6-2018

Bounded rank optimization for effective and efficient emergency response

Pallavi Madhusudan MANOHAR
Singapore Management University, pallavim@smu.edu.sg

Pradeep VARAKANTHAM
Singapore Management University, pradeepv@smu.edu.sg

Hoong Chuin LAU
Singapore Management University, hclau@smu.edu.sg

Bounded Rank Optimization for Effective and Efficient Emergency Response

Pallavi Manohar
School Of Information Systems
Singapore Management University
pallavim@smu.edu.sg

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

Hoong Chuin Lau
School Of Information Systems
Singapore Management University
hclau@smu.edu.sg

Abstract

Effective placement of emergency response vehicles (such as ambulances, fire trucks, police cars) to deal with medical, fire or criminal activities can reduce the incident response time by few seconds, which in turn can potentially save a human life. Owing to its adoption in Emergency Medical Services (EMSs) worldwide, existing research on improving emergency response has focused on optimizing the objective of bounded time (i.e. number of incidents served in a fixed time). Due to the dependence of this objective on temporal uncertainty, optimizing the bounded time objective is challenging. In this paper, we propose a new objective referred to as the bounded rank (which is the number of incidents served by a base station whose rank is below a bounded rank value) that has nice theoretical properties and serves as an indirect substitute for the bounded time objective.

To understand the theoretical properties of this new objective in the context of the spatio-temporal uncertainty associated with emergency incidents, we first provide a Poisson Point Process (PPP) model of the emergency response problem. We then formally define the bounded rank objective in the context of the model and demonstrate that the bounded rank metric is monotone submodular. Due to the monotone submodularity of the objective, we can propose a greedy approach that can provide an *a priori* guarantee of 50% from optimal and a much tighter *posteriori* guarantee. Practically and more importantly, we demonstrate that optimizing this bounded rank objective on simulators validated on real data (and not just on the abstract PPP model) provides better results than the best known approach for optimizing bounded time objective.

1 Introduction

To handle medical, fire or crime related emergencies, Emergency Response Vehicles (ERVs) like ambulances, fire rescue vehicles and police cars are strategically positioned at a set of base stations throughout the city. Since time can be critical in responding to such emergency situations, we need to optimize the placement of these ERVs so they can reach incident locations within the fastest possible time. Specifically, a Key Performance Indicator (KPI) employed by many emergency management systems worldwide is to maximize the number of incidents that have a response time less than a fixed time value (which is dependent on the nature of the emergency). This is referred to as the bounded time objective, and the challenge is to increase the number of incidents that have a response time lower than the bounded time while considering the spatio-temporal uncertainty associated with the occurrence of emergency incidents.

Emergency management systems have been extensively studied in the literature, specifically in the context of ambulance response[1]. There are two main threads of existing research. The first thread has focused on the dispatch of ambulances from base locations and the second thread has focused on the allocation and reallocation of ambulances to base stations. A survey of existing approaches (Brotcorne, Laporte, and Semet 2003) reveals that much of the previous work in EMS has been on the dispatch of ambulances from base stations (Schmid 2012; Andersson and Värbrand 2007; Ibri, Nourelfath, and Drias 2012). While dispatch is an important mechanism to improve emergency management systems, given the significant spatio-temporal uncertainty associated with occurrence of incidents as well as the uncertainty involved in ascertaining the criticality of an emergency request over phone, in almost all EMSs world wide, the ambulance dispatch procedure is fixed: *the nearest ambulance to the incident location is dispatched*.

The key focus of this paper is to advance research on the latter problem (allocation and reallocation of ERVs). Unlike Maxwell *et al.* (Maxwell et al. 2010), we focus on allocating ambulances for the entire fleet and not for an individual ambulance. As indicated earlier, we focus on improving the bounded time response and hence is different to the work on heuristic worst case planning (Andersson and Värbrand 2007). Owing to the significant spatio-temporal uncertainty in incident occurrence coupled with the city scale nature of the problem (with large number of bases, ERVs and locations), the allocation problem is computationally challenging and is typically driven by data. There have been typically two types of objectives considered in the literature while optimizing allocation using data-driven models:

1. Bounded time objective (Yue, Marla, and Krishnan 2012): The goal here is to maximize the number of

[1]Given the extensive literature in ambulance response, we will use ERVs and ambulances synonymously

incidents that have a response time less than a fixed time value.

2. Bounded risk objective (Saisubramanian, Varakantham, and Lau 2015; Ghosh and Varakantham 2016): The goal here is to minimize the response time for a fixed percentile of requests.

For bounded time objective, existing approaches (Yue, Marla, and Krishnan 2012) that are based on data-driven simulators greedily allocate ambulances to base stations based on the marginal benefit in terms of the number of incidents. For bounded risk objective, existing data-driven approaches employ a combination of linear optimization and sample average approximation (Pagnoncelli, Ahmed, and Shapiro 2009; Varakantham and Kumar 2013) to minimize response time. However, with both these objectives and threads of work, there is no optimal solution available and the focus is on approximate solutions. In this paper, our focus is on a bounded rank metric that when optimized using a greedy approach improves performance with respect to bounded time objective better than the greedy approach that optimizes bounded time objective.

The new objective, Bounded Rank (BR) has a parameter K. The goal with bounded rank objective is to maximize the number of incidents that are served by an ambulance from one of the K nearest base stations. Typically, emergency response systems dispatch the response vehicle from the nearest base station. Due to the spatio-temporal uncertainty and limited ambulances at each base station, there is a good chance that the ambulance from the nearest base station may not be available to serve all the incidents in the same area. In fact, in our real world datasets, we observe that the top 2 nearest stations typically serve most of the incidents equally. Therefore, by picking a K that ensures response time is less than the bounded time in bounded time objective, bounded rank can be made to optimize for bounded time indirectly.

It is important to note that for both bounded time and bounded rank, there are two key similarities which allow for this indirect optimization of bounded time metric using BR objective-(i) Output with both metrics is the number (or percentage) of requests satisfying a certain criterion (response time $<$ given time value for bounded time, and rank $<$ K for bounded rank) and (ii) Both these metrics have continuous (if we consider percentage of requests) values and do not result in infeasible solutions. On the other hand, for bounded risk, output is the response time satisfying a certain criterion (like percentage of requests with a lesser response time than the objective is at least 80) resulting in infeasible solutions if there is no allocation that will result in 80 (or some fixed) percentile of requests being served.

We make the following contributions in this paper:

(1) To represent the spatio-temporal uncertainty associated with emergency incidents, we first provide a Poisson Point Process (PPP) model of the emergency response problem and formally define the bounded rank objective in the context of the model.

(2) We demonstrate that the metric of incidents served from a bounded rank (e.g., nearest base is a rank of 1, second nearest is rank of 2) base station is monotone submodular. Due to the submodularity of the objective, we can propose a greedy approach that can provide an *a priori* guarantee of 50% from optimal and a tighter *posteriori* guarantee.

(3) Practically and more importantly, we demonstrate that optimizing this bounded rank objective provides better results than the state of art existing approach for optimizing bounded time objective on two simulators validated by real world data sets. We also observe that the bounded rank objective performs very well on the bounded risk metric as well.

2 Background

In this section, we first describe monotone submodularity and matroids as they are required in proving submodularity of the bounded rank objective and also in showing the guarantee of the greedy approach. Next, we describe the data driven optimization work of Yue *et al.* (Yue, Marla, and Krishnan 2012), as key technical details from that work are referenced in this paper.

Monotone Submodularity and Matroids

We now describe submodular functions and matroids.

Definition 1. *Given a finite set, Π, a **submodular** function is a set function, $F : 2^{\Pi} \rightarrow \mathbb{R}$, where 2^{Π} is the power set corresponding to Π. More importantly, $\forall X, Y \subseteq \Pi$ with $X \subseteq Y$ and for every $i \in \Pi \setminus Y$,*

$$F(X \cup i) - F(X) \geq F(Y \cup i) - F(Y)$$

*A submodular function F is **monotone** if $F(Y) \geq F(X)$ for $X \subseteq Y$.*

Monotone submodular functions are interesting because maximizing a submodular function to pick a fixed number of elements (say k) from the finite set (Π) while difficult can be approximated efficiently with a strong quality guarantee. Specifically, a greedy algorithm that incrementally generates the solution set by maximizing marginal utility provides solutions that are at least 63% $(1 - \frac{1}{e})$ of the optimal solution.

In this paper, we are also interested in maximizing a submodular function, however, under a specific constraint on the finite set (Π) and the elements that are picked. Specifically, the constraint is specified using a partition matroid. We provide the formal definitions below:

Definition 2. *For a finite ground set, Π, let $\mathcal{P}$ be a non-empty collection of subsets of Π. The system $\Gamma = (\Pi, \mathcal{P})$ is a matroid if it satisfies the following two properties:*

- *The hereditary property: $\mathcal{P}_1 \in \mathcal{P} \wedge \mathcal{P}_2 \subset \mathcal{P}_1 \implies \mathcal{P}_2 \in \mathcal{P}$. In other words, all the subsets of $\mathcal{P}_1$ must be in $\mathcal{P}$.*

- *The exchange property:* $\forall \mathcal{P}_1, \mathcal{P}_2 \in \mathcal{P} : |\mathcal{P}_1| < |\mathcal{P}_2| \implies \exists x \in \mathcal{P}_2 \setminus \mathcal{P}_1; \mathcal{P}_1 \cup x \in \mathcal{P}$.

We are specifically interested in a ground set that is partitioned as $\Pi = \Pi_1 \cup \Pi_2 \cup \ldots \cup \Pi_k$. The family of subsets, $\mathcal{P} = \{P \subseteq \Pi : \forall i, |P \cap \Pi_i| \leq 1\}$ forms a matroid called a partition matroid. This family of subsets denotes that any solution can include at most one element from each ground set partition. This is relevant in this paper, as ground set partitions represent base set for each ambulance and we need to pick one base for each ambulance.

Simulation Driven Optimization

Yue *et al.* (Yue, Marla, and Krishnan 2012) provided the event-driven simulator of Algorithm 1, which employs order of incident arrival and the nearest idle ambulance dispatch policy. We start with an event set ξ where each element $e \in \xi$ represents an emergency incident and the list is sorted based on arrival order of incident. I denotes the set of available ambulances that are allocated according to given allocation $\boldsymbol{A}$. 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 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 metrics: percentage of requests served in bounded time.

Algorithm 1: 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 $t_r(a_r)$ into ξ;
  else if *e=job completion event for* r **then**
    $I \leftarrow I \cup \{a_r\}$ // Update available ambulance;
until $(|\xi| > 0)$;
return $\{\boldsymbol{a_r}\}$

Yue *et al.* (Yue, Marla, and Krishnan 2012) also provided a greedy algorithm that incrementally considers assignment of ambulances to base stations based on the marginal benefit (in terms of number of incidents served within a fixed time) computed using the simulator above.

3 Poisson Point Process (PPP) Model of Emergency Incidents

We now describe a model to represent the problem of optimizing allocation of ERVs in emergency response. The occurrence of an emergency incident is a spatio-temporal random event. The density of incidents varies spatio-temporally and the arrivals are independent of each other. As in previous work (Yue, Marla, and Krishnan 2012; Peleg and Pliskin 2004), we represent the arrival process of emergency incidents using a Poisson distribution. We further generalize the Poisson process to incorporate a spatial variation. It is captured by a spatial non-homogeneous Poisson process with a fixed, spatially varying density function in two dimension, $\lambda(x, y)$ which is formed from the arrived incident location points. The temporal variation of the incident arrivals can be captured by considering different density functions, i.e., $\lambda_t(x, y)$ for different times of the day like peak and non-peak hours. The granularity of t could be as large as a season or a week or it could be varying every hour. Since we are providing an off-line optimization solution, granularity of t is not indeed a constraint. Thus, our interest is in optimizing the ambulance allocation over a given period of time, e.g, in an hour or over a day, under a spatial Poisson model with the given density $\lambda(x, y)$ in that time period. Henceforth, while we do not explicitly mention time period, the focus is on a given time period.

The emergency incident points in a two-dimensional space, $\{S_i\}$ form a non homogeneous Poisson process with the density $\lambda(x, y)$. The expected number of incidents is given by,

$$\bar{S} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \lambda(x, y)\, dx\, dy \tag{1}$$

For a zone Z_i, $i = 1, 2, \ldots L$, the probability that there are k incidents in zone Z_i is given by Poisson distribution as follows.

$$\boldsymbol{Pr}\{N(Z_i) = k\} = \frac{e^{-\bar{S}(Z_i)} \bar{S}(Z_i)^k}{k!} \ , \tag{2}$$

where, $\bar{S}(Z_i)$ is the expected number of incidents in zone Z_i which is given by,

$$\bar{S}(Z_i) = \int \int_{(x,y) \in Z_i} \lambda(x, y)\, dx\, dy \tag{3}$$

The density could vary in different zones of the city and can be obtained using data of historical incidents (during the time period of interest). For purposes of generality, we will refer to density using $\lambda(x, y)$ and not use a zone representation.

4 Optimizing Ambulance Allocation in PPP Model

We now formally define the ambulance allocation problem using the following tuple:

$$< \mathcal{R}, \mathcal{B}, \mathcal{N}, F_K >$$

$\mathcal{R}$ denotes a set of emergency requests. $\mathcal{B} = \{b_1, b_2, \ldots b_{|\mathcal{B}|}\}$ denotes the set of bases where ambulances can be positioned. $\mathcal{N} = \{n_1, n_2, \ldots n_{|\mathcal{N}|}\}$ represents the set of ambulances. The goal is to optimize metric F_K by computing an allocation set, $\mathcal{A}$, where

$$\mathcal{A} = \{(i,j) | i \in \mathcal{N}, j \in \mathcal{B}, \sum_{j \in \mathcal{B}} |(i,j)| = 1\}$$

Each element (i,j) represents the assignment of i^{th} ambulance to j^{th} base. Note that there can be multiple ambulances allocated to the same base but an ambulance can be allocated to one base (and hence the $\sum_{j \in \mathcal{B}} |(i,j)| = 1$).

F_K is the Bounded Rank (BR) metric. For each emergency incident request, $r \in \mathcal{R}$, we consider a set of nearest bases ranked in increasing order of their respective response times, assuming that there is an idle or available ambulance. This is typically used for implementing the nearest idle dispatch policy. In the context of the PPP model described earlier, F_K measures the expected number of incidents served from the nearest (in terms of response time) K bases (i.e., up to rank K) for a given allocation of ambulances.

Let us consider an ambulance allocation, $\mathcal{A}$. We split the spatial non-homogeneous Poisson process into two types, type-0 and type-1. Type-1 are the points or incidents which satisfy bounded rank objective (i.e., are served by a base of rank-K or lower). Our metric $F_K(\mathcal{A})$ then corresponds to the expected number of incidents of type-1 and is given as follows:

$$F_K(\mathcal{A}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \lambda(x,y) P_K^{\{\mathcal{A}\}}(x,y)\, dx\, dy \quad (4)$$

$P_K^{\{\mathcal{A}\}}(x,y)$ is the probability that the incident at (x,y) is served by a base of rank less than K in the allocation $\mathcal{A}$. This division into multiple types is referred to as splitting or thinning (Ross 2010) of a Poisson point process and the expression is justified through this connection to thinning/splitting. We now describe the computation of $P_K^{\{\mathcal{A}\}}(x,y)$.

Calculating $P_K^{\{\mathcal{A}\}}(x,y)$

The probability computation has two parts:

(1) There is a base of rank-K or lower in the set $\mathcal{A}$ for (x,y): For a given point or incident (x,y), the event of having a base, b, of rank-K or lower in the rank list of the point (or the zone to which this point belongs to) in allocation $\mathcal{A}$ is a deterministic event for a given $b, \mathcal{A}$ and (x,y). Let $\boldsymbol{I}^K_{(x,y)}(b,\mathcal{A})$ be the indicator variable that is set to 1 if $rank_{(x,y)}(b) \leq K$; $b \in \mathcal{A}$ and 0 otherwise.

(2) There is an ambulance available at a base identified in (1) above: To decide availability of ambulance at a base, we model the dynamics of emergency request service at each base using queueing theory. Specifically, the queueing model of interest is an Erlang-B or loss model, where there is no queueing of emergency requests. Queueing here essentially means requests waiting for service completion. It is a M/M/c/0 queue (Ross 2010) where there are Poisson arrivals of emergency incidents, exponential service for the incident requests and c servers or ambulances with zero buffer or no queueing. Thus, the steady state probability of an ambulance being free at a base b with c ambulances is given as follows:

$$\boldsymbol{Pr}\{Q < c\} \quad = \quad 1 - \pi_c = 1 - \frac{\frac{m^c}{c!}}{\sum_{j=0}^{c} \frac{m^j}{j!}} \quad (5)$$

where, Q is the number of customers waiting for service completion or equivalently it is the number of customers in the system and similarly m is the expected number of customers in the system or equivalently expected number of busy servers or ambulances.

As in prior literature (Larson 1974; Lee et al. 2006), service time for an emergency request can be assumed to be exponentially distributed with rate μ based on average service times from history and the expected service time is $T = \frac{1}{\mu}$. For arrival rate, λ, $m = \lambda \times T = \frac{\lambda}{\mu}$.

Let $\left\langle b_1(x,y), b_2(x,y), \ldots, b_K(x,y) \right\rangle$ be the top K nearest bases for incident at location (x,y). For notational convenience, we will refer to $b_i(x,y)$ as b_i. Given (1) and (2) above, we have:

$$\begin{aligned} P_K^{\{\mathcal{A}\}}(x,y) &= \boldsymbol{Pr}\{\text{request at } (x,y) \\ &\quad \text{served by rank} \leq K \text{ base}\} \\ &= \quad 1 - \boldsymbol{Pr}\{\text{request at } (x,y) \text{ is not} \\ &\quad \text{served by } \{b_1 \cap \ldots \cap b_K\}\} \\ &= 1 - \prod_{i=1}^{K} \boldsymbol{I}^K_{(x,y)}(b_i, \mathcal{A}) \cdot \boldsymbol{Pr}\{Q_{b_i} = c_{b_i}\} \end{aligned} \quad (6)$$

Here, c_{b_i} is the number of ambulances allocated to base b_i in $\mathcal{A}$. Q_{b_i} is a random variable representing number of emergency requests served or number of busy ambulances at base b_i. $\boldsymbol{Pr}\{Q_{b_i} = c_{b_i}\}$ is the blocking probability which means the probability that there is no free ambulance and hence the request is blocked, i.e., can not be served. This probability is essentially $\pi_{c_{b_i}}$ and from (5) is given by,

$$\boldsymbol{Pr}\{Q_{b_i} = c_{b_i}\} = \frac{\frac{m_{b_i}^{c_{b_i}}}{c_{b_i}!}}{\sum_{j=0}^{c_{b_i}} \frac{m_{b_i}^j}{j!}}, \quad (7)$$

where, $m_{b_i} = \frac{\lambda_{b_i}}{\mu}$ and,

$$\lambda_{b_i} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda(x,y) \cdot \boldsymbol{I}^K_{(x,y)}(b_i, \mathcal{A})\, dx\, dy \quad (8)$$

$\boldsymbol{I}^K_{(x,y)}(b_i, \mathcal{A})$ represents the event of having base, b_i, as the rank-K or lower ranked base in the rank list of a given point (x, y). Here, $\frac{1}{\mu}$ is the average service time (round trip time for an ambulance, i.e., from base to incidence to hospital and back to base) which is obtained for each base using historic data.

Monotone Submodularity of F_K

In order to prove monotone submodularity of a function, F_K, we have to satisfy the following requirements:

- There exists a finite ground set, E such that F_K is defined for all subsets of E.
- F_K is monotone, i.e., $\forall \mathcal{A} \subseteq \hat{\mathcal{A}} \subseteq 2^E$, we have

$$F_K(\mathcal{A}) \leq F_K(\hat{\mathcal{A}}) \quad (9)$$

- F_K is submodular i.e., $\forall \mathcal{A} \subseteq \hat{\mathcal{A}} \subseteq 2^E$ and for every, $a \in 2^E \setminus \hat{\mathcal{A}}$, we have,

$$F_K(\mathcal{A} \cup \{a\}) - F_K(\mathcal{A}) \geq F_K(\hat{\mathcal{A}} \cup \{a\}) - F_K(\hat{\mathcal{A}})$$

For the first requirement, ground set, E is the set of all possible assignments of ambulances to base stations:

$$E = \{(i,j) | \forall i \in \mathcal{N}, \forall j \in \mathcal{B}\}$$

It should be noted that F_K should be defined for all possible allocations, irrespective of whether an allocation is valid or not. The definition of F_K remains the same as in Equation 4, except that in case of an invalid allocation i.e., when an ambulance is allocated to multiple bases, we assume that there are as many copies of that ambulance (as the bases it is assigned to). For second and third requirements, we show it in proof of Proposition 1.

Proposition 1. *$F_K : 2^E \rightarrow \mathbb{R}$ defined using Equation 4 is monotone submodular.*

Proof: Let $\mathcal{A} \subseteq \hat{\mathcal{A}} \subseteq 2^E$. The superset $\hat{\mathcal{A}}$ is generated by adding more elements to the set $\mathcal{A}$. Addition of an element essentially means that an additional ambulance is allocated a base[2]

Since number of ambulances at bases either stays same or increases, it should be noted that $P_K^{\{\hat{\mathcal{A}}\}}(x,y) \geq P_K^{\{\mathcal{A}\}}(x,y)$. Therefore, $F_K(\hat{\mathcal{A}}) \geq F_K(\mathcal{A})$ and *monotonicity* of F_K is proved.

Now, to prove submodularity, we first obtain $F_K(\hat{\mathcal{A}} \cup \{a\}) - F_K(\hat{\mathcal{A}})$ as follows. From Equation (4), for every, $a \in 2^E \setminus \hat{\mathcal{A}}$, we have that

$$F_K(\hat{\mathcal{A}} \cup \{a\}) - F_K(\hat{\mathcal{A}}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda(x,y)\big(P_K^{\hat{\mathcal{A}}\cup\{a\}}(x,y) - P_K^{\hat{\mathcal{A}}}(x,y)\big)\, dx\, dy \quad (10)$$

Using probability rule of $\boldsymbol{Pr}(X \cup Y) = \boldsymbol{Pr}(X) + \boldsymbol{Pr}(Y) - \boldsymbol{Pr}(X \cap Y)$, we can write

$$P_K^{\hat{\mathcal{A}}\cup\{a\}}(x,y) - P_K^{\hat{\mathcal{A}}}(x,y) = P_K^{\{a\}}(x,y) - P_K^{\hat{\mathcal{A}}\cap\{a\}}(x,y)$$

Since an incident at point (x, y) being served by a bounded rank base from $\hat{\mathcal{A}}$ and set a are independent,

$$= P_K^{\{a\}}(x,y) - P_K^{\hat{\mathcal{A}}}(x,y) P_K^{\{a\}}(x,y) = P_K^{\{a\}}(x,y)(1 - P_K^{\{\hat{\mathcal{A}}\}}(x,y))$$

Substituting this expression in Equation (10) we get,

$$F_K(\hat{\mathcal{A}} \cup \{a\}) - F_K(\hat{\mathcal{A}}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda(x,y) P_K^{\{a\}}(x,y)(1 - P_K^{\{\hat{\mathcal{A}}\}})(x,y)\, dx\, dy \quad (11)$$

Similarly, $F_K(\mathcal{A} \cup \{a\}) - F_K(\mathcal{A})$ can be written as:

$$F_K(\mathcal{A} \cup \{a\}) - F_K(\mathcal{A}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda(x,y) \times P_K^{\{a\}}(x,y)(1 - P_K^{\{\mathcal{A}\}}(x,y))\, dx\, dy \quad (12)$$

From the above equations, relating marginal gains essentially reduces to finding the relation between probabilities, $P_K^{\{\hat{\mathcal{A}}\}}$ and $P_K^{\{\mathcal{A}\}}$. Since $P_K^{\{\hat{\mathcal{A}}\}} \geq P_K^{\{\mathcal{A}\}}$, we have:

$$F_K(\mathcal{A} \cup \{a\}) - F_K(\mathcal{A}) \geq F_K(\hat{\mathcal{A}} \cup \{a\}) - F_K(\hat{\mathcal{A}})$$

Hence *submodularity* of F_K is also proved. ■

5 Greedy Approximation

Given that F_K is monotone submodular, a greedy approach can provide a strong offline guarantee and good quality solutions. The greedy algorithm which assigns $|\mathcal{N}|$ ambulances to $|\mathcal{B}|$ bases using bounded rank metric F_K is given by Algorithm 2. At each iteration, we add an ambulance to a base that provides the maximum marginal benefit (in terms of F_K) over all the bases. The metric $F_K(\mathcal{A})$ is obtained using thinning of the non homogeneous Poisson process. $I_{(i,j)\in\mathcal{A}}$ is an indicator variable that indicates if $(i, j) \in \mathcal{A}$.

Since, we have to maximize a submodular function F_K with respect to a partition matroid constraint (i.e., one base for each ambulance), therefore, the greedy algorithm provides solutions that are at least 50% of optimal according to the following proposition due to Fisher *et al.*

[2]There is also the case of the same ambulance being allocated two bases. This is an invalid allocation, but we have to show submodularity even for this case. As indicated earlier, in such a case we define F_K assuming there are two copies of that ambulance.

Algorithm 2: BR-Greedy($\mathcal{R}, \mathcal{B}, N$)

Input: $\mathcal{R}, \mathcal{B}, \mathcal{N}$;
Output: $\mathcal{A}$ s.t. $|\mathcal{A}| = |\mathcal{N}|$,
$\forall i \in \mathcal{N} : \sum_{j \in \mathcal{B}} I_{(i,j) \in \mathcal{A}} = 1$;
begin
$\mathcal{A} \leftarrow \emptyset$;
for $i \in \mathcal{N}$ **do**
  for $j \in \mathcal{B}$ **do**
    $a_j \leftarrow (i, j)$;
    if $a_j \notin \mathcal{A}$ **then**
      $\delta_{a_j|\mathcal{A}} = F_K(\mathcal{A} \cup \{a_j\}) - F_K(\mathcal{A})$;
  $a^* \leftarrow \underset{j \in \mathcal{B}}{\operatorname{argmax}} \delta_{a_j|\mathcal{A}}$;
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{a^*\}$;
end
return $\mathcal{A}$

Proposition 2. *(Fisher, Nemhauser, and Wolsey 1978): Greedy algorithm for maximizing a monotone submodular function subject to a partition matroid yields solutions that are at least 50% of the optimal solution.*

While the *a priori* guarantee is 50% of optimal, we can provide a tighter *posteriori* guarantee as shown in proposition below.

Proposition 3. *If the optimal solution, $\mathcal{A}^*$ has m changes in allocation from a given solution $\mathcal{A}$, then*

$$F_K(\mathcal{A}^*) \leq F_K(\mathcal{A}) + m \cdot \delta_K^*(\mathcal{A})$$

where $\delta_K^*(\mathcal{A}) = max_a \Big[F_K(\mathcal{A} \cup \{a\}) - F_K(\mathcal{A}) \Big]$

Proof. The proof for the above proposition is a direct result of applying the greedy algorithm to a partition matroid (Goundan and Schulz 2007). For any monotone submodular function, $g : 2^{\prod} \rightarrow \mathbb{R}$ with optimal solution Z^*, we have:

$$g(Z^*) \leq g(Z) + \sum_{e \in Z^* \setminus Z} \delta_e(Z)$$

In the context of the metric, F_K this translates to:

$$F_K(\mathcal{A}^*) \leq F_K(\mathcal{A}) + \sum_{a \in \mathcal{A}^* \setminus \mathcal{A}} \delta_{K,a}(\mathcal{A})$$

While we do not know composition of $\mathcal{A}^*$, we can add best marginal value for every allocation to obtain upper bound for $F_K(\mathcal{A}^*)$. Since all ambulances are homogeneous and there are m such ambulances whose allocation is different (from the proposition definition), we have

$$F_K(\mathcal{A}^*) \leq F_K(\mathcal{A}) + m \cdot \delta_K^*(\mathcal{A})$$

where $\delta_K^*(\mathcal{A}) = max_a \Big[F_K(\mathcal{A} \cup \{a\}) - F_K(\mathcal{A}) \Big]$ ■

It should be noted that in the worst case, allocation for every ambulance is different in the optimal solution. Therefore, the worst case value of m is equal to $|\mathcal{N}|$, the total number of ambulances.

6 Comparing BR and BT in Simulation

Due to the dependence on time in bounded time (BT) objective, it is difficult to analytically represent the behavior of BT in steady state. On the other hand, as shown in previous sections, behavior of BR objective can be analytically studied in steady state. In order to compare the two objective values in non-steady state, we have to consider transient behavior. Therefore, we employ the simulation model as given in (Yue, Marla, and Krishnan 2012) and described in Section 2, as it can capture temporal dynamics of the underlying EMS system. In such a deterministic setting (with exact logs of emergency requests), the utility function or metric BR can be computed exactly and not in expectation. Now the metric BR is given by,

$$F_K^{BR}(\mathcal{A}) = \sum_{r \in \mathcal{R}} \sum_{i=0}^{K} F_{r,i}^{BR}(\mathcal{A}) \tag{13}$$

where,

$$F_{r,i}^{BR}(\mathcal{A}) = \begin{cases} 1 & \text{if request r served from base } b_i \in \mathcal{A} \\ 0 & \text{Otherwise} \end{cases}$$

There can be other variants of BR objective, where serving from top rank base can be prioritized ($w_i = \frac{1}{2^i}$):

$$F_K^{BRW}(\mathcal{A}) = \sum_{r \in \mathcal{R}} \sum_{i=0}^{K} w_i \cdot F_{r,i}^{BR}(\mathcal{A}) \tag{14}$$

In (Yue, Marla, and Krishnan 2012), the metric used is $F_T^{BT}(\mathcal{A}) = \sum_{r \in \mathcal{R}} F_{r,T}^{BT}(\mathcal{A})$ where,

$$F_{r,T}^{BT}(\mathcal{A}) = \begin{cases} 1 & \text{if response time for r } \leq \text{ T minutes} \\ 0 & \text{Otherwise} \end{cases}$$

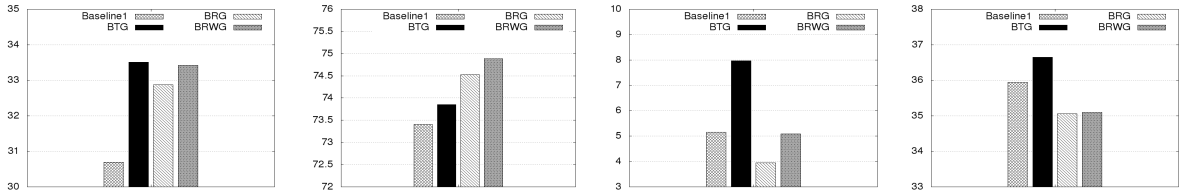
7 Experimental Results

We first describe the experimental setup and then describe the key results that demonstrate the utility of our greedy algorithm that optimizes BR objective.

We experimented with one simulated data set, *dataset-1* and a real dataset, *dataset-2* [3] is adopted from (Yue, Marla, and Krishnan 2012). Each request log in both data sets contains the following information (a) Incident 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) Round-about time for each of the feasible base stations. While these specific details might not always be readily available for real deployment, as indicated in (Ghosh and Varakantham 2016), we can estimate them using straightforward methods. We can compute set of feasible nearby bases and predict the response and round-off times for bases.

We consider three objectives, $F_K^{BR}(\mathcal{A}), F_K^{BRW}(\mathcal{A})$, and $F_T^{BT}(\mathcal{A})$ and use the greedy approximation approach of Algorithm 2 to optimize the three objectives.

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



(a) % served within 15 min. (b) % served within 30 min. (c) % not served (d) α^{th} percentile in min.

Figure 1: Performance comparison for dataset-2

	% in $\mathcal{T}$	% in $\mathcal{T}$ + 3.5 mins
BTG	5%	-0.7%
BRG	5.4%	0.0%
BRWG	6%	0.5%
	% in $\mathcal{T}$	% in $\mathcal{T}$ + 3.5 mins
BTG	5%	0.5%
BRG	4.8%	1.0%
BRWG	5.2%	1.0%
	% in $\mathcal{T}$	% in $\mathcal{T}$ + 3.5 mins
BTG	3.5%	0.3%
BRG	3.3%	0.5%
BRWG	3.7%	0.7%

Table 1: Percentage Improvement over Baseline with $|\mathcal{B}| - \exists$, $|\mathcal{B}|$ and ($|\mathcal{B}|$ + 11) Ambulances

	α	$(\alpha + 0.1)$
BTG	-0.1 mins	-0.6 mins
BRG	0.05 mins	0 mins
BRWG	0.1 mins	0.2 mins
	α	$(\alpha + 0.1)$
BTG	0.4 mins	0.25 mins
BRG	0.4 mins	0.3 mins
BRWG	0.4 mins	0.3 mins
	α	$(\alpha + 0.1)$
BTG	0.3 mins	0.4 mins
BRG	0.3 mins	0.4 mins
BRWG	0.3 mins	0.45 mins

Table 2: Reduction in α^{th} and $(\alpha + 0.1)^{th}$ percentile response time in minutes compared to Baseline with $|\mathcal{B}| - \exists$, $|\mathcal{B}|$ and ($|\mathcal{B}|$ + 11) Ambulances

The greedy algorithms that optimize these objectives are referred to as BRG (Bounded Rank Greedy), BRWG (Bounded Rank Weighted Greedy) and BTG (Bounded Time Greedy) . The objectives are then evaluated using the event-driven simulator of Algorithm 1 with respect to the bounded time and bounded risk performance metrics. As indicated earlier, the event-driven simulator employs the nearest idle ambulance dispatch policy.

Performance comparison on *dataset-1*

In this section, we provide comparison of all the approaches with respect to bounded time and bounded risk metrics on dataset-1. In this dataset, there are $|\mathcal{B}|$ base stations. We have request logs over a period of six months. We use first 3 month logs for training purpose to generate the ambulance allocation using different approaches and the performance is tested on request logs over the other 3 months. We show performance comparison between the three greedy approaches and a baseline allocation with respect to the two metrics. Baseline allocation here essentially represents an allocation that is derived based on historical load at stations.

We experimented with ambulance fleets of different sizes ($|\mathcal{B}|$-9, $|\mathcal{B}|$ and $|\mathcal{B}|$+11) allocated to $|\mathcal{B}|$ bases. Furthermore, we used different metric values for bounded time ($\mathcal{T}$ and $\mathcal{T}$+3.5 minutes)[4] and bounded risk (α and α+0.1 percentile). These results are shown in Tables 1 and 2. Here are the key observations:

- Both bounded rank greedy approaches (BRG and BRWG) perform better than the baseline approaches consistently with respect to bounded time as well as bounded risk metrics. However, bounded time greedy (BTG) fares worse than the baseline for all the objectives except bounded time of $\mathcal{T}$ minutes for which it is optimized, in the experiment with $|\mathcal{B}|$- 9 ambulances. BRWG outperforms BTG consistently with respect to both bounded time and bound risk metrics.
- The difference between BRWG and BTG in terms of bounded time (around 1.1%) and bounded risk response time (30-40 seconds) does not seem significant. However, qualitatively 1% amounts to serving about 6-7 more requests per day within the $\mathcal{T}$ minute

[4]The bounds are decided through some preliminary experiments. A more thorough theoretical and empirical analysis for setting of bounds given KPI is left for future work.

m	Online bound
$\|\mathcal{B}\| + 11$	75%
$\frac{1}{2} \cdot (\mathcal{B} + 11)$	90%
10	95%
0	100 %

Table 3: Online Bound

mark and 30 seconds improvement is equivalent to taking five ambulances out of service while retaining the response time.

Posteriori Bounds for BRG Here, we demonstrate that the *posteriori* (online) guarantee, obtained by employing Proposition 3, is significantly better than the *a priori* guarantee (of 50% from optimal) for BRG on a given training set of data. It should be noted that the number of differences with optimal will be equal to number of ambulances only in pathological cases that are created synthetically. In real problems, the value of m is expected to be much lower (as ambulances are homogeneous). Therefore, we calculated online bound for multiple different values of m.

Table 3 illustrates the quality guarantee in the case where we have $|\mathcal{B}|$+11 ambulances over $|\mathcal{B}|$ bases. It should be noted that even when $m = |\mathcal{B}| + 11$, the guarantee is 75% of optimal and if the number of differences with optimal was half of $|\mathcal{B}|$+11, then the guarantee is 90% of optimal. The online guarantees for other settings and data sets were very similar.

Performance on dataset-2

In this section, we provide comparison of all the approaches with respect to bounded time and bounded risk metrics on dataset-2. In this dataset, there are 58 ambulances and 58 base stations. The best allocation is obtained using 500 training logs, validated on 500 logs and is tested on 500 test logs. We present performance of the greedy approaches and the Baseline1 (1 ambulance at each base station) in Figure 1 over four metrics (similar to Yue et al.). Here are the key observations:

- BRG and BRWG perform better than the baseline approach with respect to all the four metrics whereas (similar to dataset-1) BTG fared worse than the baseline approach for bounded risk metric and for % of requests not served. BRG served around 1.1% more requests than the baseline.
- BRWG served 1% more requests than BTG in 30 minutes. In terms of 80th percentile as well, both BRG and BRWG outperformed baseline and BTG by around 1 and 2 minutes, respectively.

8 Conclusion

We have modeled the spatio-temporal uncertainty in emergency requests using a non-homogeneous Poisson point process. We presented a novel indirect objective based on number of incidents served from bounded rank base stations and showed it to be monotone submodular. It enabled us to provide greedy approximation for Bounded Rank (BR) objective with 50% offline guarantee and much tighter *posteriori* guarantee. More importantly, the elegant theoretical guarantee in the model also translates to improved performance on simulators validated on two real data sets. We demonstrated that our BR greedy algorithm consistently performs better than the existing and baseline approaches.

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.

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

Fisher, M. L.; Nemhauser, G. L.; and Wolsey, L. A. 1978. An analysis of approximations for maxi- mizing submodular set functions - II. *Math. Prog. Study* 8:73–87.

Ghosh, S., and Varakantham, P. 2016. Strategic planning for setting up base stations in emergency medical systems. In *ICAPS*, 385–393.

Goundan, P. R., and Schulz, A. S. 2007. Revisiting the greedy approach to submodular set function maximization. *Optimization Online*.

Ibri, S.; Nourelfath, M.; and Drias, H. 2012. A multi-agent approach for integrated emergency vehicle dispatching and covering problem. *Engineering Applications of Artificial Intelligence* 25(3):554–565.

Larson, R. C. 1974. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research* 1(1):67–95.

Lee, E. K.; Maheshwary, S.; Mason, J.; and Glisson, W. 2006. Large-scale dispensing for emergency response to bioterrorism and infectious-disease outbreak. *Interfaces* 36(6):591–607.

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.

Pagnoncelli, B.; Ahmed, S.; and Shapiro, A. 2009. Sample average approximation method for chance constrained programming: theory and applications. *Journal of optimization theory and applications* 142(2):399–416.

Peleg, K., and Pliskin, J. S. 2004. A geographic information system simulation model of ems: reducing ambulance response time. *The American journal of emergency medicine* 22(3):164–170.

Ross, S. M. 2010. *Introduction to Probability Models*. Academic Press, tenth edition.

Saisubramanian, S.; Varakantham, P.; and Lau, H. C. 2015. Risk based optimization for improving emer-

gency medical systems. In *Proceedings of Twenty-Ninth AAAI Conference on Artificial Intelligence*, 702–708.

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

Varakantham, P., and Kumar, A. 2013. Optimization approaches for solving chance constrained stochastic orienteering problems. In *International Conference on Algorithmic DecisionTheory*, 387–398. Springer.

Yue, Y.; Marla, L.; and Krishnan, R. 2012. An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment. In *Proceedings of Twenty-Sixth AAAI Conference on Artificial Intelligence*.