Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

COOPERATIVE STRATEGIES TO REDUCE AMBULANCE DIVERSION

Reidar Hagtvedt

Department of Finance and Management Science
University of Alberta School of Business
Edmonton, AB T6G 2R6, Canada

Mark Ferguson

The Georgia Tech College of Business
800 West Peachtree St, NW
Atlanta, GA 30308-0520, U.S.A.

Paul Griffin

H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive
Atlanta, GA 30332-0205, U.S.A.

Gregory Todd Jones

Georgia State University College of Law &
The Center for Negotiation and Conflict Resolution
Georgia State University
140 Decatur Street,
Atlanta, GA 30308, U.S.A.

Pinar Keskinocak

H. Milton Stewart School of Industrial and Systems Engineering
Center for Humanitarian Logistics
Georgia Institute of Technology
Atlanta, GA 30332-0205, U.S.A.

ABSTRACT

Overcrowding in the emergency departments (ED) has led to an increase in the use of ambulance diversion (AD), during which a hospital formally is not accepting patients by ambulance. We use a number of tools to considers methods by which hospitals in a metro area may cooperate to reduce diversion, including contracts and pressure from outside regulators. The tools include a birth-death process, discrete event simulations, agent-based simulation model, and some game theory to examine the potential for cooperative strategies. We use data to suggest a functional form for the payoff of such games. We find that a centralized form of routing is needed, as voluntary cooperation does not appear to be robust in the presence of noise or strategic behavior, and ethical considerations also have a significant impact.

1 INTRODUCTION

From 1992 to 2003, the number of hospital emergency rooms in the U.S. went from approximately 6000 to less than 4000, while emergency department visits increased from 89.8 to 108 million (Schafermeyer and Asplin 2003). Many hospital patients arrive through the emergency room, and when the hospital is at capacity, many patients are *boarded* in the ED until a regular bed opens up. This exacerbates crowding of the ED. To combat this problem, the hospital may go on *diversion* (or Ambulance Diversion, AD), where ambulances are sent to other facilities.

At the level of a hospital system for a given metropolitan statistical area (MSA), the emergency services must coordinate dispatching, especially when capacity is tight. This leads to the question of cooperation between the hospitals. Systems may vary from each hospital operating independently to a completely centralized system. Advances in technology enable both tighter integration on the one hand, and better information flow to aid independent hospital decision making, on the other.

We seek cooperative strategies to limit AD, between these extreme positions. The first goal in this study is therefore to identify a method to reduce AD in a cooperative manner, and to evaluate its impact from the perspective of hospitals, public, and government.

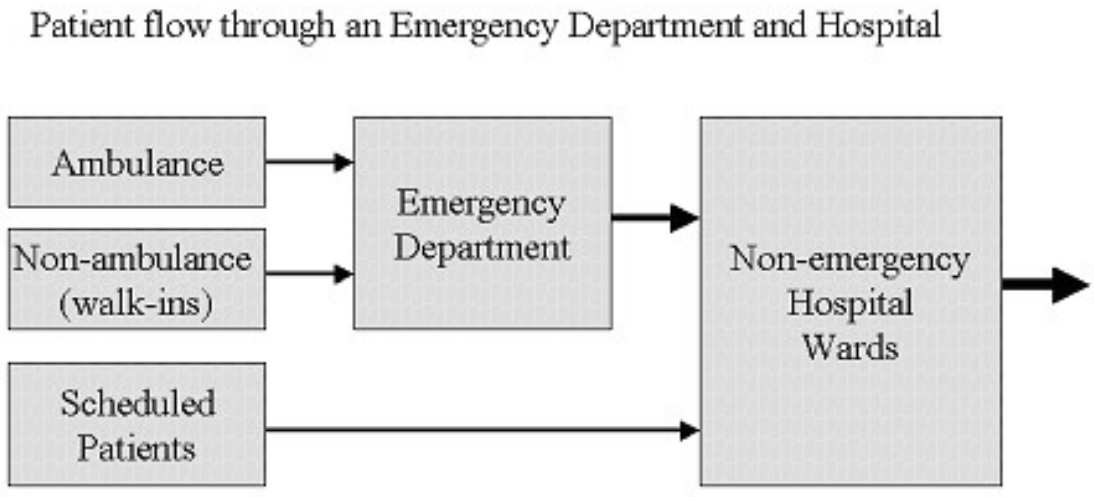

Figure 1: Patients moving through the ED and main hospital ward

The idea behind this research came from a description of how AD worked at DeKalb Medical Center (DKMC) in Atlanta, GA (Hardy and Te 2005). When the hospital reaches capacity, the hospital goes on AD, and ambulance dispatchers are informed. Going on AD is not done lightly, and DKMC remains on diversion until some slack capacity is generated, in order that a minor surge of new patients will not overwhelm the ED again. The flow of patients is illustrated in Figure 1.

The spur to extend this avenue to systems of hospitals came from descriptions of waves of diversions in the Chicago area.

To work up from patients to systems of hospitals requires several steps. We begin by considering a standard birth-death model for patient arrivals and departures, and then a discrete event simulation when tweaking the closed-form solution becomes intractable. We use these tools to examine an approach to lower both the rate and variability of new patient arrivals to the ED when capacity is tight, thereby lessening the probability that a hospital will have to go on diversion. The benefits include less time spent on diversion for hospitals, less crowding of ED's, enhanced government ability to respond to disasters, and better health service for the public.

There are three types of patients in the U.S.: uninsured, privately insured, and federally insured (Medicare and Medicaid; M/M). We examine the impact from allowing partial diversion (PD) of M/M patients to low occupancy hospitals, to see if this would increase overall patient flow, for all classes of patients.

Limiting probabilities for a Markov process are derived in order to show the potential effects, and data from DeKalb Medical Center in Atlanta is used to parameterize a numerical example, as well as to provide the framework for a simulation model.

Unfortunately, the benefits are fragile, and unlikely to hold in any system with significant noise. The scheme itself raises ethical issues, and as the benefits are ephemeral, this would appear to be a dead-end for combating AD. We therefore turn to investigating cooperative strategies between hospitals, without patients nor government directly involved.

We use an agent-based model (ABM) to provide insight into how a more complex spatial structure, a variable number of differentiated hospitals, differentiated policies, and various types of enforcement may contribute to cooperation in avoiding ambulance diversion. The ABM is constructed to be consistent with the previous portions of the paper, and it is by adding these new features that the fragility of PD is made clear.

Finally, we use a game theory approach to examine the rational incentives to hospitals to pre-emptively go on AD, even before the practical maximum patient load is reached. As before, we retain as much as possible of the preceding work, to aid comparison. We find that with sufficient patient demand, the hospitals are operating in an N-Player Prisoner's Dilemma, which makes cooperation highly unlikely without external pressure. Further, the optimal strategy is to go on AD at a threshold less than capacity, which decreases with the expected overflow from other hospitals, but increases with the square root of a penalty for going on AD.

We round off by commenting on how these (non)results may be applied to the discussion of reforming hospital legislation, and on the need for either legally binding contracts or outside regulation.

2 LITERATURE REVIEW

The literature on hospital and emergency room overcrowding is growing with the perception of incipient crisis (Institute of Medicine 2006), the core of which is a decrease in hospital capacity over the last ten years, as the number of ED visits increase. Schafermeyer and Asplin (2003) give an overview of hospital and ED overcrowding.

The scope of the problem has grown significantly in recent years (Fatovich and Hirsch 2003, Schafermeyer and Asplin 2003), with more than nine out of ten ED directors considered overcrowding a significant problem (Derlet et al. 2001). In 2003, an estimated half million diversions occurred, while the overall volume of ED visits was slightly more than sixteen million (Burt et al. 2006). Although this is primarily a patient-safety issue, lost revenues is also considered a significant problem for many hospitals (Geer and Smith 2004).

2.1 Overcrowding

The causes of overcrowding of EDs may be split into input, throughput, and output problems (Fatovich and Hirsch 2003; Asplin et al. 2003). Since the ED plays such a large number of roles in the U.S. health system, there are a number of suggested reasons for the causes of overcrowding (Derlet and Richards 2000), including more acute and complex cases presenting to the ED, increased patient volumes, and limited resources. A blockage of output from the ED to the main hospital, i.e. boarding of patients in the ED, is considered a primary cause of over-crowding (Schafermeyer and Asplin 2003).

Instead of focusing only on a single hospital, several studies suggest that AD is best seen as a system-wide problem Lagoe et al. 2003; United States General Accounting Office(GAO) 2003), since diversion at one hospital impacts nearby hospitals. If one hospital goes on diversion, other hospitals have an incentive to do so as well, to avoid an onerous increase in traffic. In a recent experiment, two hospitals agreed to measure the impact on one hospital ED, when the other committed to remain off diversion for a week (Vilke et al. 2004). The effect was remarkable: diversion hours fell from 19.4 and 27.7 to 1.4 and zero, respectively. A system-wide collaborative approach in Sacramento, where the hospitals committed to working together to manage patient flow, resulted in significant reductions in AD (Patel et al. 2006). Collaborative efforts in Rochester, New York, saw some improvement (Schneider et al. 2001), and a 24% reduction of hours on AD in Syracuse, New York (Lagoe et al. 2003).

The studies that have been conducted therefore support the contention that ED overcrowding and diversions are symptoms of a systemic problem (Pham et al. 2006). It follows that solutions to the problem may be found outside of the ED itself, a point emphasized in several studies (Lagoe et al. 2002; Fatovich and Hirsch 2003; Schafermeyer and Asplin 2003).

2.2 Proposed Solutions to Overcrowding

The most obvious way to increase throughput in an ED is to add capacity. However, building new space is remarkably expensive, one estimate putting the cost at approximately one million USD per bed (Norland 2005). In one hospital, increasing ICU beds from 47 to 67 decreased average daily time on diversion from 3.8 to 1.4 hours (McConnell et al. 2005). Better management of hospital beds, as opposed to simply increasing the number of beds, is also expected to reduce ED overcrowding in general, and AD in particular (Asplin and Magid 2007).

In addition to increasing resources, there are several suggestions for more flexible use of assets. Selective diversions, e.g. for those patients that emergency medical services (EMS) personnel believe are unlikely to need critical care (Price et al. 2005), is one possibility. A program to predict diversion may effectively preemptively divert in order to avoid formal diversion (Epstein and Tian 2006). Information technology may provide decision support systems to improve ED operations (Gordon and Asplin 2004). And in perhaps the most straight-forward application of internet technology, posting an up-to-the-minute workload schedule of EDs in the Perth area was found to reduce the time spent on diversion by more than a third from 2002 to 2003 (specifically from 1788 to 1138 hours), despite an increase in demand (Sprivulis and Gerrard 2005).

Several operations management techniques have been used to help address the problem of overcrowding (Asplin 2006, Institute of Medicine 2006). For example, Litvak et al. (2001) make the point that most capacity planning is done using average demand on resources, but ignores variability. This immediately suggests that reducing variability in demand would be a viable method to limit diversion, e.g. by scheduling elective surgeries (Litvak et al. 2001, Lane et al. 2000). Lane

et al. (2000) use a system dynamics approach to simulate admission to acute hospitals, and found that elective surgeries function as a safety valve in the UK, being canceled to allow ED patients admittance.

2.3 Current recommendations

The American College of Emergency Physicians (ACEP) have issued Guidelines for AD (Brennan et al. 2000), which suggest that EMS agencies need working agreements to coordinate, and diversion should be a temporary situation, managed systemically, and avoided as much as possible. The guidelines explicitly allow for "selective diversion", without specifying this concept further. The guidelines suggest the decision to go on diversion should reside with the emergency physician at the ED, without being based on financial considerations. We note that this point regards going on diversion in any one particular instance. Since the capacity of a hospital or ED is fundamentally dependent on financial resources, this caveat cannot extend to the system for hospital management over time.

2.4 Incentives to change

The Institute of Medicine report (Institute of Medicine 2006) explicitly points out that hospitals have few financial incentives "to reduce crowding", and to suggest that firm rules are needed to avoid diversion. However, we argue that hospitals do have a financial incentive in their capacity decisions and resource allocations. Further, McConnell et al. (2006) estimate that each hour spent on diversion cost one hospital $ 1,086 in foregone revenue.

2.5 Systems of Agents

There are a number of tools that may be used to study systems. Note that in any system that has a long-term tendency to decrease capacity when there is an excess, and political pressure to increase capacity when there is a dearth, will tend to lie in the borderland between shortage and surplus. This has been called self-organized criticality, and is most familiar from the sand-pile model (Bak et al. 1988).

Systems that suddenly shift from one state to another are well-known in physics, and have been described as cascades (Watts 2002, Arneodo et al. 1998), avalanches (Paczuski et al. 1996), or percolation (Solomon et al. 2000). The difficulty with applying these models to this setting is that hospitals are strategic actors, who are capable of changing their beliefs about the probability distribution over other hospitals actions, depending on what they observe.

In order to capture such games, settings of repeated Prisoner's Dilemma games in various spaces have been examined (Axelrod and Hamilton 1981, Axelrod 1997), and the dynamics are examined using agent based models (Doran et al. 2001, Tesfatsion 2001). The potential for cooperation in the absence of an enforcer has been studied, and in some instances cooperation can evolve (Ostrom et al. 1992, Ostrom 2000).

2.6 Contribution

The multiple methods we use to study the phenomenon of strategic ambulance diversion, and shed light on potential levers of influence, is the primary contribution of this paper. We investigate a policy that raises an interesting question of distributive justice versus efficiency, and we consider the limits of voluntary cooperation in this setting.

3 THE PATIENT BIRTH-DEATH PROCESS: ASSUMPTIONS AND MODEL FORMULATION

In order to model the hospitals occupancy as a continuous time Markov process, we require several assumptions. However, in the interest of space, we refer to the Figure 2. Patients enter, are served, and leave at a rate proportional to the number of patients present. To make the problem tractable, all arrivals and service times are exponentially distributed.

We assume two classes of patients in this model, a high revenue class consisting of the privately insured patients, and a low revenue class, made up of M/M and uninsured patients.

The hospital has a fixed number of beds for the modeling period: N. The hospital will go on full diversion only when at capacity. The hospital will remain on full diversion until some slack develops, specifically until the number of beds occupied is M.

Under the more flexible contract between the hospital and the federal government, the hospital may go on partial diversion earlier than at full capacity, in order to avoid full diversion. We specify $K < N$, and allow the hospital to selectively divert all patients except those with private insurance when the patient population has reached size K. Every state is recurrent

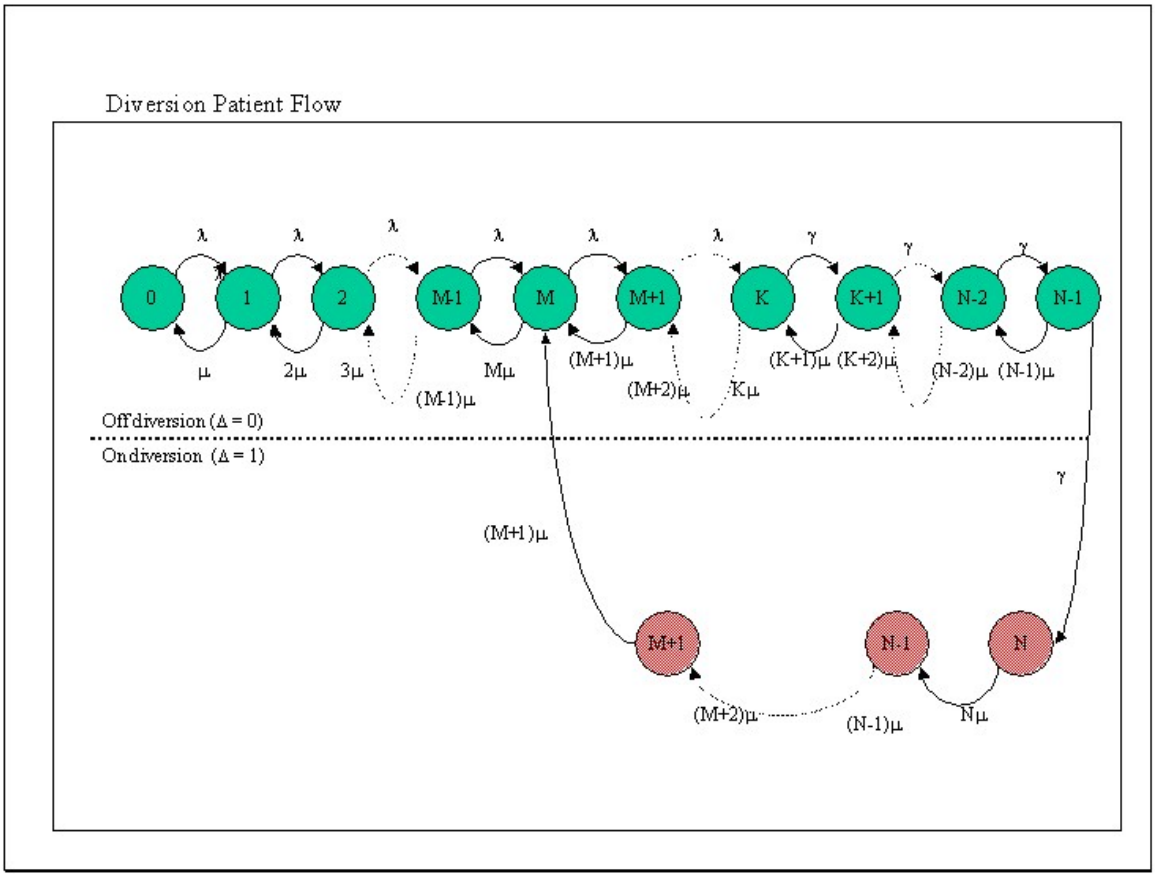

Figure 2: Diversion Patient Flow

and the continuous-time Markov chain is irreducible, so the Markov chain is ergodic. This means that if the hospital runs long enough, every bed occupancy level (X) will take place, and the hospital will spend some time on, as well as off, full diversion. This assumption assures the existence and uniqueness of the limiting probability distribution (Ross 1996).

3.1 State Space and Notation

We indicate that the hospital is on full diversion with $\Delta = 1$, and off full diversion with $\Delta = 0$. X captures the number of beds occupied. The states are fully specified by the number of beds occupied, and whether or not the hospital is on full diversion: $\Omega = \{0,1\} \times \{0,...,N\}$. The states may therefore be visualized as on a chain of nodes, with one loop for diversion (see Figure 2).

3.2 Model

The limiting probabilities are characterized by the balance equations, which are based on the fact that for an ergodic continuous-time Markov chain, the number of exits and entries to a state must equal. We specify the balance equations and present the solution as the limiting probabilities for the system, or the long-run proportion of time the system will be in a given state. We first provide a small example in order to motivate the discussion.

3.3 Small-Scale model

We illustrate our model with a simple with only two beds. The arrival rate overall is λ, and once one bed is occupied, only a subset of patients are brought in. The arrival rate with one bed occupied is therefore lower, and we denote the new arrival

rate as γ, such that $\gamma < \lambda$. Since N is equal to 2, diversion sets in when the hospital is full, and there is only one other state in the state space, namely when one bed is occupied while the hospital is on diversion (Node 1,1). This gives the following solution:

$$P_0 = \frac{2\mu(\mu+\gamma)}{2\mu(\mu+\gamma+\lambda)+3\gamma\lambda} \quad (1)$$

$$P_1 = \frac{2\lambda\mu}{2\mu(\mu+\gamma+\lambda)+3\gamma\lambda} \quad (2)$$

$$P_2 = \frac{\lambda\gamma}{2\mu(\mu+\gamma+\lambda)+3\gamma\lambda} \quad (3)$$

$$P_{1,1} = \frac{2\gamma\lambda}{2\mu(\mu+\gamma+\lambda)+3\gamma\lambda} \quad (4)$$

$$(5)$$

This small model suggests some qualitative results which may also hold in the full-scale model. If we consider that selective diversion boils down to selecting an arrival rate γ to limit the time spent on diversion, then the marginal effect of γ on the time spent on diversion, $P_{1,1}$, is as follows:

$$\frac{\partial P_{1,1}}{\partial \gamma} = \left(\frac{2\lambda}{2\mu(\mu+\gamma+\lambda)+3\gamma\lambda}\right)\left[1-\frac{\gamma(2\mu+3\lambda)}{2\mu(\mu+\gamma+\lambda)+3\gamma\lambda}\right] \quad (6)$$

Since $\gamma, \lambda, \mu > 0$, $\frac{2\mu\gamma+3\gamma\lambda}{2\mu^2+2\mu\gamma+2\mu\lambda+3\gamma\lambda} < 1$, and therefore both of the factors in Equation (6) are positive, i.e. increasing the arrival rate while on selective diversion increases the time on full diversion. This intuitive result is illustrative of the greater issue: by managing the distribution of patients such that hospitals on the verge of full diversion are partially shielded, full diversion with its concomitant problems are avoided some portion of the time. The full-scale model is qualitatively very similar to the small-scale model, and we remove it to focus on the simulations.

In order to assess the incentives for the hospitals, we take into account revenue, as well as the occupancy rates. We therefore turn to empirical data to provide the input to a simulation to assess occupancy and revenue. The rationale for using simulation is that while the state space without revenue classes is of size $2N-M$, even with only two revenue classes each occupancy level may hold from 0 to its level of one patient type, and so the size of the state space grows to $\frac{1}{2}((N+1)(N+2)+(N-M-1)(N+M+2))$. In our example, with $N=100$ beds and an occupancy level to go off diversion of $M=90$, this gives 13695 states.

3.4 Data

The data-set was comprised of 140,720 records from 27,002 anonymous patients based on 459 days from an Dekalb Medical Center in Atlanta Georgia, which supplies more than thirty thousand services with individual charges (Hardy and Te 2005). Mean lengths of stay per type of patient varied from 1 to 140 days, while the average actually collected by type of patient varied from \$7 to \$774. The mean length of stay was 5.18 days, with a standard deviation of 7.73 days, and with an average collection of \$3,516.51 (standard deviation \$7118.16), which corresponds to an average per day charge collected of \$679.41.

3.5 Discrete Event Simulation

We begin with an initial patient population drawn from the distribution of patients in the hospital data. We assume exponential arrival and service times, but use the sample mean inter-arrival and service times to calibrate. Since we use different distributions for eighteen different patient classes, the simulation has an added degree of realism, in that patients arrive at different frequencies, have different mean lengths-of-stay, and generate different average revenue per day.

The hospital beds are filled on a first-come first-served basis, and when the hospital reaches capacity of 100 full beds, it ceased to accept patients, until there are 10 free beds. We then test various occupancy levels at which to go on partial diversion, K, from 1 to 25. For brevity, we denote the number of beds reserved "Capital Protection Level", or "CapProt" in the figures. As indicated in Figure 3, there does not appear to be much effect when just a few beds are reserved, suggesting that the revenue effect is modest.

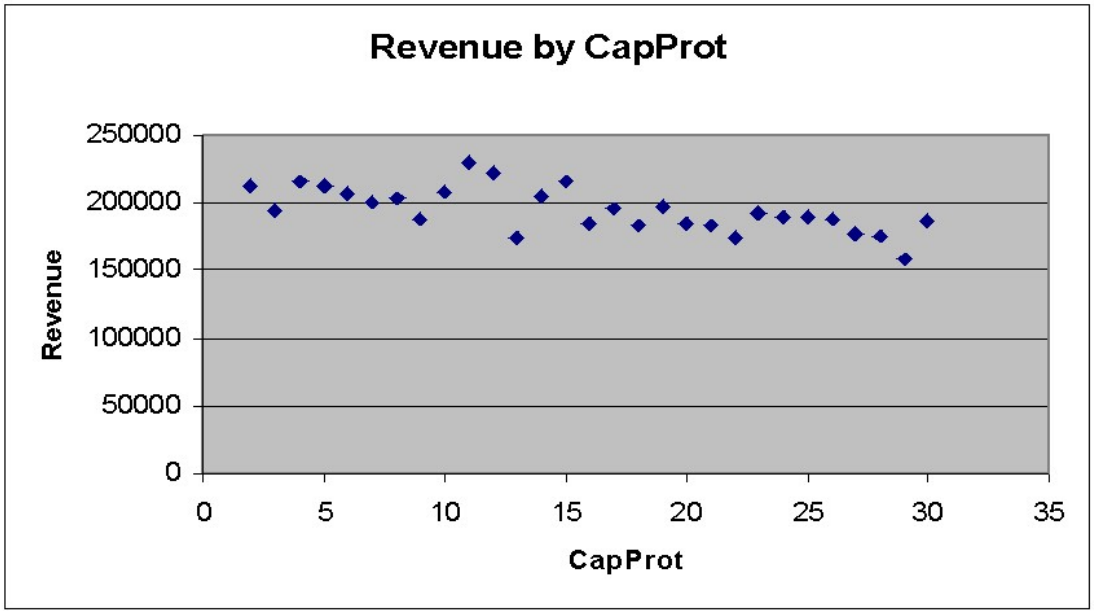

Figure 3: Revenue by Capital Protection Level

The effect on time spent on diversion is more striking. Figure 4 shows how the time spent on diversion falls to nearly zero with eight beds reserved for partial diversion. This compares to an average time on diversion without partial diversion of 12.42 days out of 100. The results for eight beds for a simulation of 100 runs of 100 days are shown in Table 1.

We run the simulation for 25 beds to illustrate that neither the hospital nor the patients benefit from reserving too many beds: the utilization rate falls, and revenue does as well. It is true that the diversion time when 25 beds are reserved falls to an actual value of zero in the simulation, but that compares to 0.14 days out of 100 with 8 beds reserved. When compared to 12.41 without this partial diversion, we consider 8 beds reserved sufficient to resolve the problem in practice.

To give the hospital an incentive to introduce this scheme, the revenue effect is crucial. With 8 beds reserved, revenue increases from $671,976 (s=170,624) to $714,217 (s=143,670).

4 Agent-Based Simulation

In order to examine the effects of varying other parameters, including spatial aspects, we created the simulation in Netlogo (Wil , Tisue and Wilensky 2004b, Tisue and Wilensky 2004a). Schelling's segregation model (Schelling 1969) was included, in order to create a spatial pattern of two groups of patients (which we could think of high and low income). We then experimented with 1260 different simulations, using patient loads between 900 and 1100, with capacity was distributed around a mean of 1000, and to our surprise, we were never able to find a set of parameters that included variability in capacity, and showed a gain for the underserved patients. This precludes any need for statistics.

Since the purpose of the exercise was to identify intervals in the parameter space that allowed a Pareto improvement, we concluded that the noise in the system drowns out any advantage that may be provided to the diverted group from an increased patient flow.

5 GAME-THEORETIC APPROACH

Having found from the Agent-Based Model that the PD approach would not be robust in the face of noise, we turned to game theory and threshold models from physics to find opportunities for cooperation in the system of hospitals. Again, we

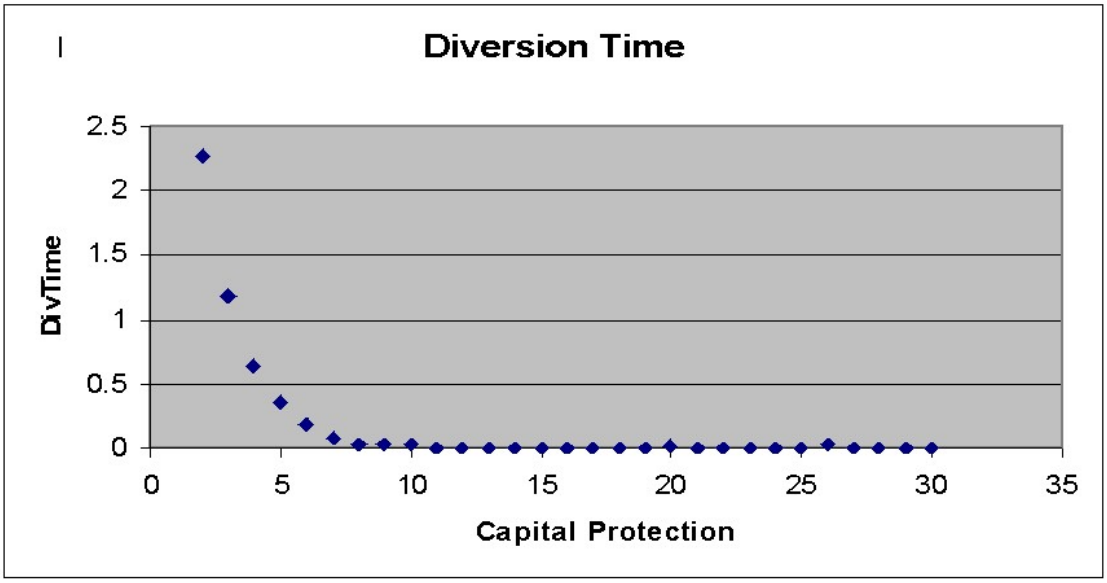

Figure 4: Diversion Time by Capital Protection Level

retain as much as possible from the previous work, and examine how the hospitals would approach each opportunity to stay open in the face of high traffic (i.e. cooperate), or go on diversion before absolutely necessary (i.e. defect).

Modeling the decision of hospitals to go on AD as a sequential game, with hospitals discovering their patient level, and observing other hospitals going on AD, was made challenging due to the ability of strategic players to update their beliefs about the other actors. The physics literature appeared to describe similar phenomena, but never with actors that are able to play strategically. Therefore, percolation, cascade, and avalanche models, all failed to be directly applicable. However, the threshold effect from raising some parameter (temperature, energy level, pressure), and seeing the system change radically at some point, seemed appropriate to hospital diversions. Specifically, at low levels of traffic, there's no diversion, and at high levels, everyone is on diversion (which really devolves to nobody being on diversion, except when there's a safety hospital).

This observation, that a system would tend to a fragile equilibrium between excess capacity and too frequent diversion, would suggest that the Self-Organized Criticality (SOC) literature may be applicable. The canonical example of SOC, the sand-pile metaphor, consists of having sand grains dropped down one at a time. At random times, avalanches occur, with the angle of the sand-pile fairly constant. This angle represents the capacity of the hospital system per population unit.

The metaphor is useful to consider why we might find hospital systems so prone to go on diversion the world over. A perception of "too many empty beds" will lead to cutbacks, while a complete failure of the system leads to added investment. One would therefore expect rational investors to balance the pain caused by ambulance diversion with the pain from investing scarce capital to avoid diversion, leading to a system constantly on the edge of seizing up. This is the hallmark of complex systems, and so opens up alternative approaches to studying ambulance diversion in systems over time. However, the strategic nature of hospitals does not lend itself to passive players such as grains of sand.

To try to capture the salient points, without making the game too simple, we used the following payoff function:

$$\pi = -(a + X - F)^2 - P \cdot I(AD) \tag{7}$$

Here, π is the payoff, a is the initial patient level, X is the added patient load, F is an ideal patient load, meant to be less than having patients in hallways and boarding in the ER; i.e. less than the absolute limit of patients that could be

Table 1: Simulation Results for DeKalb Medical Center

	Mean Number of Patients (SD)	Total HMO Patients	Total M/M Patients	Total Uninsured Patients
Without Capacity Protection	91.9 (0.61)	541	686	161
Capacity Protection of 25	74.7 (0.34)	692	383	88
Capacity Protection of 8	89.84 (0.45)	688	560	132

managed. P is a fixed penalty for going on diversion, and $I(AD)$ is an indicator function equal to 1 on ambulance diversion, and zero otherwise.

This functional form allows a maximum to be reached at some level less than overfull, with an increasing loss at very low levels of occupancy, as well as very high patient loads. The penalty allows the intrinsic costs of going on and off diversion to be included, as well as any policy-driven fee for doing so.

5.1 Prisoner's Dilemma Aspect of the two-player game

If we either consider only two hospitals in the metro area, we can study the payoffs and include the effect of having both hospitals go on diversion. In that case, to add to the pain of diversion, we assume that the hospitals are forced to take their own patients. To shorten notation, let $Y_i = a_i + X_i$. In that case, the payoffs to the patients from diverting or not are:

		Player 2	
		NAD	AD
Player 1	NAD	$-(Y_1-F)^2$,$-(Y_2-F)^2$	$-(Y_1+Y_2-2F)^2$,$-P$
	AD	$-P$, $-(Y_1+Y_2-2F)^2$	$-(Y_1-F)^2-P$,$-(Y_2-F)^2-P$

Assuming $Y_i > F$, this is a Prisoner's Dilemma if $-(Y_i+Y_j-2F)^2 < -(Y_i-F)^2 - P <_(Y_i-F)^2 < -P$. The first two inequalities always hold with our assumptions, and the third simplifies to: $Y_i^2 - 2FY_i + (F^2-P) > 0$, which again yields $a_i + X_i > F + \sqrt{P}$, consistent with what we find below. In other words, provided the patient inflow is sufficiently higher than the ideal load, the temptation to defect exists. This implies, among other things, that a larger group of hospitals will have an even harder time cooperating than a two hospital system (Hauert and Schuster 1997).

5.2 N-hospital Game

Next we expand to the N hospital game. Whenever Nature has chosen an intensity or temperature, the beliefs of the other players as functions of ones own observed patient load become difficult to find in closed-form. We therefore simplify, and assume an initial distribution a_i for each hospital. The focal hospital is just a or a_0, so that the remaining hospitals are indexed with $i\varepsilon\{1,\ldots,N\}$, for a total of $N+1$ hospitals. We assume that the initial distribution is common knowledge, as is the distribution of new arrivals X. We do not specify the distribution, however, which simplifies the math.

We use a threshold T to indicate the patient load at which a hospital will go on ambulance diversion. We envision this $T > F$, which also follows from the equations below.

For convenience, we also define Z as a column vector of size N, such that each entry is zero if the corresponding hospital did not divert anyone, and otherwise represents the number of patients diverted. AD is the corresponding vector of dummy variables, such that a one indicates a hospital on diversion, while a zero indicates the hospital accepted all the patients. If $One(N)$ is an N-vector of ones, then $\Gamma = Z'One(N)/[1+(One(N)-AD)'One(N)]$ represents the number of diverted patients, divided by the number of hospitals that did not divert. We assume that patients are evenly divided by the non-diverting hospitals, and Γ is therefore the number of additional patients the focal hospital will receive.

The distribution of X is simply $P(X=x)$, with support in the natural numbers. From this basic structure, we may characterize the expected payoff with and without diversion. Without diversion, we assume the hospital can keep exactly F patients, and so $E\pi = -P$. Without diversion, the expression is more complex:

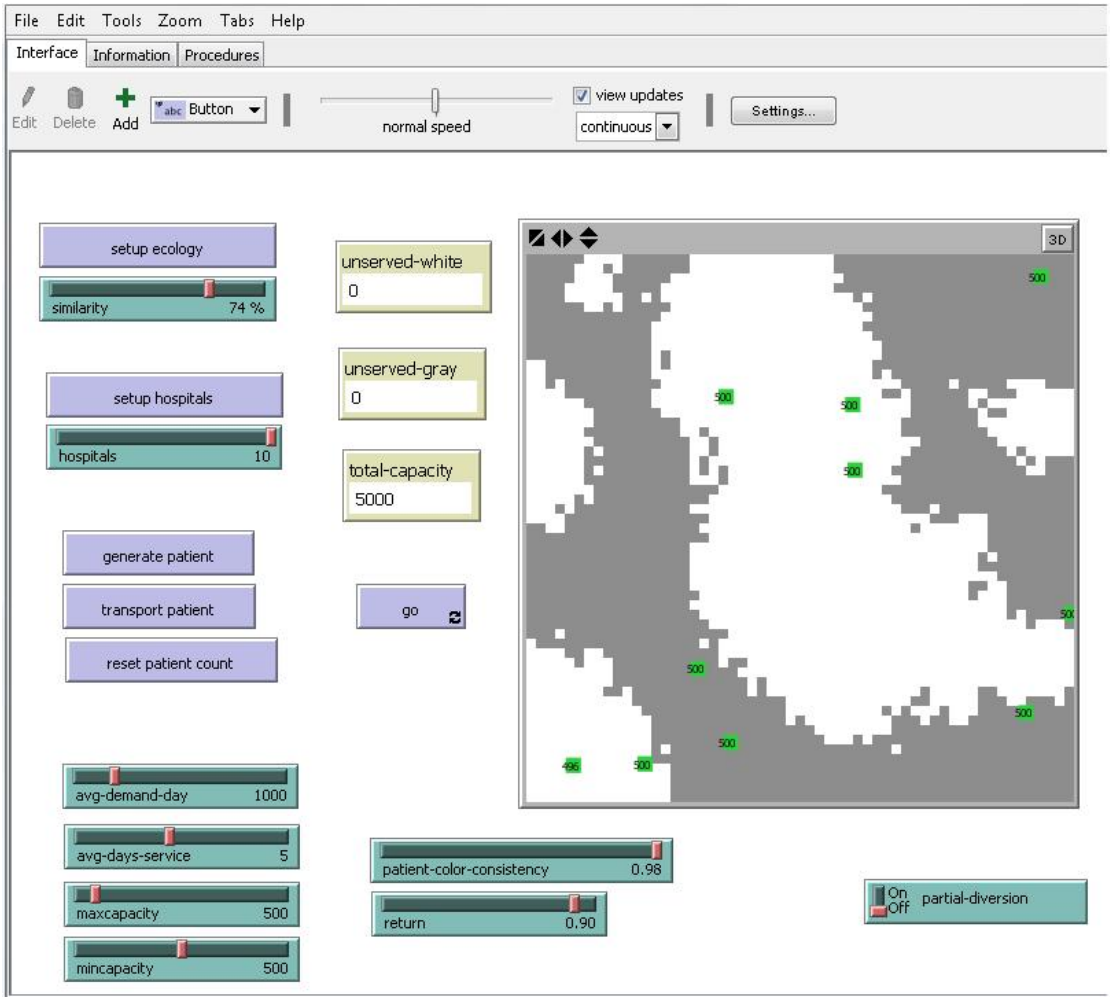

Figure 5: NetLogo Agent-Based Simulation

$$E^{NAD}\pi = \sum_{Z} -(a+X-F+\Gamma)^2 \cdot P(Z) \tag{8}$$

Z is a vector, and the probability of Z being any particular value depends on both X and T. However, we assume that all the hospitals follow the same strategy of diverting at T, and that all the patient distributions are independent, which means that for each hospital, $P(Z_i = 0) = P(X_i < T - a_i)$, $P(Z_i \varepsilon \{1, \ldots, T-F-1\}) = 0$, and for any $z \geq T-F, P(Z_i = z) = P(X_i F - a_i + z)$. Due to our simplifications, $P(Z) = \Pi_i P(Z_i)$. Since we have constructed the payoff function so that increasing the patient load when the focal hospital should have been on diversion must decrease the expected payoff, the optimal policy requires that any patient load higher than T must also lead to diversion. Similarly, any patient load lower than any other that does not require diversion, also does not require diversion. Notice that with this formulation, with N assumed large, we further simplify by not assigning the patients back to the hospitals if all the hospitals go on diversion.

We therefore seek a T such that $E^{NAD}\pi(a+X=T-1) > -P$, while $E^{NAD}\pi(a+X=T) < -P$. Multiplying the second inequality by -1 and adding gives $E^{NAD}\pi(a+X=T-1) - E^{NAD}\pi(a+X=T) > 0$. This simplifies to $T > F + \frac{1}{2} - \sum_Z \Gamma \cdot P(Z)$. In words, the threshold must exceed the ideal level F, minus the expected overflow from the other hospitals, $E\Gamma$, plus half a patient. The latter is intriguing, but stems from the unit increase in patients arriving.

This result is unsurprising, so we seek alternate ways to characterize T. Since the cutoffs are discrete patients, we must allow discontinuity at each integer, so calculus is infeasible. Second, we imagine that $T > F$, but that if the hospital is first going to incur the cost of diversion, it will divert enough patients to achieve the ideal load of F, so the amount diverted

is discontinuous at T. Due to the threshold strategy, we therefore solve for $E^{NAD}\pi(a+X=T) \approx -P$. A somewhat long quadratic equation arises, which has one economically meaningful solution:

$$T = F - E\Gamma + \sqrt{P - (E(\Gamma^2) - (E\Gamma)^2} = F - E\Gamma + \sqrt{P - Var(\Gamma)} \tag{9}$$

The radical reflects the fact that unless the penalty P is sufficiently high, i.e. greater than the variance of the expected overflow of patients, it will be dominated by the noise in the system and have no impact. However, the previous result continues to hold, and the threshold will therefore be at least $F + \frac{1}{2} - \sum_Z \Gamma \cdot P(Z)$, up to $P = Var(\Gamma) + \frac{1}{4}$, and then rising with P.

In order for this threshold policy to be a Nash Equilibrium (NE) for the set of hospitals, the policy must be in the best response set to all the other hospitals diverting at T. Note that Γ is a function of T, which we underscore by introducing $G(T) = F - E\Gamma(T) + \sqrt{P - Var(\Gamma(T))}$. As the threshold goes to zero, all patients are diverted, and provided $P \geq Var(\Gamma)$, $G(T) > 0$, and in particular, for small values of T, $G(T) > T$. If the threshold goes to infinity, no patients are ever diverted, which means $E\Gamma$ and $Var(\Gamma)$ both go to zero. Hence for large values of T, $G(T) \to F + \sqrt{P} < T$. By the mean value theorem, there exists such a T. This T may not be integer, so we allow the hospitals to randomize over strategies, which ensures the existence of a Nash Equilibrium.

5.3 Discussion

Since the variance of the overflow will increase if we split a patient flow up into many small streams, this is yet another argument for consolidating into one hospital or hospital system. However, since moving to more independent hospitals moves the game closer to a mean-field game (Johnson et al. 1998), it is not surprising that more agents leads to the need for a stronger penalty to force cooperation.

The main observation from this section is to note that for certain values of the parameter space, the game devolves to a Prisoner's Dilemma, which has been extensively studied in the evolutionary game theory literature. We simply note that without either binding contracts, enforced by an external agent, or direct regulation by the external agent, this makes cooperation very difficult.

6 CONCLUSION

In this paper we address cooperative strategies for hospitals, in order to reduce ambulance diversion. These strategies lie in between the extreme of laissez faire individualism and a central planner approach. We first use standard tools to model the flow of patients, then simulate to get more detail, and then another simulation to assess the robustness of a Pareto improving approach.

Partial Diversion allows a hospital to divert some traffic when capacity has become tight. Although the Federal government may be able to force through such a measure in theory, the ethical problems of partial diversion are troubling.

The Agent-Based model shows that with added noise, a partial diversion scheme to share load is fragile. Without being able to claim that such a scheme is a Pareto improvement, except under unrealistically narrow conditions, it seems clear that this approach is infeasible. Since that is even true even without taking into account the ethical issue this approach raises, it is safe to say that partial diversion is a non-starter.

The game theory apporach shows that without some form of binding cooperative scheme, the incentives to defect are not only strong, but will often lead to system-wide pre-emptive diversion. Not surprisingly, the hospitals operate under a Prisoner's Dilemma when the patient load is sufficient. One noteworthy aspect of the game is that the penalty, P, must be large enough to balance out the variance of the overflow before there is a an optimal threshold.

Having examined these aspects of cooperative solutions to ambulance diversion, we believe the incentives require a centralized agent to route patients, at least when the patient load is high.

6.1 Future Research

Some avenues of future research on ambulance diversion remain promising. It would be of interest to map out how and when a central planner should intervene, and what minimum rules must be in place to maximize system-wide efficiency. Also, the challenge of balancing distributive justice with efficiency is stark in the setting of partial diversion, and the legal and ethical

implications are intriguing. Although we concluded that partial diversion was infeasible, any situation with wealth-dependent care raises the same issue, so we believe it is an interesting question.

ACKNOWLEDGEMENTS

DeKalb Medical Center provided data and initial input for this study. This research was partially funded through an NSF Careers grant.

REFERENCES

Arneodo, A., J. Muzy, and D. Sornette. 1998. "Direct" causal cascade in the stock market. *The European Physical Journal B-Condensed Matter and Complex Systems* 2 (2): 277–282.

Asplin, B. 2006. Hospital-Based Emergency Care: A Future Without Boarding? *Annals of Emergency Medicine* 48 (2): 121–125.

Asplin, B., and D. Magid. 2007. If You Want to Fix Crowding, Start by Fixing Your Hospital. *Annals of Emergency Medicine* 49 (3): 273–274.

Asplin, B., D. Magid, K. Rhodes, L. Solberg, N. Lurie, and C. Camargo. 2003. A conceptual model of emergency department crowding. *Annals of Emergency Medicine* 42 (2): 173–180.

Axelrod, R. 1997. Advancing the art of simulation in the social sciences. *Complexity* 3 (2): 16–22.

Axelrod, R., and W. Hamilton. 1981. The evolution of cooperation. *Science* 211 (4489): 1390–1396.

Bak, P., C. Tang, and K. Wiesenfeld. 1988. Self-organized criticality. *Physical review A* 38 (1): 364–374.

Brennan, J., D. Allin, A. Calkins, E. Enguidanos, L. Heimbach, J. Pruden, and D. Stilley. 2000. Guidelines for ambulance diversion. *Annals of Emergency Medicine* 36 (4): 376–377.

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

Derlet, R., and J. Richards. 2000. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Ann Emerg Med* 35 (1): 63–8.

Derlet, R. W., J. R. Richards, and R. L. Kravitz. 2001. Frequent overcrowding in U.S. emergency departments. *Academic Emergency Medicine* 8:151–155.

Doran, J., S. Franklin, N. Jennings, and T. Norman. 2001. On cooperation in multi-agent systems. *The Knowledge Engineering Review* 12 (03): 309–314.

Epstein, S. K., and L. Tian. 2006. Development of an emergency department work score to predict ambulance diversion. *Academic Emergency Medicine* 13 (4): 421426.

Fatovich, D. M., and R. L. Hirsch. 2003. Entry overload, emergency department overcrowding, and ambulance bypass. *Emergency Medicine Journal* 20:406–409.

Geer, R., and J. Smith. 2004. Strategies to take hospitals off (revenue) diversion. *Healthcare Financial Management* 58 (3): 70–74.

Gordon, B., and B. Asplin. 2004. Using Online Analytical Processing to Manage Emergency Department Operations. *Academic Emergency Medicine* 11 (11): 1206.

Hardy, A., and S. Te. 2005, March. Interview at DeKalb Medical Center on march 9, 2005. Personal communcation.

Hauert, C., and H. Schuster. 1997. Effects of increasing the number of players and memory size in the iterated Prisoner's Dilemma: a numerical approach. *Proceedings of the Royal Society of London-B-Biological Sciences* 264 (1381): 513–520.

Institute of Medicine 2006. The future of emergency care in the United States health system. *Annals of Emergency Medicine* 48 (2): 115–120.

Johnson, N., S. Jarvis, R. Jonson, P. Cheung, Y. Kwong, and P. Hui. 1998. Volatility and agent adaptability in a self-organizing market. *Arxiv preprint cond-mat/9802177*.

Lagoe, R., J. Kohlbrenner, L. Hall, M. Roizen, P. Nadle, and R. Hunt. 2003. REDUCING AMBULANCE DIVERSION: AM ULTIHOSPITAL APPROACH. *Prehospital Emergency Care* 7 (1): 99–108.

Lagoe, R. J., R. C. Hunt, P. A. Nadle, and J. C. Kohlbrenner. 2002. Utilization and impact of ambulance diversion at the community level. *Prehospital Emergency Care* 6 (2): 191–198.

Lane, D. C., C. Monefeldt, and J. V. Rosenhead. 2000. Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society* 51:518–531.

Litvak, E., M. C. Long, A. B. Cooper, and M. L. McManus. 2001. Emergency department diversion: Causes and solutions. *Academic Emergency Medicine* 8 (11): 1108–1110.

McConnell, K., C. Richards, M. Daya, S. Bernell, C. Weathers, and R. Lowe. 2005. Effect of Increased ICU Capacity on Emergency Department Length of Stay and Ambulance Diversion. *Annals of Emergency Medicine* 45 (5): 471–478.

McConnell, K., C. Richards, M. Daya, C. Weathers, and R. Lowe. 2006. Ambulance Diversion and Lost Hospital Revenues. *Annals of Emergency Medicine* 48 (6): 702–710.

Norland, S. 2005. Containing costs in the ED. *Healthcare Financial Management* 59:66–73.

Ostrom, E. 2000. Collective action and the evolution of social norms. *The Journal of Economic Perspectives*:137–158.

Ostrom, E., J. Walker, and R. Gardner. 1992. Covenants with and without a sword: Self-governance is possible. *The American Political Science Review*:404–417.

Paczuski, M., S. Maslov, and P. Bak. 1996. Avalanche dynamics in evolution, growth, and depinning models. *Physical Review E* 53 (1): 414–443.

Patel, P. B., R. W. Derlet, D. R. Vinson, M. Williams, and J. Wills. 2006. Ambulance diversion reduction: the Sacramento solution. *The American Journal of Emergency Medicine* 24:206–213.

Pham, J. C., R. Patel, M. G. Millin, T. D. Kirsch, and A. Chanmugam. 2006. The effects of ambulance diversion: A comprehensive review. *Academic Emergency Medicine* 13:1220–1227.

Price, T. G., E. A. Hooker, and J. Neubauer. 2005. Prehospital provider prediction of emergency department dispostion: Implications for selective diversion. *Prehospital Emergency Care* 9 (3): 322–325.

Ross, S. M. 1996. *Stochastic processes*. 2 ed. John Wiley & Sons, Inc.

Schafermeyer, R. W., and B. R. Asplin. 2003. Hospital and emergency department crowding in the united states. *Emergency Medicine* 15:22–27.

Schelling, T. 1969. Models of segregation. *The American Economic Review*:488–493.

Schneider, S., F. Zwemer, A. Doniger, R. Dick, T. Czapranski, and E. Davis. 2001. Rochester, new york: A decade of emergency department overcrowding. *Academmic Emergency Medicine* 8 (11): 10441050.

Solomon, S., G. Weisbuch, L. de Arcangelis, N. Jan, and D. Stauffer. 2000. Social percolation models. *Physica A: Statistical Mechanics and its Applications* 277 (1-2): 239–247.

Sprivulis, P., and B. Gerrard. 2005. Internet-accessible emergency department workload information reduces ambulance diversion. *Prehospital Emergency Care* 9 (3): 285–291.

Tesfatsion, L. 2001. Introduction to the special issue on agent-based computational economics. *Journal of Economic Dynamics and Control* 25 (3-4): 281–293.

Tisue, S., and U. Wilensky. 2004a. NetLogo: A simple environment for modeling complexity. In *International Conference on Complex Systems*, 16–21.

Tisue, S., and U. Wilensky. 2004b. NetLogo: Design and implementation of a multi-agent modeling environment. In *Proceedings of Agent 2004*, 7–9. Citeseer.

United States General Accounting Office(GAO) 2003. Hospital Emergency Departments - Crowded Conditions Vary among Hospitals and Communities. Technical report, United States General Accounting Office(GAO).

Vilke, G., L. Brown, P. Skogland, C. Simmons, and D. Guss. 2004. Approach to decreasing emergency department ambulance diversion hours. *J Emerg Med* 26 (2): 189–92.

Watts, D. 2002. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99 (9): 5766.

AUTHOR BIOGRAPHIES

REIDAR HAGTVEDT is a Visiting Assistant Professor of Healthcare Operations Management at the University of Alberta School of Business, with a joint appointment to iCare, the health research arm of Capital Health. He received his Ph.D. in Business from Georgia State University in 1999, and a Ph.D. in Industrial and Systems Engineering from Georgia Institute of Technology in 2008. His email address is Hagtvedt@ualberta.ca.

MARK FERGUSON is the Steven A. Denning Professor of Technology and Management and the John and Wendi Wells Associate Professor of Operations Management in the College of Management at Georgia Tech. He received his PhD in Business Administration, with a concentration in Operations Management from Duke University in 2001. Dr. Ferguson's research interests involve many areas of supply chain management including supply chain design for sustainable operations, contracts that improve overall supply chain efficiency, pricing and revenue management, and the management of perishable products. Two of his papers have won best paper awards from the Production and Operations Management Society and several of his research

projects have been funded by the National Science Foundation. His email address is Mark.Ferguson@mgt.gatech.edu.

PAUL GRIFFIN is a professor in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Penn State University, where he serves as the Peter and Angela Dal Pezzo Department Head Chair. His research and teaching interests are in health and supply chain systems. In particular, his current research activities have focused on cost-effectiveness modeling of public health interventions, health logistics, health access and economic modeling, and supply chain coordination and control including pricing and contracting mechanisms. His email address is pmg14@psu.edu.

GREGORY TODD JONES is Faculty Research Fellow at the Georgia State University College of Law, Director of Research at the Interuniversity Consortium on Negotiation and Conflict Resolution, and Director of the Computational Laboratory for Complex Adaptive Systems. He received a Ph.D. in Decision Sciences and a J.D., both from Georgia State University, in 2003. His research focuses on emergent cooperation in computational simulation models. His email address is gtjones@gsu.edu.

PINAR KESKINOCAK received her Ph.D. in Operations Research from Carnegie Mellon University in 1997. She is an associate professor in the H. Milton Stewart School of Industrial and Systems Engineering and the co-founder and co-director of the Center for Humanitarian Logistics at Georgia Institute of Technology. Her research focuses on applications of operations research and management science with societal impact (particularly health and humanitarian applications), supply chain management, pricing and revenue management, and logistics/transportation. Her research has been published in journals such as Operations Research, Management Science, Manufacturing & Service Operations Management, Production and Operations Management, IIE Transactions, Naval Research Logistics, and Interfaces. Her email address is pinar@isye.gatech.edu.