# A game theoretic model of the behavioural gaming that takes place at the EMS - ED interface

## 1 Abstract

Emergency departments (EDs) in hospitals are usually under pressure to achieve a target amount of time that describes the arrival of patients and the time it takes to receive treatment. For example in the UK this is often set as 95% of patients to be treated within 4 hours. There is empirical evidence to suggest that imposing targets in the ED results in gaming at the interface of care between the EMS and ED. If the ED is busy and a patient is stable in the ambulance, there is little incentive for the ED to accept the patient whereby the clock will start ticking on the 4 hour target. This in turn impacts on the ability of the EMS to respond to emergency calls.

This study explores the impact that this effect may have on an ambulance's utilisation and their ability to respond to emergency calls. More specifically multiple scenarios are examined where an ambulance service needs to distribute patients between neighbouring hospitals. The interaction between the hospitals and the ambulance service is defined in a game theoretic framework where the ambulance service has to decide how many patients to distribute to each hospital in order to minimise the occurrence of this effect. The methodology involves the use of a queueing model for each hospital that is used to inform the decision process of the ambulance service so as to create a game for which the Nash Equilibria can be calculated.

## 2 Introduction

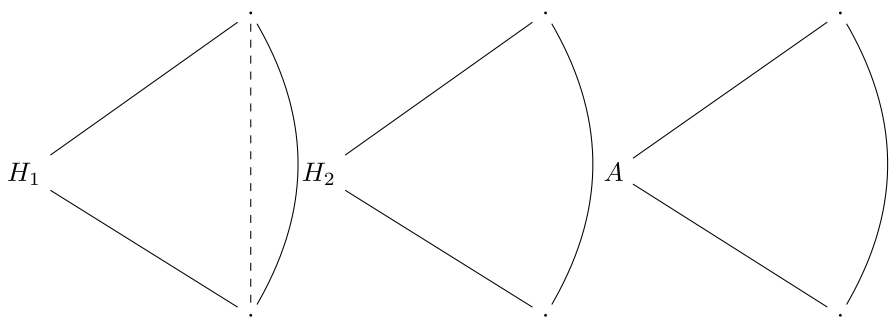



Figure 1: Ambulance Decision Problem

### States:

1. $A$ = Ambulance
2. $H_i$ = Hospital i

### Notation:

- $\Lambda$ = total number of patients that need to be hospitalised
- $p_i$ = proportion of patients going to Hospital i ($p_i\Lambda$ = number of patients going to hospital i)
- $d_i$ = distance from Hospital i
- $\hat{c}_i$ = capacity of hospital i
- $W(c, \lambda\, \mu)$ = waiting time in the system function
- $\mu_i$ = service rate of hospital i
- $\lambda_i^o$ = arrival rate of other patients to the hospital (not by ambulance)
- $C_i(p_i) = d_i + W(c = \hat{c}_i,\ \lambda = p_i\Lambda + \lambda_i^o,\ \mu = \mu_i)$

## 3 Queuing Theory component

## 4 Game Theory component:

**Players:**

- Ambulance
- Hospital A

- Hospital B

**Strategies of players:**

- Hospital i:
    1. Close doors at $\hat{c}_i = 1$
    2. Close doors at $\hat{c}_i = 2$
    3. ...
    4. Close doors at $\hat{c}_i = C_i$
- Ambulance:
    1. Choose $p_1 \in [0, 1]$

**Cost Functions:** Waiting times + the distance to each hospital.

# 5 Formulas

$$\hat{c}_i \in \{1, 2, \dots, C_i\}$$

$$\rho_i = \frac{p_i \Lambda + \lambda_i^o}{\hat{c}_i \mu_i}$$

$$(W_q)_i = \frac{1}{\hat{c}_i \mu_i} \frac{(\hat{c}_i \rho_i)^{\hat{c}_i}}{\hat{c}_i!(1-\rho_i)^2} (P_0)_i$$

$$(P_0)_i = \frac{1}{\sum_{n=0}^{\hat{c}_i - 1} \left[ \frac{(\hat{c}_i \rho_i)^n}{n!} \right] + \frac{(\hat{c}_i \rho_i)^{\hat{c}_i}}{\hat{c}_i!(1-\rho_i)}}$$

$$P(W_q > T) = \frac{(\frac{\lambda}{\mu})^c P_0}{c!(1 - \frac{\lambda}{c\mu})} (e^{-(c\mu - \lambda)T})$$

# 6 Quick Methodology

- Fix the parameters $\Lambda$, $\lambda_i^o$, $\mu_i$ and $C_i$.
- $\forall\ \hat{c_i} \in \{1, 2, \dots, C_A\}$ and $\forall\ \hat{c_j} \in \{1, 2, \dots, C_B\}$
- Calculate $p_A$ and $p_B = 1 - p_A$ s.t. $(W_q)_A = (W_q)_B$.
- Calculate the probability $P((W_q)_i \leq 4$ hours
- Fill matrix A with $U^A_{\hat{c_i},\hat{c_j}} = 1 - |0.95 - P((W_q)_A \leq 4)|$ and
- fill matrix B with $U^B_{\hat{c_i},\hat{c_j}} = 1 - |0.95 - P((W_q)_B \leq 4)|$

$$A = \begin{array}{|c|c|c|c|}\hline U^A_{1,1} & U^A_{1,2} & \cdots & U^A_{1,C_B} \\ \hline U^A_{2,1} & U^A_{2,2} & \cdots & U^A_{2,C_B} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline U^A_{C_A,1} & U^A_{C_A,2} & \cdots & U^A_{C_A,C_B} \\ \hline \end{array}$$

$$B = \begin{array}{|c|c|c|c|}\hline U^B_{1,1} & U^B_{1,2} & \cdots & U^B_{1,C_B} \\ \hline U^B_{2,1} & U^B_{2,2} & \cdots & U^B_{2,C_B} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline U^B_{C_A,1} & U^B_{C_A,2} & \cdots & U^B_{C_A,C_B} \\ \hline \end{array}$$

- Ambulance decides the proportion of people to distribute to each hospital based on optimal patient distribution.

# 7 Proper Methodology

The problem is formulated as a normal form game where the players are the two hospitals. Each hospital is given $C_A$ and $C_B$ number of strategies where $C_A$ and $C_B$ are the total capacities of the hospitals. In other words, depending on the capacity of each hospital, they may choose to stop receiving patients from arriving ambulances whenever they reach a certain capacity threshold. The goal of this problem is to satisfy the ED regulations which state that 95% of the patients should see a specialist within 4 hours of their arrival to the hospital. The mean of the random variable $W_q$ is the average waiting time in the queue for hospital i.

$$W_q(\lambda_i, \mu_i, \hat{c_i}) = \frac{1}{\hat{c_i}\mu_i} \frac{(\hat{c_i}\rho_i)^{\hat{c_i}}}{\hat{c_i}!(1-\rho_i)^2} P_0, \quad i \in \{A, B\} \tag{1}$$

Thus, the utilities of the two players should be the proportion of people that fall within the 4 hours target. This is also equivalent to the probability of the waiting time of an individual to be less than or equal to 4 hours.

$$P(W_q(\lambda_i, \mu_i, \hat{c_i}) \leq 4), \quad i \in \{A, B\} \tag{2}$$

Therefore, a sensible goal for each player should be to minimise that probability, but the actual target of the hospitals is to satisfy 95% of those patients within the 4-hour time limit. Therefore, the goal should be to get that probability as close to 0.95 as possible. Thus each player should aim to minimise:

$$|0.95 - P(W_q(\lambda_i, \mu_i, \hat{c_i}) \leq 4)|, \quad i \in \{A, B\} \tag{3}$$

The classic formulation of a normal form game looks into the maximisation of each player's payoff. Consequently the utilities can be altered such that the goal of each player is to maximise:

$$U^A_{\hat{c_A}, \hat{c_B}} = 1 - |0.95 - P(W_q(\lambda_A, \mu_A, \hat{c_A}) \leq 4)| \tag{4}$$
$$U^B_{\hat{c_A}, \hat{c_B}} = 1 - |0.95 - P(W_q(\lambda_B, \mu_B, \hat{c_B}) \leq 4)| \tag{5}$$

Finally, the problem can be expressed as a normal form game with two players where each player/hospital has $C_A$ and $C_B$ strategies respectively. The two $C_A \times C_B$ payoff matrices for the utilities of the two hospitals can be defined as:

$$\text{A} = \begin{array}{|c|c|c|c|} \hline U^A_{1,1} & U^A_{1,2} & \dots & U^A_{1,C_2} \\ \hline U^A_{2,1} & U^A_{2,2} & \dots & U^A_{2,C_2} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline U^A_{C_1,1} & U^A_{C_1,2} & \dots & U^A_{C_1,C_2} \\ \hline \end{array} \qquad \text{B} = \begin{array}{|c|c|c|c|} \hline U^B_{1,1} & U^B_{1,2} & \dots & U^B_{1,C_2} \\ \hline U^B_{2,1} & U^B_{2,2} & \dots & U^B_{2,C_2} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline U^B_{C_1,1} & U^B_{C_1,2} & \dots & U^B_{C_1,C_2} \\ \hline \end{array}$$

Once the certain strategies of the game have been selected the ambulance service can decide what would be the optimal way to distribute patients. However, the way the ambulance service distributes patients can affect the utilities of the game. So how would one solve this kind of problem?

### 7.1 Solution

As mentioned before the problem requires the construction of two queuing models that will be needed for the formulation of the normal form game. Based on those utilities the ambulance service will then decide the percentage of patients that will distribute to each hospital.

First and foremost, the model consists of several parameters that are unknown and are assumed to be fixed. The model will be run multiple times for various values of these parameters.

| | |
|---|---|
| $\Lambda$ | Number of patients that need to be distributed |
| $\lambda_i^o$ | Arrival rate of other patients that enter hospital i |
| $\mu_i$ | Service rate of hospital i |
| $C_i$ | Total capacity of hospital i |

Table 1: Fixed Parameters

Having established the fixed parameters of the model, the hospitals' utilities need to be calculated. In order to do so a backwards induction approach will be used. The EMS aims to distribute the patients such that the mean waiting time of patients is minimal. This can be further interpreted as when the mean waiting time of hospital A equals the mean waiting time of hospital B. Thus, the minimal mean waiting time can be found for the values of $p_A$ and $p_B$ that solve the following equation:

$$W_q(\lambda_A, \mu_A, \hat{c_A}) = W_q(\lambda_B, \mu_B, \hat{c_B}) \tag{6}$$

Equation (6) needs to be solved for all values of $c_i \in \{1, 2, \dots C_A\}$ and $c_j \in \{1, 2, \dots C_B\}$. Then, for every $c_i$ and $c_j$ the utility equation (4) has to be calculated for both hospitals. In order to solve it though, one must first estimate the probability $P[(W_q)_{\{A,B\}}] \leq 4]$. That is the probability that the waiting time in the queue for one of the hospitals is less than 4 hours. For a multi-server system, the distribution of the waiting time can be given by equation 7. The above expression returns the probability that the waiting time in the queue is less than some time T.

$$P(W_q > T) = \frac{(\frac{\lambda}{\mu})^c P_0}{c!(1 - \frac{\lambda}{c\mu})} (e^{-(c\mu - \lambda)T}) \tag{7}$$

Consequently when incorporating equation (7) into (4) a newer utility equation can be acquired:

$$U_{\hat{c_i}, \hat{c_j}}^{\{A,B\}} = 1 - \left| \left[ \frac{(\frac{\lambda}{\mu})^c P_0}{c!(1 - \frac{\lambda}{c\mu})} \left( e^{-(c\mu - \lambda)T} \right) \right] - 0.05 \right| \tag{8}$$

A =

| | | | |
|---|---|---|---|
| $U_{1,1}^A$ | $U_{1,2}^A$ | $\dots$ | $U_{1,C_2}^A$ |
| $U_{2,1}^A$ | $U_{2,2}^A$ | $\dots$ | $U_{2,C_2}^A$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $U_{C_1,1}^A$ | $U_{C_1,2}^A$ | $\dots$ | $U_{C_1,C_2}^A$ |

B =

| | | | |
|---|---|---|---|
| $U_{1,1}^B$ | $U_{1,2}^B$ | $\dots$ | $U_{1,C_2}^B$ |
| $U_{2,1}^B$ | $U_{2,2}^B$ | $\dots$ | $U_{2,C_2}^B$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $U_{C_1,1}^B$ | $U_{C_1,2}^B$ | $\dots$ | $U_{C_1,C_2}^B$ |

# 8 Hospital Markov chain model

The following Markov chain represents the transition between states of a hospital while capturing the EMS interaction with it. The hospital accepts both ambulance and other patients normally until a certain threshold $T$ is reached. When it is reached all ambulances that arrive will be marked as *"parked outside"* until the number of people in the system is reduced below $T$. Additionally, if the patients in the hospital keep rising, they may exceed the number of servers $C$ available, which will in turn mean that every new patient will have to wait for a server to become free. The states of the Markov chain are denoted by $(u, v)$ where:

- $u$ = number of ambulances parked outside of the hospital
- $v$ = number of patients in the hospital

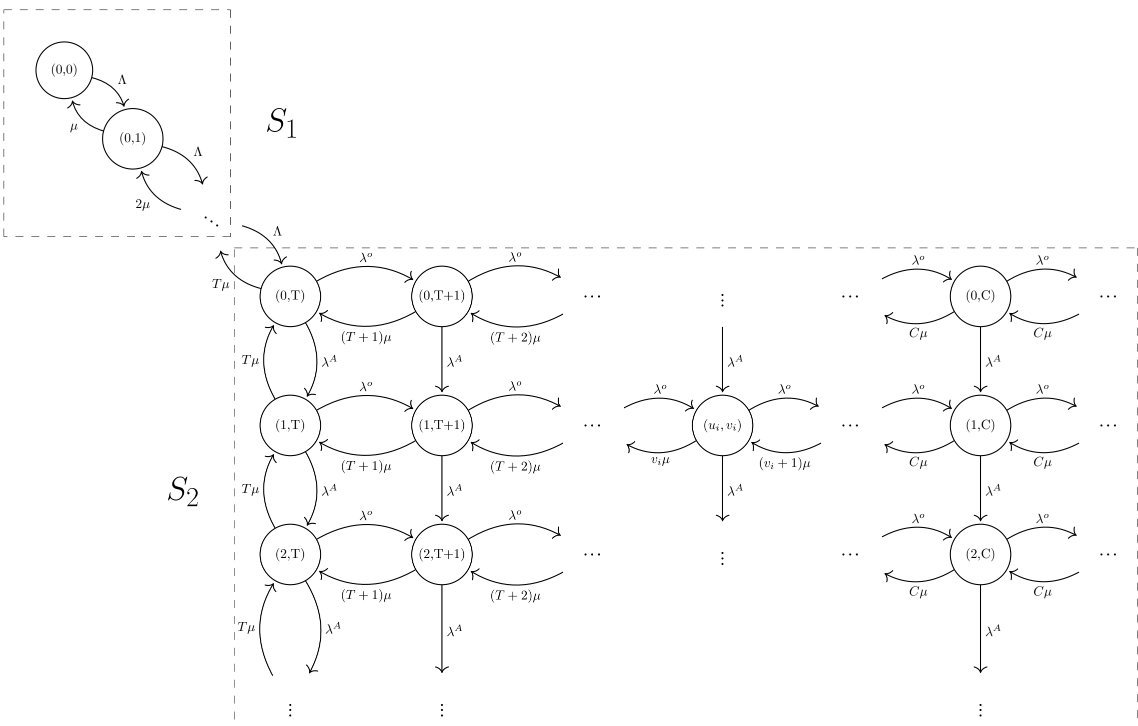



Figure 2: Markov chains

## 8.1 Markov-chain state mapping function

The transition matrix of the Markov-chain representation described above can be denoted by a state mapping function. The state space of this function is defined as:

$$
\begin{aligned}
S(T) =& S_1(T) \cup S_2(T) \text{ where:} \\
S_1(T) =& \left\{(0,v) \in \mathbb{N}_0^2 \mid v < T\right\} \\
S_2(T) =& \{(u,v) \in \mathbb{N}_0^2 \mid v \geq T\}
\end{aligned} \tag{9}
$$

Therefore, the entries of the transition matrix $Q$, can be given by $q_{i,j} = q_{(u_i,v_i),(u_j,v_j)}$ which is the transition rate from state $i = (u_i, v_i)$ to state $j = (u_j, v_j)$ for all $(u_i, v_i), (u_j, v_j) \in S$.

$$
q_{i,j} = \begin{cases}
\Lambda, & \textbf{if } (u_i,v_i) - (u_j,v_j) = (0,-1) \textbf{ and } v_i < \text{t} \\
\lambda^o, & \textbf{if } (u_i,v_i) - (u_j,v_j) = (0,-1) \textbf{ and } v_i \geq \text{t} \\
\lambda^a, & \textbf{if } (u_i,v_i) - (u_j,v_j) = (-1,0) \\
v_i\mu, & \textbf{if } (u_i,v_i) - (u_j,v_j) = (0,1) \textbf{ and } v_i \leq C \textbf{ or} \\
 & \quad (u_i,v_i) - (u_j,v_j) = (1,0) \textbf{ and } v_i = T \leq C \\
C\mu, & \textbf{if } (u_i,v_i) - (u_j,v_j) = (0,1) \textbf{ and } v_i > C \textbf{ or} \\
 & \quad (u_i,v_i) - (u_j,v_j) = (1,0) \textbf{ and } v_i = T > C \\
-\sum_{j=1}^{|Q|} q_{i,j} & \textbf{if } i = j \\
0, & \textbf{otherwise}
\end{cases} \tag{10}
$$

In order to acquire an exact solution of the problem a slight adjustment needs to be considered. The problem defined above assumes no upper boundary to the number of people that can wait for service or the number of ambulances that can be parked outside. Therefore, a different state space $\tilde{S}$ needs to be constructed where $\tilde{S} \subseteq S$ and there is a maximum allowed number of people $N$ that can be in the system and a maximum allowed number of ambulances $M$ parked outside:

$$
\tilde{S} = \{(u,v) \in S \mid u \leq M, v \leq N\} \tag{11}
$$

## 8.2 Steady State

Having calculated the transition matrix $Q$ for a given set of parameters the probability vector $\pi$ needs to be considered. The vector $\pi$ is commonly used to study such stochastic systems and it's main purpose is to keep track of the probability of being at any given state of the system. The term *steady state* refers to the instance of the vector $\pi$ where the probabilities of being at any state become stable over time. Thus, by considering the steady state vector $\pi$ the relationship between it and $Q$ is given by:

$$
\frac{d\pi}{dt} = \pi Q = 0
$$

There are numerous methods that can be used to solve problems of such kind. In this paper only numeric and algebraic approaches will be considered.

### 8.2.1 Numeric integration

The first approach to be considered is to solve the differential equation numerically by observing the behaviour of the model over time. The solution is obtained via python's SciPy library. The functions odeint and solve_ivp have been used in order to find a solution to the problem. Both of these functions can be used to solve any system of first order ODEs.

### 8.2.2 Linear algebraic approach

Another approach to be considered is the linear algebraic method. The steady state vector can be found algebraically by satisfying the following set of equations:

$$\pi Q = 0$$

$$\sum_i \pi_i = 1$$

These equations can be solved by slightly altering $Q$ such that the final column is replaced by a vector of ones. Thus, the resultant solution occurs from solving the equation $\tilde{Q}^T \pi = b$ where $\tilde{Q}$ and $b$ are defined as:

$$\tilde{q_{i,j}} = \begin{cases} 1, & \textbf{if } j = |Q| \\ q_{i,j}, & \textbf{otherwise} \end{cases}$$

$$b = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

### 8.2.3 Least Squares approach

Finally, the last approach to be considered is the least squares method. This approach is considered because while the problem becomes more complex (in terms of input parameters) the computational time required to solve it increases exponentially. Thus, one may obtain the steady state vector $\pi$ by solving the following equation.

$$\pi = \text{argmin}_{x \in \mathbb{R}^d} \|Mx - b\|_2^2$$

## 8.3 Expressions derived from $\pi$:

One may easily derive the average number of individuals that are at any given state using $pi$. The average number of individuals in state $i$ can be calculated by multiplying the number of individuals that are present in state $i$ with the probability of being at that particular state (i.e $\pi_i(u_i + v_i)$). Using this logic it is possible to calculate any performance measures that are related to the mean number of individuals in the system.

Average number of patients in the system:

$$L = \sum_{i=1}^{|\pi|} \pi_i(u_i + v_i) \tag{12}$$

Average number of patients in the hospital:

$$L_H = \sum_{i=1}^{|\pi|} \pi_i v_i \tag{13}$$

Average number of ambulances being blocked:

$$L_A = \sum_{i=1}^{|\pi|} \pi_i u_i \tag{14}$$

Consequently getting the performance measures that are related to the duration of time is not as straightforward. Such performance measures are the mean waiting time in the system and the mean time blocked in the system. Under the scope of this study two approaches have are considered; a recursive algorithm and consequently a closed-form formula.

### 8.3.1 Mean Waiting Time - Recursive

The research question that needs to be answered here is: "When an ambulance/other patient enters the system, what is the expected time that they will have to wait?". In order to formulate the answer to that question one needs to consider all possible scenarios of what state the system can be in when an individual arrives.

To calculate such times one must first identify all states $(u_i, v_i)$ that when the model is in that state and an individual arrives they will incur a wait. For this particular Markov chain, this points to all states that satisfy $v_i \geq C$ i.e. all states where the number of individuals in the hospital exceed the number of servers. It is important to note here that an this is not a strict inequality since when the model is at the state with $C$ patients, the arriving patient will proceed to wait. The set of such states is defined as *waiting states* and can be denoted as a subset of all the states, where:

$$S_w = \{(u, v) \in S \mid v \geq C\}$$

Additionally, there are certain states in the model where arrivals cannot occur. An "ambulance" patient cannot arrive in the model when the model is at any state $(M, v)\ \forall\ v \geq T$ and equivalently an "other" patient cannot arrive whenever the model is at any state $(u, N)\ \forall\ u$ where $N$ is the system capacity and $M$ is the parking capacity. In essence this indicates that the formula will behave differently for ambulance and other patients. Therefore the set of all such states that an arrival may occur are defined as *accepting states* and are denoted as:

$$S_A^A = \{(u, v) \in S \mid v < M\}$$

$$S_A^o = \{(u, v) \in S \mid u < N\}$$

Moreover, another element that needs to be considered is the expected waiting time in each state $c(u, v)$. In order to do so a variation of the Markov model has to be considered where when the individual arrives at any of the states of the model no more arrivals can occur after that. Thus, one may acquire the desired time by calculating the inverse of the sum of the out-flow rate of that state. Since we are ignoring arrivals though the only way to exit the state will only be via a service. In essence this notion can be expressed as:

$$c(u, v) = \frac{1}{\min(v, C)\mu} \tag{15}$$

Equation 15 considers two scenarios of states; whether the number of patients are more than the number of servers or not. Now, using the above equations, and considering all sates that belong in $S_w$ the following recursive formula can be used to get the mean waiting time spent in each state in the Markov model. Note that the formula is different for each type of patients. Thus, the expected waiting time of an "other" patient when they arrive at state $(u, v)$ is given by:

$$w(u, v) = \begin{cases} 0, & \textbf{if } (u, v) \notin S_w \\ c(u, v + 1) + w(0, v - 1), & \textbf{otherwise} \end{cases} \tag{16}$$

The equivalent time for "ambulance" patients is denoted in a slightly different manner. States $s_a$ and $s_n$ represent the *arriving state* and the *next state* to be visited by the recursive formula. The arriving state is the state that the individual will actually arrive in when the model is in state $(u, v)$ and the next state is the one that the patient will move down to. Note here that all *next states* move straight to the first row since no wait occurs for any ambulance patient in any row apart from the first one.

$$w_A(u, v) = \begin{cases} 0 & \textbf{if } (u, v) \notin S_w \\ c(s_a) + w(s_n) & \textbf{otherwise} \end{cases} \tag{17}$$

$$s_a = \begin{cases} (u + 1, v), & \textbf{if } v \geq T \\ (u, v + 1), & \textbf{otherwise} \end{cases} \qquad s_n = \begin{cases} (0, T), & \textbf{if } u \geq 1 \\ (0, v - 1), & \textbf{otherwise} \end{cases}$$

Thus, the overall mean waiting time can be calculated by summing over all expected waiting times of accepting states multiplied by the probability of being at that state and dividing by the probability of an individual arriving in the system:

$$W = \frac{\sum_{(u,v) \in S_A} w(u, v)\pi_{(u,v)}}{\sum_{(u,v) \in S_A} \pi_{(u,v)}} \tag{18}$$

This formula has to be performed twice, one for each type of patient and then the overall expected waiting time can be found by:

#### 8.3.2 Mean Waiting Time - Closed-form

Mean waiting time in the hospital:

$$W_q = \sum_{i=1}^{|\pi|} \frac{\max(v_i - C, 0)\ \pi_i}{\sum_{\substack{j=1\\ i\neq j}}^{|\pi|} q_{ij}} \tag{19}$$

$$W_q = \sum_{i=1}^{|\pi|} \pi_i \frac{v_i - c}{v_i \mu}, \quad v_i > c \tag{20}$$

# 9 Markov chain VS Simulation

## 9.1 Example model

Consider the Markov chain paradigm in figure 3. The illustrated model represents the unrealistically small system of a hospital with a system capacity of five and an ambulance parking capacity of three. The hospital in this particular example also has four servers and a threshold of three; meaning that every ambulance that arrives in a time that there are three or more individuals in the hospital, will proceed to the parking space.

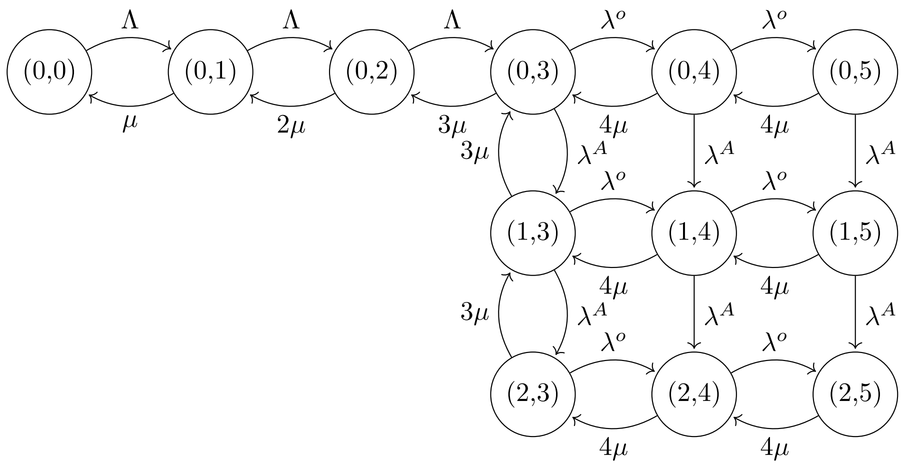



Figure 3: Markov chains: number of servers=4

In addition to the Markov chain model a simulation model has also been built based on the same parameters. Comparing the results of the Markov model and the equivalent simulation model the resultant plots arose.

The heatmaps in figure 4 represent the state probabilities for the Markov chain model, the simulation model and the difference between the two. Each pixel of the heatmap corresponds to the equivalent state of figure 3 and represents the probability of being at that state in any particular moment of time.

It can be observed that both Markov chain and simulation models' state probabilities vary from 5% to 25% and that states $(0, 1)$ and $(0, 2)$ are the most visited ones. Looking at the differences' heatmap, one may identify that the differences between the two are minimal.

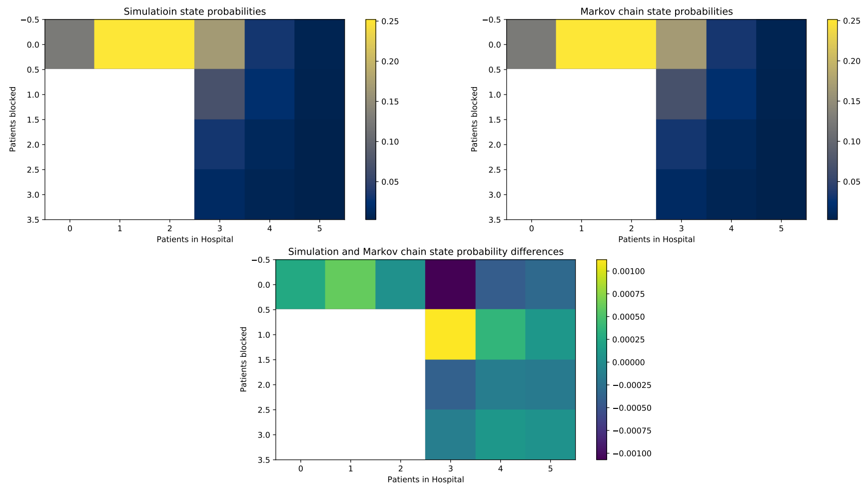
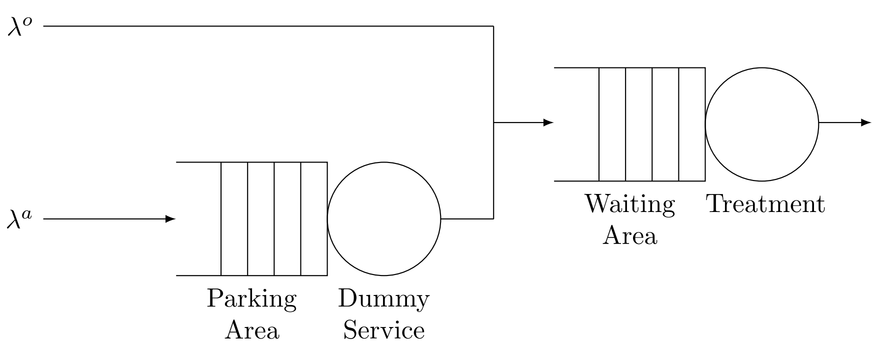



Figure 4: Heatmaps of Simulation, Markov chains and differences of the two

## 10 Figures that might be useful



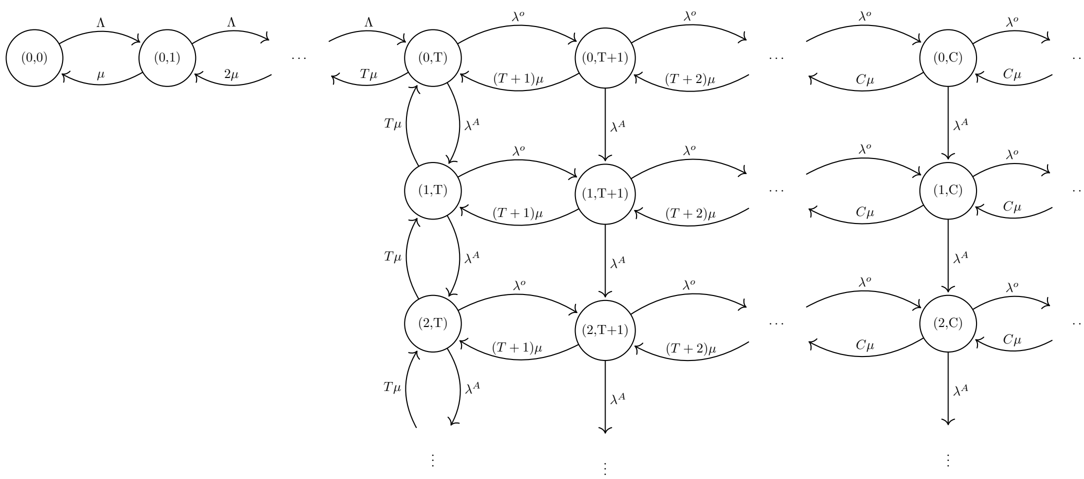



Figure 5: Markov chains

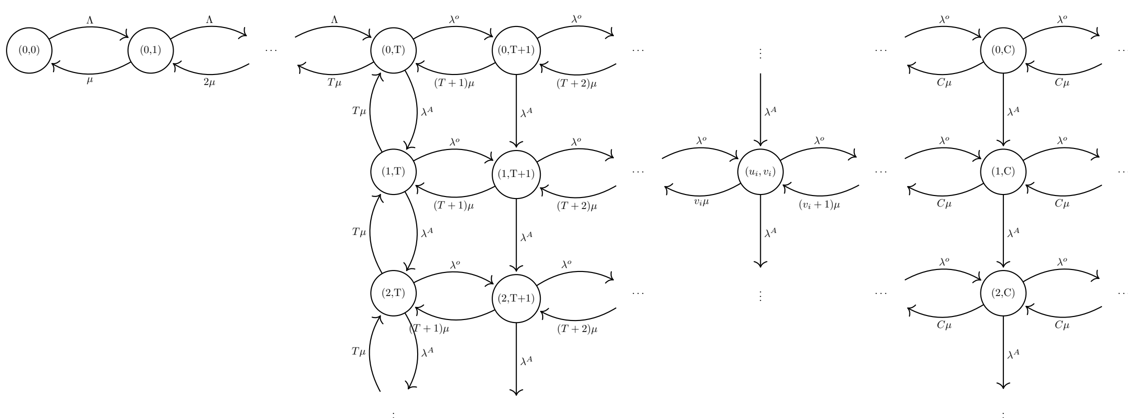



Figure 6: Markov chains

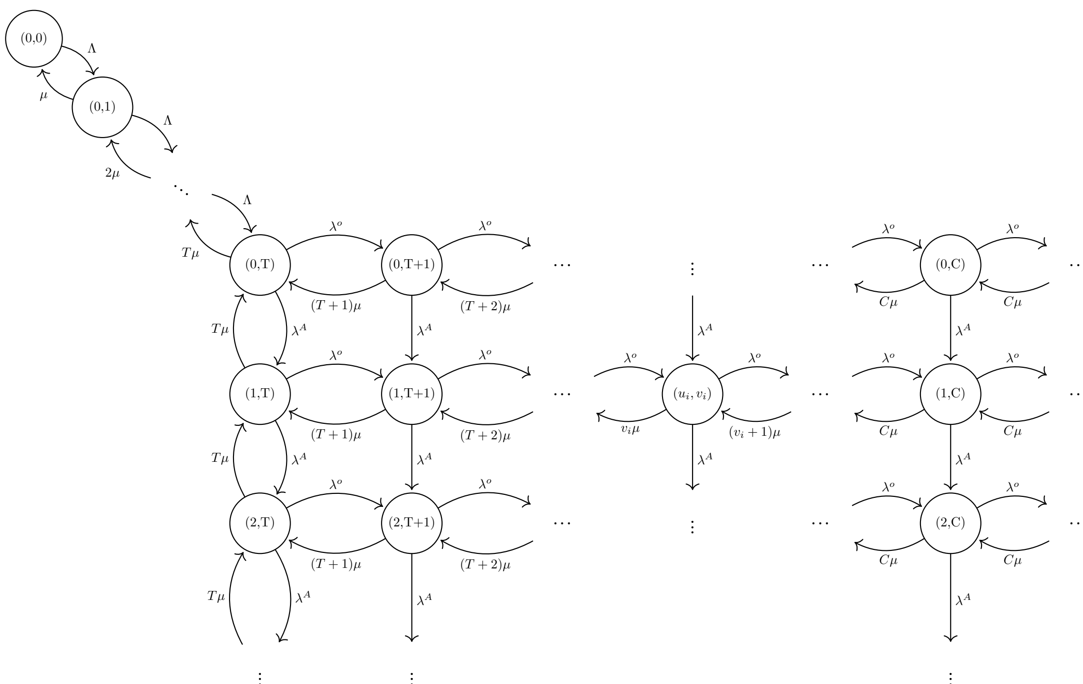



Figure 7: Markov chains