An Efficient Data-driven Approach for Emergency Medical Services

Lavanya Marla

University of Illinois at Urbana-Champaign

Collaborators: Prof. Ramayya Krishnan (CMU), Dr. Yisong Yue (Caltech)

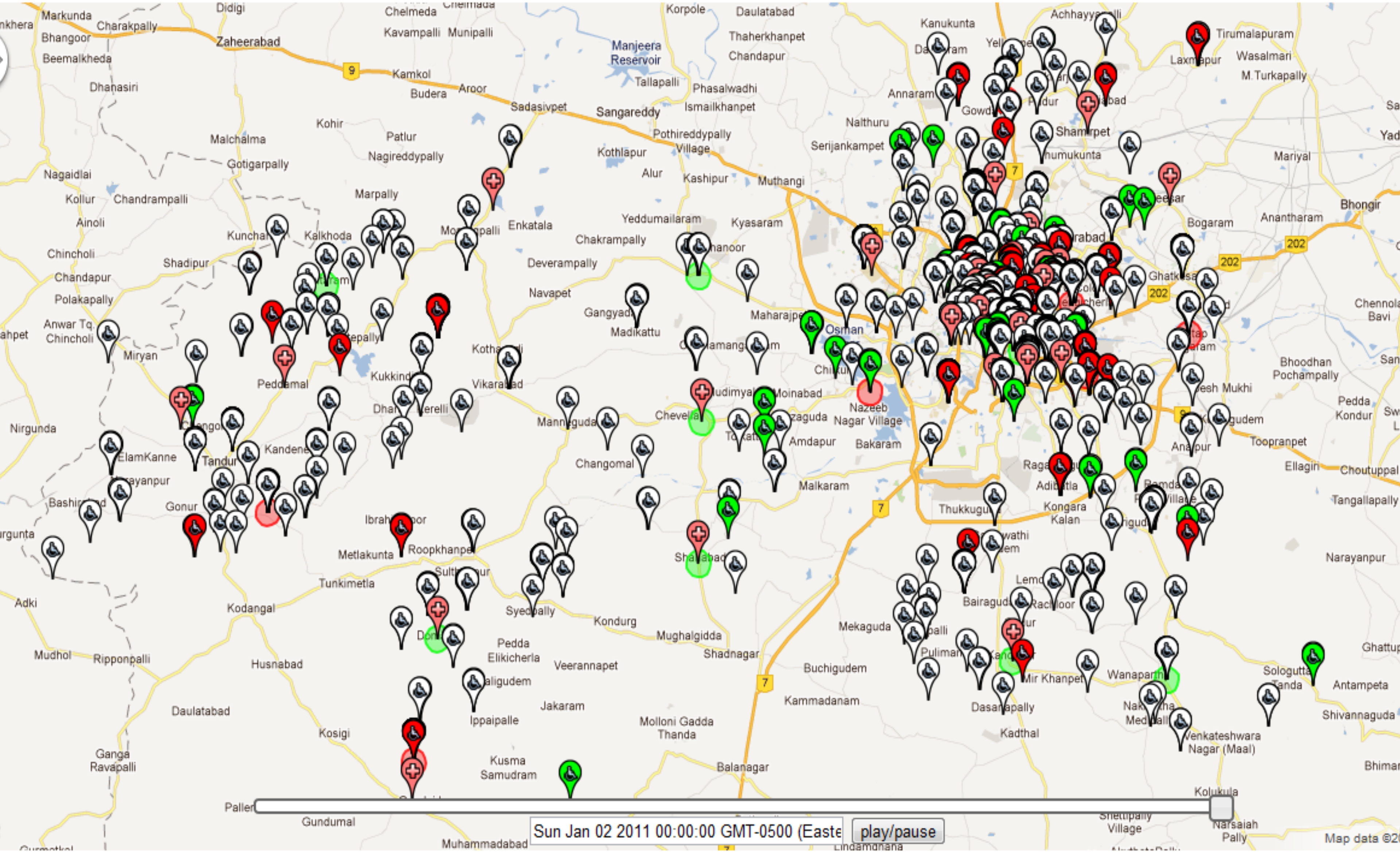
Talk outline

- Ground Realities for EMS in Emerging Economies
- Data-driven Simulation
- Mathematical Formulations
- Results
- Ongoing and Future Work

EMS in Emerging Economies: Ground Realities

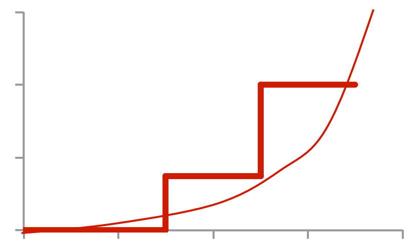
- Highly resource constrained
  - 75M people, 750 ambulance bases (AP)
- Large-scale
- Prior to this operator, no central ambulance provider
  - Hospital ambulances, taxis
- Public-private partnership
  - No fees charged for service (paid by state)
- Cell-phone-based communications
- DATA COLLECTION

Service Area, Bases and Calls



Inefficiencies in spite of sophisticated models

Existing literature: medium-scale



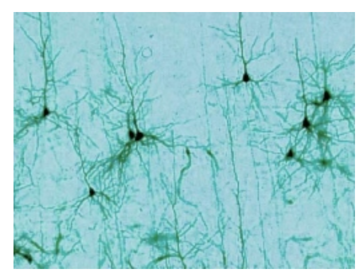
Non-linearities between survival and service time

Multiple heterogeneous resources – ALS, BLS

ALS

'Discrete'

BLS



Network effect – Propagation effect of ambulances in use

Source: http://www.castlelab.princeton.edu/transportationlogistics.htm, http://sbb.ch, www.colinfahey.com

Challenges in EMS in Emerging Economies

- Traffic congestion
  - Public acceptability
    - Clear traffic for ambulance
- Competition with ad-hoc networks
  - Decreases utilization of ambulances
- No real-time position availability
- New cities
  - New traffic patterns
  - New modes of transport

Key Questions of Interest

- How can performance be improved using existing resources (e.g., ambulances)?
  - Static allocation?
  - Dynamic redeployment?
  - Change dispatch policy?
- How to characterize the state of the system?
  - Metrics
- How to model how the system is affected by current allocation and dispatching policy?
- Can a decision support tool be developed to answer these questions?

Key concepts

- Network consists of ambulances located at bases
- Each base's coverage area is approximately a set of grids around it
- Each call has a priority queue of bases
  - Best served by first base in queue
- A served call consists of:
  - ambulance arriving from its base to the scene
  - taking the patient to a hospital
  - returning to (same/another) base

Design Principles

- Do not add extra bases or ambulances than those determined by the operator
  - Logistical challenges
- Consistency with current dispatching model
  - Calls served FCFS
  - Assign nearest free ambulance available
  - Priority queue for ambulances: learn from data logs (congestion implicit)
- Derive congestion information from data logs

Contributions

Models

- Problem-driven, data-driven models
- Problem structure, solution quality, tractability

Algorithms

- Static allocation of ambulances
- Dynamic redeployment of ambulances

Applications

- Emergency Medical Systems
- Disaster response, humanitarian logistics
- Facility location

Our approach

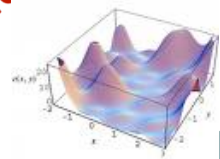
- Use data collected by the operator (call logs)
  - Capture time-dependent travel times
  - Optimize for metrics like preparedness, survival probabilities
  - Scalability
- Learn from the system data
- Build a solution that is faithful to the data (call logs)

Goal 1: Efficient and robust ambulance allocation

Goal 2: Dynamic repositioning policy

Solution Approach Summary

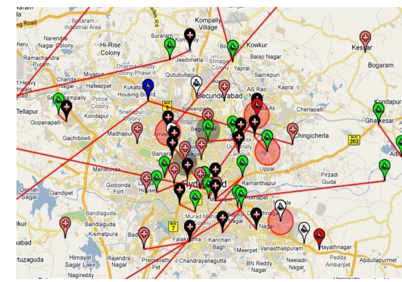
OPTIMIZATION + SIMULATION



VISUALIZATION

Ambulance Location

Dynamic Reposition , Dispatch



TRAFFIC MODELING

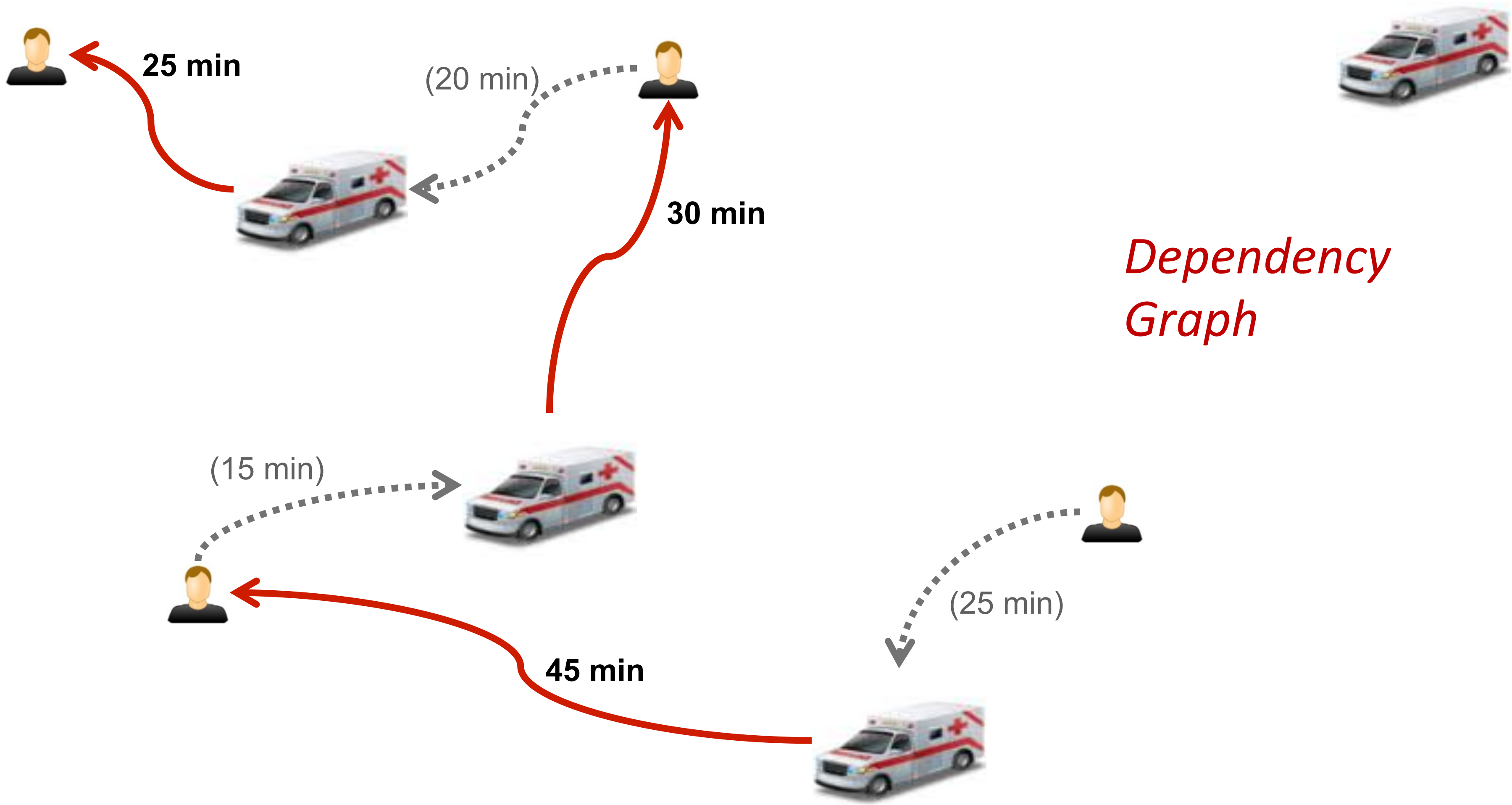
HUMAN BEHAVIOR MODELING

REAL-TIME INFORMATION

Talk outline

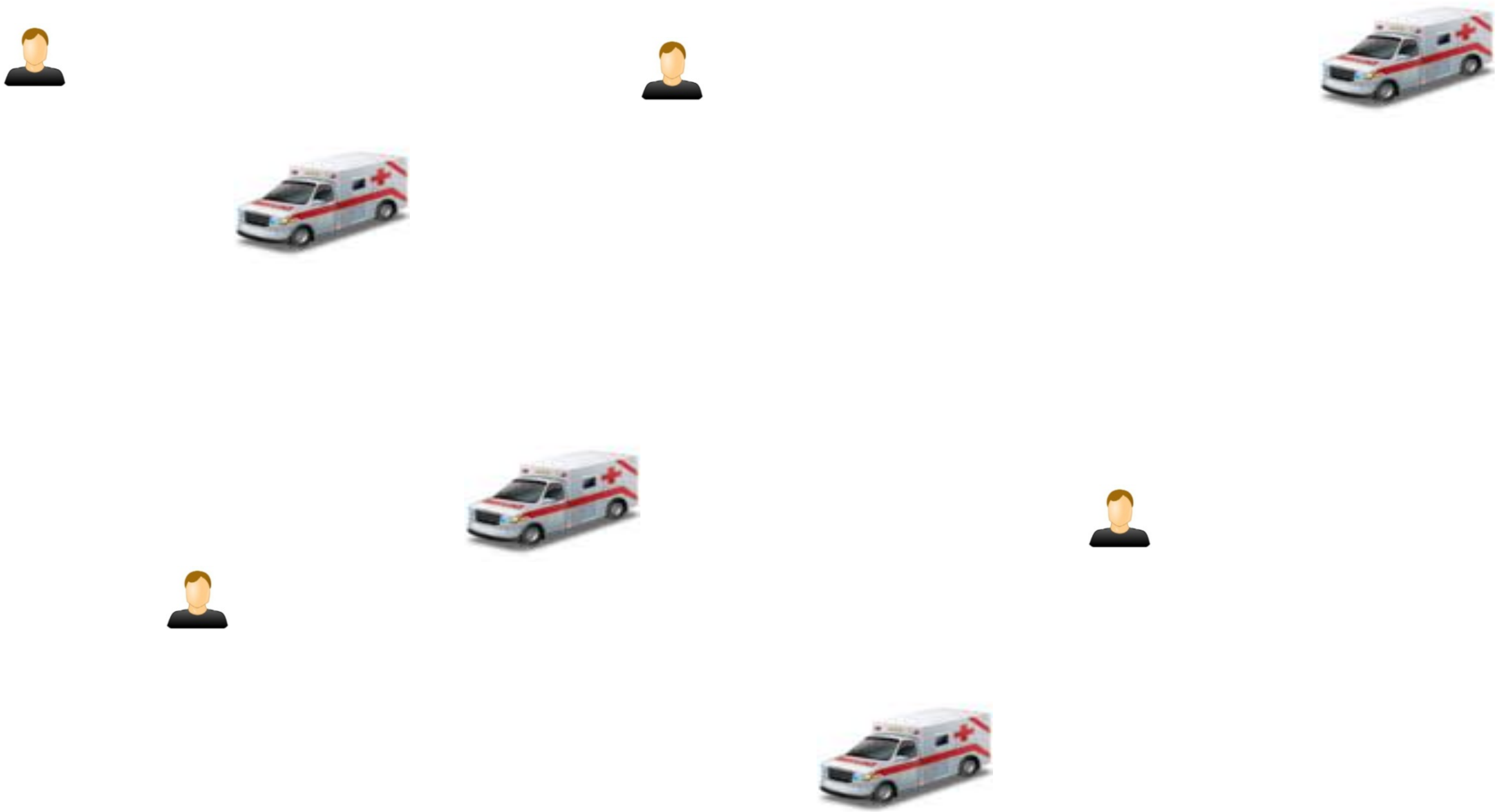
- Ground Realities for EMS in Emerging Economies
- Data-driven Simulation
- Mathematical Formulations
- Results
- Ongoing and Future Work

Modeling Concept: Chain Formation

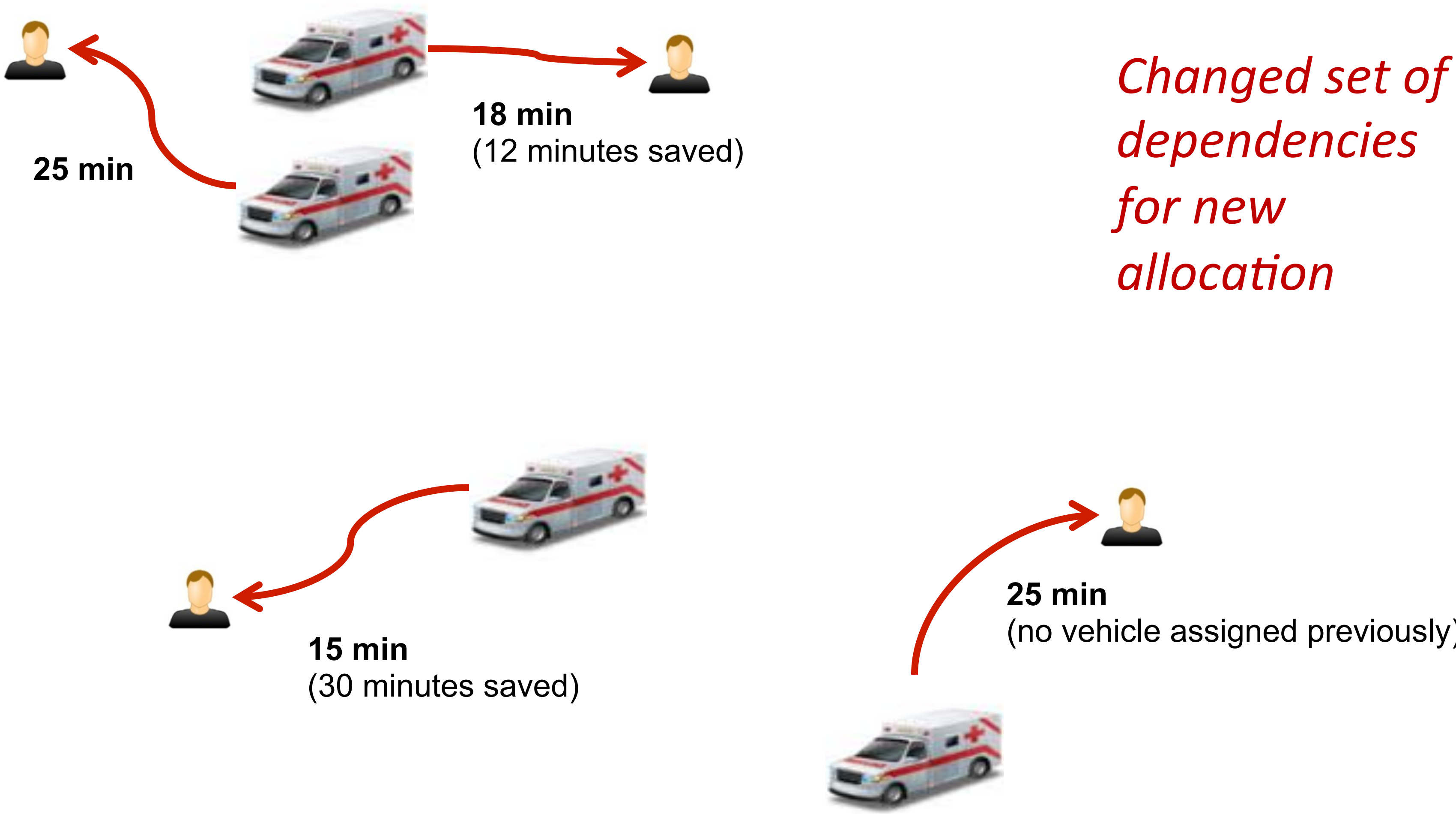


Dependency Graph

Modeling Concept: Chain Formation



Modeling Concept: Chain Formation



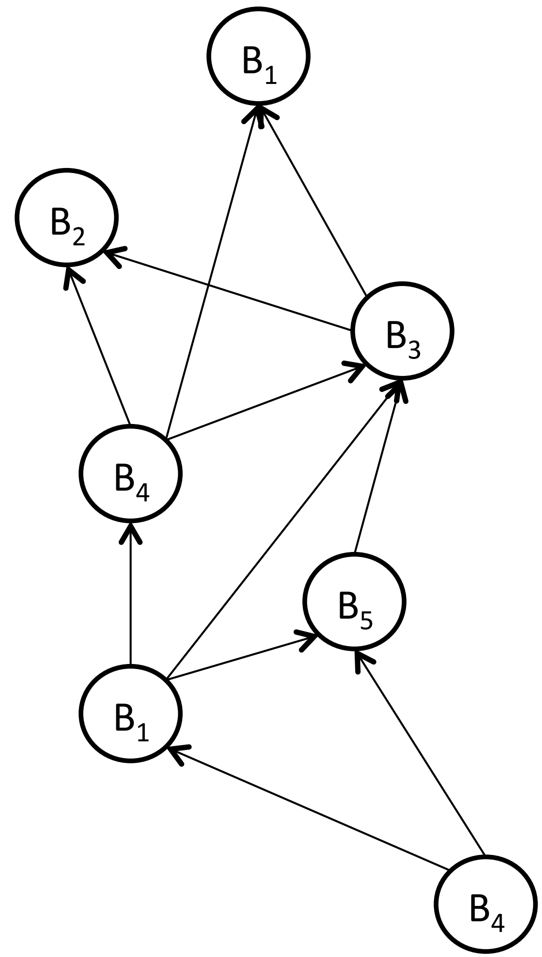
Changed set of dependencies for new allocation

Modeling Concept: Dependency Chains

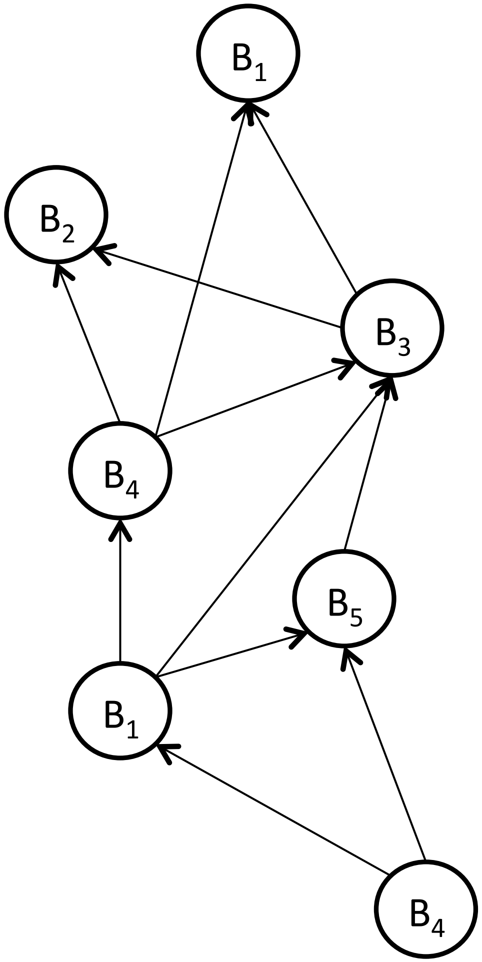
Given: One ambulance each at b1 – b5; dispatch policy; request (call) set

Request set R

Call 1	{B1, B2,..}
Call 2	{B2, B3,..}
Call 3	{B1, B2, B3,…}
Call 4	{B1, B2, B3, B4}
Call 5	{B3, B5}
Call 6	{B3, B4, B5, B1}
Call 7	{B1, B5, B4}



Simulation Framework to Compute Allocation Cost



Request set R

Simulation approach to evaluate ambulance-to-base allocations

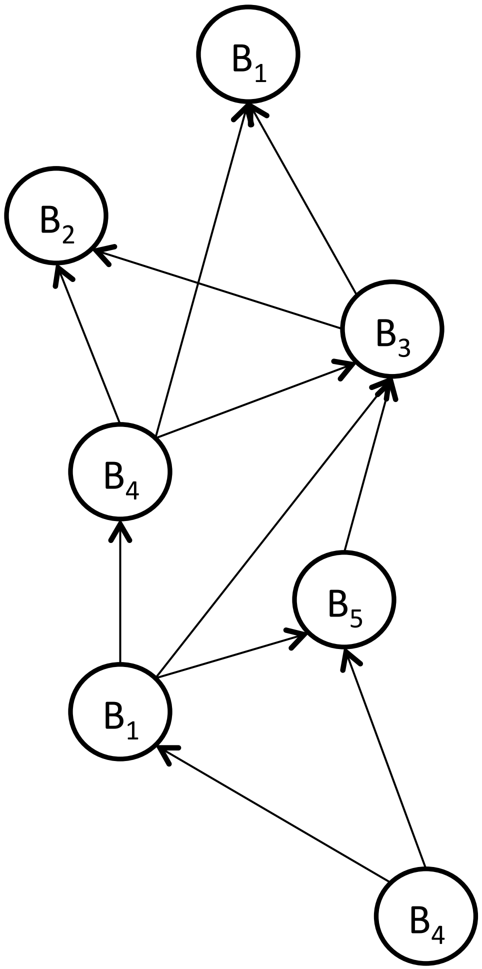
- Simulate Dispatch Officer assigning ambulances to calls
- Simulate response times and outcomes
- Data-driven approach (based on actual call logs)

$$L_R(A)=\sum_{r \in R} L_r(y_r, o_r);\ \{ y_r = ambulance\ a$$

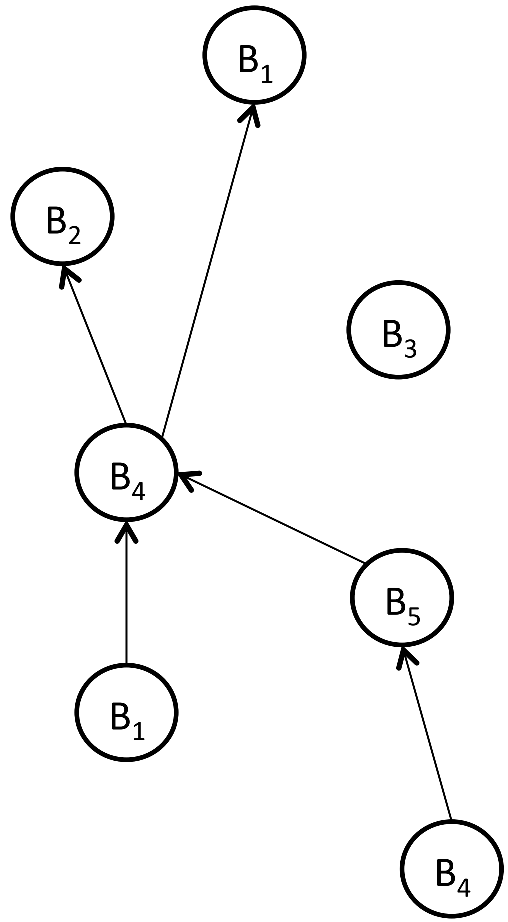
Based on call logs we can model:

- Call congestion patterns
- Chains and other long-range system effects
- Utilization of various base locations

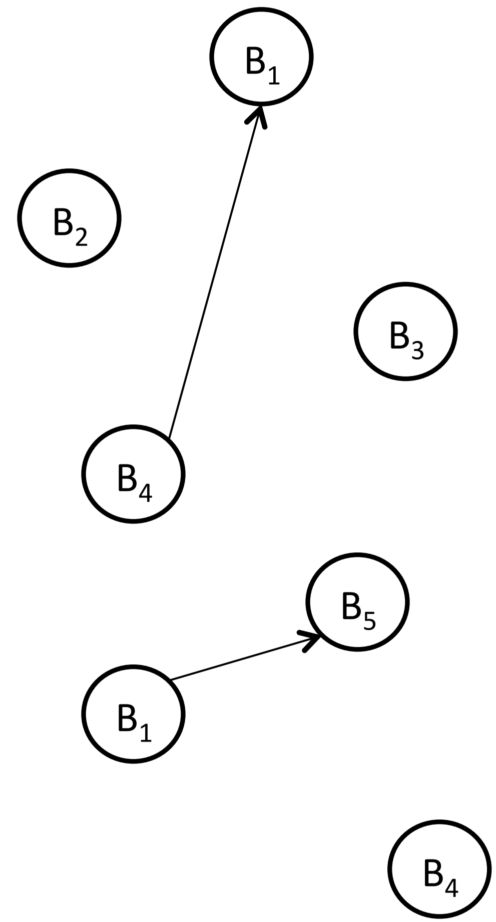
Breaking dependencies improves service



1 ambulance each at $B_1 - B_5$



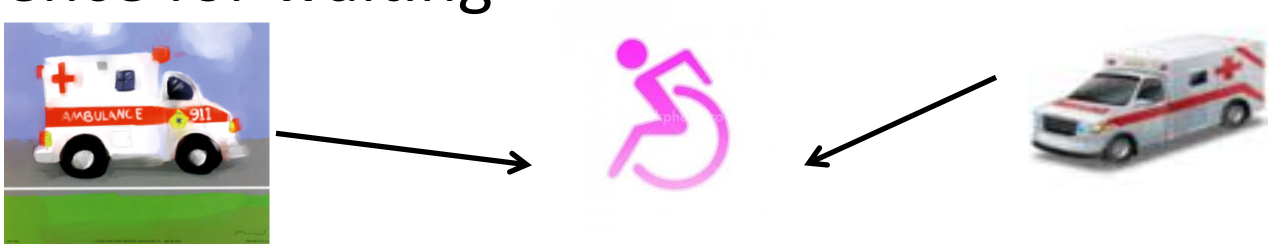
Add ambulance to B_1



Add ambulance to B_2

Modeling Abandonment

- Customer calls multiple service providers, limited patience for waiting



  - Choose the one which arrives first

- Abandonment model

$$\log \frac{Prob(abandoned)}{Prob(not\ abandoned)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

  - x_1 = 1 if request from a rural area
  - x_2 = base-to-scene * if (urban, peak hour)
  - x_3 = base-to-scene * if (rural, peak hour)

Talk outline

- Ground Realities for EMS in Emerging Economies
- Data-driven Simulation
- Mathematical Formulations
- Results
- Ongoing and Future Work

Mathematical Formulations

- R = Request set
- G_R = dependency graph
- $L_R(A)$ = total cost of allocation A for request R, from evaluating G_R

Utility of Static Allocation: $F_R(A) = L_R(\emptyset) - L_R(A)$

Static Allocation Objective: $\underset{A \in \mathcal{M}(\mathcal{A}):|A| \leq K}{\operatorname{argmax}} F_R(A)$

- s_t = state of the system at time t
- W_{s_t} = currently free allocation

Dynamic Redeployment Utility: $F_R(\pi) = E_{(s_1,\ldots,s_T)} E\left[\sum_{t=1}^{T} F(\pi(s_t) \mid s_t)\right]$

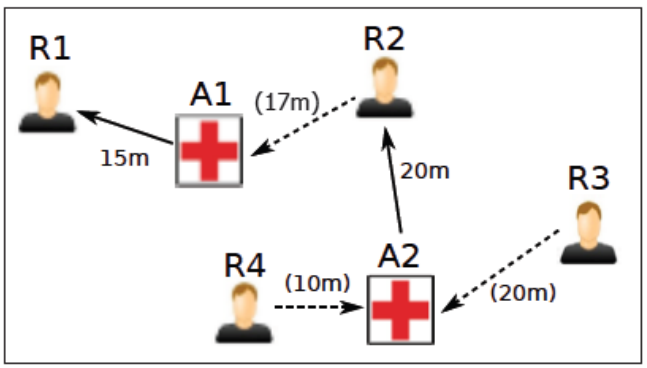
Dynamic (myopic) Redeployment: $\underset{A \in \mathcal{M}(\mathcal{A}, W_{s_t}):|A| \leq W_{s_t}}{\operatorname{argmax}} F_{R_t}(A \mid s_t)$

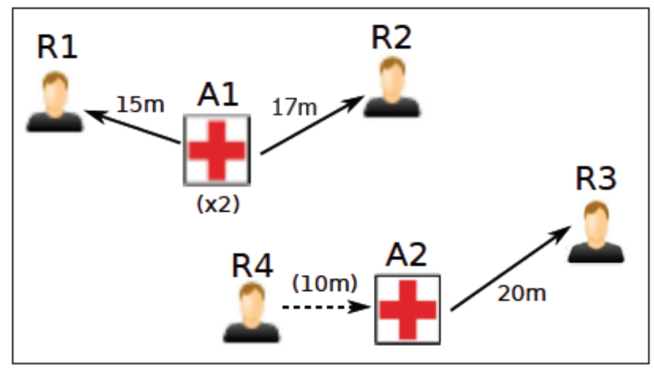
Claim: F is submodular?

- $F(A)$ is submodular *iff*

$$\forall A \subseteq B, \forall a, \ \delta_{F}\, aA \geq \delta_{F}\, aB$$

Gain of ambulance a only decreases with larger allocations



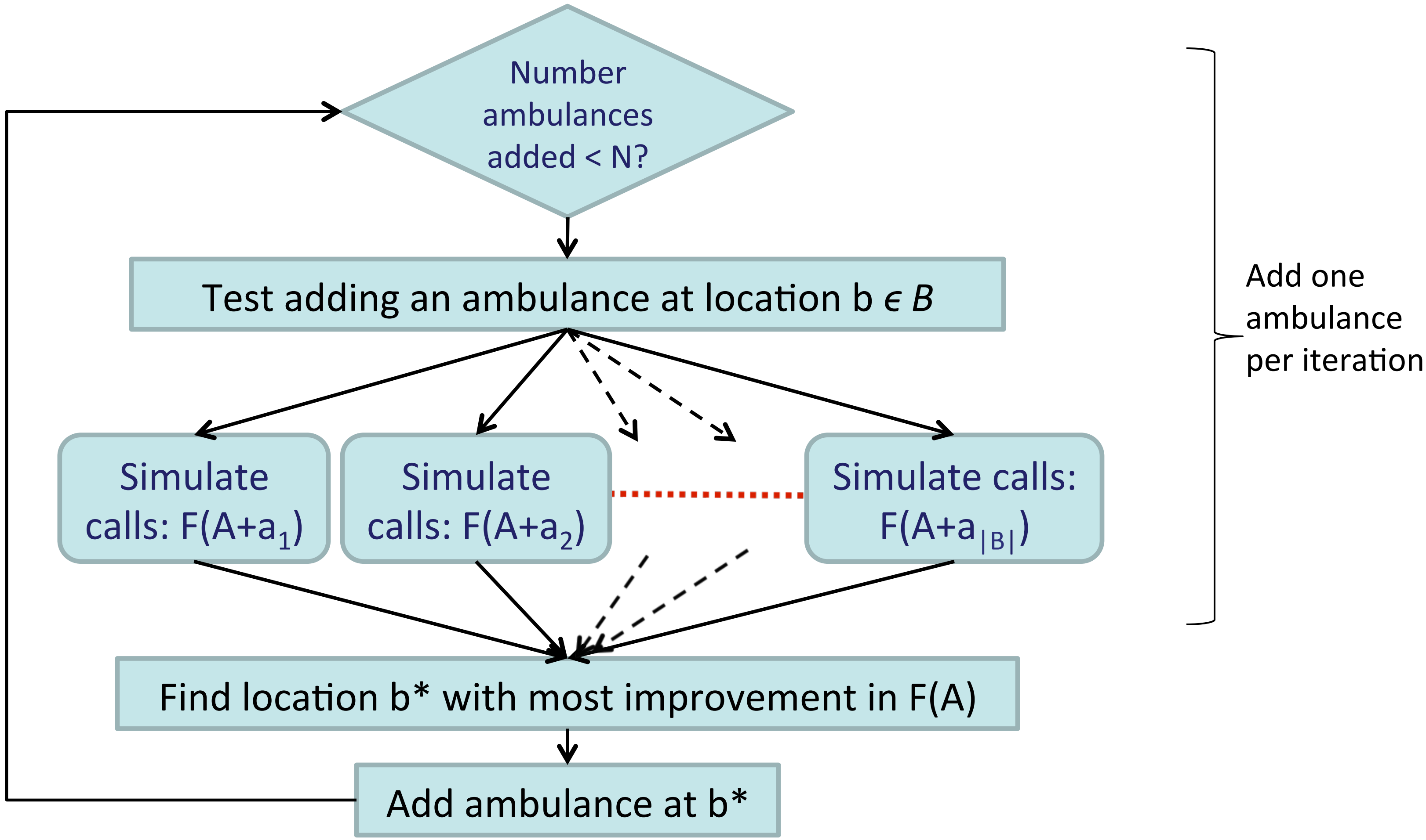


A_1	A_2	$F(A)$	A_1	A_2	$F(A)$
1	1	1	2	1	1
1	2	1	2	2	2
Gain		0	Gain		1

Rare case in data but happens nonetheless!

Simulation-Optimization (Greedy algorithm)

Goal: Allocate N ambulances among M bases



Running time = NB * O(Simulator)

Non-submodularity of *F* (static and dynamic)

- If monotone submodular, greedy algorithm returns solutions that achieve

$$F(A) \geq (1 - 1/e\ ) OPT$$

- Approximate monotonicity: $\forall A, \forall a, \delta \downarrow F$
$aA + \epsilon \downarrow m \geq 0$

- Approximate submodularity:

$$\forall A \subseteq B, \forall a, \delta \downarrow F\ aA + \epsilon \downarrow s \geq \delta \downarrow F\ aB$$

Theoretical Guarantees and Bounds (1)

Theorem: Let F be approximate submodular with additive violation and approximate monotone with additive violation . Let $A_1, \ldots, A_k$ denote the intermediate solutions of Greedy as it optimizes on F for a budget of K ambulances, the greedy algorithm produces an allocation A that satisfies

- Need to compute ϵ_s and ϵ_m
- Integer program written based on dependency chain model

Theoretical Bounds: Omniscient dispatcher

Utility of an *Omniscient dispatcher (G)*:

s.t. = 1

Maximize 'gain'

Serve each call

Ambulance count at each base

Ambulance count on network

Theorem: The objective, G, as measured by simulating an omniscient dispatcher, is monotone submodular. Furthermore, for and A and R, we have . Also, for any *A* with *|A|=K*,

.

Talk outline

- Ground Realities for EMS in Emerging Economies
- Data-driven Simulation
- Mathematical Formulations
- Results
- Ongoing and Future Work

Cost Function F

- $L_r(y) = \begin{cases} 0 & \text{if service time} \le 15 \text{ min} \\ 1 & \text{if service time} \le 30 \text{ min} \\ 2 & \text{if service time} \le 60 \text{ min} \\ 5 & \text{otherwise} \end{cases}$

Metrics and Static allocation

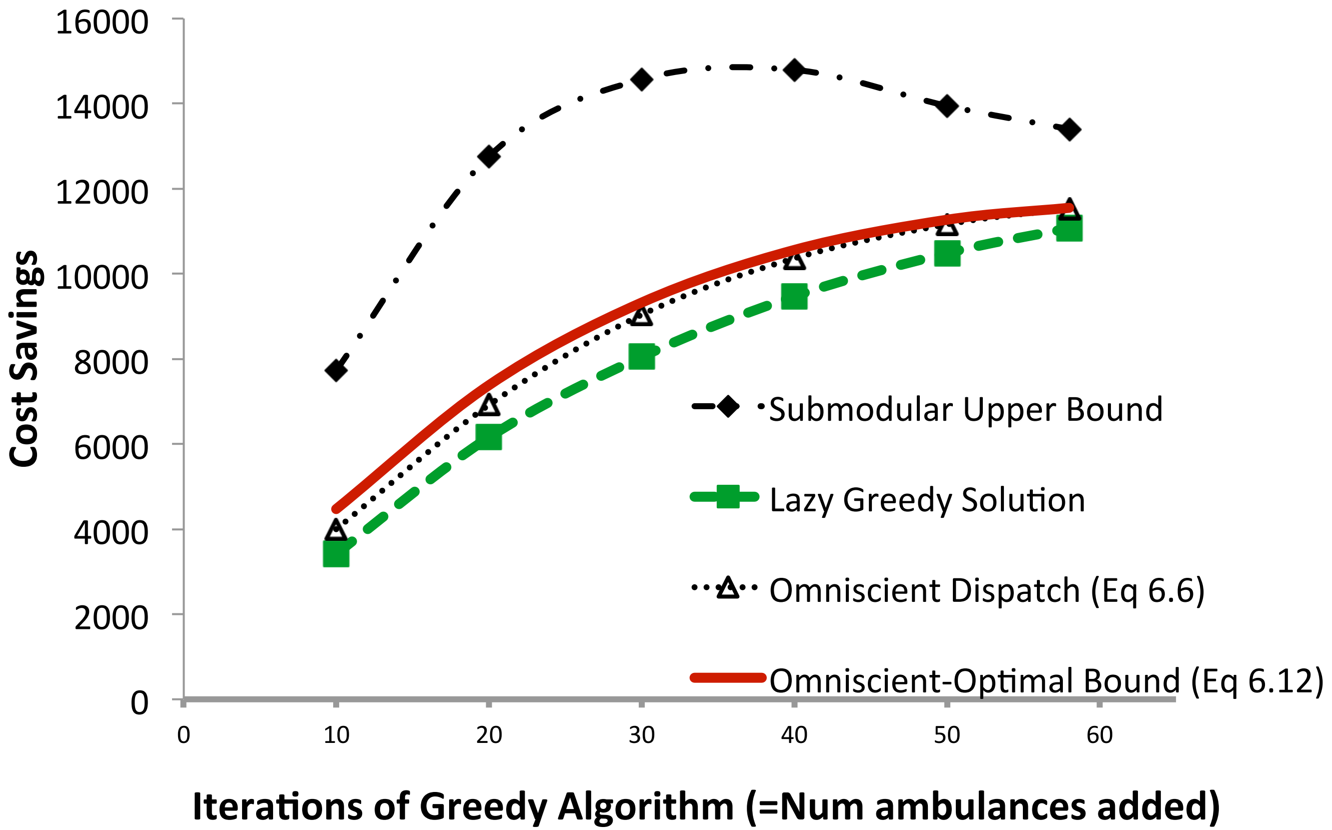
$$L_r(y) = \begin{cases} 0 & \text{if service time} \le 15 \text{ min} \\ 1 & \text{if service time} \le 30 \text{ min} \\ 2 & \text{if service time} \le 60 \text{ min} \\ 5 & \text{otherwise} \end{cases}$$

Result: Greedy solution improves upon baseline allocation of operator

Metric	Improvement over baseline allocation
# Calls w/ Base-to-scene < 15 min	6.1% (increase)
# Calls w/ Base-to-scene <30 min	3.4% (increase)
# Ambulances Busy	42.7% (decrease)
# Calls serviced by primary base	9.4%

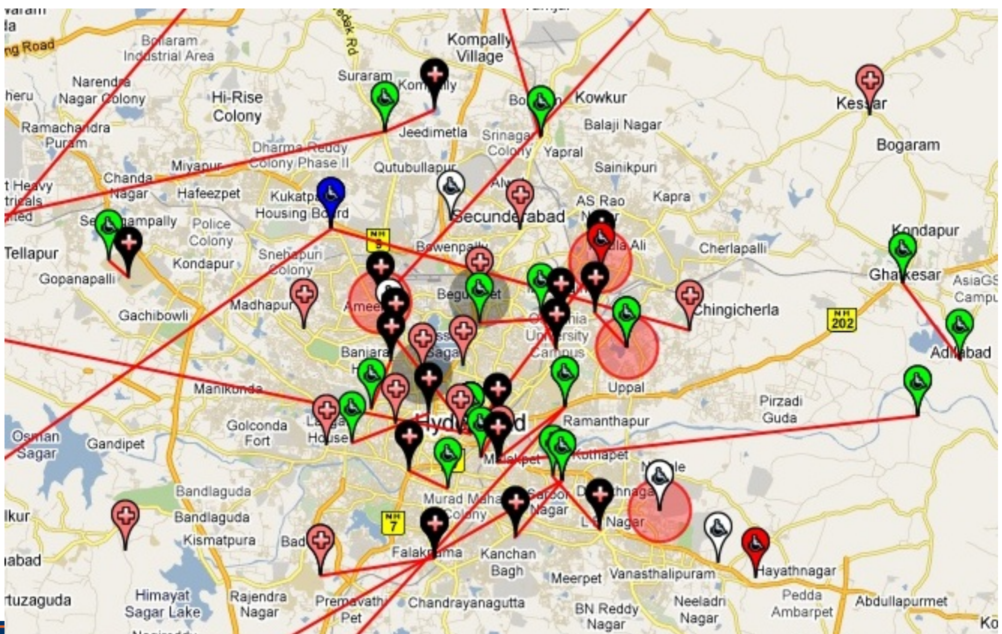
Bounds

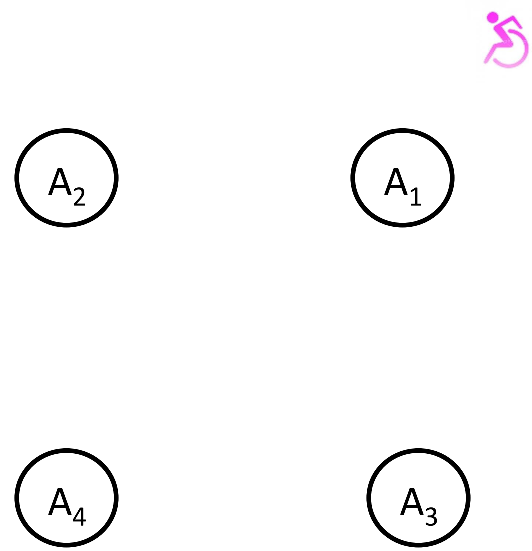
Result: Greedy solution close to bound from optimal dispatch allocation => 'close' to optimal



Dynamic repositioning

- Under high demand regions
  - ‘System stress’
- Re-position ambulances in real-time
  - Move free ambulances from ‘home’ base to nearby bases
  - Waiting on street corners

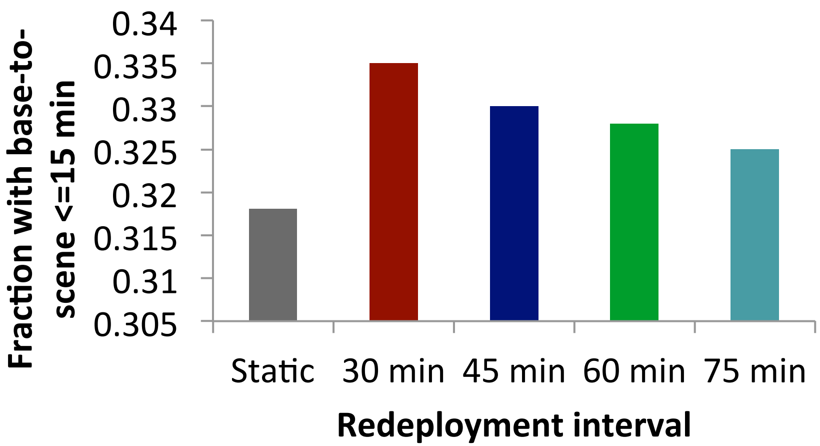


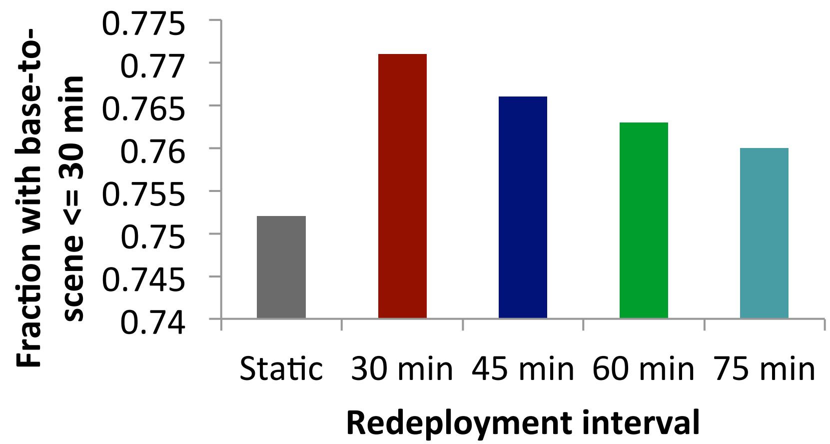


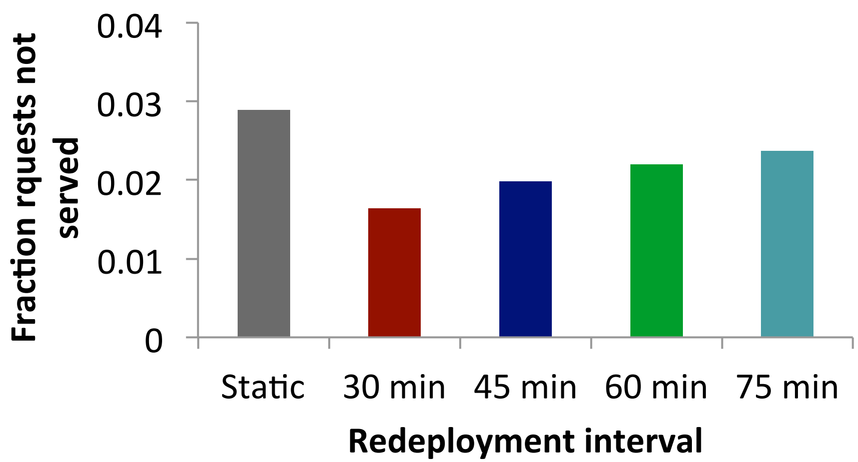
Dynamic repositioning vs. static allocation

Result 1: More redeployment produces better service

Result 2: Most impacted metric = number of calls served







Value of Dynamic Repositioning

- Value in dynamic repositioning compared to static

	Look-ahead = 45 min
Calls served with base-scene <15 min	39.32%
Calls with base-scene <30 min	-0.1%
Calls served by primary base	-1.8%
Calls not served (vehicles busy)	**-30.6%**

- Most impacted metric: calls served
- Value higher when greater flexibility in repositioning – example: more often, more ambulances allowed to be repositioned

Robustness under congestion fluctuations

Result: Even under variability in demands and travel times, the *Greedy* solution shows improvement over default.

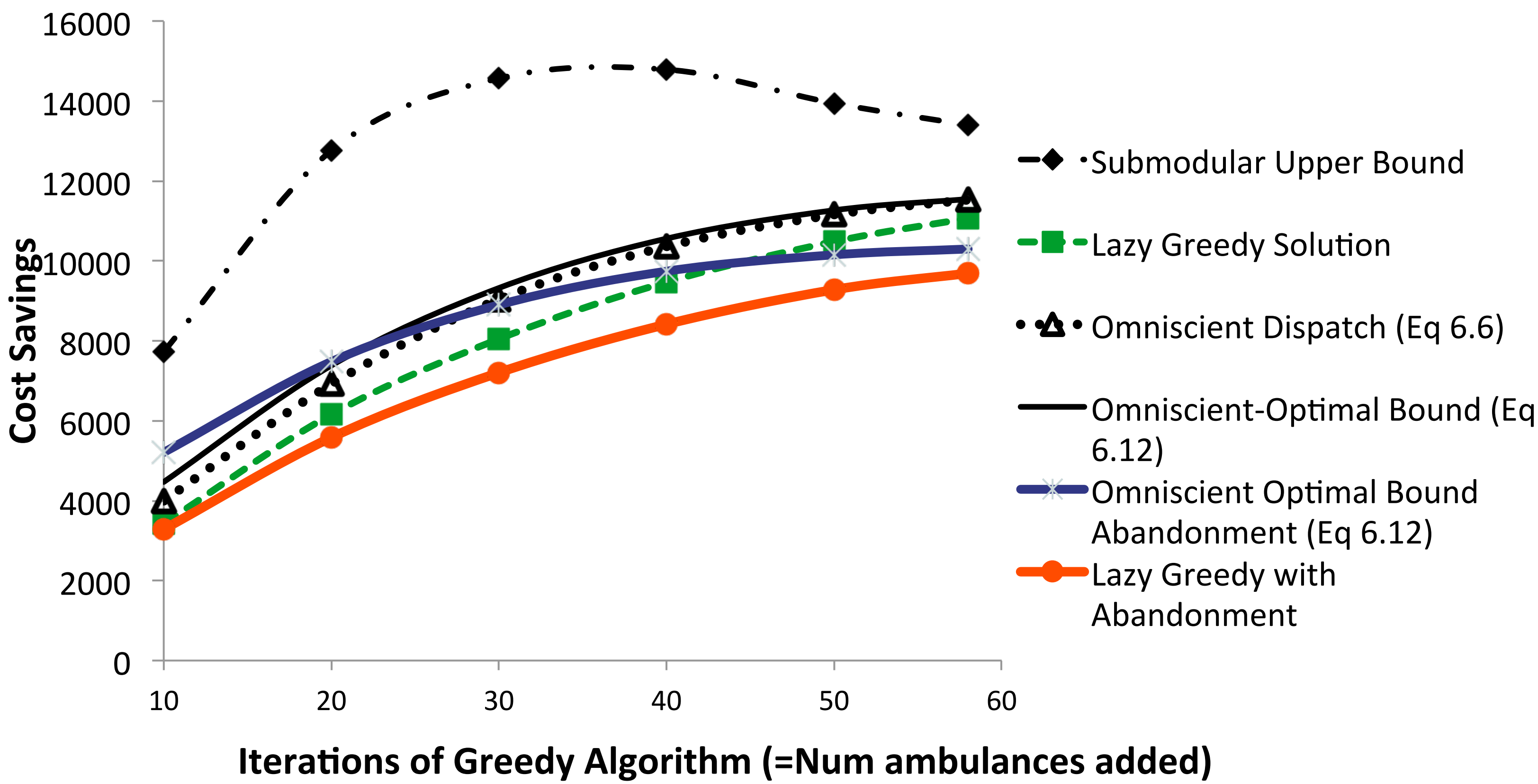
	0% increase in demand	10% increase in demand	15% increase in demand
Base-to-scene <15 min	6.1%	5.7%	5.0%
Base-to-scene <30 min	3.4%	3.5%	3.8%
Served by primary base	9.4%	10.1%	10.3%
Calls not served	-42.7%%	-36.2%	-33.3%

	0% increase in travel time	10% increase in travel time	15% increase in travel time
Base-to-scene <15 min	6.1%	5.7%	4.9%
Base-to-scene <30 min	3.4%	3.9%	3.7%
Served by primary base	9.4%	10.4%	10.5%
Calls not served	-42.7%%	-35.0%	-31.5%

*Measured using simulation on independent data, for a period of one month

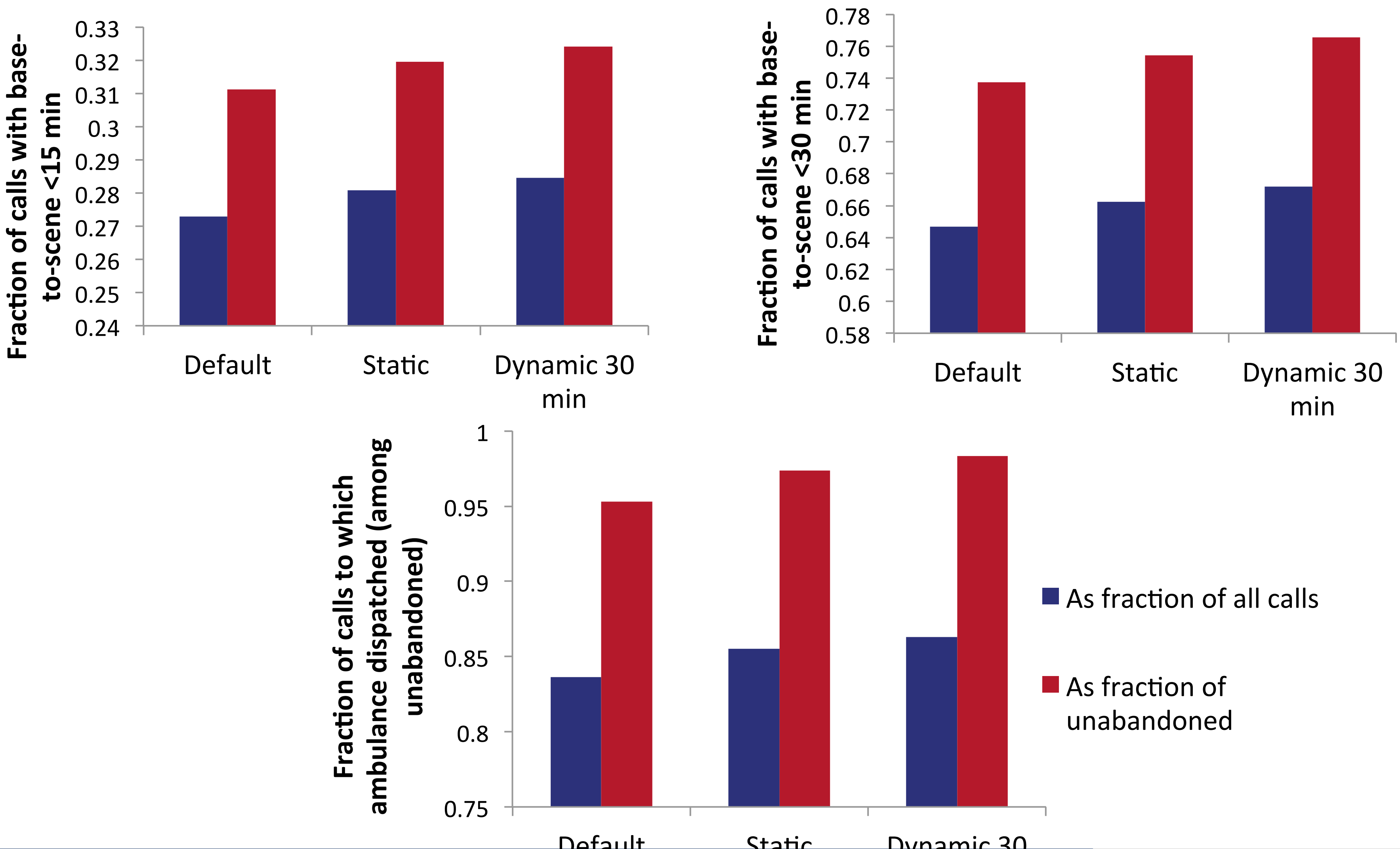
Bounds with abandonment

Result: Optimality gap remains similar in the case of abandonment => ‘close’ to optimal



Solutions with abandonment

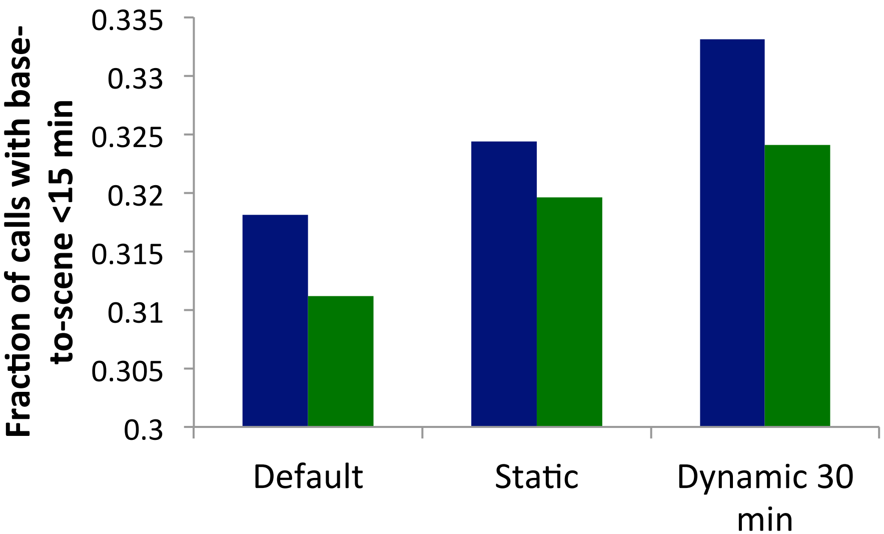
Result: Improvements with respect to all metrics

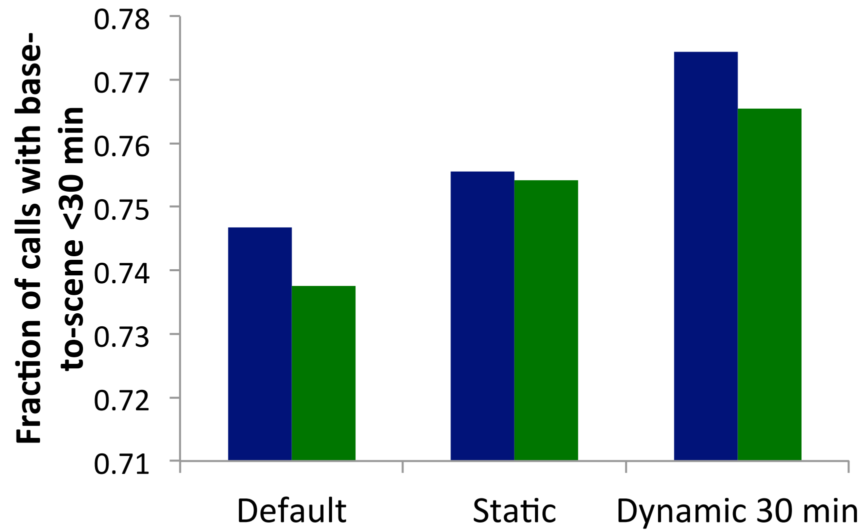


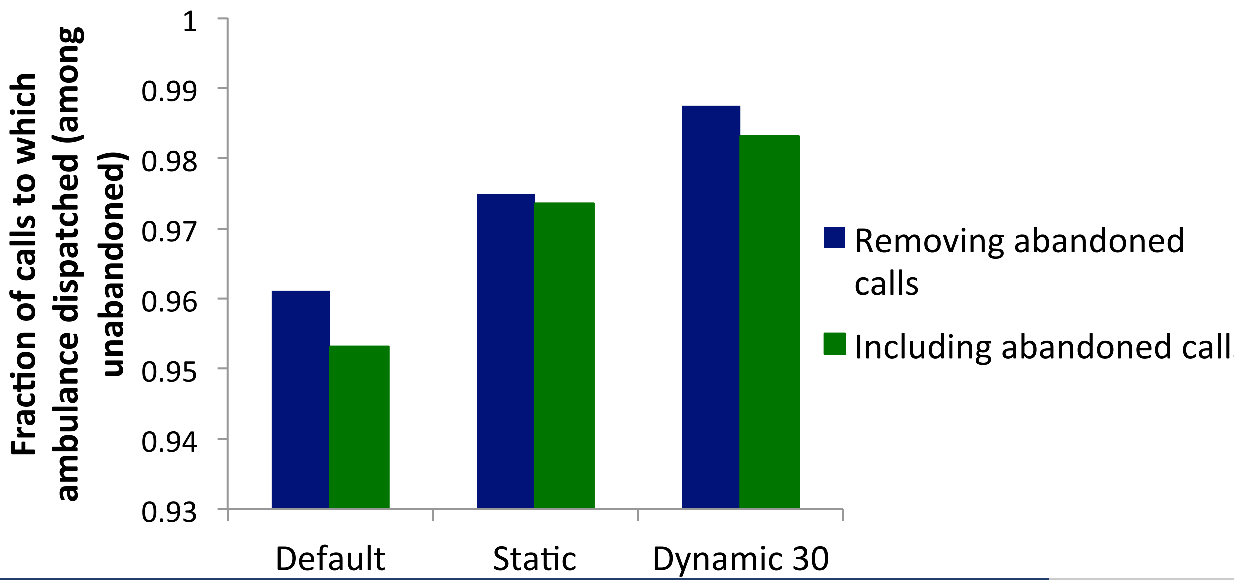
Opportunity cost of abandonment

- Abandoned calls add inefficiency to the system
- Ambulance could have served another customer (with a better service level)
- How much is lost due to abandoned calls?
  - Find optimal allocation when abandoned calls existed
  - Remove abandoned calls and measure impact of optimal allocation
- 12% calls abandoned in data set
  - ~6% improvement when abandoned calls ignored
  - Remaining 6% of calls do not reduce service level

Opportunity Cost of Abandonment







Takeaways

- Static allocation provides good results compared to baseline operations.
- More repositioning makes more ambulances available where needed; covers requests better
  - Reposition often if idle travel cost is low
- Greedy algorithm is quick, particularly for dynamic redeployment (<~10s)
- Solutions from our algorithm are robust
- Opportunity cost of abandonment is about 50% that of fraction of abandoned calls

Talk outline

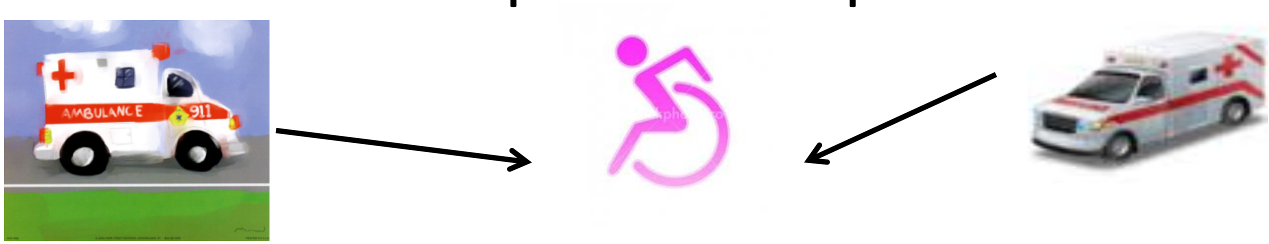
- Ground Realities for EMS in Emerging Economies
- Data-driven Simulation
- Mathematical Formulations
- Results
- Ongoing and Future Work

New/needed technology: Traffic models

- Existing routes
  - Currently use data-driven models for traffic congestion capture
  - Allows to extrapolate data for routes taken in past
- New routes?
  - Crowdsource/obtain traffic information from other ambulances
  - Communication between ambulances to share traffic data

New/needed technology: Human behavior models

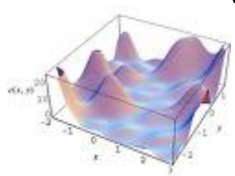
- ‘Conflict’ between existing ad-hoc networks and the operator’s network
- Customer calls multiple service providers



  - Choose the one which arrives first
- Modeled higher abandonment in select urban areas
- How to improve ambulance utilization?
  - Better dispatching models?
- What system can lead to improved social welfare?

Robust and Dynamic Approaches for Evolving Infrastructure

OPTIMIZATION

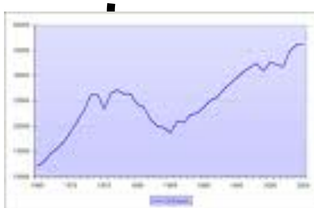


SIMULATION

Robust Planning

Dynamic Reconfig-uration

Repair/ Recovery



PREDICTION

REAL-TIME INFORMATION

ENVIRONMENTAL IMPACT

THANK YOU!