Robust Ambulance Allocation Using Risk-based Metrics

Kaushik Krishnan and Lavanya Marla
Department of Industrial and Systems Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801
Email: {kkrishn3,lavanyam} @illinois.edu

Yisong Yue
Department of Computer Science
California Institute of Technology
Pasadena, California
Email: yyue@caltech.edu

***Abstract*—This paper focuses on robust location strategies for a fleet of ambulances in cities in order to maximize service levels under unexpected demand patterns. Our work is motivated by the fact that when small parts of networks incur emergencies according to a heavy-tailed distribution, the structure of the network under resource constraints results in the entire system behaving in a heavy-tailed manner. To address this, metrics other than average-case need to be used. We achieve robust location strategies by including risk metrics that account for tail behavior and not average performance alone. Because of the exponentially large solution space for locating K ambulances in N locations on the network, our approach is based on an efficient algorithm that allows for optimizing based on these risk metrics. We show that optimizing based on risk measures can account for spatio-temporal patterns and prevent the extent of delay cascades that are typically seen in heavy-tailed arrival distributions. From our computational results based on data from a large Asian city, we show that planning with some robustness metrics as targets leads to solutions that perform well in heavy-tailed demand scenarios.**

I. Introduction

Emergency medical services (EMS) form a critical component of a city's infrastructure. These services typically use the available transportation infrastructure of the city and as such are integral components of the services that the transportation infrastructure is meant to provide. Transportation infrastructure and Emergency Medical Systems (EMS) are interdependent in that EMS-type services are necessary for the users of transportation services when emergencies or disasters occur; and similarly, transportation infrastructure is necessary for EMS services to reach their target population.

The important resource allocation question we deal with in this paper is that of the location of ambulances in a city's network to improve service levels. This is an important question because of the spatio-temporal nature of the emergencies occurring in a system. Due to this spatio-temporal nature, the resource constraints on ambulances and the network structure of the city, there will be cascading effects of the calls causing the service level of each call to be dependent on the previous calls and the ambulances assigned to them. We discuss this in detail the following section.

The cascading dependencies occurring in the system are usually modeled making the assumption of spatio-temporal Poisson arrivals, and focus on metrics of average service levels, average survival rate and average throughput. Examples of these studies include [1], [2], [3], [4]. As a result, most practitioners focus on the overall service levels and not the robustness of the allocations under uncertain conditions.

In this paper, we are interested specifically in highly resource-constrained settings such as the ones occurring in emerging economies, that lead to potentially more spatio-temporal cascading behavior due to resource limitations. Additionally, we focus on a balance between average and worst-case metrics and show that these lead to more robust allocations as defined by risk metrics.

The contributions of our paper are as follows.

1) We present the first data-driven modeling and solution approach that incorporates risk metrics in the objective function.
2) Because most risk metrics do not satisfy clear properties of consistency, we use the Conditional Value-at-Risk (CVaR) that is used in the EMS context for the first time. In particular, our objective function trades off the mean and the CVaR values in its exploration of the solution.
3) Through computational experiments we demonstrate that optimizing the ambulance allocation with objective functions that include risk metrics generates solutions that are more robust in their tail behavior than solutions optimized using the average. This results in a better prevention of cascading behavior over networks, particularly in the case of heavy-tailed distributions.

II. Literature

The study of vehicle allocation (or deployment) for Emergency Medical Services (EMS) enjoys a rich history [5], [6], [7]. Broadly speaking, the general problem setting can be described simply as computing a location, redeployment or dispatch strategy for a set of ambulances such that some measure of 'fitness' that represents the service level of the system is optimized. Two natural areas for resource allocation are in allocation of ambulance to bases, and dispatching of ambulance to requests. Our paper focuses on optimizing the location of ambulances to bases, typically because it affords the highest gains.

[7] present a survey on static ambulance location and dynamic redeployment of ambulances. Multiple types of ap-

978-1-4673-9622-6/16/$31.00 ©2016 IEEE

proaches are used - the primary ones being deterministic models, probabilistic models and heuristics. Deterministic models are based on integer programming-type techniques, which do not capture the stochastic considerations regarding ambulance availability. Probabilistic models are typically queueing-based models that capture more of the system dynamics [4], [8], [1]. The hypercube models in [9], [4] and [8] were some of the first and seminal works to use these techniques applied to emergency vehicle location and public systems. These approaches are used to quickly identify, in the large solution space, solutions that can produce good results. The solutions thus identified are then evaluated in further detail using event-based simulation techniques. [1] build on these approaches to capture the dynamics of the system more accurately, and show that the subset of solutions chosen by their approach can improve on those identified by the original hypercube model proposed by Larson. [10] use a version of the approx hypercube model that allows capturing station-specific busy probabilities, and allows multiple vehicles at a station. They provide a theoretical convergence guarantee for a restricted special case. [11] embed a hypercube model into a genetic algorithm for potentially dispatching multiple ambulances to a call as needed, and tailored for highways where only potentially the first nearest and second nearest ambulances may be dispatched to a call. [12] present models for ambulance allocation and districting on highways; using a hypercube model embedded in a genetic algorithm. Stochastic programming techniques that bridge deterministic and probabilistic models have been used by [3] to capture more of the system dynamics when ambulances are assigned to calls and are hence absent.

While probabilistic models capture the system dynamics and the impact of allocated ambulances through queueing-based models, deterministic models using techniques like integer programming-based model sometimes capture these by using notions such as 'coverage' or 'double coverage' of a demand point by an ambulance [13]. The measures and objective functions used in these models are approximations to the behavior of the system and not always related to the evaluation metrics of interest.

Thus such mathematical programming approaches often fail to characterize completely the dynamics of ambulance dispatch and emergency response in general. To evaluate the true metrics of interest, simulation-based *evaluation* is often used. Almost all modeling approaches in the literature employ simulations for final evaluation of a small selection of solutions from optimization, queueing or heuristic models [1], [14], [15].

The first motivation for our simulation-optimization approach is due to this implicit preference for simulation-based evaluation as the evaluation method of choice. Simulation as a tool to help in optimization has been used in a few works, such as [6], [16], [17]; however, theoretical guarantees on performance are not provided except in [18].

The second motivation for our research arises from the fact that the vast majority of the literature on emergency medical system planning in cities has focused on optimizing for long-term average metrics. Limited work, such as [19] and [20], has used metrics that are related to risk or tail behavior of the system. It is also well-known that clusters of emergencies can take the shape of disasters that are characterized by heavy-tailed distributions [21], and when a disaster or large casualty occurs in even a small part of the service area, it is important to evaluate its impact on the entire network.

III. Motivation

Data from various sources have shown us that inter-arrival times of emergencies follow light-tailed distributions, typically modeled as Poisson arrival rates with exponential inter-arrival times. Therefore, this is the distribution most used in the literature. The patterns of calls, interacting with the placement (locations) of ambulances and the geometry or configuration of the city, causes cascading dependencies. Previous work [18] has shown that effective placement of ambulances, as opposed to naive location, can significantly reduce these dependencies and improve overall service levels.

We illustrate our point considering the example of data from a large Asian city. The city contains 83 sub-city districts. While the city has required us to keep the data and source confidential, we see that all sub-districts in the city of interest have arrival rates that can be modeled as Poisson arrival rates. When these are plotted temporally, the entire network's call stream remains a Poisson distribution. However, when as few as six sub-districts begin to follow a heavy-tailed distribution (of inter-arrival times), the entire call stream also follows a heavy-tailed distribution (see Table I).

We now compare the performance of these allocations using a data-driven discrete-event simulator, based upon the simulator described in [18]. The procedure followed in the simulator and the embedded dispatch process are described in Algorithm 2 and Algorithm 1 respectively. The simulator captures the fact that calls are served on a first-come-first-serve basis, the nearest free ambulance is dispatched to each call, and the service level of each call as the consequence of the dispatched ambulance. Ideally, every request should be assigned to its highest priority ambulance. However, such ambulances may not be available if they are still servicing previous requests. This creates what we refer to as a *dependency* between two requests.

We first define some notation. Let $R = \{r_1, \ldots, r_N\}$ be a request log with a sequence of requests, A be the allocation vector of ambulances to bases, and DISPATCH be the nearest-free-ambulance dispatch policy described in Algorithm 1. Also let y_r be the base of ambulance dispatched to service request r ($\perp$ if no ambulance was dispatched), $r(y_r)$ denote the active call r to which ambulance y_r is dispatched, and $\bar{t}_r(y_r)$ be the completion time of request r. Then, informally, we say that request r depends on request r' if the assignment of $y_{r'}$ to r' causes r to be assigned y_r such that $y_{r'} \succ_{p_r} y_r$ [18]. The formal definitions follow.

Definition 1: There exists an active dependency $\gamma_{r,r',y_{r'}}$ from request r to request r' with label $y_{r'}$ if

1) $t_{r'} < t_r$ (r' arrives before r)

TABLE I
COMPARING CALL STREAMS FROM LIGHT-TAILED AND HEAVY-TAILED DISTRIBUTIONS

	Call Log 1	Call Log 2
SUB-CITY DISTRICTS (LIGHT-TAILED)	83	77
SUB-CITY DISTRICTS (HEAVY-TAILED)	6	0
DISTRIBUTION	Poisson	Weibull
PARAMETERS	Rate = 0.28	Shape = 0.97, Scale = 3.9

Algorithm 1 First-come First-served Dispatch Policy

input: current request r, Available ambulances W
for $a \in r.q$ in decreasing preference order **do**
  if $a \in W$ **then**
    return: a
  end if
end for
return: $\perp$

2) $\bar{t}_{r'}(y_{r'}) > t_r$ (r' completes after r arrives – this indicates that the two requests "overlap" in time)
3) $y_{r'} \succ_{p_r} y_r$ (r' is assigned an ambulance from a higher priority base, w.r.t. r's priority queue, than the ambulance ultimately assigned to r)

The dependency structure in the network is dependent on the call logs, the ambulance allocation and the dispatch policy. A 'good' allocation will reduce the cascading nature of the dependencies among the calls, allowing more calls to be served by a 'closer' ambulance with a better service level [18].

Because cascading dependency behavior is seen even in Poisson arrivals (exponential inter-arrival times), these dependencies are exacerbated even more when arrivals follow a (even partial) heavy-tailed distribution. We consider the naive allocation that the operator uses and evaluate the two call logs (see Table II).

Therefore it becomes more important to have an allocation that optimizes based on not only overall system performance but also the performance at the tail of the distribution.

IV. MODELING APPROACH

Our approach focuses on optimizing allocations on the network using a combination of expected value metrics and risk metrics. In particular, as a risk metric, we consider the metric of Conditional-Value-at-Risk (CVaR) due to its properties of coherence [22], [23]. For a general loss function described by random variable X and $0 < \alpha < 1$, $CVaR$ is defined as $CVaR_\alpha = \frac{1}{\alpha}\int_{1-\alpha}^{1} VaR_\alpha(X)d\alpha$ where VaR_α is the Value-at-risk (VaR). This can be equivalently written as:

$$CVaR_\alpha = -\frac{1}{\alpha}\left(E[X\ 1_{\{X \geq x_\alpha\}}] + x_\alpha(\alpha - P[X \geq x_\alpha])\right) \tag{1}$$

where $x_\alpha = \sup\{x \in \mathbb{R} : P(X \geq x) \geq \alpha\}$ is the upper α-quantile and $1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$ is the indicator function.

Studies that capture risk in ambulance allocation, such as [19] and [20] typically use chance-constraints ([24], [25]) to capture risk. Chance-constraints are typically satisficing constraints that are similar in properties to the Value-at-risk measure. However, VaR is not a coherent measure of risk, as is discussed in [22]. Therefore we resort to optimizing the Conditional-Value-at-Risk, which is a coherent risk measure. In particular, we want to minimize the expected value of the upper α-th quantile of the loss function of interest, or maximize the expected value of the lower α-th quantile of the corresponding gain function.

Algorithm 2 SIMULATOR: Data-driven Simulator Method

input: (R, A, π, t_d), DISPATCH
$W \leftarrow A$ *//keeps track of which ambulances are free*
$\hat{R} \leftarrow \emptyset$ *//keeps track of active requests*
initialize $Y = \{y_r\}_{r \in R}$ such that $y_r \leftarrow \perp$
initialize events $\mathcal{E} \leftarrow R$ sorted in arrival order
insert redeployment events spaced every t_d minutes to $\mathcal{E}$.
while $|\mathcal{E}| > 0$ **do**
  remove next arriving event e from $\mathcal{E}$
  if $e =$ new request r **then**
    $y_r \leftarrow$ DISPATCH(r, W, R) *//dispatch policy*
    if $y_r \neq \perp$ **then**
      $\hat{R} \leftarrow \hat{R} + r(y_r)$ *//updating active requests*
      $W \leftarrow W - y_r$ *//updating free ambulances*
      insert job completion event at time $\bar{t}_r(y_r)$ into $\mathcal{E}$
    end if
  else if $e =$ job completion event $\bar{t}_r(y_r)$ **then**
    $\hat{R} \leftarrow \hat{R} - r(y_r)$ *//updating active requests*
    $W \leftarrow W + y_r$ *//updating free ambulances*
  else if $e =$ redeployment event **then**
    $W \leftarrow \pi(W, \hat{R})$ *//redeploying free ambulances*
  end if
end while
return: Processed assignments of ambulances to requests Y

As seen in [26], CVaR can be represented as the maximum of submodular functions that are parameterized by a smooth parameter. In combination with the observation made in [18] that the loss function represented the ambulance allocation is approximately submodular, a linear combination of these functions is likely to be well-solved by algorithms that are built for optimizing submodular functions. We discuss this in detail below.

More formally, let A denote an allocation of ambulances to a set of bases $\mathcal{A}$ (there can be more than one ambulance at a base). We represent A as a multiset of elements in $\mathcal{A}$. Let $M(\mathcal{A})$ denote the multi-powerset of $\mathcal{A}$ and $L(A)$ as the cost of allocation A. Correspondingly, the gain is defined as $L(\emptyset)$ - $L(A)$ by comparing to the null allocation. $L(A)$ can incorporate any loss metric corresponding to the way calls are

TABLE II
NAIVE ALLOCATION PERFORMANCE ON LIGHT-TAILED AND HEAVY-TAILED CALL LOGS

	Call Log 1	Call Log 2
MEAN OF NUMBER OF CALLS NOT SERVED	8.87%	10.23%
90TH QUANTILE OF CALLS NOT SERVED	10.11%	12.58%

serviced by the positioning of ambulances A [18]. Examples include: (a) the fraction of requests not served, (b) the fraction of requests whose service time is above some target threshold, or (c) the fraction of requests served at each service level.

Our goal is to maximize the weighted sum of expected gain and CVaR of gain, over some distribution of requests (assumed to follow probability distribution $\mathbf{P}(R)$, and as represented by the sampled request logs). That is, we want to find A that maximizes $F(A) = (\beta * (E(L(\emptyset)) - E(L(A))) + (1-\beta) * (CVaR_\alpha L(\emptyset) - CVaR_\alpha L(A))$.

We define L using the outcomes of simulated requests over several request logs. Using the simulator in Algorithm 2, we measure the expected metrics of interest over the set of training request logs. Let $Y_R = \{y_r\}_{r \in R}$ denote the output of Algorithm 2 for request log R. Then we write the expected value of the loss function over the request logs as

$$\mathbf{E}_{L(A)} = \mathbf{E}_{R \sim \mathbf{P}(R)} \left[\sum_{r \in R} L_r(y_r) \right], \tag{2}$$

where $L_r(y)$ is the penalty of assigning request r with y_r (e.g., whether or not assigning ambulance y_r to r results in a service time above a target threshold). Similarly the CVaR at the αth protection level can be written as

$$CVaR_{L(A)} = \mathbf{E}_{R \sim \mathbf{P}(R)} \left[\left(\sum_{r \in R} L_r(y_r) | \sum_{r \in R} L_r(y_r) > VaR_\alpha L(A) \right) \right], \tag{3}$$

In practice, we resort to optimizing over a collection of request logs $\mathcal{R} = \left\{ \{R_{mn}\}_{m=1}^{M} \right\}_{n=1}^{N}$, where each request log $R_m \in \mathcal{R}$ is sampled i.i.d according to $\mathbf{P}(R)$. In our experiments, we use Sample Average Approximation [27] to bound the difference between our sample average objective and the optimal expected performance, by approximating the expectation with the sampled average and the CVaR with the CVaR over the set of sampled request logs.

Let $\delta_F(a|A)$ denote the gain of adding a to A, defined as

$$\delta_{E(L)}(a|A) = L(A) - L(A \cup a); \tag{4}$$
$$\delta_{CVaR_\alpha(L)} = CVaR_\alpha(L(A)) - CVaR_\alpha(L(A \cup a)); \tag{5}$$
$$\delta_F(a|A) = \beta * \delta_{E(L)}(a|A) + (1-\beta)\delta_{CVaR_\alpha(L)}. \tag{6}$$

$\delta_{E(L)}(A)$ corresponds to the expected value of the dependency chains broken by the allocation A compared to the null allocation, and $\delta_{CVaR_\alpha(L)}$ corresponds to the conditional-value-at-risk of the dependency costs, that is their expected loss greater than the αth quantile.

Algorithm 3 Greedy Ambulance Allocation

```
input: F, K
A ← ∅
for ℓ = 1, ..., K do
    â ← argmax_a δ_F(a|A)      //see (6)
    A ← A + â
end for
return: A
```

Given a budget of K ambulances, the static allocation goal then is to select the ambulance allocation A (with $|A| \leq K$) such that the utility $F(A)$ is maximized. More formally, we can write our optimization problem as

$$\underset{A \in M(\mathcal{A}): |A| \leq K}{\operatorname{argmax}} \quad F(A). \tag{7}$$

We employ the greedy algorithm presented in [18], because the properties of *approximate* submodularity still hold as discussed above. The greedy algorithm is described in Algorithm 3. The algorithm iteratively selects the ambulance a that has maximal incremental gain to the current solution until K ambulances have been allocated. Note that each evaluation of $\delta(a|A)$ requires running the simulator to evaluate $F(A+a)$.

V. COMPUTATIONAL RESULTS

In this section, we present our computational results on data from the large Asian city described in Section III. The usage data contains approximately ten thousand logged emergency requests over the course of one month. Each record in the request log contains the type and location of the request, the ambulance (if any) that was dispatched, and the various travel times (e.g., base to scene, scene to hospital, etc). The request arrival rates fit typically into Poisson distributions (and inter-arrival times into exponential distributions) per sub-city-district and service times fit into lognormal distributions, respectively. Request arrivals and service times all appear statistically independent. However, certain sub-city-districts also have inter-arrival distributions that might also be fit to heavy-tailed distributions (specifically the Weibull distribution) in a statistically significant manner. We will therefore examine if the difference in the assumptions behind the distributions (which makes the sampling consistent with real-world data) results in solutions that are robust to heavy tailed arrival rates.

In particular, we run our optimization model described in 3 with training call logs following the Poisson distribution, and test call logs sampled according to the following cases:

- Poisson call arrivals (perfect information about distributions)
- Weibull call arrivals in sub-city-districts that can be fit to Weibull (call arrivals are heavy-tailed)

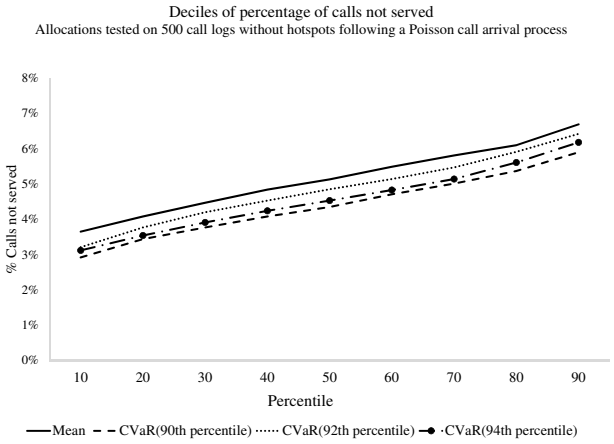

Fig. 1. Performance of risk-optimized allocation on light-tailed request logs

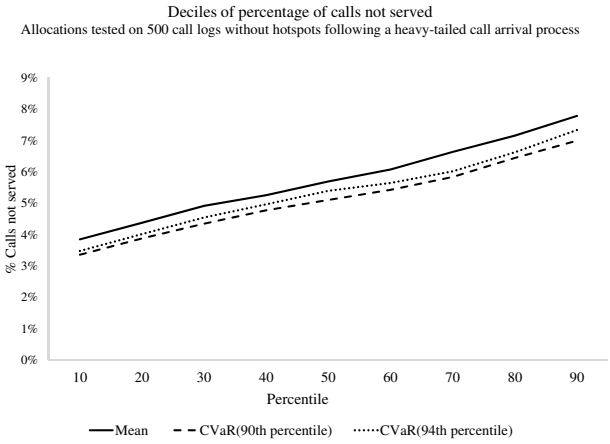

Fig. 2. Performance of risk-optimized allocation on heavy-tailed request logs

- Poisson call arrivals with hotspots in some sub-city-districts (chosen such that high arrival areas are simultaneously stressed)
- Weibull call arrivals with hotspots in some sub-city-districts (chosen such that high arrival areas are simultaneously stressed)

Using the parameters of the fitted Poisson distributions, we built a generative Poisson process model for sampling emergency requests. Our action space contains 58 bases and 58 ambulances. We evaluate our methods over a period of one week. 500 training call logs and 500 test call logs, each spanning one day, and independent of each other, are used.

We consider the following cost function in our experiments.

$$L_r(y) = \begin{cases} 1 & \text{if service time} \geq 30min \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Our metrics are the various quantiles of non-service metrics over the test request logs (in order to evaluate the tail probabilities of failure/non-service), as well as the mean performance. We use β = 0.7 and solve for varying values of α (tail CVaR values).

Figures 1 and 2 present the improvement in the tail metrics (calls not served) for varying values of protection levels α when the test call logs are Poisson and Weibull respectively. Figures 3 and 4 present the improvement in the tail metrics (calls not served) for varying values of protection levels α when the test call logs are Poisson and Weibull respectively, and hotspots in call arrivals occur, causing system stress.

Our results show that optimizing with risk-metrics provides improved results on tail-related metrics such as service failure probabilities, as compared to optimizing using expected value-based metrics (such as mean) alone. Particularly in the case of imperfect information, such as training with a distribution that is different from the test distribution, or with hotspots occurring in the test cases but not in the training cases used for optimization, we found that the results improved upon optimizing using expected-value-based metrics alone.

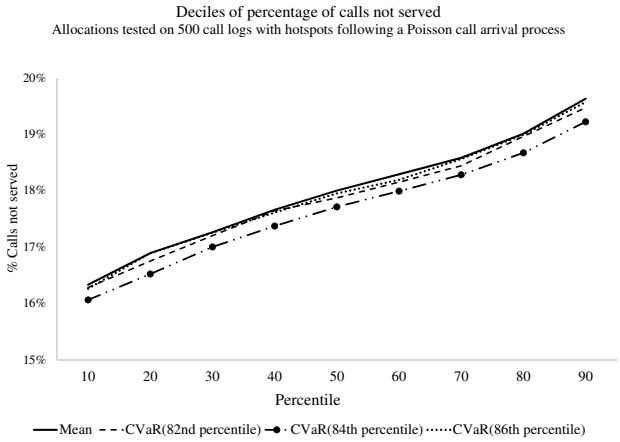

Fig. 3. Performance of risk-optimized allocation on light-tailed request logs with hotspots

We also observed showed better improvements in the case of heavy-tailed distributions, indicating the need for optimizing explicitly for such systems using risk measures.

As a caveat, we find that while choosing the right protection level α for the tail probabilities is not obvious, further research can be conducted to optimize the choice of α. Additionally, the risk-based objective function suggested in this paper should be studied for its theoretical properties and bounds similar to [18] can be proposed.

VI. Conclusion

We have presented an efficient and effective approach to ambulance fleet allocation that is data-driven and balances expected value-based and risk-based metrics. In simulation experiments based on a real EMS system in Asia, this approach improved upon tail performance metrics measured using a data-driven discrete-event simulator. This showed better improvements in the case of heavy-tailed distributions, indicating the need for optimizing explicitly for such systems. Further research will need to be conducted to examine the theoretical

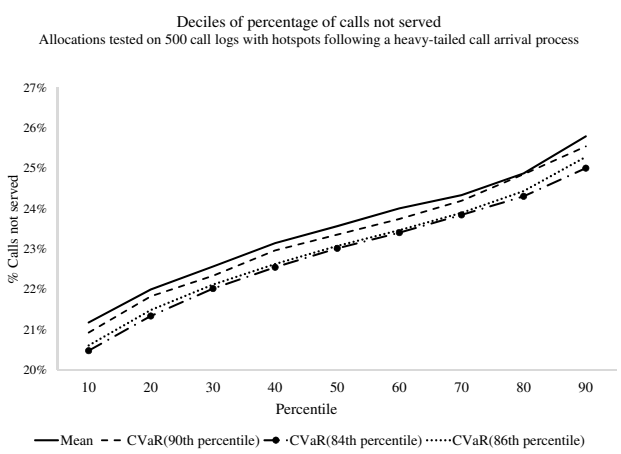

Fig. 4. Performance of risk-optimized allocation on heavy-tailed request logs with hotspots

properties of the objective function as well as to bound the performance of this algorithm under various risk metrics including worst-case-based and tail-protection-based metrics.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments that greatly improved the presentation in this manuscript.

REFERENCES

[1] M. Restrepo, S. Henderson, and H. Topaloglu, "Erlang loss models for the static deployment of ambulances," *Health care management science*, vol. 12, no. 1, pp. 67–79, 2009.

[2] M. Restrepo, "Computational methods for static allocation and real-time redeployment of ambulances," Ph.D. dissertation, Cornell University, 2008.

[3] M. O. Ball and F. L. Lin, "A reliability model applied to emergency service vehicle location," *Operations Research*, vol. 41, no. 1, pp. 18–36, 1993, special Issue on Stochastic and Dynamic Models in Transportation.

[4] R. Larson, "A hypercube queueing model for facility location and redistricting in urban emergency services," *Computers and Operations Research*, vol. 1, pp. 67–75, 1974.

[5] A. J. Swersey, "The deployment of police, fire and emergency medical units," in *Operations Research and the Public Sector. In: Handbooks in Operations Research and Management science*, S. Pollock, M. Rothkopf, and A. Barnett, Eds. Amsterdam: North-Holland, 1994, pp. 151–200.

[6] S. G. Henderson and A. J. Mason, "Ambulance service planning: simulation and data visualization," in *Operations Research and Health Care: A Handbook of Methods and Applications*, F. S. M. L. Brandeau and W. P. Pierskalla, Eds. Boston: Kluwer Academic, 2004, pp. 77–102.

[7] L. Brotcorne, G. Laporte, and F. Semet, "Ambulance location and relocation models," *European journal of operational research*, vol. 147, no. 3, pp. 451–463, 2003.

[8] R. Larson, "Approximating the performance of urban emergency service systems," *Operations Research*, vol. 23, pp. 845–868, 1975.

[9] R. Larson and K. Stevenson, "On insensitivities in urban redistricting and facility location," *Operations Research*, pp. 595–612, 1972.

[10] S. Budge, A. Ingolfsson, and E. Erkut, "Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location," *Operations Research*, 2007.

[11] A. Iannoni, R. Morabito, and C. Saydam, "A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways," *Annals of Operations Research*, vol. 157, no. 1, pp. 207–224, 2008.

[12] ——, "An optimization approach for ambulance location and the districting of the response segments on highways," *European Journal of Operational Research*, vol. 195, no. 2, pp. 528–542, 2009.

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

[14] M. Maxwell, M. Restrepo, S. Henderson, and H. Topaloglu, "Approximate dynamic programming for ambulance redeployment," vol. 22, no. 2, pp. 266–281, 2010.

[15] S. Budge, A. Ingolfsson, and E. Erkut, "Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location," *Operations Research*, vol. 57, pp. 251–255, 2009.

[16] A. Ingolfsson, E. Erkut, and S. Budge, "Simulation of single start station for edmonton ems," *Journal of the Operational Research Society*, vol. 54, pp. 736–746, 2003.

[17] M. Lubicz and Z. Mielczarek, "Simulation modeling of emergency medical services," *European Journal of Operational Research*, vol. 29, pp. 178–185, 1987.

[18] Y. Yue, L. Marla, and R. Krishnan, "An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment," in *AAAI Conference on Artificial Intelligence (AAAI)*, July 2012.

[19] S. Subramanian, P. Varakantham, and H. Lau, "Risk based optimization for improving emergency medical systems," *Computers and Operations Research*, vol. 1, pp. 67–75, 1974.

[20] N. Noyan, "Alternate risk measures for emergency medical service system design," *Annals of Operations Research*, vol. 181, no. 1, pp. 559–589, 2010.

[21] V. Pisarenko and M. Rodkin, *Heavy-Tailed Distributions in Disaster Analysis*, ser. Advances in Natural and Technological Hazards Research. Springer Netherlands, 2010. [Online]. Available: https://books.google.com/books?id=3yKQni8OLj4C

[22] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[23] ——, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[24] A. Charnes and W. W. Cooper, "Chance constrained programming," *Management Science*, vol. 6(1), pp. 73–79, 1959.

[25] ——, "Deterministic equivalents for optimizing and satisficing under chance constraints," *Operations Research*, vol. 11(1), pp. 18–39, 1963.

[26] T. Maehara, "Risk averse submodular utility maximization," *Operations Research Letters*, vol. 43, no. 5, pp. 526–529, 2015.

[27] B. Verweij, S. Ahmed, A. Kleywegt, G. Nemhauser, and A. Shapiro, "The sample average approximation method applied to stochastic routing problems: a computational study," *Computational Optimization and Applications*, vol. 24, no. 2, pp. 289–333, 2003.