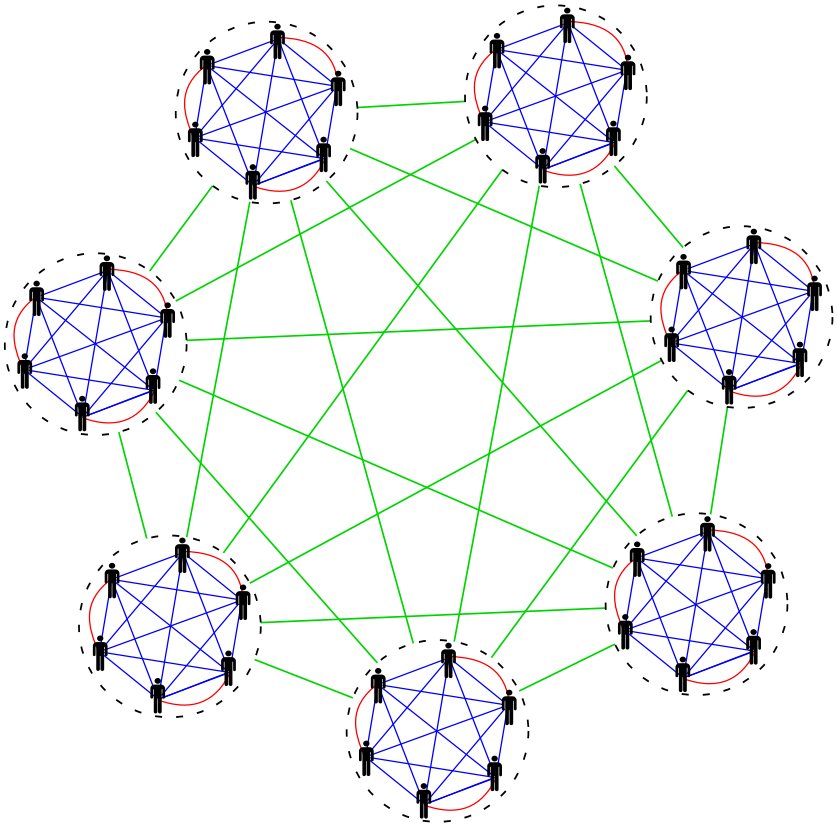
TANNEKE OUBOTER

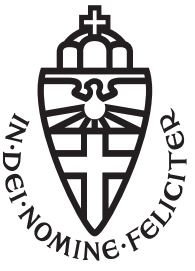
STOCHASTIC EPIDEMIC MODELS FOR POPULATIONS WITH SOCIAL STRUCTURES



STOCHASTIC EPIDEMIC MODELS FOR POPULATIONS WITH SOCIAL STRUCTURES

TANNEKE OUBOTER

Master's thesis
May, 2010



SUPERVISORS:
prof. dr. R.W.J. Meester,
dr. J.P. Trapman

Vrije Universiteit
Amsterdam

SECOND READER:
dr. W. Bosma

Radboud Universiteit
Nijmegen

Preface

In this thesis I have studied the spread of infection diseases within populations. In reality, populations contain many social structures. The degree of intimacy of a contact has a strong influence on the rate of transmission. This has inspired the construction of two mathematical models which take the degree of intimacy into account. At the same time, the models needed to be tractable for mathematical analyses. I have found it interesting to encounter the strength and limitations of mathematical analyses in the field of epidemiology.

To accomplish this thesis, my own social circle was essential. During my study, the most important part of this circle was based in Nijmegen. I started to enjoy mathematics even more because of the good atmosphere at the mathematics department of the Radboud University. I want to thank my dear fellow students for a wonderful study time. Some of you became very good friends. I have really enjoyed discovering the exiting places in Nijmegen and sharing our struggles and enthusiasm for mathematics.
I also want to thank the staff members, in particular Ronald Kortram, Wim Veldman, Klaas Landsman, Mai Gehkre and Wieb Bosma, for the valuable and personal conversations. You have been a great support for me, also in difficult times.
In the last couple of years I got inspired by the lectures of Ronald Meester. This has led me to choose to leave the warm nest of Nijmegen for the big city of Amsterdam. Ronald and Pieter I would like to thank you for being my supervisors and to introduce me to the field of stochastics and epidemiology. Ronald, you have shown me that is essential for a mathematician to be very precise, even if you know the result intuitively. I am grateful for all that I have learned from you. Pieter, from the start you gave me the confidence I needed. You were always there to discuss my progress, doubts and ideas. Even after your move to Stockholm, you were closely involved and you have been a great support. Wieb Bosma was my second reader and advisor from

Nijmegen. Thank you for your willingness to stay informed, and for the pep talks I really needed at times.
The last personal words are for my friends and family. Their loving support was of great value. Most importantly my late parents, Janke and Stefan, who have given me the unconditional love and have taught me so much. And also Bas, who have supported me in so many ways. Thanks for always standing next to me!

Contents

Chapter 1

Introduction

A basic stochastic model for the spread of an infectious disease is the standard SIR, "Susceptible → Infectious → Removed", epidemic model. In this model one assumes a closed homogeneous population, which means that the population is not influenced from outside and that the disease has the same effect on each person. In reality, some individuals have higher infectivity or are more susceptible than others. This heterogeneity of a population has been modeled by Meester and Trapman [10] and by Diekmann and Heesterbeek [5], among others. Another important assumption in the standard SIR model is uniform mixing between the individuals, which means that all individuals meet each other at equal rate. In this thesis, we will drop the assumption of uniform mixing and consider populations which contain a social structure.
The social structure in a population consists of subgroups such as households, schools, workplaces and sports clubs. These social networks overlap and the rate of disease transmission between two individuals depends on the subgroup they both belong to. For instance, you could imagine that a boy and a girl in the same household are more likely to infect each other than people who meet at most once a week in the pub.
Within subgroups, there still is homogeneous mixing. However, if we interlink these structures, a different situation arises. The spread of a disease within a school can then be influenced by the contacts that pupils have at their football club or in their households. How do these local structures connect? And what is the influence on the global spread of a disease? These questions have been the motivation for this thesis.

The first addition is the so-called household model, where only one type

of subgroups is taken into account. Much work has already been done on this [1] [2] [3] [4]. The authors have investigated how household contacts facilitate the global spread of infection.
In this thesis we will consider two social levels, households and schools, where every individual is part of precisely one household and goes to precisely one school. This model can particularly be used to model the spread of childhood diseases, such as measles, rubella and mumps.
There are different ways to interlink the households and schools. We will consider the *Hierarchical* versus the *Random* network model.

- *Hierarchical*: In this model, all children in each household go to the same school. Hence the name Hierarchical: households are fully contained in schools. The relations between the subgroups and their individuals can be represented by a tree, see figure 1.1. The lowest level represents the individuals of the population. On top of this, there is the level of households, and these households are contained in the level of schools.

- *Random*: Here, every household member goes independently of his or her sibling to an arbitrary school. Even though this is far from reality, it contrasts with the Hierarchical model. In addition, because of the uniform distribution, we have the mathematical tools to analyze this.

In both models, there is also a possibility for individuals to meet outside of their households and schools. For instance in the library, or out on the street. We call this highest level the *community*, and assume that all individuals are equally likely to meet each other via *global contact*. This is especially relevant for the Hierarchical model, because there it is the only way to transmit a disease from school to school.

Andersson has also described the Random model for different levels of subgroups in [1], but proofs are not included. We will treat the model more rigorously. The Hierarchical model has received less attention. In [13], the Hierarchical model is proposed as an important extension of the simple homogeneously mixing SIR model, because the hierarchical structure captures the basis framework of how the human population is organized and at the same time, the model remains tractable to analyze. However, their mathematical analysis is very limited.

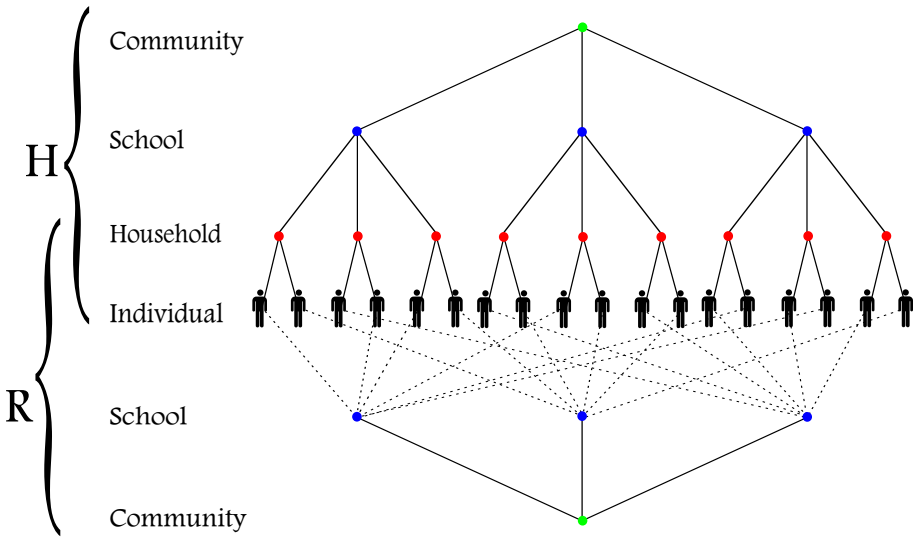


Figure 1.1: This picture shows an example of how individuals can be classified in the Hierarchical and the Random model. In this diagram schools consists of 6 individuals and households are of size 2, but this is only illustrative. The upper part shows the rigid structure of the Hierarchical model (H). In the bottom part, the dashed lines of the Random model (R), indicate that the individuals are uniformly random distributed over the schools.

Overview

In this thesis we will analyze the *Hierarchical* versus the *Random* household-school model and compare some of their outbreak characteristics such as the *expected final size*, the *probability of extinction* and the *reproduction number*. Here the final size is the total number of individuals infected during a large outbreak of a disease. The expected final size shall be determined given that a large outbreak will occur. The probability that the spread of one initial infection does not lead to a large outbreak, is called the extinction probability. The reproduction number R^* is defined as the expected direct infections by one infectious individual.

To compute these outbreak characteristics for large populations, we will observe the spread of a disease in a slightly different way than the original epidemic process evolves. We could let the time dynamics out of considera-

tion, which enables us to approximate the epidemic processes by branching processes. These approximate results are shown to be exact as the population size tends to infinity. We have compared both models on their main characteristics heuristically by proving that in the start of the epidemic, the Hierarchical model is stochastically dominated by the Random model.

In the next chapter, we will first consider infection outbreak characteristics in the standard SIR model. We will see that in a randomly mixing population, the number of infectious individuals grows exponentially in the beginning. We will introduce the branching process and show the relation to SIR models.

Chapter 2

Branching processes

Branching processes, first formulated by Galton and Watson (1874), are used to model the reproduction of a population from generation to generation. Galton and Watson have designed this model to study the extinction of family names. The evolution of a population is represented by a tree (ordered network without loops) where individuals give birth according to a fixed 'offspring distribution', independent of each other. The initial set of individuals is the '0-th generation', their children are called the 'first generation', and so on. One of the main questions in the theory of branching processes is: what is the probability that a population dies out after a certain finite number of generations?

Define μ as the expected number of children of each individual and q as the probability that the population dies out (the extinction probability). When studying the development of an infinitely large random tree network, one can observe a sharp phase transition when μ exceeds the critical number of one. We can formalize this in the following theorem.

Theorem 2.0.1. *When $\mu \leq 1$, the branching process dies out with probability one ($q = 1$), except in the case where every individual produces one child with certainty. When $\mu > 1$, the branching process grows forever with positive probability ($q < 1$).*

(For a proof of this theorem we refer to [9]).

Note that the trivial case where every individual gives birth to exactly one child with probability 1, is a deterministic process and is not interesting for our analysis. So from now on we leave this trivial case out of consideration.

In epidemiology, we are interested in the probability that an infection dies out quickly in a large population. How does this phase transition for μ relate to an epidemic process? The vertices in the branching tree can be interpreted as the infected individuals of the population in the epidemic process, and the edges as the direct infections. We will make use of the terminology of branching processes; we call the set of individuals infected by an infectious individual x_0 the 'offspring' (or 'descendants' or 'children') of x_0.

2.1 Relation with randomly mixing populations

When the population is very large and randomly mixed, the probability that an infective individual contacts an already infected individual during the first stage of the epidemic is very small. So the beginning of the epidemic process can be approximated by a branching process and the reproduction number R^* has the same threshold behavior as μ. In the following Theorem we will make this 'first stage' more formal by giving a lower bound as function of the population size n.

Theorem 2.1.1. *Consider a sequence of uniformly mixing populations growing in their size. For each $\delta > 0$ and $0 < \epsilon < \frac{1}{2}$, there exists a n, such that, within a population of size n, if the total number of infected individuals is less than $n^{1/2-\epsilon}$, then the probability that a loop appears is at most δ. So the start of an epidemic behaves with high probability as a branching process.*

Proof. First we will prove that the probability of no loops in the first k infections is equal to

$$\prod_{i=1}^{k-1} \left(1 - \frac{i}{n-1}\right)$$

The initial infectious individual of the population makes his first contact with a susceptible individual with probability one, since all the others are susceptible in the beginning of the process. Now there are two infectious individuals. The event that the next contact that one of them makes is with a susceptible has probability $1 - \frac{1}{n-1}$, since only one of the other individuals is not susceptible anymore. In the same way, we can show that if the first $k-1$ contacts were all with susceptibles (these contacts result in exactly k infected individuals), then the probability that the k-th contact is with an susceptible, given that this k-th contact occurs, is equal to $1 - \frac{k-1}{n-1}$. So the

joint probability that all the first k contacts are with susceptibles is

$$1 \cdot \left(1 - \frac{1}{n-1}\right) \cdot \left(1 - \frac{2}{n-1}\right) \cdot \ldots \left(1 - \frac{k-1}{n-1}\right) = \prod_{i=1}^{k-1} \left(1 - \frac{i}{n-1}\right)$$

Furthermore, we can prove by induction that

$$\prod_{i=1}^{k-1} \left(1 - \frac{i}{n-1}\right) \geq 1 - \sum_{i=1}^{k-1} \left(\frac{i}{n-1}\right)$$

However, $\sum_{i=1}^{k-1} \left(\frac{i}{n-1}\right) = \frac{k(k-1)}{2(n-1)}$. So if $k < n^{1/2-\epsilon}$ for $0 < \epsilon < \frac{1}{2}$, then $\frac{k(k-1)}{2(n-1)}$ converges to zero as n goes to ∞. We can conclude that for $k < n^{1/2-\epsilon}$,

$$\prod_{i=1}^{k-1} \left(1 - \frac{i}{n-1}\right) \to 1, n \to \infty$$

□

2.2 Generating functions

It will be very convenient to represent the probability distribution of a random variable in a power series, the so called *Probability generating function*. This one-to-one correspondence provides an alternative way for computations with random variables. Below, we will show how this generating function plays a leading role in the proof of Theorem 2.1 in [9], especially for finding the extinction probability. In chapter 6, we will use this tool to calculate some basic outbreak characteristics of the household-school models.

Definition The generating function of a discrete random variable X is defined as

$$f_X(s) := \mathbb{E}[s^X] = \sum_{k=0}^{\infty} \mathbb{P}[X = k] s^k$$

where s is a real variable between 0 and 1.

This function could be used to calculate the mean and the variance of X as follows:

$$\begin{aligned} \mathbb{E}[X] &= f'_X(1) \\ \text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = f''_X(1) + f'_X(1) - f'_X(1)^2 \end{aligned}$$

Let Z_n be the number of infectious individuals in the n-th generation. The probability generating function of the random variables Z_n can be represented by

$$f_{Z_n}(s) = \sum_{k=0}^{\infty} \mathbb{P}[Z_n = k] s^k$$

In a branching process, it is assumed that $Z_0, Z_1, Z_2, \ldots$ form a *Markov chain*: the size of n-th generation only depends on the $(n-1)$-st generation, not on the sizes of generations preceding the $(n-1)$-st. Branching processes have the extra property that the individuals in a particular generation do not interact with each other. So Z_n could be written as a random sum of independent identically distributed ("i.i.d.") random variables $X_1, \ldots, X_{Z_{n-1}}$, all with a common generating function f_X. A proof by induction will show that f_{Z_n} is the n-th iterate of f_X: (Discovered by Watson in 1874 [6])

$$\begin{aligned}
f_{Z_n}(s) &= \mathbb{E}[s^{Z_n}] \\
&= \mathbb{E}\left[\mathbb{E}[s^{Z_n}|Z_{n-1}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[s^{X_1+X_2+\ldots+X_{Z_{n-1}}}]\right] = \mathbb{E}\left[\mathbb{E}[s^{X_1} s^{X_2} \ldots s^{X_{Z_{n-1}}}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[(s^X)^{Z_{n-1}}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[s^X]^{Z_{n-1}}\right] \\
&= \mathbb{E}[f_X(s)^{Z_{n-1}}] \\
&= f_{Z_{n-1}}(f_X(s))
\end{aligned}$$

By induction we get $f_{Z_n}(s) = \overbrace{f_X(f_X(\ldots(f_X(s))\ldots))}^{\text{n times}}$. This iterative relation tells us everything about Z_n: if we know f_X then f_{Z_n} is determined. By using this relation it is also possible to prove the exponential growth (or decrease) of the expected final size of a branching process.

Theorem 2.2.1. *If* $\mathbb{E}[X] = \mu$, *then* $\mathbb{E}[Z_n] = \mu^n$

Proof.

$$\begin{aligned}
\mathbb{E}[Z_n] &= f'_{Z_n}(1) = (f \circ f_{Z_{n-1}})'(1) \\
&= f'(f_{Z_n}(1)) f'_{Z_{n-1}}(1) \\
&= f'(1) f'_{Z_{n-1}}(1) \\
&= \mu \mathbb{E}[Z_{n-1}]
\end{aligned}$$

Applying the iteration gives us the desired result. □

Hence, if the expected number of new infected individuals μ is larger than one, the expectation value of infected individuals grows forever. If it is smaller than one it decays to zero. This statement is consistent with Theorem 2.0.1, but to actually prove a phase transition for the extinction probability, more work is needed [9], see the example below.

Definition By *extinction* we mean the event that the sequence $\{Z_n\}$ consists of zeros for all but a finite number of n and this means that $Z_n \to 0$. Because $\{Z_n = 0\}$ is a monotonously increasing event, we have that

$$q := \mathbb{P}[Z_n \to 0] = \lim_{n\to\infty} \mathbb{P}[Z_n = 0].$$

Example Consider a Markov process $\{Z_n\}$ with the following conditional probability distribution:

$$\begin{aligned}
\mathbb{P}[Z_n = 0 | Z_{n-1} = 0] &= 1 \\
\mathbb{P}[Z_n = 0 | Z_{n-1} = n-1] &= 1 - \frac{\sqrt{n-1}}{\sqrt{n}} \\
\mathbb{P}[Z_n = n | Z_{n-1} = 0] &= 0 \\
\mathbb{P}[Z_n = n | Z_{n-1} = n-1] &= \frac{\sqrt{n-1}}{\sqrt{n}}.
\end{aligned}$$

If we start with $Z_1 = 1$ then the marginal probability distribution for Z_n is given by

$$\mathbb{P}[Z_n = 0] = 1 - \frac{1}{\sqrt{n}} \quad \text{and} \quad \mathbb{P}[Z_n = n] = \frac{1}{\sqrt{n}}.$$

One can observe that the expectation of this sequence tends to infinity while the extinction probability tends to 1, as $n \to \infty$.

The generating function plays an important role in computations of the extinction probability. We make the following observation: if the population of all descendants of a single infectious individual x_0 goes extinct, then either x_0 does not produce new infectious individuals at all, or each of the populations formed by the 'children' of x_0 goes extinct. Note that the number N of 'children' per individuals is an i.i.d. random number, with common generating function f_N. These arguments can be summarized in the following equation

$$q = \sum_{k=0}^{\infty} \mathbb{P}[N = k] q^k = f_N(q). \tag{2.1}$$

Observe that $q = 1$ will always fit, since $\sum_{k=0}^{\infty} \mathbb{P}[N = k] = 1$. In the proof of Theorem 2.1 in [9] it is shown that the extinction probability is equal to the smallest non-negative root of the equation above.

2.3 Multi-type branching processes

A generalization of the single-type branching process is a process that involves several types of individuals. This multi-type branching process can still be described by a Markov process with no interaction between the individuals and some results of the ordinary branching process can easily be extended. The theory of branching processes is comprehensively described in [6]. For completeness, in this subsection we will provide a basic structure of this topic that will be used in chapter 6 for analyzing the Random household-school model.

Suppose we have an appropriate classification of k different types such that every individual produces new types of individuals following a fixed offspring distribution. The states of the Markov process can be denoted in vector notation, using a bold font. Let $\mathbf{Z}_n^i$ be the vector with for each type the number of individuals of that type in generation n infected by an individual of type i, and interpret Z_n^{ij} as the j-th component of this vector i.e. the number of individuals of type j infected by an individual of type i. When it is assumed that the start of the process is the non-random unit-vector $\mathbf{e}^i$, then the probability mass function of the random variable $\mathbf{Z}_1^i$ can be represented by the generating function:

$$f^i(s_1, \ldots, s_k) = \sum_{r_1, \ldots r_k = 0}^{\infty} \mathbb{P}[\mathbf{Z}_1^i = (r_1, \ldots, r_k)] s_1^{r_1} \ldots s_k^{r_k}, \quad 0 \leq \|\mathbf{s}\|_\infty \leq 1.$$

If $\mathbf{Z}_{n-1} = (r_1, r_2, \ldots, r_k)$ then $\mathbf{Z}_n$ can be written as the sum of $r_1 + \ldots + r_k$ random variables, where all r_j random variables are identically distributed. Together they give a family of generating functions $(f_{n-1}^1, \ldots f_{n-1}^k)$ yielding a similarly iterative relation as before:

$$f_n^i(\mathbf{s}) = f^i(f_{n-1}^1(\mathbf{s}), \ldots, f_{n-1}^k(\mathbf{s})). \tag{2.2}$$

The mean number of new infectious individuals of type j infected by one individual of type i defines a $k \times k$ reproduction matrix M [6] with

$$M_{ij} := \mathbb{E}[Z_1^{ij}] = \frac{\partial f_i(1,\dots,1)}{\partial s_j} \qquad i,j = 1,\dots k.$$

By the *Perron-Frobenius* theorem we know that if the matrix M is *positive regular* (i.e. there exists a $N > 0$ such that all entries in M^N are strictly positive) then there exists a positive real eigenvalue ρ of this matrix M such that all other eigenvalues λ are strictly smaller in absolute value $|\lambda| < \rho$. By iteration we have that ρ^N is the largest eigenvalue of M^N. The positive regularity assumption means in our case that we have to assume that for every i there exists a j such that $\mathbb{P}[Z_1^{ij} = 0] > 0$.
This eigenvalue ρ has the same threshold behavior as μ has in the single-type branching process: if $\rho \leq 1$ then all eigenvalues are smaller than or equal to one and the epidemic dies out, if $\rho > 1$ then there is at least one *direction*, i.e. the direction of the eigenvector associated with ρ, in which the epidemic grows forever with a probability larger than zero. By *direction* we mean the scalar multiples of the particular vector. We conclude that ρ is a valid threshold measure and we will use it as reproduction number for the multi-type branching process.

The extinction probability for the multi-type branching process can be derived similarly to equation (2.1). Let $q(i)$ be the probability that the epidemic dies out given that the initial infective individual is of type i and let $\mathbf{f} = (f^1,\dots f^k)$ be the generating function. From Theorem 7.1 in [6] we know that $\mathbf{q} = (q(1),\dots q(k))$ is the solution of $\mathbf{f}(\mathbf{s}) = \mathbf{s}$ with smallest Euclidean norm. Similar to Theorem 2.0.1: if $\rho > 1$, then $q(i) < 1$ for all i and if $\rho \geq 1$, then $q(i) = 1$ for all i.
The mean extinction probability value can then easily be computed by a weighted average over the different types:

$$\bar{q} = \sum_{i=0}^{k} \mathbb{P}[Z_0 = e^i] q(i). \tag{2.3}$$

Chapter 3

Set-up for the household-school models

In the previous chapter we briefly discussed the relationship between branching processes and the spread of epidemics within uniformly mixing populations. In this chapter we will give a formal set-up of the household-school models, where individuals are not uniformly mixed anymore. To analyze the progress of an epidemic across these structures, figure 1.1 is not practical, therefore we will introduce a multi-layered graph representation. By doing *percolation* on the different levels of these graphs, we will show that the Hierarchical and the Random model both can be approximated by a (multi-type) branching process. Let us start with a formal description of the standard simple SIR model.

3.1 Standard SIR model

The standard stochastic SIR model assumes a closed population (no births, deaths and migration are considered) that is homogeneous, randomly mixing. The individuals in the population are at first 'susceptible' and after they get infected they remain 'infectious' for some period of time. During their *infectious period*, each infected individual makes contact with a given individual at the time points of a Poisson process with rate $\frac{\beta}{n-1}$, where n is the size of the population and β is called the *infection rate*. In a large population, this rate can also be interpreted as the mean number of ikndividuals an infectious individual will infect during a certain time frame, say one time unit. Contact between an infectious and a susceptible individual always results in transmission of infection. When the infectious period has

terminated, the individual remains immune to the infection for the rest of his life and is considered as *removed*, he or she is not part of the epidemic process anymore. The epidemic ceases if all the infectious individuals are removed. Hence the name SIR (Susceptible, Infectious, Removed).

3.2 Household-school models

We consider a population of n individuals, organized in a social network of households and schools, where each individual belongs to exactly one household and also to exactly one school. We assume that the school size, denoted by n_S, and household size, denoted by n_H, are relatively small compared to n. They are held fixed when $n \to \infty$. For ease of presentation we assume that the epidemic is initiated by one infectious child x_0 and that the other $n-1$ are initially susceptible. We also assume that the infectious period I is constant and the same for every individual, say $I = 1$ [time unit], such that the analyses become much more convenient.
The infectious contacts take place following the time points of a Poisson process. In the household-school models, the individuals may transmit the disease at three different levels. They make *household* contact with a given sibling at a Poisson rate β_H, *school* contact with a given schoolmate at a Poisson rate β_S. Finally, each individual can infect all $n-1$ other individuals by global contact in the community. This rate between an infectious and a given individual is set $\frac{\beta_G}{n-1}$ in order to keep the total contact rate β_G independent of the population size. All of these contacts between an infectious and a susceptible individual always result in immediate infection transmission and the rates are defined per pair of individuals. In Theorem 2.1.1 we have seen that for a large homogeneous population, the probability of contacting a given individual more than once tends to zero, as n goes to infinity. So for large n, β_G can be interpreted as the mean number of globally infected individuals, infected by one infectious individual. However this is not true for the contact rates within the households and schools. Since their sizes are relatively small compared to n, there is a substantial probability that within these subgroups an infectious individual contacts an already infected individual such that the disease will not transmit by this contact. From now on we will make a difference in terminology: a "contact" made by an infectious individual will only result in infection transmission if the receiver is susceptible. When it is given that the receiver is susceptible, we say that the infectious individual "infects" this individual and we call this an "infectious contact".

By definition of a Poisson process, the waiting time between two contacts is exponentially distributed. The cumulative exponential distribution function with rate β and waiting time X is defined as: $\mathbb{P}[X \leq t] = 1 - e^{-t\beta}$. We just assumed that one infectious contact is enough to transmit the infection. Therefore, by the assumption that the infectious period $I = 1$, the probability that an infectious individual infects a given individual within his household (resp. school and community) is given by $\bar{p}_H := 1 - e^{-\beta_H}$ (resp. $\bar{p}_S := 1 - e^{-\beta_S}$, $\bar{p}_G := 1 - e^{-\frac{\beta_G}{n-1}}$). So if we consider the spread of the disease restricted within the subgroups, we assume that the individuals are still randomly mixed. However, we are interested in how the spread of the disease behaves if we interlink these overlapping networks.

We will model the progress of an epidemic across a merged graph consisting of three different layers, where the edges represent the possible contacts or connections between the individuals within the different subgroups, see figure 3.1. A complete green graph represents the connections within the community. Households are represented by complete subgraphs with red edges and schools are represented by complete blue subgraphs. Notice that only the color of the graph, not its geometric distance, determines the rate of infection. To determine the basic characteristics of an epidemic, like the reproduction number R^*, the final size T and the extinction probability q, we are not interested in the precise time evolution, but only in the final outcome of the epidemic. Therefore, we could model the epidemic spread by a (bond) percolation model. Here, the time dynamics have been dropped and only the static cluster of infected individuals is considered. Since we assumed a fixed infectious period, the infections made by the same individual are independent. We will describe this model below.

Percolation model

In a *bond* (resp. *site*) percolation model on an infinite network structure, edges (resp. vertices) are *open* with probability p and *closed* with probability $1 - p$, independently of each other. An open edge means in our case that the connection could be used for transmitting infection. Such an edge will stay part of the network and a closed edge will be deleted. Typical questions that can be answered in percolation theory are: does the remaining graph of open edges have an infinitely large connected subgraph (also called infinite component)? And what is the critical value $0 \leq p_c \leq 1$ such that the probability of an infinitely large component, called the *survival* probability, is 0 if $p < p_c$, and strictly larger than 0 for $p > p_c$? If there exists an

infinitely large component of open edges with probability greater than zero, then what is the probability that our initial infectious individual x_0 is part of it? Like the reproduction number R^*, this phase transition p_c is an important threshold measure. Percolation theory works on infinite networks to obtain sharp phase transitions. However, in real life, structures are always of finite size. Therefore, we are interested in the asymptotic behavior of sequences of finite graphs where the population size n grows to infinity. In chapter 5, we will study this limiting behavior in more detail. The relation between epidemiology and percolation theory was earlier described in [12] [11] and gives us an important tool for modeling the spread of a disease.

We want to describe a procedure that builds a connected component of infectious individuals. In the bond percolation model we can construct such a component along the way. We start with examining all edges of different colors that are connected to our initial infectious vertex x_0. Each edge is open with a corresponding edge probabilities $\bar{p}_G$, $\bar{p}_S$ or $\bar{p}_H$, depending on the subgroup where it belongs to. The green (resp. blue, red) edge from x_0 to a given individual w is drawn if and only if x_0 will infect w by global (resp. household, school) contact during its infectious period. In this way, it is possible that w is connected to his sibling x_0 by a red, blue and green edge as well, for example. The initial infectious individual itself (x_0) is called 'generation 0'. The other endpoints of the open edges connected to x_0 are called 'generation 1'. In the next iteration, we move on to one of the infected vertices in the first generation, say x_1. Explore all the edges connected to x_1 and repeat the procedure for those edges. All the endpoints of the newly explored edges with one endpoint in generation 1, are called 'generation 2', except for those vertices of previous generations (they were already infected before). Continuing these steps for every generation results in a final set of the epidemic represented by a directed cluster. Note that this cluster has a tree structure if and only if any individual will be at most infected once.

Fixed versus random infectious period

In general, we have to perform percolation on a directed graph since the state of each edge depends on the infectious period of its starting point. However, throughout this thesis, we shall assume a constant infectious period. In this case it does not matter whether the edges are directed or not, because the event that the edge from v to w is open is independent of the 'state' of all the other edges, in particular it is independent of the state of the edge from w to v [12]. Here, the edges $v \to w$ and $w \to v$ are open

with the same probability This means that during the exploration of the cluster, if v becomes infected earlier than w then we only have to explore the edge from v to w, and visa versa. So for a fixed infectious period we could drop the direction of the edges without any consequence for size of the percolation cluster. The proofs in chapter 4 and 5 will essentially be based on the criteria that the network is undirected.

Until so far, the household-school models were described together. However, how the three layers are mixing up is different for the Random and the Hierarchical model, so from now on we shall treat them separately. In the next chapter we will see how we the final size within the schools and households could be determined in both models. The resulting distributions of the final subgroup sizes shall be used in chapter 6 to approximate the overall epidemic spread.

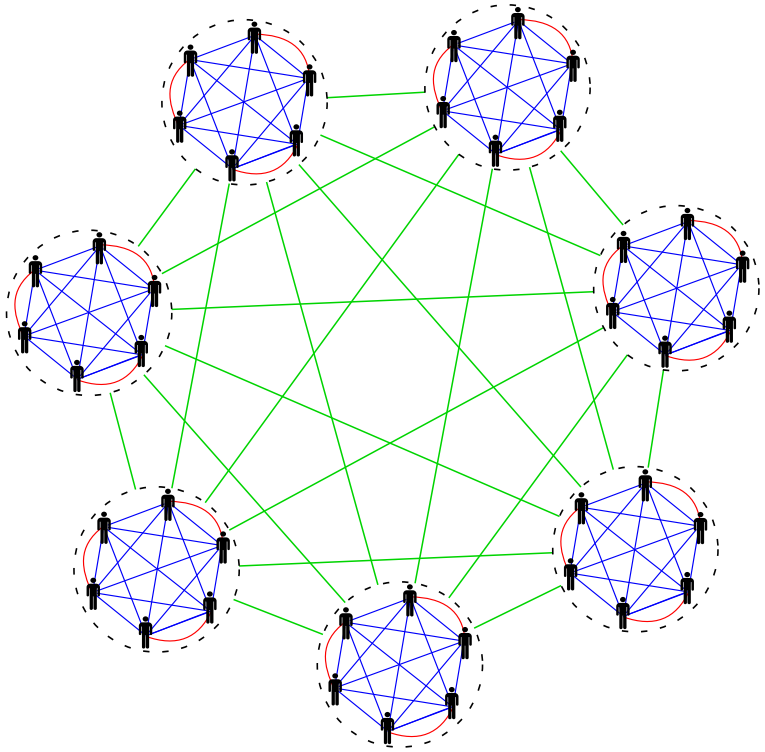


Figure 3.1: This is almost the same situation as in figure 1.1 but here the Hierarchical structure is shown from another perspective. Each green connection between two different schools represents the 36 global connections between all pairs of individuals at these two schools. For simplicity, only blue edges are shown, where there also is a green connection. To obtain the Random model from this picture, you can think of an arbitrary formation of households. In this example these households are of size two.

Chapter 4

The final size within small finite groups

In this chapter, we will calculate the *exact*, albeit implicit, probability distribution of the final epidemic size within a (small) finite population, using the recursive formula of Theorem 2.2 in [2]. We shall give a probabilistic proof of this formula for a randomly mixing population, which will be needed for the Random model in chapter 6. Note that for large finite sets, the calculations become computationally hard. In our household-school model, the largest finite subgroups are schools, so we have to assume them not to be larger than say fifty pupils. Secondly, we will give a rather similar proof for a generalized equation, stated in chapter 6 of [2], and we will apply this formula to the hierarchical household-school structure. Note that the calculations can also be used for an epidemic with random infectious period, but then the proofs are less intuitive.

4.1 Random model final size

Theorem 4.1.1. *Consider a standard SIR epidemic model, starting with a population of size $n+1$ where one individual is infectious and the other n are initially susceptible. We assume that the infectious period is fixed. We define p as the infection probability and P_k^n as the probability that the final size of the epidemic is equal to k, $0 \leq k \leq n$. Note that the final size does* ***not*** *include the initial infectious individual. Then for each l, $0 \leq l \leq n$*

$$P_l^n = \left((1-p)^{(n-l)(l+1)}\right)\binom{n}{l} - \sum_{k=0}^{l-1} P_k^n \left((1-p)^{(n-l)(l-k)}\right)\binom{n-k}{l-k} \quad (4.1)$$

Proof. We model the progress of an epidemic across a complete graph of size $n+1$, where every pair of vertices (individuals) is connected by an open (resp. closed) edge with probability p (resp. $1-p$). The final set of the epidemic, denoted by $\mathbf{T}$, could be compared with the percolation cluster of the initial infectious individual x_0. We stress that here the final set $\mathbf{T}$ does not include the initial infectious individuals.

The proof is based on a specific way of counting. Denote by $\mathbf{N}$ the total set of individuals excluding x_0. Number the individuals of $\mathbf{N}$ by $\{1,\ldots,n\}$ and consider the event that the final set of the epidemic is equal to $\mathbf{L} = \{1,2,\ldots,l\}$. The probability on this event will be denoted by $P_{\mathbf{L}}^n$. Since the population is homogeneous and uniformly mixed, there are $\binom{n}{l}$ ways to choose a set of l elements. So we have that $P_l^n = \binom{n}{l} P_{\mathbf{L}}^n$, and it is suffices to show that

$$P_{\mathbf{L}}^n = (1-p)^{(n-l)(l+1)} - \sum_{k=0}^{l-1} P_k^n \left[(1-p)^{(n-l)(l-k)}\right]\binom{n-k}{l-k} \Big/ \binom{n}{l} \quad (4.2)$$

First we remark that $\{\mathbf{T} = \mathbf{L}\}$ occurs, if and only if the following two events will occur:

$\mathcal{B}$: All of the $l+1$ elements of set $\mathbf{L} \cup \{x_0\}$ fail to infect any of the other individuals in set $\mathbf{N} \backslash \mathbf{L}$. We will call this a closed border of $\mathbf{L}$.
$\mathcal{C}$: All individuals of $\mathbf{L}$ become infected: $\mathbf{L} \subseteq \mathbf{T}$.

So

$$P_{\mathbf{L}}^n = \mathbb{P}[\mathcal{B} \cap \mathcal{C}] = \mathbb{P}[\mathcal{B}] - \mathbb{P}[\mathcal{B} \cap \mathcal{C}^c]$$

We observe that $\mathcal{B}$ has probability $(1-p)^{(n-l)(l+1)}$ since all $(n-l)(l+1)$ connections between the two subsets are closed with probability $(1-p)$, independently of each other, because a fixed infectious period is assumed.

Secondly, we write

$$\mathcal{B} \cap \mathcal{C}^c = \mathcal{B} \cap (\cup_{k=0}^{l-1} \cup_{\mathbf{S}_k \subset \mathbf{L}} \{\mathbf{T} \cap \mathbf{L} = \mathbf{S}_k\})$$

where we have taken the union over all possible sub-epidemics of size k within $\mathbf{L}$, for all $0 \leq k \leq l-1$, and $\mathbf{S}_k$ represents a particular (proper) subset of size k.
To express $\mathbb{P}[C^c]$ in P_k^n, we recall that P_k^n represents the probability that the final set of the epidemic within the total population is of size k. We are only interested in the probability that the final k infectious individuals are all within set $\mathbf{L}$. Because all individuals have equal probability to become infected we have to multiply by a fraction $\binom{l}{k}/\binom{n}{k}$: there are $\binom{l}{k}$ ways to choose k elements within $\mathbf{L}$, divided by the total number of ways to choose k elements, so only a fraction $\binom{l}{k}/\binom{n}{k}$ of sets of size k are a subset of $\mathbf{L}$. However, this fraction is exactly the same as $\binom{n-k}{l-k}/\binom{n}{l}$, only the way of choosing l and k is reversed. Here, we first fix a set $\mathbf{S}_k$ and next we choose the other $l-k$ elements such that $\mathbf{S}_k \subset \mathbf{L}$, divided by the total number of possible ways of choosing l elements. So $\binom{n-k}{l-k}/\binom{n}{l}$ is the fraction of sets of size l which contain subset $\mathbf{S}_k$.

Hence, for a given k, the probability that the epidemic results in a final set of k elements that are contained in $\mathbf{L}$, is given by

$$P_k^n \binom{n-k}{l-k} / \binom{n}{l} \tag{4.3}$$

Notice that now there automatically is a closed connection between these k elements and the elements of $\mathbf{N} \backslash \mathbf{L}$: it is included in the probability P_k^n. To determine the probability that $\mathcal{B} \cap \mathcal{C}^c$ occurs, we need a closed connection between all elements of $\mathbf{L}$ and $\mathbf{N} \backslash \mathbf{L}$. The probability that the other $l-k$ edges of the border of $\mathbf{L}$ are closed is given by

$$(1-p)^{(n-l)(l-k)} \tag{4.4}$$

Since the infectious period is fixed, connections are closed independently of each other. Hence, we can multiply (4.3) by (4.4) to obtain the desired result.

□

Theorem 2.2 in [2] is an immediate result of Theorem 4.1.1.

Corollary 4.1.2.

$$\sum_{k=0}^{l} \binom{n-k}{l-k} P_k^n / [(1-p)^{(n-l)(k+1)}] = \binom{n}{l} \tag{4.5}$$

Proof.

$$\sum_{k=0}^{l} \binom{n-k}{l-k} P_k^n / \left((1-p)^{(n-l)(k+1)}\right) =$$
$$\sum_{k=0}^{l-1} \binom{n-k}{l-k} P_k^n / \left((1-p)^{(n-l)(k+1)}\right) + P_l^n / \left((1-p)^{(n-l)(l+1)}\right) =$$
$$\sum_{k=0}^{l-1} \binom{n-k}{l-k} P_k^n / \left((1-p)^{(n-l)(k+1)}\right) + \binom{n}{l} - \sum_{k=0}^{l-1} \binom{n-k}{l-k} P_k^n / \left((1-p)^{(n-l)(k+1)}\right)$$
$$= \binom{n}{l}$$

□

4.2 Hierarchical model final size

In the Hierarchical structure, the individuals within schools do not mix uniformly at random anymore. Remember that the spread of infection within a given school can be modeled across a graph with two different layers, depicted in figure 3.1: one layer consists of a complete blue graph with edge probability $p_S = 1 - e^{-\beta_S + \frac{\beta_G}{n-1}}$ representing the school and global contacts, on top of this there are red edges between every pair of siblings, open with probability $\bar{p}_H := 1 - e^{-\beta_H}$.
We will subdivide the school-population into different types of individuals, called a *type assignment*, such that the infection probability between any individual of type i and type j, denoted by p_{ij}, is the same. Since we assume a fixed infectious period, p_{ij} and p_{ji} are automatically the same.

First, in Theorem 4.2.1, we will introduce the final size distribution for a general subgroup divided in different types. We will use the vector notation: let $\mathbf{v} \leq \mathbf{n}$ mean $v_i \leq n_i$ for all $i \leq k$,

$$\binom{\mathbf{n}}{\mathbf{v}} = \prod_{i=1}^{k} \binom{n_i}{v_i} \qquad \text{and} \qquad \sum_{\mathbf{u}=0}^{\mathbf{v}} = \sum_{u_1=0}^{v_1} \cdots \sum_{u_k=0}^{v_k}$$

Theorem 4.2.1. *Consider a population, subdivided into k different types, of size $\mathbf{n} = (n_1, \ldots, n_l + 1, \ldots n_k)$, where one individual of type l is initially infectious and the other individuals of the population are initially susceptible. Denote by p_{ij} the infection probability matrix and by $P_{\mathbf{u}}$ the probability that the final size of the epidemic is equal to $\mathbf{u} = (u_1, \ldots, u_k)$, $0 \leq \mathbf{u} \leq \mathbf{n}$. The δ_l-function indicates the initial infectious individual, $\delta_l(i) = 1$ if and only if $i = l$, otherwise $\delta_l(i) = 0$. Note that the final size does not include the initial infectious individual. Then for each $\mathbf{v}$, $0 \leq \mathbf{v} \leq \mathbf{n}$:*

$$\sum_{\mathbf{u}=0}^{\mathbf{v}} \binom{\mathbf{n}-\mathbf{u}}{\mathbf{v}-\mathbf{u}} P_{\mathbf{v}} / \prod_{i=1}^{k} \left(\prod_{j=1}^{k} (1-p_{ij})^{(n_j - v_j)} \right)^{u_i + \delta_l(i)} = \binom{\mathbf{n}}{\mathbf{v}} \tag{4.6}$$

We will prove this theorem for the two level-mixing case where we have schools of size n_S and households of size 2. One can imagine that it is not difficult to generalize this to larger household sizes and to more levels of mixing subgroups, only the computations will become more tedious, so we will not deal with this in this thesis. Below, we shall define an appropriate type assignment for the case of households of size 2.

Type assignment

In the Hierarchical model, individuals can make contact both at school and at home, each with a certain probability. To make the appropriate type assignment, we have to distinguish the events where siblings make household contact and where they do not. By household contact we mean that **if** one of the siblings will be infected, he or she will automatically infect his or her sibling. Specifically, before we explore the actual spread of a disease across the network, we can first perform percolation only on the red graph. Here, an open red edge between two siblings indicates that household contact will occur and these events are independent of the possible school-infections. Based on this percolation we can define the different types of individuals. Individuals of *type 2* make household contact with their household member, individuals of *type 1* do not. However, this definition will not result in an appropriate type assignment: if an individual of type 2 is infectious, then he will infect his sibling (also of type 2) with probability one but other individuals of type 2 with probability p_S. Therefore, we have to observe **households** of type 2 instead of **individuals** of type 2, while we observe individuals of type 1. By this type assignment we are able to define a contact

probability matrix p_{ij}, where the infection spreading still only depends on transmission within school. Observe that between all pairs of households of type 2 there are 4 blue connections through which it is possible to transmit the disease, between each pair of individuals of type 1 and households of type 2 there are 2 blue connections and between two individuals of type 1 there is only one way to transmit. Together, this results in the following matrix:

$$\begin{aligned} p_{11} &= p_S := 1 - e^{-\beta_S} \\ p_{12} &= 1 - (1 - p_S)^2 = 2p_S - p_S^2 \\ p_{21} &= 2p_S - p_S^2 \\ p_{22} &= 1 - (1 - p_S)^4 \end{aligned}$$

Proof of Theorem 4.2 in the case of two different types. Consider a school of size n_S. Denote by n_1 (resp. n_2) the random number of individuals (resp. households) of type 1 (resp. 2). As we have mentioned above, we perform percolation on the red graph to obtain a certain type partition. This is a binomial process, where we have $\frac{n_S}{2}$ number of red edges (=number of households) and each edge is closed with probability $\bar{p}_H$, independently of each other. So $n_1 \sim 2 \cdot \text{bin}(n/2, 1 - \bar{p}_H)$ and $n_2 \sim \text{bin}(n/2, \bar{p}_H)$

Consider a given realization $(\bar{n}_1, \bar{n}_2)$ of these binomial processes, and suppose that our initial infectious individual is of type 1. We construct a new complete graph where the vertices of a graph can be distinguished in three different sets: the set of individuals of type 1 (excluding the initial infective) denoted by $\mathbf{N}_1$, the set of households of type 2 denoted by $\mathbf{N}_2$, and the initial infective of type 1. The edge probabilities are determined by the contact probability matrix as defined above.
We number the individuals of type 1 and define for each $v_1 \leq \bar{n}_1$, the subset $\mathbf{V}_1 \subset \mathbf{N}_1$ to be the set $\{a_1, \ldots, a_{v_1}\}$. We can do the same for **households** of type 2 and define for each $v_2 \leq \bar{n}_2$, the subset $\mathbf{V}_2 \subset \mathbf{N}_2$ to be the set $\{b_1, \ldots, b_{v_2}\}$. Since there are $\binom{\bar{n}_1}{v_1}\binom{\bar{n}_2}{v_2}$ ways to choose v_1 elements out of $\bar{n}_1$ and v_2 elements out of $\bar{n}_2$, we have $P_{(v_1,v_2)} = \binom{\bar{n}_1}{v_1}\binom{\bar{n}_2}{v_2} P_{(\mathbf{V}_1,\mathbf{V}_2)}$

We could argue in the same way as for the homogeneous case, and gen-

eralize equation (4.2) to the hierarchical setting described above.

$$
\begin{aligned}
P_{(\mathbf{V}_1,\mathbf{V}_2)} \quad = \quad & \left(q_{11}^{n_1-v_1} q_{12}^{n_2-v_2}\right)^{v_1+1} \left(q_{21}^{n_1-v_1} q_{22}^{n_2-v_2}\right)^{v_2} - \\
& \left(\sum_{u_1=0}^{v_1-1} \sum_{u_2=0}^{v_2} P_{(u_1,u_2)} \left(q_{11}^{n_1-v_1} q_{12}^{n_2-v_2}\right)^{u_1-v_1} \left(q_{21}^{n_1-v_1} q_{22}^{n_2-v_2}\right)^{u_2-v_2} \right. \\
& \left. \cdot \binom{n_1-u_1}{v_1-u_1} / \binom{n_1}{v_1} \cdot \binom{n_2-u_2}{v_2-u_2} / \binom{n_2}{v_2} \right) - \\
& \sum_{u_2=0}^{v_2-1} P_{(v_1,u_2)} \left(q_{21}^{n_1-v_1} q_{22}^{n_2-v2}\right)^{u_2-v_2} \cdot 1 / \binom{n_1}{v_1} \cdot \binom{n_2-u_2}{v_2-u_2} / \binom{n_2}{v_2}
\end{aligned}
$$

Here, the *border* of $\mathbf{V}_1 \cup \mathbf{V}_2$ is closed with probability $(q_{11}^{n_1-v_1} q_{12}^{n_2-v_2})^{v_1+1} (q_{21}^{n_1-v} q_{22}^{n_2-v_2})^{v_2}$ since the set $\mathbf{N} \backslash (\mathbf{V}_1 \cup \mathbf{V}_2)$ splits up into two smaller sets: the individuals of type 1 and the households of type 2. Further, we have again subtracted all events where the final set of size (u_1, u_2) is a sub-epidemic of $(\mathbf{V}_1, \mathbf{V}_2)$. We have assumed that the initial infectious individual is of type 1, but we could repeat the argument for the case that the initial infectious is of type 2, except that the final size will then be at least *one*.
It is now straightforward to complete the proof. □

By using Theorem 4.2.1 and 4.1.1 we can solve the final size probabilities in the subgroups recursively. In figure 4.1, we have plotted these probabilities for the Hierarchical model, where we have chosen a specific school size and infection rates. One can see that either a few individuals or a considerably large part of the school population becomes infected. This *bimodal* behavior [2] becomes more evident as n grows large, and we will prove this in the next chapter.

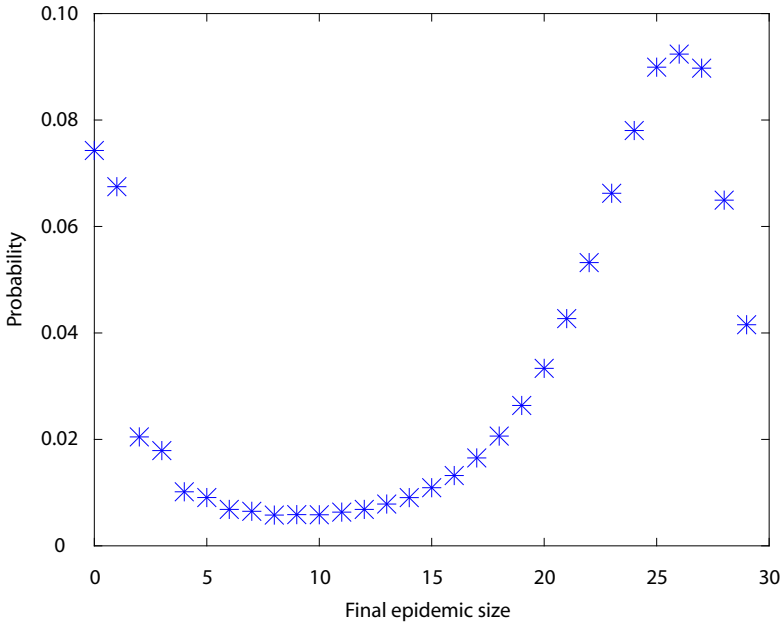


Figure 4.1: The final size distribution within a Hierarchical school of size 30 where households are of size 2. In this example, the contact-per-pair rates are $\beta_H = 1.4$ and $\beta_S = 0.1$

Chapter 5

Asymptotic behavior

One of the advantages of considering the limit of an epidemic spread within large populations that grow in size, is that only two possible scenarios can occur: the number of new infections in generation n either goes to 0 with *extinction probability* q or it goes to ∞ with probability $1 - q$. However, in reality, an infectious disease spreads within a finite population. For large finite populations, exact computations such as in chapter 4 are intractable, so we are interested in the asymptotic behavior of sequences of finite network structures that grow in size. We shall see that the sharp phase transition we observe in the limit of a sequence of randomly mixing populations, is a good indication for the phase transition in those large finite populations. Furthermore, we will see how the extinction probability and the final size are related to each other.

5.1 Intuitive introduction

In Theorem 2.1.1 we have seen that a homogeneously mixing population can be approximated by a branching process until at most $n^{1/2-\epsilon}$ (for $\epsilon > 0$) individuals are infected and removed. But what can we say about the final epidemic size, what happens after the branching approximation breaks down? In this chapter we will show that the survival probability of the disease (one minus the extinction probability) within a large finite homogeneously mixing population converges to the survival probability of the corresponding branching process. Moreover, we will show that when the reproduction number is larger than one, the proportional final size will also converge to that survival probability. The proofs are based on the lecture notes of Van der Hofstad [7].

In section 3 we have constructed the final set of the epidemic by exploring the percolation cluster step by step, starting with the initial infectious. However, we could also do this the other way around: we start with percolation on the whole network, and after that choose one node uniformly at random as starting point of the epidemic and call this vertex x_0. Since we assume a fixed infectious period, the distribution of the final cluster size is independent of its starting point. So we can choose our starting point uniformly at random. We denote the cluster of a node v by $C(v)$, and its size by $|C(v)|$. In the literature, performing percolation with edge probability p on the complete network of size n can be viewed as the same stochastic process as the Erdös Rényi graph with parameters (n, p). One remark on terminology is that *cluster* and *component* will have the same meaning.

Relation between final size and extinction probability

The main theorem of this chapter can informally be stated as: when the Poisson infection rate β is larger than one, then the component of maximum size, denoted by $C_{\max}$, is of order n with high probability, and the other clusters are much smaller, they will be at most of order $\log(n)$. Van der Hofstad has also presented a proof that if $\beta < 1$ then $C_{\max}$ is of order $\log(n)$, but this proof is not included in this thesis.

For $\beta > 1$, the result of the main theorem directly implies a relation between the final size and the extinction probability. Because we choose our starting point of the epidemic uniformly at random, and because the connected 'large' component of order n is unique, we have that the probability on a 'large' outbreak is exactly $\frac{|C_{\max}|}{n}$, i.e. the probability that we choose x_0 in $C_{\max}$. We conclude that for $\beta > 1$, the proportion of removed (i.e. eventually infected) individuals converges to the survival probability $(= 1 - q)$, as n goes to ∞.

New perspective on branching and epidemic processes

It is common to study the descendants of a branching process from generation to generation, but for our purposes in this chapter, it will be more convenient to construct a branching process by sequentially exploring the number of children of each member of the population [7]. We will describe it formally.

Consider a branching process $\{X_t\}$ with a Poisson-β offspring distribution. In the sequential construction, each individual can have three possible statuses: *neutral*, *active* and *inactive* (compare these with the epidemic labels, *susceptible*, *infectious* and *removed*). These statuses will change during the exploration of the connected component.
We will use S_t to denote the number of active vertices at time t. We start at $t = 0$ with a single initial infectious individual. This individual is called *active*, all other individuals are initially *neutral*. This means that $S_0 = 1$. We explore the children of this initial infectious individual, denoted by X_1. At $t = 1$, the initial infectious individual is set to the *inactive* status, since he is already explored, we thus have $S_1 = X_1$. In the next time step we move on to one of the active individuals, and increase t by 1. Then we explore his children, denoted by X_{t+1}. This means that after t time steps, we have S_t active and t inactive individuals (t individuals are explored). We can repeat this procedure until no *active* individuals are left over, then the epidemic is extinguished. By induction we get the following formalization:

Definition Let $X_1, X_2, \ldots$ be a sequence of independent Poisson-β random variables. Define the number of active vertices at time t as

$$S_t := S_{t-1} + X_t - 1 = X_1 + \ldots X_t - (t-1).$$

Then the total offspring of the branching process is given by

$$T = \min\{t : S_t = 0\} = \min\{t : X_1 + \ldots X_t = t - 1\}.$$

Within a randomly mixing population of size n, an epidemic process with contact rate β can also be described sequentially. The only difference is that the sequence of $X_1, X_2, \ldots$ is not i.i.d. anymore, where X_i represents the individuals infected by the i-th explored active individual. At each 'exploration time' t, the number of new infections X_t depends on the number of susceptible (neutral) individuals, denoted by N_t. Note that $N_t = n - S_{t-1} - (t-1)$. As mentioned in section 3.2, when the infectious period is assumed to be constant, the probability that an infectious individual infects a given susceptible is given by $p := 1 - e^{\frac{\beta}{n-1}}$. Therefore, conditionally on S_{t-1} we have:

$$X_t = \text{bin}(n - S_{t-1} - (t-1), p). \tag{5.1}$$

Recalling Theorem 2.1.1, we have that the sequence $X_1, X_2, \ldots$ is almost i.i.d. as long as the number of infectious (active) and removed (explored) individuals is not too large.
Further, we can see that N_t is binomial distributed as well, but with another

success parameter. Intuitively, every individual except the first infectious one, has independently of all other vertices, a probability $(1-p)^t$ to stay susceptible in the first t explorations which results in:

$$N_t = \text{bin}(n-1, (1-p)^t). \tag{5.2}$$

By using $S_t + (t-1) + N_t = n-1$, we have that the complement of N_t is also binomial distributed:

$$S_t + (t-1) = \text{bin}(n-1, 1-(1-p)^t). \tag{5.3}$$

The formulation of an epidemic process in a set of binomial distributed random variables turns out to be helpful in the proof below.

5.2 Formal proof of bimodal behavior

In this chapter we will prove the following main theorem:

Theorem 5.2.1. *Fix $\beta > 1$. Then for every $\nu \in (\frac{1}{2}, 1)$, there exists a $\delta(\nu, \beta) > 0$ such that*

$$\mathbb{P}[||C_{max}| - n\theta_\beta| \leq n^\nu] \geq 1 - n^{-\delta}.$$

We start by investigating the number of vertices in connected components of size at least $K\log(n) =: k_n$, denoted by

$$Z_{\geq k_n} = \sum_v \mathbf{1}_{\{|C(v)| \geq k_n\}}.$$

First we will show that $Z_{\geq k_n}$ contains at least a positive fraction of the population as n goes to infinity. In particular, this fraction will converge to the survival probability of the corresponding branching process.
Secondly, we will see that the points in $Z_{\geq k_n}$ are in fact in the same unique *giant* component, i.e. a connected subgraph that contains the majority of vertices of the entire graph.

In lemma 5.2.2 we will evaluate the expected size of $Z_{\geq k_n}$. This lemma can also be interpreted as an extended formulation of the branching approximation. Earlier we have shown that until $\sqrt{n}$ individuals are infected and removed, the epidemic process behaves like a branching process, with high probability. Lemma 5.2.2 tells us that the bound of $\sqrt{n}$ is not so tight but could be replaced by a more general function k_n, where $k_n/n \to 0$ and

$k_n \geq K \log(n)$, as $n \to \infty$ and K is 'large enough'.

Consider a randomly mixing population of size n. We fix a Poisson-β contact rate $\beta > 1$, and a constant infectious period $I = 1$, yielding an edge probability $p = 1 - e^{-\frac{\beta}{n}}$, as mentioned in section 3.2. We denote by $\mathbb{P}_\beta$ (resp. $\mathbb{P}_{n,p}$) the probability measure on the epidemic process where contacts are made according to a Poisson-β (resp. Binomial(n,p)) distribution. Note that for large n, these Poisson-β and Binomial(n,p) distributions are close to each other. When the context is clear, we will omit the subscripts. Let θ_β be the *survival probability* of the corresponding branching process (with the same model parameters and probability measure denoted by $\mathbb{P}^*_\beta$ and $\mathbb{P}^*_{(n,p)}$). Recall that θ_β is the probability that a given node is part of an infinitely large cluster and that it is equal to $1 - q_\beta$.

We will use the big O notation to describe the limiting behavior as n goes to infinity. Formally, $f(n) = O(g(n))$ as $n \to \infty$ if and only if there exists a positive real number M and a real number N such that for all $n > N$, $|f(n)| \leq M|g(n)|$.

Lemma 5.2.2. *For a large randomly mixing population as described above, growing to infinity, there exists a K, such that for all $k_n \geq K \log(n)$ and for every node v:*

$$\mathbb{P}_\beta[|C(v)| \geq k_n] = \theta_\beta + O\left(\frac{k_n}{n}\right). \tag{5.4}$$

Proof. First we will prove that the final size of an epidemic process is stochastically dominated by the final size of a branching process (with the same infection rate) yielding a sharp upper bound on $\mathbb{P}_\beta[|C(v)| \geq k_n]$. Secondly, we will prove that until k_n individuals are infected and removed an epidemic process can be bounded from below by a branching process with an infection rate which depends on k_n. We emphasize that the processes are non-spatial, which means that every individual has the same probability to be infected, independently of its geometric distance to the infectious individuals.

Upper bound:
Recall the sequential construction, mentioned in chapter 3. Let X_i and X_i^* denote the stochastic offspring of i-th explored individual in the epidemic process and branching process respectively. We can write

$$X_i^* = \text{bin}(n-(S_{i-1}+(i-1)),p)+\text{bin}(S_{i-1}+(i-1),p) = X_i+\text{bin}(S_{i-1}+(i-1),p).$$

Consider a realization $\bar{X}_1, \bar{X}_2, ...$ of the epidemic process with starting point x_0, and for each i add extra active individuals to $\bar{X}_i$ following a bin$(S_{i-1} + (i-1), p)$ distribution. The resulting offspring sequence $\bar{X}_1^*, \bar{X}_2^*, ...$ can be viewed as realization of the bin(n, p)-branching process. By this construction we have $\{|C(v)| \geq k_n\} \subseteq \{T \geq k_n\}$ such that we can conclude that $\mathbb{P}_\beta[|C(v)| \geq k_n] \leq \mathbb{P}^*_\beta[T \geq k_n]$.

Furthermore, we note that

$$\mathbb{P}^*_\beta[T \geq k_n] = \mathbb{P}^*_\beta[T = \infty] + \mathbb{P}^*_\beta[k_n \leq T \leq \infty].$$

So it is suffices to show that $\mathbb{P}^*[k_n \leq T \leq \infty] = O\left(\frac{k_n}{n}\right)$:

$$\begin{aligned}
\mathbb{P}^*[k_n \leq T \leq \infty] &\leq \sum_{t=k_n}^{\infty} \mathbb{P}^*[S_t^* = 0] \\
&= \sum_{t=k_n}^{\infty} \mathbb{P}^*[X_1^* + \ldots + X_t^* = t-1] \\
&\leq \sum_{t=k_n}^{\infty} \mathbb{P}^*[X_1^* + \ldots + X_t^* \leq t].
\end{aligned}$$

Where in the first inequality we have used that $T = t$ implies that $S_t = 0$. By using the Markov inequality, we can give an upper bound on $\mathbb{P}^*[X_1^* + \ldots X_t^* \leq t]$, also known as the *Chernoff* bound. Note that the sequence $X_1^*, X_2^*, \ldots$ is i.i.d. and that $\mathbb{E}^*[X_i^*] = \beta > 1$.
For every $s \geq 0$ we have:

$$\begin{aligned}
\mathbb{P}^*[X_1^* + \ldots + X_t^* \leq t] &\leq \mathbb{P}[e^{s\sum_{i=1}^t X_i^*} \leq e^{st}] \\
&\leq e^{-st}\mathbb{E}[e^{s\sum_{i=1}^t X_i^*}] \\
&= e^{-st}\left(\mathbb{E}[e^{sX_1^*}]\right)^t \\
&= \left(e^{-s+\log(\mathbb{E}[e^{sX_1^*}])}\right)^t \\
&\leq e^{-t\sup_{s\geq 0}(s-\log(\mathbb{E}[e^{sX_1^*}]))}.
\end{aligned}$$

In the second inequality, we have used the Markov inequality. Minimizing the right hand side over all $s \geq 0$, results in an upper bound which is exponentially decreasing in t, and this is precisely what we want to obtain. Since X_1^* is Poisson-β distributed we have

$$\mathbb{E}^*[e^{sX_1^*}] = \sum_{n=0}^{\infty} e^{-\beta}\frac{\beta^n}{n!}e^{sn} = e^{\beta(e^s-1)}. \tag{5.5}$$

Hence,

$$\sup_{s\geq 0}(s - \log(\mathbb{E}[e^{sX_1}])) = \beta - 1 - \log(\beta) =: I_\beta > 0.$$

Now, we can complete the upper bound. For all $k_n > (I_\beta)^{-1}\log(n)$ we have

$$\begin{aligned}\mathbb{P}^*[k_n \leq T \leq \infty] &\leq \sum_{t=k_n}^{\infty} e^{-tI_\beta} \\ &\leq \frac{e^{-k_n I_\beta}}{1-e^{-I_\beta}} \leq Ce^{-\log(n)} = O\left(\frac{1}{n}\right).\end{aligned}$$

Lower bound:
We will again use a coupling argument to obtain a lower bound on $\mathbb{P}[|C(v)| \geq k]$. For each k, we could couple the epidemic process until k individuals are infected, to a branching process with a $\text{bin}(n-k,p)$ offspring distribution, where the total offspring is denoted by T_L. The big difference with above is that this coupling explicitly depends on k.

First, we will show that for all k, $\{T_L \geq k\} \subseteq \{|C(x_0)| \geq k\}$:
Consider percolation on the complete graph. We will explore the connected component of an epidemic process and a $\text{bin}(n-k,p)$ branching process simultaneously. Recall that in the epidemic process the individuals can have three possible statuses: *neutral*, *active* and *inactive*. For this coupling we will need an extra status, some neutral vertices will be classified as *forbidden*. In the branching process we will not explore the edges connected to these forbidden individuals, such that we can held the number of *allowed* vertices fixed to $n-k$, where the *allowed* vertices are the neutral vertices that are not forbidden. Note that this can be realized until k individuals are active and inactive, then we stop the exploration and know that the event $\{|C(x_0)| \geq k\}$ occurs.
Number the individuals of the population, and start with one active vertex x_0. Initially, classify the vertices $\{n-k+2,\ldots,n\} =: F_1$ as forbidden, such that $|F_1 \cup \{x_0\}| = k$. For the epidemic process we explore *all* edges connected to the neutral vertices, where every edge is independently occupied with probability p. For the branching process we exclude the edges that are connected to the forbidden vertices. Every time that a neutral vertex is found to be occupied, we make the forbidden vertex with the largest index neutral. This keeps the number of allowed vertices fixed to $n-k$, such that the number of children of a given individual is $\text{bin}(n-k,p)$ distributed. By this construction, as long as the number of active and inactive vertices is at

most k, the branching cluster contains less points then the epidemic cluster. So we can conclude that $\{T_L \geq k\}$ only occurs if $\{|C(x_0)| \geq k\}$ occurs, and this proves the claim.

Consider the $\text{bin}(n-k_n, p)$ branching process, where p is defined by $p := 1 - e^{-\frac{\beta}{n}}$. This process can be approximated by a Poisson-branching process with infection rate $\beta_n := \frac{\beta}{n}(n-k_n)$ and survival probability θ_{β_n}.
We have that for all $k_n > I_\beta^{-1}\log(n) > I_{\beta_n}^{-1}\log(n)$:

$$\begin{aligned}\mathbb{P}^*[|C(v)| \geq k_n] &\geq \mathbb{P}^*[T_L \geq k_n] \\ &= \theta_{\beta_n} + \mathbb{P}^*[k_n \leq T_L \leq \infty] \\ &= \theta_{\beta_n} + O\left(e^{-k_n I_{\beta_n}}\right) = \theta_{\beta_n} + O\left(\frac{k_n}{n}\right).\end{aligned}$$

We claim that $q_{\beta_n} = q_\beta + O\left(\frac{k_n}{n}\right)$ which automatically implies $\theta_{\beta_n} = \theta_\beta + O\left(\frac{k_n}{n}\right)$ and this completes the proof.
The claim can be proved by the mean value theorem: In Corollary 3.17 of [7] is shown that for $\beta > 1$, the extinction probability q_β is continuously differentiable. This means that the derivative of q_β is bounded on the bounded interval (β_n, β). Furthermore, for n large enough, for all $\beta_n^* \in (\beta_n, \beta)$ we have $\beta_n^* > 1$, hence

$$q_{\beta_n} = q_\beta + O(\beta_n - \beta) = q_\beta + O\left(\frac{k_n}{n}\right)$$

This completes the proof. □

As a direct consequence of Lemma 5.2.2, we can evaluate the expected value of $Z_{\geq k_n}$:

$$\mathbb{E}[Z_{\geq k_n}] = n\mathbb{P}[|C(v)| \geq k_n] = n\theta_\beta + O(k_n), \tag{5.6}$$

since all points in the undirected network have the same probability to be contained in a given connected component.

By using the Chebyshev inequality and bounding the variance of $Z_{\geq k_n}$, we will show that with high probability, the real value of $Z_{\geq k_n}$ is 'close' to its mean value. More precise:

Lemma 5.2.3. *For all $\nu \in (\frac{1}{2}, 1)$, $k_n = K\log(n)$ and sufficiently large n, there exists a $\delta > 0$ such that*

$$\mathbb{P}[|Z_{\geq k_n} - n\theta_\beta| \leq n^\nu] \geq 1 - n^{-\delta}.$$

The proof of this lemma is based on the following observations.

Using (5.6) we get for sufficiently large n

$$\mathbb{P}[|Z_{\geq k_n} - n\theta_\beta| \leq n^\nu] \geq \mathbb{P}\left[|Z_{\geq k_n} - \mathbb{E}[Z_{\geq k_n}]| \leq n^\nu/2\right].$$

By applying the Chebyshev inequality we get:

$$\mathbb{P}[|Z_{\geq k_n} - \mathbb{E}[Z_{\geq k_n}]| \leq n^\nu/2] \geq 1 - 4\frac{\mathrm{Var}[Z_{\geq k_n}]}{n^{2\nu}}.$$

Note that the Chebyshev inequality gives in general a relatively poor bound, like the Markov inequality, but in this case it will proved to be enough since $\nu > \frac{1}{2}$ and $\mathrm{Var}[Z_{\geq k_n}] = O(n)$ as will shown below.

Lemma 5.2.4. *For every n and $k < n$,*

$$Var[Z_{\geq k}] \leq (\beta k + 1)\chi_{<k} n,$$

where

$$\chi_{<k} := \mathbb{E}[|C(v)|\mathbf{1}_{\{|C(v)|<k\}}] \leq k.$$

Proof. By definition we have

$$\mathrm{Var}[Z_{\geq k}] = \mathrm{Var}[n - Z_{<k}] = \mathrm{Var}[Z_{<k}].$$

So it is suffices to compute $\mathrm{Var}[Z_{<k}] := \mathrm{Var}\sum_v \mathbf{1}_{\{|C(v)|<k\}}$.

$$\begin{aligned}
\mathrm{Var}[Z_{<k}] &\leq \mathbb{E}[Z_{<k}^2] - \mathbb{E}[Z_{<k}]^2 \\
&= \mathbb{E}\left[\sum_{i,j} \mathbf{1}_{\{|C(i)|<k\}}\mathbf{1}_{\{|C(j)|<k\}}\right] \\
&\quad -\mathbb{E}\left[\sum_i \mathbf{1}_{|C(i)|<k}\right] \cdot \mathbb{E}\left[\sum_j \mathbf{1}_{|C(j)|<k}\right] \\
&= \sum_{i,j=1}^{n} \left(\mathbb{P}[|C(i)| < k, |C(j)| < k] - \mathbb{P}[|C(i)| < k]\mathbb{P}[|C(j)| < k]\right).
\end{aligned}$$

The following natural step is to split $\mathbb{P}[|C(i)| < k, |C(j)| < k]$ depending on whether $i \longleftrightarrow j$ or not. Since $i \longleftrightarrow j$ automatically implies that $|C(i)| =$

$|C(j)|$, one part is relatively easy to compute:

$$\begin{aligned}\sum_{i,j=1}^{n} \mathbb{P}[|C(i)|<k, |C(j)|<k, i \longleftrightarrow j] &= \sum_{i,j=1}^{n} \mathbb{E}\left[\mathbf{1}_{\{(|C(i)|<k, i\longleftrightarrow j)\}}\right] \\ &= \sum_{i}^{n}\sum_{j}^{n} \mathbb{E}\left[\mathbf{1}_{\{(|C(i)|<k\}}\mathbf{1}_{\{i\longleftrightarrow j)\}}\right] \\ &= \sum_{i}^{n} \mathbb{E}\left[\sum_{j}^{n} \mathbf{1}_{\{(|C(i)|<k\}}\mathbf{1}_{\{i\longleftrightarrow j)\}}\right] \\ &= \sum_{i}^{n} \mathbb{E}\left[\mathbf{1}_{\{(|C(i)|<k\}}|C(i)|\right] = n\chi_{<k}.\end{aligned}$$

For the second part, we write for all $l < k_n$:

$$\begin{aligned}&\mathbb{P}[|C(i)|=l, |C(j)|<k, i \nleftrightarrow j] = \\ =\ &\mathbb{P}[|C(j)|<k||C(i)|=l, i \nleftrightarrow j]\cdot\mathbb{P}[|C(i)|=l, i \nleftrightarrow j] \leq \\ \leq\ &\mathbb{P}[|C(j)|<k||C(i)|=l, i \nleftrightarrow j]\cdot\mathbb{P}[|C(i)|=l].\end{aligned}$$

Together we get

$$\begin{aligned}&\mathrm{Var}[Z_{<k}] = n\chi_{<k} + \\ &+\sum_{l=1}^{k-1}\sum_{i,j=1}^{n} \mathbb{P}[|C(i)|=l]\cdot\left(\mathbb{P}_{n,p}[|C(j)|<k||C(i)|=l, i \nleftrightarrow j] - \mathbb{P}[|C(j)|<k]\right).\end{aligned}$$

Observe that when $|C(i)| = l$ and $i \nleftrightarrow j$, the conditional probability distribution of $|C(j)|$ in a population of size n is equal to the unconditional probability distribution of $|C(1)|$ in a population of size $n-l$, both with the same edge probability p. In formula, using the subscript notation, we get:

$$\mathbb{P}_{n,p}[|C(j)|<k||C(i)|=l, i \nleftrightarrow j] = \mathbb{P}_{n-l,p}[|C(1)|<k].$$

Before we can compare the events $\{|C(1)|<k\}_{n-l,p}$ and $\{|C(1)|<k\}_{n,p}$, we have to define them on the same probability space. Consider a realization of the epidemic process within a population of size $n-l$, starting with one initial infectious individual. To extend this to an epidemic within a population of size n, we add l extra points $\{n-l+1,\ldots,n\} := \mathcal{V}$, and for every $v \in \mathcal{V}$

we draw a connection to each of the other $n-1$ points with probability p, independently of each other.
With this coupling back in mind, we will bound the probability that $\{|C(1)| < k\}_{n-l,p}$ and $\{|C(1)| \geq k\}_{n,p}$ both happens.
This event can only happen if at least one of the vertices in $\mathcal{V}$ is connected to the component of 1 within the population of size $n-l$, denoted by $C(1)_{n-l}$. By using Boole's inequality [7] twice, we get:

$$\begin{aligned}
\mathbb{P}_{n-l,p}[|C(1)| < k] - \mathbb{P}_{n,p}[|C(1)| < k] &\leq \mathbb{P}[\cup_{a\in\mathcal{V}} \cup_{b\in C(1)_{n-l}} a \longleftrightarrow b] \\
&\leq \sum_{a\in\mathcal{V}} \sum_{b\in C(j)} \mathbb{P}[a \longleftrightarrow b] \\
&\leq lkp.
\end{aligned}$$

Now we can complete the proof by using $p := 1 - e^{\beta/n} \leq \beta/n$

$$\begin{aligned}
\mathrm{Var}[Z_{<k}] &\leq n\chi_{<k} + \sum_{l=1}^{k-1} \sum_{i,j=1}^{n} \mathbb{P}[|C(i)| = l] lkp \\
&= n\chi_{<k} + kp \sum_{j,i=1} \sum_{l=1}^{k-1} \mathbb{P}[|C(i)| = l] l \\
&= n\chi_{<k} + kp \sum_{j,i=1} \mathbb{E}\left[|C(v)| \mathbf{1}_{\{|C(v)|<k\}}\right] \\
&= n\chi_{<k} + kpn^2\chi_{<k} \leq n\chi_{<k} + k\beta n\chi_{<k} = (\beta k + 1)\chi_{<k} n.
\end{aligned}$$

□

Proof. To finish the proof of lemma 5.2.3 we note that for sufficiently large n, and any $\nu \in (\frac{1}{2})$ there exists a $\delta < 1 - 2\nu$ such that

$$\mathbb{P}[|Z_{\geq k_n} - n\theta_\beta| \leq n^\nu] \geq 1 - 4n^{1-2\nu}(\beta k_n^2 + k_n) >= 1 - n^{-\delta},$$

since $k_n = K\log(n)$. □

Up to now, we have shown that the clusters with size at least k_n together contain approximately a fraction θ_β of all vertices. But what can we say about the maximal cluster? Before moving to the main theorem we state the following lemma:

Lemma 5.2.5. *Fix $k_n = K\log(n)$, $\beta > 1$ and for all $\alpha < \theta_\beta$, then for all vertices v, there exists a $J = (\alpha, \beta)$ such that*

$$\mathbb{P}[k_n \leq |C(v)| \leq n\alpha] \leq Ce^{-k_n J},$$

where $C := (1 - e^{-J})^{-1}$.

Corollary 5.2.6. *Fix $k_n = K\log(n)$ and $\alpha < \theta_\beta$. Then for K sufficiently large, with high probability there are no clusters with size in between k_n and θ_β.*

By using the Markov inequality and the fact that for all vertices u, v, $\mathbb{P}[|C(v)| < k] = \mathbb{P}[|C(u)| < k]$, we have:

$$\begin{aligned}
\mathbb{P}[\exists v : k_n \leq |C(v)| \leq \alpha n] &= \mathbb{P}[Z_{\geq k_n} - Z_{\geq \alpha n+1} \geq 1] \\
&\leq \mathbb{E}[(Z_{\geq k_n} - Z_{\geq \alpha n+1}] \\
&= n\mathbb{P}[k_n \leq |C(v)| \leq \alpha] \\
&\leq Cne^{-k_n J} = Cn^{1-JK}.
\end{aligned}$$

The following theorem shows a sharp bound on the distribution of the deviation between a binomial variable and its mean value.

Theorem 5.2.7. *Let $X \sim \text{bin}(n, p)$ and let $\mathbb{E}[X] = \beta$. Then*

$$\mathbb{P}[X \geq \mathbb{E}[X] - t] \leq \exp(-\frac{t^2}{2\beta}).$$

A proof of this theorem can be found in [7] (Theorem 2.18).

Proof of lemma 5.2.5. Fix $\alpha < \theta_\beta$. Recall the sequential construction and note that for each $t \in (0, n]$:

$$\mathbb{P}[|C(v)| = t] = \mathbb{P}[S_t = 0 \cap S_{t-1} \neq 0] \leq \mathbb{P}[S_t = 0].$$

So,

$$\mathbb{P}[k_n \leq |C(v)| \leq \alpha] \leq \sum_{t=k_n}^{\alpha n} \mathbb{P}[S_t = 0] \leq \sum_{t=k_n}^{\alpha n} \mathbb{P}[S_t \leq 0].$$

By equation (5.3) of paragraph 5.1 we have for $p = 1 - e^{-\frac{\beta}{n}}$ and $t = \gamma n$ with $\gamma \in [k_n/n, \alpha]$:

$$\begin{aligned}
\mathbb{P}[S_t \leq 0] &= \mathbb{P}[\text{bin}(n-1, 1-(1-p)^t) \leq t-1] \\
&\leq \mathbb{P}[\text{bin}(n-1, 1-e^{-\gamma\beta}) \leq \gamma n - 1] \\
&\leq \mathbb{P}[\text{bin}(n, 1-e^{-\gamma\beta}) \leq \gamma n].
\end{aligned}$$

To bound this probability we will use Theorem 5.2.7. Write $X \sim \text{bin}(n, 1 - e^{-\gamma\beta})$. By using (2.1) combined with (5.5), we have

$$\theta_\beta = 1 - e^{-\beta\theta_\beta}.$$

For $\alpha < \theta_\beta$ we have $\alpha < 1 - e^{-\beta\alpha}$. Then for all $\gamma \in [k_n/n, \alpha]$ there exists an ϵ such that

$$\mathbb{E}[X] = n(1 - e^{-\gamma\beta}) \geq n(1 + \epsilon)\gamma. \tag{5.7}$$

Using Theorem 5.2.7 and (5.7) gives for every $t := \gamma n \leq \alpha n$

$$\mathbb{P}[S_t \leq 0] \leq \mathbb{P}[X \leq \mathbb{E}[X] - \gamma\epsilon n] \leq e^{-t^2\epsilon^2/2\beta} \leq e^{-t\epsilon^2/2\beta}.$$

Define $J(\alpha, \beta)$ as $J := \epsilon^2/2\gamma$. Now we can complete the proof

$$\mathbb{P}[k_n \leq |C(v)| \leq \alpha] \leq \sum_{t=k_n}^{\alpha n} \mathbb{P}[S_t \leq 0] \leq \sum_{t=k_n}^{\alpha n} e^{-Jt} \leq [1 - e^{-J}]^{-1} e^{-k_n J}.$$

□

We are now ready to combine the results in the main theorem of this Chapter:

Proof of Theorem 5.2.1. Fix $\nu \in (\frac{1}{2}, 1)$. Choose $\delta < 2\nu - 1$, then fix $k_n = K \log(n)$ such that $\delta < KJ - 1$. By Corollary 5.2.6 and Lemma 5.2.3, we have for all $\alpha < \theta_\beta$

$$\mathbb{P}[\mathcal{A}_n] \geq 1 - n^{-\delta},$$

where

$$\mathcal{A}_n := \{\nexists v : k_n \leq |C(v)| \leq \alpha n\} \cap \{|Z_{\geq k_n} - n\theta_\beta| \leq n^\nu\}.$$

Furthermore, $\{|Z_{\geq k_n} - n\theta_\beta| \leq n^\nu\}$ implies that when n is sufficiently large, there exists at least one cluster of size larger than k_n. This means that $|C_{\max}| \leq Z_{\geq k_n}$.

On the other hand, $\mathcal{A}_n$ also implies that are no more than two connected components of size larger than k_n. This can be argued by contradiction: Suppose there are at least two components with size at least k_n. When $\alpha > \theta_\beta/2$ and $\mathcal{A}_n$ occurs, there are no connected components of size in between k_n and αn, so $Z_{\geq k_n} \geq 2\alpha n > (\theta + \epsilon)n$. But when n is large enough

this is in contradiction with $Z_{\geq k_n} \leq \theta_\beta n + n^\nu$, since $\nu < 1$. So together we conclude that $|C_{\max}| = Z_{\geq k_n}$. This gives

$$\mathbb{P}[||C_{\max}| - n\theta_\beta| \leq n^\nu] \geq \mathbb{P}[\{||C_{\max}| - n\theta_\beta| \leq n^\nu\} \cap \mathcal{A}_n] \geq \mathbb{P}[\mathcal{A}_n] \geq 1 - n^\delta.$$

□

The result of Theorem 5.2.1 could be compared with the *weak law of large numbers* that says that the average of a sequence of n random variables converges in probability to the expected value. Here, for $\beta > 1$, the sample average is equal to the proportional final size T_n^* of a disease that spreads through a population of size n, which is the same as the probability that an average individual is part of the epidemic. Then the expected value is equal to the survival probability of a given individual within an infinite population, denoted by θ_β, where the spread of the disease behaves as a branching process. We could reformulate Theorem 5.2.1 as:

For all $\epsilon \in (0, \frac{1}{2})$, there exists a $\delta > 0$ such that

$$\mathbb{P}[|T_n^* - \theta_\beta| \leq n^{-\epsilon}] \geq 1 - n^\delta.$$

Van der Hofstad [7] also shows a *Central Limit Theorem* for the proportional final epidemic size:

$$\sqrt{n}(T_n^* - \theta_\beta) \quad \xrightarrow{d} \quad \mathcal{Z}.$$

Here the sequence converges in distribution, denoted by $\xrightarrow{d}$ to a Normal random variable $\mathcal{Z}$ with mean 0 and variance $\sigma_\beta^2 = \frac{\theta_\beta(1-\theta_\beta}{(1-\beta+\beta\theta_\beta)^2}$. Theorem 5.2.1 plays an essential role in the proof. In some sense you could say that the Central Limit Theorem implies the weak Law of Large Numbers, except that convergence in distribution is a weaker convergence than convergence in probability.

Chapter 6

Model comparison for large populations

In the previous chapter we have proven a limit on the final epidemic size for sequences of finite randomly mixing populations, with probability tending to one as $n \to \infty$. In this chapter we will use this limit to approximate the basic characteristics of the Hierarchical and the Random household-school models for large finite populations that grow in size in the appropriate way. Since we are not interested in the precise time evolution, we can consider the spread of an epidemic in a slightly different way than we described in chapter 3, while the number of eventually infected individuals remains the same. By this modification we can show that both models can be approximated by a certain branching process. We will compare the models (i.e. branching approximations) on their main characteristics numerically and moreover, we will prove a strong relation between the expected final epidemic sizes of the two models. Furthermore, we will discuss the strengths and limitations of the reproduction number.

6.1 Hierarchical model

Branching approximation

In our household-school models, Theorem 2.1.1 is no longer valid. Because of the strong connections within the relatively small subgroups, there is always a substantial probability that an already infected individual will be infected again by an infectious member of his own household or school. The event that the contact made by an infectious individual results in infection

transmission, depends on the status of the receiver. By these correlations in the small subgroups, the precise time evolution of the epidemic is more complex to analyze. However, we are only interested in the final outcome of the epidemic. So we could reconstruct the percolation cluster such that the spread of an epidemic in the beginning can be approximated by a branching process, in some sense. Andersson and Ball et al. have used a similar argument for the household model in [1] [3]. Below, we shall extend this argument for the Hierarchical model.

Consider a large population of size n represented by the Hierarchical household-school graph (like figure 3.1) with corresponding edge probabilities as defined in section 3.2. We will construct a percolation cluster of infected individuals in a slightly different way as described in section 3.2, and we shall call this the *modified* cluster.
First we consider the epidemic spreading only within the school (which automatically includes the households) of the initial infective, this is what we call a *local* epidemic. By our assumption of a fixed infectious period, every individual makes global contacts following the same distribution, independently of each other. Furthermore, until $n^{1/2-\epsilon}$ individuals are infected and removed (we will cal this the beginning of the epidemic), these global contacts are with high probability made with individuals on previously uninfected schools. The proof of this statement is similar to the proof of Theorem 3.2, because the school sizes are held fixed as n goes to infinity. So in the beginning of the epidemic, the offspring of global infections made by the eventually infected individuals of this local epidemic are all dispersed across different unexplored schools. We move on to the newly infected schools and consider them in the same manner. We conclude that until $n^{1/2-\epsilon}$ individuals (or schools) are infected and removed we could, with high probability, replace each school by one vertex such that the process could be approximated by a branching process, where the offspring of each school corresponds exactly to the set of globally infected children produced by all eventually infected individuals within that particular school. We will call this a *school-to-school* branching approximation.

Reproduction number

In epidemiology, the reproduction number is a relevant quantity for practical purposes and easy to evaluate. Recalling the definition, R^* is the mean number of infections caused by one infectious individual. It is an important

threshold function that indicates whether a large outbreak will occur or not. Because of its critical behavior around 1, the reproduction number can be used to prevent a major epidemic: for a reproduction number of $R^* > 1$, a proportion $1 - \frac{1}{R^*}$ of the infective contacts must be blocked to halt the growth of an epidemic. Here, we mean only the contacts which will certainly result in new infectious. In a well mixed population, it would be enough to make a fraction $1 - \frac{1}{R^*}$ of the population immune against the infection, if a vaccine is available. This intervention reduces the number of infected individuals in the next generation by a factor $\frac{1}{R^*}$. This will result in a new reproduction number $R^* = 1$, and thus the epidemic will eventually die out. In a Hierarchically structured population we will use the 'school reproductive number' R_S, defined as the expected number of schools infected by an infectious school. Obviously, in this hierarchical model, R_S caries other information than R^*. However, both reproduction numbers are epidemic thresholds and they exceed 1 for the same model parameters. This is because, if an epidemic dies out on school level, this also happens on individual level, and vice versa. So comparable to the individual situation, temporarily closing a fraction $1 - \frac{1}{R_S}$ of all schools will halt the epidemic spread.

To calculate the actual offspring distribution of the school to school branching process, we have to incorporate the final size of the local epidemic. Consider a sequence of populations $\mathcal{H}(n)$, $n \to \infty$, where all schools are of size n_S (not growing with n) and all individuals are equally likely to meet each other outside school. Number the individuals of a school S by $s_0, s_1, s_2, \ldots, s_{n_S-1}$ where s_0 is the initial infective within the school. For each n, let C_i be the number of global neighbors infected by individual s_i, in case if s_i is infected by the initial infectious individual s_0. As we have mentioned earlier, for large n, these globally infectious contacts are with high probability all in distinct, previously uninfected households. So $C_0, \ldots, C_{n_S-1}$ are mutually independent and identically Poisson distributed, "i.i.d.", with mean β_G. Therefore, we get

$$R_S = \mathbb{E}[C_0 + \sum_{i=1}^{T} C_i] = \mathbb{E}[(T+1)C_i] = (\mathbb{E}[T] + 1)\beta_G.$$

where T is the final size of the within household-school epidemic not including the initial infective, which can be computed by equation (4.6) of chapter 4. In chapter 5, we have seen that for large n, in the beginning of the epidemic it becomes clear whether a large outbreak will occur or not, with high probability. Therefore, this reproduction number is a good approximation for the threshold measure of an epidemic within a finite population.

Extinction probability

Again, we consider a sequence of populations $\mathcal{H}(n)$ with the appropriate structure, growing in n. The extinction probability on a school level corresponds to that on an individual level, as we have seen earlier. So by Theorem 5.2.1, we can approximate the extinction probability of the epidemic process by the extinction probability of the corresponding school-to-school branching process. We use the same notation as above where C_i is the number of globally infected individuals caused by an infectious individual i, and let T be the final size of the local school-epidemic. Conditioned on the final size, $C_1, C_2, \ldots, C_T$ are mutually independent and Poisson (β_G) distributed. The offspring distribution between schools can be described by the following generating function by using equation (2.1) from chapter 2:

$$f_{\mathcal{S}}(s) = \mathbb{E}\left[\mathbb{E}\left[s^{\sum_{i=1}^{T} C_i}|T\right]\right] = f_T(f_{C_1}(s)) = \sum_{i=0}^{n_S-1} \mathbb{P}[T=i]\left(\sum_{k=0}^{n-1}\mathbb{P}[C_1=k]s^k\right)^{i+1}.$$

The extinction probability q_S is equal to the lowest root of $f_{\mathcal{S}}(s) - s$, and can be evaluated numerically.

6.2 Random model

Branching approximation

We will show that until $n^{1/2-\epsilon}$ individuals of the population are infected and removed, the epidemic spread within the Random model can be approximated by a multi-type branching process.
Similar to above, we first consider the epidemic spreading only within the school and household of the initial infective x_0. However, other than in the Hierarchical model, we claim that in the beginning of the epidemic, siblings are with high probability, member of different previously uninfected schools. This claim can be proved in the same way as Theorem 2.1.1, since the household and school size are held fixed as n goes to infinity. So we determine final school and household epidemic separately, and assign all eventually infected individuals of this household and school epidemic to x_0. Notice that by the assumption of a fixed infectious period, this rearrangement of the original percolation cluster will not influence the total final set of the epidemic. In the next step, we explore the global contacts made by x_0, and we recall that all of these individuals, in the beginning of the epidemic, are member of different, previously uninfected schools. Then we move on to one

of the individuals in the next generation and we consider them in the same manner, and so on. By this modification, self loops and parallel edges will be avoided, while the set of finally infected individuals remains the same.
Now, we could label the infected individuals such that each individual of a certain type has the same fixed multi-type offspring distribution, with high probability. We stress that during this labeling, we only consider the infected individuals, susceptible individuals are not taken into account. We say that an individual is of type H (resp. type S or type G) if it has become infected via an edge of color red (resp. blue or green). Notice that individuals of type H or S will not infect any individual of its own type, since all eventually infected individuals of the local epidemic are already assigned to the initial infectious. The precise offspring distributions for each type will be further explained below.

Reproduction number

We will define a reproduction matrix $\mathcal{R}_{ij}$, as mentioned in section 2.3 and like Andersson has done in [1]. Recall, $\mathcal{R}_{ij}$ is the mean number of infected individuals of type j infected by one individual of type i. Recall that in the early stages of the epidemic, infected individuals are seldom part of the same household and school, especially for large n.
By the randomly mixing property of the community level we know that each individual, regardless its type, infects on average β_G individuals of type G. Furthermore, for large n, an individual of type G is with high probability the initial infectious individual of his own household and school. He will on average infect a number of $\mathbb{E}[T_H]$ and $\mathbb{E}[T_S]$ members of his own household and school, where T_H and T_S are the household and school final epidemic size, not including the initial infective. The final local sizes can be calculated by equation 4.1 of chapter 4. Finally, recall that an individual of type H or S will not produce any individuals of his own type. We could summarize these observations in the following matrix:

$$\mathcal{R} := \begin{pmatrix} \beta_G & \mathbb{E}[T_S] & \mathbb{E}[T_H] \\ \beta_G & 0 & \mathbb{E}[T_H] \\ \beta_G & \mathbb{E}[T_S] & 0 \end{pmatrix}.$$

This matrix $\mathcal{R}$ is positive regular, so we will use its largest eigenvalue R_ρ as an epidemic threshold value for the Random model, see section 2.3. R_ρ, the largest root of $g(x) = \det(\mathcal{R} - xI)$, has no easy explicit expression, so we will calculate it numerically for different model parameters $(\beta_H, \beta_S, \beta_G)$.

Extinction probability

Similar to the Hierarchical model, we know that the extinction probability vector $\mathbf{q} = (q_H, q_S, q_G)$ of the corresponding multi-type branching process, is a good indication for the extinction probability within a large *finite* population. The proof of this statement is slightly different as the proof in chapter 5, but the details are not included in this Thesis. Note that the epidemic dies out with probability 1 if and only if $\mathbf{q} = 1$, i.e. for all i, $q_i = 1$. Also, if there exists an i for which $q_i < 1$ then for all $j \neq i$, $q_j < 1$.

Recall that by assumption, an individual of type S or H will never infect an individual of its own type. Furthermore, for large n, an individual of a certain type will infect individuals of different types independently of each other. These observations lead to the following set of equations, using the generating functions stated in paragraph 2.3:

$$\mathbf{q} = (q_H, q_S, q_H) = \mathbf{f}(\mathbf{q}) = (f_H(\mathbf{q}), f_S(\mathbf{q}), f_G(\mathbf{q})).$$

If $\mathbf{Z}_0$ is of type H, then $\mathbf{Z}_1^H$ will have the following generating function:

$$\begin{aligned} f^H(\mathbf{q}) &= \sum_{a_H,a_S,a_G=0}^{\infty} \mathbb{P}_H[\mathbf{Z}_1 = (a_H, a_S, a_G)](q_H)^{a_H}(q_S)^{a_S}(q_G)^{a_G} \\ &= \sum_{a_G=0}^{\infty} \sum_{a_S=0}^{n_S} \mathbb{P}_H[Z_1^S = a_S]\mathbb{P}_H[Z_1^G = a_G](q_S)^{a_S}(q_G)^{a_G}, \end{aligned}$$

where $\mathbb{P}_H[\mathbf{Z}_1 = (a_H, a_S, a_G)]$ is the probability that an individual of type H will infect a_H individuals in its own household, a_S in its own schools and a_G individuals on global level. The random variable Z_1^G is Poisson(β_G)-distributed and the distribution of Z_1^S can be evaluated by equation 4.1 of chapter 4. In the same way we can compute the generating functions f^S and f^G.

It is not possible to obtain an explicit expression for q_H, q_S and q_R, so we will determine them numerically. In the next paragraph we will compare both models in different settings, using the same parameter values $(\beta_H, \beta_S, \beta_G)$.

So in the Random model, the characteristics are determined on individual level, while in the Hierarchical model we have computed them on school level. For the reproduction numbers, this means that these cannot directly

be compared, because they contain different information. However, regardless of the level on which the reproduction number is determined, the number does indicate whether there is a probability larger than 0 that a large epidemic outbreak will occur. So, the proportional final size on individual and school level is the same. In section 5.1 we have shown a direct relation between the extinction probability and the final size, given that a large outbreak occurs. Therefore, these characteristics can well be compared between the Hierarchical and Random model. For practical purposes, the reproduction number is used more often, because it is easier to compute. But for our purposes, the reproduction number is only useful as a threshold value.

6.3 Coupling argument and numerical results

We want to investigate how the configuration of a social network influences the spread of an epidemic. We have tried to make a direct comparison between the Hierarchical and the Random model by holding their subgroup sizes identical. However, in the beginning of the epidemic, this directly implies an unequal number of *neighbors* for each individual in both models, where neighbors are the direct connections in the corresponding graphs. In the Random model every individual has with high probability $n_H - 1 + n_S - 1$ local connections since all siblings of each individual are in the beginning members of different schools, while in the Hierarchical model every individual is always connected to $n_S - 1$ local neighbors since here siblings are automatically schoolmates as well. Mainly because of this difference in number of neighbors, we could prove a strong relation between the final results of an epidemic in the Hierarchical and the Random model.

Theorem 6.3.1. *For all $\delta > 0$ there exists a n such that for all $\epsilon > 0$ and $k < n^{1/2-\epsilon}$*

$$\mathbb{P}[T^H(n) \leq k] + \delta \geq \mathbb{P}[T^R(n) \leq k],$$

where $T^H(n)$ and $T^R(n)$ are the final epidemic sizes within a population of size n for the Hierarchical and Random model respectively, with schools of size n_S, households of size n_H, and transmission per pair rates denoted by $(\frac{\beta_G}{n-1}, \beta_S, \beta_H)$.

Proof. We want to describe a procedure that builds a cluster of infected individuals in the Hierarchical model and in the Random model simultaneously, starting with the initial infectious x_0, in such a way that with high probability in the beginning of the epidemic, the number of infected and removed individuals in the Hierarchical model is smaller or equal than the

number of infected and removed individuals in the Random model.

Let $\mathcal{A}_n$ be the event that, up to $n^{1/2-\epsilon}$ are infected and removed within a population if size n, the siblings of each infectious individual in the Random model, are all members of different, previously uninfected, schools.

Let $\mathcal{B}_n$ be the event that, up to $n^{1/2-\epsilon}$ are infected and removed within a population if size n, the global contacts in the Hierarchical and the Random model, are made with members of different, previously uninfected, schools.

Fix $\delta > 0$ and $\epsilon > 0$. Similar to Theorem 2.1.1, we know that there exist a n_1 and a n_2 such that $\mathbb{P}[\mathcal{A}_{n_1}] \geq 1 - \delta$ and $\mathbb{P}[\mathcal{B}_{n_2}] \geq 1 - \delta$. This is because the school and household size held fixed as the population size grows and because the siblings are paired uniformly at random.
If we choose $n = \max(n_1, n_2)$ and define $\mathcal{D}_n := \mathcal{A}_n \cap \mathcal{B}_n$, then $\mathbb{P}[\mathcal{D}_n] \geq 1 - \delta$.

Since

$$\begin{aligned} \mathbb{P}[T^R(n) \leq k] &= \mathbb{P}[\{T^R(n) \leq k\} \cap \mathcal{D}_n] + \mathbb{P}[\{T^R(n) \leq k\} \cap \mathcal{D}_n^c] \\ &\leq \mathbb{P}[\{T^R(n) \leq k\} \cap \mathcal{D}_n] + \delta, \end{aligned}$$

it is suffices to prove that for all $k < n^{1/2-\epsilon}$

$$\mathbb{P}[T^R(n) \leq k \cap \mathcal{D}_n] \leq \mathbb{P}[T^H(n) \leq k \cap \mathcal{D}_n].$$

But this is only true if and only if

$$\mathbb{P}[T^R(n) \leq k | \mathcal{D}_n] \leq \mathbb{P}[T^H(n) \leq k | \mathcal{D}_n].$$

We are going to prove this latter statement. We start by considering the local epidemic spread within the school of x_0. This means that we explore the percolation cluster on the blue subgraph starting with x_0 and using the infection probability $\bar{p}_S := 1 - e^{-\beta_S}$, as described in section 3.2. Each time a blue edge in the Random model is marked as open or closed we will give the same mark to the corresponding edge in the Hierarchical model. Note that by construction, these blue clusters in both models are of the same size and we give the same mark (number) to corresponding vertices.
Secondly, we explore the household infections made by the already infected individuals. Obviously we only have to explore those vertices in the blue cluster of the Hierarchical model, who have at least one susceptible sibling.

This set of vertices will be denoted by $\mathcal{X}_0$. Choose $x \in \mathcal{X}_0$. We let the susceptible siblings of x in the Hierarchical model correspond to arbitrarily chosen siblings of x in the Random model by marking them with the same number. These siblings in the Random model are always susceptible. This is because when we condition on $\mathcal{A}_n$, we have for each $k < n^{1/2-\epsilon}$ that, up to k individuals are infected and removed (we call this the beginning of the epidemic), siblings in the Random model are member of previously uninfected schools. On the other hand, siblings in the Hierarchical model are always member of already infected schools. Note that the other siblings of x in the Random model will not be examined. Each time an explored red edge is marked as open or closed in the Random model we will give the same mark to the corresponding red edge in the Hierarchical model. After that, we remove the labels of those individuals who remain susceptible during these household infections. Now, two vertices in the different models with corresponding labels will also have the same number of departing edges.
Then, we explore the school infections made by one of those new infected siblings of x. Start with sibling y and let the susceptible schoolmates of y in the Hierarchical model correspond to arbitrary schoolmates of y in the Random model, which are automatically susceptible. Couple the school infections in the same way as described above, and finally remove the labels of those individuals who remain susceptible during this local school infection. Then we move on to one of the other unexplored siblings of x and we repeat the marking procedure, and so on.

In the next iteration we define a new set of infected vertices who are still not explored and whose siblings are still susceptible. We denote this set by $\mathcal{X}_1$. We choose an individual $x_1 \in \mathcal{X}_1$ and we let the cluster initiated by x_1 in the Hierarchical model, correspond to a cluster in the Random model in the same manner as described above. We could follow this procedure t iterations, where t is the first moment that $\mathcal{X}_t$ is empty.

Up to now, we have the local epidemic in the Hierarchical model, caused by school and household infections, initiated by x_0 (denoted by $\mathcal{C}_0$), coupled to an epidemic cluster in the Random model. In figure 6.1 you can see that infectious individuals in the Random model are dispersed across different schools while in the Hierarchical model, this cluster is restricted to the school of x_0. The actual local cluster in the Random model could be larger, since not all edges were explored in this model.

Now we examen the global connections departing from the individuals con-

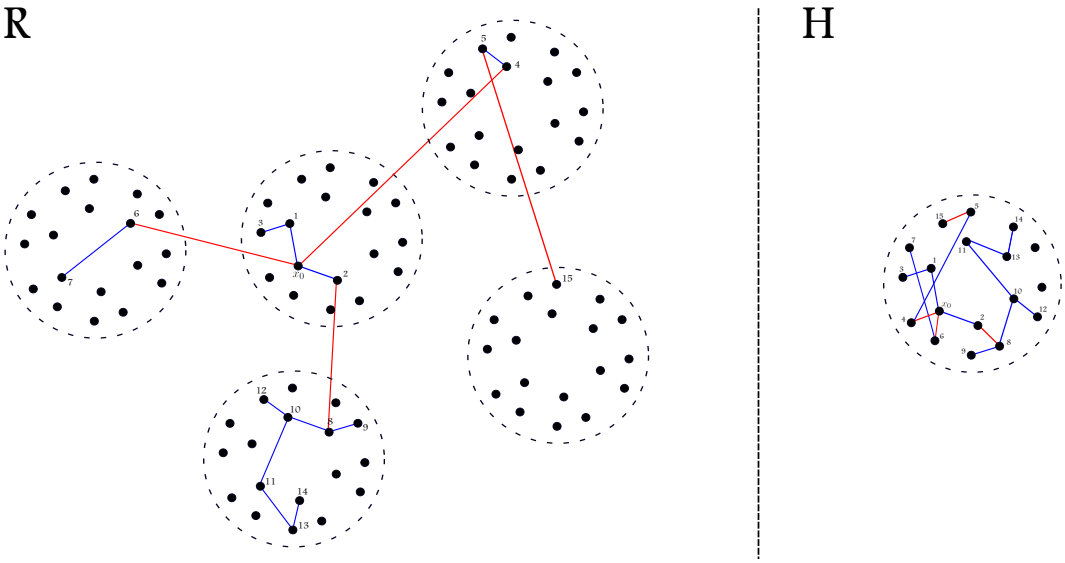


Figure 6.1: In this picture, the cluster of school and household infections in the Hierarchical model (H) $\mathcal{C}_0$, initiated by x_0, is coupled to a cluster of the same size in the Random model (R). This is an example where the schools are of size 18 and the households are of size 3. Each infected individual is assigned by a number which indicate its corresponding vertex and also, it is the order of exploration.

tained in $\mathcal{C}_0$, all open with probability $\bar{p}_G$ independently of each other. Each time a green edge between a vertex in $\mathcal{C}_0$ and one of its neighbors is marked as open or closed in the Random model, we give the same mark to the corresponding green edge in the Hierarchical model. Since we condition on $\mathcal{B}_n$, all endpoints of these green edges are in the beginning of the epidemic, members of previously uninfected school. Start with considering one endpoint, say x_i. We could explore the local epidemic $\mathcal{C}_i$ in the Hierarchical model, initiated by x_i, and we couple this to a cluster in the Random model in the same way as described above. We repeat this procedure for all endpoints x_j. The newly infected individuals in the obtained clusters $\mathcal{C}_j$ could make global connections on their turn, and so on. So we conclude that for all $k < n^{1/2-\epsilon}$, $\mathbb{P}[T^H(n) \leq k|\mathcal{D}_n] \geq \mathbb{P}[T^R(n) \leq k|\mathcal{D}_n]$. □

Corollary 6.3.2. *Let q_H and q_R be the extinction probabilities for corresponding branching processes and let $\mathbb{E}[T^H(n)]$ and $\mathbb{E}[T^R(n)]$ be the expected final size within a population of size n, for the Hierarchical and the Random model respectively. For all transmission rates $(\frac{\beta_G}{n-1}, \beta_S, \beta_H)$, we have*

$$q_H \geq q_R$$

and

$$\lim_{n\to\infty} \mathbb{E}[T^H(n)] \leq \lim_{n\to\infty} \mathbb{E}[T^R(n)].$$

Proof. By Lemma 5.2.2, we know that in a randomly mixing population with infection rate β, there exists a K such that the probability that the epidemic cluster will not be larger than $K\log(n)$ tends to the extinction probability of the corresponding branching process, as n goes to infinity. We could apply this lemma directly to the Hierarchical model, since there we have uniform mixing on school level. For Random model, Lemma 5.2.2 have to be generalized to the multi-type uniform mixing case, but the proof is not included in this thesis.
For the randomly mixing case, we fix a N and $\epsilon > 0$, such that for all $n > N$, $K\log(n) < n^{1/2-\epsilon}$. So by Lemma 6.3.1, when we choose $k \in (K\log(n), n^{1/2-\epsilon})$ and we let n grow to infinity, we get $q_H \geq q_R$. Furthermore, we have seen in Theorem 5.2.1 that, given that a large outbreak occurs, the proportional final size converges to the corresponding survival probability. So, when we generalize the results of chapter 5 to the multi-type case, the inequality for the extinction probabilities implies that $\lim_{n\to\infty} \mathbb{E}[T^H(n)] \leq \lim_{n\to\infty} \mathbb{E}[T^R(n)]$. □

The strong relation between both models is also explicitly shown in figure 6.2 and 6.3 for a finite population of size 100, where we have calculated

the reproduction number and the extinction probability for some model parameters. In these two examples we have held the relation between the transmission parameters $\beta_H : \beta_S : \beta_G$ fixed. In figure 6.2 one can see that for all $(\beta_H, \beta_S, \beta_G)$ the red line is below the blue line and this is consistent with Corollary 6.3.2. Remember that R_ρ contains different information than R_S, so it is of little use to compare their exact values, hence the intersection in graphic 6.3 has less meaning. However, R_ρ and R_S both serve as an epidemic threshold, and they exceed the critical value 1 for the same model parameters as the extinction probabilities become below 1.

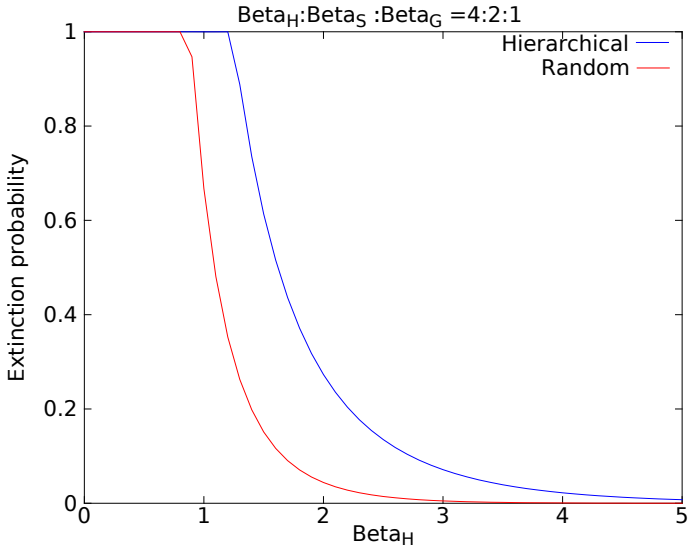


Figure 6.2: The extinction probabilities

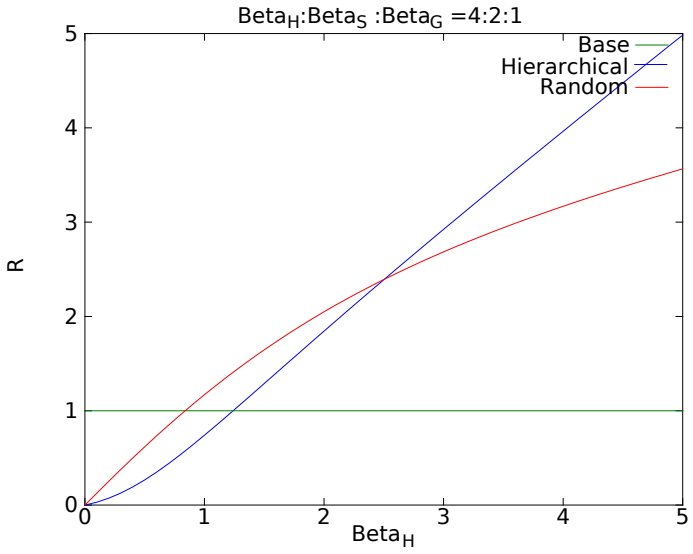


Figure 6.3: The reproduction numbers

6.4 Comparison with equal number of neighbors

The strong relation we have proven in Theorem 6.3.1, gives rise to a comparison of a Hierarchical model and a Random model where the number of neighbors for each individual are the same in the start of the epidemic. We could for instance consider a Hierarchical model with schools of size n_S and households of size 2 and a Random model where the schools are of size $n_S - 1$ and household of size 2. It is difficult to compare these models exactly, but we will explain their difference intuitively.
Firstly, we remark that although each individual in both models is connected to $(n_S - 1) + 1$ individuals, in the Hierarchical model one of these neighbors has a higher probability to become infected. This is the sibling, which can be infected here by school and household contact as well. It seems that this observation implies that the Hierarchical model will spread the epidemic more easily. However, we still have not taken into account that there is also a probability that newly infected individuals were already infected before.
Especially, in Hierarchical structure this will be more often the case. Remember that in the Random model, sibling are with high probability member of previously uninfected schools, while in the Hierarchical model household infection remains within the schools. In the Hierarchical model, infection transmission from school to school is only possible via global contact. By this reasoning, we could expect that for large β_H the Random model will spread the disease more quickly while for small β_H the Hierarchical model is more friendly for the epidemic spread. Note that a larger household size will only strengthen this effect.
These observations are also confirmed in the graphics below, see figure 6.4. We have plotted the reproduction numbers for the Hierarchical and Random model for three different values of β_H. In particular, look at the different values of β_S for which the red 'Rand3' lines intersect the base line. These intersections indicate the phase transitions in the models.

6.5 Further research

In Theorem 6.3.1 we have proved that in the start of an epidemic, when the number of new infected individuals can be approximated by a branching process, the Hierarchical model is in the limit stochastically dominated by the Random model. However, it is not clear whether this strong comparison still holds after the branching approximation breaks down. In this case, the coupling we described above will fail: it is possible that individuals

infected via household contact or global contact were already infected before in the Random model while these 'same' individuals were still susceptible in Hierarchical model. In order to prove a comparison (if it exists) between the eventually epidemic sizes in both models, a creative way of coupling between the epidemic processes have to be constructed.

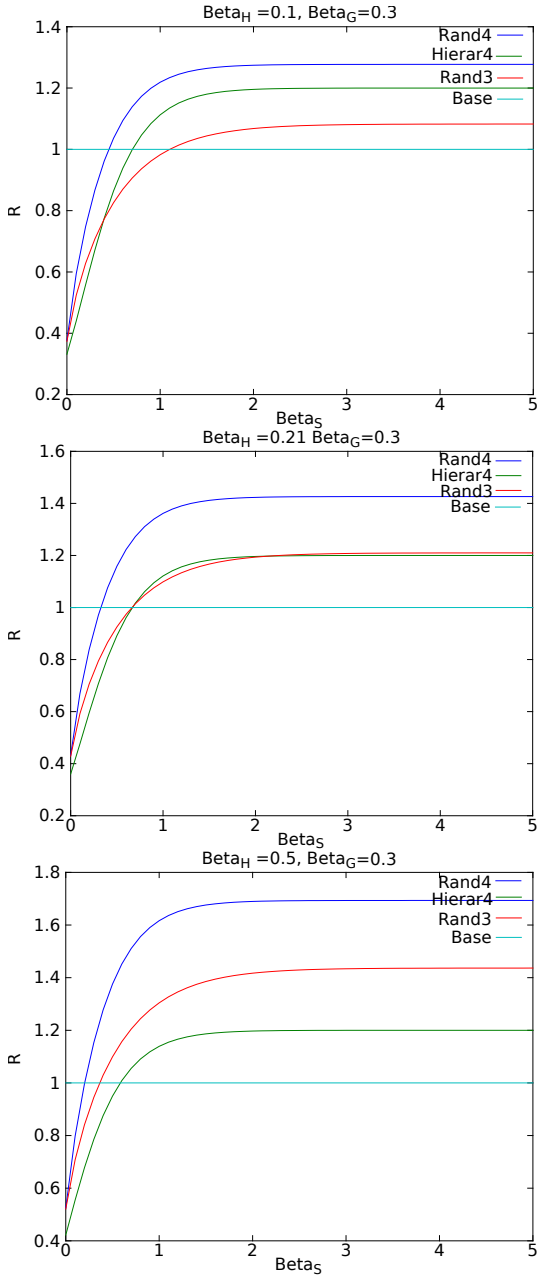


Figure 6.4: You see the reproduction numbers varying with β_S for a Random (resp. Random, Household) model where schools are of size 3 (resp. 4 , 4) and households are of size 2, called 'Rand3' (resp. 'Rand4', 'Hierar4 '). In the second plot we have tuned $\beta_H = 0.22$ such that 'Rand3' and 'Hierar4' has a phase transition for the same β_S. Here, $\beta_G = 0.3$, which is rather arbitrary.

Chapter 7

Discussion

In this chapter we will present some remarks on the assumptions we have made in the previous chapters.

For ease of presentation we have assumed that the schools (resp. households) in the Random and Hierarchical model are all of the same size. To obtain a model closer to reality we can easily generalize our results by incorporating unequal subgroup sizes [3] [4]. We can then do similar computations as in Chapter 6, but the set of different types will become much larger. Individuals in the Random model will then not only be labeled according to whether they are infected by household, school or global contact, but the labels will also have to indicate the size of the subgroup they belong to. In the Hierarchical model, the spread from school to school should then be approximated by a multi-type branching process, where the schools are labeled according to the number of households of different sizes which it contains.

Another assumption we have made, was that all individuals are infectious for a fixed period of time. Although in reality, the infectious period is different for each individual, it is proven [8] [10] that by assuming a fixed infectious period, the actual final epidemic size will be overestimated. So the vaccinating strategies we have proposed in chapter 6 will at least be sufficient to halt the spread of an epidemic. This is because when we assume a fixed infectious period, our model corresponds to the bond percolation model where the infections made by the same individual are independent. Meester and Trapman have presented a proof (based on Kuulasmaa [8]) that the bond percolation model corresponds to a worst case scenario, in the sense that the probability of an individual being part of an infinitely large cluster is

maximal, with respect to epidemics with other (random) infectious periods. See Theorem 1.1 of [10] where the susceptibility of all individuals $\bar{W} = 1$.

Finally we remark that in [1], [12], [4], and other research, the underlying social network for the spread of an epidemic via global contacts is described by random graphs (e.g. Erdös Rényi graphs) instead of by a complete graph as we have used. Because in reality it is unlikely that an individual will meet all other individuals during his or her lifetime, the random graph is used to approximate the actual social network stochastically. In the Erdös Rényi graph, the i-th and the j-th individual are neighbors of each other with a certain probability p_c (i.e. they are connected by an edge in the network), and this probability is the same for each pairs of individuals. Secondly, to model the spread of global infections across the resulting network, there is an additional probability p_β (independent of p_c) that the infection will actually transmit from i to j when i becomes infective earlier than j, or visa versa. However, in the Hierarchical and Random model, we have combined these two processes into one process. Here, the contact probability p_G could be viewed as a product of these independent 'neighbor' and 'infection' ('close contact' in literature) probabilities p_c and p_β.

Bibliography

[1] Andersson, H. Epidemic models and social networks. *Mathematical Scientist*, 24:128–147, April 1999.

[2] Andersson, H. and Britton, T. *Stochastic epidemic models and their statistical Analysis*. Spring Lecture Notes in Statistics, 2000.

[3] Ball, F.G., Mollison, D., and Scalia-Tomba, G. Epidemics with two levels of mixing. *Annals of Applied probability*, pages 46–89, 1997.

[4] Ball, F.G, Trapman, J.P., and Sirl, D. Threshold Behaviour and final outcome of an epidemic on a random network with household structure. *Advances in Applied probability*, 41:765 –796, 2009.

[5] Diekmann, O. and Heesterbeek, J.A.P. *Mathematical epidemiology of Infectious Diseases*. Chichester: John Wiley & Son, 2000.

[6] Harris, T.E. *The Theory of Branching Processes*. Doven Publications, New York, 2nd Edition, 1989.

[7] Hofstad, R. van der. Random graphs and complex networks. http://www.win.tue.nl/ rhofstad/NotesRGCN2009.pdf, 2009.

[8] Kuulasmaa, K. The spatial general epidemic and locally dependent random graphs. *Advances in Applied probability*, 1982.

[9] Meester, R.W.J. and Franscetti, M. *Random networks for communication, from statistical physics to information systems*. Cambridge University Press, 2007.

[10] Meester, R.W.J. and Trapman, J.P. Bounding basic characteristics of spatial epidemics with a new percolation model. *(Submitted)*, 2008.

[11] Mollison, D. Epidemic models: their structure and relation to data. *Journal of applied mathematics*, 1977.

[12] Trapman, J.P. *On stochastic models for the spread of infections*. PhD thesis, September 2006.

[13] Watts D.J., Muhamad, R., Medina, D.C., and Dodds, P.S. Multiscale, resurgent epidemic in a hierarchical metapopulation model. *PNAS, applied mathematics*, 7, 2005.