T \bar{N}_Q(t)\, dt.$$

Since $\bar{N}_Q(t)$ is piecewise linear, with drops of magnitude one at t_i ($i = 1, 2, ..., N$), we can easily evaluate N_Q as follows (defining $t_0 \equiv 0$):

$$\bar{N}_Q = \frac{1}{2T} \sum_{i=1}^{N} \left(t_i - t_{i-1}\right)\left[\bar{N}_Q\left(t_i-\right) + \bar{N}_Q\left(t_{i-1}+\right)\right] \tag{15}$$

Example. Drawing from our continuing N = 3 example,

$$\bar{N}_Q = \frac{1}{2}\left(\frac{1}{3}\right)\left[1.5 + \left(1.4375 + 0.5\right) + \left(1.0 + 0.4375\right)\right] = 0.8125$$

Note that $\bar{N}_Q$ is the time average queue length during the congestion period for which the departure instants are known; $\bar{N}_Q$ is not the average queue length observed by a random customer arriving during the congestion period, because the conditioning information removes the Poisson arrival assumption (!).

To find the time average queue length over larger time intervals, including multiple congestion and uncongestion periods, one simply computes appropriate (time) weighted averages.

It is well-known that Poisson arrivals see time averages [Wolff, 1981]. Assuming that the queueing system is ergodic (which would be true for instance if each congestion period is governed by the same probability laws) our computations for N_Q and incidence probabilities (see Section IV.7) when averaged over many congestion periods would approach time averages.

6. Mean Delay in Queue

The expected total number of minutes spent in queue by customers during a congestion period is

$$E\left[\int_0^T N_Q(t)dt\right] = \int_0^T \bar{N}_Q(t)\,dt = T\,\bar{N}_Q.$$

Since there are N customers arriving during the congestion period, the average amount of time spent in queue per customer is

$$E\left[W_Q\right] \equiv \overline{W}_Q = \frac{1}{N}\int_0^T \overline{N}_Q(t)\,dt = \left(\frac{T}{N}\right)\overline{N}_Q \qquad (16)$$

Since N customers arrive (depart) during the period (0, T), the quantity (N/T) is the average arrival (departure) rate of customers during the congestion period. Equation (16), when rewritten

$$\overline{N}_Q = \left(\frac{N}{T}\right)\overline{W}_Q$$

is equivalent to Little's formula $L_Q = \lambda W_Q$ [Little, 1961]. In our running numerical example, $\overline{W}_Q = 0.2708$.

7. Incidence Probabilities

In this section we wish to compute the probability distribution of the queue length upon arrival of a random customer during a congestion period. Since the congestion period commences and terminates with zero customers in queue, we use the observation that for each queue length transition from i to i + 1 during the congestion period there must be a transition from i + 1 to i (i = 0, 1, 2, ...). If we define

$$\Pi_k \equiv \text{Prob \{a randomly arriving customer finds k customers in queue\}},$$

$$k = 0, 1, 2, \ldots$$

then, Π_k can be found by computing the probability that a randomly *departing* customer leaves behind k customers in queue. (This is a familiar argument found in the analysis of M/G/1 queues. [cf. Kleinrock [1975]] .)

We can write

$$\Pi_k = \frac{1}{N}\sum_{j=1}^{N} \text{Prob}\{j^{th} \text{ departing customer leaves behind k in queue}\}$$

$$= \frac{1}{N}\sum_{j=1}^{N} \text{Prob}\{\text{exactly } j+k \text{ arrivals in } [0,t_j]\}$$

$$= \frac{1}{N}\sum_{j=1}^{N} \left[\text{Prob}\{\text{at least } j+k \text{ arrivals in } [0,T] - \text{Prob}\{\text{at least } j+k+1 \text{ arrivals in } [0,T]\right]$$

or,

$$\Pi_k = \frac{1}{N}\sum_{j=1}^{N} \left(\beta_{(j+k)j}(\underline{t}) - \beta_{(j+k+1)j}(\underline{t})\right). \tag{17}$$

For our continuing numerical example, we find the following:

$$\Pi_1 = \frac{1}{3}\left(\frac{7}{16} - \frac{1}{16} + \frac{7}{16}\right) = \frac{13}{48} \approx 0.271$$

$$\Pi_2 = \frac{1}{3}\,\frac{1}{16} = \frac{1}{48} \approx 0.021$$

$$\Pi_0 = 1 - (\Pi_1 + \Pi_2) = \frac{34}{48} \approx 0.708$$

or

$$\underline{\Pi} = \left(\frac{34}{48}, \frac{13}{48}, \frac{1}{48}\right) \approx (0.708, 0.271, 0.021)$$

The average queue length experienced by an arriving customer, call it $\bar{\ell}_Q$, is

$$\bar{\ell}_Q = 0\cdot\ \frac{34}{48} + 1\cdot\ \frac{13}{48} + 2\cdot\ \frac{1}{48} = \frac{7}{12,} \approx 0.3125$$

in this case considerably less that the time average queue length $\bar{L}_Q = 0.8125$.

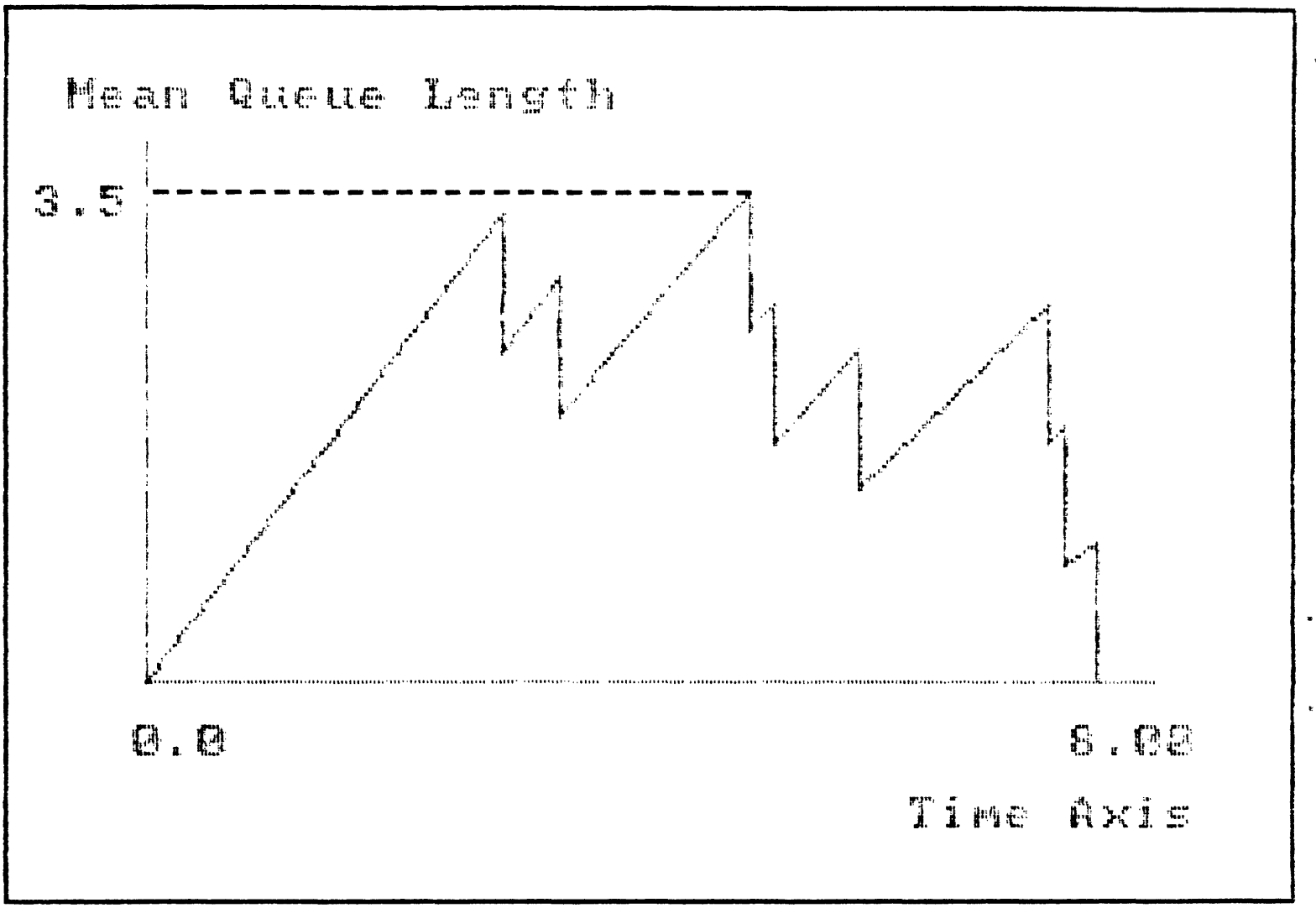
We have developed a computer program that carries out all of the computations of this paper, including plotting $\overline{N}_Q(t)$. We show in Figure 3 $\overline{N}_Q(t)$ for a congestion period having $N = 8$ simultaneous departures and service initiations as follows:

Congestion period starts at $t = 0$

$$t_1 = 3.0$$
$$t_2 = 3.5$$
$$t_3 = 5.1$$
$$t_4 = 5.3$$
$$t_5 = 6.0$$
$$t_6 = 7.6$$
$$t_7 = 7.75$$
$$t_8 = 8.0$$

The key statistics for this example are displayed in Table 1.

As a final example, we plot in Figure 4 $\overline{N}_Q(t)$ for an example having $N = 29$.



$\overline{N}_Q(t)$ for N = 8 Example

Figure 3

Matrix of the Betas

1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.9485	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.7299	0.8647	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.4220	0.5940	0.9816	1.0000	1.0000	1.0000	1.0000	1.0000
0.1696	0.2923	0.8143	0.8718	1.0000	1.0000	1.0000	1.0000
0.0446	0.0957	0.4868	0.5569	0.7984	1.0000	1.0000	1.0000
0.0070	0.0188	0.1870	0.2321	0.4445	0.9789	1.0000	1.0000
0.0005	0.0017	0.0338	0.0459	0.1227	0.7184	0.8209	1.0000

Cumulative Expected Number of Customers

3.3222	3.8673	5.5035	5.7067	6.3656	7.6972	7.8209	8.0000

Incidence Probabilities

Π_0	Π_1	Π_2	Π_3	Π_4	Π_5	Π_6	Π_7
0.2169	0.3009	0.2877	0.1322	0.0501	0.0111	0.0010	0.0001

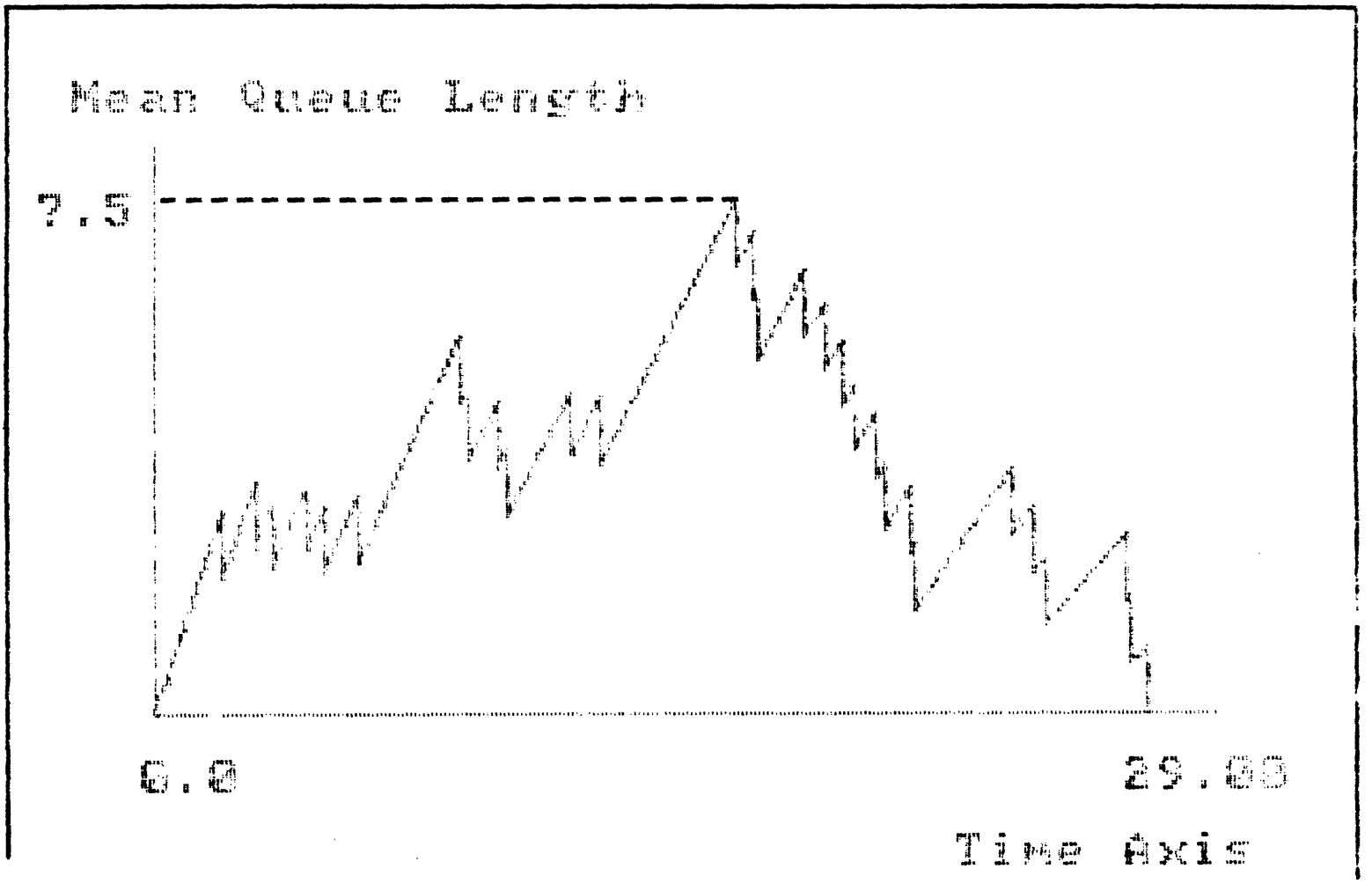
* Average Number of Customers in the Queue
as seen by a randomly arriving customer = 1.5354

* Time Average Number of Customers in the Queue = 2.0332

* Average Waiting Time for Customers in the Queue = 2.0332

Table 1

Detailed Statistics for N = 8 Example



$\bar{N}_Q(t)$ fo N = 29 Example

Figure 4

V. Lost Customers: Balking and Reneging

Using the transactional data with the (unconditional) Poisson arrival assumption, it is possible to estimate the number of customers who choose not to enter the system because the system is too congested at their time of arrival. Such prospective customers who do not even join the queue are said to balk due to congestion; others who join but later depart before entering service are said to renege. With transactional data , we are unable to distinguish between the two types of behavior.

Suppose there are M noncongestion periods, each terminated by a "first arriving customer" who initiates a congestion period. This customer enters service immediately. All others arriving during that congestion period are delayed in queue. For the i^{th} such congestion period, let the time from its commencement until the first departure of a serviced customer be τ_i. Under the Poisson arrival hypothesis, with no balking or reneging, the probability that a queue will form during $[0, \tau_i]$ is $1\text{-}e^{-\lambda\tau i}$, where λ is the Poisson rate parameter. The expected number of the M congestion periods that will be accompanied by queueing is

$$\overline{Q}\,(M) = \sum_{i=1}^{M} \left(1 - e^{-\lambda\tau_i}\right). \tag{18}$$

(When allowing the possibility of balking and reneging, the parameter λ should be estimated from the average value of the duration of the noncongestion period [which should equal λ^{-1}], during which no balking or reneging can occur.)

Suppose from the M congestion periods one observes q(M) congestion periods having queues. Suppose $q(M) < Q(M)$. Then one could perform various statistical tests to determine if the difference is statistically significant, and if it is, one could

reasonably conclude that congestion causes the measured amount of balking and/or reneging.

Suppose the balking/reneging is of a simple binary type. With probability p a Poisson arrival representing a potential customer will balk/renege whenever any congestion (i.e., delay) is to be experienced, regardless of the queue length. Then the queueing system would be driven by an alternating Poisson process, with rate parameter λ during noncongestion periods and $\lambda(1-p)$ during congestion periods. If only this simple type of balking/reneging occurs then all the results represented by Equations (2)-(17) remain valid. However, if more complicated state-dependent balking/reneging occurs, then since (unconditioned) arrivals during congestion periods are no longer Poisson, Equations (2)-(17) are not valid.

To test for the presence of state-dependent balking/reneging, one can obtain expressions using the transactional data for the number of congestion periods having length *greater* than k customers, given length *at least equal to* k, for $k = 2, 3, 4, \ldots$, under the hypothesis of no balking/reneging. One can than compare with the data to discover the extent of state-dependent balking/reneging.

As an example consider a congestion period having at least 2 customers, with τ_i defined as above and $\tau_i^{[1]}$ the time between the first and second service completions during the i^{th} congestion period. Note that only the second customer is delayed in queue. Then we can write

$$\Pr\{\text{only one arrival in } (0, \tau_i + \tau_i^{[1]}) \mid \text{at least one arrival in } (0, \tau_i)\} =$$

$$\Pr\{\text{only one arrival in } (0, \tau_i) \mid \text{at least one in } (0, \tau_i)\} \cdot \Pr\{0 \text{ arrival in } (\tau_i, \tau_i + \tau_i^{[1]})\}$$

$$= \frac{\lambda \tau_i e^{-\lambda \tau_i}}{1 - e^{-\lambda \tau_i}} e^{-\lambda \tau_i^{[1]}}$$

If there are M_1 congestion periods having at least two customers, the second of whom is delayed in queue, then under the Poisson-arrival-no-balking/reneging hypothesis, the expected number of congestion periods having *more than* two customers is

$$\bar{Q}_1(M_1) = \sum_{i=1}^{M_1} \left(1 - \frac{\lambda \tau_i e^{-\lambda(\tau_i + \tau_i^{[1]})}}{1 - e^{-\lambda \tau_i}}\right) \tag{19}$$

A similar line of reasoning can be continued to higher levels of congestion. When comparing with data, successive differences between theoretical and observed values reveal estimates of congestion-related balking and/or reneging. If the levels of balking/reneging are significant, then as stated previously Equations (2)-(17) are no longer valid. Further research is required to develop accurate queue estimation methods from transactional data in the presence of congestion-level-dependent balking/reneging.

VI. Summary and Conclusions

In this paper we have shown how to apply ideas of order statistics to deduce the behavior of Poisson-arrival queues without observing them. We simply use transactional data (i.e., times of service commencement and service completion) for each customer together with the Poisson assumption to derive time-dependent mean number in queue, mean wait in queue and the probability distribution of the number of customers in queue upon arrival of a random customer. Using the same ideas, additional performance measures could be devised if desired. The paper concluded with a proposed methodology to determine whether customers are balking and/or reneging during periods of congestion, again using only transactional data and the assumption that potential arrivals to the queue occur according to a Poisson process.

With the exception of the balking/reneging results, none of our formulas contain the rate parameter λ of the Poisson process. This is because the total number of (Poisson) arrivals over a congestion period is given as part of the conditioning information. Thus our results could be averaged over congestion periods occurring during times of different Poisson rate intensities. In fact, λ could be a slowly varying function of time, $\lambda(t)$, and our results would be approximately correct, so long as $\lambda(t)$ does not "change very much" over any congestion period.

A limitation in implementing the methods proposed herein is that evaluation of the matrix $\underline{\beta}$ requires $0(N^5)$ computations for a congestion period having N arrivals. Clearly this is not practical for very large N. However, with today's computers, such calculations are feasible certainly for $N \leqq 50$ and probably for $N \leqq 100$. As a bench mark, the average number of customers who queue in an M/G/1 system during a period of congestion is $\rho/(1-\rho)$ (where $\rho = \lambda$ E[service time]), which is less than 10 for $\rho < 0.9$ (Kleinrock [1975], p. 217). So for many important applications the fifth

order growth in computational work with N should not be an impediment to implementation. For more saturated systems, we may seek approximations or limiting results.

Acknowledgments

This research was supported by the National Science Foundation under Grant #SES 8709811. I thank Christopher Athaide for excellent research assistance, both in computer programming and in suggesting Lemma 2. I also thank for comments on an earlier draft, A. Barnett, S. Graves, and A. Odoni.

References

(1) Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D., Statistical Inference Under Order Restriction, John Wiley and Sons, New York, 1972.

(2) David, H.A., Order Statistics, John Wiley and Sons, New York, 1981.

(3) Kleinrock, L., Queueing Systems, Volumes 1 and 2, John Wiley and Sons, New York, 1975.

(4) Little, J.D.C., "A Proof of the Queueing Formula $L = \lambda W$," Operations Research, Vol. 9, pp. 383-387, 1961.

(5) Wolff, R.W., "Poisson Arrivals See Time Averages," Operations Research, Vol. 30, pp. 223-231, 1987.

Appendix I

Lemma 4: For $t \geq 0$, $N_a(t)$ is a concave function of t.

Proof From Lemma 1 we know that $N_a(t)$ is piecewise linear, continuous, monotone non-decreasing. We first prove the theorem for $N=2$, then for general N. For $x_2 > x_1$ define the "truncated ramp function"

$$\ell\left(t; x_1, x_2\right) \equiv \begin{cases} 0 & \text{for } t \leq x_1 \\ (t-x_1)/(x_2-x_2) & \text{for } x_1 < t \leq x_2 \\ 1 & \text{for } t > x_2 \end{cases}$$

Without loss of generality we can assume that the N time-conditioned arrivals occur in [0,1]. Define

$N_a(t|\Gamma) \equiv$ mean number of arrivals in [0,t], given event Γ.

$N=2$. Let the unordered arrival times be U_1, U_2. The time conditioning information is MIN $[U_1, U_2] \leq t_1$ where $0 < t_1 < 1$, and MAX $[U_1, U_2] \leq 1$. Without time conditioning, call that event A, U_1 and U_2 are i.i.d., uniformly over [0,1] and $N_a(t|A) = 2t$, $0 \leq t \leq 1$. Hence, given A, one can write

$$2t = p_1\, 2\ell(t; 0, t_1) + p_2\, 2\ell(t; t_1, 1) + p_3\,[\ell(t; 0, t_1) + \ell(t; t_1, 1)],$$

where $p_1 > 0, p_2 > 0, p_3 > 0$ represent probabilities that the two unordered (unconditioned) arrival times are (1) both in [0,t]; (2) both in $(t_1,1]$; and (3) such that one is in $[0, t_1]$ and the other is in $(t_1, 1]$. But the time conditioning information excludes possibility (2), implying that

$$\overline{N}_a(t) = \frac{p_1}{1-p_2}\, 2\ell\left(t; 0, t_1\right) + \frac{p_3}{1-p_2}\left[\ell\left(t; 0, t_1\right) + \ell\left(t; t_1, 1\right)\right]$$

Since $p_2 > 0$, we must have at $t = t_1$,

$$\bar{N}_a(t_1) > p_1\, 2\ell\left(t_1;\, 0, t_1\right) + p_3\, \ell\left(t_1;\, 0, t_1\right) = 2t_1,$$

implying $N_a(t)$ is concave.

<u>Arbitrary N.</u> (contradiction) If $N_a(t)$ is not concave they there must exist at least one k for which

$$\bar{N}\left(t_k\right) < \bar{N}\left(t_{k-1}\right) + \left[\bar{N}\left(t_{k+1}\right) - \bar{N}\left(t_{k-1}\right)\right] t_k / \left[t_{k+1} - t_{k-1}\right],$$

where $\underline{t} \equiv (t_i)$ is the vector of conditioning times such that the i^{th} smallest U_j must be less than or equal to t_i, where we assume $0 \equiv t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N \equiv 1$.

Expanding the logic shown for $N = 2$, we can write

$$\bar{N}_a(t) = \sum_{j=0}^{N-1} \sum_{i=1}^{N-j} i\,\ell\left(t; t_j; t_{j+1}\right) p_{ij} \tag{A1}$$

where the probabilities p_{ij} are conditionally multinomial. But (A1) can be written

$$\bar{N}_a(t) = \sum_{j=0}^{k-2} \sum_{i=1}^{N-j} i\,\ell\left(t; t_j; t_{j+1}\right) p_{ij}, + \sum_{j=k-1}^{k} \sum_{i=1}^{N-j} i\,\ell\left(t; t_j; t_{j+1}\right) p_{ij}$$

$$+ \sum_{j=k+1}^{N-1} \sum_{i=1}^{N-j} i\,\ell\left(t; t_j; t_{j+1}\right) p_{ij} \tag{A2}$$

For $t_{k-1} < t \leqq t_{k+1}$, the first term in (A2) contributes a positive constant to $N_a(t)$ and the third term contributes zero. Hence to determine concavity we focus on the second term and on the intervals $[t_{k-1}, t_k]$, $[t_k, t_{k+1}]$.

Suppose in any given realization of the process, we are given additional conditioning information that $N_a(t_{k-1}) = j$ for $j \geqq k$. Then of the remaining N-j time-conditioned arrivals, we may have any positive number (up to N-j) of them uniformly (conditionally) independently distributed over the joint interval $[t_{k-1}, t_{k+1}]$, with the remainder distributed appropriately (given the time conditioning information) over

$[t_{k+1}, 1]$. For each such possibility, for $t_{k-1} < t \leqq t_{k+1}$, the conditional contribution to $N_a(t)$ is a positively sloped straight line; probabilisticly weighting each possibility, the corresponding weighted sum of straight lines is a positively sloped straight line, a property that does not violate concavity.

Now focus on the conditioning information $N_a(t_{k-1}) = k\text{-}1$. Assume further (for the moment) that $N_a(t_{k+1}) = k\text{-}1+m$, i.e., m arrivals occur in $[t_{k-1}, t_{k+1}]$, for $m = 2, 3, ..., N\text{-}k+1$. If the m arrivals were uniformly independently distributed over $[t_{k-1}, t_{k+1}]$, then we could write for $t_{k-1} < t \leqq t_{k+1}$,

$$\bar{N}_a(t|\beta_m) = k-1 + \left[n/\left(t_{k+1} - t_{k-1}\right)\right]\left(t - t_{k-1}\right)$$

$$= k-1 + \sum_{i=0}^{n} p_i \left\{ i\,\ell\left(t; t_{k-1}\, t_k\right) + \left(n-i\right)\ell\left(t; t_k, t_{k+1}\right)\right\}$$

for appropriate conditional probabilities $p_i > 0$ ($i = 0, 1, ..., m$) and where $\beta_m \equiv$ event that k-1 time-conditioned arrivals are in $[0, t_{k-1}]$ and m arrivals are uniformly independently distributed in $[t_{k-1}, t_{k+1}]$. But considering $N_a(t)$, time- conditioning prohibits the event whose probability is p_0, i.e., the event having zero of the n arrivals in $(t_{k-1}, t_k]$. Let $\beta_n = \beta_n$ - {event that all n arrivals are in $(t_k, t_{k+1}]$}.

Then,

$$\bar{N}_a(t|\beta_n') = k-1 + \sum_{i=0}^{n} \frac{p_i}{1-p_0} \left\{ i\,\ell\left(t; t_{k-1}, t_k\right) + \left(n-i\right)\ell\left(t; t_k, t_{k+1}\right)\right\}$$

and at the "breakpoint" t_k we have

$$\bar{N}_a\left(t_k|\beta_n'\right) = k-1 + \sum_{i=0}^{n} \frac{ip_i}{1-p_0} > k-1 + \left[n/\left(t_{k+1} - t_{k-1}\right)\right]\left(t_k - t_{k-1}\right)$$

Hence for any n ($n = 2, 3, ..., N\text{-}k+1$) we have shown that $N_a(t|\beta_n)$ is concave over $[t_{k-1}, t_{k+1}]$. To complete the proof we multiply each $N_a(t|\beta_n)$ by the appropriate

probability, sum to obtain $N_a(t)$ over $[t_{k-1}, t_{k+1}]$, and use the fact that a sum of concave functions is concave.