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

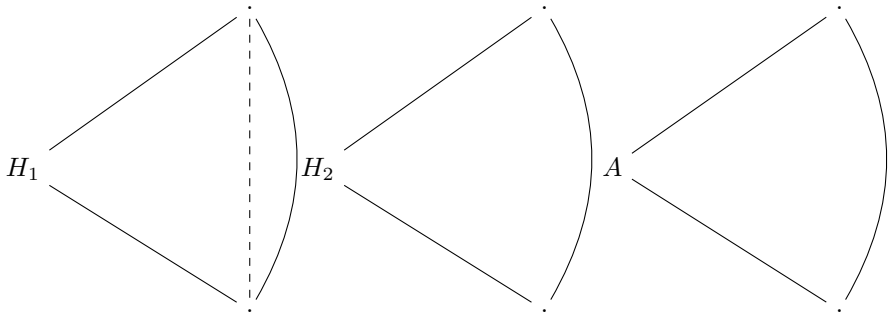


Figure 1: Ambulance Decision Problem

States:

1. A = Ambulance
2. H_i = Hospital i

Notation:

- Λ = 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
- μ_i = service rate of hospital i
- λ_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 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.

4 Quick Methodology

- Fix the parameters Λ, λ_i^o, μ_i and C_i.
- $\forall\ \hat{c_i} \in \{1, 2, \ldots, C_A\}$ and $\forall\ \hat{c_j} \in \{1, 2, \ldots, C_B\}$
- 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 $U^B_{\hat{c_i},\hat{c_j}} = 1 - |0.95 - P((W_q)_B \leq 4)|$

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

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

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

5 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?

5.1 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.

Λ	Number of patients that need to be distributed
λ_i^o	Arrival rate of other patients that enter hospital i
μ_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:

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

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

6 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

6.1 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 $(u_i, v_i), (u_j, v_j) \in S$.

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

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:

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

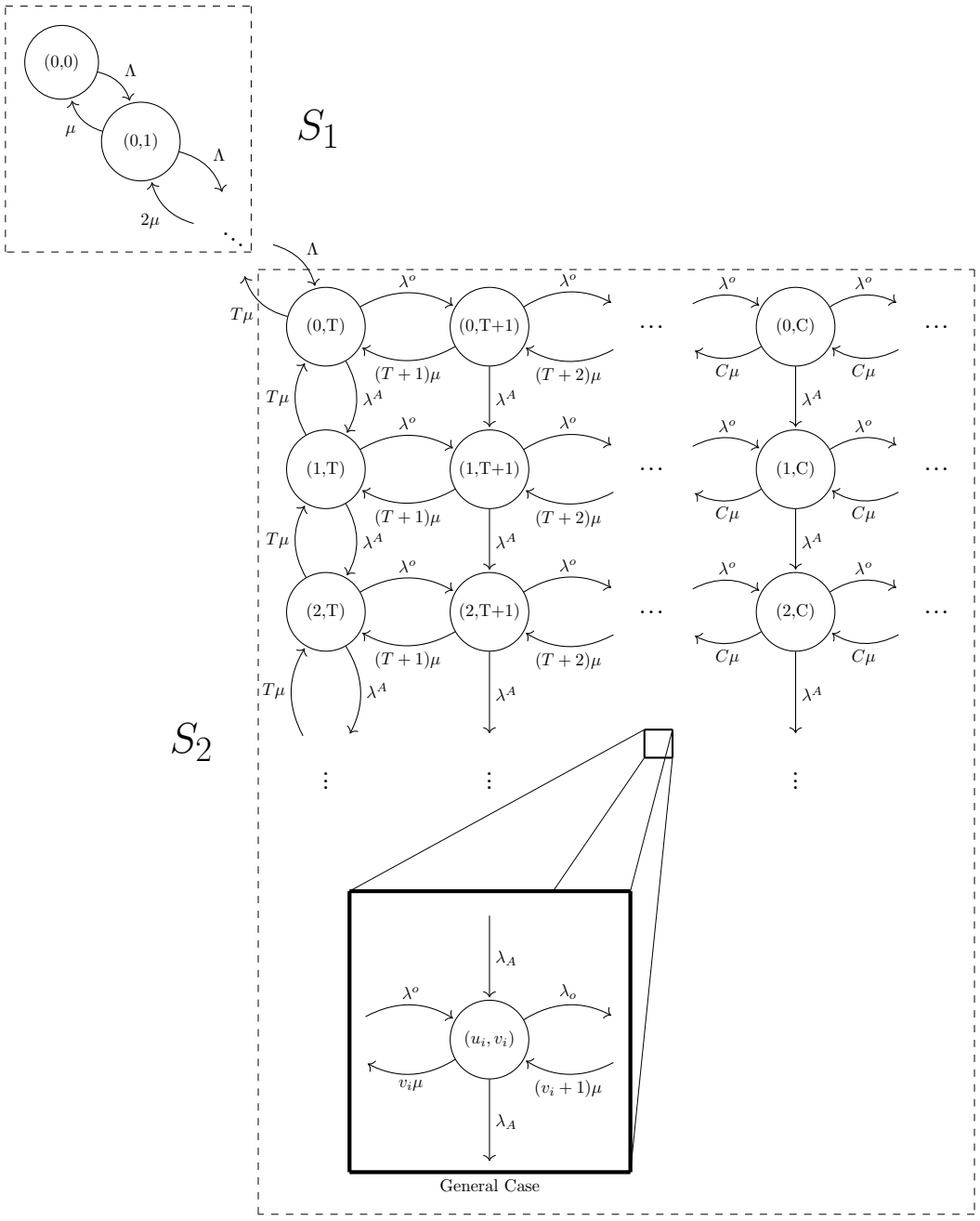


Figure 2: Markov chain

6.2 Steady State

Having calculated the transition matrix Q for a given set of parameters the probability vector π needs to be considered. The vector π 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 π where the probabilities of being at any state become stable over time. Thus, by considering the steady state vector π 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.

6.2.1 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.

6.2.2 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:

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

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

6.2.3 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 π by solving the following equation.

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

6.3 Closed-form formula for state probabilities

This section aims to describe a closed form formula that gets the state probabilities array π for a given Markov chain model.

6.3.1 Parameters

The inputs of the formula are the number of servers C, the threshold T, the system capacity N and the parking capacity M. Additional parameters of the model are the ambulance arrival rate, the others' arrival rate and the service rate, but for the purpose of this section these will remain unknown $(\lambda^A, \lambda^o, \mu)$. More specifically, the way these parameters are translated into the model are:

- **Number of servers (C):** All service rates μ in the Markov chain are multiplied by a coefficient equal to v for a state (u, v) that stops increasing at $v = C$. Thus, the coefficients of the service rate have a lower bound of 0 and an upper bound of C.
- **Threshold (T):** Determines the length of the left *arm* of the model. In essence the threshold acts as a breakpoint between states where $u = 0$ and states where $0 \leq u \leq M$. Increasing T results in having more set of states where u can only be 0.
- **System capacity (N):** Is the upper bound of v for all states (u, v).
- **Parking capacity (M):** Is the upper bound of u for all states (u, v) such that $u \geq T$.

6.3.2 Example figure of Markov Model

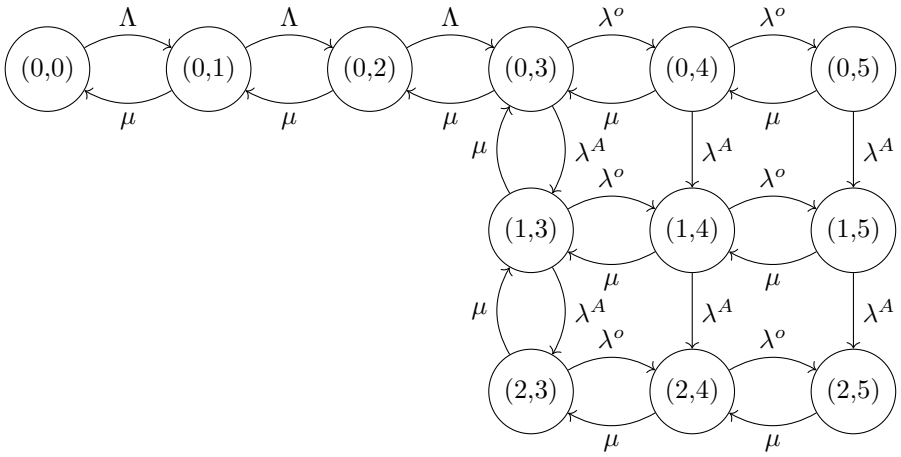


Figure 3: $C = 1, T = 3, N = 5, M = 2$

In figure 3 an example of such Markov model is shown where $C = 1$ meaning the only coefficient in front of any μ is going to be 1, $T = 3$ which means that the *left arm* of the model has a length of 3, $N = 5$ that indicates that the right-most states (u, v) are of the form $(u, 5)$ and $M = 2$ that equivalently shows that the bottom states are of the form $(2, v)$.

6.3.3 Graph theoretical approach for state probabilities

An additional approach that one may consider to get the state probabilities is the graph theoretical approach for state probabilities. Thus, it can be assumed that a Markov chain model M can be translated as a weighted directed graph G_M where every edge has a weight that corresponds to the rate of the edges of the Markov chain.

A *directed spanning tree* of a directed graph is defined as a subset of the graph that visits all the vertices of the graph and does not include any cycles. Unlike undirected spanning trees, directed ones also have a root which means that a directed spanning tree that is rooted at a vertex v has to have a path from any other vertex to vertex v. For example, consider the graph shown in figure 4. The graph points out a spanning tree that is rooted at vertex 3.

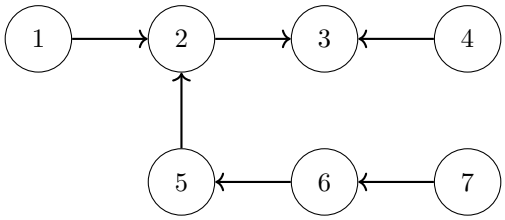


Figure 4: Spanning tree of a graph rooted at vertex 3

Additionally, let us denote the set of all spanning trees of G as $T(G)$ and the subset of $T(G)$ that includes only the spanning trees that are rooted at vertex v as $T_v(G)$. The weight of a spanning tree t can be defined as the product of the weights of the edges it contains:

$$w(t) = \prod_{e \in t} w(e)$$

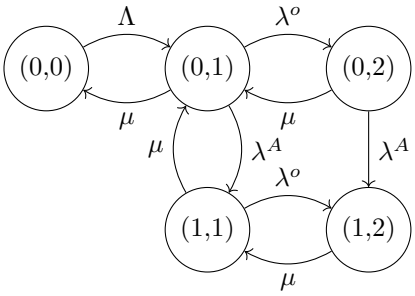
Theorem: Markov chain tree theorem
Let M be an irreducible Markov chain on n states with stationary distribution $\pi_1, \pi_2, \ldots, \pi_n$. Let G_M be the directed graph associated with M. Then the probability of being at state u is given by:

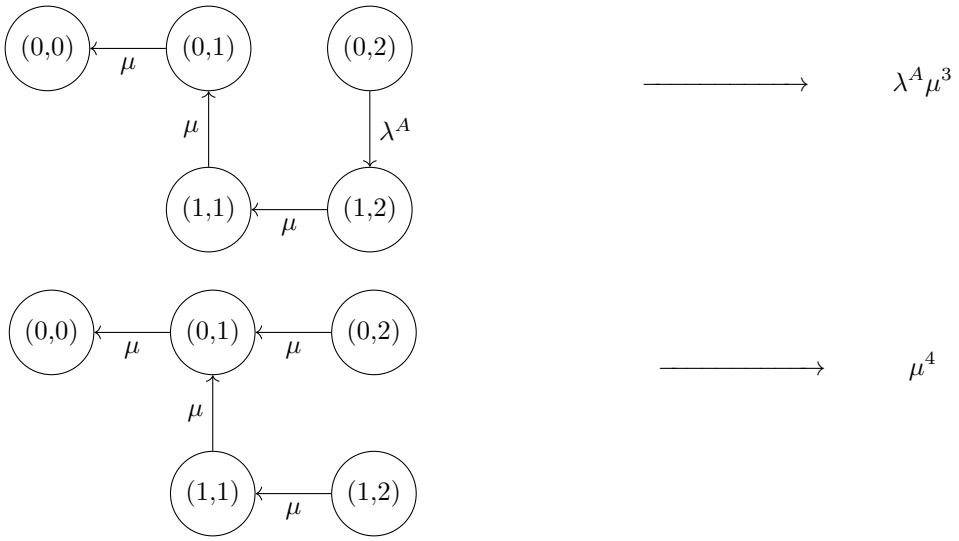
$$\pi_i = \frac{\sum_{t \in T_i(G_M)} w(t)}{\sum_{t \in T(G_M)} w(t)} \tag{12}$$

Equation 12 states that the probability of being at state u can be found by dividing the sum of the weights of all spanning trees rooted at u by the sum of the weights of all spanning trees of the graph. Let us ignore the denominator of that fraction for now and focus only on the numerator denoted as $\tilde{\pi}_i = \sum_{t \in T_i(G_M)} w(t)$

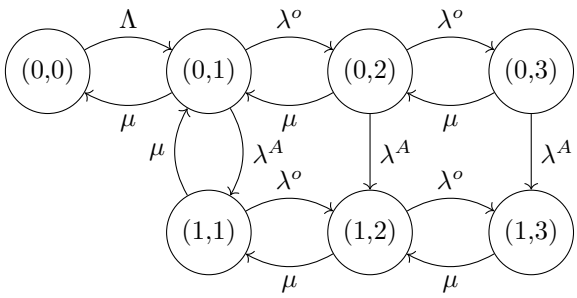
6.3.4 Spanning Trees rooted at $(0,0)$

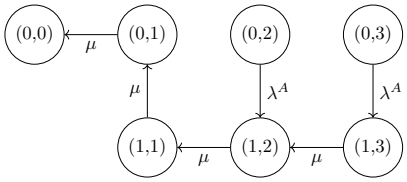
Let us now consider some examples of spanning trees that are rooted at $(0,0)$. For each of the following examples the complete model is shown, then all possible spanning trees rooted at $(0,0)$ along with the weight associated with each spanning tree and finally the value of $\tilde{\pi}_{(0,0)}$ which is the sum of all the weights of the spanning trees.



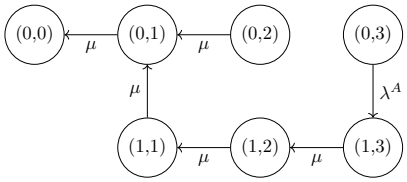


$$\tilde{\pi}_{(0,0)} = \mu^4 + \lambda^A \mu^3$$

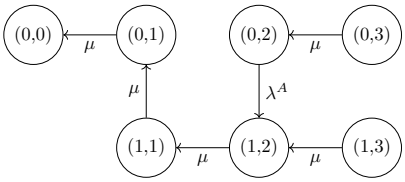




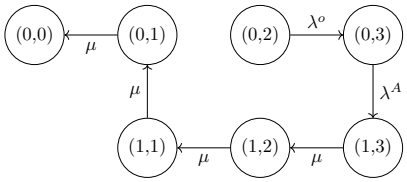
$\longrightarrow \quad (\lambda^A)^2\mu^4$



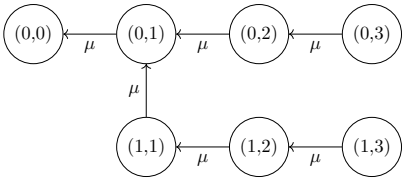
$\longrightarrow \quad \lambda^A\mu^5$



$\longrightarrow \quad \lambda^A\mu^5$

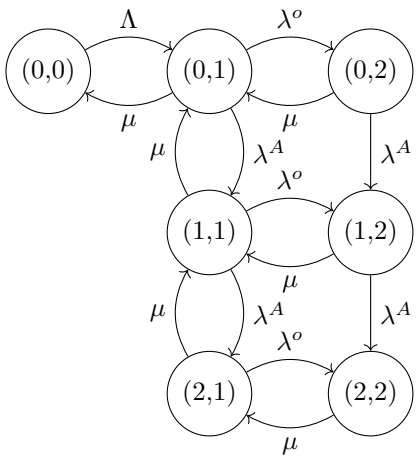


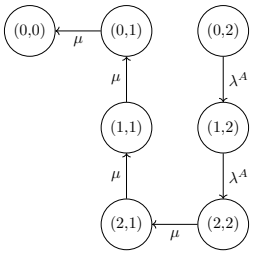
$\longrightarrow \quad \lambda^A\lambda^o\mu^4$



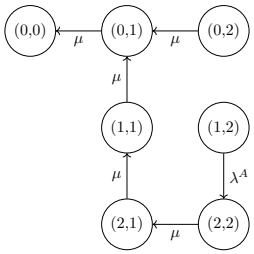
$\longrightarrow \quad \mu^6$

$$\tilde{\pi}_{(0,0)} = (\lambda^A)^2\mu^4 + 2\lambda^A\mu^5 + \lambda^A\lambda^o\mu^4 + \mu^6$$

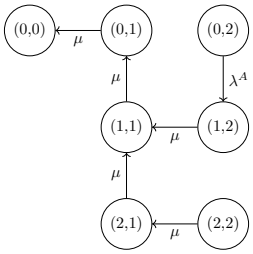




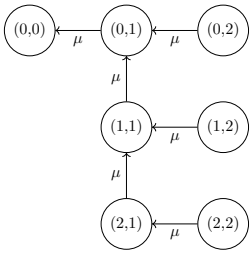
$(\lambda^A)^2\mu^4$



$\lambda^A\mu^5$



$\lambda^A\mu^5$



μ^6

$$\tilde{\pi}_{(0,0)} = (\lambda^A)^2\mu^4 + 2\lambda^A\mu^5 + \mu^6$$

6.3.5 Conjecture of adding rows

Let us consider three Markov models with the same number of servers $C = 1$, the same threshold $T = 1$, the same system capacity $N = 2$ but different parking capacity $M = 1$, $M = 2$ and $M = 3$.

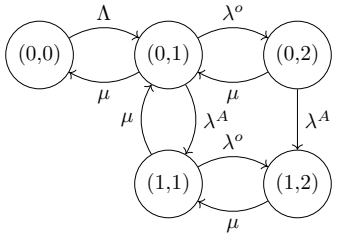


Figure 5: $M = 1$

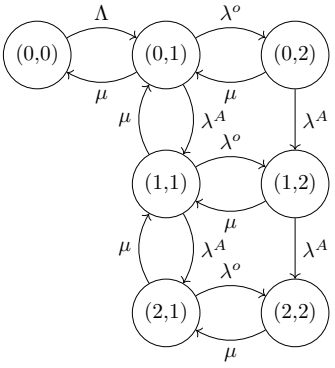


Figure 6: $M = 2$

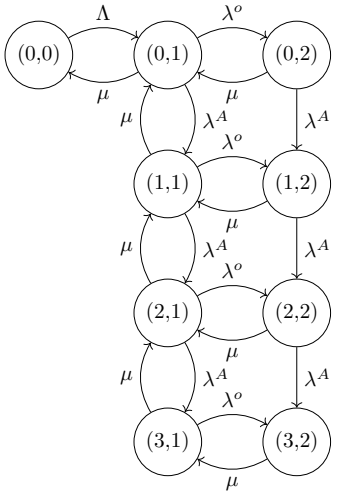


Figure 7: $M = 3$

By increasing the parking capacity of the model it can be easily observed that the number of spanning trees rooted at $(0, 0)$ increases as well since more combinations of paths can be generated using the new edges and vertices. The corresponding values of $\tilde{\pi}_{(0,0)}$ of the three models are:

$$M = 1 : \tilde{\pi}_{(0,0)} = \mu^4 + \mu^3\lambda_A = \mu^3(\mu + \lambda^A) \tag{13}$$

$$M = 2 : \tilde{\pi}_{(0,0)} = \mu^6 + 2\mu^5\lambda_A + \mu^4(\lambda^A)^2 = \mu^4(\mu^2 + 2\mu\lambda_A + (\lambda^A)^2) = \mu^4(\mu + \lambda^A)^2 \tag{14}$$

$$\begin{aligned} M = 3 : \tilde{\pi}_{(0,0)} &= \mu^8 + 3\mu^7\lambda^A + 3\mu^6(\lambda^A)^2 + \mu^5(\lambda^A)^3 \\ &= \mu^5(\mu^3 + 3\mu^2\lambda^A + 3\mu(\lambda^A)^2 + (\lambda^A)^3) \\ &= \mu^5(\mu + \lambda^A)^3 \end{aligned} \tag{15}$$

It can be observed from equations (13),(14) and (15), that there is a noticeable relationship between them. Thus, a generalised formula for the value of $\tilde{\pi}_{(0,0)}$ when $C = 1, T = 1$ and $N = 1$ is given by:

$$\tilde{\pi}_{(0,0)} = \mu^{(N+M)}(\mu + \lambda^A)^M \tag{16}$$

It is important to note here that the above property holds when the system capacity is greater than one as well ($N \geq 1$). For instance let us consider a Markov model with C number of servers, a threshold of T, a system capacity of N and a parking capacity of M. The equivalent values of $\tilde{\pi}_{(0,0)}$ can be expressed in terms of an unknown function $k(C, T, N)$ as:

$$\tilde{\pi}_{(0,0)} = \mu^{(N+M)}(k(C, T, N))^M \tag{17}$$

6.3.6 Matrix-tree theorem for directed graphs (Kirchhoff's theorem):

The number of directed spanning trees rooted at a state i can be found by calculating the determinant of the Laplacian matrix Q of the directed graph and removing row i and column i.

6.4 Expressions derived from π:

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:

$$L = \sum_{i=1}^{|\pi|} \pi_i(u_i + v_i) \tag{18}$$

Average number of patients in the hospital:

$$L_H = \sum_{i=1}^{|\pi|} \pi_i v_i \tag{19}$$

Average number of ambulances being blocked:

$$L_A = \sum_{i=1}^{|\pi|} \pi_i u_i \tag{20}$$

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 been considered to calculate these performance measures; a recursive algorithm and consequently a closed-form formula.

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. Furthermore, a different recursive formula arises for *ambulance patients* and a different one for *other patients*.

6.5 Mean waiting time

6.5.1 Recursive formula for mean waiting time of *other patients*

To calculate the mean waiting time of *other patients* one must first identify the set of states (u, v) that will imply that a wait will occur. For this particular Markov chain, this points to all states that satisfy $v > C$ i.e. all states where the number of individuals in the hospital exceed the number of servers. The set of such states is defined as *waiting states* and can be denoted as a subset of all the states, where:

$$S_w = \{(u, v) \in S \mid v > C\} \tag{21}$$

Additionally, there are certain states in the model where arrivals cannot occur. An *other patient* cannot arrive whenever the model is at any state (u, N) for all u where N is the system capacity.

Therefore the set of all such states that an arrival may occur are defined as *accepting states* and are denoted as:

$$S_A^{(o)} = \{(u, v) \in S \mid u < N\} \tag{22}$$

Moreover, another element that needs to be considered is the expected waiting time in each state $c(u, v)$, otherwise known as sojourn time. 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.

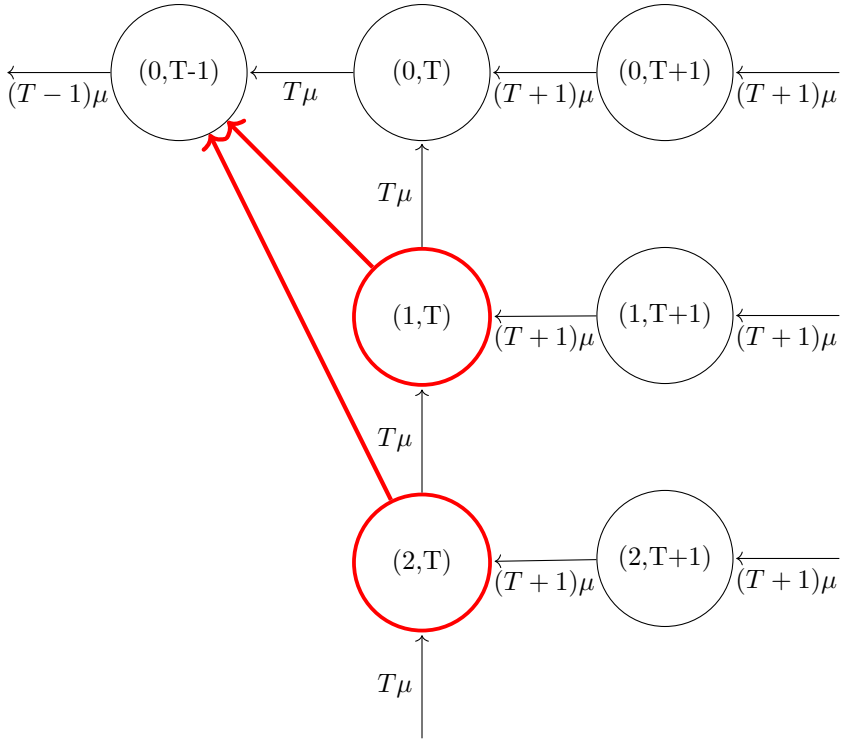


Figure 8: Markov chain - ignoring any arrivals

As illustrated in figure 8 an *other patient*, when in the threshold column, only visits one of the nodes since they are not affected by *ambulance patients*. 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:

$$c^{(o)}(u,v) = \begin{cases} 0, & \textbf{if } u > 0 \textbf{ and } v = T \\ \frac{1}{\min(v,C)\mu}, & \textbf{otherwise} \end{cases} \tag{23}$$

Note that whenever any *other patient* is at a state (u, v) where $u > 0$ and $v = T$ (i.e. all states $(1, T), (2, T) \ldots, (M, T)$) the sojourn time is set to 0. This is done to capture the trip thorough the Markov chain from the perspective of other patients. Meaning that they will visit all states of the threshold column but only the one in the first row will return a non-zero sojourn time.

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. For *other patients*, whenever the model is at state (u, v), any incoming patient will proceed to arrive at state $(u, v + 1)$. Patients will then proceed to visit all other states until they reach one which has less than C servers occupied (i.e. until a server becomes available). The formula goes through all states from right to left recursively and adds the sojourn times of all these states together until it reaches a state that is not in the set of waiting states. Thus, the expected waiting time of an *other patient* when they arrive at state (u, v) can be given by:

$$w^{(o)}(u,v) = \begin{cases} 0, & \textbf{if } (u,v) \notin S_w \\ c^{(o)}(u,v) + w^{(o)}(u-1,v), & \textbf{if } u > 0 \textbf{ and } v = T \\ c^{(o)}(u,v) + w^{(o)}(u,v-1), & \textbf{otherwise} \end{cases} \tag{24}$$

Finally, 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 being in any accepting state.

$$W^{(o)} = \frac{\sum_{(u,v)\in S_A^{(o)}} w^{(o)}(u,v)\pi_{(u,v)}}{\sum_{(u,v)\in S_A^{(o)}} \pi_{(u,v)}} \tag{25}$$

6.5.2 Recursive formula for mean waiting time of *ambulance patients*

Equivalently the mean waiting time for *ambulance patients* can be calculated in a similar manner. The set of waiting states is the same as before but there is a slight modification in the set of accepting states.

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

$$S_A^{(a)} = \begin{cases} \{(u,v) \in S \mid v < M\} & \textbf{if } T \leq N \\ \{(u,v) \in S \mid v < N\} & \textbf{otherwise} \end{cases} \tag{26}$$

The set of accepting states is modified in such a way such that an *ambulance patient* cannot arrive in the model when the model is at any state (M, v) for all $v \geq T$ where M is the parking capacity and T is the threshold. An odd situation here is when the threshold is set to a very high number that is more than the actual system capacity. In such cases the set of accepting states is defined in the same way as the *other patients* case. That is because whenever $T > N$ no ambulance will ever be blocked in the model (since that threshold can never be reached) and thus the last accepting state of the model will be state $(0, N - 1)$.

Now just like in the *other patients* case the sojourn time is needed. For *ambulance patients* whenever individuals are at any row apart from the first one they automatically get a wait time of 0 since they are essentially blocked.

$$c^{(a)}(u,v) = \begin{cases} 0, & \textbf{if } u > 0 \\ \frac{1}{\min(v,C)\mu}, & \textbf{otherwise} \end{cases} \tag{27}$$

Finally, the recursive formula and the mean waiting time equation are identical to the ones described above with the exception that they now use $c^{(a)}(u,v)$ instead of $c^{(o)}(u,v)$.

$$w^{(a)}(u,v) = \begin{cases} 0, & \textbf{if } (u,v) \notin S_w \\ c^{(a)}(u,v) + w^{(a)}(u-1,v), & \textbf{if } u > 0 \textbf{ and } v = T \\ c^{(a)}(u,v) + w^{(a)}(u,v-1), & \textbf{otherwise} \end{cases} \tag{28}$$

$$W^{(a)} = \frac{\sum_{(u,v)\in S_A^{(a)}} w^{(a)}(u,v)\pi_{(u,v)}}{\sum_{(u,v)\in S_A^{(a)}} \pi_{(u,v)}} \tag{29}$$

6.5.3 Mean Waiting Time - Closed-form

Upon closer inspection of the recursive formula a more compact formula can arise. The equivalent closed-form formula eliminates the need for recursion and thus makes the computation of waiting times much more efficient. Just like in the recursive part there are two formulas; one for *ambulance* and one for *other patients*. The formulas are given by:

$$W^{(o)} = \frac{\sum_{\substack{(u,v)\in S_A^{(o)} \\ v\geq C}} \frac{1}{C\mu} \times (v - C + 1) \times \pi(u,v)}{\sum_{(u,v)\in S_A^{(o)}} \pi(u,v)} \tag{30}$$

$$W^{(a)} = \frac{\sum_{\substack{(u,v)\in S_A^{(a)} \\ min(v,T)\geq C}} \frac{1}{C\mu} \times (\min(v+1,T) - C) \times \pi(u,v)}{\sum_{(u,v)\in S_A^{(a)}} \pi(u,v)} \tag{31}$$

Note here that the summation, in both equations 30 and 31, goes through all states in the set of accepting states of either *ambulance or other patients* respectively, where a wait incurs. In equation 30 that includes all states (u,v) in the set of accepting states of other patients such that $v \geq C$; i.e. whenever an arrival occurs and the system is at a state where the number of individuals in the system is more than or equal to C. Consequently, for the states that are included in the summation the expression $v - C + 1$ indicates the amount of people in service one would have to wait for upon arrival at the hospital.

Additionally, the minimisation function in equation 31 (*ambulance patients*) ensures that when an ambulance arrives at any state that is greater than the predetermined threshold, the wait that the individual will have to endure remains the same. In essence, the expression $\min(v+1,T) - C$ returns the number of people in line in front of a particular individual upon arrival.

6.5.4 Overall Waiting Time

Consequently, the overall waiting time should can be estimated by a linear combination of the waiting times of *other and ambulance patients*. The overall waiting time can be then given by the following equation where c_o and c_a are the coefficients of each patient's type waiting time:

$$W = c_o W^{(o)} + c_a W^{(a)} \tag{32}$$

The two coefficients represent the proportion of patients of each patient type that traversed through the model. Theoretically, getting these percentages should be as simple as looking at the arrival rates of each patient type but in practise if the hospital or the parking space is full, some patients may be lost to the system. Thus, one should account for the probability that a patient is lost to the system. This probability can be easily calculated by using the two sets of accepting states $S_A^{(a)}$ and $S_A^{(o)}$ defined earlier in equations 22 and 26. Let us define here the probability, for either patient type, that an individual is not lost in the system by:

$$P(L'_o) = \sum_{(u,v)\,\in S_A^{(o)}} \pi(u,v) \qquad\qquad P(L'_a) = \sum_{(u,v)\,\in S_A^{(a)}} \pi(u,v)$$

Having defined these probabilities one may combine them with the arrival rates of each patient type in such a way to get the expected proportions of ambulance and other patients in the model. Thus, by using these values as the coefficient of equation 32 the resultant equation can be used to get the overall waiting time. Note here that the equation below gets the overall waiting time for both the recursive and the closed-form formula.

$$W = \frac{\lambda_o P(L'_o)}{\lambda_a P(L'_a) + \lambda_o P(L'_o)} W^{(o)} + \frac{\lambda_a P(L'_a)}{\lambda_a P(L'_a) + \lambda_o P(L'_o)} W^{(a)} \tag{33}$$

7 Markov chain VS Simulation

7.1 Example model

Consider the Markov chain paradigm in figure 9. 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.

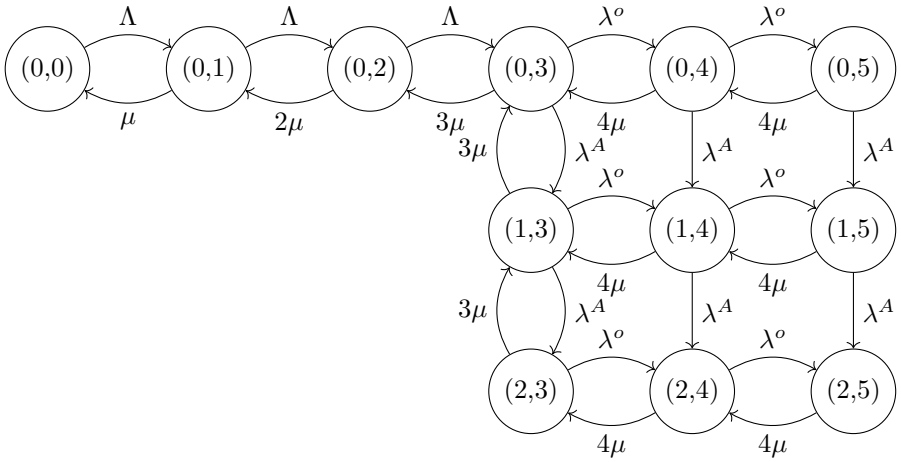


Figure 9: Markov chains: number of servers=4

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 10 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 9 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.

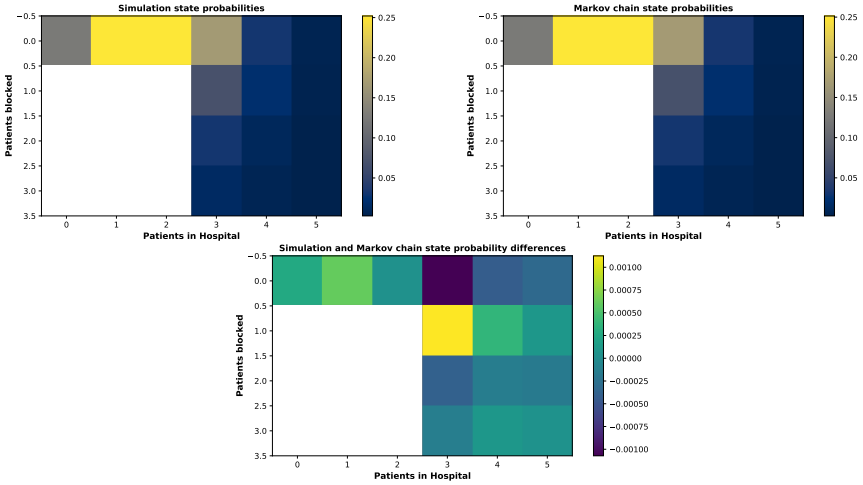
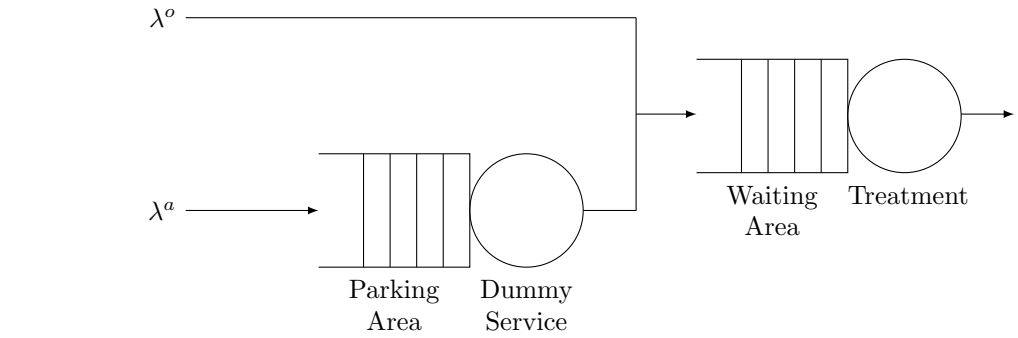
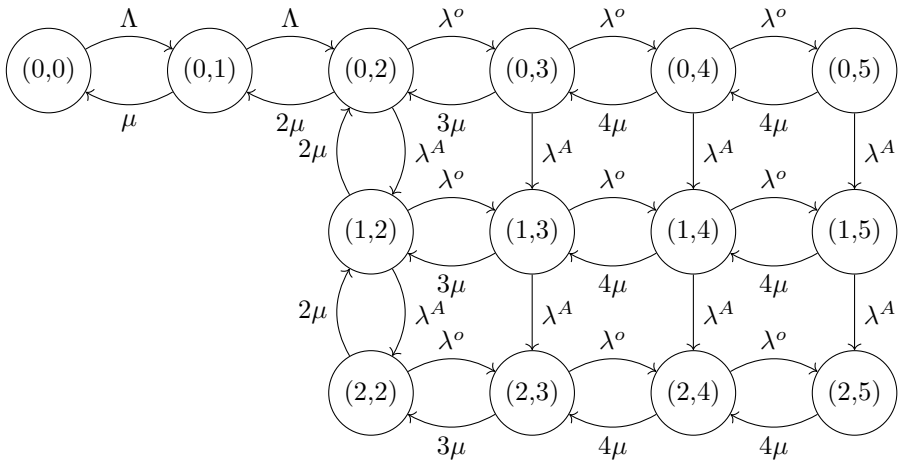


Figure 10: Heatmaps of Simulation, Markov chains and differences of the two

8 Figures that might be useful





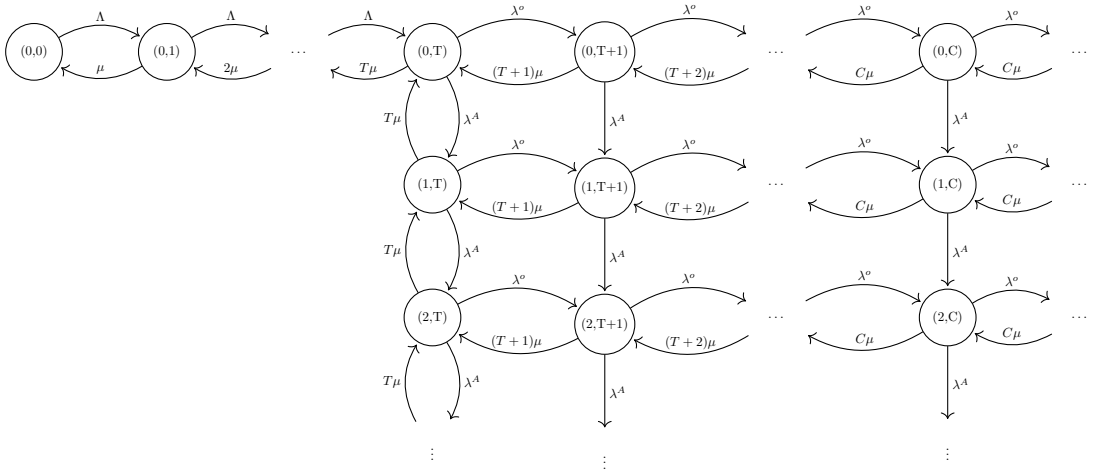


Figure 11: Markov chains

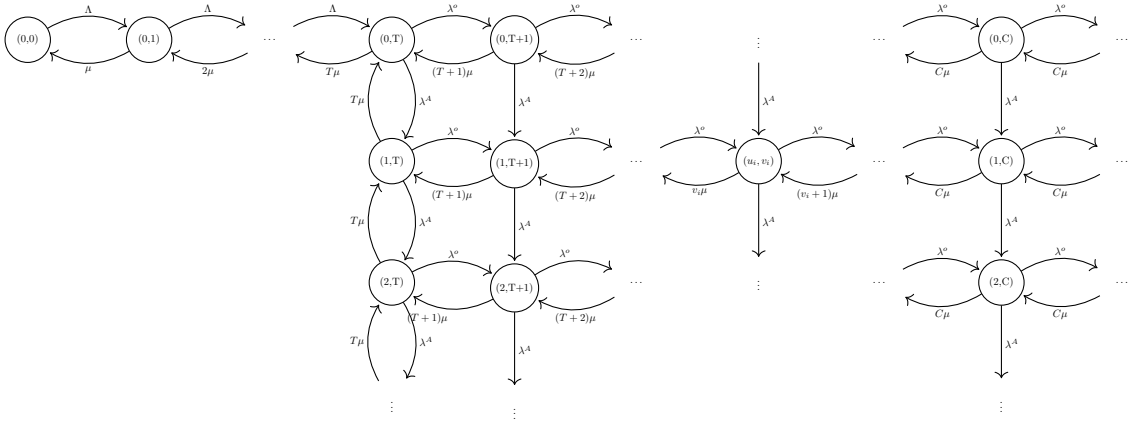


Figure 12: Markov chains

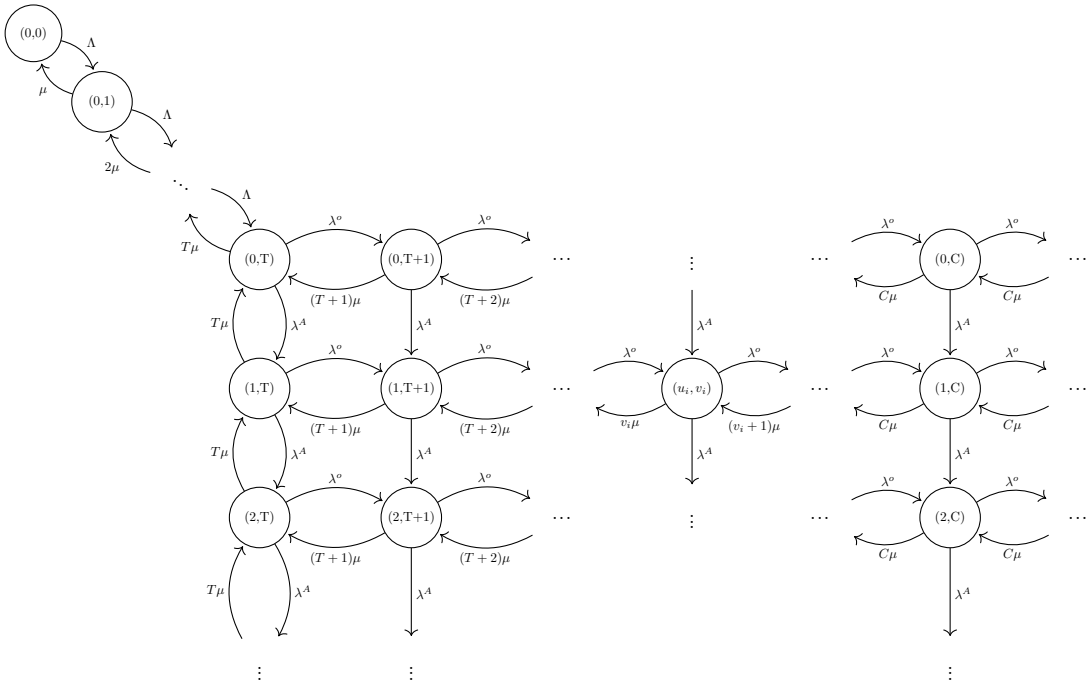


Figure 13: Markov chains

9 Formulas

$$\hat{c_i} \in \{1, 2, \ldots, C_i\}$$

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

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

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

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

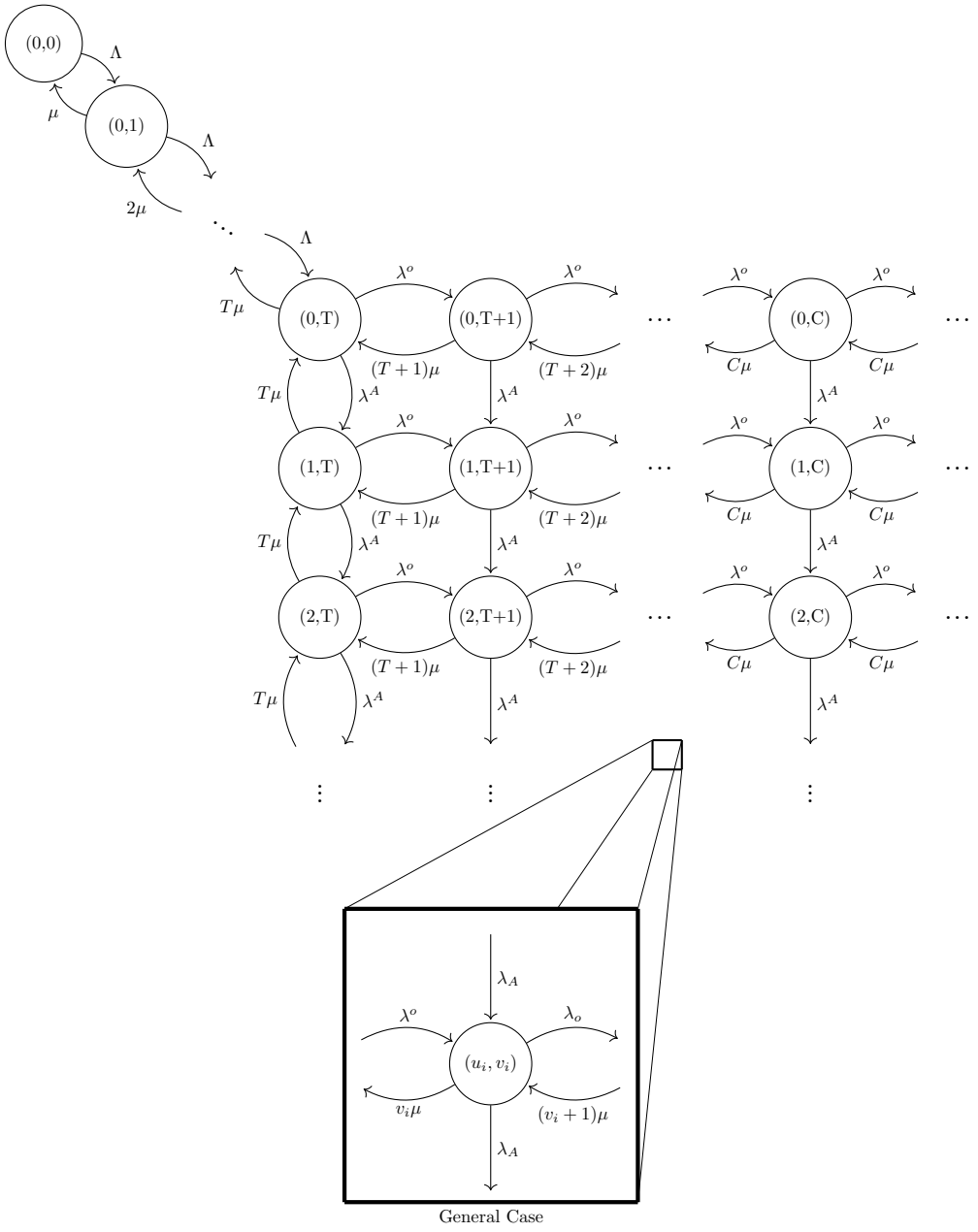


Figure 14: Markov chain