# Patient Flow Management in Emergency Departments

#### Junfei Huang

With Boaz Carmeli and Prof. Avishai Mandelbaum

#### **Outline**

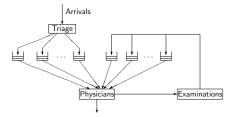
- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

# **Emergency Department (ED) in Rambam Hospital**

- Rambam Hospital is the largest hospital in northern Israel;
- ▶ The ED has 40 beds; 245 patients arriving daily;
- ▶ Patients: New vs. Work In Process (WIP);
- ► Canadian triage system:
  - New (triage) patients are classified into 5 clinical classes:
    - Resuscitation; Emergent; Urgent; Less Urgent; Non-urgent;
  - Result in several classes of WIP patients;
- Existing scheduling policy (static priority);



► ED is blocked, long sojourn time (with mean 4.5 hours, 10% over 6 hours);

#### **Research Questions**

- Objective of the project:
  - Design new scheduling policy within present triage system;
    - New: Pre-specified requirements on time till first examination
       Clinical:
    - WIP: Push them out as soon as possible Operational;
  - How do we make the tradeoff between clinical and operational considerations?
    - Minimize the congestion;
    - Subject to the deadline constraints;
- Achievement:
  - A "nearly" optimal and implementable dynamic scheduling policy;

## A Good Hospital in China



## **Emergency Departments in China**

#### 卫生部拟将急诊分三区 病人按病情分四级

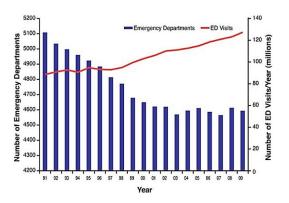
2011年09月07日 08:55:01 来源: 人民日报 【字号 大小】【收藏】【打印】【关闭】 分享到新华微博

卫生部日前公布《急诊病人病情分级试点指导原则(征求意见稿)》。卫生部拟将急诊 科从功能结构上分为红黄绿"三区",将病人的病情分为"四级",从而提高急诊病人分诊 准确率,保障急诊病人医疗安全。

征求意见稿提出,急诊病人病情的严重程度决定病人就诊及处置的优先次序。急诊病人病情分级不仅仅是给病人排序,而且要分流病人,使病人在合适的时间去合适的区域获得恰当的诊疗。

- In 2011, the MOH of China proposed to introduce a triage system to manage EDs;
  - Improve the quality of care (safety of patients);
- Patients are classified into 4 levels according to their severity;
- A natural problem is how to schedule the patients;

## **Emergency Departments in the United States**

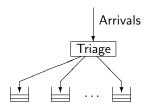


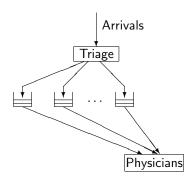
- ► ED environment has become more crowded (waiting time increased by 25% (46.5 to 58.1 mins), from 2003 to 2009);
- The need for the tradeoff becomes more pronounced;

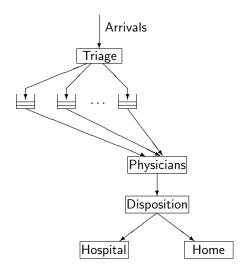
#### **Outline**

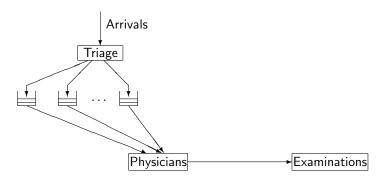
- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

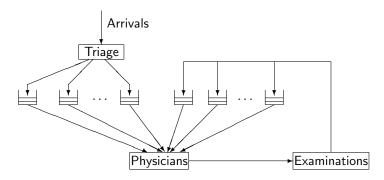
Arrivals

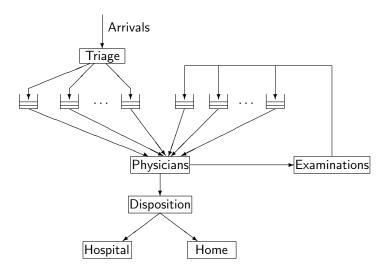


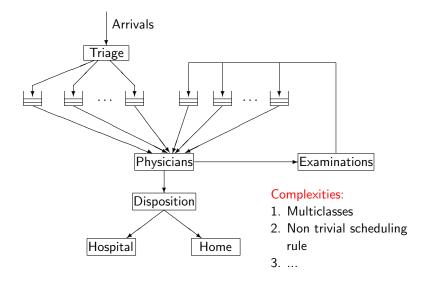






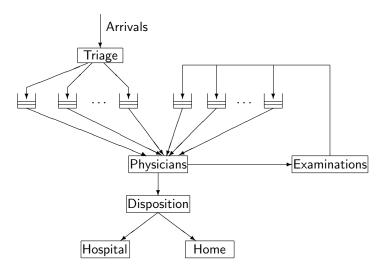


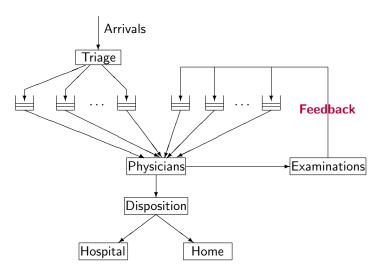


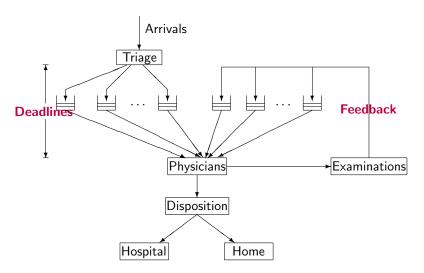


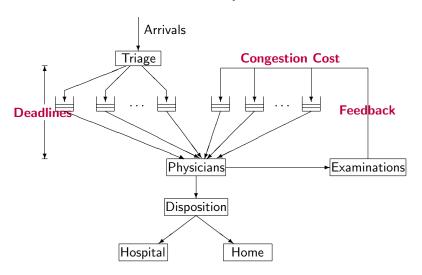
#### **Complexities and the State of the Arts**

- Many simulation-based studies;
- Few analytical models;
- An analytical model:
  - Manage patient flow in an ED, from a queueing-theory perspective;
  - Capture the most important features (what are they?);







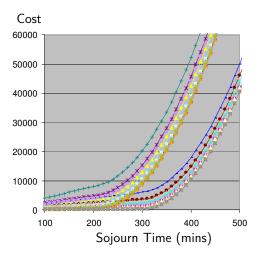


#### **Example: The ED in Rambam Hospital**

#### Feedback

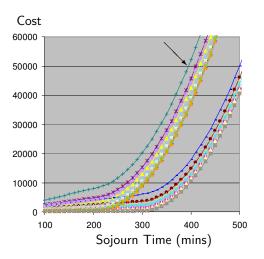
- Empirical analysis shows, on average, each patient visits the physician for at least 3 times;
- Deadlines on time-till-first-treatment:
  - Canadian Triage System patients are classified into 5 levels:
    - Resuscitation (Immediate);
    - Emergent (15 mins); Urgent (30 mins); Less Urgent (60 mins); Non-urgent (120 mins);
- Congestion costs:
  - Waiting costs; clinical costs; emotional costs; psychological costs; others (long waits increase the probability of disaster);

### **Example: The ED in Rambam Hospital**

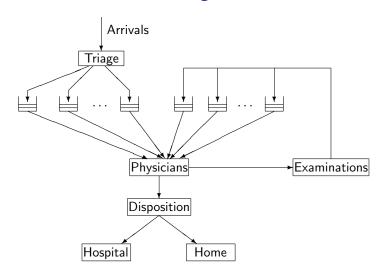


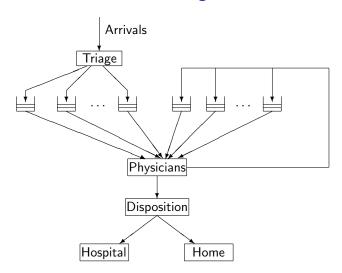
• Triage class (3); age (4); decision after treatments (2);

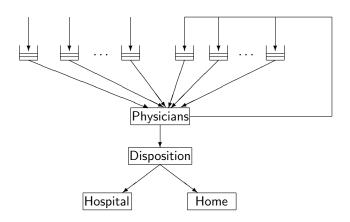
### **Example: The ED in Rambam Hospital**

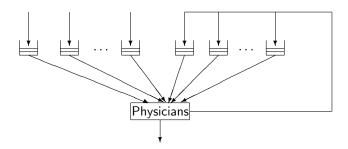


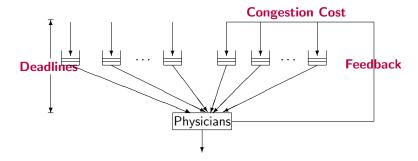
- Triage class (3); age (4); decision after treatments (2);
- Triage class 3 (urgent), Age>75, Discharged;

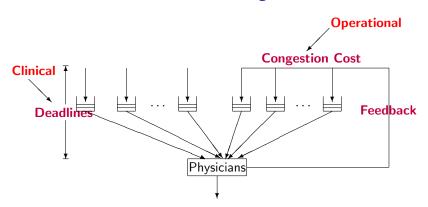


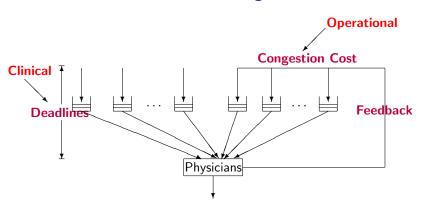












Clinical v.s. Operational – How to Balance?

#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

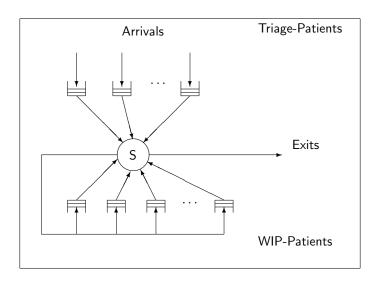
### Two ED Models

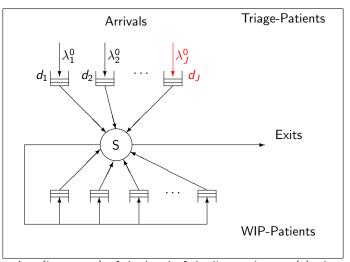
	Queue Length Model	Sojourn Time Model
Congestion Cost	Queue Length	Sojourn Time
WIP Transition	Markovian	Deterministic

#### **Outline**

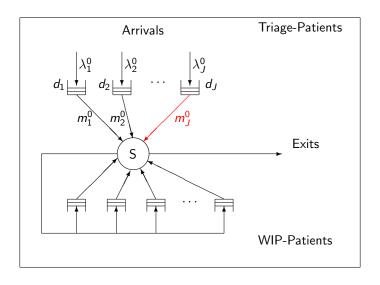
- Motivation
- Modeling ED Operations
- Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

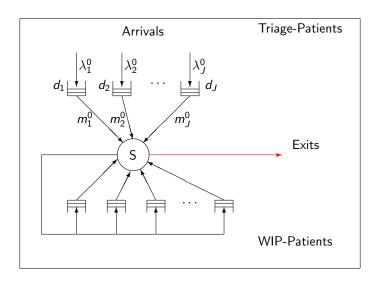
## **Queue Length Model - Structure**

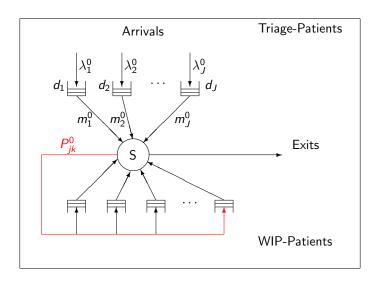


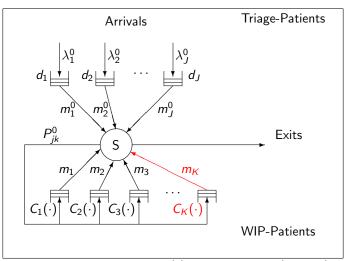


Age (in system) of the head-of-the-line patient:  $\tau_j(t)$ , then deadline constraints:  $\tau_j(t) \leq d_i, j \in \mathcal{J}$ .

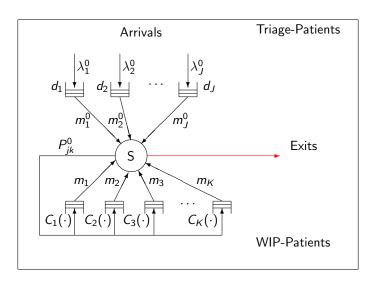


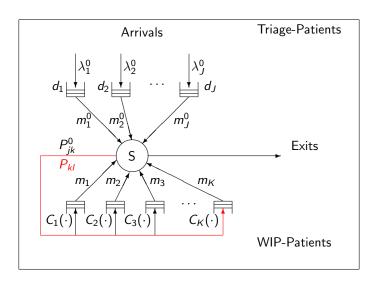






Queue length at time t:  $Q_k(t)$ , incur queueing (holding) cost at rate  $C_k(Q_k(t))$ ; the total cost rate:  $\sum_{k \in \mathcal{K}} C_k(Q_k(t))$ .





#### **Problem Formulation:**

**Constrained optimization problem**: for any T > 0,

$$egin{aligned} \min_{\pi \in \Pi} & \int_0^T \sum_{k \in \mathcal{K}} C_k(Q_k^\pi(s)) \mathrm{d}s \ \end{aligned}$$
 s.t.  $au_i^\pi(t) \leq d_i, \quad \forall j \in \mathcal{J} \quad \text{and} \quad 0 \leq t \leq T.$ 

- Example:
  - Linear holding cost:

$$C_1(x) = 2x \implies \int_0^T C_1(Q_1(t))dt = 2\int_0^T Q_1(t)dt;$$

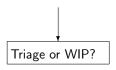
- $\tau_3(t) \leq 30$ ,  $\tau_4(t) \leq 60$ ,  $\tau_5(t) \leq 120$ ;
- ▶ Infeasibility:  $\tau_j$ ,  $j \in \mathcal{J}$  random,  $d_j$  deterministic;
- Asymptotic framework:
  - relax: "feasibility" → "asymptotical compliance (feasibility)";
  - ullet relax: "optimality" o "asymptotical optimality".

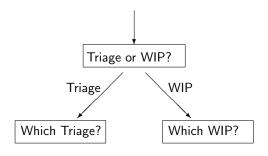
#### Main Result

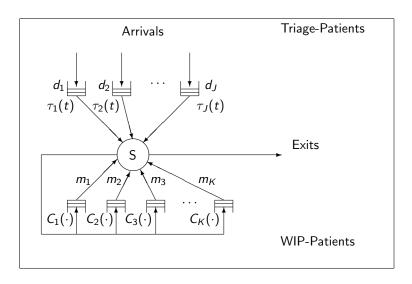
#### Main Result

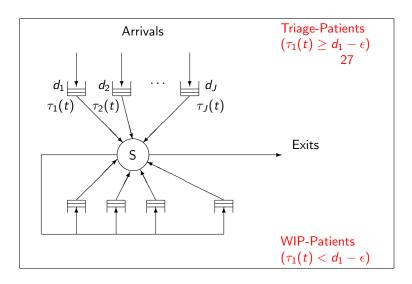
Identified a scheduling policy that is "nearly" optimal when the system is heavily loaded.

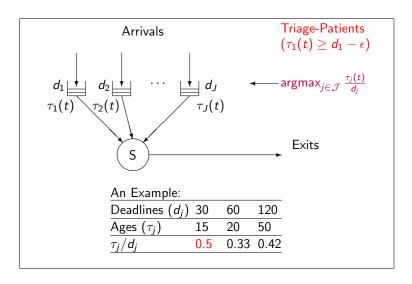
- ► The rigorous meaning of "nearly optimal" will be defined in the asymptotic framework;
- ► Heavily loaded system:
  - Intuitive: Arrival rate ≈ service capacity;
  - Rigorous: Traffic intensity  $\rho = \frac{1}{S} \sum_{j \in \mathcal{J}} \lambda_j m_j^e \approx 1$ .
  - Realistic: Crowded ED environment;
- What is the structure of the scheduling policy?

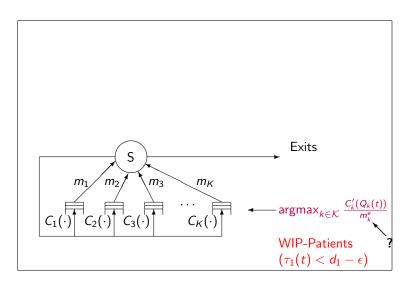


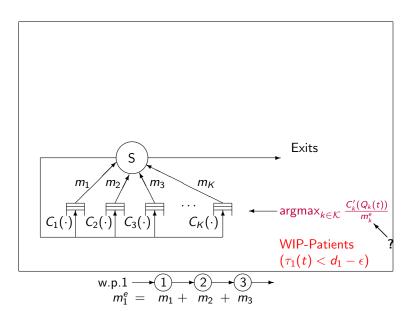


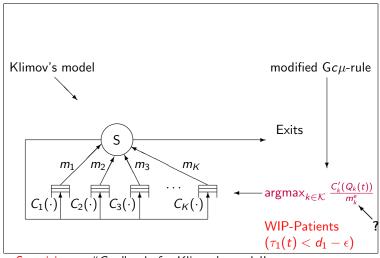




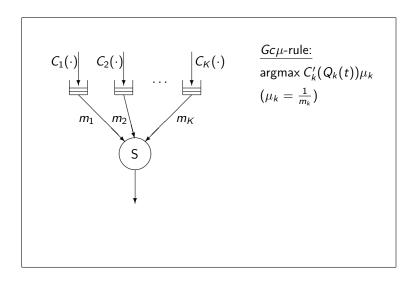


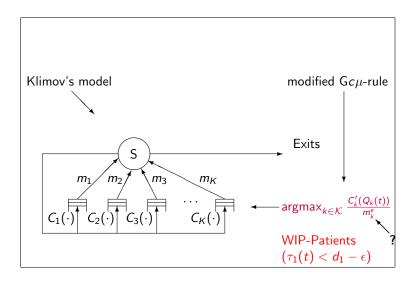






Surprising: a " $Gc\mu$ "-rule for Klimov's model!





#### **Effective Mean Service Time**

- Effective mean service time of a patient is the expectation of the total potential service requirement of that patient after entering into that class;
- An example:

$$m_1^e = m_1 + 0.4 m_2^e + 0.6 m_3^e$$

•  $M^e = (m_k^e)_{k \in \mathcal{K}}$  of WIP patients<sup>1</sup>:

$$M^e = M + PM^e \quad \Rightarrow \quad M^e = [I - P]^{-1}M;$$

▶  $M_{\mathcal{J}}^e = (m_i^e)_{j \in \mathcal{J}}$  of Triage patients:

$$\frac{M_{\mathcal{J}}^{e} = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}M^{e};}{{}^{1}M_{\mathcal{J}} = (m_{i})_{i \in \mathcal{J}}, P_{\mathcal{J}\mathcal{K}} = [P_{ik}], M = (m_{k})_{k \in \mathcal{K}}, P = [P_{kl}];}$$

#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

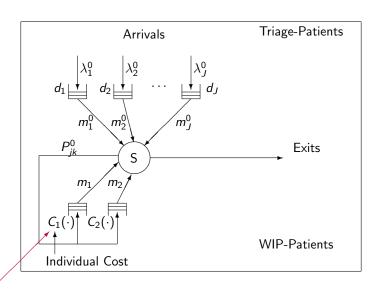
### Sojourn-Time Model

- Deadline constraints on triage patients do not change;
- For WIP visits, each patient has a deterministic routing vector;
  - Realistic: blood test → X-ray → ECG → ...;
- $\triangleright$   $C_0$ : the *starting classes* of any route;
- ▶  $C_k$ : all classes on a route starting with  $k \in C_0$ ;
- ▶  $\bigcup_{k \in C_0} C_k \setminus \{k\}$ : subsequent classes;

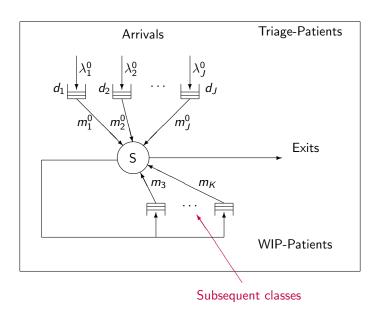
Congestion cost:

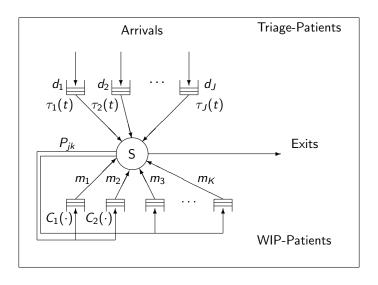
$$\mathcal{S}(t) := \sum_{k \in \mathcal{C}_0} \sum_{i=1}^{E_k(t)} C_k \left( \sum_{k' \in \mathcal{C}_k} \omega_{k'}(i) \right);$$

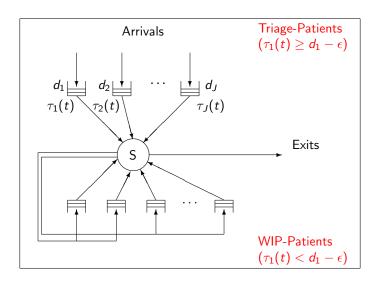
### Sojourn Time Model - Structure

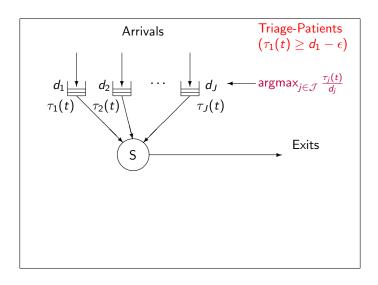


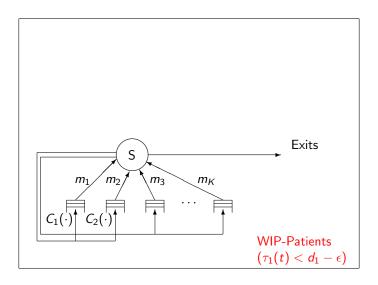
### Sojourn Time Model - Structure

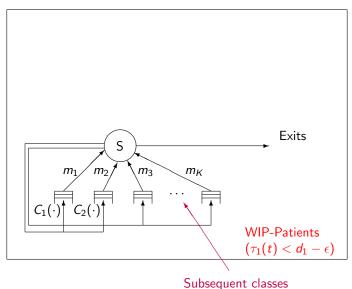


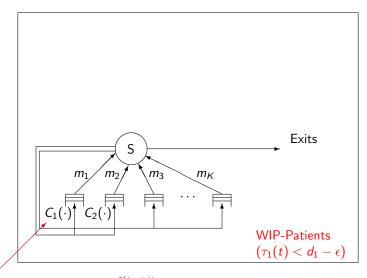












Starting classes,  $\operatorname{argmax}_{k \in \mathcal{C}_0} \frac{C_k'(\tau_k(t))}{m_k^e}$ 

### **Interim Summary**

- Analyze the emergency departments;
  - Capture the tradeoff between triage- vs. WIP-patients;
- Build two queueing models;
  - Queue Length Model;
  - Sojourn Time Model;
- Provide good and implementable scheduling policies;

Table: Comparison of Two Models

	Queue Length Model	Sojourn Time Model
Congestion Cost	Queue Length	Sojourn Time
WIP Transition	Markovian	Deterministic
WIP Policy	Queue Length	Age

#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

### An ED Case Study: Value of Information

Data is from ED in Rambam; sojourn time model;

# WIP visits	1	2	3	4	5
Proportion	0.28	0.30	0.28	0.11	0.03

A & D Status	cus Admitted		Discharged	
Proportion/Cost function	0.40, t <sup>2</sup>	2	0.60,	$2t^2$

- Is it worthy to estimate these two kinds of information upon a patient's arrival?
- ▶ Patients are classified into different classes according to the availability of these two kinds of information:

	Case 1	Case 2	Case 3
# WIP visits	N	Y	Y
A & D Status	N	N	Y
Congestion Cost	Benchmark	<b>↓18.01%</b>	<b>∜26.8%</b>

### An ED Case Study: Value of Information

Data is from ED in Rambam; sojourn time model;

# WIP visits	1	2	3	4	5
Proportion	0.28	0.30	0.28	0.11	0.03

A & D Status	Admitted		Discharged	
Proportion/Cost function	0.40,	$t^2$	0.60,	$2t^2$

- Is it worthy to estimate these two kinds of information upon a patient's arrival?
- ▶ Patients are classified into different classes according to the availability of these two kinds of information:

	Case 1	Case 2	Case 3
# WIP visits	N	Y	Y
A & D Status	N	N	Y
Congestion Cost	Benchmark	<b>↓18.01%</b>	<b>↓</b> 26.8%

► Good news: A well trained nurse can estimate both kinds of information very accurately!

#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

### Intuition for the Queue Length Model

- $\triangleright$  A(t): total potential service requirement brought into the ED;
- ightharpoonup T(t): amount of service requirement has been served;
- W(t) = A(t) T(t): total potential service requirement left:
  - minimized by work-conserving policy;
  - invariant to any work-conserving policy;
  - conditional on the queue length processes,

$$oxed{W(t)pprox \sum_{j\in\mathcal{J}} m_j^e imes Q_j(t) + \sum_{k\in\mathcal{K}} m_k^e imes Q_k(t)}$$

▶ As a result,  $\sum_{j \in \mathcal{J}} m_j^e \times Q_j(t) + \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t)$  is minimized and invariant to any work-conserving policy;

min 
$$\sum_{k \in \mathcal{K}} C_k(Q_k(t))$$

Myopic

s.t. 
$$au_j(t) \leq d_j, \quad j \in \mathcal{J};$$

$$\begin{split} \min \quad & \sum_{k \in \mathcal{K}} C_k(Q_k(t)) \\ \text{s.t.} \quad & \tau_j(t) \leq d_j, \quad j \in \mathcal{J}; \\ & \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) + \sum_{j \in \mathcal{J}} m_j^e \times Q_j(t) \approx W(t); \end{split}$$

$$\begin{split} \min \quad & \sum_{k \in \mathcal{K}} C_k(Q_k(t)) \\ \text{s.t.} \quad & \tau_j(t) \leq d_j, \quad j \in \mathcal{J}; \\ & \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) \approx W(t) - \sum_{j \in \mathcal{J}} m_j^e \times Q_j(t); \end{split}$$

min 
$$\sum_{k\in\mathcal{K}} C_k(Q_k(t))$$

s.t.

$$\sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) \approx (W(t) - \sum_{j \in \mathcal{J}} \lambda_j d_j m_j^e)^+$$

min 
$$\sum_{k\in\mathcal{K}} C_k(Q_k(t))$$

s.t.

# Analysis – Asymptotic Framework What is "nearly optimal"?

## **Asymptotic Framework**

- ▶ A sequence of systems, indexed by  $r \uparrow \infty$  :
- ▶ Arrival rate for class j triage patients  $\lambda_i^r \to \lambda_j > 0$ ,  $j \in \mathcal{J}$ ;
- Service requirement & routing behavior do not change;
- Traffic intensity:  $\rho^r = \sum_{j \in \mathcal{J}} \lambda_j^r m_i^e$ ;
- ▶ (Conventional) heavy traffic condition: there exists a  $\beta \in \mathbb{R}$ ,

$$r(\rho^r-1)\to \beta$$
, as  $r\to \infty$ .

▶ A family of control policies  $\{\pi^r\}$  is called *asymptotically* compliant (feasible) if for any fixed T > 0, as  $r \to \infty$ ,

$$\sup_{0 \le t \le T} \left[ \widehat{\tau}_j^r(t) - \frac{d_j^r}{r} \right]^+ \Rightarrow 0, \quad j \in \mathcal{J}.$$

 $\widehat{\tau}_i^r(t) := \frac{1}{r} \tau_i^r(r^2 t)$  – diffusion scaled age processes w.r.t  $\{\pi^r\}$ .

## **Asymptotic Optimization**

▶ Diffusion scaled queue length processes:

$$\widehat{Q}_k^r(t) = rac{1}{r} Q_k^r(r^2 t), \quad k \in \mathcal{K}.$$

Cumulative queueing cost:

$$\mathcal{U}^r(t) := \int_0^t \sum_{k \in \mathcal{K}} C_k\left(\widehat{Q}_k^r(s)\right) \mathrm{d}s.$$

- A family of control policies  $\{\pi_*^r\}$  is said to be *asymptotically optimal* if
  - ▶ it is asymptotically compliant and
  - it stochastically minimizes the cumulative cost:

$$\limsup_{r\to\infty}\mathcal{U}^r_*(t)\leq_{s.t.}\liminf_{r\to\infty}\mathcal{U}^r(t),$$

 $\{\mathcal{U}_*^r\}$  – queueing cost corresponding to  $\{\pi_*^r\}$ ;

 $\{\mathcal{U}^r\}$  – corresponding to any asymptotically compliant policies.

The proposed family of control policies is asymptotically optimal.

## A Roadmap for the Proofs

- There is a lower bound for any asymptotically compliant family of policies;
- ► The proposed family of scheduling policies achieves the lower bound;
  - State-Space-Collapse (SSC);

#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

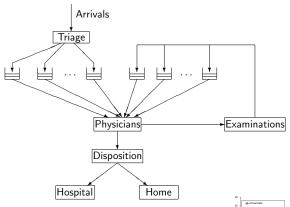
#### **Conclusions – Contributions**

- Practical: model and analyze the control of patient flow in EDs:
  - Give rise to insightful and implementable scheduling policies;
  - Capture the <u>tradeoff</u> between triage- vs. WIP-patients;
  - Enable analysis of the value of **information** in a real ED setup.
- ► Theoretical: analyze multiclass queueing systems with feedback:
  - Prove the <u>conjecture</u> in Mandelbaum and Stolyar (OR, 2004), improve upon it with simpler asymptotically optimal policies;
  - **G***c*μ **rule** for Klimov's model with convex costs (queue length, individual waiting times and cumulative sojourn times);
  - Analyze multiclass queueing systems with feedback, under <u>any</u> work-conserving policy;

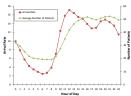
#### **Outline**

- Motivation
- Modeling ED Operations
- ▶ Two ED Models
  - Queue Length Model
  - Sojourn Time Model
- Case Study
- Intuition and Technical Ideas
- Summary and Contributions
- Future Directions

#### **Future Directions**



- ▶ Other features:
  - Time varying arrival rate;
  - Adding delays between transfers;
  - Adding global constraint on sojourn times;
  - Adding abandonment (LWBS, LAMA);



# Thank You! Questions?

min 
$$\sum_{k \in \mathcal{K}} C_k(Q_k(t))$$

s.t. 
$$au_j(t) \leq d_j, \quad j \in \mathcal{J};$$

$$\begin{split} \min \quad & \sum_{k \in \mathcal{K}} C_k(Q_k(t)) \\ \text{s.t.} \quad & \tau_j(t) \leq d_j, \quad j \in \mathcal{J}; \\ & \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) \approx W(t) - \sum_{i \in \mathcal{J}} m_i^e \times Q_j(t); \end{split}$$

$$\begin{split} \min \quad & \sum_{k \in \mathcal{K}} C_k(Q_k(t)) \\ \text{s.t.} \quad & \tau_j(t) \leq d_j, \quad j \in \mathcal{J}; \qquad \left(Q_j(t) \approx \lambda_j \tau_j(t)\right) \\ & \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) \approx W(t) - \sum_{j \in \mathcal{J}} m_j^e \times Q_j(t); \end{split}$$

$$\begin{split} \min \quad & \sum_{k \in \mathcal{K}} C_k(Q_k(t)) \\ \text{s.t.} \quad & \tau_j(t) \leq d_j, \quad j \in \mathcal{J}; \qquad (Q_j(t) \approx \lambda_j \tau_j(t)) \\ & \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) \approx \boxed{W(t) - \sum_{j \in \mathcal{J}} m_j^e \times Q_j(t);} \\ & \qquad \qquad \text{Minimize} \\ & \qquad (W(t) - \sum_{j \in \mathcal{J}} \lambda_j d_j m_j^e)^+ \end{split}$$

$$\min \quad \sum_{k \in \mathcal{K}} C_k(Q_k(t))$$
s.t.  $\tau_j(t) \leq d_j, \quad j \in \mathcal{J}; \qquad (Q_j(t) \approx \lambda_j \tau_j(t))$ 

$$\sum_{k \in \mathcal{K}} m_k^e \times Q_k(t) \approx \boxed{W(t) - \sum_{j \in \mathcal{J}} m_j^e \times Q_j(t);}$$

$$Minimize$$

$$(W(t) - \sum_{j \in \mathcal{J}} \lambda_j d_j m_j^e)^+$$

$$VS. d_1 \leftarrow \frac{\tau_j(t)}{d_j} \approx \frac{\tau_{j'}(t)}{d_{j'}} \qquad A \text{ threshold policy}$$

$$\sum_{j \in \mathcal{J}} \lambda_j m_j^e \tau_j(t) \text{ vs. } \sum_{j \in \mathcal{J}} \lambda_j m_j^e d_j$$

min 
$$\sum_{k\in\mathcal{K}} C_k(Q_k(t))$$

s.t.

$$\sum_{k \in \mathcal{K}} m_k^e imes Q_k(t) pprox (W(t) - \sum_{j \in \mathcal{J}} \lambda_j d_j m_j^e)^+$$

min 
$$\sum_{k\in\mathcal{K}} C_k(Q_k(t))$$

s.t.