

# Centralized vs. Decentralized Ambulance Diversion: A Network Perspective

Sarang Deo\* Itai Gurvich†

## Abstract

In recent years, growth in the demand for emergency medical services along with decline in the number of hospitals with emergency departments (EDs) has led to overcrowding. In periods of overcrowding, an ED can request the Emergency Medical Services (EMS) agency to divert incoming ambulances to neighboring hospitals, a phenomenon known as “ambulance diversion”. The EMS agency will accept this request provided that at least one of the neighboring EDs is not on diversion. From an operations perspective, properly executed ambulance diversion should result in resource pooling and reduce the overcrowding and delays in a network of EDs. Recent evidence indicates, however, that this potential benefit is not always realized. In this paper, we provide one potential explanation for this discrepancy and suggest potential remedies. Using a queueing game between two EDs that aim to minimize their own waiting time, we find that decentralized decisions regarding diversion explain the lack of pooling benefits. Specifically, we find the existence of a *defensive equilibrium*, wherein each ED does not accept diverted ambulances from the other ED. This defensiveness results in a *de-pooling* of the network and, in turn, in delays that are significantly higher than when a social planner coordinates diversion. The social optimum is, itself, difficult to characterize analytically and has limited practical appeal as it depends on problem parameters such as arrival rates and length of stay. Instead, we identify an alternative solution that is more amenable to implementation and can be used by the EMS agencies to coordinate diversion decisions even without the exact knowledge of these parameters. We show that this solution is approximately optimal for the social planner’s problem. Moreover, it is Pareto improving over the defensive equilibrium whereas the social optimum, in general, might not be.

*Keywords:* emergency department, ambulance diversion, game theory, queueing networks.

## 1 Introduction

In many regions of the US, emergency departments (EDs) can declare “diversion” status (a request to divert incoming ambulances), if they are overcrowded. The local emergency medical system (EMS) agency may respond by re-routing ambulances to other neighboring EDs if they are available to accept additional patients. From an operations perspective, if properly executed, ambulance diversion can help to balance capacity and demand in a network by re-routing ambulances away from overcrowded EDs to less crowded ones. In other words, it has the potential of achieving reduced waiting times through increased resource pooling. Yet, anecdotal [Chidester et al. 2009] and preliminary empirical evidence [Mihal and Moilanen 2005, Kowalczyk 2009] suggests that ambulance diversion has no beneficial impact on waiting times at EDs. This lack of benefit together with adverse effects such as delays in patient transport [Schull et al.

---

\*Kellogg School of Management (s-deo@kellogg.northwestern.edu)

†Kellogg School of Management i-gurvich@kellogg.northwestern.edu)

2003] and poorer health outcomes for patients [Schull et al. 2004] has motivated regulators in many places to place restrictions on ambulance diversion. A recent notable example is the complete ban on ambulance diversion in Massachusetts [Kowalczyk 2009].

One potential explanation for the discrepancy between operations management theory predicting reduction in waiting times and the empirical evidence indicating no such effect is the decentralized decision making by EDs. The purported operational benefits of ambulance diversion presuppose centralized coordination that can match excess capacity and excess demand. In practice, however, diversion decisions are often made by ED administrators (e.g. nurse or physician in-charge) with the objective of mitigating overcrowding at their own location, while keeping the number of diverted patients at reasonable levels.

The ED administrator assesses crowding in terms of the total number of patients in the ED or the number patients who are awaiting admission to the inpatient department (boarding patients) or some aggregation of the two and requests diversion status when this crowding measure exceeds a predetermined threshold. The EMS agency accepts the diversion request if neighboring EDs are also not on diversion. Otherwise, diversion requests are ignored and each ED is forced to accept its own ambulance arrivals. This coordination guideline, frequently referred to as “All on Diversion, Nobody on Diversion” (hereafter “ADND”), is commonly employed by EMS agencies [Fatovich et al. 2005, Chidester et al. 2009, Mihal and Moilanen 2005, MARCER 2007].

To formally investigate the impact of decentralized decision making, we develop a stylized queueing-network model comprising two EDs each receiving two arrival streams – walk-ins and ambulances – and we embed this queueing network within a static non-cooperative game. Here, each ED administrator chooses a diversion threshold based on the number of patients in the ED with the objective of minimizing the average waiting time at her location. The threshold diversion policy and the ADND guideline create a feedback structure which renders the game intractable for detailed analysis. Nevertheless, we find the existence of a *defensive equilibrium* in which both EDs choose a threshold of zero, ambulances are not diverted at all and consequently, the potential benefits of resource pooling are not realized. However, contrary to a priori intuition, we find that being defensive is not a dominant strategy due to the aforementioned feedback structure. We find that the defensive equilibrium described above persists as a Nash equilibrium in a model variant, in which we allow each ED to impose an upper bound on the number of ambulances “lost” due to diversion. This accounts for constraints that ED administrators might impose to prevent excessive lost revenues from diversion or to maintain a certain level of utilization.

To benchmark the equilibrium outcomes and suggest potential policy interventions, we consider a centralized system. Here, a benevolent social planner coordinates the threshold choices of the ED administrators with the objective of minimizing the average waiting time across the entire network. To make a fair comparison with the decentralized setting, we do not include other potential costs borne by society such as increased travel times, patients prevented from seeing their own physician in their local ED, reduced ambulance availability etc. Optimal solutions to the social planner’s problem are difficult to characterize analytically and depend on all problem parameters: arrival rates, service rates and ED capacities.

We propose a simpler solution, where each ED uses its capacity (the number of beds) as its diversion threshold. This solution has several benefits. First, we prove that these capacity-based thresholds perform

almost as well as the social optimum. Second, we prove that the expected waiting time in each ED under this approximate optimal solution is at most marginally higher compared to that under the *defensive equilibrium*. In fact, in all our numerical experiments, we find that the approximate optimal solution is (strictly) Pareto improving. Finally, our numerical results suggest that this approximate solution can be induced as an equilibrium using a relatively transparent regulation: EDs with available beds cannot declare diversion.

Comparison of the decentralized equilibrium with the (centralized) socially optimal solution allows us to synthesize two disparate observations from practice. First, anecdotal and empirical evidence [Sun et al. 2006, Allon et al. 2009] suggests that EDs overuse diversion, a practice referred to as *defensive diversion* or the *network effect of diversion*. Our result that decentralized decision making leads to a *defensive equilibrium* provides a theoretical support for this observation. Second, preliminary observations [Kowalczyk 2009] indicate that the waiting times in Massachusetts did not increase following the recent ban on ambulance diversion. This has been attributed to internal process improvements at EDs. Our results provide one alternative explanation in suggesting that one should not expect an increase in waiting time if the EDs were in a *defensive equilibrium* before the ban.

The remainder of the paper is organized as follows: §2 reviews the relevant literature. We describe the ED network model and provide initial characterization of the underlying queueing dynamics in §3. We analyze the decentralized setting in §4 and the centralized setting in §5. Numerical results are presented in §6 followed by concluding remarks in §7. All the proofs appear in the appendix.

## 2 Literature Review

Our work is related to two distinct streams of literature: the emergency medicine (EM) literature and the operations management (OM) literature. We contribute to the former by providing one explanation for the phenomenon of *defensive diversion* and proposing possible remedies. We contribute to the latter by formulating a theoretical model of ambulance diversion that incorporates both the nature of decision making (decentralized vs. centralized) and the queueing-network dynamics. Our model of decentralized admission control contributes to recent work on service competition, which has focused mostly on price and capacity.

Although there is extensive work in the EM literature on ambulance diversion [Hoot and Aronsky 2008, Pham et al. 2008, Moskop et al. 2009], only few papers provide evidence for the *network effect* of ambulance diversion or *defensive diversion*. Sun et al. [2006] find that diversion hours at neighboring EDs in LA county are correlated. Similarly, Allon et al. [2009] find that EDs that have more EDs in their proximity spend more hours on diversion, after controlling for their own congestion level. Vilkes et al. [2004] and Mihal and Moilanen [2005] provide evidence from San Diego and Los Angeles, respectively, that EDs use *defensive diversion* to protect themselves from getting ambulances from their neighboring EDs. We add to this stream by presenting the first theoretical model of queueing-network aspects of ambulance diversion.

There is extensive OM literature on ED operations that uses queueing theory to optimize patient flow within hospitals [Cochran and Roche 2009, Vassilicopoulos 1985, Bagust et al. 1999]. However, only few papers in this literature explicitly model the admission control decision associated with ambulance diversion [Allon et al. 2009, Ramirez et al. 2009, Enders et al. 2010] with only the latter modeling interaction between neighboring EDs. Enders et al. [2010] develop a simulation model incorporating multiple details

of the patient-flow dynamics. They conduct extensive numerical analysis to study various mechanisms to signal delay between EDs and the EMS agencies. In contrast, we abstract away from these details, explicitly model the game between EDs and derive theoretical results pertaining to equilibrium and social optimum. Hagtvedt et al. [2009] also consider a non-cooperative game of ambulance diversion but use reduced form specifications for the payoff function of the EDs. The payoffs in our model, in contrast, are direct consequences of the underlying queueing-network dynamics.

Game-theoretic queueing models have been studied extensively in the OM literature. Most of these models focus on settings in which the firms' decision is either price and/or capacity [Levhari and Luski 1978, Cachon and Harker 2002, Kalai et al. 1992, Cachon and Zhang 2007, Allon and Federgruen 2007]. In these models, the choice of these two variables determines the (state independent) arrival rate for each firm. In our model, in contrast, the effective arrival rate to each ED is state dependent due to diversion decisions, which correspond to admission control<sup>1</sup>. This admission control results in multi-dimensional queueing dynamics that are difficult to analyze. Chen et al. [2008] consider dynamic routing of customers to one of many servers that compete on the basis of price and real-time delay. They use heavy-traffic analysis to replace the original game with an approximate one that reduces the dimensionality of the queueing dynamics and simplifies the analysis. Such heavy-traffic approximations do not provide reduction in dimensionality in our model and hence we do not pursue this direction.

The literature on the optimal (centralized) routing in queueing networks is also quite extensive. Some relevant examples in the context of parallel server networks with multiple servers per server group (as the one we study here) Atar [2005], Gurvich and Whitt [2009], Tezcan [2008], Stolyar [2005], Adan et al. [1994]. The latter three are the closest to our ambulance diversion model in that they study settings in which routing decisions are made upon customer/patient arrival and once the customers are assigned to a queue they cannot be re-routed.

Finally, our work is related also to the vast literature on resource pooling, which has primarily focused on the operational benefits of pooling under centralized decision making (Akşin et al. [2007] contains a rich reference list on this literature, albeit focused on call centers). In contrast, we show that all the benefits of pooling are lost due to the divergence between the incentives of the decision makers.

### 3 The network model

In this section, we define the model primitives, provide a formal description of the ambulance diversion policies and characterize the underlying Markov chain for a stylized network comprising two EDs.

#### 3.1 The primitives: Patient arrivals, length of stay and ED capacity

We consider a network consisting of two EDs ( $i = 1, 2$ ) as depicted in Figure 1. With **no** diversion, ED  $i$  faces two independent stationary arrival streams: ambulances arrive according to a Poisson process with rate  $\lambda_i^a > 0$  and walk-ins according to a Poisson process with rate  $\lambda_i^w > 0$  resulting in total arrival rate  $\lambda_i := \lambda_i^w + \lambda_i^a$ . These “potential” arrivals correspond to patients for whom ED  $i$  is the preferred ED due

---

<sup>1</sup>[Lin 2003] also analyzes a model of decentralized routing, significantly different from ours, that generalizes the model of [Naor 1969] by considering multiple points of entry (or gatekeepers) to a single queue.

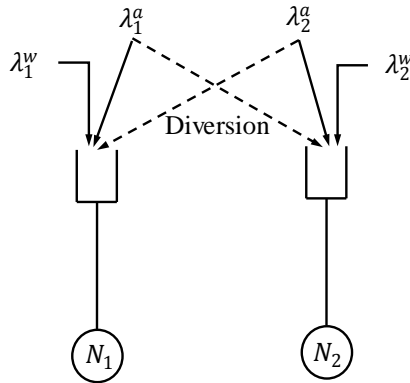


Figure 1: A network with two EDs

to insurance coverage, preferred physician, physical proximity from their residence or any combination of these factors. We refer to these as the arrivals from *catchment area*  $i$ . While EDs have to accept all walk-in arrivals according to the US federal law [General Accounting Office 2003], ambulances may be diverted. Thus, the “effective” ambulance arrivals can differ from the “potential” arrivals; see §3.2 below.

We assume that the length of stay of a patient is exponentially distributed with mean  $1/\mu$  in both EDs. The length of stay is independent across patients. ED  $i$  has  $N_i$  beds, hereafter referred to as the capacity of ED  $i$ . We assume that each ED has sufficient capacity to serve its potential arrivals if required to do so, i.e.,

$$N_i > \frac{\lambda_i}{\mu}, \text{ for } i = 1, 2. \quad (1)$$

### 3.2 Diversion policy and routing guideline

We now describe the mechanism through which EDs declare diversion status and the local EMS agency (partially) coordinates diversions in real time. ED administrators request to be “on diversion” or declare “diversion status” if they assess that crowding at their location can compromise patient care. We assume that ED administrators use the total number of patients in the ED (waiting or being attended to) as a primary measure of crowding and go “on diversion” if this number exceeds a pre-specified threshold. In accordance with the US federal law and the diversion guidelines circulated by several EMS agencies in the US [MARCER 2007, NYBEMS 2006, PEHSC 2004], we assume that the ED administrators do not account for overcrowding at neighboring EDs.

More formally, the threshold policy is as follows: The administrator at ED  $i$  declares “diversion status” at time  $t$  if the number of patients in the ED,  $X_i(t)$ , is greater than or equal to a threshold  $K_i$ , hereafter referred to as *the diversion threshold* for ED  $i$ . Thus, a diversion threshold of 0 is equivalent to the ED always being on diversion whereas a diversion threshold of  $\infty$  implies that the ED is never on diversion.

The local EMS agency, in its turn, has access to the real-time diversion status of each ED in the region and uses this information to partially coordinate the transport of ambulances: If none of the EDs is on diversion, the “effective” instantaneous ambulance arrival rate to ED  $i$  is equal to its “potential” ambulance arrival rate, i.e., to  $\lambda_i^a$ . If only ED  $i$  is on diversion (and ED  $j \neq i$  is not), then ambulances from catchment

area  $i$  are routed to ED  $j$  so that the “effective” instantaneous ambulance arrival rate to ED  $j$  becomes  $\hat{\lambda}_j^a = \lambda_i^a + \lambda_j^a$  while that of ED  $i$  is 0 (walk-ins still continue to arrive). Finally, if both EDs declare a diversion status, the “All on Diversion, Nobody on Diversion” (ADND) guideline [Asamoah et al. 2008, Mihal and Moilanen 2005, MARCER 2007] is imposed by the EMS agency and ambulance arrivals are routed to their designated EDs so that  $\hat{\lambda}_i^a = \lambda_i^a$ ,  $i = 1, 2$ . All possible outcomes are shown in Table 1.

Table 1: Effective arrival rates depending on diversion statuses

		ED 1	
		Diversion ( $X_1(t) \geq K_1$ )	No Diversion ( $X_1(t) < K_1$ )
ED 2	Diversion ( $X_2(t) \geq K_2$ )	$(\hat{\lambda}_1^a = \lambda_1^a, \hat{\lambda}_2^a = \lambda_2^a)$	$(\hat{\lambda}_1^a = \lambda_1^a + \lambda_2^a, \hat{\lambda}_2^a = 0)$
	No Diversion ( $X_2(t) < K_2$ )	$(\hat{\lambda}_1^a = 0, \hat{\lambda}_2^a = \lambda_1^a + \lambda_2^a)$	$(\hat{\lambda}_1^a = \lambda_1^a, \hat{\lambda}_2^a = \lambda_2^a)$

Thus, given a threshold pair  $K = (K_1, K_2)$ , the instantaneous arrival rate to ED  $i$  at time  $t$  is:

$$\hat{\lambda}_i(X_1(t), X_2(t)) := \lambda_i^w + \lambda_i^a(1 - \mathbb{1}\{X_i(t) \geq K_i, X_j(t) < K_j\}) + \lambda_j^a \mathbb{1}\{X_i(t) < K_i, X_j(t) \geq K_j\}.$$

We end this section with a discussion of key the modeling choices and assumptions we have made thus far.

1. *Exponential service and inter-arrival times:* In our model we assume that the arrivals follow a Poisson process and service times are exponentially distributed. While Poisson arrivals (possibly doubly stochastic) is a widely acceptable assumption for service systems (see e.g. Maman [2009] and the references therein), there is recent evidence that service times are not exponentially distributed (see e.g. Gans et al. [2010], Yom-Tov [2009] and references therein). Our proofs do rely on the Markovian assumptions but we conjecture that the essence of our results extends also to non-Markovian settings.
2. *Stationarity:* We assume that both arrival streams are stationary. While we believe that the stationary setting captures the key features of the network effect of ambulance diversion, it may be of interest to explore a more realistic model with non-stationarity.
3. *Prioritization of patients within each ED:* Since our objective function for each ED corresponds to the waiting time averaged over all patients served in this ED (see §4), it is immaterial how service is prioritized amongst patients. Our only restriction on the assignment of patients to beds is that it is done in a work-conserving manner.
4. *All ambulances are divertible:* For simplicity, our definition of the effective arrival rates assumes that diversion has an all-or-nothing effect on ambulances. When ED  $i$  is on diversion but ED  $j$  is not, ED  $i$  will not receive any ambulances. However, in reality, ED  $i$  may continue to receive some fraction of the pre-diversion ambulance arrivals owing to patient preferences, legal requirements or medical

reasons such as availability of a specific specialist or a specific medical equipment. These can be modeled by appropriately increasing  $\lambda_1^w$  and decreasing  $\lambda_1^a$ .

5. *Threshold policies:* Threshold policies are abstractions from more complex workload measures of ED crowding proposed in the emergency medicine literature, which account for additional factors such as staffing and patient severity [Reeder and Garrison 2001, Weiss et al. 2004, Bernstein et al. 2003, Epstein and Tian 2006]. Recent research shows that ED occupancy predicts overcrowding as well as these workload measures [Hoot et al. 2007] and correlates well with the staff’s perception of overcrowding [McCarthy et al. 2008].
6. *Patient boarding:* There is evidence that some hospitals base their diversion thresholds on the number of boarding patients, i.e., patients who are awaiting transfer to the inpatient department [Allon et al. 2009, Ramirez et al. 2009]. To maintain analytical tractability, we do not consider these policies here.

### 3.3 The underlying Markov chain

The variable  $X(t) = (X_1(t), X_2(t))$  captures the number of patients in each ED at time  $t$ . Under our assumptions on the primitives, the process  $X(\cdot) = (X(t), t \geq 0)$  is a Continuous Time Markov Chain (CTMC). We write  $X^K(\cdot)$  instead of  $X(\cdot)$  to make the dependence of the dynamics on the threshold pair  $K$  explicit. With the exception of  $K = (0, 0)$ , the steady-state distribution of  $X^K(\cdot)$  is intractable for closed-form characterization as shown in the following theorem.

**Theorem 1** *Suppose that (1) holds. Then, for any finite threshold pair  $K = (K_1, K_2)$ , the CTMC  $X^K(\cdot)$  has a unique steady-state distribution. Also, for all  $K \neq (0, 0)$ , the CTMC is not reversible.*

Using Theorem 1, whenever  $K$  is finite, we use  $X^K = (X_1^K, X_2^K)$  to denote a random variable that is distributed according to the unique steady-state distribution of  $X^K(\cdot)$ .

In the absence of diversion, stability is trivially guaranteed by condition (1). In the presence of diversion, the “effective” instantaneous arrival rate at ED  $i$ ,  $\hat{\lambda}_i(X_1^K(t), X_2^K(t))$  may be greater than  $\mu N_i$  in some parts of the state space so that stability is not automatically guaranteed. With finite thresholds, however, when the queue at ED  $i$  is long (so that the total number of patients is greater than the threshold  $K_i$ ), it faces at most an arrival rate  $\lambda_i$  which is, by assumption (1), less than its capacity  $\mu N_i$ .

Unfortunately, since  $X^K(\cdot)$  is not a reversible CTMC, it is not clear if there exist non-trivial parameters such that (with  $K \neq (0, 0)$ ) the steady-state distribution is of product form. Yet, we are able to analytically establish key properties of the equilibrium and the social optimum without explicitly characterizing the steady-state distribution. We compute this distribution for our numerical experiments; see §6.

Each ED uses a work conserving policy in serving its arriving patients, i.e, there can be no available beds in ED  $i$  while there are patients waiting in its queue. Hence, the steady-state queue length at ED  $i$ ,  $Q_i^K$ , satisfies  $Q_i^K = \max\{0, X_i^K - N_i\}$ . Using Little’s Law, the expected steady-state waiting time in ED  $i$  is then defined by  $\mathbb{E}[W_i(K_1, K_2)] = \mathbb{E}[Q_i^K] / \bar{\lambda}_i(K)$ <sup>2</sup>, where  $\bar{\lambda}_i(K)$  is the steady-state effective arrival rate to

<sup>2</sup>When steady-state does not exist for  $X^K(\cdot)$ , we replace the right hand side with the expected long run average  $\mathbb{E}\left[\limsup_{t \rightarrow \infty} \frac{1}{A_i(t)} \int_0^t (X_i^K(s) - N_i)^+ ds\right]$ , where  $A_i(t)$  is the number of arrivals to ED  $i$  by time  $t$ . When steady-state does exist Little’s law applies and the two definitions coincide; see e.g, Theorem 2.1 of Whitt [1991].

ED  $i$  when the threshold pair is  $K = (K_1, K_2)$  and is given by:

$$\bar{\lambda}_i(K) := \mathbb{E}[\hat{\lambda}_i(X^K)] = \lambda_i^w + \lambda_i^a(1 - \mathbb{P}\{X_i^K \geq K_i, X_j^K < K_j\}) + \lambda_j^a \mathbb{P}\{X_i^K < K_i, X_j^K \geq K_j\} \quad (2)$$

Next, we consider two systems that differ in how the threshold pair  $K$  is determined. In §4, ED administrators choose the thresholds to minimize expected waiting time at their own location. In §5 a central planner chooses the thresholds so as to minimize the average patient wait across the network.

## 4 Decentralized setting

In the decentralized setting, the administrator of ED  $i$  chooses a threshold  $K_i$  with the objective of minimizing expected waiting time at her location. Waiting time is related to other measures of delay previously used in the OM literature [Allon et al. 2009, Green 2002] and in the EM literature [Weiss et al. 2004, Bernstein et al. 2003, Epstein and Tian 2006]. It has also been described as reflecting the EDs' mission [Schull et al. 2002, Franaszek et al. 2002]. We do not include the cost and time of ambulance travel in objective function since EDs do not have direct control over and do not internalize these factors.

The waiting time at ED  $i$  depends not only on the threshold choice of ED  $i$  but also on the threshold chosen at ED  $j \neq i$  through ambulance diversion. Hence, we model the choice of thresholds in the decentralized setting as a static, non-cooperative game between the EDs, formally defined below.

**Definition 1** *The Diversion Game is the two-player game with payoffs  $\Pi_i(K_1, K_2) := \mathbb{E}[W_i(K_1, K_2)]$   $i = 1, 2$  and strategy space  $\mathbb{Z}_+ \times \mathbb{Z}_+$ .*

Figure 2 depicts the two-dimensional payoff function for one set of parameters. Evidently, the game's

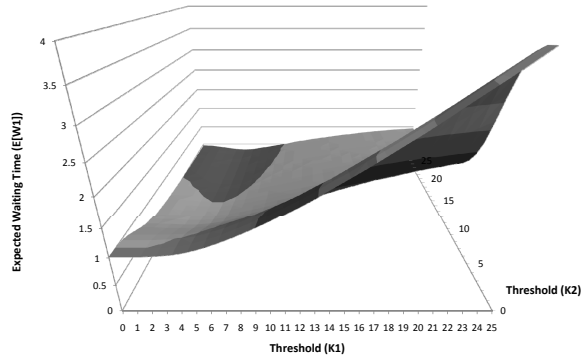


Figure 2: The function  $\mathbb{E}[W_1(\cdot, \cdot)]$  for  $(\lambda_1, \lambda_2, N_1, N_2, \mu) = (4.32, 5.3, 5, 6, 1)$ ,  $\lambda_i^a = 0.25\lambda_i$

payoff functions do not satisfy any monotonicity or convexity properties. From Figure 3, we see that the average waiting time at ED 1 is not a strictly increasing function of the threshold  $K_1$  for a given threshold  $K_2$  of ED 2. As a result, choosing a threshold of 0 is not a dominant strategy. Indeed, there exist values of  $K_2$  for which the best response of ED 1 is to choose a strictly positive threshold, thus occasionally accepting patients that are diverted from ED 2.

The fact that ED  $i$ 's pay-off is not monotone in its threshold is a result of the “non-linearity” of the

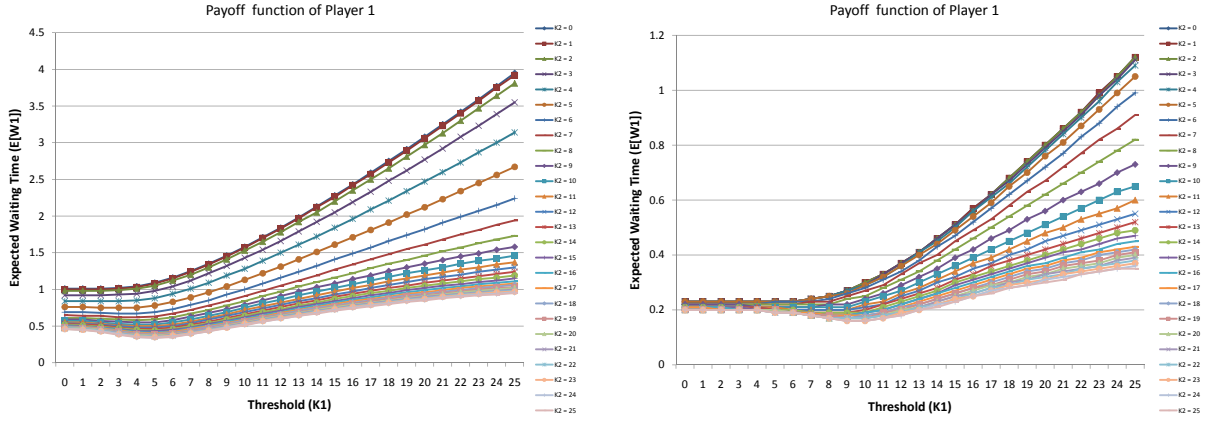


Figure 3: Payoff of ED 1 as a function of its threshold ( $K_1$ ) for different values of ED 2's threshold ( $K_2$ ): (i)  $(\lambda_1, \lambda_2, N_1, N_2, \mu) = (4.32, 5.3, 5, 6, 1)$ ,  $\lambda_i^a = 0.25\lambda_i$  (ii)  $(\lambda_1, \lambda_2, N_1, N_2, \mu) = (8.1, 9.9, 10, 10, 1)$ ,  $\lambda_i^a = 0.25\lambda_i$

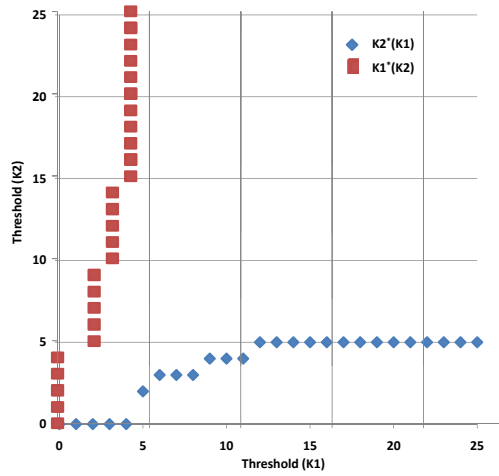


Figure 4: Best response dynamics for  $\lambda_1 = 4.32$ ,  $\lambda_2 = 5.3$ ,  $N_1 = 5$ ,  $N_2 = 6$

queueing-network dynamics. By accepting some of ED 2's patients when ED 2 is crowded (but ED 1 is not), ED 1 helps decrease the crowding in ED 2. Being less crowded, ED 2 will go on diversion less frequently and this might allow ED 1, when it is itself crowded, to divert more ambulances to ED 2. This nonlinear feedback makes it beneficial for ED 1 to occasionally accept diverted ambulances from ED 2.

The ADND guideline, designed by the EMS agency, plays an important role here since when both EDs are on diversion, each ED gets its own arrivals. ED 1 can then accept some ambulances knowing that when it is itself very crowded (and on diversion), it will not have to receive arrivals from ED 2.

Figure 4 depicts the best response functions for the two EDs for one instance of arrival rates and capacities. The “diamond” series represents the best response of ED 2 to different thresholds of ED 1. On a scale rotated at 90 degrees, the “square” series represents the best response of ED 1 to different thresholds of ED 2. As can be observed from the figure, despite  $K_i = 0$  **not** being a dominant strategy for ED  $i$ ,  $K = (0, 0)$  is a Nash equilibrium. The following theorem proves that this numerical observation holds in general.

**Theorem 2** *Suppose that (1) holds. Then,*

$$\mathbb{E}[W_1(0, 0)] < \mathbb{E}[W_1(K_1, 0)] \text{ and } \mathbb{E}[W_2(0, 0)] < \mathbb{E}[W_2(0, K_2)] \quad \forall K_1, K_2 > 0. \quad (3)$$

*In particular,  $K = (0, 0)$  is a Nash equilibrium in the diversion game.*

Thus, starting at  $(0, 0)$ , any unilateral deviation by one of the EDs strictly increases its waiting time. While it is clear that the queue length of ED 1 will be larger under  $(K_1, 0)$  than under  $(0, 0)$  for  $K_1 > 0$ , it does not follow immediately from Little’s Law that the steady-state waiting time will also be larger since the effective arrival rate also increases for  $K_1 > 0$ . In the proof we show that the queue length increases, in a sense, more than the arrival rate when moving from  $(0, 0)$  to  $(K_1, 0)$  for  $K_1 > 0$ .

**Remark 1 (“Defensive” Diversion)** The equilibrium  $(0, 0)$  literally implies that the EDs are always on diversion status. While this is not observed in reality, we interpret the threshold value of 0 as reflecting a choice of small thresholds which, in turn, can be interpreted as defensive behavior of the EDs. Since the payoff functions are fairly flat in a large enough neighborhood of  $K_i = 0$  for small enough thresholds of ED  $j$  (see Figure 3), any choice of thresholds in this neighborhood will result in payoffs (waiting times) that are similar to those under the threshold pair  $(0, 0)$ . In this light, the equilibrium  $(0, 0)$  supports the anecdotal and preliminary empirical evidence that EDs overuse diversion in a defensive manner to avoid getting significant numbers of ambulance arrivals from other EDs. ■

**Remark 2 (“De-pooling” effect)** At the  $(0, 0)$  equilibrium both EDs are always on diversion status which, due to the ADND policy, implies that no ambulances are actually diverted. Consequently, with decentralized decision making, the ED network loses all operational pooling benefits of ambulance diversion. We label this as the *de-pooling effect* of decentralized decision making. With no ambulances being diverted, the two EDs operate like independent  $M/M/N$  queues so that the equilibrium waiting time in the network with decentralized decisions is identical to that in a network in which diversion is formally disallowed. In turn, our equilibrium result provides one possible explanation (that does not hinge on internal process improvements) that the ED waiting times in Massachusetts did not significantly increase after diversion was banned completely in January 2009. ■

#### 4.1 Uniqueness and refinement

It is worth noting that, by Theorem 2,  $(0, 0)$  is an equilibrium regardless of the respective arrival rates or any other system parameters. In fact, as stated in Theorem 3, even if there exist multiple equilibria for the diversion game,  $(0, 0)$  is the only one that has this invariance property.

**Theorem 3** *For any pair  $K = (K_1, K_2) \neq 0$ , there exist values of  $\lambda_1^a, \lambda_2^a > 0$  for which  $K$  is **not** a Nash equilibrium in the diversion game.*

To gain some intuition into the result, we note that if  $\lambda_1^a$  is small (but positive), ED 1 is expected to gain less from diverting its own ambulances compared to avoiding other EDs’ ambulances. Hence it is likely that  $K_1 = 0$  will be ED 1’s best response to any threshold  $K_2$  chosen by ED 2. Showing that ED 1’s best response is to pick  $K_1 = 0$  when  $\lambda_1^a$  is small constitutes the main step in the proof of Theorem 3. The proof of Theorem 3 that appears in the appendix is a formalization of the above intuition.

The uniqueness of  $(0, 0)$  as an equilibrium with arrival-rate invariance has useful implications for ambulance diversion. As EDs experience frequent changes in their arrival rates in the course of the day and over the week, it is an expensive task for ED administrators to continually recalculate their equilibrium strategies and communicate these to the staff for implementation. It thus seems reasonable that if EDs are looking for a simple (and stable) rule-of-thumb for threshold choices,  $(0, 0)$  would be the resulting equilibrium.

The game theory literature discusses the notion of “simple equilibrium” to account for players capabilities [Baron and Kalai 1993]. They contend that simplicity is important in characterizing focal points (introduced by Schelling [1958]) in the case of multiple equilibria. From this perspective, our result in Theorem 3 can be interpreted, informally, as one that characterizes  $(0, 0)$  as the unique simple equilibrium.

While Theorem 3 shows that  $(0, 0)$  is the unique equilibrium with this property, we are not able to prove analytically that it is also unique in the standard sense, i.e, for given arrival rate parameters. Our numerical experiments (see §6), however, show that the equilibrium is not only unique but also globally stable under a Tatônnement scheme [Vives 2001, chapter 2.6] for a wide range of problem parameters.

## 4.2 A formulation with revenue considerations

In our analysis thus far, we have assumed that the objective of each ED is to minimize the average waiting time of its incoming patients. In some cases, however, hospital administrators may be concerned about the impact of diversions on lost revenue. Accordingly, we augment the original formulation by placing a lower bound on the effective arrival rate to each ED. This constraint reflects that the hospital administrator does not want to lose too many arrivals due to diversion. In this new formulation, given the threshold  $K_j$  of ED  $j \neq i$ , ED  $i$  solves the problem

$$\begin{aligned} \min_{K_i \in \mathbb{Z}_+} \quad & \mathbb{E}[W_i(K_i, K_j)] \\ \text{s.t.} \quad & \bar{\lambda}_i(K_i, K_j) \geq \lambda_i(1 - \eta), \end{aligned} \tag{4}$$

where the parameters  $\eta \in [0, 1]$  captures the hospital administrators’ tolerance to losing arrivals. For special case of  $\eta = 1$  we return to our original diversion game. Definition (1) of the diversion game is now changed to incorporate these constraints. Naturally, an equilibrium in the constrained version of the diversion game is a pair  $(K_1, K_2)$  such that no ED has a *feasible* move that would decrease its waiting time. Note that  $(0, 0)$  is feasible for this game since at  $(0, 0)$ ,  $\bar{\lambda}_i(0, 0) = \lambda_i$  for  $i = 1, 2$ . Also, a move by ED 1 to  $K_1 > 0$  is feasible (since  $K_2 = 0$ , the effective arrival rate to ED 1 increases from this move). At the same time we have already shown in Theorem 2 that any such move would strictly increase the waiting time in ED 1. Hence, ED 1 does not have a profitable deviation. A symmetric argument applies to ED 2. Thus,  $(0, 0)$  is an equilibrium also for the constrained diversion game.

**Remark 3 (Alternative modeling of revenues)** The constrained formulation considered above allows for revenue/reputation concerns by constraining the number of diverted ambulances. It is, however, most appropriate in settings in which ambulance diversion is primarily driven by overcrowding in the ED and is not fine-tuned to maximize revenues. This seems consistent with the common practice (see e.g. Franaszek et al. [2002]). Nevertheless, one could consider an alternative model in which revenues are explicitly modeled into the objective function by assigning weights to revenues and waiting times. The resulting model is

less tractable but, more importantly, the outcome of such a model would most likely depend on the specific values of the weights, which are difficult to estimate. ■

## 5 Centralized setting

To evaluate the inefficiency in the decentralized setting described in §4, we consider a setting where the social planner chooses a threshold pair  $K^* = (K_1^*, K_2^*)$  to minimize the average waiting time across all patients in the network. To that end, let  $Q_\Sigma^K$  be the steady-state total queue length across the network under the threshold pair  $K = (K_1, K_2)$ . By Little’s law, the social planner’s problem is given by:

$$\min_{K \in \bar{\mathbb{Z}}_+^2} \mathbb{E}[W(K_1, K_2)] := \min_{K \in \bar{\mathbb{Z}}_+^2} \frac{\mathbb{E}[Q_\Sigma^{(K_1, K_2)}]}{\lambda_1 + \lambda_2}, \quad (5)$$

where  $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+ \cup \{\infty\}$ <sup>3</sup>. To be consistent with the decentralized setting, we do not include the impact of ambulance diversion on transportation and other social costs. Henceforth, we use the notation  $(K_1^*, K_2^*)$  to denote an optimal solution. The statements of our results are unaffected by whether or not multiple optimal solutions exist to the social planner’s problem.

**Remark 4 (Alternative centralized routing policies)** Restricting attention to threshold policies has the important benefit of allowing us to perform a clear comparison between the centralized setting and the decentralized one. Clearly, the social planner can do better with other routing policies that take into account real-time queue-length or waiting-time information for both EDs, such as the “Join the Shortest Queue” policy [Adan et al. 1994, Whitt 1986]. While these policies are not necessarily optimal in our setting because of multiple servers in each queueing system and because only part of the arrivals (ambulances) can be actively controlled, it is possible that they perform better than the best threshold policy; see §6. ■

As argued in §4, the function  $\mathbb{E}[W(\cdot, \cdot)]$  is not amenable to exact analysis. Instead, we develop a lower bound and an upper bound to the social planner’s problem. The lower bound is given by a perfectly pooled system with instantaneous rerouting and the upper bound is given by a system where the threshold pair is  $K = (N_1, N_2)$  – capacity-based thresholds. Figure 5 depicts the system configuration for the upper bound and lower bound in addition to the decentralized and the centralized settings.

### 5.1 A perfectly pooled system with instantaneous re-routing (lower bound)

We now introduce an artificial construct which is obtained by relaxing the social planner’s problem. Specifically, we make the following allowances in patient routing:

- (i) The planner has complete freedom in routing ambulances and walk-ins and is not restricted to threshold-type policies.
- (ii) The planner can *instantaneously* re-route waiting patients (walk-ins and ambulances) from one ED to the other. For example, if ED 2 has available beds, a patient waiting in ED 1 can be moved instantaneously to ED 2.

---

<sup>3</sup>A priori, the social planner might choose to set one threshold or both thresholds to infinity which is not covered by Theorem 1. This is overcome by appropriately replacing the steady state expectation with long run averages.

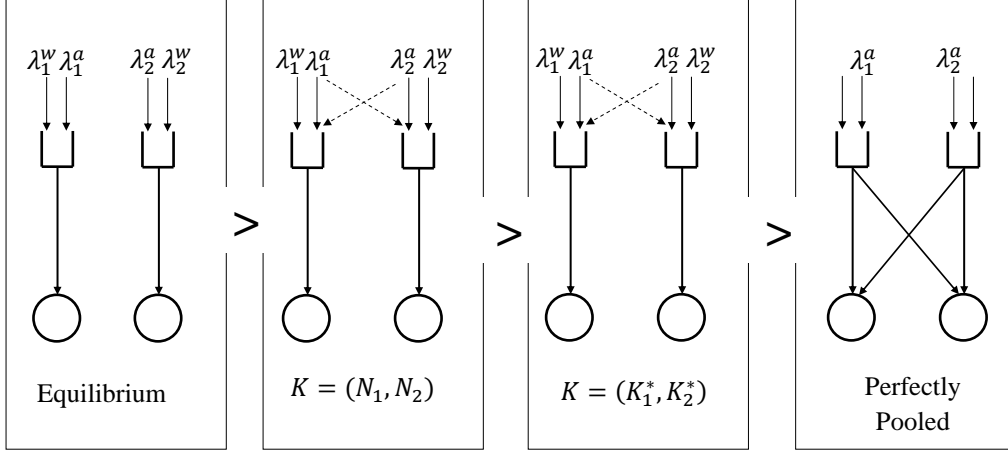


Figure 5: From equilibrium to perfectly pooled. The inequalities denote the ordering of total waiting times under different scenarios.

We refer to this as the *perfectly pooled* system since the instantaneous re-routing disallows having simultaneously empty beds in one ED and waiting patients in the other. In doing so, it essentially pools the queues and avoids the wastage of capacity in the network. Note that the perfectly pooled system is an unfair benchmark for the ED network since it has two significant degrees of freedom that the original ED network does not have: it allows for rerouting of walk-ins (and not only of ambulances) and it allows for rerouting of patients after they have been waiting in a queue. With identical service rates, any non-idling routing policy for the perfectly pooled system will induce the same distribution for the total number of patients in system, that of an  $M/M/N$  queue with arrival rate  $\lambda_\Sigma = \lambda_1 + \lambda_2$ , service rate  $\mu$  and  $N_\Sigma = N_1 + N_2$  servers.

More formally, let  $X^P(t) = (X_1^P(t), X_2^P(t))$  denote the number of patients in each ED in the perfectly pooled system at time  $t$  and we let  $X^P(\cdot) = (X^P(t), t \geq 0)$  be the corresponding stochastic process. We denote by  $Q_\Sigma^P(t)$  the corresponding total *queue* length at time  $t$  and we let  $Q_\Sigma^P$  be a random variable with the corresponding steady-state distribution. Then,  $Q_\Sigma^P$  has the steady-state distribution of the number of customers in an  $M/M/N$  queue with arrival rate  $\lambda_\Sigma$ , service rate  $\mu$  and  $N_\Sigma$  servers. This steady-state exists since condition (1) implies that  $\lambda_\Sigma < \mu N_\Sigma$ .

In turn,  $\mathbb{E}[W^P] := \mathbb{E}[Q_\Sigma^P]/\lambda_\Sigma$  is the expected waiting time in this  $M/M/N$  queue and is the same (and minimal) under all non-idling policies. Let

$$\tilde{\rho}_1 := \frac{\lambda_1^w + \mu N_2}{\mu N_1 + \lambda_2 + \lambda_1^a}, \quad \tilde{\rho}_2 := \frac{\lambda_2^w + \mu N_1}{\mu N_2 + \lambda_1 + \lambda_2^a}, \quad \text{and } C(\tilde{\rho}_1, \tilde{\rho}_2) := \left[ \frac{1}{1 - \tilde{\rho}_1} + \frac{1}{1 - \tilde{\rho}_2} \right]. \quad (6)$$

Note that  $\max\{\tilde{\rho}_1, \tilde{\rho}_2\} < 1$  if, for example,

$$\mu N_1 \leq \lambda_1 + \lambda_2^a \quad \text{and} \quad \mu N_2 \leq \lambda_2 + \lambda_1^a. \quad (7)$$

**Theorem 4** *Suppose that (7) holds. Then, given  $0 < \epsilon < 1$ , there exists  $T(\epsilon)$  that depends, in addition, only*

on  $\mu$  and the aggregate parameters  $\lambda_\Sigma$  and  $N_\Sigma$  such that

$$\mathbb{E}[W^P] \leq \mathbb{E}[W(K_1^*, K_2^*)] \leq \frac{\mathbb{E}[W^P]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)}. \quad (8)$$

The constant  $T(\epsilon)$ , defined explicitly in equation (A14) of the appendix, is the time that it takes the expected queue length in an  $M/M/N$  queue, initialized at time 0 with a carefully chosen distribution, to approach its steady-state mean. Importantly, as stated in the theorem,  $T(\epsilon)$  depends only on the aggregate parameters  $\lambda_\Sigma$  and  $N_\Sigma$  and on the service rate  $\mu$ . Also, for reasonably large systems  $T(\epsilon)/(\lambda_\Sigma/\mu)$ , is provably orders of magnitude smaller than  $\mathbb{E}[W^P]$ . See Remark A.1 in the appendix for more details.

Theorem 4 shows that, under proper conditions, the social optimum is fairly close to the lower bound. Moreover, it provides an indication of potential dependencies of the gap between the social optimum and its lower bound on different system parameters. Fixing  $\tilde{\rho}_1$  and  $\tilde{\rho}_2$ , the bound becomes tighter as the overall system size, as reflected in the parameters  $\lambda_\Sigma$  and  $N_\Sigma$ , increases. For a given system size, the bound becomes tighter as the parameters  $\tilde{\rho}_1$  and  $\tilde{\rho}_2$  and, in turn,  $C(\tilde{\rho}_1, \tilde{\rho}_2)$ , become smaller. Note that small values of  $\tilde{\rho}_1$  and  $\tilde{\rho}_2$  correspond to a large fraction of ambulances (out of the total incoming patients). Intuitively, with sufficient ambulance load, a stream of ambulances will be routed from one ED to the other, thus quickly removing the “inefficiency” of having a queue in one ED simultaneously with available beds at the other.

## 5.2 Capacity-based static thresholds (upper bound)

We consider the case where  $K = (N_1, N_2)$ , i.e., each ED uses its number of beds as its threshold. Here, when ED 1 has available beds but ED 2 has waiting patients, all (newly) arriving ambulances are diverted to ED 1. Our motivation for considering this threshold pair is that it mimics, to the extent possible within threshold policies, the response mechanism of the perfectly pooled system to local inefficiencies. When there are available beds in one ED simultaneously with queued patients at the other, it routes all the “routable” patients to the ED with available beds. Since walk-in and patients that are already queued are not “routable” in the real ED network, the extent to which this can re-balance the system is less than that in the perfectly pooled system. However, the next result shows that, despite this restriction, the diversions here are sufficient to capture most of the gap between the social optimum and the idealized system.

**Theorem 5** Fix  $\epsilon > 0$ . Suppose that the conditions of Theorem 4 hold and let  $T(\epsilon)$  be as defined there. Then,

$$\mathbb{E}[W(K_1^*, K_2^*)] \leq \mathbb{E}[W(N_1, N_2)] \leq \frac{\mathbb{E}[W(K_1^*, K_2^*)]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)}.$$

By our discussion following Theorem 4, one expects  $(N_1, N_2)$  to perform well for reasonably large systems<sup>4</sup>. Our numerical experiments in §6 show that the gap between  $\mathbb{E}[W(K_1^*, K_2^*)]$  and  $\mathbb{E}[W(N_1, N_2)]$  is small for relatively small systems too. In addition to being close to the true social optimum, the approximate

<sup>4</sup>Heavy-traffic analysis, as the one in Tezcan [2008] can, in fact, show that the threshold policy with the threshold pair  $(N_1, N_2)$  is nearly optimal for (5). The driving force in the near optimality result as in our bounds, is the relatively quick “re-balancing” driven by ambulance re-routing.

social optimum has several advantages from an implementation perspective, which we discuss below.

**Remark 5 (Implementation)** The pair  $(N_1, N_2)$ , unlike the social optimum, does not require the knowledge of the system parameters  $(\lambda_i^a, \lambda_i^w)$ ,  $i \in \{1, 2\}$  and  $\mu$ . Moreover, the simplicity of the solution makes it useful for coordination purposes. Specifically, rather than trying to enforce specific threshold values, the coordinating body (e.g. the local EMS agency) could devise the following policy: “EDs are not allowed to divert when there are empty beds.” In Figure 3, we observe that  $\mathbb{E}[W_i(K_1, K_2)]$  is decreasing in  $K_i$  for any threshold  $K_j \geq N_j$  of the other ED. Based on this numerical observation, we note that the threshold pair  $(N_1, N_2)$  will be supported as an equilibrium for a diversion game in which bed reservation is disallowed. This numerical observation, together with our near optimality result in Theorem 5 suggests that such a regulation can induce an equilibrium that is also nearly optimal for the social planner’s problem. Theorems 6 and 7 below show that  $(N_1, N_2)$  has additional appealing properties. This discussion presumes that ED administrators have correct information about the number of busy beds and that they do not manipulate this information (by purposefully delaying transfers to inpatient departments). ■

Theorem 5 is concerned with the waiting time averaged over all patients in the network. The next theorem is concerned with the waiting time in each ED separately. It compares the waiting times under the threshold pair  $(N_1, N_2)$  and under the defensive equilibrium  $(0, 0)$ . Below,  $p(\lambda, \mu, N)$  is the probability of delay (i.e. of all servers being busy) in an  $M/M/N$  queue with arrival rate  $\lambda$ , service rate  $\mu$  and  $N$  servers. Formally,  $p(\lambda, \mu, N) := \mathbb{P}\{W_{M(\lambda), M(\mu), N} > 0\}$ , where  $W_{M(\lambda), M(\mu), N}$  is the steady-state distribution of the waiting time in an  $M/M/N$  queue with parameters  $\lambda, \mu$  and  $N$  that satisfy  $\lambda < \mu N$ .

**Theorem 6 (Pareto improvement)** *Let  $(\lambda_1, \lambda_2, N_1, N_2)$  be such that*

$$p(\lambda_\Sigma, \mu, N_\Sigma) \leq \min\{p(\lambda_1, \mu, N_1), p(\lambda_2, \mu, N_2)\}. \quad (9)$$

*Fix  $\epsilon$  and let  $T(\epsilon)$  be as in Theorem 4. Then,*

$$\mathbb{E}[W_i(N_1, N_2)] \leq \mathbb{E}[W_i(0, 0)] \left(1 - 2\sqrt{\frac{\mu}{\lambda_i}}\right)^{-1} \left(1 + 2\frac{\delta_i}{\mathbb{P}\{W_i(0, 0) > 0\}}\right), \quad i = 1, 2, \quad (10)$$

where  $\delta_i := \sqrt{(\mu N_i - \lambda_i) \left(\frac{2\mu T(\epsilon) C(\hat{\rho}_1, \hat{\rho}_2)}{\lambda_i} + \frac{\epsilon \lambda_\Sigma}{\lambda_i(\mu N_\Sigma - \lambda_\Sigma)}\right)}$ .

The condition on the  $M/M/N$  delay probabilities in equation (9) requires that  $(\mu, \lambda_1, \lambda_2, N_1, N_2)$  are such that pooling two independent  $M/M/N$  queues with the respective parameters into a single  $M/M/N$  queue improves the delay probability for everyone. This holds for many, but not all, parameter combinations. In particular, it is not likely to hold in settings in which there is significant asymmetry between the EDs with respect to their parameters. For example, for the case  $\mu = 1$ , and  $N_1 = N_2 = 10$ , this condition holds for  $\lambda_1 = 7.5$  and  $\lambda_2 = 9$  but it does not hold if  $\lambda_1 = 6$  and  $\lambda_2 = 9$ .

When condition (9) is satisfied, Theorem 6 implies that the expected waiting time in each ED does not increase significantly if moving from the  $(0, 0)$  equilibrium to the capacity-based threshold pair  $(N_1, N_2)$ .

Note that  $\delta_i \leq \sqrt{\frac{1 - \rho_i}{\rho_i} 2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \epsilon \frac{\lambda_\Sigma}{\lambda_i}}$ , where  $\rho_i = \lambda_i / \mu N_i$  is the utilization of ED  $i$  in the absence of diversion. Thus, the bound on the right hand side of (10) is tight provided that ED  $i$  has a sufficiently high load,  $\lambda_i$  is not negligible compared to  $\lambda_\Sigma$  and  $T(\epsilon)$  is not large. In our numerical experiments in §6 the bound is tight and the waiting time of each ED strictly decreases when moving from  $(0, 0)$  to  $(N_1, N_2)$ .

The next result completes our analysis of the threshold pair  $(N_1, N_2)$ . It shows that the EDs do not experience a significant decrease in their arrival rate under  $(N_1, N_2)$  compared to that under the  $(0, 0)$  equilibrium.

**Theorem 7 (Effective arrivals under capacity-based thresholds)** *Recall that the effective arrival rate  $\bar{\lambda}_i(N_1, N_2)$  is given by (2). Then,  $\bar{\lambda}_i(N_1, N_2) \geq \lambda_i - 2\sqrt{\mu\lambda_i}$ ,  $i = 1, 2$ .*

Note that Theorem 7 does not impose any restrictions on the network parameters. Naturally, the bound is trivial unless  $\lambda_i > 2\sqrt{\mu\lambda_i}$ .

**Remark 6 (Re-visiting the constrained formulation)** Note that given  $\eta > 0$ , Theorem 7 guarantees that the threshold pair  $(N_1, N_2)$  is feasible for (4) for all  $\lambda_i$  sufficiently large (which requires, by (1) that  $N_i$  be sufficiently large). Thus, the implications from the analysis of  $(N_1, N_2)$  also carry over to the constrained formulation of the diversion game introduced in §4.2. In particular, in the context of the constrained version of the diversion game, the pair  $(N_1, N_2)$  is Pareto improving and nearly socially optimal. ■

## 6 Numerical results

Our numerical study has three main objectives. First, we study the impact of key system parameters on the outcomes under defensive equilibrium, socially optimal thresholds and (approximately optimal) capacity-based thresholds and on the relationship between the three. We explore the following questions: (i) how does the total safety capacity in the network impact the extent of inefficiency introduced to the system by the decentralized equilibrium (§6.3), and (ii) how does the difference in safety capacity between the EDs affect this inefficiency (§6.4). Second, for each parameter set we study the performance of the threshold pair  $(N_1, N_2)$  (§6.5). Third, we complement our results regarding the equilibrium  $(0, 0)$  in Theorems 2 and 3 by numerically testing for its global stability (§6.6).

### 6.1 Methodology

For the numerical experiments we explicitly calculate the stationary distribution of the underlying Markov chain. To numerically compute the stationary distributions, we truncate the state space by removing all states with queue length above a very large number  $B$ . We choose  $B$  to be large enough so that the numerical results do not change with further increase in its value. More formally, we truncate the CTMC  $X^K(\cdot)$  to the state space  $\mathcal{A}_B := [0, B] \times [0, B]$ . The truncation is done by removing all edges between states  $x \in \mathcal{A}_B$  and  $y \notin \mathcal{A}_B$ . The resulting process,  $X^{K,B}(\cdot)$ , is an irreducible Markov chain on a bounded state space and hence a unique steady-state distribution trivially exists and can be numerically computed by means of solving the balance equations  $\pi^B Q^B = 0$  with  $Q^B$  being the generator of the truncated chain. These computations are performed using a MATLAB code that solves the underlying system of linear equations. As expected,  $\pi^B$  converges weakly to  $\pi$  as  $B \rightarrow \infty$ , where  $\pi$  is the unique steady-state distribution of the

original (non-truncated) chain  $X^K(\cdot)$  and, moreover, the corresponding expectations converge. The proof of this fact is based on showing that the (family) of  $X^{K,B}$  of steady-state variable is tight in the proper sense (see appendix). This convergence, together with a choice of large enough truncation parameter  $B$ , guarantees that our results are representative of the original network.

## 6.2 Choice of parameters

We use two sets of parameter combinations. The first set is designed to investigate the impact of total safety capacity in the network. Here, we fix  $N_1 = 5$  and  $N_2 = 6$ ,  $\mu = 1$  and we choose arrival rates  $\lambda_1$  and  $\lambda_2$  so that the utilization (with no diversion) is equal in both EDs (i.e.  $\lambda_1/N_1 = \lambda_2/N_2 = \rho$ ). We then vary the value of  $\rho$  on  $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ . The second set is designed to investigate the impact of differences in safety capacity between the two EDs. Here, we fix  $N_1 = N_2 = 10$ ,  $\mu_1 = \mu_2 = 1$  and total arrival rate  $\lambda_\Sigma = \lambda_1 + \lambda_2 = 18$ . We then split the total arrival rate between the two EDs into different combinations of  $\lambda_1$  and  $\lambda_2$  so that the pair of (no-diversion) utilizations  $\rho_1 := \lambda_1/N_1$  and  $\rho_2 := \lambda_2/N_2$ , takes the values:  $\{(0.81, 0.99), (0.828, 0.972), (0.846, 0.936), (0.882, 0.918), (0.9, 0.9)\}$ . For all parameter combinations, we fix  $\lambda_i^a = 0.25\lambda_i$  in accordance with empirical evidence [Falvo et al. 2007].

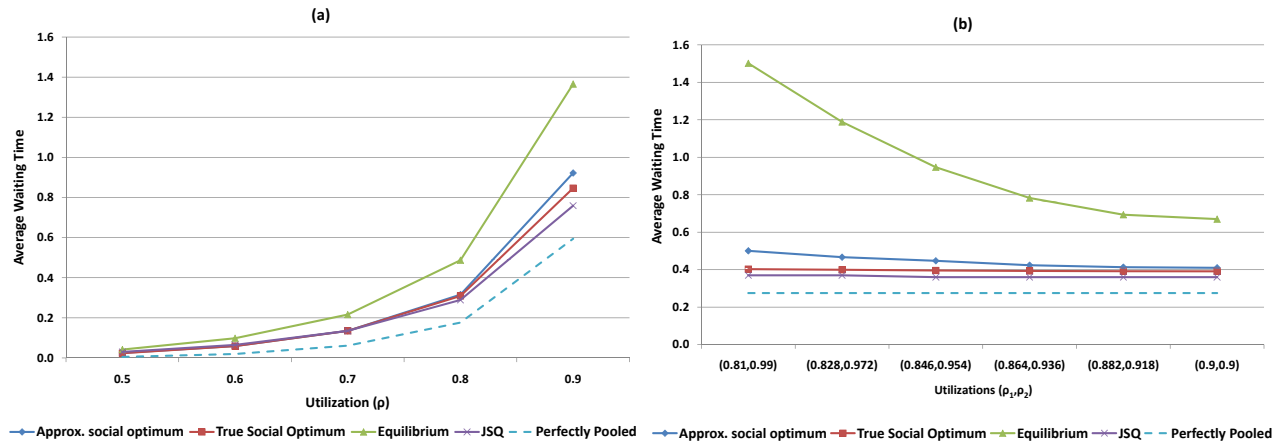


Figure 6: Waiting time as a function of utilization parameters: (a) Case of equal utilizations (b) Case of unequal utilizations

## 6.3 Impact of total safety capacity

Figure 6(a) shows the expected waiting time in the defensive equilibrium  $(0, 0)$  (the decentralized network), the social optimum (the centralized network), the approximate social optimum  $(N_1, N_2)$ , the “Join the Shortest Queue” policy (JSQ)<sup>5</sup>, and the perfectly pooled system for different values of the total utilization  $\rho$ , where both EDs have equal utilization.

As expected, the waiting time under all scenarios increases with  $\rho$ . Moreover, the gap between the social optimum and the equilibrium increases with  $\rho$ . This can be explained via the connection to pooling / de-pooling. It is known, that the benefits of pooling two independent  $M/M/N$  queues with identical service rates increases with the utilization (see, e.g., [Benjaafar 1995]). Since the equilibrium leads to complete de-pooling while the social optimization is fairly close to a perfectly pooled system, one expects that the benefit

<sup>5</sup>The performance of JSQ is obtained using an ARENA simulation model.

of social optimization will increase with an increase in  $\rho$ . Figure 6(a) also shows that the performance of the approximate and the true social optimum (restricted to static thresholds) is very close to that of the JSQ policy. The waiting time is slightly higher under JSQ for the low utilization values  $\rho = 0.5$  and  $\rho = 0.6$  and it is slightly lower than the social optimum otherwise.

Table 2 displays the socially optimal thresholds,  $K_1^*$  and  $K_2^*$ . These were found by brute force enumer-

Table 2: Socially optimal thresholds – Equal utilizations

$\rho$	0.5	0.6	0.7	0.8	0.9
$K_1^*$	4	4, 5	5	6	7, 8
$K_2^*$	5, 6	6	6	7	9

ation of all possible combinations of thresholds for values that are below the truncation level  $B$ . We note that the socially optimal thresholds are equal to the respective capacities for low utilization values but are higher than the capacity for higher system utilization. Thus, the social planner chooses higher thresholds in more congested systems. It is also interesting to note that the socially optimal threshold for the smaller ED is lower for all values of system utilization.

#### 6.4 Impact of differences in safety capacity

Figure 6(b) displays the results for parameter combinations in which we keep the total utilization  $\rho = \lambda_\Sigma/N_\Sigma$  constant and equal to 0.9 but vary the individual utilizations  $\rho_1 = \lambda_1/N_1$  and  $\rho_2 = \lambda_2/N_2$  (recall that  $\mu = 1$  throughout) thus introducing asymmetry in the network.

We find that the benefit of social optimization increases as the network becomes more asymmetric, i.e., as the difference in safety capacity between EDs decreases. This again can be explained via the connection to pooling. Since the centralized setting closely imitates the perfectly pooled system, it is fairly indifferent to the local safety capacity in each ED and depends only on the total safety capacity in the network. In the decentralized setting, however, the total waiting time decreases as the EDs become more symmetric. These numerical results are suggestive of interesting research questions regarding the impact of relative asymmetry of the network to the magnitude of pooling benefits, which to the best of our knowledge, has not been studied before in the literature.

Similar to §6.3, the performance of the approximate and true social optimum (restricted to static thresholds) is very close to JSQ policy. In fact, the performance gap between the two is less than 10% for all combinations of utilizations in our numerical experiments. Table 3 show that the socially optimal thresholds are above capacity, i.e.  $K_i^* \geq N_i$ . Moreover, the socially optimal thresholds are higher for EDs with more safety capacity thus enabling better resource pooling.

#### 6.5 Role of the approximate social optimum

In this section, we make important observations regarding the performance of the approximate social optimum  $(N_1, N_2)$ , which are common to both set of experiments under various conditions of total safety capacity and relative safety capacity described above.

Table 3: Socially optimal thresholds – Unequal Utilizations

$(\rho_1, \rho_2)$	(0.810,0.990)	(0.828,0.972)	(0.846,0.954)	(0.864,0.936)	(0.882,0.918)	(0.9,0.9)
$K_1^*$	16	15	14	13	13	12
$K_2^*$	11	11	11	11	12	12

1. **The pair  $(N_1, N_2)$  as an approximation to the optimal solution:** Figure 6 contains waiting times corresponding to the equilibrium ( $\mathbb{E}[W(0,0)]$ ), the true social optimum ( $\mathbb{E}[W(K_1^*, K_2^*)]$ ), and the approximate social optimum ( $\mathbb{E}[W(N_1, N_2)]$ ). The prescription of  $(N_1, N_2)$  performs extremely well for most parameter values. From Figure 6(b), we note that its performance deteriorates slightly with the asymmetry in the network
2. **Pareto Improvement:** As shown in Tables 4 and 5, both EDs gain in terms of waiting times under the approximate social optimum compared to the equilibrium. Such a Pareto improvement does not always hold for the social optimum  $(K_1^*, K_2^*)$ . Indeed, in the first three parameter combinations in Table 5, ED 1 has a longer waiting time under the socially optimal threshold pair than under equilibrium, i.e.,  $\mathbb{E}[W_1(K_1^*, K_2^*)] > \mathbb{E}[W_1(0,0)]$ . In these first three sets  $\rho_1$  is relatively low and it seems that the social optimization exploits this fact so that the more congested ED 2 gets greater share of the pooling benefit obtained from social optimization. Under  $(N_1, N_2)$  ED 2 still gets more benefit, but ED 1 does not experience degradation in its waiting times compared to  $(0,0)$ . As argued in §5, this observation may be important for coordination purposes because ED managers are more likely to accept coordination mechanisms that do not hurt them compared to the equilibrium.
3. **Patient inflow:** As stated in Theorem 6, the total inflow of patients to each ED does not differ significantly between  $(N_1, N_2)$  and  $(0,0)$ . Indeed, the ratio  $\bar{\lambda}_i/\lambda_i$  is less than 102% in Table 4 and is less than 105% in Table 5. Moreover, this ratio increases with the asymmetry in the network.

Table 4: Effective arrival rate and average waiting time for two EDs – Equal utilization case

$\rho$	Equilibrium				Approximate Social Optimum				Social Optimum			
	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	$\lambda_1$	$\lambda_2$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	$\lambda_1$	$\lambda_2$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	$\lambda_1$	$\lambda_2$
0.5	0.05	0.03	2.50	3.00	0.03	0.02	2.55	2.95	0.03	0.02	2.33	3.00
0.6	0.12	0.08	3.00	3.60	0.07	0.05	2.95	3.65	0.07	0.05	3.00	3.60
0.7	0.25	0.19	3.50	4.20	0.15	0.12	3.53	4.17	0.15	0.12	3.60	4.17
0.8	0.55	0.43	4.00	4.80	0.35	0.29	4.02	4.78	0.35	0.28	3.97	4.75
0.9	1.52	1.23	4.50	5.40	1.01	0.85	4.49	5.41	0.88	0.82	4.47	5.43

## 6.6 Stability of the equilibrium

As seen from Figure 4, if ED  $j$  chooses a threshold  $K_j > N_j$ , ED  $i$ 's best response is to choose a threshold  $K_i = N_i$ . However, for this choice of ED  $i$ , ED  $j$ 's best response is to choose zero, to which ED 1's

Table 5: Effective arrival rates and average waiting time for two EDs – Unequal utilizations

$(\rho_1, \rho_2)$	Equilibrium				Approximate Social Optimum				Social Optimum			
	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	$\bar{\lambda}_1$	$\bar{\lambda}_2$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	$\bar{\lambda}_1$	$\bar{\lambda}_2$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	$\bar{\lambda}_1$	$\bar{\lambda}_2$
(0.81,0.99)	0.23	2.90	8.10	9.90	0.23	0.76	8.71	9.29	0.37	0.44	8.93	9.07
(0.828,0.972)	0.28	2.17	8.28	9.72	0.25	0.67	8.76	9.24	0.36	0.43	8.16	9.74
(0.846,0.954)	0.34	1.59	8.46	9.54	0.29	0.59	8.61	9.39	0.37	0.43	9.07	8.93
(0.864,0.936)	0.42	1.16	8.64	9.36	0.32	0.52	8.81	9.19	0.37	0.42	9.07	8.93
(0.882,0.918)	0.52	0.87	8.82	9.18	0.36	0.46	8.87	9.13	0.38	0.4	8.95	9.05
(0.9,0.9)	0.67	0.67	9.00	9.00	0.41	0.41	9.00	9.00	0.39	0.39	8.97	9.03

best response is zero too thus yielding an equilibrium  $(0, 0)$ . We numerically validate this Tatônnement stability result for each of the parameter combinations by initializing the game at several finite threshold pairs  $(K_1, K_2)$ ; see Chapter 2.6 of Vives [2001] for an explanation of Tatônnement stability. For our 11 parameter sets, the maximum number of plays to reach  $(0, 0)$  is equal to 2. The global stability of  $(0, 0)$  as observed in these numerical examples strongly suggests that  $(0, 0)$  is unique.

## 7 Concluding remarks

In this paper, we propose a stylized queueing-network model of two EDs to study the network effect of ambulance diversion. Each ED aims to minimize the expected waiting time of its patients (walk-ins and ambulances) and chooses its diversion threshold based on the number of patients at its location. We model the decentralized decision making in the network as a non-cooperative game. Analysis of the game reveals that, in equilibrium, EDs declare diversion status *defensively* to avoid getting arrivals from each other. This equilibrium undermines all potential pooling benefits of ambulance diversion, a phenomenon we label as the *de-pooling effect*. These results provide one potential explanation for the evidence regarding defensive diversion and the impact of canceling ambulance diversion in Massachusetts in Jan 2009.

Given the pessimistic predictions of the equilibrium and the difficulty in characterizing the true socially optimal solution, we propose and analyze an alternative solution to the social planner’s problem in which the diversion thresholds are set to be equal to the EDs’ respective capacities. When there are available beds in one ED simultaneously with queued patients at the other, this policy routes all the “routable” patients to the ED with available beds and thus recovers most of the pooling benefits. In addition to its being easier to implement than the true social optimum, it is also a Pareto improvement over the equilibrium, i.e., it reduces the expected waiting times of both EDs.

Obvious extensions of our work include relaxing the model assumptions and choices discussed in §3.2 such as non-Markovian distributions, non-stationary arrivals, patient priorities and patient boarding. Here, we propose two directions for future research that, we believe, can significantly improve our understanding of ambulance diversion and its effective management.

**Travel time:** We do not explicitly model ambulance travel times, which is reasonable if the differences in travel times to the different EDs are negligible compared to the waiting times in the EDs. Significant travel

time heterogeneity may, however, limit the ability to divert some ambulances due to the potential impact of the increased travel time on patients' health. While it is likely that  $(0, 0)$  will remain an equilibrium in this case, the social planner's ability to extract pooling benefits from centralized diversion might be limited because travel times would delay the impact of diversion.

**Optimal trigger for diversion:** Motivated by the literature, we used the number of patients in the ED as a measure for ED crowding. Others have proposed and used number of boarding patients for this purpose. However, both measures capture only one aspect of reality and are interdependent outcomes of underlying capacity constraints in the ED and in inpatient department. Several complex workload measures have been proposed in the emergency medicine literature to measure crowding but their potential implication on ambulance diversion has not been studied. This would be an important step in the design of good crowding measures for EDs.

## Appendix

This appendix includes the proofs of the key results in the paper. In §A.1 we prove Theorem 2 and in §A.2 we prove Theorems 4 and 5. The proofs of Theorems 1, 3, 6 and 7 as well as those of auxiliary lemmas and propositions that are stated within this appendix are relegated to a technical report [[reference to authors' website would be given here](#)]. In the technical report the reader will also find the proof of existence of solutions to the social planner's problem (5) and the proof of convergence of the truncated chains that were used in our numerical experiments in §6.

We first specify the notational conventions that will be used throughout the appendix.

**State descriptors and threshold notation:** Given a threshold pair  $K = (K_1, K_2)$ ,  $X_i^K(t)$  stands for the number of patients (in beds or waiting) in ED  $i$  at time  $t$  and we let  $X^K(t) = (X_1^K(t), X_2^K(t))$ . The CTMC  $X^K(\cdot) = (X^k(t), t \geq 0)$  then captures the two dimensional behavior of the ED network. For  $i = 1, 2$ , we let  $Q_i^K(t)$  and  $Z_i^K(t)$  be, respectively, the queue length and the number of occupied beds in ED  $i$  at time  $t$  so that  $X_i^K(t) = Z_i^K(t) + Q_i^K(t)$ . The notation  $I_i^K(t) = N_i - Z_i^K(t)$  then stands for the number of available (idle) beds in ED  $i$  at time  $t$ .

We add the subscript  $\Sigma$  to denote aggregate network quantities. For example,  $Q_\Sigma^K(t) = Q_1^K(t) + Q_2^K(t)$  is the total number of patients in queue in both EDs at time  $t$  when the threshold pair  $K$  is used.

As in §5 of the manuscript, we add the superscript  $P$  to distinguish quantities corresponding to the perfectly pooled system from those corresponding to the threshold system. For example,  $Q_\Sigma^P(t)$  stands for the total number of patients in queue in the perfectly pooled system at time  $t$ .

For brevity, we use the term “*the  $(K_1, K_2)$ -threshold system*” when referring to the ED network operating under the threshold mechanism with the threshold pair  $(K_1, K_2)$  and use the term “*the pooled system*” when referring to the ED network operating under the perfectly pooled mechanism.

Two threshold pairs are of special interest and will appear repeatedly in this appendix. These are  $(0, 0)$  and  $(N_1, N_2)$ . We will abbreviate and use the superscript 0 when the threshold pair is  $(0, 0)$  and use the superscript  $N$  when the threshold pair is  $(N_1, N_2)$ .

**Steady-state quantities:** When steady-state exists for the CTMC  $X^K(\cdot)$ , steady-state quantities (queue length, number of patients etc.) appear without the time argument  $t$ . Accordingly,  $X_i^K$  stands for the *steady-*

state number of patients at ED  $i$ . The vector  $\pi^K = (\pi^K(i, j), (i, j) \in \mathbb{Z}_+^2)$  is then the distribution of  $X^K = (X_1^K, X_2^K)$ . That is, for  $(i, j) \in \mathbb{Z}_+^2$ ,  $\pi^K(i, j) = \mathbb{P}\{X_1^K = i, X_2^K = j\}$ . We let  $\pi_i^K$  be the marginal steady-state distribution of ED  $i$ , i.e.,  $\pi_i^K(j) = \mathbb{P}\{X_i^K = j\}$  for  $j \in \mathbb{Z}_+$ .

**Other notational conventions:** For  $x, y \in \mathbb{R}$  we use  $x \wedge y = \min\{x, y\}$  and  $x \vee y = \max\{x, y\}$ . We use  $\Rightarrow$  to denote weak convergence of random variables in  $\mathbb{R}^d$ . The dimension  $d$  of the random variable will be clear from the context. This should be distinguished from  $\rightarrow$  which we use to denote convergence of series of deterministic vectors.

Finally, we use  $\leq_{st}$  to denote stochastic ordering. For two random variables  $R_1$  and  $R_2$  with values in  $\mathbb{Z}_+$ ,  $R_1 \leq_{st} R_2$  if  $\mathbb{P}\{R_1 > j\} \leq \mathbb{P}\{R_2 > j\}$  for all  $j \in \mathbb{Z}_+$ . This is equivalent to having  $\mathbb{E}[f(R_1)] \leq \mathbb{E}[f(R_2)]$  for all non-decreasing functions  $f(\cdot)$  for which the expectations are defined; see e.g. §A4 of Asmussen [2003].

## A.1 Proofs for §4

### A.1.1 Proof of Theorem 2

In establishing (3) we focus on ED 2. The proof for ED 1 is identical. First, observe that using Little's law we can write  $\mathbb{E}[W_2(0, K_2)] = \mathbb{E}[Q_2^{(0, K_2)}] / \bar{\lambda}_2(0, K_2)$ . It is intuitively clear that  $\mathbb{E}[Q_2^{(0, K_2)}] \geq \mathbb{E}[Q_2^{(0, 0)}]$  because the arrivals to ED 2 will increase but it will not be able to divert ambulances to ED 1 because  $K_1 = 0$ . This can be formalized by a simple coupling argument but to prove that  $\mathbb{E}[W_2(0, K_2)] \geq \mathbb{E}[W_2(0, 0)]$ , one also needs to rule out the possibility that  $\bar{\lambda}_2(0, K_2)$  is “too big” compared to  $\bar{\lambda}_2(0, 0)$ . Specifically, we would have to show that

$$\frac{\bar{\lambda}_2(0, K_2)}{\bar{\lambda}_2(0, 0)} \leq \frac{\mathbb{E}[Q_2^{(0, K_2)}]}{\mathbb{E}[Q_2^{(0, 0)}]}.$$

Furthermore, at least one of the inequalities (for the queue or for the arrival rate) has to be strict in order to prove our result. Hence, comparison of queues through coupling will not suffice.

Instead, we work directly with expressions for the steady-state distribution of  $X^K(\cdot)$ . This is simplified by a specific characteristic of thresholds of the form  $K = (0, K_2)$ . For such a threshold, the evolution of  $X_2^K(t)$  follows a one-dimensional Birth-and-Death (BD) process so that the marginal distribution of  $X_2^{(0, K_2)}$  can be computed in closed form. Specifically, with  $K = (0, K_2)$ , the state transition diagram for  $X_2^K(\cdot)$  is as depicted in Figure 7. We use this BD process to compare  $\mathbb{E}[W_2(0, K_2)] = \mathbb{E}[Q_2^{(0, K_2)}] / \bar{\lambda}_2(0, K_2)$  with

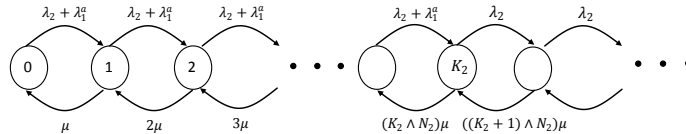


Figure 7: The Birth and Death (BD) process for ED 2 under  $(0, K_2)$

$\mathbb{E}[W_2(0, 0)] = \mathbb{E}[Q_2^{(0, 0)}] / \bar{\lambda}_2(0, 0)$ . For the rest of the argument we fix  $K_2 > 0$ . To simplify notation we let  $\pi_2^0 = \pi_2^{(0, 0)}$  and  $\pi_2^K = \pi_2^{(0, K_2)}$ .

For the argument we fix  $K_2 < N_2$  but it will be easy to see that the argument works for arbitrary  $K_2 > 0$ .

We let  $\check{\lambda} := \lambda_2 + \lambda_1^a$  and introduce the following notation:

$$d(l) := \left( \frac{\lambda_2}{N_2 \mu} \right)^l, \quad l \geq 1, \quad \bar{d}_{K_2} := \prod_{j=K_2+1}^{N_2} \left( \frac{\lambda_2}{j\mu} \right), \quad \bar{d}_{N_2} := \frac{\lambda/(N_2\mu)}{1 - \lambda/(N_2\mu)}.$$

For all  $l \leq N_2$  we let

$$f(l) := \prod_{j=1}^l \frac{1}{j\mu},$$

and set  $f(0) = 1$ . Then, by considering the balance equations for the BD in Figure 7 we have that

$$(\lambda_2 + \lambda_1^a \mathbb{1}\{l < K_2\}) \pi_2^K(l) = \mu((l+1) \wedge N_2),$$

so that

$$\pi_2^K(l) = \begin{cases} \pi_2^K(0) \check{\lambda}^{K_2} f(K_2) \bar{d}_{K_2} d(l - N_2), & l \geq N_2, \\ \pi_2^K(0) \check{\lambda}^l f(l), & l \leq K_2, \\ \pi_2^K(0) \check{\lambda}^{K_2} \lambda_2^{l-K_2} f(l), & \text{otherwise} \end{cases} \quad (\text{A1})$$

and (using  $K_2 = 0$ ) that

$$\pi_2^0(l) = \begin{cases} \pi_2^0(0) \lambda_2^{N_2} f(N_2) d(l - N_2), & l \geq N_2, \\ \pi_2^0(0) \lambda_2^l f(l), & \text{otherwise.} \end{cases} \quad (\text{A2})$$

Then,

$$\mathbb{E}[W(0, 0)] = \frac{\sum_{l \geq N_2} (l - N_2) \pi_2^0(l)}{\lambda_2 \sum_{l=0}^{\infty} \pi_2^0(l)}, \quad \text{and} \quad \mathbb{E}[W(0, K_2)] = \frac{\sum_{l \geq N_2} (l - N_2) \pi_2^K(l)}{\sum_{l=0}^{K_2-1} \check{\lambda} \pi_2^K(l) + \sum_{l=K_2}^{\infty} \lambda_2 \pi_2^K(l)},$$

so that  $\mathbb{E}[W(0, K_2)]/\mathbb{E}[W(0, 0)] = A \cdot B$ , where

$$A := \frac{\sum_{l \geq N_2} (l - N_2) \pi_2^K(l)}{\sum_{l \geq N_2} (l - N_2) \pi_2^0(l)}, \quad \text{and} \quad B := \frac{\lambda_2 \sum_{l=0}^{\infty} \pi_2^0(l)}{\sum_{l=0}^{K_2-1} \check{\lambda} \pi_2^K(l) + \sum_{l=K_2}^{\infty} \lambda_2 \pi_2^K(l)}.$$

The required result will be proved if we show that  $A \cdot B \geq 1$ . To that end, note that

$$A = \frac{\pi_2^K(0) \prod_{j=1}^{K_2} \left( \frac{\check{\lambda}}{j\mu} \right)}{\pi_2^0(0) \prod_{j=1}^{K_2} \left( \frac{\lambda_2}{j\mu} \right)} = \frac{\pi_2^K(0)}{\pi_2^0(0)} \left( \frac{\check{\lambda}}{\lambda_2} \right)^{K_2},$$

and

$$B = \frac{\pi_2^0(0)}{\pi_2^K(0)} \frac{\lambda_2 \sum_{l=0}^{\infty} \frac{\pi_2^0(l)}{\pi_2^0(0)}}{\sum_{l=0}^{K_2-1} \check{\lambda} \frac{\pi_2^K(l)}{\pi_2^K(0)} + \sum_{l=K_2}^{\infty} \lambda_2 \frac{\pi_2^K(l)}{\pi_2^K(0)}},$$

so that

$$\frac{\mathbb{E}[W(0, K_2)]}{\mathbb{E}[W(0, 0)]} = A \cdot B = \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} \cdot \frac{\lambda_2 \sum_{l=0}^{\infty} \frac{\pi_2^0(l)}{\pi_2^0(0)}}{\sum_{l=0}^{K_2-1} \check{\lambda} \frac{\pi_2^K(l)}{\pi_2^K(0)} + \sum_{l=K_2}^{\infty} \lambda_2 \frac{\pi_2^K(l)}{\pi_2^K(0)}}.$$

Letting

$$C := \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} \quad \text{and} \quad D := \frac{\lambda_2 \sum_{l=0}^{\infty} \frac{\pi_2^0(l)}{\pi_2^0(0)}}{\sum_{l=0}^{K_2-1} \check{\lambda} \frac{\pi_2^K(l)}{\pi_2^K(0)} + \sum_{l=K_2}^{\infty} \lambda_2 \frac{\pi_2^K(l)}{\pi_2^K(0)}},$$

we have that  $A \cdot B = C \cdot D$ . To treat  $D$  let

$$M_{K_2+1} := \sum_{l=K_2+1}^{N_2} \prod_{j=K_2+1}^l \left(\frac{\lambda_2}{j\mu}\right),$$

and note that, by (A1) and (A2), we have that

$$\sum_{l=K_2+1}^{\infty} \frac{\pi_2^K(l)}{\pi_2^K(0)} = \check{\lambda}^{K_2} f(K_2) M_{K_2+1}, \quad \sum_{l=K_2+1}^{\infty} \frac{\pi_2^0(l)}{\pi_2^0(0)} = \lambda_2^{K_2} f(K_2) M_{K_2+1},$$

and

$$\sum_{l=N_2}^{\infty} \frac{\pi_2^K(l)}{\pi_2^K(0)} = \check{\lambda}^{K_2} f(K_2) \bar{d}_{K_2} \bar{d}_{N_2}, \quad \sum_{l=N_2}^{\infty} \frac{\pi_2^0(l)}{\pi_2^0(0)} = \lambda_2^{K_2} f(K_2) \bar{d}_{K_2} \bar{d}_{N_2}.$$

Thus, letting  $E := M_{K_2+1} + \bar{d}_{K_2} \bar{d}_{N_2}$  and decomposing the sums in the numerator and denominator of  $D$  and using (A1) and (A2) we have, after some simplifications, that

$$D = \frac{\lambda_2 + \lambda_2 \sum_{l=1}^{K_2} \lambda_2^l f(l) + \lambda_2 \lambda_2^{K_2} f(K_2) E}{\check{\lambda} + \check{\lambda} \sum_{l=1}^{K_2-1} \check{\lambda}^l f(l) + \lambda_2 \check{\lambda}^{K_2} f(K_2) + \lambda_2 \check{\lambda}^{K_2} f(K_2) E}.$$

In turn, we have that

$$C \cdot D = \frac{\lambda_2 \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} + \lambda_2 \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} \sum_{l=1}^{K_2} \lambda_2^l f(l) + \lambda_2 \check{\lambda}^{K_2} f(K_2) E}{\check{\lambda} + \check{\lambda} \sum_{l=1}^{K_2-1} \check{\lambda}^l f(l) + \lambda_2 \check{\lambda}^{K_2} f(K_2) + \lambda_2 \check{\lambda}^{K_2} f(K_2) E}.$$

Letting  $F := \lambda_2 \check{\lambda}^{K_2} f(K_2) E$  and decomposing the second sum in the numerator we have

$$C \cdot D = \frac{\lambda_2 \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} + \lambda_2 \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} \sum_{l=1}^{K_2-1} \lambda_2^l f(l) + \lambda_2 \left(\frac{\check{\lambda}}{\lambda_2}\right)^{K_2} \lambda_2^{K_2} f(K_2) + F}{\check{\lambda} + \check{\lambda} \sum_{l=1}^{K_2-1} \check{\lambda}^l f(l) + \lambda_2 \check{\lambda}^{K_2} f(K_2) + F} \geq 1,$$

where the last inequality follows by noting that the  $i^{th}$  summand in the numerator is greater or equal than the  $i^{th}$  summand in the denominator whenever  $K_2 \geq 1$ . Actually, the inequality is strict whenever  $\check{\lambda} > \lambda_2$  i.e. when  $\lambda_1^a > 0$ . ■

A small modification of the above proof also establishes the following result which we use in the proof of Theorem 3.

**Corollary A.1** Fix  $K_1, K_2 > 0$ . Then, we have that  $\mathbb{E}[W_2(\infty, K_2)] > \mathbb{E}[W_2(\infty, 0)]$  and  $\mathbb{E}[W_1(K_1, \infty)] > \mathbb{E}[W_1(0, \infty)]$ . A threshold pair  $(K_1, K_2)$  where  $K_i = \infty$  for some  $i = 1, 2$  is **not** an equilibrium for the diversion game.

Note that, with  $K = (\infty, K_2)$ , the evolution of  $X_2^K(\cdot)$  is that of a stable Birth and Death process so that, via Little's law, the expectation  $\mathbb{E}[W_2(\infty, K_2)]$  is well defined. A similar argument applies to  $\mathbb{E}[W_1(K_1, \infty)]$ . The corollary follows from the proof of Theorem 2 since if  $K_1 = \infty$  and  $K_2 > 0$ , the (state-dependent) arrival rate to ED 2 is  $\lambda_2 = \lambda_2^a + \lambda_2^w$  whenever ED 2 is in state  $i < K_2$  and it is  $\lambda_2^w < \lambda_2$  whenever  $i \geq K_2$ . Hence, replacing, in the proof of Theorem 2,  $\check{\lambda}$  with  $\lambda_2$  and  $\lambda_2$  with  $\lambda_2^w$  establishes the desired result.

## A.2 Proofs for §5

This section has two subsections. In §A.2.1 we introduce a construction of the sample paths of the threshold system. The construction is then used in the proofs of Theorems 4 and 5 that appear in §A.2.2. The proofs of auxiliary results (Lemma A.1, Propositions A.1-A.3 and Corollary A.2 appear in the technical report [\[reference to authors' website would be given here\]](#).

### A.2.1 A sample-path construction

Here we introduce a sample path construction for the threshold system and the perfectly pooled system using dynamic thinning of Poisson process. This construction is instrumental in our proofs of Theorems 4, 5 and 6 in the manuscript.

To generate arrivals, for  $i = 1, 2$ , let  $A_i^w(\cdot)$  be a Poisson process with rate  $\lambda_i^w$  and let  $A_i^a(\cdot)$  be Poisson processes with rate  $\lambda_i^a$ . Then,  $A(t) = \sum_{i=1}^2 (A_i^a(t) + A_i^w(t))$  is the total number of arrivals (ambulances and walk-ins) by time  $t$  to both EDs. We denote by  $\tau_{i,a}^n$  the time of the  $n^{th}$  jump of the process  $A_i^a(\cdot)$ , i.e.,  $\tau_{i,a}^n := \inf\{t \geq 0 : A_i^a(t) = n\}$ .

To generate service completions, we introduce a Poisson process  $S(\cdot)$  with rate  $\mu N_\Sigma$ . We introduce two infinite sequences  $\{U_n, n \in \mathbb{Z}_+\}$  and  $\{Y_n, n \in \mathbb{Z}_+\}$  of i.i.d uniform  $[0, 1]$  random variables. The  $n^{th}$  jump of  $S(\cdot)$  is an actual service completion in the network if  $U_n \in [0, Z_\Sigma^K(\tau_s^n -)/N_\Sigma]$ , where  $\tau_s^n$  is the time of that  $n^{th}$  jump of the process  $S(\cdot)$ , i.e.,  $\tau_s^n := \inf\{t \geq 0 : S(t) = n\}$ . The sequence  $\{Y_n, n \in \mathbb{Z}_+\}$  is then used to determine which ED is the one in which the service completion happens. Specifically, if  $S(\cdot)$  jumps at time  $t$  and  $U_n \in [0, Z_\Sigma^K(t-)/N_\Sigma]$  then the resulting service completion is at ED 1 if, in addition,  $Y_n \in [0, Z_1^K(t-)/Z_\Sigma^K(t-)]$ . It is a service completion in ED 2 otherwise. The sequences  $\{Y_n\}_{n \geq 1}, \{U_n\}_{n \geq 1}$  are constructed as independent of each other and of the Poisson processes  $(A_i^w(\cdot), A_i^a(\cdot); i = 1, 2)$ , and  $S(\cdot)$ . All Poisson processes introduced thus far are assumed to have right continuous sample paths.

We call  $(A_1^w(\cdot), A_1^a(\cdot), A_2^w(\cdot), A_2^a(\cdot), S(\cdot), \{U_n\}_{n \geq 1}, \{Y_n\}_{n \geq 1})$  the *primitives*. Given an initial condition  $(Q_1(0), Q_2(0), Z_1(0), Z_2(0))$  and a realization of the primitives we can explicitly construct the patient-flow

dynamics. Specifically, for a threshold pair  $K = (K_1, K_2)$  we write

$$\begin{aligned}
X_1^K(t) &= X_1(0) + A_1^w(t) \\
&+ \sum_{n=1}^{A_1^a(t)} \left( \mathbb{1}\{X_1^K(\tau_{1,a}^n-) < K_1\} + \mathbb{1}\{X_1^K(\tau_{1,a}^n-) \geq K_1, X_2^K(\tau_{1,a}^n-) \geq K_2\} \right) \\
&+ \sum_{n=1}^{A_2^a(t)} \mathbb{1}\{X_2^K(\tau_{2,a}^n-) \geq K_2, X_1^K(\tau_{2,a}^n-) < K_1\} \\
&- \sum_{n=1}^{S(t)} \mathbb{1}\{U_n \in [0, Z_\Sigma^K(\tau_s^n-)/N_\Sigma]\} \mathbb{1}\{Y_n \in [0, Z_1^K(\tau_s^n-)/Z_\Sigma^K(\tau_s^n-)]\}. \tag{A3}
\end{aligned}$$

Here, the first line corresponds to walk-in arrivals. The second line corresponds to ambulance arrivals from catchment area 1 that are not diverted to ED 2 and the third line to arrivals of ambulances from catchment area 2 that are diverted to ED 1. The last line corresponds to service completions. A similar equation is constructed for  $X_2^K(t)$ . This construction generates the correct probability law for  $X^K(\cdot)$ .

Since each ED is work conserving (no patients are in queue while there are available beds) we have that  $Z_i^K(t) = X_i^K(t) \wedge N_i$ . In turn, the equations (A3) for  $i = 1, 2$  completely define the dynamics of the threshold system. Summing (A3) over  $i = 1, 2$  one gets

$$X_\Sigma^K(t) = X_\Sigma^K(0) + A(t) - \sum_{n=1}^{S(t)} \mathbb{1}\{U_n \in [0, Z_\Sigma^K(\tau_s^n-)/N_\Sigma]\} \tag{A4}$$

We next construct the sample paths of the total number of patients in the perfectly pooled system  $X_\Sigma^P(t) = X_1^P(t) + X_2^P(t)$  using the same primitives. Note that, in the pooled system, there is no idleness while there is queue so that  $Z_\Sigma^P(t) = X_\Sigma^P(t) \wedge N_\Sigma$  for all  $t \geq 0$ . Then, given an initial state  $(X_1^P(0), X_2^P(0))$  such that  $Z_\Sigma^P(0) = X_\Sigma^P(0) \wedge N_\Sigma$ , we write

$$X_\Sigma^P(t) = X_\Sigma^P(0) + A(t) - \sum_{n=1}^{S(t)} \mathbb{1}\{U_n \in [0, (X_\Sigma^P(\tau_s^n-) \wedge N_\Sigma)/N_\Sigma]\}. \tag{A5}$$

As expected, these dynamics are equal in law to the dynamics of an  $M/M/N$  queue with arrival rate  $\lambda_\Sigma$ , service rate  $\mu$  and  $N_\Sigma$  servers.

In our proofs we will sometimes initialize the threshold system and the pooled system with the stationary distribution of the former. In those cases, the pooled system might be initialized in a state in which the condition  $Z_\Sigma^P(0) = X_\Sigma^P(0) \wedge N_\Sigma$  is violated. We will then assume that there is instantaneous re-shuffling at time 0 so that, after the reshuffling,  $I_j(0+) = (I_j(0) + Q_j(0)) \wedge N_j$  and  $Q_j(0+) = Q_j(0) - I_j(0) \wedge Q_j(0)$  and so that  $Z_\Sigma^P(t) = X_\Sigma^P(t) \wedge N_\Sigma$  for all  $t > 0$ .

### A.2.2 Proof of Theorems 4 and 5

We prove both theorems simultaneously by showing that

$$\begin{aligned} \mathbb{E}[W^P] &\leq \mathbb{E}[W(K_1^*, K_2^*)] \leq \mathbb{E}[W(N_1, N_2)] \\ &\leq \frac{\mathbb{E}[W(K_1^*, K_2^*)]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)} \leq \frac{\mathbb{E}[W^P]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)}. \end{aligned}$$

Through most of the proof we fix the threshold pair to be  $K = (N_1, N_2)$  and omit the threshold superscript from all relevant quantities.

The proof proceeds in two main steps. The first step, stated in Proposition A.1, bounds the distance between  $Q_\Sigma(\cdot)$  and  $Q_\Sigma^P(\cdot)$  on finite intervals in terms of a process  $C(\cdot)$  that we explicitly define. The second step, Proposition A.2, provides a bound for the steady-state mean of  $C(\cdot)$ .

Recall that in the pooled system  $Z_\Sigma^P(t) = X_\Sigma^P(t) \wedge N_\Sigma$  for all  $t > 0$ . The threshold system does not satisfy this property since there might be times at which there is positive queue in one ED and available beds at the other. A somewhat weaker relation does hold for the  $(N_1, N_2)$ -threshold system as stated in the following lemma.

**Lemma A.1** *Let  $\mathcal{K}_1 := \{x : x_1 \geq N_1, x_2 < N_2\}$  and  $\mathcal{K}_2 := \{x : x_1 < N_1, x_2 \geq N_2\}$ . Then, for all  $t \geq 0$ ,*

$$Z_\Sigma(t) = X_\Sigma(t) \wedge N_\Sigma - C(t), \text{ and } Q_\Sigma(t) = (X_\Sigma(t) - N_\Sigma)^+ + C(t), \quad (\text{A6})$$

where

$$C(t) := (Q_1(t) \wedge I_2(t)) \mathbb{1}\{X(t) \in \mathcal{K}_1\} + (Q_2(t) \wedge I_1(t)) \mathbb{1}\{X(t) \in \mathcal{K}_2\}. \quad (\text{A7})$$

Lemma A.1 shows that the process  $C(t)$  captures the ‘‘inefficiency’’ of the threshold system at time  $t$ . Using (A6) we re-write (A4) as follows

$$X_\Sigma(t) = X_\Sigma(0) + A(t) - \sum_{n=1}^{S(t)} \mathbb{1}\{U_n \in [0, (X_\Sigma(\tau_s^n -) \wedge N_\Sigma - C(\tau_s^n -)) / N_\Sigma]\}. \quad (\text{A8})$$

Equations (A5) and (A8) are starting points for the comparison between the pooled system and the  $(N_1, N_2)$ -threshold system. We state three propositions that correspond to key steps in the proof of Theorem 4.

**Proposition A.1** *Fix  $x \in \mathbb{Z}_+$  and assume  $X_\Sigma(0) = X_\Sigma^P(0) = x$  and that  $Q_\Sigma^P(0) = (x - N_\Sigma)^+$ . Then,*

$$X_\Sigma^P(t) \leq X_\Sigma(t), \text{ and } Q_\Sigma^P(t) \leq Q_\Sigma(t) \text{ for all } t \geq 0, \text{ almost surely, and} \quad (\text{A9})$$

$$\mathbb{E}[X_\Sigma(t) - X_\Sigma^P(t)] \leq \mathbb{E} \left[ \mu \int_0^t C(s) ds \right], \text{ for all } t \geq 0. \quad (\text{A10})$$

An immediate corollary of (A9) is that, taking  $t \rightarrow \infty$ , we have the ordering of the steady-state quantities

as follows:

$$X_{\Sigma}^P \leq_{st} X_{\Sigma}, \quad \text{and} \quad Q_{\Sigma}^P \leq_{st} Q_{\Sigma}. \quad (\text{A11})$$

The next proposition bounds the expectation of the process  $C(t)$  in steady-state.

**Proposition A.2** *Assume that  $\max\{\tilde{\rho}_1, \tilde{\rho}_2\} < 1$  and suppose that  $(X_1(0), X_2(0))$  is distributed according to the unique steady-state distribution  $\pi$  of  $X(\cdot)$ . Then,*

$$\mathbb{E}[C(t)] \leq C(\tilde{\rho}_1, \tilde{\rho}_2) := \frac{1}{1 - \tilde{\rho}_1} + \frac{1}{1 - \tilde{\rho}_2}, \quad \text{for all } t \geq 0. \quad (\text{A12})$$

Consequently, by (A6)

$$\left| \mathbb{E}[Q_{\Sigma}(t)] - \mathbb{E}[(X_{\Sigma}(t) - N_{\Sigma})^+] \right| \leq C(\tilde{\rho}_1, \tilde{\rho}_2), \quad \text{for all } t \geq 0. \quad (\text{A13})$$

Let  $\mathcal{M}_i, i = 1, 2$  be two independent random variables with the steady-state distribution of an  $M/M/N$  queue with service rate  $\mu$ ,  $N_i$  servers and arrival rate  $\rho\mu N_i$  where  $\rho = \lambda_{\Sigma}/(\mu N_{\Sigma})$ .

By basic expressions for the  $M/M/N$  queue  $\mathbb{P}\{\mathcal{M}_i \geq N_i + q | \mathcal{M}_i \geq N_i\} = \rho^q$ ; see e.g. Chapter 3 of Asmussen [2003]. Let  $\bar{\mathcal{M}}_i, i = 1, 2$  be independent random variables with the distribution  $\mathbb{P}\{\bar{\mathcal{M}}_i \geq N_i + q\} = \rho^q$ . In particular,  $\mathcal{M}_i \leq_{st} \bar{\mathcal{M}}_i, i = 1, 2$ . Define the sums  $\mathcal{M}_{\Sigma} = \mathcal{M}_1 + \mathcal{M}_2$  and  $\bar{\mathcal{M}}_{\Sigma} = \bar{\mathcal{M}}_1 + \bar{\mathcal{M}}_2$ . Hence,  $\mathcal{M}_{\Sigma} \leq_{st} \bar{\mathcal{M}}_{\Sigma}$ . Let  $\bar{\nu}_{\Sigma}$  be the distribution of  $\bar{\mathcal{M}}_{\Sigma}$ .

In the next proposition we initialize  $X(\cdot)$  and  $X^P(\cdot)$  at time  $t = 0$  with the steady-state distribution of the former. We augment the notation and add the superscript  $\pi$  to all processes to make explicit the fact that the system is initialized at time 0 with the steady-state distribution  $\pi$ . Recall that throughout this section the threshold pair is fixed to  $K = (N_1, N_2)$ .

**Proposition A.3** *Assume that  $(X_1(0), X_2(0)) = (X_1^P(0), X_2^P(0)) \sim \pi$ . Then, for all  $t \geq 0$ ,  $X_{\Sigma, \pi}^P(t) \leq_{st} X_{\Sigma, \pi}(t) \leq_{st} \mathcal{M}_{\Sigma} \leq_{st} \bar{\mathcal{M}}_{\Sigma}$ .*

Note that the first inequality in the proposition follows Proposition A.1. The proof of second inequality appears at the end of this section.

An immediate implication of Proposition A.3 is that initializing both the  $(N_1, N_2)$ -threshold system and the pooled system with the stationary distribution of the former, the total number of patients in the pooled system is stochastically bounded, at all times, by  $\bar{\mathcal{M}}_{\Sigma}$  whose distribution,  $\bar{\nu}_{\Sigma}$ , has the sole parameter  $\rho = \lambda_{\Sigma}/(\mu N_{\Sigma})$ .

Let  $Q_{\Sigma, \bar{\nu}_{\Sigma}}^P(t)$  be the queue length in the pooled system at time  $t$  if the pooled system is initialized at time 0. Then,  $X_{\Sigma}^P(0) \sim \bar{\nu}_{\Sigma}$  and  $Q_{\Sigma}^P(0) = (X_{\Sigma}^P(0) - N_{\Sigma})^+$ . Let  $Q_{\Sigma}^P$  be a random variable with the distribution of the steady-state queue-length in the pooled system—this is the distribution of the queue length in an  $M/M/N$  queue with parameters  $\lambda_{\Sigma}, \mu$  and  $N_{\Sigma}$ .

**Corollary A.2** *Let*

$$T(\epsilon) := \inf \left\{ t \geq 0 : \mathbb{E}[Q_{\Sigma, \bar{\nu}_{\Sigma}}^P(s)] \leq \mathbb{E}[Q_{\Sigma}^P](1 + \epsilon), \quad \forall s \geq t \right\} \vee \frac{1}{\mu}. \quad (\text{A14})$$

Then, for all  $t \geq T(\epsilon)$ .

$$\mathbb{E}[Q_{\Sigma, \pi}^P(t)] \leq \mathbb{E}[Q_{\Sigma, \bar{\nu}_{\Sigma}}^P(t)] \leq \mathbb{E}[Q_{\Sigma}^P](1 + \epsilon). \quad (\text{A15})$$

**Remark A.1** Note that  $T(\epsilon)$  depends on the rate of convergence of the (time-dependent) expected queue-length in an  $M/M/N$  queue to its steady-state value. Convergence to steady-state is often stated in terms of total-variation bounds that require, in a sense, that the whole distribution converges; see e.g. §5 of Meyn and Tweedie [1993]. Several papers consider convergence to steady-state of the stable  $M/M/N$  queue. Most recent among these is Gamarnik and Goldberg [2010] that studies this question for the  $M/M/N$  queue in the Halfin-Whitt regime. Gamarnik and Goldberg [2010] and some of the references therein are concerned with identifying the exact rate of convergence to steady-state in a total variation sense. Our requirements are different and concerned with the convergence rate of the expectations. In this regard, a Lyapunov criteria for exponential ergodicity of  $M/M/N$  queue length can be established that would imply, in turn, an exponential rate of convergence of the (time dependent) queue length mean to the steady-state mean; see §6 and 7 of Meyn and Tweedie [1993]. Studying the dependence of this convergence rate on the system parameters is beyond the scope of our paper.

We do remark that for a sequence of  $M/M/N$  queues in the Halfin-Whitt many-server heavy-traffic regime, as studied in Gamarnik and Goldberg [2010], it can be shown that – with the initial distribution  $\bar{\nu}_{\Sigma}$  carefully chosen as above –  $T(\epsilon)$  is bounded by a constant that does not grow with the system size. We hence expect the term  $T(\epsilon)/\lambda_{\Sigma}$ , that appears on the right hand side of our bounds in Theorem 4, to be small for large systems. ■

**Completing the proof of Theorems 4 and 5:** Initialize the  $(N_1, N_2)$ -threshold system with its steady-state distribution  $\pi$ . We initialize the pooled system with  $X_{\Sigma}^P(0) = X_{\Sigma}(0)$  and so that  $Q_{\Sigma}^P(0) = (X_{\Sigma}^P(0) - N_{\Sigma})^+$ . Beyond this, how the patients are distributed between the EDs is immaterial to our proof.

Using Propositions A.1 and A.2 we have  $\mathbb{E}[X_{\Sigma}(T(\epsilon)) - X_{\Sigma}^P(T(\epsilon))] \leq \mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2)$ , and, in particular,  $\mathbb{E}[(X_{\Sigma}(T(\epsilon)) - N_{\Sigma})^+ - (X_{\Sigma}^P(T(\epsilon)) - N_{\Sigma})^+] \leq \mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2)$ . Using the fact that  $Q_{\Sigma}^P(t) = (X_{\Sigma}^P(t) - N_{\Sigma})^+$  for all  $t > 0$  together with (A13) we have that:

$$\begin{aligned} \mathbb{E}[Q_{\Sigma}(T(\epsilon)) - Q_{\Sigma}^P(T(\epsilon))] &= \mathbb{E}[(X_{\Sigma}(T(\epsilon)) - N_{\Sigma})^+ - Q_{\Sigma}^P(T(\epsilon))] + \mathbb{E}[Q_{\Sigma}(T(\epsilon)) - (X_{\Sigma}(T(\epsilon)) - N_{\Sigma})^+] \\ &\leq \mu C(\tilde{\rho}_1, \tilde{\rho}_2) T(\epsilon) + C(\tilde{\rho}_1, \tilde{\rho}_2). \end{aligned}$$

By definition  $T(\epsilon) \geq 1/\mu$  so that

$$\mathbb{E}[Q_{\Sigma}(T(\epsilon)) - Q_{\Sigma}^P(T(\epsilon))] \leq 2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2). \quad (\text{A16})$$

Combining (A15) with (A16) we get

$$\mathbb{E}[Q_{\Sigma}] - \mathbb{E}[Q_{\Sigma}^P] \leq 2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \epsilon \mathbb{E}[Q_{\Sigma}^P] \leq 2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \epsilon \mathbb{E}[Q_{\Sigma}], \quad (\text{A17})$$

where the last inequality follows from Proposition A.1. By Little's law

$$\mathbb{E}[W(N_1, N_2)] = \frac{\mathbb{E}[Q_\Sigma]}{\lambda_\Sigma} \leq \frac{\mathbb{E}[Q_\Sigma^P]}{\lambda_\Sigma} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma} + \epsilon \mathbb{E}[W(N_1, N_2)].$$

Hence,

$$\mathbb{E}[W(N_1, N_2)] \leq \frac{\mathbb{E}[W^P]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)}. \quad (\text{A18})$$

By definition,  $\mathbb{E}[W(K_1^*, K_2^*)] \leq \mathbb{E}[W(N_1, N_2)]$ . Also, for any threshold pair (including a socially optimal one), the total expected steady-state queue length is greater than that in the perfectly pooled system. This simple observation is proved identically to the first part of Proposition A.1. Indeed, we note that the proof of (A9) there does not hinge on the specific value of the thresholds being  $(N_1, N_2)$  rather than arbitrary  $(K_1, K_2)$ . In turn we have that  $\mathbb{E}[W^P] \leq \mathbb{E}[W(K_1^*, K_2^*)] \leq \mathbb{E}[W(N_1, N_2)]$ . Plugging this into (A18) we finally have that

$$\begin{aligned} \mathbb{E}[W^P] &\leq \mathbb{E}[W(K_1^*, K_2^*)] \leq \mathbb{E}[W(N_1, N_2)] \\ &\leq \frac{\mathbb{E}[W(K_1^*, K_2^*)]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)} \leq \frac{\mathbb{E}[W^P]}{1 - \epsilon} + \frac{2\mu T(\epsilon)C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_\Sigma(1 - \epsilon)}. \end{aligned}$$

This concludes the proof of Theorems 4 and 5. ■

## References

- I. Adan, G. van Houtum, and J. van der Wal. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research*, 48(2):197–217, 1994.
- O. Akşin, F. Karaesmen, and E. Ormeci. A review of workforce cross-training in call centers from an operations management perspective. *CRC Press*, pages 211–240, 2007.
- G. Allon and A. Federgruen. Competition in service industries. *Operation Research*, 55(1):37–55, 2007.
- G. Allon, S. Deo, and W. Lin. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Working paper, Kellogg School of Management*, 2009.
- O. Asamoah, S. Weiss, A. Ernst, M. Richards, and D. Sklar. A novel diversion protocol dramatically reduces diversion hours. *The American Journal of Emergency Medicine*, 26(6):670, 2008.
- S. Asmussen. *Applied probability and queues*. Springer Verlag, 2003.
- R. Atar. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 15(4):2606–2650, 2005.
- A. Bagust et al. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *British Medical Journal*, 319(7203):155, 1999.
- D. Baron and E. Kalai. The simplest equilibrium of a majority-rule division game. *Journal of Economic Theory*, 61(2):290–301, 1993.
- S. Benjaafar. Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research*, 87(2):375–388, 1995.
- S. Bernstein, V. Verghese, W. Leung, A. Lunney, and I. Perez. Development and validation of a new index to measure emergency department crowding. *Emergency Medicine*, 10:938–942, 2003.

- G. Cachon and P. Harker. Competition and outsourcing with scale economies. *Management Science*, 48(10):1314–1333, 2002.
- G. Cachon and F. Zhang. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science*, 53(3):408, 2007.
- Y. Chen, C. Maglaras, and G. Vulcano. Design of an aggregated marketplace under congestion effects: Asymptotic analysis and equilibrium characterization. 2008.
- C. Chidester, W. Koenig, and R. Tadeo. Personal communication with LA County EMS Agency, 2009.
- J. Cochran and K. Roche. A multi-class queueing network analysis methodology for improving hospital emergency department performance. *Computers and Operations Research*, 36(5):1497–1512, 2009.
- P. Enders, A. Scheller-Wolf, S. H. Cho, and M. Shunko. Hospital capacity effects on inbound ambulance traffic. Working paper, Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, 2010.
- S. Epstein and L. Tian. Development of an emergency department work score to predict ambulance diversion. *Academic Emergency Medicine*, 13(4):421–426, 2006.
- T. Falvo, L. Grove, R. Stachura, and W. Zirkin. The financial impact of ambulance diversions and patient elopements. *Academic Emergency Medicine*, 14(1):58–62, 2007.
- D. Fatovich, Y. Nagree, and P. Sprivulis. Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. *British Medical Journal*, 22(5):351, 2005.
- J. B. Franaszek, B. R. Asplin, and B. Brunner. Responding to emergency department crowding: A guidebook for chapters. Technical report, American College of Emergency Physicians, 2002.
- D. Gamarnik and D. Goldberg. On the Rate of Convergence to Stationarity of the M/M/N Queue in the Halfin-Whitt Regime. *Working paper, MIT*, 2010.
- N. Gans, N. Liu, A. Mandelbaum, H. Shen, and H. Ye. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Festschrift for Lawrence D. Brown*, volume 6 of *IMS Collections*, pages 99–123. 2010.
- General Accounting Office. Hospital emergency departments: Crowded conditions vary among hospitals and communities, 2003.
- L. V. Green. How many hospital beds? *Inquiry*, 39:400–412, 2002.
- I. Gurvich and W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management*, 11(2):237–253, 2009.
- R. Hagtvedt, M. Ferguson, P. Griffin, G. Jones, and P. Keskinocak. Cooperative strategies to reduce ambulance diversion. In *Winter Simulation Conference (WSC), Proceedings of the 2009*, pages 1861–1874, 2009.
- N. Hoot and D. Aronsky. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136, 2008.
- N. Hoot, C. Zhou, I. Jones, and D. Aronsky. Measuring and Forecasting Emergency Department Crowding in Real Time. *Annals of Emergency Medicine*, 49(6):747–755, 2007.
- E. Kalai, M. Kamien, and M. Rubinovitch. Optimal service speeds in a competitive environment. *Management Science*, 38(8):1154–1163, 1992.
- L. Kowalczyk. States ER policy passes checkup. *The Boston Globe*, December 14 2009.
- D. Levhari and I. Luski. Duopoly pricing and waiting lines. *European Economic Review*, 11(1):17–35, 1978.
- K. Lin. Decentralized admission control of a queueing system: A game-theoretic model. *Naval Research Logistics*, 50(7):702–718, 2003.

- S. Maman. Uncertainty in the demand for service: The case of call centers and emergency departments. Msc. Thesis, Technion-Israel Institute of Technology, 2009.
- MARCER. Community plan for ambulance diversion for the greater Kansas City Metropolitan Area, 2007.
- M. L. McCarthy, D. Aronsky, I. D. Jones, J. R. Miner, R. A. Band, J. M. Baren, J. S. Desmond, K. M. Baumlin, R. Ding, and R. Shesser. The emergency department occupancy rate: A simple measure of emergency department crowding? *Annals of Emergency Medicine*, 51(1):15–24, 2008.
- S. Meyn and R. Tweedie. Stability of markovian processes III: Foster-lyapunov criteria for continuous-time processes. *Ann. Appl. Prob.*, 25(3):518–548, 1993.
- N. Mihal and R. Moilanen. When Emergency Rooms Close: Ambulance Diversion in the West San Fernando Valley. Technical report, UC Los Angeles: The Ralph and Goldy Lewis Center for Regional Policy Studies. Retrieved from: <http://www.escholarship.org/uc/item/4g80h84v>, 2005.
- J. Moskop, D. Sklar, J. Geiderman, R. Schears, and K. Bookman. Emergency Department Crowding, Part 1—Concept, Causes, and Moral Consequences. *Annals of Emergency Medicine*, 53(5):605–611, 2009.
- P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.
- NYBEMS. Emergency patient destinations and hospital diversion, 2006.
- PEHSC. Joint position statement: Guidelines for hospital ambulance-diversion policies, 2004.
- J. Pham, J. Story, R. Hicks, A. Shore, L. Morlock, D. Cheung, G. Kelen, and P. Pronovost. National study on the frequency, types, causes, and consequences of voluntarily reported emergency department medication errors. *Journal of Emergency Medicine*, 2008.
- A. Ramirez, J. Fowler, and T. Wu. Analysis of ambulance diversion policies for a large-size hospital. In *Winter Simulation Conference (WSC), Proceedings of the 2009*, pages 1875–1886, 2009.
- T. Reeder and H. Garrison. When the Safety Net Is Unsafe: Real-time Assessment of the Overcrowded Emergency Department. *Academic Emergency Medicine*, 8(11):1070–1074, 2001.
- T. Schelling. The strategy of conflict. Prospectus for a reorientation of game theory. *Journal of Conflict Resolution*, 2(3):203, 1958.
- M. J. Schull, P. M. Slaughter, and D. A. Redelmeier. Urban emergency department overcrowding: defining the problem and eliminating misconceptions. *Canadian Journal of Emergency Medicine*, 4(2):76–83, 2002.
- M. J. Schull, L. J. Morrison, M. Vermeulen, and D. A. Redelmeier. Emergency department gridlock and out-of-hospital delays for cardiac patients. *Academic Emergency Medicine*, 10(7):709–716, 2003.
- M. J. Schull, M. Vermuelen, G. Slaughter, L. Morrison, and P. Daly. Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Annals of Emergency Medicine*, 44(6):577–585, 2004.
- A. Stolyar. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences*, 19(02):141–189, 2005.
- B. C. Sun, S. A. Mohanty, R. Weiss, R. Tadeo, M. Hasbrouck, W. Koenig, C. Meyer, and S. Asch. 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, 2006.
- T. Tezcan. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Mathematics of Operations Research*, 33(1):51–90, 2008.
- G. Vassilopoulos. A simulation model for bed allocation to hospital inpatient departments. *Simulation*, 45(5):233–241, 1985.

- G. M. Vilkes, E. M. Castillo, M. A. Metz, L. U. Ray, P. A. Murrin, R. Lev, and T. C. Chan. Community trial to decrease ambulance diversion hours: The San Diego county patient destination trial. *Annals of Emergency Medicine*, 44(4): 295–303, 2004.
- X. Vives. *Oligopoly pricing: Old ideas and new tools*. The MIT Press, 2001.
- S. Weiss, R. Derlet, J. Arndahl, A. Ernst, J. Richards, M. Fernandez-Frankelton, R. Schwab, T. Stair, P. Vicellio, D. Levy, et al. Estimating the degree of emergency department overcrowding in academic medical centers: results of the National ED Overcrowding Study (NEDOCS). *Academic Emergency Medicine*, 11(1):38–50, 2004.
- W. Whitt. Deciding which queue to join: Some counterexamples. *Operations research*, 34(1):55–62, 1986.
- W. Whitt. A review of  $L = \lambda W$  and extensions. *Queueing Systems*, 9:235–268, 1991.
- G. Yom-Tov. Queues in hospitals: Queueing networks with reentering customers in the qed regime. PhD Thesis, Technion-Israel Institute of Technology, 2009.