The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2012-14
December 28, 2012

Predicting Gridlocks in Emergency Departments: A Bayesian Discrete Time Queueing Theoretic Approach

Toros Caglar
OptumRx
Irvine, CA, USA

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

Queueing Systems manuscript No.
(will be inserted by the editor)

Predicting Gridlocks in Emergency Departments: A Bayesian Discrete Time Queueing Theoretic Approach

Toros Caglar · Refik Soyer

Received: January 1, 2013 / Accepted: date

Abstract In this work, we develop a Bayesian framework for the analysis of discrete time queueing systems to predict emergency room gridlocks caused by blocked patients due to insufficient critical care beds in the hospital. Gridlocks are considered to be the major reason for ambulance diversions, affecting the timeliness and effectiveness of the healthcare service. In doing so, we investigate the Bayesian analysis of discrete time queueing networks with an emphasis on the queueing network formed by the emergency room and the hospital. We allow the parameters of the system to have a dependent structure based on a common environment via a Markov-modulated model and illustrate our methods via a numerical example.

Keywords Emergency room gridlocks · Discrete time queueing networks · Bayesian analysis · Markov modulated models

1 Introduction

The mission of emergency rooms is to provide timely emergency care to patients in need of medical attention. After attracting the public's attention more than a decade ago, overcrowding in emergency rooms (ER) has resurfaced as a

T. Caglar
OptumRx
2300 Main Street
Irvine, CA, 92614
Tel.: +949-252-4327
E-mail: toros.caglar@optum.com

R. Soyer
The George Washington University
2201 G Street NW
Funger Hall, Suite 415C
Washington, DC, 20052
E-mail: soyer@gwu.edu

healthcare crisis in the past few years. Unfortunately, this time overcrowding is even more widespread, and in some places accepted as the standard of care [9]. According to surveys by the American Hospital Association (AHA), over 50% of the 28 hospitals studied describe their ER to be running at or over capacity [18][19]. With patient waiting times averaging as high as 96 minutes, the ERs are rarely able to comply with the first clause of their mission, timeliness.

Patients with disparate needs and varying urgencies, randomly and without prior notice, arrive to the ER either on foot, or by an emergency medical system (EMS) transport unit. Due to the nature of the demand, peaks in the system occupancy are common, and it is a challenge for the ER to respond to patient needs in a timely manner. During such chaotic periods, ERs can decrease their loads significantly through diversion.

An ER on diversion suspends arrivals that can be controlled (i.e., non-walk-in patients) by forcing some or all of the EMS transport units to search for alternate treatment facilities for their patients [3]. During this search, the valuable time that the patient is losing, and the unavailability of the EMS crew to answer other calls, are just two of the problems that the ERs inability to maintain availability is responsible for. An investigation in Houston, Texas, has revealed results about the relationship between diversions and trauma mortality rates, which revealed the acuity of this problem. During a two-year period, the mortality rate of severely injured trauma patients was almost twice as high (25% vs. 14.4%) on days when both of Houstons Level I trauma centers were on diversion [4]. Unfortunately, such studies linking ambulance diversion to unsatisfactory healthcare performance are not numerous. Policymakers do not address diversion as a healthcare priority for a variety of reasons. The most obvious reason, according to [2], a physician at the Department of Emergency Medicine at Regions Hospital in St. Paul, Minnesota, is that the public did not ask them to do so. Even though diversion is a known problem, most people do not believe they will ever be diverted. The 2004 AHA survey results present an example contradicting this belief; during the three days the survey was conducted, one third of the 28 studied hospitals were on diversion more than 20% of the time. Lack of studies on linking poor ER performance and ambulance diversions is also a factor for the insufficient attention on ambulance diversion. When initially implemented, diversion was supposed to be a temporary mechanism that was rarely used. However, nowadays diversions have almost become standard operating procedures and need to be investigated.

AHA ambulance diversion surveys show lack of critical care (CC) beds in the hospital as the major cause of the diversions [18][19]. After being treated in the ER, the patient is either discharged (no further immediate treatment is necessary), or admitted to the hospital. However, admission may not be possible due to the lack of available CC beds in the hospital. In this case, the patient occupies a bed in the ER, waiting for a CC bed to become available. This condition may lead to gridlock, which may be defined as the state of an ER when (i) all beds in the ER are occupied, and (ii) there is at least one patient in one of the beds in the ER waiting to be transferred to the hospital, but the transfer is not possible due to the lack of available beds. When these

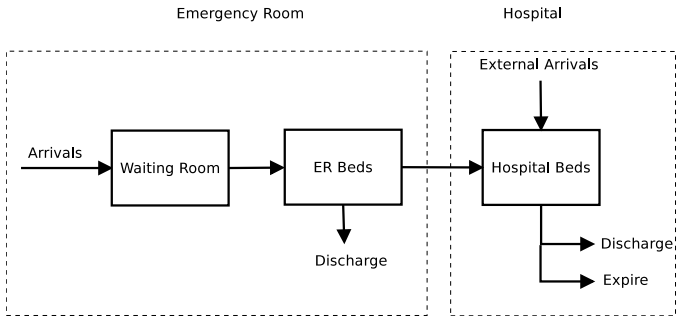

Fig. 1 Network structure between the emergency room and the hospital

conditions are reached, the idle patient that needs to be transferred is said to be on hold and is occupying a bed in the ER, preventing its use by other patients.

The objective of this research is to construct a predictive model that can provide a probability of entering gridlock within a time horizon, conditioned on the systems current state. We propose to reach our objective by developing a methodology for the Bayesian analysis of the ER-hospital queueing network.

2 Queueing network representation of the ER-Hospital relationship

The system in consideration is composed of an ER and the critical care beds in a hospital. In the most basic representation, the system is a network of two queues. Figure 1 illustrates the three main parts of the system. The patients randomly arrive to the ER. Depending on the availability of the beds and staff in the ER, the patient is either admitted for diagnosis and treatment, or sent to a waiting room. The waiting room serves as a queue for the ER beds. The patients in the waiting room are prioritized according to the severity of their condition, and they are admitted into the ER on a first come first served basis with respect to their priorities. After being treated in the ER, the patient is either discharged (no further immediate treatment is necessary), or admitted to the hospital. The second queue in the system is for hospital admissions. There are only a fixed number of critical care (CC) beds in the hospital. If a patient needs to be admitted and requires a CC bed, at least one CC bed in the hospital needs to be available, or the patient will be on hold until an occupied bed is freed. In this part of the system, the ER beds serve as a queue to the CC beds, and the CC beds are blocking the arrival stream from the ER.

The ER serves a time-dependent and multi-class arrival process. Upon entering, patients are classified according to the severity of their conditions. These classifications, called Emergency Severity Indexes (ESI), can have a number of levels depending on the healthcare provider. The most common are 3-level and 5-level ESI systems, where lower indexes denote higher severities.

Different classes of patients tend to have different arrival rates that also vary according to the time of the day and day of the week.

An ER can enter gridlock by either an ER arrival, or an ER end-of-service. According to our definition, for the system to be in gridlock, the hospital and ER beds must be occupied with at least one boarding patient (waiting for a hospital bed while occupying an ER bed). An ER arrival occupying the last bed can put the system into a gridlock if the system had full hospital beds and boarding patients in the ER. An ER end-of-service can put the system in gridlock if the hospital beds are full, the ER beds are full, and the end-of-service just resulted in a boarding patient in the ER.

3 Discrete time queues, their networks and Bayesian analysis

Applications of discrete time queues to continuous time problems are rare. [8] discuss the advantages of employing discrete time queues and point out that even though most systems behave continuously, our measurement and interpretation discretize these systems. We measure time as multiples of a convenient unit (e.g. seconds, minutes) and operational decisions are frequently made based on longer, discrete time blocks. In this application, we use ER arrival and service data to provide numerical examples for our methods. Healthcare operations in particular is a good candidate for discrete time analysis for several reasons. First, as pointed out earlier, system performance and characteristics are measured as multiples of a discrete time unit. Also, ER staff schedule is based on discrete time slots, which can be matched by the discrete time queueing model. Finally, arrival rates to the ER can change randomly, affected by traffic accidents, natural disasters or other outside factors. These random changes can be modeled through a Markov modulated version of these queues, which lend themselves more naturally to this analysis than their continuous counterparts. The complex probabilistic structure defining the ER-hospital queue can easily be modeled using our approach. The corresponding Bayesian analysis is also performed more easily under the discrete case. It is true that continuous models could be used to obtain predictions within discrete time blocks. However, their analysis is not any more convenient in such complex systems. Furthermore, gridlock prediction is the result of transient analysis, which is a lot more convenient to perform using the discrete time methods we present.

The foundations and the motivation behind Bayesian analysis of queues are covered very thoroughly in [13] and [14]. The main aspect that sets the Bayesian approach apart from the classical analysis of queues is the handling of the unknown parameters describing the system. As opposed to considering the parameters of the system describing the arrival and the service processes as fixed but unknown, the Bayesian approach describes them via probability distributions. This allows the usage of personal probabilities to describe the uncertainties about the parameters and coherently updating these uncertainties as more information is received. The probabilistic inference provided by

the Bayesian framework also allows one to attack decision making problems through applications of Bayesian utility theory. For example, system design problems deciding on number of servers and investments on increasing service rates, one can construct the appropriate utility measures and provide solutions maximizing expected utility. The Bayesian approach also allows the predictions of unobserved quantities to be made easily. These quantities are not limited to the future inter-arrival or service times, but also include performance measures of the queueing system such as the number of customers and waiting times. [1] derive closed form expressions for most performance measures of the M/M/1 queue. The authors then provide an analysis of the M/M/c queue in [1]. Although the preceding works focus on continuous time queues, the main ideas regarding the inferential statistics and the Bayesian methodologies transfer perfectly to the discrete case.

The Bayesian analysis of discrete time queues is very scarce in the literature. In [5,6], the author concentrates on non-parametric Bayesian analysis of $Geo/G/1$ queues, motivated by ATM (Asynchronous Transfer Mode) systems. Probability generating functions for the delay distributions are calculated and approximations used are shown to converge to true results under the large sample assumption. Such approximations are necessitated by the large buffer sizes the communication systems posses, which make exact computations difficult.

We continue this work by describing the queueing network representation of the ER-hospital relationship. We focus primarily on modeling the ER-hospital network which has a very specific structure that can not be modeled with the existing models in the literature without making significant assumptions. Instead, we turn to numerical results calculated by Markov chain based analysis and present the gridlock prediction technique to be used. We then move on to the Bayesian analytic methods to infer the homogeneous and Markov modulated network parameters. Finally, we cover gridlock prediction under the Bayesian framework as well as steady state network results, and provide a numerical example to illustrate this approach.

The discrete time queueing analysis of networks with inherent slotted time scales date back to mid 1960s due to [11,12], where the author utilized Markov chain based methodologies to analyze the systems. However, as a result of the complexities involved in analysis of these systems, and the convenience of continuous time approximations, the work on the discrete time queueing networks subsided until the introduction of ATM as the multiplexing technique for Broadband Integrated Services Digital Networks [7].

Applications of discrete time queues to inherently continuous systems are rare. First discussion of such applications appeared in [8] where the authors comment on the merits of discrete time models. As far as we know, in the discrete time queueing network literature, the applications have strictly been on computer and communications systems which behave in a discrete manner. In our application to the ER-hospital network, even though the timeline is continuous, in practice, problems of decision making require discretization of the timeline. For example shift scheduling problems will adhere to discrete time blocks. Also, as pointed out by [8], introducing probabilistic structures

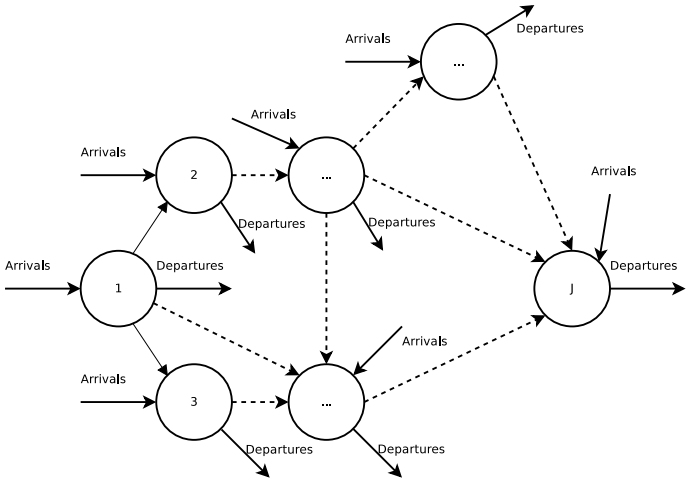

Fig. 2 General queueing network structure

to models are much less complicated in a discrete time setting. Blocking, balking, ER to hospital transfer probabilities and customer types are just some examples of structures that can easily be incorporated into a discrete time queueing network. Finally, gridlock prediction also makes most sense when modeled with a discrete time model since the communication of discrete time blocks is easier than dealing with continuous time measurements.

Discrete time queueing networks are constructed by connecting multiple queueing nodes in a network structure. Figure 2 illustrates a general queueing network structure formed by connecting J nodes, where each node is a service station and allowed its independent arrival stream. Customers entering a node are queued and served. Upon service completion, the customer can either move to one of the subsequent nodes, or depart the system. The customers routing behavior can be modeled probabilistically. The time-line is divided into discrete slots and the customers behave according to the late arrival scheme.

4 Modeling the ER-hospital network

For the ER-hospital interaction, we only need a two node network. However, we need to incorporate blocking into the system, which is not covered in the systems considered by [7] and summarized above. Let us start this discussion by considering the simplest, two node queueing network in discrete time, with state independent arrival and service, depicted in Figure 3. Outside arrival and departure streams to and from the nodes are allowed, however, for the initial model, blocking is ignored. Also, let the system have only one server per node as a starting point. Let λ_i and μ_i denote the arrival and service probabilities at node $i = 1, 2$, respectively. With probability p, customers whose service is completed in the first node are routed to the second node. With probability

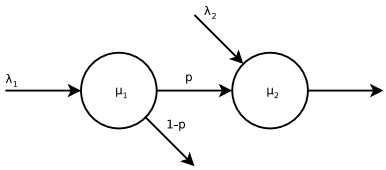

Fig. 3 A general two node network

$1-p$, they leave the system after their service completion. The transfer to the second node is instantaneous. Once a customer's service at the second node is completed, it leaves the system. We can describe the state of our network by keeping track of the number of customers in both of the nodes, (N_1, N_2), which creates a two dimensional state space with a discrete time Markov chain $\mathbf{Q}$ representing the state transition probabilities. Let $u^{(i)}$ denote an increase of i customers in a node. Similarly, let $d^{(i)}$ represent a decrease of i, and r represent neither an increase, nor a decrease in the number of customers in a node. For simplicity, also let $d^{(1)} = d$ and $u^{(1)} = u$. We will then represent most of the state transition probabilities by proper pairings of these changes in the number of customers in the two nodes of the network, conditioned on the initial state of the network. For example, the argument $(d, u^{(2)})_{(j,k)}$ represents the probability of a decrease of 1 customers in the first node and an increase of 2 customers in the second node, given that the first node had j customers and the second node had k customers. These probabilities are given in (1),(2),(3) and (4).

$$\begin{aligned} &\text{For } j = k = 0: \qquad (1)\\ &(r,r)_{(0,0)} = \bar{\lambda_1}\bar{\lambda_2}\\ &(u,r)_{(0,0)} = \lambda_1\bar{\lambda_2}\\ &(r,u)_{(0,0)} = \bar{\lambda_1}\lambda_2\\ &(u,u)_{(0,0)} = \lambda_1\lambda_2 \end{aligned}$$

When both nodes are initially empty ($j = k = 0$), the probabilities in (1) are obtained easily. There can't be a decrease in either of the nodes. So, each node either increases by one with the probability of its corresponding arrival process, or, they stay the same with the complement of its arrival probability.

For $j > 0$ and $k = 0$: (2)

$$
\begin{aligned}
(d, r)_{(j,0)} &= \bar{\lambda_1}\mu_1\bar{p}\bar{\lambda_2} \\
(r, r)_{(j,0)} &= \lambda_1\mu_1\bar{p}\bar{\lambda_2} + \bar{\lambda_1}\bar{\mu_1}\bar{\lambda_2} \\
(u, r)_{(j,0)} &= \lambda_1\bar{\mu_1}\bar{\lambda_2} \\
(d, u)_{(j,0)} &= \bar{\lambda_1}(\mu_1 p\bar{\lambda_2} + \mu_1\bar{p}\lambda_2) \\
(r, u)_{(j,0)} &= \lambda_1(\mu_1 p\bar{\lambda_2} + \mu_1\bar{p}\lambda_2) + \bar{\lambda_1}\bar{\mu_1}\lambda_2 \\
(u, u)_{(j,0)} &= \lambda_1\bar{\mu_1}\lambda_2 \\
(d, u^{(2)})_{(j,0)} &= \bar{\lambda_1}\mu_1 p\lambda_2 \\
(r, u^{(2)})_{(j,0)} &= \lambda_1\mu_1 p\lambda_2
\end{aligned}
$$

When $j > 0$ and $k = 0$, the first node can increase its number of customers only by an arrival with probability λ_1. However, there are two ways for the first node to go down by one customer. We either see a service completion that departs the system with probability $\mu_1\bar{p}$, or, a service completion that transfers to the second node with probability $\mu_1 p$. The second occurrence also is one of the two ways that could increase the number of customers in the second node. The other way is an external arrival with probability λ_2. The second node can also increase by two if there is both an external arrival and a transfer from the first node. The system will stay unchanged if the number of arrivals and departures are equal to each other in both of the nodes.

For $j = 0$ and $k > 0$: (3)

$$
\begin{aligned}
(r, d)_{(0,k)} &= \bar{\lambda_1}\bar{\lambda_2}\mu_2 \\
(u, d)_{(0,k)} &= \lambda_1\bar{\lambda_2}\mu_2 \\
(r, r)_{(0,k)} &= \bar{\lambda_1}(\bar{\lambda_2}\bar{\mu_2} + \lambda_2\mu_2) \\
(u, r)_{(0,k)} &= \lambda_1(\bar{\lambda_2}\bar{\mu_2} + \lambda_2\mu_2) \\
(r, u)_{(0,k)} &= \bar{\lambda_1}\lambda_2\bar{\mu_2} \\
(u, u)_{(0,k)} &= \bar{\lambda_1}\lambda_2\bar{\mu_2}
\end{aligned}
$$

When $j = 0$ and $k > 0$, we can only have arrivals to the first node with probability λ_1. Since there are no customers in service in node one, the only way for node two to increase by a customer is with an external arrival with probability λ_2. The number of customers in node two can also decrease by one if there is a service completion, which occurs in a time slot with probability μ_2. Once again the system remains unchanged if the number of arrivals and departures are identical in both of the nodes. The probabilities in (4) for the $j > 0, k > 0$ case can be similarly calculated with the information already given.

For $j > 0$ and $k > 0$: (4)

$$
\begin{aligned}
(d,d)_{(j,k)} &= \bar{\lambda_1}\mu_1\bar{p}\bar{\lambda_2}\mu_2 \\
(r,d)_{(j,k)} &= \bar{\lambda_1}\bar{\mu_1}\bar{\lambda_2}\mu_2 + \lambda_1\mu_1\bar{p}\bar{\lambda_2}\mu_2 \\
(u,d)_{(j,k)} &= \lambda_1\bar{\mu_1}\bar{\lambda_2}\mu_2 \\
(d,r)_{(j,k)} &= \bar{\lambda_1}\mu_1(p\bar{\lambda_2}\mu_2 + \bar{p}\bar{\lambda_2}\bar{\mu_2} + \bar{p}\lambda_2\mu_2) \\
(r,r)_{(j,k)} &= \bar{\lambda_1}\bar{\mu_1}(\lambda_2\mu_2 + \bar{\lambda_2}\bar{\mu_2}) + \lambda_1\mu_1(p\bar{\lambda_2}\mu_2 + \bar{p}\bar{\lambda_2}\bar{\mu_2} + \bar{p}\lambda_2\mu_2) \\
(u,r)_{(j,k)} &= \lambda_1\bar{\mu_1}(\lambda_2\mu_2 + \bar{\lambda_2}\bar{\mu_2}) \\
(d,u)_{(j,k)} &= \bar{\lambda_1}\mu_1(p\lambda_2\mu_2 + p\bar{\lambda_2}\bar{\mu_2} + \bar{p}\lambda_2\bar{\mu_2}) \\
(r,u)_{(j,k)} &= \lambda_1\mu_1(p\lambda_2\mu_2 + p\bar{\lambda_2}\bar{\mu_2} + \bar{p}\lambda_2\bar{\mu_2}) + \bar{\lambda_1}\bar{\mu_1}\lambda_2\bar{\mu_2} \\
(u,u)_{(j,k)} &= \lambda_1\bar{\mu_1}\lambda_2\bar{\mu_2} \\
(d,u^{(2)})_{(j,k)} &= \bar{\lambda_1}\mu_1 p\lambda_2\bar{\mu_2} \\
(r,u^{(2)})_{(j,k)} &= \lambda_1\mu_1 p\lambda_2\bar{\mu_2}
\end{aligned}
$$

where $\bar{x} = 1 - x$ for any x. To obtain the stationary probabilities, we need the $\boldsymbol{\Pi}$ that satisfies the system of equations given by $\boldsymbol{\Pi} = \boldsymbol{\Pi}\mathbf{Q}$. A closed form solution for $\boldsymbol{\Pi}$ does not exist, but a numerical approximation can be obtained by taking a large power of a truncated matrix. Once again, in systems with finite capacity such as the ER-hospital network we are considering, exact solutions can be numerically calculated via the matrix multiplications.

The context of gridlock prediction introduces additional structures to the simple network illustrated in Figure 3. Since we are dealing with an actual system with inherent capacity limits that are often met, we need to consider truncated systems with limited waiting rooms. The waiting room capacity for the ER is a fixed and positive number. However, due to the nature of this problem, a waiting room associated with the critical care beds in the hospital does not exist. Patients needing transfer to the hospital from the ER will use the ER beds as a waiting room, occupying them and causing gridlock. This occurrence introduces the idea of blocking into our model.

4.1 State space of the ER-hospital network

With the restrictions on queue capacity, the state space for the ER-hospital network becomes a finite one. During a visit, a patient can be in one of the following four sections in the system.

1. In the waiting room before the ER
2. Being treated in a bed in the ER
3. Waiting idle for a hospital bed while occupying a bed in the ER
4. Being treated in a bed in the hospital

A state space representing all these sections individually would need to be four dimensional, resulting in complex and time consuming calculations. The approach presented in [10] combines the first two positions, simplifying

Table 1 Description of the state naming approach

(n, d, h)	Description	Range
n	Patients in the waiting room and non-idle patients in the ER	$0, \ldots, n_{max}$ (n_{max} = # of ER beds + Waiting room capacity)
d	Patients on hold	$0, \ldots, d_{max}$ (d_{max} = # of ER beds)
h	Patients in the hospital (Critical care beds)	$0, \ldots, h_{max}$ (h_{max} = # of hospital beds

the state space. In this research, the labeling shown in Table 1, is used to make the manipulations comparatively more manageable. The first element describing a state, n, represents the number of patients in the waiting room and the non-idle patients in the ER. The value of n ranges from 0 to n_{max} (waiting room capacity + total number of ER beds). The second element, d, denotes the number of idle patients in the ER, and the third element, h, gives the number of patients in the hospital.

Not all of the states created using [10]'s naming approach are feasible. For a patient to be idle ($d > 0$), all the hospitals beds need to be full ($h = h_{max}$). Also, patients in the waiting room and the ER ($n+d$) can not be greater than the maximum capacity of the ER and the waiting room (n_{max}). The total number of feasible states is given in (5). For example a system with a 20 seat waiting room capacity, 10 ER beds and 4 critical care beds would have 410 feasible states.

$$\begin{aligned} \text{Number of feasible states} &= (n_{max}+1)(h_{max}+1) + \sum_{k=n_{max}-d_{max}}^{n_{max}-1} (k+1) \\ &= (n_{max}+1)(h_{max}+1) + \frac{d_{max}(2n_{max}-d_{max}+1)}{2} \end{aligned} \tag{5}$$

We can further simplify our model for illustrative purposes by introducing additional restrictions to the system. In order to do so, let us suppress the external arrivals to the critical care beds in the hospital. This may be a valid restriction in cases where there is a group of beds reserved strictly for ER transfers. Let us also force all patients whose services have been completed in the ER to need a hospital transfer. Note that we can easily relax these restrictions if the system in question requires it (an advantage of using discrete time queues as pointed out in [15]).

We also need to decide on a queue discipline since we are considering a system with batch arrivals and service completions. Let us assume the model depicted in Figure 4 which is similar to the late-arrival scheme. The arrivals to the system happen right before the end of a time slot. The departures occur in the beginning of a time slot, but they are ordered; hospital patients complete their service and depart the system first, followed by the completion of ER

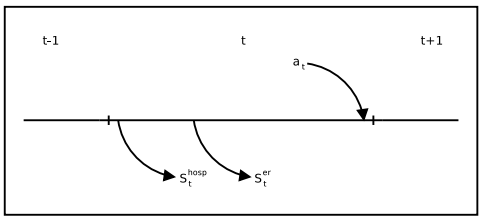

Fig. 4 Network arrival-departure scheme

services. System state variables (n_t, d_t, h_t) is read after the arrivals to system is completed.

In order to describe the transition probability matrix of our network, let us denote the following functions:

- $A(k)$: Probability that k customers arrive to the ER in a time slot
- $S^{er}(j)$: Probability of completing the service of j patients at the ER in a time slot
- $S^{hosp}(i)$: Probability of completing the service of i patients at the hospital in a time slot

Then, we can define the probabilities in the transition matrix τ as

$$P((n_t, d_t, h_t) \rightarrow (n_{t+1}, d_{t+1}, h_{t+1})) = \sum_{(i,j,k)\in H} A(k)S^{er}(j)S^{hosp}(i)$$

where H is the set of triplets (i, j, k) adhering to the following conditions:

$$\begin{aligned} n_{t+1} &= n_t + k - j \\ d_{t+1} &= d_t + j - min(j + d_t, h_{max} - h_t + i) \\ h_{t+1} &= h_t + min(j + d_t, h_{max} - h_t + i) - i \end{aligned}$$

The transition from n_t to n_{t+1} is determined by k arrivals to the ER and j service completions. All k arrivals will increase n_t and all j service completions will decrease n_t even if they can not leave the ER. This is because n_t only considers non-idle patients in the ER. The j service completions will increase the number of delayed customers d_t momentarily. If some of the delayed patients can be transferred to the hospital, then the number of delayed patients will decrease by the number of transferred patients which is given by $min(j + d_t, h_{max} - h_t + i)$. These transferred patients will increase the hospital occupancy immediately since there is no queue. However, the service completions at the hospital given by i will decrease h_t in the same time slot.

We will now consider the same ER-hospital network, with arrival and service processes that depend on a common and randomly changing environment. The system parameters change based on the state of the environment that has Markovian jumps, modulated by a transition matrix G. Consequently, the

homogeneous arrival and service rates become dependent on the state the environment is in and we have

$$\begin{aligned}\lambda(i) &: \text{Arrival rate to the ER when the environment is in state } i\\ \mu^{er}(i) &: \text{Service rate at the ER when the environment is in state } i\\ \mu^{hosp}(i) &: \text{Service rate at the hospital when the environment is in state } i\end{aligned}$$

where $i \in E$.

The state space for the Markov-modulated network can be obtained from the homogeneous network model discussed in Section 4.1. Recall that τ is the matrix containing the transition probabilities of the network states described by the $\{n_t, d_t, h_t\}$ triplets. Bringing in Markov modulation increases the state space by one more dimension, incorporating the environmental state, resulting in the 4-tuple $\{n_t, d_t, h_t, y_t\}$. The transition probability matrix for this new state space $\mathbf{Q}$ can be obtained by combining τ and the environment's transition matrix G. If we define $\tau_{k,l} \equiv G(k,l) \times \tau$, then

$$\mathbf{Q} = \begin{bmatrix} \tau_{1,1} & \tau_{1,2} & \cdots & \tau_{1,K} \\ \tau_{2,1} & \tau_{2,2} & \cdots & \tau_{2,K} \\ \cdots & \cdots & \cdots & \cdots \\ \tau_{K,1} & \tau_{K,2} & \cdots & \tau_{K,K} \end{bmatrix}$$

where K is the number of states in the environment E. This approach creates a dependent structure among the arrival and service rates. However, when conditioned on the environmental state, the rates are independent. We will exploit this conditional independence property when providing the Bayesian analysis.

In this research, we aim to develop a methodology for predicting gridlocks in an ER-hospital system, to be used as a warning system in healthcare operations. As a predictive measure, we will utilize transient probabilities of first state passage times in a discrete time Markov chain representing the states of the network. Since we have an inherently finite system, transient results can be obtained from k-step state transition matrices, which can be calculated by taking the k^{th} power of a state transition matrix.

We are interested in the probability that the network enters gridlock within a fixed time frame given the current state of the system. For any k,

$$P(\{n_{t+k}, d_{t+k}, h_{t+k}, y_{t+k}\} \in \nu | \{n_t, d_t, h_t, y_t\}) \tag{6}$$

where ν is the set of gridlock states, provides the probability of the network entering gridlock in exactly k time slots and can be obtained by looking at the row corresponding to the initial state of the k-step transition probability matrix, and adding the probabilities matching gridlock states. In order to obtain the probability of gridlocking anytime during the k time slots, and decrease the size of our matrix, we can modify the matrix $\mathbf{Q}$ to collect the states that satisfy gridlock requirements in an absorbing super state and create a new matrix $\mathbf{Q}^*$. As discussed in [17], by creating the absorbing gridlock state,

a system entering a gridlock will not leave that state and probabilities of ever gridlocking can be obtained.

According to our definition of gridlock, only the states with full ER beds ($n + d > c$) and at least one idle patient ($d > 0$) are gridlock indicators. To create the new matrix, we go through the following steps:

1. Let the set $\nu \equiv \{n, d, h, y\}$ s.t. $d > 0$ and $n + d > c$.
2. Add an additional state g^* to the ER-hospital network's state space.
3. For each row in $\mathbf{Q}^*$, sum the probabilities corresponding to transitions to states in ν. Copy this sum to the column corresponding to g^*.
4. Make the state g^* absorbing, i.e. set $\mathbf{Q}^*(g^*, g^*) = 1$.
5. Remove all the row and columns in $\mathbf{Q}^*$ that are in ν.

This operation leaves us with a proper transition probability matrix $\mathbf{Q}^*$ for the Markov chain representing the states of the ER-hospital network in a randomly modulated environment.

To illustrate how to perform gridlock predictions, let us first assume the set of system parameters are known and set them to arbitrary values $\{\bar{\lambda}(i), \bar{\mu}^{er}(i), \bar{\mu}^{hosp}(i)\}$. Let us also assume that the current environmental state is known. The prediction is performed conditioned on an initial characteristics of the ER-hospital network. As discussed in [17], by creating the absorbing gridlock state, a system entering a gridlock will not leave that state. Consequently, the probability of ever entering gridlock in k time slots can be given as

$$P(T_g \leq k | \{n_0, d_0, h_0, y_0\}) = (\mathbf{Q}^*)^k (\{n_0, d_0, h_0, y_0\}, g^*) \tag{7}$$

where T_g is the first passage time to the gridlock state.

However, we will most likely not be able to observe the environment. In that case, we would need to average the gridlock probabilities conditioned on the environmental state over the possible environmental states, which can be obtained as

$$P(T_g \leq k | \{n_0, d_0, h_0\}) = \sum_{y \in E} (\mathbf{Q}^*)^k (\{n_0, d_0, h_0, y\}, g^*) G^*(y) \tag{8}$$

where $G^*(y)$ is the steady state probability that the environment is in state y, which can be obtained from the y^{th} column of a large power of the G matrix.

We can also obtain the distribution of the first passage time (T_g) to the gridlock state g^* using the recursive formula

$$\begin{aligned} P(T_g < k | \{n_0, d_0, h_0\}) &= \mathbf{Q}^*(\{n_0, d_0, h_0\}, g^*) \\ &+ \sum_{\forall \{n,d,h\}} \mathbf{Q}^*(\{n_0, d_0, h_0\}, \{n, d, h\}) P(T_g < k - 1 | \{n, d, h\}) \end{aligned} \tag{9}$$

for $k > 3$, where $P(T_g < 2 | \{n_0, d_0, h_0\}) = \mathbf{Q}^*(\{n_0, d_0, h_0\}, g^*)$.

5 Bayesian analysis of the ER-Hospital network

5.1 Inference of the arrival and service rates

The Bayesian analysis of the simplified ER-Hospital network described above requires the inference of arrival and service process parameters jointly for the ER and the hospital. We are considering a system with blocking which require slight modifications to the Bayesian inference of the single discrete time queue.

Let us define the following parameters:

- λ^{er} - arrival rate to the ER
- μ^{er} - service rate at the ER
- μ^{hosp} - service rate at the hospital

For the reasons discussed previously, batch arrival models are more appropriate when modeling an inherently continuous system via a discrete queueing model. Let us observe the system through its three components. The arrivals to the ER are counted at each time slot providing us $a^{(T)} = \{a_1, \ldots, a_T\}$. Similarly, the service completions at the ER and the hospital are counted and recorded at each time slot resulting in the sequences $s_{er}^{(T)} = \{s_1^{er}, \ldots, s_T^{er}\}$ and $s_{hosp}^{(T)} = \{s_1^{hosp}, \ldots, s_T^{hosp}\}$, respectively.

We can obtain the state labels given in Table 1 by keeping track of the arrivals and service completions observed at the network.

$$
\begin{aligned}
n_t &= n_{t-1} + a_t - s_t^{er} \\
d_t &= d_{t-1} + s_t^{er} - min(s_t^{er} + d_{t-1}, h_{max} - h_{t-1} + s_t^{hosp}) \\
h_t &= h_{t-1} + min(s_t^{er} + d_{t-1}, h_{max} - h_{t-1} + s_t^{hosp}) - s_t^{hosp}
\end{aligned}
$$

Since the state of the system does not affect the arrival of the customers, the inference for λ^{er} is identical to the single queue case. Therefore, we will adopt the inference for the c-server batch arrival geometric queue, $Geo^x/Geo/c$. The batch sizes in this model were defined by a Poisson distribution of rate λ, which will be referred to as λ^{er} in this section. Given a sequence of arrivals $a^{(T)}$ and a conjugate $Gamma(\alpha, \beta)$ prior, we can obtain the posterior distribution for the arrival rate λ^{er} as $Gamma(\alpha + \sum_{t=1}^{n} a_t, \beta + n)$.

The ER service process on the other hand is slightly different. The dependence on the busy servers still exist. However, due to boarding (blocked) patients occupying ER beds, not all occupied servers can complete a service. Specifically, the number of service completions need to be selected from the number of *working* servers. The number of actively working servers at time slot t depends on the system occupancy and is given by $min(n_{t-1}, c - d_{t-1})$. If there are no customers in the waiting room, then n_t will provide the number of working servers. If there are patients in the waiting room on the other hand, $c - d_{t-1}$ is the number of servers that are not occupied by blocked patients, which is an upper limit on the number of possible service completions at time

slot t. We can employ a binomial process dependent on the number of working servers given as

$$P(S_t = s_t | a^{(T)}, s_{er}^{(T)}, s_{hosp}^{(T)}) \tag{10}$$
$$= \binom{min(n_{t-1}, c - d_{t-1})}{s_t} (\mu^{er})^{s_t} (1 - \mu^{er})^{min(n_{t-1}, c-d_{t-1}) - s_t}$$

which in turn results in the joint likelihood for the ER service rate μ^{er}

$$L(\mu^{er}; a^{(T)}, s_{er}^{(T)}, s_{hosp}^{(T)}) \propto (\mu^{er})^{\sum_{t=1}^{n} s_t^{er}} (1 - \mu^{er})^{\sum_{t=1}^{n} (min(n_{t-1}, c-d_{t-1}) - s_t^{er})}. \tag{11}$$

When the likelihood in (11) is combined with a conjugate $Beta(\gamma_{er}, \delta_{er})$ prior, the posterior distribution of the ER service rate parameter can be obtained as

$$P(\mu^{er} | a^{(T)}, s_{er}^{(T)}, s_{hosp}^{(T)}) \tag{12}$$
$$\sim Beta(\gamma_{er} + \sum_{t=1}^{n} s_t^{er}, \delta_{er} + \sum_{t=1}^{n} (min(n_{t-1}, c - d_{t-1}) - s_t^{er})).$$

The service process in the hospital does not have any blocking. However, it lacks a waiting room, so, the number of customers in this part of the network given by h_t is always equal to the number of busy (and working) servers. Once again, assuming a conjugate $Beta(\gamma_{hosp}, \delta_{hosp})$ prior, we can obtain the posterior distribution of μ^{hosp} as $Beta(\gamma_{hosp} + \sum_{t=1}^{n} s_t^{hosp}, \delta_{hosp} + \sum_{t=1}^{n} (h_{t-1} - s_t^{hosp}))$.

5.2 Analysis with a Markov modulated environment

For the Bayesian analysis, following an approach inspired by [16], let us denote the observations of the arrival process by $a^{(T)} = \{a_1, \ldots, a_T\}$, the ER service process by $s_{er}^{(T)} = \{s_1^{er}, \ldots, s_T^{er}\}$, and the hospital service process by $s_{hosp}^{(T)} = \{s_1^{hosp}, \ldots, s_T^{hosp}\}$. Since the arrival and the service processes are independent given the observable environmental state $y^{(T)} = \{y_1, \ldots, y_T\}$, we can obtain the likelihood functions for the realizations as

$$L(\boldsymbol{\Lambda}; a^{(T)}, y^{(T)}) \propto \prod_{t=1}^{n} G(y_{t-1}, y_t) \lambda(y_t)^{a_t} e^{-\lambda(y_t)}$$

$$L(\mathbf{M^{er}}; a^{(T)}, s_{er}^{(T)}, y^{(T)})$$
$$\propto \prod_{t=1}^{n} G(y_{t-1}, y_t) (\mu^{er}(y_t))^{s_t^{er}} (1 - \mu^{er}(y_t))^{min(n_{t-1}, c-d_{t-1}) - s_t^{er}}$$

$$L(\mathbf{M^{hosp}}; a^{(T)}, s_{hosp}^{(T)}, y^{(T)})$$
$$\propto \prod_{t=1}^{n} G(y_{t-1}, y_t) (\mu^{hosp}(y_t))^{s_t^{hosp}} (1 - \mu^{hosp}(y_t))^{h_{t-1} - s_t^{hosp}}$$

where $G(y_0, y_1) = 1$, $N_t = \sum_{i=1}^{t} a_i - \sum_{i=2}^{t} s_i$, $\boldsymbol{\Lambda} = \{\lambda(1), \ldots, \lambda(K)\}$, $\mathbf{M^{er}} = \{\mu^{er}(1), \ldots, \mu^{er}(K)\}$, $\mathbf{M^{hosp}} = \{\mu^{hosp}(1), \ldots, \mu^{hosp}(K)\}$, and K is the number of distinct environments.

We can put independent Dirichlet priors on the rows of the transition matrix G

$$P(G(i)) \propto \prod_{j \in E} G(i,j)^{\zeta_j^i - 1}. \tag{13}$$

for $i \in E$.

We can also assume conjugate $Gamma(\alpha(i), \beta(i))$ priors on the state dependent arrival process and conjugate $Beta(\gamma^{er}(i), \delta^{er}(i))$ and $Beta(\gamma^{hosp}(i), \delta^{hosp}(i))$ priors on the state dependent service processes in the ER and the hospital, respectively.

For the case where the environmental processes are observable (i.e., the data $\mathbf{D} = \{a^{(T)}, s_{er}^{(T)}, s_{hosp}^{(T)}, y^{(T)}\}$), we can analytically obtain the posterior distributions of the unknown arrival and service parameters. These distributions are given as

$$G(i)|\mathbf{D} \sim \text{Dirichlet}\left\{\zeta_j^i + \sum_{t=1}^{n-1} I(y_t = i, y_{t+1} = j); j \in E\right\} \tag{14}$$

$$\lambda(i)|\mathbf{D} \sim \text{Gamma}\left(\alpha(i) + \sum_{t=1}^{n} I(y_t = i)a_t, \beta(i) + \sum_{t=1}^{n} I(y_t = i)\right) \tag{15}$$

$$\mu^{er}(i)|\mathbf{D} \sim \text{Beta}\left(\gamma^{er}(i) + \sum_{t=1}^{n} I(y_t = i)s_t^{er}, \delta^{er}(i) + \sum_{t=1}^{n} I(y_t = i)(min(n_{t-1}, c - d_{t-1}) - s_t^{er})\right) \tag{16}$$

$$\mu^{hosp}(i)|\mathbf{D} \sim \text{Beta}\left(\gamma^{hosp}(i) + \sum_{t=1}^{n} I(y_t = i)s_t^{hosp}, \delta^{hosp}(i) + \sum_{t=1}^{n} I(y_t = i)(h_{t-1} - s_t^{hosp})\right) \tag{17}$$

where $I(.)$ is the indicator function that returns 1 if its argument is true, and 0 otherwise.

For the case where the environmental process is unobservable, we cannot obtain recognizable distributions for the parameters. Instead, we can employ an Markov chain Monte Carlo (MCMC) method such as the Gibbs sampler. For the sampler, need the full conditional distributions of the unknown parameters $\mathbf{G}, \boldsymbol{\Lambda}, \mathbf{M}^{er}, \mathbf{M}^{hosp}, \mathbf{Y}$. The first four full conditional distributions are already given above. The distribution for $\mathbf{Y}$ can be given as

$$P(Y_t|\mathbf{D}, \mathbf{Y}^{(-\mathbf{t})}, \lambda(Y_t), \mu^{er}(Y_t), \mu^{hosp}(Y_t), \mathbf{G}) \propto \tag{18}$$
$$G(Y_{t-1}, Y_t)\lambda(Y_t)^{a_t} e^{-\lambda(Y_t)}$$
$$\times (\mu^{er}(Y_t))^{s_t^{er}} (1-\mu^{er}(Y_t))^{min(n_{t-1}, c-d_{t-1})-s_t^{er}} G(Y_t, Y_{t+1})$$
$$\times (\mu^{hosp}(Y_t))^{s_t^{hosp}} (1-\mu^{hosp}(Y_t))^{h_{t-1}-s_t^{hosp}} G(Y_t, Y_{t+1})$$

for $t < n$, and

$$P(Y_t|\mathbf{D}, \mathbf{Y}^{(-\mathbf{t})}, \lambda(Y_t), \mu^{er}(Y_t), \mu^{hosp}(Y_t), \mathbf{G}) \propto \tag{19}$$
$$G(Y_{t-1}, Y_t)\lambda(Y_t)^{a_t} e^{-\lambda(Y_t)}$$
$$\times (\mu^{er}(Y_t))^{s_t^{er}} (1-\mu^{er}(Y_t))^{min(n_{t-1}, c-d_{t-1})-s_t^{er}} G(Y_t, Y_{t+1})$$
$$\times (\mu^{hosp}(Y_t))^{s_t^{hosp}} (1-\mu^{hosp}(Y_t))^{h_{t-1}-s_t^{hosp}}$$

for $t = n$.

To simulate from the joint posterior distribution of $\{\mathbf{G}, \boldsymbol{\Lambda}, \mathbf{M}^{er}, \mathbf{M}^{hosp}, \mathbf{Y}\}$, we first determine $\boldsymbol{\zeta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ to form our priors. Then we observe our data $\{a^{(T)}, s^{(T)}\}$, and set $Y_1 = 1$. We then need to initialize the Gibbs sampler for the first iteration and give values to: $Y_2^0, \ldots, Y_T^0$, $G_{1,1}^0, \ldots, G_{K,K}^0$,$\lambda(1)^0, \ldots, \lambda(K)^0$, and $\mu(1)^0, \ldots, \mu(K)^0$. Then we can set the counter $[ctr] = 0$ and start the sampler:

1. Generate $G_i^{[ctr+1]}$, $i = 1, \ldots, K$ from (14)
2. Generate $\lambda(i)^{[ctr+1]}$, $i = 1, \ldots, K$ from (15)
3. Generate $\mu^{er}(i)^{[ctr+1]}$, $i = 1, \ldots, K$ from (16)
4. Generate $\mu^{hosp}(i)^{[ctr+1]}$, $i = 1, \ldots, K$ from (17)
5. Generate $Y_t^{[ctr+1]}$, $t = 2, \ldots, T$ from (18) and (19)
6. Increment ctr by 1 and go to Step 1.

We can use these joint posterior distributions to obtain the desired inferences about the parameters of interest as well as use the generated samples in calculating gridlock probabilities in the next section. Since the ordering of the system states is arbitrary, we will concentrate on the marginal distributions of the three components of the system, n, d, and h.

5.3 Gridlock prediction

In the Bayesian framework, our inference provided us simulated samples of the network parameters. In order to obtain a gridlock probability we need to calculate the posterior predictive distribution of the gridlock probability in k time slots, given our samples. If we let $\boldsymbol{\Theta}*$ be the collection of the posterior samples of the system parameters $\lambda(i), \mu^{er}(i), \mu^{hosp}(i), \mathbf{G}$ for $i \in E$, then

$$P(T_g \leq k|\{n_t, d_t, h_t\}) = \frac{1}{J}\sum_{j=1}^{J}\sum_{y\in E}(\mathbf{Q}_\mathbf{j}^*)^k(\{n_t, d_t, h_t, y\}, g^*)G_j^*(y) \tag{20}$$

where $(\mathbf{Q}^*)_j^k$ is the k^{th} power of the matrix $\mathbf{Q}^*$ obtained by using the j^{th} member of J posterior $\boldsymbol{\Theta}^*$ samples, and $G_j^*(y)$ is the $G^*(y)$ obtained from the j^{th} sample.

6 Numerical demonstration

Since we do not have full ER-hospital network data to use in our model, for demonstrative purposes, we will employ a sequence of 1000 simulated arrival and service times. The test data is modulated by a 2-state environment governed by the transition matrix

$$\mathbf{G} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} \end{array} \tag{21}$$

and the arrival and service rates given by

- $\lambda(1) = \lambda(2) = 4$
- $\mu^{er}(1) = 0.8$, $\mu^{er}(2) = 0.9$
- $\mu^{hosp}(1) = 0.7$, $\mu^{hosp}(1) = 0.95$

for an ER-hospital network with 5 ER beds, 5 hospital beds and a waiting room capacity of 5 patients. As with all our other analysis, due to the lack of apriori information, we will employ non-informative priors.

6.1 Inference Results

The arrival rates $\lambda(1)$ and $\lambda(2)$ are known to be equal. The box plots shown in Figure 5 do not contradict this information, and depicts overlapping distributions with very similar means. The differences in the ER service rates $\mu^{er}(1)$ and $\mu^{er}(2)$, and hospital service rates $\mu^{hosp}(1)$ and $\mu^{hosp}(2)$ on the other hand can be clearly observed in Figure 6 and Figure 7, respectively.

We can also obtain the posterior distribution of $\mathbf{G}$ which is illustrated in Figure 8. As expected, the plotted histograms show that the distributions the members of $\mathbf{G}$ are centered around the true values used to create the simulated dataset.

6.2 Steady state results and gridlock prediction

For each of the joint posterior sample obtained from the Bayesian inference, we can form a transition matrix $\mathbf{Q}$ for the ER-hospital network, and a reduced matrix $\mathbf{Q}^*$ that has an absorbing gridlock state. From our analysis, we have obtained the steady state distribution of the network which is obtained by taking a large power of the probability transition matrix $\mathbf{Q}$ for each of the posterior samples, and averaging them out over the number of samples. The

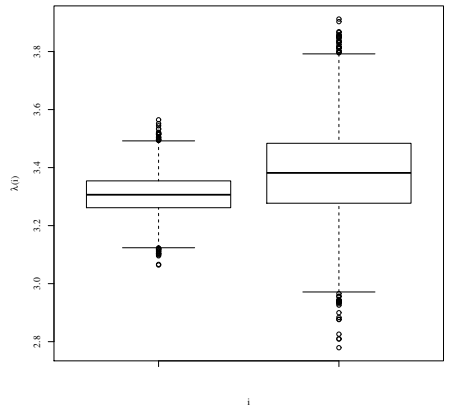

Fig. 5 Box Plots of the arrival rates $\lambda(1)$ and $\lambda(2)$

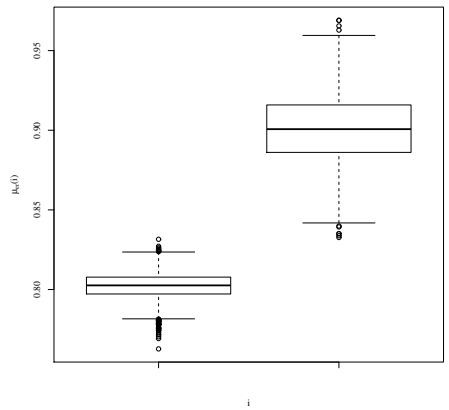

Fig. 6 Box Plots of the ER service rates $\mu^{er}(1)$ and $\mu^{er}(2)$

resulting distribution provides us the steady state probability of the system being in any one of its feasible states in the long run. However, since the ordering of the system states is arbitrary, distribution of the system states looks multi-modal and difficult to interpret. Instead, we have obtained the marginal distributions of n, d and h. These distributions are illustrated via the posterior sample histograms plotted in Figures 9, 10, and 11, respectively.

For the gridlock prediction, we need to concentrate on the reduced matrix $\mathbf{Q}^*$. In Figure 12, we are providing the probabilities of the network going into gridlock in two hours conditioned on any one of the feasible, non-gridlocked states of the system. Since we are interested in a two-hour window and time slots have been selected to be 30 minutes long, for each posterior sample,

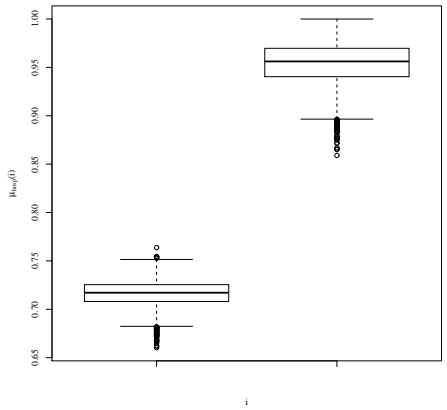

Fig. 7 Box Plots of the Hospital service rates $\mu^{hosp}(1)$ and $\mu^{hosp}(2)$

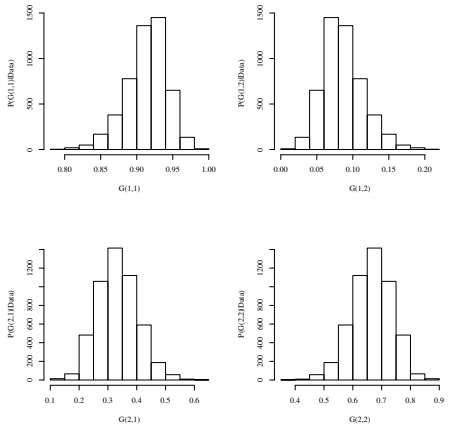

Fig. 8 Posterior distribution of **G**

the fourth power of the matrix $\mathbf{Q}^*$ is calculated. Then the column associated with the absorbing gridlock state g^* has been collected from each sample and averaged over the samples to provide the gridlocking probabilities.

We may also be interested in the relationship between the gridlock probabilities and the individual components of the system state (n, d, h). We can investigate this effect by plotting the gridlock probabilities against increasing values of n, d and h separately, keeping the remaining components constant. For the effect of n, we fixed $d = 0$ and $h = 5$. Similarly, for d, we fixed $n = 0$ and $h = 5$, and for h, we fixed $n = 10$ and $d = 0$. In Figures 13, 14 and 15, we can see the resulting distributions collected from Figure 12.

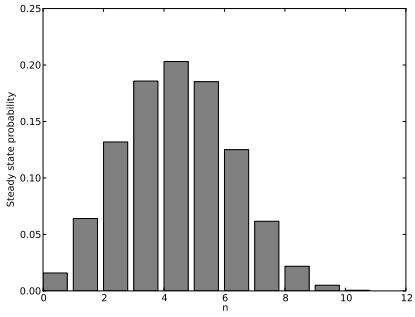

Fig. 9 Steady state distribution of n

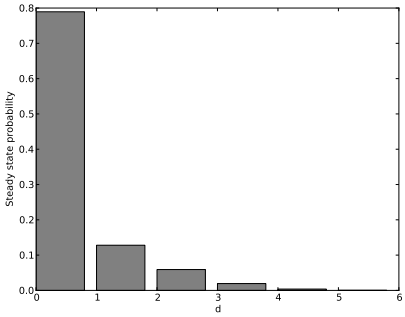

Fig. 10 Steady state distribution of d

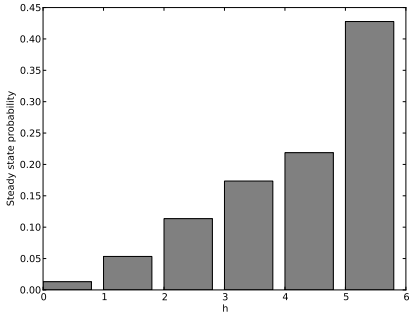

Fig. 11 Steady state distribution of h

7 Concluding remarks

In this work, we illustrate a method to conduct Bayesian analysis on discrete time networks and perform gridlock predictions on the ER-hospital network. In doing so, a Bayesian method for performing probabilistic inference on the

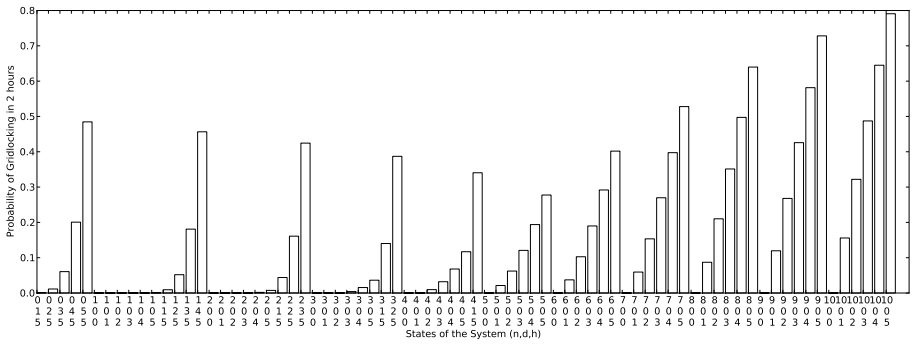

Fig. 12 Probability of gridlocking in 2 hours given current network state

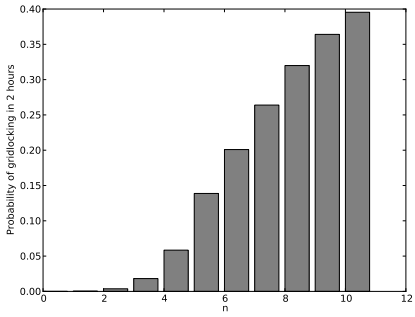

Fig. 13 Effect of n on gridlock probability

parameters of homogeneous and and Markov modulated discrete time queues is presented. Inferential methods on the networks of discrete time queues as well as any Bayesian work is not existent in the current literature as far as we know, and out methods can be used for the analysis of both continuous and discrete time systems. We also show that numerical results are fairly straightforward to obtain and allow for modeling complex systems which can not be analyzed via closed form solutions.

In the ER-hospital network considered, there are several layers of the system where decisions needs to be made. Fundamentally, system capacity decisions can be made based on server costs and customer service levels as discussed earlier. However, one can incorporate gridlock probabilities into the analysis. When considering system design problems, steady state probabilities of reaching gridlock can be a part of the penalty function by considering the associated costs (e.g. idle resources, ambulance diversions). Transient analysis discussed in this work also brings in additional questions. Since we bring an ability to provide gridlock probabilities, temporary mechanisms designed to relieve the system congestion can be built into the system. For instance, an additional room of critical care beds that are only used during high gridlock

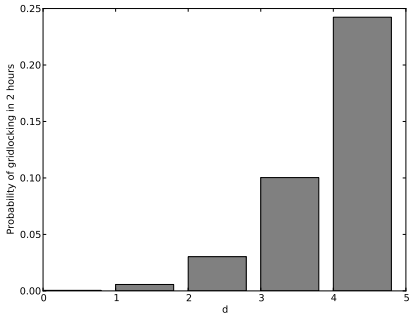

Fig. 14 Effect of d on gridlock probability

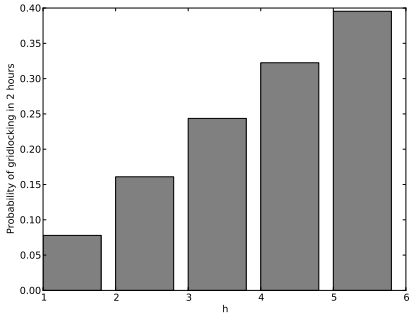

Fig. 15 Effect of h on gridlock probability

probability states can be utilized. The system may also attempt to increase its service rate by taking some costly shortcuts or temporary increases in personnel. As we have seen in our second essay, it is not unusual for the service rate to increase as a reaction to an increasing arrival rate. Allowing for the service rate to increase as a precautionary measure may avoid severe congestions and gridlocks, providing better customer service and decreasing operating costs. Since a discrete time structure assumed, results obtained can be used directly since most operations decisions are made based on a discrete time schedule.

References

1. Armero, C., Bayyari, M.J.: Bayesian prediction in m/m/1 queues. Queueing Systems **15**, 401–417 (1994)
2. Asplin, B.R.: Does ambulance diversion matter? Annals of Emergency Medicine **41**(4), 477–480 (2003).
3. Bennett, R.M.: State of emergency: modeling heals the wounds of emergency department diversions. Industrial Engineer **35**(5), 50–54 (2003)
4. Brewer, S.: Study: clogged trauma care leads to deaths. Newspaper Article (2002)
5. Conti, P.L.: Large sample bayesian analysis for geo/g/1 discrete-time queueing models. The Annals of Statistics **27**(6), 1785–1807 (1999).

6. Conti, P.L.: Bootstrap approximations for bayesian analysis of geo/g/1 discrete-time queueing models. Journal of Statistical Planning and Inference **120**(1-2), 65 – 84 (2004).
7. Daduna, H.: Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks, *Lecture notes in computer science*, vol. 2046. Springer (2001)
8. Dafermos, S.C., Neuts, M.F.: A single server queue in discrete time. Cahiers du Centre d'Etude de Recherche Operationnelle **13**, 23–40 (1971)
9. Henry, M.C.: Overcrowding in america's emergency departments: Inpatient wards replace emergency care. Academic Emergency Medicine **8**(2), 151–155 (2001)
10. Hunt, G.C.: Sequential Arrays of Waiting Lines. Operations Research **4**(6), 674–683 (1956).
11. Kleinrock, L.: Analysis of a time-shared processor. Naval Research Logistics Quarterly **11**(1), 59–73 (1964).
12. Kleinrock, L.: Time-shared systems: a theoretical treatment. J. ACM **14**(2), 242–261 (1967).
13. McGrath, M.F., Gross, D., Singpurwalla, N.D.: A subjective bayesian approach to the theory of queues i – modeling. Queueing Systems **1**(4), 317–333 (1987)
14. McGrath, M.F., Singpurwalla, N.D.: A subjective bayesian approach to the theory of queues ii – inference and information in m/m/1 queues. Queueing Systems **1**(4), 335–353 (1987)
15. Neuts, M.F.: The single server queue in discrete time-numerical analysis i. Naval Research Logistics Quarterly **20**(2), 297–304 (1973).
16. Ozekici, S., Soyer, R.: Network reliability assessment in a random environment. Naval Research Logistics **50**, 574–579 (2003)
17. Ross, S.M.: Introduction to Probability Models, Ninth Edition. Academic Press, Inc., Orlando, FL, USA (2006)
18. The Lewin Group: Emergency department overload: A growing crisis survey (2002)
19. The Lewin Group: Hospital capacity and emergency department diversion: Four community case studies (2004)