Role of Parallelism in Ambulance Dispatching

Seokcheon Lee

***Abstract*—The demand for emergency medical service (EMS) has been rising over time, leading to the need for efficient yet effective techniques for managing ambulance logistics. Ambulance dispatching decisions in EMS assign ambulances to calls such that the response time is minimized. A notion of parallelism is developed for ambulance dispatching decisions that allows considering both idle and busy ambulances in parallel rather than just idle ones. The parallelism, applied upon the centrality policy found in literature, results in the parallelized centrality policy complementing and enhancing the centrality policy. The experimental analysis evidences that the parallelism significantly reduces response time by up to 43.4% over the existing approaches that only consider idle ambulances.**

***Index Terms*—Ambulance dispatching, emergency medical service, parallelism, response time.**

I. Introduction

THE DEMAND for emergency medical service (EMS) has been rising over time [see Fig. 1 for the number of emergency department (ED) visits by ambulance between 2003 and 2010 in the U.S.], leading to the need for efficient yet effective techniques for managing ambulance logistics. Response time, which is significantly influenced by the ambulance logistics, is the time taken to reach patient after an emergency call is received, and it has been used as an important performance measure since it directly affects the welfare and safety of patients. For example, sudden cardiac arrest is a leading cause of deaths in the U.S., responsible for more than 350 000 deaths each year [2]. The effect of a 1 m reduction in response time for patients with sudden cardiac arrest is estimated to increase survival rate by 24% [3].

Three types of ambulance logistics decisions are associated with the response time: 1) ambulance location; 2) ambulance relocation; and 3) ambulance dispatching. First, ambulance location problems involve establishing optimal locations of ambulance stations in terms of coverage (see [4]–[7] for detailed reviews). Coverage is the fraction of calls that can be responded within a time limit, assuming a demand node is covered by an ambulance station if the time length between them is within the time limit. Second, relocation decisions enforce ambulances to move to different locations in order to increase the coverage based on temporal and geographical demand patterns [8]–[13].

Manuscript received April 16, 2013; revised August 26, 2013; accepted October 19, 2013. Date of publication January 16, 2014; date of current version July 15, 2014. This paper was recommended by Associate Editor N. Wu.

The author is with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: stonesky@purdue.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2013.2296280

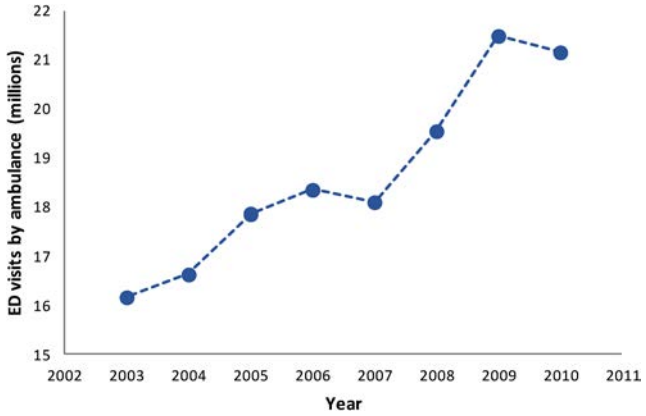

Fig. 1. ED visits by ambulance, 31% increase in 7 years and 4% increase annually [1].

Lastly, ambulance dispatching decisions assign ambulances to calls, which can be either call-initiated or ambulance-initiated [14], [15]. When a newly arriving call finds several idle ambulances, it initiates a decision of selecting a unit (ambulance) among idle units (call-initiated). If calls cannot be immediately assigned, they start being queued and a unit that has just got freed has to choose a call among those waiting, thereby initiating a dispatching decision (ambulance-initiated). The relevance of the two types of dispatching decisions depends on the busyness of the system. Call-initiated decisions are more relevant in routine emergency scenarios where system load is relatively low, whilst ambulance-initiated decisions play a primary role in high load conditions.

A greedy policy, dispatching the closest unit available or dispatching to the closest call waiting, can be used for the two types of dispatching decisions. The greedy policy, due to the computational efficiency and at the same time the capability of achieving a certain level of effectiveness, is most commonly used in EMS practice [13], [16]–[18], and it is also widely adopted in various other applications [19]–[26].

In a recent research, Lee [14], [15] introduced a novel ambulance dispatching policy for ambulance-initiated decisions, centrality policy, in response to the rising occurrence of catastrophic disasters that the world has been experiencing. If a patient chosen is located away from the rest of patients, the next response times will tend to increase. The centrality policy, therefore, prioritizes calls based on the so-called centrality that can be interpreted as the efficiency of a call site in reaching out other calls or the density of calls around a call with respect to the geographical call distribution over the service area. The experimental results show that the centrality policy reduces response time by up to 86% over the greedy policy. The centrality consideration is from the recognition of the fact

that a large portion of incidents are managed on site without need for transport to hospital. For example, the percentage of essential emergency calls that require transferring to hospital is reported to be only 25% in the U.S. [27]. In contrast, the statistics collected by the U.K. Department of Health, during April 2012 to February 2013, from different regions of U.K., state that the percentage varies from 46–78% (http://www.england.nhs.uk/statistics/ambulance-quality-indicators/ambqi-2012-13/). This implies that it is highly possible that a unit continues serving several calls before having to go to hospital. Therefore, the centrality can play an important role in guiding dispatching decisions and reducing response time.

The greedy policy and centrality policy, however, take into account only idle units in making dispatching decisions, despite the possibility that a busy unit can respond more quickly even after the completion of currently assigned service. It would therefore be possible to further improve these policies by incorporating both idle and busy units in parallel. The parallelism blurs the boundary between call-initiated and ambulance-initiated decisions, and an assignment problem in either case has to be solved that matches between multiple (idle/busy) units and multiple unassigned calls. This paper aims to propose a method of synthesizing the parallelism into the centrality policy, producing a novel policy called parallelized centrality policy, and demonstrate the impact of the parallelism on performance improvement. The analysis evidences that the parallelism can significantly reduce the average as well as variation of response time beyond existing approaches that only consider idle units, thus mitigating the risk of exposing patients to excessively tardy responses.

The rest of this paper is organized as follows. Section II introduces the centrality policy to which the parallelism is applied. The centrality policy is transformed into the parallelized centrality policy in Section III, and the impact of parallelism is evaluated in various scenarios in Section IV. In Section V, the parallelized centrality policy is more generalized by incorporating calibration parameters, based on the lessons from the experimental analysis. Finally, Section VI concludes this paper and discusses future work.

II. Centrality Policy

This section introduces the centrality policy that is recently developed by Lee [14], [15] in support of ambulance-initiated dispatching decisions, as a basis to which the parallelism principle is integrated. When an ambulance gets freed, a network can be constructed where nodes represent waiting calls that have not been assigned to any unit and an edge between every pair of calls has a value of distance (in time) between the two call sites connected by the edge. Node centrality in a network indicates the importance of a node in the operational efficiency of the network [28], [29], and this network representation of calls facilitates quantifying the centrality of calls. When calls are prioritized by the centrality and a unit is dispatched to the most central call, the unit will be given the opportunity, after the completion of immediate service, to serve the other calls around it at the maximum rate of completion. However, if calls are prioritized only by the centrality, the units would travel excessively just to reposition themselves in central nodes without enough exploitation of calls in vicinity. Therefore, it is undesirable to use the centrality alone for dispatching decision and the centrality has to be combined with a measure that provides the capability of local exploitation. The closeness that is used in the greedy policy is an appropriate measure as it enables to pursue minimizing each current response time. Now, if calls are prioritized by centrality and closeness at the same time, units can be equipped with both global exploration capability and local exploitation capability.

Based on the background briefly described above, the centrality policy is formed in four steps as follows.

1) When an ambulance v gets freed, identify all unassigned calls U.
2) Compute centrality c_u of each call $u \in U$ upon the network of calls U with the edge between every pair of calls having a value of distance τ_{ui} (in time) between them
$$c_u = \sum_{i \in U, i \neq u} \frac{1}{(1 + \tau_{ui})}. \quad (1)$$
3) Compute fitness f_{vu} between the freed unit v and a call $u \in U$ based on the centrality c_u weighted by parameter α (≥ 0) and expected response time t_{vu} for the unit v to reach the call site u
$$f_{vu} = \frac{c_u^{\alpha}}{(1 + t_{vu})}. \quad (2)$$
4) Dispatch the freed unit v to the call u^* that maximizes the fitness
$$u^* = \arg\max_{u \in U} f_{vu}. \quad (3)$$

The centrality c_u in step 2 is represented by weighted degree among others due to its computational efficiency appropriate to the real-time decisions and due to its ability of producing robust performance in various operational scenarios (see [14] for details). The weighted degree of a node is computed by the sum of the weights of connected edges when higher weight values are preferred (e.g., capacity and strength) [30], [31]. However, the weight in the call network represents distance and lower weight values are preferred. The weighted degree in this case is computed by the sum of the reciprocals of weights.

A calibration parameter in step 3, weight on centrality α, is associated with the centrality policy. The centrality policy is exactly same as the greedy policy when $\alpha = 0$; however, when the weight is positive the policy incorporates centrality into decision by the extent corresponding to the weight. As indicated in [14], the weight value has to be carefully chosen according to the operating environment; even a small weight value gives significant benefits but the performance gets considerably degraded if the weight is too large. Lee [15] provides a heuristic approach for choosing the weight value as follows. One crucial characteristic of ambulance dispatching is the uncertainty involved in the need for transferring patient to hospital, as mentioned before. The choice of the weight

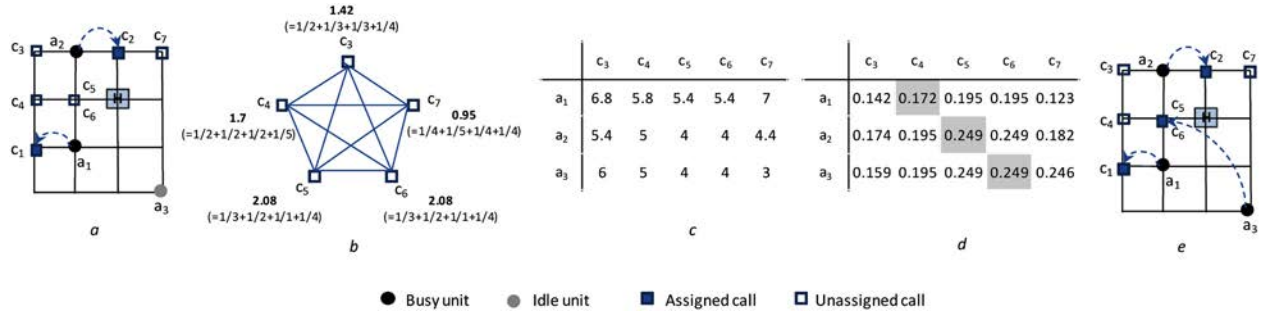

Fig. 2. Illustrative example of parallelized centrality policy. (a) a_3 gets freed (edge length = 1 m). (b) Centrality. (c) Expected response time (*hospital_prob* = 0.7). (d) Fitness (α = 0.3). (e) a_3 is dispatched to c_5 or c_6.

value will be affected by the probability of transferring as it determines the relevance of centrality consideration, i.e., the centrality will be more relevant when the probability is lower. The probability of transferring to hospital is denoted as *hospital_prob*. As the *hospital_prob* approaches zero, the centrality plays a more important role because of the higher possibility of continuing onsite services without transferring to hospital. On the other hand, if it approaches one, the role of centrality diminishes as the unit is more likely to go to hospital before serving next call. Therefore the usefulness of centrality is determined by the *hospital_prob* and thus 1− *hospital_prob* is used as the weight on centrality.

The centrality policy is designed specifically for ambulance-initiated dispatching decisions. However, for the sake of completeness of a policy and to facilitate integrating the parallelism principle, let us assume that the centrality policy adopts a greedy approach for call-initiated decisions, i.e., when a new call arrives and multiple units are available, dispatch the closest unit to the call. Therefore, two complete ambulance dispatching policies are now available: greedy policy (greedy in both call- and ambulance-initiated decisions) and centrality policy (greedy in call-initiated decisions and based on centrality in ambulance-initiated decisions).

III. Parallelism

The centrality policy significantly reduces response time by up to 86% over the greedy policy [14], [15]. The centrality policy, however, takes into account only idle units despite the possibility that a busy unit can respond more quickly even after the completion of currently assigned service. There is an opportunity here to further improve the centrality policy by incorporating both idle and busy units in parallel. The parallelized centrality policy is presented in five steps as follows, along with an illustrative example in Fig. 2(a) in which there are two busy $\{a_1, a_2\}$ and one idle $\{a_3\}$ units, two assigned $\{c_1, c_2\}$ and five unassigned $\{c_3, c_4, c_5, c_6, c_7\}$ calls, and a hospital. The unit a_3 just got freed and a decision needs to be made here on the selection of a call among unassigned calls (if appropriate).

1) When making either a call-initiated or ambulance-initiated decision, identify all unassigned calls U and all idle/busy units $V = V_{\text{idle}} \cup V_{\text{busy}}$.

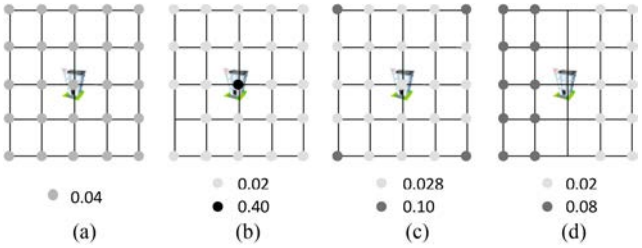

Fig. 3. Call distribution patterns. (a) Uniform. (b) Centered. (c) Cornered. (d) Bipartite.

Fig. 2(a): Unit a_3 gets freed. There are five calls waiting, $U = \{c_3, c_4, c_5, c_6, c_7\}$, and one idle and two busy units, $V_{\text{idle}} = \{a_3\}$, $V_{\text{busy}} = \{a_1, a_2\}$, $V = \{a_1, a_2, a_3\}$.

2) Compute centrality c_u of each call $u \in U$ upon the network of calls U with the edge between every pair of calls having a value of distance τ_{ui}(in time) between them

$$c_u = \sum_{i \in U, i \neq u} \frac{1}{(1 + \tau_{ui})}. \tag{4}$$

Fig. 2(b): The centrality of each call is computed upon the network of calls $U = \{c_3, c_4, c_5, c_6, c_7\}$, assuming every edge length of the grid in Fig. 2(a) is 1 m.

3) Compute fitness f_{vu} between a unit $v \in V$ and a call $u \in U$, based on the centrality c_u, weighted by parameter α ($\geq$0), and expected response time t_{vu} for the unit v to reach the call site u including, if any, the time expected to be spent on the already assigned call

$$f_{vu} = \frac{c_u^{\alpha}}{(1 + t_{vu})}. \tag{5}$$

Fig. 2(c) and (d): Assuming onsite service time = 1 m, *hospital_prob* = 0.7, and units are discharged immediately after arrival to hospital, the expected response time [Fig. 2(c)] and fitness [Fig. 2(d)] are computed for each pair between $V = \{a_1, a_2, a_3\}$ and $U = \{c_3, c_4, c_5, c_6, c_7\}$ with α=0.3.

4) Establish a one-to-one assignment by prioritizing the matches with larger fitness.

Fig. 2(d): First, the match a_2-c_5 is chosen (f_{25} is among the largest), second, a_3-c_6 (f_{36} is the largest after excluding matched units and calls, a_2 and c_5), and third, a_1-c_4

(f_{14} is the largest after excluding all matched ones, a_2, a_3, c_5, c_6), resulting in a one-to-one assignment.

5) Dispatch the idle units to their matched calls (if any) and leave unassigned the calls matched to busy units (without reservation).
   Fig. 2(e): The idle unit a_3 is dispatched to its matched call c_6, and other calls $\{c_3, c_4, c_5, c_6\}$ remain unassigned.

Step 1 considers all units (both idle and busy) and all unassigned calls when making a dispatching decision (either call-initiated or ambulance-initiated). The expected response time in Step 3 now includes the time expected to be spent on the currently assigned call (if any) as well, since busy units are also taken into consideration. The calculation of response time, therefore, requires a certain assumption on hospital selection, i.e., selecting one of hospitals available when transferring a patient to hospital (note that the parallelized centrality policy is operable along with any hospital selection policy as long as the expected response time is computable, and further details on hospital selection will be discussed in Section VI). The assignment problem in step iv is to make a one-to-one assignment that prioritizes matches with higher fitness score in order to pursue exploitation under uncertainty of future. The assignment solution has two types of matches. One is those between a busy unit and a call, and the other between an idle unit and a call. The matches of latter type are immediately executed; however, those calls matched to busy units remain unassigned with no reservation since better matches may arise in the future (see Step 5).

The execution of the parallelized centrality policy, in general, results in an assignment problem between multiple units and multiple calls. For example, suppose a new call arrives and multiple units are idle at the time, and the policy does not assign an idle unit to the call because a busy unit can respond to the call more quickly. Then, when another call arrives, two unassigned calls will be present at the same time whilst several idle units exist. Therefore, the boundary of call- and ambulance-initiated decisions becomes blurred. Also, note that keeping idle units idle in spite of the presence of waiting calls helps enhance preparedness by preventing units from being unnecessarily concentrated in a local region. However, if calls are dense in a certain area beyond its service capacity, idle units will be anyway dispatched to that area according to the assignment mechanism.

The parallelized centrality policy is exactly same as the centrality policy when $V = V_{\text{idle}}$, and it becomes the greedy policy if $V = V_{\text{idle}}$ and $\alpha = 0$. Also, let us call a special case of the parallelized centrality policy as the Parallelism Policy when only the parallelism is taken into account without centrality, i.e. when $V = V_{\text{idle}} \cup V_{\text{busy}}$ and $\alpha = 0$. The parallelism policy is the application of parallelism to the greedy policy whereas the parallelized centrality policy is the one to the centrality policy.

IV. Performance Evaluation

In this section, the parallelism is evaluated in various scenarios implemented in a discrete event simulator, by performance enhancement upon two policies: greedy policy and centrality policy.

A. Experiment Design

The service area is represented in a 5*5 square grid as shown in Fig. 3. Each vertex generates calls and ambulances move from vertex to vertex through edges. Once dispatched to a call site, the ambulance serves the patient with a service time (onsite time) that is exponentially distributed [32], [33]. The ambulance then, with a probability of *hospital_prob*, transfers the patient to a hospital located in the center of the grid, and the unit is discharged from the hospital after a certain period of time (turnaround time) that is exponentially distributed.

Four factors are associated with constructing different test conditions: 1) call distribution pattern; 2) size of ambulance fleet; 3) *hospital_prob*; and 4) time parameters. Total 12 500 calls are generated at a certain rate from an exponential distribution [13], [33], [34], and they are placed in the vertices according to one of four call distribution patterns in Fig. 3(a)–(d). A value in the figure represents the probability to allocate a call to a corresponding vertex. For example, in the cornered pattern, each vertex located in a corner gets an arriving call with probability 0.1. The four call patterns are designed to reflect various possible scenarios in reality. The size of ambulance fleet is in $\{2, 3, 4, 5\}$ (the parallelism is of no effect when the size is one; therefore this case is excluded), and the ambulances in each simulation run are initially located at random positions. The *hospital_prob* (i.e., probability of transferring patient to hospital), ranges in 0∼1 with an increment of 0.1. There are four time-related parameters, <call arrival interval, edge length, onsite time, turnaround time>, and two parameter sets are used, <1, 1, 0.5, 0> and <10, 5, 17, 40>, all expressed in minutes (including edge length). The first set is from the experimental setting used in [14] and the second set is from [15]. Especially the parameters of onsite time and turnaround time in the second set are from the statistics collected in several empirical analyses [35]–[37].

As a result, 352 test conditions (4 call distribution patterns * 4 sizes of ambulance fleet * 11 *hospital_prob* * 2 parameter sets) are established. For each test condition, four dispatching policies are applied: greedy policy, centrality policy, parallelism policy, and parallelized centrality (p-centrality) policy. The weight on centrality, α, used in centrality policy and parallelized centrality policy is set to 1-*hospital_prob* as it is suggested as a heuristic approach in [15]. Fifty simulation runs are replicated for each test scenario.

B. Performance Enhancement

Table I shows a summary of results in terms of average reduction in response time of a policy (B) over another policy (A) (represented in A: B), i.e., average reduction in response time of B over A = (average response time with A − average response time with B)/(average response time with A). As can be observed in the table, the parallelism reduces response time by up to 43.4% over greedy and centrality policies (maximums in Greedy: Parallelism and Centrality: P-Centrality). However, there are some cases where the parallelism has a

TABLE I
AVERAGE REDUCTION IN RESPONSE TIME

A:B	<1,1,0.5,0> Uniform	Centered	Cornered	Bipartite	<10,5,17,40> Uniform	Centered	Cornered	Bipartite	Total
Greedy : Parallelism	-20.5~34.2%	-0.8~27.8%	-1.2~43.4%	-2.5~34.7%	-1.2~15.7%	-4.1~10.6%	-2.2~21.4%	-3.6~16.5%	-20.5~43.4%
Centrality : P-Centrality	-2.2~34.0%	-0.1~27.6%	-15.5~43.2%	-4.3~34.9%	-8.1~15.7%	-2.1~11.6%	-6.4~20.9%	-17.4~16.3%	-17.4~43.2%
Greedy : Centrality	-0.5~89.6%	-1.7~69.7%	0.0~86.5%	-0.2~44.8%	-0.2~69.6%	-1.8~46.8%	-2.1~53.6%	-1.1~43.8%	-2.1~89.6%
Greedy : P-Centrality	-1.2~89.9%	-0.1~74.6%	-0.3~86.5%	-0.1~46.4%	-0.3~71.1%	-0.5~46.2%	-1.9~50.7%	-1.1~45.5%	-1.9~89.9%

Average reduction in response time of B over A = (average response time with A – average response time with B)/(average response time with A)

TABLE II
FREQUENCY OF REDUCTION IN AVERAGE RESPONSE TIME

	<1,1,0.5,0>												<10,5,17,40>												Total		
	Uniform			Centered			Cornered			Bipartite			Uniform			Centered			Cornered			Bipartite					
A:B	E	A	B	E	A	B	E	A	B	E	A	B	E	A	B	E	A	B	E	A	B	E	A	B	E	A	B
Greedy : Parallelism	18	1	25	11	0	33	21	0	23	24	0	20	35	0	9	37	0	7	37	0	7	38	0	6	221 (62.8%)	1 (0.3%)	130 (36.9%)
Centrality : P-Centrality	23	0	21	14	0	30	25	1	18	25	0	19	35	1	8	37	0	7	37	1	6	36	1	7	232 (65.9%)	4 (1.1%)	116 (33.0%)
Greedy : Centrality	24	0	20	35	0	9	21	0	23	27	0	17	32	0	12	35	0	9	33	0	11	34	0	10	241 (68.5%)	0 (0.0%)	111 (31.5%)
Greedy : P-Centrality	11	0	33	7	0	37	13	0	31	14	0	30	24	0	20	29	0	15	27	0	17	28	0	16	153 (43.5%)	0 (0.0%)	199 (56.5%)

E: A and B are equivalent. A: A is better (lower in average response time) than B. B: B is better (lower in average response time) than A

negative effect, increasing response time by 20.5% in the worst case (minimums in Greedy: Parallelism and Centrality: P-Centrality). To examine the frequency of this negative effect, a statistical testing (*t*-test) is conducted for the significance in the difference of two policies (say A and B) with significance level set at 0.05. Table II presents the frequencies of being equivalent (E), having A better (A), and having B better (B). As can be noticed, the effect of parallelism is negative only in 5 out of 704 cases (<1%) (Greedy: Parallelism and Centrality: P-Centrality), while being positive in 246 cases (35%), well supporting the potential of parallelism in reducing response time.

When comparing Greedy: Centrality and Greedy: P-Centrality in Table I, the effect of parallelism does not seem eminent (due to similar ranges, $-2.1\sim89.6\%$ and $-1.9\sim89.9\%$), though its benefits look considerable in Table II (increasing positive cases from 31.5% to 56.5%). This is due to the fact that the parallelism has a characteristic of effecting especially low-load conditions where centrality is not very useful. Fig. 4 shows the average reduction in response time of three policies (centrality, parallelism, and p-centrality) over the greedy policy, from the test condition with uniform call pattern and parameter set $<1,\ 1,\ 0.5,\ 0>$ (Note that the overall pattern is similar in other conditions as well). In high-load conditions (small size of fleet and high *hospital_prob*), calls are queued everywhere and idle units are likely to be assigned to the calls in vicinity regardless of whether busy units are considered or not. This characteristic of the parallelism (being more effective in low-load conditions) complements the centrality that does not have significant impact in low-load conditions (large size of fleet and low *hospital_prob*). The centrality alone is not very useful in such low-load conditions since most decisions are call-initiated. Therefore, once centrality is synthesized with parallelism, the performance gets enhanced in both low and high load conditions.

C. Performance Variation

Coverage level is another measure that can be used in evaluating dispatching policies, where coverage level corresponds to the percentage of calls responded to within a given response time threshold. However, various definitions exist for the response time threshold depending on factors such as country, urbanization, and urgency, ranging in 7–30 m [9], [38]–[43]. Also, the coverage level would make no difference in patient outcomes among scenarios within or beyond a threshold, though significant differences exist [44]. Therefore, rather than evaluating according to a specific threshold, the policies can be evaluated by variation of response time in conjunction with average performance. Average standard deviation is used to measure the variation of response time. Table III shows a summary of the reduction in variation, Table IV presents the frequency of reduction in variation by a statistical testing (F-test) with significance level set at 0.05, and Fig. 5 shows the reduction in variation of the policies over the greedy policy, similarly to the previous section but on variation data. The overall pattern in variation is similar to the one in average response time, i.e., the less the average response time is, the less the variation is, supporting the significance of parallelism even in reducing the variation.

The reduction in both average and variation implies that the policy equipped with parallelism is likely to have a

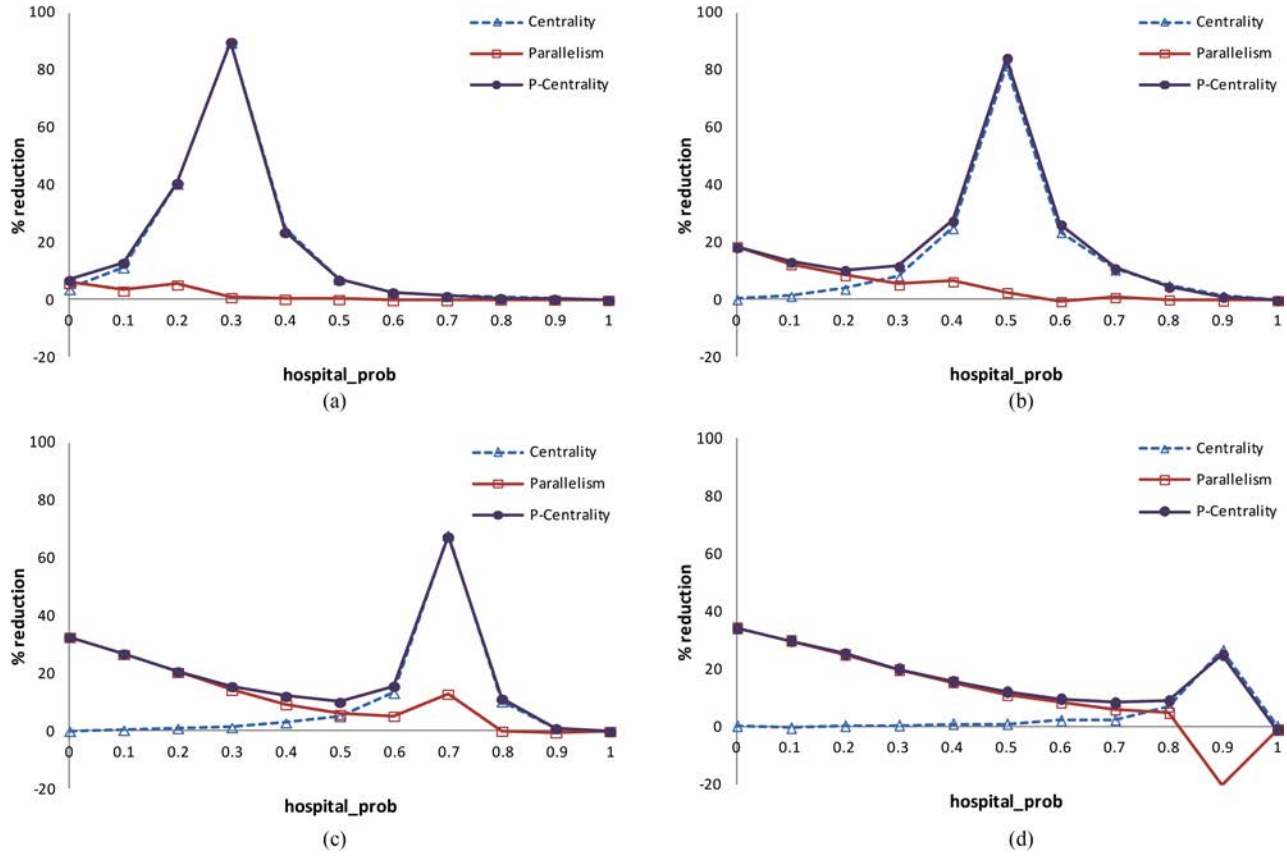

Fig. 4. Average reduction in response time over greedy policy—uniform call pattern, parameter set <1, 1, 0.5, 0>. (a) 2 units. (b) 3 units. (c) 4 units. (d) 5 units.

TABLE III
REDUCTION IN VARIATION

A:B	<1,1,0.5,0> Uniform	Centered	Cornered	Bipartite	<10,5,17,40> Uniform	Centered	Cornered	Bipartite	Total
Greedy : Parallelism	-23.2~36.6%	-0.3~27.1%	-0.4~39.2%	-2.6~37.2%	-3.4~15.1%	-4.9~10.0%	-4.6~18.0%	-2.4~15.7%	-23.2~39.2%
Centrality : P-Centrality	-3.3~36.6%	-1.8~26.8%	-15.9~39.2%	-3.6~37.5%	-11.4~14.5%	-2.1~11.4%	-6.2~17.8%	-16.2~15.8%	-16.2~39.2%
Greedy : Centrality	-0.2~93.8%	-3.3~78.1%	0.0~94.0%	-0.6~64.3%	0.0~86.0%	-3.0~70.5%	0.0~82.9%	0.0~68.2%	-3.3~94.0%
Greedy : P-Centrality	-0.9~93.9%	-0.2~81.6%	-0.1~94.0%	-0.4~64.9%	0.0~86.4%	0.0~69.9%	0.0~81.9%	-0.1~69.4%	-0.9~94.0%

Reduction in variation of B over A = (variation with A – variation with B) / (variation with A)

better coverage level for any response time threshold and mitigate the risk of exposing patients to excessively tardy responses. However, there are several instances in which the parallelism has a negative effect (increasing response time and/or variation). This means that the way of applying the parallelism needs to be more robust to different operational conditions, and the next section introduces a rule for activating parallelism by adding a new calibration parameter within the policy.

V. ACTIVATION RULE OF PARALLELISM

There could be situations where the parallelism makes performance even worse, in which cases it is better to deactivate it. This section provides an activation rule of the parallelism and incorporates it into the parallelized centrality policy by adding a new calibration parameter β. The parallelism parameter β serves as a threshold of activating the parallelism in terms of the number of unassigned calls. The short-term nature of the assignment process ignoring multihop routes could lead to solutions far from being optimal especially when the uncertainty of dynamics gets increased due to the presence of a large number of calls and involvement of busy units. The parallelism therefore would be better to be deactivated when there are unassigned calls over a certain limit and in such a case only idle units are considered.

The parallelized centrality policy presented in Section III is transformed into a policy that has two calibration parameters (centrality parameter α and parallelism parameter β) as follows.

TABLE IV
FREQUENCY OF REDUCTION IN VARIATION

A:B	<1,1,0.5,0> Uniform E	A	B	Centered E	A	B	Cornered E	A	B	Bipartite E	A	B	<10,5,17,40> Uniform E	A	B	Centered E	A	B	Cornered E	A	B	Bipartite E	A	B	Total E	A	B
Greedy : Parallelism	33	1	10	32	0	12	33	0	11	34	0	10	42	0	2	44	0	0	43	0	1	43	0	1	304 (86.4%)	1 (0.3%)	47 (13.4%)
Centrality : P-Centrality	35	0	9	34	0	10	37	1	6	37	0	7	43	0	1	44	0	0	43	0	1	42	1	1	315 (89.5%)	2 (0.6%)	35 (9.9%)
Greedy : Centrality	29	0	15	34	0	10	19	0	25	31	0	13	26	0	18	31	0	13	18	0	26	25	0	19	213 (60.5%)	0 (0.0%)	139 (39.5%)
Greedy : P-Centrality	14	0	30	25	0	19	9	0	35	18	0	26	24	0	20	31	0	13	17	0	27	24	0	20	162 (46.0%)	0 (0.0%)	190 (54.0%)

E: A and B are equivalent. A: A is better (lower in variation) than B. B: B is better (lower in variation) than A

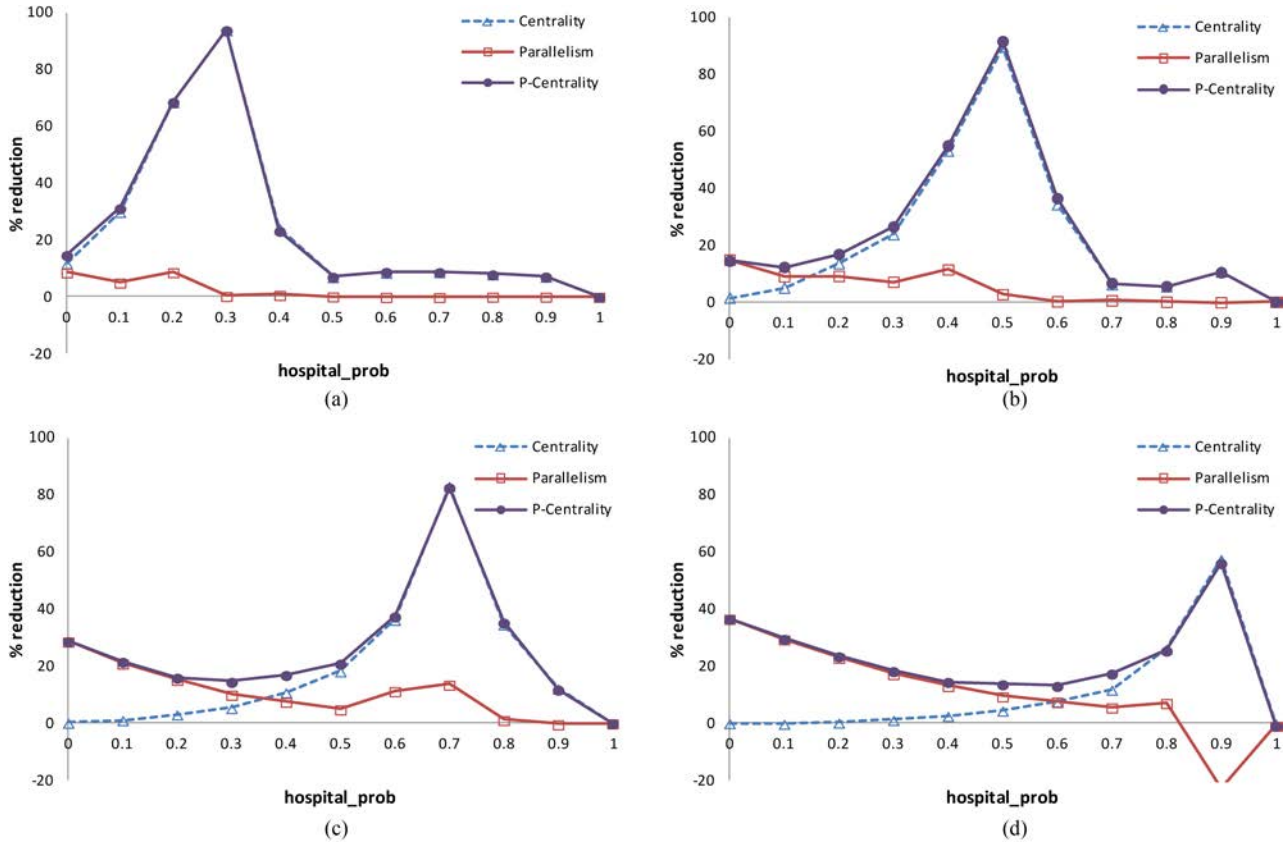

Fig. 5. Reduction in variation over greedy policy—uniform call pattern, parameter set < 1, 1, 0.5, 0 >. (a) 2 units. (b) 3 units. (c) 4 units. (d) 5 units.

1) When making either a call-initiated or ambulance-initiated decision, identify all unassigned calls U and, if $|U| \leq \beta$, all idle/busy units $V = V_{\text{idle}} \cup V_{\text{busy}}$, otherwise only idle units $V = V_{\text{idle}}$.
2) Compute centrality c_u of each call $u \in U$ upon the network of calls U with the edge between every pair of calls having a value of distance τ_{ui} (in time) between them

$$c_u = \sum_{i \in U, i \neq u} \frac{1}{(1 + \tau_{ui})}. \tag{6}$$

3) Compute fitness f_{vu} between a unit $v \in V$ and a call $u \in U$ based on the centrality c_u weighted by parameter α (≥ 0) and expected response time t_{vu} for the unit v to reach the call site u including, if any, the time expected to be spent on the already assigned call

$$f_{vu} = \frac{c_u^{\alpha}}{(1 + t_{vu})}. \tag{7}$$

4) Establish an one-to-one assignment by prioritizing the matches with larger fitness.
5) Dispatch the idle units to their matched calls (if any) and leave unassigned the calls matched to busy units (without reservation).

The only change made is in Step 1 where the parallelism parameter β is applied as activation threshold. Two calibration parameters (α and β) are now associated with the parallelized

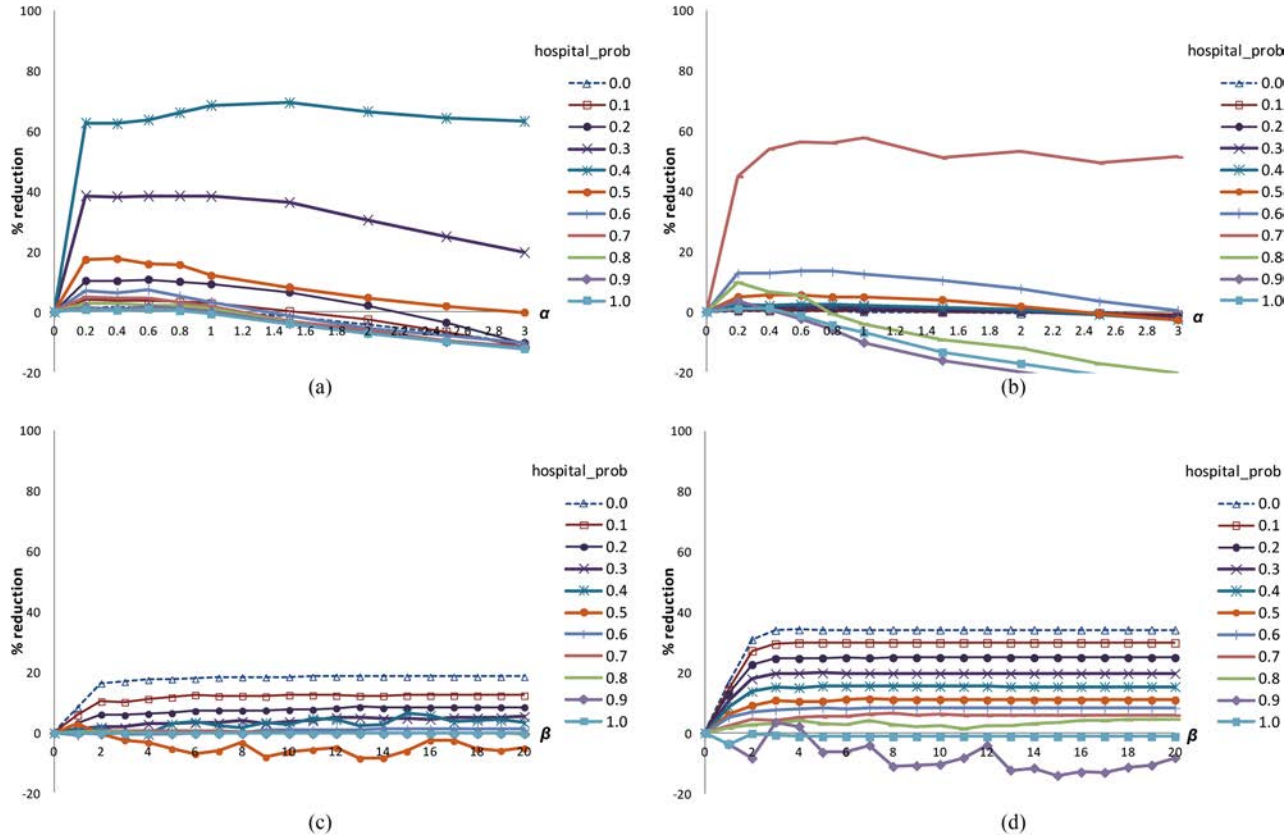

Fig. 6. Effect of calibration parameters α and β—uniform call pattern, parameter set $<1, 1, 0.5, 0>$. (a) 3 units $\beta = 0$. (b) 5 units $\beta = 0$. (c) 3 units $\alpha = 0$. (d) 5 units $\alpha = 0$.

centrality policy, and all other three policies (greedy, centrality, and parallelism) can be instantiated from this policy by adjusting these parameters: $\alpha = 0$ and $\beta = 0 \rightarrow$ greedy policy, $\beta = 0 \rightarrow$ centrality policy, and $\alpha = 0$ and $\beta = \infty \rightarrow$ parallelism policy. The calibration parameters are closely related to the performance of the policy. For example, Fig. 6 shows that the effect of centrality parameter α (by setting $\beta = 0$) and parallelism parameter β (by fixing $\alpha = 0$) in terms of average reduction in response time over greedy policy, in uniform call pattern and with parameter set $<1, 1, 0.5, 0>$. The improvement by centrality [Fig. 6(a) and (b)] is significant even with a small weight, and after reaching the peak the improvement keeps going down toward negative improvement (i.e., increase in response time). As discussed in Section II, if centrality is too much pursued, the units will travel excessively just for repositioning purpose without enough exploitation of calls in vicinity. The performance improvement by parallelism [Fig. 6(c) and (d)] tends to be larger as the parameter β increases (in other words, as the parallelism is more incorporated). In some cases, however, highly irregular patterns can be observed with sometimes falling even into negative region.

The nonlinear behaviors with the two calibration parameters give rise to the need for carefully choosing the right values according to the operating environment, in order to maximize the benefits of centrality and parallelism principles. The right choice of the parameter values would be affected by various factors such as information uncertainty, probability of transferring to hospital, size of service area, size of fleet, call distribution pattern, onsite service time, turnaround time, and so on.

VI. Conclusion and Future Work

A notion of parallelism is developed for ambulance dispatching decisions that allow considering both idle and busy units in parallel rather than just idle ones. The parallelism, applied upon the centrality policy, results in the parallelized centrality policy complementing and enhancing the centrality policy in average as well as variation of response time. The parallelized centrality policy involves two calibration parameters, centrality parameter and parallelism parameter, which enable to achieve portable performance adaptively to various operational scenarios. In order to maximize the benefits of the policy, the parameter values have to be carefully chosen according to the characteristics of the operating environment factoring in such as information uncertainty, probability of transferring to hospital, size of service area, size of fleet, and call distribution pattern.

The calibration process of identifying the best parameter values, however, is quite laborious and computationally demanding. Moreover, the parameters, in general, strongly depend on each problem instance and thus the calibration process has to be repeated for each problem instance. Therefore, optimal parameter matching rules need to be explored through theoretical and experimental analyses that relate the characteristics of operational scenarios to the best parameter

values, so that users can be guided by the rules in choosing appropriate parameter values for the scenarios of interest. The parameter matching process would require utilizing statistical analysis and data mining techniques that have been proven useful, referring to the research on parameter calibration [45]–[51].

The parallelism can be adapted and applied to the priority dispatching systems which prioritize ambulance calls in accordance with their degree of urgency. In case the parallelism is directly applied to such systems, an idle unit may keep being idle in spite of the presence of high-priority patients (with life-threatening symptoms), which would be politically impossible or subject to lawsuit. One way of overcoming such problematic situations is to first apply the parallelism policy or parallelized centrality policy among unassigned high-priority calls and idle units only (excluding busy units), and then apply the policy to the calls with lower priorities involving both idle and busy units. The high-priority calls then will be always assigned an idle ambulance (if any). Various approaches of adapting the parallelism principle can be devised, and an extensive research is required to evaluate them with respect to EMS regulations and performance.

Another interesting research topic is on hospital selection which is also necessary in calculating expected response time of busy units in the parallelized centrality policy. The hospital selection decision determines an appropriate hospital when transferring a patient to hospital, and it is closely associated with logistics efficiency because it influences the availability of ambulances to other patients. An ambulance becomes unavailable during the transfer time consisting of transportation time (from the scene to a hospital) and turnaround time (interval between arrival at the hospital and the time the ambulance becomes available to respond to another call). The majority of patients are, in reality, transferred to the nearest hospital [52], [53]. However, crowding in the ED has a direct impact on the transfer time due to the delay caused by the lack of resources (space, bed, personnel, and so on) [54]. One popular technique used to avoid crowding is the ambulance diversion that incoming ambulances are redirected to nearby, less crowded EDs. Approximately 500 000 ambulances are diverted annually in the U.S. [55]. Although ambulance diversion can reduce crowding, it can increase the transfer time of the patients being diverted and can reduce the availability of ambulances [56]. Sprivulis and Gerrard [57] and Larson [58] describe the use of real-time information on ED status (occupied spaces, emergency inpatients, waiting room patients, and so on) which helps EMS crew make more informed decisions with significant decreases in diversion hours and a more balanced workload between hospitals.

Advanced hospital selection policies can be composed based on various factors discussed above. It would be also meaningful to investigate the interactions of potential hospital selection policies with the parallelism. For example, even though a hospital selection policy is better than another without parallelism (i.e. considering only idle units), it can result in worse performance with parallelism (i.e., considering both idle and busy units), when the projected response times of busy units are subject to large variability and/or inaccuracy.

REFERENCES

[1] Center for Disease Control and Prevention. (2003–2010). *National Hospital Ambulatory Medical Care Survey: Emergency Department Survey* [Online]. Available: http://www.cdc.gov/nchs/ahcd/web_tables.htm.

[2] Heart Rhythm Foundation. (2013). *Sudden Cardiac Arrest Facts* [Online]. Available: http://www.hrsonline.org/News/Fact-Sheets/SCA-Facts#axzz2ln1VfDKe.

[3] C. O'Keeffe, J. Nicholl, J. Turner, and S. Goodacre, "Role of ambulance response times in the survival of patients with out-of-hospital cardiac arrest," *Emergency Med. J.*, vol. 28, no. 8, pp. 703–706, 2011.

[4] C. S. ReVelle and K. Hogan, "The maximum availability location problem," *Transport. Sci.*, vol. 23, no. 3, pp. 192–200, 1989.

[5] L. Brotcorne, G. Laporte, and F. Semet, "Ambulance location and relocation models," *Eur. J. Oper. Res.*, vol. 147, no. 3, pp. 451–463, 2003.

[6] J. B. Goldberg, "Operations research models for the deployment of emergency services vehicles," *EMS Manage. J.*, vol. 1, no. 1, pp. 20–39, 2004.

[7] H. Jia, F. Ordonez, and M. Dessouky "A modeling framework for facility location of medical services for large-scale emergencies," *IIE Trans.*, vol. 39, no. 1, pp. 41–55, 2007.

[8] P. Kolesar and W. Walker, "An algorithm for the dynamic relocation of fire companies," *Oper. Res.*, vol. 22, no. 2, pp. 249–274, 1974.

[9] M. Gendreau, G. Laporte, and F. Semet, "A dynamic model and parallel tabu search heuristic for real-time ambulance relocation," *Parallel Comput.*, vol. 27, no. 12, pp. 1641–1653, 2001.

[10] L. V. Green and P. J. Kolesar, "Improving emergency responsiveness with management science," *Manage. Sci.*, vol. 50, no. 8, pp. 1001–1014, 2004.

[11] M. Gendreau, G. Laporte, and F. Semet, "The maximal expected coverage relocation problem for emergency vehicles," *J. Oper. Res. Soc.*, vol. 57, no. 1, pp. 22–28, 2006.

[12] R. Nair and E. Miller-Hooks, "Evaluation of relocation strategies for emergency medical service vehicles," *Transport. Res. Rec.: J. Transport. Res. Board*, vol. 2137, pp. 63–73, 2009.

[13] R. Alanis, A. Ingolfsson, and B. Kolfal, "A Markov chain model for an EMS system with repositioning," *Prod. Oper. Manage.*, vol. 22, no. 1, pp. 216–231, 2013.

[14] S. Lee, "The role of centrality in ambulance dispatching," *Decision Support Syst.*, vol. 54, no. 1, pp. 282–291, 2012.

[15] S. Lee, "Centrality-based ambulance dispatching for demanding emergency situations," *J. Oper. Res. Soc.*, vol. 64, no. 4, pp. 611–618, 2013.

[16] J. M. Chaiken and R. C. Larson, "Methods for allocating urban emergency units: A survey," *Manage. Sci.*, vol. 19, no. 3, pp. 110–130, 1972.

[17] J. Hayes, A. Moore, G. Benwell, and B. Wong, "Ambulance dispatch complexity and dispatcher decision strategies: Implications for interface design," *Computer Human Interaction (LNCS)*, vol. 3101. Berlin, Germany: Springer, 2004, pp. 589–593.

[18] S. F. Dean, "Why the closest ambulance cannot be dispatched in an urban emergency medical services system," *Prehospital Disaster Med.*, vol. 23, no. 2, pp. 161–165, 2008.

[19] P. J. Egbelu and J. M. A. Tanchoco, "Characterization of automatic guided vehicle dispatching rules," *Int. J. Prod. Res.*, vol. 22, no. 3, pp. 359–374, 1984.

[20] D. J. Bertsimas and G. van Ryzin, "A stochastic and dynamic vehicle routing problem in the Euclidean plane," *Oper. Res.*, vol. 39, no. 4, pp. 601–615, 1991.

[21] R. J. Mantel and H. R. A. Landeweerd, "Design and operational control of an AGV system," *Int. J. Prod. Econom.*, vol. 41, nos. 1–3, pp. 257–266, 1995.

[22] E. Østergaard, M. J. Matarić, and G. S. Sukhatme, "Distributed multi-robot task allocation for emergency handling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2001, pp. 821–826.

[23] B. P. Gerkey and M. J. Matarić, "Sold!: Auction methods for multi-robot coordination," *IEEE Trans. Robot. Autom.*, vol. 18, no. 2, pp. 758–768, Oct. 2002.

[24] M. B. M. de Koster, T. Le-Anh, and J. R. van der Meer, "Testing and classifying vehicle dispatching rules in three real-world settings," *J. Oper. Manage.*, vol. 22, no. 4, pp. 369–386, 2004.

[25] M. B. Dias, "TraderBots: A new paradigm for robust and efficient multirobot coordination in dynamic environments," Ph.D. dissertation, Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2004.

[26] P. B. Sujit and R. Beard, "Distributed sequential auctions for multiple UAV task allocation," in *Proc. Amer. Contr. Conf.*, 2007, pp. 3955–3960.

[27] E. A. Blackstone, A. J. Buck, and S. Hakim, "The economics of emergency response," *Policy Sci.*, vol. 40, no. 4, pp. 313–334, 2007.

[28] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.
[29] M. Barthelemy, "Betweenness centrality in large complex networks," *Eur. Phys. J. B*, vol. 38, no. 2, pp. 163–168, 2004.
[30] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 11, pp. 3747–3752, 2004.
[31] M. E. J. Newman, "Analysis of weighted networks," *Phys. Rev. E*, vol. 70, no. 5, 056131, 2004.
[32] E. Erkut, A. Ingolfsson, and G. Erdoğan, "Ambulance location for maximum survival," *Naval Res. Logist.*, vo. 55, no. 1, pp. 42–58, 2008.
[33] L. A. McLay and M. Mayorga, "A dispatching model for server-to-customer systems that balances efficiency and equity," *Manuf. Serv. Oper. Manage.*, vol. 15, no. 2, pp. 205–220, 2013.
[34] M. Singer and P. Donoso, "Assessing an ambulance service with queuing theory," *Comput. Oper. Res.*, vol. 35, no. 8, pp. 2549–2560, 2008.
[35] B. G. Carr, J. M. Caplan, J. P. Pryor, and C. C. Branas, "A meta-analysis of prehospital care times for trauma," *Prehospital Emergency Care*, vol. 10, no. 2, pp. 198–206, 2006.
[36] S. Budge, A. Ingolfsson, and D. Zerom, "Empirical analysis of ambulance travel times: The case of calgary emergency medical services," *Manage. Sci.*, vol. 56, no. 4, pp. 716–723, 2010.
[37] S. Vandeventer, J. R. Studnek, J. S. Garrett, S. R. Ward, K. Staley, and T. Blackwell, "The association between ambulance hospital turnaround times and patient acuity, destination hospital, and time of day," *Prehospital Emergency Care*, vol. 15, no. 3, pp. 366–370, 2011.
[38] M. O. Ball and L. F. Lin, "A reliability model applied to emergency service vehicle location," *Oper. Res.*, vol. 41, no. 1, pp. 18–36, 1993.
[39] J. Holloway, G. Francis, and M. Hinton, "A vehicle for change? a case study of performance improvement in the 'new' public sector," *Int. J. Public Sector Manage.*, vol. 12, no. 4, pp. 351–365, 1999.
[40] K. McGrath, "The Golden Circle: A way of arguing and acting about technology in the London ambulance service," *Eur. J. Inf. Syst.*, vol. 11, no. 4, pp. 251–266, 2002.
[41] P. T. Pons and V. J. Markovchick, "Eight minutes or less: Does the ambulance response time guideline impact trauma patient outcome?" *J. Emergency Med.*, vol. 23, no. 1, pp. 43–48, 2002.
[42] M. Woollard, D. Lewis, and S. Brooks, "Strategic change in the ambulance service: Barriers and success strategies for the implementation of high-performance management systems," *Strategic Change*, vol. 12, no. 3, pp. 165–175, 2003.
[43] J. J. M. Black and G. D. Davies, "International EMS systems: United Kingdom," *Resuscitation*, vol. 64, no. 1, pp. 21–29, 2005.
[44] L. A. McLay and M. E. Mayorga, "Evaluating the impact of performance goals on dispatching decisions in emergency medical service," *IIE Trans. Healthcare Syst. Eng.*, vol. 1, no. 3, pp. 185–196, 2011.
[45] J. Grefenstette, "Optimization of control parameters for genetic algorithms," *IEEE Trans. Syst., Man, Cybern.*, vol. 16, no. 1, pp. 122–128, Jan. 1986.
[46] A. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms," *IEEE Trans. Evol. Comput.*, vol. 3, no. 2, pp. 124–141, Jul. 1999.
[47] O. Francois and C. Lavergne, "Design of evolutionary algorithms—A statistical perspective," *IEEE Trans. Evol. Comput.*, vol. 5, no. 2, pp. 129–148, Apr. 2001.
[48] A. Czarn, C. MacNish, K. Vijayan, B. Turlach, and R. Gupta, "Statistical exploratory analysis of genetic algorithms," *IEEE Trans. Evol. Comput.*, vol. 8, no. 4, pp. 405–421, Aug. 2004.
[49] T. Bartz-Beielstein, *Experimental Research in Evolutionary Computation: The New Experimentalism*. Berlin, Germany: Springer, 2006.
[50] W. A. de Landgraaf, A. E. Eiben, and V. Nannen, "Parameter calibration using meta-algorithms," in *Proc. IEEE Congr. Evol. Comput.*, 2007, pp. 71–78.
[51] T. Bartz-Beielstein, M. Chiarandini, L. Paquete, and M. Preuss, *Empirical Methods for the Analysis of Optimization Algorithms*. Berlin, Germany: Springer, 2009.
[52] E. Auf der Heide, "The importance of evidence-based disaster planning," *Ann. Emergency Med.*, vol. 47, no. 1, pp. 34–49, 2006.
[53] B. T. Squire, A. Tamayo, and J. H. Tamayo-Sarver, "At-risk populations and the critically ill rely disproportionately on ambulance transport to emergency departments," *Ann. Emergency Med.*, vol. 56, no. 4, pp. 341–347, 2010.
[54] M. Eckstein, S. M. Isaacs, C. M. Slovis, B. J. Kaufman, J. R. Loin, *et al.*, "Facilitating EMS turnaround intervals at hospitals in the face of receiving facility overcrowding," *Prehospital Emergency Care*, vol. 9, no. 3, pp. 267–275, 2005.
[55] C. W. Burt, L. F. McCaig, and R. H. Valverde, "Analysis of ambulance transports and diversions among US emergency departments," *Ann. Emergency Med.*, vol. 47, no. 4, pp. 317–326, 2006.
[56] A. J. E. Carter and R. Grierson, "The impact of ambulance diversion on EMS resource availability," *Prehospital Emergency Care*, vol. 11, no. 4, pp. 421–426, 2007.
[57] P. Sprivulis and B. Gerrard, "Internet-accessible emergency department workload information reduces ambulance diversion," *Prehospital Emergency Care*, vol. 9, no. 3, pp. 285–291, 2005.
[58] G. Larson, "Ambulance destination determination system for ambulance distribution as an alternative to ambulance diversion," *J. Emergency Nurs.*, vol. 34, no. 4, pp. 357–358, 2008.

Seokcheon Lee received the B.S. and M.S. degrees in industrial engineering from Seoul National University, Seoul, Korea, in 1991 and 1993, respectively, and the Ph.D. degree in industrial engineering from Pennsylvania State University, University Park, PA, USA, in 2005.

He is currently an Associate Professor with the School of Industrial Engineering, Purdue University, West Lafayette, IN, USA. His current research interests include distributed control of large-scale complex network systems (e.g., supply network coordination, dynamic resource allocation, and emergency logistics), based on the fundamental decision principles designed from multidisciplinary perspectives.