

Epidemics on social network graphs

Maria Deijfen *

Submitted in partial fulfillment of the Master of Science degree.

Supervisor: Håkan Andersson.

January 2000

Abstract

This work is concerned with relaxing the assumption of homogeneous mixing in a Reed-Frost epidemic model. After a thorough presentation of the standard Reed-Frost model we introduce social structure in the population by aid of a social network graph and investigate the effects on the epidemic spread. A number of possible choices for the underlying social network are presented: Bernoulli graphs, 3-clique graphs, Markov graphs, Household graphs and small-world networks. For each of these structures we investigate if it can be regarded as a good model for a social network. We also derive asymptotic formulas for epidemiological quantities like the basic reproduction number and the final size of the epidemic. When exact results are beyond reach computer simulations are presented.

*Department of Mathematics, Stockholm University, SE-10691 Stockholm, Sweden.

Contents

1	Introduction	4
1.1	Sketch of the problem	4
1.2	Short history	5
1.3	Outline of the contents	6
2	The Reed-Frost model	8
2.1	Description of the model	8
2.2	Branching process approximation	9
2.3	Asymptotic behaviour of the Reed-Frost epidemic	12
2.4	Summarizing discussion	16
3	Epidemics on graphs	18
3.1	Graph theoretic notation	18
3.2	Social network graphs	19
3.3	Epidemic spread on a fixed graph	19
3.4	Properties of a social network graph	21
4	Bernoulli random graphs	23
4.1	Construction of the network	23
4.2	Epidemic behaviour	24
4.3	Properties of the graph	27
5	The 3-clique model	29
5.1	Construction of the network	29
5.2	Epidemic behaviour	30
5.3	Properties of the graph	34
6	Markov graphs	36
6.1	Description of the model	36
6.2	Properties of the model	39

6.3	Simulations	40
7	The household model	43
7.1	Description of the model	43
7.2	Epidemic behaviour	44
7.3	Properties of the model	46
8	Small-world networks	48
8.1	The small-world phenomenon	49
8.2	Construction of the network	51
8.3	Tuning the parameters	52
8.4	Epidemic behaviour	53
9	Epilogue	56

Chapter 1

Introduction

In this introductory chapter we give a quick sketch of the problem treated in the work. A very short history of epidemic modelling is followed by an outline of the contents.

1.1 Sketch of the problem

A mathematical model of the spread of an infectious disease is a description of the epidemic process made in a way that enables us to quantify important information about the process. For example we want the modelling framework to make it possible for us to analyze the conditions under which a large outbreak is possible, to calculate the probability of such an outbreak and to say something about the size of it, in case it occurs. There are two main model types: deterministic and stochastic ones. Deterministic models do not take randomness in the disease transmission into account whereas stochastic models do. Since the natural way to describe the spread of a disease is stochastic - we talk about the *probability* of disease transmission between two individuals rather than knowing for certain that a transmission will take place - a stochastic model is of course more likely to catch the features of a real-life epidemic. The price for this higher level of realism is that a stochastic model is often much more complicated to analyze than a deterministic one. We are forced to impose several simplifying assumptions to deal with the analysis and this in turn decreases the practical interest of the model. The vast majority of the stochastic models for epidemics treated in the literature incorporates three basic restrictions on the population in which the epidemic takes place:

- No births, deaths, immigration or emigration occur. The population

is *closed*.

- All individuals are of the same type; no age-dependence, no differences in susceptibility etc. The population is *homogeneous*.
- A given individual makes contacts with all other individuals with the same probability. The population is *homogeneously mixing*.

This work is concerned with relaxing the assumption of homogeneous mixing in a stochastic epidemic model, while in all other respects the simplest possible setup is chosen. The homogeneous mixing assumption is a very basic assumption which greatly simplifies the analysis, because it enables population structure to be ignored. However, in real life there is not likely to exist any population where the mixing is completely random. On the opposite; a high level of selectivity in the contact process (and thereby also in the transmission process of an infectious disease) is present. If a person catches a cold she is of course more likely to infect her best friend than someone who is a complete stranger to her. To introduce the desired social structure in the epidemic model we proceed as follows: first we generate a graph describing the relations in the population and then we let the disease spread along the social network so obtained. The problem treated here is to find a graph that captures the social formations in a human population in a satisfactory way and is yet simple enough to lend itself to mathematical analysis.

Ability to predict the outcome of a real-life epidemic would of course be very useful from a public-health point of view. This makes it an important task to develop the epidemic models of today and make them reflect the complexity of the underlying reality in a better way.

1.2 Short history

Epidemic modelling is relatively new as an area of research in the mathematical sciences. The first complete model that received attention in the literature was created by Kermack and McKendrick in 1927. This was a deterministic model where the outcome of the epidemic process depended only on the number of initially infected individuals in the population. In 1928 Reed and Frost presented the first stochastic model, the chain binomial model; a model that, assuming the disease to spread in generations, specifies the probability law of the epidemic in the next generation, given the present. This is the model used in this work. The following two decades

not much effort was put into the analysis of stochastic epidemic modelling. In 1949, though, Bartlett came up with a stochastic version of the Kermack-McKendrick model and since then the interest in the area has more or less exploded. Important work has been made by Bailey (1975), Sellke (1983), von Bahr and Martin-Löf (1980) and Scalia-Tomba (1985, 1990).

As was mentioned in the previous section, the literature concerning epidemic models has so far been dominated by models that focus on populations in which random mixing between individuals is assumed. However, in recent years various attempts have been made to take population structure into account, see e.g. Altmann (1996), Diekmann, de Jong and Metz (1998), Ball, Mollison and Scalia-Tomba (1997), Kretzschmar and Morris (1996) and Andersson (1998). These works cover ideas such as introducing two different levels of mixing (Ball *et al*), analyzing the spread of the disease as a function of the concurrency of relationships in the population (Kretzschmar and Morris) and using random graphs as models for social networks (Andersson).

1.3 Outline of the contents

An SIR epidemic model is a model where there are just three possible states for an individual; susceptible (S), infected (I) and removed (R) and the only possible transitions are $S \rightarrow I$ and $I \rightarrow R$. The formulation of the model usually includes an assumption of homogeneous mixing in the population. This work aims at investigating what happens if this assumption is dropped in a very simple SIR-model with constant infectious periods; the so called Reed-Frost model. To achieve this we generate a random graph that can serve as a model for a social network and study the epidemic spread along this graph according to a modified version of the Reed-Frost model.

In Chapter 2 the Reed-Frost model is presented together with a branching process approximation of the initial stage, valid in large populations. The epidemiological quantities that we will be concerned with are defined and derived and the asymptotic behaviour (i.e. the behaviour when the population size tends to infinity) is exploited. At the end of the section the drawbacks of the model are formulated more precisely. Chapter 3 starts with some basic notation from the graph theoretical area. We modify the Reed-Frost model from Chapter 2 to suit our purpose and describe the epidemic spread on a fixed graph. A discussion where we list some of the properties a graph describing social interactions should exhibit closes the section. The rest of the work presents a number of possible choices for the underlying social network

graph and investigates its impact on the epidemic process. As a warm-up the Bernoulli random graph is introduced in Chapter 4 and generalized to the 3-clique model in Chapter 5. Chapters 6-8 deals with Markov graphs, Household models and Small-world networks, respectively. A few new ideas and results are discussed, see in particular Chapter 5 and 8.

The mathematics is throughout kept on a fairly basic level and rigorous proofs are replaced by heuristic arguments. Some knowledge about basic concepts in probability theory is assumed, but no previous experience with epidemic modelling is required.

Chapter 2

The Reed-Frost model

In this chapter we describe the Reed-Frost model of the spread of an infectious disease. We motivate a very useful approximation of the initial stage of the epidemic, valid in large populations, and define some important epidemiological quantities. In the last section we comment on the advantages and drawbacks of the Reed-Frost model.

2.1 Description of the model

First we introduce some notation and formulate the dynamics of a Reed-Frost epidemic.

The Reed-Frost model: We consider a closed, homogeneous population consisting of N individuals. Let n denote the number of susceptible individuals at time $t = 0$ and m_n the number of infected individuals, where $n + m_n = N$ and m_n/n small. The dynamics of the epidemic runs as follows: Assume that an individual i is infected at time t . A given individual j is contacted by i with probability γ/n , $\gamma > 0$, and if j is susceptible then j becomes infected at time $t + 1$. At time $t + 1$ also, i becomes removed (by immunity or death) and plays no further part in the epidemic process. The epidemic ceases when there are no infectious individuals present in the population. All contacts are assumed to be independent of each other.

According to the definition above the Reed-Frost model assumes a constant latent period (where the time is re-scaled, making this period last for one unit of time) and a short infectious period (represented by a single point in time) during which a given infective infects a given susceptible with probability

γ/n (of course we have to assume that the population size is greater than γ in order for this expression to define a probability, but this assumption will always be met here since we are only concerned with epidemics in large populations). This probability should be thought of as a product of the contact probability and the probability of a disease transmission in case of a contact, thus giving the probability of a contact resulting in a new infective. The reason why this parameter has to be scaled by n is twofold: We want to prevent the epidemic from immediately exploding and, with probability one, striking out the entire population and we want to make the average number of secondary cases generated by a given infectious individual independent of the population size. This will hopefully be more clear in light of the branching process approximation presented later in this section.

Let $E_{n,m_n}(\lambda)$ denote the entire epidemic process as described in the definition. Later in the work we will be concerned with the *asymptotic* behaviour of the process, i.e. the behaviour as the size of the susceptible population becomes large. Hence we are going to study a *sequence* of processes $E_{n,m_n}(\lambda)$ and investigate what happens as n tends to infinity. The reason why the number of initially infectious individuals, m_n , is indexed by n is that sometimes we want this number to grow with n . In a situation where the disease is brought into the susceptible population by aid of visiting friends of the n members of the population for example we certainly expect the number of visiting friends to increase with the population size. We can also think of scenarios where the number of initial infectives stays constant as n increases, i.e. $m_n \equiv m$. Which of these cases that is at hand is important to know if we want to investigate the asymptotic behaviour of the epidemic.

Before closing this subsection we note that the Reed-Frost model is indeed an SIR-model (encountered in the Introduction). The only possible states for an individual are susceptible (S), infected (I) and removed (R) and the only possible transitions are $S \rightarrow I$ (a susceptible becomes infected) and $I \rightarrow R$ (an infected becomes removed).

2.2 Branching process approximation

Throughout this work we focus on the epidemic spread in *large* populations (for example we might consider the population consisting of all inhabitants in a country, or part of a country). This means that we assume the population size N to be a very large number and hence we can expect asymptotic results (i.e. results derived as N tends to infinity) to be valid. This makes life more easy for us, since two of the most important probabilistic results, the law of

large numbers and the central limit theorem, are of asymptotic nature. Here we describe how the initial stages of the generation process of infectives in a large population can be approximated by a so called branching process. This approximation allows us to make use of standard results from branching process theory, which will prove useful when deriving certain epidemiological quantities.

A branching process is described by the following model:

Branching process: At time $t = 0$ there exists an initial population consisting of m individuals, called ancestors. During its life span, every individual gives birth to a random number of children. During their life spans, these children give birth to a random number of children, and so on. At time $t = 1$ the population consists of the children of the ancestors (the ancestors themselves excluded), at time $t = 2$ the population consists of the children of these children and, in general, at time $t = k$ the population consists of the progeny of the members in generation $k - 1$. The branching process at time $t = k$ simply tells us the size of the population at that time. The reproduction rules are that all individuals give birth according to the same probability law independently of each other and that the number of children produced by an individual is independent of the number of individuals in its generation.

Here is the branching process interpretation of the Reed-Frost epidemic spread: The m ancestors in the branching process are our m_n initial infectives in the epidemic process and a birth in the branching process corresponds to the appearance of a new infective in the epidemic. Note that the infectious contacts made by a given infective in the epidemic process are assumed to be independent of the contacts made by all other infectives and likewise the number of children produced by a given member in the branching process is independent of the number of children produced by any other member. Hence, so far so good. But what about the distribution of the number of new infectives generated by a given infectious individual in the epidemic? To deal with this problem, consider one of the m_n initial infectives. According to the model she will, independently, infect each of the n susceptible with probability γ/n . This could be thought of as she performs n trials, each of which is a success with probability γ/n (where a success means that a transmission of the disease takes place). Hence the number of new infectives is binomially distributed with parameters $(n, \gamma/n)$, i.e. the probability that the number of new infectives generated by one of

the initially infectives equals k is given by

$$b_k = \binom{n}{k} (1 - \gamma/n)^{n-k} (\gamma/n)^k \quad k = 0, 1, \dots, n.$$

If n is large, this distribution can be approximated by a Poisson distribution. In general, a binomial distribution with parameters (n, p) , where n is large and p small, is close to a Poisson distribution with parameter np . In our case $p = \gamma/n$ is small if n is large, and $np = \gamma$. Hence, if the population size is large b_k can be replaced by p_k , where p_k is a Poisson (γ) probability, i.e.

$$p_k = \frac{\gamma^k}{k!} e^{-\gamma} \quad k = 0, 1, 2, \dots$$

One of the advantages of this approximation is that p_k , as opposed to b_k , does not involve any factorial expressions. Factorials often lead to cumbersome calculations and numerical problems and we are happy to get rid of them if we can. The approximation is justified by the following calculation:

$$b_k = \binom{n}{k} (1 - \gamma/n)^{n-k} (\gamma/n)^k \approx \frac{n^k}{k!} (1 - \gamma/n)^{n-k} \frac{\gamma^k}{n^k} \longrightarrow \frac{\gamma^k}{k!} e^{-\gamma} \quad \text{as } n \rightarrow \infty.$$

So far we have argued that in a large population the number of new cases generated by one of the initially infectives in the epidemic process is approximately Poisson distributed with expected value γ (in a Poisson distribution the expected value equals the value of the parameter). Now consider an infective in some later stage of the epidemic. The number of new infectives generated by her is no longer binomial $(n, \gamma/n)$ distributed, since some of the n initially susceptible individuals are now no longer susceptible. Either they are infected or they have been so in some earlier stage and are removed from the process by now. An attempt to infect any of these individuals is doomed to fail, i.e. the probability of a success is 0 instead of γ/n . But let us assume that we are still in the beginning of the epidemic process and that the population size is large. The fraction already infected will then be very small and we can approximate the distribution of new cases generated by each infective with a binomial $(n, \gamma/n)$ distribution, which in turn can be approximated by a Poisson (γ) distribution.

To conclude we have "proved" that the number of infectious individuals in a large population initially behaves like a branching process, where each member gives birth to a Poisson (γ) distributed number of children. Perhaps this can explain why the infection probability p_{inf} has to be scaled by n . Suppose $p_{inf} \equiv p$. This would imply that the number of new cases generated

by an infective in the beginning of the process was binomial (n, p) distributed with expected value np . But np tends to infinity with n , causing the process to explode with probability 1 in a large population. This of course yields a model of not much interest. The advantage of setting $p_{inf} = \gamma/n$ should now be clear: it makes the expected number of infectious contacts made by a given infective equal to γ , which is independent of n and thus stays bounded as n tends to infinity.

2.3 Asymptotic behaviour of the Reed-Frost epidemic

Let us now study the epidemic spread in some large population according to the Reed-Frost model. We start by defining the epidemiological quantities that we will be interested in.

- *The basic reproduction number, R_0 .* This is a critical parameter indicating whether or not a large outbreak is possible. In the Reed-Frost case it is defined as the expected number of secondary cases generated by one infectious individual in a large susceptible population. In more complicated models though, it is not always obvious how R_0 should be defined. In general we define R_0 to be any nonnegative function of the model parameters such that if $R_0 \leq 1$ then the asymptotic probability of a large outbreak is zero, while if $R_0 > 1$ then there is, asymptotically, a strictly positive probability of a large outbreak. (By a large outbreak we mean an outbreak such that the size of the outbreak is of the same order as the whole population.)
- *The final size of the epidemic, τ .* If we let Z_n denote the number of individuals ever infected in a susceptible population of size n , then $\tau = Z_n/n$, i.e. τ is the proportion of the initially susceptible population that has ever experienced the disease when the epidemic ceases. Note that the m_n initially infectives are traditionally not included in Z_n .

Before attacking the Reed-Frost model we state a classical result from branching process theory that will prove useful to us.

Theorem 2.3.1 *Consider a discrete-time branching process with m ancestors and the average number of children produced by a given member equal to γ (the number of children produced by a given member is not allowed to be constant). Let Z =total progeny in the process, $Z' = Z + m$. The following holds:*

(i) $\gamma \leq 1 \Rightarrow P(Z' < \infty) = 1$

(ii) $\gamma > 1 \Rightarrow P(Z' < \infty) = q^m$ where $q < 1$ satisfies a certain equation.

In words the theorem tells us that the process will exhibit two different behaviours depending on whether γ is below or above one. If $\gamma \leq 1$ the total progeny is finite almost surely, whereas if $\gamma > 1$ there is a strictly positive probability $1 - q^m$ that the total progeny is infinite. For a proof of the theorem see e.g. Gut (1991).

Basic reproduction number

In light of the branching process approximation from Section 2.2 the above result immediately gives us R_0 in the Reed-Frost model. If we let $E_{n,m_n}(\gamma)$ denote the standard Reed-Frost epidemic process, then, if n is large, the initial stage in $E_{n,m_n}(\gamma)$ can be approximated by a branching process, initiated by m_n ancestors, where the average number of children produced by each member equals γ . Applying the above theorem to this process yields that a large outbreak is possible only if $\gamma > 1$. To see this we note that if $\gamma \leq 1$ the total progeny in the approximating process is finite with probability one, i.e. the process goes extinct with probability one, making a large outbreak in the epidemic process impossible. Indeed, for a large outbreak to be possible we require some kind of expanding behaviour in the beginning of the time course and it should be intuitively clear that this can be obtained only if the average number of new cases generated by each infective in the initial stage of the epidemic is strictly greater than one, i.e. if $\gamma > 1$. To conclude $R_0 = \gamma$ in the Reed-Frost model.

Final size equation

Let us move on to the final size of the epidemic, τ . Remember that τ is defined as the proportion of the n initially susceptible individuals that is ultimately infected. To find an equation for τ we express the probability to escape infection in two different ways and equate. First we note that the proportion of the n initially susceptible individuals that escapes the epidemic equals $1 - \tau$, since τ is defined to be the proportion that does not escape. Now assume that $m_n/n \rightarrow \mu$ as $n \rightarrow \infty$ (i.e. μ is a measure of the asymptotic proportion of initially infected individuals present in the population) and consider the probability that a given susceptible individual escapes infection, p_{esc} . The probability that she escapes infection from a given infective equals $1 - \gamma/n$. To escape the whole epidemic she has to escape infection from all the $n\mu$ initially infected individuals and all the $n\tau$ individuals ultimately infected in the initially susceptible population, i.e.

$$p_{esc} = (1 - \gamma/n)^{n(\tau+\mu)} \rightarrow e^{-\gamma(\tau+\mu)} \quad \text{as } n \rightarrow \infty.$$

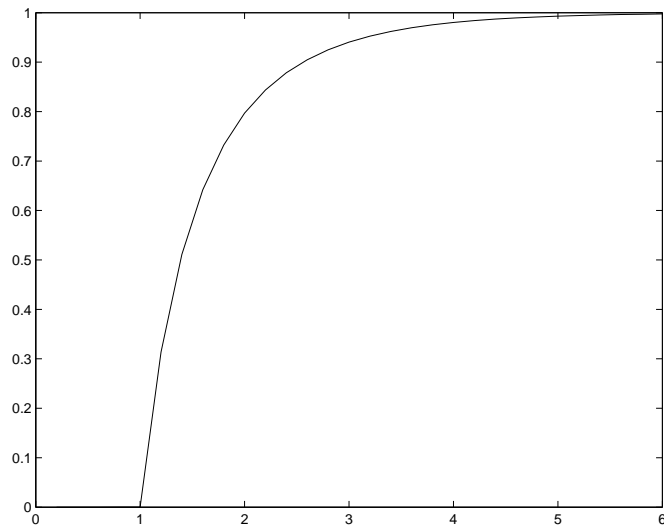


Figure 2.1: Final size against γ for $\mu = 0$.

Asymptotically the proportion that escapes the epidemic should equal the probability that a given susceptible escapes infection. Thus the final size is given by the solution to the equation

$$1 - \tau = e^{-\gamma(\tau+\mu)}. \quad (2.1)$$

This equation can not be solved explicitly but numerical solutions are easily obtained for given values of γ and μ . In Figure 2.1 τ is plotted against γ for $\mu = 0$.

Threshold limit theorem

We now present the famous threshold limit theorem for the Reed-Frost epidemic. This theorem gives us complete information about the distribution of the final size of the epidemic. We have seen that if a small number of infectives is introduced in a large susceptible population the number of infected individuals initially behaves like a branching process. If the basic reproduction number, R_0 , is less than or equal to 1 only smaller outbreaks are possible, whereas if R_0 is strictly greater than 1 there is a positive probability that the approximating branching process explodes, implying that the branching process approximation breaks down after some time. It seems reasonable that in this case the asymptotic distribution of the final size will consist of two parts, one branching process part close to zero and one other

part concentrated around the large outbreak size. Indeed this is exactly what the threshold theorem tells us. As $n \rightarrow \infty$ one of two possible scenarios can occur: either only few individuals will ever get infected, or else a deterministic positive proportion (with some Gaussian noise of smaller order) will have experienced the disease at the end of the epidemic.

Before stating the theorem let us remind the reader of some notation. Consider the Reed-Frost epidemic spread in a population initially consisting of n susceptible individuals.

$$\begin{aligned} m_n &= \text{number of initial infectives introduced in the population} \\ Z_n &= \text{number of initially susceptible individuals ultimately infected} \\ Z'_n &= Z_n + m_n \\ \mu &= \lim_{n \rightarrow \infty} m_n/n \\ E_{n,m_n}(\gamma) &: \text{denotes the entire process.} \end{aligned}$$

The threshold limit theorem is concerned with the *asymptotic* distribution of the final size Z_n , i.e. we study a *sequence* of processes $E_{n,m_n}(\gamma)$ and investigate what happens with Z_n as n tends to infinity, i.e. as the size of the susceptible population becomes large. As mentioned in Section 2.1 the number of initially infected individuals will have impact on the final size of the epidemic. In some situations we expect the number of initial infectives to increase with the population size in some linear way, i.e. $m_n = n\mu$ in the limit, where $\mu > 0$, while in other situations the number of initial infectives stays constant as n increases, i.e. $m_n \equiv m$, implying that $\mu = 0$. Which of these cases that is at hand together with whether R_0 is below or above 1 determines the behaviour of the final size according to the threshold theorem now stated:

Theorem 2.3.2 *Consider a sequence of epidemic processes E_{n,m_n} . Define τ as the nontrivial solution to the final size equation (2.1).*

(i) *Assume $m_n/n \rightarrow \mu > 0$ as $n \rightarrow \infty$.*

Then the sequence $\sqrt{n}(Z'_n/n - (\tau + \mu))$ converges to a normally distributed random variable with mean 0.

(ii) *Assume $m_n \equiv m$ for all n , i.e. $\mu = 0$.*

If $\gamma \leq 1$ then with probability one Z_n converges to Z , where $P(Z < \infty) = 1$ and Z is the total progeny in a discrete time branching process, initiated by m ancestors, in which individuals on average give birth to γ children.

If $\gamma > 1$ then Z_n still converges to Z , but now $P(Z < \infty) = q^m$, where q^m is the extinction probability of the branching process. With probability $1 - q^m$,

the sequence $\sqrt{n}(Z'_n/n - (\tau + \mu))$ converges to a normally distributed random variable with mean 0.

In words the theorem tells us that in case of a large number of initial infectives the final size proportion Z_n/n converges in distribution to a point mass at τ , the solution to the final size equation (2.1) with $\mu > 0$. The fluctuations around the limit are Gaussian of order $1/\sqrt{n}$. In case of a small number of initial infectives we have two possible scenarios. If the basic reproduction number exceeds 1, then Z_n/n converges in distribution to a random variable with mass q^m at the point 0 and mass $1 - q^m$ at the point τ , where τ is the nontrivial solution of equation (2.1) with $\mu = 0$. Again the fluctuations around τ are Gaussian of order $1/\sqrt{n}$ and the mass around 0 is distributed as the total progeny in the approximating branching process. If the basic reproduction number is less than or equal to 1 all probability mass lies in the branching process part, i.e. a large outbreak is not possible. The proof of the theorem is complicated and therefore omitted, see e.g. Scalia-Tomba (1985).

2.4 Summarizing discussion

The Reed-Frost model was introduced approximately seventy years ago as a mathematical model for the spread of measles in a human population. It was the first epidemic model using a stochastic setting. The advantage of the model lies primarily in its very simple structure, which makes it relatively easy to work with. As seen in the previous section various kinds of calculations can be performed and we are able to derive the asymptotic behaviour of the epidemic. The drawback of the model lies in a number of simplifying assumptions included in the formulation, drastically decreasing the realism and applicability. The course of the disease described by the model is assumed to consist of a latent period of constant length followed by a very short infectious period. Immediately after the infectious period symptoms appear and the infective individual is isolated from the susceptible population. With these assumptions the spread of the disease appears in generations of infectives, separated by a latent period of length one. It is of course easy to think of diseases that do not fit into this framework and indeed one weakness in the Reed-Frost model is that it is not very flexible in its description of the disease. The model can be generalized though, to the so called *standard SIR epidemic model*, in which each infective generates new cases according to a Poisson process during an arbitrarily distributed infectious period. The restrictions imposed on the population in which the

epidemic takes place however remains also in this generalized version. The first restriction is that the total population is assumed to stay constant in size during the spread of the disease, i.e. we assume that no births, deaths, immigration or emigration occur during the epidemic. If we consider a short time epidemic this might be approximately true, while we have to be more careful if we are working with a wider time prospect. The second restriction is that all members of the population are assumed to be of the same type with regard to the disease, i.e. all individuals are assumed to be equally susceptible to the disease and equally infectious if infected. In real-life epidemics this assumption is often violated; usually children are more susceptible to flues for example, and sometimes individuals with previous history to the disease have acquired partial immunity. Varying infectivity might be present for example if the disease is sexually transmitted, in that some individuals are more promiscuous than others. The third restriction imposed on the population is that it is assumed to be uniformly mixing, i.e. a given individual has contact with each other individual in the population with *equal* probability ($=\gamma/n$, i.e. a very small number if the population is large). In a human population this is of course rarely the case. Social structures, such as households and friendship relations, cause the contact probability to be much greater within small subgroups in the population.

In this work we introduce heterogeneous mixing in the population by aid of an underlying social network graph. To be more precise, we generate a graph to represent the social structure in the population and then we study the epidemic spread on this graph, where a given individual contacts each *neighbor* in the graph with probability p . This should yield a more realistic model than the one assuming homogeneous mixing, since it stipulates that individuals mix primarily with individuals to which they are socially related in some way. Furthermore, by choosing a network graph where the neighborhood size is bounded, we do not have to scale the contact probability by n to make the model work. The contacts with the graph neighbors can take place with a "normal" probability p , which of course is more realistic. Indeed, the probability that a person meet with her best friend does not decrease as the population size grows.

We find that the concept of heterogeneous mixing has a fundamental influence on the basic modelling assumptions. Other effects, caused by for instance multi-type populations, births and deaths, immigration and emigration, partial or temporary immunity, could presumably be incorporated at later stages.

Chapter 3

Epidemics on graphs

In this chapter we set up the modelling assumptions that will be used during the rest of the work. First we introduce some graph theoretic notation and show how this notation can be used to describe a social network. We modify the Reed-Frost model from the previous chapter and describe the epidemic spread on a fixed graph. Finally we define some properties that we want a social network graph to exhibit.

3.1 Graph theoretic notation

A labelled *graph* \mathcal{G} consists of two sets of information: a set of labelled *vertices*, $\mathcal{V} = (v_1 \dots v_N)$, and a set of *edges*, $\mathcal{E} = (e_1 \dots e_E)$, between pairs of vertices. There are N vertices and E edges, where each edge can be written as a pair of vertices, $e_k = (v_i, v_j)$. In this work we will only be concerned with *undirected* graphs, i.e. graphs where each edge consists of an *unordered* pair of vertices, implying that the edge between v_i and v_j is identical to the edge between v_j and v_i . We will not allow edges between a vertex and itself, (v_i, v_i) . Such edges are called *loops*. Also we do not allow duplicate edges, i.e. a pair of vertices can not be included more than once in \mathcal{E} . A graph like this, that has no loops and includes no more than one undirected edge between a pair of vertices, is called a *simple, undirected graph*.

More terminology: Two vertices, v_i and v_j , are *adjacent* if the pair (v_i, v_j) is in the set of edges, i.e. if there is an edge between v_i and v_j . There is a *path* between v_i and v_j if there is a collection of edges $(v_i, v_{k_1}), (v_{k_1}, v_{k_2}), \dots, (v_{k_n}, v_j)$ connecting v_i and v_j . All edges in the path are assumed to be distinct. We say that two vertices belong to the same *component* if there exists a path between them. This implies that any graph consists of a number of disjoint

components. Finally we define the *degree* of a vertex v_i to be the number of vertices adjacent to it.

3.2 Social network graphs

When a graph is used to describe a social network the vertices are used to represent the members of the population and the edges are used to represent ties between the members. If we want to describe the social structure in a population consisting of N individuals we proceed as follows:

1. Represent each individual i with a labelled vertex v_i , $i = 1, \dots, N$.
2. Draw an edge between vertices v_i and v_j if individuals i and j are acquainted with each other in some way.

The graph obtained in this way is easily seen to be a description of the social relations in the population.

If two vertices, v_i and v_j , in a social network graph are adjacent we call the corresponding individuals, i and j , *neighbors*. This means that i and j are friends or have some other kind of social relation to each other. The existence of a path between two vertices means that there is a friendship chain connecting the corresponding individuals in the population and the degree of a vertex v_i in a friendship graph tells us the number of friends of individual i .

3.3 Epidemic spread on a fixed graph

We now modify the Reed-Frost model from Chapter 2 slightly and define the epidemic spread along a fixed graph.

Reed-Frost epidemic on a fixed graph: We consider a closed, homogeneous population, consisting of N individuals, where the neighborhood structure is represented by an undirected, labelled graph \mathcal{G} , which is assumed to be fixed during the course of the epidemic. Let n denote the number of initially susceptible individuals and m_n the number of initially infectious individuals, where $n + m_n = N$ and m_n/n small. At $t = 0$ we pick the m_n initial infectives at random from the population. The dynamic of the epidemic then runs as follows: Assume that an individual i is infected at time t . A given *neighbor* in \mathcal{G} ,

j , is contacted by i with probability p and if j is susceptible then j becomes infected at time $t+1$. At time $t+1$ also, i becomes removed (by immunity or death) and plays no further part in the epidemic process. The epidemic ceases when there are no infectious individuals present in the population. All contacts are assumed to be independent of each other.

At a first glance this model may seem very similar to the one presented in Chapter 2. Indeed, they both assume a constant latent period lasting for one unit of time and an infectious period reduced to a single point in time. But the above model differs from the model defined in Chapter 2 on two crucial points:

1. A given infective can only infect her *neighbors* in the social network graph.
2. The infection probability, p , is not scaled by n .

Provided that we choose a graph that describes the social structure in the population accurately this should be a more realistic framework than one in which a given infective infects every other individual in the population with the same probability, scaled by n . In real-life infectives can of course only infect individuals with whom they have some kind of contact and it seems reasonable that individuals have contact primarily with individuals to which they are socially related in some way, i.e. with the neighbors in the social network graph. In the above model we have introduced this kind of heterogeneous mixing in the population by stipulating the infection probability to be different for neighbors ($=p$) and non-neighbors ($=0$). The infection probability should be thought of as a product of the contact probability and the probability of a disease transmission in case of a contact, thus giving the probability of a contact resulting in a new infective. Since the population is assumed to be homogeneous with respect to the disease (i.e. all individuals are equally susceptible and equally infective if infected) the probability of a disease transmission in case of a contact is the same over the entire population. Hence, the difference in the infection probability between neighbors and non-neighbors is caused by different contact probabilities, i.e. the population is heterogeneously mixing. This justifies the fact that the model is sometimes referred to as the *heterogeneous Reed-Frost model*.

In the model defined in Chapter 2 the infection probability was scaled by n . The scaling was motivated from technical reasons: since a given infective

could infect all susceptible individuals in the population the model would otherwise have exploded. In the above model, where infectives can only infect their neighbors in the underlying social network graph, no scaling is necessary provided we choose a network graph where the neighborhood size is bounded. Bounded neighborhood size is indeed a very natural restriction to impose on a graph describing social relations. The average neighborhood size tells us the average number of friends of a member in the population and this number should of course be bounded: no one has infinitely many friends.

3.4 Properties of a social network graph

The population size is throughout this work assumed to be large. Typically we consider a population consisting of all inhabitants in a city or a in country. This means that it is an impossible task to delineate the social relations in detail. We have to rely on *models* of the social structures. In this work we introduce different types of *random graphs* as models for social networks. Assume that we are considering a population consisting of N individuals. As before we represent the N individuals with labelled vertices v_1, \dots, v_N . The edges representing the neighborhood structure in the population are then generated according to some random procedure. The resulting graph is our social network model. The problem is to find random procedures that yields graphs that are complicated enough to catch something of the irregular contact pattern in a human population and yet simple enough to lend themselves to mathematical analysis. Let us list a few properties that we want the graph to exhibit:

- *Bounded neighborhood size.* As explained above, one restriction on the graph, necessary to keep the model from exploding, is for the average neighborhood size to be bounded, i.e. the average number of vertices adjacent to a vertex in the graph must remain finite as $N \rightarrow \infty$. In a friendship graph this restriction reflects the fact that friendship circles are limited in size also in large populations.
- *Transitivity.* We want the graph to be highly transitive, i.e. we want it to contain a certain amount of triangles. This is a consequence of the fact that we expect many of our friends to be friends also of each other. Hence, we are looking for a model where the probability that a certain edge is present given the absence of presence of all other edges is larger if its end vertices are second neighbors than if they are not.

- *Realistic dependence structure.* The graph must not contain unnatural dependence structures among the edges. For example the absence or presence of a certain edge should not depend on information regarding parts of the graph that are located far away from the edge in question. Indeed, the social behaviour of a person is often influenced by the behaviour of her friends but it is not likely to be influenced by people that she does not have any kind of relation to, i.e. by people located far away from her in the social network graph.

In the rest of this work we present a number of possible choices for the social network graph and study its effects on the spread of infection according to the modified Reed-Frost model. We will be particularly interested in the basic reproduction number and the final size of the epidemic. Since we most often consider large populations when dealing with epidemic modelling, it is important for the graph to "behave well" asymptotically. That is, we want the graph to exhibit the kind of behaviour described above as the population size tends to infinity. It will become clear that the mathematical analysis is made considerably more difficult as the complexity of the graph increases. For epidemic models to approach reality though, it is crucial to find ways to attack also these more complicated model structures.

Chapter 4

Bernoulli random graphs

A so called Bernoulli random graph is constructed by adding edges uniformly over a set of vertices. It is the simplest possible choice of underlying network in our epidemic model and we present it here to illustrate the techniques that we will use when attacking more complicated structures. Bernoulli random graphs was introduced in the late fifties by Erdős and Renyi and an extensive treatment of the model can be found in e.g. Bollobas (1985).

4.1 Construction of the network

Given a set of N labelled vertices, a Bernoulli graph is generated by connecting each given pair of vertices by drawing an edge between them with some probability r . That is, we consider all possible pairs, (v_i, v_j) $i, j = 1, \dots, N, i \neq j$, and independently of each other we include each of them in \mathcal{E} with probability r . Since there are $\binom{N}{2}$ possible edges between N given vertices and each of these edges is either present or not present in a Bernoulli graph, there are $2^{\binom{N}{2}}$ different Bernoulli networks of order N . We use $\mathcal{G}(N, r)$ to denote this class of graphs.

The number of vertices adjacent to a given vertex v_i in a Bernoulli graph is binomially distributed with parameters $N - 1$ and r . To see this we note that there are $N - 1$ possible edges that includes v_i as one endpoint and each of these edges is present in the graph with probability r . This implies that the average degree of a vertex is equal to $(N - 1)r$. Remember though that one of our restrictions on the network graph, formulated in the previous section, was for the average neighborhood size to stay bounded as the population size, N , tends to infinity. Thus, to make the $\mathcal{G}(N, r)$ graph a possible candidate as a model for the social network, we put $r = \lambda/N$

for some $\lambda > 0$. Recalling the Poisson approximation from Section 2.2 we deduce that if N is large the degree of a vertex in a $\mathcal{G}(N, \lambda/N)$ graph is approximately Poisson distributed with parameter λ .

4.2 Epidemic behaviour

Given an outcome \mathcal{G} belonging to $\mathcal{G}(N, \lambda/N)$ we run a Reed-Frost epidemic on \mathcal{G} , i.e. we infect a number of individuals at random in the population and let the disease spread according to the rule that each infective infects each of her neighbors in \mathcal{G} with probability p .

Basic reproduction number

First we calculate the basic reproduction number. Remember that in the standard Reed-Frost model the basic reproduction number, R_0 , is defined as the average number of new cases generated by one infectious individual in a large, susceptible population. Let us find an expression for this quantity in a heterogeneous Reed-Frost epidemic where the underlying social network is modelled as a Bernoulli graph. Consider a given infective, i , living in a large, susceptible population where the neighborhood structure is represented by a graph $\mathcal{G} \in \mathcal{G}(N, \lambda/N)$. Let D_i denote the number of neighbors of i in \mathcal{G} . Each of these neighbors is, independently of each other, infected by i with probability p . Hence the number of new infectives generated by i is binomially distributed with parameters D_i and p . It can be shown that the expected value in this distribution is equal to $E[D_i]p$. Since the population is assumed to be large, D_i is approximately Poisson distributed with parameter λ and thus $E[D_i] = \lambda$. Hence the expected number of secondary cases generated by i is equal to λp .

Provided that the population is large, contacted individuals in the epidemic process are susceptible with high probability in the beginning of the time course. Thus the initial stage in the generation process of infectives is well approximated by a branching process with reproduction mean equal to the average number of new cases generated by a given infective. According to Theorem 2.3.1 there is a positive probability for this branching process to explode if and only if the reproduction mean is strictly greater than one. Translating this into epidemic terms yields that a major outbreak is possible if and only if an infective in the beginning of the time course infects more than one individual on the average. Thus, the quantity λp works as a flag, indicating whether a major outbreak is possible or not. Hence, we put $R_0 = \lambda p$ if $\mathcal{G} \in \mathcal{G}(N, \lambda/N)$.

As an alternative, R_0 can be found using random graph methods. Let us state a well-known result from graph theory:

Theorem 4.2.1 *Consider the $\mathcal{G}(N, r)$ model. Assume $r = r_N = \beta/N$, some $\beta > 0$, $N \rightarrow \infty$. The following holds:*

(i) *If $\beta \leq 1$ then, with probability one, a vertex chosen at random will belong to a component of relative size $O(1/N)$ (i.e. the relative size of the component tends to zero as N tends to infinity).*

(ii) *If $\beta > 1$ then the relative size of the largest component converges in probability to some constant c , $0 < c < 1$. A vertex chosen at random will belong to this giant component with probability c and to a component of relative size $O(1/N)$ with probability $1 - c$.*

To apply this theorem to our epidemic model, for a short while we let r denote infection probability instead of network probability, i.e. an edge between vertices v_i and v_j means, not only that individuals i and j are neighbors, but also that a disease transmission has actually taken place. We stress that this is a different way of using the $\mathcal{G}(N, r)$ model than in the rest of this section; an outcome \mathcal{G} is now used to get a picture of the epidemic spread, not as model for the social network. To calculate $r = r_N$ we consider the probability that a given susceptible individual i is infected by a given infective j . For this to happen, firstly j has to be a neighbor of i , i.e. there must be an edge connecting vertices v_i and v_j in the social network graph. Since this network is a Bernoulli $(N, \lambda/N)$ graph, the pair (v_i, v_j) is included in the set of edges with probability λ/N . Secondly, given that i and j are neighbors, a disease transmission has to take place. This happens with probability p . Hence the probability that i is infected by j is equal to $\lambda p/N$ and thus a $\mathcal{G}(N, \lambda p/N)$ graph provides a model for the epidemic spread. In this graph an edge between vertices v_i and v_j should be interpreted so that *if* individual i ever gets infected, then i will infect individual j . The above theorem with $\beta = \lambda p$ now immediately gives us R_0 : Given an outcome $\mathcal{G} \in \mathcal{G}(N, \lambda p/N)$ we pick m vertices at random from the graph and infect the corresponding individuals. These initially infected individuals will cause their entire components in the graph to be infected and hence the final size of the epidemic is equal to the total relative size of the components to which the initial infectives belongs. According to the theorem, if λp does not exceed one, each initial infective will, with probability one, belong to a component with asymptotic relative size equal to zero, yielding only smaller outbreaks. If λp is strictly greater than one though, there is a nonzero probability c that an initial infective belongs to a giant component (i.e. a component with size

of the same order as the population). This induces a major outbreak in the epidemic. Hence a major outbreak is possible if and only if λp exceeds one, implying that $R_0 = \lambda p$.

Final size equation

Let us heuristically derive an equation for the final size of the epidemic, τ , valid in large populations. Remember that τ is defined as the proportion of the n initially susceptible individuals that is ultimately infected. First we note that the proportion of the n initially susceptible individuals that escapes the epidemic equals $1 - \tau$, since τ is defined to be the proportion that does not escape. Now let A denote the event that a given susceptible individual i escapes infection and let Y denote the number of neighbors of i in \mathcal{G} . Conditioning on Y yields

$$P(A) = E[P(A|Y)] = E[P(A_j)^Y]$$

where A_j denotes the event that i escapes infection from a given neighbor j . If j is infected she infects i with probability p . Hence

$$P(A_j) = 1 - pq_{inf}$$

where q_{inf} denotes the probability that j is ever infected during the epidemic. Asymptotically the probability that a given individual is infected should equal the proportion of the population that is ultimately infected in the epidemic. Since the number of individuals ultimately infected is equal to the sum of the $n\mu$ initially infected individuals and the $n\tau$ individuals ultimately infected in the initially susceptible population we have

$$q_{inf} = (n\tau + n\mu)/N = (\tau + \mu)/(1 + \mu)$$

where the second equality follows from the identity $N = n(1 + \mu)$. Hence

$$P(A_j) = 1 - p(\tau + \mu)/(1 + \mu)$$

implying that

$$P(A) = E[(1 - p(\tau + \mu)/(1 + \mu))^Y]. \quad (4.1)$$

In general, the probability generating function of a random variable X is defined as $\varphi_X(k) := E[k^X]$. In particular, if X is Poisson (β) distributed $\varphi_X(k) = e^{-\beta(1-k)}$. Using the facts that the right-hand side of equation (4.1) is equal to the probability generating function of Y (the number of neighbors

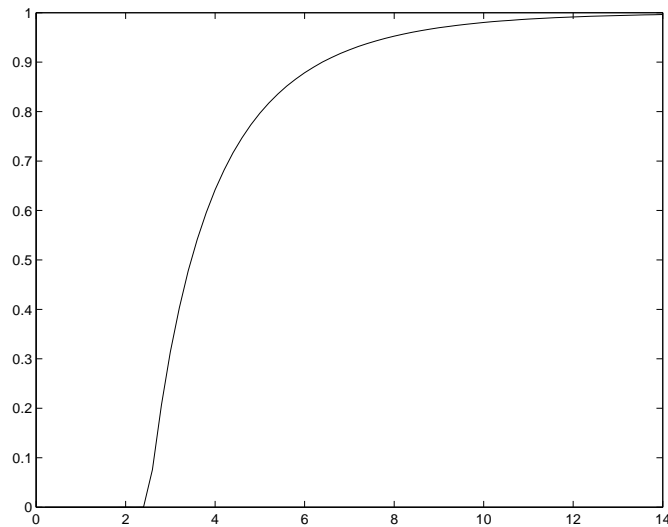


Figure 4.1: Final size against λ for $\mu = 0$.

of i in \mathcal{G}) evaluated at $1 - p(\tau + \mu)/(1 + \mu)$ and that Y is asymptotically Poisson (λ) distributed, yields

$$P(A) = e^{-\lambda p(\tau + \mu)/(1 + \mu)}.$$

Asymptotically the proportion that escapes the epidemic should equal the probability that a given susceptible escapes infection. Thus the asymptotic final size of the epidemic is given by the solution to the equation

$$1 - \tau = e^{-\lambda p(\tau + \mu)/(1 + \mu)}.$$

This equation can not be solved explicitly but numerical solutions are easily obtained. In Figure 4.1 τ is plotted against λ for $p = 0, 4$. We have assumed a small number of initial infectives, i.e. $\mu = 0$.

4.3 Properties of the graph

Is the Bernoulli random graph a good model to represent a social network? The construction procedure is very simple and maybe this should lead one to suspect that the resulting graph is not complicated enough to catch the often very complex social formations in a human population. Since the edge probability is scaled by the population size, the average neighborhood size

is indeed bounded, but what about the other properties defined in Section 3.4? Let us investigate the transitivity of the graph. Remember that social networks are highly transitive, due to the fact that the probability for two people to make friends with each other increases if they have a mutual acquaintance. In a Bernoulli graph though, the probability that individuals i and j are neighbors given that they are second neighbors is equal to λ/N , since all connections are made independently of each other. This is a result that is not desirable for us, since it implies that asymptotically the graph contains very few triangles. To conclude, the Bernoulli random graph fails to catch a very important feature present in the human contact pattern and can thus not be regarded as a good model for a social network.

Chapter 5

The 3-clique model

In graph theory the concept of an *m-clique* is used to denote a complete subgraph of order m , i.e. a set of m vertices between which all $\binom{m}{2}$ possible edges are present. Thus, a 2-clique is a single edge, a 3-clique is a triangle, and so on. An *m-clique graph* is obtained by randomly constructing cliques of order $2, \dots, m$ in an initially empty graph, i.e. a graph with initially no edges in it. The Bernoulli random graph (see Chapter 4), where single edges are randomly added over a set of vertices, is another name for the 2-clique graph. In this chapter we describe the 3-clique graph, where, besides single edges, also triangles are included. This is an attempt to increase the transitivity of the model as compared to the Bernoulli random graph, in which the amount of triangles was found to be asymptotically very small. A model based on the addition of complete subgraphs of order higher than 3 can hardly be regarded as a realistic model of friendship formation: it is hardly true that one is more likely to make friends with a friend of a friend of a friend than with anyone else. A thorough description of the 3-clique graph can be found in Karlberg (1997) and criticism of it as a model for a social network is brought up by Jonasson (1997).

5.1 Construction of the network

Given a set of N labelled vertices a 3-clique graph is constructed by randomly adding edges and triangles. More formally; first we consider all pairs of vertices (v_i, v_j) $i, j = 1, \dots, N, i \neq j$, and with probability r we include the pair in the set of edges, i.e. we draw an edge between vertices v_i and v_j . Then we consider all triples of vertices (v_i, v_j, v_k) $i, j, k = 1, \dots, N, i \neq j \neq k$, and with probability \tilde{r} we include the corresponding triangle (i.e. the pairs

(v_i, v_j) , (v_j, v_k) and (v_k, v_i)) in the set of edges. All edges and triangles are included independently of each other, with duplicate edges forbidden. This construction procedure generates a graph \mathcal{G} that can be written as the union of two independent graphs \mathcal{G}_1 and \mathcal{G}_2 , where \mathcal{G}_1 is a Bernoulli graph with edge probability r and \mathcal{G}_2 is a graph with all possible triangles present independently of each other with probability \tilde{r} .

Given a vertex v in a 3-clique graph $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$ we let X denote the number of single edges with v as one endpoint and \tilde{X} the number of triangles in which v is used as a corner point, i.e. X is the degree of vertex v in \mathcal{G}_1 and \tilde{X} is the number of triangles in \mathcal{G}_2 involving vertex v . The total number of vertices adjacent to v is equal to $X + 2\tilde{X}$, since each edge in \mathcal{G}_1 including v as an endpoint provides one adjacent vertex and each triangle in \mathcal{G}_2 that v is a part of provides two adjacent vertices. Hence, for the average neighborhood size in \mathcal{G} to stay bounded in large populations it is necessary that both $E[X]$ and $E[\tilde{X}]$ are asymptotically finite. There are $N - 1$ single edges with v as one endpoint and each of these possible edges is, independently of each other, present in \mathcal{G}_1 with probability r . Hence X is binomially distributed with parameters $N - 1$ and r , implying that $E[X] = (N - 1)r$. To keep this number bounded as N tends to infinity, we put $r = \lambda/(N - 1)$ for some $\lambda > 0$. Thus $E[X] = \lambda$ for all N and according to the Poisson approximation from Section 2.2 X is asymptotically Poisson (λ) distributed. Moving on to \tilde{X} we note that to construct a triangle including v as a corner point we have to pick two other vertices among the remaining $N - 1$ vertices in the graph to serve as the other two corner points. This can be done in $\binom{N-1}{2}$ different ways, implying that there are $\binom{N-1}{2}$ possible triangles involving v . Each of these triangles is, independently of all other triangles, present in \mathcal{G}_2 with probability \tilde{r} . Hence, \tilde{X} is binomially distributed with parameters $\binom{N-1}{2}$ and \tilde{r} . To keep the expected value in this distribution bounded we put $\tilde{r} = \tilde{\lambda}/\binom{N-1}{2}$ for some $\tilde{\lambda} > 0$. Thus, $E[\tilde{X}] = \tilde{\lambda}$ for all N and \tilde{X} is asymptotically Poisson distributed with parameter ($\tilde{\lambda}$).

5.2 Epidemic behaviour

Given a graph \mathcal{G} generated according to the random procedure described above we let a Reed-Frost epidemic spread along \mathcal{G} . Remember that the graph is assumed to be fixed during the course of the epidemic.

Basic reproduction number

Consider a given infective i in a large susceptible population where the social

structure is represented by a 3-clique graph $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$. Let R_1 denote the expected number of secondary cases generated by i in \mathcal{G}_1 and let R_2 denote the expected number of secondary cases generated by i in \mathcal{G}_2 . R_0 is defined as the total number of secondary cases generated by i , i.e.

$$R_0 = R_1 + R_2.$$

The number of individuals connected to i in \mathcal{G}_1 is equal to X (defined above). Each of these individuals is infected by i with probability p . Hence,

$$R_1 = E[X]p = \lambda p.$$

To calculate R_2 ; let Z denote the number of secondary cases generated by i in a fixed triangle $\{(v_i, v_j), (v_j, v_k), (v_k, v_i)\}$. In this notation we have

$$R_2 = E[\tilde{X}]E[Z]$$

where \tilde{X} is the number of triangles in \mathcal{G}_2 involving v_i . By construction, $E[\tilde{X}] = \tilde{\lambda}$. Hence it remains to calculate $E[Z]$. To do this we first note that $Z \in \{0, 1, 2\}$. The probability of one new infective is equal to $2p(1-p)^2$, since for exactly one new infective to appear a disease transmission has to take place along one of the edges (v_i, v_j) and (v_i, v_k) but not along the remaining two edges. In case of two new infectives, i.e. in case both j and k are infected, we have two possibilities:

1. Both j and k are infected directly by i . In this case disease transmission takes place along the edges (v_i, v_j) and (v_i, v_k) . The probability for this event is equal to p^2 .
2. The infection takes place in two steps, i.e. i infects $j(k)$ and $j(k)$ infects $k(j)$. In this case disease transmission takes place along the edges $(v_i, v_{j(k)})$ and $(v_{j(k)}, v_{k(j)})$ but not along the edge $(v_i, v_{k(j)})$. The probability for this event is equal to $2p^2(1-p)$.

To conclude, the probability of two new infectives equals $p^2 + 2p^2(1-p)$. Hence

$$E[Z] = 2p(1-p)^2 + 2(p^2 + 2p^2(1-p)) = 2(p + p^2 - p^3)$$

implying that

$$R_2 = E[\tilde{X}]E[Z] = 2\tilde{\lambda}(p + p^2 - p^3).$$

Putting all this together yields

$$R_0 = R_1 + R_2 = \lambda p + 2\tilde{\lambda}(p + p^2 - p^3).$$

Final size equation

We turn to find an equation for the final size proportion of the epidemic, τ . For the sake of simplicity we consider only the case when $\mu = 0$, i.e. we assume a small number of initial infectives. Just as when deriving the final size equation for the Bernoulli random graph we will set about it by expressing the probability to escape infection in two different ways and equate. First we introduce some notation. Consider a given susceptible individual i and let

$$\begin{aligned} A &= \{i \text{ escapes infection}\} \\ A' &= \{i \text{ escapes infection along single edges}\} \\ A'' &= \{i \text{ escapes infection along triangle edges}\} \end{aligned}$$

Due to independence we have

$$P(A) = P(A')P(A''). \tag{5.1}$$

First we consider the probability that i escapes infection along single edges, $P(A')$. Conditioning on X (the number of neighbors of i in \mathcal{G}_1) yields

$$P(A') = E[P(A'|X)] = E[P(A'_j)^X],$$

where A'_j denotes the event that i escapes infection from a given neighbor j in \mathcal{G}_1 . If j is infected during the epidemic the probability that i escapes infection from j is equal to $1 - p$. If j escapes the epidemic the probability that i escapes infection from j is of course equal to 1. Asymptotically the probability that a given individual is infected by the epidemic is equal to the proportion of the population that is ultimately infected, τ . Hence

$$P(A'_j) = (1 - \tau) + \tau(1 - p) = 1 - \tau p$$

implying that

$$P(A') = E[(1 - \tau p)^X].$$

The right-hand side in this equality can be recognized as the probability generating function of X evaluated at $1 - \tau p$. Using the fact that X is asymptotically Poisson (λ) distributed yields

$$P(A') = e^{-\lambda\tau p}. \quad (5.2)$$

Moving on to $P(A'')$ we obtain by conditioning on \tilde{X} (the number of triangles in \mathcal{G}_2 involving vertex v_i)

$$P(A'') = E[P(A''|\tilde{X})] = E[P(A''_{\Delta})^{\tilde{X}}], \quad (5.3)$$

where A''_{Δ} denotes the event that i escapes infection in a fix triangle $\Delta = \{(v_i, v_j), (v_j, v_k), (v_k, v_i)\}$. To find an expression for $P(A''_{\Delta})$ we condition on whether $j(k)$ brings the disease on to $k(j)$ or not if $j(k)$ is infected in the epidemic, i.e.

$$P(A''_{\Delta}) = P(B)P(A''_{\Delta}|B) + P(B^c)P(A''_{\Delta}|B^c) \quad (5.4)$$

where

$$B = \{\text{an infection link is present between vertices } v_i \text{ and } v_j\}.$$

In the case when both j and k are infected as soon as one of them is (i.e. in case B occurs) we have three possible scenarios:

1. Neither j nor k is ever infected during the epidemic. Since the asymptotic probability for an individual to escape infection is equal to $(1-\tau)$ this occurs with probability $(1-\tau)^2$. In this case the probability that i escapes infection from j and k is of course equal to 1.
2. One of j and k is infected. The probability for this event is equal to $2\tau(1-\tau)$. Since we are assuming that the infected individual brings the disease on to the uninfected individual the probability that i escapes infection from both j and k is equal to q^2 (where $q = (1-p)$) in this case.
3. Both j and k are infected. This occurs with probability τ^2 . The probability that i escapes infection from both j and k is equal to q^2 .

Thus

$$P(A''_{\Delta}|B) = (1-\tau)^2 + 2\tau(1-\tau)q^2 + \tau^2q^2.$$

In the case when there is no infection link between vertices v_i and v_j we reason in much the same way as above, the only difference being that when only one of j and k is infected the probability that i escapes infection from

j and k equals q instead of q^2 , since the infected individual does not bring the disease on to the uninfected individual. Hence

$$P(A''_{\Delta}|B^c) = (1 - \tau)^2 + 2\tau(1 - \tau)q + \tau^2q^2.$$

Since $P(B) = p$ we obtain from equation (5.4):

$$\begin{aligned} P(A_{\Delta}) &= p((1 - \tau)^2 + 2\tau(1 - \tau)q^2 + \tau^2q^2) + \\ &\quad q((1 - \tau)^2 + 2\tau(1 - \tau)q + \tau^2q^2) \\ &= 1 - (2\tau(2 - \tau)p + \tau(2 - 3\tau)p^2 - 2\tau(1 - \tau)p^3). \end{aligned} \quad (5.5)$$

Now return to equation (5.3). The right-hand side can be recognized as the probability generating function of \tilde{X} evaluated at $P(A''_{\Delta})$. Using the expression for $P(A''_{\Delta})$ derived in equation (5.5) and the fact that \tilde{X} is asymptotically Poisson ($\tilde{\lambda}$) distributed yields

$$P(A'') = \exp \{-\tilde{\lambda}(2\tau(2 - \tau)p + \tau(2 - 3\tau)p^2 - 2\tau(1 - \tau)p^3)\}. \quad (5.6)$$

Finally we substitute the expressions for $P(A')$ and $P(A'')$ (equations (5.2) and (5.6)) into equation (5.1) and obtain the following expression for the asymptotic probability that a given susceptible individual escapes infection

$$P(A) = \exp \{-\lambda\tau p - \tilde{\lambda}(2\tau(2 - \tau)p + \tau(2 - 3\tau)p^2 - 2\tau(1 - \tau)p^3)\}.$$

Asymptotically the probability that a given susceptible escapes infection should equal $1 - \tau$, the proportion that escapes the epidemic. Thus the final size of the epidemic is given by the solution to the equation

$$1 - \tau = \exp \{-\lambda\tau p - \tilde{\lambda}(2\tau(2 - \tau)p + \tau(2 - 3\tau)p^2 - 2\tau(1 - \tau)p^3)\}.$$

5.3 Properties of the graph

As explained in Section 3.4, when modelling a social network one wants to study random graphs which contains a certain amount of triangles. In a 3-clique graph the amount of triangles can be increased by choosing the parameter λ large. Hence, the model certainly includes an element of transitivity. It turns out though, that the dependence structure in the graph has an undesirable property, in that the conditional probability that a certain edge is present given the rest of the graph exhibits an unnatural behaviour. To

understand this, consider the probability that a fixed pair (v_i, v_j) is included in the set of edges in the graph \mathcal{G}_2 given the rest of \mathcal{G}_2 (where \mathcal{G}_2 is the "triangle part" of a 3-clique graph $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$). If the edge would not form any triangle this probability is zero. If it would form a triangle the conditional probability for the edge to be present is $(N-2)\tilde{\lambda}/\binom{N-1}{2} = 2\tilde{\lambda}/(N-1)$ if its end vertices are already parts of other triangles and one if not. This means that to determine the probability that individuals i and j are neighbors in \mathcal{G}_2 requires information not only about whether they share a mutual friend or not, but also about the further acquaintances of the mutual friend. This information is indeed not likely to be relevant in friendship formations. When including single edges in the model (i.e. when mixing the graph \mathcal{G}_2 with \mathcal{G}_1) the mentioned effect is slightly weakened, but not completely ruled out. This unnatural dependence structure between the edges somewhat decreases the suitability with the 3-clique graph as a model for a social network.

Chapter 6

Markov graphs

The Markov random graph model is an attempt to create a model which has an element of transitivity but which does not share the drawback of the 3-clique graph mentioned above, i.e. a model with a large amount of triangles but without intricate dependencies between the edges. The definition of the model is given in form of a probability measure which makes it possible for us to reward transitivity and stipulates that all edges not sharing a node should be independent. At first this may seem to be a promising construction, but this is flawed. It turns out that the model exhibits a very undesirable asymptotic behaviour (Strauss (1986)). Also; the rather implicit way in which the model is defined makes it a difficult task to derive results about the epidemic spread. However, such results would perhaps not be of much interest, since in epidemic modelling we most often assume the population size to be large and the asymptotic properties of the Markov graph severely decreases its applicability as a model for the social network in a large population. Here we will concentrate on describing the model and illustrate the asymptotic behaviour by aid of some simulation results.

6.1 Description of the model

A graph \mathcal{G} on N vertices is described by the *edge indicators* I_{ij} $i, j = 1 \dots N$, defined by

$$I_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

i.e. I_{ij} is equal to one if there is an edge between vertices v_i and v_j and zero if not. There are $\binom{N}{2}$ possible edges in an undirected, simple graph on N vertices. Hence, we need $\binom{N}{2}$ indicators to describe it. In a random graph

the edge indicators are random variables and we can define a *dependence graph*, \mathcal{D} , that specifies the dependence structure between the edges. This is a non-random graph obtained in the following way:

1. Represent each edge indicator I_{ij} $i, j = 1 \dots N, i \neq j$, as a vertex.
2. Draw an edge between two vertices iff the corresponding indicators are dependent conditional on all other indicators (i.e. conditional on the rest of the graph \mathcal{G}).

This means that the vertices of \mathcal{D} are the possible edges of \mathcal{G} and the edges of \mathcal{D} are the pairs of edges in \mathcal{G} that are conditionally dependent. As an example, in the Bernoulli random graph the dependence graph is empty (i.e. it contains no edges), since all edges are independent.

The following theorem specifies the probability function of a general random graph \mathcal{G} with dependence structure \mathcal{D} . More formally, it defines a probability measure on the space of all random graphs on N vertices with dependence graph \mathcal{D} . At least theoretically, this makes it possible for us to generate samples of random graphs with arbitrary dependencies between the edges. A proof of the theorem can be found in Preston (1974).

Theorem 6.1.1 *A given undirected random graph \mathcal{G} on N vertices with dependence structure \mathcal{D} has probability*

$$P(\mathcal{G}) = c^{-1} \exp\left\{\sum_{A \subseteq \mathcal{E}} \alpha(A)\right\}$$

where $\alpha(A)$ is an arbitrary constant if A is a clique of \mathcal{D} and $\alpha(A) = 0$ otherwise and c is a normalizing constant.

According to the theorem, the only properties that we are able to control in a random graph \mathcal{G} with dependence structure \mathcal{D} are the ones that are connected to the cliques of \mathcal{D} , i.e. a certain property in a random graph can be rewarded only if it is in some way reflected by the cliques in its dependence graph. That a certain property is rewarded means that an outcome \mathcal{G} is assigned a larger probability if it possesses this particular property than if it does not. In our social network model we want to be able to reward transitivity, i.e. we want to define a probability measure that is possible for us to manipulate so that graphs with a large amount of triangles are made more likely than graphs containing few triangles. It turns out that the *Markov random graph measure* fulfills this wish, at the same time as it

is based on a dependence structure that is very natural when interpreted in terms of friendship formations.

We will say that a random graph \mathcal{G} is a *Markov graph*, or has *Markov dependence*, if its dependence graph \mathcal{D} contains no edges between indicators I_{ij} and I_{kl} belonging to distinct pairs of vertices in \mathcal{G} , i.e. if non-incident edges in \mathcal{G} are conditionally independent. Thus, in a Markov graph, the cliques of \mathcal{D} correspond to sets of edges such that any pair of edges within the set must be incident. It is readily seen that the only sets in which this is the case are triangles and stars; that is, triangles

$$T_{v_i v_j v_k} = \{(v_i, v_j), (v_j, v_k), (v_k, v_i)\}$$

and k -stars

$$S_{v_{i_0} \dots v_{i_k}} = \{(v_{i_0}, v_{i_l}); l = 1 \dots k\}, \quad k = 1 \dots N - 1.$$

To state Theorem 6.1.1 in the special case when \mathcal{G} is a Markov graph, let us introduce the notation

$$\begin{aligned} \alpha(T_{v_i v_j v_k}) &= \xi_{ijk} \\ \alpha(S_{v_{i_0} \dots v_{i_k}}) &= \sigma_{i_0 \dots i_k}. \end{aligned}$$

To simplify the model and reduce the number of parameters we impose a homogeneity condition: we assume that all isomorphic graphs have the same probability. This means that we are assuming the vertices to be a priori indistinguishable, and we do not define parameters specific to the different vertices. In effect, instead of introducing $\binom{N}{3}$ different triangle parameters (one for each possible triangle) we introduce one single parameter controlling the total number of triangles in the graph and, analogously, for each k we introduce one single parameter controlling the total number of k -stars in the graph. Thus

$$\begin{aligned} \alpha(T_{v_i v_j v_k}) &= \xi_{ijk} = \xi \\ \alpha(S_{v_{i_0} \dots v_{i_k}}) &= \sigma_{i_0 \dots i_k} = \sigma_k, \quad k = 1 \dots N - 1. \end{aligned}$$

The probability function of a Markov random graph is now obtained as a simple corollary to Theorem 6.1.1:

Corollary 6.1.1 *A given homogeneous undirected Markov graph has probability*

$$P(\mathcal{G}) = c^{-1} \exp\{\xi t(\mathcal{G}) + \sum_{k=1}^{N-1} \sigma_k s_k(\mathcal{G})\} \quad (6.1)$$

where $t(\mathcal{G})$ is the number of triangles in \mathcal{G} and $s_k(\mathcal{G})$ is the number of k -stars in \mathcal{G} .

The star parameters σ_k are hard to interpret jointly. For example, every k -star contains $\binom{k}{j}$ j -stars for all $j < k$. To get rid of this multiple counting we replace $s_k(\mathcal{G})$ by $d_k(\mathcal{G})$ where $d_k(\mathcal{G})$ is defined as the number of vertices of degree k in \mathcal{G} . We have

$$s_k(\mathcal{G}) = \sum_{j \geq k} \binom{j}{k} d_k(\mathcal{G}).$$

Hence, if we introduce new parameters

$$\delta_j = \sum_{k \leq j} \binom{j}{k} \sigma_k$$

the measure (6.1) can be written as

$$P(\mathcal{G}) = c^{-1} \exp\{\xi t(\mathcal{G}) + \sum_{j=1}^{N-1} \delta_j d_j(\mathcal{G})\}. \quad (6.2)$$

Samples from this probability distribution can be generated using MCMC-techniques.

6.2 Properties of the model

The Markov random graph model is based on a dependence graph in which edges are independent if they do not share a node. This is a dependence structure that is very suitable to describe a social network. Interpreted in terms of friendship formations it stipulates that the social behaviour of an individual should be influenced only by her friends, not by individuals to whom she does not have any kind of relation. Furthermore, the model contains a parameter ξ that makes it possible for us to control the number of triangles in the graph: by choosing $\xi > 0$ we obtain a model with a bias towards transitivity. So far the Markov random graph appears to be a promising candidate as a model for a social network. But, as mentioned in the introduction to this section, this is flawed. Strauss (1986) shows that the model is degenerate in the sense that if ξ is strictly positive (however small) then, as the number of vertices increases, the probability tends to one that an arbitrarily large fraction of the edges will coalesce into a clique,

i.e. a complete subset of the graph. In terms of a social network this means that in a large population all ties between the individuals will, with large probability, be collected in a single giant friendship circle in which everybody is acquainted to everybody else. This unnatural asymptotic behaviour makes the Markov random graph very unsuitable as a model for the underlying social network in our model, since we are to a large extent concerned with precisely asymptotic results. The proof of the degeneracy is technical and not very instructive and it is therefore omitted. Instead we present some simulation results to illustrate the behaviour.

6.3 Simulations

Our aim here is to generate Markov random graphs with fixed average degree and show that if $\xi > 0$ almost all edges in the graph are coalesced in one single clique. As mentioned above it can be shown analytically that this is the case in large graphs but we will see that the behaviour occurs already in graphs of quite small order. To simplify the simulations we set $\delta_j = 0$ for all j , i.e. the model (6.2) is reduced to

$$P(\mathcal{G}) = c^{-1} \exp\{\xi t(\mathcal{G})\}. \quad (6.3)$$

Since our purpose is to use the graph as a model for a social network we condition on the average degree. This is indeed a natural thing to do when modelling a social network. It simply means that we fix the average number of friends of a member of the population. We note that conditioning on the average degree is roughly the same as conditioning on the number of edges: In a graph with N vertices and average degree d there are approximately $Nd/2$ edges and vice versa. Consequently, our first task is to construct a Markov graph with N vertices and $Nd/2$ edges. To achieve this we use a simplified version of the *Metropolis method*. A graph \mathcal{G}_0 is chosen uniformly from all graphs with N vertices and $Nd/2$ edges and a sequence $\{\mathcal{G}_k\}$ is generated inductively as follows:

1. At step k , pick at random a pair $(v_i, v_j) \in \mathcal{E}_{\mathcal{G}_k}$ and another pair $(v_{i'}, v_{j'}) \notin \mathcal{E}_{\mathcal{G}_k}$, i.e. pick two vertices, v_i and v_j , in \mathcal{G}_k with an edge between them and pick two vertices, $v_{i'}$ and $v_{j'}$, with no edge between them. Let \mathcal{G}' be the graph in which the edge between v_i and v_j is replaced by an edge between $v_{i'}$ and $v_{j'}$, i.e. $\mathcal{G}' = \mathcal{G}_k - (v_i, v_j) + (v_{i'}, v_{j'})$ with some abuse of notation.

2. Compute $\Delta t = t(\mathcal{G}_k) - t(\mathcal{G}')$, where $t(\mathcal{G})$ is defined above as the number of triangles in a graph \mathcal{G} .
3. Now pick \mathcal{G}_{k+1} according to the following rules:
 - If $\xi \Delta t \leq 0$; set $\mathcal{G}_{k+1} = \mathcal{G}'$.
 - If $\xi \Delta t > 0$; set $\mathcal{G}_{k+1} = \mathcal{G}'$ with probability $e^{-\xi \Delta t}$ and set $\mathcal{G}_{k+1} = \mathcal{G}_k$ with probability $1 - e^{-\xi \Delta t}$.

It may be seen that this procedure generates a Markov chain where the stationary distribution is obtained from (6.3) by conditioning on the number of edges, i.e.

$$\mathcal{L}(\mathcal{G}_k) \Rightarrow P(\mathcal{G}) \mid \{|\mathcal{E}_{\mathcal{G}}| = Nd/2\} \quad \text{as } k \rightarrow \infty.$$

Figure 6.1 below shows plots of the expected number of triangles in the graph \mathcal{G} , $E[T(\mathcal{G})]$, against ξ for different values of N (the reason for not studying N larger than 30 is that the convergence of the algorithm then turns out to be very slow). The average degree d is set to 5. When calculating $E[T(\mathcal{G})]$ we have used the fact that under certain regularity conditions the time average in a Markov chain converges almost surely to the expected value, i.e.

$$\frac{1}{n} \sum_{k=1}^n T(\mathcal{G}_k) \rightarrow E[T(\mathcal{G})] \quad \text{a.s. as } n \rightarrow \infty.$$

It is readily seen that in a graph with a fixed number of edges the largest possible number of triangles is achieved if all edges are collected in one single clique. If the number of edges is $Nd/2$ the size of the maximal clique, k_{max} , can be obtained from the relation $\binom{k_{max}}{2} = Nd/2$. Approximately we have $k_{max} = \sqrt{Nd}$ and hence the maximal number of triangles in the graph is $\binom{\sqrt{Nd}}{3}$. The degenerate behaviour of the Markov graph is illustrated in that the number of triangles in the graphs are approaching the maximal levels (the dashed lines in Figure 6.1) as soon as $\xi > 0$, i.e. almost all edges are collected in one single clique in a Markov graph with positive ξ .

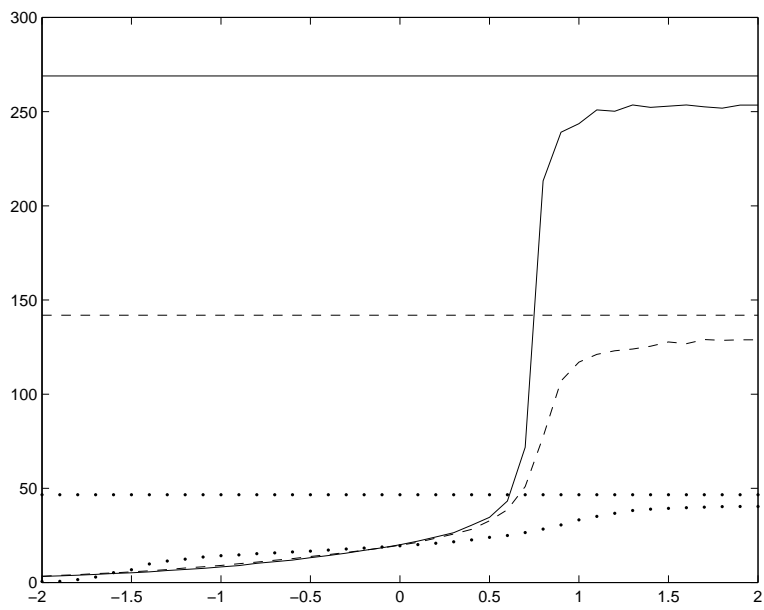


Figure 6.1: Expected and maximal number of triangles against ξ in a Markov graph with 10, 20 and 30 vertices (dotted, dashed and solid line, respectively) and average degree 5.

Chapter 7

The household model

The contact structure in a human population often contain several overlapping systems of small social groups with a high level of mixing. A few examples are households, schools and work places. The spread of infection is usually greatly facilitated by the presence of these groups and hence it is desirable to take them into account when modelling the social network. In this chapter we describe the so called household model in which the social network is modelled as the union of a small-group structure, representing e.g. the system of households in the population, and a global structure, representing e.g. the formation of friendships. A thorough treatment of the model can be found in Ball, Mollison and Scalia-Tomba (1997). The graph approach developed here is borrowed from Andersson (1998).

7.1 Description of the model

First we describe the underlying network in the household model. Given a set of N labelled vertices, let \mathcal{G}_1 be a graph consisting of N/k disjoint k -cliques, where $k \ll N$, and let \mathcal{G}_2 be a Bernoulli graph with edge probability r , i.e. $\mathcal{G}_2 \in \mathcal{G}(N, r)$. The social network \mathcal{G} is formed as the superposition of the graphs \mathcal{G}_1 and \mathcal{G}_2 , i.e. $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$. To keep the neighborhood size in \mathcal{G} bounded we have to scale the edge probability in \mathcal{G}_2 by N , i.e. we have to put $r = \lambda/N$ for some $\lambda > 0$.

The graph \mathcal{G}_1 divides the population into small mutually exclusive groups representing for example households, while the graph \mathcal{G}_2 is meant to describe the friendship relations in the population. When considered separately the household structure \mathcal{G}_1 does not allow any disease spread worth mentioning, since there are no connections between the households. However, when the

friendship structure \mathcal{G}_2 is added the households are linked together and a different situation arises. The disease can now spread between the households by aid of friendship connections. Observe that in the above setting we assume all households to be of the same size. The case of unequal household sizes can be treated similarly but the notation becomes cumbersome (Ball *et al*).

7.2 Epidemic behaviour

Basic reproduction number

Just as when the social network was modelled as a Bernoulli graph/3-clique graph we aim at using a branching process approach to find an expression for the basic reproduction number. To achieve this we will use a *clumped Reed-Frost model*. First we consider only local infection links, i.e. infection links using edges in the household structure \mathcal{G}_1 . An infection link between two members, i and j , in a household should be interpreted so that *if* $i(j)$ ever becomes infected then she will bring the disease on to $j(i)$. The infection links partition each household into a number of connected components - clumps - each with the property that as soon as one member of it is infected the entire clump is infected. Assuming that the number of clumps is large the early stages of the epidemic can be approximated by a branching process in which individuals correspond to clumps and the offspring of a given clump are the components that are directly infected by aid of global infection links (i.e. infection links using edges in the friendship structure \mathcal{G}_2) from infectives in that clump. The reproduction mean in the branching process is equal to the average number of clumps that is infected by a given infective clump in the beginning of the time course. To find an expression for the reproduction mean, let H denote the clump size, i.e. let H denote the final size of a Reed-Frost epidemic in a group of k individuals, where we start by infecting one single individual and where the infection probability is p . We will make an exception from the rule that initially infected individuals are not included in the final size of an epidemic and include also the initial infective in H . If the population size is large the number of neighbors of a given individual in the global structure \mathcal{G}_2 is approximately Poisson (λ) distributed and, using the fact that $k \ll N$, the neighbors belong with large probability to distinct clumps. This implies that in a large susceptible population the average number of new clumps that is infected by each infective in a local epidemic (i.e. an epidemic within a household) equals λp . Hence the reproduction mean in the branching process is equal to $E[H]\lambda p$, since the average size

of a local epidemic is equal to $E[H]$. According to Theorem 2.1 there is a positive probability for the branching process to explode if and only if $E[H]\lambda p > 1$. Thus a good candidate for the basic reproduction number is given by

$$R_0 = E[H]\lambda p,$$

since explosion in the branching process is equivalent to a major outbreak in the epidemic. Closed expressions for $E[H]$ does not exist. It is possible though to derive a triangular linear system of equations for $\mathbf{P}_1^k = (P_{10}^k, P_{11}^k, \dots, P_{1k}^k)$ where P_{ij}^k is the probability that j individuals (including the initial infectives) are ultimately infected in a Reed-Frost epidemic in a group of k individuals with i initial infectives and infection probability p (Andersson and Britton (1999)). This enables us to calculate $E[H]$.

If $k = 1$ the household graph \mathcal{G}_1 is empty and the social network is reduced to a Bernoulli graph, implying that $R_0 = \lambda p$. If $k > 0$ we have $E[H] > 1$ and $R_0 = E[H]\lambda p > \lambda p$. Hence the formation of households have an "amplification" effect on the spread of infection in that the basic reproduction number is increased from the individual-to-individual value to a larger clump-to-clump value.

Final size equation

If the number of initial infectives is large, i.e. if $\mu > 0$, the final size equation for the household model is complicated to derive. Hence, in what follows we will assume that $\mu = 0$. Remember that the final size of an epidemic, τ , is defined as the proportion of the initially susceptible population that is ultimately infected. If we assume that $\mu = 0$ though, this proportion is the same as the proportion of the entire population that is ultimately infected, since asymptotically the proportion of initially infected individuals is zero. Here we will derive an equation for the asymptotic final size by considering the average proportion of the members in a given household that is ultimately infected. Let q_j denote the asymptotic proportion of households in which j individuals are ultimately infected. Then the average number of individuals that are ultimately infected in a given household is asymptotically given by $\sum_{j=0}^k j q_j$. Hence

$$\tau = \frac{1}{k} \sum_{j=0}^k j q_j. \quad (7.1)$$

To find an equation for q_j we note that since $\mu = 0$ the asymptotic proportion of the households that contains any initially infected individuals is zero, i.e.

asymptotically all individuals in a given household are initially susceptible with probability one. This implies that

$$q_j = \sum_{i=0}^j P(X = i) P_{ij}^k \quad (7.2)$$

where X is the number of individuals in a given household that are infected from outside (i.e. by neighbors in the friendship graph \mathcal{G}_2) and P_{ij}^k is defined as above. The number of individuals in a given household that are asymptotically infected from outside is binomially distributed with parameters k and p_{inf} where p_{inf} denotes the asymptotic probability that a given individual is infected from outside. To find an expression for p_{inf} we note that the probability to escape global infection from a given member in the population is equal to $1 - \lambda p/N$. To escape the entire epidemic one has to escape infection from all the $N\tau$ ultimately infected individuals. Hence

$$1 - p_{inf} = (1 - \lambda p/N)^{N\tau} \longrightarrow e^{-\lambda p\tau} \quad \text{as } N \rightarrow \infty.$$

Thus

$$P(X = i) = \binom{k}{i} (1 - e^{-\lambda p\tau})^i (e^{-\lambda p\tau})^{k-i}.$$

Substituting this into equation (7.2) yields

$$q_j = \sum_{i=0}^j \binom{k}{i} (1 - e^{-\lambda p\tau})^i (e^{-\lambda p\tau})^{k-i} P_{ij}^k. \quad (7.3)$$

Equations (7.1) and (7.3) together yield an implicit equation for τ when $\mu = 0$. For the general case we refer to Ball *et al.*

7.3 Properties of the model

The household network is constructed to catch the formation of small social groups with complete mixing. In the above setting only one small-group structure, meant to describe the system of families, is included but the model can be generalized to contain several structures in which the groups correspond to for example school classes and work places (Andersson (1998)). The networks presented previously in this work focuses exclusively on global friendship relations. The fact that also local group formations are taken into account in the household network must be regarded as a great advantage with the model, especially since the groups turn out to have a significant

effect on the spread of infection. The friendship structure in the household model is modelled as a Bernoulli graph where the edge probability is scaled by the population size and in this fact lies the main disadvantage with the model. As explained in Section 4.3 the transitivity in a Bernoulli $(N, \lambda/N)$ graph is asymptotically zero, implying that in a large population there is practically no transitivity between the households in the household network. As indicated above it is a quite complicated task to derive results about an epidemic among households already when the friendship structure is chosen to be the simplest possible, i.e. a Bernoulli graph. However, if one wants to increase the applicability of the model it is necessary to combine the household structure with more realistic friendship structures.

Chapter 8

Small-world networks

The experience of meeting a complete stranger and finding that you share a mutual acquaintance, is one with which many of us are familiar - "it is a small world!" we say. A few years ago this phenomenon gave rise to the concept of Six Degrees of Separation, which is based on the notion that everyone in the world is connected to everyone else through a chain of at most six mutual acquaintances. If two people have one mutual acquaintance, then they are said to have one degree of separation. The estimate of six degrees of separation can be understood heuristically from the following calculation: Suppose that a person has 25 friends and each of them have 25 new friends and so on. Then in seven steps this person would be connected to six billion people. Thus, assuming there are six billion people on the Earth, seven connections or six degrees of separation are enough to link any two people together. There is of course a big mistake in this calculation: we are assuming that none of a persons friends know each other. In reality this is rarely the case, since friendship circles are often strongly overlapping. Indeed the concept of Six Degrees of Separation should be viewed as somewhat anecdotal, but the underlying message should be taken seriously: The length of the shortest friendship chain connecting two people in a human friendship structure is often surprisingly short, even in very large populations.

The *small-world networks* were introduced by Strogatz and Watts (1998) as an attempt to create a class of highly transitive graphs with short path lengths. The coexistence of high transitivity and short paths between the vertices in a social network graph will be referred to as the *small-world phenomenon*. This concept is explained in greater detail and made more formal in the beginning of the section. We then describe a construction procedure aimed at producing networks exhibiting the small-world phenomenon. In

order to facilitate analytical treatment of the model this procedure differs slightly from the one suggested by Strogatz and Watts. The behaviour of the resulting network is illustrated by aid of computer simulations. Finally we present a few results about the epidemic spread along small-world networks.

8.1 The small-world phenomenon

As encountered in the introduction to this section the paths between people in a network representing friendship relations in a human population are often surprisingly short, that is; almost every person in the population is somehow close to almost every other person, even those that are perceived as likely to be far away (Milgram (1967), Kochen (1989)). The world is small. Let us list a few reasons why this should be surprising in the first place:

1. The network is numerically large. In this work the population typically consist of all inhabitants in a country or part of a country, implying that the population size N is a very large number.
2. The network is sparse in the sense that the average degree of a vertex is very small compared to the number of vertices, that is; the average number of friends of a member in the population is small compared to the population size.
3. The network is decentralized in that it contains no central vertex to which most other vertices are adjacent. Indeed, the set of acquaintances of even the most socially active person in the population involves only a tiny fraction of the entire population. This implies that not only the average degree, but also the maximal degree over all vertices is much smaller than the number of vertices N .

All these properties concurrent to make the short paths in social networks remarkable. However, if the edges were drawn independently (as in a Bernoulli graph for example) then it follows from random graph theory (Bollobas (1985)) that most vertices would be only a few degrees of separation apart even for very large N . In a good model for a social network though, the edges are not independent. On the opposite, most friendship circles are strongly overlapping, i.e. we expect many of our friends to be friends also of each other. To summarize, we are looking for a graph that satisfies the three criteria above and at the same time possesses both transitivity and

short path lengths. Such a graph will be said to exhibit the small-world phenomenon. Let us introduce two graph characteristics to help us make this concept more precise:

- *The c -path length fraction, L_c .* For a given graph of order N , let S_c be the number of pairs of vertices between which the shortest path consists of more than c edges, where $0 \leq c \leq N$, $c \in \mathbf{N}$, and define

$$L_c = \frac{S_c}{\binom{N}{2}}$$

i.e. L_c is the share of all pairs of vertices in the graph in which it takes more than c steps to reach from one vertex to the other.

In this section the focus will be on investigating how the features of a graph of given order N and with fixed average degree changes as certain parameters in the construction procedure described below varies. This implies that the exact value of L_c is not our main interest, but rather the changes in it as we manipulate the structure of the graph. Hence, the index c can be picked in a quite arbitrary way as long as we avoid obviously unsuitable choices. Setting $c = 1$ or $c = 500$ in a graph with 1000 vertices and average degree 10 would for example cause us to miss most changes in the structure, since L_1 is approximately 1 and L_{500} is approximately 0 for all possible constellations of the edges. However, if one wants to be able to use the c -path length fraction to compare graphs of different order it is intuitively clear that c has to be chosen as some increasing function of N . Indeed, the length of the shortest friendship chain connecting two arbitrary residents in a city is likely to be larger if the city has 100 000 inhabitants than if it has 10 000 inhabitants. The order of the growth is probably highly dependent on N though: The change in path length is presumably larger if we compare a city with 10 000 inhabitants and one with 100 000 inhabitants, than if we compare a city with one million inhabitants and one with ten million inhabitants. In this work we will set $c = \lfloor \log N \rfloor$. This is a subjective choice made to catch some of the features mentioned above. To simplify notation we will drop the index and write $L_{\lfloor \log N \rfloor} = L$.

- *The clustering coefficient, C .* Suppose that the degree of a vertex v_i in a given graph is equal to k , i.e. suppose that individual i has k neighbors. Then, if $k \geq 2$, at most $\binom{k}{2}$ edges can exist between these neighbors (this occurs when every neighbor of i is connected to every other neighbor of i). Let E_{v_i} denote the number of these possible edges that actually exist and

define the local clustering of vertex v_i as

$$C_{v_i} = \begin{cases} \frac{E_{v_i}}{\binom{k}{2}} & \text{if } k \geq 2 \\ \frac{|\mathcal{E}| - k}{\binom{N}{2} - (N-1)} & \text{if } k < 2. \end{cases}$$

(The definition of C_{v_i} for $k < 2$ might seem a bit awkward at first sight, but it is necessary to get around the division-by-zero problem and can be interpreted as the overall clustering in the rest of the graph.) Now define the clustering coefficient in the graph, C , as the average clustering over all vertices i.e.

$$C = \frac{1}{N} \sum_{i=1}^N C_{v_i}.$$

In a friendship network C_{v_i} reflects the extent to which friends of i are also friends of each other, and thus C measures the cliquishness of a typical friendship circle. Equivalently C can be regarded as a measure of the transitivity in the network.

Stated in terms of these quantities a small-world network is a graph with large clustering coefficient and small path length fraction. In order to decide what values for C and L that counts as large/small it is necessary to determine the ranges over which C and L can vary. Clearly, the largest value that C can attain is $C = 1$ for a complete graph and the smallest conceivable value of C is $C = 0$ for an empty graph. These two graphs also have extremal path lengths. However, this is not a very instructive comparison, since it is obvious that clustering and path length will change as more edges are added to any graph. As mentioned above, the approach here will be to study how the properties of a graph with a fixed number of edges are affected as the distribution of the edges over the graph is changed. Our aim is to find a constellation that causes C to be large and L to be small, that is; large/small relative to the largest/smallest possible values in a graph with this particular number of edges.

8.2 Construction of the network

To begin with we arrange the N vertices in a circle. Let \mathcal{G}_1 be a graph constructed by connecting each vertex to each of its k nearest neighbors in a clockwise sense, one at a time, with probability r . We move clockwise around the circle, considering one vertex in turn until one lap is completed. This yields a graph in which each vertex is connected to each of its k nearest

neighbors in both directions on the circle with probability r . Let \mathcal{G}_2 be a Bernoulli graph with edge probability λ/N , constructed independently of \mathcal{G}_1 . The network \mathcal{G} is obtained as the superposition of \mathcal{G}_1 and \mathcal{G}_2 , i.e. $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$. We will assume k to be much smaller than the order of the graph, N . This implies that in a large population the probability that any of the edges drawn from vertex v_i as a part of \mathcal{G}_1 is also included in \mathcal{G}_2 is very small. Hence, for large N the average degree d in \mathcal{G} is equal to the sum of the average degrees in \mathcal{G}_1 and \mathcal{G}_2 , i.e. $d = 2kr + \lambda$.

In the above construction the graph \mathcal{G}_1 is meant to represent the local friendship formations, for example we can think of the social network obtained as people get acquainted to people living in the same area. To generate this local structure, for each vertex we need a definition of which vertices are the possible local neighbors. We have chosen to arrange the vertices in a circle and stipulate that the possible neighbors of a vertex are the k nearest vertices in both directions on the circle. The circle representation can be motivated from the fact that it exhibits minimal structure, in that no vertices are to be identified as special in any way. The graph \mathcal{G}_2 represents global contacts, i.e. friendship connections between people who do not have any natural ties between them. Two strangers from different parts of a country that meet and make friends with each other is an example of a situation meant to be described by an edge in \mathcal{G}_2 .

8.3 Tuning the parameters

Can a graph constructed according to the procedure described above be made to exhibit the small-world phenomenon? That is, can the parameters r and λ be chosen in such a way that high clustering and short path lengths occur at the same time? To investigate this we set the average degree in the graph equal to $2k$, i.e.

$$2kr + \lambda = 2k. \tag{8.1}$$

This implies that the number of edges in the graph is approximately equal to Nk . Our task is to bring about the coexistence of high clustering and short paths between the vertices by distributing the edges rightly over the graph. To be more precise, we want to find r and λ satisfying equation (8.1), such that both C is large and L small relative to the largest/smallest possible values in a graph with Nk edges. We will rely on computer simulations to show that it is indeed possible to find such parameter values, i.e. the answer to the question addressed in the beginning of this section is positive.

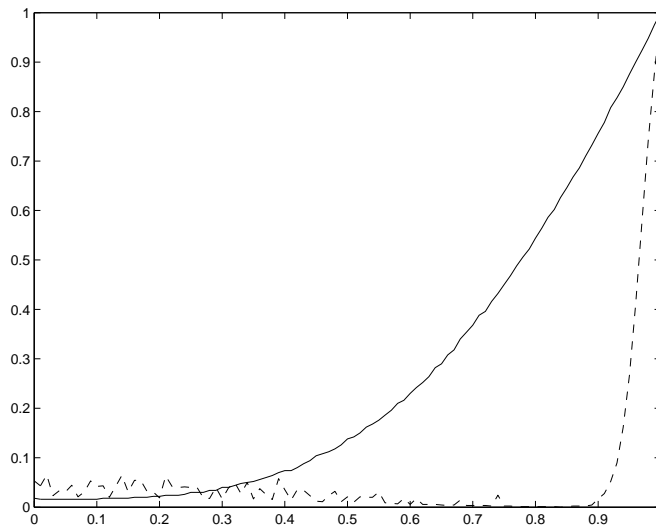


Figure 8.1: Clustering coefficient (solid line) and path length fraction (dashed line), scaled by their values at $r = 1$, against r .

The data shown in Figure 8.1 below are averages over 100 realizations of the construction procedure described in the previous section for different values of r . For each r the global parameter λ can be obtained from (8.1). The graphs has $N = 500$ vertices and average degree 4, i.e. $k = 2$. As r decreases we can see a rapid drop in $L(r)$ followed by an interval for r where $L(r)$ is small and $C(r)$ large at the same time. For r in this interval (say $0,8 < r < 0,9$) the graphs exhibit the small-world phenomenon. As r is decreased further $C(r)$ decreases to its low value for the pure Bernoulli graph and the small world behaviour is destroyed. The result implies that introducing only a few global short cuts in the graph causes the path lengths to drop quite drastically while the clustering remains essentially unchanged, i.e. the transition into a small world is almost undetectable at the local level.

8.4 Epidemic behaviour

Basic reproduction number

We have not succeeded in finding a general formula for the basic reproduction number in the small world model. Here we give an expression in the special case when $k = 1$. Unfortunately the formula cannot be generalized

in any obvious way. For $k \geq 2$ the derivations become cumbersome resulting in extensive expressions.

When $k = 1$ the local network \mathcal{G}_1 is constructed by connecting each vertex to its clockwise nearest neighbor with probability r . To find the basic reproduction number we will use the same technique as for the Household model. First we consider only local infection links, i.e. infection links using edges in \mathcal{G}_1 . These links partition the population in a number of connected components. In a large population the components will be small compared to the population size, implying that the global neighbors of the members in a component with large probability belong to distinct components. Hence, the initial stage of the epidemic can be described by a branching process in which individuals correspond to components. Using Theorem 2.3.1 the basic reproduction number is obtained as the reproduction mean in this process. To derive the reproduction mean we fix a vertex v_i in the circle. The number of vertices in each direction on the circle that can be reached from v_i by aid of local infection links is geometrically distributed with parameter $1 - pr$, since each infection link is present with probability pr . Hence, if individual i is infected she will generate on the average $pr/(1 - pr)$ new cases in each direction, implying that the average size of a local epidemic is equal to $2pr/(1 - pr) + 1$. If the population is large the number of global neighbors of a given individual is approximately Poisson (λ) distributed and the neighbors belong with large probability to distinct local components. Thus, in a large susceptible population the average number of components that is infected by each infective in a local epidemic equals λp . Hence,

$$R_0 = \left(2 \frac{pr}{1 - pr} + 1 \right) \lambda p.$$

Final size

We have not put in any effort to derive exact equations for the final size of a small world epidemic. Such equations might be possible to find for small k . For the graph to provide a realistic model for a social network though, at least we have to pick $k \geq 10$. Hence, final size equations for $k = 1$ and $k = 2$ are not very interesting from a practical point of view. For larger k the equations are presumably complicated to derive. Here we present a few simulations of the final size.

Figure 8.2 shows the mean size of the largest component in the infection graph as a function of the local contact probability r (the global parameter λ can be obtained from (8.1)). This should be interpreted as the worst

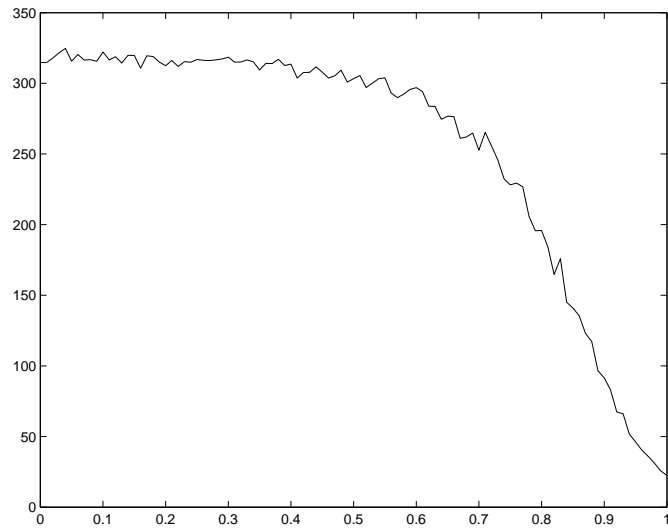


Figure 8.2: Mean size of the largest component in the infection graph against r for $p = 0,4$.

possible epidemic if we introduce one initial infective in the population. The graph has $N = 500$ vertices and average degree 4, i.e. $k = 2$. The infection probability p is equal to 0,4. The graph reveals that the final size increases as r decreases, the increase being very rapid in the interval $0,8 < r < 0,9$ where the small-world behaviour sets in.

Chapter 9

Epilogue

The purpose of this work has been to present a number of random graph constructions and investigate their impact on the epidemic spread when used as models for the underlying social network in a heterogeneous Reed-Frost model. When interpreted as models for social networks all graphs that we have considered are impaired by various drawbacks, in that they fail to catch some of the features present in the human contact pattern. The simplest construction, the Bernoulli random graph, exhibits asymptotically no transitivity, i.e. it contains asymptotically very few triangles. This disqualifies it as a model for a social network, since transitivity is maybe the most characteristic property with friendship formations. In the 3-clique graph the amount of triangles is increased but the model contains a slightly unnatural dependence structure between the edges. The Markov random graph possesses both transitivity and a nice dependence structure and it seems to be a promising candidate at first. But it turns out that the edges are distributed in an unnatural way over the graph, severely decreasing its applicability as a social network model. The Household model incorporates the formation of small social groups with complete mixing. In the present formulation of the model though, there is no transitivity between the households, which of course would be desirable. The last graph model presented is the class of small-world networks. These have turned out to be good models for social networks. For example one has been able to reproduce in a satisfactory way the structure of certain real-life social networks by aid of small-world simulations. However, although we have introduced a probabilistically more simple construction procedure than the one originally suggested by Strogatz and Watts it seems to be a difficult task to perform exact calculations within the model.

Random graphs as models for social networks is a quite new idea in mathematical epidemic modelling. The models that have received most attention so far from a mathematical point of view are the Bernoulli random graph model and the Household model. This can be explained from the fact that they are rather easy to analyze by aid of analytical methods. The 3-clique graph has not been used previously in epidemic modelling although it shares the simplicity in construction of the Bernoulli graph. The Markov graph is included here since it is very popular among social network people. In light of its degenerate asymptotic behaviour though, we feel that it is not worth putting to much effort into this model. The small-world networks also originate from the sociologists and hopefully this model can be developed further in the future. A few examples of possible social network models not included in this work are random graphs with prescribed degrees (Andersson (1998)) and random intersection graphs (Karonski, Scheinerman and Singer-Cohen (1999)).

If assigned to choose between the models presented here we would at present suggest the 3-clique graph. It is indeed true that it includes a somewhat unnatural dependence structure between the edges, but it certainly contains a satisfactory amount of triangles and it does not exhibit the kind of degenerate behavior as does the Markov graph. Furthermore, it is a quite simple construction that makes it possible for us to derive exact expressions for several epidemiological quantities and this must be regarded as a great advantage with the model. The Household graph should also be kept in mind as a possible candidate as a social network model. It is desirable though to develop the model further so that the transitivity between the households is increased, that is; the Bernoulli graph representing the global contact structure should if possible be replaced by a graph that contains a larger amount of triangles. Finally, the small-world networks are of course very interesting, since they have turned out to be such good descriptions of social networks. In the present formulation though, the model is cumbersome to deal with analytically. It remains an open problem to find a way of constructing small-world networks that lends themselves better to mathematical analysis.

References

- M. Altmann (1996): Network measures for Epidemiology, in *Models for infectious human diseases: Their structure and relation to data*, V. Isham and G. Medley, Eds., Cambridge University Press.
- H. Andersson (1998): Epidemic models on graphs and lattices, *Research Report No.204*, Stockholm University, Department of Mathematical Statistics.
- H. Andersson and T. Britton (1999): Stochastic epidemic models, *U.U.D.M. Lecture Notes 1999:2*, Uppsala University, Department of Mathematics.
- B. von Bahr and A. Martin-Löf (1980): Threshold limit theorems for some epidemic processes, *Adv. Appl. Prob.* **12**, 319-349.
- N. Bailey (1975): *The mathematical theory of infectious diseases*, Griffin, London.
- F. Ball, D. Mollison and G. Scalia-Tomba (1997): Epidemics with two levels of mixing, *Ann. Appl. Prob.* **7**, 46-89.
- A. Barbour and D. Mollison (1990): Epidemics and random graphs, in *Stochastic processes in epidemic theory*, J. Gabriel, C. Lefevre and P. Picard, Eds., *Lecture notes in Biomathematics* **86**, 86-89.
- B. Bollobas (1995): *Random graphs*, Academic Press, New York.
- K. Faust and S. Wasserman (1994): *Social network analysis*, Cambridge University Press.
- O. Frank and D. Strauss (1986): Markov graphs, *J. Amer. Stat. Assoc.* **81**, 832-842.
- J. Jonasson (1997): The random triangel model, Preprint.
- M. Karlberg (1997): Triad count estimation and transitivity testing in graphs and digraphs, Stockholm University Ph.D. thesis, Department of Statistics.

- M. Karonski, E. Scheinerman and K. Singer-Cohen (1999): On random intersection graphs, *Comb. Prob. Comp.* **8**, 131-159.
- M. Kochen (1989): Toward structural sociodynamics, in *The Small World*, edited by M. Kochen, Norwood, Ablex, 52-64.
- M. Kretzschmar and M. Morris (1996): Measures of concurrency in networks and the spread of infectious disease, *Math. Bio.* **133**, 165-195.
- S. Milgram (1967): The small world problem, *Psychology Today* **2**, 60-67.
- C. Preston (1974): *Gibbs states on countable sets*, Cambridge University Press.
- G. Scalia-Tomba (1985): Asymptotic final size distribution for some chain-binomial processes, *Adv. Appl. Prob.* **17**, 477-495.
- G. Scalia-Tomba (1990): On the asymptotic final size distribution of epidemics in heterogeneous populations, in *Stochastic processes in epidemic theory*, J. Gabriel, C. Lefevre and P. Picard, Eds, *Lecture notes in Biomathematics* **86**, 189-196.
- D. Strauss (1985): On a general class of models for interaction, *SIAM Review* **28**, 513-527.
- S. Strogatz and D. Watts (1998): Collective dynamics of small-world networks, *Letters to nature* **393**, 440-442.