Branching Processes in Queuing Theory

Yoram Clapper

July 11, 2017

Bachelor’s thesis mathematics
Supervisor: prof. dr. Rudesindo Núñez Queija

Korteweg-de Vries Instituut voor Wiskunde
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam
Abstract

In this thesis, well-known results of queuing theory will be demonstrated based on the Galton-Watson branching process. These results concern a queuing system with a Poisson input, a general service time distribution, and one server, also known as the M/G/1 queue. A detailed branching structure is provided that describes how the busy period of the M/G/1 queue (with an arbitrary order of service) and a Galton-Watson process are related. The idea of using branching processes in queuing theory is not new, but the construction of the branching structure used in this thesis is. This structure is used to derive an implicit formula for the Laplace-Stieltjes Transform of the length of the busy period for the M/G/1 queue. The M/G/1 queue following the last-come first-served pre-emptive resume (LCFS-PR) discipline will be considered in more depth. The structure of the Galton-Watson process is used to obtain the result that the marginal distribution of the queue length for a non-empty M/G/1 LCFS-PR queue in equilibrium is a geometric distribution. This result will be extended to the joint distribution of the queue length and the residual service times of the customers for the discrete-time version of the M/G/1 LCFS-PR queue.
# Contents

1 Introduction 4

2 Preliminaries 6
   2.1 Markov chains and Poisson processes 6
   2.2 Queuing systems: the M/G/1 model 8
   2.3 Branching processes: the Galton-Watson model 9

3 The busy period of the M/G/1 queue 13
   3.1 The busy period regarded as a Galton-Watson process 13
   3.2 Moments of the length of a busy period 16

4 The M/G/1 LCFS-PR queue 18
   4.1 The Galton-Watson process for the M/G/1 LCFS-PR queue 18
   4.2 The equilibrium distribution of the queue length 20

5 The joint equilibrium distribution of the M-G-1 LCFS-PR queue 22
   5.1 Discrete-time queuing systems 22
   5.2 Embedded Markov chain 23
   5.3 Invariant distribution 24

6 Conclusion 27

7 Layman’s summary 28

Bibliography 30
1 Introduction

The Markov Chains course was the main motivation for the choice of the topic of this thesis. A Markov chain is a stochastic process that is used for modelling an object moving from one state to another, where the probability that the object reaches a certain state is completely determined by the current state of the object, and is independent of the path the object has taken to get to its current state. A simple example of a Markov chain is the popular board game Monopoly: the next square the token lands on is completely determined by the current square it is on and by the throw of the dice. More sophisticated applications of Markov chains include queuing theory and branching processes. Although the idea of an object manoeuvring from state to state guided by the laws of probability is beautiful in itself, envisioning this in the context of queuing theory and branching processes makes it considerably more lively.

The process of branching should be understood as a probabilistic model representing the growth and decay of a population of objects; e.g., the development of an infectious disease in a population or the propagation of particles in a chain reactor. The reason that such processes are called branching processes is because one can often envision these processes as a tree graph. In queuing theory, we are mainly interested in understanding the probabilistic behaviour of the arrival and departure of objects into and from a system. These objects arrive at random moments in time, and all require the use of a particular type of equipment for a random length of time. Queues form whenever the resource of such equipment is scarce. An obvious example of a queuing system is that of a store where all customers require the service of the personnel. Queuing systems also arise in computer science and traffic engineering.

It was only at the end of the course that these applications of the theory of Markov chains were introduced and as a result it felt as though these subjects were only briefly touched upon. At first glance it didn’t seem that queuing theory and branching processes had a lot in common, except for their Markovian nature. It therefore came as a pleasant surprise to discover that some queues bear the structure of a branching process.

The aim of this thesis is to explore this structure with respect to queues. It is not possible to cover the entire theory of branching processes in queuing theory in a single text. To narrow this down, the topic will be explored using a particular type of queue. Properties of this queue will be studied using the structure of a branching process. In so doing some well-known results in queuing theory will be derived. These results concern the expected queue length and the stationary distribution of the queue length process. Here “stationary” refers to a statistical equilibrium of a process, in the sense that when starting with this stationary distribution, this distribution is preserved throughout time.

These results are obtained within four core chapters. In Chapter 2, the preliminary knowledge needed for this thesis is provided. In Chapter 3, the relation between a
branching process and a particular type of queuing system is made explicit, and is used to find the expected queue length. This relationship is further utilised in Chapter 4 by deriving the marginal stationary distribution of the queue length of a typical queuing system. Chapter 4 serves as a motivation to extend this result to the joint stationary distribution of the queue length and to the residual service times of the customers. In Chapter 5, the extension of this result is addressed for the discrete-time version of the relevant queue.
2 Preliminaries

In this chapter we will introduce the concepts and results that are needed to understand the following chapters. To start off we will give the formal definition of a Markov chain and state some results concerning Markov chains. The Poisson process and the Markov property of the Poisson process will then be introduced. Since it is assumed that the reader is already familiar with both concepts, we shall not dwell too long upon these subjects. Next, the concept of a queuing system will be explored, and the so-called M/G/1 queue will be introduced. Finally, we will elaborate on a branching process called the Galton-Watson process in some depth.

A great deal of the information gathered in Section 2.1, if not all, comes from [4] where these subjects are treated more extensively. The concepts presented in Section 2.2 are based on those introduced in [2] and [5]. The notation established in Section 2.3 is mainly derived from [3].

2.1 Markov chains and Poisson processes

A Markov chain is a discrete-time stochastic process \((X_n)_{n=0}^{\infty}\) for which only the present value of the process is relevant in predicting the future. The past values of the process provide no further information and are therefore irrelevant. In this sense a Markov chain has no memory of the past. A Markov chain takes values in a so-called state-space \(I\) which is assumed to be countable; an element \(i \in I\) is called a state.

The definition of a Markov chain requires an initial distribution \(\zeta = (\zeta_i)_{i \in I}\) and a matrix \(P = (p_{ij})_{i,j \in I}\) such that \(\sum_{j \in I} p_{ij} = 1\) for all \(i \in I\), called the transition matrix.

**Definition 2.1** (Markov chain). A discrete-time stochastic process \((X_n)_{n=0}^{\infty}\) taking values in a state-space \(I\), with initial distribution \(\zeta\) and transition matrix \(P\), such that for all \(i, i_0, i_1, \ldots, i_{n+1} \in I\) the following holds

(i) \(P(X_0 = i) = \zeta_i\),

(ii) \(P(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = p_{i_n i_{n+1}}\),

is called a *Markov chain*.

**Theorem 2.1** (Markov property, Theorem 1.1.2 in [4]). Let \((X_n)_{n=0}^{\infty}\) be a Markov chain with transition matrix \(P\) and initial distribution \(\zeta\). Then conditional on \(X_m = i\), the process \((X_{m+n})_{n=0}^{\infty}\) is a Markov chain with transition matrix \(P\) and initial distribution \((\delta_{ij})_{j \in I}\) (here \(\delta_{ij}\) is the Kronecker delta).

Next we introduce the notion of an invariant distribution and state two results which help to explain the role of invariant distributions in the theory of Markov chains.
Definition 2.2 (Invariant distribution). A distribution $\pi = (\pi_i)_{i \in I}$ is said to be invariant to a matrix $P = (p_{ij})_{i,j \in I}$ if $\pi P = \pi$.

Theorem 2.2 (Theorem 1.7.1 in [4]). If $(X_n)_{n \geq 0}$ is a Markov chain with an initial distribution $\pi$ such that $\pi$ is invariant to the transition matrix $P$. Then $(X_{m+n})_{n \geq 0}$ is a Markov chain with initial distribution $\pi$ and transition matrix $P$.

The Markov chains that satisfy the conditions stated in Theorem 2.2 are prominent throughout this thesis.

Definition 2.3. A Markov chain $(X_n)_{n \geq 0}$ is said to be in equilibrium if the initial distribution is invariant to the transition matrix.

Theorem 2.3 (Convergence to equilibrium, Theorem 1.8.3 in [4]). If $(X_n)_{n \geq 0}$ is an irreducible and aperiodic Markov chain and $\pi$ is an invariant distribution to the transition matrix $P$. Then

$$P(X_n = j) \to \pi_j, \quad \text{as } n \to \infty \text{ for all } j.$$  

In this context the term “irreducible” means that each state can be reached from any other state with positive probability, and the term “aperiodic” means that, given any state, there is a positive probability to return from the state back to itself for all sufficiently large number of steps (see [4]).

It is also possible to define continuous-time Markov chains, better referred to as Markov processes. Although we won’t give the general definition of a Markov process in continuous time, we will focus on a special type of Markov process known as the Poisson process. The Poisson process is often used to model a process that involves counting discrete events occurring throughout continuous time. In order to introduce the Poisson process we need the notion of stationary and independent increments. Let $(X_t)_{t \geq 0}$ be a continuous-time process; where we consider its increments $X_t - X_s$ over any interval $(s, t]$. We say that $(X_t)_{t \geq 0}$ has stationary increments if the distribution of $X_{s+t} - X_s$ only depends on $t \geq 0$, and we say that $(X_t)_{t \geq 0}$ has independent increments if its increments over any finite collection of disjoint intervals are independent.

Definition 2.4 (Poisson process). Let $(X_t)_{t \geq 0}$ be an increasing, right-continuous, integer-valued process starting at 0 and let $\lambda \in \mathbb{R}_{>0}$. If $(X_t)_{t \geq 0}$ satisfies one of the following (equivalent) conditions then $(X_t)_{t \geq 0}$ is called a Poisson process of rate $\lambda$,

(i) $(X_t)_{t \geq 0}$ has stationary independent increments and $X_t$ is distributed according to a Poisson distribution with parameter $\lambda t$ for every $t \geq 0$;

(ii) $(X_t)_{t \geq 0}$ has independent increments and as $h \downarrow 0$, uniformly in $t$,

$$\mathbb{P}(X_{t+h} - X_t = 0) = 1 - \lambda h + o(h), \quad \mathbb{P}(X_{t+h} - X_t = 1) = \lambda h + o(h).$$  

Conditions (i) and (ii) are called the transition probability definition and the infinitesimal definition respectively.

Theorem 2.4 (Markov property of the Poisson process, Theorem 2.4.1 in [4]). Let $(X_t)_{t \geq 0}$ be a Poisson process of rate $\lambda$. Then for any $s \geq 0$ the process $(X_{s+t} - X_s)_{t \geq 0}$ is a Poisson process of rate $\lambda$, independent of $(X_r : r \leq s)$.
2.2 Queuing systems: the M/G/1 model

A queuing system is a model for a particular class of processes of adding and removing objects from some space in a prescribed way. The objects are often referred to as customers and the space is often referred to as the system. It is said that a customer arrives in the system when it is added to the system, and that the customer leaves the system when it is removed from the system. Throughout this thesis we will consider a specific type of queuing system. It is therefore convenient to first introduce the underlying model for this queuing system, and discuss all quantities and concepts we are interested in based on this model.

Example 2.1. Consider a queuing system consisting of a waiting room and one server (see Figure 2.1). Customers arrive into the system and require a certain amount of work to be done by the server. If an arriving customer finds the system to be empty, the customer occupies the server immediately, otherwise the customer waits in the system until service of the customer begins. Depending on the order of service, the server chooses which customer to serve next. As soon as the work is done, the customer leaves the system. The length of time that a customer occupies the server is called the service time of the customer. The prescription of the arrivals of the customers is called the arrival process. In our model it is assumed that the inter-arrival times between the customers are independent and equally distributed. Furthermore, the service time of a customer is assumed to be independent and equally distributed for each customer. This model is called the G(eneral) I(ndependent)/GI/1 queue.

The sojourn time of a customer is the length of time between the moment that the customer arrives into the system and the moment the customer leaves the system. Note that the service time and the time spend waiting in the system both add up to the sojourn time. The number of customers present in the system (these include the customers in the waiting room as well as the customer occupying the server) is called the queue length. The period between the instant a customer enters a previously empty system, until the next instant that the system is completely empty, is called the busy period. The GI/GI/1 queue is said to follow the first-come first-served (FCFS) discipline if the order in which the customers are served is the same as the order in which the customers arrive. The GI/GI/1 queue is said to follow the last-come first-served pre-emptive resume (LCFS-PR) discipline if the order in which the customers are served is as follows: a newly arriving customer is served immediately; if upon arrival of the new customer there is a customer already being served, then this service is interrupted, to be resumed the moment the new customer leaves the system. The new customer, in turn, can also be interrupted by a newly arriving customer, and so on. Note that when measuring the service time of a customer, as soon as the service of the customer is interrupted, the measuring must be put on hold, only to be resumed when the service of the customer is resumed.

In the following chapters we will study a special case of the GI/GI/1 queue called the M(arkov)/G/1 queue. For the M/G/1 queue, the arrival process is further specified by
assuming that the counting process of the customers arriving in the system throughout time is a Poisson process of rate $\lambda$.

A queuing system can sometimes be related to a Markov chain corresponding to a quantity of the model; e.g., the queue length of the system. The queuing system is said to be in equilibrium if the corresponding Markov chain is in equilibrium. A well-known result in queuing theory is Little’s law. Although Little’s law holds for more general queuing models, it will be stated in terms of the M/G/1 queue.

**Theorem 2.5** (Little’s law, [2]). Consider the M/G/1 queue in equilibrium. Denote $L$ for the queue length and denote $T$ for the sojourn time of a customer. Then

$$E[L] = \lambda E[T].$$

### 2.3 Branching processes: the Galton-Watson model

A branching process is a probabilistic model often used to represent the growth and decay of a population of objects. A well-known branching process is the so-called Galton-Watson process which forms the core of this thesis. The corresponding Galton-Watson model is the most basic model used for branching processes and was introduced by Galton and Watson in their study of the extinction of family names. Therefore it is hardly surprising that this model describes the population in terms of generations. In the Galton-Watson model every individual in the population generates a random number of offspring. In this context, and throughout this thesis, the term “offspring” is used to refer to the immediate descendants of the individual. It is assumed that the number of offspring is independent and equally distributed according to some distribution $F$ for each individual. The process begins with a certain starting population which we will call the 0th-generation. The aggregate offspring of all individuals from the 0th-generation is called the 1st-generation, the offspring of the individuals from the 1st-generation form the 2nd-generation etcetera. The number of individuals in the $n$th-generation will be denoted with $Z_n$. Formally a Galton-Watson process can be defined by means of a Markov chain.

**Definition 2.5** (Galton-Watson process). Let $F$ be a probability distribution on $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. A **Galton-Watson process** is a Markov chain $(Z_n)_{n \geq 0}$ with $Z_0 = 1$, state-space $I = \mathbb{N}_0$ and the transition matrix $P$ defined such that, for every $n \in \mathbb{N}_0$, conditionally...
on \( Z_n = k \) the following relation holds

\[
Z_{n+1} \overset{d}{=} \sum_{i=1}^{k} N_i,
\]

with \( N_i \), for every \( 1 \leq i \leq k \), independent and distributed according to \( F \) (see Remark 2.2 for the case \( k = 0 \)).

**Remark 2.1.** The symbol " \( \overset{d}{=} \) " means that the left-hand side and the right-hand side of the equation have the same probability distribution.

**Remark 2.2.** We will use the convention that the empty sum equals zero; so if \( Z_n = 0 \) for some \( n \in \mathbb{N} \) then \( Z_m = 0 \) for all \( m > n \); i.e., if a generation contains no individuals then the subsequent generations contain no individuals either.

There is a useful way to identify the individuals of a population uniquely with the elements of the set \( \Gamma = \bigcup_{k=1}^{\infty} \mathbb{N}^k \). Denote the single individual from the 0th-generation with (1). If we refer to a single individual of the offspring of a member of the population as a child of that member, then the \( j \)-th-child of individual (1) is denoted with (1, \( j \)), and the \( k \)-th-child of individual (1, \( j \)) will be denoted with (1, \( j, k \)), etcetera. This way every individual in the population corresponds uniquely with an element of \( \Gamma \). This identification allows us to describe the generations of a population in terms of subsets of \( \Gamma \). We introduce a collection of independent and equally distributed random variables \( N_\gamma \) for each \( \gamma \in \Gamma \) with distribution \( F \). The random variable \( N_\gamma \) will denote the number of offspring of an individual \( \gamma \in \Gamma \) (note that \( N_1 \) from Definition 2.5 is equal in distribution to \( N_{(1)} \) although isn’t the same object). The consecutive generations can be described recursively as follows. Let \( i_0 = 1 \) and define the sets

\[
I_0 = \{(1)\},
I_n = \{(i_{i_0}, i_{i_1}, \ldots, i_{i_{n-1}}) : (i_{i_0}, i_{i_1}, \ldots, i_{i_{n-1}}) \in I_{n-1} \text{ and } 1 \leq i_{i_n} \leq N_{(i_{i_0}, i_{i_1}, \ldots, i_{i_{n-1}})}\},
\]

for \( n \in \mathbb{N} \). If there exists some \( n \in \mathbb{N} \) such that \( N_\gamma = 0 \) for all \( \gamma \in I_{n-1} \), then \( I_n = \emptyset \) and we take \( I_m = \emptyset \) for all \( m > n \). The set \( I_n \), for \( n \in \mathbb{N}_0 \), is called the \( n \)-th-generation.

**Remark 2.3.** Note that the set \( I_n \) depends on the random variables \( (N_\gamma)_{\gamma \in I_{n-1}} \) and is therefore a stochastic set. The construction of \( I_n \) is only possible if \( N_\gamma \) is known for all \( \gamma \in I_{n-1} \).

From the definition of the generation sets it follows that whenever we find a generation to be empty, all subsequent generations will be empty as well. Since \( Z_n = 0 \) if and only if \( I_n = \emptyset \), this agrees with Remark 2.2. In this light, we introduce the following definition.

**Definition 2.6.** Let \((Z_n)_{n\geq0}\) be a Galton-Watson process. The Galton-Watson process is said to have ended if there exists an \( n \in \mathbb{N} \) such that \( Z_n = 0 \).

There is an insightful, schematic way of visualising a Galton-Watson process by means of a tree graph. This visualisation is illustrated in Figure 2.2 together with Example 2.2. This representation is useful to keep in mind when working with Galton-Watson processes.
Example 2.2. In Figure 2.2 a realisation of a Galton-Watson process is illustrated: the starting population consists of individual (1). The offspring of individual (1) are the individuals (1, 1) and (1, 2). The individuals (1, 1, 1), (1, 1, 2) and (1, 2, 1), (1, 2, 2), (1, 2, 3) are the offspring of the individuals (1, 1) and (1, 2) respectively. Individual (1, 2, 1, 1) is the offspring of individual (1, 2, 1). The individuals (1, 1, 1), (1, 1, 2), (1, 2, 2), (1, 2, 3) and (1, 2, 1, 1) have no offspring. Hence, the Galton-Watson process has ended. The generations thus become

\[ I_0 = \{(1)\}, \]
\[ I_1 = \{(1, 1), (1, 2)\}, \]
\[ I_2 = \{(1, 1, 1), (1, 1, 2), (1, 2, 1), (1, 2, 2), (1, 2, 3)\}, \]
\[ I_3 = \{(1, 2, 1, 1)\}. \]

Figure 2.2: A realisation of a Galton-Watson process.

A useful property of the Galton-Watson process is that every individual in the population generates its own respective Galton-Watson process. For example, in Figure 2.2 we could focus on the process only involving individual (1, 2) and its descendants. To make this precise, consider an individual \( \gamma = (1, i_1, \ldots, i_m) \in I_m \) to be present in the population and let \( i_0 = 1 \) once again. Define the sets

\[ I_{\gamma,0} = \{\gamma\}, \]
\[ I_{\gamma,n} = \{(i_0, i_1, \ldots, i_m + n) : (i_0, i_1, \ldots, i_{m+n-1}) \in I_{\gamma,n-1} \text{ and } 1 \leq i_{m+n} \leq N_{(i_0, i_1, \ldots, i_{m+n-1})}\}, \]

for \( n \in \mathbb{N} \), the same way as we defined the generation sets. Denote the number of elements in \( I_{\gamma,n} \) with \( Z_{\gamma,n} \) for \( n \in \mathbb{N}_0 \). Then

\[ Z_{\gamma,n+1} = \sum_{\delta \in I_{\gamma,n}} N_\delta \tag{2.1} \]

11
and \((Z_{\gamma,n})_{n \geq 0}\) is indeed a Galton-Watson process. The process \((Z_{\gamma,n})_{n \geq 0}\) is called the \textit{Galton-Watson process generated by} \(\gamma\) and the set \(I_{\gamma,n}\) is called the \textit{nth-generation generated by} \(\gamma\). Note that the original Galton-Watson process is the same as the Galton-Watson process generated by \((1); i.e., I_n = I_{(1),n}\). An important feature of the Galton-Watson process is that the Galton-Watson process generated by an individual \(\gamma\) is equally distributed for every individual \(\gamma \in \Gamma\). Moreover, if an individual \(\gamma\) doesn’t belong to any of the generations generated by some individual \(\delta\) and vice versa, then the respective Galton-Watson processes generated by \(\gamma\) and \(\delta\) are independent of one another. All of this follows from Equation (2.1) together with the assumption that all \((N_{\gamma})_{\gamma \in \Gamma}\) are independent and equally distributed.
3 The busy period of the M/G/1 queue

The reason that the Galton-Watson process was introduced in the last chapter is because, rather surprisingly, a busy period of the M/G/1 queue can effectively be viewed as a Galton-Watson process. In this chapter we will be concerned with showing how the connection between a busy period of the M/G/1 queue and a Galton-Watson process is made. The construction that connects the busy period and the Galton-Watson process can be used for any order of service. Although the construction is fairly simple, matters can get confusing without a concrete example. Therefore we will first state the rules of the construction and then explain the construction by using the M/G/1 FCFS queue as an example. After that it won’t be hard to understand that the construction can be used for the M/G/1 queue for any order of service. We will use the construction to find the moments of the length of the busy period for the M/G/1 queue given any order of service.

3.1 The busy period regarded as a Galton-Watson process

Consider the M/G/1 queue and suppose that the system is empty. The first customer to arrive into the system (and therefore to start a busy period) can, in terms of the Galton-Watson process, be viewed as the single individual in the 0th-generation of a Galton-Watson process. The customers that arrive while the first customer occupies the server can be viewed as the offspring of the first customer. More generally, still during the busy period initiated by the first customer, customers that arrive into the system while the server is occupied can be viewed as the offspring of the customer occupying the server at that moment. This way, with the notation introduced in Section 2.3, each customer that contributes to the busy period can be identified with an element of $I_n$, the number of customers in the $n$th-generation will be denoted by $Z_n$.

Example 3.1. Consider the M/G/1 FCFS queue and suppose that a customer arrives into a previously empty system. As said, this customer can be viewed as the single individual in the 0th-generation of a Galton-Watson process, hence we will label this customer as customer (1). Since customer (1) arrived into an empty system it is served immediately. Now suppose that while customer (1) occupies the server a new customer arrives. This customer can, in terms of a Galton-Watson process, be viewed as a child of customer (1) and thus be labeled customer (1,1). Since customer (1,1) is a child of customer (1), customer (1,1) is considered a member of generation $I_1$. The next customer that arrives while customer (1) occupies the server is again viewed as a child of customer (1), is this time labeled as customer (1,2), and is also considered a member of generation...
In general, the \( i \)-th customer that arrives while customer (1) occupies the server, is viewed as a child of customer (1), is labeled as customer (1, \( i \)), and is considered a member of generation \( I_1 \). Therefore the offspring of customer (1) are all customers that arrive while customer (1) occupies the server. This process continues until the service of customer (1) is completed and after that generation \( I_1 \) remains unchanged. As soon as the service of customer (1) is completed, customer (1) leaves the system and customer (1,1) occupies the server. Now, in the same way as for customer (1), every customer that arrives in the system while customer (1,1) occupies the server is viewed as a child of customer (1,1). So the \( i \)-th customer that arrives while customer (1,1) occupies the server is labeled as customer (1,1, \( i \)) and is considered a member of generation \( I_2 \). When the service of customer (1,1) is completed and customer (1,1) leaves the system, the service of customer (1,2) starts and all customers that arrive while customer (1,2) occupies the server are labeled accordingly and are considered members of generation \( I_2 \) as well. The construction of generation \( I_2 \) is completed if the last child of customer (1) leaves the system, or in other words, if all members of generation \( I_1 \) have left the system. In general, if a customer arrives while a certain customer \( (1, i_1, i_2, \ldots, i_n) \) occupies the server, the arriving customer is labeled as customer \((1, i_1, i_2, \ldots, i_n, i_{n+1})\) for some \( i_{n+1} \in \mathbb{N} \) and is considered a member of generation \( I_{n+1} \). The offspring of customer \((1, i_1, i_2, \ldots, i_n)\) therefore consist of all customers that arrive while customer \((1, i_2, i_2, \ldots, i_n)\) occupies the server. The construction of generation \( I_{n+1} \) is completed if all members of generation \( I_n \) left the system.

**Remark 3.1.** As we have already mentioned in the introduction of this chapter, the construction of the Galton-Watson process remains valid for any order of service. If we picked an arbitrary customer \( \gamma \) present in the system and let it occupy the server, the offspring of customer \( \gamma \) would still be the customers that arrive during the period that customer \( \gamma \) occupies the server. This means that, just as in Example 3.1, if \( \gamma = (1, i_1, i_2, \ldots, i_n) \in I_n \), the \( i \)-th customer that arrives during the service of customer \( \gamma \) is labeled as customer \((1, i_1, i_2, \ldots, i_n, i)\) and is considered a member of generation \( I_{n+1} \).

**Remark 3.2.** It needs to be emphasised that a busy period can only correspond to one Galton-Watson process, namely, the Galton-Watson process generated by the customer initiating the busy period. Whenever a busy period is initiated by a customer, a new Galton-Watson process starts which is unrelated to the Galton-Watson processes corresponding to previous busy periods.

Throughout the rest of this chapter we will consider the M/G/1 queue with an arbitrary order of service.

**Example 3.2.** To familiarise ourselves with the the Galton-Watson process in the context of the M/G/1 queue with an arbitrary order of service, a realisation is illustrated using Figure 2.2: customer (1) arrived into an empty system initiating a busy period, thus customer (1) is the starting population. Customers (1,1) and (1,2) arrived while customer (1) occupied the server and are therefore the offspring of customer (1). The customers (1,1,1), (1,1,2) and (1,2,1), (1,2,2) arrived while customers (1,1) and (1,2) occupied the server respectively, and are therefore the offspring of customers.
(1, 1) and (2, 2) respectively. Customer (1,2,1,1) arrived while customer (1,2,1) occupied the server and is therefore the offspring of customer (1,2,1). No customers arrived while customers (1, 1, 1), (1, 1, 2), (1, 2, 2), (1, 2, 3) and (1, 2, 1, 1) occupied the server and they therefore have no offspring.

In order to verify that the process \((Z_n)_{n \geq 0}\) satisfies the conditions of a Galton-Watson process we introduce random variable \(N_\gamma\) as the number of customers that arrive while the server is occupied by a customer \(\gamma \in \Gamma\). The assumption that the arrival process for the M/G/1 queue is a Poisson process of rate \(\lambda\) makes it possible to determine the distribution of \(N_\gamma\). First, it follows from the Markov property of the Poisson process that \(N_\gamma\) is independent and equally distributed for every customer \(\gamma \in \Gamma\). Second, if we write \(N(t)\) for the number of customers arriving in the system over a period of time \(t \geq 0\) and denote the service time of customer \(\gamma\) with \(B_\gamma\), then from the Markov property of the Poisson process, it follows that \(N_\gamma \overset{d}{=} N(B_\gamma)\). Hence, given \(B_\gamma = t\), the random variable \(N_\gamma\) is distributed according to a Poisson distribution of parameter \(\lambda t\). Since the offspring of a customer \(\gamma\) is defined as the customers that arrive during the server occupancy by customer \(\gamma\), the random variable \(N_\gamma\) represents the number of offspring as well. Therefore from the definition of the generation sets \((I_n)_{n \in \mathbb{N}_0}\) the following relation holds for the size of the \((n+1)\)th-generation:

\[
Z_{n+1} = \sum_{\gamma \in I_n} N_\gamma
\]

where \(n \in \mathbb{N}_0\). This shows that \((Z_n)_{n \geq 0}\) indeed follows a Galton-Watson process.

The structure of the Galton-Watson process provides an effective method for analysing the length of the busy period. Consider the system in a busy period and its corresponding Galton-Watson process. Let \(S\) be the length of the busy period and define \(M = \bigcup_{n=0}^\infty I_n \subset \Gamma\) as the population of customers that contribute to the busy period. Since the length of the busy period is the sum of the service times of all customers contributing to the busy period it follows that

\[
S = \sum_{\gamma \in M} B_\gamma. \quad (3.1)
\]

Now suppose that \(N_{(1)} > 0\). For \(1 \leq i \leq N_{(1)}\) define \(M_i = \bigcup_{n=0}^\infty I_{(1),n}\), which is the set containing all customers constituting the Galton-Watson process generated by customer \((1, i)\). Define

\[
S_i = \sum_{\gamma \in M_i} B_\gamma. \quad (3.2)
\]

Since \(M = \{(1)\} \cup \left(\bigcup_{i=1}^{N_{(1)}} M_i\right)\) and \(M_i \cap M_j = \emptyset\) whenever \(i \neq j\) we obtain the crucial relation

\[
S = B_{(1)} + \sum_{i=1}^{N_{(1)}} S_i. \quad (3.3)
\]
Now $M$ is just the set of all customers constituting the Galton-Watson process generated by customer (1); the Galton-Watson process generated by a customer $\gamma$ has the same distribution for every $\gamma \in \Gamma$, therefore it follows that $|M| \overset{d}{=} |M_i|$ and thus from Equations (3.1) and (3.2) we obtain $S \overset{d}{=} S_i$. Moreover $M_i \cap M_j = \emptyset$ whenever $i \neq j$, hence $S_i$ and $S_j$ are independent whenever $i \neq j$.

3.2 Moments of the length of a busy period

Expression (3.3) makes it possible to obtain the moments of the length of the busy period. We already concluded that $N(1) \overset{d}{=} N(B(1))$. Furthermore, since the service time of customer (1) is equal in distribution to the service time $B$ of a customer in general; i.e, $B(1) \overset{d}{=} B$, we have $N(1) \overset{d}{=} N(B)$ and therefore

$$S \overset{d}{=} B + \sum_{i=1}^{N(B)} S_i.$$  \hfill (3.4)

In order to obtain insight into the moments of $S$, we first define the Laplace-Stieltjes Transform of $S$

$$\pi(x) = \mathbb{E}[e^{-xS}], \quad x \in \mathbb{R}_{\geq 0}.$$  

Furthermore define the Laplace-Stieltjes Transform of $B$

$$\beta(x) = \mathbb{E}[e^{-xB}], \quad x \in \mathbb{R}_{\geq 0}.$$  

It follows from expression (3.4) that

$$\mathbb{E}[e^{-xS}|B = t] = e^{-xt}\mathbb{E}[e^{-x \sum_{i=1}^{N(t)} S_i}].$$  \hfill (3.5)

Since the random variable $N(t)$ is distributed according to a Poisson distribution with parameter $\lambda t$ it follows that

$$\mathbb{E}[e^{-x \sum_{i=1}^{N(t)} S_i}] = \sum_{k=0}^{\infty} \mathbb{E}[e^{-x \sum_{i=1}^{k} S_i}] \mathbb{P}(N(t) = k) = \sum_{k=0}^{\infty} \mathbb{E}[e^{-xS}]^k \mathbb{P}(N(t) = k)$$

$$= \sum_{k=0}^{\infty} \pi(x)^k \mathbb{P}(N(t) = k) = \mathbb{E}[\pi(x)^{N(t)}] = e^{-\lambda t(1 - \pi(x))}.$$  

The second equality follows since $S_1, S_2, \ldots, S_k$ are independent of one another and equal in distribution to $S$. The fourth equality follows by evaluating the probability generating function of a Poisson distributed random variable in $\pi(x)$. By combining this result with (3.5) we obtain

$$\pi(x) = \mathbb{E}[e^{-xS}] = \mathbb{E}[\mathbb{E}[e^{-xS}|B]] = \mathbb{E}[e^{-(x+\lambda(1-\pi(x)))B}] = \beta(x + \lambda(1 - \pi(x))).$$  

16
Although the relation that we have found for $\pi(x)$ is implicit, it is still possible to obtain the moments of $S$. As an illustration we will compute the first moment of $S$. We have
\[
\pi'(x) = \beta'(x + \lambda(1 - \pi(x)))(1 - \lambda\pi'(x)),
\]
therefore, using $\pi(0) = 1$,
\[
\pi'(0) = \beta'(0)(1 - \lambda\pi'(0)),
\]
hence
\[
\mathbb{E}[S] = -\pi'(0) = \frac{-\beta'(0)}{1 + \lambda\beta'(0)} = \frac{\mathbb{E}[B]}{1 - \lambda\mathbb{E}[B]}.
\]
4 The M/G/1 LCFS-PR queue

In this chapter we will use the structure of the Galton-Watson process specifically to find the distribution of the queue length of a non-empty M/G/1 LCFS-PR queue in equilibrium. The queue length distribution of the M/G/1 queue in equilibrium is dependent of the order of service. Hence the structure of the Galton-Watson process for the LCFS-PR discipline is utilised in a more direct manner when finding the mentioned distribution. It is therefore important, as well as for the sake of establishing notation, to gain some more insight into how the Galton-Watson process is formed for the LCFS-PR discipline. For this reason we will start off by reconsidering the Galton-Watson process for the M/G/1 LCFS-PR queue. After that we will derive some structural properties of the Galton-Watson process for the M/G/1 LCFS-PR, which we will use to obtain the aforementioned distribution.

The intuition for finding the geometric distribution as described in [5] is formalised in this chapter by means of the Galton-Watson process.

4.1 The Galton-Watson process for the M/G/1 LCFS-PR queue

Consider the M/G/1 LCFS-PR queue in equilibrium and suppose that a new customer \( \gamma \in \Gamma \) arrives into the system. The LCFS-PR discipline prescribes that customer \( \gamma \) is served immediately. In terms of the Galton-Watson process, customer \( \gamma \) can be viewed as the 0th-generation of the Galton-Watson process generated by customer \( \gamma \); i.e., \( \gamma_0 := \gamma \in I_{\gamma,0} \). If a customer arrives while customer \( \gamma_0 \) is occupying the server then the service of customer \( \gamma_0 \) is interrupted and the service of the customer that just arrived starts. Since this customer arrived while customer \( \gamma_0 \) was occupying the server it belongs to generation \( I_{\gamma,1} \) and will be identified as customer \( \gamma_1 \in I_{\gamma,1} \). The service of customer \( \gamma_0 \) resumes once customer \( \gamma_1 \) leaves the system. Note that the service of customer \( \gamma_1 \) can also be interrupted, etcetera. Every next customer that arrives while customer \( \gamma_0 \) occupies the server follows the same process. In general, for a customer \( \gamma_k \in I_{\gamma,k} \), the arrival of a customer while customer \( \gamma_k \) occupies the server induces a customer \( \gamma_{k+1} \in I_{\gamma,k+1} \). The service of customer \( \gamma_k \) is interrupted, the service of customer \( \gamma_{k+1} \) starts, and once customer \( \gamma_{k+1} \) leaves the system the service of customer \( \gamma_k \) resumes.

Remark 4.1. In this construction for the LCFS-PR discipline the Galton-Watson process generated by customer \( \gamma_0 \) only ends once customer \( \gamma_0 \) leaves the system. After all, as long as customer \( \gamma_0 \) is present in the system, arriving customers keep adding to the generations generated by customer \( \gamma_0 \). A consequence of this is that customer \( \gamma_0 \) is present in the system for as long as the Galton-Watson process is active, thus customer...
\( \gamma_0 \) is the last customer that leaves the system of the respective Galton-Watson process. Moreover, if we consider customer \( \gamma_0 \) to be a customer that initiated a busy period it follows that the sojourn time of customer \( \gamma_0 \) is equal in distribution to the length of a busy period.

To get a better idea of what is going on, a realisation of the process just described is illustrated following Example 4.1 and Figure 4.1.

**Example 4.1.** In Figure 4.1 customer \( \gamma_0 \) arrived in the system and thus initiated a Galton-Watson process. The service of customer \( \gamma_0 \) was interrupted by the arrival of customer \( \gamma_1 \) and service of customer \( \gamma_1 \) began immediately. In the tree graph, \( \gamma_1 \) does not have any offspring, signifying that the service of \( \gamma_1 \) was not interrupted. Once customer \( \gamma_1 \) left the system, the service of customer \( \gamma_0 \) was resumed. The service of customer \( \gamma_0 \) was again interrupted, this time by the arrival of customer \( \gamma_1' \) who immediately occupied the server. While customer \( \gamma_1' \) was occupying the server, customer \( \gamma_2 \) first interrupted the service of customer \( \gamma_1' \), and after customer \( \gamma_2 \) left the system, giving way to service of customer \( \gamma_1' \) again, customer \( \gamma_2' \) interrupted the service of customer \( \gamma_1' \) once more. After that, customer \( \gamma_2', \gamma_1' \) and \( \gamma_0 \) subsequently completed their service and left the system, thereby ending the Galton-Watson process generated by customer \( \gamma_0 \).

![Figure 4.1: A realisation of the Galton-Watson process for the M/G/1 LCFS-PR queue.](image)

Denote the number of customers present in the system with \( N \). From here on we will assume that \( N > 0 \) and consequently that the system is in a busy period initiated by a customer \( \gamma \in \Gamma \); i.e., customer \( \gamma \) arrived into an empty system. This means that the Galton-Watson process generated by customer \( \gamma \) is active. Therefore all customers will be considered as individuals corresponding to a generation \( I_{\gamma,n} \) generated by customer \( \gamma \).

**Lemma 4.1.** If a customer \( \gamma_n \in I_{\gamma,n} \) for \( n \in \mathbb{N}_0 \) is present in the system, then it is the only member of generation \( I_{\gamma,n} \) that is present.

**Proof.** It needs to be shown that if a customer \( \gamma_n \in I_{\gamma,n} \) is present in the system, then there exists no customer \( \gamma'_n \in I_{\gamma,n} \setminus \{\gamma_n\} \) that is present in the system at the
same time, for every \( n \in \mathbb{N}_0 \). The case \( n = 0 \) is immediate since there is only one customer in generation \( I_{\gamma,0} \) by definition of \( I_{\gamma,0} \). Now suppose that \( n > 0 \) and that a customer \( \gamma_n \in I_{\gamma,n} \) is present in the system. The presence of customer \( \gamma_n \) means that it has interrupted the service of a customer \( \gamma_{n-1} \in I_{\gamma,n-1} \). The service of customer \( \gamma_{n-1} \) can only continue when customer \( \gamma_n \) leaves the system, and it is only then that a new customer can interrupt the service of customer \( \gamma_{n-1} \) and therefore be added to the generation \( I_{\gamma,n} \). This shows that not more than one member per generation can be present in the system at the same time.

Lemma 4.2. Suppose that \( N \geq m \) for \( m \in \mathbb{N} \) and let \( k = m - 1 \). Then a customer \( \gamma_k \in I_{\gamma,k} \) is present in the system.

Proof. The proof is by induction. The case \( m = 1 \) follows from Remark 4.1. Next, let \( m \geq 2 \) and consider the situation \( N \geq m \). Suppose that, if \( N \geq m - 1 \), a customer \( \gamma_{k-1} \in I_{\gamma,k-1} \) is present in the system. From Lemma 4.1 it follows that \( \# \{ \delta \in I_{\gamma,n} : 0 \leq n \leq k - 1 \text{ and present in the system} \} \leq k \). Since \( N \geq m > k \) there exists an \( n \geq k \) such that a customer \( \gamma_n \in I_{\gamma,n} \) is present in the system. Thus \( \gamma_n \in I_{\gamma,n-k} \) for some \( \gamma_k \in I_{\gamma,k} \). Remark 4.1 now shows that there exists a customer \( \gamma_k \in I_{\gamma,k} \) that is present in the system.

4.2 The equilibrium distribution of the queue length

Combining Lemma 4.1 and Lemma 4.2 shows that, under the assumption that \( N \geq m \), a customer \( \gamma_n \in I_{\gamma,n} \) is present in the system for every \( n \in \{0, 1, 2, \ldots, k = m - 1\} \) and that this is the only customer present per respective generation. Therefore the customers present in the system necessarily form a path from customer \( \gamma_0 \) up to the customer in service in the tree graph corresponding to the Galton-Watson process generated by customer \( \gamma_0 \). Now, given that a customer \( \gamma_n \in I_{\gamma,n} \) is present in the system, define \( C_{\gamma_n} \) as the number of customers on the path from customer \( \gamma_n \) up to the customer in service without counting customer \( \gamma_n \) itself, then

\[
N = C_{\gamma_0} + 1 = C_{\gamma_1} + 2 = \cdots = C_{\gamma_k} + m.
\]

Therefore we obtain

\[
P(C_{\gamma_k} = j | N \geq m) = P(N = m + j | N \geq m).
\]

for \( j \in \mathbb{N}_0 \). Note that \( C_{\gamma_n} \) is just the number of customers present in the system that constitute the Galton-Watson process generated by customer \( \gamma_n \), without counting customer \( \gamma_n \) itself. Since the Galton-Watson process generated by a customer \( \gamma \in \Gamma \) has the same distribution for every customer \( \gamma \in \Gamma \), it follows that \( P(C_{\gamma_k} = j | N \geq m) \) is independent of \( m \) and therefore that \( P(N = m + j | N \geq m) \) is independent of \( m \) as well. This result makes it possible to derive the distribution of \( N \) given that \( N > 0 \). This is done inductively: define \( \rho = P(N = 1 | N > 0) \). We will prove that

\[
P(N = n | N > 0) = (1 - \rho)^{n-1} \rho,
\]

20
for \( n \in \mathbb{N} \). The case \( n = 1 \) follows trivially by definition of \( \rho \). Now let \( n > 1 \) and suppose that

\[
P(N = n - 1|N > 0) = (1 - \rho)^{n-2}\rho.
\]

Then

\[
P(N = n|N > 0) = P(N > 1|N > 0)P(N = n|N > 1) + P(N = 1|N > 0)P(N = n|N = 1)
\]

\[
= P(N > 1|N > 0)P(N = n|N > 1)
\]

\[
= P(N > 1|N > 0)P(N = n|N \geq 2)
\]

\[
= P(N > 1|N > 0)P(N = n - 1|N \geq 1)
\]

\[
= P(N > 1|N > 0)P(N = n - 1|N > 0)
\]

\[
= (1 - \rho)^{n-1}\rho,
\]

proving the result. From this it follows that \( N \) given \( N > 0 \) has a geometric distribution with parameter \( \rho \). To derive an expression for \( \rho \) we note that for the M/G/1 LCFS-PR queue the sojourn time of a customer is equal in distribution to that of the length of a busy period \( S \). Combining this with Little’s law gives us on the one hand

\[
E[N|N > 0] = \lambda E[S],
\]

while on the other hand we have

\[
E[N|N > 0] = \frac{1}{\rho}.
\]

Therefore we obtain

\[
\rho = \frac{1}{\lambda E[S]}.
\]
5 The joint equilibrium distribution of the M-G-1 LCFS-PR queue

In the previous chapter we obtained the equilibrium distribution of the queue length for the M/G/1 LCFS-PR queue given that it is non-empty. This was done under the assumption that such an equilibrium distribution exists. Although this result is interesting in itself, it also serves as a motivation to explore if it is possible to extend this result. In this chapter we reach the most important result of this thesis by combining the results obtained in previous chapters. We will try to find the joint distribution of the queue length and residual service times of a non-empty M/G/1 LCFS-PR queue in equilibrium, in a way that shows that the aforementioned assumption is indeed correct. However, to make this mathematically precise we need the notion of a continuous-time Markov chain on a continuous state-space. The knowledge of such a Markov chain is not assumed in this thesis and to avoid this we will first introduce the idea of a discrete-time queuing system based on the continuous-time queuing system as described in [1]. From there on we will introduce the discrete version of the M/G/1 LCFS-PR queue and show that it “converges” to the continuous-time version of the M/G/1 LCFS-PR queue. The latter is done using a formally imprecise but intuitively clear argument. Finally, we will find the said distribution for the discrete-time version of the M/G/1 LCFS-PR queue.

5.1 Discrete-time queuing systems

Identify the positive time line with \( \mathbb{R}_{>0} \) and partition \( \mathbb{R}_{>0} \) with the collection of intervals \( \{(\frac{j-1}{k}, \frac{j}{k}) : j \in \mathbb{N}\} \), where \( k \in \mathbb{N} \) is fixed. The interval \( (\frac{j-1}{k}, \frac{j}{k}) \) is referred to as the \( j \)th-slot. Denote \( \Delta_k = \frac{1}{k} \) for the length of the slots. In the discrete-time queuing model it is still possible for the customers to arrive any moment in time. The most important difference between the continuous-time queuing model and the discrete-time queuing model is that in the discrete-time queuing model the systems “actions” only take place at the boundary points of the slots, more precisely,

- it is only possible to observe the number of customers present in the system at the boundary points of the slots;
- it is only possible for a customer to start service or leave the system at the boundary points of the slots;
- it is only possible to measure the length of the service time of a customer in numbers of slots.
This last condition might need some explanation. Suppose that the (uninterrupted) service of a certain customer starts at time $t_0$ and is completed at time $t \in \mathbb{R}_{>t_0}$ with $\frac{n-1}{k} < t \leq \frac{n}{k}$ for some $n \in \mathbb{N}$. From the second condition it follows that $t_0 = \frac{m}{k}$ for some $m \in \mathbb{N}$ with $m < n$, and that the customer leaves the system at time $\frac{n}{k}$. Hence the customer occupied the server for $n - m$ slots, which means that the service time of the customer is $\frac{n-m}{k}$.

Just as the continuous-time queuing model can be specified with the GI/GI/1 model, the discrete-time queuing model can be specified with the so-called GI-GI-1 model. This model prescribes the arrival and service process in the discrete-time queuing model. The prescription is as follows,

- the system contains only one server;
- the number of customers arriving in a slot is, for every slot, equally distributed and independent of one another;
- the number of slots that any customer needs for service is, for every customer, equally distributed and independent of one another.

The LCFS-PR discipline has a discrete analogue as well: the last customer that arrived into a slot $j$ seize the server at time $\frac{j}{k}$; the customer occupying the server in slot $j$ (if there is any) is then placed back into the queue (unless the service of this customer is completed in slot $j$, in which case it leaves the system). The order in which customers are served is determined by the order in which customers arrive. A customer that arrived at a time $t$ has priority over a customer that arrived at time $s < t$.

Now it is possible to introduce the discrete-time variant of the M/G/1 queue and thus the discrete-time variant of the M/G/1 LCFS-PR queue. The M/G/1 model is characterised by the assumption that the arrival process is a Poisson process of rate $\lambda$. The infinitesimal definition of the Poisson process provides a way to introduce the discrete-time version of the M/G/1 model. Define $A_j$ as the number of customers that arrive in the $j$th-slot, such that for $k \uparrow \infty$ and $n \in \mathbb{N}_0$ the following holds

$$P(A_j = 1) = \lambda \Delta_k,$$
$$P(A_j = 0) = 1 - \lambda \Delta_k.$$

Since $k \uparrow \infty$ implies that $\Delta_k \downarrow 0$, it is intuitively clear that this corresponds with the infinitesimal definition of the Poisson process. Moreover, as $k \uparrow \infty$ the distance between the boundary points on which the system is observed becomes arbitrarily small. Thus the GI-GI-1 model with its arrival process described by the random variables $A_1, A_2, \ldots$ approximates (intuitively speaking) the M/G/1 model. The discrete-time version of the M/G/1 model is therefore defined as the GI-GI-1 model with its arrival process described by the random variables $A_1, A_2, \ldots$ and which we will call the M-G-1 model.

### 5.2 Embedded Markov chain

Obtaining the desired equilibrium distribution involves finding a Markov chain corresponding to the M-G-1 LCFS-PR queue. This Markov chain must contain information
about the number of customers present in the system as well as the residual service times
of these customers. Consider the M-G-1 LCFS-PR queue and define $N_j$ as the number
of customers that are present in the system observed at the boundary of slot $j$. Now
suppose that on the boundary point of slot $j$ it is observed that $n$ customers are present
in the system, denote $R_{j,1}, R_{j,2}, \ldots, R_{j,n}$ for their respective residual service times. As
a consequence of the discrete-time queuing model, the residual service time $R_{j,i}$ of the
$i$th-customer is measured in units of $1/k$; for example, if the $i$th-customer occupies the
server during the $(j+1)$th-slot and $R_{j,i} = \frac{m}{k}$ for some $m \in \mathbb{N}$ then $R_{j+1,i} = \frac{m-1}{k}$.
Furthermore, note that as soon as $R_{j,i} = 0$ the $i$th-customer leaves the system. Given
that $A_j = 1$, denote the service time of the customer arriving during slot $j$ with $B_j$.
Now define

$$X_j = (N_j, R_{j,1}, R_{j,2}, \ldots, R_{j,N_j}),$$

and let $X_0$ be distributed according to an appropriate initial distribution, then

$$N_{j+1} = \begin{cases} 
N_j + 1 \{A_j = 1\}, & \text{if } N_j > 0 \text{ and } R_{j,N_j} > \frac{1}{k}, \\
N_j - 1 \{A_j = 0\}, & \text{if } N_j > 0 \text{ and } R_{j,N_j} = \frac{1}{k}, \\
1 \{A_j = 1\}, & \text{if } N_j = 0,
\end{cases}$$

and

$$R_{j+1,N_{j+1}} = \begin{cases} 
1 \{A_j = 1\} B_j + 1 \{A_j = 0\} (R_{j,N_j} - \frac{1}{k}), & \text{if } N_j > 0 \text{ and } R_{j,N_j} > \frac{1}{k}, \\
1 \{A_j = 1\} B_j + 1 \{A_j = 0\} R_{j,N_j-1}, & \text{if } N_j > 0 \text{ and } R_{j,N_j} = \frac{1}{k}, \\
1 \{A_j = 1\} B_j, & \text{if } N_j = 0.
\end{cases}$$

Thus, knowing $X_j$ provides sufficient information to determine the distribution of $X_{j+1}$. With this result and with the use of the independence of the aforementioned random variables it is possible to show formally that $(X_j)_{j \geq 0}$ is indeed a Markov chain. This
endeavour doesn’t contribute to the whole and is therefore omitted.

5.3 Invariant distribution

Now that it is clear that $(X_j)_{j \geq 0}$ is a Markov chain, it becomes possible to investigate
whether $(X_j)_{j \geq 0}$ has an equilibrium distribution. One method of doing this is to pos-
tulate a certain probability distribution $\pi = (\pi_\alpha)_{\alpha \in I}$ and show that it is invariant with
respect to the transition matrix $P$; i.e.,

$$\sum_{\alpha \in I} \pi_\alpha p_{\alpha \beta} = \pi_\beta. \quad (5.1)$$

Now the first step is to find a suitable candidate for $\pi$. First of all, the results obtained in
the previous chapter show that the equilibrium distribution of $N_j$ given that the system
is non-empty is a geometric distribution with parameter $\rho$. Furthermore, by means of
renewal theory, it is possible to show that the equilibrium distribution of the residual service time \( R \) of a customer in general satisfies

\[
\mathbb{P}(R = r) = \frac{\mathbb{P}(B \geq r)}{\mathbb{E}[B]}.
\]

(See [2]). With the marginal equilibrium distributions in mind, we postulate the following probability distribution \( \pi = (\pi_\alpha)_{\alpha \in I} \): for \( \alpha = (n, r_1, r_2, \ldots, r_n) \) with \( n > 0 \) define

\[
\pi_\alpha = (1 - \rho)^{n-1} \rho \prod_{m=1}^{n} \frac{\mathbb{P}(B \geq r_m)}{\mathbb{E}[B]}.
\]

The next step is to prove that for \( k \uparrow \infty \) the distribution \( \pi \) is asymptotically invariant by checking if Equation (5.1) holds. Let \( \beta = (n, r_1, r_2, \ldots, r_n) \). At first sight it seems that the states \( \alpha \in I \) such that \( p_{\alpha \beta} > 0 \) are

\[
\begin{align*}
\alpha_1 &= (n, r_1, r_2, \ldots, r_n + 1/k), \\
\alpha_2 &= (n - 1, r_1, r_2, \ldots, r_n - 1/k), \\
\alpha_3 &= (n + 1, r_1, r_2, \ldots, r_n, 1/k),
\end{align*}
\]

with

\[
\begin{align*}
p_{\alpha_1 \beta} &= p_{\alpha_3 \beta} = 1 - \lambda \Delta_k, \\
p_{\alpha_2 \beta} &= \lambda \Delta_k \mathbb{P}(B = r_n),
\end{align*}
\]

for \( k \uparrow \infty \). However, state \( \alpha_3 \) can't be attained for \( k \uparrow \infty \) and thus \( p_{\alpha_3 \beta} = 0 \) for \( k \uparrow \infty \). To show that state \( \alpha_3 \) can't be attained note that

\[
\{N_j = n + 1; R_{j,n+1} = 0\} = \emptyset,
\]

and that probability is a continuous set function for increasing/decreasing sequences of sets. From this it follows that

\[
\begin{align*}
\lim_{k \uparrow \infty} \mathbb{P}(N_j = n + 1, R_{j,1} = r_1, \ldots, R_{j,n} = r_n, R_{j,n+1} = 1/k) \\
\leq \lim_{k \uparrow \infty} \mathbb{P}(N_j = n + 1, R_{j,n+1} = 1/k) \\
\leq \lim_{k \uparrow \infty} \mathbb{P}(N_j = n + 1, R_{j,n+1} \leq 1/k) \\
= \mathbb{P}(N_j = n + 1, R_{j,n+1} \leq 0) = \mathbb{P}(N_j = n + 1, R_{j,n+1} = 0) = 0.
\end{align*}
\]

Hence, checking whether Equation (5.1) holds comes down to checking if the following equation holds

\[
\pi_{\alpha_1} p_{\alpha_1 \beta} + \pi_{\alpha_2} p_{\alpha_2 \beta} = \pi_\beta.
\]
Using $\lambda \Delta_k \downarrow 0$ for $k \uparrow \infty$ it follows that

$$\pi_{a_1} \alpha_{1\beta} + \pi_{a_2} \alpha_{2\beta} = \lim_{k \uparrow \infty} (1 - \rho)^{n-1} \rho \left( \prod_{m=1}^{n-1} \frac{\mathbb{P}(B \geq r_m)}{\mathbb{E}[B]} \right) \frac{\mathbb{P}(B \geq r_n + \frac{1}{k})}{\mathbb{E}[B]} (1 - \lambda \Delta_k)$$

$$+ \lim_{k \uparrow \infty} (1 - \rho)^{n-2} \rho \left( \prod_{m=1}^{n-2} \frac{\mathbb{P}(B \geq r_m)}{\mathbb{E}[B]} \right) \frac{\mathbb{P}(B \geq r_{n-1} + \frac{1}{k})}{\mathbb{E}(B)} \lambda \Delta_k \mathbb{P}(B = r_n)$$

$$= (1 - \rho)^{n-1} \rho \prod_{m=1}^{n} \frac{\mathbb{P}(B \geq r_m)}{\mathbb{E}[B]}$$

$$= \pi_{\beta}.$$
6 Conclusion

As we have seen, the structural properties of the Galton-Watson process are of use in deriving results concerning the M/G/1 queue. The structure provides a way to identify customers and categorise them into generations accordingly. It gives a better understanding of the arrival and service processes of the queue, because it now makes it possible to relate customers to one another. This gives more insight into the dynamics of the busy period. Once the connection between the busy period and the Galton-Watson process is made, a strong property of the Galton-Watson process can be readily applied to the busy period. The population of customers can be divided into sub-populations, such that the sizes of the sub-populations are equally distributed relative to the size of the main population, and such that the sub-populations are independent of one another. The structure of the Galton-Watson process also provides further information about the customers present in the system at the same time in a M/G/1 LCFS-PR queue during a busy period. In particular, the customers present in the system at the same time necessarily form a path in the tree graph to the corresponding Galton-Watson process between the customer that initiated the busy period and the customer that is being served.

Detailed results for the M/G/1 queue are derived using the aforementioned properties. The first property is used to derive the expected length of the busy period for a M/G/1 queue (with arbitrary order of service) and both properties are used to obtain the result that the marginal distribution of the queue length for a non-empty M/G/1 LCFS-PR queue in equilibrium is a geometric distribution.

The latter result served as a motivation to postulate a joint equilibrium distribution of the queue length and the residual service times of the customers. It turned out that the residual service times are independent and identically distributed. The confirmation of the postulated distribution required the application of a continuous-time Markov chain on a continuous state-space. To avoid this, a discrete-time version of the M/G/1 LCFS-PR queue was introduced in a formally imprecise but intuitively clear manner. The postulated distribution was successfully confirmed for the discrete-time version of the M/G/1 LCFS-PR queue.

A suggestion for a further study of the subject would be to improve the introduction of the discrete-time version of the M/G/1 queue by making it formally precise.
A branching process is a mathematical model that is often used to study the growth and decay of a population. In this model it is assumed that every individual of the population has a random number of offspring. A population is described in terms of generations. Consider a starting population made up of one individual, and call this starting population the 0th-generation. The offspring of this single individual are called the 1st-generation; the offspring of the individuals from the 1st-generation form the 2nd-generation, etcetera. The result of such a process can be visualised as in Figure 7.1.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{branching_process.png}
\caption{Schematic representation of a branching process.}
\end{figure}

In Figure 7.1 the nodes resemble the individuals: the 0th-generation consists of one individual, the 1st-generation consists of three individuals, and the 2nd-generation consists of two individuals.

In queuing theory, we are mainly interested in understanding the probabilistic behaviour of customers in a queue. Customers arrive at random moments in time and require random amounts of service. The basic queuing system consists of a waiting room and a server (see Figure 7.2). The order of service of a queuing system may vary per queue. An order of service one often encounters is the first-come first-served serving discipline. In this order of service, the order in which customers are served is the same as the order in which they arrive. A more exotic order of service is the last-come first-served pre-emptive resume discipline, or LCFS-PR for short. In this order of service customers that enter the system immediately seize the server, moving the customer currently occupying the server (if there is any) back into the queue. The service of the customer that is moved back into the queue is resumed as soon as the customer that interrupted the service leaves the system. Note that the service of the customer that just seized the server can also be interrupted by the arrival of a new customer.

The structure of a branching process can be readily applied to a queuing system. Consider a customer entering an empty system as the starting population, and consider
customers that arrive during the service of another customer as the offspring of the customer in service. In this way the generations can be defined as for a branching process. This structure provides a better understanding of the arrival and service processes of the queue, for it is now possible to relate the customers to one another. For instance, the branching structure provides further information about the customers that are present in the system. This information makes it possible to find the expected period of time between the moment that a customer arrives in a previously empty system and the moment the system is completely empty.

The branching structure also proves itself useful in the analysis of the long-term behaviour of a queue, especially for a queuing system that follows the LCFS-PR discipline. In this context “long-term behaviour” refers to the probability distributions of the values of the quantities we are interested in, which change over time. One such a quantity is the number of customers that are present in the queue, or in other words: the queue length. The long-term behaviour of the queue length for a queue that follows the LCFS-PR discipline approaches a so-called geometric distribution. This result can be demonstrated using the branching structure. Another quantity of interest is the remaining time that a customer has to occupy the server, also known as the residual service time. The latter result induces the search for a joint distribution of the queue length and the residual service times of the customers present in the queue. This distribution is hard to obtain if one considers these quantities throughout continuous time. However, if these quantities can only be observed at some fixed points in time, the analysis becomes a great deal easier. In this case it is possible to show that the residual service times of the customers become independent of one another and are all identically distributed as time progresses.

To conclude, it is beautiful to see that, although branching processes and queuing processes are seemingly unrelated, branching processes can be used to provide information about queuing processes. As may already be clear from the aforementioned results, the information provided can lead to some quite deep results about queuing systems.
Bibliography