Stockholms universitet

End of an SIR epidemic on a configuration model

Ana Serafimović

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2014:1
Matematisk statistik
Februari 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2014:1**
http://www.math.su.se

End of an SIR epidemic on a configuration model

Ana Serafimović*

February 2014

Abstract

We consider an SIR (Susceptible-Infective-Recovered) epidemic on a configuration model network of contacts. To approximate the spread of the disease, we use branching processes. The main objective is to make inference about the final stages of the epidemic. We derive the distribution of the ultimately susceptible individuals (those who escape the epidemic) and calculate the fraction of their neighbors which are also ultimately susceptible. The most important result is that, under certain regularity conditions, the distribution of the time until the epidemic is completely eradicated has an exponentially declining tail. This will lead us to conclude that, even though a disease might take off quickly, it will nevertheless persist for a much longer time.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: anaserafimovic@yahoo.com. Supervisor: Pieter Trapman.

Acknowledgements

I would like to most genuinely thank my supervisor Dr. Pieter Trapman for formulating the problem and letting me work for and learn from him. Writing this thesis has been an invaluable experience. Also, I would like to thank Prof. Tom Britton for reading the text and for the final remarks.

I also want to thank Stockholm University and Serbian Ministry of Youth and Sports, for the financial support.

Many thanks to my family, for more than the past couple of years and for more than words can say.

Contents

1 Introduction

The research interest in graph theory has over time shifted from questions regarding local properties of small graphs/networks (starting with The Seven Bridges of Königsberg problem, solved by Euler in 1735) to the properties of large graphs. The motivation is that large data sets have recently emerged, like the World Wide Web and genomic aberrations. As well as that, graphs are now normally used for modeling relations which are inherently stochastic. In particular, we will consider applications to epidemiology, where it is, for instance, sensible to assume that a contact causes infection with some probability. On the other hand, we now have technological advances which were not available in the 18.th century and which enable us to analyze such large data sets. Also, the vast theory of probability mainly consists of asymptotic results, which has ultimately given rise to this field of study. Accordingly, theory of random graphs is a fairly new branch of mathematics, introduced in [10], in 1959.

In this work, we are interested in the time it takes until an epidemic is completely eradicated. We estimate for how long control measures should be prolonged. For example, if an animal disease is spreading within a country and there are transport bans on meat, due to the economic loss and the safety of those who import, it is essential to know when the country can safely export again. This is of particular interest when it comes to lethal or otherwise severe diseases that spread quickly. We will prove that, even though a disease might take off quickly, that is, cover most of the individuals infected during the epidemic within a short period of time, the disease will nevertheless persist for a much longer time.

We will study random graphs with certain social structure, in particular random graphs with specified vertex degrees, which means that each individual has his or her fixed number of acquaintances with whom they can interact. The neighboring vertices are chosen at random (this will be discussed in more detail in Section 3). We then impose a disease spreading along the network of contacts, which we approximate using a branching process. We assume that the disease propagates as an SIR (Susceptible - Infective - Removed) epidemic. That is to say, an individual can only transit from S to I, or from I to R state. There is a small number of the initially infective (in most of the literature only one) and the rest of the population, say n individuals, are susceptible to start with. At any time an infective can contact his or her neighbors according to a Poisson process with per pair intensity β. These processes are independent of each other. If the contacted vertex is still susceptible, it acquires the infection and instantly becomes infectious (can further spread the disease). He or she can infect others for a fixed or random time which we denote as I and call infectious period. We assume that the infectious periods are random and identically distributed. These periods are also assumed to be independent for different individuals and each I is independent of the Poisson contact processes. When the infectious period is over, the individual either dies or becomes permanently immune to the given disease, and can thus be shifted to the

cohort Removed.

An additional simplifying assumption we make is that the population is closed - there are no births, deaths or migration. Therefore, the removed stay in the system, but become "ghosts", as transferring infection to them is ignored. If we assume that the population is large, then in the early stages of an epidemic outbreak the probability of infecting a "ghost" will be very small. That justifies the branching process approximation, until the number of ever infected individuals reaches a value that makes it highly probable for such loops in transmission paths to occur. In Section 5.1 we prove that, for the number of the infected of smaller order than $\sqrt{n}$, the probability of infecting a ghost is arbitrarily small.

However, our main goal is to make inference about the final stages of an epidemic, if one occurs. We say that an epidemic, or a large outbreak, occurs if the number of ultimately infected individuals is of the same order as the number of individuals in the population, n. In this setup branching processes should be used with caution. We define epidemic generated graph and susceptibility sets and study the epidemic on them. We then derive the distribution of the ultimately susceptible individuals (those who escape the epidemic) and calculate the fraction of their neighbors which are also ultimately susceptible. The main question is to provide a distribution for how much time it takes for an epidemic to cease. In order to do so, we make use of the standard branching processes theory, in continuous time [14].

For an SIR epidemic in a randomly mixing population (in which every individual contacts every other independently and with the same intensity), the following differential equations are used to describe the deterministic approximation of the time dynamics of the epidemic. If $s(t)$ is the fraction of the susceptible at time t, $i(t)$ the fraction of the infected and $r(t)$ the fraction of the recovered in a population, the equations are

$$\begin{aligned}
\frac{ds(t)}{dt} &= -\beta s(t)i(t), \\
\frac{di(t)}{dt} &= \beta s(t)i(t) - \gamma i(t), \\
\frac{dr(t)}{dt} &= \gamma i(t).
\end{aligned}$$

Parameter β is the rate at which an infectious individual contacts the susceptible ones and γ is the recovery rate of the infected individuals. Note that this means that per-pair infectious contact rate equals $\frac{\beta}{n}$.

According to these equations, the curve corresponding to the number of infected individuals through time is shaped as in Figure 1 - it increases until it peaks and then the number of infections declines until the end. It was proved in [7] that the above introduced social structure (the configuration model) yields a very similar shape of the epidemic curve.

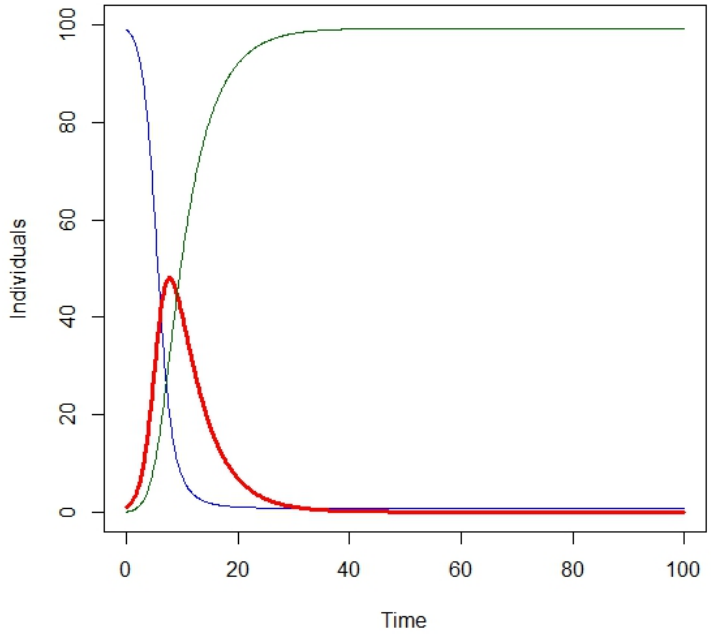

Figure 1: An example of time dynamics of an SIR epidemic in a closed population obtained from the above differential equations. Red and emphasized - infected individuals, blue - susceptible individuals, green - recovered individuals. The infection rate is $\beta = 1$, the recovery rate is $\gamma = 0.2$.

The remainder of the thesis is organized as follows.
Section 2 introduces the notation.
Section 3 is about branching processes - discrete time (Galton Watson) and continuous time (the general branching process).
In Section 4 we introduce the structure of the social network, on which a disease will be imposed. In particular, the random graph of contacts is a configuration model. Here we motivate and outline the assumptions that hold throughout the thesis. In Subsection 4.3 we outline the small world results for random graphs, which have motivated this work.
In Section 5 we prove that the branching approximation is valid until the number of infections is of smaller order than n. We define the basic reproduction number and calculate it for a configuration model. In the last two subsections, some results from [22] are generalized so that they hold not only for a constant infectious period, but also when the infectious period is random.
In Section 6 we obtain the distribution for the time until the epidemic dies out completely. In order to do so, we first derive the degree distribution of the individuals that escape the epidemic and the fraction of the neighbors of the escaping vertices that also escape the epidemic.

2 Notation

An *undirected graph* is an ordered couple $G = (V, E)$ of a set of vertices or nodes, $V = \{v_i\}$ and a set of edges, E, whose elements are of the form $\{v_i, v_j\}$. If edges are defined as ordered couples (v_i, v_j), then G is a *directed graph*. A graph is random if its edges are chosen according to a certain random mechanism.

The asymptotic notations are defined as follows.

$$f(x) = O(g(x)), \text{ as } x \to a \text{ if } \limsup_{x\to a} |f(x)/g(x)| < \infty,$$
$$f(x) = o(g(x)), \text{ as } x \to a \text{ if } \lim_{x\to a} f(x)/g(x) = 0,$$
$$f(x) = \theta(g(x)), \text{ as } x \to a \text{ if } 0 < \liminf_{x\to a} |f(x)/g(x)| < \limsup_{x\to a} |f(x)/g(x)| < \infty,$$
$$f(x) \approx g(x), \text{ as } x \to a \text{ if } \lim_{x\to a} (f(x) - g(x)) = 0,$$
$$f(x) \sim g(x), \text{ as } x \to a \text{ if } \lim_{x\to a} \frac{f(x)}{g(x)} = 1.$$

Further we define the convergence of a sequence of random variables (to a limiting random variable). There are several types of stochastic convergence. Let X and X_n, $n = 1, 2, ...$ be random variables defined on a probability space $(\Omega, \mathscr{F}, P)$.

Let F_n and F be the distribution functions of the variables X_n and X, respectively. A sequence $X_1, X_2, ...$ *converges in distribution*, or *weakly*, or *converges in law*, if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every x at which F is continuous. We write $X_n \xrightarrow{D} X$.

The sequence of random variables (X_n) converges *in probability* towards X if for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

We write $X_n \xrightarrow{P} X$.

The sequence (X_n) converges *almost surely*, or *strongly*, or *with probability* 1 if

$$P\left(\omega \in \Omega | \lim_{n\to\infty} X_n(\omega) = X\right) = 1.$$

We write $X_n \xrightarrow{a.s.} X$.

Assume $r > 0$. The sequence (X_n) converges *in r.th mean* or *in* L^r *norm* if the r.th absolute moments of X_n and X exist and

$$\lim_{n\to\infty} E(|X_n - X|^r) = 0,$$

where E denotes expectation. We denote the L^r convergence by $X_n \xrightarrow{L^r} X$.
One can find more details on the topic of stochastic convergence in Chapter 7 of [11], for example.

We will also need a part of d'Alembert's ratio test for series convergence and we prove it here (see Section 9.2 in [1]).

Theorem 2.1. *Let* a_k *be positive for* $k = 1, 2, ...$
1. If there exists k_0 *such that for all* $k > k_0$ *we have* $\frac{a_{k+1}}{a_k} \leq l < 1$*, then* $\sum a_k < \infty$.
2. If $\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = l < 1$*, then* $\sum a_k < \infty$.

Proof. 1. For all $k > k_0$ it holds that $a_{k+1} \leq l \cdot a_k$. This implies that $a_{k+m} \leq l^m a_k$, for $k > k_0$. Therefore, the tail of the series $\sum_k a_k$ is bounded by the tail of a convergent series $a_{k_0} \sum_k l^k$ and we may infer that $\sum_k a_k < \infty$.

2. We pick an $\epsilon > 0$ such that $l + \epsilon < 1$. Since $\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = l$, there exists a k_1 such that for all $k > k_1$ it holds that $\frac{a_{k+1}}{a_k} < l + \epsilon$. The result follows from 1. □

3 Branching processes

The results presented in this chapter are already well known. We follow the approach of [12], [3] and [14] and one can consult these for the detailed proofs.

The theory of branching processes was first introduced with Galton-Watson branching processes, in order to solve the following problem ([12]):

Let $p_0, p_1, p_2, ...$ be the probabilities that a man has $0, 1, 2...$ sons respectively, and let each son have the same probability for sons of his own, and so on. What is the probability that a male line goes extinct after r generations, and more generally, what is the probability for any given number of descendants in the male line in any given generation?

The early stages of an infectious disease propagating through a large population can be approximated by a branching process. We assume that there are a large number of susceptibles and a very small number of infectives in the beginning of an outbreak. The argument that makes the approximation valid is that given such assumptions, the probability of infecting a person who has already been infected is very small at the early stages.

At first we only keep track of the generation process, that is without concerning ourselves with the time of birth, or in our case time of infecting a neighbor. This background is enough when we try to answer the questions about the distribution of the ultimately susceptible and the fraction of neighbors of an ultimately susceptible which are also ultimately susceptible - as these questions do not depend on the real time of infection occurrences. In other words, we first consider the Galton-Watson branching process.

In order to estimate the time needed for the epidemic to die out, we use the general branching process.

3.1 The Galton-Watson branching process

We assume that all the individuals are of the same type and that they spread the infection independently of how many other infectives there are at that point of time. We denote by $Z_0, Z_1, Z_2...$ the number of individuals in the 0.th, first... generation. Assume that $Z_0 = 1$. The sequence is defined as a Markov chain, or in other words, the number of infecteds in $(i+1)$.st generation only depends on the number of i.th generation infecteds and not on generations preceding i. Transition probabilities do not change over time.

We denote $P(Z_1 = k) = p_k$, the probability that a person infected in generation i infects k people in generation $(i + 1)$. Distribution (p_k) is independent of i.

Given that individuals of the same generation spread the infection independently, if there are k people in generation i, then Z_{i+1} will be the sum of k independent random variables distributed as Z_1. If $Z_i = 0$, then $Z_j = 0$ for every $j > i$ (so, 0 is an absorbing

state).

3.1.1 Generating functions of the generation sizes

Branching processes can be analyzed using the properties of a generating function, $G(z) = \sum_{k=0}^{\infty} p_k z^k$, $|z| \leq 1$. We will always assume $p_0 + p_1 < 1$, so the generating function is strictly convex. Also, we assume the first and second moment of (p_k) to be finite.

Iterates will be denoted with a subscript, $G_0(z) = z$, $G_{i+1}(z) = G(G_i(z))$. It can be proved by induction that also $G_{i+j}(z) = G_i(G_j(z))$ and in particular, $G_{i+1}(z) = G_i(G(z))$. We denote the generating function of Z_i by $G_{(i)}(z)$. As we will see in the following theorem, $G_{(i)} = G_i$.

Note that the generating function of the sum of two independent random variables is the product of the generating functions of those random variables. Again, using induction one can infer that the generating function of a sum of independent discrete random variables is the product of the generating functions of the summands.

Theorem 3.1. *The generating function of the size of the i.th generation, Z_i is the i.th iterate $G_i(z)$.*

Proof. Note that, if there are k infectives in generation i, the generating function of Z_{i+1} is $(G(z))^k$, for $k = 0, 1, ...$ Therefore,

$$G_{(i+1)}(z) = \sum_{k=0}^{\infty} P(Z_i = k)(G(z))^k = G_{(i)}(G(z)).$$

By definition we have that $G_{(0)}(z) = G_0(z)$, and by induction $G_{(i)}(z) = G_i(z)$. □

We now have the distribution of the i.th generation size and can calculate the moments. Define $\mu = EZ_1$, $\sigma^2 = Var(Z_1) = EZ_1^2 - \mu^2$. Note that $\mu = G'(1)$ and $\sigma^2 = G''(1) - \mu^2 + \mu$. Differentiating the iterates, we get

$$G'_{i+1}(1) = G'(G_i(1))G'_i(1) = G'(1)G'_i(1).$$

Again induction yields $G'_i(1) = \mu^i$. If $G''(1) < \infty$,

$$G''_{i+1}(1) = G'(1)G''_i(1) + G''(1)\left(G'_i(1)\right)^2. \tag{1}$$

Repeating 1, $Var(Z_i)$ can be calculated. Namely, the following theorem holds.

Theorem 3.2. *The expected value of i.th generation number of infectives is $EZ_i = \mu^i$. If $Var(Z_1) < \infty$, then*

$$Var(Z_i) = EZ_i^2 - (EZ_i)^2 = \begin{cases} \frac{\sigma^2\mu^i(\mu^i-1)}{\mu^2-\mu} & \text{if } \mu \neq 1, \\ n\sigma^2 & \text{if } \mu = 1. \end{cases}$$

□

3.1.2 The extinction probability

Extinction of a branching process is the event that all but finitely many Z_i values are 0. Since $P(Z_{i+1} = 0|Z_i = 0) = 1$, we have

$$\begin{aligned} q := P(Z_i \to 0) &= P(\exists i, Z_i = 0) \\ &= P[(Z_1 = 0) \cup (Z_2 = 0)...] \\ &= \lim_{i\to\infty} P[(Z_1 = 0) \cup ... \cup (Z_i = 0)] \\ &= \lim_{i\to\infty} P(Z_i = 0) = \lim_{i\to\infty} G_i(0). \end{aligned}$$

Since $\lim_{i\to\infty} G_i(0) = \lim_{i\to\infty} G_{i+1}(0)$ and G is continuous, the extinction probability q satisfies $q = G(q)$.
This has an interpretation that, in order for the process to go extinct, it is necessary and sufficient that all the families of the first generation vertices go extinct. Moreover, the following holds.

Theorem 3.3. *If $\mu \leq 1$, the extinction probability q equals 1. If $\mu > 1$, the extinction probability is the unique solution in $[0, 1)$ of the equation*

$$s = G(s).$$

□

For a detailed proof, we refer to Chapter 1 in [12].

Another important property is the instability of the process.

Theorem 3.4. *Regardless of the mean value μ, every state $k = 1, 2, ...$ of the process (Z_i) is transient, or $\lim_{i\to\infty} P(Z_i = k) = 0$. Moreover, $Z_i \to \infty$ with probability $1 - q$ and $Z_i \to 0$ with probability q.* □

This behavior is not in line with most biological systems, that tend to reach (quasi) equilibrium after a certain time. Nevertheless, branching processes work well for describing the first stages of many biological processes.

If $\mu < 1$, we call the branching process *subcritical*, if $\mu = 1$ *critical* and if $\mu > 1$ *supercritical*. The last will be of interest in this thesis, as it corresponds to the case when an epidemic takes off and affects a major fraction of the population. Also, subcritical branching processes will be used in Section 6.3.

3.1.3 Scaling of the generation sizes

For Galton-Watson processes it holds that $E(Z_i) = E(Z_0)\mu^i$. Therefore, if the process is supercritical, the expected generation size grows exponentially. In the case of a generation size itself, Z_i, the following holds. Let $W_i = Z_i/\mu^i$.

Theorem 3.5. *If $E(Z_1 \log(Z_1)) < \infty$ the martingale W_i converges almost surely and in L^1 norm to a non-degenerate random variable W which is 0 if and only if $Z_i \to 0$.*

3.2 The general branching process

The theory of Galton-Watson processes deals with the number of infections in a generation. However, it is possible that an infective individual infects one of his or her neighbors, and continues to be infective for some time longer. In other words, we may have two individuals of different generations being infective at the same time point. Furthermore, it is not possible to observe the number of individuals infected in a generation. We would rather be able to obtain the number of the infected individuals at a point of time. Therefore, we need a more refined model to deal with the questions regarding time and for that purpose we consider the general process, as it was done in Chapter 6 of [14].

We will assume that there is only one ancestor, denoted by 0. An individual x who is the j_k.th child of the j_{k-1}.th child of... of the j_1.st child of the ancestor will be denoted by $x = (0, j_1, ...j_k)$. The set of all possible individuals is

$$J = \left\{ 0 \cup \bigcup_{k=1}^{\infty} \{(0, x) | x \in \mathbb{N}^k\} \right\}.$$

Note that J is countable. However, not every possible individual is necessarily realized. For example, if the ancestor has one child, there are no individuals of the form $(0, 2, j_2, ..., j_k)$.

The realized individuals that belong to $\left\{(0, x) | x \in \mathbb{N}^k\right\}$ constitute the k.th generation of the embedded Galton-Watson process [14]. Note that the general process dies out if and only if the embedded Galton-Watson process does.

In this thesis only the subcritical general branching processes will be considered.

3.2.1 The subcritical case

We introduce the following notation. We also comment on Malthusian parameter and calculate it for our model.

L is the distribution function of the life time l.
φ is the reproduction process of an arbitrary individual.
μ is $E(\varphi(\cdot))$, the reproduction function.

α is the *Malthusian parameter*, that is, the unique solution, if one exists, of the following integral equation

$$\int_0^{\infty} e^{-\alpha t} \mu(dt) = 1. \tag{2}$$

The parameter α can be interpreted as the exponential growth rate. If the infectious rate is constant, say β, during the infectious period, then the mean number of infections

during the interval $[t, t+dt)$ equals

$$\mu(dt) = \beta P(l > t)dt.$$

Assume that $L(0) = 0$. Then, by partial integration we obtain

$$\int_0^\infty e^{-\alpha t}\mu(dt) = \frac{\beta}{\alpha}\left(1 - \int_0^\infty e^{-\alpha t}L(dt)\right)$$

and the Malthusian parameter is the solution of the equation

$$\frac{\beta}{\beta - \alpha}\int_0^\infty e^{-\alpha t}L(dt) = 1. \tag{3}$$

If a solution exists, it must be unique [12]. This is because the integral is monotone, as a function of α. In supercritical case, α exists and it is positive. In subcritical case, α may not exist and if it does, then it must be negative [14] [12]. If $\alpha < 0$, the integral diverges if the rate at which the tail of the life time density L' declines is slower than exponential, roughly speaking. If it diverges, the Malthusian parameter does not exist.

Example. We will calculate the Malthusian parameter under the following assumptions. Let an individual make infectious contacts according to a Poisson process with per pair intensity β. We will denote the mean number of neighbors by ν. Let the infectious period I be exponentially distributed, with parameter λ. The mean number of infections in the interval $[t, t+dt)$ equals

$$\mu(dt) = (\nu - 1)\beta P(l > t)dt,$$

as an individual can infect all of his or her neighbors, apart from the one that infected him or her. Then (2) becomes

$$(\nu - 1)\beta\int_0^\infty e^{-\alpha t}P(l > t)dt = 1,$$

which yields

$$\alpha = (\nu - 1)\beta - \lambda. \quad \square \tag{4}$$

$\hat{\varphi}$ is the Laplace-Stiltjes transform of φ. For example,

$$\hat{\varphi}(\alpha) = \int_0^\infty e^{-\alpha t}\varphi(dt).$$

z_t or z_t^∞ denotes the number of individuals alive and younger than ∞ at time t.

Theorem 3.6. *Consider a subcritical non-lattice process with Malthusian parameter α. Assume that the following two conditions hold.*

$$\int_0^\infty te^{-\alpha t}L(dt) < \infty$$

and

$$\int_0^\infty te^{-\alpha t}\mu(dt) < \infty.$$

Then

$$\lim_{t\to\infty} e^{-\alpha t}P(z_t > 0)$$

exists. The limit is strictly positive if

$$E(\hat{\varphi}(\alpha)\log(\varphi(\infty))) < \infty.$$

□

4 Configuration model

4.1 The random graph

Although it is certain that a model is wrong, we would nevertheless like to make ours useful. One of the first characteristics of a real world network that we would like the model to replicate is the degree distribution. The degree of a vertex v (sometimes also called the valence) is the number of edges incident to v. For a very detailed and mathematically rigorous overview of the terminology in graph theory, we refer to the first chapter of [8].

One of the most simple and best studied models in the field of random graphs is that of Erdős and Rényi. The most common version of this graph assumes that there are n vertices, and between any two of them an edge is present with probability p, independently of the existence of the other edges. This means that the degree of v will be equal to k with probability $p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}$, for $k = 0, 1, 2, ...$ If we assume $p = \lambda/(n-1)$, the probability p_k approaches $\frac{\lambda^k e^{-\lambda}}{k!}$ as $n \to \infty$. Therefore, the degree distribution in an Erdős and Rényi graph is binomial, or asymptotically Poisson, as we want p to decrease with n increasing, in order for the mean degree of an arbitrary vertex to be fixed. However, in many real-world networks the degree distribution is not Poisson and they are better fitted with a power law degree distribution, which means $p_k \approx Ck^{-\beta}$, for large n and k [20]. In order to circumvent the problem, random graphs with given degree sequence, or configuration models, have been studied (Figure 2).

The configuration model is constructed as follows. First we draw degrees for the vertices d_i $(i = 1, ..., n)$ from the distribution of a random variable D. We denote $p_i = P(D = i)$. Even though drawn as independent numbers from the distribution of D, the degrees are not strictly mutually independent. Apparently, $l_n = \sum_{i=1}^{n} d_i$ must be even for the sequence to lead to a proper graph when n is finite (Handshaking lemma, Euler 1736). In particular, $2m = \sum_{i=1}^{n} d_i$, where m is the number of edges. Therefore, we would have to condition on this event. If d_i is even with probability $p \in (0, 1)$, then the sum is even with a probability that approaches $1/2$, as $n \to \infty$ [13].

However, that will not make much difference. We could, for example add a half-edge to the vertex labeled as n, if l_n is odd. That way instead of D, we consider the sequence of degrees drawn from the distribution of the new variable, $M = D + I_{\{l_n odd, i=n\}}$. Nevertheless, if we pick a vertex uniformly at random,

$$P(D = M) = P(I_{\{l_n odd, i=n\}} = 0) = 1 - \frac{1}{2n} \to 1, \quad n \to \infty,$$

and M converges to D in probability. In a similar manner we could prove that, given $ED < \infty$ and $ED^2 < \infty$, $EM \to ED$ and $EM^2 \to ED^2$, respectively. Therefore, we omit conditioning.

A multigraph with given degree sequence can be constructed as follows (a multigraph is defined in the following paragraph). Once $(d_1, ..., d_n)$ is specified, vertex i is assigned d_i half edges emerging from i. Further, in each step, if not already occupied, two of the half edges are paired and labeled as occupied. The procedure is terminated when there are no more unoccupied half edges. As every half-edge can be paired with each of the remaining half-edges, self loops and multiple edges can occur. (Figure 2)

A *self-loop* is an edge with the same starting and ending vertex, which we can interpret as a vertex being in contact with itself. *Multiple edges* are edges that share both end-vertices, and we can interpret that the existence of multiple edges between a couple of vertices means increased probability of contact. If a graph does not contain self-loops or multiple edges, it is called *simple*. If both are permitted, we call it a *multigraph*. In this work we allow for both self loops and multiple edges, since under mild conditions, there will "not be many" of either. Namely, the following theorem holds.

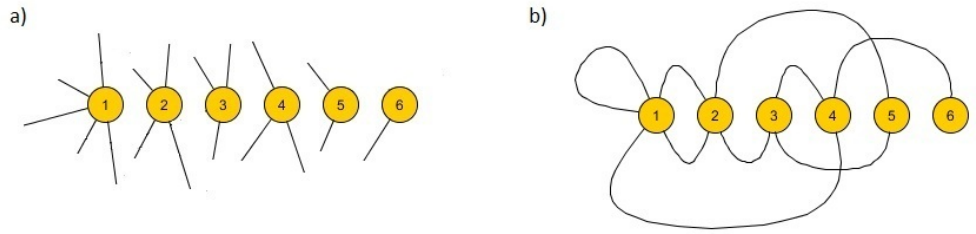

Figure 2: a) Sequence of degrees. b) A graph with corresponding degree sequence. Vertex 1 contains a self loop. There are multiple edges connecting vertices 1 and 2.

Theorem 4.1. *Assume that distribution (p_i) of the random variable D has finite second moment and denote $\nu = \frac{E(D(D-1))}{ED}$. When $n \to \infty$, the number of self-loops S_n and the number of multiple edges M_n are asymptotically independent $Poisson(\nu/2)$ and $Poisson((\nu/2)^2)$ random variables.* □

The proof can be found in Chapter 7 of [13]. An outline of the same proof is also available in Chapter 3 in [9]. Theorem 4.1 has an important corollary. The notation is the same as above.

Theorem 4.2. *The probability that a configuration with degree sequence (d_i) gives a simple graph is asymptotically equal $e^{-\nu/2-\nu^2/4}$.*

Proof. Follows from theorem 4.1, as a graph is simple if and only if $S_n = M_n = 0$. The probability of that event converges to the probability that $S = M = 0$, which equals $e^{-\nu/2-\nu^2/4}$. □

One can also construct a random graph with sequence of degrees (d_i) in the following way. First, make l_n vertex copies, d_i of which are identified with vertex i. Pick a starting vertex copy and then pick its pair uniformly from the rest of the vertex copies. Take the

following unpaired copy and take its pair uniformly at random from the vertex copies that have not been matched yet. The procedure ends when there are no more copies to match. Each graph obtained in that way (with l_n vertices, each of degree 1) is called a *configuration*, and hence the same name for the initial model.

Note that such configuration corresponds to an instance of a random graph with the same degrees and that each configuration is equally probable. From the construction it follows that there are $(l_n - 1)!!$ configurations with the same number of half-edges, l_n. (Recall that $x!!$ denotes the product of all the odd natural numbers smaller or equal than x, if x is odd and the product of all the even natural numbers smaller or equal than x, if x is even.)

Denote by g_{ij} the number of edges connecting vertices i and j. This also means g_{ii} is the number of self-loops of the vertex i. Then it holds $d_i = g_{ii} + \sum_j g_{ij}$. Thus, a self loop adds 2 to the degree of a vertex that contains one. As it has been proven in [13], the probability of a multigraph that has edges determined by the values g_{ij} is then given by

$$\frac{1}{(l_n - 1)!!} \frac{\prod_i d_i!}{\prod_i 2^{g_{ii}} \prod_{1 \le i \le j \le n} g_{ij}!}. \tag{5}$$

To see that this holds, recall that each configuration has probability $\frac{1}{(l_n-1)!!}$. Note that if the edges incident to the same vertex are permuted, the graph does not change. However, the configuration will change, unless we permute the multiple edges. We can also permute the self-loops and this will not change the configuration either. Therefore, we divide with $\prod_{1 \le i \le j \le n} g_{ij}!$. Additionally, if we have a self-loop, the configuration does not change if the vertices at the ends of the same edge change places, which explains the term $\frac{1}{\prod_i 2^{g_{ii}}}$. To sum up, in order to obtain the probability of a multigraph, we add the probabilities of the different configurations that give rise to that multigraph.

Note that the formula above yields that all the simple graphs are equally probable, since for a graph to be simple, it is necessary and sufficient that the following two conditions hold: $g_{ij} \in \{0, 1\}$ and $g_{ii} = 0$, for $i, j = 1, 2, ...$

Therefore, by the above procedure, for the given sequence of degrees, we pick one of the equally probable configurations, and emphasize that not all multigraphs are equally probable. In fact, from (5) we see that among the graphs with the given degree sequence (d_i), the most probable are the simple ones. Compare this to the result of the Theorem 4.2, that the asymptotic probability of a graph with given degree sequence to be simple is positive (namely, equal to $e^{-\nu/2-\nu^2/4}$), in the case of finite ν.

4.2 Phase transition

A *giant component* of an undirected graph is a connected component which contains the number of vertices of order $O(n)$, as $n \to \infty$. Otherwise a component is *small*.

There exists a threshold value above which a graph has a (single) giant component and below which it only contains small components ([18], [13]). Let $|\cdot|$ denote the number of vertices of a graph or a component, $v_k(\cdot)$ the number of vertices of degree k and $E(\cdot)$ the number of edges of a graph or a component. In Chapter 10 of [13] the following was proved. Note that the theorem also states that, if there is a giant component, it must be unique.

Theorem 4.3. *Suppose $E(D) < \infty$ and consider the random graph on n vertices with degrees distributed as the random variable D. Let n tend to ∞. Let C_{max} be the largest and $C_{(2)}$ the second largest component. Then the following holds.*

1. *If $\nu = \frac{E[D(D-1)]}{E(D)} > 1$, then there exist $\xi \in [0,1)$ and $\varsigma \in (0,1]$ such that*

$$\begin{aligned} |C_{max}|/n &\xrightarrow{P} \varsigma, \\ v_k(C_{max})/n &\xrightarrow{P} p_k(1-\xi^k), \text{ for every } k > 0, \\ E(C_{max})/n &\xrightarrow{P} \frac{1}{2}E(D)(1-\xi^2), \end{aligned}$$

   while $|C_{(2)}|/n \xrightarrow{P} 0$ and $E(C_{(2)})/n \xrightarrow{P} 0$.

2. *If $\nu = \frac{E[D(D-1)]}{E(D)} \leq 1$, then $|C_{max}|/n \xrightarrow{P} 0$ and $E(C_{max})/n \to 0$.*

□

4.2.1 Branching process interpretation

We consider the following branching process (Z_n). The first generation vertices have degree distribution (p_i), the same as the random variable D, and the mean value will be denoted by μ. Such graphs with finite mean degree are called *sparse*. Here we will also assume that the second moment of the distribution (p_i) is finite. Most of this subsection is explained in [13].

If a vertex is chosen as a neighbor, it is more likely to be of higher degree. Recall the definition of a branching process from Section 3, where the ancestor cannot be the offspring as well. Therefore we ignore the fact that a new infected can send the infection back to the one who infected him or her. That being said, a vertex must be of degree $(k+1)$ in order to have k "descendants", so the second and subsequent generation vertices have the degree distribution

$$\tilde{p}_k = \frac{(k+1)p_{k+1}}{\sum kp_k}.$$

We denote the new mean $\nu = \sum_{k=0}^{\infty} k\tilde{p}_k = \sum_{k=1}^{\infty} k(k-1)p_k/\mu$. Therefore, we actually consider a two-phase branching process.

One problem that arises with this approximation is that in reality the distribution of the offspring will change because the number of half-edges to connect to is decreasing and thus vertices with smaller degree become more probable. However, at least $o(n)$ half-edges can be connected before the distribution changes: let N_k denote the number of k degree vertices. After we have connected $o(n)$ half-edges, the number of unexposed vertices of degree k is in the interval $[N_k - o(n), N_k + o(n)]$ and we have

$$\frac{N_k - o(n)}{n - o(n)} \to \frac{N_k}{n} \quad \text{and} \quad \frac{N_k + o(n)}{n - o(n)} \to \frac{N_k}{n},$$

so the new fraction of k degree vertices is the same as the old one, in the limit of large n. Compare this to the result in Section 5.1.
The generating functions of (p_i) and $(\tilde{p}_j)$ are related. Let $G_0(z) = \sum_k p_k z^k$ and $G_1(z) = \sum_k \tilde{p}_k z^k$, then

$$\frac{G_0'(z)}{\mu} = \sum_{k=1}^{\infty} \frac{k p_k}{\mu} z^{k-1} = \sum_{k=1}^{\infty} \tilde{p}_{k-1} z^{k-1} = G_1(z). \tag{6}$$

The total progeny has the following expectation.

$$E\left(\sum_{k=0}^{\infty} Z_k\right) = 1 + \sum_{k=1}^{\infty} \mu \nu^{k-1} = \begin{cases} 1 + \frac{\mu}{1-\nu} & \text{if } \nu < 1, \\ \infty & \text{if } \nu \geq 1. \end{cases}$$

Note that $\nu = \sum_k k(k-1)p_k/\mu$ and the threshold value are the same as in Theorem 4.3.

Let ξ be the smallest fixed point of the function G_1 in $[0, 1]$. Thus defined ξ is the extinction probability of the homogenous branching process with offspring distribution $(\tilde{p}_n)$. The two phase branching process will go extinct if and only if all the families of a first generation vertex go extinct. The probability of that event is $\sum_{k=0}^{\infty} p_k \xi^k = G_0(\xi)$. Therefore, the two phase branching process survives with probability $1 - G_0(\xi)$. This corresponds to ς in Theorem 4.3.

We pick a vertex of degree k. The families of its neighbors all go extinct with probability ξ^k. The probability that the branching process starting in that vertex will survive is equal $p_k(1 - \xi^k)$, corresponding to the fraction of vertices in the large component, $\frac{v_k(C_{max})}{n}$.

Pick an arbitrary edge. It will belong to a surviving branching process if and only if at least one of its end vertices is a starting point of a surviving branching process. That happens with probability $(1 - \xi^2)$. The overall number of edges is $\frac{l_n}{2} \approx \frac{n\mu}{2}$. The chosen edge is part of a surviving process with probability $\frac{1}{2}\mu(1-\xi^2)$, equal to the probability that it belongs to the giant component, given in Theorem 4.3.

To sum up, ς can be interpreted as survival probability, ξ as extinction probability of the homogenous branching process with offspring distribution $(\tilde{p}_k)$.

4.3 Small world results

The *distance* between two vertices in a graph is the length of the shortest path between them. The *diameter* is the longest distance in the graph. The following definitions are taken from [13]. We denote by H_n the distance between two uniformly chosen connected vertices (which belong to the same component). We call H_n the *typical distance* of the graph.

The limiting random graph is a *small world* if there exists a constant K such that

$$\lim_{n\to\infty} P(H_n \leq K \log n) = 1.$$

The following results are about typical and longest distances between vertices in a configuration model. They are proved in [13], in Chapter 10. We cover the small world results for both finite and infinite variance. The following theorem provides the limiting distribution of typical distances in a configuration model with finite variance.

Theorem 4.4. *Assume that in a configuration model on n vertices both μ and ν are finite and assume that $\nu > 1$. Let H_n be the typical distance between two vertices in the model. Conditionally on $H_n < \infty$, the following holds*

$$\frac{H_n}{\log n} \xrightarrow{P} \frac{1}{\log \nu}. \tag{7}$$

□

The typical distance in the case of finite mean and infinite variance of the distribution (p_k) is given in the following theorem. First we define the *cumulative distribution function* $F_n(x)$ of (p_k), as the proportion of vertices having degree at most x. We assume that there exists a $\tau \in (2,3)$ and that for all $\delta > 0$ exist $c_1 = c_1(\delta)$ and $c_2 = c_2(\delta)$ such that

$$c_1 x^{-(\tau-1+\delta)} \leq [1 - F_n](x) \leq c_2 x^{-(\tau-1-\delta)}, \tag{8}$$

where the upper bound is required to hold for every $x \geq 1$ and the lower bound only for $1 \leq x \leq n^{\alpha}$, for some $\alpha > \frac{1}{2}$. The following theorem holds.

Theorem 4.5. *Assume that in a configuration model on n vertices the mean μ is finite. Conditionally on $H_n < \infty$, the following holds*

$$\frac{H_n}{\log\log n} \xrightarrow{P} \frac{2}{|\tau - 2|}. \tag{9}$$

□

Recall that ξ is the extinction probability of the homogenous branching process with offspring distribution $(\tilde{p}_k)$. Further we define

$$\eta = G_1'(\xi) = \sum_{k=0}^{\infty} \tilde{p}_k k \xi^{k-1}.$$

For the diameter of a configuration model on n vertices the following holds, for both finite and infinite variance case.

Theorem 4.6. *Assume that in a configuration model on n vertices the mean μ is finite. Let p_1 and p_2 denote the proportion of degree one and degree two vertices in the graph, respectively. Then,*

$$\frac{diam(G)}{\log n} \xrightarrow{P} \frac{1}{\log \nu} + \frac{2 - I_{\{p_1=0\}} - I_{\{p_2=0\}}}{|\log \eta|}.$$

□

Remark. Theorem 2.1 implies that, if $\xi < 1$, then the series $\sum k\xi^{k-1}$ converges and therefore also $\eta < \infty$. We will prove later that with $\xi < 1$ it must hold that $\eta < 1$ (see the proof of Theorem 6.3, where η corresponds to R_0^{end} in the case $\psi = 1$). If $\xi = 1$, then $\eta = \nu$. However, we are only interested in the case $\xi < 1$, when a large epidemic outbreak is possible.

From the results listed in this section, it is of interest to note that in the case of finite mean and infinite variance of (p_k), even though the typical distances are of order $O(\log \log n)$, the diameter is of order $O(\log n)$. In the infectious disease spread terminology, even though most get infected within the first $O(\log \log n)$ generations, the time required for the infection to completely die out is $O(\log n)$. It is such distributions (more precisely, the distributions satisfying the condition (8)) that have shown to be the best fit for real world networks, according to the simulations presented in [20]. The condition that the variance is infinite corresponds to the presence of enough vertices with high degrees (so called *hubs*), that will acquire the infection in the early stages and further spread it relatively quickly. The presence of enough hubs results in a lower typical distance, but it does not have much impact on the diameter.

5 SIR epidemic

We consider a population that consists of n individuals and assume that they are connected according to a configuration model contact network. An SIR epidemic spreads along the network, starting with a very small number of initially infective, compared to n. A person can only go from the state Susceptible to Infected and from Infected to Recovered. If one is recovered, he or she plays no further role in the epidemic spread. Suppose we have a large outbreak, that is the size of the epidemic is of order $O(n)$. The number of susceptibles is initially large and it decreases throughout the epidemic. The number of infected individuals is initially small, then it grows until it peaks and starts to decline. At the end it becomes zero, when all the individuals that have ever acquired the disease recover (Figure 1). The number of recovered individuals is initially zero and it increases throughout the epidemic. The number of the ultimately recovered is the size of the epidemic. At the beginning and at the end of the epidemic, while the number of infected individuals is small (of order $o(\sqrt{n})$, see Section 5.1), we approximate the spread by a branching process. While the number of infected and the number of susceptible individuals are both large, differential equations are used in order to describe the time dynamics of the process. The beginning and the middle part, until we reach the generations of size $O(\epsilon n)$ in the declining part of the epidemic, have already been well studied (for instance in [7], [17] and [24], the last two references focusing on the differential equations approximation).

In this section we prove that in a configuration model a branching process is a sensible approximation for an epidemic until the number of infections reaches $o(\sqrt{n})$, $n \to \infty$. Then, we introduce R_0 and calculate it for a configuration model. R_0 has the same role as the mean degree in a Galton-Watson branching process. It thus yields a criterion for determining when an outbreak must be small. In Section 5.4 we provide the same criterion, in yet another way. In Section 5.3 we calculate the probability of a large outbreak. Also, we emphasize a mistake that might occur should one not take into account the dependence of the number of infecetds by a fixed infective individual on the duration of his or her infectious period, in the case of a random infections period (this will be explained in detail later).

5.1 Branching process approximation

In this section we prove that if the number of infectious contacts N is of order $o(\sqrt{n})$, then the probability of a repeated person in transmission of the disease can be made arbitrarily small, for n large enough.

We obtain this result in a similar manner as we would solve the Birthday problem, which is the following: *There are N people in a room. What is the probability that no two people have the same birthday?* We assume that all the days of the year are equally

probable to be a birthday. If $N \leq 365$, the probability we want is

$$1 \cdot \left(1 - \frac{1}{365}\right) \cdot ... \cdot \left(1 - \frac{N-1}{365}\right).$$

We now precede to the problem of loops in a branching process. Let X_i be the number of neighbors of the i.th exposed individual. Recall that μ denotes the mean degree of the first generation vertex and ν the mean degree of the second and subsequent generation vertices in a branching process and that we assume both μ and ν to be finite. In each step we try to avoid the already infected persons, or the half edges with the already infected as their end vertices. Note that in order to have a branching process we need to allow for choosing a half-edge more than once, as otherwise dependencies appear. We show that choosing an already connected half-edge happens only late in the process. The probability of no loops is

$$\begin{aligned}
&\prod_{i=1}^{N}\left(1 - \frac{\sum_{k=1}^{i} X_k}{n\mu}\right) \\
&\geq \prod_{i=1}^{N}\left(e^{-\frac{\sum_{k=1}^{i} X_k}{n\mu}} - \left(\frac{\sum_{k=1}^{i} X_k}{n\mu}\right)^2\right) \\
&\approx \prod_{i=1}^{N} e^{-\frac{\sum_{k=1}^{i} X_i}{n\mu}} - \sum_{i=1}^{N}\left(\frac{\sum_{k=1}^{i} X_k}{n\mu}\right)^2 e^{-\sum_{j=1}^{N}\sum_{k=1}^{j}\frac{X_k}{n\mu}+\sum_{k=1}^{i}\frac{X_k}{n\mu}} \\
&\geq e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}} - \sum_{i=1}^{N}\left(\frac{\sum_{k=1}^{N} X_k}{n\mu}\right)^2 e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}+\sum_{k=1}^{i}\frac{X_k}{n\mu}} \\
&= e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}} - \left(\frac{\sum_{k=1}^{N} X_k}{n\mu}\right)^2 e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}} \sum_{i=1}^{N} e^{\sum_{k=1}^{i}\frac{X_k}{n\mu}} \\
&\geq e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}} - \left(\frac{N\frac{1}{N}\sum_{k=1}^{N} X_k}{n\mu}\right)^2 e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}} N \cdot e^{\frac{N\frac{1}{N}\sum_{k=1}^{N} X_k}{n\mu}} \\
&= e^{-\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu}}\left(1 - \left(\frac{N\frac{1}{N}\sum_{k=1}^{N} X_k}{n\mu}\right)^2 N \cdot e^{\frac{N\frac{1}{N}\sum_{k=1}^{N} X_k}{n\mu}}\right) \\
&=: P.
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{\sum_{k=1}^{N}(N-k+1)X_k}{n\mu} &= \frac{N(N-1)\frac{1}{N(N-1)}\sum_{k=1}^{N}(N-k+1)X_k}{n\mu} \\
&\leq \frac{N(N-1)\frac{1}{N-1}\sum_{k=1}^{N} X_k}{n\mu}.
\end{aligned}$$

The last inequality holds because

$$\sum_{k=1}^{N}(N-k+1)X_k \leq \sum_{k=1}^{N} NX_k$$

implies

$$\frac{1}{N(N-1)}\sum_{k=1}^{N}(N-k+1)X_k \leq \frac{1}{N-1}\sum_{k=1}^{N} X_k.$$

Therefore,

$$P \geq e^{-\frac{N(N-1)\frac{1}{N-1}\sum_{k=1}^{N} X_k}{n\mu}} \left(1 - \left(\frac{N\frac{1}{N}\sum_{k=1}^{N} X_k}{n\mu}\right)^2 N \cdot e^{\frac{N\frac{1}{N}\sum_{k=1}^{N} X_k}{n\mu}}\right).$$

According to the Law of Large Numbers, the expression $\frac{1}{N-1}\sum_{k=1}^{N} X_k = \frac{N}{N-1}\frac{1}{N}\sum_{k=1}^{N} X_k$ tends to ν, as n tends to infinity. Therefore, if N is $o(\sqrt{n})$, the probability of no loops will approach 1 in the limit of large n.

The first inequality of the above calculations holds as $e^{-x} \leq 1 - x + x^2$, according to the Taylor polynomial approximation of e^{-x} in the neighborhood of $x_0 = 0$. We also remark that the approximation in the third line holds with the assumption that N is $o(n)$, and that it means that the difference of the expressions on the right and left hand side approaches 0, in the limit on large n (see Section 2).

5.2 Basic reproduction number

The basic reproduction number R_0 can be defined in more than one way. We define it as the average number of secondary infections caused by an infected in the early stages of an epidemic (but not the initial case). This definition corresponds to the branching process definition of ν. The approximating branching process survives with positive probability if and only if $R_0 > 1$. Note that R_0 is not unique in having the property. Indeed, the same holds if we consider $f(R_0)$ instead, where f is an increasing function such that $f(1) = 1$ [23], [2]. We consider a second generation vertex, as in the early stages most vertices spread the infection following that distribution. Let

$$\psi = \int_0^\infty (1 - e^{-\beta t}) f_I(t) dt \tag{10}$$

be the probability of disease transmission between a susceptible-infected pair, during the infectious period of the infected. The basic reproduction number is defined as

$$\begin{aligned} R_0 =& \psi \sum_{k=1}^{\infty} (k-1) \cdot \frac{kp_k}{\sum_{k'=1}^{\infty} k' p_{k'}} \\ =& \psi\nu. \end{aligned}$$

In other words,

$$R_0 = \psi\left(E(D) + \frac{Var(D) - E(D)}{E(D)}\right).$$

Note that it is possible for R_0 to be large due to a large variance and even though the mean of the degree distribution is rather small. Compare this to the final remark in Section 4.3.

Example. We will calculate R_0 as a function of β instead of ψ in the following case. Let the infectious period be exponentially distributed with parameter λ. As we have assumed so far, the infectious contacts are made according to a Poisson process with per pair intensity β. Then,

$$\begin{aligned} R_0 &= \nu\left(1 - \lambda\int_0^\infty e^{-\beta t}e^{-\lambda t}dt\right) \\ &= \frac{\beta\nu}{\beta+\lambda}. \end{aligned}$$

Compare this result to the Malthusian parameter in (4). Note that $R_0 < 1$ if and only if $\alpha < 0$. □

5.3 Probability of a large outbreak

Recall that β is per pair intensity of the Poisson contact process and I is random infectious period, with density f_I. The generating function of the number of neighbors infected by one fixed infective is

$$\hat{G}(z) = \sum_{j=0}^{\infty}\left(\int_0^\infty \sum_{k=j}^{\infty} r_k \binom{k}{j}(1-e^{-\beta t})^j e^{-\beta t(k-j)} f_I(t)dt\right) z^j, \tag{11}$$

where (r_j) is either sequence (p_j) or $(\tilde{p}_j)$. Newman in [22] does not take into account the assumption of random infectious period, since he treats all per contact infections as independent. However, infections of a fixed infective person are in this case dependent via the infectious period: being known that he or she has already infected someone, the infective is more likely to infect the next one of the neighbors, as his or her infectious period is likely to be longer [16]. We show in the same manner as Durrett did in [9] that Newman's large outbreak probability is larger than the correct outbreak probability. In (11) the first sum and the integral change places (Lebesgue's monotone convergence

theorem, also known as Levi's theorem) and we obtain

$$\begin{aligned}
\hat{G}(z) &= \int_0^\infty \left(\sum_{k=0}^{\infty} r_k \sum_{j=0}^{k} \binom{k}{j} (1-e^{-\beta t})^j e^{-\beta t(k-j)} z^j \right) f_I(t) dt \\
&= \int_0^\infty \left(\sum_{k=0}^{\infty} r_k (z(1-e^{-\beta t}) + e^{-\beta t})^k f_I(t) \right) dt \\
&= EG(z(1-e^{-\beta I}) + e^{-\beta I}) \\
&\geq G(zE(1-e^{-\beta I}) + Ee^{-\beta I}) \\
&= G(z\psi + (1-\psi)) = \tilde{G}(z),
\end{aligned}$$

where the inequality holds with the assumption that $p_0 + p_1 < 1$ (because then G is convex), according to Jensen's inequality. Recall that the value ψ, transmissibility T in Newman's notation, is the overall probability of transmission between two fixed individuals,

$$1 - \psi = \int_0^\infty e^{-\beta t} f_I(t) dt.$$

Thus,

$$\hat{G}_0(z) > \tilde{G}_0(z) \quad \text{and} \quad \hat{G}_1(z) > \tilde{G}_1(z),$$

where $\hat{G}_0(z)$ and $\hat{G}_1(z)$ correspond to the expression in (11), with the sequence (r_j) equal to (p_j) and $(\tilde{p}_j)$, respectively. Recall that if ξ is the smallest fixed point of G_1 in $[0, 1]$, it corresponds to the extinction probability of homogenous process with offspring distribution $(\tilde{p}_i)$. As all the branching processes starting form a first generation infective have to go extinct, we infer that extinction probability of the two phase process is $\sum_{k=0}^{\infty} p_k \xi^k = G_0(\xi)$. As we now have $\hat{G}_0(\hat{\xi}) > \tilde{G}_0(\tilde{\xi})$, Newman's probability of a large outbreak, equal to $1 - \tilde{G}_0(\tilde{\xi})$, is greater than the correct outbreak probability.

5.4 Epidemic phase transition

In this section we first provide the implicit form of the generating function for the size of branching approximations of the small epidemic outbreaks. Then we calculate the mean of those approximations of outbreak sizes explicitly and infer the epidemic phase transition value for the probability ψ. We will see the connection with R_0. The calculations are derived using the correct generating functions. Nevertheless, the mean and the transition value are in line with Newman's results.

Let $H_0(x; \beta)$ denote the generating function of the outbreak size in the two-phased process and $H_1(x; \beta)$ the generating function of the outbreak size in the homogenous branching process where the underlying graph has degree distribution $\tilde{p}_k$.

Recall that the generating function of a sum of independent discrete random variables is the product of the generating functions of the summands (see Section 3.1.1). If there are k offspring in the first generation, then the size of the outbreak is the sum $Y = 1 + Y_1 + Y_2 + ... + Y_k$, where Y_i's represent the total offspring of the ancestor's child i and Y represents the total progeny. The variables Y_i are independent and identically distributed [9]. If the process is homogenous, Y_i's have the same distribution as the outbreak size Y. For a homogenous process, the following holds

$$H_1(x;\beta) = x\sum_{j=0}^{\infty}\left(\int_0^{\infty}\sum_{k=j}^{\infty}\tilde{p}_k\binom{k}{j}(1-e^{-\beta t})^j e^{-\beta t(k-j)} f_I(t)dt\right) H_1(x;\beta)^j,$$

where $\int_0^{\infty}\sum_{k=j}^{\infty}\tilde{p}_k\binom{k}{j}(1-e^{-\beta t})^j e^{-\beta t(k-j)} f_I(t)dt$ is the probability of j infections in the first generation. If there are j infections, the number of neighbors k must be $k \geq j$. We multiply by x, as that is the generating function of a random variable equal to 1 with probability 1.

Therefore, the generating functions H_0 and H_1 are the solution of the following equations.

$$\begin{aligned} H_1(x,\beta) &= x\hat{G}_1(H_1(x,\beta);\beta), \\ H_0(x,\beta) &= x\hat{G}_0(H_1(x,\beta);\beta). \end{aligned}$$

where the generating functions $\hat{G}_0$ and $\hat{G}_1$ correspond to (11), with (r_j) equal to (p_j) and $(\tilde{p}_j)$, respectively.

Now we can calculate the mean of the approximated small outbreak sizes, as the derivative in 1 of the appropriate generating function, $H_0^{'}(1;\beta)$. Differentiating the equations above, we obtain

$$\begin{aligned} H_0^{'}(1;\beta) =& 1 + \hat{G}_0^{'}(1;\beta)H_1^{'}(1;\beta), \\ H_1^{'}(1;\beta) =& 1 + \hat{G}_1^{'}(1;\beta)H_1^{'}(1;\beta), \end{aligned}$$

since $H_1(1;\beta) = 1$ and $\hat{G}_0(1;\beta) = 1$. Therefore,

$$\begin{aligned} H_0^{'}(1;\beta) =& 1 + \hat{G}_0^{'}(1;\beta)H_1^{'}(1;\beta), \\ H_1^{'}(1;\beta) =& \frac{1}{1-\hat{G}_1^{'}(1;\beta)}. \end{aligned}$$

The last two equations imply that the mean size we are looking for equals

$$H_0^{'}(1;\beta) = 1 + \frac{\hat{G}_0^{'}(1;\beta)}{1-\hat{G}_1^{'}(1;\beta)}. \tag{12}$$

We differentiate the generating functions (11) after applying the monotone convergence theorem and obtain

$$\hat{G}_0'(z;\beta) = \int_0^\infty \left(\sum_{k=0}^{\infty} k p_k (z(1-e^{-\beta t}) + e^{-\beta t})^{k-1} (1-e^{-\beta t}) f_I(t) \right) dt$$

and similarly for $\hat{G}_1'$, which then yields

$$\begin{aligned}
\hat{G}_0'(1;\beta) =& \int_0^\infty \left(\sum_{k=0}^{\infty} k p_k (1-e^{-\beta t}) f_I(t) \right) dt \\
=& \sum_{k=0}^{\infty} k p_k \left(\int_0^\infty (1-e^{-\beta t}) f_I(t) dt \right) \\
=& \mu\psi.
\end{aligned}$$

Similarly we would derive

$$\hat{G}_1'(1;\beta) = \nu\psi.$$

Now the equation (12) implies that the mean size of the small outbreak branching approximations equals

$$H_0'(1;\beta) = 1 + \frac{\mu\psi}{1-\nu\psi}. \tag{13}$$

This yields the same phase transition as that suggested by R_0 (see Section 5.2). The critical value for per-pair transmission probability is $\psi = \frac{1}{\nu}$. If ψ is greater than the critical value, a large outbreak will occur with positive probability equal to $1 - \hat{G}(\hat{\xi})$, as calculated in Section 5.3. If ψ is smaller than $\frac{1}{\nu}$, only the small outbreaks are possible, and their (approximate) mean size equals to the expression in (13).

6 End of the epidemic

First we construct the *epidemic generated graph* as in [23] or [6]. In the random graph of contacts G we replace all the undirected edges by two directed ones, pointing at opposite directions. Every vertex is assigned a value x_i from the distribution of I, a realized value of its (possible) infectious period, in case the vertex gets infected. Since infectious contacts are made according to the Poisson process with per pair intensity β, to get the epidemic graph, we thin G by deleting independently the edge emanating from vertex i with probability $e^{-\beta x_i}$. If the edge from i to j is deleted, that means that if i gets infected, we know it will not pass the infection on to vertex j. The vertices of the directed graph that we can reach starting from the initially infectious correspond to the ultimately recovered ones.

The *susceptibility set* of vertex v is the set of all vertices such that if they were initially infected then v would be ultimately recovered [4]. This set can be approximated by the *backward branching process*, which we construct using the epidemic generated graph in the following way. The first generation of the process comprises the vertices w such that there is a directed edge in the epidemic graph, leading from w to v. Repeating this for the first generation vertices instead of v, we obtain the second generation of the backward process, and so on.

We pick a starting vertex v at random, for the forward branching process we approximate the epidemic with. Also at random, pick a vertex u different form v, whose susceptibility set will be of interest. If the forward process survives, the epidemic reaches size $\theta(n)$ and the susceptibility set of u will merge with this large component with high probability if it reaches the size of order $\theta(\log n)$ (with probability that approaches 1 as $n \to \infty$). If its susceptibility set is smaller, the probability that it will merge with the forward process goes to 0. In other words, u will almost surely be ultimately recovered in the given large outbreak only if its susceptibility set reaches the size of order $\theta(\log n)$. Namely, the following theorem holds.

Let $\Re^{(n)}$ be the number of ultimately recovered individuals and let ρ and ρ^b represent survival probabilities of the forward and backward branching processes, respectively.

Theorem 6.1. *For every $0 < \epsilon < \rho^b$ the following holds*

$$\lim_{n\to\infty} P\left(\left|\frac{|\Re^{(n)}|}{n} - \rho^b\right| < \epsilon\right) = \rho. \quad \square$$

A proof for random intersection graphs (a related case) can be found in [6]. It would be derived in the context of this thesis in a similar manner, but rigorous derivations are beyond the scope of this work.

As the theorem suggests, in the event of a large outbreak, the fraction of the infected

population can be approximated by the probability of survival of the backward branching process and the fraction of those who escape the epidemic by the probability of extinction of the backward branching process. In the following section, we explore the susceptibility set of a vertex in order to obtain the probability that the vertex escapes the epidemic.

6.1 Degree distribution of the ultimately susceptible

Note that, in a backward branching process even when infectious periods are random, the probability of a vertex being ultimately recovered or susceptible does not depend on its infectious period. We use this to derive the probability of a vertex being ultimately susceptible and of degree k (as in [2]), which then yields the degree distribution of the ultimately susceptible.

Assume the epidemic takes off (which occurs with the same probability as the survival of the approximating forward branching process). Let a be the fixed number of the initially infective individuals (not growing with n). As we take n to be large, we have that $a/n \approx 0$, so a vertex is initially susceptible with probability approximately equal to 1. Then the probability that an arbitrary vertex v escapes the epidemic is

$$\xi = \sum_{k=0}^{\infty} \xi_k p_k,$$

where ξ_k is probability that a vertex of degree k does not acquire the infection until the end of the epidemic. I is the random infectious period with density f_I and per pair transmission probability ψ (defined in (10)). Then, for a fixed susceptible - infected pair, the probability of escaping infection is $1 - \psi$. In other words, if the starting vertex gets infected, it will pass the infection to its susceptible neighbor with probability ψ. We denote by w a neighbor of the initially fixed vertex v. Let s be the probability that w escapes the epidemic (s will be determined later). We have

$$\xi_k = \sum_{l=0}^{k} (1-\psi)^l \binom{k}{l} s^{k-l}(1-s)^l = (1 - \psi + \psi s)^k \,. \tag{14}$$

It follows

$$\xi = G_0 \left(1 - \psi + \psi s\right), \tag{15}$$

where G_0 is the generating function of degree distribution (p_i).
Similarly, we have

$$s = \sum_{k=0}^{\infty} \tilde{\xi}_k \tilde{p}_k,$$

where $\tilde{p}_k = kp_k / \left(\sum_{k'=0}^{\infty} k' p_{k'}\right)$ and

$$\tilde{\xi}_k = \sum_{l=0}^{k-1} (1-\psi)^l \binom{k-1}{l} s^{k-1-l}(1-s)^l = (1-\psi+\psi s)^{k-1}\,.$$

We consider only $k-1$ of the k neighbors of w because we are exploring the second generation of the susceptibility set of v, and the possibility that v infects w is not taken into account. The last three equations yield

$$s = \frac{G_0'\left(1-\psi+\psi s\right)}{G_0'(1)}. \tag{16}$$

Remark. Note that, if per pair transmission probability is $\psi = 1$, then $\tilde{\xi}_k = s^{k-1}$ and, due to the connection between G_0 and G_1 given in (6),

$$s = \frac{G_0'(s)}{G_0'(1)} = G_1(s). \tag{17}$$

Therefore, if $\psi = 1$, then s equals the extinction probability of the homogenous branching process with offspring distribution $\tilde{p}_k = kp_k / \left(\sum_{k'=0}^{\infty} k' p_{k'}\right)$.

With s implicitly given as the smallest solution of (16), from (15) we get the probability of escaping the epidemic. Also, from (14), we get the probability of escape for a vertex of degree k. This yields the probability of degree k, given that the individual is ultimately susceptible, which equals

$$u_k = \frac{\xi_k p_k}{\xi} = \frac{(1-\psi+\psi s)^k}{G_0(1-\psi+\psi s)} \cdot p_k. \tag{18}$$

Here ξ is the normalizing constant. The degree distribution of the ultimately susceptible is thus generated by

$$U(z) = \sum_{k=1}^{\infty} u_k z^k = \frac{G_0(z(1-\psi+\psi s))}{G_0(1-\psi+\psi s)}.$$

Lemma 6.2. *All moments of the degree distribution of ultimately susceptible vertices (u_k) are finite, regardless of the moments of the distribution (p_k).*

Proof. To prove this, we use tests for series convergence. Namely, the l.th moment of the distribution (u_k) is

$$\begin{aligned}\sum_{k=1}^{\infty} k^l \cdot u_k &= \sum_{k=1}^{\infty} k^l \cdot \frac{(1-\psi+\psi s)^k p_k}{G_0(1-\psi+\psi s)} \\ &= \frac{1}{G_0(1-\psi+\psi s)} \sum_{k=1}^{\infty} k^l \cdot (1-\psi+\psi s)^k p_k.\end{aligned}$$

It holds that

$$\lim_{k\to\infty} \frac{(k+1)^l(1-\psi+\psi s)^{k+1}}{k^l(1-\psi+\psi s)^k} = 1-\psi+\psi s$$
$$= 1-\psi(1-s) < 1,$$

for $s < 1$ and according to the ratio test (see Theorem 2.1), the sum $\sum_{k=1}^{\infty} k^l \cdot (1-\psi+\psi s)^k$ converges. Since $k^l \cdot (1-\psi+\psi s)^k p_k < k^l \cdot (1-\psi+\psi s)^k$, we have

$$\sum_{k=1}^{\infty} k^l \cdot u_k = \frac{1}{G_0(1-\psi+\psi s)} \sum_{k=1}^{\infty} k^l \cdot (1-\psi+\psi s)^k p_k < \infty.$$

If $s = 1$, then (15) implies that the probability of an arbitrary vertex escaping the epidemic equals $\xi = G_0(1) = 1$. However, this case is not of interest to us, as then a large outbreak cannot occur. □

Note that, in order to derive this, we have not used any information about the moments of (p_k). So, the mean of the second generation vertex degrees $\nu = \sum_{k=1}^{\infty} k(k-1)p_k/\mu$ can be infinite. In that case, $R_0 = \nu\psi = \infty$.

6.2 Fraction of ultimately susceptible neighbors of the ultimately susceptible

Let v be an arbitrary vertex of degree k and w one of its neighbors. The fraction of neighbors of the ultimately susceptible which are also ultimately susceptible is calculated as the following conditional probability.

$$\begin{aligned} &P(w \text{ is ultimately susceptible} \mid v \text{ is ultimately susceptible}) \\ =&\frac{P(v \text{ and } w \text{ are ultimately susceptible})}{P(v \text{ is ultimately susceptible})} \\ =&\frac{s \cdot \tilde{\xi}_k}{\xi_k} = \frac{s}{1-\psi+\psi s}. \end{aligned} \tag{19}$$

Note that s is the probability that the initially susceptible neighbor w escapes the infection from all its neighboring vertices, apart from perhaps v. ξ_k is the conditional probability that vertex v does not acquire the infection until the end of the epidemic, given that its degree is k. Therefore ξ_k also equals the probability that a k-degree vertex is ultimately susceptible, since a vertex is initially susceptible with probability 1. In particular, this is the probability that v is ultimately susceptible. $\tilde{\xi}_k$ is the probability that a k-degree vertex does not acquire the infection until the end of the epidemic, from a fixed group of size $(k-1)$ of its neighboring vertices (in other words, all but one of its neighbors).

In (19), in order to calculate the probability for the neighbors v and w to both be ultimately susceptible, we consider them as a pair and calculate the probability that

both are initially susceptible (this probability equals 1) and that both escape the infection from the rest of their neighbors. Note that the result does not depend on the degree k of the vertex v.

6.3 Time until the end of the epidemic

In the final stages of the epidemic, when there are $o(n)$ ultimately recovered vertices yet to be infected, that is, the difference between the number already infected and ultimately removed is $o(n)$, the degree distribution of the vertices that are still susceptible must be the same as the degree distribution of the ultimately susceptible. This is because from that point of time until the end of the epidemic the proportion of susceptible k-degree vertices remains unchanged, for every $k = 1, 2, 3...$

We pick an arbitrary vertex v in the final stages of the epidemic when there are $o(n)$ vertices yet to be infected and approximate the (final) part of the epidemic which originates from v by a branching process. Now we can calculate the exact R_0^{end} of such process and infer that it is subcritical.

Theorem 6.3. *R_0^{end}, the offspring mean of the final stages branching process approximation, starting from an ultimately recovered vertex v defined above is given by*

$$R_0^{end} = \frac{d}{dx} \sum_{k=0}^{\infty} \tilde{p}_k (1 - \psi + \psi x)^{k-1}|_{x=s},$$

where $(\tilde{p}_k)$ is the second generation vertex degree distribution of the vertices in the beginning of the outbreak. Also,

$$R_0^{end} < 1.$$

Proof. Recall that the probability of a neighbor of an ultimately susceptible being ultimately susceptible as well equals $\frac{s}{1-\psi+\psi s}$. Similarly, the probability that a neighbor of v is not yet infected in the generation when v acquires the disease is also equal to $\frac{s}{1-\psi+\psi s}$. Denote this neighbor of v by w. If w is of degree k, then it is k times more probable to be chosen as v's neighbor than a vertex of degree 1 is. Therefore, in the new branching process approximation, second degree vertices have the size-biased distribution given with

$$\tilde{u}_k = \frac{k \cdot u_k}{\sum l u_l}.$$

The neighbor w will get infected via the edge which connects it to v with probability ψ.

Therefore, the R_0^{end} is given by

$$\begin{aligned}
R_0^{end} &= \psi \sum_{k=0}^{\infty} \tilde{u}_k (k-1) \frac{s}{1-\psi+\psi s} \\
&= \psi \sum_{k=0}^{\infty} \frac{k p_k (1-\psi+\psi s)^k}{\sum_{l=0}^{\infty} l p_l (1-\psi+\psi s)^l} (k-1) \frac{s}{1-\psi+\psi s} \\
&= \psi \frac{\sum_{k=0}^{\infty} (k-1) k p_k (1-\psi+\psi s)^{k-2} s}{\sum_{l=0}^{\infty} l p_l (1-\psi+\psi s)^{l-1}} \\
&= \psi \frac{\sum_{k=0}^{\infty} \frac{(k-1) k p_k}{\mu} (1-\psi+\psi s)^{k-2} s}{\sum_{l=0}^{\infty} \frac{l p_l}{\mu} (1-\psi+\psi s)^{l-1}} \\
&= \psi \frac{\sum_{k=0}^{\infty} \frac{(k-1) k p_k}{\mu} (1-\psi+\psi s)^{k-2} s}{\sum_{l=0}^{\infty} \tilde{p}_l (1-\psi+\psi s)^{l-1}}
\end{aligned}$$

Recall from Section 6.1 that $s = \sum_{l=0}^{\infty} \tilde{p}_k \tilde{\xi}_k = \sum_{l=0}^{\infty} \tilde{p}_k (1-\psi+\psi s)^{k-1}$, which is the denominator of the above expression. As before, $\tilde{p}_k$ is the size biased distribution of the second and subsequent degree vertices at the beginning of the epidemic, $\tilde{p}_k = \frac{k p_k}{\sum l p_l}$, and μ denotes the mean of the starting degree distribution, $\mu = \sum k p_k$. Therefore,

$$\begin{aligned}
R_0^{end} &= \sum_{k=0}^{\infty} \frac{(k-1) k p_k}{\mu} \psi (1-\psi+\psi s)^{k-2} \\
&= \sum_{k=0}^{\infty} (k-1) \tilde{p}_k \psi (1-\psi+\psi s)^{k-2} \\
&= \frac{d}{dx} \sum_{k=0}^{\infty} \tilde{p}_k (1-\psi+\psi x)^{k-1} |_{x=s}
\end{aligned}$$

To see that R_0^{end} must be smaller than one, note that

$$\begin{aligned}
&\sum_{k=0}^{\infty} \tilde{p}_k (1-\psi+\psi x)^{k-1} \\
&= \sum_{k=0}^{\infty} \frac{k p_k (1-\psi+\psi x)^{k-1}}{\mu} \\
&= \frac{G_0'(1-\psi+\psi x)}{G_0'(1)},
\end{aligned}$$

where by $G_0'(z)$ we denote the derivative in z. Since all the derivatives of a generating function are positive, then $\frac{G_0'(1-\psi+\psi x)}{G_0'(1)}$ is convex. Note that s, as calculated in (16) equals

$$s = \frac{G_0'(1-\psi+\psi s)}{G_0'(1)}.$$

This means that $\frac{G_0'(1-\psi+\psi x)}{G_0'(1)}$ intersects $y = x$ in the point $x = s$, which is less than 1. Thus the derivative $\frac{d}{dx}\sum_{k=0}^{\infty}\tilde{p}_k(1-\psi+\psi x)^{k-1}$ is smaller than the derivative of $y = x$, which equals 1, for all $x \in [0, s]$. Therefore, $R_0^{end} = \frac{d}{dx}\sum_{k=0}^{\infty}\tilde{p}_k(1-\psi+\psi x)^{k-1}|_{x=s} < 1$ (see Figure 3).

Recall from (17) and the corresponding remark that, in the case $\psi = 1$, s equals to the extinction probability ξ of the homogenous branching process with offspring mean $\tilde{p}_k$. Then, from (6) we obtain

$$R_0^{end} = \frac{d}{dx}\left(\frac{G_0'(x)}{G_0'(0)}\right)|_{x=\xi} = G_1'(\xi).$$

We are interested only in the case $\xi < 1$, when a large outbreak is possible. Since the generating function G_1 is convex (all its derivatives are positive) and ξ is the fixed point of G_1 in $[0, 1)$, in the same manner as in the previous paragraph, we infer that

$$R_0^{end} = G_1'(\xi) < 1.$$

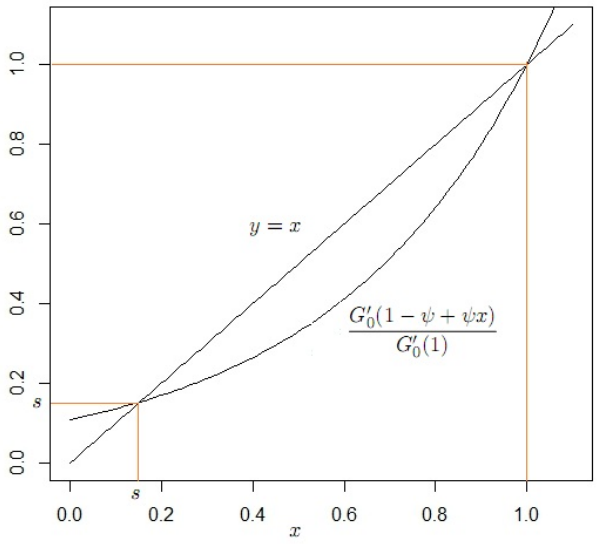

Figure 3: The generating function $\frac{G_0'(1-\psi+\psi x)}{G_0'(1)}$, with a fixed point s smaller than 1.

□

The martingale convergence theorem from Section 3.1.3 has an important corollary which is used in the analysis of the first stages of an epidemic, until the branching approximation breaks down. Namely, if the size of a generation Z_i is of order $O(f(n))$ as $n \to \infty$, then the total number of individuals who have lived until the i.th generation is also of order $O(f(n))$. More precisely,

$$\frac{Z_i}{\sum_{k=0}^{i} Z_k} \overset{a.s.}{\to} c,$$

where c is a positive constant. The following lemma says that a similar result holds for the end of the epidemic.

First, we explain why this is of interest. Let $\epsilon > 0$ be an arbitrary number close to zero. It has been proven in [7] that the generations at the peak of the epidemic consist of $O(n)$ individuals. Given this, from Figure 1 it is intuitively clear that the size of a generation will be of order $O(n^{1-\epsilon})$ in two stages of the epidemic: once while the number of infected individuals is increasing and once while it is decreasing. We consider the second time this happens. Since the underlying graph is sparse, the number of half-edges to connect to the currently infected individuals is also of order $O(n^{1-\epsilon})$. In other words, the number of the susceptibles to infect in the following generation is $O(n^{1-\epsilon})$. Similarly to the corollary discussed in the previous paragraph, the number of new infections until the end is of order $O(n^{1-\epsilon})$. Therefore, the overall number of the individuals infected from that point on is of order $o(n)$, which is the condition of Theorem 6.3. In the following lemma we outline a proof of this result.

Consider a time when there are $O(n^{1-\epsilon})$ infected in a generation in the declining part of the epidemic curve. Let the infimum of such time be denoted by t^*. Let v be an individual who is susceptible at time t^* and who acquires the infection in the following generation.

Lemma 6.4. *Assume we have a large outbreak. The overall number of the vertices infected after t^* is of order $o(n)$.*

Idea of proof. Since the population is closed, the number of susceptibles decreases over time. Therefore, the number of infected individuals after the peak must have a declining mean, until the end of the epidemic. This implies that the average number of further infections that v passes on must be smaller than 1, since otherwise the number of the infected could not decline. Thus, if we were to approximate the infections that originate from v by a branching process, the process would have to die out because its effective R_0 is smaller than 1. This means that from this point on we have $O(n^{1-\epsilon})$ small outbreaks, which are all of size $O(1)$. Therefore, the overall number of the individuals infected after t^* is bounded above by $O(n^{1-\epsilon})$, and thus it is $o(n)$. (It would be precisely of order $O(n^{1-\epsilon})$ if the processes were independent. This issue will be discussed later in the text.) □

Theorem 6.5. *The branching process (approximating the final part of the epidemic) that originates from v is subcritical.*

Proof. According to Lemma 6.4, at time t^* there are $o(n)$ individuals more to be infected. The theorem now follows from Theorem 6.3. □

6.3.1 Discrete time

Recall the definition of the embedded Galton Watson process from Section 3.2. To get an intuition about how much time it takes for the epidemic to die out, counting from the point t^*, we first consider the discrete time (the number of generations until the end). It holds that

$$E(Z_i) = E(Z_0)(\nu_*)^i,$$

where Z_0 denotes the number of the infected at time t^* and it is of order $O(n^{1-\epsilon})$. Then, ν_* is the mean of the branching processes approximating the epidemic at these stages. It follows from Theorem 6.5 that ν_* must be smaller than 1. Z_i denotes the size of the i.th generation (counting from t^*). The formula above implies that, if i is an integer larger than $\frac{\log n}{|\log \nu_*|}$, the expected i.th generation size is of order $O\left(\frac{1}{n^\epsilon}\right)$, which goes to 0 in the limit of large n.

However, the $O(n^{1-\epsilon})$ processes are not independent, as they may use the same susceptibles. In order to circumvent the problem, we consider three different processes. In each, we use only the half-edges that are attached to a still susceptible vertex at time t^*. The underlying graphs have the same degree distribution as the ultimately susceptible individuals.

First is the process approximating the epidemic, where we erase a half-edge if it connects to a vertex that has already been infected. If a half-edge is erased, the vertex at its end cannot have any offspring. In this process the above mentioned dependencies exist, and we will approximate it by another process. We denote it by Z_ϵ.

Second is a collection of $O(n^{1-\epsilon})$ independent branching processes with the appropriate reproduction mean. This process is denoted by Z.

Let $q \in (0,1)$ be an arbitrary probability. The third process is a collection of $O(n^{1-\epsilon})$ processes approximating the epidemic, where each half-edge is erased independently with probability q. In the case of erasing a half-edge, the vertex at its end cannot have any offspring. We call this process Z_q.

The epidemic approximation process, Z_ϵ, is stochastically smaller than the independent processes union, Z. The erased model, Z_q, is stochastically smaller than the epidemic approximation process, Z_ϵ, until the overall number of infected vertices reaches the value $q \cdot n_*$, where n_* is the initial number of susceptible vertices, counting time from t^*. When

the number of the overall infected (that is, infectious or recovered) reaches $q \cdot n_*$, the probability of erasing a half-edge in the epidemic approximation process Z_ϵ equals q and it further increases as we continue with connecting the half-edges. This is when the coupling of the epidemic process Z_ϵ and the erased process Z_q breaks down. On the other hand, if q is chosen so that $q \to 0$ as $n \to \infty$, then the process Z_q and the independent process Z can be coupled so that their difference is small and the survival probability and the final size are the same. Since the epidemic process Z_ϵ is stochastically in between those two, it can be coupled with either the erased or the independent process so that their difference is very small, in the limit of large n.

The above is a rough explanation and the details and proofs can be found in [5]. See also [6].

Now we approximate the epidemic branching process by the independent process and according to the previous formula, as indicated earlier, we infer that it will take at least $\frac{\log n}{|\log \nu_*|}$ generations until the epidemic dies out completely, in the sense that the expected value of a generation size after more than $\frac{\log n}{|\log \nu_*|}$ generations equals $O\left(\frac{1}{n^\epsilon}\right)$.

This is in line with the result proven in Lemma 6.2 and the small world results listed in Section 4.3. Namely, even in the case of almost sure per-pair transmission, $\psi = 1$, the moments of the degree distribution in the process that originates from v are all finite. Consequently, the typical distance and the diameter of the corresponding graph are of order $O(\log n)$. Accordingly, we have obtained that it takes $\frac{\log n}{|\log \nu_*|}$ generations until the epidemic ceases.

6.3.2 Continuous time

Time until the end cannot be replaced by the number of generations until the end, as we may have individuals of different generations that are infective at the same time point. Thus, instead of the Galton Watson processes we consider the general branching processes and use Theorem 3.6. For the epidemic to stop it is necessary and sufficient that all of the $O(n^{1-\epsilon})$ independent processes die out.

In the following text, we label the number of independent processes by $N(= O(n^{1-\epsilon}))$. We denote the number of the infected at time t in the (individual) independent processes by $z_t^{(i)}$, $i = 1, 2, ...N$. Let z_t denote the number of the infected in their union, at time t. (The notation is corresponding to that of Theorem 3.6.) The following theorem is the main result of this work.

Theorem 6.6. *If the contact process and the infectious period distribution are such that the conditions of Theorem 3.6 hold, then the random time T that passes from t^* until the end of the epidemic has a distribution with an exponentially declining tail.*

Proof. The probability that all the independent subcritical branching processes die out

can be approximated in the following way.

$$P(z_t = 0) = P(z_t^{(i)} = 0 | i = 1, ..., N) \sim (1 - c \cdot e^{\alpha t})^N,$$

as $t \to \infty$. In the above expression, α is the Malthusian parameter (see Section 3.2). Further, we have

$$\begin{aligned} P(z_t = 0) &\approx 1 - Nce^{\alpha t} + o(e^{\alpha t}) \\ &\approx 1 - Nce^{\alpha t} \\ &= 1 - O(n^{1-\epsilon})e^{\alpha t}, \end{aligned}$$

if we let $t \to \infty$ and then $n \to \infty$. That is,

$$P(z_t > 0) \approx O(n^{1-\epsilon})e^{\alpha t}, t \to \infty, n \to \infty.$$

If the time of extinction of the process Z is denoted by T, then

$$P(T > t) = O(n^{1-\epsilon})e^{\alpha t}, t \to \infty, n \to \infty$$

and the approximate cumulative distribution function of T is given by

$$P(T \leq t) = 1 - O(n^{1-\epsilon})e^{\alpha t}, t \to \infty, n \to \infty.$$

□

Note that, if $t \geq -\frac{1}{\alpha} \log n$, the extinction probability equals

$$P(z_t = 0) = 1 - O\left(\frac{1}{n^\epsilon}\right) \to 1, n \to \infty, \tag{20}$$

which confirms the discrete time result. (Note that then also $t \to \infty$.)

Remark. We could have as well started exploring the ending part from the point when there are $O(\epsilon n)$ infected vertices in the declining part of the epidemic curve and the inference would be similar. This way we would have concatenated to the work done in [7]. Instead, we start from the point when $O(n^{1-\epsilon})$ individuals are infected for the second time, which happens a little later. This is because it is of interest to see that the time it takes for an epidemic to be completely eradicated is longer than one might anticipate. Starting from a later point of time thus strengthens the result.

7 Concluding remarks

In this thesis we have explored the end of an SIR epidemic on a configuration model, when there exists a Malthusian parameter for the infectious contact process and the given life time distribution. The main result is Theorem 6.6. Namely, the time that passes until the epidemic dies out is at least $t^* - \frac{1}{\alpha}\log n$. (Recall that t^* is the infimum of time when there are $O(n^{1-\epsilon})$ infected individuals and $\alpha < 0$ is the Malthusian parameter.)

It would be of interest to examine more closely when the Malthusian parameter exits, that is, the existence of a solution of the singular integral equation (2). For example, we have seen in Section 3.2.1 that, in the case of a constant infectious rate during the infectious period, the solution does not exist if the life time density declines slowly (the decline rate is slower than exponential). However, that means that a single individual is likely to be infected for a long time. In that case, one would expect the epidemic to persist for a long time. Here we only emphasize that the diseases with such properties, for example, are not well fitted by our model.

Further, we remark that the branching process approximation (Section 5.1) was only justified in the case of finite variance. Also, the presence of a small number of self-loops and multiple edges (which is why we allow for both in the model) was only proved under the assumption of finite variance of the degree distribution (Theorem 4.1).

The final remark is regarding the configuration model as a model for real world networks. Apart from the number of self-loops and multiple edges being sparse, there are also not many triangles (cycles of length 3) [20]. However, the real world networks contain many triangles - as two friends of a person are likely to be friends as well. Therefore, it would be interesting to address the question of time in a context of a more realistic model of a contact network.

References

[1] Adnadjević D., Kadelburg Z. (2004) Matematička analiza 1, Matematički fakultet, Beograd.

[2] Andersson H. (1999) Epidemic models and social networks, The Mathematical Scientist 24 128–147.

[3] Athreya K.B., Ney P.E. (1972) Branching Processes, Springer.

[4] Ball F., Lyne O.D. (2001) Stochastic multi-type SIR epidemics among a population partitioned into households, Adv. Apl. Prob. 33 (1), 99-123.

[5] Ball F., Sirl D., Trapman P. (2009) Threshold behaviour and final outcome of an epidemic on a random network with household structure, Adv. Apl. Prob. 41, 765-796.

[6] Ball F., Sirl D., Trapman P. (2012) Epidemics on random intersection graphs, arXiv:1011.4242.

[7] Barbour A. D., Reinert G. (2013) Approximating the epidemic curve, Electron. J. Probab. 18 (2013), no. 54, 1–30.

[8] Diestel R. (2005) Graph Theory, Springer-Verlag Heidelberg, New York.

[9] Durrett R. (2006) Random Graph Dymanics Cambridge University Press.

[10] Erdős P., Rényi A. (1959) "On Random Graphs I" in Publ. Math. Debrecen 6, 290–297.

[11] Grimmett G., Stirzaker D. (2001) Probability and Random Processes, Oxford University Press.

[12] Harris T. E. (1963) The theory of branching processes, Springer.

[13] Hofstad van der R. (2013) Random Graphs and Complex Networks, Lecture Notes, http://www.win.tue.nl/ rhofstad/NotesRGCN2013.pdf

[14] Jagers P. (1975) Branching Processes with Biological Applications, London: John Wiley Sons.

[15] Kimmel M., Axelrod D.E. (2002) Branching Processes in Biology, Springer.

[16] Meester R., Trapman P. (2011) Bounding Basic Characteristics of Spatial Epidemics with a New Percolation model, Adv. Apl. Prob. 43, 335-347.

[17] Miller J. C. (2011) A note on a paper by Erik Volz: SIR dynamics in random networks. J. Math. Biol. 62, 349–358.

[18] Molloy M., Reed B. (1995) A critical point for random graphs with a given degree sequence, Rand. Struct. Alg. 6 161-179.

[19] Molloy M., Reed B. (1998) The size of the giant component of a random graphs with a given degree sequence, Comb. Prob. Comp. 7 295-305.

[20] Newman M.E.J., Random Graphs as Models of Networks, cond-mat/0202208.

[21] Newman M. E. J., Strogatz S. H., Watts D. J. (2001) Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E, 64 026118.

[22] Newman M. E. J. (2002) Spread of epidemic disease on networks, Phys. Rev. E, 66 016128.

[23] Pellis L., Ball F., Trapman P. (2012) Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R_0, Math. Biosci. 235 85–97.

[24] Volz E. (2008) SIR dynamics in random networks with heterogeneous connectivity. J. Math. Biol. 56, 293–310.