Modeling the Emergency Care Delivery System Using a Queueing Approach

by

Xu Sun

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis considers a regional emergency care delivery system that has a common emergency medical service (EMS) provider and two hospitals, each with a single emergency department (ED) and an inpatient department (ID). Patients arrive at one of the hospital EDs either by ambulance or self-transportation, and we assume that an ambulance patient has preemptive priority over a walk-in patient. Both types of patients can potentially be admitted into the ID or discharged directly from the ED. An admitted patient who cannot access the ID due to the lack of available inpatient beds becomes a boarding patient and blocks an ED server. An ED goes on diversion, e.g., requests the EMS provider to divert incoming ambulances to the neighboring facility, if the total number of its ambulance patients and boarding patients exceeds its capacity (the total number of its servers). The EMS provider will accept the diversion request if the neighboring ED is not on diversion. Both EDs choose its capacity as its diversion threshold and never change the threshold value strategically, and hence they never game. Although the network could be an idealized model of an actual operation, it can be thought of as the simplest network model that is rich enough to reproduce the variety of interactions among different system components. In particular, we aim to highlight the bottleneck effect of inpatient units on ED overcrowding and the network effects resulting from ED diversions. A continuous time Markov chain is introduced for the network model. We show that the chain is irreversible and hence its stationary distribution is difficult to characterize analytically. We identify an alternative solution that builds on queueing decomposition and matrix-analytic methods. We demonstrate through discrete-event simulations the effectiveness of this solution on deriving various performance measures of the original network model. Moreover, by conducting extensive numerical experiments, we provide potential explanations for the overcrowding and delays in a network of hospitals. We suggest remedies from a queueing perspective for the operational challenges facing emergency care delivery systems.

Acknowledgements

A thesis is a rather formal document. I feel it necessary, however, to slightly bend this formality in order to express my gratitude to the people who deserve my sincere acknowledgement.

I am most grateful to my co-supervisors Dr. Qi-Ming He and Dr. Stanko Dimitrov for their guidance and support. Their knowledge, sharpness, dedication and personalities have made the past two years not only a constant learning experience, but also quiet bit of fun. What I learned from them has transformed the way I think, work and write. They serve as a "proof-by-example" that academic research and fun are not mutually exclusive and I have found their passion for research rather contagious.

I wish to express my gratitude to all my friends in Management Sciences who make my stay at Waterloo an enjoyable experience. Special thanks are given to Jill, Shuo Xu and Xiang Gao.

During the two years in Canada I received from my parents all the support and love that one could ask for. None of these was compromised by the hardship that the physical distance has imposed on.

Finally, I am indebted to my girlfriend Lisha who filled me with love, strength and encourage. I dedicate this work to her.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

An emergency department (ED) provides an extraordinarily important public service mission by offering medical treatment for a broad spectrum of illness and injuries to patients who arrive either by ambulance or in person. However, overcrowding and long waits in EDs have raised public concerns on whether hospitals with EDs can deliver effective and timely emergency care and ensure patient safety and satisfaction. In the 2008/2009 fiscal year, there were approximately 5.4 million visits to the 160 hospital emergency departments in the Province of Ontario in Canada (McCarter (2010)). A main cause of the ever increasing demand for emergency care has been found to be the inappropriate use of hospital EDs by walk-in patients with minor illness. Putting the situation in conventional queueing terms, we say that the system arrival rate exceeds the effective system service rate over extended periods. As a result, the queueing system resides in a state of dysfunctional equilibrium where the only element that keeps the queue from growing indefinitely long is customer balking (refusing to join the queue) and reneging (leaving the queue after entering). For those who opt to leave the queue, some may return later, possibly in a worsened state, while others may never receive needed treatment.

1.1 Motivation

ED overcrowding has led to system-wide problems, one of which is offload delays in clinical handover from paramedics to ED staff. If an ED reaches full capacity or lacks ED beds, ambulances carrying patients often queue at the ED, and the paramedics stay with and continue to provide care until an empty ED bed becomes available and ED nurses can accept the patient (Almehdawe et al. (2013)). This kind of delays impair the ability of an emergency medical service (EMS) system to provide adequate service by tying up ambulance resources, reducing response time (the period between when an emergency call is recorded and the time the first ambulance resource arrives at the scene, Burt et al. (2006)). The negative impact on ambulance availability can be significant. A survey conducted in Ontario, Canada stated that all 14 provincial EMS providers surveyed complained of long offload delays, which resulted in fewer or even no ambulance available to respond to new emergency calls (McCarter (2012)). Another study in Los Angeles reported a total of $21,240$ incidents where ambulances were out of service due to waiting in front of an ED to transfer their patients to an ED bed, 8.4% of which

exceeded for 1 hour (Eckstein and Chan (2004)).

Another result of ED overcrowding is ambulance diversion, a common practice initially attempted to ameliorate demand-supply imbalance by rerouting patient flows, however now considered to have negative impacts on patient safety. When overwhelmed by overcrowding, EDs often declare diversion status and redirect ambulances to neighboring facilities (Burt et al. (2006)). While this practice is intended to ensure critically ill patients get timely medical care, this can lead to serious negative consequences if nearby hospitals happen to be in the same situation (Kolb et al. (2007)) upon patient arrival, as the transit time is longer yet waiting time at the ED is not reduced. Prolonged transportation time, discontinuity of patient care, and higher mortality in severely injured trauma patients have been reported as a result of ambulance diversion. Thereafter many local authorities have placed restrictions on the use of ambulance diversion or completely prohibited such practices. However, it is suggested, in the academic literature, that ambulance diversion, if properly executed, can reduce waiting times for a network of hospitals (Ramirez et al. (2009), Ramirez-Nafarrate et al. (2011)).

While the diversion decisions are made independently by ED administrators based on crowding measures at their own location (Deo and Gurvich (2011)), a significant and positive correlation between diversion hours of neighboring hospitals has been found by anecdotal and empirical observations, providing evidence of a network effect in emergency care systems (Sun et al. (2006)). Based on this observation, we think that previous studies which isolate a single hospital ED inevitably introduce some bias while deriving performance measures.

While a vast amount of simulation and queueing models focus on the patient flows between the EMS and emergency departments, many do not take into account the bottleneck effect of IDs on ED operations (Au-Yeung et al. (2006), Hagtvedt et al. (2009), Ramirez et al. (2009), Almehdawe et al. (2013), Deo and Gurvich (2011)). Empirical investigations discover strong correlation between inpatient hospital occupancy and ED congestion (Forster et al. (2003), Kolker (2008)). In effect, lack of hospitals beds and existence of boarding patients, e.g., admitted patients who stay in the ED due to lack of inpatient beds, have been cited as the major causes of ED overcrowding and frequent vehicle diversions (Lane et al. (2000), Proudlove et al. (2003), Shi et al. (2012), Broyles and Cochran (2011)).

Although the majority of patients are directly discharged from the ED, approximately 20% of ED patients are admitted into the hospital (Broyles and Cochran (2011)). If a patient is admitted to an ID and the ID does not have a bed available to accommodate patients, then the patient has to wait in an ED bed for an inpatient bed to become available. Patient boarding reduces ED capacity by consuming ED beds with boarding patients, creating a variety of undesirable effects such as admitted patients not being placed in an appropriate ward or placed in the hallway, prolonged waiting time for incoming patients, and patients leaving without being seen. Notably, delays in receiving inpatient care can lead to longer length of stay in the hospital, which in turn causes delays in other admitted ED patients, and so forth, exaggerating the patient crowding situation in the ED (Chan et al. (2013)).

1.2 Objectives

WE formulate a queueing network that is just sufficiently complex to render the interactions among the EMS provider, hospital EDs and their IDs nontrivial. More precisely we examine the flow of patients in a regional emergency care system that incorporates a common EMS provider that manages ambulance services and two hospitals each with a single ED and a single ID. We use our model to investigate the impact of various system parameters including diversion threshold values, size of EDs and inpatient units, and inpatient discharge rates on performance of the entire system. For each ED, we look at the fraction of time that an ED puts on *effective diversion*, average waiting time for incoming ED patients and boarding patients. We then use existing data to calibrate our model.

We consider a regional emergency care delivery system where patients arrive at one of the EDs either by ambulance or through self-transportation. For each ambulance request, there is a predetermined hospital as the destination whose catchment area covers the scene of the request. If the number of ED patients of a destination hospital exceeds a predetermined threshold while the number of patients at the other hospital does not, the incoming ambulance will be diverted to the less crowded facility. Otherwise, the predetermined hospital will accept the incoming patient. Each ED patient is discharged directly from the ED or admitted to the ID. If no inpatient bed is available, the admitted patient will stay in an ED bed waiting for an inpatient bed to become available, preventing the ED bed from accepting incoming patients.

We adopt the matrix-analytic method to analyze the proposed network model. We note that a direct application of this method involves 6 system variables and hence leads to a very large infinitesimal generator, which makes it computationally burdensome to find matrix-geometric solutions. To achieve dimensional reduction, we propose an iterative queueing decomposition method via *effective* arrival rate and ED capacity. We first introduce a two-station queue as a subnetwork to mimic the dynamics in one hospital with an ED and an ID, we find stationary distributions and derive performance measures using a matrix-geometric approach. Then we construct a simplified network model that focuses on the interplay of the EMS provider and multiple EDs by not accounting for the IDs. Finally we apply the proposed queueing decomposition approach to combine all the components into a single network.

1.3 Contributions

To summarize, our contribution in this thesis is twofold.

1. We develop a queueing network with blocking to describe the patient flows in a hospital with a single ED and a single inpatient unit. We provide an exact solution for the network by using the matrix-analytic method.

2. We quantify the network effect of ambulance diversions among a network of hospitals using queueing approach and highlight the bottleneck effect of IDs on ED overcrowding and ambulance diversions.

These two findings have significant implications in the design and optimization of an emergency care network, as external factors to an ED, such as the inpatient hospital size, and the operations of neighboring EDs are all driving forces of patient crowding in an ED. To effectively address the problem of ED overcrowding, one needs to take a systematic perspective, which we do by incorporating multiple hospitals each with a single ED and a single ID.

In addition, our iterative queueing decomposition approach for analyzing network models with routings based on threshold policy is novel. It provides a possible way to reduce the system dimension for networks that are similar to what we introduce in our study. To our knowledge, similar methods have been used for the analysis of networks of fluid queues (Liu and Whitt (2011)).

1.4 Thesis outline

The rest of the thesis proceeds as follows. Chapter 2 discusses related work and mathematical background. Chapter 3 presents the original queueing network model which incorporates all the essential components for a regional emergency care system. We analyze a single hospital as a subsystem by describing it as a two-station queueing network in Chapter 4. In Chapter 5, we formulate a simplified network model that connects the EMS provider and multiple EDs. Chapter 6 displays the iterative queueing decomposition method for analyzing the original queueing network. In Chapter 7, we extend our study by adopting diffusion approximation and examine the case where the number of inpatient beds is large. We conclude in Chapter 8.

Chapter 2

Literature Review and Mathematical Background

Numerous studies have been undertaken to measure the detrimental effect of resource shortages on ED overcrowding and diversions. In this Chapter, we present an overview of the related works and classify those works into three main categories - empirical, simulation and queueing.

2.1 Empirical observations

Derlet et al. (2001) conduct empirical analysis and identify increased patient acuity and lack of hospital beds as major contributors towards ED congestion, while increased ED visits, radiology/laboratory delays, and limited ED space are found to be minor causes of ED overcrowding. Similarly, Schull et al. (2003) finds the number of inpatient beds to be the major contributor of ambulance diversion when studying the Toronto Emergency Medical Service. We also examine the effects of number of hospital beds, patient volumes and ED capacity on ED overcrowding. Our numerical results highlight the tremendous impact of inpatient unit capacity on ED congestion and ambulance diversion. Forster et al. (2003) discover a significant association between hospital bed occupancy and ED length-of-stay (LOS) by analyzing the longitudinal data of one academic hospital. Their study suggest that ED LOS is heavily affected by hospital bed occupancy level. We consider the mean boarding time in our study as a crucial indicator of ED LOS. Our numerical experiments suggest that the influence of the inpatient unit capacity on the mean boarding time is significant. Richardson (2002) conducts a retrospective study of all patients admitted to an inpatient hospital via an ED to investigate the relationship between access block (the number of admitted patient who cannot receive a physical inpatient bed) and inpatient LOS. He finds the blocked patients have a mean ED duration of more than 10 hours while no-blocked group have a duration of less than 4 hours. In addition, mean inpatient LOS is found to be 4.2 days and 5.9 days for blocked patients and no-blocked group respectively, indicating an association between patient boarding and inpatient LOS. Similar conclusions are reported by Chan et al. (2013). Yet to maintain analytic simplicity, we do not consider the effect of patient boarding on inpatient LOS. A network effect of ED diversions is identified by Sun et al. (2006), who find that hospital closure increased the hours on diversion of nearby EDs. Meanwhile the faction of time on diversion in neighboring EDs are positively correlated. In this dissertation, we investigate the correlation between

the diversion probability of two hospital EDs.

2.2 Simulation studies

More recently, substantial papers have formulated detailed simulation models to analyze ED operations. Au-Yeung et al. (2006) develop a Markovian queueing network model to examine patient flows and calibrate the model using real data. They advocate that by giving priority to minor illness over major illness, the mean waiting time for low acuity patients can be substantially reduced while the average waiting time for high acuity is not greatly affected. For ease of analysis, we group patients by the mode of arrival. Specifically we consider ambulance patients and walk-in patients. We assign preemptive priority to all ambulance patients over walk-in patients. Kolker (2008) introduces a flexible and versatile simulation model to examine the impact of ED patient LOS on ED diversion. He claims that diversion can be eliminated if patients discharged directly from an ED stay less than 5 hours, while patients admitted into the hospital stay no more than 6 hours. Though his study contributes ED diversion to prolonged LOS at ED, his model excludes external factors that lead to long ED LOS as in our model, such as lack of inpatient beds and patient boarding. Kolb et al. (2007) construct a simulation model to investigate the relationship between ED overcrowding and inpatient hospital utilization. Similar to our ED-ID model introduced in later chapters, they assume patient arrivals consist of ambulance patients and walk-in patients. ED patients are either released directly from the ED or admitted into the ID. By varying the admission rate, they detect a linear correlation between patient crowding and admission rate. As we incorporate the inpatient unit as a system component, we introduce the admission rates of ED patients into our input parameters. Powell et al. (2012) conduct a cross-sectional study on the impact of discharge timing on ED patient boarding using discrete-event simulations. Three discharge policies are virtually implemented and compared. They suggest that patient boarding can be ameliorated or even eliminated by evening the number of discharges at different times. We consider a queueing network without time-varying parameters. Yet we test the impact of inpatient discharge rate on various performance measures in our numerical experiments. Shi et al. (2012) also consider the influence of hospital discharging policies on patient boarding. They build simulation programs to approximate inpatient operations of a hospital in Singapore. They use the model to evaluate the impact of operational policies on patient waiting times and proportions of overflow. They find that reducing the utilization by 10% through increasing bed capacity can reduce the overflow proportion by 9%, though it does not help much in reducing waiting times. Hagtvedt et al. (2009) investigate strategic ambulance diversions using multiple tools, including a birth-and-death process, simulation and game theory to study possible cooperative strategies. Their model assume N beds in a hospital. The hospital will go on full diversion when the number of patients reaches N, and will remain on diversion status until $N - M$ beds are available where M is a

predetermined decision variable. We intend to conduct a cross-sectional study on an emergency care system rather than determine an optimal diversion policy. Therefore we describe a simple scenario where each hospital ED chooses its capacity as the diversion threshold. Most related to our analysis is the work of Nafarrate et al. (2010). The authors propose a simulation-optimization approach on a regional emergency deliver system and use a genetic algorithm to determine the best combination of ambulance diversion policies that give minimum average-patient non-value added time for all the hospitals in the region. in our model, we also consider an emergency care system with a common EMS provider and multiple hospitals. We develop an analytic model and use computer simulation to validate the analytic solutions.

2.3 Queueing models and queueing approximations

In the model introduced in later chapters, we consider how to incorporate the measured blocking effect of an inpatient unit into our queueing network. There have been a number of works in queueing applications which consider patient boarding in healthcare operations. Koizumi et al. (2005) establish a queueing network model with blocking to analyze the blocking phenomena between three types of mental facilities. The effective service rate is proposed to capture the congestion impact of a downstream facility on upstream stations. The authors found that the system-wide congestion is primarily due to the deficit of one facility (bottleneck station) instead of the cumulative effect of shortages across the system. Similarly Broyles and Cochran (2011) present an application of queueing network modeling in analyzing the effect of ED patient boarding on ED patient delays. They model ED and an ID as a two-station Markovian queue and decompose the system by considering the ED LOS as the sum of service time without boarding and delay for boarding. In our work, we consider the pooling effect of ambulance diversions among multiple facilities.

Deo and Gurvich (2011) consider a stylized queueing-network model of two EDs with each ED trying to minimize the expected waiting time of their own location by choosing a diversion threshold value. The authors first model the decentralized setting as a non-cooperative game, in which each ED puts on diversion status defensively to avoid patient shifts to each other in equilibrium. Due to the complexity of characterizing and analyzing the true socially optimal solution, the authors turn to an alternate solution in which diversion thresholds are set to be equal to the two EDs capacities respectively, and prove the policy to be near-optimal. We also consider diversions in our analysis. In contrast, we study a system where inpatient units are included and characterized as bottleneck stations. We use matrix-analytic methods to find stationary probabilities and derive performance measures.

Allon et al. (2013) apply fluid and diffusion approximations to analyze the impacts of inpatient capacity and ED size on the extent of ambulance diversion respectively. They express the fraction of time an ED is on diversion as a function of ED and hospital sizes as well as other system parameters.

Their stylized models suggest that inpatient capacity and ED size both contribute to ambulance diversion. Again, our model differs from theirs in that we consider ambulance offload delays and incorporate the network effect among multiple hospitals. We use matrix-geometric approach and queueing decomposition method to solve our model.

Through this dissertation, we consider an ED server as a combination of several medical resources (e.g. a bed, a physician, and a ED nurse). We consider an inpatient bed as a server in IDs. We do not distinguish different inpatient units and thus collect all the inpatient units in a single unit. For special attentions on intensive care unit (ICU), we refer the reader to McManus et al. (2004), Chan et al. (2013), and Zhu et al. (2013). Other studies that use queueing model to analyze ED staffing are elaborated in Green et al. (2006), Green (2008), and Yankovic and Green (2011).

2.4 Matrix analytic method

Introduced in 1970s by Marcel Neuts, matrix-analytic method serves as a direct and powerful tool to construct and analyze Markov chain in a unified and computationally tractable way. Initially developed to examine the embedded Markov chain of standard $G/M/1$ and $M/G/1$ queue, this technique has been successfully applied to model a wide range of stochastic models. Early monographs on matrix-analytic methods include Neuts (1981) and Neuts (1989). Developments in the 1980s and 1990s are summarized in Latouche and Ramaswami (1999). A most up-to-date book for researchers in this area is prepared by He (2013). For specialized topics and fields of application, we refer to Ost (2001), Breuer and Baum (2005) and Tian and Zhang (2006), Artalejo and Gómez-Corral (2008).

We restrict our attention to a special kind of Markov chain called the quasi-birth-and-death (QBD) process. To define a QBD process, we first describe a continuous time Markov chain.

2.4.1 Continuous time Markov Chains

Let X_t be a family of random variables, parameterized by $t \in [0, +\infty)$ and with value taken from a discrete set S. $\{X_t; t > 0\}$ is a continuous time Markov chain if

$$P[X_{s+t} = j | X_s = i, X_{s_n} = i_{s_n}, \ldots, X_{s_1} = i_{s_1}] = P[X_{s+t} = j | x_s = i] \tag{2.1}$$

for all $t > 0, s > s_n > \cdots > s_1 \geq 0$ and $i, j, i_k \in S$. A process that satisfies equation (2.1) is said to have the Markov property and the quantities $P[X_{s+t} = j | x_s = i]$ are referred as transition probabilities. Since a Markov chain can have an infinite number of states, the transition probabilities can potentially be represented by a infinite matrix P^s_{s+t}. With an additional assumption of time-homogeneity, the notation P^s_{s+t} reduces to P_t. Hence for continuous time Markov chain, the family of matrices P_t replaces the single transition matrix P of a classical Markov chain. An analogous statement for the

continuous time Markov chain is

$$P_{s+t} = P_t P_s. \tag{2.2}$$

Equation (2.2) is known as the semi-group property, and naturally $P_0 = I$. Moreover it is helpful to explicitly describe the structure of the underlying sample space Ω of a continuous time Markov chain. Here Ω is the space of step functions defined on the nonnegative real line with values taken in the state space S. With the additional requirement of right continuity in the form

$$\lim_{t \to a^+} \omega(t) = \omega(a)$$

imposed on $\omega \in \Omega$, $X_t(\omega)$ is a right continuous function of t for each fixed ω. Under this context, $P_{ij}^{(t)}$ is interpreted as the probability on the set of paths ω with $\omega(t) = j$ given that $\omega(0) = i$.

A continuous time Markov chain is determined by the matrices P_t. The fact that we now have a continuous time parameter allows us to apply notions from calculus to the continuous time process in a way that was not feasible for the discrete time version. However it also creates technical issues, a rigorous treatment of which invokes full machinery from functional analysis. For simplicity, we assume transition probability matrix P_t is right continuous and therefore

$$\lim_{t \to 0^+} P_t = I. \tag{2.3}$$

We define the infinitesimal generator of the continuous time Markov chain as the one-sided derivative

$$Q = \lim_{t \to 0^+} \frac{P_t - I}{t}. \tag{2.4}$$

For the time being, there is no guarantee that the limiting matrix Q exists, yet we proceed as if it does, and define the define the derivative of P_t at time t as

$$\frac{\mathrm{d}P_h}{\mathrm{d}t} = \lim_{t \to 0^+} \frac{P_{t+h} - P_t}{h}. \tag{2.5}$$

Using the semi-group property, one gets the *Kolmogorov backward* and *Kolmogorov forward* equations

$$\frac{\mathrm{d}P_t}{\mathrm{d}t} = QP_t, \qquad \frac{\mathrm{d}P_t}{\mathrm{d}t} = P_t Q. \tag{2.6}$$

Recall that the stationary distribution is crucial in the study of limiting behavior of a Markov chain. For a continuous time Markov chain, a stationary distribution is defined as a row vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ that satisfies

$$\boldsymbol{\pi} P_t = \boldsymbol{\pi} \ \forall t \geq 0, \quad \sum \pi_j = 1, \ \pi_j \geq 0.$$

It is well known that the condition $\boldsymbol{\pi} P_t = \boldsymbol{\pi}$ is equivalent to $\boldsymbol{\pi} Q = 0$ where again Q denotes the infinitesimal generator. It is common practice that one first constructs the infinitesimal generator a continuous time Markov chain and then derives the stationary by solving the equation $\boldsymbol{\pi} Q = 0$.

2.4.2 Quasi-birth-and-death process

A QBD process is one type of continuous time Markov chain with a state space § = $\{(i,j), i = 0, 1 \ldots, 1 \leq j \leq m\}$ where i denotes the *level* and j denotes the phase. The infinitesimal generator Q of a QBD process $\{X(t); t \geq 0\}$ takes the form of the following:

$$Q = \begin{pmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_{1,1} & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2.7}$$

where $(A_{0,0} + A_{0,1})\boldsymbol{e} = (A_{1,0} + A_{1,1} + A_0)\boldsymbol{e} = (A_2 + A_1 + A_0)\boldsymbol{e} = 0$ and $\boldsymbol{e}$ is by convention a vector of ones. The restrictions imply that matrix $A = A_2 + A_1 + A_0$ is also an infinitesimal generator with a finite state space.

2.4.3 Matrix-geometric solutions

The unique structure of a QBD process makes it possible to develop a parsimonious procedure to compute the stationary distribution. In particular we have Neuts' theorem (Neuts (1981)) that states the process $\{X(t); t \geq 0\}$ is positive recurrent iff all the eigenvalues of the minimal solution matrix R to the matrix-quadratic equation

$$A_0 + RA_1 + R^2A_2 = 0 \tag{2.8}$$

are within the unit disk and the set of equations

$$\begin{aligned} x_0(A_{0,0} + RA_{1,0}) &= \mathbf{0} \\ x_0(I - R)^{-1}\boldsymbol{e} &= 1 \end{aligned} \tag{2.9}$$

has a positive solution x_0.

If the matrix A is irreducible, then the spectral of rate matrix R is less than 1 iff

$$\boldsymbol{\theta}A_2\boldsymbol{e} > \boldsymbol{\theta}A_0\boldsymbol{e}, \tag{2.10}$$

where $\boldsymbol{\theta}$ is the stationary distribution vector of the infinitesimal generator A. The stationary distribution $\boldsymbol{\pi} = [\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots]$ satisfies the set of equations

$$\begin{aligned} \boldsymbol{\pi}Q &= 0, \\ \boldsymbol{\pi}\boldsymbol{e} &= 1, \end{aligned} \tag{2.11}$$

and has the matrix-geometric form

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 R^{n-1}, \quad n \geq 1. \tag{2.12}$$

The boundary probabilities $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)$ can be obtained by solving the following equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1) \begin{pmatrix} A_{0,0} & A_{0,1} \\ A_{1,0} & A_{1,1} + RA_2 \end{pmatrix} = 0, \tag{2.13}$$

$$\boldsymbol{\pi}_0 \boldsymbol{e} + \boldsymbol{\pi}_1 (I - R)^{-1} \boldsymbol{e} = 1. \tag{2.14}$$

In this dissertation, we select one state variable, e.g. the number of walk-in patients, to be the level variable, and consider the other state variable(s) phase variable(s), e.g. the numbers of admitted patients and ambulance patients, as phase variable(s). We construct an infinitesimal generator Q as the form in (2.7), based on which we derive stationary distribution vector and various performance measures.

2.5 Heavy traffic approximation

In this section, we review the basics of fluid and diffusion limits of queues, which becomes useful in the latter part of our modeling. We take the $GI/G/1$ queue as an example to illustrate the idea behind a heavy traffic approximation.

2.5.1 Constructing the queue-length and workload processes: a sample-path approach

Consider a queue with a single server. The queue starts with $Q(0)$ customers in the system. There is an input stream of customers arriving in the system. Denote by U_1 the arrival epoch of the first arrival and by U_i the inter-arrival time between the $(i-1)$th and ith arrivals, $i = 2, 3, \ldots$. Suppose that V_i is the time required by the ith customer to be served. Let

$$U(0) := 0, U(k) := \sum_{i=1}^{k} U_i, V(0) := 0, V(k) = \sum_{i=1}^{k} V_i;$$

and define two counting (jump) processes, namely the arrival process and the service process by

$$A(t) := \sup\{k | U(k) \leq t\}, \quad S(t) := \sum_{i=1}^{k} \{k | V(k) \leq t.\}$$

By definition, $A(t)$ counts the number of arrivals during $(0, t]$ while $S(t)$ gives number of customers that the server can potentially serve during $(0, t]$. Let $Q(t)$ be the number of customers in the system at time t and $B(t)$ denote the cumulative amount of time when the server is busy over the interval $[0, t]$. Under a work-conserving discipline, e.g., the server cannot be idle as long as a customer stays in the system, then $Q(t)$, the queue-length process $Q(t)$ and $B(t)$ jointly satisfy the relations

$$Q(t) = Q(0) + A(t) - S(B(t)), \tag{2.15}$$

$$B(t) = \int_0^t \mathbf{1}_{Q(s)>0} \mathrm{d}s \tag{2.16}$$

Based on the construction of process $B(t)$, another process I, termed as idle time process, can be expressed as:

$$I(t) = t - B(t) = \int_0^t \mathbf{1}_{Q(s)=0} \mathrm{d}s. \tag{2.17}$$

From the definition of $V(k)$, it is easy to see that $V(Q(0) + A(t))$ gives the total amount of time required by all customers that show up in the system up to time t. Therefore the workload, e.g., the amount of time required to clear up all customers that present at time t, would be calculated as

$$Z(t) = V(Q(0) + A(t)) - B(t).$$

2.5.2 Reflection mapping

The fact that $A(t)$ and $S(t)$ are asymptotically close to μt and λt with mild regularity conditions motivates the use of a 'centering' technique to the queue-length process. By writing $Q(t)$ as

$$Q(t) = X(t) + Y(t), \tag{2.18}$$

where

$$\begin{aligned} X(t) &:= Q(0) + (\lambda - \mu)t + [A(t) - \lambda t] - [S(B(t)) - \mu B(t)], \\ Y(t) &:= \mu I(t), \end{aligned}$$

one establishes the following relations:

$$Q(t) \geq 0, \tag{2.19}$$

$$\mathrm{d}Y(t) \geq 0, Y(0) = 0, \tag{2.20}$$

$$Q(t)\mathrm{d}Y(t) = 0. \tag{2.21}$$

A common approach for establishing limit theorems is known as reflection mapping, see e.g., Theorem 6.1 in Chen and Yao (2001), which says that $\forall x \in \mathcal{D}$, $\exists!(y, z) \in \mathcal{D}^2$ s.t.

$$z = x + y \geq 0 \tag{2.22}$$

$$\mathrm{d}y \geq 0, y(0) = 0, \tag{2.23}$$

$$z\mathrm{d}y = 0. \tag{2.24}$$

More precisely,

$$y(t) = \sup_{0 \leq s \leq t} [-x(s)]^+, \tag{2.25}$$

$$z(t) = x(t) + \sup_{0 \le s \le t} [-x(s)]^+. \tag{2.26}$$

Denote the mappings by $y = \Psi(x)$ and $z = \Phi(x)$. It can be shown that both mappings are Lipschitz continuous. Applying the above result to the dynamic system described through (2.18), (2.19), (2.20) and (2.21), one has $Q = \Phi(X)$ and $I = \frac{1}{\mu}\Psi(X)$. Intuitively, if X can be approximated by a limiting process $\mathcal{X}$, then Q can be approximated using $\Phi(\mathcal{X})$.

2.5.3 Fluid and diffusion limit

To establish a fluid approximation of the single-server queue, one starts with an assumption on the two primitive processes $A(t)$ and $S(t)$, which states that

$$\bar{A}_n(t) \to \lambda t, \quad \bar{S}_n(t) \to \mu t, u.o.c., \quad n \to \infty, \tag{2.27}$$

where

$$\bar{A}_n(t) = \frac{1}{n}A(nt), \quad \bar{S}_n(t) = \frac{1}{n}S(nt), \tag{2.28}$$

and 'u.o.c.' denotes uniformly on compact subsets.

Denote by $\{Q_n, n = 1, 2, \ldots\}$ a sequence of queues and $Q_n(0)$ the initial queue length of the nth queue. Let Z_n and B_n be the corresponding performance measures. With an additional assumption that

$$\bar{Q}_n(0) := \frac{1}{n}Q_n(0) \to \bar{Q}(0), \qquad n \to \infty,$$

A fluid approximation is concerned with the limits of

$$\bar{Q}_n(t) := \frac{1}{n}Q_n(nt), \bar{Z}_n(t) := \frac{1}{n}Z_n(nt), \bar{B}_n(t) := \frac{1}{n}B_n(nt),$$

and it says

$$(\bar{Q}_n, \bar{Z}_n, \bar{B}_n) \to (\bar{Q}, \bar{Z}, \bar{B}), \quad u.o.c., \quad n \to \infty, \tag{2.29}$$

where

$$\bar{Q} = [\bar{Q}(0) + (\lambda - \mu)t]^+, \tag{2.30}$$

$$\bar{Z} = \frac{1}{\mu}\bar{Q}, \tag{2.31}$$

$$\bar{B} = t - \frac{1}{\mu}[-\bar{Q}(0) - (\lambda - \mu)t]^+. \tag{2.32}$$

Just as central limit theorem is a refinement of strong law of large numbers, a diffusion approximation is a refinement of fluid approximation and concerned with the weak convergence of the processes

$$\hat{Q}_n(t) := \sqrt{n}[\bar{Q}_n(t) - (\lambda - \mu)^+ t], \tag{2.33}$$

$$\hat{Z}_n(t) := \sqrt{n}[\bar{Z}_n(t) - (\rho - 1)^+ t], \tag{2.34}$$

$$\hat{B}_n(t) := \sqrt{n}[(\rho \wedge 1)t - \bar{B}_n(t)]. \tag{2.35}$$

If

$$(\hat{A}_n, \hat{S}_n) \xrightarrow{d} (\hat{A}, \hat{S}), \quad n \to \infty \tag{2.36}$$

holds with

$$\begin{aligned} \hat{A}_n(t) &:= \sqrt{n}[\bar{A}_n(t) - \lambda t], \\ \hat{S}_n(t) &:= \sqrt{n}[\bar{S}_n(t) - \mu t], \end{aligned}$$

a functional central limit theorem for $(\hat{Q}_n(t), \hat{Z}_n(t), \hat{B}_n(t))$ is established as (see e.g., Theorem 6.8 in Chen and Yao (2001)):

$$(\hat{Q}_n, \hat{Z}_n, \hat{B}_n) \xrightarrow{d} (\hat{Q}, \hat{Z}, \hat{B}), \tag{2.37}$$

where the form of $(\hat{Q}, \hat{Z}, \hat{B})$ depends on utilization ρ. In particular, for $\rho = 1$,

$$\hat{Q} = \hat{A} - \hat{S} + \Psi(\hat{A} - \hat{S}), \tag{2.38}$$

$$\hat{Z} = \frac{1}{\mu}\hat{Q}, \tag{2.39}$$

$$\hat{B} = \frac{1}{\mu}\Psi(\hat{A} - \hat{S}). \tag{2.40}$$

2.5.4 Approximations for $GI/G/1$ queue

Suppose that the inter-arrival times U_i and service times V_i are independent identically distributed random variables with a finite variance (2.27). Further denote by C_a and C_s the coefficients of variation of U_i and V_i, respectively. It follows from (2.27) that

$$\begin{aligned} \hat{A}_n(t) = \sqrt{n}[\frac{1}{n}A(nt) - \lambda t] &\overset{d}{\approx} \hat{A}(t), \\ \hat{S}_n(t) = \sqrt{n}[\frac{1}{n}S(nt) - \mu t] &\overset{d}{\approx} \hat{S}(t), \end{aligned} \tag{2.41}$$

where '$\overset{d}{\approx}$' denotes 'approximately equal in distribution'. Replacing nt by t and multiply both sides by $\sqrt{n}$, we rewrite the above equation by

$$\begin{aligned} A(t) - \lambda t &\overset{d}{\approx} \sqrt{n}\hat{A}(t/n) \overset{d}{=} \hat{A}(t), \\ S(t) - \mu t &\overset{d}{\approx} \sqrt{n}\hat{S}(t/n) \overset{d}{=} \hat{S}(t) \end{aligned} \tag{2.42}$$

where the equality comes from the scaling property of Brownian motion. Recalling (2.18) and the fluid limit for $B(t)$, the approximations in the above equation leads to an approximation for X, which is

$$\hat{X}(t) = Q(0) + \hat{A}(t) - \hat{S}((\rho \wedge 1)t) + (\lambda - \mu)t. \tag{2.43}$$

It is easy to observe that $\hat{X}$ is a Brownian motion starting at $Q(0)$ with drift parameter $\lambda - \mu$ and variance $\sigma^2 = \lambda C_a^2 + (\lambda \wedge \mu) C_s^2$. And hence by reflection mapping, the queue-length process can be approximated by a reflected Brownian motion, denoted as $RBM_{Q(0)}(\lambda - \mu, \lambda C_a^2 + (\lambda \wedge \mu) C_s^2)$. It follows from the fact that a reflected Brownian motion with a negative drift has a stationary distribution, which is exponentially distribution, that the stationary distribution of the queue length can be approximated by an exponential distribution with rate

$$\frac{2(1-\rho)}{\rho(C_a^2 + C_s^2)}.$$

Chapter 3

Modeling of a Centralized Emergency Care Delivery System

We formulate a parsimonious representation of the patient flow through the emergency care network in a geographical region to capture the essential characteristics of a real system. We mention that the modeling approach adopted here can be applied to a real system with arbitrary number of hospitals. For the cases with three or more hospitals however, one needs to specify a destination policy that dictates which neighboring facility should the ambulance be diverted to if the corresponding ED is full while there are more than one facilities that have ED beds available. One could follow a *Nearest Hospital* policy as is described in Ramirez-Nafarrate et al. (2011). Suppose that we are dealing with three hospital EDs labeled as a, b and c, each of which owns a catchment area in the region, and at the moment a patient enters the system from area a. If ED a goes on diversion while ED b and c do not, the patient would be diverted to ED b, given b is closer to a. If ED b also goes on diversion while c does not, then the patient would be diverted to ED c even b is closer to a. If all hospitals go on diversion, then ED a must accept the patient even it goes on diversion. Without losing generosity and operational insight, we narrow our analysis to a network of two hospitals.

3.1 Model description

Fig.3.1 depicts the model of patient flows through the regional emergency care delivery system where there is a common EMS provider and two hospitals, each of which has its own catchment area in the region. All hospitals in the model share a similar structure. They include an ED and one ID (ID). While it is conceivable that the network shown in Fig.3.1 could be an idealized model of an actual operation, the reader will be best served by thinking of this model as the simplest network model that is rich enough to reproduce the variety of interactions between all the system components.

3.1.1 Priority discipline

We view each ED as a multi-server station. Each server can be considered as the combination of resources (e.g. nurse, physician, and bed) needed to care for an individual patient. ED $r(r = 1, 2)$

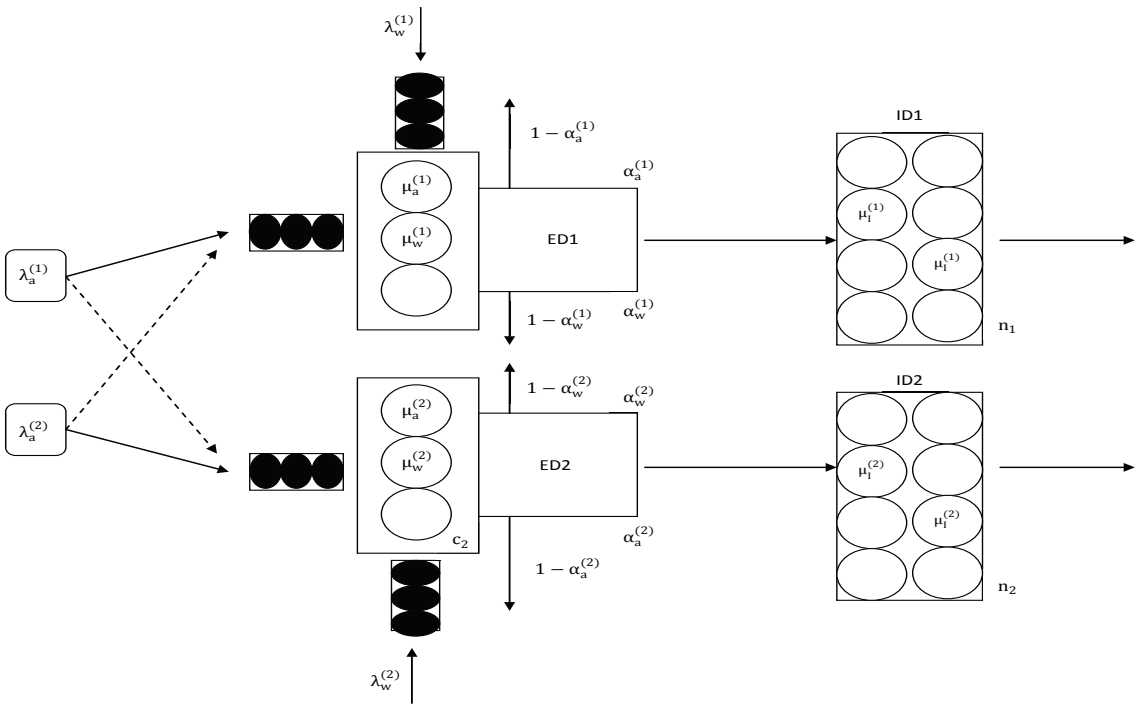

Figure 3.1.1: The queueing network model.

has c_r servers and each server operates independently of others. From the ED perspective, there are two arrival streams: ambulance patients and walk-in patients. When patients arrive at an ED, they are triaged according to the illness acuity. Normally, patients who arrive by ambulance are in life-threatening situation or a potential threat to life, and therefore require immediate medical attention. For this reason we have assumed that ambulance patients have preemptive priority over walk-in patients. By this assumption, the treatment of a walk-in patient will be preempted when an ambulance patient arrives while all servers are busy. In real applications, a walk-in patient can be diagnosed to be in severe condition and assigned the same acuity level as an ambulance patient. Since this group of patients constitute a small fraction of walk-in patients, and we simplify our model by making this simplifying assumption.

3.1.2 Admission and patient boarding

After ED service is complete, an ED patient will either be discharged directly from the ED or be admitted into the inpatient unit. The admission probability is associated with the patient severity and hence connects to the arrival mode. We use α_a and α_w to denote the admission probability of ambulance patients and walk-in patients respectively. It follows that $(1-\alpha_a)$ of ambulance patients and $(1-\alpha_w)$ of walk-in patients will be discharged immediately from the ED upon service completion.

We consider each ID as a multi-server station where n_r denotes the total number of inpatient beds in ID r. An ID receives two streams of patients: emergency patients who are admitted from the ED and patients who are admitted directly. We inherit the assumption used in Allon et al. (2013) that a fraction of the inpatient beds are reserved for emergency patients. Without introducing additional notations, we let n_r be the number of beds reserved for the emergency patients in ID r. An admitted patient is blocked in the ED if the ID does not have beds available to accommodate new patients. In other words, an admitted patient has to stay in the ED and wait for an inpatient bed to become available, even though ED service is complete. We refer this kind of patients as boarding patients and assume that they cannot be bumped out by new ambulance patients. Boarding patients reduce ED efficiency by consuming ED beds and nurses. ED boarding not only impedes patient flow but also raises safety issues that come up as is observed in practice if boarding patients are placed in the hallway (Hall (2006)). This situation can be modelled by adding a finite buffer between the ED and the ID. However, to maintain as much simplicity as possible, we assume a zero buffer for our model.

3.1.3 Diversion policy

To ensure access to emergency care, the local EMS system develops a diversion policy that governs transport-to-destination once an ambulance request enters the system. We first assume that a hospital ED goes on and off diversion based on the number of ambulance patients and boarding patients. As diversion should occur only after the hospital has exhausted all internal resources, we assume that each ED uses its capacity (number of servers) as its diversion threshold. In addition, we follow, in our modeling, a coordination guideline as is employed in most EMS agencies referred as " All on Diversion, Nobody on Diversion" (Deo and Gurvich (2011)), which implies that all hospitals must open when the entire health care system is overloaded. We label the catchment area of first hospital as area 1 and that of the second as area 2. Suppose all the servers in ED 1 have been occupied by ambulance or boarding patients, while ED 2 has extra capacity to accept new ambulance arrivals, then a demand from area 1 will be diverted to ED 2, given there is an ambulance available. Otherwise ED 1 is forced to accept the patient. To incorporate the effect of ED overcrowding on ambulance availability, we adopt an approach very similar to the work of Almehdawe et al. (2013) where the local EMS provider operates N ambulances in total, and an ambulance is tied up (offload delays) if the ED is full when the new ambulance patient arrives.

3.1.4 Arrival and service pattern

Since we intend to construct a Markovian queueing model which is easy to analyze, we make the following assumptions with respect to arrival patterns and service times. We assume that ambulance requests are generated independently in two catchment areas according to Poisson processes with

rates $\lambda_a^{(1)}$ and $\lambda_a^{(2)}$ respectively. We also assume the arrivals of walk-in patients at ED r follow Poisson process with rate $\lambda_w^{(r)}$. In practice, patient arrivals may be nonhomogeneous and depends on the time of the day, day of the week, or presents seasonal variations. But if we consider a short period of time, say half an hour (see the arrival rate curve to ED in Shi et al. (2012)), the Poisson process is a reasonable approximation. We also assume that the service time for an ambulance patient at ED r is an exponential random variable with parameter $\mu_a^{(r)}$ while the service time for a walk-in patient at ED r is exponentially distributed with parameter $\mu_w^{(r)}$. Inpatient LOS at hospital r is subject to exponential distribution with parameter $\mu_I^{(r)}$. It is worth noting that an inpatient normally has length of stay in days and an ED patient typically has length of stay in hours whereas the transit time to the hospital is at a magnitude of minutes (Budge et al. (2010)), which is fairly small by comparison. Therefore we neglect the transit time of ambulances yet still consider the potential offload delays in front of an ED. This simplification allows us to obtain important operational insights without overly complicating our analysis. More importantly, the biggest challenge facing today's EMS providers is ambulance offload delays causes by insufficient hospital beds and subsequent ED overcrowding and patient boarding Eckstein and Chan (2004), Almehdawe et al. (2013)).

3.1.5 System variables and performance measures

To describe the state of a hospital over time, one needs to keep track of the numbers of ambulance patients, walk-in patients and admitted patients. As we consider a emergency care system with two hospitals where ambulance transit time is negligible, there needs to be six state variables so as to obtain a complete description of the system. For $r \in \{1, 2\}$, we denote the system variables as follows:

1. $q_a^{(r)}(t)$: The number of ambulance patients both waiting and in service in hospital r at time t.
2. $q_w^{(r)}(t)$: The number of walk-in patients both waiting and in service in hospital r at time t.
3. $q_I^{(r)}(t)$: The number of admitted patients in hospital r, including patients in the ID and boarding patients in the ED at time t.

For $r \in \{1, 2\}$, if we denote the number of boarding patients in ED r and the number of ambulances experiencing offload delays in front of it as $q_B^{(r)}$ and $q_O^{(r)}$, then $q_B^{(r)} = \max(0, q_I^{(r)} - n_r)$ and $q_O^{(r)} = \max(0, q_B^{(r)} + q_a^{(r)} - c_r)$. The exponential assumptions imposed on patient arrivals and service times entail that the process $X(t) \equiv \{(q_a^1(t), q_w^1(t), q_I^1(t), q_a^2(t), q_w^2(t), q_I^2(t))\}$ is a CTMC. If further $X(t)$ satisfies ergodicity, then there exists a stationary distribution for the process $X(t)$. Frow now on, P denotes the probability measure on X in steady state without indicated otherwise. We define the *effective diversion probability* $P_D^{(r)}$ for $r \in \{1, 2\}$ and the *loss probability* P_L due to ambulance

unavailability:

$$\begin{aligned} P_D^{(1)} &= P\{q_a^{(1)} + q_B^{(1)} \geq c_1, q_a^{(2)} + q_B^{(2)} < c_2\}, \\ P_D^{(2)} &= P\{q_a^{(2)} + q_B^{(2)} \geq c_2, q_a^{(1)} + q_B^{(1)} < c_1\}, \\ P_L &= P\{q_O^{(1)} + q_O^{(2)} = N\}, \end{aligned} \tag{3.1}$$

We show through the following theorem, a steady-state distribution for the CTMC is difficult to characterized analytically, therefore we resort to other approaches that will be discussed in later chapters.

Theorem 3.1.1. *The CTMC is not reversible.*

Proof. By Kolmogorov criteria, to prove the irreversibility it is sufficient to find a number $L \in Z_+$ and a sequence of states $\{x_m \in Z_+^6, m = 0, \ldots, L\}$ such that $\prod_{m=1}^{L} q_{x_{m-1},x_m} q_{x_L,x_0} \neq q_{x_0,x_L} \prod_{m=1}^{L} q_{x_m,x_{m-1}}$, where q_{x_i,x_j} is the transition rate from x_i to x_j. To that end, consider the four states $x_0 = (c_1 + 2, 0, 0, c_2 - 1, 0, 0), x_1 = (c_1 + 1, 0, 0, c_2 - 1, 0, 0), x_2 = (c_1 + 1, 0, 0, c_2, 0, 0), x_3 = (c_1 + 2, 0, 0, c_2, 0, 0)$. Then

$$\prod_{m=1}^{3} q_{x_{m-1},x_m} q_{x_3,x_0} = c_1 \mu_a^{(1)} (\lambda_a^{(1)} + \lambda_a^{(2)}) \lambda_a^{(1)} c_2 \mu_a^{(2)},$$

while

$$q_{x_0,x_3} \prod_{m=1}^{3} q_{x_m,x_{m-1}} = (\lambda_a^{(1)} + \lambda_a^{(2)}) c_1 \mu_a^{(1)} c_2 \mu_a^{(2)} \cdot 0 = 0.$$

Thus the chain is irreversible. □

To conclude our discussion on system modeling, we summarize all main assumptions that have been made so far:

- Patients arrive at ED either by ambulance or self-transportation. An ambulance patient can preempt a walk-in patient in service if no ED server is available upon her arrival.
- Each ED patient will be admitted into the ID with a given probability. An admitted patient who cannot access into the inpatient unit due to lack of inpatient beds will stay in the ED and block an ED server.
- The EMS agency employs an "All on Diversion, Nobody on Diversion" policy.
- Ambulance transit time is zero.
- Each hospital ED goes on and off diversion based on the number of ambulance patients and boarding patients. Each ED sets its diversion threshold as its capacity and EDs don't game the system as in Deo and Gurvich (2011).
- An ambulance is tied up in the ED if there is no server available to treat ambulance patients.

- An ambulance patient is lost if all ambulances are tied up in hospital EDs.
- For each catchment area, (ambulance) demands are generated according to a Poisson process. Arrivals of walk-in patients follows a Poisson process.
- Service time for ED patients and inpatients are exponentially distributed.

3.2 Queueing decomposition

To construct an infinitesimal generator for the process $X(t)$ is feasible in theory. Yet to obtain a matrix-geometric solution using conventional matrix-analytic method is rather difficult, as there are two system variables each with a infinite number of states. Fortunately dimensional reduction can be achieved by decomposing the original network into subsystems which are mathematically tractable. By doing so, we are striking a balance between reducing the dimensionality and slightly changing the model.

Consider a situation where the practice of diversion is strictly prohibited and each hospital receives ambulance patients from its own catchment area. The flow of patients in both hospitals evolve *almost* independently except that they share the same EMS resources. It is easy to see the arrival pattern for ambulance patients at each ED resembles that of requests generated from the corresponding area. With the diversion mechanism added into the system, the EMS agency helps balance the supply-and-demand in a network of hospital EDs by re-routing demand from an overcrowded facility to a less crowded one. Therefore, the actual demands eventually taken to ED 1 consist of the non-diverted demands from area 1 and the demands diverted from area 2. The analogous situation holds for ED 2. Using the state variables, we express the instantaneous arrival rates of ambulance patients to ED 1 and ED 2 as

$$\begin{aligned}
&\hat{\lambda}_a^{(1)}(q_a^{(1)}, q_B^{(1)}, q_a^{(2)}, q_B^{(2)}) = \mathbf{1}\{\max(0, q_B^{(1)} + q_a^{(1)} - c_1) + \max(0, q_B^{(2)} + q_a^{(2)} - c_2) < N\} \\
&\times [\lambda_a^{(1)}(1 - \mathbf{1}\{q_a^{(1)} + q_B^{(1)} \geq c_1, q_a^{(2)} + q_B^{(2)} < c_2\}) + \lambda_a^{(2)}\mathbf{1}\{q_a^{(2)} + q_B^{(2)} \geq c_2, q_a^{(1)} + q_B^{(1)} < c_1\}], \\
&\hat{\lambda}_a^{(2)}(q_a^{(1)}, q_B^{(1)}, q_a^{(2)}, q_B^{(2)}) = \mathbf{1}\{\max(0, q_B^{(1)} + q_a^{(1)} - c_1) + \max(0, q_B^{(2)} + q_a^{(2)} - c_2) < N\} \\
&\times [\lambda_a^{(2)}(1 - \mathbf{1}\{q_a^{(2)} + q_B^{(2)} \geq c_2, q_a^{(1)} + q_B^{(1)} < c_1\}) + \lambda_a^{(1)}\mathbf{1}\{q_a^{(1)} + q_B^{(1)} \geq c_1, q_a^{(2)} + q_B^{(2)} < c_2\}].
\end{aligned} \tag{3.2}$$

Due to the 'admission control', the arrival rate to each ED is state-dependent, and thus the arrival process at each ED is no longer a Poisson process. But still we can define the *effective arrival rate* to each ED which is independent of the system state by

$$\begin{aligned}
\hat{\lambda}_a^{(1)} &= [(1 - P_D^{(1)})\lambda_a^{(1)} + P_D^{(2)}\lambda_a^{(2)}] \cdot (1 - P_L), \\
\hat{\lambda}_a^{(2)} &= [(1 - P_D^{(2)})\lambda_a^{(2)} + P_D^{(1)}\lambda_a^{(1)}] \cdot (1 - P_L),
\end{aligned} \tag{3.3}$$

where $P_D^{(1)}, P_D^{(2)}$ and P_L are defined in (3.1). By definition, $\hat{\lambda}_a^{(1)}$ and $\hat{\lambda}_a^{(2)}$ are contingent on $P_D^{(1)}, P_D^{(2)}$ and P_L. A problem arises when it comes to estimating $P_D^{(1)}, P_D^{(2)}$ and P_L. It is easy to see that an ID

affects its ED only through boarding patients. We take a step further by rewriting (3.1) using law of total probability:

$$
\begin{aligned}
P_D^{(1)} &= \sum_{i,j} P\{q_B^{(1)} = i, q_B^{(2)} = j\} P\{q_a^{(1)} \geq c_1 - i, q_a^{(2)} < c_2 - j | q_B^{(1)} = i, q_B^{(2)} = j\}, \\
P_D^{(2)} &= \sum_{i,j} P\{q_B^{(1)} = i, q_B^{(2)} = j\} P\{q_a^{(2)} \geq c_2 - j, q_a^{(1)} < c_1 - i | q_B^{(1)} = i, q_B^{(2)} = j\}, \\
P_L &= \sum_{i,j} P\{q_B^{(1)} = i, q_B^{(2)} = j\} P\{[0, q_a^{(1)} + i - c_1]^+ + [0, q_a^{(2)} + j - c_2]^+ = N | q_B^{(1)} = i, q_B^{(2)} = j\}.
\end{aligned}
\tag{3.4}
$$

For $r \in \{1, 2\}$, we define *effective ED capacity* as $\hat{c}_r \equiv c_r - q_B^{(r)}$. It follows that both $\hat{c}_1$ and $\hat{c}_2$ depend on the system state and therefore change over time. An ED goes on diversion whenever the number of ambulance patients reaches its effective ED capacity. Consider time periods when $\{q_B^{(1)} = i, q_B^{(2)} = j\}$. Removing all the time periods in which the numbers of blocked servers are not (i, j), then one sees (i, j) servers are always blocked. For such periods of time, the conditional diversion probabilities, given $(\hat{c}_1, \hat{c}_2)$, can be determined by solving a model that only has the EMS provider and hospital EDs with ED capacities being $(\hat{c}_1, \hat{c}_2)$. It remains to consider the marginal distribution of $(q_B^{(1)}, q_B^{(2)})$.

Here we slightly change the original model by seeing the input process of ambulance patients at each ED as a Poisson process with rate $\hat{\lambda}_a^{(1)}$ for ED 1 and $\hat{\lambda}_a^{(2)}$ for ED 2. With this small change, we can actually isolate a hospital and treat it as a two-station queue with two types of customers. The stationary distribution over $q_B^{(r)}$ becomes a performance measure for $r \in \{1, 2\}$ if the limiting distribution to the queue can be found. Noticeably, to determine $\hat{\lambda}_a^{(1)}$ and $\hat{\lambda}_a^{(2)}$ one still needs to approximate $P_D^{(1)}, P_D^{(2)}$ and P_L.

We find ourselves in an endless loop of finding $[P_D^{(1)}, P_D^{(2)}, P_L]$. To break the loop we develop an iterative approach to determine those values. As the first step towards our goal, we analyze a single hospital by describing it as a two-station queueing network in Chapter 4. Then we proceed in Chapter 5 by formulating a simplified network model that connects the EMS provider and multiple EDs without the IDs being involved.

Chapter 4

A Tandem Queue with ED Patient Boarding ("ED-ID" Model)

This chapter is dedicated to the modeling of a SINGLE hospital by constructing a two-station queueing network, which we refer as the "ED-ID" model. This model is introduced as a building block for analyzing the original network as in Fig.3.1. Yet it can be used by a hospital administrator who has an intimate knowledge of her facility to analyze the flow of patients. The model parameters defined in this chapter are essentially the same as those in Chapter 3 but all superscripts being omitted because we only consider one hospital.

4.1 Stochastic model

Emergency departments are the centrepiece of an emergency care delivery system. An ED can simultaneously be considered as a downstream station of EMS agency, and an upstream station of an ID (ID). A typical ED of a major public hospital has large patient volume and high utilization rate. It treats all types of patients, ranging from minor-ill to trauma patients (Hall (2006)). Fig.4.1 illustrates the flow of patients through the ED to ID.

We start by a brief summary of the assumptions with which we develop our model. (i) The ED has c servers and receives two types of patients - the ambulance patients and walk-in patients. (ii) Patient arrivals follows a Poisson process with rates λ_a for ambulance patients and λ_w for walk-in patients. An incoming ambulance patient preempts a walk-in patient in service whenever all servers are busy. (iii) There is finite buffer for ambulance patients with size B, while the queue length of walk-in patients can be infinite. (iv) ED service times are exponentially distributed with parameters μ_a and μ_w accordingly. (v) An ambulance (walk-in) patient is admitted into the inpatient unit with a probability of α_a (α_w). (vi) A patient is blocked in the ED if she is admitted while there is no inpatient beds available. (vii) Once an inpatient bed becomes available, the boarding patient at the head of the waiting line occupies the empty bed immediately. (viii) The ID has n beds in total and each bed acts as a server in the ID. Finally, we list all the parameters:

- λ_a: Arrival rate of ambulance patients to the ED

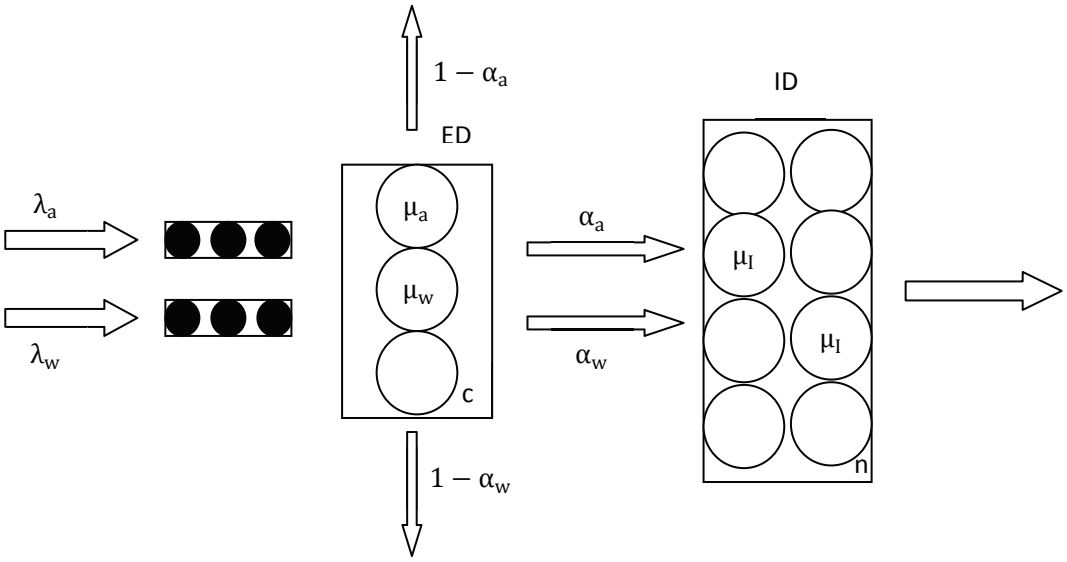

Figure 4.1.1: The subnetwork model.

- λ_w: Arrival rate of walk-in patients to the ED
- μ_a: Service rate per server for an ambulance patient at ED
- μ_w: Service rate per server for an walk-in patient at ED
- α_a: Admission probability for ambulance patients
- α_w: Admission probability for walk-in patients
- μ_I: Service rate per inpatient bed for both types of patients in the ID
- c: Number of servers in the ED
- n: Number of servers in the ID
- B: Buffer size for ambulance patients

To analyze the queue, we establish a CTMC. Under regularity conditions, we are able to find the stationary distribution which we later use to derive various performance measures. We start by introducing three state variables:

1. $q_a(t)$: The number of ambulance patients in service or waiting in the ED at time t.

2. $q_w(t)$: The number of walk-in patients in service or waiting in the ED at time t.

3. $q_I(t)$: The number of admitted patients including patients in the inpatient unit and boarding patient in the ED at time t.

For an arbitrary time t_0, the sojourn time for the state $[q_w(t_0), q_I(t_0), q_a(t_0)]$ is exponentially distributed. Therefore the process $Y(t) = \{[q_w(t), q_I(t), q_a(t)], t > 0\}$ is a CTMC. We organize the state space of the CTMC in such a way that q_w is the level variable, q_I and q_a are the phase variables. The infinitesimal generator takes the form of the following:

$$Q = \begin{pmatrix} A_{(0,0)} & A_0 & & & & & \\ A_{(1,0)} & A_{1,1} & A_0 & & & & \\ & A_{(2,1)} & A_{(2,2)} & A_0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & A_{(c-1,c-2)} & A_{(c-1,c-1)} & A_0 & \\ & & & & A_2 & A_1 & A_0 \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.1}$$

Based on this arrangement, $Y(t)$ is a quasi-birth-and-death (QBD) process. We provide details of the matrix blocks in Q and computational steps for the matrix-geometric solution in Appendix A.

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$ the stationary probability vector of the QBD process. The stationary distribution exists if and only if the CTMC is ergodic. Since the CTMC introduced is irreducible and has QBD structure, the theorem below is an immediate result of Neuts' theorem (Neuts (1981)).

Theorem 4.1.1. *If $\boldsymbol{\theta} A_0 \boldsymbol{e} < \boldsymbol{\theta} A_2 \boldsymbol{e}$, where $A = A_0 + A_1 + A_2$ and $\boldsymbol{\theta}$ satisfies $\boldsymbol{\theta} A = 0$ and $\boldsymbol{\theta} \boldsymbol{e} = 1$. There exists a unique non-negative vector $\boldsymbol{\pi}$ that satisfies the linear system:*

$$\boldsymbol{\pi} Q = 0, \qquad \text{and} \quad \boldsymbol{\pi} \boldsymbol{e} = 1. \tag{4.2}$$

Further, the matrix-geometric solution is given by

$$\begin{aligned} \boldsymbol{\pi}_i &= \boldsymbol{\pi}_{i-1} R, \quad \text{for } i \geq c, \\ \boldsymbol{\pi}_i &= \boldsymbol{\pi}_{i-1} R_i, \quad \text{for } 1 \leq i < c, \end{aligned} \tag{4.3}$$

where the rate matrix R is the minimal nonnegative solution to the nonlinear equation:

$$A_0 + R A_1 + R^2 A_2 = 0, \tag{4.4}$$

and the rate matrices $R_i, i = 1, \ldots, c-1$ can be solved recursively as

$$\begin{aligned} R_{c-1} &= A_0 (A_{(c-1,c-1)} + R A_2)^{-1}, \\ R_j &= A_0 (A_{(j,j)} + R_{j+1} A_{j,j+1})^{-1} \quad \text{for } j = 1, \ldots, c-2. \end{aligned} \tag{4.5}$$

The boundary probability $\boldsymbol{\pi}_0$ can be found through the boundary balance equation and the normalization condition:

$$\begin{aligned}&\boldsymbol{\pi}_0(A_{(0,0)} + R_1 A_{(1,0)}) = 0;\\ &\boldsymbol{\pi}_0(\boldsymbol{e} + R_1\boldsymbol{e} + \cdots + R_1\cdots R_{c-2}\boldsymbol{e} + R_1\cdots R_{c-1}(1-R)^{-1}\boldsymbol{e}) = 1.\end{aligned} \tag{4.6}$$

4.2 Performance measures

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$ the stationary distribution of the CTMC $Y(t)$. We are able to derive a number of performance measures in terms of $\boldsymbol{\pi}$. The performance measures are derived based on the assumption that the system has reached the state of equilibrium. Denote

$$\pi_{i,j,k} = P\{q_w = i, q_I = j, q_a = k\}.$$

1. The stationary distribution of q_I.
$$\pi_I(j) = \sum_{i,k} \pi_{i,j,k} \tag{4.7}$$

2. The probability distribution of the number of boarding patients in the ED. We slightly abuse the notation here by writing q_B as the number of boarding patients. We note that an admitted patient is blocked if and only if $q_I \geq n$. Naturally, one has $q_B = \max(0, q_I - n)$. The probability distribution of boarding patients can be expressed as follows:
$$\pi_B(m) = P(q_B = m) = \begin{cases} \sum_{j=0}^{n} \pi_I(j), \text{ for} & m = 0; \\ \pi_I(m+n), \text{ for} & 0 < m \leq c. \end{cases} \tag{4.8}$$

3. Boarding probability:
$$P_B = 1 - \pi_B(0). \tag{4.9}$$

4. The mean number of boarding patients can be calculated as
$$E[N_B] = \sum_{m=0}^{c} m \cdot \pi_B(m). \tag{4.10}$$

5. Define $\boldsymbol{\vartheta} = [\pi_I(n), \ldots, \pi_I(n+c-1)]/(1-\pi_I(n+c))$. The ith component of $\boldsymbol{\vartheta}$ gives the probability that a boarding patient has to wait for the service completion of i inpatients. It is easy to see that the boarding time b (the waiting period for an inpatient bed to become available) follows a generalized Erlang distribution with a phase-type representation $(\boldsymbol{\vartheta}, n\mu_I J_c)$, where
$$J_c = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}_{c\times c}. \tag{4.11}$$

The distribution function of b is given by

$$P(b < t) = 1 - \boldsymbol{\vartheta} \exp\{-n\mu_I J_c t\}\boldsymbol{e}. \tag{4.12}$$

By fundamentals of phase-type distribution, we obtain the mean boarding time $E[b]$ as:

$$E[b] = -\frac{1}{n\mu_I} \cdot \boldsymbol{\vartheta} J^{-1}\boldsymbol{e} \tag{4.13}$$

For now we derive the distribution over the number of boarding patients in the ED, along with boarding time distribution in terms of stationary distribution vector. We later relate those performance measures to the queueing decomposition method.

4.3 Waiting time distribution of walk-in patients

A walk-in patient can potentially be pushed out by ambulance patients for a number of times before her service is complete. Thus, attention here is naturally focused on the sojourn time distribution of a walk-in patient in the queue. The queue length process of walk-in patients is a QBD process, as is shown by (4.1) where transitions of the numbers of admitted patients and ambulance patients are encoded in the matrix blocks, e.g., $J(t) = [q_I(t), q_a(t)]$. However, in general the sojourn time distributions are complicated. For the ease of our analysis, we consider the time spent by a tagged walk-in patient waiting until the number of walk-in patients in front of her drops below a certain level. Specifically we choose the level to be equal to the ED capacity. We start by showing the stationary distribution of $Y(t)$ right after a walk-in patient enters the system.

We denote by $\hat{Y}(n) = [\hat{q}_w(n), \hat{q}_I(n), \hat{q}_a(n)]$ the state of the QBD process right after the nth arrival epoch. The process $\{\hat{Y}(n)\}$ then becomes a discrete-time $GI/M/c$ type Markov chain. Let $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \ldots)$ denote the the stationary distribution of $\{\hat{Y}(n)\}$. The following result is a direct consequence of Theorem 1 in Ozawa (2006).

Proposition 4.3.1. *$\hat{\boldsymbol{\pi}}$ is given as*

$$\begin{aligned}
\hat{\boldsymbol{\pi}}_1 &= \varphi\boldsymbol{\pi}_0 A_0, \\
\hat{\boldsymbol{\pi}}_l &= \varphi\boldsymbol{\pi}_{l-1} A_0 = \hat{\boldsymbol{\pi}}_{l-1} R_{l-1}, \quad 2 \le l \le c, \\
\hat{\boldsymbol{\pi}}_l &= \varphi\boldsymbol{\pi}_{l-1} A_0 = \hat{\boldsymbol{\pi}}_{l-1} R, \quad l > c,
\end{aligned} \tag{4.14}$$

where φ is the normalizing constant given as

$$\varphi = \{\boldsymbol{\pi}_0[I + R_1 + R_1R_2 + \cdots + R_1 \cdots R_{c-1}(I - R)^{-1}]\boldsymbol{e}\}^{-1}.$$

We proceed by considering the number of departures in $(0, t]$ denoted by $O(t)$ for those (walk-in) patients seen by the tagged customer upon her arrival. We introduce the matrix $P(k, t)$ with the (i, j)

element being $P(O(t)=k, J(t)=j|J(0)=i)$. Further we denote $D_0 = A_0 + A_1$ and $D_1 = A_2$. Given that the number of walk-in patients in front of the tagged customer is no less than the ED capacity, the matrices $P(k,t), k=0,1,\ldots$ satisfy the set of differential equations

$$\begin{aligned} \frac{\partial}{\partial t}P(0,t) &= P(0,t)D_0, \\ \frac{\partial}{\partial t}P(k,t) &= P(k,t)D_0 + P(k-1,t)D_1, \quad k \geq 1, \end{aligned} \tag{4.15}$$

with the initial conditions $P(0,0) = I$ and $P(k,0) = O, k \geq 1$, where O is a matrix of 0's.

We define W_c as the waiting time of the tagged walk-in patient to see the queue length in front of her drop below the ED capacity c. Noting that

$$P(W_c > t|\hat{q}_w \geq c) = \sum_{n=c}^{\infty} \hat{\boldsymbol{\pi}}_n \sum_{k=0}^{n-c} P(k,t)\boldsymbol{e},$$

we obtain the following result as an immediate result of Theorem 2 in Ozawa (2006)

Theorem 4.3.1. *The conditional complementary distribution of W_c given that the number of walk-in patients seen by the tagged patient is greater or equal to c is calculated by*

$$P(W_c > t|\hat{q}_w \geq c) = (\boldsymbol{e}^\top \otimes \boldsymbol{\kappa}) \exp([D_0 \otimes I + D_1 \otimes R]t)\boldsymbol{\xi}, \tag{4.16}$$

where

$$\boldsymbol{\kappa} = \boldsymbol{\pi}_0 R_1 \cdots R_{c-1}(I-R)^{-1},$$

and $\boldsymbol{\xi}$ is defined by

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_{d_A} \end{pmatrix}.$$

where d_A denotes the number of rows (columns) in matrix A, and $\boldsymbol{e}_j$ is a vector of length d_A with the jth element being 1 and other elements being zero.

From theorem 4.3.1 and total probability, we obtain the following.

Corollary 4.3.1. *The probability distribution of W_c can be expressed by*

$$F_{W_c}(t) = P(W_c \leq t) = \begin{cases} 0, & t < 0, \\ 1 - \boldsymbol{\kappa}\boldsymbol{e}, & t = 0, \\ 1 - \boldsymbol{\kappa}\boldsymbol{e} + \boldsymbol{\kappa}\boldsymbol{e}\{1 - (\boldsymbol{e}^\top \otimes \boldsymbol{\kappa}) \exp([D_0 \otimes I + D_1 \otimes R]t)\boldsymbol{\xi}\}, & t > 0. \end{cases} \tag{4.17}$$

As is shown in the corollary, the W_c has a mixture distribution, a finite mixture of a deteriorate distribution defined on a single point 0 and a matrix-exponential distribution with a support $(0, +\infty)$.

Chapter 5

A Model with Ambulance Diversions and Offload delays ("EMS-ED" Model)

In this chapter, we construct a simplified network model without the inpatient units being involved. More specifically we aim to investigate the patient flows from a single EMS provider into one of two EDs. To do that we consider the effective ED capacity $\hat{c}_r$ instead of the potential capacity c_r used in Chapter 3. Throughout this chapter, we see $\hat{c}_r$ as a fixed value rather than a random variable.

5.1 Stochastic model

To capture the characteristics of ambulance diversions and offload delays, we formulate a queueing network based on the following assumptions: (i) There are two catchment areas that generate requests for ambulance service. The demands are generated according to Poisson process with rate $\lambda_a^{(1)}$ and $\lambda_a^{(2)}$ respectively. Patients from area r will be transported to ED r if ED r does not go on diversion. (ii) The LOS of ambulance patients at ED r follows exponential distribution with parameter $\mu_a^{(r)}$. (iii) Ambulance patients have preemptive priority over walk-in patients. (iv) The time for an ambulance to transport a patient to an ED is negligible, e.g. the transit time is zero in this model. (v) Each ED uses $\hat{c}_r$ as the diversion threshold. Once the number of ambulance patients reaches the threshold value, the corresponding ED declares diversion status. (vi) If all EDs go on diversion, then all the EDs go off diversion. Since an ambulance patient can preempt a walk-in patient, the network can be considered as a single class queue from the point of view of high priority customers. To describe the system, we introduce two system variables as follows:

1. $q_1(t)$: The number of ambulance patients in service or waiting in ED 1 at time t.
2. $q_2(t)$: The number of ambulance patients in service or waiting in ED 2 at time t.

By assumption (v), if $q_r(t) \geq \hat{c}_r (r = 1, 2)$, then there is at least one ambulance patient waiting in the queue and $q_r(t) - \hat{c}_r$ ambulances experiencing ambulance offload delays at front of ED r. If $\sum_r \max[0, q_r(t) - \hat{c}_r] = N$ where N is the total number of ambulance the EMS provider operates, then any incoming request for ambulance service will be lost.

Consider the process $\{q_1(t), q_2(t)), t > 0\}$. It is easy to verify that $\{q_1(t), q_2(t)), t > 0\}$ is a CTMC. The fact that $q_1 + q_2 \leq N + \hat{c}_1 + \hat{c}_2$ implies $q_1(t)$ and $q_2(t)$ are finite. Therefore $\{(q_1(t), q_2(t)), t > 0\}$ is a CTMC with a finite state space. We construct the infinitesimal generator for the process $\{q_1(t), q_2(t)), t > 0\}$. We arrange the state space of the CTMC in such a way that $q_2(t)$ is the level variable while $q_1(t)$ is the phase variable. Because each of the state variables changes its value by at most one whenever an arrival or service completion occurs, $\{(q_1(t), q_2(t)), t > 0\}$ is a level dependent quasi-birth-and-death (QBD) process with a finite number of levels. The infinitesimal generator Q_N for the process $\{q_1(t), q_2(t)), t > 0\}$ has the following general structure:

$$Q_N = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ N + \hat{c}_2 - 1 \\ N + \hat{c}_2 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & & \cdots & N + \hat{c}_2 - 1 & N + \hat{c}_2 \end{array} \\ \begin{pmatrix} A_{(0,0)} & A_{(0,1)} & & & & \\ A_{(1,0)} & A_{(1,1)} & A_{(1,1)} & & & \\ & \ddots & \ddots & \ddots & & \\ & & A_{(N+\hat{c}_2-1,N+\hat{c}_2-2)} & A_{(N+\hat{c}_2-1,N+\hat{c}_2-1)} & A_{(N+\hat{c}_2-1,N+\hat{c}_2)} \\ & & & A_{(N+\hat{c}_2,N+\hat{c}_2-1)} & A_{(N+\hat{c}_2,N+\hat{c}_2)} \end{pmatrix} \end{array}. \tag{5.1}$$

We put the details of the matrix blocks in Q and computational steps for the matrix-geometric solution into Appendix B.

5.2 Performance measures

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{\hat{c}_2+N})$ the stationary distribution of the Markov chain $\{q_1(t), q_2(t)), t > 0\}$. We are able to derive a number of system performance measures based on the stationary distribution $\boldsymbol{\pi}$. Performance measures for ED 1 can be obtained by changing the role of ED 1 and ED 2 in the analysis. The performance measures mentioned above are derived based on the assumption that the system has reached the state of equilibrium. Denote

$$\pi_{i_1,i_2} = \hat{P}(q_1 = i_1, q_2 = i_2).$$

1. We define the random variable $q_O(t)$ as the number of ambulances in offload delays in the network at time t. By model description, there are ambulances experiencing offload delays if and only if there is at least one ED where the number of ambulance patients exceeds its total capacity. Therefore the probability distribution for the number of ambulances in offload delays can be calculated as:

$$\pi_O(m) = \hat{P}\{q_O = m\} = \begin{cases} \sum_{(i_1,i_2)\in\Omega: i_1 \leq \hat{c}_1, i_2 \leq \hat{c}_2} \pi_{i_1,i_2}, & \text{for } m = 0; \\ \sum_{(i_i,i_2)\in\Omega: \sum_{r=1}^{2} \max[0, i_r - \hat{c}_r] = m} \pi_{i_i,i_2}, & \text{for } 0 < m \leq N. \end{cases} \tag{5.2}$$

2. The loss probability: We define the loss probability as the chance of all ambulances being busy

(in transit or in offload delays), denoted as $\hat{P}_L$. Then the loss probability is calculated by

$$\hat{P}_L = \pi_O(m) = \sum_{(i_i,i_2)\in\Omega:\sum_{r=1}^{2}\max[0,j_r-\hat{c}_r]=N} \pi_{i_i,i_2}. \tag{5.3}$$

3. A particular interest in this simplified network is to see to what extent an ED puts on *effective* diversion. We define the effective diversion request as requests not ignored by the EMS provider. Therefore the diversion status of ED 2 is effective if and only if ED 1 does not declare diversion request and there are ambulance available to serve the emergency call. Let $\hat{P}_D^{(2)}$ denote the probability that ED 2 puts on effective diversion. Then $\hat{P}_D^{(2)}$ is calculated by

$$\hat{P}_D^{(2)} = \frac{\sum_{(i_1,i_2)\in\Omega:i_1<\hat{c}_1,\hat{c}_2\le i_2<\hat{c}_2+N}\pi_{i_1,i_2}}{1-P_L}. \tag{5.4}$$

Similar we can calculate the percentage of time that ED 1 is on effective diversion status.

$$\hat{P}_D^{(1)} = \frac{\sum_{(i_1,i_2)\in\Omega:\hat{c}_1\le i_1<\hat{c}_1+N,i_2<\hat{c}_2}\pi_{i_1,i_2}}{1-P_L}. \tag{5.5}$$

Now we have defined the stationary distribution over the number of ambulances experiencing offload delays, the loss probability, and more importantly the effective diversion probabilities for both EDs. We exploit those performance measures in the next chapter.

Chapter 6

An Iterative Queueing Decomposition Method

We describe some of our initial observations regarding the original network model before formalizing our queueing decomposition method. Suppose the whole system is now at equilibrium and now we shrink the size of the patient department in hospital 1 by removing some of the inpatient beds. The reduction of service capacity of ID 1 will result in more severe patient boarding at ED 1, making ED 1 go on diversion more frequently and divert more patients to ED 2. The reduced number of arrivals at ED 1 attenuates the impact of inpatient bed shortages at hospital 1. On the other hand, the increased number of arrivals at ED 2 leaves hospital 2 more heavily-loaded, resulting in more boarding patients and the ED 2 will be on diversion more frequently as well.

6.1 Computational procedure

To create an estimate for $P_D^{(1)}$, $P_D^{(2)}$ and P_L, we condition on the effective ED capacity, e.g., the number of unblocked ED servers. For each fixed pair $(\hat{c}_1, \hat{c}_2)$, we find the diversion and loss probabilities using the "EMS-ED" model. It follows from the law of total probability that

$$\begin{aligned}
\bar{P}_D^{(1)} &= \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \pi_B^{(1)}(i)\pi_B^{(2)}(j)\hat{P}_D^{(1)}(c_1-i, c_2-j), \\
\bar{P}_D^{(2)} &= \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \pi_B^{(1)}(i)\pi_B^{(2)}(j)\hat{P}_D^{(2)}(c_1-i, c_2-j), \\
\bar{P}_L &= \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \pi_B^{(1)}(i)\pi_B^{(2)}(j)\hat{P}_L(c_1-i, c_2-j),
\end{aligned} \tag{6.1}$$

where $\pi_B^{(r)}$ denotes the stationary distribution over the number of boarding patients in ED r while $\hat{P}_D^{(1)}$, $\hat{P}_D^{(2)}$ and $\hat{P}_L$ are defined by (5.3), (5.5) and (5.4) respectively.

It remains to determine $\pi_I^{(r)}$. To that end, we need to find $\hat{\lambda}_a^{(r)}$ and bring it into the "ED-ID" model. Recall that $\hat{\lambda}_a^{(1)}$ and $\hat{\lambda}_a^{(2)}$ are contingent on $P_D^{(1)}$, $P_D^{(2)}$ and P_L.

We propose an iterative approach for locating $[\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L]$. To facilitate our discussions, we define a mapping T that maps from $\mathbb{R}^2$ onto itself, where $\mathbb{R}^2$ is a two-dimensional Euclidean space. To

construct the mapping T, we introduce two mappings T_1 and T_2. We start with a vector $[\tilde{\lambda}_a^{(1)}, \tilde{\lambda}_a^{(2)}]$, potential with $[\lambda_a^{(1)}, \lambda_a^{(2)}]$, and substitute the element $\tilde{\lambda}_a^{(1)}$ ($\tilde{\lambda}_a^{(2)}$) into the "ED-ID" model to derive the stationary distribution $\pi_B^{(1)}$ ($\pi_B^{(2)}$). For each fixed pair $[c_1 - i, c_2 - j]$, we turn to solve the "EMS-ED" models and derive $\hat{P}_D^{(1)}$, $\hat{P}_D^{(2)}$, and $\hat{P}_L$ as performance measures. We calculate $[\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L]$ using (6.1).

$$T_1 : \mathbb{R}^2 \to \mathbb{R}^3 \quad [\tilde{\lambda}_a^{(1)}, \tilde{\lambda}_a^{(2)}] \to [\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L]. \tag{6.2}$$

We update the arrival rates $[\bar{\lambda}_a^{(1)}, \bar{\lambda}_a^{(2)}]$ by (3.3) with $[\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L]$ that is obtained from the previous step.

$$T_2 : \mathbb{R}^3 \to \mathbb{R}^2 \quad [\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L] \to [\bar{\lambda}_a^{(1)}(\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L), \bar{\lambda}_a^{(2)}(\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L)]. \tag{6.3}$$

Finally we define T as the a composition of mappings:

$$T = T_2 T_1 : \mathbb{R}^2 \to \mathbb{R}^2 \quad [\tilde{\lambda}_a^{(1)}, \tilde{\lambda}_a^{(2)}] \to [\bar{\lambda}_a^{(1)}, \bar{\lambda}_a^{(2)}]. \tag{6.4}$$

To establish a fixed-point theorem for T can be challenging. For now we assume that T has a unique fixed point, and denote by $[\lambda_a^{(1)*}, \lambda_a^{(2)*}]$, e.g.,

$$[\lambda_a^{(1)*}, \lambda_a^{(2)*}] = T([\lambda_a^{(1)*}, \lambda_a^{(2)*}]), \tag{6.5}$$

it follows that

$$\begin{aligned} \lambda_a^{(1)*} &= [(1 - \bar{P}_D^{(1)*})\lambda_a^{(1)} + \bar{P}_D^{(2)*}\lambda_a^{(2)}] \cdot (1 - \bar{P}_L^*), \\ \lambda_a^{(2)*} &= [(1 - \bar{P}_D^{(2)*})\lambda_a^{(2)} + \bar{P}_D^{(1)*}\lambda_a^{(1)}] \cdot (1 - \bar{P}_L^*), \end{aligned} \tag{6.6}$$

and

$$\begin{aligned} \bar{P}_D^{(1)*} &= \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \pi_B^{(1)*}(i)\pi_B^{(2)*}(j)\hat{P}_D^{(1)}(c_1 - i, c_2 - j), \\ \bar{P}_D^{(2)*} &= \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \pi_B^{(1)*}(i)\pi_B^{(2)*}(j)\hat{P}_D^{(2)}(c_1 - i, c_2 - j), \\ \bar{P}_L^{*} &= \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \pi_B^{(1)*}(i)\pi_B^{(2)*}(j)\hat{P}_L(c_1 - i, c_2 - j), \end{aligned} \tag{6.7}$$

where $\pi_B^{(1)*}$ ($\pi_B^{(2)*}$) is obtained by solving the "ED-ID" model using $\lambda_a^{(1)*}$ ($\lambda_a^{(2)*}$). Ideally $[\lambda_a^{(1)*}, \lambda_a^{(2)*}]$ can be close to $[\hat{\lambda}_a^{(1)}, \hat{\lambda}_a^{(2)}]$ and therefore $[\bar{P}_D^{(1)*}, \bar{P}_D^{(2)*}, \bar{P}_L]$ is close to $[P_D^{(1)*}, P_D^{(2)*}, P_L]$.

We formalize the computational procedure of find the effective diversion probabilities in the following algorithm.

Algorithm 6.1.1. *Computations of* $(\lambda_a^{(1)*}, \lambda_a^{(2)*})$.

1. *Set* $\hat{\lambda}_a^{(1)} := \lambda_a^{(1)}$ *and* $\hat{\lambda}_a^{(2)} := \lambda_a^{(2)}$.
2. *Solve the "ED-ID" model using* $\hat{\lambda}_a^{(1)}$ ($\hat{\lambda}_a^{(2)}$) *as an input parameter to obtain the stationary distribution* $\pi_B^{(1)}$ ($\pi_B^{(2)}$). *Then update* $(\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L)$ *using (6.1).*

3. *Update* $\hat{\lambda}_a^{(1)}$ *using (3.3) with* $(P_D^{(1)}, P_D^{(2)}, P_L)$ *replaced with* $(\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L)$.

4. *Repeat steps 2-3 until the difference of two consecutive values for* $\hat{\lambda}_a^{(1)}$ *and* $\hat{\lambda}_a^{(2)}$ *are sufficiently small, e.g., less or equal to a predetermined tolerance threshold* ε.

In the next section, we undertake 5 case studies in which we use Algorithm 6.1.1 to find $\hat{\lambda}_a^{(1)}$ and $\hat{\lambda}_a^{(2)}$ and the corresponding $(\bar{P}_D^{(1)}, \bar{P}_D^{(2)}, \bar{P}_L)$. Although we have no convergence proof for the algorithm, we haven't seen counterexamples where the algorithm fails to converge with a tolerance of $10E-5$.

6.2 Numerical investigations

We analyze 5 cases in which a regional emergency care delivery system is modeled by the queueing network described in Chapter 3. We first examine in case 1 the bottleneck effect of an ID on ED boarding and diversions by varying the number of inpatient beds of one hospital. We explore the network effect of ED diversions by drawing the diversion probability curves for both hospitals. The second case considers the impact of ED size on performance measures. Case 3 corresponds to the design scenario where one hospital introduces new discharge policies that speed up the discharge process of its inpatients. In case 4 and 5, sensitivity analysis is performed to assess the effects of input rates.

For each case we calculate various performance measures using the queueing decomposition. To benchmark the performance of our approximation, we rely throughout on discrete-event simulations and construct a 95% confidence intervals (CIs) for the diversion probabilities in each case. Other types of CIs, 99% CIs for example, can be constructed as well. In our case, the width of the 99% CIs is almost the same as that of the 95% CIs. Therefore we only display the 95% CIs for all cases. We compare the resulting diversion probability curves achieved from the queueing decomposition and simulations. As is seen in (3.3) an effective estimate of the diversion probability ensures the accuracy for estimating other performance measures.

6.2.1 Parameter selection

We lack operational data of a real regional emergency care system, we determine system parameters based on some prior knowledge. For example, medical reports emphasize that walk-in patients take up the most of the ED population. In some places walk-in patients account for 60% to 70% of the ED population. Other empirical studies suggest that the admission probabilities for ambulance patients and walk-in patients are approximately 40% and 15% respectively (see Hall (2006)). We mention that the parameter selection is somewhat subjective. Yet the numerical results capture the essential elements of the patient flow without losing operational insights. We believe that our method can be

used by agencies of a regional emergency care delivery system to aid them in strategic decision as they will have intimate knowledge of the system parameters.

The hospital sizes vary from place to place. For illustration purposes, we consider in all the cases a scale-down network with two hospitals each of which has a single ED and a single ID.

6.2.2 Case study 1

The system parameter used in this case are recorded in Table 6.2.2. In order to explore the impact of inpatient hospital size on the performance of the entire system, we vary the number of inpatient beds in the first hospital. In particular, we will examine what fraction of patients are *actually* diverted from one hospital to the other and how boarding probability and mean boarding time are affected by the size of the inpatient unit.

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(24, 24)$	N	8
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(36, 36)$	(c_1, c_2)	$(12, 12)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	$(0.75,\ 0.75)$	(n_1, n_2)	$(*, 24)$

Table 6.2.1: System parameters for case 1

	Queueing decomposition						Simulation	
	$P_D^{(1)}$	$P_D^{(2)}$	$P_B^{(1)}$	$P_B^{(2)}$	$E[b_1]$	$E[b_2]$	$P_D^{(1)}$	$P_D^{(2)}$
$n_1 = 20$	0.0997	0.0302	0.6770	0.3311	0.2674	0.0913	[0.0914 0.1004]	[0.0269 0.0306]
$n_1 = 21$	0.0721	0.0276	0.5546	0.2965	0.1973	0.0811	[0.0623 0.0695]	[0.0233 0.0271]
$n_1 = 22$	0.0502	0.0255	0.4353	0.2704	0.1402	0.0735	[0.0416 0.0476]	[0.0211 0.0246]
$n_1 = 23$	0.0340	0.0239	0.3287	0.2519	0.0965	0.0681	[0.0277 0.0315]	[0.0192 0.0225]
$n_1 = 24$	0.0227	0.0227	0.2394	0.2394	0.0645	0.0645	[0.0178 0.0208]	[0.0179 0.0212]
$n_1 = 25$	0.0152	0.0220	0.1686	0.2313	0.0419	0.0621	[0.0118 0.0142]	[0.0173 0.0199]
$n_1 = 26$	0.0103	0.0215	0.1151	0.2261	0.0266	0.0606	[0.0079 0.0098]	[0.0167 0.0198]
$n_1 = 27$	0.0072	0.0212	0.0763	0.2229	0.0165	0.0597	[0.0060 0.0071]	[0.0165 0.0197]
$n_1 = 28$	0.0054	0.0210	0.0491	0.2210	0.0100	0.0592	[0.0046 0.0054]	[0.0163 0.0192]

Table 6.2.2: Performance measures for case 1

The values reported in Table 6.2.2 are the diversion and boarding probabilities as well as mean

boarding times of two hospitals when the number of inpatient beds in hospital 1 increases from 20 to 28. This provides an interpretation of the bottleneck of the ID in the patient flow process. We find a monotonic decrease for all the measures of hospital 1. As we have expected, we see a downward trend in all the measures of hospital 2 as well, indicating a network effect of ED diversions. We append two columns on the right that correspond to the results obtained from discrete-event simulation.

A graphical comparison on the results of queueing decomposition method and discrete-event simulation is displayed in Fig.6.2.2.

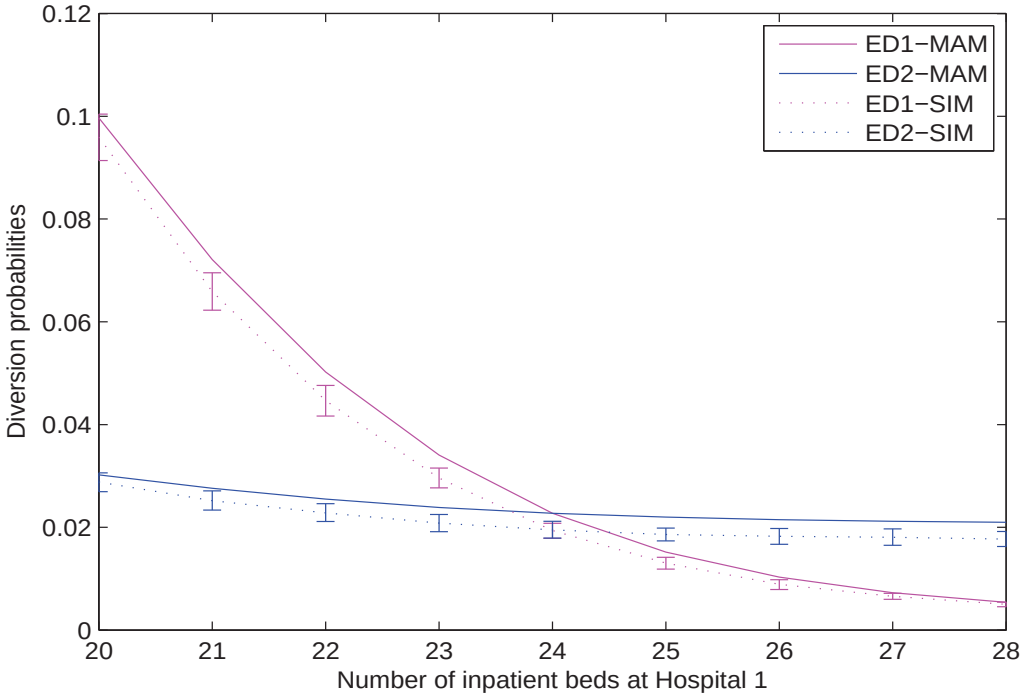

Figure 6.2.1: Case 1

As is observed, non of the solutions obtained from queueing decomposition fall into the 95% confidence interval constructed using simulations, yet the curves from the decomposition method follow the same pattern as that from simulations. Moreover the gap between two sets of curves are generally small. It is easily seen that for both hospitals the diversion probability decays almost exponentially in the number of inpatient beds, given other system parameters remain fixed.

6.2.3 Case study 2

In this case, we investigate the impact of ED size on system performance. To do so, we vary the ED capacity of hospital 1 while keeping other input parameters as a constant. System parameters

used are summarized in Table 6.2.3. The results, which are displayed in Table 6.2.3, show how various performance measures change in the number of ED servers.

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(24, 24)$	N	8
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(36, 36)$	(c_1, c_2)	$(*, 12)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	$(0.75,\ 0.75)$	(n_1, n_2)	$(24, 24)$

Table 6.2.3: System parameters for case 2

	Queueing decomposition						Simulation	
	$P_D^{(1)}$	$P_D^{(2)}$	$P_B^{(1)}$	$P_B^{(2)}$	$E[b_1]$	$E[b_2]$	$P_D^{(1)}$	$P_D^{(2)}$
$c_1 = 10$	0.0392	0.0245	0.2051	0.2578	0.0478	0.0698	[0.0340 0.0379]	[0.0198 0.0230]
$c_1 = 11$	0.0291	0.0234	0.2258	0.2464	0.0566	0.0665	[0.0241 0.0276]	[0.0184 0.0218]
$c_1 = 12$	0.0227	0.0227	0.2394	0.2394	0.0645	0.0645	[0.0181 0.0214]	[0.0179 0.0216]
$c_1 = 13$	0.0184	0.0223	0.2482	0.2348	0.0715	0.0631	[0.0143 0.0169]	[0.0174 0.0205]
$c_1 = 14$	0.0153	0.0220	0.2541	0.2314	0.0776	0.0622	[0.0113 0.0142]	[0.0172 0.0202]

Table 6.2.4: Performance measures for case 2

Again we identify a monotonic decrease for the faction of time on effective diversions for both hospitals, although the diversion probability of hospital 2 is not significantly affected by the change in ED capacity of hospital 1. Surprisingly however, the boarding probability and the mean boarding time increase in hospital 1 as one expand its ED size, while that in hospital 2 decrease slightly. A justification for this phenomena might be that more ambulance patients are routed from hospital 2 to hospital 1, leaving the inpatient unit at hospital 1 more congested. This is partly in conformance with the findings from previous studies that hospitals with a larger ED experience more severe access block to the inpatient unit.

Fig.6.2.3 displays the solutions from both queueing decomposition and simulations. Again a gap is noticeable between these two methods. The difference comes partly from the fact that we change by applying queueing decomposition the original network model. Recall the original network where an ambulance patient can always be assigned to an ED server whenever an ED has beds available. Using queueing decomposition, we assume the diversion probabilities to be routing probabilities in a network of queues. Hence we create a possibility that an ambulance patient is waiting at one facility while there

are servers being idle (in the eye of an ambulance patient) at the other one. Another source of error can potentially be the choice of stopping criterion in the iterative algorithm of queueing decomposition. Recall the threshold value ε specified before the algorithm is implemented.

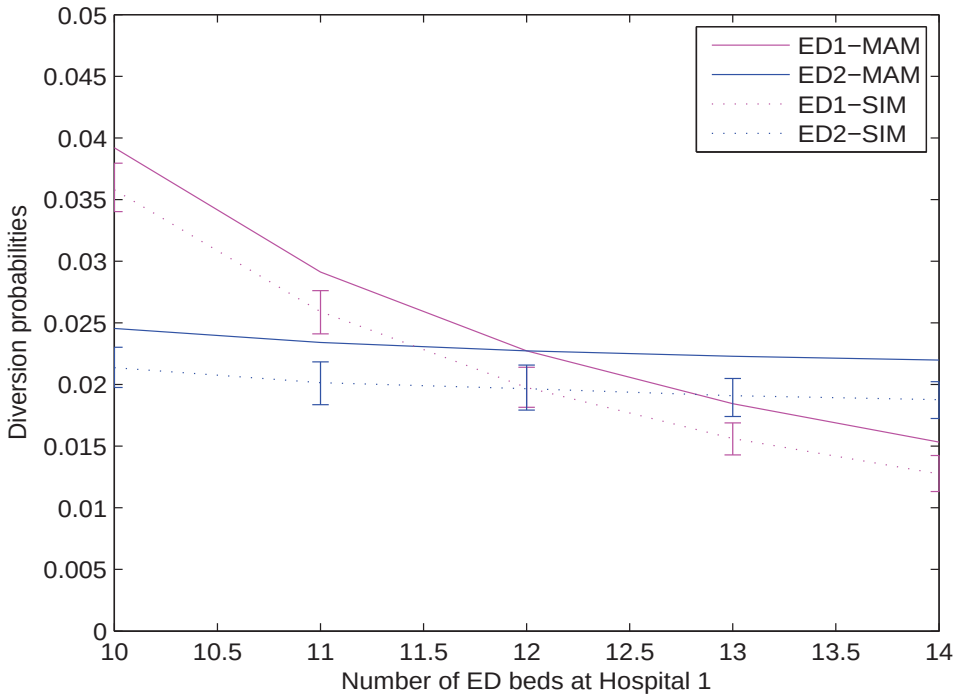

Figure 6.2.2: Case 2

One sees that the diversion probability of hospital 1 decays exponentially while the curve for hospital 2 is almost flat.

6.2.4 Case study 3

Inpatient discharge policies have been extensively discussed in previous studies. We present a case study that explores the impact of discharge rate on system performance. Still we look at various measures for both hospitals. we parameterize the model using the data in Table 6.2.4, and present the results in Table 6.2.4.

We observe a monotonic decrease for all the performance measures, as we speed up the discharge process. We find that a moderate improvement (14%) for the discharge process can achieve 66% reduction on diversion probability and 70% reduction on the mean waiting time for boarding patients.

We draw the plot of the diversion probability curves for both hospitals in Fig.6.2.4. We find that the diversion probabilities for both hospitals are concave and monotonically decreasing functions with

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(24, 24)$	N	8
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(36, 36)$	(c_1, c_2)	$(12, 12)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	(*, 0.75)	(n_1, n_2)	$(24, 24)$

Table 6.2.5: System parameters for case 3

	Queueing decomposition						Simulation	
	$P_D^{(1)}$	$P_D^{(2)}$	$P_B^{(1)}$	$P_B^{(2)}$	$E[b_1]$	$E[b_2]$	$P_D^{(1)}$	$P_D^{(2)}$
$\mu_I^{(1)} = 0.70$	0.0420	0.0247	0.3801	0.2610	0.1179	0.0707	[0.0344 0.0391]	[0.0199 0.0234]
$\mu_I^{(1)} = 0.71$	0.0372	0.0242	0.3483	0.2554	0.1049	0.0691	[0.0299 0.0344]	[0.0195 0.0226]
$\mu_I^{(1)} = 0.72$	0.0329	0.0238	0.3181	0.2507	0.0931	0.0677	[0.0263 0.0304]	[0.0191 0.0219]
$\mu_I^{(1)} = 0.73$	0.0291	0.0234	0.2900	0.2464	0.0825	0.0665	[0.0234 0.0269]	[0.0186 0.0221]
$\mu_I^{(1)} = 0.74$	0.0257	0.0230	0.2638	0.2426	0.0730	0.0654	[0.0206 0.0240]	[0.0183 0.0215]
$\mu_I^{(1)} = 0.75$	0.0227	0.0227	0.2394	0.2394	0.0645	0.0645	[0.0179 0.0208]	[0.0179 0.0210]
$\mu_I^{(1)} = 0.76$	0.0201	0.0225	0.2169	0.2365	0.0569	0.0636	[0.0159 0.0188]	[0.0182 0.0212]
$\mu_I^{(1)} = 0.77$	0.0178	0.0222	0.1962	0.2341	0.0502	0.0629	[0.0140 0.0165]	[0.0175 0.0207]
$\mu_I^{(1)} = 0.78$	0.0158	0.0220	0.1771	0.2320	0.0442	0.0623	[0.0127 0.0151]	[0.0174 0.0200]
$\mu_I^{(1)} = 0.79$	0.0141	0.0219	0.1597	0.2301	0.0389	0.0618	[0.0110 0.0132]	[0.0172 0.0204]
$\mu_I^{(1)} = 0.80$	0.0126	0.0217	0.1439	0.2285	0.0342	0.0613	[0.0098 0.0120]	[0.0171 0.0202]

Table 6.2.6: Performance measures for case 3

respect to the inpatient discharge rate in hospital 1, highlighting a positive interaction between two IDs.

6.2.5 Case study 4

In previous cases, we assume steady (time-invariant) input streams for both hospitals. In order to examine the impact of the time-varying factors on the network model, it is plausible to perform sensitivity analysis to the input processes. To do so, we vary the demand rate in the catchment area of hospital 1 while keeping other input parameters fixed. We also assume that the two hospitals have equal sizes (the same number of ED servers and inpatient beds). We then calculate the performance measures with different (ambulance) demand rates. We list the parameters in Table 6.2.5 and summarise the

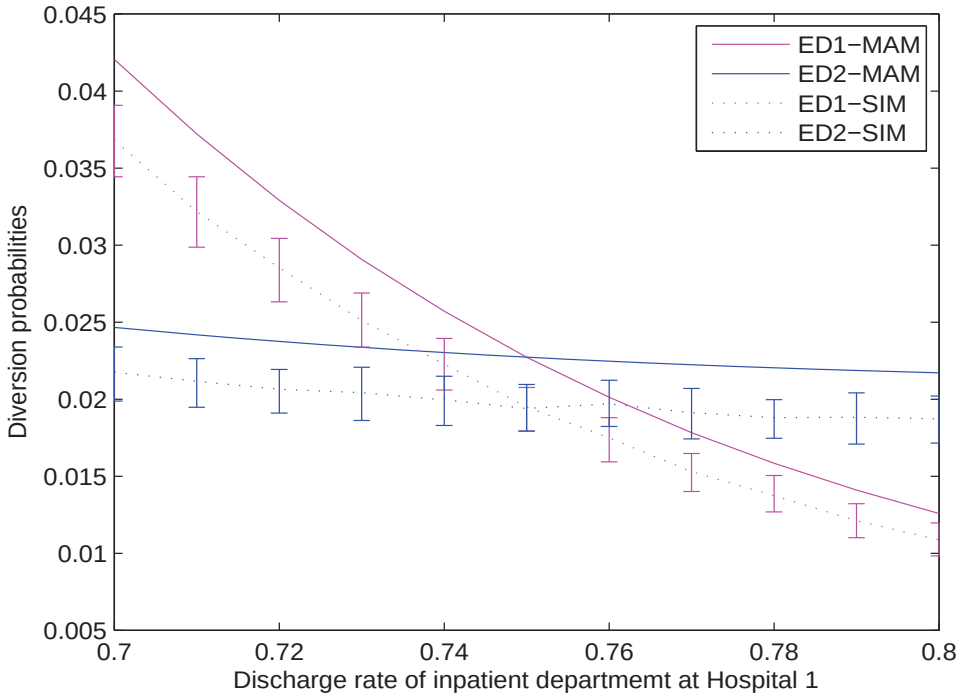

Figure 6.2.3: Case 3

results in Table 6.2.5.

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(*, 24)$	N	8
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(36, 36)$	(c_1, c_2)	$(10, 10)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	$(0.75,\ 0.75)$	(n_1, n_2)	$(24, 24)$

Table 6.2.7: System parameters for case 4

A close examination of Table 6.2.5 and Fig.6.2.5 suggests that diversion probability of hospital 1 can range from less than 1% to nearly 5% if $\lambda_a^{(1)}$ varies within $[21, 27]$. Boarding probability and mean boarding time triple as $\lambda_a^{(1)}$ increases from the lowest value to the highest. The set of measures at hospital 2 also increase, but not significantly.

	Queueing decomposition						Simulation	
	$P_D^{(1)}$	$P_D^{(2)}$	$P_B^{(1)}$	$P_B^{(2)}$	$E[b_1]$	$E[b_2]$	$P_D^{(1)}$	$P_D^{(2)}$
$\lambda_a^{(1)} = 21$	0.0081	0.0211	0.1270	0.2227	0.0327	0.0597	[0.0061 0.0076]	[0.0166 0.0198]
$\lambda_a^{(1)} = 22$	0.0118	0.0215	0.1602	0.2267	0.0419	0.0608	[0.0091 0.0112]	[0.0169 0.0200]
$\lambda_a^{(1)} = 23$	0.0166	0.0220	0.1979	0.2321	0.0526	0.0624	[0.0128 0.0154]	[0.0179 0.0206]
$\lambda_a^{(1)} = 24$	0.0227	0.0227	0.2394	0.2394	0.0645	0.0645	[0.0180 0.0210]	[0.0180 0.0213]
$\lambda_a^{(1)} = 25$	0.0301	0.0237	0.2839	0.2489	0.0774	0.0672	[0.0245 0.0280]	[0.0190 0.0222]
$\lambda_a^{(1)} = 26$	0.0388	0.0249	0.3305	0.2607	0.0912	0.0706	[0.0313 0.0362]	[0.0201 0.0231]
$\lambda_a^{(1)} = 27$	0.0486	0.0264	0.3781	0.2751	0.1053	0.0749	[0.0409 0.0461]	[0.0217 0.0253]

Table 6.2.8: Performance measures for case 4

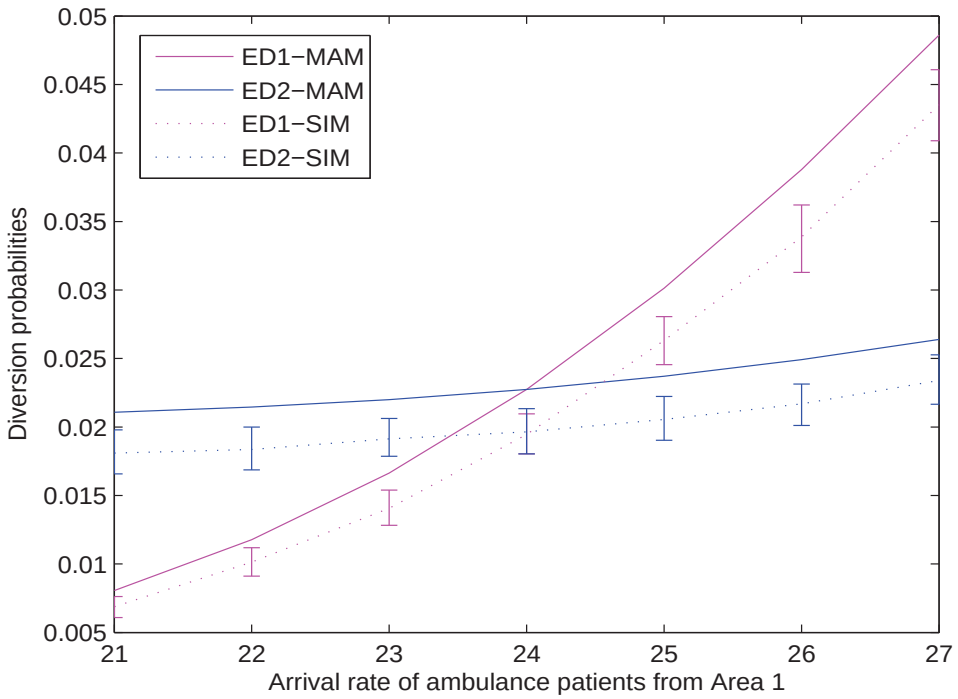

Figure 6.2.4: Case 4

6.2.6 Case study 5

We assume throughout our analysis that ambulance patients have preemptive priority over walk-in patients in the ED, which means that an ambulance patient sees no walk-in patients. On the other hand, both types of patients can potentially be admitted into the inpatient unit and patients who

cannot access into the inpatient unit will be backlogged in the ED. Therefore, we speculate that if a hospital ED is experiencing a higher volume of walk-in patients, an ambulance patient can be more easily diverted to neighboring hospitals as well. To test our hypothesis, we perform a sensitivity analysis to assess the impact of walk-in patients on ambulance diversion and ED boarding. To do so, we vary the arrival rate of walk-in patient in hospital 1 while fixing other input parameters. Again both hospitals are assume to have equal sizes. Table 6.2.6 shows the system parameters in this case.

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(24, 24)$	N	8
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(*, 36)$	(c_1, c_2)	$(10, 10)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	$(0.75,\ 0.75)$	(n_1, n_2)	$(24, 24)$

Table 6.2.9: System parameters for case 5

	Queueing decomposition						Simulation	
	$P_D^{(1)}$	$P_D^{(2)}$	$P_B^{(1)}$	$P_B^{(2)}$	$E[b_1]$	$E[b_2]$	$P_D^{(1)}$	$P_D^{(2)}$
$\lambda_w^{(1)} = 32$	0.0163	0.0221	0.1751	0.2325	0.0460	0.0625	[0.0129 0.0154]	[0.0174 0.0204]
$\lambda_w^{(1)} = 33$	0.0177	0.0222	0.1900	0.2340	0.0502	0.0629	[0.0141 0.0167]	[0.0176 0.0206]
$\lambda_w^{(1)} = 34$	0.0193	0.0224	0.2057	0.2357	0.0547	0.0634	[0.0153 0.0183]	[0.0176 0.0207]
$\lambda_w^{(1)} = 35$	0.0210	0.0226	0.2222	0.2374	0.0595	0.0639	[0.0167 0.0194]	[0.0178 0.0207]
$\lambda_w^{(1)} = 36$	0.0227	0.0227	0.2394	0.2394	0.0645	0.0645	[0.0181 0.0215]	[0.0179 0.0214]
$\lambda_w^{(1)} = 37$	0.0246	0.0229	0.2574	0.2415	0.0697	0.0651	[0.0190 0.0227]	[0.0177 0.0211]
$\lambda_w^{(1)} = 38$	0.0266	0.0231	0.2762	0.2436	0.0753	0.0657	[0.0210 0.0247]	[0.0181 0.0216]
$\lambda_w^{(1)} = 39$	0.0287	0.0233	0.2958	0.2460	0.0811	0.0664	[0.0232 0.0265]	[0.0187 0.0213]
$\lambda_w^{(1)} = 40$	0.0310	0.0236	0.3160	0.2485	0.0871	0.0671	[0.0247 0.0284]	[0.0187 0.0221]

Table 6.2.10: Performance measures for case 5

The results, which are summarized in Table 6.2.6 and Fig.6.2.6, indicate that the diversion probability of hospital 1 changes less drastically compared to the fourth case, ranging from 1.6% to 3.1% as $\lambda_w^{(1)}$ varies from 32 to 40. The diversion probability of hospital 2 is not significantly affected. Boarding probability of hospital 1 almost doubles as $\lambda_w^{(1)}$ increases from the lowest value to the highest, while the mean boarding time shows no drastic changes.

We end this section with a discussion of the key results that we have observed so far. If one hospital

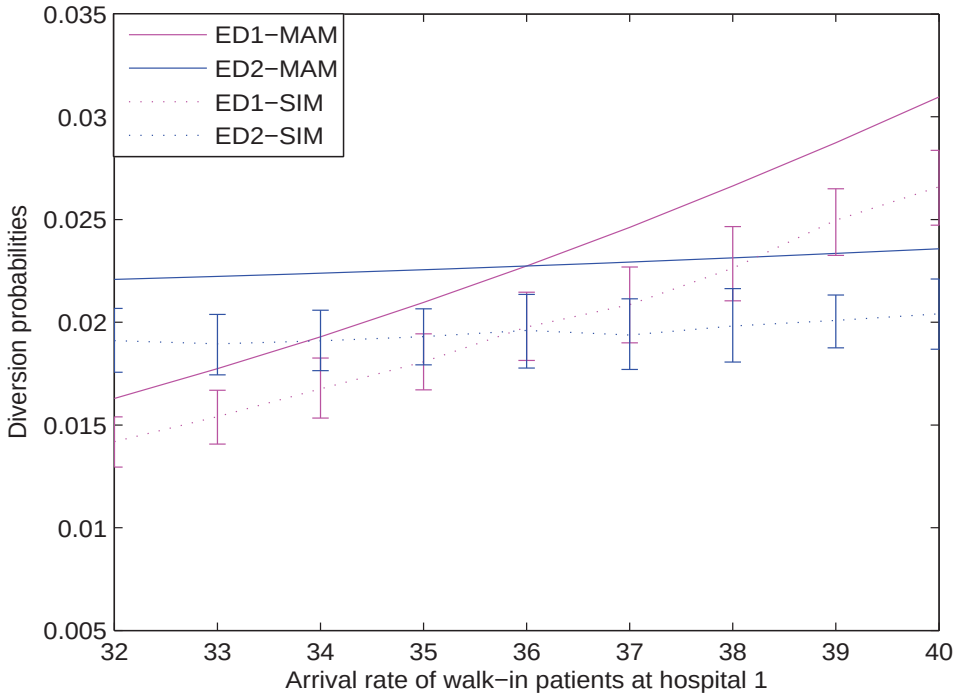

Figure 6.2.5: Case 5

expands its inpatient unit capacity or speeds up its inpatient discharge process, a less number of ambulances will be turned away from its catchment area and meanwhile the number of boarding patients can be reduced. If one hospital enlarges the size of its ED but keeps the inpatient unit capacity unchanged, the faction of time that the hospital ED is on effective diversion decreases but an admitted patient has a higher chance of being boarded in the ED. Finally higher arrival volumes of both types of patients can result in a drastic increase of ambulance diversions and boarded patients in the ED. The table below summarizes the outcomes of all cases that we investigate in the section.

Case #	The changing parameter	Impact on performance measures
1	$n_1 \uparrow$	$P_D^{(1)} \downarrow$ $P_D^{(2)} \downarrow$ $P_B^{(1)} \downarrow$ $P_B^{(2)} \downarrow$ $E[b_1] \downarrow$ $E[b_2] \downarrow$
2	$c_1 \uparrow$	$P_D^{(1)} \downarrow$ $P_D^{(2)} \downarrow$ $P_B^{(1)} \uparrow$ $P_B^{(2)} \downarrow$ $E[b_1] \uparrow$ $E[b_2] \downarrow$
3	$\mu_I^{(1)} \uparrow$	$P_D^{(1)} \downarrow$ $P_D^{(2)} \downarrow$ $P_B^{(1)} \downarrow$ $P_B^{(2)} \downarrow$ $E[b_1] \downarrow$ $E[b_2] \downarrow$
4	$\lambda_a^{(1)} \uparrow$	$P_D^{(1)} \uparrow$ $P_D^{(2)} \uparrow$ $P_B^{(1)} \uparrow$ $P_B^{(2)} \uparrow$ $E[b_1] \uparrow$ $E[b_2] \uparrow$
5	$\lambda_w^{(1)} \uparrow$	$P_D^{(1)} \uparrow$ $P_D^{(2)} \uparrow$ $P_B^{(1)} \uparrow$ $P_B^{(2)} \uparrow$ $E[b_1] \uparrow$ $E[b_2] \uparrow$

Chapter 7

Diffusion Approximation for the Inpatient Unit

We demonstrate the use of decomposition method for solving the network model as in Fig.3.1. Our numerical tests suggest that finding the matrix-geometric solution for the "EMS-ED" model is computationally expensive, especially when the hospital size is huge. Using a 2.50 GHz processor, we can only compute the solution for a model with up to 30 ED servers and 60 inpatient beds for each hospital. In the light of this observation, we consider diffusion approximations.

Diffusion models present tractable approaches to model unpredictable variability via first two moments of the service distribution rather than the entire distribution in classical queueing theory. Therefore we naturally generalize the exponential assumption on inpatient LOS to an arbitrary probability distribution. Although some clinical studies find the coefficient of variation of inpatient LOS is approximately equal to 1, suggesting the exponential assumption of LOS is very reasonable (Green and Nguyen (2001)), other studies suggest that inpatient LOS distribution can be more versatile. Marshall and McClean (2004) advocate the use of Coxian phase-type distribution in modeling inpatient duration in hospital. Kolker (2008) identifies log-logistic to be the best fit distribution for inpatient LOS. Shi et al. (2012) fit the LOS data with log-normal distribution.

7.1 Model specifications

We consider the queueing model as in Fig.4.1. To simplify our analysis, we add a few more assumptions regarding arrival pattern and patient blocking.

1. We approximate the arrival process of inpatients via ED by a Poisson process with rate $\lambda_I = \alpha_a\lambda_a + \alpha_w\lambda_w$. To see if this approximation works well, we refer readers to the work of Allon et al. (2013) for more detailed explanations.

2. We approximate the size of waiting buffer for ID by the number of ED servers c. This approximation is reasonable because all the boarding patients will stay in the ED while waiting for a bed, and there number cannot exceed the total number of ED servers.

We consider the ID as an $M/G/n/n+c$ queue with patient arrivals following Poisson process with rate λ_I, and a general distribution G for inpatient service duration. Allon et al. (2013) use a similar

approximation method as we describe here. However, their approximation and ours differ in several aspects. While they consider a diversion policy based on the number of boarding patients only, our model uses a diversion policy based on the number of boarding patients and high acuity patients in the ED. Their use of fluid and diffusion approximations aim to derive performance measures such as probabilities of diversion and patient delays in the ED, while we introduce diffusion approximation to obtain stationary distribution for the number of boarding patients, based on which we apply equation (6.1).

Attentions on approximations for the multiserver queues with a finite waiting room are few; e.g., see Whitt (1984), Whitt (2004) and references therein. Because we are more interested in the probability distribution for the number of boarding patients, we resort to Kimura (2003), and throw the technical details into Appendix C.

7.2 Numerical analysis

We analyze two examples numerically based on diffusion approximation. In both cases we consider an emergency care system that has two hospitals with different sizes. We assume that the larger one receives a higher patient volume. Similar to the first two cases that are discussed in the previous chapter, case 6 and case 7, which will be discussed shortly, investigate the impact of inpatient unit size and ED size on the system performance respectively.

7.2.1 Case study 6

We consider two hospitals and assume hospital 1 is larger than hospital 2 in terms of ED size and number of inpatient beds and therefore has higher patient volume. The system parameters used in this case are reported in Table 7.2.1. Performance measures are illustrated in Fig.7.2.1. We observe

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(88, 100)$	N	8
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(100, 135)$	(c_1, c_2)	$(27, 30)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	$(0.75,\ 0.75)$	(n_1, n_2)	$(*, 80)$

Table 7.2.1: System parameters for case 1

that from Fig.7.2.1 the fraction of time that hospital 1 is one diversion decreases from 4.5% to less than 2% if 7 inpatient beds are added based on the existing 87 beds. The diversion probability of

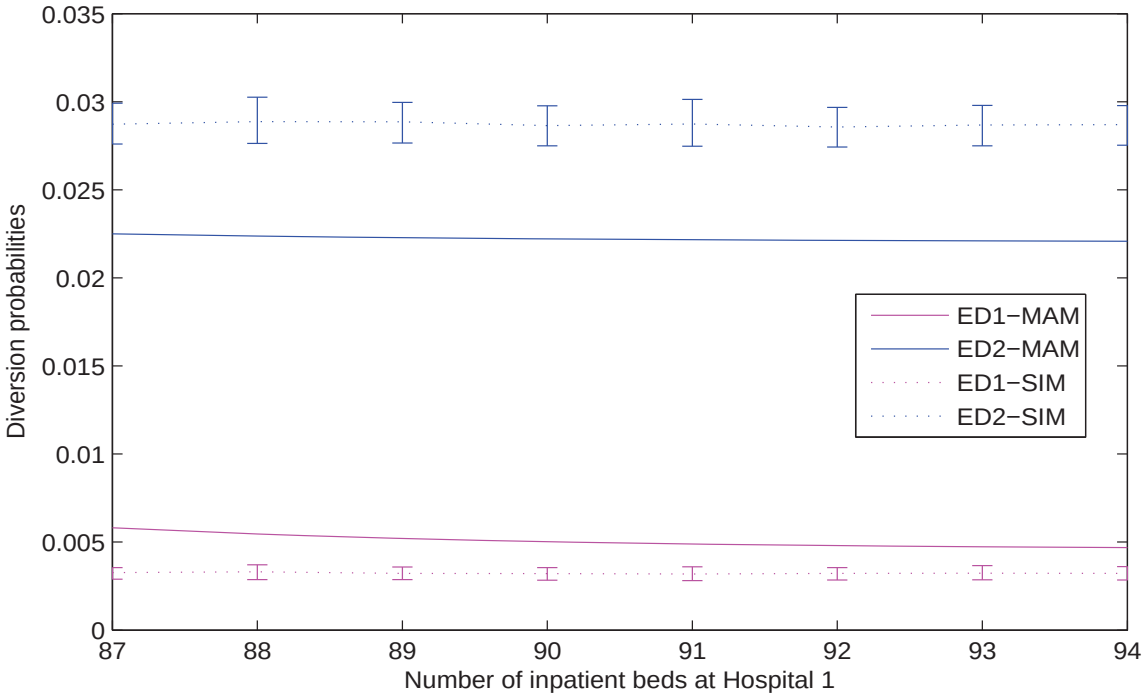

Figure 7.2.1: Case 6

hospital 2 also decreases, but very slightly.

7.2.2 Case study 7

Again we consider two hospitals with hospital 2 being smaller and receiving a lower patient volume. The system parameters used in this case are reported in Table 7.2.2. Numerical results are reflected in Fig. 7.2.2. We find that from Fig.7.2.2, the fraction of time that hospital 1 is one diversion decreases

System parameters	Value	System parameters	Value
$(\lambda_a^{(1)}, \lambda_a^{(2)})$	$(80, 100)$	N	12
$(\lambda_w^{(1)}, \lambda_w^{(2)})$	$(120, 135)$	(c_1, c_2)	$(*, 27)$
$(\mu_a^{(1)}, \mu_a^{(2)})$	$(5, 5)$	$(\alpha_a^{(1)}, \alpha_a^{(2)})$	$(0.4, 0.4)$
$(\mu_w^{(1)}, \mu_w^{(2)})$	$(15, 15)$	$(\alpha_w^{(1)}, \alpha_w^{(2)})$	$(0.15, 0.15)$
$(\mu_I^{(1)}, \mu_I^{(2)})$	$(0.75, 0.75)$	(n_1, n_2)	$(80, 90)$

Table 7.2.2: System parameters for case 1

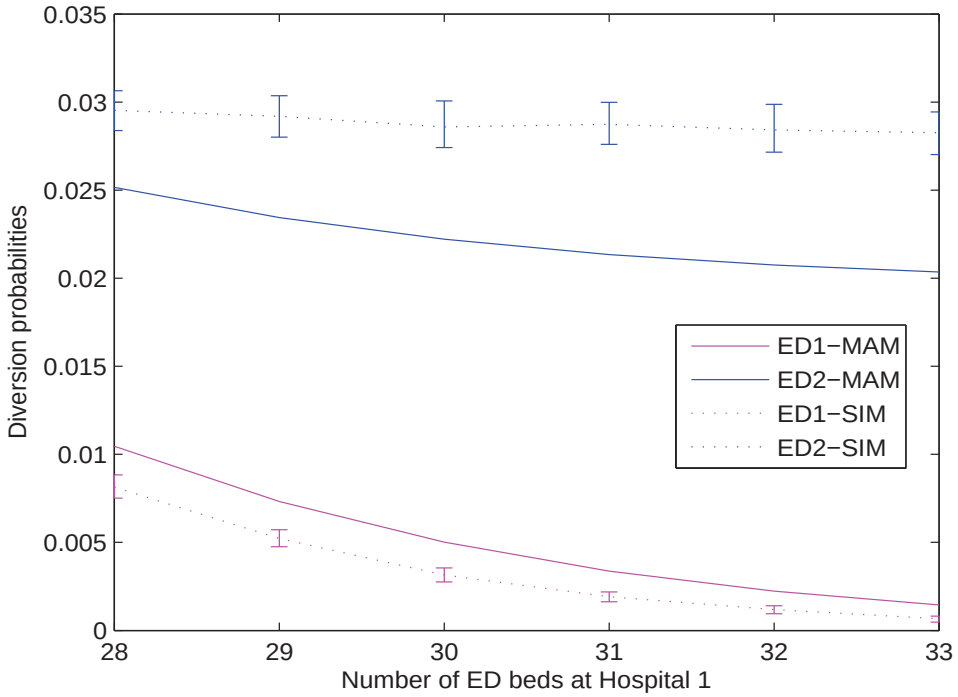

Figure 7.2.2: Case 7

from 4% to 2% if 6 ED beds are added based on the existing 28 beds.

Chapter 8

Conclusion

In this dissertation we provide a theoretical analysis of the patient flows in a regional emergency care system. We analyze the phenomenon of ED patient boarding and ambulance diversions from a network point of view by formulating a queueing network model with three components, e.g., a common EMS provider, hospital EDs and IDs. Our analysis demonstrates that the chance of an admitted patient being blocked and the probability that an ambulance is diverted from one hospital to a neighboring facility are more sensitive to the inpatient unit capacity and discharge rate than to the ED capacity and input rates. In addition, we theoretically highlight a network effect of ambulance diversions, i.e., diversion at one facility affects the patient flows and consequently the diversion probability of the neighboring facilities.

8.1 Contribution to the queueing theory

We construct a multi-server priority queue with blocking to describe the patient flow of a general hospital. Using a matrix-geometric solution, we find that the time spent by a blocked customer entering the downstream station follows phase-type distributions. Further, we show that the time spent by a low priority customer waiting until she reaches a server for the first time can be approximated by a matrix-exponential distribution.

We develop an iterative method for queueing decomposition and dimensional reduction, and develop simulation programs to test the effectiveness of our iterative approach. To date, we have not found a counter-example where the iterative algorithm fails to converge. Meanwhile solutions from the discrete-event simulation are very close to that of the iterative method.

8.2 Implication for health operations management

Our network model helps policy makers gain insight when making hospital/EMS strategic decisions on the design of an emergency care system. Given a set of operational objectives and the cost of adding an ED server/inpatient bed, our model helps answer questions like if extra ED server(s)/inpatient bed(s) is needed in a hospital, or if an inpatient bed should be moved from one hospital to another.

To strike a balance between the operating costs and the risk of ambulance unavailability due to offload delays, a proper ambulance fleet size can be determined using our model. In addition, destination and diversion policies need to be periodically reviewed as hospital size and discharge policies may change over time, and those changes affect patient flows of all hospitals in the region.

8.3 Limitations and future research

Our queueing approximation method has several limitations that stem from the approximation for the actual input process. Even the requests for ambulance are generated according to a Poisson process at each catchment area and the ambulance transit time is negligible, the actual input process of ambulance patients does not follow a Poisson process as patient arrival rates are state-dependent when ED diversions are taken into account.

In this dissertation, we also add a preemptive priority assumption for ED patients. This assumption makes our analysis more convenient. However it can be problematic if one considers practical situations where medical treatment is rarely interrupted as long as a patient starts service. It would be valuable to consider non-preemptive priority for ambulance patients. In order to do so, one has to incorporate the number of walk-in patients as system variable in the "EMS-ED" model. Since the number of walk-in patients at each hospital can be infinite, finding matrix-geometric solutions can be challenging.

One open problem is whether the iterative algorithm for queueing decomposition will always converge, and if it does not, under what circumstances will the algorithm converge? This question is left to future studies.

Appendix

Appendix A

The details of boundary matrices $A_{(k,k)}, k = 0, \ldots, c-1$ are as follows:

$$A_{(k,k)} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ n \\ \vdots \\ n+c-1 \\ n+c \end{array} \begin{pmatrix} a_0^{(k)} & b_0 & & & & & & \\ d_1 & a_1^{(k)} & b_1 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & d_n & a_n^{(k)} & b_n & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & d_{n+c-1} & a_{n+c-1}^{(k)} & b_{n+c-1} & \\ & & & & & d_{n+c} & a_{n+c}^{(k)} \end{pmatrix} \quad \text{for } k = 0, \ldots, c-1 \tag{8.1}$$

The details of boundary matrices $A_{(k,k-1)}, k = 1, \ldots, c-1$ are as follows:

$$A_{(k,k-1)} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ n \\ \vdots \\ n+c-1 \\ n+c \end{array} \begin{pmatrix} g_{0,0}^{(k)} & g_{0,1}^{(k)} & & & & & & \\ & g_{1,1}^{(k)} & g_{1,2}^{(k)} & & & & & \\ & & \ddots & \ddots & & & & \\ & & & g_{n,n}^{(k)} & g_{n,n+1}^{(k)} & & & \\ & & & & \ddots & \ddots & & \\ & & & & & g_{n+c-1,n+c-1}^{(k)} & g_{n+c-1,n+c}^{(k)} & \\ & & & & & & g_{n+c,n+c}^{(k)} \end{pmatrix} \quad \text{for } k = 1, \ldots, c-1 \tag{8.2}$$

The rate at which the number of walk-in patients increases is captured in the matrix A_0. The details of A_0 are as follows:

$$A_0 = \begin{array}{c} 0 \\ 1 \\ \vdots \\ n \\ \vdots \\ n+c \end{array} \begin{pmatrix} I_{(c+B+1)\times(c+B+1)} & & & & \\ & I_{(c+B+1)\times(c+B+1)} & & & \\ & & \ddots & & \\ & & & I_{(c+B+1)\times(c+B+1)} & \\ & & & & \ddots \\ & & & & & I_{(B+1)\times(B+1)} \end{pmatrix} \cdot \lambda_w \tag{8.3}$$

The matrix A_1 includes transitions that do not affect the number of walk-in patients in the hospital; it includes admission and departure of inpatients and ambulance patient arrivals and service completion. The details of A_1 are as follows:

$$A_1 = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ n \\ \vdots \\ n+c-1 \\ n+c \end{array} \begin{pmatrix} a_0^{(c)} & b_0 & & & & & \\ d_1 & a_1^{(c)} & b_1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & d_n & a_n^{(c)} & b_n & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & d_{n+c-1} & a_{n+c-1}^{(c)} & b_{n+c-1} \\ & & & & & d_{n+c} & a_{n+c}^{(c)} \end{pmatrix} \tag{8.4}$$

The matrix A_2 represents service completion of walk-in patients. Because those patients have lower priority than ambulance patients, they cannot start their service when all the servers are occupied by ambulance patients or boarding patients. The details of A_2 are as follows:

$$A_2 = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ n \\ \vdots \\ n+c-1 \\ n+c \end{array} \begin{pmatrix} g_{0,0}^{(c)} & g_{0,1}^{(c)} & & & & & \\ & g_{1,1}^{(c)} & g_{1,2}^{(c)} & & & & \\ & & \ddots & \ddots & & & \\ & & & g_{n,n}^{(c)} & g_{n,n+1}^{(c)} & & \\ & & & & \ddots & \ddots & \\ & & & & & g_{n+c-1,n+c-1}^{(c)} & g_{n+c-1,n+c}^{(c)} \\ & & & & & & g_{n+c,n+c}^{(c)} \end{pmatrix} \tag{8.5}$$

Let $l_i = \max(0, i-c)$, we are able to express the matrices $a_i^{(k)}$, $b_i^{(k)}$, $d_i^{(k)}$, $g_{i,i}^{(k)}$ and $g_{i,i+1}^{(k)}$.

$$a_i^{(k)} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ c \\ \vdots \\ c+B-1 \\ c+B \end{array} \begin{pmatrix} * & \lambda_a & & & & & \\ \bar{\alpha}_a\mu_a & * & \lambda_a & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & c\bar{\alpha}_a\mu_a & * & \lambda_a & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & c\bar{\alpha}_a\mu_a & * & \lambda_a \\ & & & & & c\bar{\alpha}_a\mu_a & * \end{pmatrix} \quad \text{for } i = 0, \ldots, c; \tag{8.6}$$

$$
a_i^{(k)} = \begin{array}{l} 0 \\ 1 \\ \vdots \\ c-l_i \\ \vdots \\ c+B-l_i-1 \\ c+B-l_i \end{array}
\begin{pmatrix}
* & \lambda_a & & & & & \\
\bar{\alpha}_a\mu_a & * & \lambda_a & & & & \\
& \ddots & \ddots & \ddots & & & \\
& & (c-l_i)\bar{\alpha}_a\mu_a & * & \lambda_a & & \\
& & & \ddots & \ddots & \ddots & \\
& & & & (c-l_i)\bar{\alpha}_a\mu_a & * & \lambda_a \\
& & & & & (c-l_i)\bar{\alpha}_a\mu_a & *
\end{pmatrix}
\quad \text{for } i = c+1, \ldots, c+n; \tag{8.7}
$$

where $*$ is calculated such that the rows of the matrix Q sum to zero.

$$
b_i = \begin{array}{l} 0 \\ 1 \\ \vdots \\ c \\ \vdots \\ c+B \end{array}
\begin{pmatrix}
0 & & & & & \\
\alpha_a\mu_a & 0 & & & & \\
& \ddots & \ddots & & & \\
& & c\alpha_a\mu_a & 0 & & \\
& & & \ddots & \ddots & \\
& & & & c\alpha_a\mu_a & 0
\end{pmatrix}
\quad \text{for } i = 0, \ldots, c-1; \tag{8.8}
$$

$$
b_i = \begin{array}{l} 0 \\ 1 \\ \vdots \\ c-l_i \\ \vdots \\ c+B-l_i \end{array}
\begin{pmatrix}
0 & & & & \\
\alpha_a\mu_a & & & & \\
& \ddots & & & \\
& & (c-l_i)\alpha_a\mu_a & & \\
& & & \ddots & \\
& & & & (c-l_i)\alpha_a\mu_a
\end{pmatrix}
\quad \text{for } i = c, \ldots, c+n-1; \tag{8.9}
$$

$$
d_i = i\mu_I \cdot I_{(c+B+1)\times(c+B+1)} \quad \text{for } i = 1, \ldots, c; \tag{8.10}
$$

$$
d_i = n\mu_I \cdot [I_{(c+B-l_i+1)\times(c+B-l_i+1)}\ 0] \quad \text{for } i = c, \ldots, c+n; \tag{8.11}
$$

$$
g_{i,i}^{(k)} = \begin{array}{l} 0 \\ \vdots \\ c-k \\ c-k-1 \\ \vdots \\ c \\ \vdots \\ c+B \end{array}
\begin{pmatrix}
k\bar{\alpha}_w\mu_w & & & & & & \\
& \ddots & & & & & \\
& & k\bar{\alpha}_w\mu_w & & & & \\
& & & (k-1)\bar{\alpha}_w\mu_w & & & \\
& & & & \ddots & & \\
& & & & & 0 & \\
& & & & & & \ddots \\
& & & & & & 0
\end{pmatrix}
\tag{8.12}
$$

for $i = 0, \ldots, c$;

$$g_{i,i}^{(k)} = \begin{array}{c} 0 \\ \vdots \\ c - l_i - k \\ c - l_i - k - 1 \\ \vdots \\ c - l_i \\ \vdots \\ c + B - l_i \end{array} \begin{pmatrix} k\bar{\alpha}_w\mu_w & & & & & & \\ & \ddots & & & & & \\ & & k\bar{\alpha}_w\mu_w & & & & \\ & & & (k-1)\bar{\alpha}_w\mu_w & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & \ddots \\ & & & & & & 0 \end{pmatrix} \tag{8.13}$$

for $i = c + 1, \ldots, c + n$;

$$g_{i,i+1}^{(k)} = \begin{array}{c} 0 \\ \vdots \\ c - k \\ c - k - 1 \\ \vdots \\ c \\ \vdots \\ c + B \end{array} \begin{pmatrix} k\alpha_w\mu_w & & & & & & \\ & \ddots & & & & & \\ & & k\alpha_w\mu_w & & & & \\ & & & (k-1)\alpha_w\mu_w & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & \ddots \\ & & & & & & 0 \end{pmatrix} \tag{8.14}$$

for $i = 0, \ldots, c - 1$;

$$g_{i,i+1}^{(k)} = \begin{array}{c} 0 \\ \vdots \\ c - l_i - k \\ c - l_i - k - 1 \\ \vdots \\ c - l_i \\ \vdots \\ c + B - l_i - 1 \\ c + B - l_i \end{array} \begin{pmatrix} k\alpha_w\mu_w & & & & & & \\ & \ddots & & & & & \\ & & k\alpha_w\mu_w & & & & \\ & & & (k-1)\alpha_w\mu_w & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & \ddots \\ & & & & & & 0 \\ & & & & & & 0 \end{pmatrix} \tag{8.15}$$

for $i = c, \ldots, c + n - 1$;

Appendix B

The diagonal matrices $A_{(k,k)}$ include transitions that do not affect the number of ambulance patients at ED 2; it includes ambulance patient arrivals and departures at ED 1. Denote $m_k = \max(0, k - \hat{c}_2)$. The details of matrices $A_{(k,k)}$ are given as follows:

$$A_{(k,k)} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ \hat{c}_1 - 1 \\ \hat{c}_1 \\ \vdots \\ \hat{c}_1 + N - 1 \\ \hat{c}_1 + N \end{array} \begin{pmatrix} * & \lambda_a^{(1)} & & & & & & \\ \mu_a^{(1)} & * & \lambda_a^{(1)} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & (\hat{c}_1 - 1)\mu_a^{(1)} & * & \lambda_a^{(1)} & & & \\ & & & \hat{c}_1\mu_a^{(1)} & * & 0 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \hat{c}_1\mu_a^{(1)} & * & 0 \\ & & & & & & \hat{c}_1\mu_a^{(1)} & * \end{pmatrix} \quad \text{for } k = 0, \ldots, \hat{c}_2 - 1; \tag{8.16}$$

$$A_{(k,k)} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ \hat{c}_1 - 1 \\ \hat{c}_1 \\ \vdots \\ \hat{c}_1 + N - m_k - 1 \\ \hat{c}_1 + N - m_k \end{array} \begin{pmatrix} * & \lambda_a^{(1)} + \lambda_a^{(2)} & & & & & & \\ \mu_a^{(1)} & * & \lambda_a^{(1)} + \lambda_a^{(2)} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & (\hat{c}_1 - 1)\mu_a^{(1)} & * & \lambda_a^{(1)} + \lambda_a^{(2)} & & & \\ & & & \hat{c}_1\mu_a^{(1)} & * & \lambda_a^{(1)} & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \hat{c}_1\mu_a^{(1)} & * & \lambda_a^{(1)} \\ & & & & & & \hat{c}_1\mu_a^{(1)} & * \end{pmatrix} \quad \text{for } k = \hat{c}_2, \ldots, \hat{c}_2 + N; \tag{8.17}$$

The upper diagonal matrices $A_{(k,k+1)}$ of Q_N reflect ambulance patient arrivals to ED 2. The details of matrices $A_{(k,k+1)}$ are given as follows:

$$A_{(k,k+1)} = \begin{array}{c} 0 \\ \vdots \\ \hat{c}_1 - 1 \\ \hat{c}_1 \\ \vdots \\ \hat{c}_1 + N - 1 \\ \hat{c}_1 + N \end{array} \begin{pmatrix} \lambda_a^{(2)} & & & & & & \\ & \ddots & & & & & \\ & & \lambda_a^{(2)} & & & & \\ & & & \lambda_a^{(1)} + \lambda_a^{(2)} & & & \\ & & & & \ddots & & \\ & & & & & \lambda_a^{(1)} + \lambda_a^{(2)} & \\ & & & & & & 0 \end{pmatrix} \quad \text{for } k = 0, \ldots, \hat{c}_2 - 1; \tag{8.18}$$

$$A_{(k,k+1)} = \begin{matrix} 0 \\ 1 \\ \vdots \\ \hat{c}_1 - 1 \\ \hat{c}_1 \\ \vdots \\ \hat{c}_1 + N - m_k - 1 \\ \hat{c}_1 + N - m_k \end{matrix} \begin{pmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & \lambda_a^{(2)} & & & \\ & & & & \ddots & & \\ & & & & & \lambda_a^{(2)} & \\ & & & & & & 0 \end{pmatrix} \quad \text{for } k = \hat{c}_2, \ldots, \hat{c}_2 + N - 1; \tag{8.19}$$

The lower diagonal matrices $A_{(k,k-1)}$ of Q_N describe service completions of ED 2.

$$A_{(k,k-1)} = k\mu_a^2 \cdot I_{(\hat{c}_1+N+1)\times(\hat{c}_1+N+1)} \quad \text{for } k = 1, \ldots, \hat{c}_2; \tag{8.20}$$

$$A_{(k,k-1)} = \hat{c}_2\mu_a^{(2)} \cdot [I_{(\hat{c}_1+N-m_k+1)\times(\hat{c}_1+N-m_k+1)}\ 0] \quad \text{for } k = \hat{c}_2 + 1, \ldots, \hat{c}_2 + N; \tag{8.21}$$

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{\hat{c}_2+N})$ the stationary distribution of Q_N. Since the CTMC is irreducible, $\boldsymbol{\pi}$ exists and is the unique non-negative solution for the linear system:

$$\boldsymbol{\pi} Q_N = 0; \qquad \boldsymbol{\pi}\boldsymbol{e} = 1, \tag{8.22}$$

where $\boldsymbol{e}$ is a column vector of ones. Since the infinitesimal generator Q_N has a block tri-diagonal structure, a matrix-geometric solution can be obtained. First, for the levels $\hat{c}_2 + N$ and $\hat{c}_2 + N - 1$, we obtain

$$\boldsymbol{\pi}_{\hat{c}_2+N} = \boldsymbol{\pi}_{\hat{c}_2+N-1} R_{\hat{c}_2+N} \tag{8.23}$$

where

$$R_{\hat{c}_2+N} = -A_{\hat{c}_2+N-1,\hat{c}_2+N}(A_{\hat{c}_2+N,\hat{c}_2+N})^{-1}. \tag{8.24}$$

Level $\hat{c}_2 + N - 1$ to level 1 can be solved recursively using

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i-1} R_i, \tag{8.25}$$

where

$$R_i = -A_{i-1,i}(A_{i,i} + R_{i+1}A_{i+1,i})^{-1}. \tag{8.26}$$

In order to solve $\boldsymbol{\pi}$, we need to specify the boundary constraints for $\boldsymbol{\pi}_0$. The boundary balance equations and the normalization condition lead to the following linear system for $\boldsymbol{\pi}_0$:

$$\begin{aligned} &\boldsymbol{\pi}_0(A_{(0,0)} + R_1 A_{(1,0)}) = 0; \\ &\boldsymbol{\pi}_0[\boldsymbol{e} + R_1\boldsymbol{e} + R_1R_2\boldsymbol{e} + \cdots + R_1 \cdots R_{\hat{c}_2+N}\boldsymbol{e}] = 1. \end{aligned} \tag{8.27}$$

Appendix C

We present the steady-state distribution for the $M/G/n/n+c$ queue based on diffusion approximation.

Let μ and C_S denote the mean and CV (coefficient of variation) of the service time. Define

$$a_k = \begin{cases} \lambda + k\mu C_S^2, & k = 1, \ldots, n-1, \\ \lambda + n\mu C_S^2, & k = n, \ldots, n+c-1; \end{cases} \tag{8.28}$$

$$a_k^* = \lambda + k\mu, \qquad k = 1, \ldots, n; \tag{8.29}$$

$$\alpha_k = \frac{a_k^*}{a_k} = \frac{\lambda + k\mu}{\lambda + k\mu C_S^2}, \qquad k = 1, \ldots, n; \tag{8.30}$$

and

$$\xi_k = \frac{1}{a_k} \prod_{j=1}^{k} \left(\frac{a_j^*}{a_{j-1}^*} \frac{n\rho}{j} \right)^{\alpha_j}, \qquad k = 1, \ldots, n. \tag{8.31}$$

Then we obtain the approximate distribution for the $M/G/n/n+c$ queue as

$$p_k = \begin{cases} p_0 \xi_k, & k = 1, \ldots, n-1, \\ p_0 \xi_n \hat{\rho}^{k-n}, & k = n, \ldots, n+c-1, \\ \frac{1}{\rho} \left\{ \rho - 1 + p_0 \sum_{j=0}^{n-1} (1 - \frac{j}{n}) \xi_j \right\}, & k = n+c, \end{cases} \tag{8.32}$$

where $\hat{\rho} = \rho^{\alpha_n}$ and the empty probability p_0 is given by

$$p_0 = \begin{cases} \frac{1}{\sum_{k=0}^{n-1} (\rho + 1 - k/n) \xi_k + ((1-\hat{\rho}^c)/(1-\hat{\rho})) \rho \xi_n}, & \rho \neq 1, \\ \frac{1}{\sum_{k=0}^{n-1} (2 - k/n) \xi_k + c\xi_n}, & \rho = 1. \end{cases} \tag{8.33}$$

Bibliography

Allon, G., Deo, S., Lin, W., 2013. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. Operations Research 61 (3), 544–562.

Almehdawe, E., Jewkes, B., He, Q.-M., 2013. A markovian queueing model for ambulance offload delays. European Journal of Operational Research 226 (3), 602–614.

Artalejo, J. R., Gómez-Corral, A., 2008. Retrial queueing systems: a computational approach. Vol. 318. Springer Berlin.

Au-Yeung, S., Harrison, P., Knottenbelt, W., 2006. A queueing network model of patient flow in an accident and emergency. 20th Annual European and Simulation Modelling Conference, 60C–67.

Breuer, L., Baum, D., 2005. An introduction to queueing theory and matrix-analytic methods. Springer.

Broyles, J. R., Cochran, J. K., 2011. A queuing-base statistical approximation of hospital emergency department boarding. Proceedings of the 41st International Conference on Computers & Industrial Engineering, 122–127.

Budge, S., Ingolfsson, A., Zerom, D., 2010. Empirical analysis of ambulance travel times: the case of calgary emergency medical services. Management Science 56 (4), 716–723.

Burt, C. W., McCaig, L. F., Valverde, R. H., 2006. Analysis of ambulance transports and diversions among us emergency departments. Annals of emergency medicine 47 (4), 317–326.

Chan, C. W., Farias, V. F., Escobar, G., 2013. The impact of delays on service times in the intensive care unit, preprint.

Chen, H., Yao, D. D., 2001. Fundamentals of queueing networks: Performance, asymptotics, and optimization. Vol. 46. Springer.

Deo, S., Gurvich, I., 2011. Centralized vs. decentralized ambulance diversion: A network perspective. Management Science 57 (7), 1300–1319.

Derlet, R. W., Richards, J. R., Kravitz, R. L., 2001. Frequent overcrowding in us emergency departments. Academic Emergency Medicine 8 (2), 151–155.

Eckstein, M., Chan, L., 2004. The effect of emergency department crowding on paramedic ambulance availability. Annals of emergency medicine 43 (1), 100–105.

Forster, A. J., Stiell, I., Wells, G., Lee, A. J., Van Walraven, C., 2003. The effect of hospital occupancy on emergency department length of stay and patient disposition. Academic Emergency Medicine 10 (2), 127–133.

Green, L. V., 2008. Using operations research to reduce delays for healthcare. Tutorials in Operations Research.

Green, L. V., Nguyen, V., 2001. Strategies for cutting hospital beds: the impact on patient service. Health Services Research 36 (2), 421.

Green, L. V., Soares, J., Giglio, J. F., Green, R. A., 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine 13 (1), 61–68.

Hagtvedt, R., Ferguson, M., Griffin, P., Jones, G. T., Keskinocak, P., 2009. Cooperative strategies to reduce ambulance diversion. Proceedings of the 2009 Winter Simulation Conference, 1861–1874.

Hall, R. W., 2006. Patient flow: reducing delay in healthcare delivery. Vol. 91. Springer.

He, Q., 2013. Fundamentals of Matrix-Analytic Methods. Springer.

Kimura, T., 2003. A consistent diffusion approximation for finite-capacity multiserver queues. Mathematical and computer modelling 38 (11), 1313–1324.

Koizumi, N., Kuno, E., Smith, T. E., 2005. Modeling patient flows using a queuing network with blocking. Health Care Management Science 8 (1), 49–60.

Kolb, E. M. W., Lee, T., Peck, J., 2007. Effect of coupling between emergency department and inpatient unit on the overcrowding in emergency department. In: Simulation Conference, 2007 Winter. IEEE, pp. 1586–1593.

Kolker, A., 2008. Process modeling of emergency department patient flow: Effect of patient length of stay on ed diversion. Journal of Medical Systems 32 (5), 389–401.

Lane, D. C., Monefeldt, C., Rosenhead, J., 2000. Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. Journal of the Operational Research Society, 518–531.

Latouche, G., Ramaswami, V., 1999. Introduction to matrix analytic methods in stochastic modeling. Vol. 5. Siam.

Liu, Y., Whitt, W., 2011. A network of time-varying many-server fluid queues with customer abandonment. Operations research 59 (4), 835–846.

Marshall, A. H., McClean, S. I., 2004. Using coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. Health Care Management Science 7 (4), 285–289.

McCarter, J., December 2010. Office of the general auditor of ontario, 2010 annual report. http://www.auditor.on.ca/en/reports_2012_en.htm.

McCarter, J., December 2012. Office of the general auditor of ontario, 2012 annual report. http://www.auditor.on.ca/en/reports_2010_en.htm.

McManus, M. L., Long, M. C., Cooper, A., Litvak, E., 2004. Queuing theory accurately models the need for critical care resources. Anesthesiology 100 (5), 1271–1276.

Nafarrate, A. R., Fowler, J. W., Wu, T., 2010. Bi-criteria analysis of ambulance diversion policies. In: Simulation Conference (WSC), Proceedings of the 2010 Winter. IEEE, pp. 2315–2326.

Neuts, M. F., 1981. Matrix-geometric solutions in stochastic models: an algorithmic approach. Courier Dover Publications.

Neuts, M. F., 1989. Structured stochastic matrices of M/G/1 type and their applications. Vol. 701. Marcel Dekker New York.

Ost, A., 2001. Performance of communication systems: a model-based approach with matrix-geometric methods. Springer.

Ozawa, T., 2006. Sojourn time distributions in the queue defined by a general qbd process. Queueing Systems 53 (4), 203–211.

Powell, E. S., Khare, R. K., Venkatesh, A. K., Van Roo, B. D., Adams, J. G., Reinhardt, G., 2012. The relationship between inpatient discharge timing and emergency department boarding. The Journal of emergency medicine 42 (2), 186–196.

Proudlove, N., Gordon, K., Boaden, R., 2003. Can good bed management solve the overcrowding in accident and emergency departments? Emergency Medicine Journal 20 (2), 149–155.

Ramirez, A., Fowler, J. W., Wu, T., 2009. Analysis of ambulance diversion policies for a large-size hospital. In: Simulation Conference (WSC), Proceedings of the 2009 Winter. IEEE, pp. 1875–1886.

Ramirez-Nafarrate, A., Fowler, J. W., Wu, T., 2011. Design of centralized ambulance diversion policies using simulation-optimization. In: Proceedings of the Winter Simulation Conference. Winter Simulation Conference, pp. 1251–1262.

Richardson, D. B., 2002. The access-block effect: relationship between delay to reaching an inpatient bed and inpatient length of stay. Medical Journal of Australia 177 (9), 492–495.

Schull, M. J., Morrison, L. J., Vermeulen, M., Redelmeier, D. A., 2003. Emergency department gridlock and out-of-hospital delays for cardiac patients. Academic emergency medicine 10 (7), 709–716.

Shi, P., Chou, M., Dai, J., Ding, D., Sim, J., 2012. Hospital inpatient operations: mathematical models and managerial insights, preprint.

Sun, B. C., Mohanty, S. A., Weiss, R., Tadeo, R., Hasbrouck, M., Koenig, W., Meyer, C., Asch, S., 2006. Effects of hospital closures and hospital characteristics on emergency department ambulance diversion, los angeles county, 1998 to 2004. Annals of emergency medicine 47 (4), 309–316.

Tian, N., Zhang, Z. G., 2006. Vacation queueing models: theory and applications. Vol. 93. Springer.

Whitt, W., 1984. Heavy traffic approximations for service systems with blocking. AT&T Bell Laboratories Technical Journal 63 (5), 689–708.

Whitt, W., 2004. A diffusion approximation for the $g/gi/n/m$ queue. Operations Research 52 (6), 922–941.

Yankovic, N., Green, L. V., 2011. Identifying good nursing levels: A queuing approach. Operations research 59 (4), 942–955.

Zhu, B., Armony, M., Chan, C. W., 2013. Critical care in hospitals: When to introduce a step down unit?, to appear.