Proceedings of the Twenty-Sixth International Conference on
Automated Planning and Scheduling (ICAPS 2016)

Strategic Planning for Setting Up Base Stations in Emergency Medical Systems

Supriyo Ghosh
School of Information Systems
Singapore Management University
supriyog.2013@phdis.smu.edu.sg

Pradeep Varakantham
School of Information Systems
Singapore Management University
pradeepv@smu.edu.sg

Abstract

Emergency Medical Systems (EMSs) are an important component of public health-care services. Improving infrastructure for EMS and specifically the construction of base stations at the "right" locations to reduce response times is the main focus of this paper. This is a computationally challenging task because of the: (a) exponentially large action space arising from having to consider combinations of potential base locations, which themselves can be significant; and (b) direct impact on the performance of the ambulance allocation problem, where we decide allocation of ambulances to bases. We present an incremental greedy approach to discover the placement of bases that maximises the service level of EMS. Using the properties of submodular optimisation we show that our greedy algorithm provides quality guaranteed solutions for one of the objectives employed in real EMSs. Furthermore, we validate our derived policy by employing a real-life event driven simulator that incorporates the real dynamics of EMS. Finally, we show the utility of our approaches on a real-world dataset from a large asian city and demonstrate significant improvement over the best known approaches from literature.

Introduction

Emergency Medical Systems (EMSs) are an integral part of public health-care services. A typical EMS employs a set of Emergency Response Vehicles, ERVs (ex: ambulances, fire rescue vehicles) that provide timely care to patients (with injuries or illnesses) who seek immediate attention. In an EMS, a set of base stations are strategically placed throughout the city and a fixed number of ERVs are allocated to each base. On arrival of an emergency request, an ambulance from the nearest base is dispatched to assist the victim. The ambulance returns back to the same base after transferring the patient to a nearby hospital.

In order to sustain and maintain the efficiency of an EMS, there are typically two levels of decision making: (a) Operational Level, i.e., day-to-day decisions associated with ambulance dispatching and allocation policy; and (b) Strategic Level, i.e., long-term decisions on number of ambulances, number of bases and locations of bases.

Most papers (Yue, Marla, and Krishnan 2012; Saisubramanian, Varakantham, and Chuin 2015; Andersson and Värbrand 2007; Bjarnason et al. 2009) in improving EMSs have focussed on operational level decision making and develop strategies to improve the system efficiency by optimising performance metrics such as bounded time response (ex: percentage of requests served within 15 minutes) and bounded risk response (ex: least response time within which 80% of the requests are assisted). Although facility location problems in large-scale disaster response systems for rare and catastrophic events (ex: earthquake and hurricane) enjoy a rich history (Toregas et al. 1971; Church and Velle 1974; Jia, Ordóñez, and Dessouky 2007; Huang, Kim, and Menezes 2010), progress remains slow for strategic planning in EMSs. Unlike decision making in large scale disaster response systems for rare and catastrophic events, we are focussed on strategic level planning for EMSs where incidents happen everyday and the patterns of how incidents happen change over time.

Specifically, we are interested in the problem of setting up new bases (how many and where?). It is an extension of k-center facility location problem which is a well known NP-Hard problem (Hochbaum and Shmoys 1985). Given the exponentially large space of possibilities (subsets of potential base stations that can be built in a given budget) and the direct dependence of the selected base set on optimal allocation of ambulances to bases, this is a computationally challenging problem. Furthermore, the budget for resources (ex: expense for setting up new bases or funds for new ambulances) is dynamic and arrives over time in different chunks and thus makes it difficult to plan all base locations well in advance.

Towards addressing the above mentioned challenges, our key contributions are as follows:

- We provide an incremental greedy algorithm where bases are added as long as the marginal gain is significant. We also show that for one of the objectives typically employed in EMS, the optimisation function is monotone submodular, there by guaranteeing atleast 63% of optimal performance.
- We present an accelerated version of the greedy algorithm, referred to as lazy greedy and show that it can be utilised to optimise widely used performance objec-

tives, namely bounded time response and bounded risk response.

- We employ a real-life event driven simulator to evaluate the performance of our approaches in comparison with existing benchmark approaches.

Extensive empirical results on real-world dataset from a large asian city demonstrate that our techniques (that utilise a significantly smaller number of bases) either outperform or provide highly competitive results in comparison with the best known approaches from literature.

Related Work

Given the practical importance, a wide range of disciplines have studied problems associated with EMSs. We focus on three relevant threads of research in this paper. The first thread of papers focus on improving operational strategies for EMS. (Andersson and Värbrand 2007; Schmid 2012) develop techniques to optimally generate dispatching policy for ambulances and also provide a relocation model that dynamically suggests a destination base for ambulances after job completion. However due to inherent complexity of the process (such as congestion of ambulances at certain bases or problem with conceiving the critically of request by the operator), many EMSs prefer a fixed allocation of ambulances and follow the nearest ambulance dispatch policy. (Brotcorne, Laporte, and Semet 2003; Gendreau, Laporte, and Semet 2006) exploit mathematical models by incorporating performance metrics as a parameter of the model and provide optimisation or local search based heuristics to solve the allocation problem. (Maxwell et al. 2010) focus on optimal allocation and dynamic redeployment model for single ambulance. But optimisation models often fail to capture the dynamics of EMS such as congestion pattern in road or response time from base to scene that varies over time. Recent works (Saisubramanian, Varakantham, and Chuin 2015; Yue, Marla, and Krishnan 2012; Restrepo, Henderson, and Topaloglu 2009) overcome these caveats by employing a real-life event-driven simulator to evaluate the resulting policy. All the papers in this thread presume a fixed set of bases, while we consider the ambulance allocation problem in conjunction with discovering optimal placement for bases.

The second thread of research focuses on strategic planning for rare and large-scale disaster response (ex: fire, vehicle accident or natural disaster). The traditional model for facility location in large scale disaster response is based on the covering problem such as location set covering problem [LSCP] (Toregas et al. 1971), that aims to provide coverage to all the demand points; and maximal covering location problem [MCLP] (Church and Velle 1974), that maximises the coverage for a given a budget. P-median (Hakimi 1964) (minimises average distance between demand point and nearest facility) and P-center (Sylvester 1857) (minimises the worse case response time) models are also widely adopted in literature. Recently, (Jia, Ordóñez, and Dessouky 2007; Huang, Kim, and Menezes 2010) propose mathematical model for large-scale disaster response and solve it using optimisation method or dynamic programming. Due to the rare occurrence of the catastrophic events, these papers are focussed on robust objectives that plan for the absolute worst case. In contrast, incidents in EMSs happen every day and objectives consider softer notions of robust decision making (ex: maximise number of requests served within 15 minutes, minimise time taken to serve 80% of requests). We take a data-driven approach to find the minimal set of bases in EMS and evaluate the performance of solution on a diverse set of demand scenarios.

The last thread of research which is complimentary to this work is on optimisation of monotone submodular functions (Leskovec et al. 2007; Nemhauser, Wolsey, and Fisher 1978). Some popular application domains are: dynamic conservation planning (Golovin et al. 2011), maximising information gain in sensor placement (Krause, Singh, and Guestrin 2008) and content recommendation (Yue and Guestrin 2011). The key reason behind this extensive adoption is that a greedy approach provides $(1 - \frac{1}{e})$ approximation guarantee in case of monotone submodular functions.

Ambulance Allocation Problem

Ambulance allocation problem can be formally defined using following tuple:

$$< \mathcal{R}, \mathcal{B}, \mathcal{A}, \boldsymbol{T}, L >$$

$\mathcal{R}$ denotes a set of emergency requests, where each request $r \in \mathcal{R}$ is tagged with a tuple $< t, s, h >$. t is the arrival time, s is origin location and h is destination hospital of the request r. $\mathcal{B}$ denotes the set of possible base locations. A fleet of ambulances is represented by $\mathcal{A}$. $\boldsymbol{T}$ is a two-dimensional matrix that provides travel time between any two base locations. More specifically, T_{l_1,l_2} is the time required to move from source location l_1 to destination l_2. L is the utility function which will be explained in details later.

In this paper, we consider two main objectives:
(a) Maximise number of requests that are satisfied within a given threshold response time (ex: 15 minutes), referred to as *Bounded Time Response*;
(b) Minimise the response time for a fixed percentage (ex: 80%) of requests, referred to as *Bounded Risk Response*.

Bounded Time Response

Given a sample of training requests, our goal with this objective is to find an allocation policy for ambulances $\mathcal{A}$ into given set of bases such that maximum number of requests can be served efficiently. For this objective, the optimisation model for finding an optimal allocation of ambulances to a given set of bases $\mathcal{B}$ is compactly represented using a Mixed Integer Linear Program [MILP] in Table (1); a simple extension of the MILP provided in (Yue, Marla, and Krishnan 2012). A request $r \in \mathcal{R}$ can be served from a feasible set of nearby bases $\{\mathcal{B}_r \cup \perp\}$, where $\perp$ denotes the null assignment or lost request. x_{rl} is a binary decision variable and is set to 1 if request r is served from base $l \in \{\mathcal{B}_r \cup \perp\}$. a_l denotes the number of ambulances allocated to base $l \in \mathcal{B}$.

Intuitively, one unit of reward is provided if a request is served within 15 minutes. Let L be a function that facilitates

this reward and is defined as follows:

$$L_{rl} = \begin{cases} 1 & \text{if } T_{l,r.s} \leq 15 \text{ minutes} \\ 0 & \text{Otherwise} \end{cases}$$

$$\max_{\boldsymbol{a},\boldsymbol{x}} \sum_{r\in\mathcal{R}} \sum_{l\in\mathcal{B}_r} x_{rl} L_{rl} \quad (1)$$

$$\textbf{s.t.} \sum_{l\in\{\mathcal{B}_r\cup\perp\}} x_{rl} = 1, \qquad \forall r \in \mathcal{R} \quad (2)$$

$$x_{rl} + \sum_{j\in P_r^l} x_{jl} \leq a_l, \qquad \forall r \in \mathcal{R}, l \in \mathcal{B}_r \quad (3)$$

$$\sum_{l\in\mathcal{B}} a_l = |\mathcal{A}| \quad (4)$$

$$a_l \geq 0, x_{rl} \in \{0,1\} \quad (5)$$

Table 1: FindAllocation$(\mathcal{R}, \mathcal{B}, \mathcal{A})$

$$\max_{\boldsymbol{a},\boldsymbol{x}} M - \delta \quad (6)$$

$$\textbf{s.t.} \ \frac{\delta^r - \delta}{M} \leq z^r, \qquad \forall r \in \mathcal{R} \quad (7)$$

$$\frac{\sum_{r\in\mathcal{R}} z^r}{|\mathcal{R}|} \leq \alpha \quad (8)$$

$$\sum_{l\in\{\mathcal{B}_r\cup\perp\}} x_{rl} = 1, \qquad \forall r \in \mathcal{R} \quad (9)$$

$$x_{rl} + \sum_{j\in P_r^l} x_{jl} \leq a_l, \qquad \forall r \in \mathcal{R}, l \in \mathcal{B}_r \quad (10)$$

$$\sum_{l\in\mathcal{B}} a_l = |\mathcal{A}| \quad (11)$$

$$\delta^r \geq \sum_{l\in\mathcal{B}_r} x_{rl} \cdot T_{l,r.s} + x_{r\perp} \cdot \hat{M}, \qquad \forall r \in \mathcal{R} \quad (12)$$

$$a_l \geq 0, x_{rl} \in \{0,1\}, z^r \in \{0,1\}, \delta, \delta^r \geq 0 \quad (13)$$

Table 2: RiskAllocation$(\mathcal{R}, \mathcal{B}, \mathcal{A}, \alpha)$

Our objective (delineated in equation (1)) is to maximise the number of requests that are assisted within 15 minutes. Constraints (2) ensure that a request can be served from one base station only. P_r^l denotes the set of parents of request r for base l. A request $j \in P_l^r$ is considered as the parent of request r if it arrives before r, completes after r has arrived and base l belongs to both the feasible base set $\mathcal{B}_r$ and $\mathcal{B}_j$. Therefore, constraints (3) enforce the condition that a request can only be served from a base station if there is an available ambulance. Finally, constraints (4) ensure the equivalence between total number of allocated and available ambulances.

Bounded Risk Response

The notion of bounded risk (Saisubramanian, Varakantham, and Chuin 2015) is an important and alternative performance metric which is employed by many real world EMSs. The optimisation model for calculating the utility for a given set of bases is compactly represented using the MILP in Table (2) and is a more efficient variant of the one provided in (Saisubramanian, Varakantham, and Chuin 2015). δ^r denotes the response time for request $r \in \mathcal{R}$. δ denotes the α-response time or alternatively the percentage of requests whose response time is greater than δ should be less than the input parameter α. z^r is a binary variable that is set to 1 if response time for request r is greater than δ.

Our goal is to find an allocation of ambulances to a given set of bases, $\mathcal{B}$ such that α-response time is minimised. M represents a sufficiently large number such that objective value is always positive. We set the objective function (delineated in equation (6)) positive such that it is consistent with the objective of MILP of Table (1). Constraints (7) ensure that z^r is set to 1 if response time for request r exceeds δ. Constraints (8) enforce that the percentage of requests whose response time exceeding δ is less than the input parameter α. Another key differentiating constraints that has not been used earlier is constrains (12). These constraints ensure that the response time for request r is equals to the travel time from base (dispatched ambulance location) to scene or a relatively high number $\hat{M}$ for null assignment.

Theoretical Analysis of Objectives

In this section, we show that bounded time response objective is monotone submodular and bounded risk response objective is not submodular. Let $\mathcal{B}$ denotes a set of bases and $F(A)$ denotes the objective function for a given subset of bases $A \in 2^{\mathcal{B}}$, where objective function, $F : 2^{\mathcal{B}} \rightarrow \mathbb{R}$ is defined for a given set of requests $\mathcal{R}$, a fleet of ambulances $\mathcal{A}$ and a set of bases A.

Let A and B be two set of bases where $A \subset B \subseteq \mathcal{B}$. Let $\Delta(A|b)$ denotes the marginal gain in function F for adding a new base $b \in \mathcal{B} \setminus B$ to the current set of bases A. So, $\Delta(A|b) = F(A \cup \{b\}) - F(A)$. The objective function F is submodular if the marginal gain for adding a new base b in subset A is always higher than the gain for adding b in superset B, i.e.,

$$\Delta(A|b) - \Delta(B|b) \geq 0$$

.

Proposition 1 *F function is monotone submodular for bounded time response objective.*

Proof Sketch. Let $S_i \subseteq \mathcal{R}$ denotes the set of requests that can be served within 15 minutes from base i, then bounded time response function $F(A)$ for a given set of bases A and for optimal allocation of ambulances to A (analogous to the objective of MILP of Table (1)) is equivalent to $|\cup_{i\in A} S_i|$.

Let us have two sets of bases A and B, where B is the superset of A and represented as $\{A \cup a\}$, then Figure (1) shows the graphical proof of submodularity of bounded time response function F by employing simple properties of set union. Formally,

$$\begin{aligned} &\Delta(A|b) - \Delta(B|b) \\ &= F(A \cup \{b\}) - F(A) - F(A \cup a \cup \{b\}) + F(A \cup a) \\ &= F(a \cap \{b\}) - F(A \cap a \cap \{b\}) \geq 0 \end{aligned}$$

Hence, the bounded time response function F for a given set of requests $\mathcal{R}$ is submodular. ■

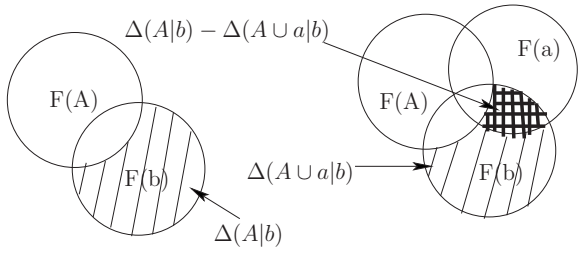

Figure 1: Bounded time response objective is Submodular

We now show that the bounded risk response objective is monotone but non-submodular. In Figure (2), we provide a simple counter example to show the non-submodularity of risk-based objective. For the ease of understanding we consider 5 requests each of which is represented by a circle. We have 3 bases represented as square. We consider a fleet of 5 ambulances. Numbers associated with each line denote the response time from the base to scene location. Let A denotes the subset and $A \cup \{a\}$ represents the superset. We are interested to find the marginal gain in α-response time for adding a new base b in both the cases. Let the tuneable parameter α is given as 0.2, therefore 80% (or 4) requests have to be served within δ. We assume the value of M as 100.

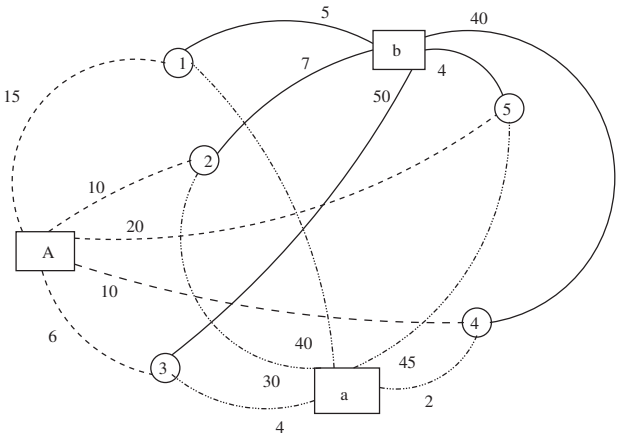

Figure 2: Non-submodularity of risk-based objective

Using only base A, we can serve 4 requests within 15 minutes, so, the value of δ is 15 and our objective, $F(A)$ is 85. If we add the new base b to A, then we observe the following optimal assignment; request 1, 2 and 5 are served from base b and request 3, 4 are served from base A. The above assignments indicate that 4 requests are served within 7 minutes, so, $F(A \cup \{b\})$ is 93. The marginal gain denoted by $\Delta(A|b)$ is (93-85)=8. In case of superset $(A \cup \{a\})$, request 1, 2 and 5 are served from base A and request 3, 4 are served from base a. So, 4 requests are assisted within 15 minutes and $F(A \cup \{a\})$ is 85. If we add the base b in superset, then request 1, 2 and 5 are served from base b while request 3, 4 are served from base a. In this case, 4 requests are served within 5 minutes, thus, $F(A \cup \{a\} \cup \{b\})$ is 95. The marginal gain $\Delta(A \cup \{a\}|b)$ is 10. Therefore, $\Delta(A|b) < \Delta(A \cup \{a\}|b)$, which proves that the bounded risk response objective is not submodular.

Strategic Planning using Greedy Approach

In this section, we outline our approach for strategic planning to decide on the number and the exact set of bases to be used. We employ the well known greedy algorithm that guarantees to provide 63% of optimal objective (Nemhauser, Wolsey, and Fisher 1978) for monotone submodular functions. Algorithm (1) provides the details of the greedy algorithm. We start with a null base set E. In each iteration we calculate the utility μ_s and optimal allocation of ambulances, $\boldsymbol{A}$ for adding each of the possible bases $s \in \mathcal{B}$ to active base set E. Then we add the base s^* (that provides maximum marginal gain) into E and remove it from possible base set $\mathcal{B}$. The process continues until the marginal gain for adding a new base is significantly higher.

Algorithm 1: SolveGreedy$(\mathcal{R}, \mathcal{B}, \mathcal{A})$

Initialize: $E \leftarrow \{\perp\}, it \leftarrow 0$;
repeat
  $\mu_s, \boldsymbol{A} \leftarrow$ FindAllocation$(\mathcal{R}, E \cup \{s\}, \mathcal{A}), \forall s \in \mathcal{B}$;
  $s^* \leftarrow \underset{s \in \mathcal{B}}{\operatorname{argmax}}\, \mu_s$;
  $E \leftarrow E \cup \{s^*\}$;
  $\mathcal{B} \leftarrow \mathcal{B} - \{s^*\}$;
until $(\max_{s \in \mathcal{B}} \mu_s \leq \epsilon)$;
return $E, \boldsymbol{A}$

Minor modification to the greedy approach can easily tackle the real-life deployment issues such as political influences that are bound to occur in the planning of EMS. Because of the political influences, a subset of bases might already be determined before the planning process. In that scenario, we need to initialise the active base set E with the pre-determined set of bases rather than an empty set and incrementally add the best possible bases until the given budget constraint is satisfied.

Lazy Greedy Algorithm

Evaluating F function or **FindAllocation()** (which requires solving MILP of Table (1)) is typically expensive even with a subset of bases and thus applying greedy algorithm (which requires evaluation of F function for every bases) can be computationally very expensive. Therefore, we employ a variant of greedy algorithm called lazy greedy (Minoux 1978) to accelerate the convergence.

The details of lazy greedy process is shown in Algorithm (2). Let $\mathcal{B}$ be the set of available base stations and E be the current set of active bases. We initialise E with a default base $\{\perp\}$ where assignment of a request to $\perp$ indicates a null assignment. In the first iteration we calculate the gain μ_s for every possible bases $s \in \mathcal{B}$ (analogous to

the greedy approach). We insert the base s^* with maximum marginal gain into E and remove s^* from available base set $\mathcal{B}$. In the subsequent iterations instead of computing gain $\Delta(E|s)$ for every base $s \in \mathcal{B}$ (which requires $O(|\mathcal{B}|)$ computations of function F), the lazy greedy keeps an upper bound μ_s for every available base. In each iteration it extracts the base ($s \in \operatorname{argmax}_{s' \in \mathcal{B}} \mu_{s'}$) with highest upper bound. Then it computes the marginal gain, $\Delta(E|s)$ for adding base s to existing base set E (i.e., the difference between the utilities $F(E \cup \{s\}) = g^{it}$ and $F(E) = g^{it-1}$) and update the upper bound μ_s as $\Delta(E|s)$. After this update if $\mu_s \geq \mu_{s'}$ for all $s' \in \mathcal{B}$, then greedy finds the best element with maximum gain (without computing gain for a large number of elements s') and insert base s into resulting base set E. This process iterates until there are no available bases whose marginal gain is higher than a predefined threshold value ϵ.

Algorithm 2: SolveLazyGreedy($\mathcal{R}, \mathcal{B}, \mathcal{A}$)

Initialize: $E \leftarrow \{\perp\}, it \leftarrow 0$;
$\mu_s, \boldsymbol{A} \leftarrow \text{FindAllocation}(\mathcal{R}, E \cup \{s\}, \mathcal{A}), \quad \forall s \in \mathcal{B}$;
$g^0 \leftarrow \max_{s \in \mathcal{B}} \mu_s$;
$s^* \leftarrow \operatorname{argmax}_{s \in \mathcal{B}} \mu_s$;
$E \leftarrow E \cup \{s^*\}$;
$\mathcal{B} \leftarrow \mathcal{B} - \{s^*\}$;
repeat
  $it \leftarrow it + 1$;
  repeat
    $s^* \leftarrow \operatorname{argmax}_{s \in \mathcal{B}} \mu_s$;
    $g^{it}, \boldsymbol{A} \leftarrow \text{FindAllocation}(\mathcal{R}, E \cup \{s^*\}, \mathcal{A})$;
    $\mu_{s^*} \leftarrow g^{it} - g^{it-1}$;
    if $\{\mu_{s^*} \geq \mu_s, \forall s \in \mathcal{B}\}$ **then**
      $E \leftarrow E \cup \{s^*\}$;
      $\mathcal{B} \leftarrow \mathcal{B} - \{s^*\}$;
      Break;
  until ***True***;
until $(\max_{s \in \mathcal{B}} \mu_s \leq \epsilon)$;
return $E, \boldsymbol{A}$

Proposition 2 *(Leskovec et al. 2007) For a placement of bases $E \in \mathcal{B}$ with a given fleet of ambulances $\mathcal{A}$, request log $\mathcal{R}$, and for each base $s \in \{\mathcal{B} \setminus E\}$ let $\Delta_s = F(E \cup s) - F(E)$. Then*

$$\max_{\mathcal{B}, \mathcal{A}, \mathcal{R}} F(\mathcal{B}) \leq F(E) + \sum_{s \in \{\mathcal{B} \setminus E\}} \Delta_s$$

By using Proposition (2) we can compute how far any given solution $F(E)$ is from the optimal solution, which can also be utilised for determining convergence.

We apply a similar lazy greedy approach to solve the bounded risk response objective, except that we calculate the F function using MILP of table (2). Even without the submodularity property of bounded risk response objective, we empirically show that lazy greedy is highly competitive with existing benchmark approaches and provide a good quality solution by utilising a significantly less number of bases.

Experimental Settings

We conduct experiments on a real world dataset[1] from a large asian city (adopted from (Yue, Marla, and Krishnan 2012)). The dataset contains a fleet of 58 ambulances and 58 base stations. We have 1500 weeks of request logs which are generated using *Poisson distribution* (Ross 1983) with the parameters estimated from real usage data over a period of one month. Each request log contains the following information (a) Origin location; (b) Arrival time; (c) A set of feasible nearby bases from where the request can be assisted; (d) Response time from each of the feasible base to scene location; and (e) Total time required for an ambulance to return back to the origin base after serving the request. In case of real deployment, the above mentioned details may not be readily available for new base locations, however, it is possible to estimate them using a straightforward method. We know the geographical locations of the requests and hospitals from the historical data. The geographical locations of the set of possible bases are also provided by the respective authority. Therefore, we can find the set of feasible nearby bases for each request and estimate the expected response and round off time for each of the possible nearby bases.

We evaluate the performance of our policy by employing a real-life event-driven simulation model (Yue, Marla, and Krishnan 2012) based on the nearest ambulance dispatch policy. We use Sample Average Approximation [**SAA**] (Verweij et al. 2003) for validation and performance estimation. Specifically, we generate 10 policies using a training dataset consisting of request logs for 10 weeks. Then we identify the policy with best validation performance over 500 weeks of request logs. Finally, we evaluate the performance of the validated policy on 3 test datasets each of which contains 300 weeks of request logs. We compare our approach with three existing benchmark approaches from literature (a) Greedy approach provided by (Yue, Marla, and Krishnan 2012); (b) Risk-based optimisation approach [RBO] (Saisubramanian, Varakantham, and Chuin 2015); and (c) A baseline approach where 1 ambulance is allocated to every base.

Simulation Model

We evaluate the performance of ambulance allocation policy on the resulting base set using a real-life event-driven simulation model (courtesy: (Yue, Marla, and Krishnan 2012)) based on the nearest ambulance dispatch policy. The pseudo code for the event-driven simulator is shown in Algorithm (3). We start with an event set ξ where each element $e \in \xi$ represents a request and the list is sorted based on arrival order of requests. I denotes the set of available ambulances that are allocated according to given policy $\boldsymbol{A}$. a_r denotes the ambulance id that is assigned for request $r \in \mathcal{R}$. Initially each request is tagged as null assignment. In each

[1] http://projects.yisongyue.com/ambulance_allocation/

iteration we pop the first element e from the event list ξ. If the event e is a new request then we dispatch the nearest available ambulance a_r for the request and remove the ambulance from available ambulance set I. We also insert a job-completion event in the event list at time $t_r(a_r)$, where $t_r(a_r)$ denotes the time when ambulance a_r will return back to base after completing the job r . On the other hand, if the popped element e is a job completion event for request r, then we add the ambulance a_r to the set I such that it can be used to serve a new request. This process continues until the event list becomes empty. Once the process is finished, we can use the assignment results to measure the responsiveness of the system such as bounded time response or bounded risk response time for the given sample requests. We use this simulation model to compute the performance metrics for all the benchmark algorithms.

Algorithm 3: EDSimulator($\mathcal{R}, \mathcal{B}, \boldsymbol{A}$)

Initialize: $it \leftarrow 0$;
$I \leftarrow \boldsymbol{A}$ // Initialise set of available ambulance;
$\xi \leftarrow \mathcal{R}$ sorted in arrival order;
$\boldsymbol{a} = \{a_r | a_r \leftarrow \perp\}$ //Initialise as null assignment ;
repeat
  Pop next arriving event e from ξ;
  if e =*New Request* r **then**
    $a_r \leftarrow Dispatch(r, I)$ // Dispatch nearest free ambulance;
    $I \leftarrow I - \{a_r\}$ // Update available ambulance;
    Push job completion event at time $t_r(a_r)$ into ξ;
  else if *e=job completion event for* r **then**
    $I \leftarrow I \cup \{a_r\}$ // Update available ambulance;
until $(|\xi| > 0)$;
return $\{\boldsymbol{a_r}\}$

Sample Average Approximation (SAA)

We employ Sample Average Approximation (Verweij et al. 2003) for policy validation and performance estimation. We generate M minimal base sets $B_1, ..., B_M$ and allocation policies $A_1, ..., A_M$ for M sample of request logs. Then we validate those policies on N_{valid} samples and select the best allocation policy A^* and base placement B^*, which has maximum validation performance. Finally we test the performance of policy (A^*, B^*) on a separate collection of N_{test} samples and report the performance statistics. We measure the performance metrics by taking average over all the samples. For e.g., if we have N sample of request logs $\mathcal{R} = \{R_1, ..., R_N\}$, then the expectation is computed using Equation (14) by taking average over all the N samples.

$$F_{\mathcal{R}}(A^*, B^*) = \frac{1}{N} \sum_{i=1}^{N} \sum_{r \in R_i} F_r(A^*, B^*) \qquad (14)$$

Benchmark 1: (Yue, Marla, and Krishnan 2012) The primary goal of this paper is to efficiently allocate an entire fleet of ambulances to a predetermined set of bases such that the percentage of requests served within a certain threshold time bound is maximised. They used a greedy approach to find the optimal allocation for ambulances in each iteration using a real-life event driven simulator and incrementally added the ambulances until the entire fleet is allocated efficiently. In addition,they showed that the proposed allocation policy can be effectively utilised for the dynamic redeployment of free ambulances.

Benchmark 2: (Saisubramanian, Varakantham, and Chuin 2015) This paper proposes to minimise the bounded risk (i.e., the time bound within which $\alpha\%$ requests are served), a metric employed by many EMSs, by efficiently allocating a fleet of ambulances to a given set of bases. Due to the inherent complexity of the introduced MILP w.r.t. number of requests, they employ Lagrangian Dual Decomposition (LDD) to improve the scalability and approximately solve the assignment problem by considering a large sample of requests. Finally, they evaluate the performance of their allocation policy using an event driven simulator.

Experimental Results

We compare our approach with respect to performance metrics such as (a) Runtime; (b) Bounded time response: percentage of requests served within 15 minutes; and (c) Bounded risk response: α-response time (unless otherwise stated we use α value as 0.2). We provide five thread of results on real world dataset (a) Gain in runtime for lazy greedy over general greedy approach; (b) Experimental validation of the submodularity of bounded time response and non-submodularity of bounded risk response; (c) Effect of external parameter such as risk tolerance level [α] on strategic planning; (d) Effect of external budget such as size of ambulance fleet on two objective functions as well as on the strategic planning (number of required bases); and (e) Performance comparison with the benchmark approaches on three test datasets, each contains 300 weeks of requests.

Runtime Results : Figure (3) plots the runtime comparison between lazy greedy and general greedy approach. Figure 3(a) depicts the runtime for bounded time response objective on a sample of around 3000 requests. X-axis denotes the iteration number and Y-axis represents the runtime in seconds in a logarithmic scale. Greedy approach is unable to finish more than 20 iterations within the cut-off time of 2 hours, while lazy greedy approach provides a significant gain over greedy and completes the process within 10 minutes. Figure 3(b) shows the runtime for bounded risk response objective. While greedy approach is unable to complete 18 iterations within the threshold time of 2 hours, lazy greedy significantly accelerates it and finish the process within the cut-off time. Note that the runtime for both the greedy and lazy greedy for initial 12 iterations was equal. This is so because we cannot serve 80% of the requests (i.e., $\alpha = 0.2$) using less than 13 base stations (because a request can only be assisted from a subset of nearby bases), and therefore in the initial iterations upper bound was equal for every possible bases (i.e., $\mu_s = M, \forall s \in \mathcal{B}$). So, the lazy greedy essentially search over all the possible bases, which is equivalent to general greedy approach.

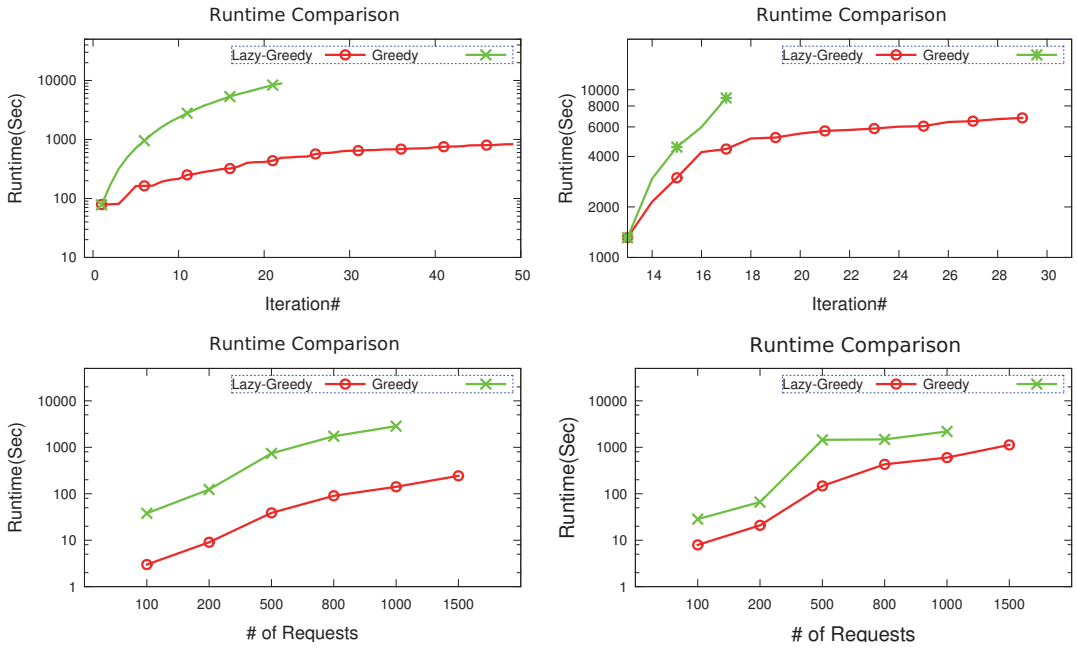

Figure 3: Runtime: Greedy vs. Lazy greedy (a) Iterations wise for Bounded time response; (b) Iterations wise for Bounded risk response; (c) With varying request for Bounded time response; (d) With varying request for Bounded risk response

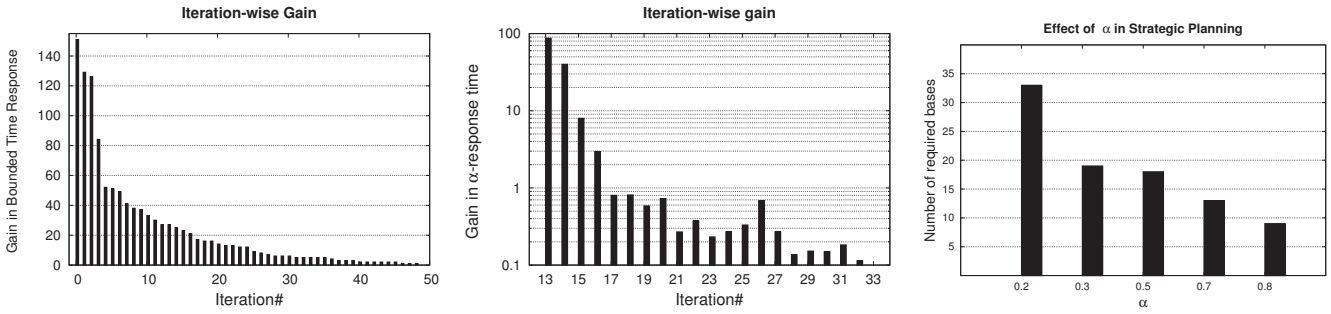

Figure 4: (a) Iteration-wise gain for bounded time response; (b) Iteration-wise gain for bounded risk response; (c) Effect of α on strategic planning.

Figure 3(c) demonstrates the gain in runtime for lazy greedy approach where we vary the number of requests in the X-axis. The complexity of greedy approach grows exponentially as the number of requests increases. This is so because the dependency between requests increases for densely populated request logs. Greedy cannot solve problems with more than 1000 requests within the cut-off time, while lazy greedy solves the problem with 1500 requests within 2 minutes. In the same direction, Figure 3(d) demonstrates that lazy greedy significantly outperforms greedy approach in case of bounded risk response objective.

Submodularity Results : Figure 4(a),4(b) depict the marginal gain for adding a base in each iteration for both the objective functions. Figure 4(a) clearly shows that marginal gain decreases monotonically in each iteration which validates the submodularity property of the bounded time response objective. Figure 4(b) delineates the iteration wise gain of α-response time in a logarithmic scale. As expected, due to the non-submodularity, in few cases the marginal gain in later iteration is slightly higher.

Effect of external parameter α: Figure 4(c) depicts the effect of parameter α in strategic planning for the bounded risk response objective on a fixed sample of requests. Note that increasing α value indicates that less number of requests need to be served within α-response time. Therefore, the size of resulting base set reduces as we increase the α value.

Results on varying budget : Our model can be employed to find the right location for a small set of new ambulances in addition to an existing fleet of ambulances. For e.g., if a new budget arises for p ambulances at certain point of time, and q number of ambulances already exists in system, then we can use our algorithm with $(p + q)$ ambulances to find

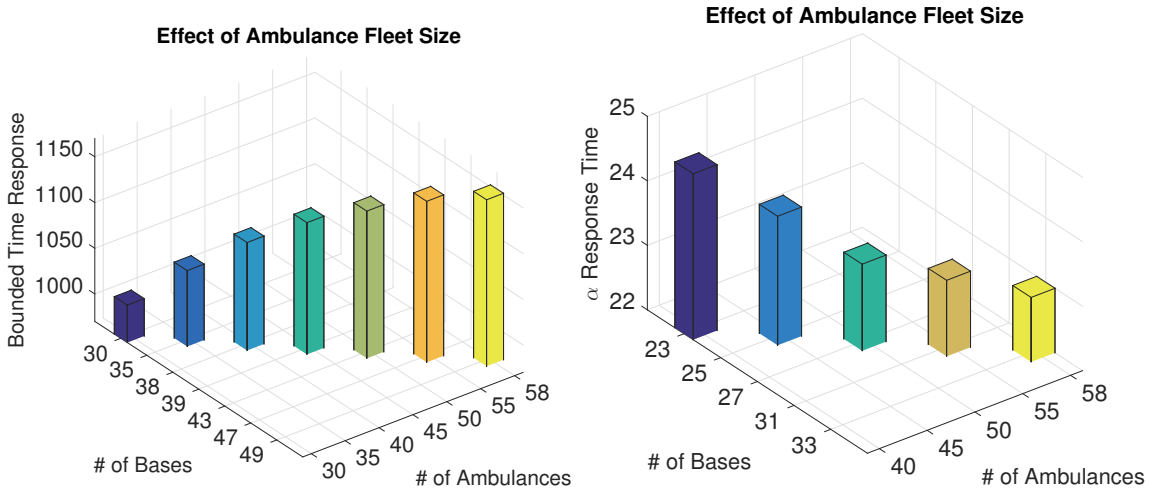

Figure 5: Effect of fleet size for optimising (a) Bounded time response; (b) Bounded risk response.

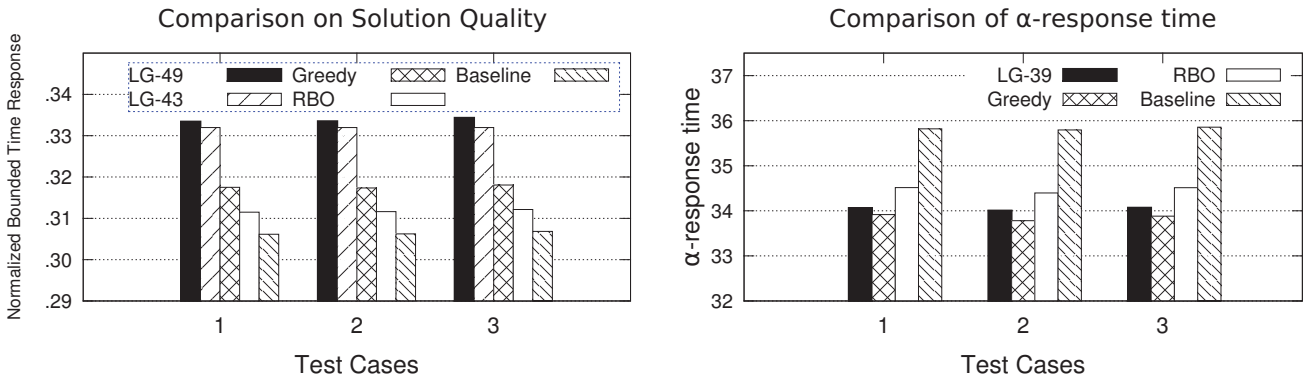

Figure 6: Quality : (a) Bounded time response; (b) Bounded risk response.

the minimal subset of bases such that the entire fleet can be allocated efficiently. Figure (5) show the performance with respect to varying fleet size on a sample of around 3000 requests. We show the effect of varying fleet size in the strategic planning (ex: number of required bases) as well as in objective value (ex: bounded time response or α-response time). We vary the ambulance fleet size in X-axis, while Y-axis shows the size of active base set and Z-axis denotes the utility. We observe the pattern is consistent, i.e., bounded time response increases with number of ambulances (Figure 5(a)) and bounded risk response is inversely proportional to fleet size (Figure 5(b)). For both the objectives, as we increase the number of ambulances, we need additional bases to effectively allocate the entire fleet of ambulances.

Results on test cases : The last and most important thread of results demonstrate the performance comparison between all the benchmark approaches on the test instances. We provide performance for two of our allocation policies. LG-49 represents an allocation policy (generated using lazy greedy) where the process continues until the marginal gain is positive and it produces a resulting base set of size 49. LG-43 symbolises an allocation policy with 43 bases where we stop the process if the marginal gain is less than or equals to 2. It indicates a crucial advantage of our approach in strategic planning as we have the flexibility to generate strategy based on the expectation of EMS operators and the availability of budget to construct the base stations. Figure 6(a) plots the normalised bounded time response value for all the test cases. Y-axis represents the percentage of requests served within 15 minutes. As each of the test cases involves 300 weeks of request logs, we report the average utility using SAA. In all the test cases our allocation policy (even with lesser number of bases) outperforms the existing benchmark approaches and provide almost 2% gain in bounded time response.

Figure 6(b) illustrates the performance comparison on α-response time. LG-39 symbolises an allocation policy with 39 bases that is generated using lazy greedy. Interestingly by utilising less than 70% of total bases, our approach significantly outperforms the baseline approach and is highly competitive with other two benchmark approaches.

Conclusion

In this paper we present a promising approach for placement of bases and ambulances in EMS. We employ an incremental greedy approach that identifies the base with maximum marginal gain in each iteration and add it to the resulting base set. A lazy greedy approach is further utilised to accelerate the convergence and the derived policy is evaluated using a real-world event driven simulator. We show that our approach can be utilised to optimise crucial performance metrics such as bounded time response and bounded risk response. The empirical results on real-world dataset demonstrate that our approach significantly improves the service level of EMS over existing benchmark approaches.

Acknowledgements

This research is supported by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

References

Andersson, T., and Värbrand, P. 2007. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society* 58(2):195–201.

Bjarnason, R.; Tadepalli, P.; Fern, A.; and Niedner, C. 2009. Simulation-based optimization of resource placement and emergency response. In *IAAI*.

Brotcorne, L.; Laporte, G.; and Semet, F. 2003. Ambulance location and relocation models. *European journal of operational research* 147(3):451–463.

Church, R., and Velle, C. R. 1974. The maximal covering location problem. *Papers in regional science* 32(1):101–118.

Gendreau, M.; Laporte, G.; and Semet, F. 2006. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society* 57(1):22–28.

Golovin, D.; Krause, A.; Gardner, B.; Converse, S. J.; and Morey, S. 2011. Dynamic resource allocation in conservation planning. In *AAAI*, volume 11, 1331–1336.

Hakimi, S. L. 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research* 12(3):450–459.

Hochbaum, D. S., and Shmoys, D. B. 1985. A best possible heuristic for the k-center problem. *Mathematics of operations research* 10(2):180–184.

Huang, R.; Kim, S.; and Menezes, M. B. 2010. Facility location for large-scale emergencies. *Annals of Operations Research* 181(1):271–286.

Jia, H.; Ordóñez, F.; and Dessouky, M. 2007. A modeling framework for facility location of medical services for large-scale emergencies. *IIE transactions* 39(1):41–55.

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research* 9:235–284.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429. ACM.

Maxwell, M. S.; Restrepo, M.; Henderson, S. G.; and Topaloglu, H. 2010. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing* 22(2):266–281.

Minoux, M. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*. Springer. 234–243.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming* 14(1):265–294.

Restrepo, M.; Henderson, S. G.; and Topaloglu, H. 2009. Erlang loss models for the static deployment of ambulances. *Health care management science* 12(1):67–79.

Ross, S. 1983. *Stochastic processes*, volume 23. John Wiley & Sons New York.

Saisubramanian, S.; Varakantham, P.; and Chuin, L. H. 2015. Risk based optimization for improving emergency medical systems. In *AAAI*.

Schmid, V. 2012. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research* 219(3):611–621.

Sylvester, J. J. 1857. A question in the geometry of situation. *Quarterly Journal of Pure and Applied Mathematics* 1.

Toregas, C.; Swain, R.; ReVelle, C.; and Bergman, L. 1971. The location of emergency service facilities. *Operations Research* 19(6):1363–1373.

Verweij, B.; Ahmed, S.; Kleywegt, A. J.; Nemhauser, G.; and Shapiro, A. 2003. The sample average approximation method applied to stochastic routing problems: a computational study. *Computational Optimization and Applications* 24(2-3):289–333.

Yue, Y., and Guestrin, C. 2011. Linear submodular bandits and their application to diversified retrieval. In *Advances in Neural Information Processing Systems*, 2483–2491.

Yue, Y.; Marla, L.; and Krishnan, R. 2012. An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment. In *AAAI*.