NUMERIC METHODS FOR STOCHASTIC DISEASE SPREAD MODELS

by

Zeynep Gökçe İşlier

B.S., Industrial Engineering, Boğaziçi University, 2012

M.S., Industrial Engineering, Boğaziçi University, 2014

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Industrial Engineering

Boğaziçi University

2020

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to Assoc. Prof. Wolfgang Hormann for his support and guidance throughout my entire graduate study, and I feel myself lucky to have the opportunity to work with him. I'd further like to thank Prof. Refik Güllü for his insightful comments and encouragements throughout this work. I would like to thank Assist. Prof. Enis Kayış for taking part in my thesis committee, and for his valuable and directing comments.

I wish to thank all my friends, and the former and current members of BUFAIM team for their support and friendship. I want to thank my colleagues Kübra, Mert, Gökalp, and Bahadır for the great times we spent together in BUFAIM.

Lastly my family, they deserve to take credit for all my accomplishments. I would like to thank Turgut for his support on any issue without hesitation, and always be there for me. I thank my sister for being a friend to me, my mother for her unlimited patience and encouragement, and my father for giving me the perspective I have followed.

This work is supported by Boğaziçi University Research Fund Grant Number 11920. I also thank TUBITAK for their financial support during my doctoral studies under BIDEB-2211 programme.

ABSTRACT

NUMERIC METHODS FOR STOCHASTIC DISEASE SPREAD MODELS

Disease spread models are important in controlling the new infectious diseases that suddenly threaten the public health. Mathematical tools are especially important to understand the spread of the disease since experiments are not possible in the area. This study focusses on a stochastic SIR (susceptible-infected-recovered) model for a finite population. We first assume a homogeneous population and study Markov modelling of disease spread for an exponential infectious period. We present the algorithms that compute the expected duration of an epidemic, the final outbreak size distribution and the maximum number of simultaneously infected individuals distribution. After stating the problems with exponential infectious period, we assume an Erlang distributed infectious period allowing us to use Markov chains. The Markov disease spread model proposed for it uses the *remaining stages* as state variables and treats the Erlang distributed infectious period as simply exponential. This enables us to compute the exact final outbreak size distribution for large populations efficiently. Moreover, we propose an approximation for the distribution of the maximum epidemic size using the exact distribution of the *remaining stages*. We also consider a mixture of Erlangs so that by using the first two moments of an infectious period one can fit a corresponding mixture. Furthermore, by considering a mixture of Erlangs distribution for the infectious period and assuming two types of infected individuals as symptomatic and asymptomatic, our proposed models are implemented with the parameters similar to those reported for COVID-19 spread. Finally, a stochastic SIR model for a non homogeneous population is considered. The notion of R_0 for heterogeneous populations is discussed and individual R_0 as the expected number of secondary cases produced by a unique given initially infected individual. We propose a general formula for individual R_0 and use it for the assessment and development of the intervention methods.

ÖZET

STOKASTİK HASTALIK YAYILMA MODELLERİ İÇİN SAYISAL YÖNTEMLER

Halk sağlığını tehdit eden yeni bulaşıcı hastalıkların kontrolünde hastalık yayılım modelleri çok önemlidir. Alanda deneyler de mümkün olmadığından matematiksel araçlar hastalığın yayılmasını anlamak için özellikle önemlidir. Bu çalışma, sınırlı bir nüfus için stokastik SIR (susceptible-infected-recovered) modeline odaklanmaktadır. Başlangıç olarak homojen bir nüfusu düşünüyor ve bulaşıcı dönemin üstel dağılıma sahip olduğunu varsayarak hastalık yayılımının Markov modellemesini çalışıyoruz. Bir salgının beklenen süresini, salgında enfekte olmuş toplam kişi sayısının dağılımını ve aynı anda enfekte olmuş maksimum kişi sayısının dağılımını hesaplayan algoritmaları veriyoruz. Üstel dağılan bulaşıcı dönemin sorunlarını belirttikten sonra Markov zincirlerini kullanmamıza izin veren Erlang dağılan bulaşıcı dönemi varsayıyoruz. Önerdiğimiz Markov hastalık yayılım modeli kalan aşamaları durum değişkenleri olarak kabul ederek üstel dağılıma benzetilir. Bu büyük popülasyonlar için nihai salgın büyüklüğü dağılımını verimli bir şekilde hesaplamamızı sağlar. Ayrıca, kalan aşamaların tam dağılımını kullanarak maksimum enfekte sayısının dağılımı için bir yaklaşım öneriyoruz. Ayrıca bulaşıcı dönem için bir Erlang karışımı düşünüyoruz ki ilk iki momenti verilen herhangi bir bulaşıcı döneme karşılık gelen dağılımı kullanabiliyoruz. Erlang karışıma sahip bulaşıcı dönem ve semptomatik ile asemptomatik olarak iki tip enfekte bireyin varlığını varsayarak önerdiğimiz modelleri COVID-19 yayılımı için uyguluyoruz. Son olarak homojen olmayan bir nüfus için stokastik SIR model düşünülmektedir. Heterojen nüfuslar için R_0 kavramını tartışıyoruz ve bireysel R_0 kavramını enfekte bir birey tarafından üretilen beklenen ikincil vaka sayısı olarak tanımlıyoruz. Bireysel R_0 için genel bir formül öneriyoruz ve bunu hastalığa müdahale yöntemlerinin değerlendirilmesi ve geliştirilmesi için kullanıyoruz.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i
ABSTRACT ii
ÖZET iii
LIST OF FIGURES vii
LIST OF TABLES xi
LIST OF SYMBOLS xiii
LIST OF ACRONYMS/ABBREVIATIONS xv
1. INTRODUCTION 1
1.1. Basics for Our Stochastic Disease Spread Models 2
1.2. Motivation 3
1.3. Organization 5
2. LITERATURE REVIEW 7
2.1. Markov Modeling of Disease Spread 7
2.1.1. Markov Modeling with Exponential Infectious Period 8
2.1.2. Markov Modelling with Alternative Infectious Period Distributions 10
2.1.3. Discussion 13
2.2. Disease Spread for Non Homogeneous Populations 13
2.2.1. Agent Based Simulation 14
2.2.2. Metapopulation 17
2.2.3. Contact Network 20
2.2.4. Discussion 25
3. MARKOV MODELING OF DISEASE SPREAD WITH EXPONENTIAL INFECTIOUS PERIOD 27
3.1. Model Definition 28
3.2. Expected Duration of an Epidemic with Exponential Infectious Period 30
3.3. Final Outbreak Size Distribution with Exponential Infectious Period . . 32
3.4. Maximum Epidemic Size Distribution with Exponential Infectious Period 38
3.5. SIS Markov Model and Expected Duration of an Epidemic 40
3.6. Discussion 43

4. MARKOV MODELING OF DISEASE SPREAD WITH ERLANG INFECTIOUS PERIOD 45
4.1. Model Definition 47
4.2. Final Outbreak Size Distribution with Erlang Distributed Infectious Period 48
4.2.1. Extension to Mixtures of Erlang Distributed Infectious Period . 57
4.3. Distribution of The Maximum Number of Disease Stages and Approximation to Maximum Epidemic Size Distribution 59
4.4. Numerical Results 62
4.4.1. Effect of Infectious Period Variability 64
4.4.2. An Infinite Population Approximation for the Outbreak Probabilities 67
4.4.3. Effect of k on Approximation to Maximum Epidemic Size Distribution 69
4.5. Discussion 71
5. COVID-19 SPREAD: ANALYSIS USING A MODIFIED STOCHASTIC SIR MODEL 75
5.1. Model Definition 77
5.2. Determination of Model Parameters 80
5.3. Numerical Results 83
5.3.1. Final Outbreak Size Distribution for COVID-19 83
5.3.2. Final Outbreak Size Distribution with Changing Contact Rates 89
5.4. Discussion 92
6. R_0 NOTION AND ASSESSMENT OF INTERVENTION STRATEGIES FOR NON HOMOGENEOUS POPULATIONS 94
6.1. The Notion of R_0 for Non Homogeneous Models 95
6.1.1. Use of Individual R_0 on Intervention Analysis 100
6.2. Some Non Homogeneous Population Structures for Influenza Spread . . 101
6.2.1. Model with Multiple Cities 102
6.2.2. Model with a Population of Households 103
6.3. Intervention Analysis 104
6.3.1. Intervention by Vaccination 105

6.3.2. Intervention by Social Distancing 108
6.3.3. Intervention by Use of Antiviral Drugs 109
6.4. A Model with Overlapping Mixing Groups 110
6.5. Discussion 115
7. CONCLUSIONS AND FUTURE RESEARCH 117
7.1. Main Contributions 118
7.2. Possible Future Research Directions 119
REFERENCES 121
APPENDIX A: POPULATION MATRIX GENERATION 133

LIST OF FIGURES

Figure 3.1. State transition diagram for SIR 28

Figure 3.2. The state transitions of process {I(t), S(t)} for SIR with $N = 5$ . . 29

Figure 3.3. Expected duration of an epidemic for exponential infectious period 31

Figure 3.4. Final outbreak size distribution for exponential infectious period . 34

Figure 3.5. Distribution of $I_{\max}$ for exponential(μ) infectious period 39

Figure 3.6. State transition diagram for SIS 40

Figure 3.7. The state transitions of process for SIS with $N = 5$ 40

Figure 4.1. State transition diagram for SI_kR 48

Figure 4.2. The state transitions of process $\{(V(t), S(t))\}$ for $N = 3$ and $k = 2$ 52

Figure 4.3. Exact final size distribution for Erlang distributed infectious period (k, μ) 54

Figure 4.4. Comparison of final size distributions for discrete infectious period and mixture of Erlangs distributed infectious period 58

Figure 4.5. Distribution of $V_{\max}$ for Erlang(k, μ) distributed infectious period . 60

Figure 4.6. Final outbreak size distribution for $k = 1$ based on implementation of algorithm in Figure 4.3 65

Figure 4.7. Final outbreak size distribution for $k = 2$ based on implementation of algorithm in Figure 4.3 66

Figure 4.8. Final outbreak size distribution for $k = 5$ based on implementation of algorithm in Figure 4.3 66

Figure 4.9. Final outbreak size distribution for $k = 10$ based on implementation of algorithm in Figure 4.3 67

Figure 4.10. Cumulative final outbreak size distribution for $k = 1$ 69

Figure 4.11. Cumulative final outbreak size distribution for $k = 2$ 70

Figure 4.12. Cumulative final outbreak size distribution for $k = 5$ 70

Figure 4.13. Cumulative final outbreak size distribution for $k = 10$ 71

Figure 4.14. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 1$ 72

Figure 4.15. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 2$ 72

Figure 4.16. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 5$ 73

Figure 4.17. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 10$ 74

Figure 5.1. State transition diagram for COVID-19 77

Figure 5.2. Final outbreak size distribution for $R_0 = 2.24$ and $k_a = 2$ 84

Figure 5.3. Final outbreak size distribution for $R_0 = 2.24$ and $k_a = 4$ 84

Figure 5.4. Final outbreak size distribution for $R_0 = 3.58$ and $k_a = 2$ 85

Figure 5.5. Final outbreak size distribution for $R_0 = 3.58$ and $k_a = 4$ 85

Figure 5.6. Cumulative final outbreak size distribution for $R_0 = 2.24$ and $k_a = 2$ 86

Figure 5.7. Cumulative final outbreak size distribution for $R_0 = 2.24$ and $k_a = 4$ 87

Figure 5.8. Cumulative final outbreak size distribution for $R_0 = 3.58$ and $k_a = 2$ 87

Figure 5.9. Cumulative final outbreak size distribution for $R_0 = 3.58$ and $k_a = 4$ 88

Figure 5.10. Final outbreak size distribution if $R_c = 0.7$ when at least 5% simultaneous infectives 90

Figure 5.11. Final outbreak size distribution if $R_c = 0.95$ when at least 5% simultaneous infectives 90

Figure 5.12. Final outbreak size distribution if $R_c = 0.7$ when at least 10% simultaneous infectives 91

Figure 5.13. Final outbreak size distribution if $R_c = 0.95$ when at least 10% simultaneous infectives 91

Figure 6.1. Intelligent Vaccination Strategy 106

Figure 6.2. The frequency of individual R_0s without household quarantine . . 112

Figure 6.3. The frequency of individual R_0s after vaccination without household quarantine 114

Figure 6.4. The frequency of individual R_0s after 80% household quarantine . 115

LIST OF TABLES

Table 3.1. Time required to calculate final outbreak size distribution with exponential infectious period 38

Table 4.1. Comparison of final size probability values for different population sizes for k=5 55

Table 4.2. Time required to calculate exact final outbreak size distribution (in seconds). 56

Table 4.3. $E[V_{max}]$, $\frac{2}{k+1}E[V_{max}]$ and $E[I_{max}]$ for different k and R_0 values . . . 63

Table 4.4. Numerical descriptors for final outbreak size distribution with $R_0 = 1.5$ and $\lambda = 3$ for different k values 64

Table 4.5. Comparison of outbreak probability from final outbreak size distribution and infinite population approximation 68

Table 5.1. Published estimates of R_0 for COVID-19 and our corresponding λ 82

Table 5.2. Average percentage of the individuals who have been infected during epidemic (%) 88

Table 5.3. Average percentage of the individuals who have been infected during epidemic (%) 92

Table 6.1. Population sizes and infection probabilities for the population with multiple cities. 103

Table 6.2. Individual R_0 based vaccination with different number of vaccinated individuals for the population with multiple cities 107

Table 6.3. Individual R_0 based vaccination with different number of vaccinated individuals for the population partitioned into households 108

Table 6.4. Quarantine after first day of infection for the population partitioned into households 109

Table 6.5. Use of antiviral drugs with different reduction factors without household quarantine for the population partitioned into households . . 110

Table 6.6. Use of anti viral drugs and 50% household quarantine with different reduction factors for the population partitioned into households . . 110

Table 6.7. Population matrix for a model with overlapping mixing groups . . 111

Table 6.8. Maximum individual R_0 values after random vaccination and vaccination based on individual R_0 without household quarantine . . . 113

LIST OF SYMBOLS

e_j	The unit vector of size $k$with 1 at jth entry and 0 at other entries
$f_D(d)$	The probability mass function of infectious period
i	The current number of infected individuals
$I_i(t)$	The number of infected individuals at time t who needs to goes through i more stages before becoming recovered
$I_M(j)$	The indicator function that is 1 if j is an element of mixing group M of i, 0 otherwise
I_{max}	The maximum number of simultaneously infected individuals
$I(t)$	The number of infected individuals at time t
$\tilde{I}(t)$	The vector of disease stages at time t
k	The shape parameter for an Erlang distributed infectious period
N	Size of population
p_{ij}	The probability of infection between two individuals i and j per time unit
p_{ijt}	The probability of infection between two individuals i and j at time t
$\tilde{p}_{ij}$	The probability of infection between two individuals i and j during infectious period
$P_m(i, s)$	The probability that the final number of recovered individuals is m given that starting with i infectives and s susceptibles
$P_m(w, s)$	The probability that the final number of recovered individuals is m given that starting with w infectives at each stage and s susceptibles
$Q_m(v, s)$	The probability that the maximum number of stages is m given that starting with v stages and s susceptibles
r_I	The average number of infected per time unit
r_V	The average number of infection stages per time unit
R_0	Basic reproduction number

$R_0(i)$	The expected secondary cases for staring with infected individual i
$R(t)$	The number of recovered individuals at time t
s	The current number of susceptible individuals
$S(t)$	The number of susceptible individuals at time t
SI_kR	The SIR model with an Erlang distributed infectious period that is the sum of k exponential variables with parameter μ
T_i	The infectious period of the i^{th} infected individual
v	The current number of disease stages
V_{max}	The maximum number of stages during the epidemic
$V(t)$	The total number of disease stages at time t
w	The vector denoting the number number of infectives at each stage
$\mathcal{X}$	The state space for SI_kR model
λ	The contact rate per individual
λ_{is}	The rate for the total number of contacts ending up with an infection when there exist i infectives and s susceptibles
μ	The rate parameter for an exponential and also an Erlang distributed infectious period
$\Pi_m(v, s)$	The probability that the final number of recovered individuals is m given that starting with v stages and s susceptibles
τ	The termination time of the disease
τ_{is}	The time to extinction of an epidemic given that starting with i infectives and s susceptibles
$\Phi_m(i, s)$	The probability that the maximum number of simultaneously infected individuals is m given that starting with i infectives and s susceptibles

LIST OF ACRONYMS/ABBREVIATIONS

MSEIR	Maternally derived immunity - Susceptible - Exposed - Infectious - Recovered
MSEIRS	Maternally derived immunity - Susceptible - Exposed - Infectious - Recovered - Susceptible
MSIR	Maternally derived immunity - Susceptible - Infectious - Recovered
SEIR	Susceptible - Exposed - Infectious - Recovered
SIR	Susceptible - Infectious - Recovered
SIS	Susceptible - Infectious - Susceptible

1. INTRODUCTION

Epidemiology can be defined as the study of how diseases occur in different populations and why they occur. There exist a considerable number of studies on epidemics mostly done by medical doctors. Since new infectious diseases that are more dangerous have emerged recently, more people are attracted by this area. Therefore, mathematicians are also interested in the area of epidemics and mathematical epidemiology has become an important part of epidemiology. Mathematical epidemiology has used mathematical tools to understand the spread of infectious diseases in populations. Because experiments are not possible in this area, mathematical models have become so important. That allows to compare different intervention strategies against anticipated epidemics or pandemics. They are used in order to understand the dynamics of disease spread.

Compartmental models suggested by Kermack and McKendrick (1927) are mostly used for mathematical modelling of infectious diseases. In compartmental models, the population is divided into compartments assuming that every individual in the same compartment has the same characteristics based on their disease symptoms. The classical epidemic models are usually investigated through ordinary differential equations which are deterministic and indicate time rate of transfer from one compartment to the other. Later, a stochastic framework was considered which allows us more realistic but also more complicated analysis.

The SIR model is one of the simplest compartmental models consisting of three compartments, S denotes the number susceptible, I indicates the number of infectious, and R denotes the number of recovered or immune. In SIR model, if a susceptible individual is infected, it transfers to compartment I and when it is recovered, it transfers to compartment R. This model is for infectious diseases where recovery confers lasting resistance, such as measles. Some infections do not confer any long lasting immunity and thus the SIS (Susceptible-Infectious-Susceptible) model is appropriate for these infections. For many infections, there is a significant incubation period during which the

individual has been infected but is not infectious. Those are called exposed individuals and SEIR (Susceptible-Exposed- Infectious-Recovered) model is used for them. There are also other elaborations on the basic SIR model including MSIR (Maternally derived immunity-Susceptible- Infectious-Recovered), MSEIR (Maternally derived immunity-Susceptible-Exposed- Infectious-Recovered), MSEIRS (Maternally derived immunity-Susceptible-Exposed- Infectious-Recovered-Susceptible).

In this study, we focus on stochastic SIR disease spread models. At first, we consider a finite population in which all individuals are homogeneous and mix uniformly. We propose a Markov model to characterize important properties of disease dynamics. Then, we consider a non homogeneous population and investigate the notion of computable R_0 for heterogeneous models and develop and assess intervention strategies using R_0. The following sections provide more detail on the basics for our stochastic disease spread models, the motivation of the dissertation, and the structure of the manuscript.

1.1. Basics for Our Stochastic Disease Spread Models

We consider an SIR model for finite populations with size N during the thesis. Thus, there are no births and deaths for our model only infection and recovery. SIR is used where individuals infect each other directly. And, an infected individual who recovers from the disease is also modeled to have perfect immunity to disease thereafter.

For Markov modeling of a stochastic SIR model, a homogeneous population and perfect mixing are assumed. It is also assumed that the number of contacts leading to an infection has a Poisson distribution. An infected individual remains infectious during a random time period that is generally assumed to be exponential.

Chain binomial model is another important model in the analysis of disease spread in small groups like households. In chain binomial model, the number of newly infected individuals at time t depends on the number of infected individuals and the number of susceptible individuals at time $t-1$. Further, the number of new infected individuals

at time t follows a binomial distribution.

There are some important quantities that describe the disease spread behaviour like probability of an outbreak, final outbreak size distribution, distribution of maximum number of simultaneously infected individuals. In epidemiology, epidemics are called "outbreak" if the total number of infected individuals goes to infinity for an infinite population size. However, an outbreak probability cannot be calculated exactly for finite population models since it is not clear how an outbreak should be declared.

In the literature, the basic reproduction number, R_0 is defined as the expected number of secondary cases that one infected case would produce in an entirely susceptible population. R_0 was first defined and calculated for deterministic disease spread models with homogeneous populations. The special case of the deterministic model proposed by Kermack and McKendrick in 1927 is the baseline model that has a simple formula for R_0 (Brauer *et al.*, 2008). In epidemiology, R_0 is important since it is directly related to the outbreak possibility. When it is smaller than 1, then the infection is going to disappear before involving a significant number of the population. When R_0 is greater than 1, then there is a probability of a large outbreak.

Attack rate is another important quantity in epidemiology and is calculated by dividing the total number of new cases in the population from the beginning of disease to its end to the total number of persons at risk in the population. Thus, final outbreak size distribution also gives us attack rate in epidemiology.

1.2. Motivation

Investigation of dynamic properties of disease spread is important to assess the threatening effects of infectious diseases on human life and evaluate the intervention strategies that may prevent disease spread. Deterministic models can be solved exactly using different mathematical tools like matrix operations, differential equation systems etc. However, there are some problems with deterministic disease spread models. Firstly, the number of infected cases is actually integer but in deterministic

modelling, differential equations treat them as continuous. In addition, approximations in deterministic studies are good only if both the population size and the number of initial hosts are large. However, the number of infected individuals in the population is typically small at the beginning of the disease. Therefore, the mass action assumption required by differential equation systems cannot be valid anymore and other models have to be considered that model the stochastic nature of the contact pattern of individuals. Another problem deterministic epidemic models encounter is about R_0. For deterministic models when R_0 is greater than 1, an outbreak is certain even if in reality this is not the case.

Stochastic epidemic models can handle most of the problems that deterministic models have. Stochastic models are more realistic but also more complex to analyze. Under the stochastic approach, achieving numerical solutions is very difficult even for homogeneous populations and it is almost impossible for non homogeneous populations. During our literature survey, we observe that some important quantities for properties of stochastic epidemic models like attack rate are calculated via approximations and simulations even for homogeneous populations.

In this study, we target to obtain numerical solutions for some important characteristics of disease spread like final outbreak size, time until extinction and maximum number of simultaneously infected individuals for Markov models. We also enhance our study with Markov chains considering more realistic infectious period than exponential infectious period. However, our Markov modelling is only appropriate for homogeneous populations.

Analysis of stochastic disease spread models for non homogeneous populations are mostly done by simulation. In the literature, Longini *et al.* (2004) estimate the basic reproduction number for overlapping mixing groups via agent based simulation claiming that it cannot be directly calculated. Motivated by that wrong statement we investigated the notion of R_0 for non homogeneous populations and developed a simple formula for R_0. We also claim that a notion of R_0, newly defined for non-homogenous distributions, can be used to assess the effect of intervention strategies and thus to

develop better intervention strategies.

1.3. Organization

The rest of the thesis starts with a literature review. Section 2.1 and Section 2.2 summarize the related literature in Markov modeling of stochastic disease spread and disease spread for non homogeneous populations, each of which are followed by a discussion.

Chapter 3 is dedicated to our studies related with Markov modelling of disease spread for homogeneous populations. The model properties are given and important results for epidemiology like the expected duration of epidemic, the distribution of the total number of infected individuals during epidemic, the distribution of the maximum number of simultaneously infected individuals are derived. Algorithms for exact solutions are proposed and explained in detail. As an extension, a formula for a SIS model to calculate the expected duration of an epidemic is given. Finally, infectious period distributed by Erlang is suggested as an extension. Chapter 3 ends with a discussion showing that the use of the exponential distribution for the infectious period is not realistic.

Thus, we present a model where infectious period is distributed as an Erlang random variable and propose an efficient computational procedure for the final outbreak size in Chapter 4. We also consider a mixture of Erlangs so that using the first two moments of infectious period one can fit a corresponding distribution. Finally, we suggest an approximate distribution for the maximum epidemic size distribution and implement numerical studies to demonstrate the practical importance of the proposed algorithms.

Chapter 5 includes the implementation of our methods suggested in Chapter 4 for COVID-19 spread. We integrate the recent COVID-19 epidemiological data into a stochastic SIR model considering two types of infectives as asymptomatic and symptomatic. Furthermore, we also investigate the effect of the timing and the intensity

of social distancing on the final outbreak size distribution by using state dependent probabilities.

Chapter 6 is devoted to disease spread models for non homogeneous populations. In this part, we investigate the notion of R_0 for heterogeneous populations and introduce individual R_0 as the expected number of secondary cases produced by a given unique initially infected individual. A general formula to calculate individual R_0 values is presented and some intervention methods are assessed by using it. Optimal intervention strategies are also suggested and numerical studies are carried on. Finally, Chapter 7 concludes the thesis and points out some possible future research directions.

2. LITERATURE REVIEW

The number of studies published in the area of mathematical epidemiology is increasing fast and there are many studies that approach the problem from different aspects. Studying the relevant literature, the deterministic modeling approach is the oldest in the mathematical theory of epidemics. Kermack and McKendrick (1927) is the oldest landmark paper in the area. They study the deterministic models and the formulas they offered are still used today in deterministic modelling implementations. However, we focus on stochastic disease spread models and in this chapter, a brief literature survey is provided on this topic.

In the following sections, the papers are examined in two groups. The first group includes the papers that consider the Markov modelling of disease spread. The second group contains the papers which study disease spread for non homogeneous populations especially the notion of R_0 and the assessment of intervention strategies are considered. At the end of each section, a discussion that motivates our study is provided.

2.1. Markov Modeling of Disease Spread

In this section, we investigate mostly the papers working on Markov modelling of stochastic disease spread models. Markov modeling of disease spread can be useful to calculate important results for epidemiology like total number of infected individuals during an epidemic, the duration of an epidemic and the distribution of infected individuals for an epidemic. We analyse the papers on this topic in two different groups. The first group includes the studies assuming exponential infectious period while the second group consists of the papers that criticize exponential infectious period and present other approaches for Markov modelling of disease spread with non exponential infectious period.

2.1.1. Markov Modeling with Exponential Infectious Period

In the literature, Markov modelling of disease spread assuming exponential infectious period is popular since they make the computation of important epidemic properties like the distribution of the total number of infected individuals during an epidemic and the duration of an epidemic, possible (Lloyd, 2001).

Gani and Jerwood (1971) are among the earliest researchers who apply Markov chain methods for epidemic models. They demonstrate that both Greenwood and Reed Frost chain binomial models form Markov chains and they can be used to obtain probabilities for the duration of time and the total number of infected individuals in an epidemic. They carry out a study of chain binomial models as Markov chain embedded in continuous time processes.

Gibson and Renshaw (1998) implement Markov chain Monte Carlo methodology to estimate parameters for SIR models. They develop methods based on the Metropolis Hasting algorithm and illustrate their use though simulated realizations of the process. They show that the methods presented can be useful to provide meaningful estimates for parameters and parameter uncertainty by comparing estimated likelihoods with theoretical results.

Daley and Gani (2001) employ embedded Markov chain to compute the probability distribution associated with the final size of recovered individuals. They compute the transition matrix for the embedded matrix by considering the probabilities to implement their suggested method.

Naasell (2002) studies the quasi stationary distribution and time to extinction for stochastic SI, SIS, and SIR models. He states that the stochastic models are too complex to allow explicit solutions. The quasi stationary distribution is especially important since the expected time to extinction is expressed in terms of it. Since an explicit solution for the quasi stationary distribution cannot be determined, a derivation based on a diffusion approximation is developed.

Allen (2008) reviews three different types of stochastic SIS and SIR model formulations including discrete time Markov chain, continuous time Markov chain and stochastic differential equations. They investigate the disease properties of stochastic models like probability of disease extinction, probability of disease outbreak, quasi stationary probability distribution, final size distribution and expected duration of an epidemic.

Gómez *et al.* (2010) introduce a probabilistic discrete time Markov chain formulation for contact based spreading of diseases in complex networks. They construct a dynamical system that generalizes from an individual contact process to the case where all connections are concurrently used. In order to show the validity of their approach, they obtain more details at the individual level of description based on Monte Carlo simulations.

Keeling and Ross (2007) suggest to use Kolmogorov forward equation to understand stochastic disease dynamics since it allows to simultaneously consider the probability of each possible state occurring. Kolmogorov forward equation is linear and has a matrix formulation providing disease spread dynamics. They describe the advantages of matrix formulation of dynamics to compute the expected time until extinction and to compare expected total cost of control strategies.

Artalejo *et al.* (2010) are interested in the distribution of the number of recovered individuals for SIR not only during the time until the absorption but also in transient state. They also consider SIS epidemic model and study the number of recovered individuals. They apply the generating functions, factorial moments and moment generating functions to derive final size distribution. Artalejo *et al.* (2010) also study both the maximum number of infected individuals during an epidemic and the distribution of the current number of infected individuals given that the maximum number of infected during epidemic is not above a certain value for a SIS model. They compute the distribution of maximum number of infected individuals in transient regime and until absorption by numerically inverting the Laplace transforms.

House *et al.* (2013) compare methods to compute the probability mass function of the total number of infections during an epidemic. Methods evaluated by them cover Markov chain methods including Brute force methods, Bailey's method, Path integral and Null space method. Then, numerical efficiency of Markov chain methods are compared.

Kuhnert *et al.* (2014) introduce a birth death SIR model for estimating parameters like R_0. The model approximates a classical stochastic SIR presenting trajectories of the number of susceptible, infected and recovered individuals. In their Markov chain Monte Carlo implementation of the birth death SIR model, they show that the tree likelihood is suitable, by illustrating that they can infer parameters from simulated phylogenies with high accuracy.

Dadlani *et al.* (2016) develop a continuous time Markov chain model based on the retrial queuing notion in which infected nodes return back to become susceptible only after being served by one idle recovery unit. Their aim is to assess the impact of infected nodes retrying for recovery controlled by limited resources on the transient behaviour of SI model assuming both homogeneous and non homogeneous contacts.

Amador and Lopez-Herrero (2017) focus on two important measures for the severity of an epidemic that are the total number of infectious cases and the maximum number of simultaneously infectious individuals. They consider a stochastic SEIR model and present a recursive algorithm for determining the final size distribution. They also calculate the probability of having at least m simultaneously infected individuals recursively from linear equations.

2.1.2. Markov Modelling with Alternative Infectious Period Distributions

Lloyd (2001) states that most mathematical models describe the infectiousness period by an exponential distribution. He investigates the effects of more realistic descriptions of the infectiousness period on two epidemiological consequences. He implements Monte Carlo simulation for every single transition and observes that as the

variability increases, the number of infected individuals is more likely to decrease. Thus, the disease persistence is diminished and less stable behaviour is observed.

Trapman and Bootsma (2009) establish a relationship between disease spread behaviour and dynamics of M/G/1 queues with process sharing. They constitute a link between the relation of disease spread and epidemics and the relation of M/G/1 queues and birth and death processes.

Hernández-Suárez *et al.* (2010) are interested in expected time to extinction and quasi stationary distribution for SIS and SEIS epidemic models by applying queuing theory. In their study, they derive approximations for expected time to extinction and quasi stationary distribution and all derived approximations assume a general distribution of the infectious period duration. Finally, they compare the approximations with the results obtained from simulations.

Fackrell (2009) claims that exponential distribution suffers from its lack of versatility, being characterized by only one parameter. They introduce phase type distributions that can be used in the healthcare industry. Phase type distributions are applicable to Markov chains and it includes a number of popular distributions like exponential distribution, the order p generalized Erlang distribution, the order p hyper exponential distribution and the order p Coxian distribution.

Gamma distributed infectious period is popular for epidemic models with Markov chains since it is identical to the distribution of the sum of independent exponential distributed random variables. Anderson and Watson (1980) are among the earliest researchers who consider gamma distribution for infectious period and obtain a number of important results regarding the progress of the epidemic. They implement deterministic approximation and derive equations for the final size distribution. They also use a branching process approximation for minor outbreaks and martingale central limit theorem approximation for major outbreaks.

Andersson and Britton (2000) also assume gamma distributed infectious period

and study stochastic SIR and SEIR models in dynamic populations to compute time to extinction starting from quasi stationary states. They claim that to find the exact distribution of the epidemic process is not manageable so they implement deterministic approximation and diffusion approximations. It is also stated that population size determines how the outbreak occurs.

Ma and Earn (2006) consider arbitrarily distributed transmission rates of infectious individuals and claim that this complexity has no impact on the final size formula. They replace an exponential distribution with mean $1/\gamma$ by a Gamma distribution with parameters k and $k\gamma$. It is suggested to replace a single infectious stage with k identical exponentially distributed substages with mean $1/k\gamma$.

Nishiura *et al.* (2012) use a branching process model to compute the final outbreak size distribution for minor outbreaks and apply the method to final size data of pneumonic plague with a basic reproduction number above 1. They compute the probability of extinction by using probability generating functions of branching process assuming gamma distributed infectious period.

Craft *et al.* (2013) analyse simulated data for early stages of disease to obtain important insights for predicting major outbreaks. Their study, based on simulation, applies to two approaches which are trajectory matching and discriminant function analysis assuming both exponential and gamma distributions with different parameters.

Leclerc *et al.* (2014) focus on incubation period since it is important to detect how disease pathogens spread and infects susceptible individuals. They fit alternative probability models to data to demonstrate how incubation period changes with host age. They also apply to simulation to examine the sensitivity of the lag between epidemics of cryptic infection and the associated epidemics of symptomatic disease for different distributional assumptions like exponential and Gamma.

2.1.3. Discussion

We consider a SIR disease spread model for a homogeneous population by using Markov chains. Thus, we will begin our discussion of the literature by considering the studies modelling Markov chains for disease spread. We review these studies by separating them into two groups.

The first group of studies assumes exponential infectious period and derives important quantities like probability of an outbreak, attack rate and R_0 that describes the disease spread behaviour. The second group includes the papers that apply to Markov chain modelling but assume infectious period other than exponential distribution. Among alternative distributions, Erlang(k, μ) distribution becomes more popular since it is identical to the distribution of sum of k independent exponential(μ) random variables.

Moreover, all studies presented in Section 2.1 consider disease spread models for homogeneous populations and implement Markov chain modelling approach to analyse disease spread behaviour numerically.

2.2. Disease Spread for Non Homogeneous Populations

In this section, we study disease spread for non homogeneous populations. We mainly focus on R_0 notion for heterogeneous populations and assessment of intervention strategies. We identified three main groups of papers on this topic. The first mainly authored by medical doctors and practitioners uses agent based simulation to assess disease spread behaviour. A second group considers meta population models that are especially useful for modelling pandemics. In the third group, mostly mathematicians have investigated epidemic models by considering contact networks.

2.2.1. Agent Based Simulation

Our interest in epidemiology started from agent based simulation models in epidemics. Agent based simulations are developed in order to understand the disease spread mechanism in small populations. They are actually simple but allow modelling of complex phenomena.

Longini *et al.* (2004) use stochastic epidemic simulations in order to measure the effects of antiviral agents and vaccine. They make a simulation for a heterogeneous population separated by age and mixing groups. They show how average R_0 can be computed via simulation for heterogeneous populations if the required infection transition probabilities are known. Longini *et al.* (2005) also present how the basic reproductive number R_0 affects the effectiveness of targeted antiviral prophylaxis, quarantine, and pre vaccination by specifying threshold values.

Lipsitch *et al.* (2003) also implement agent based simulation to assess the impact of public health efforts on reducing the size of the epidemic for dangerous SARS (severe acute respiratory syndrome). They report that R_0 per index case for SARS is highly variable if some control measures like isolation of SARS cases and quarantine of their asymptomatic contacts are implemented. Thus, they estimate R_0 values changing by time. They conclude that large variation in R_0 by each index case will decreases the probability of large epidemic compared to small variation in R for the same R value.

Patlolla *et al.* (2004) model the dynamics of the spread of infectious diseases by using agents to model the interaction between people and pathogens. The dynamics of the spread are analysed by using the graphical output generated via simulation and the behaviour of the spread is exactly the same as for the classical SIR model. They also indicate the threshold levels of the SIR model by using the graphical outputs.

Dunham (2005) gives the design and the implementation of an agent based epidemiological simulation system by using the MASON toolkit. His simulation model is so robust that it can be implemented for epidemic models including SIS, SIR and even

more complicated SEIR (Susceptible-Exposed-Infected-Removed) or SLIR (Susceptible-Latent-Infected-Removed) models. He concludes that his simulation tool can be used for realistic studies if proper parametrization can be done.

Ferguson *et al.* (2005) carry on a simulation analysis to assess H5N1 influenza spread after policies to increase social distance. Germann *et al.* (2006) introduce and use a large scale stochastic simulation model to study the spread of a pandemic strain of influenza virus for different R_0 values. They consider some intervention strategies including targeted antiviral prophylaxis, dynamic mass vaccination, school closure, social distancing and reductions in travel.

Wu *et al.* (2006) consider a SEIR disease spread model for a population with households. They assume that initial basic reproduction rate is 1.8 and it is considered to be dependent on the behaviour of the host population and vary during time. They implement a sensitivity analysis to estimate the average decrease in attack rates after household based intervention methods

Gustafsson and Sternad (2007) propose a Poisson simulation model for both micro and macro simulation models. Epidemic models generally consist of many discrete entities and there are two main types of simulation for these models. Micro model that describes each entity with its attribute and behavior and macro model which describes population with the number of entities in different states. Agent based model can be regarded as a micro model but due to its complexity, it cannot be implemented to macro models. Poisson simulation technique in this brings consistency for these two models by adapting agent based simulation to large models.

Auchincloss and Diez Roux (2008) focus on the study of how features of residential environments or neighbourhoods affect health. Then, they use agent based simulation since it allows the system dynamics to follow a system consisting of heterogeneous entities. They investigate the determinants of the spatial patterning of physical activity so they prepare a simple and informative agent based model. Therefore, they conclude that agent based approaches make developments of sophisticated theoretical models

possible.

Milne *et al.* (2008) propose an individual based model of a small community in the developed world with details including exact household structure, household demographics and movement within the community and individual contact patterns. They analyze the effects of some interventions including school closure, increased case isolation, workplace non-attendance and community contact reduction on the spread of pandemic influenza. They state that earlier and continuous implementations of social distancing become effective even for great R_0 values up to 2.5. They obtain consistent results with the others who also study epidemics models with individual based approaches.

Degli Atti *et al.* (2008) make a study on the spread of influenza in Italy. They use an individual based SEIR model based on demographic data since individual based models give the most reliable estimates of infectious disease spreads. They use the 2001 census data while generating the individual based simulation and evaluate the effectiveness of several control methods for different basic reproduction values. The results they obtained have become consistent with reality because the model requires detailed information on the population characteristics and the data they used are routinely collected by national statistics in Italy.

Bobashev *et al.* (2007) demonstrate a hybrid model starting with an agent based model and switching to equation based after the number of infected individuals is large. This model takes the advantage of agent based models' power in describing structured epidemiological processes and the advantage of equation based model for better tractability.

Perez and Dragicevic (2009) develop an agent based modelling approach integrating geographic information system to simulate the spread of a communicable disease in an urban environment. Their model is implemented on data obtained from Metro Vancouver and census data from Statistics Canada for the municipality of Burnaby. They state that agent based simulations is beneficial for better understanding of disease

spread dynamics improving control of an epidemic outbreak.

Andradóttir *et al.* (2011) introduce a stochastic pandemic influenza model calibrated by documented illness attack rates and basic reproductive number estimates. An agent based approach is used for simulation; intervention strategies are recommended based on the obtained simulation results for economic cost estimations. According to the results, higher vaccine efficacy and greater vaccine coverage decrease the average attack rate but vaccine coverage with disruptive social distancing strategies seem even more effective.

2.2.2. Metapopulation

Another technique for modelling epidemics in a heterogeneous population is metapopulation. In metapopulation modelling, different levels of relations are constituted between homogeneous groups so it is an approach similar to agent based modelling. Actually, recent research is more interested in structured metapopulation models rather than detailed agent based models. Metapopulation models are more proper for the study of worldwide epidemics than other modelling approaches since agent based modelling is easily implemented only for small sized populations.

In early modelling of worldwide epidemic, Hufnagel *et al.* (2004) present a probabilistic model in order to describe the spread of infectious diseases through the world and forecast the geographical spread of epidemics. In this epidemic model, a stochastic local infection dynamics among individuals and a stochastic traffic in a worldwide network are combined. Then, predictions for the future of infectious diseases are done and endangered regions in the world are identified by simulations.

Watts *et al.* (2005) state that many spatial and network models are proposed to give the aspects of interaction structure among individuals but they suffer from limited tractability and incapability of yielding general results due to their complexity. Therefore, they propose a meta population model in which homogeneous mixing is assumed in local subpopulations and a nested hierarchy of subpopulations exists. They

claim that the nested hierarchy of subpopulations allows to model large populations and even global pandemics. Their results also support that the final size and duration of an epidemic are highly sensitive to the structure of the population even if the R_0 value is constant. Moreover, similar distributions of epidemic size can correspond to very different values of R_0.

One point that makes meta population model so important recently is the increase of airline traffic in the world. Thus, some regions in the world are under bigger threat of a global pandemic than others having less traffic. Colizza *et al.* (2006) study the structure of the population flows between different geographical regions. They state that as the airline traffic has been increasing, the geographical space shrinks that diseases can spread. Therefore, the epidemic model is analysed by using an information theory approach allowing the quantitative characterization of the heterogeneity level and the predictability of disease spreading patterns. In another study of global epidemics, the properties of the airline transportation network specifying the diffusion pattern of diseases and the reliability of outputs based on stochasticity of disease transmission and traffic flows (Colizza *et al.*, 2006). Then, Colizza *et al.* (2007) study a metapopulation stochastic epidemic model considering airline travel flow information. A sensitivity analysis of pandemic is done for different virus infectiousness and different initial conditions. They compute R_0 and find the effects of basic reproduction number on a global pandemic through metapopulation technique. For metapopulation models, a threshold value R_* is also provided below which the epidemic cannot spread to a macroscopic fraction of subpopulations. A study for better understanding the effect of travel restrictions in epidemic containment is also carried on (Colizza and Vespignani, 2007). Colizza and Vespignani (2008) show that the metapopulation network exhibits a global threshold for the subpopulation invasion.They derive an explicit analytic expression for the invasion threshold that determines the minimum number individuals traveling among subpopulations.

Balcan *et al.* (2009) investigate the seasonal transmission potential and the peak time of influenza worldwide. They use a global structured metapopulation model. They regard the mobility of individuals and the transportation data worldwide. The

model parameters are estimated by maximum likelihood estimation technique based on past data. Based on the output of the global structured metapopulation model, the R_0 estimation is done effectively and the peak times are determined. This is beneficial for specifying the time and the extent of the intervention strategies. Barthélemy *et al.* (2010) show that the possibility of global disease spread is described in terms of bond on the network so they give a lower bound for the global epidemic threshold for all model parameters and all networks.

Ajelli *et al.* (2010) compare agent based models with structured metapopulation models in their study. They generate an agent based model and structured metapopulation model for a pandemic in Italy and whole Europe with identical initial conditions. Then, they compare the results obtained from both modelling techniques. Based on their results, the two models are consistent in the peak time of the epidemics. Moreover, the fraction of the population who become infected is greater for the metapopulation modelling than agent based modelling because the contact structure is more detailed in agent based modelling and homogeneity assumption is stronger in metapopulation models. One drawback of the agent based modelling is the difficulties in obtaining reliable data for most regions due to its detailed structure. Therefore, Ajelli *et al.* (2010) offer a hybrid model that make predictions at the global scale by implementing a metapopulation technique and representing individuals by employing the agent based approach. Furthermore, Merler and Ajelli (2009) study how different levels of population heterogeneity and different human mobility patterns affect the influenza pandemic. According to the results, the epidemics affect the European countries differently due to the huge differences in socio-demographic properties of European countries. Cumulative attack rate, peak daily attack rate and R_0 depend on socio-demographic parameters.

Singh (2014) also studies epidemics in metapopulation networks. He describes a set of mixed subpopulations that are coupled together with some links. He analyses different network cases including fully connected and strongly connected. Then, the distribution of outbreak size, epidemic threshold, probability of a large outbreak, and size of a large outbreak are found on both strongly connected metapopulation networks

and weakly connected metapopulation networks. Finally, Ball *et al.* (2015) state that even if theory and application of meta population models made great progress recently, there are still important challenges that remain for future study. They present seven major challenges they observe for meta population models of epidemics.

2.2.3. Contact Network

In the literature, the calculation of the basic reproduction number R_0 has always been considered important. R_0 is of interest not only because it is used in calibrating infection transmission probability but also it directly quantifies the possibility of a pandemic. Contact network is another important issue in disease spread modelling because it depends highly on both intrinsic features of the pathogen and the properties of the network structure.

Chowell *et al.* (2003) present how network characteristics can be used to generate social networks. They propose that modelling on individual interactions can be useful to analyse dynamic processes like slowing the spread of infectious diseases. They represent a network whose edge weights are derived from the daily movements of individuals between locations. The number of edges between nodes are determined by power-law distributions based on school, work and social activities.

Barrat *et al.* (2004) study the networked structures in transportation infrastructures, social phenomena, and biological systems. They state that the networked structures have gained importance in recent years due to their complex structure. The networked structures are specified by both their topology and the information dynamics on them. In this networks a weight is assigned to each edge while a link can only exist or not for standard models.

Keeling and Eames (2005) review the basis of epidemiology and network theory and study some idealized network types and approximation techniques.. Then, they introduce different methods that allow an approximation to the network to be ascertained.

Bansal *et al.* (2006) implement contact network epidemiology while investigating the optimal vaccination strategy under vaccine supply shortage. They compare two different vaccination strategies that are mortality based strategy targeting high risk populations and morbidity based strategy targeting high prevalence populations. They conclude that the mortality based strategy is preferred for high R_0 values and the morbidity based strategy is better for moderate R_0 values.

Stroud *et al.* (2006) implement power law scaling of new infection rates for SIR and SEIR model. They state that the expected number of new infections per day per infectious person has power law scaling with respect to the susceptible fraction in the population. However, in traditional approach there exists a linear relationship between them assuming a homogeneous mixing assumption. However, some people have greater chance to get infection in real life like highly connected people. This social contact structure is regarded by the power law scaling approach. Moreover, it shows that the total number of recovered people is considerably lower when using social contact structure than for the traditional approach.

Wallinga *et al.* (2006) use data of social contacts in order to estimate the age specific transmission parameters. They derive possible contact patterns based on age distributions and household sizes. It is also stated that there exist surveys that gives information on social contacts. They conclude that transmission parameters based on social contacts yield better estimations for the age specific infection patterns because the pattern of social contacts indicate that people have generally more contacts with others of similar age.

Meyers (2007) suggests contact network epidemiology as a more powerful approach for modelling disease spreads. She considers the SARS case whose basic reproduction number was first estimated to be between 2.2 and 3.6. Therefore, a worldwide pandemic was expected but only a limited spread of SARS was observed. Meyers (2007) explains this by using contact network epidemiology. She asserts that R_0 estimates for SARS were done based on the data from a hospital and a crowded apartment building where the contact rates were extremely high. Therefore, it is inappropriate

to expect the same R_0 value for a large population having a different contact network. She emphasizes that recording both the number of new infections and the total number of contacts during the infectious period is important in order to estimate the average transmissibility; estimating the basic reproduction number R_0 is not enough.

Trapman (2007) describes an infection spread by approximating the network by a random graph. He considers two different infection types. The first one assumes that an infected individual can infects its neighbours independently during his infectious period. And, the second one assumes that an infected individual either infects all of its neighbours or no one.

Ajelli and Merler (2008) also study the structured and unstructured contacts in individual based epidemic modelling. They try to evaluate the effectiveness of intervention strategies in heterogeneous populations and assert that the diversity in the choice of unstructured contacts yields different model outputs and different effective rates. It is suggested that not only the socio economic data are used for structured contacts but also data including activity levels, distance information and time to improve the model and to decrease the variance in outputs.

Ajelli and Merler (2009) claim that the development of a dynamic contact network is required for a long term study of individual based epidemic studies while they study hepatitis A transmission. Based on the dynamic contact network they study, the intervention methods and their effectiveness are determined to prevent hepatitis A in Italy. Dimitrov and Meyers (2010) also introduce a contact network epidemiology to address public health policies and develop powerful computational methods to optimize epidemic control strategies.

Paarporn *et al.* (2015) also investigate whether the available information about epidemics is used to determine the level of individuals' interactions with their neighbours thus actually changing the contact network. They generate a stochastic SIS model on an arbitrary connected network. Then, they use their contact network model and study on how the awareness of agent is effective in reducing the disease spread.

The computation of R_0 value has always been important as it informs people about the possibility of an outbreak. Contact network approach indicates that the value of R_0 without contact network might be misleading. However, the R_0 value is still important and in case the contact network is known, the basic reproduction number can be still computed (Andradóttir *et al.*, 2011).

There are also others who define basic reproduction number in different ways for the models with heterogeneous contacts. Inaba (2012) introduces a new definition of R_0 based on the generation evolution operator that can be applied to all type of heterogeneous populations. He states that the spectral radius of next generation operator is R_0 and it can be calculated as the positive eigenvalue.

Keegan and Dushoff (2016) calculate a finite population reproduction number considering different types of heterogeneity that are heterogeneity in mixing rate and heterogeneity in probability per contact. Then, they investigate the effects of heterogeneity on finite population reproduction number and conclude that heterogeneity decreases the finite population reproduction number when R_0 is large relative to the population size.

The Markov chain approach is especially popular for computing the basic reproduction number for heterogeneous populations since it makes the calculation of R_0 easier as the level of heterogeneity increases (Hernandez-Suarez, 2002). Allen and Burgin (2000) compare deterministic and stochastic SIS and SIR models where discrete time stochastic models are Markov chains. They conclude that the disease is eliminated when R_0 is smaller than 1 while the disease exists in the population when R_0 is greater than 1 for deterministic models. However, the disease ends whatever R_0 values is for stochastic models since Markov chain diagram for stochastic SIS and SIR includes at least one absorbing class and finite state space. Therefore, they make their analysis conditioned on non extinction of disease and conclude that there exist a quasi stationary probability distribution when $R_0 > 1$ whose mean is consistent with deterministic equilibrium.

Artalejo and Lopez-Herrero (2013) are concerned with the number of secondary cases of disease modelled by a Markov chain. However, they introduce new concepts that are the exact reproduction number R_{e0} and the population transmission number R_p since R_0 overestimates the average number of secondary cases.

Economou *et al.* (2015) study a stochastic SIS model for a population size N where N is small and each individual has heterogeneous contacts. They also model the problem as a continuous time Markov chain and calculate the number of infected individuals, the length of an outbreak, and the maximum number of infected individuals. They also define the basic reproduction number as random variable rather than an expectation.

López-García (2016) models a small population with N individuals by considering a graph. Each edge has a weight β_{ij} representing the interaction level from i to j. He assumes that an infectious contact on network occurs after an exponentially distributed random time and analyzes the effectiveness of control strategies by this modelling approach.

The heterogeneous disease spread models considered by Markov chains assume small population size to derive explicit results. For great population size, the level of heterogeneity is requited to decrease. Ball and Neal (2002) consider a SIR model only with two levels of mixing and define two possible contacts that are *local* and *global* contacts. They compute a threshold parameter R_* that governs whether a global epidemic occurs and the average fraction of recovered individuals. Their theoretical contact network is specialized to the households model, the overlapping groups model and the great circle model. Furthermore, R_* computed for a population partitioned into households is suggested to be used for developing optimal vaccination policies. They consider optimality in terms of vaccination cost and derive explicit results from a constructive method to describe optimal vaccination strategies (Ball and Lyne, 2002).

Andradóttir *et al.* (2011) also suggest that the fundamental aim of most disease spread models is to reduce R_0 below 1. This can be possibly reached by changing the

contact networks. Household prophylaxis, school closure and general social distancing are some possible ways that change contact network and allow to reduce R_0. Possible public health interventions other than changing contact network might be vaccination and antiviral treatment.

2.2.4. Discussion

We can mainly categorize the disease spread models for non homogeneous groups into three as the ones that consider agent based simulation, meta population and contact network.

Our interest in epidemiology starts from agent based simulation technique. Agent based simulations are developed in order to understand the disease spread mechanism in small populations. Agent based simulation is actually simple but allows the modelling of complex phenomena. Probably for that reason it seems to be popular especially among medical doctors.

Meta population is another method that constitutes different levels of relations between homogeneous groups and is mostly implemented by mathematicians. It is a similar approach to agent based modelling but it can be implemented for huge populations while agent based simulation is used for small size populations. Moreover, contact structure is more detailed in agent based modelling and homogeneity assumption is stronger in Meta population. Thus, meta population is more proper for the study of worldwide epidemics.

Contact network has more detailed contact structure than meta population so it is similar to agent based modelling approach. However, the researchers who apply to graphs to model disease spread obtain exact numerical results rather than simulation results. They generally consider small size populations. As their heterogeneity assumption increases, the population size decreases. Markov modelling is especially an important method that has been used in contact network literature recently.

The majority of the studies outlined in Section 2.2 includes main modeling approaches that are used in epidemiology to model disease spread for non homogeneous populations. The decision on which approach is implemented depends on the problem types.

3. MARKOV MODELING OF DISEASE SPREAD WITH EXPONENTIAL INFECTIOUS PERIOD

In epidemiology, stochastic models are especially important due to their more realistic assumptions as compared to deterministic models. Markov chains for modelling the stochastic dynamics in understanding of an epidemic has also become quite popular recently since it allows derivation of important quantities including probability of an outbreak, quasi-stationary probability distribution, final size distribution of an epidemic, and expected duration of an epidemic.

Even if Markov chain modelling of disease spread is not a new approach in the literature, it became quite popular in recent years as it allows to predict the behaviour of stochastic disease spread explicitly using a set of equations. By solving these equations, it is possible to derive important results for discrete time disease spread models and continuous time disease spread models. There exist great similarities between continuous time disease spread models and queuing theory where births and deaths correspond to infection and recovery in disease spread models. Markov chain models promise great improvement for epidemiology as the yield important results on epidemics. First of all, they are used to calculate basic reproduction number R_0 in stochastic epidemic models and to estimate parameter values by using a given R_0 value (Hernandez-Suarez, 2002; Artalejo and Lopez-Herrero, 2013). It is also possible to know the probability of an outbreak, the quasi stationary distribution of an epidemic, and its final size by using Markov chains (Allen, 2008).

In this part of our thesis, we deal with continuous time Markov chain modelling of SIR disease spread model with constant population size to find important properties of disease dynamics. We start with the expected time to extinction of an epidemic for a SIR model. Even if in the literature results for the expected duration of an epidemic were obtained using differential equation approximations, factorial moments and generating functions, we implement first step analysis to calculate the expected

time to disease extinction and also for the final outbreak size distribution and for the maximum number of simultaneously infected individuals. Further, we consider a SIS model and calculate the expected time to disease extinction for it.

3.1. Model Definition

We consider a finite population with size N where all individuals are homogeneous and mix uniformly. For SIR disease spread models, the population is divided into three classes as susceptible, infected and recovered. The flow diagram for each member of the population is displayed in Figure 3.1. Let $S(t)$, $I(t)$, and $R(t)$ be the number of susceptible, infected, and recovered at time t, respectively. Since this is a finite population model, we have $N = R(t) + I(t) + S(t)$, so any one of the variables is implied by the other two.

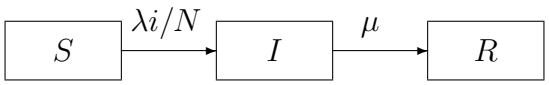

Figure 3.1. State transition diagram for SIR

The transition rate from the susceptible state to the infected state is dependent on both the number of susceptible cases, and the number of infected cases so it is denoted by λ_{is} where i and s denote the current number of infected and susceptible individuals, respectively. It is assumed that each infected individual has close contact with others in the population according to a Poisson process with parameter λ. Since every close contact would be with a susceptible individual with probability s/N and there are i infected cases, the total number of contacts (per unit time) that will end up with an infection follows a Poisson process with parameter $\lambda_{is} = is\lambda/N$ (Hernández-Suárez *et al.*, 2010; Naasell, 2002). We assume that the infectious period has an exponential distribution with parameter μ. The recovery rate for the states with i infected cases becomes $i\mu$. Therefore, it is possible to model this infection using Markov chains where the state is the combination of the number of infected individuals and the number of susceptible individuals, $(I(t), S(t))$. The Markov chain diagram with transition rates depending on both the number of infected individuals and the number of susceptible

individuals for a population size $N = 5$ becomes

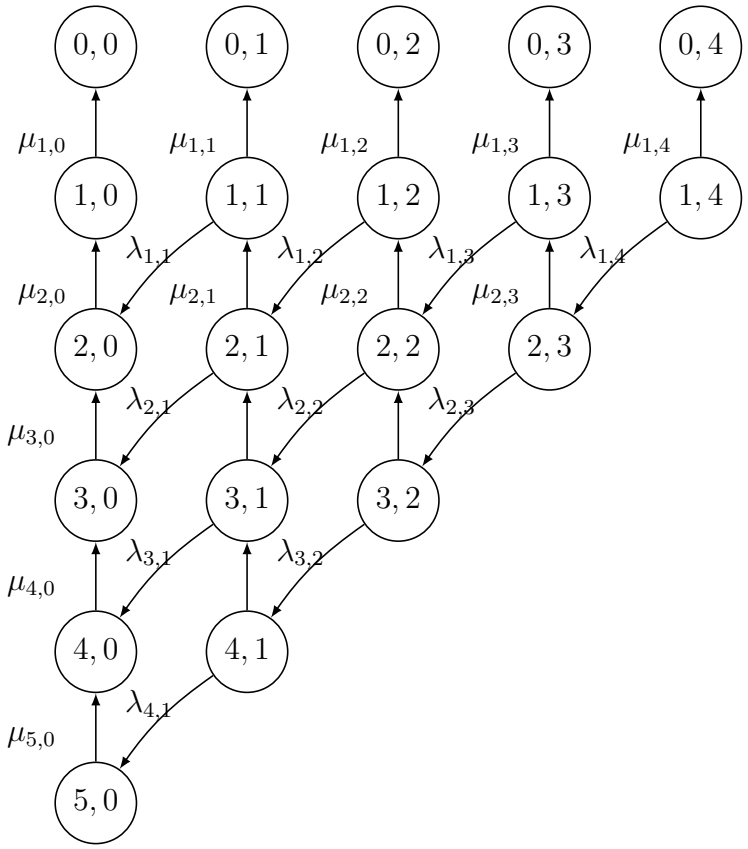

Figure 3.2. The state transitions of process {I(t), S(t)} for SIR with $N = 5$

The Markov chain in Figure 3.2 shows that there are N absorbing states for stochastic SIR models for which the number of infected individuals are zero but the number of susceptible individuals can take integer values from 0 to $N - 1$. There are questions to be answered. One interesting question is the time until absorption (the time until disease ends). Another question is in which state the process is absorbed so what is the number of susceptible individuals who can manage to escape from the disease (total number of recovered individuals). It is also important to observe the maximum number of infected individuals on the path the model follows (maximum number of simultaneously infected individuals).

3.2. Expected Duration of an Epidemic with Exponential Infectious Period

Let $\tau_{i,s}$ be the time to extinction for a population size N starting with i infected individuals and s susceptible individuals. We know that $\tau_{i,s}$ values become zero whenever there are no infected individuals.

$$\tau_{0,s} = 0 \text{ for } s = 0, 1, ..N. \tag{3.1}$$

By considering (3.1) as boundary equations for Markov SIR models, we can calculate the expected duration of an epidemic for any values of i and s. We start with calculating $E[\tau_{i,s}]$ for the states with zero susceptible individuals that is

$$E[\tau_{i,s}] = \sum_{j=1}^{i} \frac{1}{\mu_j}. \tag{3.2}$$

Then, we can benefit from the first step analysis of Markov models while specifying equations of expected time to extinction for the states with one remaining individual that are

$$E[\tau_{i,1}] = \frac{1}{\lambda_{i,1} + \mu_i} + \frac{\lambda_{i,1}}{\lambda_{i,1} + \mu_i} E[\tau_{i+1,0}] + \frac{\mu_i}{\lambda_{i,1} + \mu_i} E[\tau_{i-1,1}]. \tag{3.3}$$

In Equation 3.3, $1/(\lambda_{i,1} + \mu_i)$ gives us the expected time elapsed in state $(i, 1)$ and first step analysis is considered. We already know the value of $E[\tau_{i+1,0}]$ and there is only one term left for which we write Equation 3.4.

$$E[\tau_{i-1,1}] = \frac{1}{\lambda_{i-1,1} + \mu_{i-1}} + \frac{\lambda_{i-1,1}}{\lambda_{i-1,1} + \mu_{i-1}} E[\tau_{i,0}] + \frac{\mu_{i-1}}{\lambda_{i-1,1} + \mu_{i-1}} E[\tau_{i-2,1}]. \tag{3.4}$$

So, we continue to specify such equations until the time when the number of infected individuals becomes zero. and we obtain Equation 3.5 that is

$$\begin{aligned} E[\tau_{i,1}] &= \frac{1}{\lambda_{i,1}+\mu_i} + \frac{\lambda_{i,1}}{\lambda_{i,1}+\mu_i} E[\tau_{i+1,0}] \\ &+ \sum_{j=1}^{i-1} \left\{ \left(\frac{1}{\lambda_{j,1}+\mu_j} + \frac{\lambda_{j,1}}{\lambda_{j,1}+\mu_j} E[\tau_{j+1,0}] \right) \left(\prod_{l=j+1}^{i} \frac{\mu_l}{\lambda_{l,1}+\mu_l} \right) \right\}. \end{aligned} \quad (3.5)$$

In general for s susceptible, we can write

$$\begin{aligned} E[\tau_{i,s}] &= \frac{1}{\lambda_{i,s}+\mu_i} + \frac{\lambda_{i,s}}{\lambda_{i,s}+\mu_i} E[\tau_{i+1,s-1}] \\ &+ \sum_{j=1}^{i-1} \left\{ \left(\frac{1}{\lambda_{j,s}+\mu_j} + \frac{\lambda_{j,s}}{\lambda_{j,s}+\mu_j} E[\tau_{j+1,s-1}] \right) \left(\prod_{l=j+1}^{i} \frac{\mu_l}{\lambda_{l,s}+\mu_l} \right) \right\}. \end{aligned}$$

If we replace $\lambda_{i,s}$ with $is\lambda/N$ and μ_i with $i\mu$, we can calculate the expected duration of an epidemic for given parameter values λ, μ, and N for any number of infected individuals i and susceptible individuals s by using Equation 3.6 that is

$$E[\tau_{i,s}] = \sum_{j=1}^{i} \left\{ \left(\frac{N\mu}{s\lambda+N\mu} \right)^{i-j} + \left(\frac{N}{js\lambda+jN\mu} + \frac{s\lambda}{s\lambda+N\mu} E[\tau_{j+1,s-1}] \right) \right\}. \quad (3.6)$$

Thus, we obtain a set of equations and it is possible to calculate the expected time to extinction by solving these equations recursively using the algorithm in Figure 3.3.

Algorithm 1
1. *Set* $\tau_{0,s} = 0$ for $s = 0, 1, \ldots, N$
2. *Set* $s = 0$
3. **for** $i = 1, 2, \ldots, N$
4. *Compute* $\tau(i, s)$ *using Equation* (3.2)
5. **end for**
6. *Set* $s = s + 1$
7. **for** $i = 1, 2, \ldots, N - s$
8. *Compute* $\tau(i, s)$ *using Equation* (3.6)
9. **end for**
10. *Set* $s = s + 1$. If $s \leq N - 1$ go to step 7. Otherwise, stop the algorithm.

Figure 3.3. Expected duration of an epidemic for exponential infectious period

3.3. Final Outbreak Size Distribution with Exponential Infectious Period

In epidemiology, final outbreak size distribution indicates the distribution of total number of recovered individuals when the disease ends. Define τ as the termination time of the disease: the time when no infected individuals remain,

$$\tau = \inf\{t \geq 0 : I(t) = 0\}.$$

Given that we start with $(I(0), S(0)) = (i, s)$, the probability that the final number of recovered individuals equals m when the outbreak ends can be obtained as

$$P_m(i, s) = \Pr\{R(\tau) = m | (I(0), S(0)) = (i, s)\}.$$

To calculate $P_m(i, s)$ values, we firstly specify both the absorbing states and the transient states in our Markov model. The absorbing states for SIR include only the states with zero infected individuals so in our model there are $N+1$ different absorbing states like $(0, s)$ where s can take any integer value from 0 to N. And the remaining states are considered as transient. Further, the transition rate matrix for transient states becomes

$$M = \begin{array}{c} \\ \vdots \\ (i-1,s) \\ (i,s) \\ (i+1,s-1) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccccccc} \scriptstyle ... & \scriptstyle ... & \scriptstyle (i-2,s) & \scriptstyle (i-1,s) & \scriptstyle (i,s) & \scriptstyle ... & \scriptstyle (i,s-1) & \scriptstyle (i+1,s-1) & \scriptstyle ... & \scriptstyle (i+2,s-2) & \scriptstyle ... \end{array} \\ \begin{bmatrix} \vdots & \ddots & & & & & & & & & \\ 0 & 0 & \mu_{i-1} & 0 & 0 & \dots & \lambda_{i-1,s} & 0 & \dots & 0 & \dots \\ 0 & 0 & 0 & \mu_i & 0 & \dots & 0 & \lambda_{i,s} & \dots & 0 & \dots \\ 0 & 0 & 0 & 0 & \mu_{i+1} & \dots & 0 & 0 & \dots & \lambda_{i+1,s-1} & \dots \\ \vdots & \ddots & & & & & & & & & \end{bmatrix} \end{array}$$

.

Actually, every $P_m(i,s)$ value is equal to the probability of being absorbed by absorbing state $(0, N-m)$. Further, it is possible to implement first step analysis in order to calculate $P_m(i,s)$ values which is given in Equation 3.7 using the transition rate matrix.

$$P_m(i,s) = \frac{\lambda_{is}}{\lambda_{is}+\mu_i} P_m(i+1, s-1) + \frac{\mu_i}{\lambda_{is}+\mu_i} P_m(i-1, s). \tag{3.7}$$

If the transition rates λ_{is} and μ_i are taken as $(is\lambda)/N$ and $i\mu$ respectively, Equation 3.7 becomes

$$P_m(i,s) = \frac{\lambda s}{\lambda s+\mu N} P_m(i+1, s-1) + \frac{\mu N}{\lambda s+\mu N} P_m(i-1, s). \tag{3.8}$$

For this Markov model, the next event can be either a new infection or a recovery. Equation 3.8 shows that neither the probability of a new infection $\lambda s/(\lambda s+\mu N)$ nor the probability of recovery $\mu N/(\lambda s+\mu N)$ depends on the number of infected individuals at the current state. These probabilities change only with the number of susceptible individuals. Moreover, we need to compute $P_m(i,s)$ for $s = N-m, N-m+1, ..., N$ to determine the final outbreak size distribution since $P_m(i,s)$s are equal to zero for s values less than $N-m$. Therefore, it is possible to set the boundary conditions such that

$$P_m(0,s) = 1 \quad \text{for } s = N-m \quad \text{and} \quad P_m(0,s) = 0 \quad \text{for } s < N-m. \tag{3.9}$$

The recursive method that combines Equation 3.8 and the boundary conditions given in (3.9) is summarized in the pseudo-code given in Figure 3.4.

Algorithm 2

1. *Set* $m = 1$
2. *Set* $P_m(i, s) = 0$ *for* $s < N - m$ *and* $i = 0, 1, ...(N - s)$
3. *Set* $s = N - m$
4. *Set* $P_m(0, s) = 1$
5. **for** $i = 1, 2, \ldots, N - s$
6. *Compute* $P_m(i, s)$ *from Equation* 3.8
7. **end for**
8. *Set* $s = s + 1$
9. *Set* $P_m(0, s) = 0$
10. **for** $i = 1, 2, \ldots, N - s$
11. *Compute* $P_m(i, s)$ *from Equation* 3.8
12. **end for**
13. *Set* $s = s + 1$. If $s \leq N - 1$ go to step 9. Otherwise, go to step 14.
14. *Set* $m = m + 1$. If $m \leq N$ go to step 2. If $m = N + 1$ stop algorithm.

Figure 3.4. Final outbreak size distribution for exponential infectious period

We start to calculate $P_m(i, s)$ values for the states with $N - m$ susceptible. Since a new infection reduces the number of susceptible for these states to $N - m - 1$ for which $P_m(i + 1, N - m - 1)$ is zero, $P_m(i, N - m)$ is calculated as

$$P_m(i, N - m) = \frac{\mu N}{\lambda(N - m) + \mu N} P_m(i - 1, N - m) = \left(\frac{\mu N}{\lambda(N - m) + \mu N} \right)^i . \quad (3.10)$$

Further, we compute all $P_m(i, s)$ values recursively by increasing the number of susceptible one by one. Since we already know $P_m(i, N - m)$'s for all i values, we

continue with calculating $P_m(i, N-m+1)$'s such that

$$\begin{aligned}
& P_m(i, N-m+1) = \\
= & \sum_{a_1=0}^{i-1} \left\{ \left(\frac{\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1} \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N} P_m(i-a_1+1, N-m) \right\} \\
= & \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N} \\
& \sum_{a_1=0}^{i-1} \left\{ \left(\frac{\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1} \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i-a_1+1} \right\} \\
= & \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N} \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+1} \sum_{a_1=0}^{i-1} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1}.
\end{aligned}$$

Then, we calculate $P_m(i, N-m+2)$'s as

$$\begin{aligned}
& P_m(i, N-m+2) = \\
= & \sum_{a_2=0}^{i-1} \left\{ \left(\frac{\mu N}{\lambda(N-m+2)+\mu N} \right)^{a_2} \frac{\lambda(N-m+2)}{\lambda(N-m+2)+\mu N} \right. \\
& \left. P_m(i-a_2+1, N-m+1) \right\} \\
= & \frac{\lambda(N-m+2)}{\lambda(N-m+2)+\mu N} \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N} \sum_{a_2=0}^{i-1} \left\{ \left(\frac{\mu N}{\lambda(N-m+2)+\mu N} \right)^{a_2} \right. \\
& \left. \left(\left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i-a_2+2} \sum_{a_1=0}^{i-1} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1} \right) \right\} \\
= & \frac{\lambda(N-m+2)}{\lambda(N-m+2)+\mu N} \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N} \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+2} \\
& \sum_{a_2=0}^{i-1} \left\{ \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+2)+\mu N} \right)^{a_2} \sum_{a_1=0}^{i-a_2} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1} \right\}.
\end{aligned}$$

In general, we can write the equation for the final outbreak size m as

$$\begin{aligned}
& P_m(i, N-m+j) = \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+j} \\
& \prod_{k=1}^{j} \left\{ \frac{\lambda(N-m+k)}{\lambda(N-m+k)+\mu N} \sum_{a_k=0}^{i-\sum_{l=k+1}^{j+1}(a_l-1)} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+k)+\mu N} \right)^{a_k} \right\} \quad (3.11)
\end{aligned}$$

for $j = 1, 2, ...m - i$ where $a_{j+1} = 1$.

We will prove Equation 3.11 using induction.

i) For j=1

$$P_m(i, N-m+1) = \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N}$$
$$\sum_{a_1=0}^{i-1} \left\{ \left(\frac{\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1} \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i-a_1+1} \right\}$$
$$= \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+1} \frac{\lambda(N-m+1)}{\lambda(N-m+1)+\mu N} \sum_{a_1=0}^{i-1} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+1)+\mu N} \right)^{a_1}.$$

ii) For j=t

$$P_m(i, N-m+t) = \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+t}$$
$$\prod_{k=1}^{t} \left\{ \frac{\lambda(N-m+k)}{\lambda(N-m+k)+\mu N} \sum_{a_k=0}^{i-\sum_{l=k+1}^{t+1}(a_l-1)-1} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+k)+\mu N} \right)^{a_k} \right\}.$$

iii) For j=t+1

$$
\begin{aligned}
&P_m(i, N-m+t+1) = \\
&= \sum_{a_{t+1}=0}^{i-1} \Bigg\{ \left(\frac{\mu N}{\lambda(N-m+t+1)+\mu N} \right)^{a_{t+1}} \frac{\lambda(N-m+t+1)}{\lambda(N-m+t+1)+\mu N} \\
&P_m(i-a_{t+1}+1, N-m+t) \Bigg\} \\
&= \sum_{a_{t+1}=0}^{i-1} \Bigg\{ \left(\frac{\mu N}{\lambda(N-m+t+1)+\mu N} \right)^{a_{t+1}} \frac{\lambda(N-m+t+1)}{\lambda(N-m+t+1)+\mu N} \\
&\Bigg\{ \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i-a_{t-1}+1+t} \\
&\prod_{k=1}^{t} \left(\frac{\lambda(N-m+k)}{\lambda(N-m+k)+\mu N} \right)^{i-a_{t+1}+1-\sum_{l=k+1}^{t+1}(a_l-1)-1} \sum_{a_k=0} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+k)+\mu N} \right)^{a_k} \Bigg) \Bigg\} \Bigg\} \\
&= \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+t+1} \frac{\lambda(N-m+t+1)}{\lambda(N-m+t+1)+\mu N} \\
&\sum_{a_{t+1}=0}^{i-1} \Bigg\{ \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+t+1)+\mu N} \right)^{a_{t+1}} \\
&\prod_{k=1}^{t} \left(\frac{\lambda(N-m+k)}{\lambda(N-m+k)+\mu N} \right)^{i-a_{t+1}+1-\sum_{l=k+1}^{t+1}(a_l-1)-1} \sum_{a_k=0} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+k)+\mu N} \right)^{a_k} \Bigg) \Bigg\} \\
&= \left(\frac{\mu N}{\lambda(N-m)+\mu N} \right)^{i+t+1} \\
&\prod_{k=1}^{t+1} \Bigg\{ \frac{\lambda(N-m+k)}{\lambda(N-m+k)+\mu N} \Bigg)^{i-\sum_{l=k+1}^{t+1}(a_l-1)-1} \sum_{a_k=0} \left(\frac{\lambda(N-m)+\mu N}{\lambda(N-m+k)+\mu N} \right)^{a_k} \Bigg\}
\end{aligned}
$$

In the Markov model generated to calculate final outbreak size distribution, there are N^2 states for exponential infectious period. Since we compute $P_m(i,s)$ values for every m value from $m=1$ to $m=N$, the time required to find the final outbreak size distribution increases like N^3.

Table 3.1 gives the time required to calculate final size distribution for different population sizes. However, these results are computed for the cases with every possible initial state (i,s) so the time to calculate final size distribution for given initial state

Table 3.1. Time required to calculate final outbreak size distribution with exponential infectious period

Population Size N	100	500	1000
Time (seconds)	2.02	240.52	1886.14

will become much smaller.

3.4. Maximum Epidemic Size Distribution with Exponential Infectious Period

Another stochastic variable of interest is the maximum epidemic size, as the treatment resources required for an epidemic are proportional to this maximum size. Actually, intervention methods and control strategies aim at reducing the peak epidemic size. Define

$$I_{\max} = \max_{0 \le t \le \tau} I(t)$$

as the maximum number of infected cases, or the peak epidemic size, until the extinction of the disease (τ). Given that we start with $(I(0), S(0)) = (i, s)$, the probability that the maximum number of simultaneously infected individuals during epidemic equals m is denoted as

$$\Phi_m(i, s) = \Pr\{max(I(t)) = m | (I(0), S(0)) = (i, s), t \le \tau\}.$$

To calculate $\Phi_m(i, s)$ values, we firstly state the probabilities for some initial

states for which we already know the maximum epidemic size probabilities which are

$$\begin{aligned}\Phi_m(i,0) &= 1, \text{ for } i = m \\ \Phi_m(i,0) &= 0, \text{ for } i \neq m \text{ and} \\ \Phi_m(i,s) &= 0 \text{ for } i+s < m < i.\end{aligned}$$

Because the next event can be either a new infection or a recovery, we can implement the first step analysis again which is

$$\Phi_m(i,s) = \frac{\lambda_{is}}{\lambda_{is}+\mu_i}\Phi_m(i+1,s-1) + \frac{\mu_i}{\lambda_{is}+\mu_i}\Phi_m(i-1,s). \tag{3.12}$$

If the transition rates λ_{is} and μ_i are taken as $(is\lambda)/N$ and $i\mu$ respectively, Equation 3.12 becomes

$$\Phi_m(i,s) = \frac{\lambda s}{\lambda s+\mu N}\Phi_m(i+1,s-1) + \frac{\mu N}{\lambda s+\mu N}\Phi_m(i-1,s). \tag{3.13}$$

In the implementation of (3.13), we first calculate $Q_m(1,1)$ for the maximum possible value of m: $1+1$. Then, we decrease m one by one down to its minimum possible value: the current value of i. The procedure to calculate the distribution of maximum number of simultaneously infected individuals is given by the algorithm in Figure 3.5. Using this algorithm, we can find the maximum epidemic size distribution.

Algorithm 3

1. *Set* $\Phi_m(0,s) = 0$ *for* $m \neq 0$ *and* $\Phi_m(i,s) = 0$ *for* $i+s < m < i$
2. *Set* $\Phi_m(i,s) = 1$ *for* $m = i$ *and* $\Phi_m(i,s) = 0$ *for* $m \neq i$
3. *Set* $s = 1$
4. *Set* $i = 1$
5. **for** $m = i+s, i+s-1, \ldots, i+1$
6. *Compute* $\Phi_m(v,s)$ *using Equation* 3.13
7. **end for**
8. *Set* $\Phi_i(i,s) = 1 - \sum_{m=i+1}^{i+s}\Phi_m(i,s)$
9. *Set* $i = i+1$. If $i \leq N-s$ go to Step 5. Otherwise, go to step 10.
10. *Set* $s = s+1$. If $s \leq N-1$ go to Step 4. Otherwise, if $s = N$ stop the algorithm.

Figure 3.5. Distribution of $I_{\max}$ for exponential(μ) infectious period

3.5. SIS Markov Model and Expected Duration of an Epidemic

In this section, we also consider susceptible-infected-susceptible models for which the flow diagram for each member of the population is presented in Figure 3.6. Because

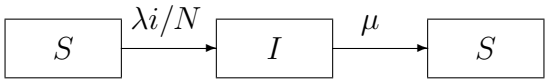

Figure 3.6. State transition diagram for SIS

the summation of the number of infected individuals I and the number of susceptible individuals S equals N, the disease dynamics in SIS epidemic models depend only on the number of infected individuals $I(t)$ and the states of the Markov model become $I(t)$. The Markov chain diagram with transition rates depending on the number of infected individuals for a population size 6 is shown in Figure 3.7.

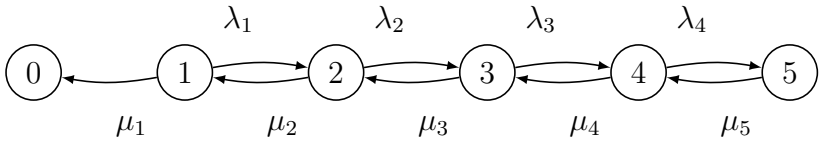

Figure 3.7. The state transitions of process for SIS with $N = 5$

The Markov chain in Figure 3.7 indicates that the stochastic SIS model is equivalent to a simple birth and death process. It is obvious that there is only one absorbing state when the population is free of infected individuals $I(t) = 0$. Therefore, determination of final outbreak size does not make sense since disease ends with 0 infected and N susceptible individuals with probability one and we are only interested in calculating the expected duration of an epidemic for SIS models.

Let τ_i be the time to extinction with initially i infected individuals. If we denote the time to go from state i to state $i-1$ with T_i, the expected duration of an epidemic

starting with i infected individuals, E$[\tau_i]$ can be calculated as

$$E[\tau_i] = \sum_{k=1}^{i} E[T_k]. \tag{3.14}$$

We can calculate $E[T_i]$ by conditioning on the next event that is either a recovery or an infection. Lets define I_i

$$I_i = \begin{cases} 1, & \text{if if first state change is due to a recovery.} \\ 0, & \text{if first state change is due to an infection.} \end{cases}$$

Then, the expected time to go from state i to state $i-1$ conditioning on the next event becomes

$$\begin{aligned} E[T_i|I_i = 1] &= \frac{1}{\lambda_i + \mu_i}, \\ E[T_i|I_i = 0] &= \frac{1}{\lambda_i + \mu_i} + E[T_{i+1}] + E[T_i] \end{aligned}$$

where the probability of the next event is

$$\begin{aligned} P[I_i = 1] &= \frac{\mu_i}{\lambda_i + \mu_i}, \\ P[I_i = 0] &= \frac{\lambda_i}{\lambda_i + \mu_i}. \end{aligned}$$

Thus, $E[T_i]$ becomes

$$E[T_i] = \frac{\mu_i}{\lambda_i + \mu_i}\frac{1}{\lambda_i + \mu_i} + \frac{\lambda_i}{\lambda_i + \mu_i}\left(\frac{1}{\lambda_i + \mu_i} + E[T_{i+1}] + E[T_i]\right)$$

and hence we can obtain a relationship between $E[T_i]$ and $E[T_{i+1}]$ such that

$$E[T_i] = \frac{1}{\mu_i} + \frac{\lambda_i}{\mu_i}E[T_{i+1}].$$

We suppose that there will be at most N infected individuals and we know E$[T_N]$

exactly that is

$$E[T_N] = \frac{1}{\mu_N}.$$

Thus, we can also calculate E$[T_{N-1}]$ as

$$E[T_{N-1}] = \frac{1}{\mu_{N-1}} + \frac{\lambda_{N-1}}{\mu_{N-1}} \frac{1}{\mu_N}.$$

Therefore, we derive recursive equations to calculate $E[T_i]$ that is

$$E[T_i] = \frac{1}{\mu_i} + \sum_{k=1}^{N-i} \left(\frac{1}{\mu_{i+k}} \prod_{l=0}^{k-1} \frac{\lambda_{l+i}}{\mu_{l+i}} \right). \tag{3.15}$$

We assume that contact rate is $(\lambda i(N-i))/N$ and recovery rate is iμ. Then, if we replace λ_i and μ_i in 3.15with $(\lambda i(N-i))/N$ and iμ respectively, Equation 3.15 becomes

$$E[T_i] = \frac{1}{i\mu} + \sum_{k=1}^{N-i} \left(\frac{1}{(i+k)\mu} \prod_{l=0}^{k-1} \frac{\lambda(N-l-i)}{N\mu} \right)$$

that is

$$E[T_i] = \frac{1}{i\mu} + \sum_{k=1}^{N-i} \left(\frac{1}{(i+k)\mu} \frac{(N-i)!}{(N-i-k)!} \left(\frac{\lambda}{N\mu} \right)^k \right). \tag{3.16}$$

While modelling SIS using Markov chains, the states include only the current number of infected individuals and have one dimension. Therefore, the expected duration of an epidemic can be calculated by a single formula given in Equation 3.16 rather than recursive algorithms.

3.6. Discussion

For a stochastic SIR model with Exponential infectious period, we have studied the expected duration of an epidemic, the distribution of the total number of recovered individuals and the distribution of the maximum number of individuals who are simultaneously infected until the end of disease by obtaining recursive algorithms. As an extension, we also derive a formula to calculate expected duration of an epidemic for SIS.

We assume that the exponential infectious period assumption for stochastic disease spread modelling by Markov chains is mainly popular due to its simplicity. Its simplicity comes from its single parameter μ and its memoryless property. But these properties are also the main problem. The memoryless property is unrealistic and only one parameter implies a lack of versatility.

Therefore, different distributions for infectious period have been offered in the literature recently. Lloyd (2001) studies the effects of more realistic distributions of the infectious period for SIR models and finds a major effect on disease dynamics. There are others who benefit from previously established results in queuing theory using e.g. M/G/1 and M/G/N type queues for which infectious periods do not have to be exponential. They derive approximate results for disease spread dynamics with different distributions of infectious period (Trapman and Bootsma, 2009; Hernández-Suárez *et al.*, 2010). Craft *et al.* (2013) apply simulation for a wide range of outbreak sizes by considering both exponential and gamma distributed transition times. Leclerc *et al.* (2014) also examine the sensitivity of infectious period to exponential versus gamma by implementing simulation considering both incubation and infectious period as Gamma distributed. Fackrell (2009) introduces phase type distributions that are used in healthcare services.

We also suggest to use Erlang distribution for infectious period as a more versatile class of distributions. Erlang distribution can be appropriate for disease spread modelling since its coefficient of variation is less than or equal to one and it allows anal-

yses using a Markovian compartment model due to the fact that Erlang distribution is identical to the distribution of the sum of k independent Exponential(μ) random variables.

4. MARKOV MODELING OF DISEASE SPREAD WITH ERLANG INFECTIOUS PERIOD[1]

Markov chain modelling of an epidemic employs the assumption that infectious period is exponential. However, different distributions for infectious period that are considered more realistic are suggested in the literature recently. Erlang distribution is especially popular for realistic infectious period assumptions. Many studies assume Erlang distribution for infectious period but they mostly derive approximations such as the deterministic large population approximation, branching process approximation and central limit theorem approximation (Anderson and Watson, 1980; Bailey, 1964). Andersson and Britton (2000) consider Erlang distribution for both latent period and infectious duration. They analyse their model by using a Markovian compartment model relying on the fact that Erlang distribution with rate parameter λ and shape parameter n is identical to the distribution of the sum of n independent exponential random variables with rate parameter λ. However, their results for some quantities like quasi-stationary distribution, and the time to extinction of the disease are also based on deterministic and diffusion approximation. Black and Ross (2015) calculate the final size distribution for exponential distribution and extend their model to a phase type distribution by splitting the infectious period into k stages. By accepting the advantage of Ball method for handling any infectious period, they claim that their method is the most efficient for values of k typically of interest since the time to compute a final size distribution grows linearly with the size of the state space. However, the size of their state space is $\binom{N+k+1}{N}$ so their computational time grows exponentially with k and their method can be calculated in practice implemented only for only small populations with sizes not larger than 100.

Therefore, SI_kR model with an Erlang distributed infectious period is computationally challenging since one needs to keep track of the system state, which can be very memory intensive and time consuming. Our aim is to provide a fast and efficient

[1]This chapter is published in İşlier *et al.* (2020a).

computational method for obtaining the distribution of the number of recovered individuals for the susceptible-infected-recovered (SIR) stochastic model over the entire duration of the epidemic. Our method relies on the concept of total remaining disease stages and does not have to keep track of the number of infected individuals at each stage of the SI_kR model. It transforms the problem in such a way that many interesting quantities can be obtained even for large size problem instances.

Another important feature of an epidemic is its maximum size since it influences the planning of control methods for dealing with epidemics. Daniels (1974) obtains an approximation for the distribution of the maximum number of infected individuals by considering a diffusion approximation of the imbedded random walk. Artalejo *et al.* (2010) also consider the maximum number of infected individuals for SIS epidemic models. Amador and Lopez-Herrero (2017) are interested in both the distribution of the final size epidemic and the maximum number of simultaneously infected individuals for a stochastic SEIR model and suggest to use the first step methodology. They discuss the distributions in transient states and the quasi stationary distributions of the maximum number of infected individuals until absorption. Amador *et al.* (2019) are interested in a multi type SIR epidemic model for exponentially distributed recovery time and derive the joint distribution of the random vector (X_{max}, T_{max}) describing the maximum number of simultaneously infected individuals and the time to reach this maximum number for the first time. They present algorithmic solutions for the mass functions of X_{max} rather than analytical formulas, by using matrix algebra.

In this part of our thesis, we firstly present a model that uses the total number of *remaining stages* as the state variable for the case when the infectious period is distributed as an Erlang random variable, resulting in an efficient computational procedure that can be used to find the final outbreak size distribution. This is a considerable improvement on previous models using the Erlang infectious period. Essentially, our state transformation enables us to treat the Erlang distributed infectious period as simply exponential. Although our model is more restrictive than Ball (1986) in terms of the infectious period distribution, our result and the structure of our expression allows us to compute the final outbreak size distribution for population sizes of thou-

sands if not more with a very high level of precision. We feel that the generality of an Erlang distributed infectious period, coupled with our method of exact computation of the outbreak distribution for large populations make our model a strong competitor to existing methods. We also demonstrate that our method can be extended to a mixture of Erlangs so that by using the first two moments of an infectious period one can easily fit a corresponding mixture. Then, we consider maximum epidemic size distribution and present an algorithm for computing the distribution of the maximum disease stages over the course of an infection. We obtain an approximation for the maximum epidemic size distribution that gives exact results for exponential infectious period. Lastly, we present a computational study for understanding the contribution of various factors affecting the final outbreak size and understanding the effect of k on our maximum epidemic size distribution.

4.1. Model Definition

We again assume a finite population with size N where all individuals are homogeneous and mix uniformly. Further, we have $N = R(t) + I(t) + S(t)$, so any one of the variables is implied by the other two. It is also assumed that the total number of contacts (per unit time) that will end up with an infection follows a Poisson process with parameter $\lambda_{is} = is\lambda/N$ (Hernández-Suárez *et al.*, 2010; Naasell, 2002). And, the infectious period has an Erlang distribution with parameters k and μ.

Because the Erlang random variable can be represented as the sum of k independent and identically distributed exponential random variables with parameter μ, the model is commonly referred as SI_kR. The flow diagram for each member of the population can be represented as in Figure 4.1 where the box with I_i denotes the compartment in which an infected individual needs to go through i more stages before becoming recovered and arrows denote the transitions between compartments.

In an Erlang distributed infectious period, each infected individual is assumed to go through k stages, before becoming recovered, and the time spent in each stage is exponentially distributed with parameter μ. Accordingly, let $I_i(t)$ be the number of

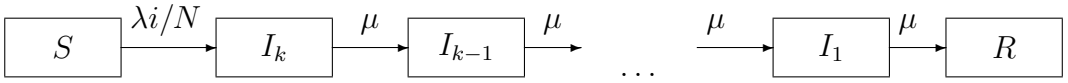

Figure 4.1. State transition diagram for SI_kR

infected individuals at time t that need to go through i more stages before recovery. Hence,

$$I(t) = \sum_{j=1}^{k} I_j(t).$$

Let $\tilde{I}(t) = (I_1(t), I_2(t), \ldots, I_k(t))$ be the vector of disease stages at time t. Then the state of the SI_kR at time t is given by $(\tilde{I}(t), S(t))$, and $\{(\tilde{I}(t), S(t)), t \geq 0\}$ is a Markov process on the state space

$$\tilde{\mathcal{X}} = \{(i_1, i_2, \ldots, i_k, s) : i_j \in \{0, 1, \ldots, N\}, 1 \leq j \leq k, s \in \{0, 1, \ldots, N\}, \sum_{j=1}^{k} i_j + s \leq N\}.$$

4.2. Final Outbreak Size Distribution with Erlang Distributed Infectious Period

Define τ as the termination time of the disease: the time when no infected individuals remain,

$$\tau = \inf\{t \geq 0 : I(t) = 0\}.$$

Let $\omega = (i_1, i_2, \ldots, i_k)$ be the current state of disease stages. Given that we start with $(\tilde{I}(0), S(0)) = (\omega, s)$, the probability that the final number of recovered individuals equals m when the outbreak ends can be obtained as

$$P_m(\omega, s) = \Pr\{R(\tau) = m | (\tilde{I}(0), S(0)) = (\omega, s)\}.$$

In principle, finding the distribution of the final outbreak size (that is, obtaining $P_m(\omega, s)$ values) is not difficult, as it corresponds to finding absorption probabilities of a Markov process. Note that the process $\{(\tilde{I}(t), S(t))\}$ makes two types of transitions.

Given that the current state is (ω, s), a transition of type-I occurs when the disease stage of an infected individual decreases by one. In this case, the state of the system moves from (ω, s) to $(\omega - e_j + e_{j-1}, s)$ for some $j = 1, 2, \ldots, k$, where e_j is a unit vector of size k with 1 at jth entry, and 0 at other entries. By convention e_0 is a zero vector of size k. A transition from j to $j-1$ occurs at rate $i_j\mu$. A type-I transition changes the state of the disease stages, and the aggregate rate of type-I transition is $\mu \sum_{j=1}^{k} i_j$. Note that $j = 1$ corresponds to a recovery.

Defined similarly, a type-II transition indicates a new infection. The state of the system moves from (ω, s) to $(\omega + e_k, s - 1)$, as a new infection brings k more disease stages, and decreases the number of susceptibles by one. A type-II transition occurs at rate $(\sum_{j=1}^{k} i_j)\lambda s/N$.

We realize that the size of the state space, and hence the computational burden in finding the absorption probability $P_m(\omega, s)$ increases exponentially with k, the number of stages. As it can be observed from the state space $\tilde{\mathcal{X}}$ of $\{(\tilde{I}(t), S(t)), t \geq 0\}$, the memory and computation requirement grows proportional to N^{k+1}, where N is the total population size, and k is the number of disease stages. Therefore, we propose a transformation of the state space in such a way that the computational burden of finding the outbreak size distribution is considerably reduced.

Let $V(t)$ be the total number of disease stages at time t:

$$V(t) = \sum_{j=1}^{k} jI_j(t).$$

Consider the process $\{(V(t), S(t)), t \geq 0\}$. Define,

$$\Pi_m(v, s) = \Pr\{(V(\tau), S(\tau)) = (0, N - m) | (V(0), S(0)) = (v, s)\}.$$

Starting with a total of v disease stages and s many susceptible, $\Pi_m(v, s)$ is the probability that at the end of the outbreak (at time τ), the total number of stages is zero, and the total number of susceptible remaining is $N - m$. The following result is our main finding in this section.

Proposition 1. *For $m = 0, 1, \ldots, N$, we have $\Pi_m(v, s) = P_m(\omega, s)$ where $\omega = (i_1, i_2, \ldots, i_k)$, and $v = \sum_{n=1}^{k} n i_n$.*

Proof. We first note that $V(t) = 0 \iff \tilde{I}(t) = (0, 0, \ldots, 0) \iff I(t) = 0$. Therefore, at the absorption time τ, $V(\tau) = 0$. Moreover, the jumps of the process $\{(\tilde{I}(t), S(t))\}$ correspond to the jumps of the process $\{(V(t), S(t))\}$. For a given current state $(\tilde{I}(t), S(t)) = (\omega, s)$, let $(V(t), S(t)) = (v, s)$, where $v = \sum_{n=1}^{k} n i_n$. If there is a transition of type-I with a move from j to $j - 1$ stages, then $V(t) = v$ jumps to

$$v' = \sum_{n \neq j, j-1} (n i_n) + j(i_j - 1) + (j - 1)(i_{j-1} + 1) = v - j + j - 1 = v - 1.$$

On the other hand, if there is a type-II transition, then $V(t)$ jumps from v to $v + k$, since ω moves to $\omega + e_k$. As this corresponds to a new infection, $S(t)$ moves from s to $s - 1$. Hence, the processes $\{(\tilde{I}(t), S(t))\}$ and $\{(V(t), S(t))\}$ make jumps at the same time, and their absorption times are equivalent. Therefore, starting from equivalent states, they have the same absorption probabilities. □

Proposition 1 enables us to compute the outbreak size distribution using a two-state process instead of a $k + 1$-state process by noting that there are several states in the original process leading to the same state in the transformed process. For example, state $(i_1, i_2, s) = (2, 0, 1)$ and state $(i_1, i_2, s) = (0, 1, 1)$ correspond to the same state

$(v, s) = (2, 1)$ in the transformed process. Let's consider the following state transitions of the original process for state (2, 0, 1),

$$\begin{aligned}(2,0,1) &\rightarrow (1,0,1) \text{ with rate } 2\mu,\\ (2,0,1) &\rightarrow (2,1,0) \text{ with rate } 2\lambda/3,\end{aligned}$$

and for state (0, 1, 1),

$$\begin{aligned}(0,1,1) &\rightarrow (1,0,1) \text{ with rate } \mu,\\ (0,1,1) &\rightarrow (0,2,0) \text{ with rate } \lambda/3.\end{aligned}$$

Although these rates are different, one step transition probabilities are the same for both states that are $\frac{2\mu}{(2\mu+2\lambda/3)} = \frac{\mu}{(\mu+\lambda/3)}$ for recovery and $\frac{2\lambda/3}{(2\mu+2\lambda/3)} = \frac{\lambda/3}{(\mu+\lambda/3)}$ for infection. Thus, the state transitions for state $(v, s) = (2, 1)$ in the transformed process become

$$\begin{aligned}(2,1) &\rightarrow (1,1) \text{ with probability } \tfrac{\mu}{(\mu+\lambda/3)},\\ (2,1) &\rightarrow (4,0) \text{ with probability } \tfrac{\lambda/3}{(\mu+\lambda/3)}.\end{aligned}$$

Moreover, Figure 4.2 displays the state transitions of the process $\{(V(t), S(t))\}$ for $N = 3$ and $k = 2$ where the arrows indicate the possible transitions and the initial state $(V(0), S(0))$=(2, 2) is coloured in black. The state (0, 3) coloured in red corresponds to all individuals being susceptible and having no infected individuals. In what follows, we formally describe the process $\{(V(t), S(t))\}$.

Let $t_1, t_2, \ldots$ be the times of transitions of the process $\{(V(t), S(t))\}$. After the ith transition, let $(V(t_i), S(t_i)) = (v, s)$ be the current system state with $v = \sum_{n=1}^{k} n i_n$. Then, if the next transition is of Type-I,

$$\begin{aligned}\Pr\{(V(t_{i+1}), S(t_{i+1})) = (v-1, s) | (V(t_i), S(t_i)) = (v, s)\} &= \frac{\mu \sum_{j=1}^{k} i_j}{\mu \sum_{j=1}^{k} i_j + \frac{s\lambda}{N} \sum_{l=1}^{k} i_l}\\ &= \frac{N\mu}{N\mu + s\lambda}. \qquad (4.1)\end{aligned}$$

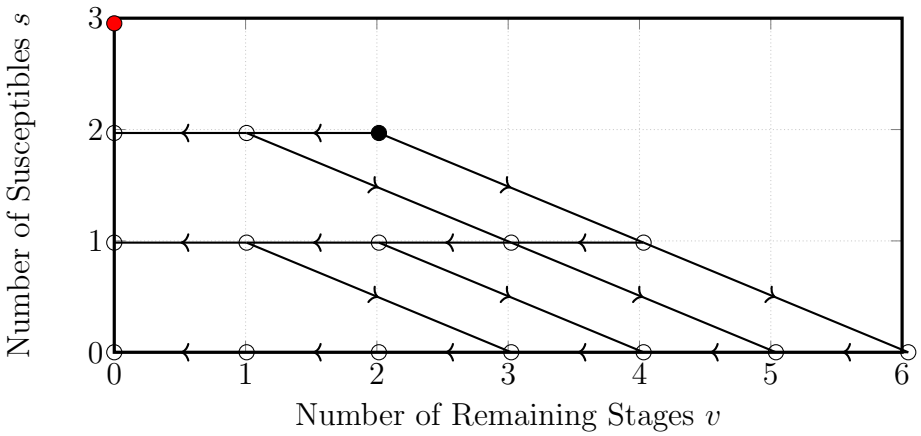

Figure 4.2. The state transitions of process $\{(V(t), S(t))\}$ for $N = 3$ and $k = 2$

If the next transition is of Type-II,

$$\Pr\{(V(t_{i+1}), S(t_{i+1})) = (v + k, s - 1) | (V(t_i), S(t_i)) = (v, s)\} = \frac{s\lambda}{N\mu + s\lambda}. \qquad (4.2)$$

Neither of the Equations 4.1 or 4.2 depend on the total number of disease stages v or the total number of infected cases $\sum_{j=1}^{k} i_j$. Therefore, the imbedded process $\{(V(t_i), S(t_i)), i = 0, 1, \ldots\}$ is a discrete time Markov chain on

$$\mathcal{X} = \{(v, s) : v \in \{0, 1, \ldots, kN\}, s \in \{0, 1, \ldots, N\}, \frac{v}{k} + s \leq N\}.$$

with transition probabilities given by (4.1) and (4.2). Note that $\{(V(t_i), S(t_i)), i = 0, 1, \ldots\}$ is not a continuous time Markov chain, as the knowledge of $(V(t), S(t))$ is not sufficient to describe changes (particularly, rates of changes) in the process. We can only describe the jumps and the probabilities of jumps of the process $\{(V(t), S(t))\}$. Although this is sufficient to obtain exact outbreak size distribution, more detailed statistics of the process $\{I(t), t \geq 0\}$, such as $\{\max I(t), 0 \leq t \leq \tau\}$ and $E[\tau]$ cannot be obtained from the imbedded Markov chain.

In order to obtain the final outbreak size distribution, we first note the boundary

conditions for $\Pi_m(v,s)$:

$$\Pi_m(v,s) = 0 \text{ for } s = 0,1,\ldots,(N-m-1),\ v = 0,1,\ldots,k(N-s), \quad (4.3)$$

$$\Pi_m(0,N-m) = 1. \quad (4.4)$$

Observe that $v = 0$ corresponds to the case where there are no infections (the extinction of the disease). Equation 4.3 follows, since if there are less than $N-m$ susceptible cases, the final outbreak size has already exceeded m so it cannot be m. Similarly, Equation 4.4 follows, since if $s = N-m$, and if there are no infections, the final outbreak size must be m. By conditioning on the first step (Durrett, 1999), after state (v,s) for $v = 0,1,\ldots$, and $s = 0,1,\ldots,N$ we have

$$\Pi_m(v,s) = \frac{\lambda s}{\lambda s + \mu N}\Pi_m(v+k, s-1) + \frac{\mu N}{\lambda s + \mu N}\Pi_m(v-1,s). \quad (4.5)$$

Considering boundary conditions and (4.5) for $s = N-m$, we obtain:

$$\Pi_m(v,s) = \left(\frac{\mu N}{\lambda s + \mu N}\right)^v. \quad (4.6)$$

After using (4.5) iteratively, and by conditioning on the first time a new infection occurs for $s > N-m$, we obtain:

$$\Pi_m(v,s) = \sum_{j=0}^{v-1}\left(\frac{\mu N}{\lambda s + \mu N}\right)^j \frac{\lambda s}{\lambda s + \mu N}\Pi_m(v-j+k, s-1). \quad (4.7)$$

The recursive method to calculate the final outbreak size distribution for the models with Erlang distributed infectious periods is given by the algorithm in Figure 4.3. For a given m, we first state our boundary conditions (given by (4.3) and (4.4)) in steps 2 to 6. Starting with $s = N-m$, we calculate $\Pi_m(v,s)$ using (4.6) in steps 7 to 9 and by increasing the number of susceptible cases one by one, we continue to calculate $\Pi_m(v,s)$ using (4.7) in steps 10 to 15.

Algorithm 4

1. *Set* $m = 1$
2. **if** $m < N$ **then**
3. *Set* $\Pi_m(v, s) = 0$ *for* $s = 0, 1, \ldots, (N - m - 1)$ *and* $v = 0, 1, \ldots, k(N - s)$
4. **end if**
5. Set $s = N - m$
6. Set $\Pi_m(0, s) = 1$
7. **for** $v = 1, 2, \ldots, k(N - s)$
8. *Compute* $\Pi_m(v, s)$ using Equation 4.6
9. **end for**
10. Set $s = s + 1$. If $s < N$ continue. Otherwise, go to Step 16.
11. Set $\Pi_m(0, s) = 0$
12. **for** $v = 1, 2, \ldots, k(N - s)$
13. *Compute* $\Pi_m(v, s)$ using Equation 4.7
14. **end for**
15. *Set* $s = s + 1$. If $s < N$ go to Step 11. Otherwise, go to Step 16.
16. *Set* $m = m + 1$. If $m \leq N$ go to step 2. Otherwise, stop the algorithm.

Figure 4.3. Exact final size distribution for Erlang distributed infectious period (k, μ)

In order to demonstrate the usefulness of the expression given in (4.5), we present a comparison with the recursion of Ball (1986). Using the expression given in Andersson and Britton (2012), the probability P_l^{N-m} that for an epidemic starting with $N - m$ susceptible the final size equals to l for $0 \leq l \leq N - m$ can be computed using the equations:

$$\sum_{j=0}^{l} \binom{N-m-j}{l-j} P_j^{N-m} / [\Phi(\lambda(N-m-l)/(N-m))]^{j+m} = \binom{N-m}{l}, \qquad (4.8)$$

where m is the initial number of infected and $\Phi(x)$ is the moment generating function of infectious period.

As the system of equations in 4.8 is triangular, the final outbreak size distribution can be calculated recursively. However, the solution of the set of equations using 4.8 is not stable even for population sizes as small as 100, and negative probabilities and probabilities greater than one occur frequently in the results. Some numerical results

for direct comparison of (4.8) and (4.5) are displayed in Table 4.1. Equation 4.8 and Equation 4.5 yield the same results for $N = 50$ but implementation of (4.8) has problems for larger population sizes. Equation 4.8 yields erroneous results starting from $m = 18$ for $N = 100$ and $m = 10$ for $N = 500$.

Table 4.1. Comparison of final size probability values for different population sizes for k=5

		Calculated Probability Value	
N	Final Size Value	Implementation of (4.8)	Implementation of (4.5)
50	1	0.276	0.276
	20	0.010	0.010
	40	0.013	0.013
	50	3.631e-06	3.631e-06
100	18	-0.006	0.003
	19	0.036	0.003
	20	-0.104	0.003
	100	-0.203	2.841e-11
500	10	-0.001	0.005
	11	0.135	0.004
	12	-2.392	0.004
	500	NaN	2.738e-50

The reason for the numerical instability of (4.8) is the large binomial coefficients in the equations. As the solution of these equations includes differences of large positive numbers, catastrophic cancellation occurs (see Goldberg (1991)). Note that the numeric instability of (4.8) is also mentioned in Andersson and Britton (2012). In practice it is often suggested to use simulation to compute the final outbreak size distribution for large N or to use asymptotic approximations. However, using (4.5), we can compute the exact distribution of the final outbreak size for population sizes larger than 2000.

As it has been emphasized before, for the SI_kR model, modelling the system

using $\{(\tilde{I}(t), S(t))\}$ results in a state space with N^{k+1} states, and computing the final outbreak size distribution requires extensive time and memory. For instance, with a population size $N = 100$ and $k = 5$, the model requires a storage capacity for 100^6 states, and is therefore not implementable. However, if we use the number of stages transformation, $V(t)$, then the number of states grows proportional to $(kN)^2$ which is considerably more efficient than the original model. Assuming there is one initially infected individual, the computation time required (in seconds) to find the exact final outbreak size distribution for different population sizes N and for different k values are given in Table 4.2 (calculations are performed on a 3.6 GHz PC using R). A large size problem with $N = 2000$ and $k = 10$ can be computed in not much more than an hour. As we can see in Table 4.2 the execution time increases approximately linear with the number of Erlang stages k and quadratic with the population size N.

Table 4.2. Time required to calculate exact final outbreak size distribution (in seconds).

	Population Size N			
k	**100**	**500**	**1000**	**2000**
2	2.02	48.92	189.14	797.44
5	5.10	123.89	482.89	2054.22
10	10.45	254.35	1022.08	4065.32
20	21.07	510.28	2032.92	8031.63

It should be noted that the execution time increases linearly with k and quadraticly with N for our recursive method, while the execution time for naive Monte Carlo simulation increases linearly with the number of repetitions. Therefore, naive Monte Carlo simulation is expected to be faster for large size problems with $N > 5000$. However, we have difficulties in calculating the probabilities of rare events efficiently via simulation while our method enables us to calculate very small tail probabilities with high efficiency (see results based on implementation of (4.5) in Table 4.1). Although we do not claim to completely replace the need for simulation in applied modelling to obtain approximate results, our proposed method can be used for large size populations

to obtain exact and numerically stable results with higher precision.

4.2.1. Extension to Mixtures of Erlang Distributed Infectious Period

The Markov SIR model we developed for Erlang distributed infectious period is also applicable to problems with more general infectious period. A mixture of Erlang distributions can be described as follows. Consider a model where with probability α_i an infected individual remains infected for an Erlang distributed period with shape parameter k_i, $i = 1, 2, \ldots, b$, where $\sum_{i=1}^{b} \alpha_i = 1$. In this case, it is easily seen that (4.7) becomes:

$$\Pi_m(v, s) = \sum_{j=0}^{v-1} \left(\frac{\mu N}{\lambda s + \mu N} \right)^j \frac{\lambda s}{\lambda s + \mu N} \sum_{i=1}^{b} \alpha_i \Pi_m(v - j + k_i, s - 1). \tag{4.9}$$

Suppose that we only have the knowledge of the mean (μ_T) and the standard deviation (σ_T) of the infectious period, T, so that its coefficient of variation is equal to $c_v = \sigma_T/\mu_T < 1$. In this case, we can construct a mixture of Erlang distributions matching these empirical parameters. Let k be such that $c_v^2 \in [1/(k-1), 1/k]$. Then, we can take an infectious period which takes Erlang$(k-1, \mu)$ and Erlang(k, μ), with respective probabilities α, and $1 - \alpha$, where (see Adan and Resing (2002))

$$\alpha = \frac{1}{1 + c_v^2} (k c_v^2 - \{k(1 + c_v^2) - k^2 c_v^2\}^{1/2}), \tag{4.10}$$

$$\mu = \frac{k - \alpha}{\mu_T}. \tag{4.11}$$

In this case Equation 4.9 simplifies to

$$\begin{aligned} \Pi_m(v, s) &= \sum_{j=0}^{v-1} \left(\frac{\mu N}{\lambda s + \mu N} \right)^j \frac{\lambda s}{\lambda s + \mu N} \{\alpha \Pi(v - j + k - 1, s - 1) \\ &+ (1 - \alpha) \Pi(v - j + k, s - 1)\}. \end{aligned}$$

Let's consider the discrete infectious period distribution in the study of Longini

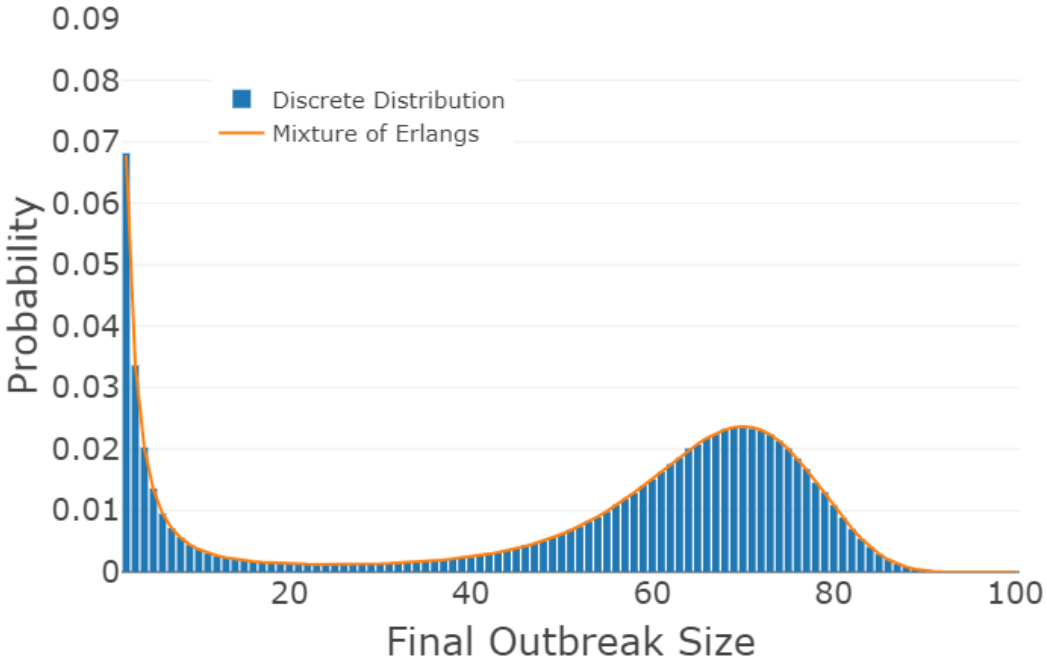

Figure 4.4. Comparison of final size distributions for discrete infectious period and mixture of Erlangs distributed infectious period

et al. (2004) with mean 4.1 and variance 0.89, so the coefficient of variation for this infectious period is equal to 0.23 (Longini *et al.*, 2004). Using mixture of Erlangs, we can find that for $k = 19$ we have $0.23^2 \in [\frac{1}{k}, \frac{1}{k-1}]$. Then,

$$\begin{aligned} \alpha &= \frac{1}{1+0.23^2}(19(0.23^2) - \{19(1+0.23^2) - (19^2)(0.23^2)\}^{1/2}) = 0.0495, \\ \mu &= \frac{19-\alpha}{4.1} = 4.622. \end{aligned}$$

Thus, it is possible to calculate the cumulative distribution of the final outbreak size for the mean infectious period 4.1 and the variance 0.89 with $R_0 = 1.69$ by mixing Erlangs with given α and μ values.

In Figure 4.4, we compare the final outbreak size distribution for discrete infectious period obtained by simulation with the exact final outbreak size distribution for mixture of Erlangs distributed infectious period with the same mean and variance.

4.3. Distribution of The Maximum Number of Disease Stages and Approximation to Maximum Epidemic Size Distribution

Another stochastic variable of interest is the maximum epidemic size, as the treatment resources required for an epidemic are proportional to this maximum size. Actually, intervention methods and control strategies aim to reduce the peak epidemic size. Define

$$I_{\max} = \max_{0 \le t \le \tau} I(t)$$

as the maximum number of infected cases, or the peak epidemic size, until the extinction of the disease (τ). Since finding the distribution of $I_{\max}$ is computationally prohibitive (as argued in Section 2), we again use the auxiliary process of total disease stages, and define

$$V_{\max} = \max_{0 \le t \le \tau} V(t).$$

Finding the distribution of V_{max} is also important since it is useful for measuring the severity of an epidemic and for appraising the effect of control methods. In what follows we provide a procedure for finding the distribution of $V_{\max}$. Let $Q_m(v, s)$ be the probability that $V_{\max} = m$ at the end of the epidemic when we start with $(V(0), S(0)) = (v, s)$. Clearly, we have

$$Q_m(v, 0) = 1, \text{ for } v = m \tag{4.12}$$

$$Q_m(v, 0) = 0, \text{ for } v \neq m \tag{4.13}$$

$$Q_m(0, s) = 1, \text{ for } m = 0 \tag{4.14}$$

$$Q_m(0, s) = 0, \text{ for } m \neq 0 \text{ and} \tag{4.15}$$

$$Q_m(v, s) = 0 \text{ for } m < v \text{ and } v + sk < m. \tag{4.16}$$

For other values of (v, s), by using the first step analysis (as in Equation 4.5),

$$Q_m(v,s) = \frac{\lambda s}{\lambda s + \mu N} Q_m(v+k, s-1) + \frac{\mu N}{\lambda s + \mu N} Q_m(v-1, s). \tag{4.17}$$

The procedure to calculate the distribution of V_{max} is given by the algorithm in Figure 4.5. To calculate $Q_m(v, s)$, we first state our boundary conditions (4.12) and (4.13) in steps 1 to 5 and (4.14) and (4.15) in steps 8 to 9. The boundary condition (4.16) is stated in 11, and $Q_m(v, s)$ is computed using Equation 4.17 in steps 12 to 14. Lastly, $Q_m(v, s)$ for $m = v$ is computed in step 15. Thus, $Q_m(v, s)$ values are computed recursively by first increasing v (step 10) and then by increasing s (step 17). The execution time for Algorithm 5 increases linearly with k and quadraticly with N so it is faster than Monte Carlo simulation for reasonably large size populations.

Algorithm 5

1. *Set* $s = 1$
2. **for** $v = 0, 1, \ldots, kN$
3. *Set* $Q_m(v, s) = 0$ for $m = 0, 1, \ldots, kN$ and $m \neq v$
4. *Set* $Q_m(v, s) = 1$ for $m = v$
5. **end for**
6. *Set* $s = 1$
7. *Set* $v = 0$
8. *Set* $Q_m(v, s) = 0$ for $m = 1, \ldots, kN$
9. *Set* $Q_m(v, s) = 1$ for $m = 0$
10. *Set* $v = v + 1$
11. *Set* $Q_m(v, s) = 0$ for $m = 0, \ldots, v-1$ and $m = v + ks + 1, \ldots, k(N - s)$
12. **for** $m = v + 1, v + 2, \ldots, v + ks$
13. *Compute* $Q_m(v, s)$ using Equation 4.17
14. **end for**
15. *Set* $Q_m(v, s) = 1 - \sum_{m=v+1}^{v+ks} Q_m(v, s)$ for $m = v$
16. *Set* $v = v + 1$. If $v \leq k(N - s)$ go to step 11. Otherwise, go to Step 17.
17. *Set* $s = s + 1$. If $s \leq N - 1$ go to step 7. Otherwise, if $s = N$ stop the algorithm.

Figure 4.5. Distribution of $V_{\max}$ for Erlang(k, μ) distributed infectious period

The distribution of $I_{\max}$ cannot be obtained from the distribution of $V_{\max}$, but there exists a relationship between $I_{\max}$ and $V_{\max}$. Similar to the study of Watson (1980)

where a random scale transformation is used to approximate the size distribution and to estimate the critical R_0, we consider a transformation of time scale depending on the current number of infectives $I(t)$ and the current number of remaining stages $V(t)$. Thus, we have the following.

Proposition 2. *Define* $r_I = E[\int_0^\tau I(t)dt]/E[\tau]$ *as the average number of infected per unit time, and similarly define* $r_V = E[\int_0^\tau V(t)dt]/E[\tau]$ *as the average number of infection stages per unit time. Then,* $r_I = \frac{2}{k+1} r_V$.

Proof. First note that $\int_0^\tau I(t)dt = \sum_{i=1}^{R(\tau)} T_i$, where $R(\tau)$ is the number of recovered infections until extinction, and T_i is the infectious period of the i^{th} infected individual. $T_1, T_2, \ldots$ are independent and identically distributed random variables. We note that for any $j = 1, 2, \ldots$ the event $\{R(\tau) \leq j\}$ only depends on $\{T_1, T_2, \ldots, T_j\}$, or specifically $\{R(\tau) \leq j\}$ is independent of $\{T_{j+1}, T_{j+2}, \ldots\}$. Then, by Wald's Equation (Ross *et al.*, 1996, Theorem 3.3.2):

$$E\left[\int_0^\tau I(t)dt\right] = E\left[\sum_{i=1}^{R(\tau)} T_i\right] = E[R(\tau)]\frac{k}{\mu}.$$

Similarly,

$$\int_0^\tau V(t)dt = \sum_{i=1}^{R(\tau)} \int_0^\tau V_i(t)dt,$$

where $\{V_i(t), 0 \leq t \leq T_i\}$ is the number of stages present in the system due to i^{th} infected individual during its infectious. Using similar arguments as above, we have

$$\begin{aligned} E\left[\int_0^\tau V(t)dt\right] &= E[R(\tau)]E\left[\int_0^{T_{i,1}} k dt + \int_{T_{i,1}}^{T_{i,1}+T_{i,2}} (k-1)dt + \cdots \int_{T_{i,1}+\ldots T_{i,k-1}}^{T_{i,1}\ldots T_{i,k}} 1 dt\right] \\ &= E[R(\tau)]\left(\frac{k}{\mu} + \frac{k-1}{\mu} + \cdots + \frac{1}{\mu}\right) \\ &= E[R(\tau)]\frac{k(k+1)}{2\mu} \end{aligned}$$

where $T_{i,j}$ is the time spent at stage $j, 1 \leq j \leq k$ by the i^{th} infected individual . □

Note that for $k = 1$, $V(t)$ and $I(t)$ are identical processes and implementation of Algorithm 5 gives the exact maximum size distribution for an exponential infectious period. Further, using Algorithm 5, we can find $E[V_{\max}] = \sum_{m=v}^{N} mQ_m(v, s)$, and using Proposition 2, we can approximate $E[I_{\max}]$ as:

$$E[I_{\max}] \approx \frac{2}{k+1} E[V_{\max}]. \tag{4.18}$$

The comparison of $E[I_{max}]$, $E[V_{max}]$ and $\frac{2}{k+1}E[V_{max}]$ for different combinations of k and R_0 is presented in Table 4.3. We estimate $E[I_{max}]$ via simulation with 10000 replications for $N = 1000$ and exactly calculate the expectation of V_{max} using the distribution of V_{max} obtained by implementing Algorithm. Table 4.3 indicates that we can estimate the expectation of the maximum number of simultaneously infected individuals using V_{max} distribution. We also observe that the percentage error calculated as

$$\frac{\frac{2}{k+1}E[V_{max}] - E[I_{max}]}{E[I_{max}]}$$

becomes higher for large k values and our approximation $\frac{2}{k+1}E[V_{max}]$ is greater than $E[I_{max}]$ when k is greater than 1. We also decide to check the relationship between the distribution of I_{max} and the distribution of V_{max} to find an approximation for the distribution of I_{max} and compare them empirically in Section 4.4.3.

4.4. Numerical Results

We present several numerical results to illustrate the precision of our method to compute the final outbreak size distribution and the performance of our maximum epidemic size approximation. All calculations are performed using R. We emphasize that our results are exact (not simulation output) and obtained using the algorithms

Table 4.3. $E[V_{max}]$, $\frac{2}{k+1}E[V_{max}]$ and $E[I_{max}]$ for different k and R_0 values

R_0	k	$E[V_{max}]$	$\frac{2}{k+1}E[V_{max}]$	$E[I_{max}]$	Error
1.5	1	27.56	27.56	27.50	0.21%
	2	66.42	44.28	42.29	4.70%
	3	107.48	53.74	50.60	6.20%
	5	191.53	63.84	59.62	7.07%
2	1	84.09	84.09	84.30	0.24%
	2	204.01	136.01	132.28	2.81%
	3	329.36	164.68	158.66	3.79%
	5	584.20	194.73	183.13	6.33%
2.5	1	147.77	147.77	148.39	0.41%
	2	353.18	235.45	229.20	2.72%
	3	565.26	282.63	271.11	4.25%
	5	993.53	331.17	311.06	6.46%

in Figure 4.3 or Figure 4.5.

Numerical scenarios have population sizes $N = 100$, $N = 1000$ and $N = 2000$. Let R_0 denote the expected number of infections caused by one infected individual in a totally susceptible population. For our models, R_0 is calculated as follows

$$R_0 = \frac{\lambda(N-1)}{N}\frac{k}{\mu} \approx \frac{\lambda k}{\mu}.$$

In this section, we first observe the effect of infectious period variability on the final outbreak size distribution. Then, we calculate the outbreak probability by assuming infinite population size and show that it is possible to obtain outbreak probability using final outbreak size distribution and checking the plateau in the cumulative final outbreak size graphs. Moreover, we compare our approximation for the maximum outbreak size distribution with simulation results and observe how k affects our approximation.

4.4.1. Effect of Infectious Period Variability

First, we compare the probabilistic behavior of final outbreak size for $k = 1, 2, 5$ and 10. To show the precision of our method, we consider $N = 2000$. It is assumed that $R_0 = 1.5$ and mean infectious period is $k/\mu = 0.5$ for all cases. There is only one initially infected individual so the process starts with $(V(0), S(0)) = (k, N - 1)$. Numerical descriptors of the final outbreak size distribution as in the study of Amador and Lopez-Herrero (2017) are: the probability of having a single infection, the probability that all the individuals get the infection, the median, the mean and the standard deviation for the final outbreak size. Thus, these results are displayed in Table 4.4.

Table 4.4. Numerical descriptors for final outbreak size distribution with $R_0 = 1.5$ and $\lambda = 3$ for different k values

$R_0 = 1.5$	$k = 1, \mu = 2$	$k = 2, \mu = 4$	$k = 5, \mu = 10$	$k = 10, \mu = 20$
Var. of Infec. Period	0.250	0.125	0.050	0.025
$\Pr\{R(\tau) = 1\}$	0.400	0.327	0.269	0.247
$\Pr\{R(\tau) = 2000\}$	4.751e-167	1.878e-188	9.120e-206	1.178e-212
Median of $R(\tau)$	3	5	996	1071
$E[R(\tau)]$	385.714	492.501	589.215	630.033
$\sigma_{R(\tau)}$	547.199	574.847	582.115	580.338

Notice that, as the variance of infectious period decreases, the probability of having a mild outbreak with a single infected decreases while the expected final outbreak size increases. Thus, more people are expected to be infected for larger k values. This is an important observation since in practice Markov disease models assume often $k = 1$, and therefore may underestimate the threat of epidemic to society by underestimating the expected outbreak size, and eventually overestimating the probability of no outbreak. Furthermore, we deliberately keep $\Pr\{R(\tau) = 2000\}$, the probability that the full population is infected in the table, as obtaining this probability through simulation with a reasonable degree of confidence, would be very difficult, if not impossible.

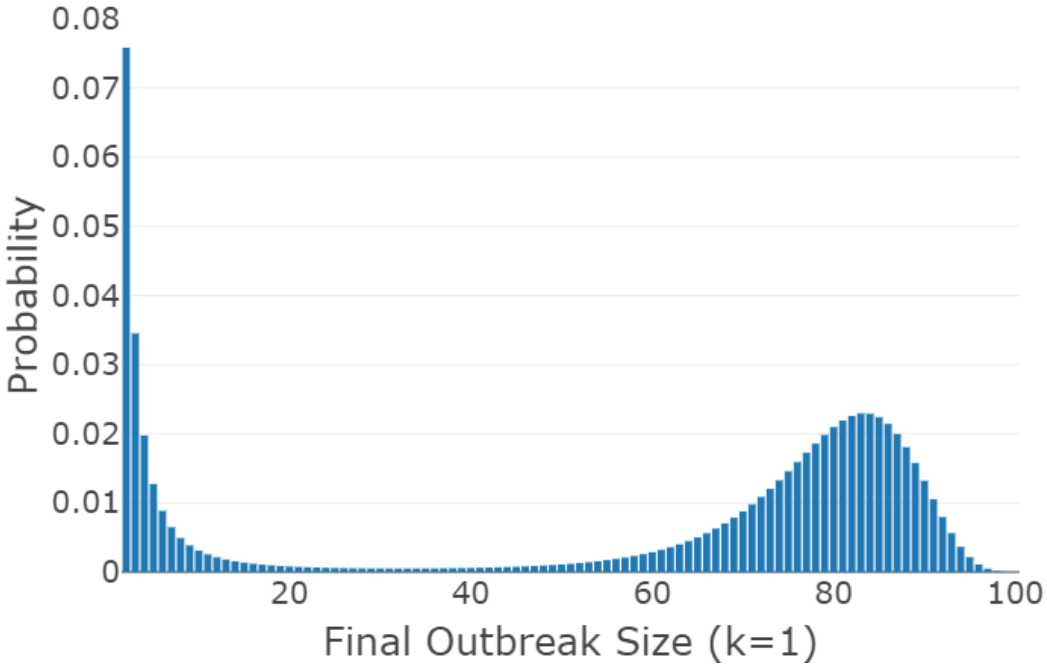

Figure 4.6. Final outbreak size distribution for $k = 1$ based on implementation of algorithm in Figure 4.3

Next, we describe the distribution of final outbreak size to assess the influence of k on disease spread behaviour. We implement Algorithm 4 for $N = 100$, $R_0 = 2$, $I(0) = 1$ and $k = 1, 2, 5$ and 10 and have consistent results with the empirical final outbreak size distribution obtained by Anderson and Watson (1980).

In Figures 4.6, 4.7, 4.8, and 4.9, $P(R(\tau) = 1)$ is not displayed for the sake of scaling since this probability is much greater than 0.08. Moreover, we display how the final outbreak size distribution changes with k noting that the mean of the infectious period ($k/\mu = 1$) is the same for all four cases but the variance of the infectious period decreases with increasing k. Figures 4.6, 4.7, 4.8, and 4.9 show that larger k values lead to a greater final outbreak size with higher probability. This is in agreement with the results summarized in Table 4.4. We also observe that the disease affects either a small proportion of the population or a significant proportion of the population. Given that a significant proportion of the population is affected by the disease, the variance of the final outbreak size distribution becomes smaller as k increases.

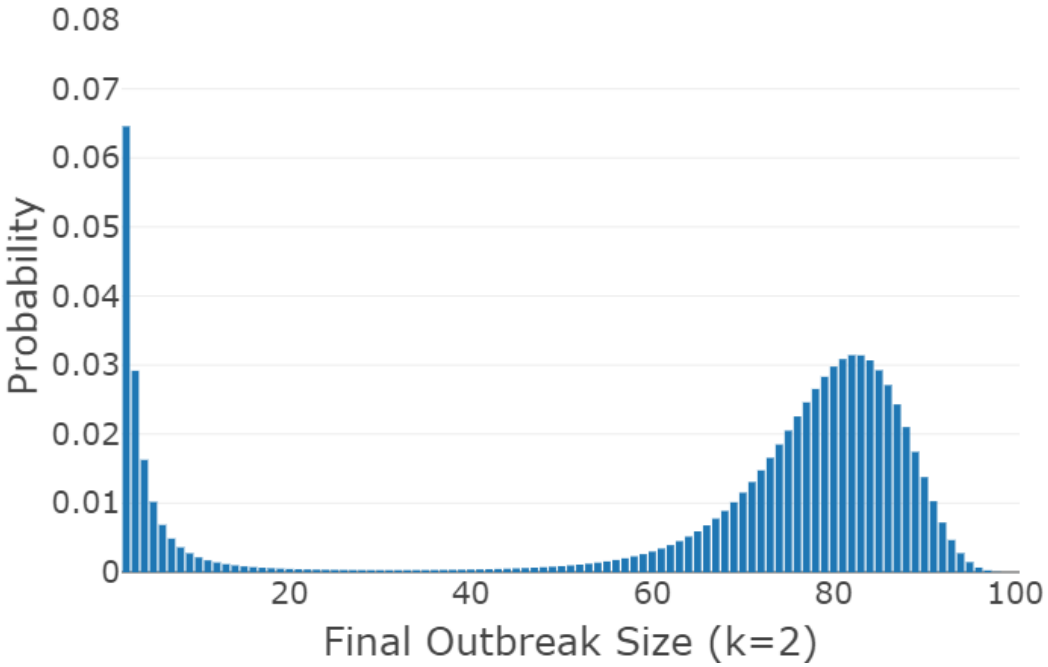

Figure 4.7. Final outbreak size distribution for $k = 2$ based on implementation of algorithm in Figure 4.3

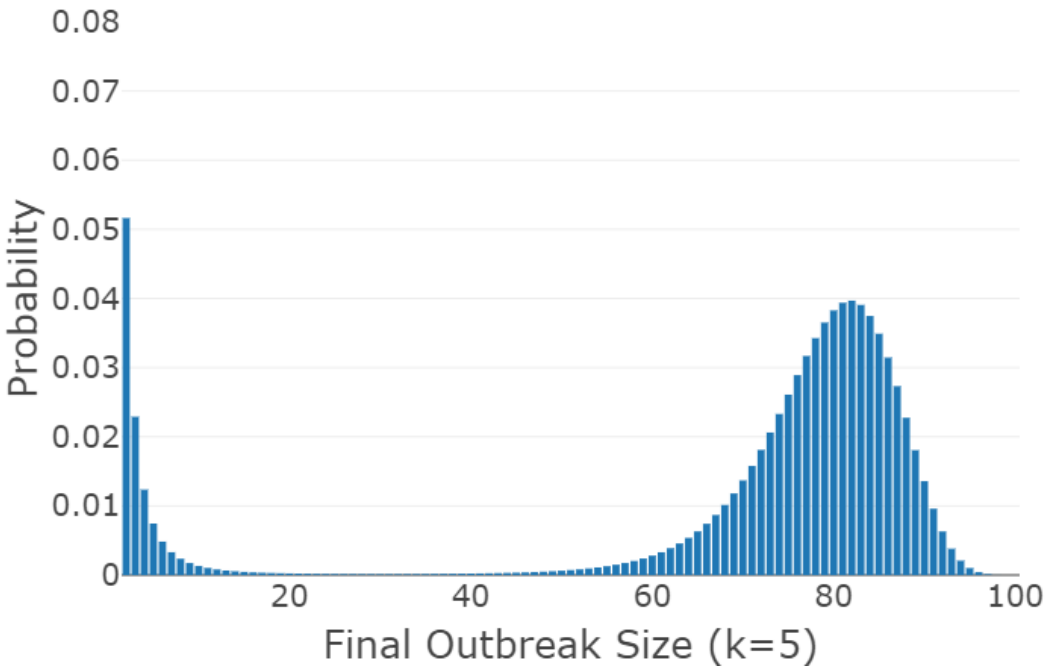

Figure 4.8. Final outbreak size distribution for $k = 5$ based on implementation of algorithm in Figure 4.3

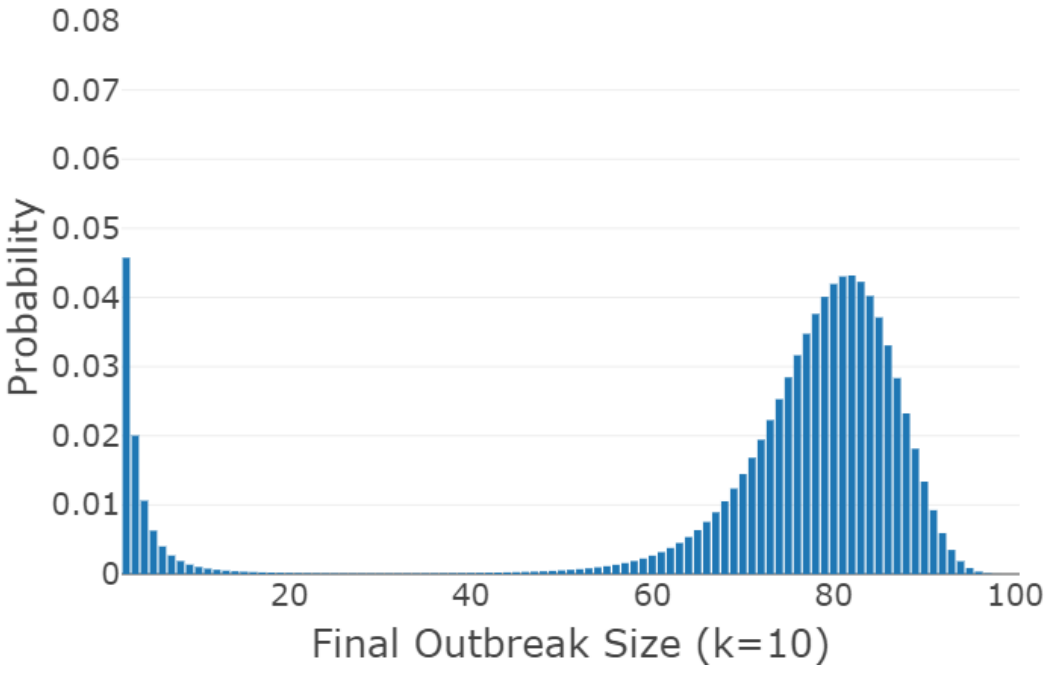

Figure 4.9. Final outbreak size distribution for $k = 10$ based on implementation of algorithm in Figure 4.3

4.4.2. An Infinite Population Approximation for the Outbreak Probabilities

It is important to know whether or not an outbreak will occur. In a deterministic and infinite population disease spread model R_0 exceeding one ensures an outbreak (Diekmann and Heesterbeek, 2000). For stochastic finite population models, an outbreak probability cannot be calculated exactly since it is not clear what percentage of the population should be infected in order an outbreak is to be declared. Assuming an infinite population size, branching process approximation is useful in predicting an outbreak (Allen and van den Driessche, 2013). In what follows we approximate the outbreak probability by considering the method proposed in the study of Andersson and Britton (2012).

Let X be the number of individuals infected by a single infected during an Erlang distributed infectious period T. Noting that as N goes to infinity, $\lambda Ts/N = \lambda T(N - i)/N$ goes to λT we write the generating function of X as

$$E[u^X] \quad = \quad E[e^{-\lambda T(1-u)}]. \tag{4.19}$$

Since T has an Erlang distribution with parameters k and μ,

$$E[u^X] = \left(\frac{\mu}{(1-u)\lambda+\mu}\right)^k.$$

The extinction probability of a branching process with one initial seed (infected) is given by (4.20) as the smallest solution u different than 1.

$$\left(\frac{\mu}{(1-u)\lambda+\mu}\right)^k = u. \tag{4.20}$$

Therefore, it is plausible that $1-u$ can be used as an approximation of the outbreak probability. Table 4.5 shows the computed u and $1-u$ values for $N=100$, $R_0=\lambda=1.5, 2$, and 2.5 and $k=1,2,5$ and 20

Table 4.5. Comparison of outbreak probability from final outbreak size distribution and infinite population approximation

R_0	k	Plateau Value from the Figure	u	1-Plateau Value	$1-u$
1.5	1	0.669	0.667	0.331	0.333
	2	0.576	0.575	0.424	0.425
	5	0.485	0.484	0.515	0.516
	10	0.431	0.431	0.569	0.569
2	1	0.502	0.500	0.498	0.500
	2	0.382	0.381	0.618	0.619
	5	0.283	0.282	0.717	0.718
	10	0.244	0.244	0.756	0.756
2.5	1	0.402	0.400	0.598	0.600
	2	0.275	0.271	0.725	0.729
	5	0.180	0.179	0.820	0.821
	10	0.143	0.143	0.857	0.857

Since we calculate the final outbreak size distribution, we can also calculate the outbreak probability. For better understanding of an outbreak probability, Figures 4.10, 4.11, 4.12, and 4.13 indicate the cumulative final outbreak size distribution for

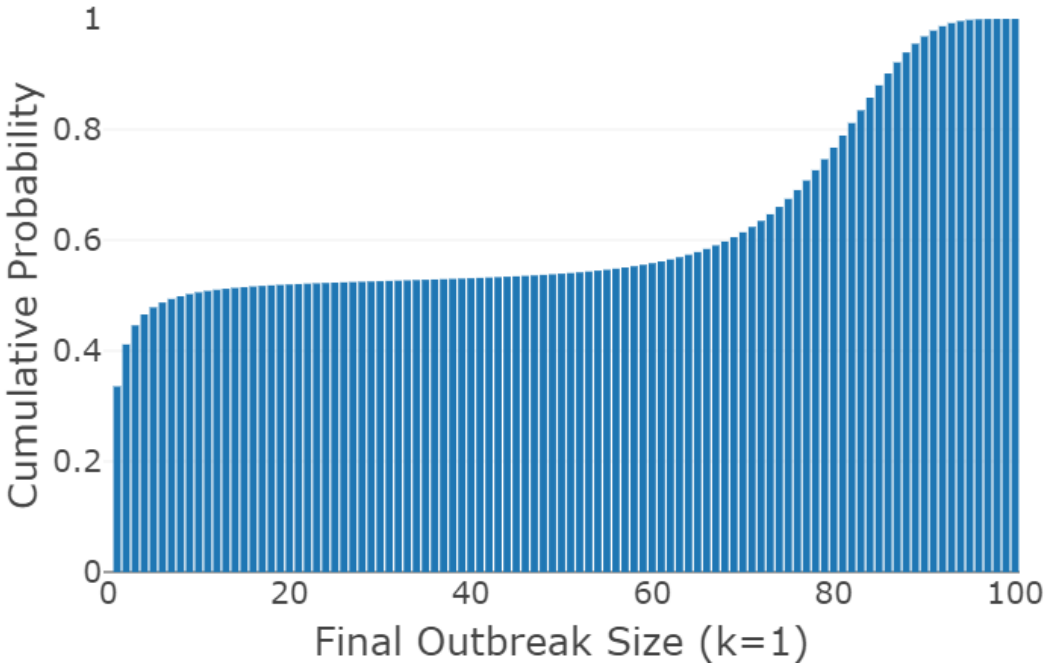

Figure 4.10. Cumulative final outbreak size distribution for $k = 1$

$R_0 = 2$. In the figures, the cumulative probability for the final outbreak size makes a jump at small values of number of infected, and then stays almost flat (the plateau of the distribution), and finally converges to one (faster or slower, depending on the disease parameters) after a large number of infected individuals.

As we observe in Figures 4.10, 4.11, 4.12, and 4.13, the cumulative probability corresponding to the plateau in the figures is very close to u, the solution of (4.20). Consequently, $1-u$ (approximate outbreak probability) and one minus the cumulative probability at the plateau are very close. We present these results in Table 4.5.

4.4.3. Effect of k on Approximation to Maximum Epidemic Size Distribution

Lastly, to observe the relationship between the distribution of I_{max} and the distribution of V_{max}, we employ the approximation in (4.18). The probability mass function of I_{max} is obtained via simulation, and the probability mass function of V_{max} is obtained by using the algorithm in Figure 4.5. We present our findings in Figures 4.14, 4.15, 4.16, and 4.17 for $k = 1, 2, 5$ and 10 respectively.

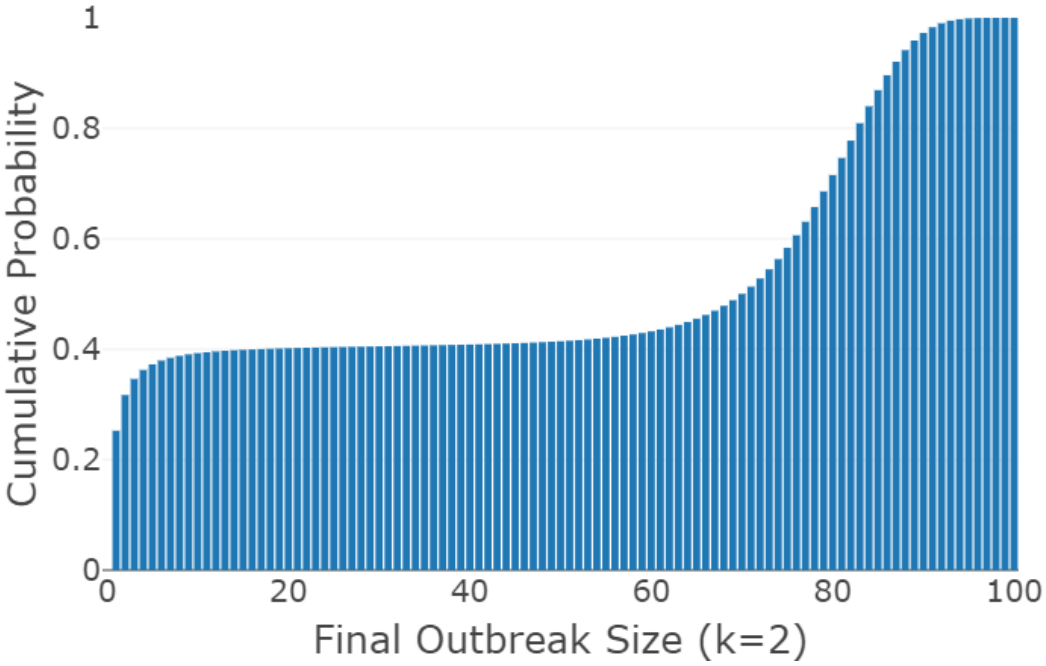

Figure 4.11. Cumulative final outbreak size distribution for $k = 2$

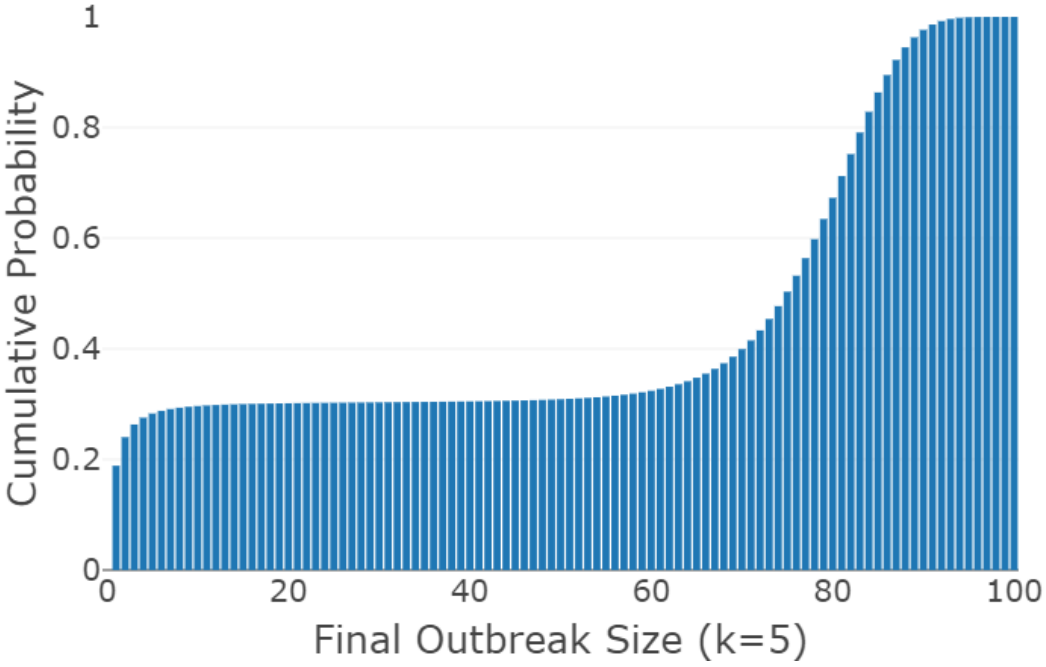

Figure 4.12. Cumulative final outbreak size distribution for $k = 5$

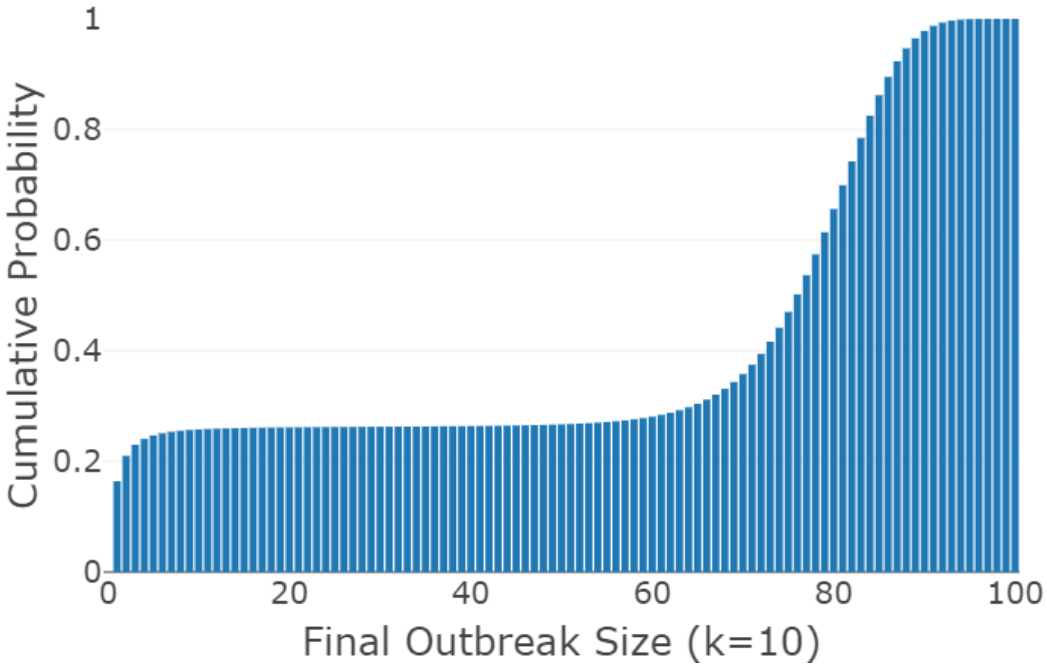

Figure 4.13. Cumulative final outbreak size distribution for $k = 10$

For the parameters that we used, the distribution of maximum number of simultaneously infected individuals can be approximated using the distribution of maximum disease stages. Figures 4.14, 4.15, 4.16, and 4.17 display that the approximation is getting worse as k increases.

4.5. Discussion

For a stochastic SIR model with an Erlang distributed infectious period, we studied the distribution of the total number of recovered individuals and the distribution of the maximum number of individuals who are simultaneously infected until the end of the disease.

We obtained recursive algorithms for the distribution of the final outbreak size and the distribution of the maximum number of stages until the end of the disease. Our algorithms can be implemented for a large size populations. Our state transformation enabled us to treat an Erlang-distributed infectious periods as a simple exponential. We also presented results showing that our recursive algorithms can be implemented for the infectious periods distributed as a mixture of Erlangs. Later, we examined

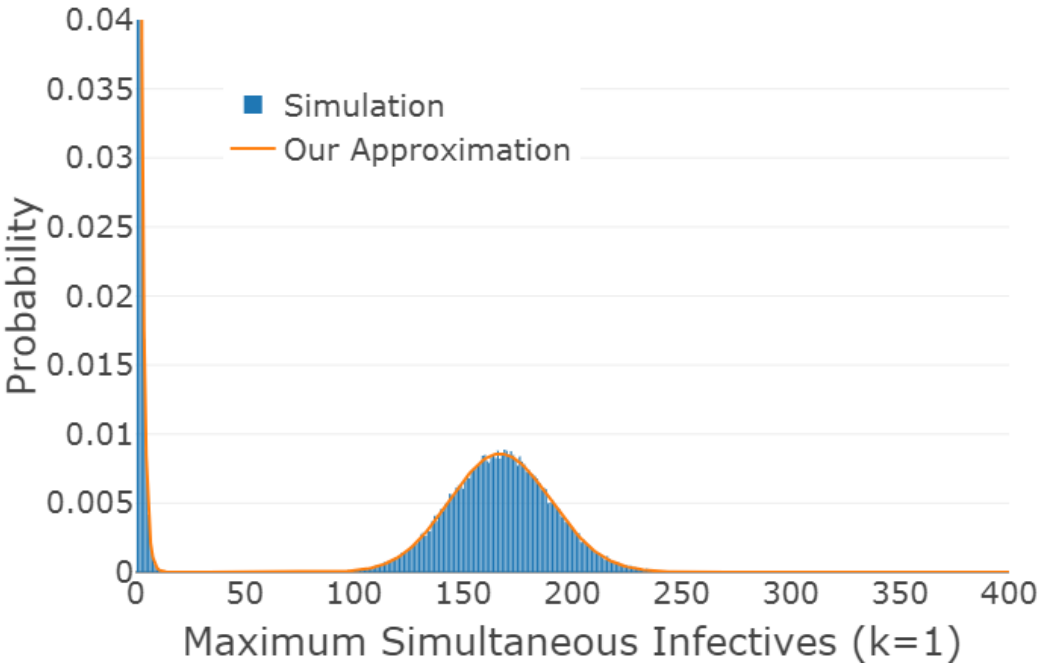

Figure 4.14. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 1$

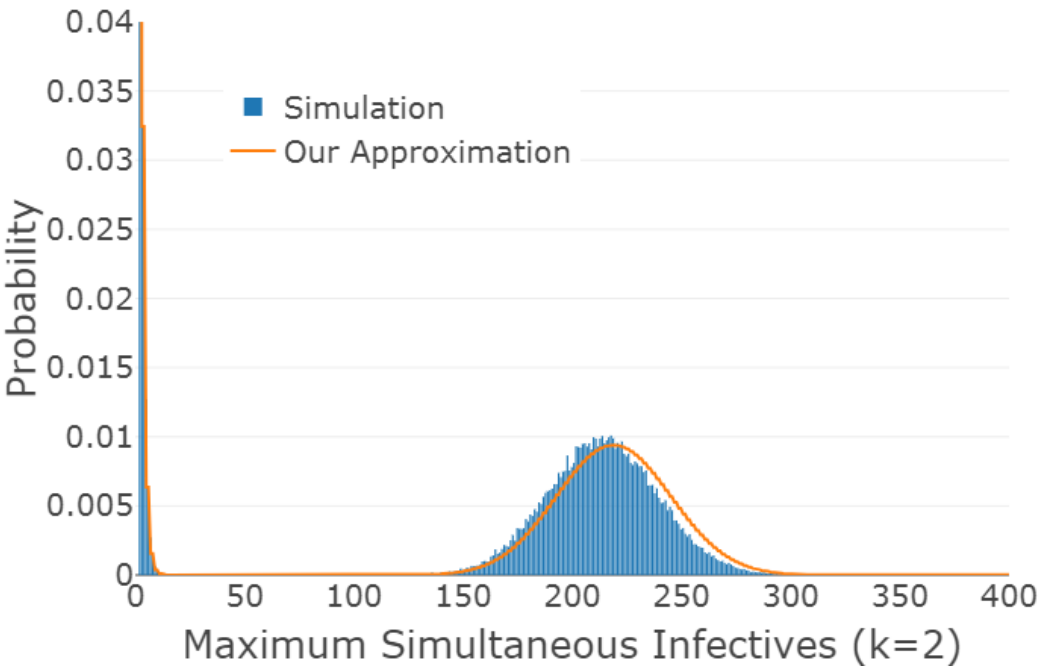

Figure 4.15. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 2$

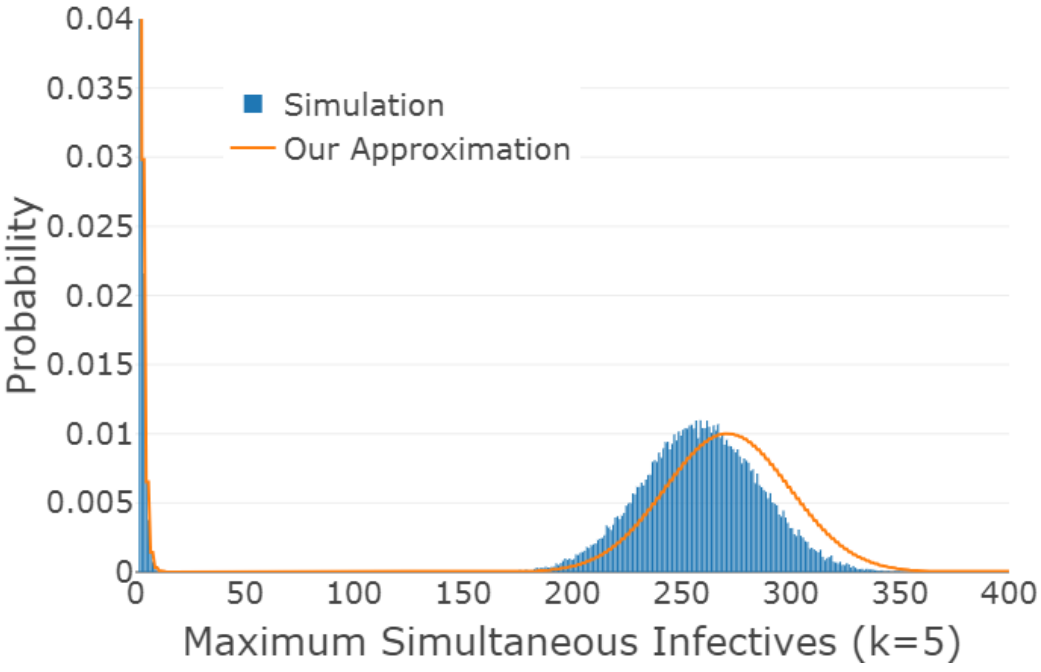

Figure 4.16. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 5$

the relationship between the maximum number of stages and the maximum number of simultaneously infected individuals, and accordingly suggested to use the distribution of V_{max} as an approximation for the distribution of the maximum number of simultaneously infected individuals. Lastly, we produced numerical results for population sizes up to 2000 individuals at a low computational cost and within a reasonable time-frame.

This study can be extended in several directions. By using our algorithms and their limiting behaviour, as the population size increases, we can calculate the probability that a large outbreak occurs. Furthermore, given that an epidemic occurs we can also obtain the conditional distribution of its size. Moreover, it seems to be an interesting question whether our results could be generalized to non-homogeneous population models, e.g. models with a household structure or with a number of large homogeneous sub-populations.

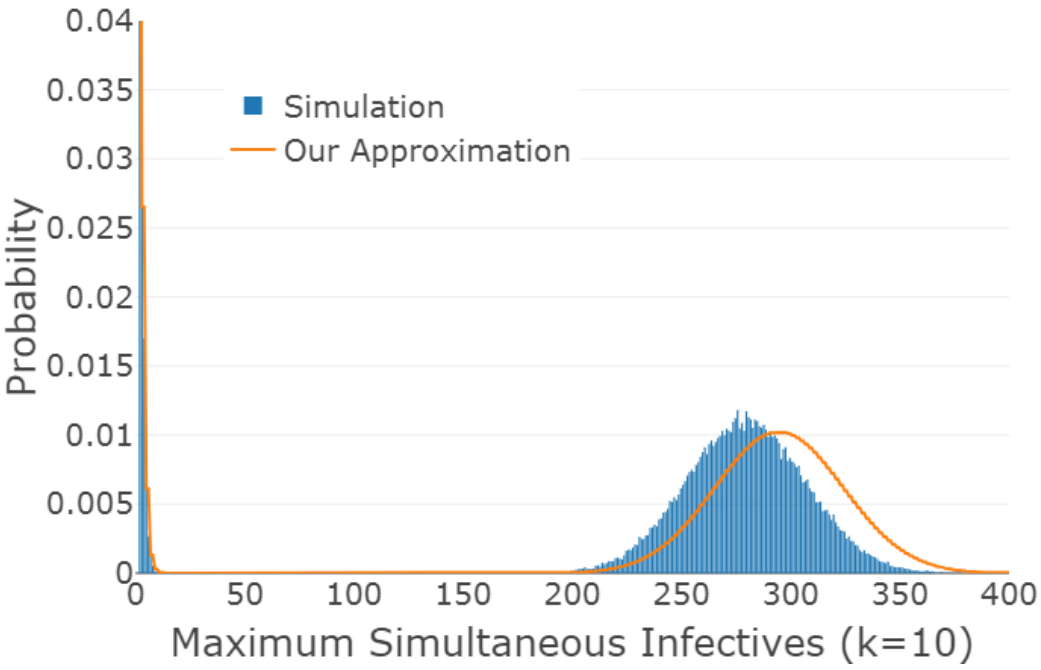

Figure 4.17. The comparison of I_{max} probability mass function obtained by simulation and our approximation using V_{max} for $k = 10$

5. COVID-19 SPREAD: ANALYSIS USING A MODIFIED STOCHASTIC SIR MODEL

In December 2019 an outbreak of COVID-19 started in Wuhan, China spreading rapidly around the world. The reason for its faster spread seems to be its relatively long incubation period, and a long asymptomatic infectious period for some cases (Radulescu and Cavanagh, 2020). By the levels of spread and severity, WHO made the assessment that COVID-19 can be characterized as a pandemic. Consequently, unprecedented intervention strategies are implemented in order to control the outbreak such as cities are quarantined, travel restrictions are implemented, and public spaces are closed. However, these intervention methods result in a significant disruption to the economic activities in the world. While social distancing decreases the need for workload and causes many jobs to be lost, the demand in food sector increases due to stockpiling of food products. Because the socio economic effects of COVID-19 are substantial, it is crucial to determine the level of community social distancing measures when containment like contact tracing is no longer sufficient decreasing the peak of the epidemic to protect healthcare capacity.

To analyse the spread of COVID-19, Yang *et al.* (2020) use a modified susceptible-exposed-infected-removed (SEIR) model that includes the migration data by introducing move-in and move-out parameters for the susceptible (S) and exposed (E) population. They consider that the exposed population is asymptomatic and infectious while infected population (I) is symptomatic and infectious. They estimate the transmission rate parameters for a deterministic model and assume that the contact rate for the susceptible and infected is considerably smaller than the contact rate for the susceptible and exposed because a symptomatic infectious will be quarantined. They discuss various time series problems to predict the number of new infections over time. Radulescu and Cavanagh (2020) study whether the control measures are properly timed and are enough to control COVID 19. They also consider a deterministic SEIR model assuming the transmission rate from susceptible to exposed is $\beta S(I + qE)/N$ where

$q < 1$ because the latent individuals are assumed to have lesser impact and N is the population size. Moreover, they also account for the significant asymptomatic spread of illness by using a large q value. Hou *et al.* (2020) also employ a well mixed SEIR model to describe the dynamics of the COVID 19 and to explore the effectiveness of the quarantine of the Wuhan city by considering infected individuals as contagious during their latency period.

There are also others who consider SIR model for COVID-19. Toda (2020) estimates the classical SIR epidemic model for COVID-19 to assess the economic impact of the epidemic. Moreover, Chen *et al.* (2020) extend the classical SIR model and use a time dependent SIR with undetectable infected persons for COVID-19. They consider two types of infected individuals: detectable infected persons and undetectable infected persons and assume different contact rates and recovery rates for different types of infected individuals. Calafiore *et al.* (2020) also study a SIR model for COVID-19 in Italy by assuming that it is possible to detect only a portion of infected individuals. They provide an effective explanatory model for prediction of the future evolution of the disease. Simha *et al.* (2020) model the evolution of COVID-19 infections using a stochastic SIR model and express the dynamics of the spread using stochastic differential equations. They implement simulation to obtain projections for India.

In this part of the thesis, we present a stochastic SIR model considering two types of infected individuals as symptomatic and asymptomatic, using a Markov chain formulation. While the studies on COVID-19 in the literature are mostly deterministic, the main advantage of our model is that its stochasticity is better in regarding uncertainties and accounting for real variabilities. Moreover, we compute the exact final outbreak size for COVID-19 for given incubation and infectious period approximated by our mixing of Erlangs distribution. Then, we check how state dependent contact rates affect final outbreak size distribution to assess the timing and the intensity of intervention methods.

5.1. Model Definition

Data for COVID-19 indicate that there are some cases who have contact the virus and transmit the virus to other people but never exhibit symptoms. These cases are called asymptomatic infectious (I_a) and should be considered in the studies for COVID-19 spread. Let p be the fraction of asymptomatic cases among all positive cases. Furthermore, the cases who exhibit symptoms (I_s) has a significant incubation period, also known as presymptomatic period, during which they do not exhibit symptoms yet but can infect others. The symptomatic cases during their incubation period are called as presymptomatic infectious (I_p). Because it is not possible to detect all carriers of COVID-19 due to lack of widespread testing, the presence of asymptomatic cases and the long incubation period are more dangerous for the public health. Therefore, we modify the original SIR model to account for the long incubation period and asymptomatic cases. The state transition diagram for COVID-19 is presented in Figure 5.1.

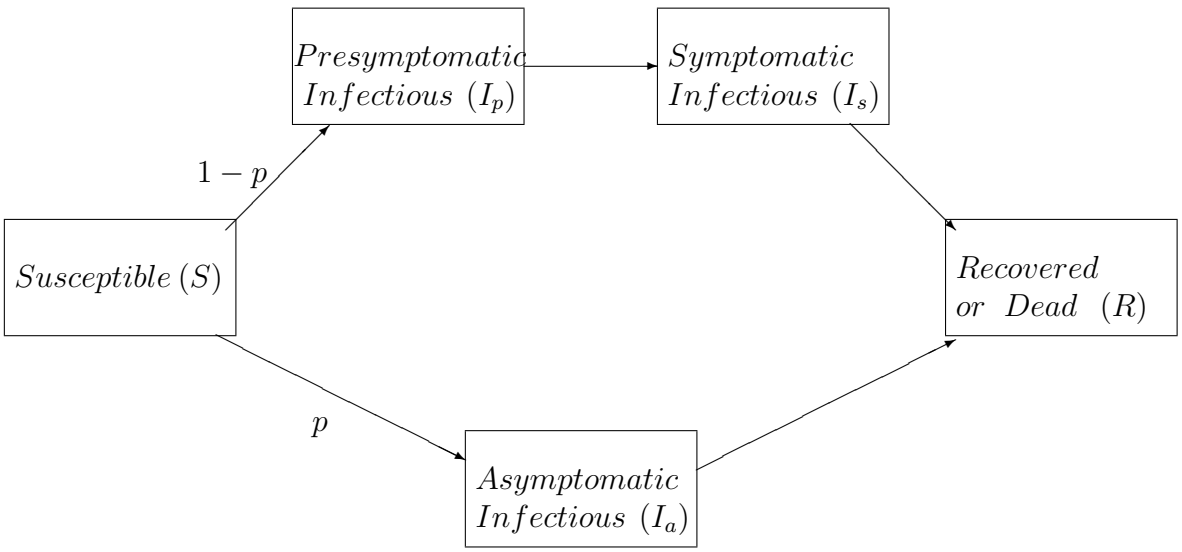

Figure 5.1. State transition diagram for COVID-19

We consider a finite population with size N where all individuals mix uniformly. The COVID-19 spread is modelled by a Markov process for our modified SIR model in which both symptomatic and asymptomatic individuals are considered. Because COVID-19 spreads as easily from individual to individual as influenza and more deadly than the flu, we assume that all cases who exhibit symptoms (I_s) will be quarantined

and cannot transmit the disease. Therefore, only the presymptomatic infectious individuals (I_p) and the asymptomatic infectious individuals (I_a) can transmit the disease. Furthermore, the total number of contacts (per unit time) ending up with an infection is assumed to follow a Poisson process with rate $\lambda s(i_p+i_a)/N$. Let the contagious time be the time period for infected individuals during which they transmit the disease. Therefore, the contagious time for I_p is the incubation period of COVID-19. However, it is not possible to estimate the contagious period for asymptomatic cases exactly because they never exhibit symptoms.

In the literature, both the incubation period and the infectious period are estimated by discrete distributions. Since we model the spread as a Markov process, we assume a mixture of Erlangs distributed contagious period that can approximate the discrete distributions suggested in the literature. Let's consider a given mean (μ_T) and a standard deviation (σ_T) of an infectious period, T, so that its coefficient of variation is $c_v = \sigma_T/\mu_T < 1$. By using the approximation in Section 4.2.1, if $c_v^2 \in [1/(k_p-1), 1/k_p]$, the contagious time for a presymptomatic infectious takes Erlang(k_p-1,μ) and Erlang(k_p,μ), with probabilities α, and $1-\alpha$ respectively, where

$$\alpha = \frac{1}{1+c_v^2}(k_p c_v^2 - \{k_p(1+c_v^2) - k_p^2 c_v^2\}^{1/2}), \tag{5.1}$$

$$\mu = \frac{k_p - \alpha}{\mu_T}. \tag{5.2}$$

Therefore, we assume a mixture of Erlang distributions for the contagious time of presymptomatic cases. However, the estimation of contagious period for asymptomatic cases is challenging because they mostly cannot be detected. Thus, we assume an Erlang distribution for the contagious period of asymptomatic cases with rate parameter μ and shape parameter k_a. Supposing $k_a < k_p$, we first consider that the expected contagious time for asymptomatic cases is shorter than the contagious time for presymptomatic cases. Then, we also consider longer contagious time for asymptomatic cases by supposing $k_a > k_p$.

Transformation of the state space in Proposition 2 enables us to transform the original Markov process for COVID-19 to the imbedded process having the same absorption probabilities. Let us define,

$$\Pi_m(v,s) = \Pr\{(V(\tau),S(\tau)) = (0,N-m)|(V(0),S(0)) = (v,s)\}.$$

Starting with a total of v disease stages and s many susceptible, $\Pi_m(v,s)$ is the probability that at the end of the outbreak (at time τ), the total number of stages is zero, and the total number of susceptible remaining is $N-m$. Then, we can compute final outbreak size (m) distributions by implementing the first step analysis for our modified model such that

$$\begin{aligned}\Pi_m(v,s) &= \frac{\mu(I_p+I_a)}{\lambda s(I_p+I_a)/N+\mu(I_p+I_a)}\Pi_m(v-1,s)\\ &+ \frac{\lambda s(I_p+I_a)/N}{\lambda s(I_p+I_a)/N+\mu(I_p+I_a)}\Bigg\{p\Bigg(\alpha\Pi_m(v+k_p-1,s-1)\\ &+ (1-\alpha)\Pi_m(v+k_p,s-1)\Bigg)+(1-p)\Pi_m(v+k_a,s-1)\Bigg\}. \qquad (5.3)\end{aligned}$$

Equation 5.3 does not depend on the number of I_p and I_s, so we do not need to follow them up and Equation 5.3 can be written as

$$\begin{aligned}\Pi_m(v,s) &= \frac{\mu N}{\lambda s+\mu N}\Pi_m(v-1,s)+\frac{\lambda s}{\lambda s+\mu N}\Bigg\{p\Bigg(\alpha\Pi_m(v+k_p-1,s-1)\\ &+ (1-\alpha)\Pi_m(v+k_p,s-1)\Bigg)+(1-p)\Pi_m(v+k_a,s-1)\Bigg\}. \qquad (5.4)\end{aligned}$$

Because both the incubation period and the infectious period are uncontrollable factors for COVID-19 and introduction of COVID-19 vaccine takes time, the only effective way to control disease spread seems to reduce contact rate, λ, by social distancing. Because the peak is important for health capacity needs and treatment availability, it makes sense to determine timing of control measures by considering the number of active cases.

Let λ be equal to λ_0 without any control measures. We assume that λ_0 is reduced to λ_c by implementing control measures if at least $c\%$ of the population becomes simultaneously infectious but not detected. Assuming when control measures are introduced and lifted dependent on the current number of cases, we can define λ as

$$\lambda = \begin{cases} \lambda_0, & I_p(t) + I_a(t) \leq i_{critical} \\ \lambda_c, & I_p(t) + I_a(t) > i_{critical}. \end{cases} \tag{5.5}$$

However, our transformed process does not follow up the number of presymptomatic and asymptomatic individuals but v. Because we cannot observe v directly, we use the approximation given in Proposition 2 and change contact rates by considering $v_{critical}$ such that

$$\lambda(v, s) = \begin{cases} \lambda_0, & \text{v} \leq v_{critical} \\ \lambda_c, & \text{v} > v_{critical} \end{cases} \tag{5.6}$$

where

$$v_{critical} = \frac{Nc}{100}\frac{(k+1)}{2}. \tag{5.7}$$

Moreover, Markov chain transition probabilities become state dependent and Equation 5.4 can be written as

$$\begin{aligned} \Pi_m(v, s) &= \frac{\mu N}{\lambda(v, s)s + \mu N}\Pi_m(v-1, s) + \frac{\lambda(v, s)s}{\lambda(v, s)s + \mu N}\Bigg\{p\Bigg(\alpha\Pi_m(v + k_p - 1, s-1) \\ &+ (1-\alpha)\Pi_m(v + k_p, s-1)\Bigg) + (1-p)\Pi_m(v + k_a, s-1)\Bigg\}. \end{aligned} \tag{5.8}$$

5.2. Determination of Model Parameters

In this section, we decide on the values of λ, k_p, k_a, and μ based on the empirical results in the literature. We also need to know p to analyse the disease behaviour. More

recent parameter estimates for COVID-19 have been made by using the information from early clusters of COVID-19 cases. Even if the estimates are different depending on the population used, they are close to each other for the incubation period and the infectious period. Moreover, the mean incubation period is found as 7.1 days for Singapore and 9 days for Tianjin (Tindale *et al.*, 2020). In this study, we set the mean and the standard deviation of the time period during which presymptomatic cases can infect susceptible individuals to 8 and 4.75 days respectively based on the incubation period distribution of COVID-19 in You *et al.* (2020). Thus, its coefficient of variation is equal to 0.59. Using mixture of Erlangs, we can find that for $k_p = 3$ we have $0.59^2 \in [\frac{1}{k_p}, \frac{1}{k_p - 1}]$. Then,

$$\begin{aligned} \alpha &= \frac{1}{1+0.59^2}(3(0.59^2) - \{3(1+0.59^2) - (3^2)(0.59^2)\}^{1/2}) = 0.0665, \\ \mu &= \frac{3-\alpha}{8} = 0.3666. \end{aligned}$$

Thus, we assume that the contagious period for symptomatic cases has an Erlang (2, 0.3666) distribution with probability 0.0665 and an Erlang (3, 0.3666) distribution with probability 0.9335 corresponding to average infectious period 8 and standard deviation 4.75. Noting that asymptomatic cases are typically hard to detect, there are no direct records for the infectious period of asymptomatic cases. In order to assess how asymptomatic cases affects the final outbreak size distribution, we assume that the contagious period of asymptomatic cases has an Erlang distribution with parameters $(k_a = 2, \mu = 0.3666)$ and $(k_a = 4, \mu = 0.3666)$ for both shorter and longer contagious time than the contagious time of presymptomatic cases.

It is also hard to determine the real proportion of asymptomatic cases because there is no widespread testing yet. However, Mizumota *et al.* (2020) conduct a statistical modelling analysis to estimate the proportion of asymptomatic individuals among those who tested positive in a cruise ship called Diamond Princess underwent a 2 week quarantine. Based on their estimates, we assume that asymptomatic proportion is 17.9% and set p equal to 0.821.

We specify the contact rate λ according to R_0 estimates for COVID-19. The studies using stochastic and statistical methods to estimate R_0 is consistent with WHO's estimates from 1.4 to 2.5. Moreover, R_0 have stabilized at around 2-3 in recent studies that are consistently above WHO' point estimates for R_0. Assuming a finite population with size N, R_0 is calculated as follows

$$R_0 = \frac{\lambda(N-1)}{N}E[T] = \frac{\lambda(N-1)}{N}\left(p\left(\alpha\frac{k_s-1}{\mu} + (1-\alpha)\frac{k_s}{\mu}\right) + (1-p)\frac{k_a}{\mu}\right)$$

where T is the contagious time. As N goes to infinity R_0 becomes

$$R_0 \approx \lambda\left(p\left(\alpha\frac{k_p-1}{\mu} + (1-\alpha)\frac{k_p}{\mu}\right) + (1-p)\frac{k_a}{\mu}\right). \tag{5.9}$$

We consider different R_0 estimates reported to range between 2 and 4 and calculate the corresponding λ values using Equation 5.9 to demonstrate how the spread is affected by the changing contact rates. Table 5.1 displays the R_0 values and corresponding contact rates λ that are considered for COVID-19 spread in this study.

Table 5.1. Published estimates of R_0 for COVID-19 and our corresponding λ

Study	Location	R_0	$\lambda(k_a = 2)$	$\lambda(k_a = 4)$
Li *et al.* (2020)	Chine	2.24	0.30	0.26
Wu *et al.* (2020)	Wuhan	2.68	0.36	0.31
Read *et al.* (2020)	Chine	3.11	0.41	0.36
Zhao *et al.* (2020)	Chine	3.58	0.47	0.42

The estimates for R_0 are too large before control measures come in. Moreover, R_0 can be reduced below one by introducing social distancing and even full lock down. Let R_c be the basic reproduction number after control measures are introduced. We consider two possible values of R_c that are 0.7 and 0.95 respectively.

5.3. Numerical Results

In this section, we analyse and predict COVID-19 spread by using our modified SIR model with parameters chosen in Section 5.2. First, we compute the exact final outbreak size distribution for COVID-19 through implementation of first step analysis using Equation 5.4 that enables us to check whether it is possible to control COVID-19 without intervention. And, we determine the ratio of the population that becomes infected when disease ends if COVID-19 cannot be controlled. The computation of final outbreak size after intervention strategies are introduced is also important. Because decreasing R_0 from 3 to below 1 requires huge effort causing economic problems, the usual tendency is to loosen restrictions when the epidemic is under control. The number of cases is important here for deciding when to loosen and tighten the restrictions. Moreover, we compute the final outbreak size distribution by considering state dependent Markov chain probabilities to understand how the timing and the intensity of social distancing affect it.

5.3.1. Final Outbreak Size Distribution for COVID-19

First, we calculate the final outbreak size distribution for the smallest and largest R_0 values in Table 5.1 considering $k_a = 2$ and $k_a = 4$ and present the results in Figures 5.2, 5.3, 5.4, and 5.5. We observe that the proportion of the population who have been ever infected increases dramatically as R_0 increases. Therefore, reduction of R_0 is extremely important for COVID-19 control. Moreover, we assume two possible k_a values for the contagious time of asymptomatic cases such that the average contagious time for $k_a = 4$ is two times greater than the average contagious time for $k_a = 2$. However, we do not observe a significant difference in the final outbreak size distributions of two outbreaks with different contagious time but equal R_0s.

In an epidemic model, another crucial question is whether the disease can be controlled or a certain fraction of total population is infected. To check if the disease can be controlled, we display the cumulative final outbreak size distribution in Figures Figures 5.6, 5.7, 5.8, and 5.9 and observe that the probability of an outbreak is af-

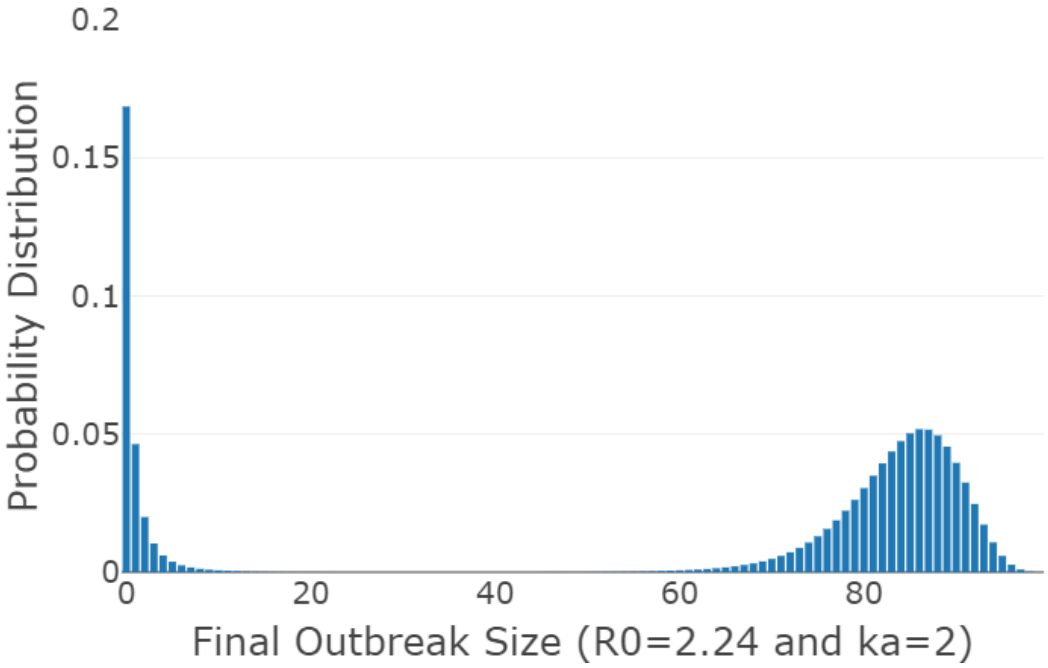

Figure 5.2. Final outbreak size distribution for $R_0 = 2.24$ and $k_a = 2$

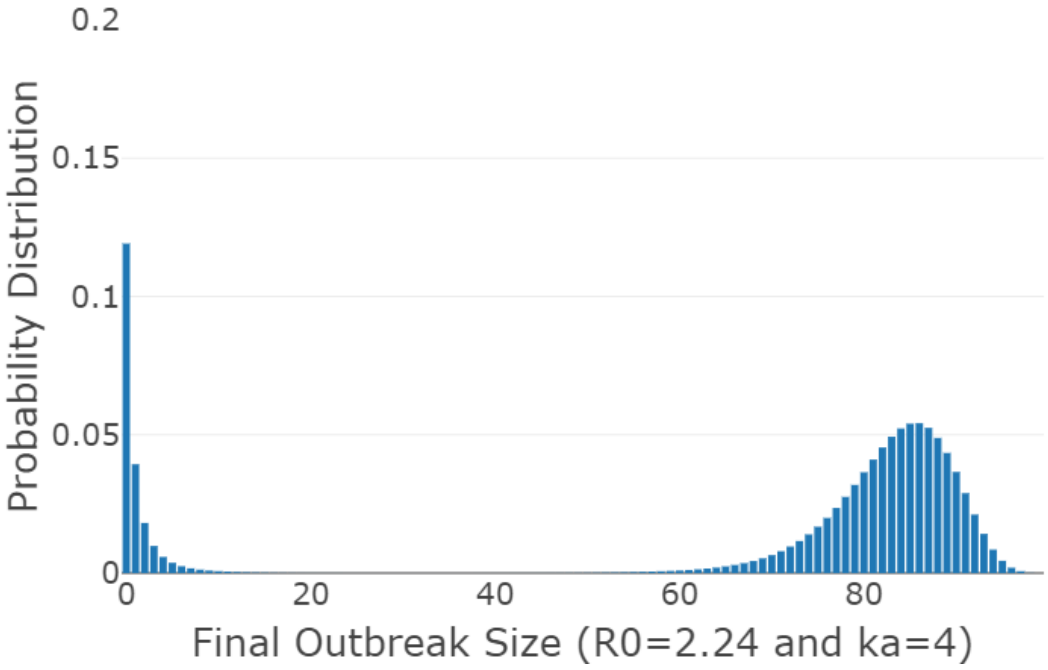

Figure 5.3. Final outbreak size distribution for $R_0 = 2.24$ and $k_a = 4$

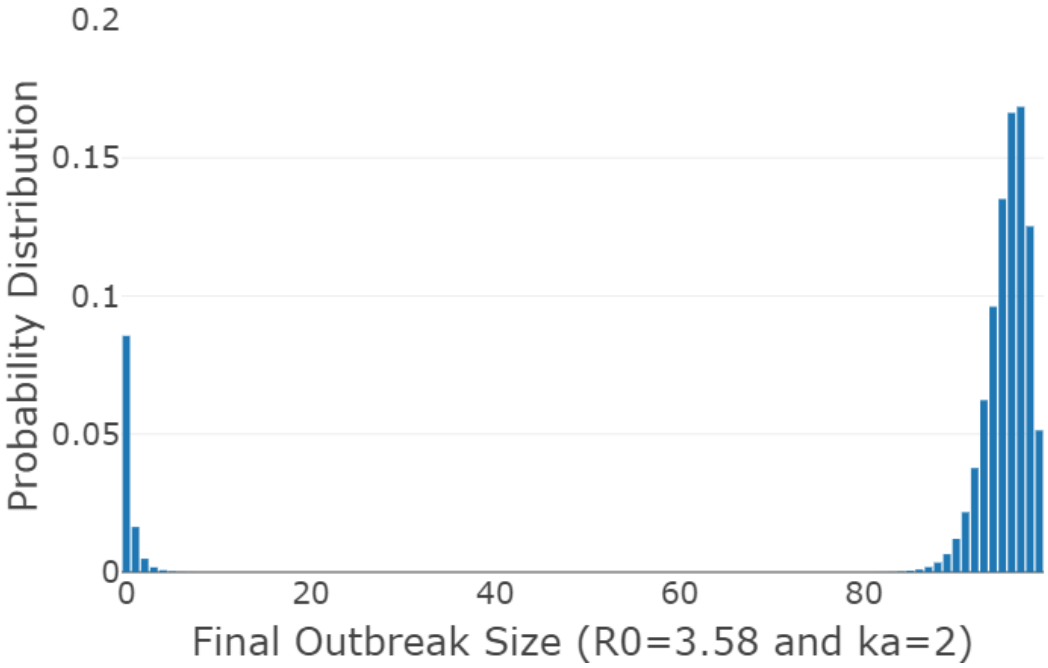

Figure 5.4. Final outbreak size distribution for $R_0 = 3.58$ and $k_a = 2$

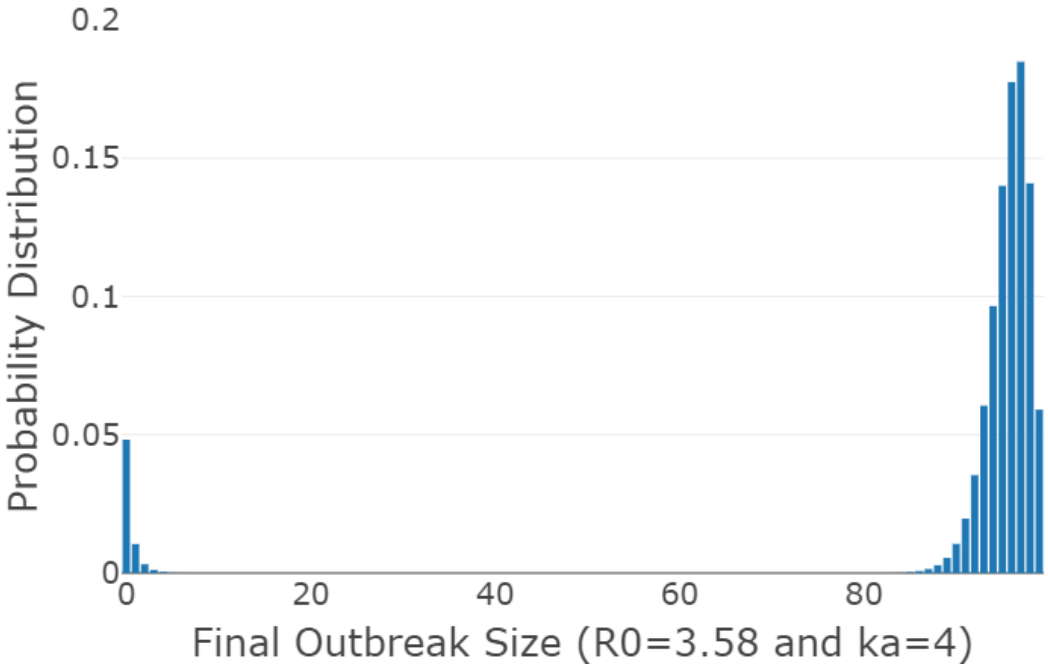

Figure 5.5. Final outbreak size distribution for $R_0 = 3.58$ and $k_a = 4$

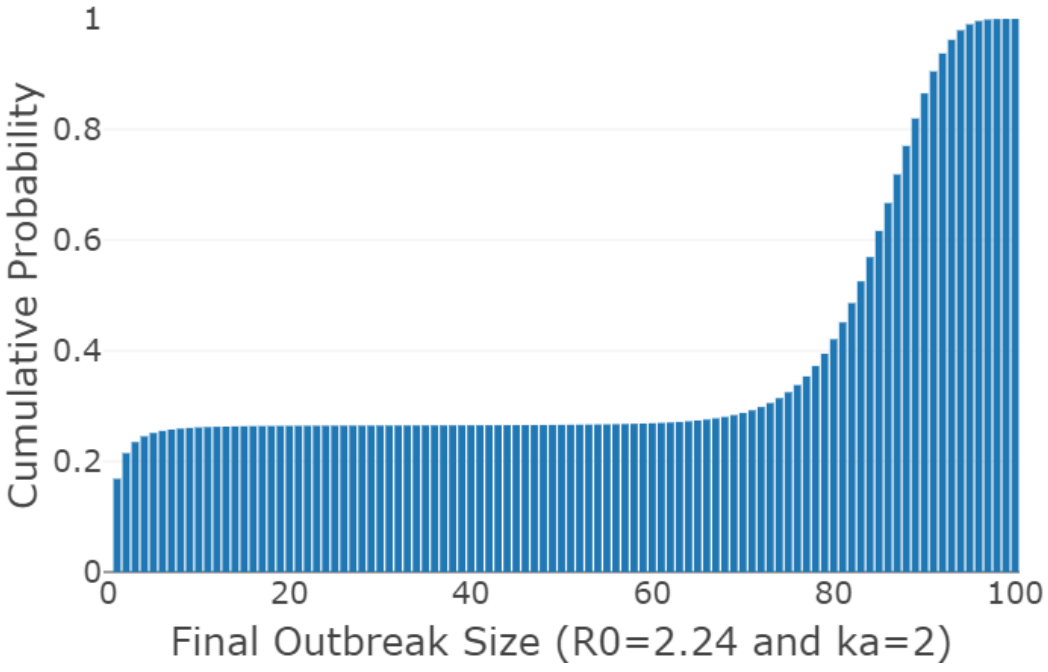

Figure 5.6. Cumulative final outbreak size distribution for $R_0 = 2.24$ and $k_a = 2$

fected by both R_0 and the contagious time of asymptomatic cases. Because we assume that symptomatic cases are quarantined when they exhibit symptoms, asymptomatic cases become more dangerous when their contagious time is longer than the contagious time of symptomatic cases. Moreover, we compute both the final outbreak size distribution and cumulative final outbreak size distribution by assuming one initially infected individual, so an outbreak is almost certain if there are more initially infected individuals.

Because the probability of an outbreak is large and an outbreak is almost certain even if there remains a single infected individual, the important question is, what is the fraction of individuals who have ever had the infection when the disease ends since it is directly proportional to the total number of deaths. To address this question, we check the average percentage of the population having COVID-19 during pandemic given that at least 10% of the population became infected by

$$E[R(\tau)] = \frac{\sum_{m=N/10}^{N} P(R(\tau) = m)m}{\sum_{m=N/10}^{N} P(R(\tau) = m)}. \tag{5.10}$$

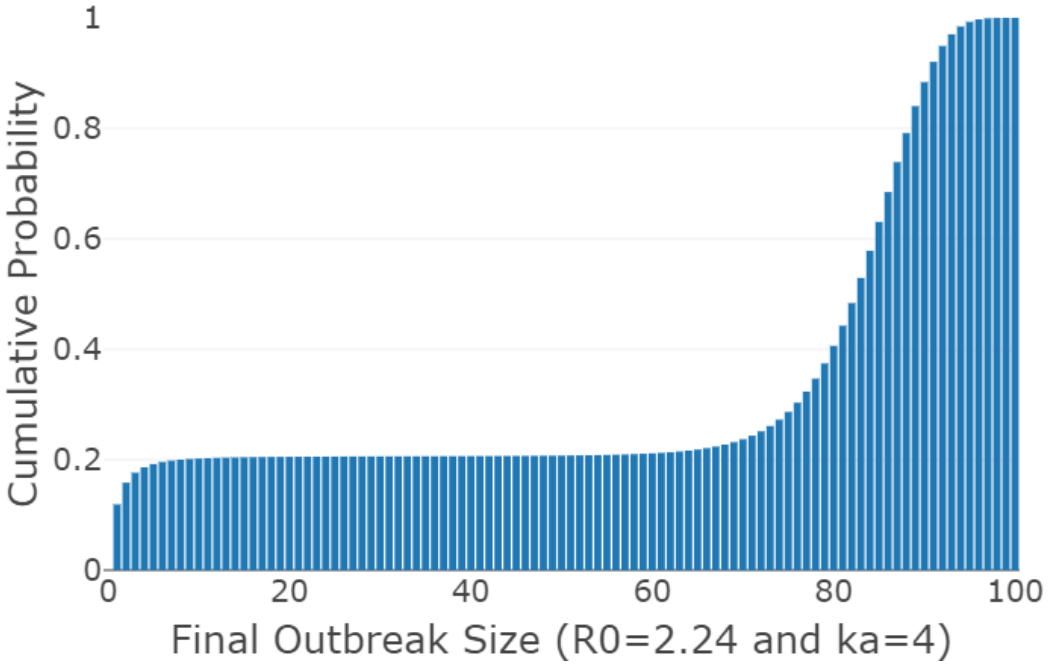

Figure 5.7. Cumulative final outbreak size distribution for $R_0 = 2.24$ and $k_a = 4$

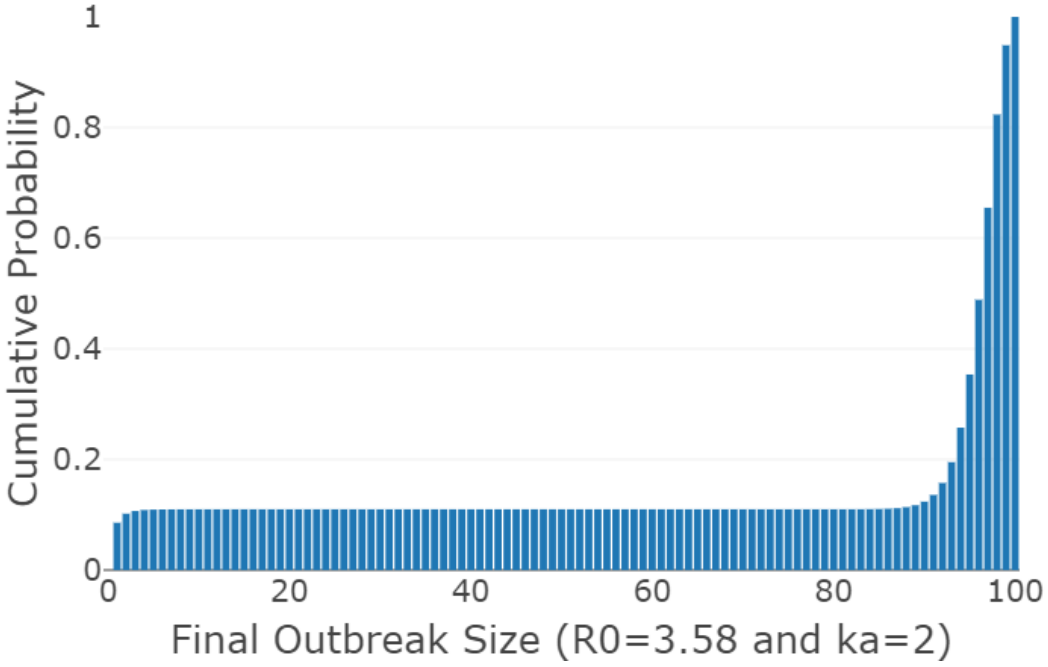

Figure 5.8. Cumulative final outbreak size distribution for $R_0 = 3.58$ and $k_a = 2$

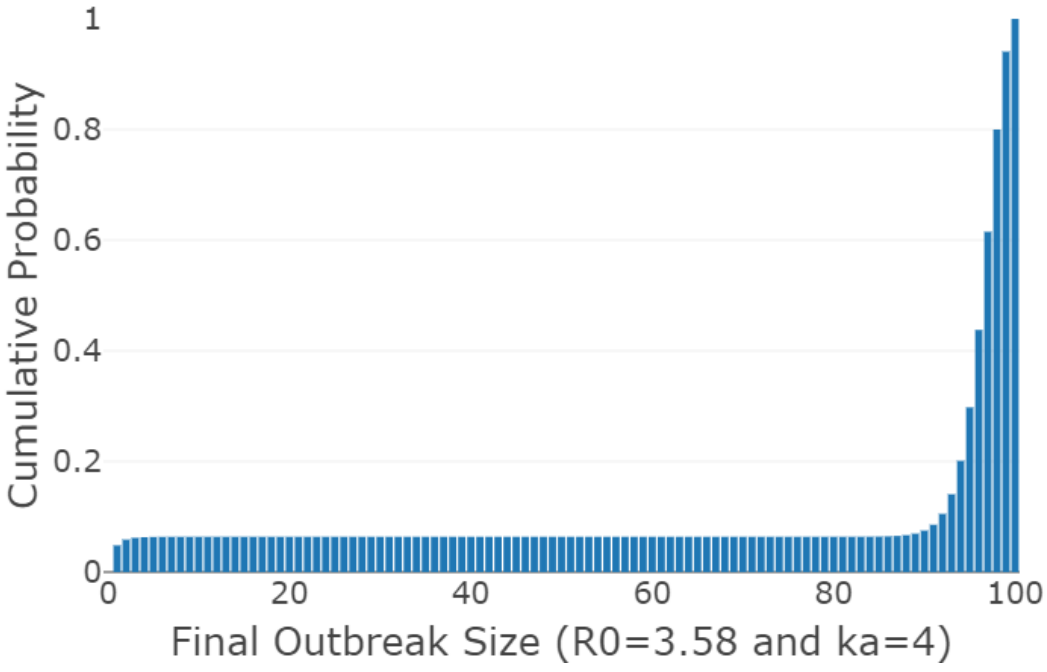

Figure 5.9. Cumulative final outbreak size distribution for $R_0 = 3.58$ and $k_a = 4$

We compute the average percentage of the population having COVID-19 for $N = 500, 1000$ and 2000 by using Equation 5.10 and present the results in Table 5.2.

Table 5.2. Average percentage of the individuals who have been infected during epidemic (%)

	$R_0 = 2.24$		$R_0 = 2.68$		$R_0 = 3.11$		$R_0 = 3.58$	
N	$k_a = 2$	$k_a = 4$	$k_a = 2$	$k_a = 4$	$k_a = 2$	$k_a = 4$	$k_a = 2$	$k_a = 4$
500	85.46	84.56	91.64	90.88	94.58	94.44	94.59	96.85
1000	85.53	84.63	91.68	90.92	94.62	94.47	96.75	96.86
2000	85.56	84.66	91.70	90.94	94.64	94.48	96.76	96.87

The average percentage of the population having COVID-19 when the outbreak ends does not change significantly with changing population size so we can estimate it for an infinite size population. It seems that simply waiting for herd immunity to COVID-19 is not an option since a huge fraction of individuals has the infection resulting in great number of deaths. Therefore, the most effective way to reach herd immunity seems to find vaccines and reduce the number of susceptible individuals.

However, because it takes time, other methods to control the spread of COVID-19 must be considered.

5.3.2. Final Outbreak Size Distribution with Changing Contact Rates

Because the incubation period and the infectious period are uncontrollable factors for COVID-19 and introduction of COVID-19 vaccine takes time, the only way to decrease R_0 seems to reduce λ by social distancing. It is possible to reduce R_0 around 0.7 by imposing strict measures causing significant socio economic problems. However, it is also not possible to control the outbreak without dropping R_0 below one.

Because individuals' behaviour changes and adapts to the disease, R_0 can decrease even if control measures do not change. Therefore, we assume $R_0 = 2.24$ that is the smallest R_0 value considered in Section 5.2. And, we compare four different ways to implement intervention strategies and to understand how the timing and intensity of social distancing affect the final outbreak size distribution.

First, we consider to impose strict measures and decrease the contact rates dramatically such that R_c becomes 0.7 if more than 5% and 10% of the population becomes simultaneously infected, respectively. Then, we also consider $R_c = 0.95$ if more than 5% and 10% of the population become simultaneously infected. Because the peak is important for the health capacity needs and treatment availability, timing of control measures is determined by considering the number of active cases. We compare the final outbreak size distributions in Figures 5.10, 5.11, 5.12, and 5.13 for $k_a = 4$ when the the timing and intensity of social distancing are different.

Assuming that control measures are introduced and lifted according to the current number of cases, the timing of social distancing affects the final outbreak size distribution more than the intensity of social distancing as long as R_c is smaller than one. The final outbreak size distributions can be useful for selecting both schedule and severity of control measures. From the figures, we observe that it is crucial to introduce intervention methods when the number of cases is small. Moreover, the final outbreak

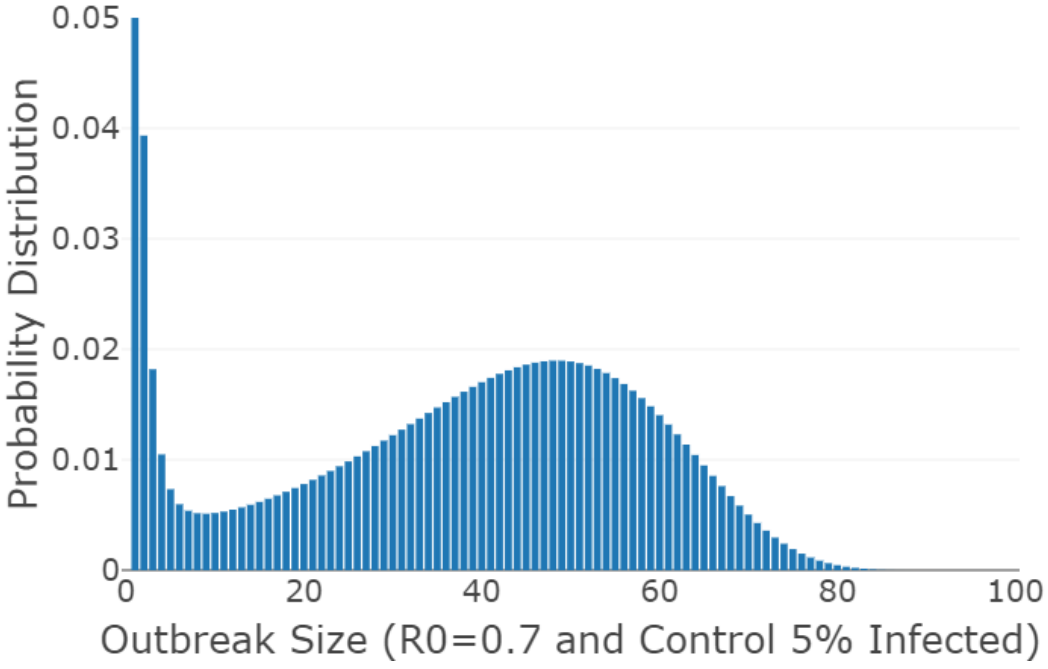

Figure 5.10. Final outbreak size distribution if $R_c = 0.7$ when at least 5% simultaneous infectives

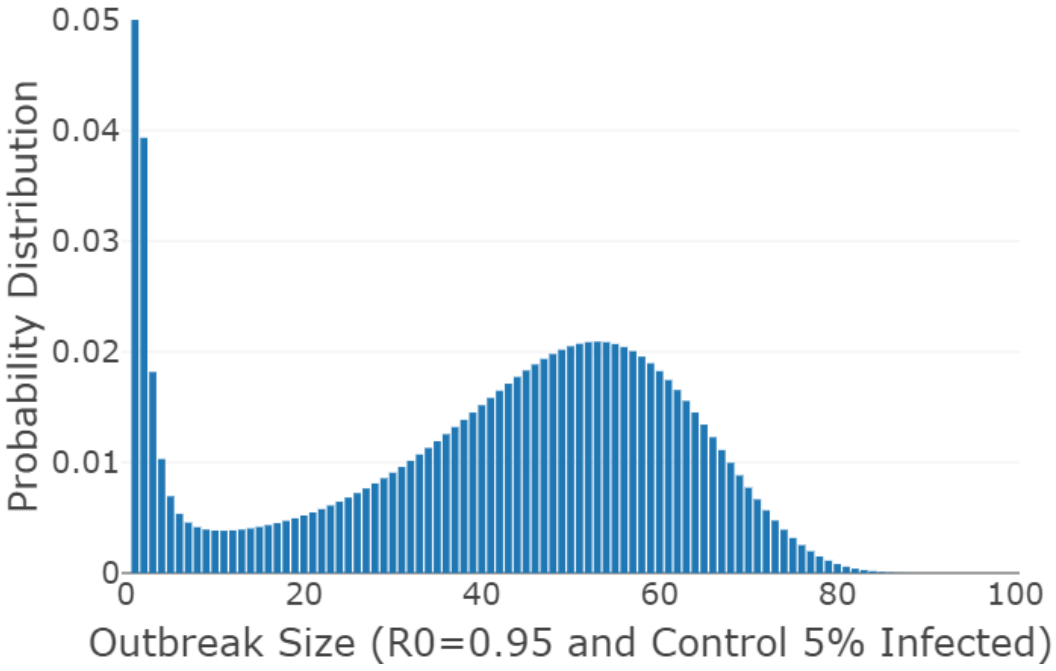

Figure 5.11. Final outbreak size distribution if $R_c = 0.95$ when at least 5% simultaneous infectives

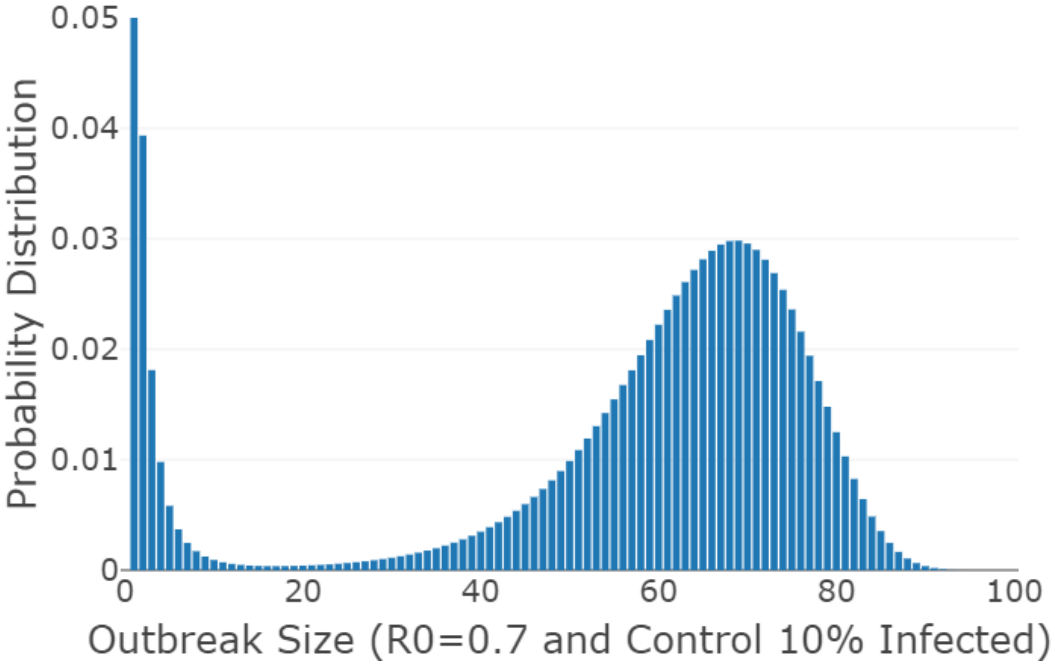

Figure 5.12. Final outbreak size distribution if $R_c = 0.7$ when at least 10% simultaneous infectives

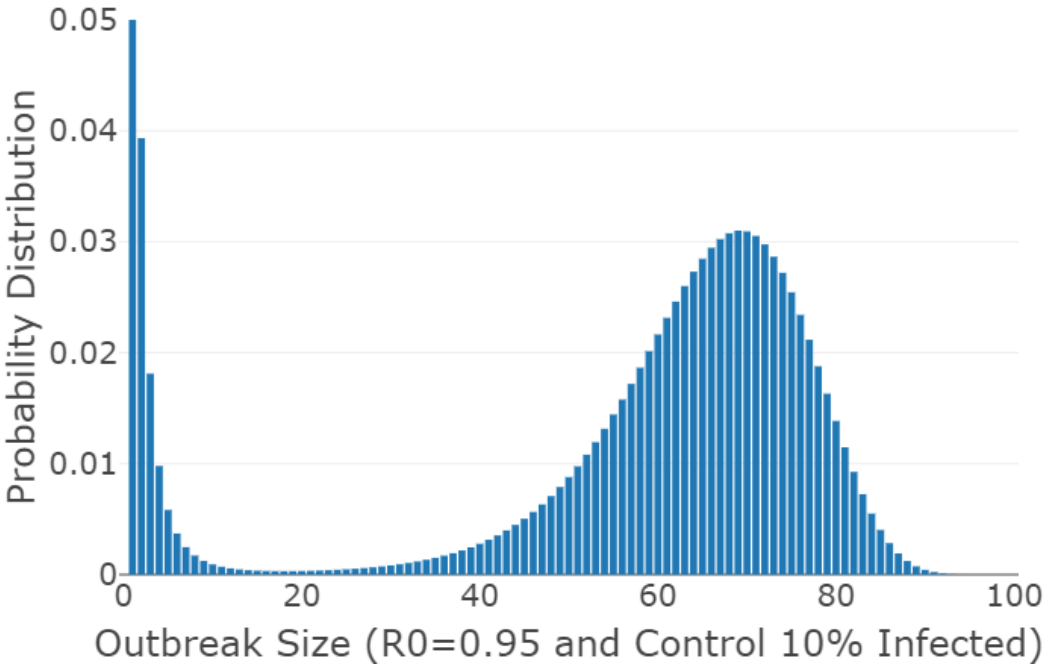

Figure 5.13. Final outbreak size distribution if $R_c = 0.95$ when at least 10% simultaneous infectives

size distributions for $R_c = 0.7$ and $R_c = 0.95$ are not different, so it is not necessary to reduce the contact rate much more after reducing R_c below one.

Lastly, we check the average fraction of the population who have ever had the infection when the disease ends given that there is an outbreak. The results are presented in Table 5.3. We observe that the expected total number of infected individuals increases significantly causing the total number of deaths to increase if control measures are implemented late. Furthermore, the expectation of the total number of infected individuals reduces to 67% if control measures are implemented when 5% of the population is simultaneously infected and to 74% if control measures are implemented when 10% of the population is simultaneously infected.

Table 5.3. Average percentage of the individuals who have been infected during epidemic (%)

	Control if 5% Simultaneous Infectives and $R_c = 0.7$		Control if 5% Simultaneous Infectives and $R_c = 0.95$		Control if 10% Simultaneous Infectives and $R_c = 0.7$		Control if 10% Simultaneous Infectives and $R_c = 0.95$	
N	$k_a = 2$	$k_a = 4$	$k_a = 2$	$k_a = 4$	$k_a = 2$	$k_a = 4$	$k_a = 2$	$k_a = 4$
500	66.56	65.37	66.74	65.53	73.80	72.80	73.87	72.86
1000	68.59	67.56	68.65	67.62	74.71	73.74	74.74	73.77
2000	69.59	68.59	69.62	68.61	75.23	74.27	75.25	74.28

5.4. Discussion

For a stochastic SIR model, we consider two types of infected individuals that are symptomatic and asymptomatic. We approximate the incubation period and the infectious period distributions by using a mixture of Erlang distributions to model a Markov chain disease spread model. We calculate the exact final outbreak size distribution and the approximate maximum epidemic size distribution by implementing first step analysis.

Since we can compute the exact final outbreak size distributions for large size

populations, we estimate the outbreak probability for different R_0s by checking their cumulative probability distribution. And, we calculate the average percentage of the population that has ever been infected assuming an outbreak exists. Since an outbreak is certain for COVID-19, we investigate how the final outbreak size distribution changes with different strategies by controlling the timing and the intensity of social distancing. We consider state dependent Markov chain probabilities while considering different timing and intensity of social distancing.

We can extend this study to assess the influence of quarantine and isolation procedures on final outbreak size distribution and the maximum epidemic size distribution. Moreover, we can conduct a sensitivity analysis to investigate how uncertainty in the disease parameters affects the dynamics of the disease spread. It is also possible to assess how the proportion of asymptomatic cases affects the spread of the disease.

6. R_0 NOTION AND ASSESSMENT OF INTERVENTION STRATEGIES FOR NON HOMOGENEOUS POPULATIONS[2]

Epidemiological information is mostly used to plan and evaluate strategies that prevent disease spread by identifying risk factors. Therefore, various disease spread models were developed in the literature. The basic reproduction number R_0 as the expected number of secondary cases produced by a single infected case in a totally susceptible population is important for the determination of the outbreak probability under homogeneous mixing assumption (Hernandez-Suarez, 2002; Craft *et al.*, 2013; Allen and Burgin, 2000). Moreover, R_0 is generally used in the literature for analyzing the possibility of an outbreak even if it is also possible to use it for intervention strategy analysis. In the literature, to develop and analyze epidemic control strategies, different mathematical approaches are implemented like introducing contact network epidemiology (Dimitrov and Meyers, 2010), implementing optimal control tools (Sharomi and Malik, 2017) and simulating scenarios (Wu *et al.*, 2006; Carvalho *et al.*, 2019). However, there are some recent papers considering the use of R_0 to analyze and develop epidemic control methods. Ball and Lyne (2002) use the average R_0 for optimal vaccination policies in a population partitioned into households. Artalejo and Lopez-Herrero (2013) also suggest to use R_0 to design control strategies for prevention of an outbreak. However, their analysis requires exponential infectious period.

In this part of the thesis, we consider a stochastic SIR among non homogeneous populations. We are concerned with the notion of computable R_0 for heterogeneous models. Our aim is to calculate the expected number of secondary cases produced by a given infected case and use it to develop effective intervention strategies and assess the intervention strategies without simulation. Therefore, we firstly introduce individual R_0 as the expected number of secondary cases produced by a given initially infected individual in a totally susceptible population. Individual R_0 is the expectation of $R_{e,0}$

[2]This chapter appears in İşlier *et al.* (2020b) as a part of its content.

rather than the exact number of secondary cases so we can propose a general formula for individual R_0 in this study, that is applicable to all types of heterogeneous populations with any size. Our second major contribution is to present that it is possible to assess intervention strategies by using the exact formula for individual R_0 without reverting to simulation. It is possible to assess the impact of intervention strategies by their capability to reduce individual R_0 values. Also, a maximum individual R_0 value smaller than 1 guarantees that an outbreak is impossible. Lastly, optimal intervention strategies can be identified based on individual R_0 values. We propose a vaccination strategy such that the individual with greatest individual R_0 are vaccinated first. In order to choose the individual who is vaccinated next, we recalculate individual R_0 values for the unvaccinated individuals and choose the individual with the greatest individual R_0 again. Thus, our vaccination strategy is to vaccinate individuals one by one by choosing the susceptible having the largest individual R_0.

6.1. The Notion of R_0 for Non Homogeneous Models

Stochastic SIR models in large non homogeneous populations grew popular among practitioners in recent years, see eg. (Longini *et al.*, 2005; Ajelli *et al.*, 2010). The infection probability between an infected and a susceptible individual is modelled with a comparatively small number of parameters assuming mixing in overlapping mixing groups. The detailed structure of a population is generated such that the mixing groups match in size and age those of real world census data. As mixing groups typically households, neighbourhoods, communities, schools and work places are considered. In several papers it is assumed without any discussion that the only way to assess the behaviour of such models is simulation. This fact attracted our attention and we aim to develop here an approach to assess the behaviour of such models for large populations using a properly defined basic reproduction number R_0 that can be calculated easily also for large populations.

As individuals in a non homogeneous population are not identical, R_0 for non homogeneous populations depends on the initially infected individual that is chosen. Thus, different R_0 values occur for different initially infected individuals. In agent

based simulation literature, the value of R_0 for the entire non homogeneous population is generally estimated by assuming "the random case". Thus, the initially infected individual is selected among the population with equal probability for every individual. Then, they use an average to calculate R_0 (Longini *et al.*, 2005). It is also suggested to estimate age dependent R_0 values and calculate overall R_0 as a weighted average of age dependent attack rate patterns (Germann *et al.*, 2006). The studies which use branching process methods to calculate R_0 also considers R_0 for non homogeneous populations as a mean of different secondary cases for different initially infected individuals (Ball and Lyne, 2002). In these studies, populations with different mixing levels and moderate size are considered based on census data. However, average R_0 is estimated via simulation without exact solution and it cannot be used to assess the possibility of an outbreak anymore.That this is not a sensible approach can be demonstrated with a very simple example:

A population with $N = 200$ is composed of two sub-populations of equal size A and B. An infective from A infects a susceptible from A with probability 0.003 and a susceptible from B with probability 0.0005 during his total infectious period. Furthermore, an infective from B infects a susceptible from B with probability 0.015 and a susceptible from A with probability 0.0005 during his total infectious period. We can easily calculate: The expected number of secondary cases for a single starting infective of A is $99(0.003)+100(0.0005)=0.347$ and for a starting infective of B it is 1.535. Taking the average over all individual we get 0.941. A value of R_0 below one should indicate that an outbreak is impossible, but in our little example it is clear that an outbreak in group B is likely if the first infective is of group B. And in such a case also several individuals of group A are likely to be infected.

R_0 for non homogeneous models is studied especially by using Markov models since they allow to calculate R_0 exactly in the literature. However, there are some problems with Markov modeling of disease spread. Markov chain processes requires exponential infectious period which is clearly unrealistic. Meanwhile, the complexity of Markov models for non homogeneous populations increases exponentially due to the size of state space, so the exact distribution of $R_0(i)$ can only be calculated for very

small populations. It is clear that even for moderate sized populations the state space for such a model is huge. This makes numerical calculations so difficult that López-García (2016), who considers a similar but continuous time model with exponential infectious period and develops numerical methods to calculate the distribution of important stochastic descriptors, stresses even in the title of the paper that this is only possible for small networks.

Following the simulation literature, we consider a simple discrete time stochastic model. Important is that we allow a very general mixing structure assuming that in a finite population of size N we know all probabilities p_{ij} that within one time-step (in practice typically one day) an infected individual "i" transmits the disease to a susceptible individual "j". It turns out that it is also sometimes necessary to allow the possibility that the infection probabilities change with time. In such cases we will write p_{ijt}.

The estimation p_{ij} for each pair of i and j is possible for small population sizes like hospitals etc. Laskowski *et al.* (2011) implement an agent based modelling for the spread of influenza like disease in an emergency department. They model patients as occupying a circular space with a radius of 60cm and define different contact types like close and casual contacts based on the distance between agents. Moreover, they consider a basic patient flow model throughout which agents come into contact with each other and the probability of infection is found based on the agent distance during contact and the duration of the contact. However, the estimation of p_{ij} is very difficult for large population sizes. The overlapping mixing groups approach is mainly suggested for estimating p_{ij} in large populations. Individual based models for disease spread have been implemented during the last 50 years, but it has been popular recently due to the lack of both data and advanced computational availability in the past Yang *et al.* (2008). Carley *et al.* (2006) propose a scalable city wide multiagent network numerical model where agents are embedded in social, health, and professional networks. The model allows to define heterogeneous population mixing by agent and social networks characteristics based on real data from census, school districts, general social surveys, etc. Bian (2004) presents a conceptual framework for an individual based spatially

explicit epidemiological model based on the following assumptions: (1) individuals are different so age groups are needed; (2) an individual has contacts with a finite number of individuals in different clusters like home and workplace; (3) individuals travel between clusters; and (4) the individuals have different contact rates such as fewer contacts for retired individuals than employed individuals. Thus, two types of contacts are defined: those within a group and those between groups. Moreover, the shift from population based models to individual based models is explained by the rapid improvements in computing power and availability of spatial data. Longini *et al.* (2004) compare the efficiency of the use of anti viral drugs and vaccination for a population with 2,000 persons who are stochastically generated by the age distribution and approximate household size published by the US Census Bureau. Yang *et al.* (2008) study an individual space time activity model for the target city, Eemnes in the Netherlands based on an activity survey data, a synthesized household data, land use data, and PC6 statistical data.

The agent based models collect the infection data at individual level and become more realistic. However, its increased complexity also brings too much a burden for model structure and requires simulation. Moreover, p_{ij}s are required to be computed for individual based models by considering the infection probabilities p_f, p_s, p_n, and p_c in mixing groups of "households", "school and play groups", "neighbourhoods" and "communities", respectively that are also changing with age groups. Then, if infection events between different mixing groups are assumed to be independent, the probability of infection between two individuals i and j during a day is calculated as

$$p_{ij} = 1 - (1 - p_c)^{I_C(j)}(1 - p_n)^{I_N(j)}(1 - p_w)^{I_W(j)}(1 - p_f)^{I_F(j)} \tag{6.1}$$

where the indicator function of a subset is defined as

$$I_M(j) = \begin{cases} 1 & j \in Mixing\ Group\ M\ of\ i \\ 0 & otherwise. \end{cases}$$

That implies that individuals i and j can mix in different mixing groups in set M

(community, neighborhood, school, work, family etc.). So the indicator function is one for several j.

The state of our model is described by the state vector holding the state S, I or R for all N individuals. In one time step a susceptible individual j is infected by a single infected individual i with probability p_{ij}. If there is more than one infected individual the assumption that these infections are independent of each other leads to the total infection probability for individual j :

$$p_j = 1 - \prod_{i \in I} (1 - p_{ij}) \,,$$

where I denotes the set of infected entities. The new infections are thus a sequence of $|S|$ independent Bernoulli trials with probabilities $\{p_j | j \in S\}$, where S denotes the set of all susceptible individuals. To pass from state I to R we use the model assumption that the infectious periods of all individuals are independent and follow a discrete distribution with probability mass function $f_D(d)$ with $d = 1, 2, \ldots$.

For non-homogeneous mixing it is obvious that we need, like suggested in López-García (2016), a definition of R_0 that considers which individual is the single starting infected. As we consider here large populations, we use the simple classical definition of R_0 and define:

$$R_0(i) = \mathrm{E}[\text{secondary cases for starting with a unique infected individual } i]$$

and call it individual R_0.

One important advantage of individual R_0 is that it can be calculated easily also for large populations. To develop the formula we first have to calculate the probability $\widetilde{p}_{ij}$ that susceptible j is infected by infectious i during the total infectious period of i.

This is easily done using conditioning on the infectious period D:

$$\widetilde{p}_{ij} = \sum_{d=0}^{\infty}[f_D(d)(1-(1-p_{ij})^d)]. \tag{6.2}$$

It is also sometimes possible that the infection probabilities change with time written as p_{ijt}. Then, the probability $\widetilde{p}_{ij}$ can be calculated as

$$\widetilde{p}_{ij} = \sum_{d=1}^{\infty}[f_D(d)(1-\Pi_{t=1}^{d}(1-p_{ijt}))]. \tag{6.3}$$

Note that also for an infectious period distribution with unbounded domain it is not difficult to calculate a close approximation of $\widetilde{p}_{ij}$ as the error committed by a cut off of the sum after $d = d_m$ is obviously always smaller than $1-F_D(d_m)$ and can thus be easily controlled. Individual $R_0(i)$ is then simply the "column sum of the matrix $\widetilde{p}_{ij}$" or more precisely the sum of all $\widetilde{p}_{ij}$'s for i fixed and $j = 1, 2, \ldots, i-1, i+1, i+2, \ldots, N$:

$$R_0(i) = \sum_{j:j\neq i} \widetilde{p}_{ij}. \tag{6.4}$$

The complexity of calculating $R_0(i)$ in (6.4) for $i = 1, 2, \ldots, n$ is in total $O(d_m\, N^2)$, where d_m denotes the size of the domain of the infectious period D for bounded infectious period or the cut off value of the infinite sum for the case that D has an unbounded domain.

6.1.1. Use of Individual R_0 on Intervention Analysis

A main aim of building agent based simulation models for influenza spread is the assessment of interventions. How is the spread of the disease changed for instance, when

- 15 percent of all individuals are vaccinated;

- anti-viral drugs are given to all members of a household when one member turns out to be infected;
- when 50 percent of all infected would stay at home after the first day of the disease.

How can the calculation of all $R_0(i)$ values help to asses the behaviour of the disease spread? As we have demonstrated with the help of a simple example above, the average of all $R_0(i)$ values does not allow a direct assessment. But it is easy to see that if $\max_i R_0(i)$ is smaller than one, an outbreak is impossible. If that value is above one the behavior is not certain but an outbreak is possible.

Like for many other interventions also for the second and third intervention example above it is obviously necessary to assume that the p_{ij} values change with time and are denoted by p_{ijt} on day t. The $\widetilde{p}_{ij}$'s are obviously calculated using Equation 6.3. To obtain the $R_0(i)$ we need again the column sums given in (6.4).

To quantify the influence of such interventions it is first necessary to decide how the parameters of the model are changed by the intervention. Here it may be necessary to make assumptions (or guesses) how the infection probabilities are changed; if we consider the case that when 50 percent of all infected would stay at home after their first day of infection is an example where it is clear that people staying at home have infection probabilities of zero with all individuals not belonging to their household.

6.2. Some Non Homogeneous Population Structures for Influenza Spread

It is possible to calculate individual R_0 values exactly for all non homogeneous models using Equation 6.4. In this part, we calculate $\widetilde{p}_{ij}$ and individual R_0 values for some non homogeneous population structures in the literature that are well applicable for airborne diseases like influenza. We need a discrete disease time for influenza and assume like Longini *et al.* (2004) that the probability mass function of 3, 4, 5 and 6 days with probabilities 0.3, 0.4, 0.2, and 0.1 respectively . Moreover, we consider two different non homogeneous population models. Then, in Section 6.3, we evaluate some

intervention methods applied for them.

6.2.1. Model with Multiple Cities

We consider a network of cities around the world connected by transportation. This model is commonly referred as meta population model in the literature suggested by Levins (1968). It includes several sub populations in which perfect mixing is assumed. Individuals travel between the cities leading to disease spread according to probabilistic rules based on the population size and the travel frequency between the cities. Population size and travel data can be obtained from different available sources (e.g. Population Division, U.S. Census Bureau 2004).

Consider now three cities, numbered 1, 2 and 3. Assuming symmetry in travel, we consider a function p_{ij} given in Table 6.1 and compute individual R_0 values by applying Equation 6.4. In big cities, it is standard to assume that R_0 is the same in a homogenous population. Thus, the expected number of individuals infected by a single infected individual is not increasing with the size of the population and the probability to meet and potentially to infect another individual is reduced as the population size increases (Lund *et al.*, 2013). Therefore, we assume greater infection probabilities within a city with smaller population sizes. To obtain the infection probabilities between cities is more complicated and challenging and it is beyond the scope of this work (Lund *et al.*, 2013). Here, we take a simplistic view and assume that the travel frequency is the greatest between city 1 and city 2 and the smallest between city 1 and city 3 by considering the distances between cities and population sizes. Furthermore, the infection probabilities between cities are considered to be around 2 percent of the infection probabilities within the same city. However, it is also possible to obtain travel frequency data for better estimation. Moreover, the reported R_0 values for the basic reproduction number in a fully susceptible population is in the range of 1.6 to 2.4 for influenza (Germann *et al.*, 2006). Thus, while setting the infection probabilities, we target to obtain average R_0 1.7 like in the study of Longini *et al.* (2004). We estimate the infection probabilities by dividing target $R_0 = 1.7$ over expected disease time and the number of susceptible individuals. We also include some super spreaders in this

Table 6.1. Population sizes and infection probabilities for the population with multiple cities.

	Pop. Size	City 1	City 2	City 3	Super Spreaders	$R_0(i)$
City 1	746	5.20e-4	1.30e-5	8.66e-6	1.00e-3	1.673
City 2	500	1.30e-5	7.80e-4	1.04e-5	1.50e-3	1.714
City 3	746	8.66e-6	1.04e-5	5.2e-4	1.00e-3	1.668
Super-spreaders	8	1.00e-3	1.50e-3	1.00e-3	0	9.174

example supposing there are some individuals who often travel. Because individuals in the same city have identical characteristic, the number of different individual R_0 values in this case is equal to the number of cities plus one for the super spreaders. The corresponding individual R_0 values are given in the last column of Table 6.1 computed by using Equation 6.4 consistent with the simulation results.

6.2.2. Model with a Population of Households

We also consider a population partitioned into several households similar to Ball and Lyne (2002) since the household based public health interventions are important to prevent the spread of infectious diseases. Moreover, the two levels of mixing is also important for the behaviour of the epidemic.

Lets consider that an infected individual infects a household member with probability p_h and other individuals with probability p_c. p_h is selected considerably higher than p_c since individuals in the same household have closer contacts. If we denote the family members of individual i as set $N(i)$, the infection probabilities for individual i are

$$p_{ij} = \begin{cases} p_h, & \text{if j} \in \text{N(i)} \\ p_c, & \text{otherwise.} \end{cases} \tag{6.5}$$

We assume that p_h and p_c are 0.0001 and 0.06 respectively for our intervention analysis

and we consider 498 households each consisting of four individuals. Further, there might be some individuals in the population who meet with other people more frequently than other individuals eg. due to their work. We call such people super spreaders. In the literature, super spreaders are defined as the individuals infecting more contacts than others. We assume that each infected super spreader infects with probability $p_s = 0.0008$ and that there are 8 super spreaders in the population.

6.3. Intervention Analysis

Intervention methods aim to change the characteristics of the spread of a disease by changing the infection probabilities p_{ij}. We suggest to assess the impact of intervention strategies by calculating and comparing the individual R_0 values of the different scenarios. We consider the models described in Section 3 where the individuals within the same group are assumed to behave homogeneously. Colizza and Vespignani (2008) define the usual R_0 as a function of disease parameters for each group while a subpopulations reproductive number R_* as a function depending on the diffusion rate of individuals among subpopulations. Thus, a group specific basic reproductive number is considered for a deterministic metapopulation system and the epidemic behaviour on both the global scale and the local scale is determined by R_* and R_0, respectively. Barthélemy *et al.* (2010) consider a stochastic metapopulation model by taking into account both temporal and topological fluctuations. Moreover, individual R_0 computed in this section is also a group specific basic reproduction number by considering both infection among the population members of each group and between the members of different groups instead of two different basic reproduction numbers as in the study of Colizza and Vespignani (2008). However, individual R_0 values can be generalized for every non homogeneous population model like individual based models. Furthermore, we illustrate some numerical results to demonstrate the use of individual R_0 for both developing and assessing intervention strategies including vaccination, social distancing and use of antiviral drugs.

6.3.1. Intervention by Vaccination

For the evaluation of vaccine efficacy, it is assumed that vaccination takes place before the infection starts to spread and that all vaccinated individuals develop immunity. Therefore, vaccinated individuals are not considered as susceptible anymore. For the vaccination as an intervention strategy, it is possible to assume random vaccination in which the individuals who are vaccinated are selected randomly with equal probabilities within the population. However, it is better to use the vaccine efficiently to attain herd immunity by vaccinating a smaller number of individuals.

Ball and Lyne (2002) develop optimal vaccination policies for a population with two levels of mixing and consider optimality in terms of the cost of vaccination program including vaccine, administration, and travel. Here, we propose an intelligent vaccination strategy when assuming that the cost of vaccine is considerably larger than the cost of vaccination. In other words, the aim is to obtain for a fixed number of vaccines the greatest reduction for the maximum individual R_0 value . In this vaccination strategy, individuals with large individual R_0 are vaccinated first because we both eliminate the greatest individual R_0 and obtain the greatest total reduction in the other individual R_0 values if p_{ij}s are symmetric. Therefore, as a next step all individual R_0 values must be recalculated and their values are arranged in non increasing order. Then, the individual who is vaccinated is selected from the top of the list and the individual R_0 values for unvaccinated individuals are recalculated. Thus, our intelligent vaccination policy is to vaccinate individuals one by one choosing the susceptible having the largest individual R_0. By taking the population matrix, *popm* and the target number of vaccinated individuals, $v_{critial}$ as input parameters, the algorithm in Figure 6.1 presents the intelligent vaccination strategy.

In the theory of branching process where m is the expected number of children of each individual, $m < 1$ implies the ultimate extinction with probability one. If a non homogeneous branching process is considered, m values are different for different individuals (Antreya, 2006). If the maximum m is smaller than one, then the process will be also extinct with probability one. Since we consider non homogeneous populations

Algorithm 6
1. *Set* $v = 0$
2. *Compute* $\tilde{p}_{ij}$ using Equation 6.2
3. **for** $i = 1, 2, \ldots, N - v$
4. *Compute* $R_0(i)$ from largest to smallest
5. **end for**
6. *Order* $R_0(i)$ from largest to smallest
7. *Remove individual i with the largest* $R_0(i)$ *from the population matrix*
8. *Set* $v = v + 1$. If $v < v_{critical}$ go to step 3. Otherwise, stop the algorithm.

Figure 6.1. Intelligent Vaccination Strategy

yielding different individual R_0 values, we guarantee that there will be no outbreak by reducing all individual R_0 values below one. The algorithm in Figure 6.1 for the intelligent vaccination policy is a greedy heuristic for heterogeneous populations and it approximates to the optimal vaccination policy as the heterogeneity level decreases.

If we consider the population with three cities described in Section 3.1, the intelligent vaccination strategy requires to vaccinate individuals from different cities. The simulation results indicate that a significant proportion of the population has been infected with probability 0.662 without vaccination. To guarantee that the infection is going to disappear before involving a significant number of the population by implementing Algorithm 6, we observe that individual R_0 values in all cities reduce below 1 if 292 individuals from city 1, 200 individuals from city 2, 290 individuals from city 3 and all 8 super spreaders are vaccinated. Therefore, the minimum number of required vaccinated individuals reducing all individual R_0 values to under 1 is found to be 790 where there will be no outbreak controlled by computing the final outbreak size through simulation. As 'in the city infection probabilities', p_{ii} are considerably greater than 'between the cities infection probabilities', p_{ij} where $i \neq j$ for a model with multiple cities, intelligent vaccination strategy based on sequential vaccination also gives us the optimal vaccination strategy for reducing all individual R_0 values to under one. Let 291 individuals be vaccinated from city 1 instead of 292 individuals, then more than one individual have to be vaccinated from the other cities to decrease individual R_0 value of city 1 below one since p_{11} is considerably greater than p_{12} and

Table 6.2. Individual R_0 based vaccination with different number of vaccinated individuals for the population with multiple cities

	8 Vaccinated Individuals		108 Vaccinated Individuals		208 Vaccinated Individuals		308 Vaccinated Individuals		790 Vaccinated Individuals	
City	Vacc. Ind.	$R_0(i)$	Vacc. Ind.	$R_0(i)$	Vacc. Ind.	$R_0(i)$	Vacc. Ind.	$R_0(i)$	Vacc. Ind.	$R_0(i)$
1	0	1.644	35	1.563	73	1.479	111	1.397	292	0.999
2	0	1.671	32	1.563	56	1.479	81	1.396	200	0.999
3	0	1.639	33	1.562	71	1.480	108	1.397	290	0.999

p_{13}. This also holds for city 2 and city 3. However, it does not always yield the optimal strategy. If we consider an individual based model where each individual has its unique $R_0(i)$, we need to decide which individuals are vaccinated in one step by considering all relationships. Even if it is not possible to vaccinate enough people to reach herd immunity intelligent vaccination is still important in order to have the greatest possible reduction of the individual R_0 values. Table 6.2 shows how individual R_0 values change for an increasing number of vaccinated individuals when using the intelligent vaccination strategy.

For the population partitioned into households described in Section 3.2, the individual R_0 value for the 1992 individuals living in households is 1.509. The sequence of intelligent vaccination starts with the super spreaders. Then, one individual is vaccinated from every family. To reduce the maximum individual R_0 value below 1, vaccination of one individual from every family is not sufficient so the vaccination continues with the vaccination of second individuals from each family. It is easy to calculate that the maximum of individual R_0 values drops below one when two individuals are vaccinated in 139 families while only one individual is vaccinated from the remaining 359 families. The two resulting individual R_0 values are 0.999 (for 1077 individuals) and 0.777 (for 278 individuals). The minimum number of vaccinated individuals required for herd immunity is thus found to be 645 and. Furthermore, Table 6.3 presents the number of susceptibles with their corresponding individual R_0 values if 8, 257, 506 and 755 individuals are vaccinated respectively. Table 6.3 indicates that individual R_0 is 1.483 for unvaccinated individuals if 8 individuals are vaccinated while individual R_0

Table 6.3. Individual R_0 based vaccination with different number of vaccinated individuals for the population partitioned into households

8 Individuals Vaccinated		In 50% families one member vaccinated		In all families one member vaccinated		In 50% families two and in 50% families one member vaccinated	
$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.
0	8	0	257	0	506	0	755
1.483	1992	1.159	747	1.057	1494	0.732	498
-	-	1.381	996	-	-	0.955	747

is reduced to 1.159 and 1.381 for 747 and 996 unvaccinated individuals, respectively if 257 individuals are vaccinated. If 506 individuals are vaccinated, individual R_0 is reduced to 1.057 for all unvaccinated individuals. Moreover, the last two columns of Table 6.3 show that individual R_0s are reduced to 0.732 and 0.995 for 498 and 747 unvaccinated individuals if 755 individuals are vaccinated, so vaccinating more than 645 individuals decreases individual R_0 much lower than 1.

If intelligent vaccination is compared to random vaccination, it is observed that the minimum number of required individuals to be vaccinated to reach herd immunity is much higher for random vaccination and its performance under limited vaccination supply is also clearly worse.

6.3.2. Intervention by Social Distancing

The simplest intervention strategy that can be considered as a method of social distancing is household quarantine. The effectiveness of household quarantine depends on many additional disease parameters like the time between the start of the infection and the start of the symptoms and the compliance rate indicating the percentage of symptomatic influenza cases who remain at home. Household quarantine can be implemented only some time after the infection starts so we assume that it is implemented after the first day of the disease.

Table 6.4. Quarantine after first day of infection for the population partitioned into households

Without Quarantine		Quarantine with Compliance Rate 50%		Quarantine with Compliance Rate 80%	
$R_0(i)$	Number of Ind.	$R_0(i)$	Number of Ind.	$R_0(i)$	Number of Ind.
1.509	1992	1.191	1992	0.999	1992
8.084	8	0	8	0	8

We consider the population partitioned into households only since it is not possible to implement social distancing by the nature of a model with multiple cities. To demonstrate the impact of household quarantine, it is assumed that individuals stay at home after the first day of infection with probability 0.5 suggested in the study of Wu *et al.* (2006). Table 6.4 shows the resulting changed individual R_0 values. The important point in Table 6.4 is the reduction in the individual R_0 values of the household members. Therefore, it is possible to decrease individual R_0 values by increasing compliance rate. It may be possible to increase the compliance rate if a viable diagnostic support including virological testing is available. Thus, we search the compliance rate to attain herd immunity for the household model. We observe that household quarantine must be accepted by at least 80% of the infected to make an outbreak impossible for the population partitioned into households described in Section 3.2.

6.3.3. Intervention by Use of Antiviral Drugs

Antiviral drugs can be both of prophylactic and therapeutic importance. The use of antiviral drugs that is evaluated here prevents infection given exposure. Therefore, it is assumed that antiviral drugs reduce the probability of transmission to others and the probability of being infected given exposure. There are no direct estimates of how much antiviral drug will reduce the probability that an infected individual will develop influenza symptoms compared with an infected person who is not using antiviral drugs but these parameters are inferred from household studies of antiviral drugs in the literature (Longini *et al.*, 2004). Therefore, considering that family members of the initially infected individual use antiviral drugs, we check how their individual R_0 values

Table 6.5. Use of antiviral drugs with different reduction factors without household quarantine for the population partitioned into households

10% Reduction		20% Reduction		30% Reduction		40% Reduction	
$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.
1.448	1992	1.386	1992	1.323	1992	1.258	1992
7.942	8	7.797	8	7.649	8	7.498	8

Table 6.6. Use of anti viral drugs and 50% household quarantine with different reduction factors for the population partitioned into households

10% Reduction		20% Reduction		30% Reduction		40% Reduction	
$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.	$R_0(i)$	Num. of Indiv.
1.130	1992	1.068	1992	1.005	1992	0.940	1992
5.477	8	5.333	8	5.184	8	5.034	8

change for assuming different reduction factors of anti viral drugs. The results in Table 6.5 indicate that, as expected, the effectiveness of the use of anti viral drugs is strongly influencing to the reduction capability for infection probabilities.

Furthermore, we check how effective is the combination of anti viral drugs and household quarantine. The results are given in Table 6.6. We observe that assuming a compliance rate of 50% and reduction rate 40% it is possible to prevent an outbreak by using the combined strategy even if this is not possible when using household quarantine and anti viral drugs alone.

6.4. A Model with Overlapping Mixing Groups

The overlapping mixing group model tries to imitate the disease spread in a real world community using census data. It requires only a moderate number of parameters. The average R_0 for these models are calculated using simulation in the literature (see Longini *et al.* (2004)). The model uses several mixing groups like "households",

Table 6.7. Population matrix for a model with overlapping mixing groups

Individual ID	Family ID	Size of Family	Neighbor ID	Community ID	Age Group	School-Work ID
1	10	1	100	1	6	9001
2	11	2	100	1	6	9001
3	11	2	100	1	3	3001

"school and play groups", "neighbourhoods" and "communities" with their respective infection probabilities p_f, p_s, p_n, and p_c changing with age groups to model all infection probabilities p_{ij} that can be calculated by using Equation 6.1.

Moreover, following Longini *et al.* (2004) we also consider asymptomatic cases for the overlapping mixing group case as a feature of influenza in real world. Asymptomatic cases are the infected individuals who do not have symptoms. Their infection probabilities are also considered to be smaller than the ones for symptomatic cases. The implementation of intervention strategies like household quarantine and the use of anti viral drugs is impossible for them due to lack of symptoms. However, the result of vaccination is not influenced by adding asymptomatic cases. To calculate individual R_0 for the models with both symptomatic and asymptomatic cases, two $\widetilde{p}_{ij}$ values for both the symptomatic and asymptomatic cases have to be calculated using Equation 6.3. Then, $R_{0,s}(i)$ and $R_{0,a}(i)$ are calculated using the corresponding $\widetilde{p}_{ij}$ values. The final $R_0(i)$ values are obtained by taking the weighted average of $R_{0,s}(i)$ and $R_{0,a}(i)$.

In the study of Longini *et al.* (2004), a population of 2000 persons in four identical neighbourhoods is considered. Each individual mixes with people in community, neighbourhood, family and play groups. Family sizes differ between one and seven. We have a similar model in the study of Longini *et al.* (2004) but we also added a mixing group work for adults. We constitute a population matrix each row of which includes the ID of community, neighbourhood, family, school-work and the age group of an individual similar to the rows of Table 6.7. So the number of rows of that population matrix is 2000. The details of the R code for generating such a population matrix based on

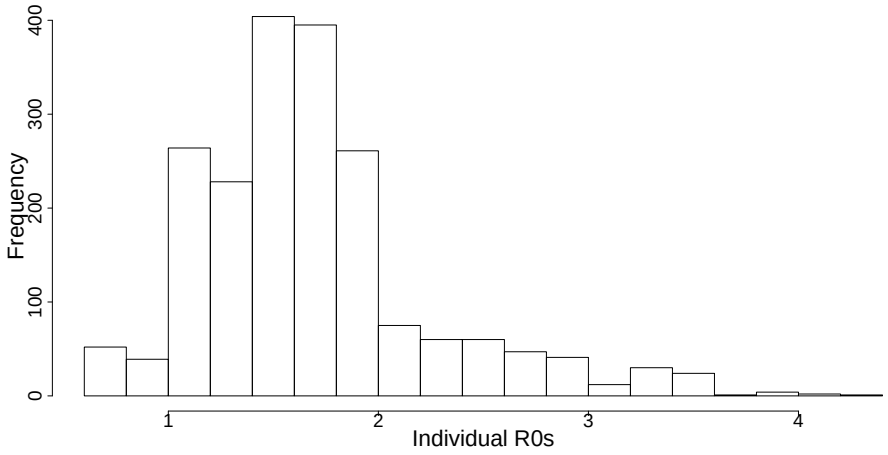

Figure 6.2. The frequency of individual R_0s without household quarantine

census data is available in Appendix A. A major practical problem for this model is the calibration of the probability of infection within each mixing group. We consider the same infection probabilities as in the study of Longini *et al.* (2004). As infectious period, 3, 4, 5 and 6 days with probabilities 0.3, 0.4, 0.2, and 0.1 is again assumed. In the study, we assume that an infected person is symptomatic with probability 0.67 and an asymptomatic infection is only half as infectious as a symptomatic infection.

In Figure 6.2, we present the histogram of all individual R_0 values of the population. The figure indicates that only a small number of individuals in the population have individual R_0 values smaller than 1 while most of the population has individual R_0 values between 1 and 2. Moreover, Longini *et al.* (2004) estimate R_0 as the average of all secondary cases that the randomly selected initial infective person would infect over all mixing groups he belongs to. They empirically calculate R_0 and find it as 1.7 with a range of secondary cases from zero to 17. We compute $R_0(i)$ values with Equation 6.4 where $\tilde{p}_{ij}$ is computed by considering the the probability of infection within each mixing group and the population matrix. Moreover, the average of $R_0(i)$ values for $i = 1, 2, ...2000$ is 1.69, so we compute R_0 suggested in the study of Longini *et al.* (2004) without implementing simulation. Furthermore, there are many possible $R_0(i)$ values giving the same average and the importance of $R_0(i)$ values increase as the heterogeneity level increases.

In studies with overlapping mixing groups, we found only the suggestion of ran-

dom vaccination and of random vaccination of children (Longini *et al.*, 2004; Germann *et al.*, 2006). In these studies, the results of vaccination are analysed estimating attack rates via simulation. However, we suggest to assess the intervention strategies without simulation also for individual based models. Since we can compute individual R_0 value for each member of the population exactly, see the histogram in Figure 1, we can compare the frequency histograms of individual R_0 values after different intervention strategies are implemented. Moreover, simulation is not needed while vaccinating individuals based on their individual R_0 values by implementing the algorithm in Figure 6.1. In this section, we apply to simulation only for random vaccination in order to compare intelligent vaccination strategy with random vaccination where the vaccinated individual is selected randomly. We compare the performance of intelligent vaccination and random vaccination by considering 30%, 50% and 80% of the population vaccinated respectively. Moreover, we record the maximum individual R_0 value for the intelligent vaccination. However, we record the minimum, median and maximum of maximum individual R_0 values for random vaccination because each run yields a different maximum individual R_0 value. In Table 6.8, we present the results.

Table 6.8. Maximum individual R_0 values after random vaccination and vaccination based on individual R_0 without household quarantine

Vaccination Percentage	Min of Maximum Individual R_0s in 1000 Repetitions	Median of Maximum Individual R_0s in 1000 Repetitions	Max of Maximum Individual R_0s in 1000 Repetitions	Max R_0 After Individual R_0 Based Vaccination
0	4.27	4.27	4.27	4.27
30	2.40	3.08	3.73	1.32
50	1.70	2.23	2.98	0.91
80	0.60	0.96	1.54	0.48

Table 6.8 indicates that 50% random vaccination of the population cannot reduce maximum individual R_0 below 1 in 1000 repetitions while 50% vaccination based on individual R_0 values of the population reduces maximum individual R_0 much lower than 1. Thus, similar to the model with multiple cities and the model with a population

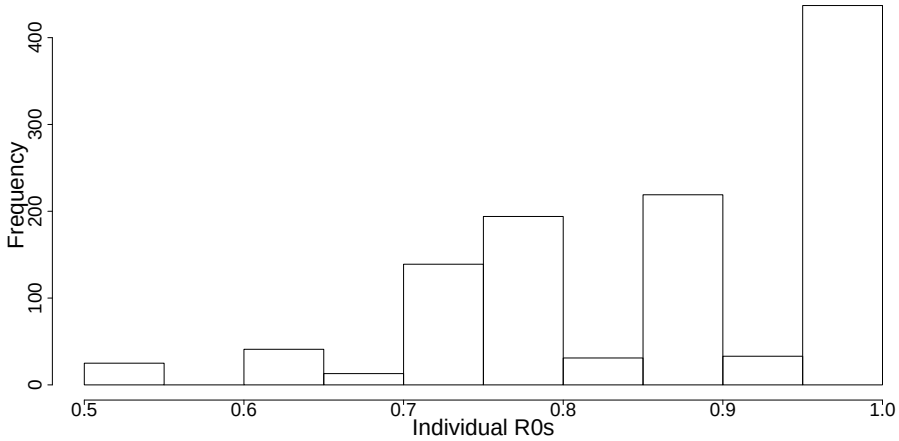

Figure 6.3. The frequency of individual R_0s after vaccination without household quarantine

partitioned into households, we take the advantages of our R_0 formula. Furthermore, the minimum required number of vaccinated individuals to guarantee herd immunity is 869. In Figure 6.3, we present individual R_0 values of the unvaccinated population after vaccination of 869 individuals based on their individual R_0s.

Finally, we also check how individual R_0 values change if 80% of the symptomatic cases stay home after their first day of infection. Figure 6.4 shows that even if a considerable reduction in individual R_0 values is obtained, herd immunity cannot be guaranteed since one third of the infectious cases are considered to be asymptomatic and household quarantine cannot be implemented for asymptomatic cases. This is also true for 100% compliance rate since the actual compliance rate can be at most 67% that is the percentage of symptomatic cases.

The results certainly depend on disease parameters and population structures but we expect that recursive individual R_0 based vaccination gives consistently a better performance. So we can see that the calculation of individual R_0 can be a useful tool to assess the performance of vaccination strategies and also to develop vaccination strategies for stochastic models with arbitrary heterogeneous contact structures.

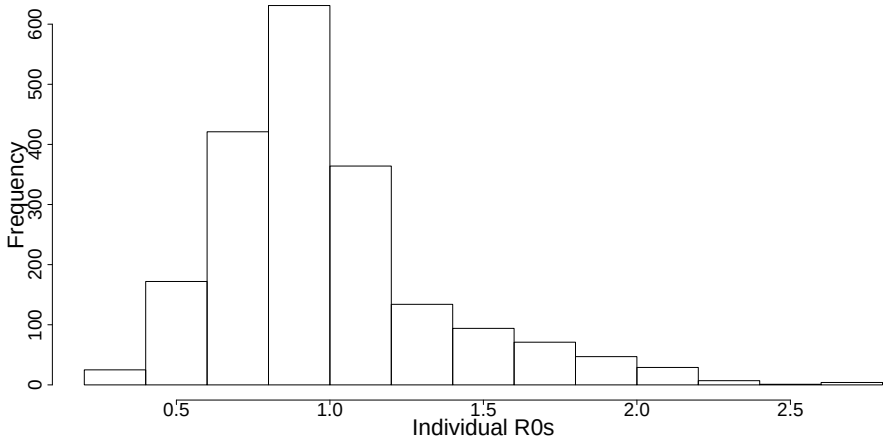

Figure 6.4. The frequency of individual R_0s after 80% household quarantine

6.5. Discussion

In this chapter, we consider a discrete time stochastic SIR model for non homogeneous populations and make three main contributions. Firstly, we introduce individual R_0 and propose a general formula for it that is applicable to all types of heterogeneous populations with any size. Our other major contribution is the assessment of intervention strategies by using the formula for individual R_0 without reverting simulation. Lastly, we define intelligent intervention strategies based on individual R_0 values.

As we have studied the notion of R_0 for non homogeneous populations, we introduced individual R_0 as the expected number of secondary cases produced by a unique given initially infected individual. In the literature, R_0 for non homogeneous populations is either calculated by using Markov chains assuming exponential disease time and small population size or estimated via simulation. Here, we propose a general formula for exact calculation of individual R_0 that is applicable to an arbitrary mixing structure and large population size.

Furthermore, the evaluation of intervention strategies is of practical importance. The effectiveness of these strategies is evaluated by simulation studies comparing the average attack rates or similar characteristics. However, we show that it is possible to assess the impact of the intervention strategies by using directly the individual R_0 formula. We analyze the effectiveness of different intervention strategies by their abil-

ity to decrease the maximum individual R_0 value below one. This method is more accurate than descriptive simulation results to decide how to make an outbreak impossible. However, it is only possible to evaluate strategies that are implemented before infection and immediately after one case is infected by using the individual R_0 values of implementing vaccination, household quarantine or the use of antiviral drugs.

Finally, an intelligent vaccination policy is developed based on individual R_0 values. Here, the aim is to obtain the greatest reduction in the maximum individual R_0 value for a fixed number of vaccines. It is observed that the number of required individuals to be vaccinated for herd immunity is much higher for random vaccination than vaccination based on individual R_0.

7. CONCLUSIONS AND FUTURE RESEARCH

In this thesis, we first considered a stochastic SIR disease spread model with a finite and homogeneous population. For an exponential infectious period we derived formulas for the expected duration of an epidemic, the distribution of the total number of recovered individuals during epidemic, and the maximum number of simultaneously infected individuals.

Because the use of exponential distribution for the infectious period is not realistic, we assume an Erlang distributed infectious period that is a more versatile class of distribution and enables us to use Markov modelling. We propose an efficient computational procedure for the distribution of the final outbreak size and an approximation for the distribution of the maximum number of simultaneously infected individuals. Moreover, we also consider a mixture of Erlangs for the infectious period.

By considering a mixture of Erlangs distribution for infectious period and assuming two types of infected individuals as symptomatic and asymptomatic, we implement our proposed methods for COVID-19 spread. We calculate the average fraction of the individuals who have been infected during the epidemic and approximately determine its maximum size distribution. Moreover, by comparing final outbreak size distributions under different control strategies, we observe that the timing rather than the intensity of intervention is more critical.

Finally, a non homogeneous population is considered for a disease spread model. We investigate the notion of R_0 for heterogeneous populations and introduce individual R_0. After a general formula for R_0 is presented, we suggest to use it for the assessment and development of intervention methods.

The main contributions of the dissertation are given in Section 7.1, and Section 7.2 state the possible future research directions.

7.1. Main Contributions

We can first discuss the contributions regarding Markov modelling of disease spread. Markov modelling of disease spread enables us to obtain exact and computable results for important properties of disease spread. The following features summarize the novel contributions of our study implementing Markov chains for disease spread models.

- Algorithms for important properties of disease spread like expected disease duration, the distribution of the final outbreak size and the distribution of the maximum epidemic size are given for exponential infectious periods.
- For an Erlang distributed infectious period, a model is proposed that uses the total number of *remaining stages* as the state variable. It is a considerable improvement since our state transformation enables us to treat the Erlang distributed infectious period as simply exponential.
- Our method to compute the exact final outbreak size distribution for an Erlang distributed infectious period is implementable for large populations that makes our model a strong competitor to existing methods.
- Our method is extended to a mixture of Erlangs so that by using the first two moments of an infectious period one can easily fit a corresponding mixture.
- An approximation is proposed for the maximum epidemic size distribution for Erlang infectious period that gives exact results for exponential infectious period.

Then, we implement our method for COVID-19 spread by assuming two types of infected individuals as symptomatic and asymptomatic, and we make the following contributions.

- It is better in quantifying uncertainties and accounting for real variabilities compared to deterministic models.
- The incubation and the infectious period can be approximated by using mixing of Erlangs distribution for given mean and variance, so the final outbreak size

distribution is exactly calculated.

- We assess how different timing and intensity of social distancing affect final outbreak size distributions and observe that early interventions for COVID-19 is too important compared to intensity of interventions.

Our disease spread model for non homogenous populations aims to compute the basic reproduction number in heterogeneous populations and assess the effect of intervention methods by using it. We summarize the contributions of our study considering non homogeneous populations as follows.

- We introduce individual R_0 as the expected number of secondary cases produced by a given unique initially infected individual.
- A general formula for individual R_0 is proposed that is applicable to all types of heterogeneous populations with any size.
- Intervention strategies are assessed by using the exact formula for individual R_0 without reverting to simulation. The impact of intervention strategies is assessed by their capability to reduce individual R_0 values.
- Intelligent intervention strategies are developed considering individual R_0 values.

7.2. Possible Future Research Directions

The first future research direction is to investigate the limiting behaviour of our algorithms as the population size increases. Given that a large outbreak occurs, the conditional distribution of the final outbreak size yielding herd immunity can be also investigated.

Focusing on the limitations of Markov disease spread models is another research direction. Markov disease spread models assume a homogeneous population which may be unrealistic for some analysis. We can therefore study an extension where our results could be generalized to non homogeneous models, e.g. models with a household structure of with a number of large sub populations. It might be also interesting to assess the impact of intervention methods by using Markov disease spread models.

Finally, the distribution of individual R_0 can be calculated for non homogeneous populations. Minimizing the cost of intervention methods by using the distribution of individual R_0s is another future research direction.

REFERENCES

Adan, I., and J. Resing, 2002, "Queueing theory", *Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, Netherlands.* .

Ajelli, M., and S. Merler, 2008, "The impact of the unstructured contacts component in influenza pandemic modeling", *PLoS One*, Vol. 3(1), pp. e1519.

Ajelli, M., and S. Merler, 2008, "An individual-based model of hepatitis A transmission", *Journal of Theoretical Biology*, Vol. 259(3), pp. 478–488.

Ajelli, M., B. Gonçalves, D. Balcan, V. Colizza, H. Hu, J. J. Ramasco, S. Merler, and A. Vespignani, 2010, "Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models", *BMC Infectious Diseases*, Vol. 10(1), pp. 190.

Allen, L. J., 2008, *An introduction to stochastic epidemic models: in mathematical epidemiology*, Springer, Berlin, Heildelberg.

Allen, L. J., and A. M. Burgin, 2000, "Comparison of deterministic and stochastic SIS and SIR models in discrete time", *Mathematical Biosciences*, Vol. 163(1), pp. 1–33.

Allen, L. J., and P. van den Driessche, 2013, "Relations between deterministic and stochastic thresholds for disease extinction in continuous-and discrete-time infectious disease models", *Mathematical Biosciences*, Vol. 243(1), pp. 99–108.

Amador, J., and M. Lopez-Herrero, 2017, "Cumulative and maximum epidemic sizes for a nonlinear SEIR stochastic model with limited resources", *Discrete & Continuous Dynamical Systems-B*, Vol. 23(8), pp. 3137.

Amador, J., D. Armesto, and A. Gómez-Corral, 2019, "Extreme values in SIR epidemic models with two strains and cross-immunity", *Mathematical Biosciences and Engineering: MBE*, Vol. 16(4), pp. 1992–2022.

Anderson, D., and R. Watson, 1980, "On the spread of a disease with gamma distributed latent and infectious periods", *Biometrika*, Vol. 67(1), pp. 191–198.

Andersson, H. and T. Britton, 2000, "Stochastic epidemics in dynamic populations: quasi-stationarity and extinction", *Journal of Mathematical Biology*, Vol. 41(6), pp. 559–580.

Andersson, H. and T. Britton, 2012, *Stochastic epidemic models and their statistical analysis*, Springer Science & Business Media, New York.

Andradóttir, S., W. Chiu, D. Goldsman, M. L. Lee, K. L. Tsui, B. Sander, D. N. Fisman, and A. Nizam, 2011, "Reactive strategies for containing developing outbreaks of pandemic influenza", *BMC Public Health*, Vol. 11(1), pp. S1.

Antreya, K. B., 2006, "Branching process", *Encyclopedia of Environmetrics*, Vol. 1.

Artalejo, J. R., A. Economou, and M. J. Lopez-Herrero, 2010, "On the number of recovered individuals in the SIS and SIR stochastic epidemic models", *Mathematical Biosciences*, Vol. 228(1), pp. 45–55.

Artalejo, J. R., A. Economou, and M. J. Lopez-Herrero, 2010, "The maximum number of infected individuals in SIS epidemic models: Computational techniques and quasi-stationary distributions", *Journal of Computational and Applied Mathematics*, Vol. 233(10), pp. 2563–2574.

Artalejo, J. R. and M. J. Lopez-Herrero, 2013, "On the exact measure of disease spread in stochastic epidemic models", *Bulletin of Mathematical Biology*, Vol. 75(7), pp. 1031–1050.

Auchincloss, A. H., and A. V. Diez Roux, 2008, "A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health", *American Journal of Epidemiology*, Vol. 168(1), pp. 1–8.

Bailey, N. T., 1964, "Some stochastic models for small epidemics in large populations", *Applied Statistics*, Vol. 13(1), pp. 9–19.

Balcan, D., H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, C. Vittoria, and A. Vespignani, 2009, "Seasonal transmission potential and activity peaks of the new influenza A (H1N1):

a Monte Carlo likelihood analysis based on human mobility", *BMC medicine*, Vol. 7(1), pp. 45.

Ball, F., 1986, "A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models", *Advances in Applied Probability*, Vol. 18(2), pp. 289–310.

Ball, F., and P. Neal, 2002, "A general model for stochastic SIR epidemics with two levels of mixing", *Mathematical Biosciences*, Vol. 180(1-2), pp. 73–102.

Ball, F. G., and O. D. Lyne, 2002, "Optimal vaccination policies for stochastic epidemics among a population of households", *Mathematical Biosciences*, Vol. 177, pp. 333–354.

Ball, F., T. Britton, T. House, V. Isham, D. Mollison, L. Pellis, and G. S. Tomba, 2015, "Seven challenges for metapopulation models of epidemics, including households models", *Epidemics*, Vol. 10, pp. 63–67.

Bansal, S., B. Pourbohloul, and L. A. Meyers, 2006, "A comparative analysis of influenza vaccination programs", *PLoS medicine*, Vol. 3(10), pp. e387.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, 2004, "The architecture of complex weighted networks", *Proceedings of the National Academy of Sciences*, Vol. 101(11), pp. 3747–3752.

Barthélemy, M., C. Godreche, and J. M. Luck, 2010, "Fluctuation effects in metapopulation models: percolation and pandemic threshold", *Journal of Theoretical Biology*, Vol. 267(4), pp. 554–564.

Bian, L., 2004, "A conceptual framework for an individual-based spatially explicit epidemiological model", *Environment and Planning B: Planning and Design*, Vol. 31(3), pp. 381–395.

Black, A. J., and J. Ross, 2015, "Computation of epidemic final size distributions", *Journal of Theoretical Biology*, Vol. 367, pp. 159–165.

Bobashev, G. V., D. M. Goedecke, F. Yu, and J. M. Epstein, 2004, "A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches", in *2007 Winter Simulation Conference*, IEEE.

Brauer, F., P. Driessche, and J. Wu, 2008, *Lecture notes in mathematical epidemiology*, Springer, Berlin.

Calafiore, G. C., C. Novara, and C. Possieri, 2020, "A modified SIR model for the COVID-19 Contagion in Italy", *arXiv preprint arXiv:2003.14391*.

Carley, K. M., D. B. Fridsma, E. Casman, A. Yahja, N. Altman, L. Chen, B. Kaminsky, and D. Nave, 2006, "BioWar: scalable agent-based model of bioattacks", In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 36(2), pp. 252–265.

Carvalho, S. A., S. O. da Silva, and I. da Cunha Charret, 2019, "Mathematical modeling of dengue epidemic: control methods and vaccination strategies", In: *Theory in Biosciences*, Vol. 138(2), pp. 223–239.

Chen, Y., P. Lu, C. Chang, and T. Liu, 2020, "A Time-dependent SIR model for COVID-19 with undetectable infected persons", *ArXiv Preprint ArXiv:2003.00122*.

Chowell, G., J. M. Hyman, S. Eubank, and C. Castillo-Chavez, 2003, "Scaling laws for the movement of people between locations in a large city", *Physical Review E*, Vol. 68(6), pp. 066102.

Colizza, V., A. Barrat, M. Barthélemy, and A. Vespignani, 2006, "The modeling of global epidemics: Stochastic dynamics and predictability", *Bulletin of Mathematical Biology*, Vol. 68(8), pp. 1893–1921.

Colizza, V., A. Barrat, M. Barthélemy, and A. Vespignani, 2006, "The role of the airline transportation network in the prediction and predictability of global epidemics", *Proceedings of the National Academy of Sciences*, Vol. 103(7), pp. 2015–2020.

Colizza, V., A. Barrat, M. Barthélemy, A. J. Valleron, and A. Vespignani, 2007, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions", *PLoS Medicine*, Vol. 4(1), pp. e13.

Colizza, V., and A. Vespignani, 2007, "Invasion threshold in heterogeneous metapopulation networks", *Physical Review Letters*, Vol. 99(14), pp. 148701.

Colizza, V., and A. Vespignani, 2008, "Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations", *Journal of Theoretical Biology*, Vol. 251(3), pp. 450–467.

Craft, M. E., H. L. Beyer, and D. T. Haydon, 2013, "Estimating the probability of a major outbreak from the timing of early cases: an indeterminate problem?", *PLoS One*, Vol. 8(3), pp. e57878.

Dadlani, A., M. S. Kumar, K. Kim, and F. D. Sahneh, 2016, "Transient analysis of a resource-limited recovery policy for epidemics: A retrial queueing approach", in *2016 IEEE 37th Sarnoff Symposium*, IEEE.

Daley, D. J., and J. Gani, 2001, *Epidemic modelling: an introduction*, Cambridge University Press, Cambridge.

Daniels, H. E., 1974, "The maximum size of a closed epidemic", *Advances in Applied Probability*, Vol. 6(4) pp. 607–621.

Degli Atti, M. L. C., S. Merler, C. Rizzo, M. Ajelli, M. Massari, P. Manfredi, C. Furlanello, G. C. Tomba, and M. Iannelli, 2008, "Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios", *PloS One*, Vol. 3(3), pp. e1790.

Diekmann, O., and J. A. P. Heesterbeek, 2000, *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, John Wiley & Sons, New York.

Dimitrov, N. B., and L. A. Meyers, 2010, "Mathematical approaches to infectious disease prediction and control", *Risk and Optimization in an Uncertain World*, INFORMS.

Dunham, J. B., 2005, "An agent-based spatially explicit epidemiological model in MASON", *Journal of Artificial Societies and Social Simulation*, Vol. 9(1) pp. 17.

Durrett, R., 1999, *Essentials of stochastic processes*, Springer, New York.

Economou, A., A. Gómez-Corral, and M. López-García, 2015, "A stochastic SIS epidemic model with heterogeneous contacts", *Physica A: Statistical Mechanics and its Applications*, Vol. 421, pp. 78–97.

Fackrell, M., 2009, "Modelling healthcare systems with phase-type distributions", *Health Care Management Science*, Vol. 12(1), pp. 11.

Ferguson, N. M., D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke, 2005, "Strategies for containing an emerging influenza pandemic in Southeast Asia", *Nature*, Vol. 437(7056), pp. 209.

Gani, J., and D. Jerwood, 1971, "Markov chain methods in chain binomial epidemic models", *Biometrics*, Vol. 27(10), pp. 591–603.

Germann, T. C., K. Kadau, I. M. Longini, and C. Macken, 2006, "Mitigation strategies for pandemic influenza in the United State", *Proceedings of the National Academy of Sciences*, Vol. 103(15), pp. 5935–5940.

Gibson, G. J., and E. Renshaw, 1998, "Estimating parameters in stochastic compartmental models using Markov chain methods", *Mathematical Medicine and Biology: A Journal of the IMA*, Vol. 15(1), pp. 19–40.

Goldberg, D., 1991, "What every computer scientist should know about floating-point arithmetic", *ACM Computing Surveys (CSUR)*, Vol. 23(1), pp. 5–48.

Gómez, S. and A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, 2010, "Discrete-time Markov chain approach to contact-based disease spreading in complex networks", *EPL (Europhysics Letters)*, Vol. 89(3), pp. 38009.

Gustafsson, L., and M. Sternad, 2008, "Bringing consistency to simulation of population models–Poisson Simulation as a bridge between micro and macro simulation", *Mathematical Biosciences*, Vol. 209(2), pp. 361–385.

Hernandez-Suarez, C. M., 2002, "A Markov chain approach to calculate R0 in stochastic epidemic models", *Journal of Theoretical Biology*, Vol. 215(1), pp. 83–93.

Hernández-Suárez, C. M., C. Castillo-Chavez, O. M. López, and K. Hernández-Cuevas, 2010, "An application of queuing theory to SIS and SEIS epidemic models", *Mathematical Biosciences Engineering*, Vol. 7(4), pp. 809–823.

Hou, C., J. Chen, Y. Zhou, L. Hua, J. Yuan, S. He, Y. Guo, S. Zhang, Q. Jia, C. Zhao, and others, 2020, "The effectiveness of quarantine of Wuhan city against the Corona Virus Disease 2019 (COVID-19): A well-mixed SEIR model analysis", *Journal of Medical Virology*.

House, T., J. V. Ross, and D. Sirl, 2013, "How big is an outbreak likely to be? Methods for epidemic final-size calculation", *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 469(2150), pp. 20120436.

Hufnagel, L., D. Brockmann, and T. Geisel, 2004, "Forecast and control of epidemics in a globalized world", *Proceedings of the National Academy of Sciences*, Vol. 101(42), pp. 15124–15129.

İşlier, Z. G., R. Güllü, and W. Hörmann, 2020a, "An exact and implementable computation of the final outbreak size distribution under Erlang distributed infectious period", *Mathematical Biosciences*, pp. 108363.

İşlier, Z. G., W. Hörmann, and R. Güllü, 2020b, "Assessing intervention strategies for non homogeneous populations using a closed form formula for R0", *ArXiv Preprint ArXiv:2008.05218*.

Inaba, H., 2012, "On a new perspective of the basic reproduction number in heterogeneous environments", *Journal of Mathematical Biology*, Vol. 65(2), pp. 309–348.

Keegan, L. T., and J. Dushoff, 2016, "Estimating finite-population reproductive numbers in heterogeneous populations", *Journal of Theoretical Biology*, Vol. 397, pp. 1–12.

Keeling, M. J., and K. T. D. Eames, 2005, "Networks and epidemic models", *Journal of the Royal Society Interface*, Vol. 2(4), pp. 295–307.

Keeling, M. J., and J. V. Ross, 2007, "On methods for studying stochastic disease dynamics", *Journal of the Royal Society Interface*, Vol. 5(19), pp. 171–181.

Kermack, W. O., and A. G. McKendrick, 1927, "A contribution to the mathematical theory of epidemics", In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, Vol. 115(772), pp. 700–721.

Kühnert, D., T. Stadler, T. G. Vaughan, and A. J. Drummond, 2014, "Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model", *Journal of the Royal Society Interface*, Vol. 11(94), pp. 20131106.

Laskowski, M., B. C. Demianyk, J. Witt, S. N. Mukhi, M. R. Friesen, and R. D. R. McLeod, 2011, *IEEE Transactions on Information Technology in Biomedicine*, Vol. 15(6), pp. 877–889.

Leclerc, M., T. Doré, C. A. Gilligan, P. Lucas, and J. Filipe, 2014, "Estimating the delay between host infection and disease (incubation period) and assessing its significance to the epidemiology of plant diseases", *PloS One*, Vol. 9(1), pp. e86568.

Levins, R., 1968, *Evolution in changing environments: some theoretical explorations*, Princeton University Press, New Jersey.

Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, and X Xing, 2020, "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia", *New England Journal of Medicine*, Vol. 382(13), pp. 1199–1207.

Lipsitch, M., T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray, 2003, "Transmission dynamics and control of severe acute respiratory syndrome", *Science*, Vol. 300(5627), pp. 1966–1970.

Liu, Y., A. A. Gayle, A. Wilder-Smith, and J. Rocklöv, 2020, "The reproductive number of COVID-19 is higher compared to SARS coronavirus", *Journal of Travel Medicine*.

Llyoyd, A. L., 2001, "Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics", *Theoretical Population Biology*, Vol. 60(1) pp. 59–71.

Longini, I. M., M. E. Halloran, A. Nizam, and Y. Yang, 2004, "Containing pandemic influenza with antiviral agents", *American Journal of Epidemiology*, Vol. 159(7), pp. 623–633.

Longini, I. M., A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. T. Cummings, and E. Halloran, 2005, "Containing pandemic influenza at the source", *Science*, Vol. 309(5737), pp. 1083–1087.

López-García, M., 2016, "Stochastic descriptors in an SIR epidemic model for heterogeneous individuals in small networks", *Mathematical Biosciences*, Vol. 271, pp. 42–61.

Lund H., L. Lizana and I. Simonsen, 2013, "Effects of city-size heterogeneity on epidemic spreading in a metapopulation: a reaction-diffusion approach", *Journal of Statistical Physics*, Vol. 151(1-2), pp. 367–382.

Ma, J., and D. J. Earn, 2006, "Generality of the final size formula for an epidemic of a newly invading infectious disease", *Bulletin of Mathematical Biology*, Vol. 68(3), pp. 679–702.

Merler, S., and M. Ajelli, 2009, "The role of population heterogeneity and human mobility in the spread of pandemic influenza", *Proceedings of the Royal Society B: Biological Sciences*, Vol. 277(1681), pp. 557–565.

Meyers, L., 2007, "Contact network epidemiology: Bond percolation applied to infectious disease prediction and control", *Bulletin of the American Mathematical Society*, Vol. 44(1), pp. 63–86.

Milne, G. J., J. K. Kelso, H. A. Kelly, S. T. Huband and J. McVernon, 2008, "A small community model for the transmission of infectious diseases: comparison of school closure as an intervention in individual-based models of an influenza pandemic", *PloS One*, Vol. 3(12), pp. e4005.

Mizumoto, K., K. Kagaya, A. Zarebski, and G. Chowell, Gerardo, 2020, "Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020", *Eurosurveillance*, Vol. 25(10), pp. 2000180.

Nåsell, I., 2002, "Stochastic models of some endemic infections", *Mathematical Biosciences*, Vol. 179(1) pp. 1–19.

Newman, M. E., 2002, "Spread of epidemic disease on networks", *Physical Review E*, Vol. 66(1) pp. 016128.

Nishiura, H., P. Yan, C. K. Sleeman, and C. J. Mode, 2012, "Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks", *Journal of Theoretical Biology*, Vol. 294, pp. 48–55.

Paarporn, K., C. Eksin, J. S. Weitz, and J. S. Shamma, 2015, "Epidemic spread over networks with agent awareness and social distancing", in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE.

Pandey, G., P. Chaudhary, R. Gupta, and S. Pal, 2020, "SEIR and Regression Model based COVID-19 outbreak predictions in India", *ArXiv Preprint ArXiv:2004.00958*.

Patlolla, P., M. Lombardi, V. Gunupudi, A. R. Mikler, and R. T. Jacob, 2004, "Agent-based simulation tools in computational epidemiology", in *International Workshop on Innovative Internet Community Systems*, Springer.

Perez, L., and S. Dragicevic, 2009, "An agent-based approach for modeling dynamics of contagious disease spread", *International journal of health geographics*, Vol. 8(1), pp. 50.

Radulescu, A., and K. Cavanagh, 2020, "Management strategies in a SEIR model of COVID 19 community spread", *ArXiv Preprint ArXiv:2003.11150*.

Read, J. M., J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell, 2020, "Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions", *MedRxiv*, pp. 10.

Ross, S. M., J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, and V. L. Bristow, 1999, *Stochastic processes*, Wiley, New York.

Sharomi, O., and T. Malik, 2017, "Optimal control in epidemiology", *Annals of Operations Research*, Vol. 251(1-2), pp. 55–71.

Simha, A., R. V. Prasad, and S. Narayana, 2020, "A simple stochastic SIR model for COVID 19 infection dynamics for Karnataka: Learning from europe", *ArXiv Preprint ArXiv:2003.11920*.

Singh, S., 2014, *Branching processes in disease epidemics*, Ph.D. Thesis, Cornell University.

Stroud, P. D., S. J. Sydoriak, J. M. Riese, J. P. Smith, S. M. Mniszewski, and P. R. Romero, 2006, "Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing", *Mathematical Biosciences*, Vol. 203(2), pp. 301–318.

Tindale, L., M. Coombe, J. E. Stockdale, E. Garlock, W. Y. V. Lau, M. Saraswat, Y. B. Lee, L. Zhang, D. Chen, J. Wallinga, and others, 2020, "Transmission interval estimates suggest pre-symptomatic spread of COVID-19", *MedRxiv*.

Toda, A. A., 2020, "Susceptible-infected-recovered (sir) dynamics of covid-19 and economic impact", *ArXiv Preprint ArXiv:2003.11221*.

Trapman, P., 2007, "On analytical approaches to epidemics on networks", *Theoretical Population Biology*, Vol. 71(2), pp. 160–173.

Trapman, P., and M. C. J. Bootsma, 2009, "A useful relationship between epidemiology and queueing theory: the distribution of the number of infectives at the moment of the first detection", *Mathematical Biosciences*, Vol. 219(1), pp. 15–22.

Wallinga, J., P. Teunis, and M. Kretzschmar, 2006, "Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents", *American Journal of Epidemiology*, Vol. 164(10), pp. 936–944.

Watson, R., 1980, "A useful random time-scale transformation for the standard epidemic model", *Journal of Applied Probability*, Vol. 17(2), pp. 324–332.

Watts, D. J., R. Muhamad, D. C. Medina, and P. S. Dodds, 2005, "Multiscale, resurgent epidemics in a hierarchical metapopulation model", *Proceedings of the National Academy of Sciences*, Vol. 102(32), pp. 11157–11162.

Wu, J. T., S. Riley, C. Fraser, and G. M. Leung, 2006, "Reducing the impact of the next influenza pandemic using household-based public health interventions", *PloS Medicine*, Vol. 3(9), pp. e361.

Wu, J. T., K. Leung, and G. M. Leung, 2020, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study", *The Lancet*, Vol. 395(10225), pp. 689–697.

Yang, Y., P. Atkinson, and D. Ettema, 2008, "Individual space–time activity-based modelling of infectious disease transmission within a city", *Journal of the Royal Society Interface*, Vol. 5(24), pp. 759–772.

Yang, Z., Z. Zeng, K. Wang, S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, and Z. Mai, 2020, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions", *Journal of Thoracic Disease*, Vol. 12(3), pp. 165.

You, C., Y. Deng, W. Hu, J. Sun, Q. Lin, F. Zhou, C. H. Pang, Y. Zhang, Z. Chen, and X. Zhou, 2020, "Estimation of the time-varying reproduction number of COVID-19 outbreak in China", *International Journal of Hygiene and Environmental Health*, pp. 113555.

Zhao, S., Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, and M. H. Wang, 2020, "Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak", *International Journal of Infectious Diseases*, Vol. 92, pp. 214–217.

APPENDIX A: POPULATION MATRIX GENERATION

In this section, we present the R code to generate a population matrix for overlapping mixing group model. We also explain the code in detail. Let first define the input vectors to calculate probability of infection between a susceptible and an infected (pinf) as

- *EInfComm:* Vector containing expected number of infected in the whole community in each agegroup by a single infected of arbitrary agegroup
- *EInfNei:* Vector containing expected number of infected in Neighborhood in each agegroup
- *EinfSWork:* Vector containing expected number of infected for schoolwork for each agegroup

It is possible to generalize the above using eg. a different EInfComm vector depending on the agegroup of the infected. Thus an *EInfComm* matrix is produced. *EInfected-Family* should be also calculated from the pinf-values for families discussed in Longini et al. (2004) nut this is not a vector but matrix. Lastly, *DFpinf* is a vector of day factors that are used to multiply the all pinf values from day 1, 2 till maximal disease length for the baseline *DFpinf* is assumed to be one.

A model similar to Longini et al. (2004) considering families, neighbourhood, and community: The age codes for the groups pre elementary school, elementary school, middle school, high school, adults and seniors are 1, 2, 3, 4, 5, and 6 respectively. The below code assumes that there are only age group adult and senior (5 and 6) for family size of 1 and two adults, one adult and one senior or two seniors for family size of 2. Moreover, the number of two adults is fixed for families of size greater than 2.

	Size	Fam	Pre	Elem	Middle	High	Adult	Senior
[1,]	1	70	0	0	0	0	35	35
[2,]	2	73	0	0	0	0	100	46
[3,]	3	27	8	7	4	8	54	0
[4,]	4	21	13	12	8	9	42	0
[5,]	5	15	8	12	11	14	30	0
[6,]	6	5	4	6	5	5	10	0
[7,]	7	2	2	3	2	3	4	0

```
sizeage<-t(matrix(c(
,1, 70, 0, 0, 0, 0, 35, 35
,2, 73, 0, 0, 0, 0, 100, 46
,3, 27, 8, 7, 4, 8, 54, 0
,4, 21, 13, 12, 8, 9, 42, 0
,5, 15, 8, 12, 11, 14, 30, 0
,6, 5, 4, 6, 5, 5, 10, 0
,7, 2, 2, 3, 2, 3, 4, 0), nrow=8))
colnames(sizeage)<-c("size","fam","pre","elem","middle",
"high","adult","senior")
IDmL<- list( IDpr=cbind(1000+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(2000+1:3,c(0.5,0.8,1)), IDmi=cbind(3000+1:2,c(0.5,1)),
IDhi= cbind(4000+1:2,c(0.9,1)),
IDwo=cbind(5000+1:10,cumsum(c(0.3,0.2,rep(0.5/8,8)))) )

generateSchoolworkID <- function(nagent,IDNOv,CDF){
# assigns the ID numbers (contained in IDNOv) to ngroup people,
# randomly using according to the probabilities prob
# nagent ...  number of agents for which ID numbers are assigned
# IDNOv ...  vector of the ID numbers that can be assigned
# CDF ...  to assign the ID numbers contained in IDNOv
```

```
d <- length(IDNOv)
if(d==1) return(rep(IDNOv,nagent))
res <- numeric(nagent)
U <- runif(nagent)
res[U<= CDF[1]] <- IDNOv[1]
for(i in 2:d){
res[U>CDF[i-1] & U<= CDF[i]]<- IDNOv[i]
}
res
}
generateSchoolworkID(nagent=20,IDNOv=c(1,7,15),CDF=c(0.5,0.8,1))

mpdatnei<-function(sat=sizeage,IDneig=1,IDcomm=1,
IDmatrixList=NULL){
# IDmatrixList ...  List holding for each for age groups 1 to 5
#a matrix with first Column IDNOv and second column CDF
#TODO make the code general, that for example also kids
#with a single parent can live in a household
sat <- data.frame(sat)
np <- sum(sat$size*sat$fam)
prd <- matrix(ncol=6,nrow=np)
colnames(prd)<- c("agegroup","fsize","IDfam","IDneig",
"IDcomm","IDworkschool")
prd <- data.frame(prd)
prd[,1]<- 5 ;
#all ages set to adult 5, later changed for children and senior
prd[,4]<- IDneig ;
prd[,5]<- IDcomm ; prd[,6]<- NA ;
# IDworkschool, later set based to age; remains NA for senior
nf1 <- sat[1,2];
nadu1 <- sat[1,7];
```

```
nsen1 <- sat[1,8];
prd[1:nadu1,1]<- 5 ; # adult singles
prd[(nadu1+1):nf1,1]<- 6 ; # senior singles
prd[1:nf1,2]<- 1 ; #all have size 1
prd[1:nf1,3]<- 1:nf1 ;
nf2<-sat[2,2];
nadu2 <- sat[2,7];
nsen2 <- sat[2,8];
if(nadu2 >= nf2){
prd[nf1+(1:(nf2*2)),1]<- 5; # all set to adult
prd[nf1+nadu2-nf2+(1:nsen2),1]<- 6;
# couples with both senior or both adult, one couple may be mixed
}
else{
prd[nf1+(1:(nf2*2)),1]<- 6; # all set to senior
prd[nf1+nsen2-nf2+(1:nadu2),1]<- 6;
# couples with both senior or both adult, one couple may be mixed
}
prd[nf1+(1:(nf2*2)),2]<- 2; #family size
prd[nf1+(1:(nf2*2)),3]<- rep(nf1+(1:nf2),each=2) ;# family ID
nf<-numeric(7)
nf[1:2]<-c(nf1,nf2)
for(j in 3:7){
j0 <- sum(sat$fam[1:(j-1)]*(1:(j-1)))
nf[j]<- sat$fam[j]
prd[j0+(1:(nf[j]*j)),2]<- j; #family size
prd[j0+(1:(nf[j]*j)),3]<- rep(cumsum(nf)[j-1]+(1:nf[j]),each=j);
# family ID
#to select the children we just produce the vector holding
#the ages of all children
agec<- rep(1:4,sat[j,3:6])[sample(sum(sat[j,3:6]))]
```

```
# randomly sample from the ages
ci <- j0 + rep((1:nf[j])*j,j-2)-rep(0:(j-3),each=nf[j])
prd[ci,1]<- agec;
}
#schoolworkID
for(j in 1:length(IDmatrixList)){
prd[prdagegroup == j,6] < -generateSchoolworkID(nagent = sum(prdagegroup==j),
IDNOv=IDmatrixList[[j]][,1],CDF=IDmatrixList[[j]][,2])
}
prd
}

res<-mpdatneig(sat=sizeagetab,IDneig=1,IDcomm=1,IDmatrixList=IDmL)
IDmLNei1<- list( IDpr=cbind(1000+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(2000+1:3,c(0.5,0.8,1)), IDmi=cbind(3000+1:2,c(0.7,1)),
IDhi= cbind(4000+1:2,c(0.9,1)),
IDwo=cbind(5000+1:10,cumsum(c(0.3,0.2,rep(0.5/8,8)))))
IDmLNei2<- list( IDpr=cbind(1100+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(c(2101,2102,2003),c(0.5,0.9,1)), IDmi=cbind(c(3101,3002),c(0.6,1)),
IDhi= cbind(c(4101,4002),c(0.8,1)),
IDwo=cbind(5000+1:10,cumsum(c(0.3,rep(0.5/8,8),0.2))))
IDmLNei3<- list( IDpr=cbind(1200+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(c(2201,2202,2003),c(0.5,0.95,1)), IDmi=cbind(c(3201+3002),c(0.8,1)),
IDhi= cbind(4201,4002),c(0.75,1)),
IDwo=cbind(5000+1:10,cumsum(c(rep(0.5/8,7),0.25,0.25,0.5/8))) )

mpdatcomm <- function(sat=sizeagetab,IDneig=1:3,IDcomm=1,IDmLL){
# IDmLL ...  list of school-work IDlist and CDF for each neighborhood
# for which a number is given in IDneig
for(i in 1:length(IDneig)){
new <- mpdatneig(sat=sat,IDneig=IDneig[i],IDcomm=IDcomm,
```

```
IDmatrixList=IDmLL[[i]])
if(i==1) res<-new
else res <- rbind(res,new)
}
res
}
IDmLNei21<- list( IDpr=cbind(21000+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(22000+1:3,c(0.5,0.8,1)), IDmi=cbind(23000+1:2,c(0.7,1)),
IDhi= cbind(24000+1:2,c(0.9,1)),
IDwo=cbind(25000+1:10,cumsum(c(0.3,0.2,rep(0.5/8,8)))) )
IDmLNei22<- list( IDpr=cbind(21100+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(c(22101,22102,22003),c(0.5,0.9,1)),
IDmi=cbind(c(23101,23002),c(0.6,1)),IDhi= cbind(c(24101,24002),c(0.8,1)),
IDwo=cbind(25000+1:10,cumsum(c(0.3,rep(0.5/8,8),0.2))) )
IDmLNei23<- list( IDpr=cbind(21200+1:6,cumsum(c(0.3,rep(0.7/5,5)))),
IDel=cbind(c(22201,22202,22003),c(0.5,0.95,1)),
IDmi=cbind(c(23201,23002),c(0.8,1)),IDhi= cbind(c(24201,24002),c(0.75,1)),
IDwo=cbind(25000+1:10,cumsum(c(rep(0.5/8,7),0.25,0.25,0.5/8))) )
#IDmLNei1
res1 <- mpdatcomm(sat=sizeagetab,IDneig=1:3,IDcomm=1,
IDmLL= list(IDmLNei1,IDmLNei2,IDmLNei3))
res2 <- mpdatcomm(sat=sizeagetab,IDneig=4:6,IDcomm=2,
IDmLL= list(IDmLNei21,IDmLNei22,IDmLNei23))
res2$IDfam <- res2$IDfam +20000
res <- rbind(res1,res2)
```