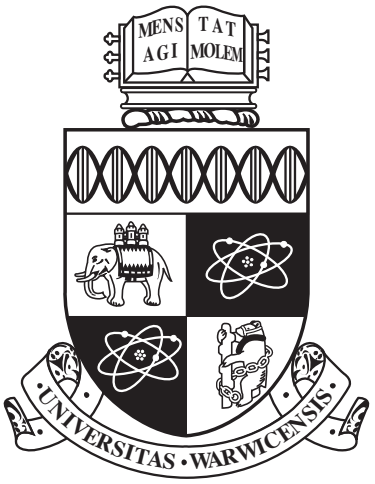


The impact of heterogeneity in contact structure on the spread of infectious diseases

by

Matthew Graham

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Complexity Science

Complexity Science Doctoral Training Centre

December 2014

THE UNIVERSITY OF
WARWICK

Contents

List of Figures

Acknowledgments

The research detailed in this thesis was funded by the Engineering and Physical Sciences Research Council (EPSRC) via the Complexity Science Doctoral Training Centre at the University of Warwick. Additionally two visits were made to the Network Dynamics and Simulation Science Laboratory at Virginia Tech, which was made possible by funding supplied by this hospitable and helpful group.

During my time of study at Warwick many people have provided me with support. First and foremost is my supervisor for this work, Thomas House, who has a seemingly inexhaustible well of both knowledge and patience. I also thank the staff and administration of the Complexity Science DTC for their help in all matters, and the WIDER group at Warwick for interesting discussion.

I was privileged to be surrounded by many generous, interesting and friendly people in this department who increased my enjoyment of research considerably. These include, Peter Dawson, Chris Oates, Ben Collyer, Anthony Woolcock, Sergio Morales, Anas Rana, Ellen Webborn, Marcus Ong, Dan Sprague, Mike Irvine, Yu-Xi Chau, Davide Michieletto, Dario Papavassiliou and Tom Machon. I thank Anthony Woolcock and Adrienne Davies, Ben Collyer and Marcus Ong, Peter Dawson and Jen Lawson along with Chris and Lucy Oates for their hospitality on numerous occasions and Davide Michieletto, Tom Machon, Anthony Woolcock, James Porter, Peter Dawson and Chris Oates for helping me hone my squash skills and to a much lesser extent my tennis skills. I also owe a large debt of gratitude to Justin Lessler for employment over the last few months of my PhD write-up.

I thank my family for their support and for not asking how the thesis writing is going too frequently. Finally I thank Hannah, whose cooking has improved beyond measure due to all the occasions on which I've left it to her to feed me, along with her unwavering help, support and friendship.

Declarations

Parts of this thesis have been published elsewhere:

- M Graham & T House (2012), "Dynamics of stochastic epidemics on heterogeneous networks". *J. Math. Biol.*

The presented thesis is MG's own work, except where research involving collaboration is concerned. In these circumstances, MG's contributions are indicated. This theses has not been submitted for a degree at any other university.

Abstract

Contact structure between individuals in a population has a large impact on the spread of an epidemic within this population. Many techniques and models are used to investigate this, from heterogeneous age-age mixing matrices to the use of network models in order to quantify the heterogeneity in the populations contacts.

For many diseases, the probability of infection per contact, along with the exact contact structure are unknown, compounding the difficulty of identifying accurate contact structures.

In this thesis, the impact that the contact structure has on the epidemic is examined in several different ways. Analytical expressions for the variance in the spread of an epidemic in its early exponential growth phase on heterogeneous networks are derived, showing that the third moment of the degree distribution is needed to fully specify this variance. This quantifies the impact that very well connected individuals can have on the early spread of an epidemic through a network.

The dependency of the potential epidemic on the heterogeneity in workplace sizes and transmission rates is examined. It is shown that large workplaces can increase the expected size of the epidemic significantly, along with increasing the effectiveness of control strategies enacted during the early stages of an epidemic.

In addition to this, a synthetic population is constructed for England and Wales from available datasets, in an attempt to model the spread of an epidemic through a realistic network of comparable size to the true population. The contact structure that is derived from this is compared with that taken from two surveys of contact structure in the same population, using simple models, and qualitative differences are seen to exist between the surveyed structures and the synthetic population structure.

Chapter 1

Introduction

Since the work of Kermack and McKendrick [Kermack and McKendrick, 1927], the study of epidemics using mathematical modelling has become widespread. One of the most prominently studied models is of a disease which has three compartments or classes which individuals can pass through, namely: a susceptible class, S, containing those individuals who can contract the disease; the infectious class, I, who currently have the disease and can transmit it to those in the susceptible class; and the removed, R, which contains those who have previously contracted the infection and have now been removed from the dynamics of the epidemic, be that by recovery to immunity from the infection or by death. This type of model is often referred to as the SIR model for obvious reasons.

Following this example, a large amount of study of epidemics via the use of mean-field dynamics has been undertaken [Anderson and May, 1992]. For the mean-field dynamics, a key assumption is that all members of the population have equal probability of spreading an epidemic to any other member of the population at all times. Though a very useful assumption, in reality this is not what is observed, and the heterogeneity that is present in the contact structures of populations has a large impact on the spread of an epidemic [Keeling, 2005; Rohani et al., 2010] and has become the subject of numerous survey studies [Read et al., 2008; Mossong et al., 2008; Danon et al., 2012; Read et al., 2011, 2014]. Additionally, methods to infer contact structures through phylogenetic analysis of disease strains in populations have recently been developed [Volz et al., 2009; Leventhal et al., 2012; Frost and Volz, 2013].

The theme of heterogeneity in contact structure is the main focus of this thesis. In §3, the variance of early growth period of an epidemic on a heterogeneous network is considered. To investigate this, the neighbourhoods of susceptible and infected individuals are considered, and the epidemic is shown to be density dependent. The work of Kurtz [1970, 1971] relating to density dependent processes is then used to calculate the variance of the number of infected individuals during the early growth period. The theoretical expression that is derived is then compared with that seen from simulating epidemics on networks and a good agreement between the two is seen.

In §4 data detailing the workplace size distribution of the UK is considered. Workplaces are where a lot of heterogeneity in the number of contacts that people have is generated, as the number of contacts made in the home or in schools show much less variance than the workplace. How this distribution, coupled with transmission rates which are modified to alter the average infectivity of an individual in different sized workplaces, impacts the

spread of an epidemic is examined. This is achieved by fitting several different distributions to fit the workplace size distribution, which are then combined with the transmission rates along with the standard final size equation in a mean-field model to estimate the potential epidemics size for these fitted distributions. The overall and secondary attack rates, which give the overall proportion of at risk individuals who become infected, and the proportion of these infected by the initial infected respectively are also considered for the modified transmission rates. It is shown that for large workplaces sizes, the increased presence of which increase the predicted final size of the epidemic, coincide with the lowest secondary attack rates. This has implications for possible control methods, as many infections can be averted by acting early in this scenario.

In §5 the construction of a synthetic population, whose contact network represents England and Wales is described. This is a network model which has the same number of individuals as there are in England and Wales, which is constructed to align with several statistics taken from census data. There have been several similar studies in the last few years which have focused on the USA [Eubank et al., 2004] and Italy [Iozzi et al., 2010]. This involves bringing together many different data sources, such as census data, and diary style information to attempt to create a representative contact network for England and Wales. Once constructed, this can be used to compare the efficacy of possible intervention strategies on a national scale. Various measures of this synthetic populations contact structure are then compared with two contact structures derived from surveying the population of Great Britain. These two surveys are POLYMOD [Mossong et al., 2008] and the UK contact survey [Danon et al., 2012].

Finally in §6, simulations conducted using the synthetic population are compared with simulations using simpler models (pairwise approximation, who-acquires-infection-from-whom and meta-population models), where the contact structure is defined by data gathered from the synthetic population, along with POLYMOD and the UK contact survey. This is in order to quantify how different these contact structures are, along with what is gained by including so much detail in the full synthetic population.

Firstly in §2, we introduce basic ideas behind mathematical models of infectious disease spread, from deterministic mean-field models up to stochastic heterogeneous network models. This leads into §3 which involves theory of stochastic heterogeneous network models. Each subsequent chapter has its own introduction detailing the necessary background material.

Chapter 2

Background

2.1 General modelling techniques

The use of mathematical models to aid in understanding the spread of infectious diseases began in the 18th Century [Bernoulli, 1766], and has become an important tool in the study and prevention of epidemics. The beginnings of modern mathematical modelling of diseases can be seen to have begun with Kermack and McKendrick [1927]. Since this early investigation, there has been much study of epidemic models and they now take a variety of mathematical forms [Anderson and May, 1992; Keeling and Rohani, 2008], and are routinely used to inform policy on disease control and contribute towards public health plans [Ferguson et al., 2003; Riley et al., 2003; Tildesley et al., 2006; Baguelin et al., 2010].

The most popular modelling approach is to generate a set of differential equations which describe the infection process for a population, and then examine what the consequences of this model are. Individuals in the population are put into separate 'compartments', which describe the state of the infection within the individual in question. A member of the population will begin in a 'susceptible' state, and once exposed to infection will progress through a number of different compartments as time progresses. The most prominent example of this is the susceptible-infectious-removed (SIR) model, first formulated in Kermack and McKendrick [1927] and since described in Dietz [1967], Keeling and Rohani [2008], Bailey [1975] and many other texts.

In this model, at any point in time everyone is either susceptible to the disease, infectious with it, or removed from the future disease dynamics. This model is obviously a simplification of reality, but provides a useful starting point for modelling diseases where previous infection confers long-lasting immunity, for example outbreaks of childhood diseases like measles, some respiratory illnesses like pandemic influenza, and historical pathogens such as smallpox [Keeling and Rohani, 2008]. It is often assumed that the time scales of the infection and the epidemic are such that the population size will be the unaltered throughout the epidemic, with the exception of death caused by the infection, i.e. births and death by other causes are irrelevant.

In the simplest models, the dynamics are taken to be mean-field, meaning that the rate of encounters between susceptibles and infecteds is given by the proportion of the population who are in these compartments. This implies that any susceptible in the population has

an equal probability of contracting the infection at any given time, and that they can be infected by any infectious member of the population, meaning that members of the population are interchangeable. For the simplest, deterministic, form of this model, the governing differential equations are given by the following:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \ , \end{aligned} \tag{2.1}$$

where S, I and R are the proportion of individuals in the susceptible, infected and removed compartments respectively, $S + I + R = 1$, γ is the removal rate of infected individuals and βI is the rate at which the infection is passed on to each susceptible. This is an example of 'frequency dependent' transmission, as as the population size increases, the number of individuals infected by a randomly chosen individual will not increase. As we have $S + I + R = 1$, we can in practice work simply with equations for S and I, as the value of R will be defined by this relationship.

To see an increase in the number of infected individuals we require that $\mathrm{d}I/\mathrm{d}t > 0$, which implies that $\beta/\gamma > 1/S$. If we begin in an entirely susceptible population, then we require that $\beta/\gamma > 1$ in order to see an epidemic in the population. This number (β/γ) is referred to as the *basic reproduction number* and is denoted by R_0. This is equal to the average number of people that a typical individual infects in an entirely susceptible population.

There are many ways in which this simple model can be amended to reflect reality more closely. The transmission rates are time independent, which implies that it is equally likely to that you will transmit an infection to someone else at all points of your own infection, which is not what is seen in reality [Hall et al., 1979; Lee et al., 2009]. Some infections, such as influenza, are also known to be seasonal, meaning that the probability of infection will increase or decrease depending on the time of year. The assumption of just three compartments is also often questionable, as it is unlikely that at the moment of infection, an individual will become infectious themselves, implying that the addition of a latent infected class is desirable. There is a great deal of work also using this susceptible-exposed-infectious-removed ($SEIR$) model e.g. Hethcote and Tudor [1980]; Longini [1986]; Li and Muldowney [1995]; Li et al. [1999] amongst many others.

The advantage of using models with additional compartments is that more complex disease dynamics can be included in the model, along with the ability to examine increasingly complex interventions. For example using an $SEIR$ model and modelling the impact of separating infected individuals from a proportion of their contacts, an individual in the exposed class can be allowed to transmit the disease to their contacts at a lower rate than those who are infectious, but will not be identified as infectious and therefore will not be separated from any of their contacts. In theory any number of compartments can be added to the model to describe different states of the disease that individuals are in along with age specific compartments. However, doing this increases the difficulty of parameterising these models and interpreting their output.

As mentioned previously, the most simple models are deterministic, meaning that if the same initial conditions are used, the dynamics of the disease will always be identical. This

is not what we would expect to observe in reality, as there are many occasions during a real epidemic where chance events occur, resulting in a different pattern of disease spread. Along with this, for (2.1), if the value $R_0 = \beta/\gamma > 1$, we will observe an epidemic, which will infect a significant fraction of the population, whilst if $R_0 < 1$, then this will not occur. Again this is not what we would expect to see on all occasions, as supercritical epidemics can die out, and those with values of $R_0 < 1$ can infect a significant number of people. If a population is very large, then it is reasonable to expect that the final number of infections would be similar if we could repeat the whole epidemic process with no additional control interventions, as all random events get averaged out to impart no significant impact on the spread of the disease. However if we consider smaller populations, the random events will have a much greater impact on the final outcome [Bartlett, 1957; Lloyd, 2004; Britton, 2010]. Therefore the use of stochastic models is common place, which allow us to examine the role that the uncertainty inherent in any epidemic has. The disadvantage of using such models is the increase in difficulty of extracting meaningful statistics from them due to this unpredictability.

One of the first stochastic models for epidemics is the Reed-Frost model [Wilson and Burke, 1942; Abbey, 1952; Bailey, 1975], which takes place in discrete time, to describe an epidemic spreading through a population. Again this falls into an SIR type model, where individuals are infectious for one time step, before being removed. Here the number of susceptible individuals during the next time step is given by a binomial distribution; $S(t+1) = \text{Bin}(S(t), q^{I(t)})$, where q is the probability of contact between any susceptible-infectious pair in the population. Hence the model is sometimes referred to as a 'chain binomial model', as we are effectively picking from a chain of binomial distributions. The number of infectious individuals at time $t+1$ is then given by $I(t+1) = S(t) - S(t+1)$.

For a simple stochastic SIR continuous time epidemic, we again have $S(t)$, $I(t)$ and $R(t)$ denoting the number of susceptible, infected and removed members of the population at time t, but these are now random variables. In much research on stochastic epidemics, the length of a infection of an individual until removal is assumed to be exponentially distributed, with removal rate γ per unit time (equivalent to removal being a Poisson process with rate γ). Contacts between members of the population also take place at the points of a Poisson process with rate β/N, where N is the population size. This choice of infection length results in the epidemic process $\{(S(t), I(t), R(t)) : t > 0\}$ having the Markov property [Bailey, 1975], as the next event to take place, be that an infection or a recovery, depends only on the current state, and not the history of infections.

A common technique for investigating stochastic models is to explicitly realise numerous stochastic trajectories which are defined by the dynamics of the epidemic [House et al., 2012]. There are many methods that are used to do this including Sellke's construction [Sellke, 1983] and Gillespie's algorithm [Gillespie, 1977], which both give results equivalent to the stochastic model. Alternatively, the tau-leap method [Gillespie, 2001] gives an approximation to the stochastic model, but is appreciably faster than the Gillespie algorithm. In practice, many simulations are performed, from which it is possible to extract statistics such as the expected number of infections or the variance possible in the size of the epidemic at any point in time.

Along with this, there are several analytical methods which are used to describe stochastic models. One example is diffusion approximations, which examine the fluctuations from the deterministic trajectory which the stochastic epidemic converges to as the population size tends to infinity, by defining an appropriate diffusion process which can be used to

describe the stochastic fluctuations around this deterministic limit [Clancy et al., 2001; Ross, 2006; Dangerfield et al., 2009; Nåsell, 2002]. This work is based on the technical results derived by Kurtz [1970, 1971] and described again by Ethier and Kurtz [1986], Andersson and Britton [2000] and many other texts.

Another analytical method for examining this particular epidemic is that of directly considering the master equation (also called Kolmogorov forward equations) of the Markov process which governs the probability of seeing a particular state for $p_{si}(t) = P(S(t) = s, I(t) = i)$ [Keeling and Ross, 2008; Nåsell, 2002]. This allows us to analyse the probability of any possible state occurring with one model realisation per parameter set, whilst with Monte Carlo simulation, this requires a large number of realisations to achieve. This method requires the use of differential equations to describe all possible states of the epidemic, which is $\frac{1}{2}(N+1)(N+2)$. Therefore as the population size increases, the computational cost of this method becomes infeasible, meaning that it is quicker to simulate the epidemic multiple times directly via Gillespie's algorithm or an equivalent approach, and draw from these simulations, conclusions about the probabilities.

Stochastic moment-closure models are also used to describe the behaviour of the epidemic by taking the moments of the differential equations describing the epidemics [Isham, 1995; Herbert and Isham, 2000; Krishnarajah et al., 2005; Keeling, 2000; Nåsell, 2003; Keeling and Rohani, 2008]. For example we can calculate the mean, variance and higher moments of the number of infective individuals in the population at time t using this method. This method allows us to, in theory, describe the behaviour of an infinite number of simulations, but, in practice, it is made difficult by the need to increase the number of moments to exactly describe the system at any given level. For example the second moment is needed to describe the evolution of the first moment, and the third moment is needed to describe the second. Therefore it is often the case that any moment above the second is set to zero, or at a certain level, the moment is approximated by a combination of lower moments. This is a great advantage over deterministic models, as these essentially follow this process, but set any moment above the first to zero. Again, however, it is often preferable to directly simulate a number of realisations, with the necessary number of simulations increasing as the order of the moment increases.

The addition of stochasticity is one step towards reality. However the assumption that all members of the population are equally likely to come into contact any other member of the population is still used in the models described so far. In many populations, not all contacts that are made are made at random, for example in human populations many interactions which would be described as contact take place in the home or at work. Network models do not make this assumption.

2.2 Network models

The use of networks as a generalisation from homogeneous mixing is becoming one of the most widely used in epidemiological modelling. Contact between two individuals of the population we are considering forms a link between them. Once a link is established, the infection can be passed along it in either direction. Specifically what contact is depends on the disease, i.e. it is different for a respiratory infection compared to a sexually transmitted infection. There have been many examples of using networks to study the spread of disease and the review papers Bansal et al. [2007] and Danon et al. [2011] compare several different

approaches to network modelling.

A network can be described by an $N \times N$ matrix $\mathbf{A}$, which is called the adjacency matrix. The matrix will be symmetric if the network is undirected, meaning that if node i has node j as a contact, then node j has node i as a contact. If the network is unweighted, i.e. all links are of equal strength, then the entries of $\mathbf{A}$ will be binary, where the entry $A_{i,j}$ will be 1 if nodes i and j have a link between them and 0 otherwise. For weighted networks, entry $A_{i,j}$ gives the strength of contact between nodes i and j, which can signify, for example, the length of time which i and j spend together. The degree of a node is the number of contacts that it has in the network. The degree of node i is denoted k_i and is given by $k_i = \sum_j A_{ij}$.

In this context, a network gives the set of contacts made by all members of the population in question, with whom it is possible to receive or transmit an infection. Some of the earliest uses of networks in this way detailed the spread of sexually transmitted diseases e.g. Klovdahl [1985]; May and Anderson [1987]. Such diseases are ideal for study using networks due to the well defined mechanisms required for transmission to occur, unlike with many other diseases where short-lived interactions can result in transmissions, for example with measles [Paunio et al., 1998] or respiratory infections like influenza.

In the network, a pair is two nodes i and j, who are neighbours of each other. A triple is given by three nodes i, j and l, where i and j are neighbours and j and l are neighbours. If i and l are also neighbours, then this triple forms a triangle in the network.

Bearing this in mind, the progression of an epidemic on a pre-specified network avoids the problem of how to reconstruct a contact network. In this scenario the epidemic will take place on a predefined or static network, or one which is evolving as defined by a set of given rules.

2.2.1 Network properties and impact on disease spread

There are several generic properties of networks, each of which can have an impact on the spread of a disease through the network.

2.2.2 Degree distribution

The degree distribution of a network is the distribution of the number of neighbours that the nodes in the network have. This is defined by a function $P(k)$, which gives the probability that a node selected uniformly at random will have k neighbours. It is clear that the higher the degree of a given node, the more likely it is to become infected during an epidemic, and it is also more likely to spread the infection once it has become infected. It has also been shown, that as the population size diverges, that if there is a large variance in the degree distribution, such as in a scale-free network then the infection can spread very quickly through the network [Barthélemy et al., 2004], and if the variance is also divergent, then there will be no epidemic threshold in the network, meaning that no matter what the ratio between removal and transmission rates, the disease will always infect a non-zero proportion of the population [Boguñá et al., 2003; Pastor-Satorras and Vespignani, 2001; Chatterjee and Durrett, 2009].

In more realistic networks however this increase in variance of degree distribution can help to control the spread, since if the individuals who have many more than the average number of contacts, termed "super spreaders", can be identified and removed from the dynamics, then this can be much more effective than random control methods [Lloyd-Smith et al., 2005; Meyers et al., 2005] and even prevent large outbreaks from occurring [Crépey et al., 2006]. In the case of sexually transmitted diseases or injecting drug users, it is more realistic to expect that these people can be identified and attempts can be made to remove them from the dynamics [Magiorkinis et al., 2013].

Along with super-spreaders, who have a large number of contacts, for diseases which are spread through contaminated droplets, such as SARS and influenza "super-spreading events" have been observed, which environmental conditions can be responsible for [Riley et al., 2003; Galvani and May, 2005; Lipsitch et al., 2003]. These events are unpredictable, which makes preventing them difficult, but we can model them by including more people with greater numbers of contacts if we desire.

In reality the exact network upon which a disease spreads is unknown, and in many cases, such as respiratory diseases in humans, essentially unknowable. However without assuming something about the contact network we are unable to progress at all. There have been many attempts made to characterise contact patterns, for example Mossong et al. [2008], Liljeros et al. [2001] and Danon et al. [2012], which lend weight to the opinion that there is a heavy tail in the distribution of the number of contacts that people have. This means that in general it is not expected that the distribution of number of contacts would be Gaussian or binomial, but is more likely to be negative binomial or power law distribution.

2.2.3 Clustering

The level of clustering in the model gives us the probability that two contacts of a randomly chosen individual are contacts of each other. It is often denoted by ϕ. If the level of clustering in our model is 0, then the probability of two of my contacts contacting each other is 0, whilst if it is 1, then it is certain that they will be contacts of each other. Informally, this is equal to 3 times the number of triangles in the network, divided by the total number of triples in the network. To calculate it for a given network we perform the following calculation,

$$\phi = \frac{\sum_{i,j,k} A_{ij}A_{jk}A_{ki}}{\sum_{i,j,k\neq i} A_{ij}A_{jk}} \ . \tag{2.2}$$

An increase in clustering will reduce the extent to which an infection will spread [Watts and Strogatz, 1998; Eames and Keeling, 2003; Kiss et al., 2005; House and Keeling, 2011a] along with increasing the time to reach the peak of the epidemic [House and Keeling, 2011a]. This is due to the fact that for the infection to spread, it is necessary for an infected individual to have susceptible contacts. Additionally, it is obvious that to become infected in the first place, one of your contacts must have been infected before you. Therefore in a highly clustered network, the probability of having a large number of susceptible contacts rapidly decreases as the epidemic progresses, since to get infected in the first place, a number of your own neighbours are likely to have been infected before you. In contrast, in a network with low or no clustering, the depletion of susceptible contacts that an infected person will have is slower, and will be $O(I)$, due to the fact that your neighbours are not

also neighbours of the infected you, or the node which infected that node and so on, and are therefore less likely to have picked up the infection previously.

In terms of combating a particular infection, contact tracing is a common tool for controlling and assessing the spread of an epidemic. This is the practice of tracing the contacts which were made by an infectious individual, as the likelihood that this will lead to an infected individual is greater than choosing from the population at random. Once traced, these contacts can be quarantined if necessary, or in the case of animal diseases, where farms are taken to be nodes rather than individual animals, the farms can be prevented from further export or import of animals and cordoned off. This has been seen to be successful in identifying infected individuals for sexually transmitted diseases [Götz et al., 2005; Fish et al., 1989] and it was somewhat successful when used in Great Britain during the spread of foot-and-mouth disease in 2001 [Ferguson et al., 2001a,b] and the SARS outbreak of 2002-03 [Lipsitch et al., 2003].

This can be effectively incorporated into a network model; contacts of infected nodes can be identified, with a certain efficacy, and then quarantined if infected. This contact tracing is usually accompanied by a specified efficacy within a model, as in reality it is unlikely that for certain disease types, all contacts will be found. It has been shown [Eames and Keeling, 2003; Kiss et al., 2005; House and Keeling, 2010], that as clustering increases, the levels of efficacy needed to produce a given reduction in disease spread decreases. Again this is because a randomly selected, infectious node is likely to have a greater number of infectious contacts in a highly clustered network than in a network with no clustering, meaning that the average number of infectious individuals found each time contact tracing is performed will be higher. However, as efficacy increases towards 1 this result is reversed [House and Keeling, 2010], which demonstrates how complicated and subtle the interaction of clustering and the spread of the epidemic is.

In reality, we may expect that the level of clustering will be non-zero and a survey of the UK population [Danon et al., 2012] which had over 5,000 responses reports the level of clustering in the population as being as high as 0.38.

2.2.4 Degree assortativity

A degree assortative network is one in which the high degree nodes are contacts of other high degree nodes and low degree nodes are neighbours of low degree nodes more than would be expected at random. A degree disassortative network has nodes with high and low degree as neighbours more commonly than would be expected at random. This is often denoted by r. To calculate this for a given network, the following is calculated,

$$r = \frac{\sum_{j,k} jk(e_{jk} - q_j q_k)}{\sum_k k^2 q_k - (\sum_k k q_k)^2} \ , \tag{2.3}$$

where $q_k = \frac{(k+1)p_{k+1}}{\sum_j j p_j}$, which is referred to as the remaining degree, and e_{jk} is the joint probability distribution of the remaining degrees of the nodes at either end of a randomly chosen edge [Newman, 2002b]. This lies between -1 and 1, which define a perfectly disassortative and assortative network respectively.

If an epidemic has only one introduction of an infected individual, then the maximum number of infections that can occur in an epidemic is equal to the size of the largest

connected component of the network. In a network with a divergent number of nodes, if the largest connected component is a non-zero fraction of the total network size, then this is referred to as the giant component of the network. It was shown in Newman [2002b], that all other things being equal, the probability that a giant component exists is greater for an assortative network than for a random network, and that in turn the probability is greater for a random network than for a disassortative network. This suggests that combating disease on assortative networks can be more difficult than other networks, due to the fact that high degree nodes are connected to each other, removing these nodes will be redundant until a high proportion of them are removed. This increases the difficulty of combating the spread of diseases where the networks are assortative, as is the case for sexual activity, compounding the impact of many sexually transmitted diseases [Potterat et al., 1985; Granath et al., 1991], along with the practice of needle sharing and the number of injecting partners in injecting drug users [Mills et al., 2012].

The opposite is true for disassortative networks in that they can be easily broken up by removing the high degree nodes. This means that these types of networks are especially vulnerable to targeted attacks and as many networks which are considered valuable, such as the internet and food webs are disassortative [Martinez, 1991; Zhang et al., 2012], these may be more vulnerable than anticipated. In terms of disease control, a disease transmitting on a disassortative network, may be more readily controlled if the high degree nodes can be identified and removed, though in practice this would not be straightforward.

An example of a disassortative network is given in Kiss et al. [2006a], where the difference between a network describing sheep movements which was derived from data regarding movements within the UK is compared to a randomly constructed network of the same size and degree. The nodes of these networks are all places which sheep move to and from, so include sheep markets along with farms. These data driven networks are disassortative, meaning that nodes with many connections are more likely to contact nodes with lower degree. This is due to the fact that the most likely route of movement is from a sheep market to any number of farms, which results in the market having a higher degree than the farms to which sheep travel.

This comparison shows that the proportion of nodes which become infected is higher for random networks. This can be explained by the disassortativity of the data driven networks, as the linking of high degree nodes with low degree ones implies that it may take longer on average to reach highly connected nodes in the network, and from there will transmit infection to nodes with lower than expected degree, slowing the spread of the epidemic. Additionally these data derived networks also have a longer path length than the randomly constructed networks, the impact of which is discussed below.

2.2.5 Average shortest path length

Path length (or shortest path length) in a network is the number of steps required to get from one node to another. The average shortest path length is the average of this number over all pairs of nodes which comprise the network. This is defined per connected component, since otherwise we would get ∞ for all networks which are not fully connected. One can also average the shortest path length from one specific node to all others in the network, which gives an indication of how likely this node is to become infected during the epidemic; as this average increases, more and more individuals must become infected before the epidemic will spread to this node. Generalising this, as the overall average

shortest path length increases, the speed of spread of the epidemic will be decreased and the rate of spatial spread will also decrease [Watts and Strogatz, 1998].

To move away from the mean-field assumption, a simple way to introduce some structure into our model is to consider the spread of an epidemic on a lattice [Sato et al., 1994]. Here contacts are defined to be the neighbours of each node on the lattice. This a realistic assumption to make in certain populations, such as in the spread of fungal parasites from plant to plant [Otten et al., 2004], but for animal and human populations this is often not a close approximation of reality. This can be extended to lattices of more than one dimension and also include links to k nearest neighbours, rather than simply immediate neighbours.

As the lattice has connections only between nearest neighbours (or some number k nearest neighbours) the average shortest path lengths will be high. For example to get from one end of the lattice to the other, an infection must pass through every connecting node.

To avoid this problem the lattice can be re-wired by removing links between certain neighbours, and attaching to a randomly chosen node in the network. This will significantly decrease these large path lengths which occur. This was popularised by Watts and Strogatz [1998], and these networks were known as 'small-world' networks, due to the fact that as the path lengths were significantly shortened, there were far fewer nodes needed to be passed through in order to reach anyone in the population. This explains the 'small-world' problem examined by Milgram [Milgram, 1967] which gave rise to the notion of six degrees of separation, as if this is assumed to be representative of the human population, any one person could then reach any other person, by going through (for arguments sake) six or fewer intermediaries.

2.2.6 Models and modelling techniques

Having discussed the properties of networks that impact on disease spread, we will now discuss different network models and techniques used to investigate disease spread on networks.

The first step away from mean-field mixing, where everyone in the population is able to contact everyone else, is where every individual in the population has a given number n contacts. These types of networks are called regular networks. The spread of an epidemic on these networks have been compared to the mean-field models [Keeling, 2005] and the impact of small-world effects [Santos et al., 2005]. As noted previously, regularity is a reasonable assumption for certain populations for [Otten et al., 2004], however for networks of human interaction, this is a poor assumption Wadsworth et al. [1993]; Liljeros et al. [2001]; Mossong et al. [2008]; Danon et al. [2012].

Heterogeneity in the number of contacts that individuals have is included in many network models. For example in Erdös-Rényi (ER) random graphs the link between every pair of nodes is present with a given probability p. For our purposes, when the construction of a network is needed, the configuration model [Molloy and Reed, 1995] is used to construct the network from the degree distribution.

The final size of an epidemic with a given degree distribution along with the mean degree of the individuals infected has been derived [Newman, 2002a] and aspects related to the ability of an epidemic to invade the population, including the invasion threshold of an

epidemic has been calculated [Keeling, 1999; Diekmann and Heesterbeek, 2000].

In most cases, the network upon which either theoretical or simulation based investigation of epidemics is performed are static networks, fixed at the starting point of the epidemic e.g. Newman [2002a], Volz [2008] and Mossong et al. [2008] to name a few. However there are also examples where this network is allowed to evolve over time, interacting with the spread of the epidemic e.g. Kamp [2010], Miller and Volz [2012] and Pastor-Satorras and Vespignani [2001]. In the case of Pastor-Satorras and Vespignani [2001] along with May and Lloyd [2001] and some models of Miller and Volz [2012] the network of contacts is changing constantly. In reality, the contact network for a respiratory infection will have some links which are always present (made often or daily), and some which are more short lived, such as those made whilst using public transport. This has been considered by including a 'global' infection term [Kiss et al., 2006b; Ball and Neal, 2008], which allows the disease to be passed between members of the population that do not have more than one interaction.

Similar techniques to those developed for stochastic mean-field models have been used for investigating stochastic network models, with moment-closure methods e.g. Taylor et al. [2012], Rand [1999] and Rogers [2011] along with Kolmogorov-forward equation techniques [Allen et al., 2008; Simon et al., 2011].

Additionally, we note the use of a probability generating function method [Volz, 2008; Miller, 2010], which allows us to succinctly denote properties of the network which influence the dynamics, to derive a small number of nonlinear ODEs that describe the dynamics of an SIR infection on a random heterogeneous network.

In general to consider the dynamics of the epidemic in continuous time, a Markov chain model is used. For an SIR-type epidemic model on an arbitrary graph with N nodes, Markov chain would involve 3^N ODEs, which quickly becomes computationally intractable. The method proposed by Ball and Neal [2008] involves creating a configuration model network at the same time as the epidemic tree, which can be defined by $2M$ ODEs in the deterministic large N limit, where M is the maximum number of contacts that any one node has.

By making an assumption about the neighbourhoods of individuals on the network a far smaller equation set has been derived [Volz, 2008]. A sophisticated convergence proof [Decreusefond et al., 2012] has demonstrated the exactness of this assumption, and hence the equation set, in the large N limit.

To investigate the inherent stochasticity of epidemics without a large increase in dimensionality has led researchers to consider the diffusion limit. This general approach to stochastic processes is typically either attributed to N G van Kampen [1992] or Kurtz [1970, 1971]. Here the stochastic Markov model can be approximated by a set of deterministic ODEs, with the stochasticity being characterised by scaled white noise processes with magnitudes defined by the transition rates between the states of the Markov model.

Such methods have been used to derive a low dimensional model in which properties of the noise in a stochastic epidemic model can be investigated analytically [Alonso et al., 2007; Black et al., 2009], by Ross [2006] to obtain expressions for the mean and variance of a meta-populations model, and by Colizza et al. [2006] to model the effect of air travel on the spread of epidemics in a large-scale network. These models are attractive since they have the same dimensionality as the deterministic limit [Volz, 2008], but allow investigation of the stochasticity of the epidemic.

These methods are employed in the next chapter in order to investigate the variance in the early growth period of a stochastic epidemic on a heterogeneous network.

Chapter 3

Early growth variance of epidemics on heterogeneous networks

In this chapter we will apply the results of Kurtz [1970, 1971] to SIR-type epidemic dynamics on a configuration model network, as was done for SIS dynamics on a regular graph by Dangerfield et al. [2009]. Using this we obtain a four-dimensional set of stochastic ODEs, from which we derive an analytical expression for the variance of the asymptotic early growth of an epidemic on a network given its degree distribution. We simulate epidemics on various networks to confirm the utility of our analytical results.

3.1 Model description

The following sections are an extended look at the results published in Graham and House [2013]. Firstly we discuss the construction of a network model and the limitations of using network models in the way that we have used them.

There are N individuals connected to each other on a configuration model network. This implies that there is no clustering in the population and so there are no short loops in the population. This means that the consideration of depletion of susceptible contacts is made simpler due as we can be sure that when an individual i is infected by individual j, that the neighbors of i will not have been infected by individual j.

Individuals are compartmentalised by their disease state S, I, or R, and their number of neighbours on the network, their degree, k.Individuals of type S_k become I_k at a rate equal to the product of the transmission rate τ and their number of infectious neighbours. Individuals of type I_k become R_k at a rate γ.

As the aim of this work is to get a theoretical result using diffusion methods, I am interested in the large N regime, in which $[S_k]$ denotes the expected number of susceptibles of degree k, $[I_k]$ the expected number of infectious individuals of degree k, and $[AB]$ for the number of connected pairs of individuals on the network where one is type A and the other is type B. Omission of a subscript denotes implicit summation, e.g. $[S] = \sum_k [S_k]$. Proportions of the population who are, say, susceptible and of degree k are represented by the bracket less equivalent of the number of the population sharing the same description, therefore in this case by S_k.

For the diffusion limit, the population size N is allowed to increase towards ∞, whilst keeping the proportions of susceptibles and infecteds of different degrees constant. What follows is a description of the deterministic process that the epidemic approaches when the process can be described as 'density-dependent'. Using appropriate theory [Kurtz, 1970, 1971], the stochasticity of the system can be characterised, and used to calculate the variance of the epidemic during its early growth phase.

Limitations of this approach are now considered.

3.1.1 Limitations of network models

The main limitation of network models is that the true network on which an epidemic will take place is unknown. Using a degree distribution to describe a full contact network is limited and considering an unweighted network means that much subtlety is lost in the description of the contact structure of a population.

As mentioned previously, the network is assumed to be unclustered, which is known to be a poor assumption [Liljeros et al., 2001; Danon et al., 2012]. Additionally, this has a significant impact on the dynamics of an epidemic as described in §2.2.3.

It is also assumed that the network upon which the epidemic is spread is a static network, meaning that the network is the same every day, which is again unrealistic.

Next notation and some standard approaches to the SIR on a network are described.

3.1.2 Notation, mean field and pairwise models

The degree distribution is given by $P(k)$ as is noted above, and d_k is used to denote the proportion of nodes which have degree k. Nd_k gives the number of nodes which have degree k. Also of use to the analysis of this system is the probability generating function (pgf) of the degree distribution. This is denoted by $g(x)$ and $g(x) = \sum d_k x^k$. Note that this gives a simple way of expressing many aspects of the system, e.g. $g'(1) = \sum k d_k = \bar{n}$ where $\bar{n}$ is the mean of the degree distribution.

Another assumption about the system is that for susceptible nodes, infection across each link is independent of all other links that the node in question has. The implication of this, is that if we calculate the probability that a node with one link which is selected uniformly at random, conditioned on having only 1 link, is susceptible at time t, and label this θ, then the probability that a node of degree k is susceptible at time t is given by θ^k. If there are no nodes of degree 1, or if we were considering a complete graph or a regular network, then we can think of this value as being the probability that infection will have passed down a specific link.

Using this variable, it is clear that $\theta = [S_1]/Nd_1$ and that $[S_k] = Nd_k\theta^k$. This therefore also gives that $[S_k] = Nd_k([S_1]/Nd_k)^k$. Using the pgf allows us to express $[S]$ as, $[S] = \sum[S_k] = \sum Nd_k\theta^k = Ng(\theta)$. It follows that instead of writing down equations that allow $[S]$ to be tracked, θ can be used in its place.

The progress of the epidemic can be described by a continuous time Markov chain, as the state of the system at a future time only depends on its current state. Ostensibly this can be described using a three-dimensional Markov chain, where we track the values of $[S]$ (or

θ), $[I]$ and $[R]$. However as the process is running on a network, the rates of change of these state variables is dependent on the network itself. For example to become infected, a susceptible node must have an infected neighbour, which therefore means that the rate of $[S] \to [I]$ depends on $[SI]$.

In fact it is simple to write down the evolution of these three state variables.

$$\begin{aligned} [\dot{S}] &= -\tau[SI], \\ [\dot{I}] &= \tau[SI] - \gamma[I] \ , \\ [\dot{R}] &= \gamma[I] \ . \end{aligned} \tag{3.1}$$

We therefore see that this set of equations is unclosed, as we need to know the evolution of $[SI]$ to fully specify the system. Therefore rather than a three-dimensional Markov chain, a four-dimensional one must be calculated.

Indexing the process of infection or recovery in terms of degree is useful to write down the rate of changes for the pairwise variables. This is done in Eames and Keeling [2002]. For example to gain an $[SI]$ pair, we can gain one whose susceptible individual has k neighbours and the infected individual has l i.e. an $[S_k I_l]$ pair. Once these have all been calculated, simply summing over the indexes leads to the differential equations needed to track the system. When this is done, the problem of needing to keep track of more variables again occurs, as we have terms involving the number of triples in the system. As can be seen in House and Keeling [2011b], the differential equation for the full set of equations at pair level is given by:

$$\begin{aligned} [\dot{S}] &= -\tau[SI] \\ [\dot{I}] &= \tau[SI] - \gamma[I] \\ [\dot{SS}] &= -2\tau[SSI] \\ [\dot{SI}] &= \tau([SSI] - [ISI] - [SI]) - \gamma[SI] \\ [\dot{II}] &= 2\tau([ISI] + [SI]) - 2\gamma[II] \ . \end{aligned} \tag{3.2}$$

meaning that the five-dimensional system must become a seven-dimensional one as $[SSI]$ and $[ISI]$ must also be tracked. This need to increase the size of the system is one which continues ad-nauseam as to give the exact dynamics at the n-th level requires the inclusion of $(n+1)$-th level variables (singles depends on pairs, pairs on triples, ...).

To make progress in this direction, an assumption about the neighbourhood must be made. An approach is to derive pairwise equations by approximating the triples by some function of pairs or lower variables such as θ. This is often done using moment-closure techniques [Rand, 1999; Rogers, 2011; House and Keeling, 2011b].

To derive pairwise equations, the method used here is to exploit an assumption about the neighbourhood of each node. This is due to the fact that when an infection or recovery takes place, the number of pairs of type $[AB]$ will be changed in a way which is dependent on the neighbours that the node has. This is why extra consideration is needed to write down a low-dimensional form for this process as the population size becomes large, since this requires the distribution of neighbours of each node.

Note that whatever assumptions are made, the equations that result from it, must agree with (3.2) at pair level. For example when calculating the differential equation for $[S]$, the

result must be $[\dot{S}] = -\tau[SI]$ and the recovery term for $[II]$ must be given by $-2\gamma[II]$.

The following assumption is made about the neighbourhoods around a susceptible of degree k: the distribution of susceptible, infected or removed neighbours of this node is independent of k. This is the same assumption that was made in Volz [2008] and was proven to be asymptotically correct in Decreusefond et al. [2012]. Defining n_S, n_I and n_R to be the number of susceptible, infected and recovered neighbours respectively, the probability of a neighbourhood is given by a multinomial distribution as follows:

$$P(n_S = x, n_I = y|k) = D^S_{x,y,k} = \binom{k}{x, y, k-x-y}(1 - p_S - q_S)^{k-x-y} p_S^x q_S^y \ , \tag{3.3}$$

where,

$$p_S = \frac{[SS]}{\sum_k k[S_k]} \ , \ q_S = \frac{[SI]}{\sum_k k[S_k]} \ . \tag{3.4}$$

Here the term $\binom{k}{x,y,k-x-y} = k!/x!y!(k-x-y)!$, is the multinomial coefficient. The fact that p_S and q_S are given by (3.4) means that no matter where on the network the susceptible node is, the distribution of its neighbours will be the same. This implies that after the epidemic begins, some time must be allowed to pass, in which the initial conditions of the system are forgotten before this assumption is accurate.

Note that we have not yet made an assumption regarding the distribution of neighbours around an infected node. This is due to the fact that this is far more complicated than the neighbourhood of a susceptible, as the longer a node is infected, the more infected (or at least non-susceptible) neighbours it is likely to have. This is analysed in detail in §3.2.

3.1.3 Density dependent processes

To calculate the variance of the epidemic process on the network, the work of Kurtz [1970, 1971] is used. There are a few conditions that the process must satisfy in order for this to be used, the first one of which is that the process can be thought of as a density dependent one. The definition of a density dependent process is given in Kurtz [1970], and is conveniently set out in Ross [2006].

To begin this definition note the following; the epidemic process is a continuous-time Markov chain, which is denoted X_N, with a discrete state space labelled $E_N \subset \mathbb{Z}^D$, where D is the dimension of the state space. The rate of transition between states j and $j+l$, with $j, j+l \in E_N$ is given by $q_N(j, j+l)$. The following is the definition of a density-dependent process given in Kurtz [1970]:

Definition 1. *A one parameter family of Markov chains, $X_N(t)$, with state space $E_N \subset \mathbb{Z}^D$ is called density dependent if and only if there exist continuous functions $f(x, l)$, where $x \in \mathbb{R}^D$, $l \in \mathbb{R}^D$, such that the rates of transitions corresponding to $X_N(t)$ are given by*

$$q_N(j, j + l) = Nf(j/N, l), \quad l \neq 0 \ .$$

Defining $Y_N(t) = X_N(t)/N$ as the density process, and $F(x) = \sum_l lf(x, l)$. In Kurtz [1970] it is shown that Y_N satisfies

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}(Y_N(s)) = \mathbb{E}(F(Y_N(s))) \ , \tag{3.5}$$

and that for all $\epsilon > 0$

$$\lim_{N\to\infty} P\Big(\sup_{s\leq t} |Y_N(s) - Y(s)| > \epsilon \Big) = 0 \ , \tag{3.6}$$

where $Y(s)$ is the solution of

$$\frac{\mathrm{d}}{\mathrm{d}s} Y(s) = F(Y(s)) \ . \tag{3.7}$$

Essentially, this definition tells us that if the transition rates in the density process $Y_N(t)$ depend only on the current state through the density j/N, then the Markov process $X_N(t)$ is density dependent.

This tells us that even though the process Y_N is stochastic, as $N \to \infty$ it can be approximated by a set of deterministic differential equations, here defined by Y, such that (3.7) holds.

Along with this calculation of the deterministic approximation, Dangerfield et al. [2009] shows how the variance of the process during the early growth period can be calculated using the work of Kurtz [1970, 1971], which is the goal of this analysis. Therefore it remains to show that the epidemic process in question is density dependent.

For the system in question let $X_N = ([S_1], [S_2], \ldots, [S_M], [I_1], [I_2], \ldots, [I_M], [SS], [SI], [II])$, where M is the number of neighbours that the most connected node in the system has. This therefore defines $Y_N = ([S_1]/N, \ldots, [S_M]/N, [I_1]/N, \ldots, [I_M]/N, [SS]/N, [SI]/N, [II]/N)$. We can calculate the change in the variables of X_N, denoted above by l, by considering the events that can occur. Namely, these are the infection of a susceptible of degree k, who has x susceptible neighbours and y infected neighbours getting infected. The other event that can occur is an infected of degree k who has again x susceptible neighbours and y infected neighbours being removed from the dynamics via recovery or death.

The changes in the variables X_N caused by the first event is given by

$$l_{\tau,x,y,k} = (-\delta_{k,1}, -\delta_{k,2}, \ldots, -\delta_{k,M}, \delta_{k,1}, \delta_{k,2}, \ldots, \delta_{k,M}, -2x, x-y, 2y) \ . \tag{3.8}$$

The delta functions are needed because, for example, $[S_1]$ and $[I_1]$ only change if the degree of the susceptible node is 1. There will always be an increase of the number of infecteds by 1, and the fact that the central susceptible has x susceptible neighbours giving x $[SS]$ pairs, these are double counted which gives the $-2x$ change in $[SS]$. The y infected neighbours, which make y $[SI]$ pairs, become $[II]$, which get double counted explaining the $2y$ change in $[II]$, whilst the x $[SS]$ pairs become $[SI]$ pairs, explaining the $x-y$ change in $[SI]$. For the second event, following a similar process,

$$l_{\gamma,x,y,k} = (0, 0, \ldots, 0, -\delta_{k,1}, -\delta_{k,2}, \ldots, -\delta_{k,M}, 0, -x, -2y) \tag{3.9}$$

First consider the transition rates related to the change in variables given by $l_{\tau,x,y,k}$, which is denoted by $q_N(j, j + l_{\tau,x,y,k})$. The assumption given at (3.3) is used to calculate this. For the infection event, the rate of transition is given by

$$q_N(j, j + l_{\tau,x,y,k}) = \tau y [S_k] D^S_{x,y,k} \ , \tag{3.10}$$

as for this event to occur, a susceptible of degree k with x susceptible nodes and y susceptible nodes must be infected, which gives the $[S_k]D^S_{x,y,k}$ term as this is assumed to be

the number of such nodes. The rate that these nodes are infected is τ multiplied by the number of neighbours who are infected, hence the τy part. When this is written out in full, using the value of $D^S_{x,y,k}$ given in (3.3), the following expression is obtained:

$$\begin{aligned} q_N(j, j + l_{\tau,x,y,k}) =& \tau y [S_k] \binom{k}{x, y, k-x-y} \left(\frac{[SS]}{\sum_k k[S_k]} \right)^x \left(\frac{[SI]}{\sum_k k[S_k]} \right)^y \left(1 - \frac{[SS] + [SI]}{\sum_k k[S_k]} \right)^{k-x-y} , \\ =& N \left(\tau y \frac{[S_k]}{N} \binom{k}{x, y, k-x-y} \left(\frac{[SS]}{N} \frac{1}{\sum_k k \frac{[S_k]}{N}} \right)^x \left(\frac{[SI]}{N} \frac{1}{\sum_k k \frac{[S_k]}{N}} \right)^y \right. \\ & \left. \left(1 - \frac{[SS] + [SI]}{N} \frac{1}{\sum_k k \frac{[S_k]}{N}} \right)^{k-x-y} \right) , \\ :=& N f(j/N, l_{\tau,x,y,k}) , \end{aligned} \tag{3.11}$$

where $f(j/N, l_{\tau,x,y,k)}$ is given by the term which is preceded by the N after the second equality sign above and is simply given by $\tau y([S_k]/N) D^S_{x,y,k}$. Therefore the infection events can be thought of as being density dependent as they satisfy definition 1.

Using (3.7), the development of the system of variables $Y_N(t)$ due to transmissions can be approximated by the following deterministic calculation:

$$\frac{\mathrm{d}}{\mathrm{d}t} Y(t) = \sum_k \sum_{x,y} l_{\tau,x,y,k} f(j/N, l_{\tau,x,y,k}) \tag{3.12}$$

As an example consider what occurs for each $[S_k]/N$ term. From (3.12), it can be seen that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \frac{[S_n]}{N} =& \sum_k \sum_{x,y} -\delta n, k f(j/N, l_{\tau,x,y,k}) \\ =& -\tau \frac{[S_n]}{N} \sum_{x,y} y D^S_{x,y,n} , \end{aligned} \tag{3.13}$$

where the summation over x and y leads to simply calculating the average number of infected partners of a node with degree n. Using (3.3) this is given by $n[SI]/\sum_k k[S_k]$. The differential equation governing $[S_n]/N$ is therefore:

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{[S_n]}{N} = -\tau \frac{[S_n]}{N} \frac{n[SI]}{\sum_k k[S_k]} . \tag{3.14}$$

Summing over all values of n will give the differential equation for $[S]/N$, which should agree with (3.1). This gives:

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{[S]}{N} = -\tau [SI] N \frac{\sum_n n[S_n]}{\sum_k k[S_k]} = -\tau \frac{[SI]}{N} , \tag{3.15}$$

as is expected from (3.1).

Making similar calculations for $[I]/N$ gives that for the transmission events

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{[I]}{N} = \tau [SI] N \frac{\sum_n n[S_n]}{\sum_k k[S_k]} = \tau \frac{[SI]}{N} , \tag{3.16}$$

which again agrees with (3.1).

Now consider $[SS]/N$. Following the same method gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{[SS]}{N} = \sum_k \sum_{x,y} (-2x)\tau y \frac{[S_k]}{N} D^S_{x,y,k} = -\frac{2\tau}{N}\sum_k k(k-1)[S_k]\frac{[SS][SI]}{(\sum_k k[S_k])^2} \ . \tag{3.17}$$

Again using the variable θ, which is the proportion of degree 1 nodes which are still susceptible this expression can be re-written to involve the pgf $g()$. Remembering that $[S_k] = Nd_k\theta^k$ gives,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{[SS]}{N} = -\frac{2\tau}{N}\sum_k k(k-1)[S_k]\frac{[SS][SI]}{(\sum_k k[S_k])^2} &= -\frac{2\tau}{N}\sum_k k(k-1)Nd_k\theta^k \frac{[SS][SI]}{(\sum_k kNd_k\theta^k)^2} \\ &= -\frac{2\tau N}{N}\frac{[SS]}{N^2}\frac{[SI]}{N^2}\frac{\theta^2\sum_k k(k-1)d_k\theta^{k-2}}{(\theta\sum_k kd_k\theta^{k-1})^2} \\ &= -2\tau\frac{[SS]}{N}\frac{[SI]}{N}\frac{g''(\theta)}{g'(\theta)^2} \ .\end{aligned} \tag{3.18}$$

For ease of notation, instead of writing $[A]/N$ to denote the density of a variable, such as $[S]/N$, define $[A]/N = (A)$. Lumping together the differential equations for $[S_k]$'s and $[I_k]$'s and performing similar calculations for $[SI]/N$ and $[II]/N$ ((SI) and (II)) gives the following set of equations which govern the evolution of the system due to transmission processes, which are indexed by τ to signify that they only include transmission terms:

$$\begin{aligned}(\dot{S})_\tau &= -\tau(SI) \\ (\dot{I})_\tau &= \tau(SI) \\ (\dot{SS})_\tau &= -2\tau(SS)(SI)\frac{g''(\theta)}{g'(\theta)^2} \\ (\dot{SI})_\tau &= \tau(SI)\Big(\frac{g''(\theta)}{g'(\theta)^2}\Big((SS)-(SI)\Big)-1\Big) \\ (\dot{II})_\tau &= 2\tau(SI)(SI)\frac{g''(\theta)}{g'(\theta)^2} \ .\end{aligned} \tag{3.19}$$

Note that this set of equations requires knowledge of the change in θ as time increases. As $S = g(\theta)$, it turns out that it is easier to keep track of θ instead of S, as the calculation of S given θ is simpler than the reverse. When the differential equation governing θ is calculated, it is seen that

$$\dot{\theta} = -\tau\frac{(SI)}{g'(\theta)} \ . \tag{3.20}$$

This is easy to show from (3.14), if $n = 1$ and noting that $\theta = (S_1)/d_1$ and $(S_k) = d_k\theta^k$.

For the recovery events,

$$q_N(j, j+l_\gamma) = \gamma[I_k]D^I_{x,y,k} = N\gamma\frac{[I_k]}{N}D^I_{x,y,k} \ , \tag{3.21}$$

For this to satisfy the definition of a density dependent process, $\gamma\frac{[I_k]}{N}D^I_{x,y,k}$ must be a function of variables in Y_N, which is denoted $f(j/N, l_{\gamma,x,y,k})$. One way to ensure that this is the case is to make an analogous definition for $D^I_{x,y,k}$ as is made for $D^S_{x,y,k}$. This would mean defining

$$D^I_{x,y,k} = \binom{k}{x,y,k-x-y}\left(\frac{[SI]}{\sum_k k[I_k]}\right)^x\left(\frac{[II]}{\sum_k k[I_k]}\right)^y\left(1-\frac{[SI]+[II]}{\sum_k k[I_k]}\right)^{k-x-y}, \quad (3.22)$$

where all bracketed terms from X_N, such as $[SI]$, can be replaced by their equivalent term from Y_N, as the division by N would be cancelled in each term.

This approach however does not correctly capture the neighbourhoods of the infecteds, namely that the longer that a node has been infected, the more infectious neighbours it is likely to have. Attempting to make a more accurate approximation for $D^I_{x,y,k}$ is a non-trivial problem, and will be returned to later in this section.

Note that in (3.2), the terms which involve recovery events (those multiplied by γ) are all at the level of pairs or lower. Hence whatever assumption is made about $D^I_{x,y,k}$ to generate the pairwise approximation terms, the terms generated must agree with those in (3.2).

To calculate the recovery terms according to (3.7), the following calculation is made

$$\frac{\mathrm{d}}{\mathrm{d}t}Y_N(t) = \sum_k\sum_{x,y} l_{\gamma,x,y,k}f(j/N, l_{\gamma,x,y,k}) \ . \quad (3.23)$$

This gives no terms for (S) or (SS), but does for the rest of Y_N.

For (I_n),

$$(\dot{I_n}) = \sum_k\sum_{x,y} -\gamma\delta n, k(I_k)D^I_{x,y,k} = -\gamma(I_n)\sum_{x,y}D^I_{x,y,n} = -\gamma I_n \ . \quad (3.24)$$

Summing over n then gives $(\dot{I}) = -\gamma(I)$, which agrees with (3.2).

Considering (SI) leads to evaluating

$$(\dot{SI}) = -\gamma\sum_k(I_k)\sum_{x,y}xD^I_{x,y,k} \ . \quad (3.25)$$

To agree with (3.2), this also gives the constraint that if the $D^I_{x,y,k}$ must satisfy both

$$\sum_k(I_k)\sum_{x,y}D^I_{x,y,k} = (I) \text{ and } \sum_k(I_k)\sum_{x,y}xD^I_{x,y,k} = (SI) \ . \quad (3.26)$$

Note that the first condition is automatically satisfied for any sensible assumption about $D^I_{x,y,k}$, as $\sum_{x,y}D^I_{x,y,k} = 1$, due to the fact that this sums over all possible arrangements for neighbours of a degree k infected node, and therefore must be equal to 1.

To finish this calculation consider the differential equation governing the recovery events involving (II). This gives

$$(\dot{II}) = -2\gamma\sum_k(I_k)\sum_{x,y}yD^I_{x,y,k} = -2\gamma(II) \ , \quad (3.27)$$

which again must match with (3.2), giving $\sum_k (I_k) \sum_{x,y} y D^I_{x,y,k} = (II)$.

Putting together the transmission event equations with the recovery event equations from (3.2) leads to the following closed, but inexact set of equations:

$$\begin{aligned}
\dot{\theta} &= -\tau \frac{(SI)}{g'(\theta)} \\
(\dot{I}) &= \tau(SI) - \gamma(I) \\
(\dot{SS}) &= -2\tau(SS)(SI)\frac{g''(\theta)}{g'(\theta)^2} \\
(\dot{SI}) &= \tau(SI)\Big(\frac{g''(\theta)}{g'(\theta)^2}\Big((SS) - (SI)\Big) - 1\Big) - \gamma(SI) \\
(\dot{II}) &= 2\tau(SI)(SI)\frac{g''(\theta)}{g'(\theta)^2} + 2\tau(SI) - 2\gamma(II) \ .
\end{aligned} \tag{3.28}$$

Note that the only place that terms involving (II) appear is in the differential equation for (II), which means that this is not needed to fully specify the system. Further, the calculation of this variable is the only place in which knowledge of the number of infected partners of an infected node is needed, therefore this is not needed to understand the system dynamics either. Therefore instead of considering $D^I_{x,y,k}$ i.e. the probability of having x susceptible neighbours and y infected neighbours for an infected node of degree k, we can ignore the number of infected neighbours so $D^I_{x,k} = \sum_y D^I_{x,y,k}$ can be used instead.

This leaves a closed, density dependent system of dimension four which is governed by the first four differential equations of (3.28).

The deterministic approximation to the stochastic process has now been calculated, and can be written succinctly if instead of considering Y_N as defined previously, the system $Z_N = \big(\theta, (I), (SS), (SI)\big)$ is considered, as it has been shown that Y_N can be written in terms of the variables in Z_N. The deterministic system of equations in (3.28) can be summarised as

$$\dot{Z_N} = \sum_e l_e f_e(Z_N, l_e) \ , \tag{3.29}$$

where e refers to any event that can occur, i.e. a susceptible with a given degree and a given number or susceptible and infected neighbours getting infected, or an equivalent infected recovering. l_e is the change in the Z_N variables from the event e and $f_e(Z_N, l_e)$ is the rate of this event given by $f_{\tau,x,y,k}(Z_N, l_{\tau,x,y,k)} = \tau y(S_k) D^S_{x,y,k}$ or $f_{\gamma,x,y,k}(Z_N, l_{\gamma,x,y,k}) = \gamma(I_k) D^I_{x,y,k}$.

For the transmission event (degree k node, x susceptible neighbours, y infected neighbours), the change in Z_N is given by

$$l_{\tau,x,y,k} = (-\delta_{k,1}/d_1, 1, -2x, (x-y)) \ , \tag{3.30}$$

and for the similar recovery event

$$l_{\gamma,x,y,k} = (0, -1, 0, -x) \ . \tag{3.31}$$

Using the same notation, by Dangerfield et al. [2009], the full stochastic system can be

written as

$$\dot{Z_N} = \sum_e l_e f_e(Z_N, l_e) + \sum_e l_e \sqrt{f_e(Z_N, l_e)}\xi_e \ , \tag{3.32}$$

where ξ_e is an independent standard Gaussian noise process associated with the event e. This is not detailed in full here as the as the square root in the $\sum_e l_e \sqrt{f_e(Z_N, l_e)}$ term, means that we cannot simplify the expressions by taking moments of (3.3) as we can for the $\sum_e l_e f_e(Z_N, l_e)$ term (which was done to derive (3.28)).

3.1.4 Early growth behaviour

The starting point with any analysis of the behaviour of the system in the early growth period is assuming that the early growth of the infection is exponential, with rate r. That is,

$$[I] = N\tilde{I}\exp(rt) \ , \tag{3.33}$$

or equivalently $(I) = \tilde{I}\exp(rt)$ where $\tilde{I}$ is a constant related to the prevalence of the infection as the early asymptomatic behaviour commences. We consider the case where $\tilde{I}$ is a small parameter, i.e. $\tilde{I} \ll 1$. When approximating other variables during the early growth period, only the lowest order terms of (I) will be considered, where we ignore higher order terms due to the fact that $I \ll 1$. Additionally in taking the diffusion limit $\tilde{I}$ should be significantly larger than 1 but also significantly smaller than the total population size N. In the large N limit, this stage of the epidemic is potentially infinitely long, but for smaller populations, this is likely to be a short period. This is because in the earliest stage of the epidemic a certain length of time must be allowed to pass, and a certain number of infections must take place before the growth of the epidemic will be exponential with rate given by r.

At the disease-free equilibrium of the system the variables we are considering take the following values,

$$\theta = 1,\ (I) = 0,\ (SS) = \bar{n} \text{ and } (SI) = 0. \tag{3.34}$$

By examining the early growth period, we consider the dynamics which follow from beginning a small perturbation from this equilibrium position at $t = 0$. If the parameters that we have dictate that the equilibrium is unstable, then we will diverge from the equilibrium and the growth in the number of infected individuals grows exponentially at rate r as stated in (3.33). This is essentially equivalent to standard linear stability analysis. We use this as an Ansatz from which we can derive the behaviour of the remaining variables.

We begin by considering θ. It is clear that the value of θ in this early stage is very close to 1 and that the amount that this differs from 1 by a factor that is linear in $[I]/N = (I)$. This is because to alter θ, a member of the population must become infected, and as θ is a proportion, this results in the division by N. Therefore in the early growth of the epidemic, θ can be approximated by:

$$\theta = 1 - K_\theta(I) \ , \tag{3.35}$$

where K_θ is a constant to be evaluated.

Following this chain of logic on, we consider functions of θ, such as $g(\theta)$. Now $g(\theta) =$

$g(1-K_\theta(I))$. Using Taylor's theorem, this can be written as

$$g(\theta) = g\Big(1 - K_\theta(I)\Big) = g(1) - K_\theta(I)g'(1) + O(I^2) \ . \tag{3.36}$$

Taking only the lowest order terms in (I) gives $g(\theta) = g(1)$, so any function of θ,

$$g(\theta) = g(1) \ , \tag{3.37}$$

in the early growth period.

Next the differential equation for (I) from (3.28) is considered. As $(I) = \tilde{I}\exp(rt)$, $(\dot{I}) = r(I)$. Additionally (3.28) gives, $(\dot{I}) = \tau(SI) - \gamma(I)$, therefore

$$(\dot{I}) = r(I) = \tau(SI) - \gamma(I) \implies (SI) = \frac{r+\gamma}{\tau}(I) \ . \tag{3.38}$$

Substituting this result into the equation for θ in (3.28) and using (3.37) gives

$$\dot{\theta} = -\frac{(r+\gamma)(I)}{g'(1)} \ . \tag{3.39}$$

This is integrable as $(I) = \tilde{I}\exp(rt)$ and when the integration is performed, using the initial conditions $I(0) = 0$ and $\theta(0) = 1$ we get the result that

$$\theta = 1 - \frac{r+\gamma}{rg'(1)}(I) \ . \tag{3.40}$$

Comparing this with (3.35) shows that $K_\theta = (r+\gamma)/rg'(1)$.

Next consider (SS). When all the population is susceptible $(SS) = \bar{n} = g'(1)$, as every linked pair in the population is an (SS) pair. To reduce the number of (SS) pairs, someone must become an infected, therefore (SS) is approximated by

$$(SS) = g'(1) - K_{(SS)}I \ , \tag{3.41}$$

where $K_{(SS)}$ is a constant to be found.

Considering again (3.28), the equation for (SI) can be re-written, noting that (SI) is $O\big((I)\big)$ and that $(\dot{SI}) = r(SI)$ by (3.38), to give,

$$(\dot{SI}) = \tau(SI)(SS)\frac{g''(1)}{g'(1)^2} - \tau(SI) - \gamma(SI) = r(SI) \ . \tag{3.42}$$

Using this equation r can be written in terms of the parameters τ, γ and properties of the network $g'(1)$ and $g''(1)$. Substituting (3.41) into this equation gives,

$$r(SI) = \tau\frac{g''(1)}{g'(1)}(SI) - \tau\frac{g''(1)}{g'(1)^2}K_{(SS)}(SI)(I) - \tau(SI) - \gamma(SI) \ . \tag{3.43}$$

Cancelling (SI) from both sides, and ignoring the term involving $K_{(SS)}(I)$, as there are terms of a lower order in (I), gives,

$$r = \tau\left(\frac{g''(1)}{g'(1)} - 1\right) - \gamma \ . \tag{3.44}$$

This relationship for r agrees with that given in Diekmann and Heesterbeek [2000].

Now (3.38) can be written as,

$$SI = \left(\frac{g''(1)}{g'(1)} - 1\right) I \ . \tag{3.45}$$

Finally to evaluate $K_{(SS)}$, the differential equation for SS from (3.28) is considered.

$$\dot{(SS)} = -K_{(SS)}\, r(I) = -2\tau (SI)(SS)\frac{g''(1)}{g'(1)^2} \ , \tag{3.46}$$

where (3.41) is used to give the first equality. Substituting (3.41) and (3.45) into the right hand side yields,

$$-K_{(SS)}\, rI = -2\tau \frac{g''(1)}{g'(1)}\left(\frac{g''(1)}{g'(1)} - 1\right)(I) + 2\tau \frac{g''(1)}{g'(1)^2}\left(\frac{g''(1)}{g'(1)} - 1\right) K_{(SS)} I^2 \ . \tag{3.47}$$

Cancelling an (I) on both sides and again ignoring the final term gives,

$$K_{(SS)} = 2\frac{\tau}{r}\frac{g''(1)}{g'(1)}\left(\frac{g''(1)}{g'(1)} - 1\right) = 2\tau g''(1)\left(\frac{g''(1)}{g'(1)} - 1\right)\frac{1}{\tau\big(g''(1) - g'(1)\big) - \gamma g'(1)} \ . \tag{3.48}$$

The following is the set of equalities that we have for the early growth period of the epidemic:

$$\begin{aligned} (I) &= \tilde{I} \exp rt \ , \\ \theta &= 1 - K_\theta (I) \ , \\ (SS) &= g'(1) - K_{(SS)}(I), \\ (SI) &= \big(g''(1)/g'(1) - 1\big)(I) \ , \end{aligned} \tag{3.49}$$

with,

$$\begin{aligned} r &= \tau\left(g''(1)/g'(1) - 1\right) - \gamma \ , \\ K_\theta &= (r+\gamma)/rg'(1) \ , \\ K_{(SS)} &= \frac{2\tau}{r}\big(g''(1)/g'(1)\big)\big(g''(1)/g'(1) - 1\big) \ . \end{aligned} \tag{3.50}$$

Note that for self-consistency of the Ansatz (3.33) we require that the value of r must be positive, as it is the growth parameter of the epidemic. This implies that,

$$\tau\left(\frac{g''(1)}{g'(1)} - 1\right) > \gamma. \tag{3.51}$$

3.1.5 Early growth variance

As detailed previously, the aim of this analysis is to calculate the variance of the amount of infection in the population in the early growth period of the epidemic. To progress towards this goal additional entities must first be defined.

Firstly let $\boldsymbol{\sigma}^2$ be the time dependent covariance matrix of our state variables Z_N. Define by $\mathbf{B}$ the Jacobian of the deterministic limit of our system (3.28), where the expression for the variables evaluated during the early growth period, detailed in (3.49), are input. Finally $\mathbf{G}$ is the local covariance matrix associated with an event i.e. when a susceptible node becomes infected, or an infected node recovers. This gives the covariance between all the state variables in $Z_N = (\theta, (I), (SS), (SI))$, so, for example the (1,2) entry of $\mathbf{G}$ gives the covariance between θ and (I).

Now if it is possible to write $\mathbf{G}$ as $\hat{\mathbf{G}}(I)$ where $\hat{\mathbf{G}}$ along with $\mathbf{B}$ being constant then the following equation from Dangerfield et al. [2009] may be used, which is again derived from the theoretical work of Kurtz [1970, 1971],

$$r\boldsymbol{\sigma}^2 - \mathbf{B}\boldsymbol{\sigma}^2 - \boldsymbol{\sigma}^2\mathbf{B}^T = [\hat{\mathbf{G}}\exp(rt) - \exp(\mathbf{B}t)\hat{\mathbf{G}}\exp(\mathbf{B}t)^T]\tilde{I} \ , \tag{3.52}$$

where r and $\tilde{I}$ are as previously defined.

The $\mathbf{B}$ matrix as stated above is the Jacobian matrix of the deterministic limit of the system (3.28), evaluated using (3.49). The Jacobian of (3.28) is given by

$$\mathbf{B} = \begin{pmatrix} -\tau(SI)\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{g'(\theta)} & 0 & 0 & -\frac{\tau}{g'(1)} \\ 0 & -\gamma & 0 & \tau \\ -2\tau(SS)(SI)\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{g''(\theta)}{g'(\theta)^2} & 0 & -2\tau(SI)\frac{g''(\theta)}{g'(\theta)^2} & -2\tau(SS)\frac{g''(1)}{g'(1)} \\ \tau(SI)\big((SS)-(SI)\big)\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{g''(\theta)}{g'(\theta)^2} & 0 & \tau(SI)\frac{g''(\theta)}{g'(\theta)^2} & \tau\Big(\frac{g''(\theta)}{g'(\theta)^2}\big((SS)-(SI)\big)-1\Big)-\gamma \end{pmatrix} . \tag{3.53}$$

Inputting the early growth approximations for the variables and working at $O(1)$, (rather than $O\big((I)\big)$) gives,

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & -\frac{\tau}{g'(1)} \\ 0 & -\gamma & 0 & \tau \\ 0 & 0 & 0 & -2\tau\frac{g''(1)}{g'(1)} \\ 0 & 0 & 0 & \tau\Big(\frac{g''(1)}{g'(1)}-1\Big)-\gamma \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & -\frac{\tau}{g'(1)} \\ 0 & -\gamma & 0 & \tau \\ 0 & 0 & 0 & -2\tau\frac{g''(1)}{g'(1)} \\ 0 & 0 & 0 & r \end{pmatrix} . \tag{3.54}$$

As shown in Kurtz [1970] and Ross [2006], $\mathbf{G}$ is calculated using the following expression,

$$\mathbf{G}_{ij} = \sum_e f_e(Z_N, l_e) l_{e,i} l_{e,j} \ , \tag{3.55}$$

where $l_{e,i}$ is the i-th entry in the change in variables vectors given by (3.30) and (3.31) and $f_e(Z_N, l_e)$ is the rate of the event e, which were also calculated previously. The aim is to find an expression for $\boldsymbol{\sigma}^2$, as this will give us the variance of the epidemic during the early growth phase.

Now follows specific details of how the various matrices in the (3.52) can be calculated. To calculate $\hat{\mathbf{G}}$, we calculate the matrix $\mathbf{G}$ and then divide by (I).

We begin by defining $\mathbf{F}_\tau$ and $\mathbf{F}_\gamma$ to be the outer products of (3.30) or two terms from

(3.31) These are given by,

$$\mathbf{F}_\tau = \frac{1}{N^2}\begin{pmatrix} \frac{\delta_{1,k}}{d_1^2} & -\frac{\delta_{1,k}}{d_1} & \frac{2x\delta_{1,k}}{d_1} & \frac{(y-x)\delta_{1,k}}{d_1} \\ -\frac{\delta_{1,k}}{d_1} & 1 & -2x & x-y \\ \frac{2x\delta_{1,k}}{d_1} & -2x & 4x^2 & 2x(y-x) \\ \frac{(y-x)\delta_{1,k}}{d_1} & x-y & 2x(y-x) & (x-y)^2 \end{pmatrix} \quad , \quad \mathbf{F}_\gamma = \frac{1}{N^2}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & x \\ 0 & 0 & 0 & 0 \\ 0 & x & 0 & x^2 \end{pmatrix} \ . \tag{3.56}$$

We can then decompose $\mathbf{G}$ as

$$\mathbf{G} = \tau \sum_{k,x,y} y(S_k) D^S_{x,y,k} \mathbf{F}_\tau + \gamma \sum_{k,x} (I_k) D^I_{x,k} \mathbf{F}_\gamma \ . \tag{3.57}$$

As, for example, (I) is the 2nd entry of Z_N and (SS) is the 3rd, the (2,3) entry of $\mathbf{G} = \mathbf{G}_{(I),(SS)}$. As an example we calculate $\mathbf{G}_{(I),(SS)}$ explicitly. As stated this is the (2,3) entry of $\mathbf{G}$, therefore we use the (2,3) entries of $\mathbf{F}_\tau = -2x$ and $\mathbf{F}_\gamma = 0$. Therefore,

$$\mathbf{G}_{(I),(SS)} = -2\frac{\tau}{N^2} \sum_{k,x,y} (S_k) xy D^S_{x,y,k} \ . \tag{3.58}$$

The sum over k can be separated from the sum over x and y, and the fact that the (1,1) moment of a multinomial distribution with variables k, p and q is given by $k(k-1)pq$ is used. Therefore

$$\mathbf{G}_{I,SS} = -2\frac{\tau}{N^2} \sum_{k,x,y} S_k \ k(k-1) \ \frac{[SS]}{\sum_k k[S_k]} \frac{[SI]}{\sum_k k[S_k]} \ . \tag{3.59}$$

Using that $[S_k] = N d_k \theta^k$ and $(S_k) = d_k \theta^k$ and then rearranging:

$$\mathbf{G}_{(I),(SS)} = -2\frac{\tau}{N^2} (SS)(SI) \frac{g''(\theta)}{g'(\theta)^2} \ . \tag{3.60}$$

When all the terms for the $\mathbf{G}$ matrix have been evaluated then inputting (3.49) gives the correct matrix for the early growth period. Doing this for $\mathbf{G}_{(I),(SS)}$ ignoring any terms of higher order than $(I)/N^2$, i.e. any term which is $O((I)^2/N^2)$ will be ignored. This gives the following expression for $\mathbf{G}_{(I),(SS)}$:

$$\mathbf{G}_{(I),(SS)} = -2\frac{\tau(I)}{N^2} \frac{g''(g''-g')}{(g')^2} + O((I)^2/N^2) \ , \tag{3.61}$$

using the fact that $g'(\theta)$ and $g''(\theta)$ will become $g'(1)$ and $g''(1)$ in the early growth period, and defining $g^{(n)} \equiv g^{(n)}(1)$. This then gives,

$$\hat{\mathbf{G}}_{(I),(SS)} = -2\frac{\tau}{N^2} \frac{g''(g''-g')}{g'^2} \ . \tag{3.62}$$

With the exception of the final entry, all entries of $\hat{\mathbf{G}}$ can be similarly evaluated, ma'king use of the assumption relating to the distribution of the states of neighbors of susceptible nodes, given by $D^S_{x,y,k}$ and what is known about $D^I_{x,k}$ from (3.26). The final entry is

an exception, as it is necessary to evaluate $\sum_{k,x,y} x^2 I_k D^I_{x,k}$. If the analogous assumption to the susceptible neighbourhoods is made for the neighbours of infected individuals, as detailed in (3.22), then this is easy enough to calculate by simply calculating moments. However, as previously mentioned, this assumption is known to be incorrect due to the age distribution of the infecteds. A more careful approach to this problem is given in the next section.

3.2 Neighbourhoods around an infected node

Firstly to give some more idea of why we must be more careful with the neighbours of infected nodes, we give the following argument.

3.2.1 Explanation of infected neighbourhood problem

The standard approach to simulating epidemics on networks is as follows. Begin by using the configuration model process, we let individual i have K_i stubs, and at time 0 the network adjacency matrix $A_{i,j}(t=0) = 0, \forall i, j$. Then the process is defined to take

$$(K_i, K_j, A_{i,j}, A_{j,i}) \to (K_i - 1, K_j - 1, A_{i,j} + 1, A_{j,i} + 1) \text{ at rate } \propto K_i K_j \ . \tag{3.63}$$

Running this process until the absorbing state $K_i = 0, \forall i$ gives the adjacency matrix of a configuration model network, although as previously described, different corrections of $O(N^{-1})$ due to self-edges and multiple-edges will arise in a finite population [Durrett, 2007].

Once the network is established and the epidemic process occurs upon it, individuals are in one of the compartments S, I or R and interact with each other as defined by the adjacency matrix of the network. The rate of infection to a susceptible with y infected neighbours is then given by τy and infectious nodes recover at rate γ.

A different way of understanding these two separate processes has been presented [Ball and Neal, 2008] in which they were combined into one process which assembles the network and spreads the infection at the same time. In this construction, every node in the network is given a number of half-links, which are then paired up as the epidemic progresses to form contacts between individuals in the following two ways:

- An infected with l remaining half-links makes contacts at rate τl; if it links to a susceptible then the infection will be transmitted with probability 1.
- When an infected recovers (which happens at rate γ) all of its remaining half-links will be paired off with randomly selected half-links in the population.

During the early growth period of the epidemic, if this process is paused at a time t and then the network is completed using configuration model methods, then this will be equivalent to constructing the network using the configuration model first and then running the epidemic on it until time t using the Gillespie algorithm, or some other suitable method. If the distribution of the disease state of neighbours around the infected nodes is then studied, it is clear from this explanation, that a node of a given degree k which has been infected for a given time a, will have a different distribution of neighbours than a node of equal degree which has been infected for a time b where $b \neq a$.

3.2.2 Neighbourhoods of infected nodes

Essentially, the reason that this is of interest is that to calculate $\mathbf{G}_{(SI),(SI)}$, the evaluation of the following is required

$$\chi := \sum_{k,x} x^2 I_k D^I_{x,k} \ , \tag{3.64}$$

and so the task is to determine the number of infectives of degree k, and the second moment of the distribution of the number of susceptibles around such infectives.

At the time of infection of a degree k node, the distribution of susceptible neighbours is given by $D^S_{x,k} = \sum_y D^S_{x,y,k}$, which assumes that the distribution of susceptible, infected and recovered neighbours are given by a multinomial distribution with probability of choosing a susceptible node, given by $[SS]/\sum_k k[S_k]$ or equivalently $(SS)/\sum_k k(S_k)$. However, we must take into consideration that the node which infected the central node, must be infected, therefore it will have at least 1 infected neighbour, so the number of susceptible neighbours is distributed as follows:

$$P(n_S = x) = \mathrm{Bin}\left(n_s = x | n = k-1, p = \frac{(SS)}{\sum_k k(S_k)} \right) \ , \tag{3.65}$$

where Bin implies that this is a binomial distribution with $n = k-1$ and $p = (SS)/\sum_k k(S_k)$.

(SS) is known in the early growth period (3.49), but (S_k) has not been calculated yet. To calculate (S_k), use again the fact that $(S_k) = d_k \theta^k$, which implies that

$$(\dot{S_k}) = d_k \dot{\theta^k} = d_k k \theta^{k-1} \dot{\theta} \ , \tag{3.66}$$

which implies that,

$$\int_0^t (S_k)(a) \mathrm{d}a = \int_0^t d_k k \theta (t-a)^{k-1} \dot{\theta}(t-a) \mathrm{d}a \ . \tag{3.67}$$

Inputting (3.28) and (3.49) to give $\dot{\theta}$ and θ respectively gives

$$\theta (t-a)^{k-1} \dot{\theta}(t-a) = -(1-K_\theta(I))^{k-1} \tau \left(\frac{g''(1)}{g'(1)} - 1 \right) \frac{(I)(t-a)}{g'(1)} \approx -\tau \frac{(I)(t) e^{-ra}}{g'} \left(\frac{g''(1)}{g'(1)} - 1 \right) . \tag{3.68}$$

(3.67) now becomes,

$$\int_0^t (S_k)(a) \mathrm{d}a = -\tau k d_k(I) \frac{(g''(1) - g'(1))}{g'(1)^2} \int_0^t e^{-ra} \mathrm{d}a \ . \tag{3.69}$$

The initial conditions are given by $(S_k)(0) = d_k$ and $(I)(0) = 0$. Performing the integrations and then using these initial conditions gives,

$$(S_k)(t) = d_k - \tau k d_k(I) \frac{(g''(1) - g'(1))}{r g'(1)^2} (1 - e^{-rt}) \ . \tag{3.70}$$

Therefore $(SS)/\sum_k k(S_k)$ becomes

$$\frac{(SS)}{\sum_k k(S_k)} = \frac{g'(1) - K_{(SS)}(I)}{\sum_k k\big(d_k - \tau k d_k(I)\frac{(g''(1)-g'(1))}{rg'(1)^2}(1-e^{-rt}\big)} = 1 + O\big((I)\big) \ , \tag{3.71}$$

where we have used that $g'(1) = \sum_k k d_k$.

The infected node infects its neighbours at rate τ, therefore if we consider a neighbour of an infected node of age a, the probability that they have not been infected by that specific infected node is given by $e^{-\tau a}$. However they may have been infected by another of their neighbours, which will be called global infection. If the neighbour is of degree l, then there are $l-1$ other links to consider. The probability that they have not been infected by a different neighbour is given by θ^{l-1}, as this is the probability of avoiding global infection down these $l-1$ links. Therefore the probability that they are still susceptible is given by $\theta^{l-1}e^{-\tau a}$.

In addition to this, the probability of a given neighbour having degree l is given by $ld_l/\sum_k kd_k$, as ld_l is the total number of links which belong to nodes of degree l and $\sum_k kd_k$ is the total number of nodes.

The probability that a given neighbour of an infected node is still susceptible can be worked out using the law of total probability,

$$\sum_l P(\text{neighbour is susceptible}|\text{degree} = l)P(\text{degree} = l) = \frac{1}{\bar{n}}\sum_l l d_l \theta^{l-1}e^{-\tau a} \ . \tag{3.72}$$

Now use the fact that $\sum_l l d_l \theta^{l-1} = g'(\theta)$ and that in the early growth period, $g'(\theta) \approx g'(1) = \bar{n}$, to see that (3.72) is approximately equal to $e^{-\tau a}$.

Therefore the probability of an infected of age a and degree k having x susceptible neighbours is given by:

$$P(n_s = x|\text{infected time } a \text{ ago}) = \text{Bin}\bigg(n_s = x|n = k-1, p = e^{-\tau a}\bigg). \tag{3.73}$$

Going back to (3.64), we must also calculate (I_k) in order to evaluate χ. To calculate (I_k), we note that the number of (I_k) nodes is given by the number of nodes of degree k which are no longer (S_k) and have not yet recovered (divided by the population size N).

$$-\frac{\mathrm{d}}{\mathrm{d}t}(S_k) = -d_k\frac{\mathrm{d}}{\mathrm{d}t}\theta^k = -Nd_k k\dot{\theta}\theta^{k-1} \ . \tag{3.74}$$

As the infected nodes have been infected at different time, this can be incorporated as the probability that an infective of age a (a node which was infected length of time a ago) is still infected is given by $e^{-\gamma a}$, as recovery takes place at rate γ. Therefore the value of (I_k) at a given time t is given by evaluating the following integral:

$$\begin{aligned}(I_k) =& -d_k k\int_0^t \dot{\theta}(t-a)(\theta(t-a))^{k-1}e^{-\gamma a}\,\mathrm{d}a \\ =& \tau k d_k(I)\frac{(g''(1)-g'(1))}{g'(1)^2}\int_0^t e^{-ra}e^{-\gamma a}\mathrm{d}a \ ,\end{aligned} \tag{3.75}$$

where the second equality is given by comparing this integral to (3.69).

This implies that equation (3.64) is given by the following expression:

$$\begin{aligned}\chi =& \sum_k \tau k d_k(I)(t) \frac{(g''(1)-g'(1))}{g'(1)^2} \int_0^t e^{-(r+\gamma)a} \Big(\sum_x x^2 \text{ Bin}(x|k-1, e^{-\tau a}) \Big) \mathrm{d}a \ , \\ =& \tau(I)(t) \frac{(g''(1)-g'(1))}{g'(1)^2} \int_0^t \sum_k k(k-1) d_k e^{-(r+\gamma)a} e^{-\tau a} \Big(1 + (k-2)e^{-\tau a} \Big) \mathrm{d}a \ ,\end{aligned} \tag{3.76}$$

where the second line is given by taking the second moment of the binomial distribution. Moving the summation over k and the terms k and d_k inside the integral allow us to use the pgf in order to simplify this expression as $g''(1) = \sum_k k(k-1)d_k$ and $g'''(1) = \sum_k k(k-1)(k-2)d_k$. We therefore have that,

$$\chi = \tau(I) \frac{(g''(1)-g'(1))}{g'(1)^2} \int_0^t g''(1) e^{-(r+\gamma+\tau)a} + g'''(1) e^{-(r+\gamma+2\tau)a} \mathrm{d}a \ . \tag{3.77}$$

This can be integrated to give the following result,

$$\sum_{k,x} x^2(I_k) D^I_{x,k} = (I) \frac{(g''(1)-g'(1))}{g'(1)^2} \left(\frac{\tau}{r+\gamma+\tau} g''(1) + \frac{\tau}{r+\gamma+2\tau} g'''(1) \right) \ , \tag{3.78}$$

at the leading order. $\mathbf{G}_{(SI),(SI)}$ is then given by (3.78) divided by N^2.

We have now calculated $\mathbf{G}$ by making an assumption about the infected neighbourhoods. However, this assumption is constrained by (3.26). The first equation is trivially satisfied as it is a probability distribution, however we must check that the second equation $\sum_k (I_k) \sum_{x,y} x D^I_{x,y,k} = (SI)$ is satisfied.

3.2.3 Satisfaction of constraints

We begin by defining the constraint to be given by ζ, where $\zeta := \sum_k (I_k) \sum_{x,y} x D^I_{x,y,k}$. Now comparing this with χ, we see that we have a very similar expression, so we therefore follow a similar process in order to evaluate it,

$$\begin{aligned}\zeta =& \sum_k \tau k d_k(I) \frac{(g''(1)-g'(1))}{g'(1)^2} \int_0^t e^{-(r+\gamma)a} \Big(\sum_x x \text{ Bin}(x|k-1, e^{-\tau a}) \Big) \mathrm{d}a \ , \\ =& \tau(I) \frac{(g''(1)-g'(1))}{g'(1)^2} \int_0^t \sum_k k(k-1) d_k e^{-(r+\gamma)a} e^{-\tau a}\end{aligned} \tag{3.79}$$

Again, simplifying by using the pgf and then integrating, we see

$$\zeta = \tau(I) \frac{g''(1)-g'(1)}{g'(1)^2} \frac{g''(1)}{r+\gamma+\tau} \ . \tag{3.80}$$

As $r = \tau\Big(\frac{g''(1)}{g'(1)} - 1\Big) - \gamma$ we have that $r+\gamma+\tau = \tau \frac{g''(1)}{g'(1)}$. Substituting this in to (3.80),

we get the following expression,

$$\zeta = \tau(I)\frac{g''(1) - g'(1)}{g'(1)^2}\frac{g'(1)g''(1)}{\tau g''(1)} = (I)\left(\frac{g''(1)}{g'(1)} - 1\right). \tag{3.81}$$

Comparing this with the expression for (SI) in the early growth period, given in (3.49), we see that they are equal, and therefore the assumption made about the neighbourhoods of infecteds preserve the self-consistent solution produced by the Ansatz (3.33).

Now that this remaining entry of $\hat{\mathbf{G}}$ has been calculated and has been shown to be consistent with the rest of the approach taken, we give the matrix in the results section which follows.

3.3 Results

Firstly we give the full $\hat{\mathbf{G}}$, and then the results gained from solving (3.52) to calculate the variance of the number of infections in the early stage of the epidemic.

3.3.1 Full $\hat{\mathbf{G}}$ matrix

$$\begin{aligned}
\hat{\mathbf{G}}_{\theta,\theta} &= \frac{\tau(g'' - g')}{g' d_1 N^2}, \\
\hat{\mathbf{G}}_{\theta,I} = \hat{\mathbf{G}}_{I,\theta} &= -\frac{\tau(g'' - g')}{g' N^2}, \\
\hat{\mathbf{G}}_{\theta,SS} = \hat{\mathbf{G}}_{SS,\theta} &= 0, \\
\hat{\mathbf{G}}_{\theta,SI} = \hat{\mathbf{G}}_{SI,\theta} &= \frac{\tau(g'' - g')}{g' N^2}, \\
\hat{\mathbf{G}}_{I,I} &= \frac{1}{N^2}\left(\tau\left(\frac{g''}{g'} - 1\right) + \gamma\right), \\
\hat{\mathbf{G}}_{I,SS} = \hat{\mathbf{G}}_{SS,I} &= -\frac{2\tau}{N^2}\frac{g''(g'' - g')}{g'^2}, \\
\hat{\mathbf{G}}_{I,SI} = \hat{\mathbf{G}}_{SI,I} &= \frac{1}{N^2}\left(\frac{g''}{g'} - 1\right)\left(\tau\left(\frac{g''}{g'} - 1\right) + \gamma\right), \\
\hat{\mathbf{G}}_{SS,SS} &= \frac{4\tau}{g' N^2}\left(\frac{g''}{g'} - 1\right)(g'' + g'''), \\
\hat{\mathbf{G}}_{SS,SI} = \hat{\mathbf{G}}_{SI,SS} &= \frac{2\tau}{g' N^2} g'''\left(\frac{g''}{g'} - 1\right), \\
\hat{\mathbf{G}}_{SI,SI} &= \frac{1}{N^2}\left(\frac{g''}{g'} - 1\right)\left(\tau\left(\frac{g''' - g'' + g'}{g'}\right)\right) \\
&\quad + \gamma\frac{(g'' - g')}{N^2 g'^2}\left(\frac{\tau}{r + \gamma + \tau} g'' + \frac{\tau}{r + \gamma + 2\tau} g'''\right).
\end{aligned} \tag{3.82}$$

This is the remaining piece needed to fully detail (3.52) which contains the covariance between all of the variables in Z_N, and specifically the variance of (I). We can now solve this algebraically.

3.3.2 Early growth variance

As described above (3.52) must be solved for $\boldsymbol{\sigma}^2$. This is too complicated to do by hand, and so computer algebra in Mathematica was used to complete this calculation. The full expression for the variance during the early growth period is extremely complicated. This is seen below in (3.83), though it is not informative due to its complexity:

$$\begin{aligned}\mathrm{Var}(I) =& \frac{\tilde{I}}{N^2(2\gamma+r)}\left(\frac{\tau e^{-2\gamma t}(e^{t(\gamma+r)}-1)}{\gamma+r}\left(\frac{\tau^2(g'-g'')(e^{t(\gamma+r)}-1)(g'-\frac{g''\tau}{\gamma+\tau}+\frac{\gamma g'''}{\gamma+2\tau}+g''')}{g'^2(\gamma+r)}\right)\right.\\
&\frac{\tau e^{-2\gamma t}(e^{t(\gamma+r)}-1)}{\gamma+r}\Big(\frac{g''}{g'}-1\Big)\Big(\gamma+\tau\Big(\frac{g''}{g'}-1\Big)\Big)+e^{rt}\Big(\gamma+\tau\Big(\frac{g''}{g'}-1\Big)\Big)+\\
&\frac{2\tau e^{rt}}{\gamma}\frac{\tau^2(g'-g')e^{\gamma(-t)}(e^{t(\gamma+r)}-1)(g'-\frac{g''\tau}{\gamma+\tau}+\frac{\gamma g'''}{\gamma+2\tau}+g''')}{g'^2(\gamma+r)}-\\
&\frac{2\tau e^{rt}}{\gamma}\frac{\tau^2(g'-g'')(e^{rt}-1)(g'(\gamma+2\tau)(\gamma+\tau)-g''\tau(\gamma+2\tau)+2g'''(\gamma+\tau)^2)}{g'^2r(\gamma+\tau)(\gamma+2\tau)}+\\
&\frac{2\tau e^{rt}}{\gamma}\left(\frac{(g'-g'')e^{\gamma(-t)}(g'(\gamma-\tau)+g''\tau)}{g'^2}+(\frac{g''}{g'}-1)(\gamma+\tau(\frac{g''}{g'}-1))\right)-\\
&\left.\frac{e^{-2\gamma t}(g'(\gamma-\tau)+g''\tau)(\gamma g'+\tau(g''-g')e^{t(\gamma+r)}+g'r+g'\tau-g''\tau)}{g'^2(\gamma+r)}\right)\end{aligned} \tag{3.83}$$

As the population size approaches infinity, this can be simplified by taking a large t limit. This is not appropriate in smaller populations, as the length of time until we leave the early growth period due to susceptible depletion is too small to take this limit. As the population gets larger however, this length of time will increase enough that the dynamics are dominated by the large t limit.

Defining t_{early} as the time at which the epidemic begins to grow at the rate predicted in Diekmann and Heesterbeek [2000], and the time at which the depletion of susceptibles affects the growth rate and we leave the early growth phase at t_{depleted}. When the current time satisfies $t_{\text{early}} \ll t \ll t_{\text{depleted}}$, the expression for the variance of the number of infecteds can be simplified. This regime will exist in an extremely large network if the initial amount of infection in the network $\tilde{I}$ is sufficiently small, but in a smaller population, this regime does not exist, again due to the length of time spent in the early growth phase.

In this limit, the mean and variance of prevalence obey the following expression,

$$\begin{aligned}&\mathrm{Mean}(I) \to \tilde{I}e^{rt}\ ,\ \text{ for }\quad r=\tau(\frac{g''}{g'}-1)-\gamma\ ,\\
&\frac{\mathrm{Var}(I)}{\mathrm{Mean}(I)^2} \to \frac{\tau g'(2g'''(\gamma+\tau)^2+(\gamma+2\tau)((\gamma+\tau)g'-\tau g''))}{N^2(\gamma+\tau)(\gamma+2\tau)(g'-g'')((\gamma+\tau)g'-\tau g'')}\ .\end{aligned} \tag{3.84}$$

where $\tilde{I}$ is a constant related to the prevalence of infection as the early asymptotic behaviour commences at t_{early}, and $g^{(n)} \equiv g^{(n)}(1)$.

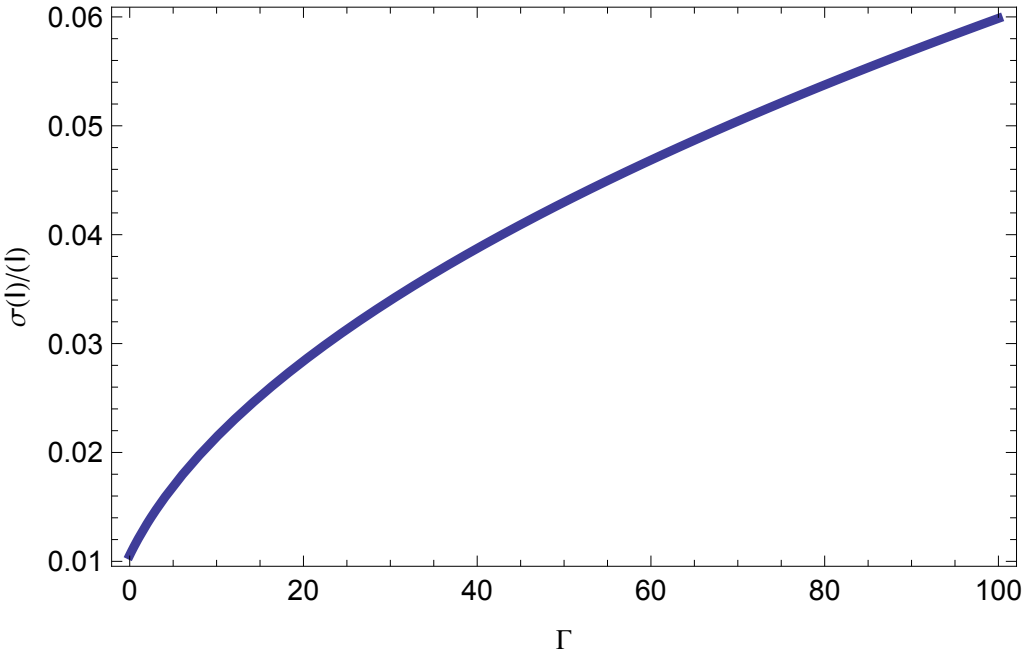


Figure 3.1: Early asymptotic dependence of the standard deviation of infection prevalence, divided by infection prevalence, on the skew Γ of the network degree distribution. The curve is plotted from equation (3.84). Parameter values are: mean degree = 10; variance in degree = 100; transmission rate $\tau = 0.05$; recovery rate $\gamma = 0.1$. Skewness of degree is varied between realistic values (0 and 100) to see how the variance of the asymptotic early prevalence of infection is affected. We see that as the skewness is increased we get a higher variability of prevalence. This is as expected, since the higher the skew, the more neighbours the most connected individuals of the population have, reducing the predictability of the epidemic due to chance events amongst this small but epidemiologically influential group.

Figure 3.1 shows how the variance in prevalence changes as skew in degree increases. Seeing that as the skew increases we get an increase in the variance of the epidemic during early growth, we would therefore see that if we had a network whose degree distribution is a power law, would show greater variation than a negative binomially distributed network with the same mean and variance would show. We think of this increase in variation being caused by the members of the population who are very well connected in the network. These people have been called super-spreaders in the past [Stein, 2011; Galvani and May, 2005; Meyers et al., 2005]. If the disease reaches these people in the early growth phase of the epidemic then we can expect a rapid increase in disease prevalence, as there will be many SI pairs through which the disease can be passed, whilst if they do not get it in this early phase, there will not be this rapid increase, which generates the large variance in this stage of the epidemic. We note that for the $\tilde{I}$ term in (3.84) we have no analytical traction on what this should be for a given epidemic. If we wish to compare this result to simulation, we will fit the value of $\tilde{I}$ so that the analytical prediction and the simulated results agree at a given point. As this is simply multiplied by the other terms, fitting it simply scales the prediction up or down rather than fine tuning the prediction itself.

3.4 Comparison with simulation

To see the accuracy of the predictions detailed, we compare them to simulations on networks. To construct these networks, we use the configuration model [Molloy and Reed, 1995]. This process is described next.

3.4.1 Network construction

To construct an uncorrelated network with a given degree distribution, $P(k)$, the configuration model is used [Molloy and Reed, 1995; Newman, 2003].

To realise this for finite N, the following method is employed:

- Draw as many numbers from the degree distribution $P(k)$ as there are members of the population. If the sum of these numbers is odd, then randomly select a previously drawn number and reduce it by 1.
- Assign each member of the population one of these drawn numbers which then corresponds to the number of 'half-links' they are given.
- Starting with the first node, which has k half-links say, pair each of the k half-links up with another from the population. This corresponds to adding 1 to the corresponding entry in the adjacency matrix $\mathbf{A}$, so if we join up a link from node i to node j, then we add one to $A(i,j)$ and $A(j,i)$ as the links are undirected.
- Repeat until all half-links have been paired up.

This produces an uncorrelated network as all pairs are generated independently, conditioned on degree. Therefore if we know that nodes l and m are connected and that nodes m and n are connected, then the probability of nodes l and n being connected is equal to that of any two randomly chosen nodes one of which has the same degree as node l and the other the same degree as node n.

In theory, if this process is carried out on an infinite population, then no repeated edges will be produced. However for a finite network, multiple edges between nodes or self-edges will be generated at a rate proportional to $1/N$. This will correspond to a value greater than one in $\mathbf{A}$ or a non-zero entry on the diagonal of $\mathbf{A}$.

Self links in the network are not permitted and in this disease context would have no meaning, as you cannot infect yourself if you are susceptible. Therefore allowing self edges would simply alter the effective degree distribution. As an example, if everyone had one of their edges to themselves, comprising of 2 half-links, then the disease would effectively be spreading on a network whose mean number of neighbours was 2 lower than if these links were to others members of the population. Multiple links between nodes (which would be indicated by integer values > 1 in the adjacency matrix) are also not included in this model. We therefore need a method to eliminate these.

3.4.2 Removal of network defects

There are several available methods for dealing with these network defects, the method we use follows. The first step in the process of removing these defects is to obtain a list of

all the nodes which have multiple connections between them and self-edges. The method for the multiple edges is slightly different from the self edge case.

For the multi edge case, one at a time the extra links between nodes are broken, say between node i and j, and then a randomly selected and connected pair of nodes will also be selected, say nodes i' and j'. We then connect i and i' together and j and j' together, which leaves the degree distribution unaltered. This is then repeated until there are no repeated- and self-edges. This process can be seen in figure 3.2.

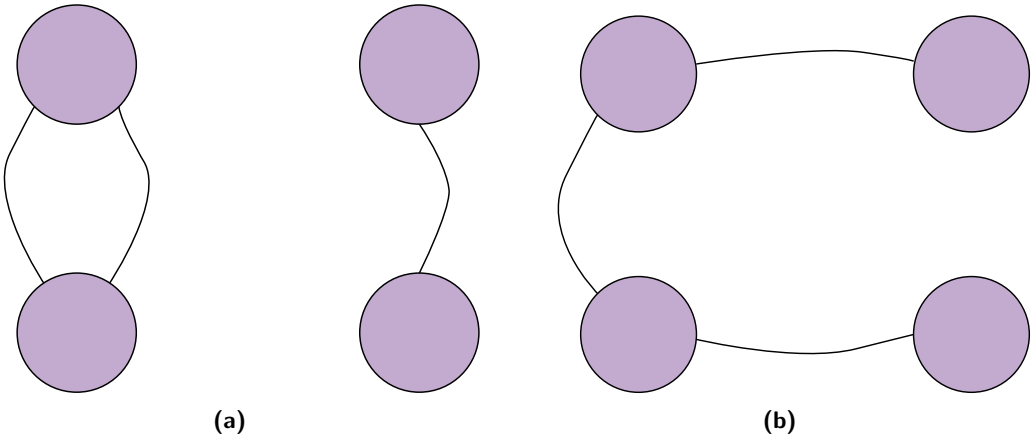


Figure 3.2: Demonstration of removal of multiple connections defect. (a) shows two nodes which have more than one connection between them, along with another randomly selected pair of nodes. In (b) one of the repeated connections between the first pair of nodes is broken, along with the connection between the randomly selected pair of nodes. The broken half-links from the first pair of nodes are then paired with the broken half-links from the second pair of nodes.

To remove self-links from a node i, a pair of nodes j and k are selected. This link is then broken and both j and k are connected to node i. This again leaves the degree distribution unaltered and the process is demonstrated in figure 3.3.

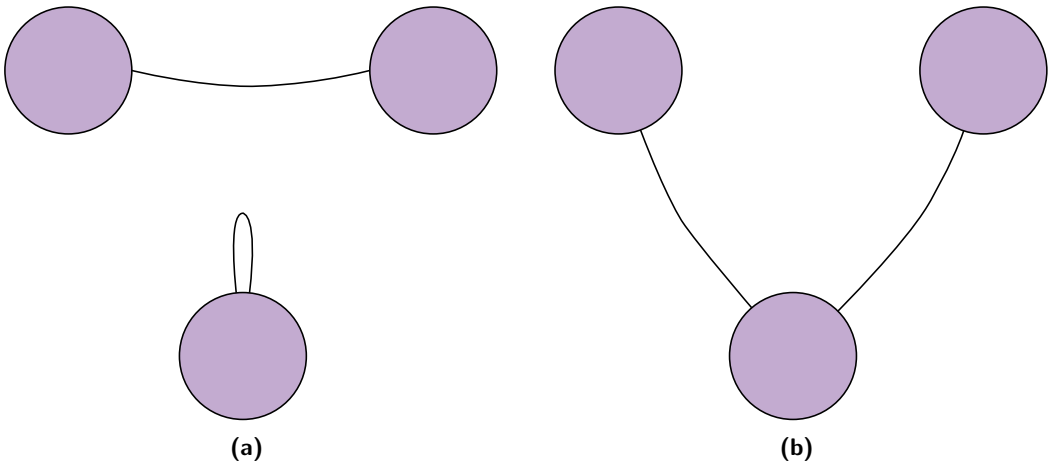


Figure 3.3: Demonstration of removal of self-link defect. (a) shows a node with a self-link and a pair of connected nodes. (b) The self-link is broken, leaving its two constituent half-links, which are then connected to the two half-links that are made from breaking the link between the two connected nodes.

In theory, it would be possible for the process explained above to fail at some point due to lack of available node combinations needed to alter the defects of the network, though the probability of this depends on the degree distribution and the size of the network. For example if the degree distribution allows nodes to have more connections than the population size, if a node did have more connections than the population size, then this process would obviously fail to produce a network with no multiple links in it. However for any networks that we consider here, this process never fails, as the probability of failure is ver low, due to the size of the population and the degree distributions considered.

3.4.3 Results of simulation

The way that we have compared our analytical results to simulation is by considering the impact of the skew of the degree distribution on the variance of the number of infecteds in the early growth period. To do this we consider networks whose degree distributions have the same mean and variance but different skews. It is expected that the network which has the larger skew will also demonstrate a larger variance during this early phase of the epidemic. We also compare the analytical prediction to the one which is generated by our simulations to see how accurate the analytical prediction is.

There are several difficulties to achieving this that we wish to note. Firstly, the analytical results depend on the network being extremely large. There is obviously a limit to the size of network that we can consider and the simulations here were run on networks of size 10^5, which was sufficiently large for our purposes.

Secondly, the analytical results only work for the early growth phase of the infection, which can be defined as the time when the pool of susceptibles is not significantly depleted. Whilst in an arbitrarily large network this can last for an arbitrarily long time, in a finite network, this is not true and will vary depending on the degree distribution of the network along with its size N. Along with this, as previously noted, we have to allow some time at the beginning of the epidemic in which the impact of the initial conditions

have diminished and we are in a period which the system has approached its average early asymptotic behaviour. Together these imply that in a finite network, this period may be very short, or may not exist at all. In the networks that we consider, this period does exist and is defined to be the period in which the growth in the number of infecteds is the same as predicted in Diekmann and Heesterbeek [2000] (given in (3.84)).

Finally, the assumptions made for the analytical system such as how the number of triples in the network can be approximated by doubles and the pgf $g(x)$, will be inexact in the finite network that we construct. There is no guarantee that this will be appropriately accurate; convergence is $O(N^{-1})$ but we do not have knowledge of the pre factors.

To simulate the epidemics, the Gillespie algorithm [Gillespie, 1977] on networks of size 10^5 is used. To converge onto the average early growth, we allow each simulation to achieve a certain number of infections (10^2) and then we set the simulation time to zero and let the epidemic progress from there. In the system of size 10^5, allowing 10^2 infections is also small enough that the susceptibles will not be depleted significantly enough to affect the rate of growth of the epidemic. We ran 10^3 simulations on two networks with the same mean and variance but with different skew.

Figure 3.4 shows the result of these simulations compared with the prediction given by (3.83). We can see that as predicted the network with the higher skew exhibits more variance in number of infecteds in the early growth stage and also we see that the analytical prediction and the simulations have a good agreement.

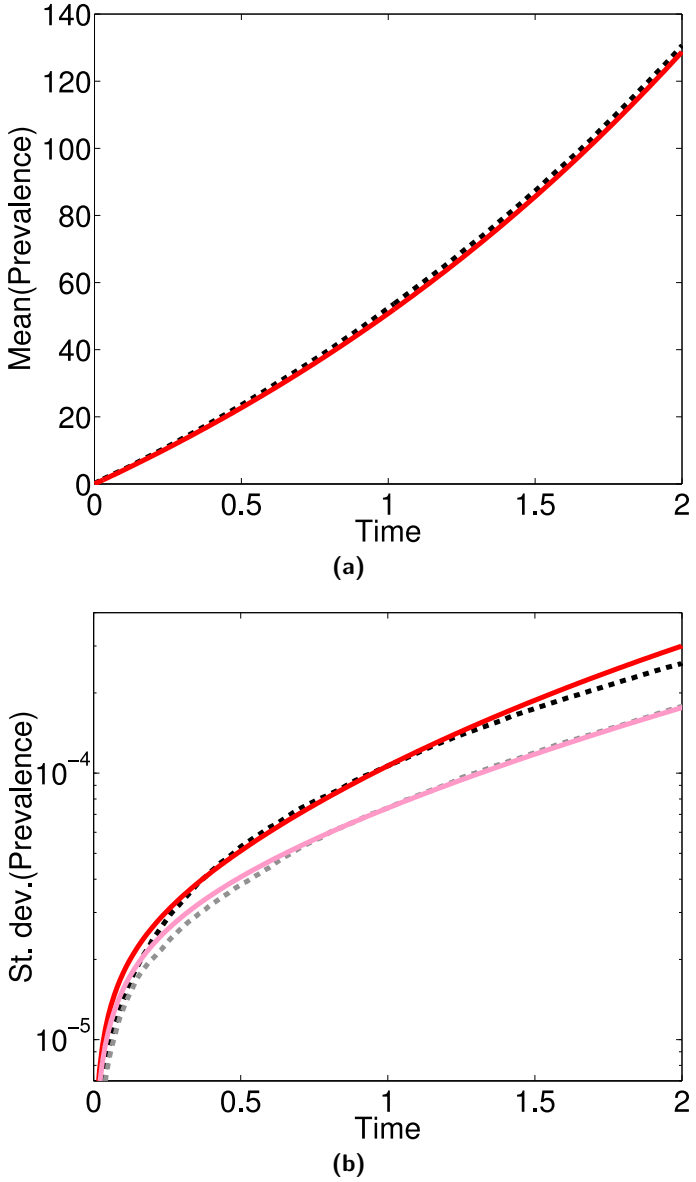


Figure 3.4: Comparison of simulated results to analytical predictions. Dashed lines are simulations and full lines are analytical predictions. Simulations are on two different networks, which have the same mean and variance for their degree distribution (mean $\approx$ 5.4 and variance $\approx$ 67.2) but different skewness: 24.3 for red / black lines and 6.7 for the pink / grey lines. Transmission and recovery rates are $\tau = 0.0408$ and $\gamma = 0.1$ respectively. (a) shows a period of time at which we have agreement in the growth of the number of infecteds between the two networks that is strongly in agreement with the theoretical prediction from Diekmann and Heesterbeek [2000]. (b) is also taken for this time and we can see that the theoretical prediction that we have described previously in this paper deviates slightly from simulation.

3.5 Branching process approximation

Along with the pairwise assumption that has just been detailed, it is possible to approach this problem of the early growth period by imbedding a branching process in the full dynamics. We can then derive results about what variance we would observe in the k-th generation of infecteds.

We define the degree of a node to be denoted by a random variable D, which follows some specified distribution. The probability that a neighbour of yours has degree k is then given by $kP(D = k)/\mathbb{E}(D)$, where $P(D = k) = d_k$ and $\mathbb{E}(D) = \bar{n}$. Then we consider a discrete time Reed-Frost epidemic on the configuration model of the network given by our degree distribution. This is a discrete time model, meaning that we are concerned with the number of invectives in successive "generations" of the disease. We note that if you are an initial infective, then you are in generation 0 of the epidemic, if you are infected by an initial infective, then you are in generation 1 and so on. The number of people that each infected individual with degree k in generation n infects, is then chosen from a binomial distribution, with $n = k - 1$ and probability p. As we are assuming no clustering in the population, and are in the early stage of the epidemic, then the only non-susceptible neighbours that an infected individual has is the individual who infected them. Again corrections to this are $O(1/N)$.

Suppose that we have I_n infected individuals in generation n. Then in generation $n + 1$, there will be $\sum_{i=1}^{I_n} B_i$ infectives, where B_i is the number of 'children' of each infective i.

The B_i's are independent and identically distributed. To progress from here we first calculate the probability generating function for the B's. To do this we calculate,

$$\mathbb{E}(s^B) = \sum_{b=0}^{\infty} s^b P(B = b). \tag{3.85}$$

Using the law of total probability we get that:

$$\begin{aligned} P(B = b) &= \sum_{k=0}^{\infty} P(B = b|D = k)P(D = k), \\ &= \sum_{k=0}^{\infty} \binom{k-1}{b} p^b (1-p)^{k-1-b} k \frac{d_k}{\mathbb{E}(D)} \ . \end{aligned} \tag{3.86}$$

Substituting this into (3.85) and swapping the order of summation to get,

$$\begin{aligned} \mathbb{E}(s^B) &= \frac{1}{\mathbb{E}(D)} \sum_{k=0}^{\infty} \left(\sum_{b=0}^{k-1} \binom{k-1}{b} (sp)^b (1-p)^{k-1-b} \right) k d_k \ , \\ &= \frac{1}{E(D)} \sum_{k=0}^{\infty} k d_k (1 - p + sp)^{k-1} \ , \\ &= \frac{1}{E(D)} g'(1 - p + sp) \ . \end{aligned} \tag{3.87}$$

From this, we get that the mean and variance in B are given by

$$\begin{aligned}\mathbb{E}(B) =& \frac{p}{\mathbb{E}(D)} g''(1) = \frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)} \ , \\ \text{var}(B) =& \frac{\mathbb{E}(D(D-1)(D-2))p^2}{\mathbb{E}(D)} + \frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)} - \left(\frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)}\right)^2\end{aligned} \tag{3.88}$$

The expectation for the number of invectives in generation k follows and is given by,

$$\mathbb{E}(I_k) = \mathbb{E}\sum_{i=1}^{I_k-1} B_i = \mathbb{E}(B)E(I_{k-1}) = \mathbb{E}(B)^k I_0 \ . \tag{3.89}$$

The law of total variance then tells us that the variance of I_k is given by,

$$\begin{aligned}\text{var}(I_k) &= \mathbb{E}\left(\text{var}\left(\sum_{i=1}^{I_{k-1}} B_i | I_{k-1}\right)\right) + \text{var}\left(\mathbb{E}\sum_{i=1}^{I_{k-1}} B_i | I_{k-1}\right) \ , \\ &= \mathbb{E}\left(\sum_{i=1}^{I_{k-1}} \text{var}(B_i)\right) + \text{var}(E(B_1)I_{k-1}) \ , \\ &= \mathbb{E}(I_{k-1})\text{var}(B) + \text{var}(I_{k-1})\mathbb{E}(B)^2 \ , \\ &= I_0\mathbb{E}(B)^{k-1}\text{var}(B) + \text{var}(I_{k-1})\mathbb{E}(B)^2 \ .\end{aligned} \tag{3.90}$$

Substituting this expression for $\text{var}(I_i)$ for decreasing values of i gives us

$$\begin{aligned}\text{var}(I_k) &= I_0\mathbb{E}(B)^{k-1}(1 + \mathbb{E}(B) + \ldots + \mathbb{E}(B)^{k-1})\text{var}(B) \ , \\ &= I_0\mathbb{E}(B)^{k-1}\frac{\mathbb{E}(B)^k - 1}{\mathbb{E}(B) - 1}\text{var}(B) \ .\end{aligned} \tag{3.91}$$

When we substitute in the the expression for the mean and variance of B from (3.88) we get,

$$\begin{aligned}\text{var}(I_k) =& I_0\left(\frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)}\right)^{k-1}\left(\left(\frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)}\right)^k - 1\right)\frac{1}{\frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)} - 1} \\ & \left(\frac{\mathbb{E}(D(D-1)(D-2))p^2}{\mathbb{E}(D)} + \frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)} - \left(\frac{\mathbb{E}(D(D-1))p}{\mathbb{E}(D)}\right)^2\right) \\ =& I_0\left(\frac{g''p}{g'}\right)^{k-1}\left(\left(\frac{g''p}{g'}\right)^k - 1\right)\frac{1}{\frac{g''p}{g'} - 1}\left(\frac{g'''p^2}{g'} + \frac{g''p}{g'} - \left(\frac{g''p}{g'}\right)^2\right) \ .\end{aligned} \tag{3.92}$$

Now, letting k become large, and setting $E(B)^k = e^{rt}$, $p = \tau/(\tau+\gamma)$, to correspond as closely as possible to continuous-time results, we obtain

$$\text{var}(I_k) \to I_0 e^{2rt}\frac{g'g'''\tau^2 + g'g''\tau(\tau+\gamma) - (g'')^2\tau^2}{g'(\tau+\gamma)(\tau g'' - (\tau+\gamma)g')} \ , \tag{3.93}$$

which is not in agreement with (3.84). This prediction shows pure exponential growth in the variance of the number of infected individuals as t increases, which is not what is seen in the previous prediction, or in the simulations performed.

This is as we would expect, because the reasoning in §3.1 above depends critically on the Markovian nature of the dynamics. In contrast, the branching process can account for non-Markovian dynamics, but does not allow for the fact that high-degree infective nodes create new infections more quickly than low-degree infective nodes. The two approaches are therefore best seen as complementary.

3.6 Summary

In this chapter we have derived an approximation to the variance in the number of infected individuals during the early growth period of an epidemic, on a heterogeneous network. This involved making a previously formulated assumption [Volz, 2008], which has since been shown to be correct [Decreusefond et al., 2012] about the neighbourhoods of susceptible individuals. It also involved the formulation of an assumption related to the number of susceptible neighbours that infected individuals have, and the demonstration that this was consistent with constraints that came from the exact unclosed dynamics of an epidemic on a network (3.2).

Once this was done, the prediction that we derived was tested against simulations preformed using Gillespie's algorithm [Gillespie, 1977] on a configuration model network [Molloy and Reed, 1995], and was shown to have a strong agreement with the variance displayed by these simulations.

Finally, a branching process approximation was also derived, and this was seen to be qualitatively different from the one derived using our previous assumptions. This was as we expected, as by taking the branching process approximation, we are ignoring crucial information about the timings of events, and therefore something as time dependent as the variance during the early growth would not be expected to have much agreement with results derived using this process.

Chapter 4

The impact of workplace size distribution on disease spread

4.1 Introduction

As we have seen in the previous chapter, the heterogeneity which is contained within the network, in terms of its degree distribution can have a large impact on the epidemic. For human populations, when it comes to diseases which can be spread without the need for intimate contact, or the sharing of bodily fluids, the main source of this heterogeneity in contact patterns is in the workplaces. This is due to the fact that the number of contacts that a person has in work will depend on many things, including the size of the workplace, along with the type of work undertaken, whilst households are generally of a similar size and have a much smaller maximum size than workplaces, and school class sizes are again less variable than workplace sizes.

Workplace contacts are simply any contacts made at a place of work. For our purposes, the interactions between children and between children and adults at schools are not included here, but those between staff at schools are. Workplace contacts will, in most cases, differ from contacts made at home, as there will be limited or no physical contact. However, time spent together will be larger than in other situations such as travelling and shopping (as evidenced by the UK time use survey [UK Data Service]), implying greater epidemiological impact than these situations. Additionally, many contacts will be repeated, so the probability of an infection passing between two people who make contact in a workplace is larger than two people who make a transitory contact.

Workplaces are also the place where the majority of adult-adult mixing will occur (schools for child-child, homes for child-adult), and so can be epidemiologically important in terms of control of epidemics. Alongside households and schools, they are thought of (with good reason) as the main magnifiers of disease, and in general the mixing of a large number of people can have a significant affect on the spread of an infection [Lloyd-Smith et al., 2005; Meyers et al., 2005] and have been included in studies of the spread of epidemics e.g. Mossong et al. [2008]; Ferguson et al. [2006]; Pellis et al. [2008]. The potential impact of these large gatherings is clear, due to the fact that the more people who are contacted by an infectious person, the greater the probability is that the infection will be spread.

This chapter considers the impact of the distribution of workplace sizes on the possible final

size of an epidemic which takes place in workplaces. We are interested in this question, as in many cases, when data is not available, modellers are forced to choose a workplace size distribution based on data which is from a different country, which may or may not be representative of the newly studied country [Ferguson et al., 2006]. This chosen distribution may have a great influence on the epidemic, as in some way, it imposes the contact patterns that take place within workplaces and may lead to a large over or underestimation of the possible spread within workplaces.

To tackle this problem we consider how postulated, size dependent, infection rates in workplaces can change the predicted final size of an epidemic in workplaces. Along with this we investigate the distribution of the number of employees in workplaces throughout the UK, which is given by a dataset provided by Blue Sheep [Bluesheep data source]. This is a private data source, which combines data from 20 separate sources to give, amongst many other datasets, the distribution of workplace sizes in the UK. This dataset is updated monthly and a version from February 2012 was used for this work.

As the impact of this distribution on the spread of disease is the question of interest, we discuss methods which can be used to fit different distributions to the data, anticipating likely choices that modellers would take if faced with modelling workplaces with no data to inform the distribution of workplace sizes. Next, we combine these fits to distributions with the dependence on transmission rates on the spread of the epidemic, and calculate the expected number of infected individuals in each circumstance, allowing us to quantify the effect of these distributions on the spread of the disease. Finally, the sensitivity of the attack rate of the epidemic within workplaces to the workplace size distribution is considered, and the attack rates are compared with those expected in differently distributed collections of individuals (such as in households) with similar transmission rates.

4.2 Transmission rates and their impact on the final size of an epidemic

We are interested in how the distribution of workplace sizes can impact the spread of an epidemic. In particular we are interested in how much difference having a large number of workplaces with a high number of employees can make to the spread of infection. It may be expected that these may be multipliers of infection, as if there are more people to contact and spread the epidemic onto, then there will be more chance of spreading the epidemic, along with increasing the probability of super-spreading events.

To investigate this we consider the spread of an SIR type infection through a workplace as being modelled by the standard set of SIR equations with removal rate γ and transmission rate given by $f(N)$, where N is the population size. This is modelled by a set of three differential equations, also given previously in (2.1), though with a different transmission rate here,

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -f(N)SI \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= f(N)SI - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= -\gamma I \ . \end{aligned} \tag{4.1}$$

Here we have that the value of $R_0 = f(N)/\gamma$. For the classic mean field SIR equations of Kermack and McKendrick [1927], which is the form given in (2.1), $f(N) = \beta/N$. This term can be thought of as the product of the contact rate between individuals and the probability of transmission between contacts, which is assumed to be invariant to any change in population size. For this formulation of $f(N)$, the contact rate can is inversely proportional to the population size. This is described as frequency dependent transmission i.e. the rate of transmission does not alter when the population density increases, it only depends on the number of contacts, as in this scenario, increasing the population N will increase the numbers of susceptibles to contact proportionally, the division by N will counteract this change, leaving the total rate of transmission unchanged.

In large populations, the assumption of frequency dependent transmission is intuitive, and has been shown to be representative, for example, with the spread of measles in England and Wales [Bjørnstad et al., 2002]. This is due to the fact that it is unexpected that on average people will infect greatly more people if they live in a large city such as London as opposed to a smaller city or town. This is because patterns of behaviour relating to contact are similar, and so will be averaged out over the whole population, giving a value of R_0 which is similar in both large and small towns or cities.

Unlike for a large population like a town or city, this is not necessarily a good assumption for a workplace, as it is likely that there is some density dependence in the transmission rate, leading to more infected individuals as the number of potential infectees (N) increases e.g. on large shop floors or open plan offices where there are more opportunities for contact. When the size of the population is increased towards infinity, R_0 is usually kept constant, by using a frequency dependent transmission term meaning that an infected person will, on average, continue infecting the same number of people in a large workplace as a small one, which is unlikely to be accurate for workplaces. To take an extreme example a workplace of 4 people means that the maximum number of people you can ever infect in the workplace is 3, whilst if you work in a large workplace, then there is much more opportunity to spread the epidemic and infect more than 3 people.

By altering $f(N)$ we can model different amounts of transmission in the population. This can be used to investigate the impact of workplace size on the level of infection in the population and can lead to larger or smaller epidemics than the one that would be observed if we consider frequency dependent transmission, where $f(N) = \beta/N$. We are therefore interested in considering the final size of epidemics in a population.

The final size of the epidemic in a population can be derived from (4.1) and is indicative of the level of spread achieved by the infection. To derive this, we divide the equation for I by the one for S and integrate to derive the well established equation for the final size of an SIR epidemic in a population. This is denoted by R_∞, and the result is given by the following implicit equation,

$$R_\infty = (1 - e^{-R_0 R_\infty}) \ , \tag{4.2}$$

where $R_0 = f(N)/\gamma$ is the reproduction number. To derive this equation, the conditions of $S(0) = N$, $I(0) = 0$ and $I(\infty) = 0$ are used along with the fact that $R_\infty = S(0) - S(\infty) = N - S(\infty)$. To find the proportion of the population which will be infected by the end of the epidemic, R_∞, root finding techniques are used on (4.2).

We note here that we are treating each workplace as a separate population in which the epidemic takes place in. By doing this we can calculate the value of R_∞ for each individual workplace and then work out what this means in terms of number of infecteds at this

workplace. This is at best a crude approximation as in general the mean field equations are held to be true as the population size tends towards infinity. This is clearly not the case for the workplaces (the largest workplaces are capped to 7,500 to be in agreement with the Blue Sheep data), but is a useful tool that we can use to investigate how the distribution of workplace sizes can have an influence on the number of cases of an epidemic.

To investigate the effect of the transmission rate on the epidemic, we define $f(N) = \beta/N^{1-\epsilon}$. This means that the mean field equations are adjusted to the following,

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta\frac{SI}{N^{1-\epsilon}} \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta\frac{SI}{N^{1-\epsilon}} - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= -\gamma I \ . \end{aligned} \tag{4.3}$$

The addition of the ϵ to the equations adds the ability to tune the affect of increasing the workplace size. From this set of ODEs we have that $R_0 = \beta/\gamma N^{1-\epsilon}$, allowing us the adjust the number of expected infections in a workplace of size N, by altering ϵ. By changing ϵ from 0 to 1, we tune between mass action, or frequency dependent transmission, and density dependent transmission. Primarily we are interested in the change to R_0 and R_∞ as ϵ is increased. In reality, the actual number of contacts you make will have a strong dependence on the size of workplace that you are in, but there may be some threshold at which you may not make any more contacts. For example you are unlikely to make more contacts in a workplace with 5,000 employees as opposed to 1,000, though no limit on the number of possible contacts has been modelled here.

Figure 4.1a shows the increase in R_0 due to increasing ϵ for various populations of size 100, 1,000 and 10,000.

For a given value of N, we set $\epsilon = 0$ and consider the scenario where $R_0 = 1$. With $\epsilon = 0$, we have that $R_0 = \beta/\gamma N$. This therefore defines the relationship between β, γ and N, as we must have $\beta = \gamma N$.

For non-zero values of ϵ, we have that $R_0 = \beta/\gamma N^{1-\epsilon}$. As we have constrained β above to be given by $\beta = \gamma N$, the value for R_0 is then given by $R_0 = \gamma N/\gamma N^{1-\epsilon} = N/N^{1-\epsilon}$, thereby leading to an increase in the value of R_0.

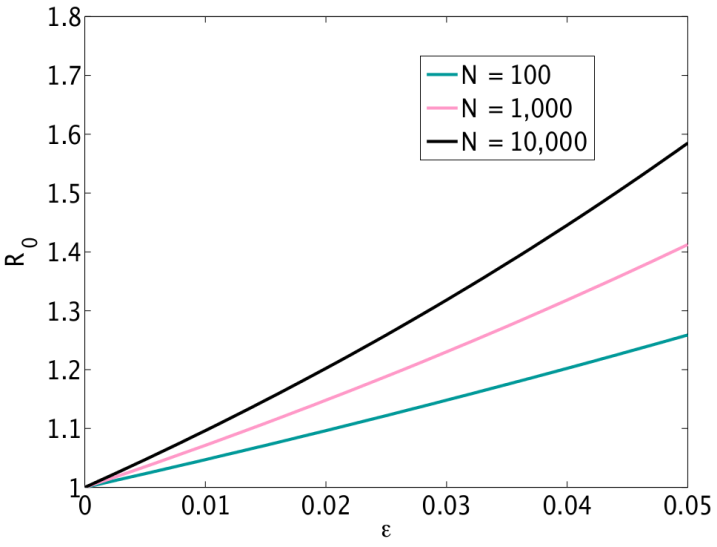
Obviously increasing ϵ increases R_0, and the affect is also dependent on N, meaning that the amplification of R_0 due to ϵ will be larger in a large population than a small one. To plot figure 4.1a, we vary the values of β and γ so that in the $\epsilon = 0$ case, R_0 would have the same value for each population size.

We can also study what happens to the final size of the epidemic when we increase ϵ. To do this we differentiate (4.2) with respect to ϵ.

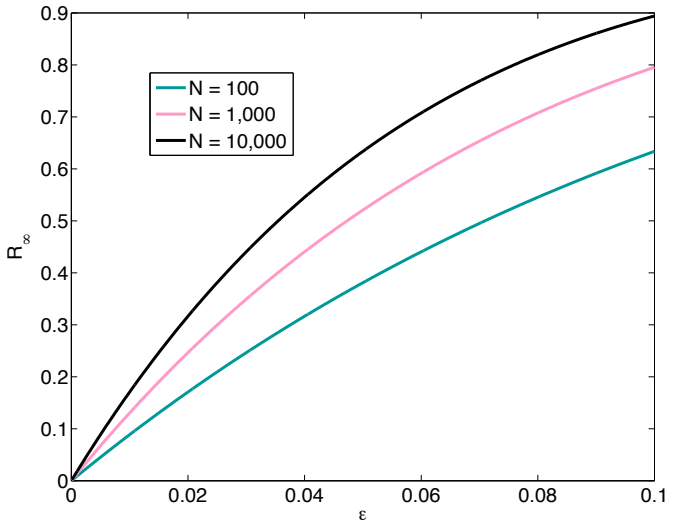
$$\frac{\partial R_\infty}{\partial \epsilon} = -\frac{\partial \exp(-R_\infty R_0)}{\partial \epsilon} \tag{4.4}$$

Using the chain rule, we can write this as follows:

$$\frac{\partial R_\infty}{\partial \epsilon} = \frac{\partial R_\infty R_0}{\partial \epsilon}\frac{\mathrm{d}\exp(-R_\infty R_0)}{\mathrm{d}R_\infty R_0} \ . \tag{4.5}$$



(a)



(b)

Figure 4.1: (a) Change in R_0 as ϵ and N are increased. β and γ are defined by $\beta = \gamma N$, as this means that we will be considering the $R_0 = 1$ situation when $\epsilon = 0$. In (b) we see the final size of an epidemic where we alter ϵ for various values of N. We calculate R_0 using $R_0 = N/N^{1-\epsilon}$. Then for a small non-zero ϵ we use root finding techniques to find the value of R_∞ from (4.2). Equation (4.8) is then numerically integrated to obtain the values for this figure.

We then use the product rule for differentiation to evaluate the right hand side of this equation along with the fact that R_0 can be written as $R_0 = \beta \exp(\epsilon \ln N)/N\gamma$ to allow us to differentiate R_0 with respect to ϵ giving

$$\frac{\partial R_0}{\partial \epsilon} = R_0 \ln N \ . \tag{4.6}$$

Once this is done, we find that,

$$\frac{\partial R_\infty}{\partial \epsilon} = -\left(\frac{\partial R_\infty}{\partial \epsilon} R_0 + R_\infty R_0 \ln N\right)(-\exp(-R_\infty R_0)) \ , \tag{4.7}$$

We now note that from (4.2) $\exp(-R_\infty R_0) = 1 - R_\infty$ and also that $R_0 = \beta/\gamma N^{1-\epsilon}$ so that when we rearrange (4.7) we obtain the following differential equation,

$$\frac{\mathrm{d}R_\infty}{\mathrm{d}\epsilon} = \frac{N^\epsilon \ln(N)\beta(1-R_\infty)R_\infty}{\gamma N - (1-R_\infty)\beta N^\epsilon} \ . \tag{4.8}$$

This cannot be solved analytically, so we solve numerically and do so using fourth order Runge-Kutta. To do this though, we need an initial condition. This can be found by inputting a small value for R_0 into (4.2) and using root finding techniques to calculate the corresponding value of R_∞ from which we can then evolve the system to calculate the change in R_∞ as we increase ϵ. For any value of ϵ, we can find the final size by using root finding techniques on (4.2), meaning that we do not require this differential equation to achieve our goals, but by deriving this we can more easily see the impact of ϵ directly.

Figure 4.1b demonstrates how R_∞ changes when we increase ϵ for different values of N. This shows that there is a large increase in the final size of the epidemic if we increase ϵ by any small amount. It also clearly shows that increasing the size of the population will increase the final size non-linearly, as can be seen by comparing the curves for $N = 100$ to $N =$ 1,000 or 10,000. For the study of the workplaces, this will mean that the large workplaces will have a proportionately larger influence on the total final size of the epidemic than expected. Therefore when we try to fit a distribution to the workplace data, if the fit is good in the bulk of the distribution, and poor at the tail, then we may get a large disagreement in the number of people who would be infected in total from this fitted distribution and the true data. As previously discussed, the use of likelihood methods to fit the distribution in a descriptive way may be compromised, as the fact that the majority of the data occurs for low sizes of workplaces means that we will likely get a good fit to the distribution in this region, which may result in a poor fit to the tail.

Figure 4.2 displays the number of workplaces that have a large number of employees (> 100), by ward. As we would expect, these cluster around big cities such as London, Birmingham and Manchester. We would therefore anticipate that the spread of epidemics through workplaces would be more severe in these areas than in other areas where there are fewer large workplaces.

Figure 4.3 shows the effect of ϵ on the final size of an epidemic for a population, for various populations. The x-axis gives the value of R_0 for the $\epsilon = 0$ case, though this is not the same as the value of R_0 for the when $\epsilon > 0$. For example if the x-axis has a value 2, this means that $\beta = 2\gamma N$, so the true value for R_0 in the $\epsilon > 0$ case is given by $2N/N^{1-\epsilon}$. Even considering the smallest difference in ϵ at $\epsilon = 0.01$ and for a value of $N = 300$, we see that the final size raises to over 10% for the case when $R_0 = 1$ for $\epsilon = 0$, meaning that

we will potentially have 30 more cases than expected in the $\epsilon = 0$ regime.

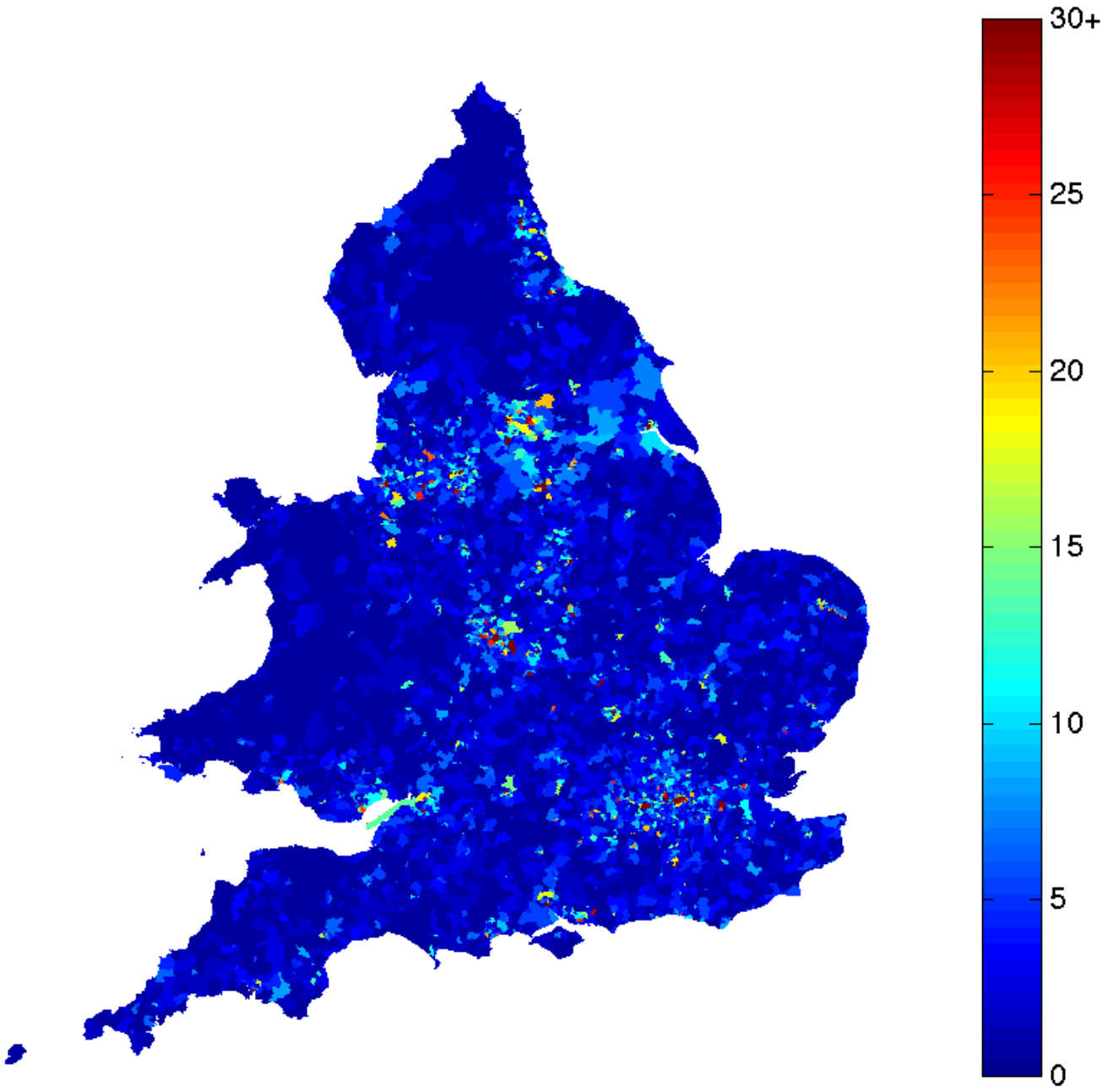


Figure 4.2: Number of workplaces with a high number of employees, showing concentration in main urban areas.

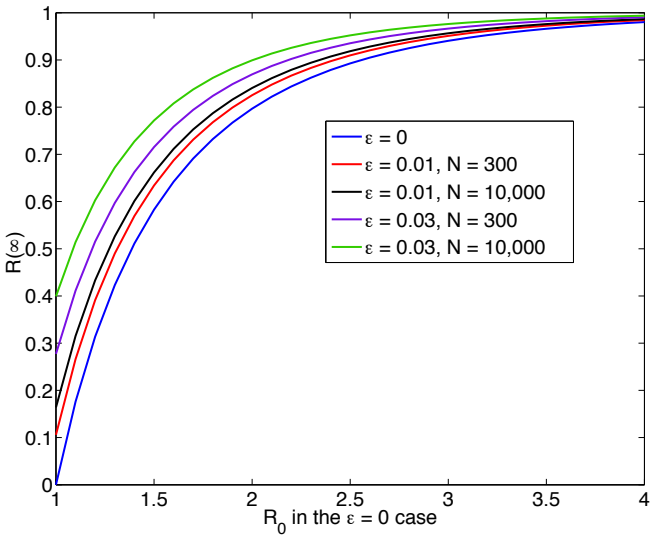


Figure 4.3: Final size of an epidemic where we alter ϵ and N. Even a small change from $\epsilon = 0$ to $\epsilon = 0.01$ can give a significant increase in the final size of the epidemic.

As we have been preparing for so far, the aim is to see how the distribution of workplace sizes can alter the spread of a disease in the population. To examine this we consider the distribution of workplace sizes in the UK, which is discussed next.

4.3 Blue Sheep data

The dataset we use is a proprietary dataset, which is compiled from over 20 separate sources, which gives the postcode location and size of workplaces in the UK [Bluesheep data source]. This is similar to the dataset used by Ferguson et al. [2006] which considered the distribution of US workplaces. This dataset along with the methods used to attempt to fit distributions to it is described next.

4.3.1 Description of data and fitting methods

The workplace data provided by Blue Sheep gives the number of employees in over 2,000,000 workplaces around the UK. The mean of this distribution is 12.23, it has a large variance of 829 and has a value of skewness of a 39.5, meaning that it is right-skewed, which is sometimes described as a 'heavy tail'. We wish to see how the number of employees is distributed and will therefore try to fit to the data with heavy tailed distributions e.g. a power law or log-normal distribution. In figure 4.4 we can see the number of workplaces with between 1 and 100 employees. This clearly demonstrates a problem in fitting this data to an appropriate probability distribution, as it shows that there has been significant rounding of the data, seen most prominently by the increases in numbers of workplaces whose number of employees is a multiple of 5 or 10. This occurs as the number of employees who are physically in a workplace from day to day is, in most cases, not a constant, along with the fact that people are more likely to report round numbers than

other numbers, an oft-observed phenomenon. This means that if the number of employees is around 20 then it is more likely that the number 20 will be reported than a neighbouring integer.

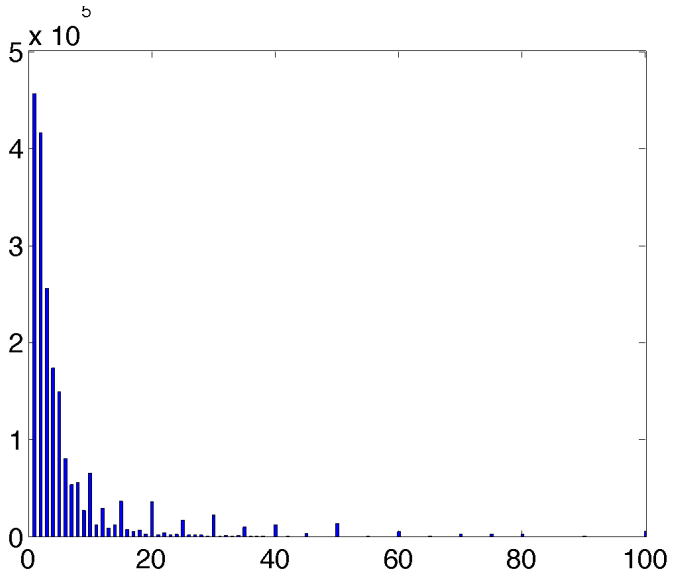


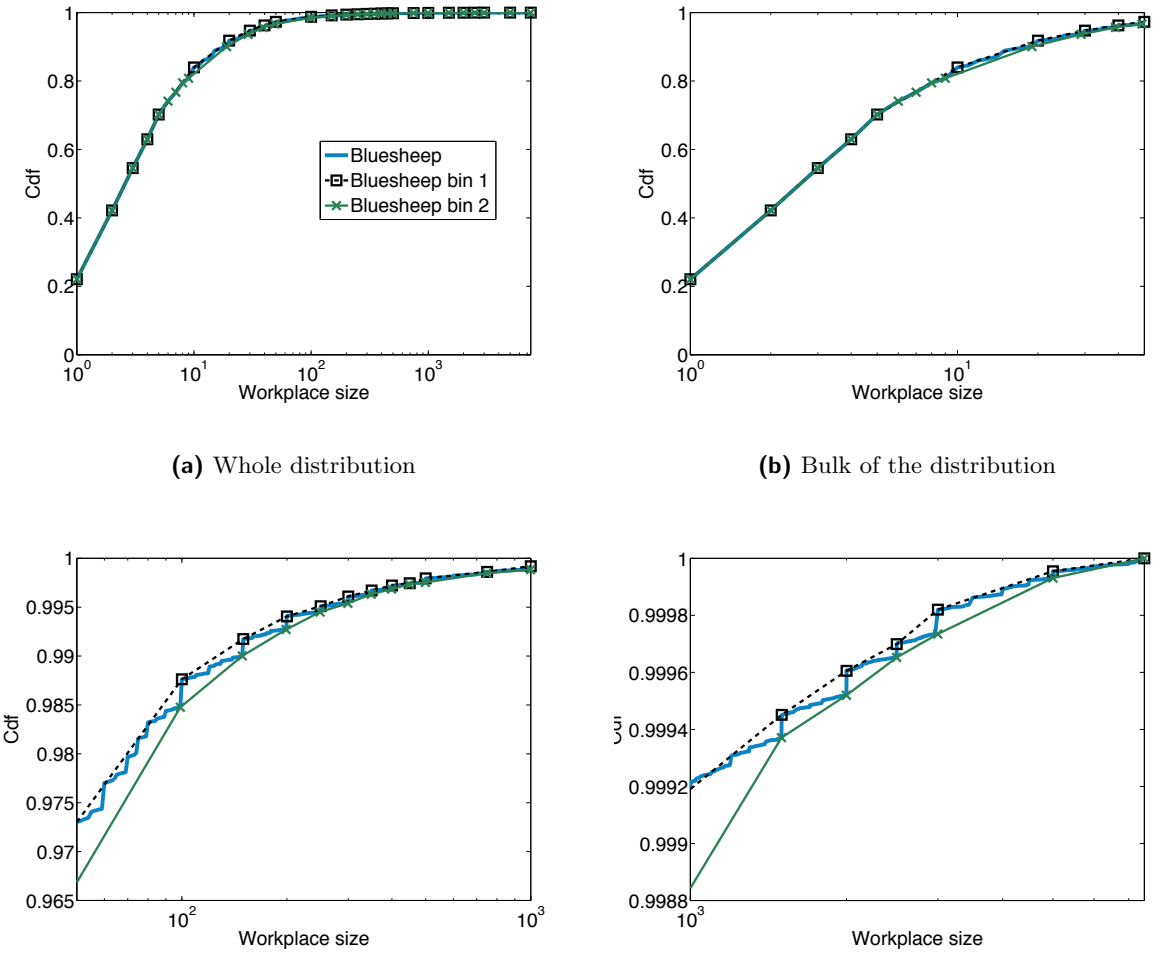
Figure 4.4: Number of workplaces containing between 1 and 100 employees.

To get around this problem, we use methodology which was been developed in Virkar and Clauset [2012]. This methodology allows us to test the likelihood of different distributions representing a binned data set, i.e. a dataset where the counts for discrete values are grouped together and summed. The aim of this is to allow us to get around the fact that the data has been rounded to multiples of 10 (or 50 or 100 when the number of employees gets higher). By binning the data appropriately we hope to output a less biased dataset, which would be indistinguishable from applying the same set of bins to the "true" non rounded dataset. To see this we can consider the following example. Take the number of workplaces which have between 27 and 33 employees. In our dataset there are significantly more which have 30 employees than any of the other values we are considering. This is because of the rounding that has occurred. If we assume that any workplace which in reality (if this was able to be counted reliably and no rounding was applied) has below 27 or above 33 employees will not be rounded to 30, then by binning the counts for 27 to 33 in our dataset, then this will give the exact same figure as if we were to bin the true figures.

We now introduce notation similar to that used by Virkar and Clauset [2012], we have k bin boundaries denoted by the set $B = \{b_1, b_2, \dots b_k\}$ and $k-1$ counts of occurrences, $H = \{h_1, h_2, \dots h_{k-1}\}$. We denote by h_i the count of the number of times that workplaces have b_i to $b_{i+1} - 1$ employees. There are obviously many different binnings that we can choose to use to examine this distribution. We impose the the constraint that the distance between consecutive bin boundaries is monotonically increasing in an attempt to reflect the actual method at which the rounded values are arrived at. We try to choose bin boundaries so that there is only one relatively highly rounded to value between consecutive bin boundaries. For example, we would not want to include 100 and 200 in the same count as they are both frequently rounded to.

Along with fitting to the entire distribution, we are interested in how the tail of the distribution can alter the expected number of infections, which means that we wish to investigate how fitting only to a portion of the distribution can change the quality of the fit, along with the number of infected individuals. We denote by $x_{\min}$ the value at which we wish to consider our fitting of the data to begin at, so if we wish to concentrate on fitting to our dataset from 10 onwards, then we would set $x_{\min} = 10$.

As previously mentioned, there are numerous sensible binnings that could be chosen to investigate the data. To limit the number of calculations and figures required to describe the attempted fittings to the data, we choose to consider two very different binnings, which we expect to bound the results of binning in general. Figure 4.5 shows the two binnings that have been chosen to investigate the possible distribution of the Blue Sheep data. Binning 1 is derived by taking each consecutive bin to end on a number that is preferentially rounded to, so we have bins such as 6-10 and 51-100. Binning 2 on the other hand always begins with a preferentially rounded to number, so has bins such as 10-19 and 50-99. Binning 1 therefore has a cumulative distribution function (cdf) which point to point appears greater than binning 2, which means that when we fit to these different bins, we will potentially end up with a distribution which results in having a greater number of large values in case 2 than in case 1. This is expected to result in the final size of epidemics being larger in the second case than the first, as the larger workplaces act as magnifiers of the disease.



(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.5: The two different binnings used. Binning 1 is always higher than binning 2, meaning that there are more workplaces with fewer employees in them in this scenario. In blue is the raw Blue Sheep data.

Next, limitations of this data source are discussed.

4.3.2 Limitation of Blue Sheep data

A large limitation of this data is that it is a private data source, and so there is little opportunity to assess the accuracy of the data that is provided. As previously stated this combines data from 20 separate sources increasing the uncertainty in the accuracy of the data. If a company name was provided for each record, then some checking could be done, but as this is not included, there is no way of easily verifying the data.

As a comparison, we can consider an open source dataset from the Office for National Statistics (ONS) which provides the number of workplaces by district from VAT and PAYE records. This is called the UK Business: Activity, size and location, and can be freely downloaded from the ONS website [UK Business]. This dataset is at district level, of

which there are 545 in this dataset and is a higher level than CAS ward. Therefore, there is therefore far less spatial resolution in this dataset than the Blue Sheep data, which is at post code level. Additionally, rather than the sizes of workplaces, the ONS data simply gives the number of workplaces in these districts, meaning that this data is not suitable for the purposes we are using the Blue Sheep data for here. We can however use this data to assess how accurate the Blue Sheep data is, at least by the measure of number of workplaces.

In figure 4.6 we see the comparison between these two datasets at a district level for the number of workplaces. It can be seen that the number of workplaces given in the Blue Sheep data is generally greater than for the UK Business dataset. One explanation for this could be duplication of workplaces in the Blue Sheep data, as this is compiled from many different sources. It is difficult to assess these differences in much detail due to the lack of transparency in the Blue Sheep data, along with the fact that simply reporting the number of workplaces, as is the case for the UK Business data, is relatively uninformative. For example it is impossible to tell if there is a regular underreporting of small workplaces here when compared with the Blue Sheep data. Overall, the difference in the number of workplaces is not so large as to cast overwhelming doubt on the accuracy of the Blue Sheep data, and if we are interested in considering the workplace size distribution, then there is little option but to use the available data.

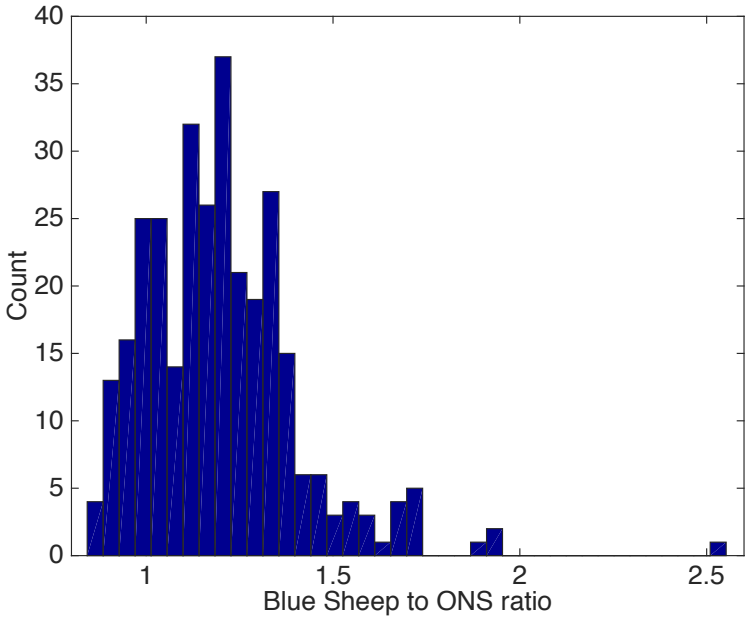


Figure 4.6: A histogram showing the count of districts which have a given ratio for the number of workplaces in the Blue Sheep data, compared with the ONS UK Business dataset.

4.3.3 Calculating the final size of epidemic in workplaces

Figure 4.7 shows how the change in ϵ could alter the progression of the epidemic in the different wards using the Blue Sheep data. For these maps, we use values for β and γ such

that if we were in the $\epsilon = 0$ case, the values of R_0 would be 1. This means that we will scale β such that $\beta = \gamma N$ for each population size N in the data set. To get the final sizes of the epidemics in the workplaces, we decide on a value for ϵ and then for each of the workplaces in the data set we will calculate the value of R_0 decided by N, ϵ, β and γ and then find R_∞ by using (4.2) and then multiply this value by N to get the expected number of people who would be infected in the workplace.

As we can see in figure 4.7, there is a great difference in the size of the epidemic for $\epsilon = 0.01$ and $\epsilon = 0.1$, as in the first case, the majority of wards have very few infecteds in the workplaces, whereas in the latter case, infection is widespread and there are many wards which have over 5,000 infecteds. Note that these maps are for illustrative purposes only, as there has been no attempt to model the spatial spread of a hypothetical epidemic in the UK here, simply the spread throughout workplaces has been considered.

We also wish to generate workplace sizes using different distributions, such as the power law, log-normal and Zipf-like distribution taken from Ferguson et al. [2006]. We will truncate so that we will not allow any of these distributions to select a sample which is higher than the maximum contained in the Blue Sheep data, which is 7,500. This is complicated by the fact that when we sample from these distributions we wish to sample (approximately) the same total number of people as are contained in the Blue Sheep data. This will require us to generate different numbers of samples, based on the distribution we are considering, as the mean of the numbers generated will vary from distribution to distribution.

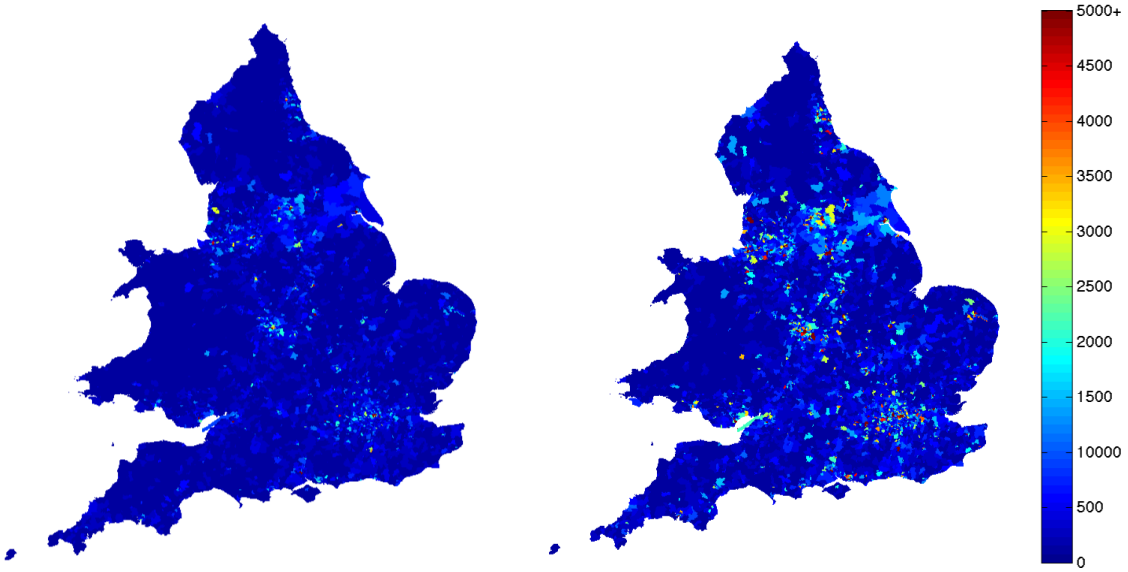
To do this we will calculate what the mean of the distribution in question is given the parameters, and then generate as many workplaces as would be needed to give the same number of total employees in the dataset if each workplace had the mean number of employees. We then consider the cumulative sum of these generated workplaces, and only include workplaces until the point at which the cumulative sum is greater than the total number of employees in the dataset. If the sum of our generated distribution is smaller, we generate more sizes until the sum is greater than the Blue Sheep dataset.

4.4 Model selection

4.4.1 Description of methods and distributions considered

There are many methods which can be used to select models, which vary wildly in their statistical sophistication. The crudest methods such as the Kolmogorov-Smirnov test and the mean-squared error simply consider the distance between two distributions in order to select one model over another. There are many more sophisticated techniques, such as the Akaike Information Criterion [Akaike, 1974] which aims to estimate the Kullback-Leibler divergence [Kullback and Leibler, 1951] between the proposed distribution and the actual distribution in question, by calculating the likelihood of the data given the proposed distribution. Here we compare the fit provided by simple distance-based techniques such as the average absolute deviance between the cdf of the proposed distribution and the true distribution, alongside likelihood-based techniques in order to see which gives us a better fit to the Blue Sheep data measured using the estimated final size of the outbreak in the workplaces.

It is anticipated that with the problem we are considering, simply calculating the model



(a) $\epsilon = 0.01$ **(b)** $\epsilon = 0.03$

(c) $\epsilon = 0.05$ **(d)** $\epsilon = 0.1$

Figure 4.7: Summed final sizes in workplaces for different values of ϵ by ward. Workplace size distribution is the Blue Sheep data.

with the highest value of the likelihood may not give us the most informative result. This is because we are interested in how the change in transmission rates will affect the size of the epidemics in the workplaces. The way that this change in transmission rates will take effect will be in the increase in the number of infecteds from workplaces with a high number of employees. Therefore one of the major factors that we must consider is the probability of selecting a 'high' number of employees. The fact that there are relatively few of these in the Blue Sheep data ($\approx 84\%$ are 1-10, $\approx 98.75\%$ are 1-100), means that the likelihood calculations will be weighted heavily by accuracy in the bulk part of the distribution, which will therefore mean that we can achieve the greatest likelihood with a very inaccurate approximation of the tail of the distribution.

As mentioned above, along with calculating likelihoods, another simple test would be to simply sum the absolute error in the cdf of the distribution in question to the cdf of our empirical data and to then choose the set of parameters which minimised this. This approach will enable us to choose the distribution which is most accurate to the Blue Sheep data over the whole range of the data necessarily meaning that the quality of the fit must be, in at least a relative sense, good in all regimes of the data, but biased to the tail due to the use of the cdf rather than the probability mass function (pmf).

If we wish to fit the data to this distribution where we use a value for $x_{\min} \neq 1$, then we must adjust the above method to do this. To do this, we firstly work out the cdf of the whole dataset. Then to fit to the data for any distance-based method, using for example least absolute error of cdf, we then only take into account the differences in the cdf's from $x_{\min}$ and up. This results in us choosing parameters which line up closely with the data from the value of $x_{\min}$ onwards but without the constraint of minimising the error for lower values.

To adapt likelihood methods, we could weight the likelihood provided from the larger workplaces more heavily than for lower values. However we have chosen to simply calculate the likelihood from the data which is greater than or equal to $x_{\min}$, as there is no way of incorporating the information contained in the data below $x_{\min}$ without explicitly including it in our calculations. When we then take the parameters which have the greatest likelihood in this regime, and try to produce the whole of the workplace distribution, it is possible for the resulting distribution to differ greatly from the Blue Sheep data. This is due to that fact that the majority of the distribution can be entirely ignored if $x_{\min} \gg 1$.

We are interested in fitting the workplace data to a distribution, in order to investigate how the assumption of different distributions for the workplace sizes can affect the way an epidemic will spread throughout the workplaces. As mentioned previously we use methodology from Virkar and Clauset [2012] to do this. Code for these calculations is also provided [Clauset and Virkar, 2012]. However this paper, deals with real valued data and allows us to fit a distribution to any dataset, without the use of a maximum value. As we wish to fit discrete data and have also imposed a maximum value onto the distribution, we develop our own techniques to do this.

In various situations to get the 'best' fit for a power law, or another distribution to a dataset, often some portion of the data that we are trying to fit is ignored in favour of fitting the tail of the data to a distribution [Clementi and Gallegati, 2005; Willinger and Paxson, 1998; Redner, 1998], which is discussed, along with the appropriateness of these fits by Clauset et al. [2009]. In general we are interested in characterising the whole of the distribution as well as just to the epidemiologically important tail, though will consider fitting only a portion of the data, as noted above, denoted by choosing a value of

$x_{\min} > 1$.

Note that due to the use of bins, it causes problems when $x_{\min} > 1$. This is because if we set our value for $x_{\min}$ to be inside a bin, it is unclear what the best course of action is, as we will be unable to split the number of occurrences within the bin to ones below $x_{\min}$ and those above. The course of action taken is described with the following example: if we have a bin which groups readings from, 51 to 100 and $x_{\min} = 100$, we define this to mean that we include the readings in this bin in our calculation. Now when we are trying to fit our distribution using distance-based methods, we normalise for values from 51 to $x_{\max}$.

The distributions we have considered for the workplace size distribution are: 1) 'offset truncated power law' distributions (which are put forward as a good candidate for workplace sizes by Ferguson et al. [2006] based on a model from Riley and Ferguson [2006]), 2) discrete power laws and 3) log-normal distributions. There are of course many different distributions which we have not considered which could be as good or better candidates for modelling this dataset. However, the distributions above are often cited as candidates for modelling heavy-tailed distributions [Crovella et al., 1998; Mitzenmacher, 2004; Clauset et al., 2009] and so are also considered likely candidates here.

In each of the three following subsections of the thesis, one the distributions which we are attempting to fit to the data (offset truncated power law, discrete power law and log-normal) is first described. This leads into the calculation of the likelihood of the distribution in question, which is followed by the fitting of this distribution to the Blue Sheep data using the likelihood method, along with the minimum total absolute error in the cdf. The number of predicted infections from the best fitting distributions is then calculated and compared to the predicted infections for the Blue Sheep data for various forms of the transmission rates, defined by the selection of ϵ. To do this we introduce a single infected individual into each workplace, and then calculate the expected final size using (4.2) for the comparative value of R_0 defined by ϵ, which is again constrained by requiring $R_0 = 1$ in the $\epsilon = 0$ case.

Therefore each of the following subsections contains methods and results, which is done to allow for full understanding of the fitting of each distribution in turn. Following these three subsections will be a summary of the results and a discussion of the relative success and failures of these distributions in the fitting to the Blue Sheep data.

4.4.2 Offset truncated power law

Workplace sizes from the USA has been fitted, using likelihood-based methods, to a 'Zipf-like' distribution previously [Ferguson et al., 2006]. More descriptively, this has also been called an offset truncated power law distribution. This was assumed to be representative for the UK by Ferguson et al. [2006], though data was not available to determine parameters. As we have this data we can examine how the assumed parameters in this study would affect the spread of the infection and also what parameters we can determine from the data to be the best fit.

The form of the cdf for this distribution is,

$$P(X \leq x) = 1 - \frac{\left(\frac{1+x_{\max}/a}{1+x/a}\right)^c - 1}{(1+x_{\max}/a)^c - 1} \ , \tag{4.9}$$

where $x_{\max}$ is the largest value we will allow the distribution to take and c and a are parameters dictating the shape of the distribution. In Ferguson et al. [2006] the values used are $a = 5.36$, $c = 1.34$ and $x_{\max} = 5,920$. We note that the maximum value in the Blue Sheep data is 7,500, so we will use $x_{\max} = 7,500$, but a and c are both values we wish to fit using likelihood methods.

We use (4.9) to calculate the probability of seeing a specific value, which is given by the following expression,

$$P(X = x) = P(X \leq x) - P(X \leq x - 1) = \frac{\left(\frac{1+x_{\max}/a}{1+(x-1)/a}\right)^c - \left(\frac{1+x_{\max}/a}{1+x/a}\right)^c}{(1 + x_{\max}/a)^c - 1} \tag{4.10}$$

where (4.9) gives the right hand side of this equation.

If we are taking a value of $x_{\min} > 1$, which implies we fit the power law from $x_{\min}$ to $x_{\max}$, then we also need to calculate a normalising constant, which we denote $\phi(a, c, x_{\min}, x_{\max})$. This is given by evaluating the following:

$$\begin{aligned}\phi(a, c, x_{\min}, x_{\max}) = \phi = &\sum_{x=x_{\min}}^{x_{\max}} P(X = x) \\ =&\frac{\left(\frac{1+x_{\max}/a}{1+(x_{\min}-1)/a}\right)^c - 1}{(1 + x_{\max}/a)^c - 1},\end{aligned} \tag{4.11}$$

as all terms other than the first and the last are cancelled out in the sum. It is simple to show that this is equal to 1 if $x_{\min} = 1$.

As we are attempting to fit this distribution to binned data, it is required to calculate the probability of seeing a value within a certain region, defined by the choice of bins. This is achieved by taking the difference of the two values of the cdf at the end points of each bin. For example to calculate the probability of seeing a value in the bin which includes values from b_{i-1} to $b_i - 1$, this is done as follows,

$$P(b_{i-1} \leq X < b_i) = \frac{1}{\phi} \frac{\left(\frac{1+x_{\max}/a}{1+(b_{i-1}-1)/a}\right)^c - \left(\frac{1+x_{\max}/a}{1+(b_i-1)/a}\right)^c}{((1 + x_{\max}/a)^c - 1)} \tag{4.12}$$

This means that the likelihood of this distribution is calculated as follows,

$$P(H|B, a, c, x_{\min}, x_{\max}) = \prod_{i=1}^{k} \left(\frac{1}{\phi} \frac{\left(\frac{1+x_{\max}/a}{1+(b_{i-1}-1)/a}\right)^c - \left(\frac{1+x_{\max}/a}{1+(b_i-1)/a}\right)^c}{(1 + x_{\max}/a)^c - 1} \right)^{h_i}. \tag{4.13}$$

The log-likelihood is then given by,

$$\begin{aligned}L(H|B, a, c, x_{\min}, x_{\max}) = &- n \ln \left((1 + x_{\max}/a)^c - 1\right) - n \ln \phi \\ &+ \sum_{i=1}^{k} h_i \ln \left(\left(\frac{1 + x_{\max}/a}{1 + (b_{i-1} - 1)/a} \right)^c - \left(\frac{1 + x_{\max}/a}{1 + (b_i - 1)/a} \right)^c \right).\end{aligned} \tag{4.14}$$

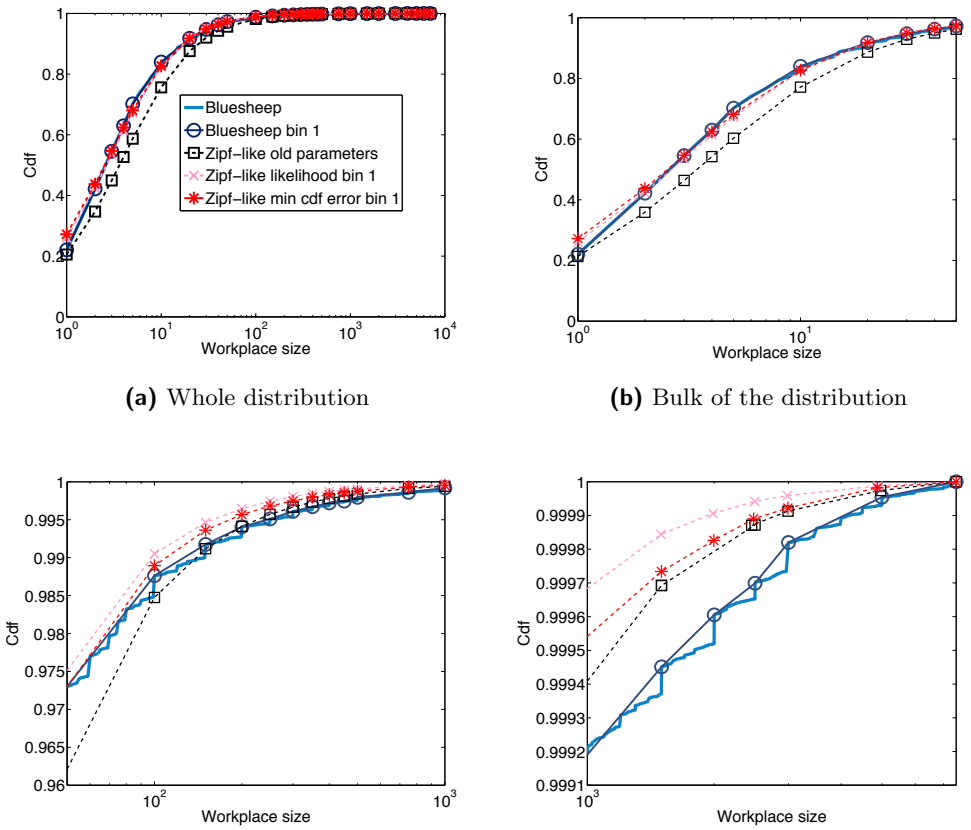
We cannot maximise this analytically, so in practice, we use a brute force gridding method by selecting a range of different values for a and c and choosing the maximum numerically. This method is repeated for all other distributions that we wish to investigate.

When comparing log-likelihoods of two different models, it is often useful to refer to the relative log-likelihood between the two models to demonstrate the loss in terms of likelihood that choosing the model with the lower log-likelihood has. This is given simply the difference in the value of the log-likelihood for the two models.

Figure 4.8 shows the fits from likelihood analysis and the minimum absolute error in cdf for binning 1 along with the parameters which are taken from Ferguson et al. [2006]. To generate this figure we plot the cdf of the raw data along with the binning in question. We then generate 10^6 samples for each set of parameter values, and then the cdf was calculated at the corresponding points from the bins from these samples.

We can see that the distributions that we have fitted are a much better fit for the cdf of the Blue Sheep data in the bulk of the distribution than the values which are taken from Ferguson et al. [2006], which as previously discussed, we would expect. We also note that from workplace size of 20, both the likelihood and min cdf error fits have larger values of cdf than the binning to which they were fitted, meaning that proportionately more workplaces have been allocated at each point in these fittings than in the dataset. As the numbers of employees in the workplaces increases past 20 the parameters taken from Ferguson et al. [2006] begin to more closely match the data, and from about workplace size 200 fit more closely than the newly fitted parameters. This is due to the differences in the distribution of workplace sizes in the US as opposed to the UK. The fit of these parameters suggests that in the USA there are proportionately fewer workplaces with fewer than 20 employees and more of a larger size.

Figure 4.9 shows similar distributions as 4.8, but the fitting is done for binning 2. Again the two fitted distributions are closer to the data than the parameters from Ferguson et al. [2006] in the bulk of the data. In the tail of the distribution, the fit to the raw Blue Sheep data in figure 4.9 is closer than that seen in figure 4.8. The parameters from Ferguson et al. [2006] again approach the data as we reach the tail (as neither of them have changed) but this time the newly produced fit using the minimum error in the cdf is as close or closer to the data at all times.



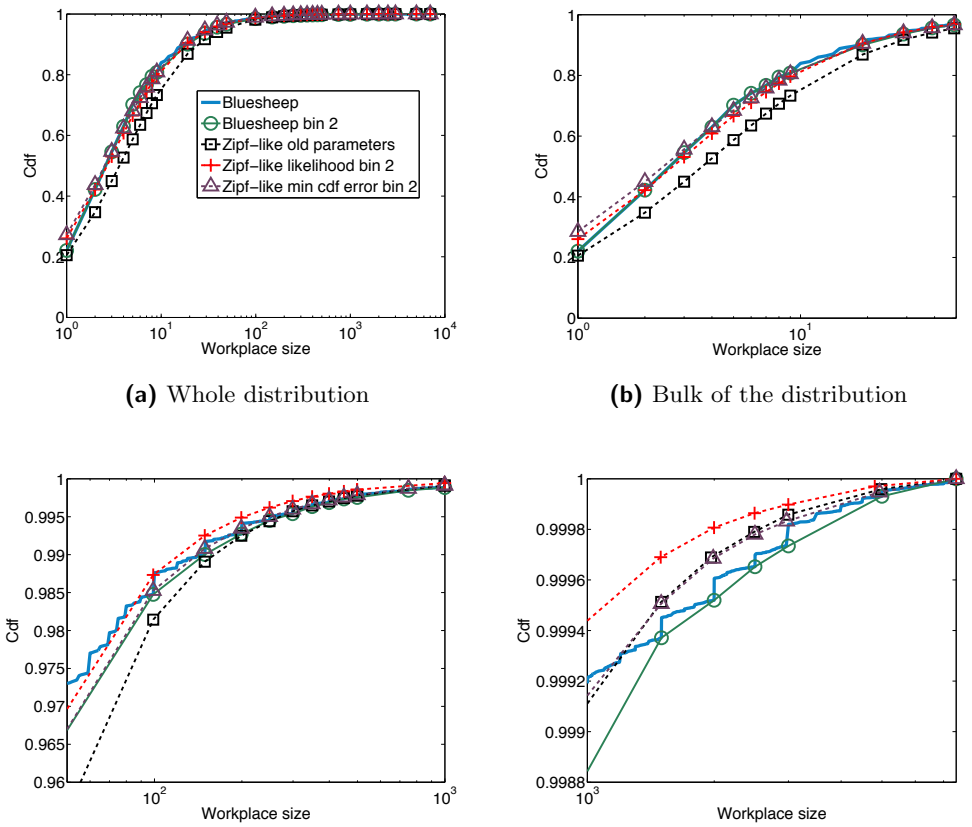
(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.8: Cdf's for the distribution of number of employees in a workplace, taken from Blue Sheep data source and compared to fitted offset truncated power laws (labelled as Zipf-like in figures). Old parameters are $a = 5.36$, $c = 1.34$ and are taken from Ferguson et al. [2006], whilst fitted parameters are $a = 4.40$, $c = 1.47$ with a log-likelihood of -4,457,400 for likelihood fitted parameters and $a = 3.8$, $c = 1.36$ with a relative log-likelihood of -2117, when compared to maximum likelihood, for minimum absolute difference between the cdf of the binned Blue Sheep data and the fitted cdf.



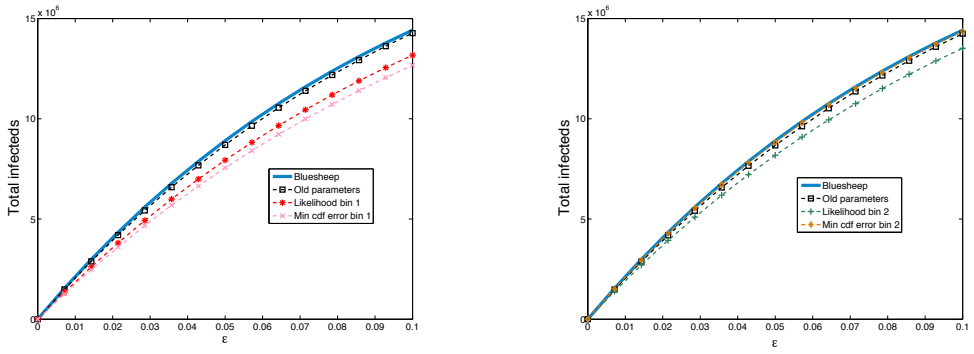
(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.9: Cdf's for the distribution of number of employees in a workplace. Old parameters are $a = 5.36$, $c = 1.34$ whilst fitted parameters are $a = 3.94$, $c = 1.34$ for likelihood fitted parameters with a log-likelihood value of -4,839,900 and $a = 3.10$, $c = 1.20$ and relative log-likelihood value of -5,542 for minimum absolute difference between the cdf of the binned Blue Sheep data and the fitted cdf.



(a)

(b)

Figure 4.10: (a) Binning 1. Parameters as given in figure 4.8. (b) Binning 2. Parameters as given in figure 4.9.

In figure 4.10 we can see the total number of infecteds in the population for various different values of ϵ when we perform calculations as described at the end of § 4.4.1. We perform this for the raw Blue Sheep data which gives the actual distribution of workplace sizes, along with several offset truncated power law distributions. Figure 4.10a shows this for the raw Blue Sheep data along with the offset truncated power law distribution with parameters taken from fitting to binning 1 and the set of parameters taken from Ferguson et al. [2006]. We can see that as ϵ is increased the total infecteds for the parameters taken from Ferguson et al. [2006] is a better approximation to the number of infecteds given by examining the Blue Sheep data, whilst at low values of ϵ the agreement is closer for the other sets of parameters. This implies that as ϵ increases, the agreement in the tail of the distributions becomes more important than the fit for the bulk, as it is in this regime of the cdf for which the non-fitted parameters have a better agreement to the data.

Figure 4.10b is similar but the analysis is performed for the fitted distributions using binning 2. Again the old parameters do better than the parameters obtained by maximising the likelihood as ϵ increases, whilst the set of parameters which are obtained from minimising the absolute error in cdf do better than these old parameters for all values of ϵ as we would expect as they are consistently closer to the cdf of the Blue Sheep data than all other parameter sets.

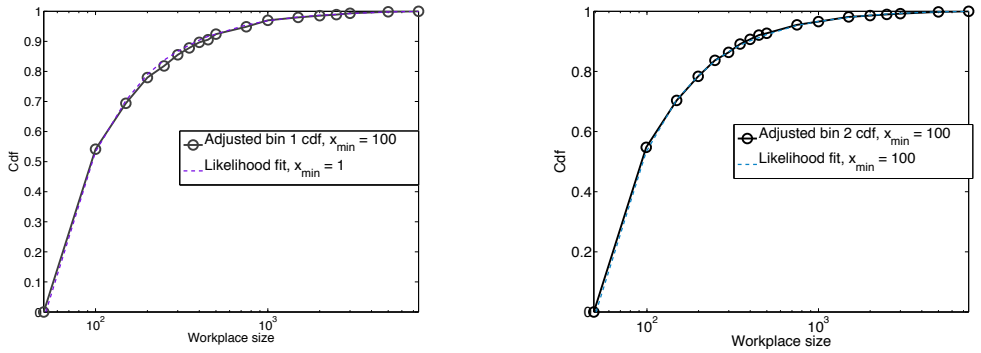


Figure 4.11: (a) Adjusted cdf for Blue Sheep compared to likelihood fitting of offset truncated power law for $x_{\min} = 100$ for binning 1. The parameter values are $a = 1.00$, $c = 1.10$. (b) Adjusted cdf for Blue Sheep compared to likelihood fitting of offset truncated power law for $x_{\min} = 100$ for binning 2. The parameter values are $a = 1.00$, $c = 1.13$.

We are interested in finding the most useful fit to the Blue Sheep data when we consider this in terms of the final size of the epidemic as dictated by (4.2) when we allow ϵ to be non-zero. As has been shown for the offset truncated power law, the parameters taken from Ferguson et al. [2006] give a good agreement to the total number of infecteds in the population despite having a far larger error in the part of the cdf corresponding to small workplaces. Therefore it is arguably more important to get a good fit to the cdf in the tail of the distribution, as it is here that the increase the numbers of infecteds produced due to the increase in ϵ becomes more important. One way of fitting to the tail of the data is include a value of $x_{\min} > 1$. We can then fit to the data using likelihood methods or any distance-based technique.

As briefly described in § 4.4 to fit for the likelihood, we simply calculate the likelihood for the data $\geq x_{\min}$. In figure 4.11a we display the adjusted cdf for the first bin along with the best likelihood offset truncated power law distribution for this region. By the adjusted cdf we mean that we remove all data below the value of $x_{\min}$ and then work out the cdf for the remaining data. For this example, we have used $x_{\min} = 100$. To calculate this adjusted cdf we remove all data below 51, as there is a bin from 51-100 and then work out the cdf, meaning that at 50 the cdf is 0. We also use this remaining data to fit the distribution with the best likelihood. As can be seen in the figure, the agreement between these two distributions is strong. Figure 4.11b shows the same for binning 2.

When we then generate the total distribution of workplace sizes using the parameters from this fit due to the fact that the likelihood method must necessarily ignore the data beneath $x_{\min}$, we are not guaranteed a good fit to the cdf of the data at any point, even in the region greater than $x_{\min}$. This is due to the fact that we are fitting to the adjusted cdf, all we can be sure of is that what remains of the cdf from $x_{\min}$ up will have the correct proportion assigned to each bin.

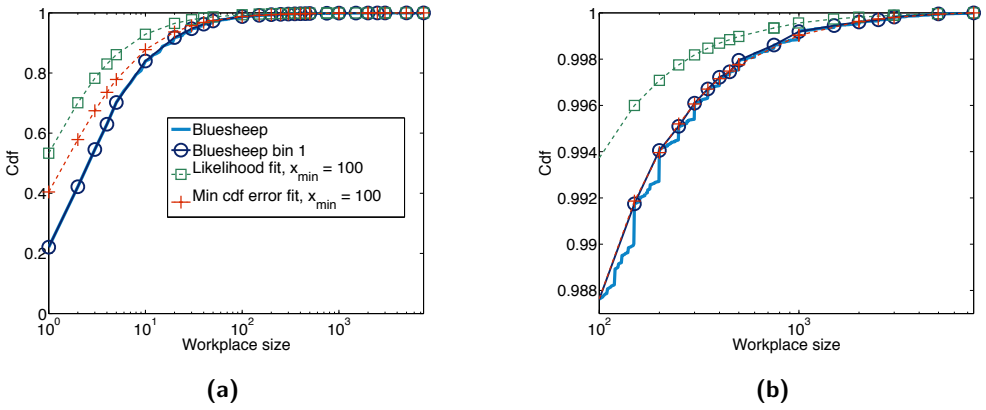


Figure 4.12: (a) Cdf of likelihood and minimum cdf error fitting of offset truncated power law for $x_{\min} = 100$ for binning 1. The parameter values are $a = 1.00$, $c = 1.10$ for the likelihood fit and $a = 1.57$, $c = 1.05$ for the minimum cdf error fit. This figure shows the whole distribution. (b) From $x_{\min} = 100$ to tail.

As can be seen in 4.12 we see the comparison of the cdf for the likelihood fits for binning 1 along with the min cdf error fit with the Blue Sheep data along with binning 1. We can see in both 4.12a and 4.12b the difference in these for the whole cdf and in more detail the region in greater than $x_{\min}$. The cdf for the best likelihood fit assigns too many workplaces a low number of employees, whilst the minimum cdf error is consistently closer to the cdf of the data than the likelihood fit and is an extremely good fit to the data from the value of $x_{\min}$ up. Figure 4.13 shows the same for binning 2.

Due to the fact that we have a poor fit to the empirical cdf for the best likelihood parameters, especially in the tail of the distribution where compared to the Blue Sheep data, there are too few workplaces, we expect the total number of infecteds in the population to be lower than the total given by the data by a significant amount. Figure 4.14 shows this to be the case, for both binnings, and also shows that the total number of infecteds for the min cdf error fits have a good agreement with the number of infecteds. On the evidence

of this analysis, when considering a non-zero value of $x_{\min}$ we will not perform fits to the data using likelihood methods in future, as this lends nothing to the analysis.

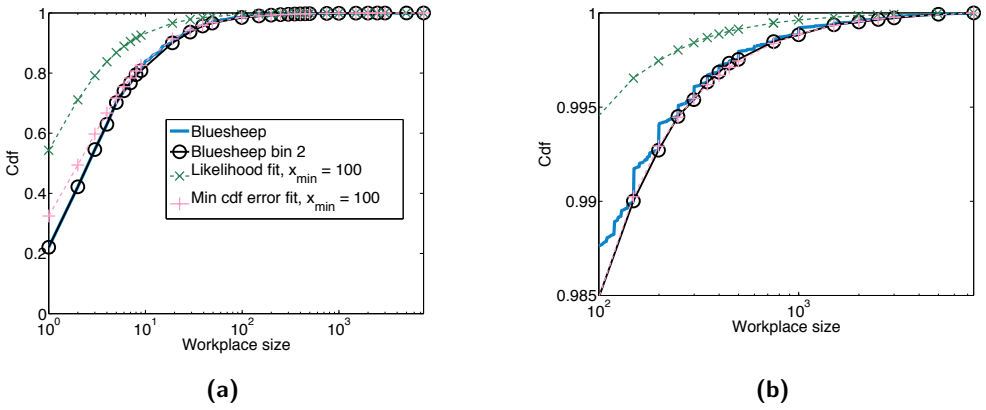


Figure 4.13: (a) Cdf of likelihood and minimum cdf error fitting for $x_{\min} = 100$ for binning 2. The parameter values are $a = 1.00$, $c = 1.13$ for the likelihood fit and $a = 2.36$, $c = 1.10$ for the minimum cdf error fit. This figure shows the whole distribution. (b) From $x_{\min} = 100$ to tail.

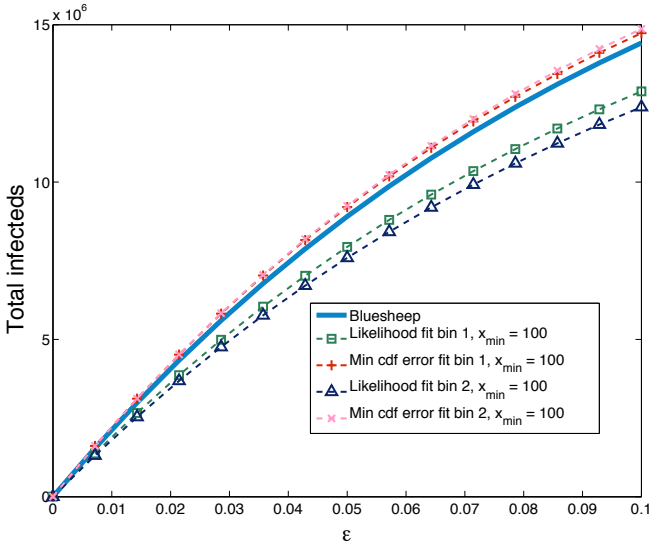


Figure 4.14: Total infecteds for different values of ϵ for different offset truncated power law distribution fits for $x_{\min} = 100$. Parameter values are as given in figs 4.12 and 4.13.

The next distribution that we wish to consider is the discrete power law, which is seen as a good candidate for fitting to many empirical datasets, and discussed next.

4.4.3 Discrete power law

As is described by Virkar and Clauset [2012], the pdf for a discrete power law, with exponent α and minimum possible value $x_{\min}$, is given by the following,

$$P(X = x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \ , \tag{4.15}$$

where $\zeta(\alpha, x_{\min}) = \sum_{x=x_{\min}}^{\infty} x^{-\alpha}$, is the normalising constant, which in this case is the Hurwitz zeta function. However as we wish to limit our distribution by the maximum of the Blue Sheep data, we must redefine this distribution (4.15). We firstly no longer use the Hurwitz zeta function, but alter this to take a third argument, $x_{\max}$, such that $\zeta(\alpha, x_{\min}, x_{\max}) = \zeta = \sum_{x=x_{\min}}^{x_{\max}} x^{-\alpha}$. The pdf is now given by,

$$P(X = x) = \begin{cases} \dfrac{x^{-\alpha}}{\zeta}, & \text{if } x_{\min} \leq x \leq x_{\max} \text{ and } x \in \mathbb{N} \\ 0, & \text{otherwise.} \end{cases} \tag{4.16}$$

To calculate the probability of selecting a value between two values b_i and b_{i+1}, we simply add up all the probabilities in between these two values,

$$P(b_i \leq X < b_{i+1}) = \sum_{x=b_i}^{b_{i+1}-1} P(X = x) \ . \tag{4.17}$$

If we assume that the recorded observations in the Blue Sheep data are taken from a discrete power law, then with k bin boundaries the set of which are denoted by B and $k-1$ observation counts collected in the set H, the likelihood of observing said counts is given by,

$$P(H|B, \alpha, x_{\min}, x_{\max}) = \prod_{i=1}^{k-1} \left(\sum_{x=b_{i-1}+1}^{x=b_i} \frac{x^{-\alpha}}{\zeta} \right)^{h_i} \ , \tag{4.18}$$

where we assume that $b_0 = x_{\min} - 1$. Like for the offset truncated power law, log-likelihood is used in favour of likelihood, which is obtained simply by taking the logarithm of (4.18). When this is done we get the following expression for the log-likelihood,

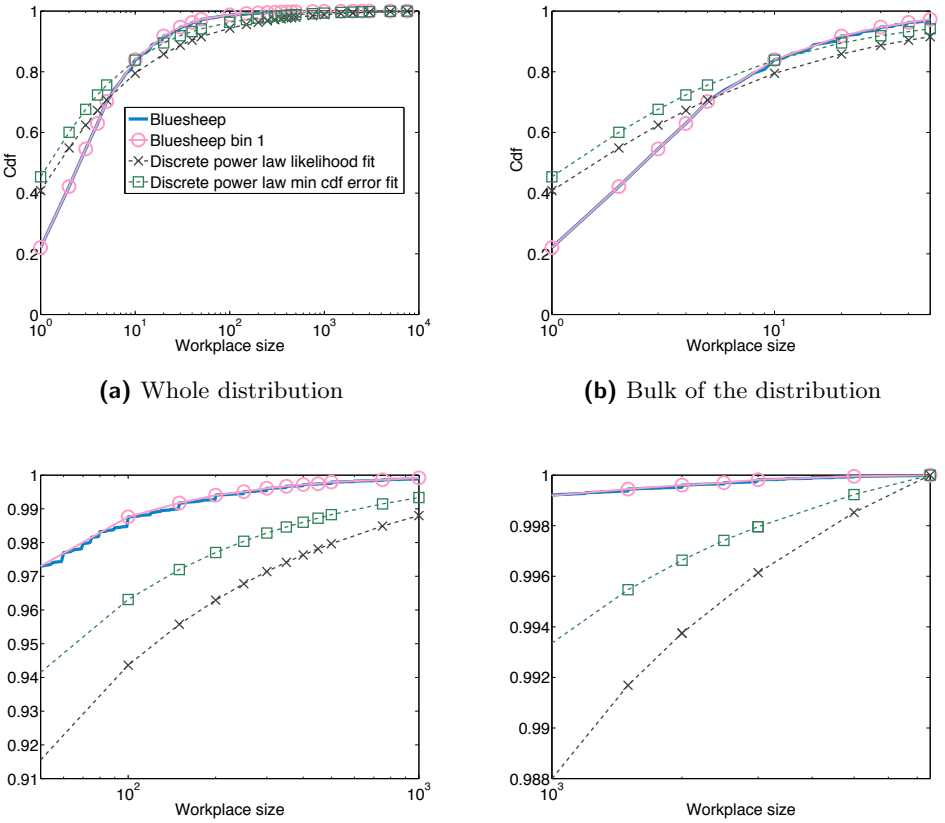
$$\begin{aligned} L(H|B, \alpha, x_{\min}, x_{\max}) &= \sum_{i=1}^{k-1} \left(-h_i \ln(\zeta) + h_i \ln \left(\sum_{x=b_{i-1}+1}^{b_i} x^{-\alpha} \right) \right) \\ &= -n \ln(\zeta) + \sum_{i=1}^{k-1} h_i \ln \left(\sum_{x=b_{i-1}+1}^{b_i} x^{-\alpha} \right) \ , \end{aligned} \tag{4.19}$$

where n is the sample size.

Again as we cannot maximise this analytically, we choose a range of values of α and select the one which gives us the largest value for the likelihood.

When we fit to this distribution, we get a different profile than the fit we got for the offset truncated power law distribution. Figures 4.15 and 4.16 demonstrate this. Again we fit these distributions for the largest value of the likelihood and also minimising the error in

the cdf. From workplace sizes of about 10 until the end, the cdfs for all the discrete power law distributions have smaller lower values than that for the empirical cdf. This means that we will have more workplaces with a large number of people working there resulting in the expectation of larger final sizes for the epidemics in comparison to the Blue Sheep data.



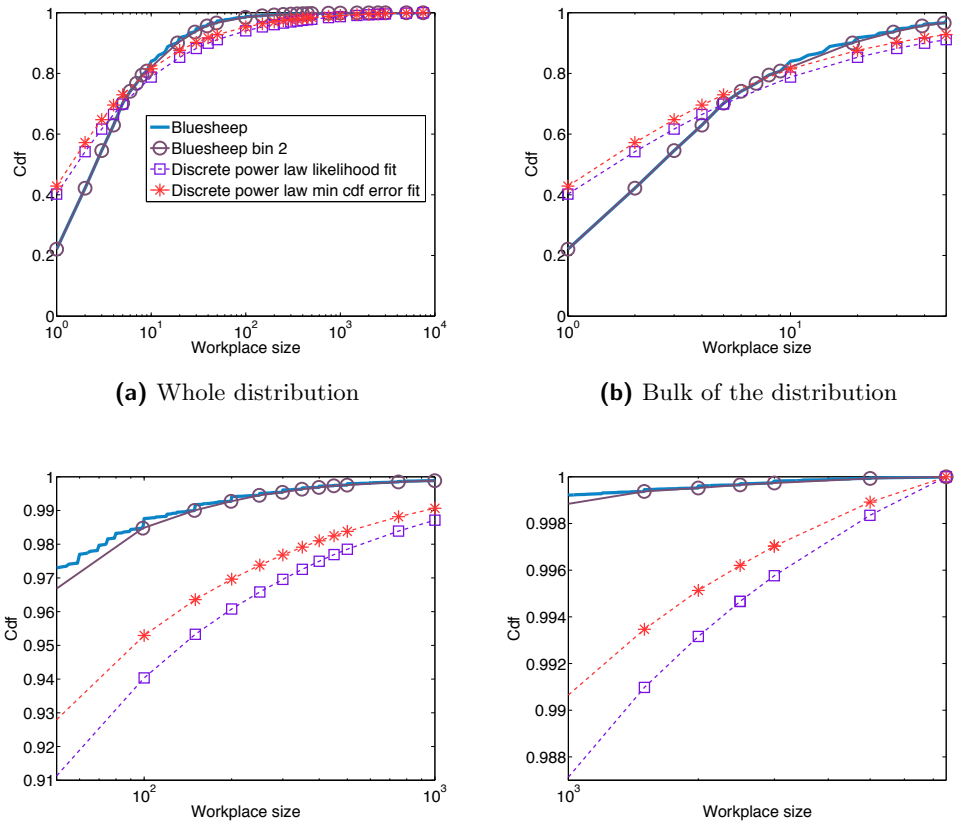
(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.15: Cdf's for the number of employees in a workplace from Blue Sheep data and fitted discrete power law distributions for binning 1. The likelihood fit has parameter $\alpha = 1.54$ with log likelihood $L = -4,735,319$ and for the min cdf error fit $\alpha = 1.63$ with relative log likelihood $L = -22,499$.



(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.16: Cdf's for the number of employees in a workplace from Blue Sheep data and fitted discrete power law distributions for binning 2. The likelihood fit has parameter $\alpha = 1.53$ with log likelihood $L = -5,091,888$ and for the min cdf error fit $\alpha = 1.58$ with relative log likelihood $L = -7,524$ when compared to the maximum.

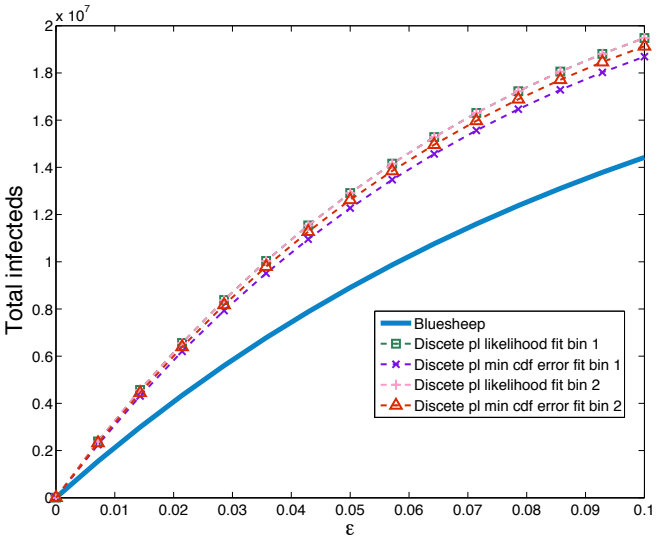
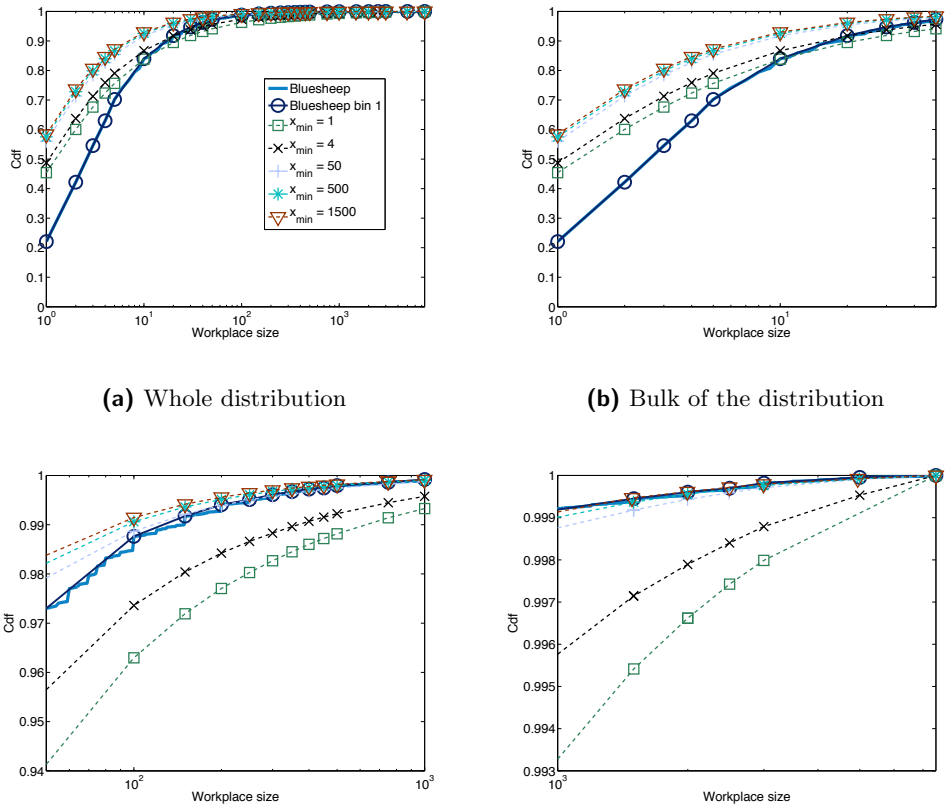


Figure 4.17: Total number of infecteds in workplaces for various values of ϵ when considering the fits to the discrete power law from 4.15 and 4.16.

We see that this is the case in figure 4.17, with all of the discrete power law distributions predicting significantly more infecteds than the Blue Sheep data for all values of $\epsilon \gtrsim 0.01$.

We repeat this fitting for the minimum cdf error for values of $x_{\min} > 1$. We choose $x_{\min}$ which lie at bin boundaries for each set of bins, and choose values up to 1,500. The result of these fits can be seen in figures 4.18 and 4.19. As we would expect, as $x_{\min}$ is increased we get a stronger agreement between the fitted distribution's cdf's and the empirical cdf at larger values of the workplace size with the fits getting better with increasing $x_{\min}$, in exchange for poorer fits in the bulk of the distribution. This translates into a far better approximation of the Blue Sheep data in terms of the size of the epidemic we would expect to see in the population as $x_{\min}$ is increased, which can be seen in figure 4.20.



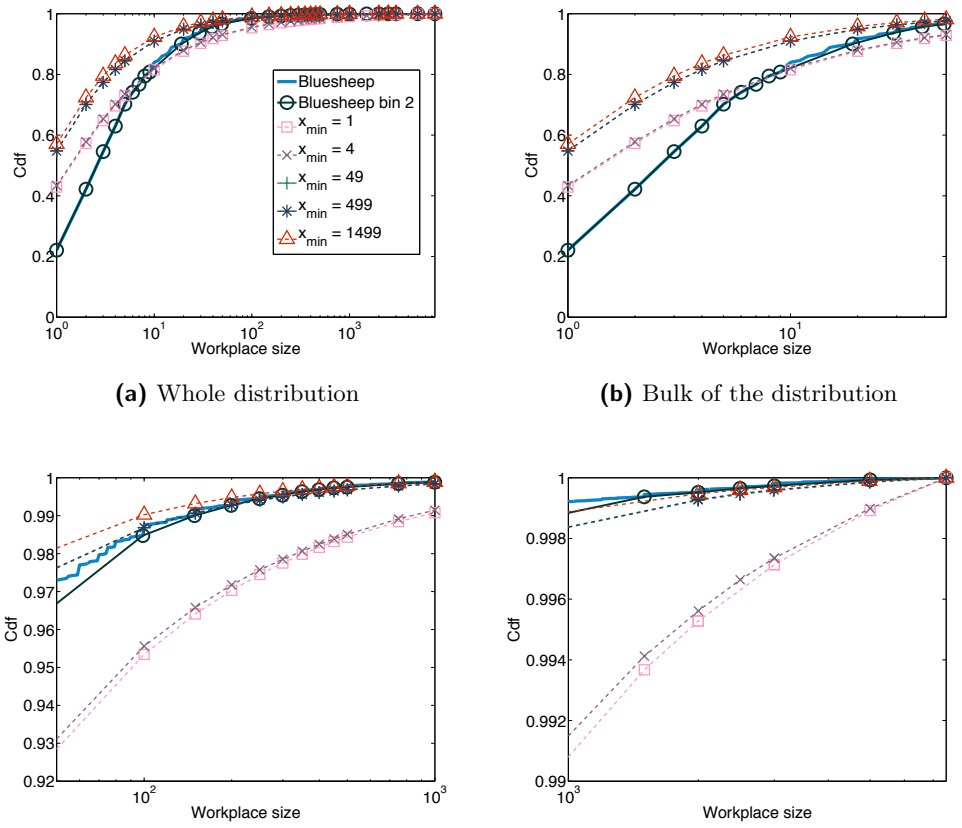
(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.18: Cdf's for the number of employees in a workplace from Blue Sheep data and fitted discrete power law distributions for various values of $x_{\min}$. For $x_{\min} = 1$, $\alpha = 1.63$, $x_{\min} = 4$ has $\alpha = 1.70$, $x_{\min} = 50$ has $\alpha = 1.87$, $x_{\min} = 500$ has $\alpha = 1.91$ and $x_{\min} = 1,500$ has $\alpha = 1.93$.



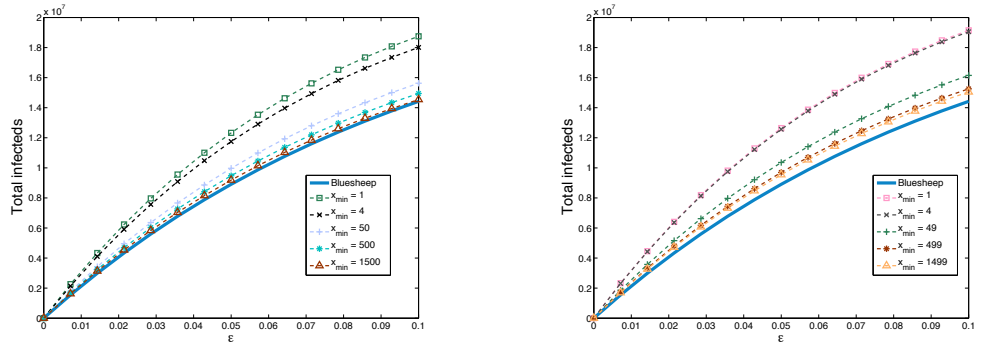
(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

(d) Tail of the distribution

Figure 4.19: Cdf's for the number of employees in a workplace from Blue Sheep data and fitted discrete power law distributions for various values of $x_{\min}$. For $x_{\min} = 1$, $\alpha = 1.58$, $x_{\min} = 4$ has $\alpha = 1.59$, $x_{\min} = 49$ has $\alpha = 1.84$, $x_{\min} = 499$ has $\alpha = 1.89$ and $x_{\min} = 1,499$ has $\alpha = 1.90$.



(a)

(b)

Figure 4.20: (a) Comparison of total infected in population for discrete power law distributions, for various values of $x_{\min}$. Parameters are selected by minimising the absolute error in the cdf of the power law distribution in comparison to binning 1. (b) Here the fitting takes place in comparison to binning 2.

Finally we consider the log-normal distribution, which is often proposed as an alternative distribution to the power law distribution for heavy-tailed empirical distributions.

4.4.4 Log-normal distribution

The log-normal distribution is a continuous probability distribution which exhibits a heavy tail, which has two parameters, a mean μ and a standard deviation σ. The pdf for this distributions is,

$$P(X = x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(- \frac{(\ln x - \mu)^2}{2\sigma^2} \right) . \tag{4.20}$$

However as we previously noted we wish to consider a discrete distribution and only allow values up to $x_{\max}$. We therefore define a normalising constant $\xi(\mu, \sigma, x_{\min}, x_{\max}) = \xi = \sum_{x=x_{\min}}^{x_{\max}} \frac{1}{x} \exp\big(- \frac{(\ln x - \mu)^2}{2\sigma^2} \big)$. The pdf for the discrete log-normal is given by,

$$P(X = x) = \begin{cases} \dfrac{1}{x\xi} \exp\left(- \dfrac{(\ln x - \mu)^2}{2\sigma^2} \right), & \text{if } x_{\min} \leq x \leq x_{\max} \text{ and } x \in \mathbb{N} \\ 0, & \text{otherwise.} \end{cases} \tag{4.21}$$

Again the probability of seeing a value between two bin boundaries is simply a sum of the discrete probabilities for that range. This gives us the following expression for the likelihood of observing our set of observation counts H given our chosen binning B,

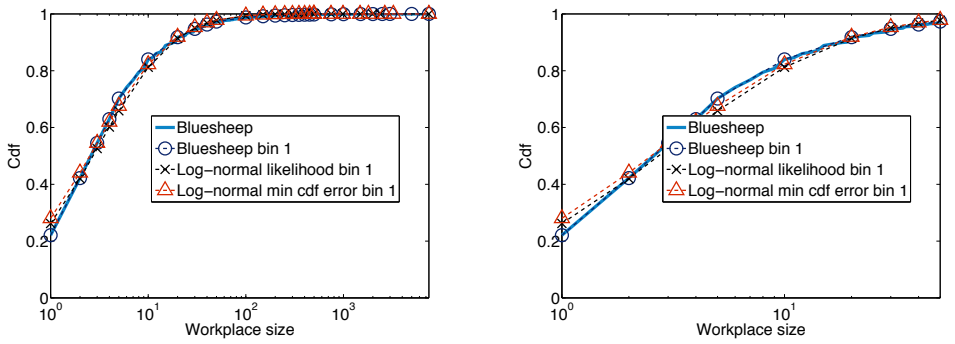
$$P(H|B, \mu, \sigma, x_{\min}, x_{\max}) = \prod_{i=1}^{k-1} \left(\frac{1}{\xi} \sum_{x=b_{i-1}+1}^{b_i} \frac{1}{x} \exp\left(- \frac{(\ln x - \mu)^2}{2\sigma^2} \right) \right)^{h_i} , \tag{4.22}$$

where we define $b_0 = x_{\min} - 1$. The log-likelihood is given by the following,

$$\begin{aligned} L(H|B, \mu, \sigma, x_{\min}, x_{\max}) = & - n \ln(\xi) + \sum_{i=1}^{k-1} h_i \ln \left(\sum_{x=b_{i-1}+1}^{b_i} \exp\left(- \frac{(\ln x - \mu)^2}{2\sigma^2} \right) \right) \\ & - \sum_{i=1}^{k-1} h_i \ln \left(\sum_{x=b_{i-1}+1}^{b_i} x \right) . \end{aligned} \tag{4.23}$$

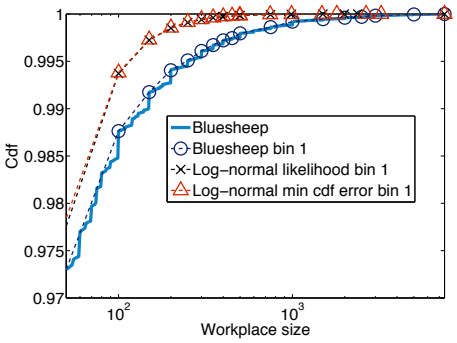
The fits provided by this distribution are seen in figures 4.21 and 4.22. We see very close fits to the Blue Sheep data in the bulk of the distribution, but as we move into the tail, the log-normal distributions cdf's are much higher than the actual data. This implies that we will observe smaller epidemics than for the Blue Sheep data, which is confirmed by figure 4.23.

As for the discrete power law, we could fit these distributions using values of $x_{\min} > 1$, which would give a better fit to the tail of the Blue Sheep distribution. For the sake of brevity however, these have been omitted.



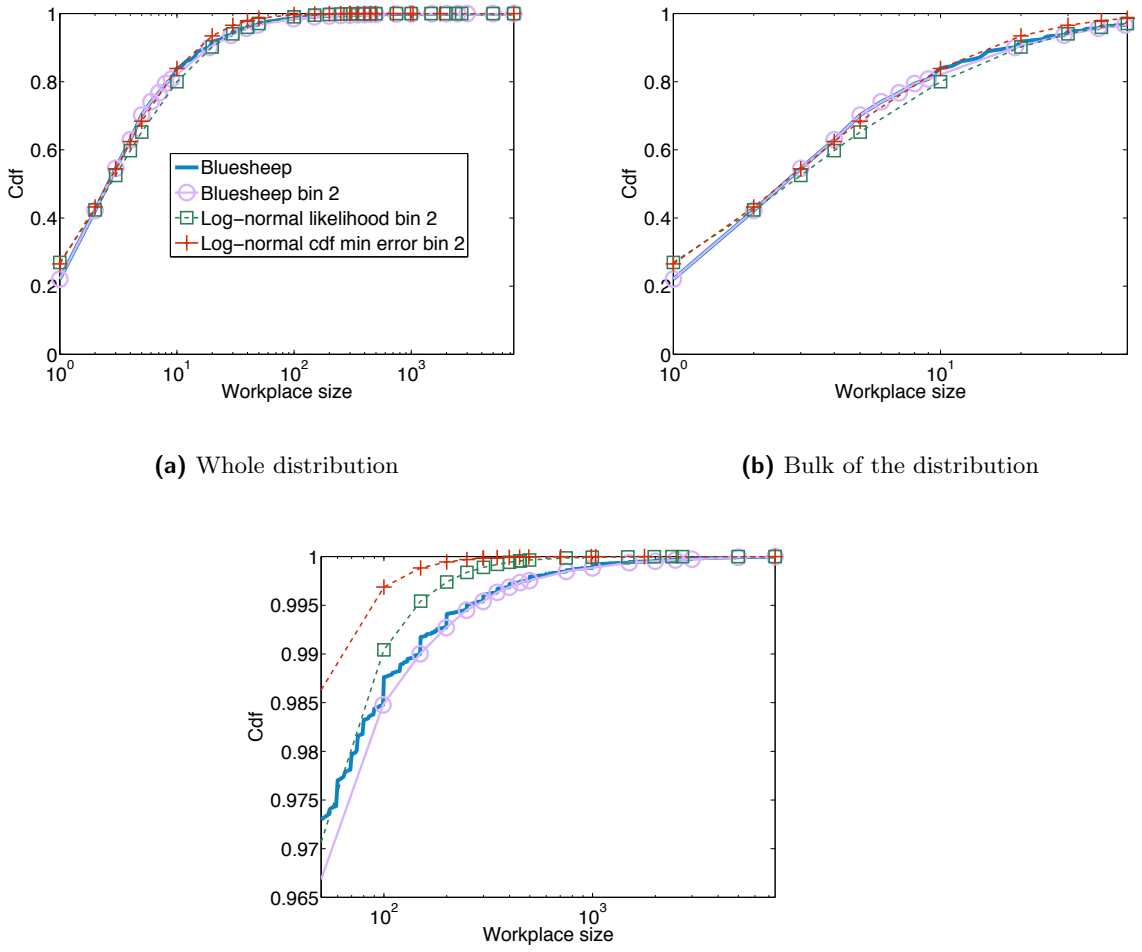
(a) Whole distribution

(b) Bulk of the distribution



(c) Bulk to tail

Figure 4.21: Cdf's for the number of employees in a workplace where we attempt to fit a log-normal distribution to the Blue Sheep data. The distribution with the best likelihood has parameters $\mu = 0.92$ and $\sigma = 1.45$ and a log-likelihood value of -5,830,900, whilst the set of parameters which give the minimum error in the cdf of the distribution are $\mu = 0.79$ and $\sigma = 1.49$ and a relative log-likelihood value of -2,372 compared with the maximum likelihood.



(a) Whole distribution

(b) Bulk of the distribution

(c) Bulk to tail

Figure 4.22: Cdf's for the number of employees in a workplace where we attempt to fit a log-normal distribution to the Blue Sheep data. The distribution with the best likelihood has parameters $\mu = 0.82$ and $\sigma = 1.57$ and a log-likelihood value of -5,923,000, whilst the set of parameters which give the minimum error in the cdf of the distribution are $\mu = 0.93$ and $\sigma = 1.33$ and a relative log-likelihood value of -23,471.

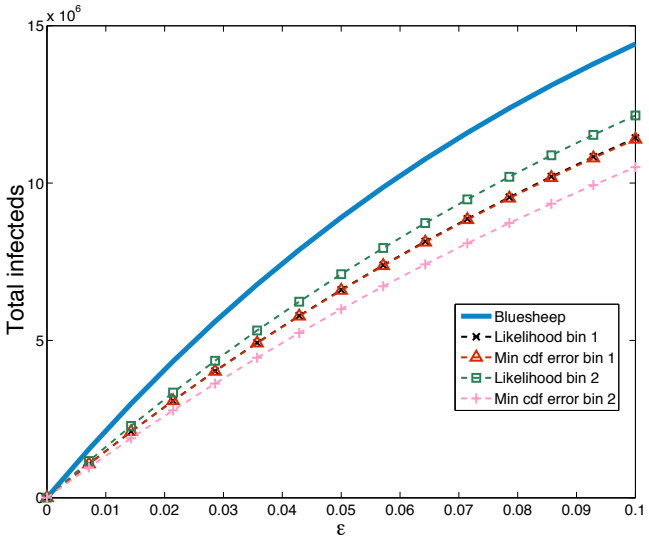


Figure 4.23: Total infecteds for the population for fitted log-normal distributions in comparison to raw Blue Sheep data.

4.5 Discussion

When we consider the best fit in terms of likelihood value, the offset truncated power law distribution gives us the best fit to the data. The set of parameters which have the greatest likelihood are found when we fit to the data using likelihood methods for the first set of bins. This gives a log-likelihood value of -4,457,400. For the discrete power law, the best log-likelihood is -4,735,419 and for log-normal it is -5,833,272. This means that simply considering likelihoods, the offset truncated power law is the best fit to the data by a substantial margin with values of the parameters for this distribution being $a = 4.40$ and $c = 1.47$.

However as we want to be able to say something about the seriousness of a potential epidemic in the population, it may be more useful to consider which set of parameters will give us the closest agreement with the Blue Sheep data in terms of the final epidemic size.

To make this comparison, we take sum the absolute values of the difference from the Blue Sheep prediction to our fitted prediction at 15 different values of ϵ, spread evenly from 0.00001 to 0.1. We can then choose the set of parameters that minimise this for each distribution and then compare these values to each other to find the 'best' fit by these criteria.

When we perform this analysis for $x_{\min} = 1$ we find that the offset truncated power law distribution is again by far and away the best choice. However the set of parameters which give us the closest agreement to the Blue Sheep data are gained by minimising the difference in the cdf's between the Blue Sheep data and the distribution for binning 2, and the parameter values are $a = 3.10$ and $c = 1.20$. The minimum cumulative difference we find over the 15 points is around 1,300,000 on average for these parameters. In fact using

the parameters which give the best likelihood give the greatest discrepancy in the number of infections when considering only the offset truncated power law and the difference in the final sizes is 17,300,000 on average.

Despite this large leap between the best and worst fits, the set of parameters which performs worst from the offset truncated power law distribution still beats all parameter fits for the log-normal and discrete power law distributions. For the log-normal, the best fit is for the likelihood fit of binning 2 and the difference is 23,200,000 whilst for the discrete power law, our best fit is from the minimum cdf error to binning 1 and the error is 43,500,000. This shows how poor the log-normal and discrete power law are when it comes to this measure, which tells us that the offset truncated power law distribution is definitely more descriptive of the data.

To achieve a more predictive fit to the amount of infection in the population for the discrete power law, we increased $x_{\min}$ and then attempt to minimise the absolute error in the cdf between the distribution in question and the Blue Sheep data. Doing this, we can get a set of parameters which is far more accurate in terms of the number of infecteds as can be seen best in figure 4.20a. Here using $x_{\min} = 1,500$ for the discrete power law, the cumulative error is 3,500,000 which is far better than using lower values of $x_{\min}$ but the best offset truncated power law distribution is still more than twice as good.

We have seen that these chosen distributions fit the cdf of the data with varying degrees of success, well for the offset truncated power law to poorly for the discrete power law. This translated into a strong or poor agreement with the Blue Sheep data for the final sizes of an epidemic in a population for various values of ϵ. It was shown that the strength of the agreement here was highly dependent on the tail rather than the bulk of the distribution. In terms of fitting the data as closely as possible, we therefore have to decide what measure is best to measure the success of the fit by. If we were simply interested in fitting the data to a certain distribution, so that we could say that the data followed this distribution, then we could simply choose the distribution and set of parameters which gave us the largest value for the likelihood.

We can therefore conclude that the offset truncated power law distribution fits the data more satisfactorily than the discrete power law or log-normal distribution as it outperforms both by a considerable margin in terms of likelihood and predictive power.

However we note that the set of parameters that we would report as fitting the data best depends on what we are interested in doing with the distribution. If we simply wish to report the most probable set of parameters in terms of the ability to describe the cdf of the data, we would report that $a = 4.40$ and $c = 1.47$. However if we were interested in studying what the effect of the parameters is in terms of final size of an epidemic, we would report that $a = 3.10$ and $c = 1.20$. Therefore what is to be done with the information is therefore an important consideration to keep in mind when attempting to fit a distribution to data.

To increase the agreement between the worst performing distribution, the discrete power law, and the Blue Sheep data, in terms of the predicted number of infecteds, we have also fitted to the cdf for different values of $x_{\min}$. For a value of $x_{\min} = 1,500$, we achieved a good agreement between the power law and the actual data for the predicted number of infecteds. This was much improved when compared to any set of parameters we get from using $x_{\min} = 1$. This tells us that the tail of the distribution is of great importance in order to characterise the possible spread of the infection through the workplace population,

and it is not instructive to simply find a parameter set which fits well in the bulk of the distribution and be confident that this is describing the data well in a way which you are interested in.

In general, it is interesting to see how the profile of the different distributions affects not only the fit of the distributions to the cdf, but also changes the way in which the total number of infections predicted differs from that of the Blue Sheep data. For the discrete power law the large workplace sizes are over sampled when we choose parameters to match the empirical cdf. On the other hand, the log-normal and offset truncated power law distributions have select more small workplace sizes. This means that the number of infecteds that these distributions predict can be greater than the data suggests (discrete power law, 4.17) or fewer (offset truncated power law, 4.10 and log-normal, 4.23).

If we are interested in including workplaces in the spread of epidemics in countries for which we have no data on the workplace size distribution and no idea what the distribution may be, then it is plausible that we may select a discrete power law as this will in all likelihood, not underestimate the severity of a potential epidemic, though this may produce a worst case scenario which is difficult to believe. However as it has been shown that for the UK the offset truncated power law gives us the best fit (of distributions considered), and this was the distribution produced for the US in Ferguson et al. [2006], it is likely that, for economically developed Western countries at least, this is a fair choice of workplace size distribution.

4.6 Attack Rates

In the previous sections of this chapter we have used the deterministic mean-field approximations to calculate the total expected final size in the workplaces when we consider different possible distributions of workplace size along with non-frequency dependent transmission rates. The use of non-frequency dependent transmission rates implies that for distributions which contain greater numbers of large workplaces, the mean transmission rate will be larger than for others distributions.

The implication of this is as follows. If we consider all of the workplaces to form a population, then if we select an individual chosen at random from this population and calculate the average number of people they will infect, then this will be higher in the distribution which has more large workplaces. This is essentially the definition of R_0. Therefore consideringtwo different workplace distributions is equivalent to considering spread of two diseases with different values of R_0. Though this is caused by the increased average workplace size, this increase in the average transmission rate means that we are not exploring the impact of this increase of average workplace size in a fair way.

A fairer way to compare the impact of the distribution of workplaces on the spread of the epidemic is to keep the mean transmission rate constant across distributions. One way of achieving this the same is to use frequency-dependent transmission. However, this will not allow us to investigate the difference between small and large workplaces, as they will be homogeneous in the spread of the infection from person to person. If we keep the mean transmission rate constant in different distributions by scaling all of the transmission rates in a consistent way, then this will allow us to investigate the dependence on the distribution in terms of how large workplaces behave differently to small workplaces if average behaviour is the same. To do this we can consider the overall attack rate or secondary attack rate in a single workplace from a single introduced infection, and then see how this scales up to population level when we keep the mean transmission rate constant.

The overall attack rate gives us the proportion of individuals who are infected in a population by an initial infected individual, and gives us information on the potential impact of an epidemic in the population. When referring to the secondary attack rate, which has been defined in multiple ways, we mean the proportion of people who are directly infected by the initial individual.

A discussion of how to scale the transmission rates precedes a discussion of attack rates, and follows here.

4.6.1 Scaling transmission rates

As mentioned above, we wish to scale transmission rates so that the mean over all workplaces is the same, but we still have variation for each individual workplace size. To do this we begin by defining $\beta_{k,\epsilon}$ to be the transmission rate for workplace size k and a given value of ϵ. As in the previous sections, we still want the transmission rate to be approximately inversely proportional to k^ϵ. As we are concerned with the spread within individual workplaces of a given size k, the number of contacts in the workplace will be at most $k-1$. The average number of contacts over the whole population of workplaces is therefore given by $\sum_l d_l(k-1)$. Therefore assuming a rate of contact proportion inversely

proportional to the number of possible contacts with these individuals as in the above sections, we have that the average contacts made by members of the population is given by $\sum_l d_l(k-1)/(k-1)^{1-\epsilon} = \sum_l d_l(k-1)^{\epsilon}$.

As we still wish the transmission rates to be proportional to the inverse of the number of contacts in the workplace, we define

$$\begin{aligned} \beta_{k,\epsilon} =& \rho,\ \epsilon = 0 \\ \beta_{k,\epsilon} =& \frac{\theta_\epsilon}{(k-1)^{1-\epsilon}} \frac{\rho}{(\sum_k d_k(k-1)^{\epsilon})},\ \epsilon > 0 \end{aligned} \tag{4.24}$$

where ρ can be changed in order to increase or decrease the amount of transmission we introduce to the population and θ_ϵ is used to appropriately scale the values of the transmission rates when necessary.

When we have frequency dependent transmission $\epsilon = 0$ this gives that $\beta_{k,0} = \rho/(k-1)$. It is worth noting that when $\epsilon = 0$, ρ is equal to the value of R_0, as the expected number of cases produced by this initial infected will be $\beta_{k,0}(k-1) = (k-1)\rho/(k-1) = \rho$.

When $\epsilon > 0$ we then use θ to scale all of the values of $\beta_{k,\epsilon}$ for any particular value of ϵ so that $N\sum_k d_k\beta_{k,\epsilon}$ is equal to the value of $N\sum_k d_k\beta_{k,0}$, resulting in the same mean transmission rate.

4.6.2 Overall attack rate

The overall attack rate is defined as the proportion of at risk individuals in a population who are infected during an epidemic. Here we assume that the whole population, and specifically all members of a workplace excluding the initial infectious individual, is at risk. To calculate the overall attack rate we consider the final size of an epidemic in a workplace if we introduce a single infectious person into it.

In a workplace of any size k, we can use Bailey's method [Bailey, 1957] to calculate at machine precision, without use of Monte Carlo simulation, the probability that the final size will be z, i.e.

$$p_z(k|\epsilon) := \lim_{t\to\infty} Pr[(S(t), I(t)) = (k-z, 0)|(S(0), I(0)) = (k-1, 1); \beta_{k,\epsilon}; \gamma] \ , \tag{4.25}$$

for z between 0 and $k-1$, where $\beta_{k,\epsilon}$ is the transmission rate in a workplace of size k for the value of ϵ in question, and γ is the recovery rate. Then to calculate the overall attack rate (OAR), we evaluate the following,

$$OAR_\epsilon = \frac{\sum_k d_k \sum_z z p_z(k|\epsilon)}{\sum_l d_l(l-1)} \ , \tag{4.26}$$

where d_k is the proportion of workplaces of size k and z ranges from 0 to $k-1$.

Figure 4.24a shows the results gained from calculating the overall attack rate for different workplace distributions. To do this we calculate the $\beta's$ as in (4.24), and input the appropriate values of $\beta_{k,\epsilon}$ into (4.25). Finally we use (4.26) to get the values of the overall attack rate. We do this using the Blue Sheep data, and for comparison, the best fitting discrete power law distribution for $x_{\min} = 1$ from 4.4.3.

This can be explained by considering figure 4.24b, which shows the expected overall attack rate by population size, which is calculated in the same way as stated previously. For a low value of ρ, 0.5 here, the OAR begins at about 33% for a population size of 2, and by the time the population size is 10 this is down to 8%, and quickly tends towards 0. The ratio between the OAR at population size 2 and 10 is approximately 4.2. Between 2 and 100, it is 34.79. As the power law distribution contains more large workplaces when compared to the workplace data this results in the Blue Sheep data having a higher OAR.

Making the same comparison between population size 2 and 10 when $\rho = 2$, only gives a ratio of 1.51 between the two OAR's, and even comparing 2 with 100, the ratio is still only 1.74. Therefore the different distributions are not dissimilar enough to cause a large difference in the overall attack rate considering the whole distribution.

As ϵ increases, the attack rate for the discrete power law is larger than that for the Blue Sheep data, again the reason for this can be seen in figure 4.24b. Taking the extreme case of $\epsilon = 1$ we see that as the population size increases, so does the value of the OAR. This is as we would expect as for $\epsilon = 1$ the transmission rate per neighbour stays constant as the size of the population increases, meaning that the force of infection is greater. This increase in OAR for larger workplaces is large enough that where the OAR for the power law when $\epsilon = 0.5$ is very close to values for the $\epsilon = 1$ case in the Blue Sheep data set.

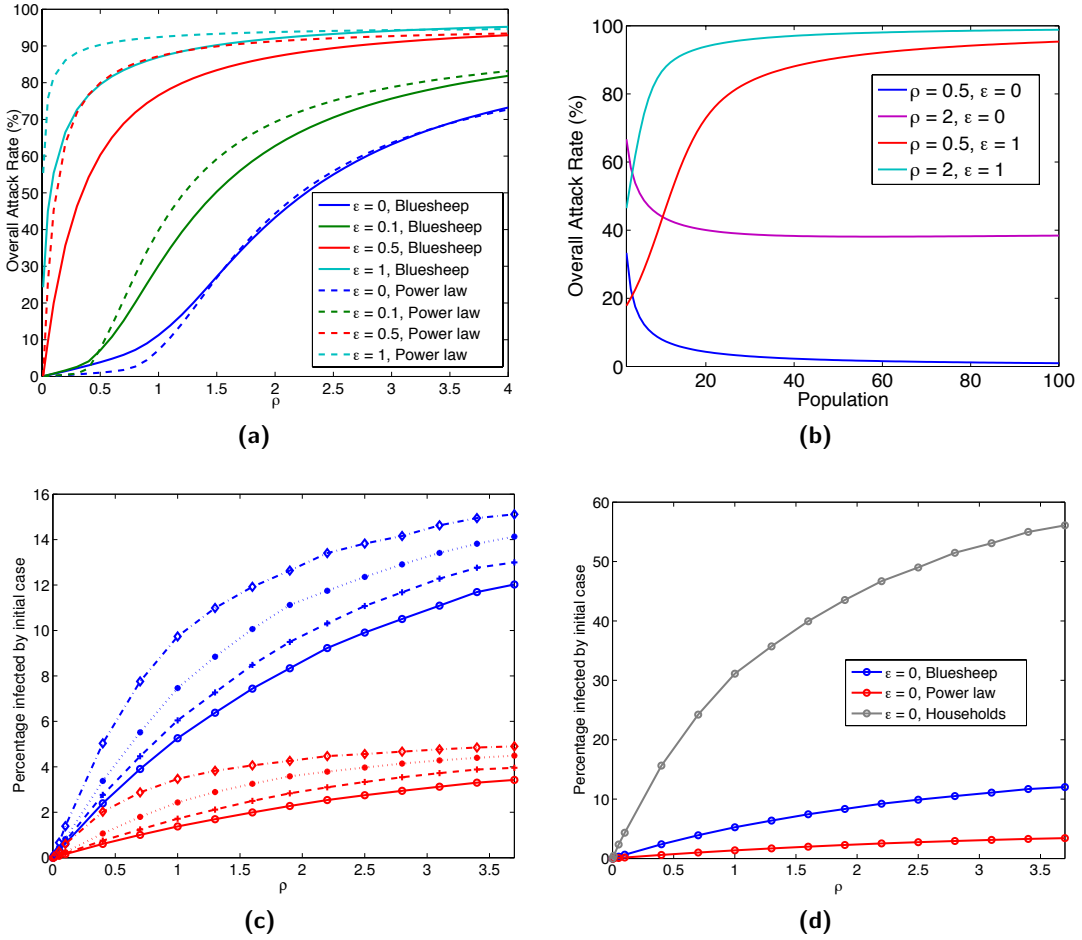


Figure 4.24: Comparison of attack rates for different datasets produced by use of Bailey's method and simulation using Gillespie's algorithm. (a) shows the overall attack rate for the workplace data and the best fitting power law distribution from §4.4.3. (b) shows the changes in the overall attack rate in a single location whose population size is given by the x-axis for two different values of ρ, when we consider frequency dependent transmission compared with density dependent transmission. For frequency dependent transmission, as population size increases, OAR decreases, whilst for density dependent transmission, the OAR increases with population size. (c) compares the value of the expected SAR for Blue Sheep data in blue, against the power law distribution in red. The four different lines for each distribution show the different values of SAR for $\epsilon = 0$, 0.1, 0.25 and 0.5 from lowest to highest in each colour. (d) shows the value of the SAR for the Blue Sheep data, power law distribution and household distribution for the UK.

4.6.3 Secondary attack rate

The secondary attack rate is the often defined as the proportion of cases among susceptible contacts of the primary infected individual [CDC]. However as we are assuming that into each workplace a primary infected is introduced, and that all people in the workplaces are connected to each other, using this definition is equivalent to the overall attack rate. Therefore instead of this definition, we define the secondary attack rate (SAR) to be the proportion of at risk individuals that are directly infected by the initial infected. Again

we are assuming that the whole of the population is at risk. Note that the average number of people directly infected by the initial infected person is equal to the value of R_0, which we are keeping constant across populations and values of ϵ here.

There is no deterministic calculation for this definition, as Bailey's method is for the final size, so to calculate this, simulation is required.

This can be achieved by using the Gillespie algorithm [Gillespie, 1977], whilst keeping track of who infected whom. This is easily achieved in a fully connected workplace of size k by defining, I_{ij} to be the number of people infected by node i in workplace j. As the workplace is fully connected, we can always make the initial infected node 1, and then make the n-th infection be on the $(n+1)$th node, e.g. the first infection (if it occurs before the initial infected recovers) will always be on node 2. This implies that the index for the I_{ij} variable will only have to go up to $k-1$, as the k-th node will never have anyone left to infect.

Defining N to be the number of workplaces in our dataset, or in a dataset generated from a given approximating distribution, we also define n_j to be the number of people in workplace j. The secondary attack rate is then given by,

$$SAR = \frac{\sum_{j=1}^{N} I_{1j}}{\sum_{j=1}^{N}(n_j - 1)} = \frac{\sum_{j=1}^{N} I_{1j}}{\sum_l d_l(l-1)} \ . \tag{4.27}$$

This is however not an ideal method for calculating the SAR across all possible distributions, as it requires us to run this calculation many times for each specific dataset to get an average across the entire dataset. Instead we run the Gillespie algorithm until the initial node is recovered 1,000 times for all values of k from 2 to the maximum value of the Blue Sheep data which is 7,500. This then gives us an expected value for the secondary attack rate in any workplace of size k, given the value of ϵ which we can call $SAR_{k,\epsilon}$. The value of the SAR for any dataset given ϵ is then,

$$SAR_\epsilon = \sum_k d_k SAR_{k,\epsilon} \ . \tag{4.28}$$

This is an interesting value to look at, as along with the overall attack rate, this will tell us what level of infection can be prevented by acting quickly and intervening at the first sign of infection in an individual. For example if the overall attack rate is 50%, and the secondary attack rate is 45%, then the benefit of intervention may be different than it would be if the secondary attack rate is only 20%.

Figure 4.24c shows the values of the secondary attack rate for the Blue Sheep data for various values of ϵ and ρ in blue, and the same for the power law distribution in red. We can see that the SAR for the Blue Sheep data is larger than for the power law distribution, despite the fact that the reverse is often true for the overall attack rate by figure 4.24a. This is due to the same fact that gives the power law distribution a larger OAR, the greater number of large workplaces. Even though, for non-zero values of ϵ, the transmission rates are increased in large workplaces, the fact that there are a greater number of people in them implies that more time is needed to infect all of them. As was shown in § 4.6.2, this increase in transmission rates increases the overall attack rate for values of $\epsilon \geq 0.1$, the implication is that more transmission is done by secondarily infected individuals.

In addition to calculating the secondary attack rate in workplaces, we can look at the SAR

in other places. One obvious distribution worth considering is that of households, as along with workplaces, households have a large impact on the spread of an epidemic. The UK census collects information about the size distribution of households [2001 census], so we use this distribution to consider the SAR in households. In figure 4.24d this is calculated for the $\epsilon = 0$ case and is given by the grey line. This was then compared with the value of SAR for Blue Sheep data in blue and the power law distribution in red, also for the $\epsilon = 0$ case. We can see that the secondary attack rate is far greater for households than either the Blue Sheep of power law distribution. This demonstrates the point above that an increase in the number of contacts will decrease the value of the SAR. This is due to the fact that the size of households necessarily have a much lower upper limit than workplaces, along with many more small households. For example, if the mean household size is a third the size of the mean workplace size, then for the same value of R_0, we would expect to observe a value of the secondary attack rate in the households which is triple that of workplaces.

4.6.4 Discussion & Limitations

In this section of the chapter we have investigated the change in overall and secondary attack rates which is brought on by a change in workplace size distribution, when R_0 is kept constant. This differs from the first part of the chapter where the average value of R_0 throughout all workplaces was allowed to increase as the proportion of the population who worked in large workplaces increased.

We have shown that the attack rates throughout a set of workplaces are keenly dependent on the distribution of the sizes of the workplaces. For low values of ϵ (< 0.1) the overall attack rates seen in large workplaces are smaller than for smaller workplaces, but this is reversed when ϵ increases.

For the secondary attack rate, in general the larger a workplace is the smaller the attack rate will be, due to the sheer weight of numbers in the large workplaces. Even in the situation where we have fully density dependent transmission, $\epsilon = 1$, the increase in the force of infection for large workplaces is insufficient to make the value for the power law distribution larger than for the Blue Sheep data.

We have seen that the overall attack rate, which can be thought of as defining a worst case scenario may be worse in a country where there are a lot of workplaces with many people interacting with each other. However due to the decrease in secondary attack rates in these same places if equally effective controls can be put in place at the first sign of infection, the worst case scenario can be successfully avoided. In fact the impact may be smaller than in countries with improved worse case scenarios, due to the inability of the initial infected case to infect a large proportion of individuals.

There are several limitations with this approach. Firstly, it is currently unknown how people actually mix in the workplace, which results in us being unable to properly characterise the interactions which take place in the workplace. Assuming that the population is well mixed is a gross simplification of the problem, but can be seen as providing us with a worst case scenario. If an infection is extremely transmissible, this may even effectively be the case, therefore the results that we obtain from assuming this can be informative.

The interaction with how infections are spread in households and through contacting people at random is also of key importance, and we have ignored this. It is clear that

people do not spend all of their time in a workplace. However, we can also think of this analysis as comparing the spread of an infection on different contact networks, where the numbers of contacts that people have is the same as a specific workplace distribution.

Chapter 5

Modelling of large human populations

5.1 Introduction

In the modelling of large human populations, there is much heterogeneity in contact patterns which we wish to capture in a systematic and descriptive way. Attempting to capture these contact patterns is extremely difficult as there are many complicating heterogeneous factors in the behaviour of humans which need to be considered. For example, this heterogeneity can come from the fact that contacts are made within different sub-populations of the population.

The workplaces examined in the previous chapter form a set of sub-populations within which people mix and form contacts, separated from the rest of the population or their own home or social contacts. This therefore suggests the idea that splitting the population into distinct areas and allowing the epidemic to spread in these populations may capture an interesting aspect of disease spread amongst humans. Each person can be assigned to multiple sub-populations, each one describing a different context in which contacts are made. When this is done, the resulting population can be referred to as an individual-based model (IBM).

What follows in this chapter is a review of models that are used in an attempt to describe the spread of epidemics in realistic human populations and ways in which attempts have been made to collect data which is relevant to these models. I then describe the construction of a large data-derived synthetic population of England and Wales, which was a large part of the work done during my PhD.

5.1.1 Meta-population models

In ecology one of the first uses of meta-population models was developed to study the migration of birds from a mainland to islands in the ocean [MacArthur and Wilson, 1967; Hanski, 2001], where the proportion of islands colonised is tracked, and the rates at which species become extinct and colonise islands are taken to be constants.

Rather than considering the number of areas colonised, the populations of different sub-populations can be tracked [Fulford et al., 2002], leading to SIR like equations for the

population sizes.

Migration in terms of human populations has minor effects on the spread of infection, as it is a relatively rare phenomenon. Therefore for humans the use of meta-populations is usually used to describe the mixing of people in different environments (home, work, school ...), to describe the mixing of different sub-populations (such as towns or villages) areas with different characteristics in terms of disease susceptibility (such as prevalence of vaccination or immunity to infection).

In Sattenspiel and Dietz [1995] a population is divided into n separate regions with rates of moving from region to region defining a set of differential equations for each region. This can be thought of as an expression of how commuting takes place in a population, and can incorporate migration, by having different total rates in and out of regions. This approach is described as a 'mobility process' in the paper, as it simply describes the movement of people between regions. A disease model can easily be constructed from this mobility process, as the rates of movements between regions can carry an infection between regions at a rate dependent on the number of infective individuals in each region. A measles model is constructed using these techniques, and the impact of connecting two cities, one of which has a value of R_0 above the epidemic threshold and another below this threshold when they are disconnected, can have on these values of R_0 is studied by Arino and van den Driessche [2003].

Along with a local epidemic threshold, R_0, which indicates the extent spread of the epidemic in each region of the population, Colizza and Vespignani [2008] calculate the global invasion threshold $R_\star$, which must be greater than 1 if the epidemic is going to spread into an increasing number of regions in the global population, and is dependent on the disease dynamics, along with the amount of movement between regions.

The use of meta-populations also allows the investigation of heterogeneity in the population in ways that affect the spread of a disease. In Metcalf et al. [2013] (along with many earlier sources such as Gotelli [1991], Ostfeld and Keesing [2000] and Gilbert et al. [2001]), the re-introduction of diseases into sub-populations is termed "rescue effects", which was first used by Brown and Kodric-Brown [1997] to describe the effect that immigration has on the extinction of species living on islands. For four childhood diseases; measles, mumps, rubella and whooping cough, it was shown that the probability of extinction of the disease was roughly twice as high on island nations than non-islands, due to the decrease in global connectivity associated with being on an island, hence leading to a decrease in the regularity of rescue effects.

Additionally meta-populations can model the spread between different species, such as between rodents and humans for the spread of bubonic plague by Keeling and Gilligan [2000]. Here it was shown that the spread among the rodent population, without imports from outside the population in question can explain the intermittent spread among humans. Another obvious candidate for such analysis is that of the spread of vector-borne (mosquito spread) diseases such as malaria, chikungunya or dengue fever. It has been shown that the spread of these diseases can be explained by the movement of humans amongst static mosquito populations [Adams and Kapan, 2009]. Along with this, the movement of humans, which is taken from mobile phone probes [Gonzalez et al., 2008], combined with changing the distance that mosquitos travel, can drastically increase the epidemic height and decrease the time to peak in a vector borne disease epidemic [Moulay and Pigné, 2013].

A drawback of using differential equation models for meta-populations is that the number of equations needed increases with the number of regions or species included. This can become unwieldy if there are a large number of regions. However direct simulation of these populations is possible with the use of stochastic simulation techniques as described in §2 and is fast for this type of model.

A step up from meta-population models in terms of complexity is that of individual based models, which track each member of the population individually. These are becoming more relevant to the study of epidemics, as the increase in available data means that rather than assigning an average behaviour to a group of people, we can instead select behaviour from data-derived distributions which will describe the spread of the infection more accurately.

Next I describe data which is typically collected in order to inform models of disease spread such as individual based models. Following that is a discussion of individual based models which have been used previously.

5.1.2 Attempts to characterise contact structure

The main difficulty in using network models, in terms of attempting to represent reality, is that it is difficult to know the exact network on which an epidemic is spreading. The difficulty of this varies from disease to disease, as the nature of the contact required to spread the infection influences the difficulty of creating a representative network. For sexually transmitted diseases or those which are spread by needle sharing in injecting drug users, respondent driven sampling [Heckathorn, 1997; Mills et al., 2012] can aid in the construction of realistic networks, though whether the biases inherent in these samples can be overcome is not clear [McCreesh et al., 2012].

Surveys of the population are also widely used to understand the network of sexual contacts that exist in the population. For example the National Health and Social Life Survey conducted in 1992 sampled 1,511 men and 1,921 women in the USA and was used to examine the apparent over representation of STDs amongst the African American community, estimated to be more than 10 times more prevalent, when compared to other ethnic groups in the US [Laumann and Youm, 1999]. It was shown that the concentration of STDs in this community was in part due to the fact that the choices of partner are more likely to stay within the same community than for other communities. Along with this, the network defined by the sexual contacts formed by African Americans was found to be far more degree disassortative than for others. This implies that people who have small numbers of partners are likely to make contacts in the "core group". This group is defined by Hethcote and Yorke [1984] to be the group who are extremely sexually active and efficient transmitters of STDs. The degree disassortative mixing observed allows infections to spill over into the entire African American population more easily. This core group has been seen to exist in many networks of sexual interactions [McKusick et al., 1985; Handsfield et al., 1989; Wadsworth et al., 1993; Liljeros et al., 2001; Erens et al., 2001]. The number of partners of this group is characterised by a large variance and has been shown to follow a power-law distribution [Schneeberger et al., 2004], and to have an important impact on the spread of diseases [Anderson and May, 1988; Gómez-Gardeñes et al., 2008]. The impact of these degree heterogeneities have been studied extensively [Eames and Keeling, 2002; Gómez-Gardeñes et al., 2008].

For diseases which are not sexually transmitted, influenza for example, it is unknown exactly what type of contact is enough to spread the infection on. This is obviously a major obstacle to obtaining accurate networks for these types of diseases, and can lead to having an average degree which is higher or lower than the true value. A commonly used method to attempt to capture these networks is again to survey people. Here people are asked to self-report numbers of people they meet each day that fit a given definition of a contact, or are asked to keep diary type data for extended periods of time. There have been several such studies of this form.

The POLYMOD study [Mossong et al., 2008] surveyed 7,290 participants who recorded characteristics of 97,904 different contacts, including demographic information such as age about the people with whom they contacted. A contact was defined as physical contact or a two-way conversation with our without physical contact. In Danon et al. [2012] over 5,000 people in the UK were again asked to give information on the number of people with whom they had had a face to face conversation with, whether there was any physical contact and how long these interactions took place for. There was also the possibility to enter groups of people at once, to allow people to record instances where there were a lot of interactions occurring simultaneously, and information was also given on which of an individuals contacts also met each other, allowing a measurement of the clustering of the network to be made.

In Read et al. [2008], 49 people, who were all staff or students of the University of Warwick, were asked to fill in a survey for 14 non-consecutive days with intervals of 10 days. 47 of these people completed this task for at least 9 days and 8,661 individual contacts involving 3,528 unique individuals were recorded along with the situation type of the interaction, with 'Home', 'Work', 'Shopping', 'Travel' and 'Social' being the categories for the interactions and whether the interactions were one-off or repeated. Again a contact implied a face to face conversation, and incidence of physical contact was also recorded.

The impact of the number of repetitions of a given contact, along with the situation in which the contact took place were investigated by Read et al. [2008] and Eames et al. [2009]. Contacts which were repeated were seen as more important to the spread of the epidemic and were therefore given more weight in the network. To consider this, weighted and unweighted networks were constructed and compared to weighted and unweighted mean field models. The predicted final size of the epidemic, along with the impact of different vaccination strategies was compared and it was seen that the addition of weights to the network for a given transmission rate could lead to a smaller predicted epidemic size, whilst also making targeted vaccination strategies difficult to manage.

An urban and rural population of 1,821 participants from 856 households, across 40 communities near Guangzhou China was surveyed with 12,147 total contact events, including 33,789 people in 4,803 different locations [Read et al., 2014]. Duration of contact, along with distance from home was also recorded, allowing the impact of location along with duration on number of contacts and clustering of links to be investigated. The results of this survey broadly agree with findings of the POLYMOD contact survey [Mossong et al., 2008] and was in agreement with [Danon et al., 2012] in that the level of clustering between contacts increases with distance from home.

These are obviously very worthwhile studies of the network on which a disease such as influenza is spread on, and the fact that there is qualitative agreement between many studies in different countries and types of population is encouraging. However it was also shown from the survey performed in China, that the contact structure was not sufficient to

describe the levels of immunity to influenza A viruses [Lessler et al., 2011], meaning that even if we can characterise the interactions that people have on a day-to-day basis, there are still factors that determine the extent of disease spread that are not captured.

Along with this, the accuracy of self-reported contacts is also an issue to consider, whether people misremember or miscount the amount of contact that they had on a given day and whether the type of contact needed to spread a specific infection differs from what is asked for in the survey. For an infection such as measles it is likely that speaking with someone for any period of time is enough to transfer the infection to them [Paunio et al., 1998], but for diseases which are less transmissible it is less clear what contact is sufficient to cause an infection to pass from one person to another [Killingley et al., 2012].

5.1.3 Individual based models

Along with the increase in available data, the availability of computing memory and speed has greatly increased, which allows large, memory-intensive simulations to be run quickly enough to be realistically implemented.

One way in which the data has improved is in the use of Radio Frequency Identification (RFID) devices. These are used to track the locations of the wearers and to generate person-to-person interaction networks via proximity between subjects [Cattuto et al., 2010; Machens et al., 2013]. Once these networks are formed, it is then possible to simulate an epidemic on them.

Along with this the ubiquity of mobile phones has lead to some large data sets [Gonzalez et al., 2008], which gives information regarding the movements that people make.

The idea of using many different datasets to model real world populations is a well established method in an attempt to characterise the spread of epidemics seen in Ferguson et al. [2005, 2006]. These papers used datasets related to household size, age structure, school and workplace size data along with commuting data to generate individually based simulations where the spatial density of people is in agreement with reality. The focus here relates to the prevention of serious flu epidemics, through investigating the effectiveness of various interventions and combinations of intervention strategies.

These have also been used to create synthetic populations, which is an established method in the use of networks to model disease spread [Eubank et al., 2004]. To construct a synthetic population, data sources including census data, diary based activity surveys and workplace data are all combined in an attempt to construct a realistic contact network for a given area, which has the same number of individuals as the true population in that area. Once this is completed the spread of epidemics on the population can be investigated and different control strategies can be tested. An outbreak of smallpox in Portland Oregon was the first hypothetical outbreak considered on a synthetic population constructed using this method [Barrett et al., 2005; Eubank et al., 2004], but since multiple different areas have been considered, for example Boston [Lewis et al., 2013] and Washington DC [Parikh et al., 2013b]. The synthetic populations generated have even been used to investigate the impact of human behaviour on the size of disaster which follows a nuclear explosion [Parikh et al., 2013a].

To simulate on such large networks, the creation of efficient algorithms has been required [Barrett et al., 2008]. Using these methods allows many simulations to be run in order to

generate useful statistics relating to the use of intervention and surveillance strategies to combat epidemics [Lewis et al., 2013].

A similar approach has been used to construct the Little Italy model [Iozzi et al., 2010]. This makes us of survey detailing how the Italian public spend their time, which was performed by the Istituto Nazionale di Statistica (ISTAT) [ISTAT homepage] of 55,773 individuals from 21,075 households to construct a synthetic population. This survey consists of 144, 10 minute intervals over a 24 hour day, in which the type of activity being performed and the type of location that the person was in is given. This was then combined with a '"minimally" complex set of rules' [Iozzi et al., 2010] to generate a synthetic population. This set of rules involved including individuals who filled in the form for a weekday, therefore limiting the number of responses to 18,085 and rather than trying to scale up to a population level simply constructing a population with this number of people in it. The survey data was combined with data about household size and composition, school class size and workplace size for towns of a similar size to Little Italy.

To generate the contact matrix people are placed in households as defined by the household data, and are then assigned to workplaces/schools throughout the day, with contacts made in each of these places to give an adjacency matrix for the population. Three different methods are then used to give the final contact matrix: Type 1 weights contacts by time together, Type 2 weights contacts by number of times they meet each other in a day and Type 3 is unweighted.

The benefit of constructing a population in this manner, is that population size is small meaning that simulation is relatively quick in comparison to large populations. However the small population is also a large limitation as this means that the investigation of realistic control strategies is not possible. Comparison to actual epidemics is possible and is done by Iozzi et al. [2010]. The Little Italy population is compared with contact structures derived from POLYMOD [Mossong et al., 2008], along with 'Big Italy', a contact structure which gives the average number of contacts between all possible ages in the household, workplace and general community contexts. This was generated to match data regarding household composition family structures, school, university and workplace structure along with homogeneous mixing for the community level interactions.

Which method produces the most accurate results to observed epidemics is then considered. However this is not answered satisfactorily, as the spread of two epidemics through Italy are considered, Varicella and ParvoVirus, and different methods give the best fit (as measured by AIC) in each case (POLYMOD for Varicella and Little Italy Type 2 for ParvoVirus).

This demonstrates clearly the fact that contact structure can have a large impact on the spread of an epidemics, and also the difficulty of choosing what the best assumptions are to make for a given disease. POLYMOD for example gives information regarding the age, sex, location, duration, frequency, and occurrence of physical contact between people, which seems to be a sensible list of variables to collect, but is a worse fit for the ParvoVirus outbreak than all of the Little Italy models, along with Big Italy, so is obviously not a perfect method for constructing contact networks. In the paper it is suggested that the high level of assortativity present in the Little Italy models allows for a good fit for ParvoVirus, but also prevents it from fitting the Varicella well. This implies that assortativity is an important measure for achieving a realistic model, but also one which is difficult to get right and must be fine tuned in order to fit to data satisfactorily.

As mentioned previously, a large portion of my PhD was dedicated to the construction of a synthetic population for England and Wales which was done in collaboration with the Network Dynamics and Simulation Science Laboratory at Virginia Tech (NDSSL). The process for this construction is described next.

5.2 Description of synthetic population construction

In order to represent England and Wales, we assign a set of activities to individuals according to some data-derived distributions, all of which will have some location in the simulation. People in the population who share the same physical location at the same time are then defined as being contacts of each other.

To do this, we first assign the number of people in the population that we are modelling to households, which they share with certain other individuals in the population. Households are given locations again according to a data-derived distribution. People will then be assigned a programme of activities throughout a 24 hour period, which will define the contact network on which the spread of epidemics can then be simulated on. The activities given to people fall into categories such as 'Work', 'Retail' and 'Other' and are based on diary data for the UK. Note that this assignment of activities is done only once, and then the same network will be used for each day of the simulation. This process produces a weighted network, as the time that people spend in the same place is tracked in the construction of the network and is then used to model the likelihood of one person infecting another more accurately than using an unweighted network.

The construction of households was undertaken entirely by the author, whilst assigning activities to the population was completed by collaborators at NDSSL.

5.2.1 Datasets

This section gives a brief explanation about what datasets are used to construct the synthetic population, along with descriptions about how they are used in the population construction.

The construction of the synthetic population for England and Wales takes place on the level of Census Area Statistics (CAS) wards, which were created for 2001 Census outputs [Office for National Statistics, b]. There are 8,850 of these wards in England and Wales and this is the level of the aggregate data which is supplied from the 2001 census. In making the population our first step is to construct households of people, which match certain statistics from census or other data sources, such as the age distribution of people in wards, or the number of households containing 2 adults and 2 children.

The household distribution data set (HHD) gives us details about how many households of different compositions there are by CAS ward. This is a table that was commissioned from the Office for National Statistics [Office for National Statistics, a] and is discussed in House and Keeling [2009]. This dataset gives the number of households in each ward, which have 1-8+ occupants and 0-8+ dependent children in these households. We follow the information given in this dataset to choose the number of houses with the possible household types in our synthetic version of the population. We can also calculate the

number of adults and children in the population from this. When we do this we get 11,665,495 children in England and 39,407,463 adults.

We note that as we are only given information about houses with 8+ inhabitants and 8+ children, the information here is incomplete in terms of these large households. More information related to these large households can be found in the 1% household sample of the population from the census which we discuss this dataset next.

The 1% household sample of anonymous records (HSAR) [Office for National Statistics, c] gives us data direct from the census on 1% of the households in the UK population. This is an anonymized data set which gives complete census records from households in the UK. This is not an open data source and to access it, registration along with confirmation from UK Data Service [UKDS] is required. A large limitation with this dataset is location data is not included, which decreases the amount of information in this data in a significant way.

For each household, the person who filled out the census form for the household is called the household representative person (HRP) and the HHD dataset includes complete households worth of data, in that records of all the people who share a household are included in the dataset. We note that HRPs can be any age from 16 up. There are 60 separate variables in this dataset, but we are not interested in all of them. As previously mentioned, to construct the contact network for the UK we must assign people activities throughout the day. Therefore we are mainly interested in variables which give us the best information to choose the activities that people perform during the day. For example, the number of hours someone works per week tells us whether someone is unemployed or in work part or full time, and therefore when they are likely to be in work. The variables that are used are described in Table 5.1.

Variable	**Possible Values**
Household number	1–total households in dataset.
Person number within the household	1–number of people in household.
Relationship status from person to HRP	HRP; Spouse; partner; child; step-child; sibling; parent; step-parent; grandchild; grandparent; other related; unrelated; unknown.
What type of accommodation the household is	Detached; semi-detached; terraced house or bungalow; purpose built flats; converted flat or shared house; flat or maisonette in converted building; mobile or temporary structure.
Age of person	0, 2, 4 ...; 80+.
Description of family	Lone parent male; lone parent female; married couple no children; married couple children all belong to both parents; married couple children do not all belong to both members; cohabiting couple no children; cohabiting couple children all belong to both parents; cohabiting couple children do not all belong to both parents; ungrouped individual (not in a family).

Is this person the HRP?	Yes; no.
Hours worked per week	No work; number of hours worked in integers.
What gender the person is	Male; female.
Tenure of accommodation	Owns with or without a mortgage; rents from council or housing authority; private rental.
Transport to work method	Not in work; work at home; Underground, metro, light rail, tram or tube; train; bus, minibus, coach; Motor cycle, scooter or moped; Driving a car or van; Passenger in a car or van; Taxi or minicab; Bicycle; on foot; other; Car or van pool.
Number of people who work in same place as the person	No job record; 1–9 people; 10–24 people; 25–499 people; 500 or more people.
Age of the HRP	0, 2, 4 ...; 80+.

Table 5.1: Variables, and their possible values, that we use from the 1% H-SAR sample.

From this dataset, we can also extract the size of each house and the number of adults and children in the house. This will be used later on in the process of population construction, as we are interested in the difference in the types of people who live in houses of the different sizes, e.g. a young adult who lives with several other young adults is likely to behave differently to a middle aged adult who lives with another middle aged adult and 3 children. Also as mentioned above, we consider the construction of houses of the 8+ type from the HHD dataset. When we consider the houses with 8+ children, we find that roughly 20% of these houses have 1 adult in them, 70% have 2 adults and 10% have 3 adults. When houses of this type are populated, we will pick the number of adults randomly according to these percentages.

The distribution of numbers of people by age differs from ward to ward. We use a dataset which gives the number of people of each age in every CAS ward, and refer to this as the for population by age and location dataset (PBAL) . An example of this can be seen in figure 5.1d for an anonymous ward. This dataset simply gives us the number of people of each age group in the given CAS ward. We note that the elder statistics are grouped together in the following groups; 75 to 79, 80 to 84, 85 to 89, 90 to 94, 95 to 99, 100 years and over. As we create the population, we will select the ages of people in order to approximate the distribution of ages given by this dataset, as the age of a person can have a large impact on their likelihood of being part of the epidemic.

This dataset is aggregate data derived from the UK census in 2001. This is downloaded from the Office for National Statistics using the InFuse tool [InFuse]. To download this data, the 2001 census data option was selected from the main InFuse page. The filter 'Age' can then be selected followed by the age ranges desired along with the geography level (e.g. selecting the populations by CAS ward for this study). At the time of population construction, the data from 2011 was unavailable, though this data has now been made available.

We can also use this dataset to calculate the number of adults and children in the popu-

lation. If we take an adult to be 16 or over and dependent children to be 15 or younger then we get 10,482,326 children and 41,514,706 adults. Note that this is fewer children and a greater number of adults than the HHD dataset, which suggests that by dependent child in the HHD dataset, they do not mean only people under 16, but people under 18. If instead we count the number of people under 18 as children in the PBAL data set then we get 11,788,034 children and 40,208,998 adults, which is closer to the numbers from HHD. We will therefore label anyone under 18 as a child.

To assign people to households we use another dataset from the census, which gives the number of occupied households at each post code location in the census. This is obtained from the Office for National Statistics and can downloaded from the NOMIS website [NOMIS].

Postcodes are also assigned a geographical location, and a dataset containing these is available from the Ordnance Survey (OS) and is called Code-Point Open. This can be downloaded from the OS website [Code-Point Open]. As described on the support website for this data, the location of each post code is obtained by taking the average of the coordinates of all the individual addresses in a postcode, and then reporting the location of the nearest building to this point. Postcodes can be linked with wards using another dataset from the Office for National Statistics geoportal site [Geoportal] called the National Statistics Postcode Directory. As we have the number of people living in each ward from the PBAL dataset we can then place our households according to the NOMIS dataset at the location given by the Code-Point Open dataset.

Locations and sizes of workplaces are given in the previously detailed Blue Sheep data. Locations and numbers of students at schools are taken from the National Pupil Database [NPD]. This is not an open data source and to be used an application for access must first be made. This dataset provides the home postcode and school post code for over 7 million school students in the UK, and from this the size and location of schools can be derived by summing the the numbers of pupils attending school in given postcodes.

In general, limitations with these datasets is that there is data that is missing that would be helpful to have. For example, there is no attempt made to include data giving spatial information relating to employment, as this is something which can be highly correlated and also impacts the contact structures that people will develop. Along with this, data on income of households and the location of different households in the HSAR dat set would be informative for choosing activities for individuals.

These are the datasets which are used in the construction of the synthetic population. In the next section, the methods used to construct this population are described.

5.2.2 Population construction

The process of constructing the population is essentially the following:

1. Select a HRP from a data derived distribution.
2. Select a relationship to another adult in the house by choosing from a data derived distribution.
3. Choose the age of this adult.
4. Repeat for remainder of adults.

5. Repeat equivalently for each child in the household.

6. Go to 1.

As is to be expected, there are few houses that are very large in the HSAR dataset e.g. there are only 41 examples of houses which have greater that 7 children in them. Therefore we combine the data for houses that we class as being large, where the cut off for a large house is chosen to be any house with more than 5 children in them. The process we go through to construct the households differs between the large and non-large households. We will describe the process for the non-large households first.

The process of deriving the distributions to choose from is mainly done through looking at the HSAR dataset. The task is first split into household types. We begin by identifying which households have a given type, e.g. find all households which have 3 people 2 of which are children. We then go through this set of data person by person. If we encounter a HRP, then we simply record their age. We will later create a cumulative density function (cdf) for the ages of HRP's in each household, an example of which can be seen in figure 5.1a, where the data has been perturbed while maintaining the qualitative properties which differentiate the different scenarios.

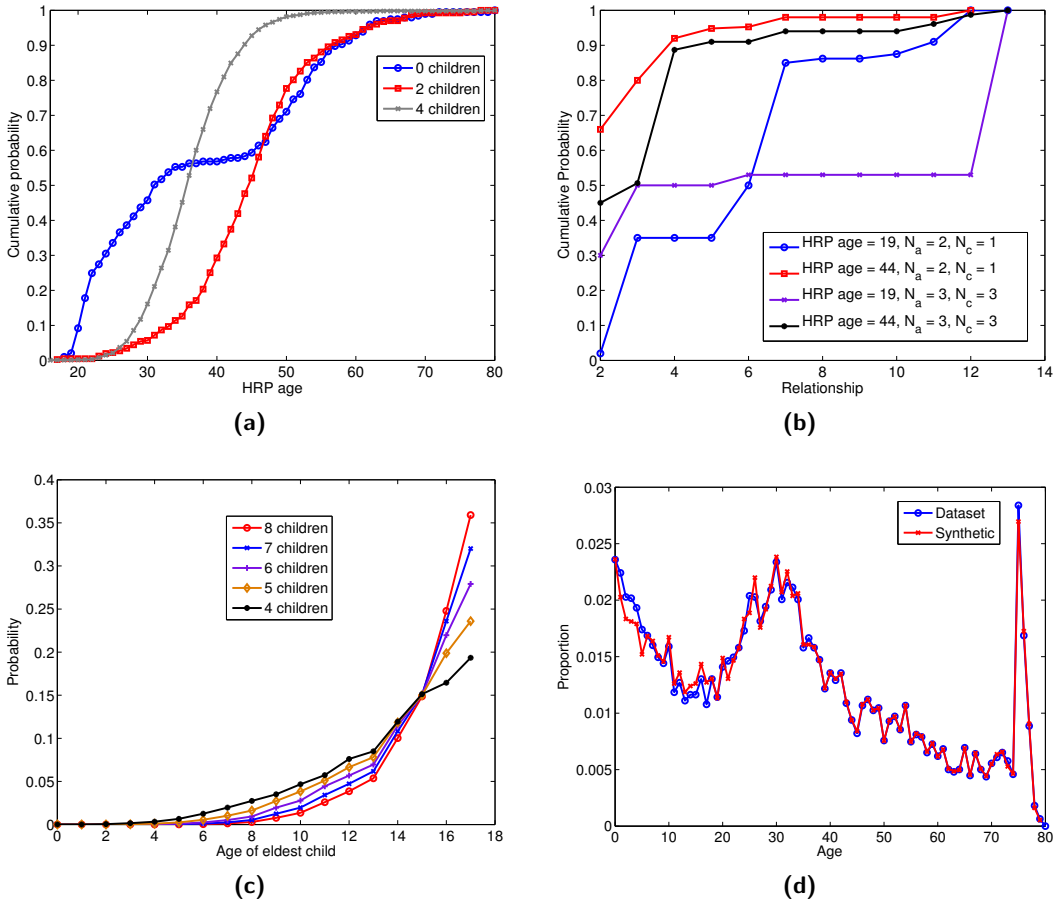


Figure 5.1: Visualisation of data sources used in construction of the synthetic population.

If we consider the shape of these cdf's it gives us some intuitive and interesting information. The 3 different curves are for the same household size, 6, but different number of children. Looking at the curve for 0 children, we see that there are about half of these households have HRP's who for young adults, from age 16 to about 30. There is then a period of about 20 years when there are very few households of this type, presumably as most people at the age of 30 will move out of large house shares with other adults and into more family orientated houses. For 4 children, the vast majority of these houses have a HRP between 30 and 50 whilst for 2 children, there are more houses with HRP's between the ages of 40 and 60.

For non-HRP's, we note several of their attributes. We note the relationship type to the HRP along with the difference in age from the HRP. We then construct cdf's for relationship type which is indexed by HRP age and house type along with cdf's for age gap from HRP to other person by household type, HRP age and relationship type. This is done similarly for adults and children. Along with this, we also construct cdf's for relationship type for adults if we do not include partners or spouses. This is because these make up a large proportion of the relationships in the population, and there will be only one of them per household. Therefore when we choose the population, we will choose from this if we have already selected a partner or spouse from the other cdf.

Examples of these cdf's can be seen in Figures 5.1b, where again the figure shows data which has been perturbed.

There are 99,991 men who are HRP's and 46,993 women who are HRP's. Therefore we pick more HRP men in accordance with this ratio. When we are constructing the synthetic population we choose from these cdf's to populate households in the wards. Initially we choose the HRP from the cdf for HRP age by household type, and then we choose the required number of adults needed for the household, first choosing their relationship type by HRP age and household type and then choosing the age gap between them and the HRP by relationship, HRP age and household type. Once this has been completed for the adults in the household we do the same for the children.

Note, that when we select a person of a given age to be a member of the synthetic population, a HRP say, we choose an individual record from the HSAR, who matches the characteristics that we are looking for, in terms of age and sex. We therefore select each record multiple times. This is due to the fact that we do not have any other detailed information about the people who make up the true population.

For large households, adults are chosen in the same way, but children are selected differently. This is because we want to construct realistic households, and if we simply choose several children from the cdf for age distribution, we may get an over representation in the number of households with multiple children of the same age living in them. This may also have an impact on the dynamics of the epidemics as, for example, it is possible that children who are under 5, will have different numbers of contacts and contact patterns than older children.

The process for selecting the children is as follows:

1. Choose the eldest child in the household from a data derived distribution.
2. Select all other children in the household from another data derived distribution

To choose the age of the eldest child we use order statistics from which we can calculate the distribution of the k-th smallest sample from a given distribution. We begin by considering

the distribution of all the ages of children in the 1% sample, who live in large households. If we know all the ages of children associated with large households, we can construct the pdf (f) and cdf (F) of this distribution. We then wish to consider the distribution of the kth smallest value, which we will denote as $X_{(k)}$, from n values selected from the pdf. This is obviously the same as working out the probability that there are $(n-k)$ samples larger. We use order statistics to derive the probability that $X_{(k)} = x$, which is given by the following formula:

$$P(X_{(k)} = x) = \sum_{j=0}^{n-k} \binom{n}{j} \left((1-F(x))^j F(x)^{n-j} - (1-F(x)-f(x))^j (F(x)-f(x))^{n-j} \right) . \quad (5.1)$$

We wish to use this to constrain the selection of the eldest child, therefore we just require the distribution of $X_{(n)}$, which is given by the following:

$$P(X_{(n)} = x) = F(x)^n - (F(x) - f(x))^n \quad (5.2)$$

This can be seen for households with different numbers of children in figure 5.1c. This shows that the more children in the household, the older the eldest child is likely to be. This obviously is what we would expect as for example it is unlikely that there will be 6 children in a house, the eldest of whom is 4 years old.

Once we have chosen the eldest child, we then choose the rest of the children by selecting their ages from the cdf detailing the distribution of age differences amongst children in the large households. The reason that we have chosen to do this in this circumstance is to try and limit the number of children of the same age that we select for the population, as we do not observe that many twins/triplets etc. in the true population.

We note that the fact that we use the cdf's to select the population, and then deplete from the population by age and location dataset means that we can select a person with a certain age, when according to this dataset, we have already chosen enough people of that age for that location. To try and minimise this, if we find that we have selected someone of an age which has already been fully depleted, we will look either side of this age to try and find someone of a similar age, which has not yet been fully depleted. We do not look further than 10 years either side from this initially selected age for adults and 7 either side for children, as we wish to use the datasets fully, rather than randomising the selection of the synthetic population too much. This means that the resulting age distributions may not match the PBAL distributions exactly.

Once this is done we can then compare how alike our two populations are, i.e. the synthetic one and the one from the data, at least in terms of the population by age. An example of this can be seen in Figure 5.1d. We see that though there is not an exact match between the two, the synthetic populations are a good match for the population from the data by this measure.

As mentioned previously we try to not choose a wildly unrealistic number of twins in the population. Website Twins UK gives details about how many twins are born each year, and there are roughly 10,000 sets of twins per year. This means that we would expect roughly 180,000 sets of twins aged between 0 and 17. When we look in our population we have 433,095 sets of twins, which is slightly less than 2.5 times the expected amount. We could try to constrain the number of twins, but currently have not done this. The

question of whether this is an issue that we need to worry about is open. The main goal of constructing this population is to construct a contact network from it, which we can then run epidemics on. The actual ages of the population are therefore not the most important thing to get right, as we expect that having two 14 year old twins in a household, is no different in terms of the activities assigned and the spread of disease, as having a 12 year old and a 14 year old in the household.

There are obviously many different ways that this process could be done differently and other data sources could be used to complement this process. For example, we could use the proportion of people in full time employment in each ward or proportion of people who own their own homes but this has not been done here.

Once all of the houses are constructed for each ward, they are randomly assigned postcodes from that ward, which have an accompanying latitude and longitude, allowing them to be located on a map of England and Wales. The process for constructing a synthetic population from this point follows that described by Eubank et al. [2004] and Eubank et al. [2006] and was carried out by collaborators from the Network Dynamics and Simulation Science Laboratory at Virginia Tech. This is described briefly below.

Every person is assigned a set of activities and times for these activities that occupy their day in the synthetic population. The types of activities performed are based on those included in the United Kingdom Time Use Survey [UK Data Service]. This is a diary based survey which was filled out by households in the year 2000 and records data such as age of people in the household, income, household size amongst other information. Along with this people indicate what activity they were performing for each of the 10 minute intervals in a 24 hour period, along with who they were with and where they were. For the synthetic population, people were assigned activities from this via the use of a classification and regression tree (CART) which chooses activities for individuals based on the demographic information which was selected for them during the construction of the synthetic household population. People are assigned to the activities 'home', 'work', 'school', 'shopping' or "other" during the day.

The previously discussed Blue Sheep data which contains work location and workplace size is then used to place people who are partaking in the activity work, into a workplace. This is done via a gravity model where the distance from a persons home and the size of the workplace define the probability that a person will be placed in each workplace. For people placed in large workplaces, people are assigned to sublocations of this workplace so that the number of contacts that these people have for workplace activities is not excessively large. The size of these sublocations is arbitrarily set to be 35.

For the shopping activity, the Blue Sheep data is again used to provide the probability of someone visiting a certain location for shopping. This is done by selecting all workplaces whose description is related to retail. We investigate the use of this method in the maps shown in figures 5.2 and 5.3. In the first figure, we can see a map of central London, with features such as the Thames, Hyde Park on the left of the map and Regent's Park directly above, visible due to the lack of workplaces in these locations. The density of shopping locations, along with workplaces in general is concentrated in several areas, most noticeably just to the East of the Hyde Park. Figure 5.3 zooms into this region to show in more detail the concentration of retail locations in this region. This is reassuring as this area contains Oxford Street, Picadilly Circus and Regent Street, amongst other well known retail and leisure areas in London, implying that the method used for selecting shopping locations is picking out areas which line up with our expectation of reality. To

assign people to each location for shopping a gravity model is again used.

We note that there are postcodes which appear to be inside the Thames. Comparing these locations with maps of London, it appears that these line up with piers which will overhang the river, which explains these postcode locations. Additionally, inside the park there are various buildings including cafes which explains the presence of work and retail postcodes within the park grounds.

Once this is done, contacts are made during activities, and are given a specific duration according to the length of time that the people involved in a particular contact are in the same place for. This defines the contact network on which an epidemic will spread.

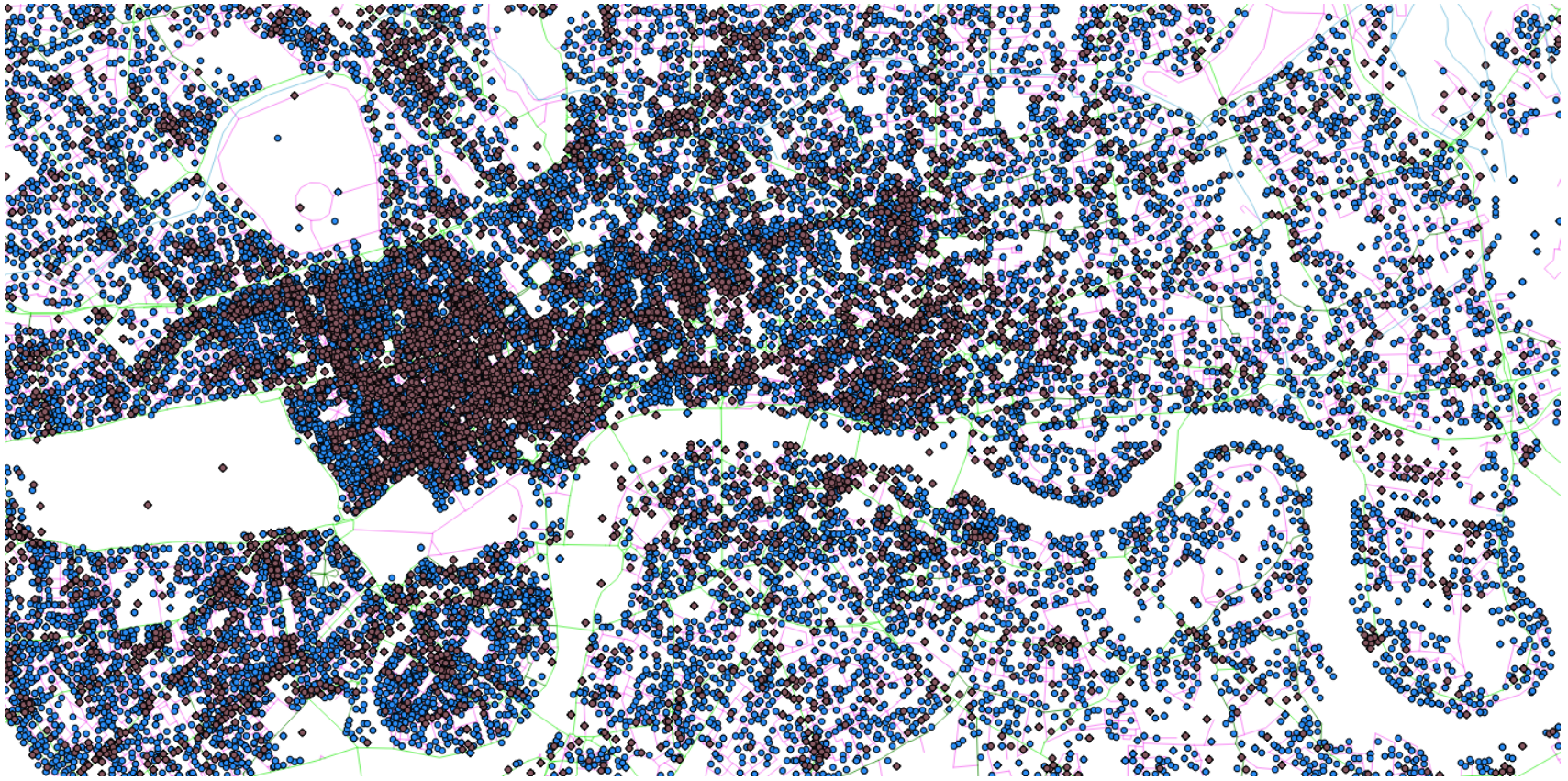


Figure 5.2: Map showing the centre of London, with Blue Sheep workplaces seen as blue dots, and the Blue Sheep derived retail shops in brown.

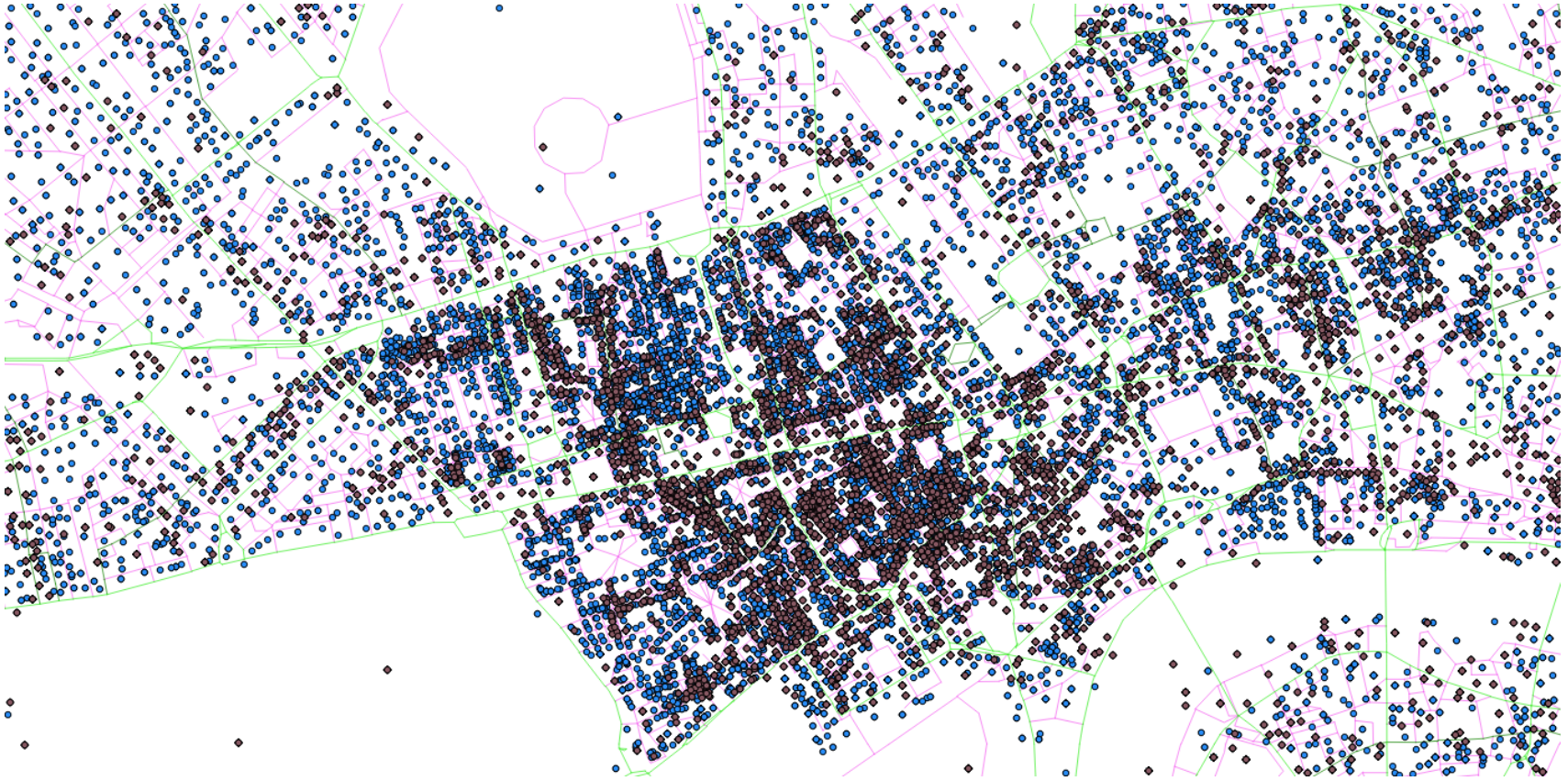


Figure 5.3: Map showing the area of London around Oxford Street, with Blue Sheep workplaces seen as blue dots, and the Blue Sheep derived retail shops in brown.

5.3 Analysis of Population

Once population construction has been completed, we can investigate the properties that the network has. As this is a very large network, with ≈50 million nodes this is a non-trivial task, and requires the use of specially constructed data structures and algorithms [CINET: Cyber-Infrastructure for Network Science].

We are also interested in comparing the contact structure of this population with other contact structures which exist for the UK. To do this we will compare to POLYMOD [Mossong et al., 2008] and the UK contact survey [Danon et al., 2012]. These are both surveys of the UK population. There is a potential for bias here, as the surveys of the populations have not been completed proportionally by each age group, and therefore some age groups are under represented (such as the young) and others are over represented (such as the old), whilst the synthetic population is representative of age. On top of this, the surveys cover the UK, whilst the synthetic population is only for England and Wales, which may be another way that bias could be introduced. This hasn't been adjusted for in any of the comparisons made here or in the following chapter which discusses comparisons of these datasets in several different modeling frameworks.

One of the main items of interest when attempting to construct realistic networks is the age-age mixing in the network, as it is expected that the contact patterns and ages of contacts made will differ according to many things, the easiest of which to account for being age. Figure 5.4 shows this information for the synthetic population through the who-acquires infection from whom matrix (WAIFW), and offers a comparison to the same data taken from the POLYMOD survey.

Figures 5.4a and 5.4b shows the who acquires infection from whom matrix (WAIFW) which is defined by the synthetic population. In figure 5.4a we see the numbers of contacts made between people of different ages in the population. We can see that there is a high number of contacts amongst children, given by households and schools, and between adults and children which are given by the household structure. Along with this, there are also a large number of contacts amongst individuals between the ages of 30 and 45, which are given by the mixing of people in workplaces.

Figure 5.4b gives a similar WAIFW, but here the contacts are weighted by duration. The high intensity amongst adults between 30 and 45 is now greatly diminished, whilst those contacts that are defined by households or schools are maintained. This implies that transmission between any two individuals is more likely to happen if they are contacts in homes or school, as the duration of contacts here is greater.

In figures 5.4c and 5.4d we can see the WAIFW for the POLYMOD contact matrices. Here the diagonal which gives contacts between children of school age is by far the most prominent feature of the matrix. The difference in the two matrices can be explained due to the way that activities are assigned in the synthetic population, as the assignment of people to classrooms is less strict than in reality and allows for more mixing between children of different ages.

The POLYMOD heatmaps in figures 5.4c and 5.4d show a strong diagonal component, especially for individuals aged between 0 and 18, indicating that individuals of these ages have comparatively high amounts of mixing with people of the same age when compared with others of different ages. This represents mixing within schools, and for older people, within workplaces. There are also off-diagonals, which indicate mixing of individuals of

with other individuals who are aged approximately 30 years older. This indicates children mixing with their parents in households.

Figure 5.5 displays the empirical degree distributions for the synthetic population, POLYMOD and UK social contact survey. These are also split by activity type and the survival function, which is 1 - the cumulative distribution function of the degree distributions, is also displayed. In figure 5.5a, we see the full degree distributions for values between 0 and 100. The synthetic population displays a secondary peak in degree probability at around 35, which is caused by the workplaces being split into sublocations. Considering the plots of the survival function seen in figures 5.5b, 5.5f and 5.5h, we see that the mean and variance of the degree distribution when all contacts are combined, along with the work/school contacts and other contacts is much larger for the synthetic population. For home contacts, the mean and variance are smaller for the synthetic population (figure 5.5d). In general there is much more heterogeneity in the degree of individuals in the synthetic population than for POLYMOD or the UK contact survey. However, since much of this heterogeneity in numbers of contacts in the synthetic population are from "other" activities, which are likely to be of shorter duration than contacts in home or school/work, how much of an impact these have on the epidemic is unknown.

In figure 5.6 we can see the clustering in the synthetic population as compared to that from the contact survey performed in the UK [Danon et al., 2012]. Here the clustering coefficient is weighted to take account of contact time between individuals. To calculate this for an individual i, who has contact with individuals j and k, for times t_j^i and t_k^i, we define $C_{j,k} = 1$ if j and k contact each other and 0 otherwise. The weighted clustering coefficient for individual i, ϕ_i is then given by,

$$\phi_i = \frac{\sum_{j,k} t_j^i t_k^i C_{j,k}}{\sum_{j \neq k} t_j^i t_k^i} \ . \tag{5.3}$$

This is done for all individuals and then averaged by degree. Doing this for the synthetic population and and UK contact survey produces the figures seen in figure 5.6.

Figures 5.6b and 5.6d show the clustering evident for individuals with low degree (up to 250 for the synthetic population and up to 200 for the contact survey) and figures 5.6c and show this for the full range of degrees. For the contact survey we see that the level of clustering, as reported by individuals is roughly equal across all degrees. This seems counter intuitive, as we would expect the level of clustering to decrease as degree increases. However constant levels of clustering have been shown to exist in other networks, such as the Internet, once degree-correlation biases are eliminated from the definition of Soffer and Vázquez [2005]. This is because if a highly connected node has neighbours who have low degree, then it is impossible for the standard definition of clustering to give a large value, as without increasing the number of neighbours that these low degrees have, they cannot all be connected to each other. As in the contact survey, when the degree gets high, this is reported usually as a group and the question of whether these people met each other is independent of contacts made in other circumstances, this degree-correlation bias is also sidestepped. Therefore we can expect clustering to stay constant through all degrees.

When considering the synthetic population however, the clustering does not stay constant at all degrees. As seen in figure 5.6a, the level of clustering increases from the lowest degree until ≈ 35, when it then begins to decrease until the degree reaches 100 at which

point the level of clustering begins to increase again and figure 5.6c shows that it continues to increase until it reaches a maximum level of ≈ 0.75 and maintains this high level until it begins to decrease at an approximately linear rate from degree 2000. The increases in the clustering can be explained due to the process of assigning contacts to people when they are in the same location. If people are in the same location at the same time then they are considered contacts of each other [Eubank et al., 2006]. At locations with large numbers of people there at the same time, such as in the case of busy shopping areas, it is possible that a large number of people can be linked to all others, meaning that we can get large degrees and levels of clustering coinciding. It is possible that if we included a constraint that the contact must be of a certain duration that the level of clustering may decrease, as the contacts at large shopping areas will be mainly fleeting ones, but this has not been included in the analysis of the population here.

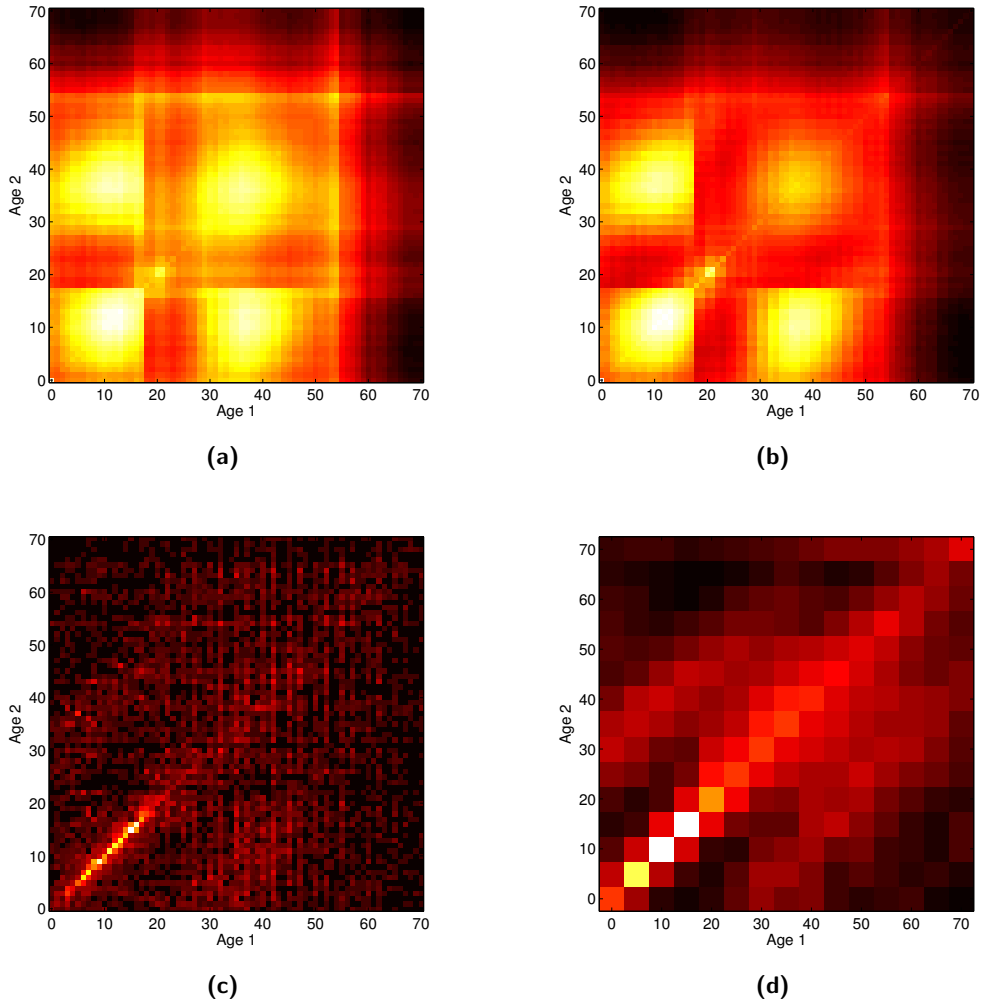


Figure 5.4: Different measures of 'who acquires infection from whom' (WAIFW) by age. (A) Count of all contacts in synthetic network; (B) contacts in synthetic network weighted by duration; (C) count of British contacts in the raw POLYMOD data; (D) processed mixing matrix from POLYMOD data provided by Mossong et al. [2008]. For (C) and (D), Age 1 is respondent and Age 2 is contact. Heatmap intensity is proportional to $\sqrt{\text{value}}$.

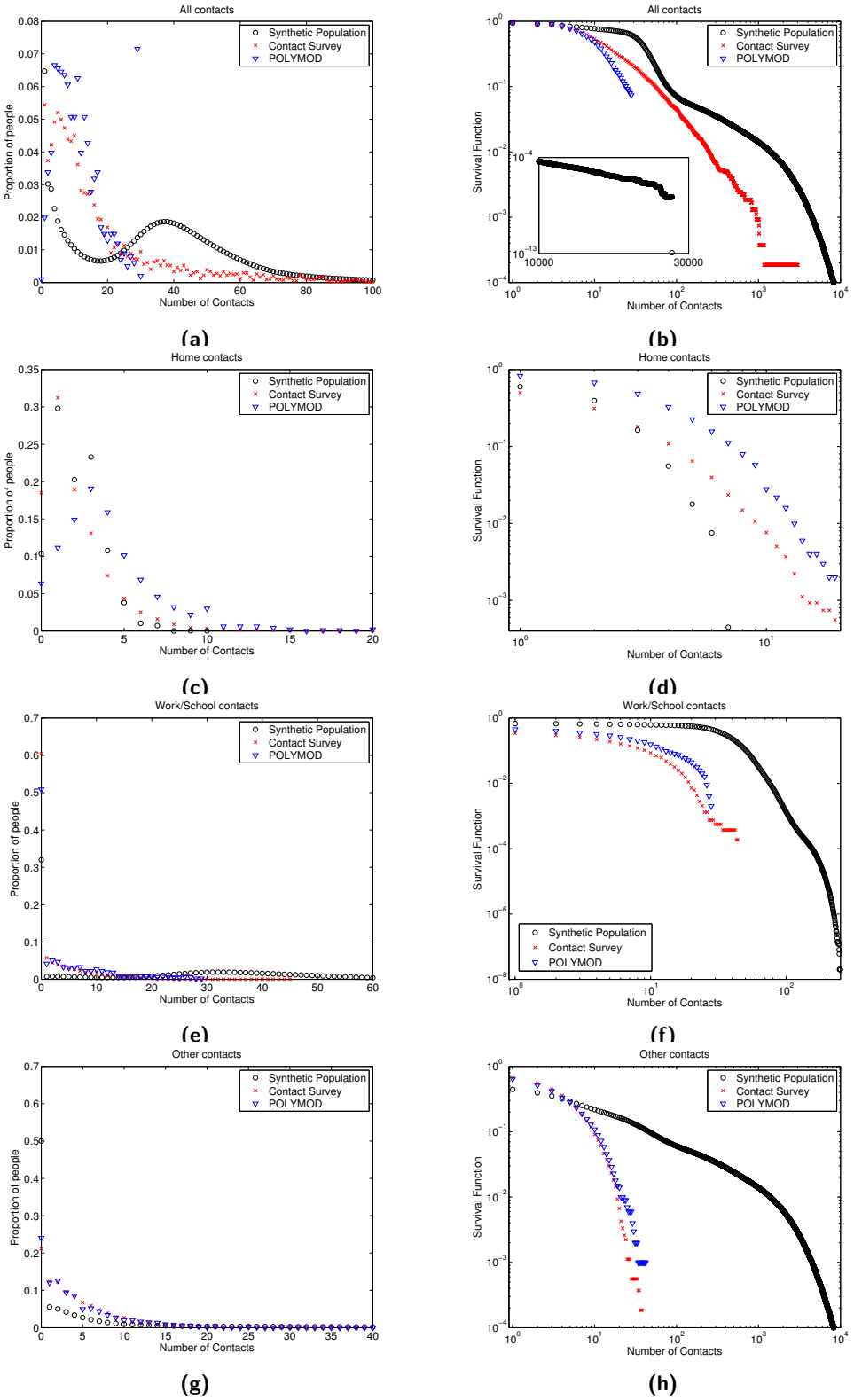


Figure 5.5: Heterogeneity in numbers of contacts by activity for synthetic population, Social Contact Survey [Danon et al., 2012, 2013] and POLYMOD [Mossong et al., 2008].

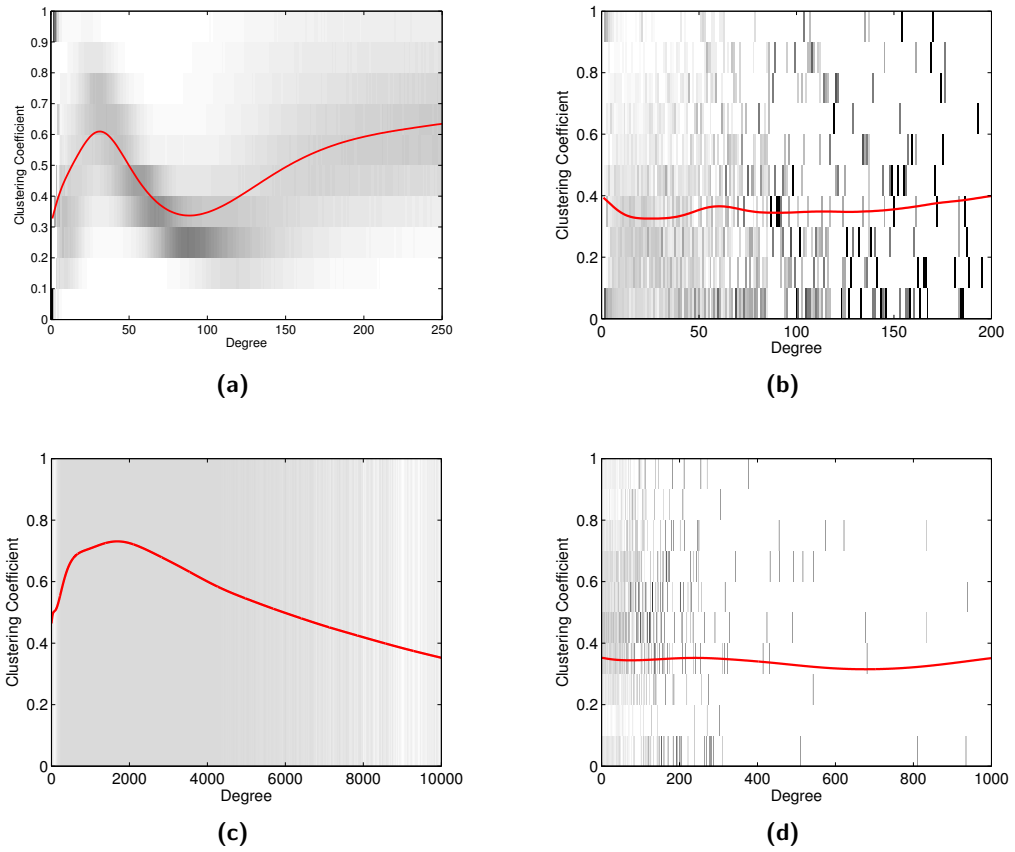


Figure 5.6: Clustering coefficient versus node degree for (A,C) synthetic population versus (B,D) the Social Contact Survey [Danon et al., 2012, 2013]. Histograms at each number of contacts (greyscale shading) are shown together with smoothing cubic spline plots (red lines).

5.4 Summary

In this chapter, the process of constructing a data-derived synthetic population for England and Wales was described. This involved the use of several data sets, along with statistical and mechanistic techniques used in the construction of synthetic populations in the USA [Eubank et al., 2006].

Once constructed, this population was analysed and compared with two survey based models of the population in the UK [Mossong et al., 2008; Danon et al., 2012]. It was shown that the level of heterogeneity in terms of the numbers of contacts that individuals have was much higher in the synthetic population and mixing was less concentrated to within an individuals own age group than that seen in Mossong et al. [2008]. In terms of clustering of contacts, we saw that the level of clustering by degree was larger than that derived from Danon et al. [2012]. The impact of these observations on the spread of an epidemic is discussed in the following chapter.

Chapter 6

Comparison of disease spread on different data-derived populations

6.1 Introduction

The use of synthetic model populations to investigate the spread of an epidemic in a population is well established, and has been the goal in the creation of the synthetic population for England and Wales. Despite large increases in computation power, simulations for populations of the size of our synthetic population ($N \approx 5 \times 10^7$ with more than 10^9 edges) pose a significant challenge, although as previously mentioned, efficient techniques have been created which allow for simulation on these large networks [Barrett et al., 2008]. Along with this, a web-based tool has been created called Interface to Synthetic Information Systems (ISIS), which allows for the parameters of an $SEIR$ or SIR model to be tuned and interventions to be studied [ISIS VBI]. This is used in order to investigate which interventions are most likely to decrease the spread of the epidemic effectively in the network.

It is worth noting that along with the final size of an epidemic, whether with or without control interventions, the time to the peak of the epidemic, along with the size of the peak are important aspects of disease control. This is due to the fact that the size of the peak is related to the burden that will be put on the healthcare systems of the country in question, for example there may not be enough hospital beds for influenza patients if the peak in flu infections is very high. The time to the peak is also important, as the longer that this is, the more time there is to impose potential control strategies, along with preparing healthcare systems and facilities.

To investigate the utility of constructing such large populations which, even with efficient algorithms, are difficult to use, we compare the output of this model to simpler models which are derived from data. We will consider the impact of workplace closure and vaccination, when these are possible given the model used, on the final size of the predicted epidemic.

To begin with we consider the spread of epidemics on the full synthetic population for several disease parameters. Vaccination and closure of workplaces is considered for the

synthetic population and these are compared with each other to see which would is predicted to have the greatest impact on the spread of the epidemic.

We will look at the pairwise approximation ODEs from (3.28) to a network defined by the degree distribution of the synthetic population along with the degree distributions of the POLYMOD and UK contact surveys [Mossong et al., 2008; Danon et al., 2012]. We consider the impact of workplace closure here along with no controls.

In the following section we use the WAIFW matrix produced by the synthetic population, along with the one for POLYMOD and uniform mixing. We consider only vaccination here, as closing workplaces will change these matrices in an unknown way.

Finally we produce a meta-population model which is built using similar principles to the synthetic population, to produce population models in agreement with the synthetic population, POLYMOD and UK contact survey. Here we consider the impact of both workplace closure and vaccination on the final size of the epidemic.

This will allow us to investigate the impact on the spread of a disease that having these different datasets give us, along with how the different levels of heterogeneity in each of the populations and each modelling technique interact with each other.

6.2 Synthetic population

We begin by considering simulations which were done using the full synthetic population. Due to the size and complexity of this network, simulation on it takes a large amount of time and therefore, while this limits the number of realisations, we gain increased complexity and realism in the contact structure.

6.2.1 Synthetic population simulations

Figure 6.1 shows the result of simulations on the synthetic population. Here the model is for a disease with transmission rate $\tau = 1.25 \times 10^{-4}$ days^{-1}, and a final size of 66% of the model population. We compare the impact of closing workplaces to using no control methods at all. In 6.1, as we would expect, we can see that the closing of workplaces has a large impact on the amount of infections in the population (the dark grey lines are individual simulations for the work closure, and the light grey ones are for when no control methods are used). We can also see that the time to the peak is longer in the case where there are workplace closures, which is due to the fact that the decrease in the number of contacts that people have will decrease the speed at which the epidemic is able to spread, as effectively transmission rate is decreased.

The time until the infection is eradicated from the population is also longer. The reason for this difference in time is that in the case where there are no controls, the disease infects more people the population, and therefore the amount of susceptible contacts that people have is diminished more quickly. This means that once the peak amount of infectivity in the population has been reached, the drop off is much quicker in terms of the number of infecteds, as there are less people still susceptible to become infected.

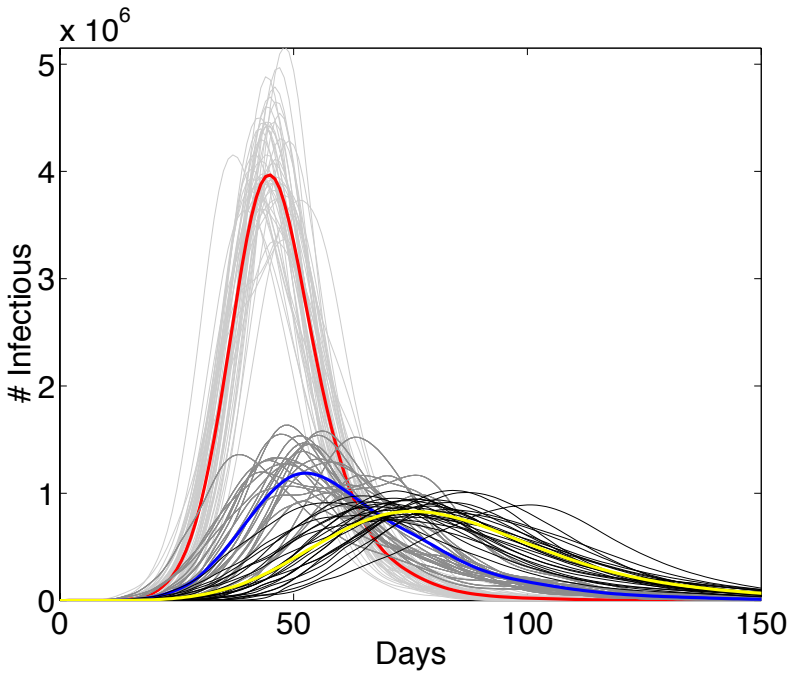


Figure 6.1: Simulations from the ISIS tool on the synthetic population for the UK. The red line is the average of the light grey simulations which have no control, where the transmissibility $\tau = 1.25 \times 10^{-4}$ days^{-1}. The blue line is the average of the dark grey simulations, which have workplace contacts removed and the same level of transmissibility as the light grey simulations. The yellow line is the average of the simulations in black. These have no control placed upon them, but the transmissibility is $\tau = 6 \times 10^{-5}$ days^{-1}. The cumulative number of infections for the black and dark grey simulations is approximately the same.

Figure 6.1 also has simulations from a much less transmissible infection ($\tau = 6 \times 10^{-5}$ days^{-1}) with no controls placed upon it. These are shown in black with the yellow line being the average of these simulations. We note that the total number of infections is approximately equal for these simulations as for the simulations where the workplaces have been closed in the more infectious simulations (≈17,850,000 to ≈17,900,000 respectively). This implies that the decrease in the number of contacts combined with the increase in the contact rate in the dark grey simulations is sufficient to give the same level of transmissibility throughout the full network when compared to the simulations in black. However the peak is lower and the time to reach the peak is greater in the case where the infectivity is lower and no controls are used. This is in line with the effect of clustering on the spread of the disease described in §2.2.3, as the network where there is no control will have a higher level of clustering in it, as the workplaces have a high level of clustering in them (figure 5.6) and theoretical results dictate that the time to the peak will be longer and the peak will be smaller if clustering is increased [House and Keeling, 2011a].

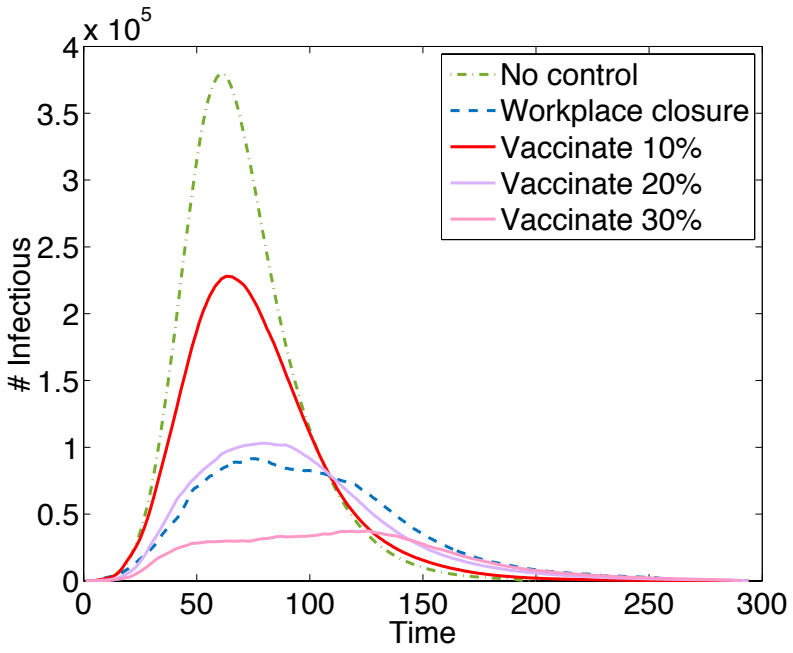


Figure 6.2: Average infection trails of simulations from the ISIS tool on the synthetic population for the UK. The transmissibility term here is $\tau = 6 \times 10^{-5}$ days^{-1}. We see the comparison between closing workplaces and vaccinating 10%, 20% and 30% of the population and introducing no controls on the population.

In figure 6.2 we can again see the average spread of the epidemic in the population with no controls alongside the spread when seen when closing workplaces and randomly vaccinating 10, 20 and 30% of the population. The transmissibility of the disease here is 6×10^{-5} days^{-1} in all five scenarios. When no controls are introduced, approximately 20,360,000 are infected with the disease. For the two cases where the workplaces are closed and 20% of the population is vaccinated, the final size is approximately 9,800,000. We can see that along with the final size being comparable in both control strategies, the trajectory of the number of infecteds is also comparable. The decrease of the final size of the epidemic to vaccinating 20% is marginally over 50%.

As the vaccination program considered here is enacted before the epidemic spreads through the population, it is clear that the number of susceptibles which remain in the population will be smaller in the case where we vaccinate than when workplaces are closed. The same number were removed by the epidemic, as the final size is the same, but there is an additional 20% of removed individuals who were vaccinated at the beginning of the epidemic. Vaccination would therefore make it more difficult for the same disease to re-invade the population and cause another epidemic, and if this did occur, the number of infected individuals would be lower.

Vaccinating 10% and 30% of the population decreases the final size of the epidemic to approximately 15, 150,000 and 4,900,000, which signify a decrease of approximately 25% and 75% in these two cases. Combining this with the 50% decrease by vaccinating 20%, we see that the linear relationship between vaccination percentage and decrease in size of the epidemic that is seen in Eames et al. [2009] is maintained for this model population.

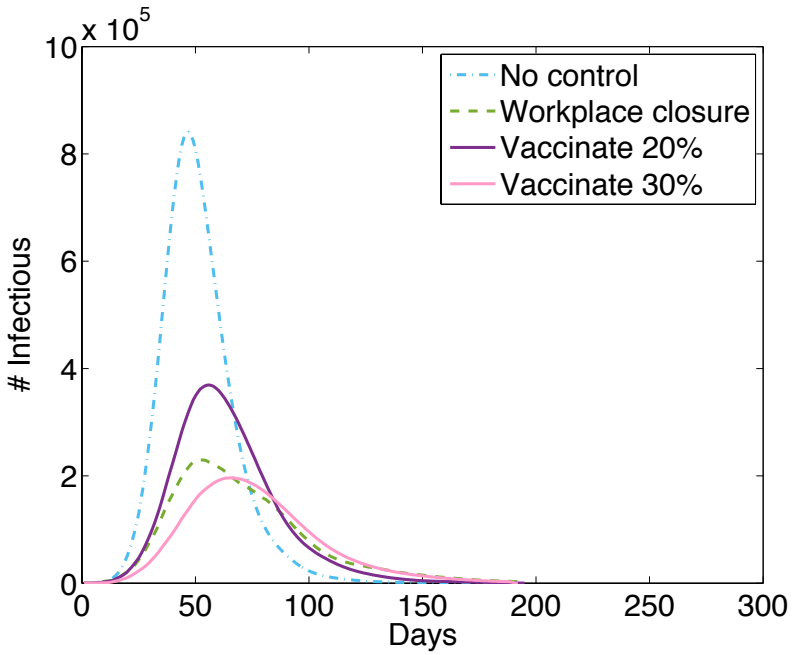


Figure 6.3: Average infection trails of simulations from the ISIS tool on the synthetic population for the UK. The transmissibility term is $\tau = 9 \times 10^{-5}$ days^{-1}. We see the comparison between closing workplaces and vaccinating 20% or 30% of the population and introducing no controls on the population.

Figure 6.3 displays the same epidemic spreads as given in figure 6.2, but for a higher transmission rate of 9×10^{-5} days^{-1}. This higher transmission rate results in a quicker epidemic, which can be seen by comparing the spread in 6.3 to 6.2, along with a larger overall epidemic. For no control here we get a final size of approximately 28,000,000. For this transmission rate it can be seen that the closure of workplaces is more successful in stopping the spread of the disease than vaccinating at 20%, as the final size is approximately 14,200,000 for the workplace closure, compared with 18,050,000 for vaccinating, which is a 50% decrease in comparison to a 36% decrease. However when compared to vaccinating 30% of the population, the vaccination is more successful reducing the final size to approximately 12,700,000.

Figure 6.4 again increases the transmission rate, this time to 1.25×10^{-4} days^{-1}. Here we have only considered the difference in the number of infected individuals for closing workplaces and random vaccination at 20%. It is clear that the impact of closing the workplaces is greater for this infection parameter set, here averting 45% of cases as compared to 30% for vaccinating.

When we consider the impact of vaccinating 20% of the population generated in Eames et al. [2009], the reduction in final size is approximately 40%, when the size of the outbreak in the population with no interventions was 50%. We have shown that when the outbreak in the synthetic population with no control is 40% (figure 6.2), there is a 50% decrease due to this level of vaccination, whilst when the outbreak with no controls is 56% (figure 6.3), the decrease in final size is approximately 36%. This implies that if the outbreak in the population with no controls was the same as in Eames et al. [2009], then the impact of this level of vaccination would be in broad agreement in the synthetic population. We have also seen that (at least for one set of infection parameters), that the increase in efficacy of

vaccination as a control measure is linear in the proportion of the population vaccinated, which again agrees with Eames et al. [2009].

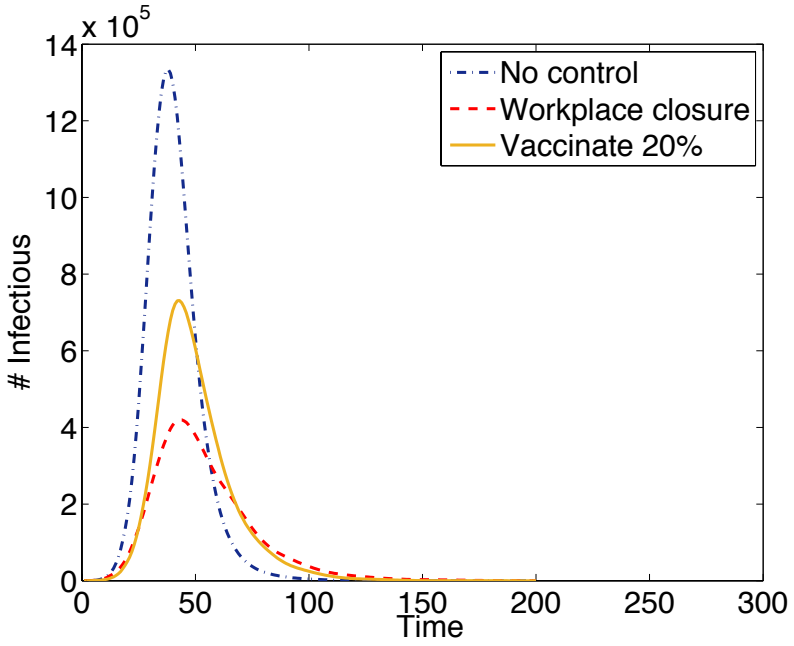


Figure 6.4: Average infection trails of simulations from the ISIS tool on the synthetic population for the UK. The transmissibility term is $\tau = 1.25 \times 10^{-4}$ days^{-1}. We see the comparison between closing workplaces and vaccinating 20% of the population and introducing no controls on the population.

6.2.2 Synthetic population simulations - summary & limitations

We have seen that the impact of interventions on the synthetic population is highly dependent on the expected final size of the epidemic in the case where no controls are enforced. For low values of the final size, it is likely that vaccinating 20% of the population would, along with averting more infections during the current epidemic, would also reduce the potential for a future epidemic by a large amount, than closing workplaces. As the final size increases however, the closure of workplaces begins to become far more successful at reducing the numbers of infecteds seen.

We also see a potential agreement to work done on clustered networks [Eames et al., 2009] in terms of the effectiveness of random vaccination.

There are several limitations to this approach. The fact that the synthetic population is so large and complicated means that the amount of conclusions that we have been able to draw from it are limited in scope, as we would need to explore many more parameter values before being able to draw meaningful conclusions, which has not been possible so far. Additionally, it has been shown that swapping 20% of the edges in the synthetic population leaving the degree distribution unchanged and then simulating an epidemic on the original and altered network with the same parameters yields epidemics which are dissimilar to each other [Bisset and Marathe, 2009]. This is unsurprising, as swapping edges will disturb the structure in the network in unpredictable ways. However, as much of the construction is based on seemingly sensible matching criteria, these techniques have

not been shown to represent reality accurately and ultimately there is no way to be sure how accurate the resulting network is.

This leads us to want to consider what other techniques can be used to model this and similar populations in a way that will allow us to draw more concrete conclusions about the similarities and differences between different models of contact structure, by allowing a larger parameter set to be explored. This begins with the pairwise approximation of the synthetic population and the UK and POLYMOD contact surveys.

6.3 Pairwise approximation of degree distribution model populations

In this section, we consider the degree distributions which are taken from the synthetic population, POLYMOD survey for the UK and the UK contact survey. We use the deterministic pairwise approximation ODEs (3.28) to investigate the difference between these degree distributions when it comes to the spread of an epidemic. This is equivalent to constructing configuration model type, undirected and unweighted networks corresponding to each degree distribution, and taking the average of a large number of simulations performed using the Gillespie algorithm.

The datasets all have contacts categorised by contact type, home contacts, work/school contacts and "other" contacts. To compare directly with the control methods used when directly simulating on the constructed synthetic population, the control method we will consider is that of closing workplaces. As the contacts are split by context, this is simple enough to implement by simply removing any contacts categorised as work.

As the POLYMOD and UK contact surveys have far fewer responses than the population considered in the synthetic population we draw 1,000,000 times from each degree distribution to construct our model populations. This is an arbitrary choice, as the equations used are independent of population size, but for the sake of direct comparison between the expected epidemic trails at a given time it is helpful to have relatable numbers of infections, rather than fractions of the population infected at a given time.

In figure 6.5 we can see the degree distributions that are used for the configuration networks of the synthetic population, POLYMOD and the contact survey. All three of these populations have contacts categorised by contact type, home contacts, work/school contacts and "other" contacts. Taking this information we work out the distribution of these contacts by type and then select the degree of each member of the population by choosing from these three categories.

This leads to a degree distribution similar to the one shown in 6.5a. We can see that the synthetic population has a greater proportion of nodes with degree greater than 25 than either the synthetic population or POLYMOD. The maximum degree for the contact survey and POLYMOD are around 75, whilst for the synthetic population it is around 20,500. As we see in 5.5 there is a secondary peak in the degree distribution of the synthetic population at around degree 40. This is a by-product of the population construction, as during workplace assignment large workplaces are split into sections with a maximum number of people in them (≈ 35). Therefore there is this increased number of people who will have this number of workplace contacts.

When we extract the workplace contacts, we get the distributions seen in 6.5b. The shape of the distributions for POLYMOD and contact survey are left relatively unchanged when compared to the full distribution, whilst the secondary peak in the synthetic population is removed, greatly altering the shape of the distribution.

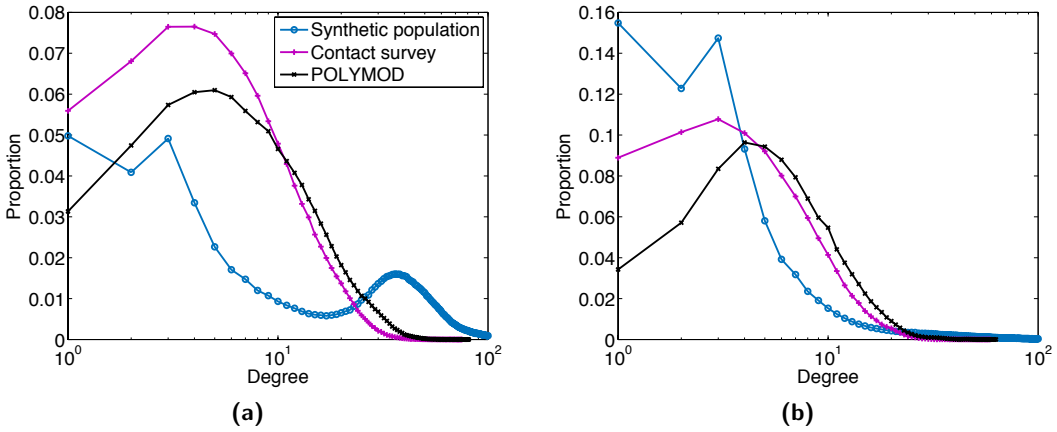


Figure 6.5: Degree distributions for synthetic population, contact survey and POLYMOD. (a) shows the full degree distributions for the three different populations. (b) has the degree distribution for the populations when the workplace contacts are ignored.

Figure 6.6 shows the progress of the epidemic through the populations. In each case the final size of the epidemic is approximately 0.6 times the population size for the full degree distribution, and the epidemics are also run with the same parameters for the scenario where there are no workplace contacts. Figures 6.6a, 6.6b and 6.6c show the epidemics for the synthetic, UK contact survey and POLYMOD model populations respectively, whilst 6.6d shows the three next to each other. Note that for the first figure, the x-axis goes from 0 to 20, whilst for the second and third both go from 0 to 200. We have considered the impact of closing the workplaces and this is given by the dashed line in each of the first three figures. In all simulations the transmission rate $\tau = 0.03$ and the recovery rate γ is varied in order to give the desired final size. For the synthetic population $\gamma = 0.76$ and for the contact survey and POLYMOD $\gamma = 0.15$ and 0.21 respectively. The course of the epidemic is fairly similar for the UK contact survey and POLYMOD, but is very dissimilar for the model population from the synthetic population.

The most obvious difference between the two model populations derived from the surveys and the synthetic population, is the speed at which the spread occurs in the early growth period. The different speeds can be accounted for by the fact that the early growth rate is dependent on the network, as given in (3.84). Calculating what the early growth rate is for each of the three different model populations in turn (synthetic population, UK contact survey and POLYMOD) for the no control and workplace removed scenarios we get: $r = 40.23$, $r = 55.40$, $r = 0.23$, $r = 0.11$, $r = 0.28$ and $r = 0.11$, which explains this vast difference in the early period.

The fact that the value of r is larger in the situation where workplaces are closed for the synthetic population degree distribution then when they are open, displays one of the difficulties with simply using the number of contacts that each person has in each location type as the degree of the person. In the synthetic population, a large proportion of the variance in degree distribution as given in §3), which defines the value of r when

the transmission and recovery rates are kept constant, increases due to the decrease in the mean degree by closing workplaces. In reality the contacts made in the "other" locations are likely to be fleeting and much less likely to produce additional infections as contacts made at home or at work.

In figure 6.7 we can see the impact of closing workplaces in each model population for a given final size when considering the full degree distribution. Again the two model population derived from surveys behave similarly to each other, with the closing of workplaces being somewhat more effective in the model population derived from the UK contact survey, but both following a similar pattern to each other. In both cases, the intervention is completely successful for final sizes below 0.4.

The impact of closing workplaces for the synthetic population degree distribution follows a much different trajectory, with the intervention increasing in effectiveness more sharply than for the other two populations as the final size decreases away from 1 in the full degree distribution. The increase then plateaus and between a final size of 0.5 and 0.6, the intervention becomes less effective than for the other populations. Unexpectedly by the time the final size is approaching 0 for the full population, the impact is decreasing for the synthetic population.

One reason for this is similar to the one which explains the increase in the early growth of the epidemic when workplaces are closed in 6.6, in that the contacts made in "other" locations are more important in the early growth period. When the final size is relatively low, the early growth period makes up a larger fraction of total infections and so the closure of workplaces has less time in which to make a big difference to the final size of the epidemic.

Comparing the impact of workplace closures in the ISIS simulations, as seen in 6.2, 6.3 and 6.4, where when the final size in the population is 0.4 and 0.56 of the entire population when no interventions are seen, we get a decrease of around 50% of cases by closing workplaces, whilst for final size of approximately 66%, the decrease is approximately 45%. The decrease seen in figure 6.7 is in agreement with these three figures. Whether there would be an agreement between these methods for more values of final size is currently unknown and would require extensive simulation time to test. However we would not expect to see a similar trajectory of intervention impact in the full synthetic population, as the contact time between individuals is included in the full model, which would diminish the impact of contacts made in "other" locations.

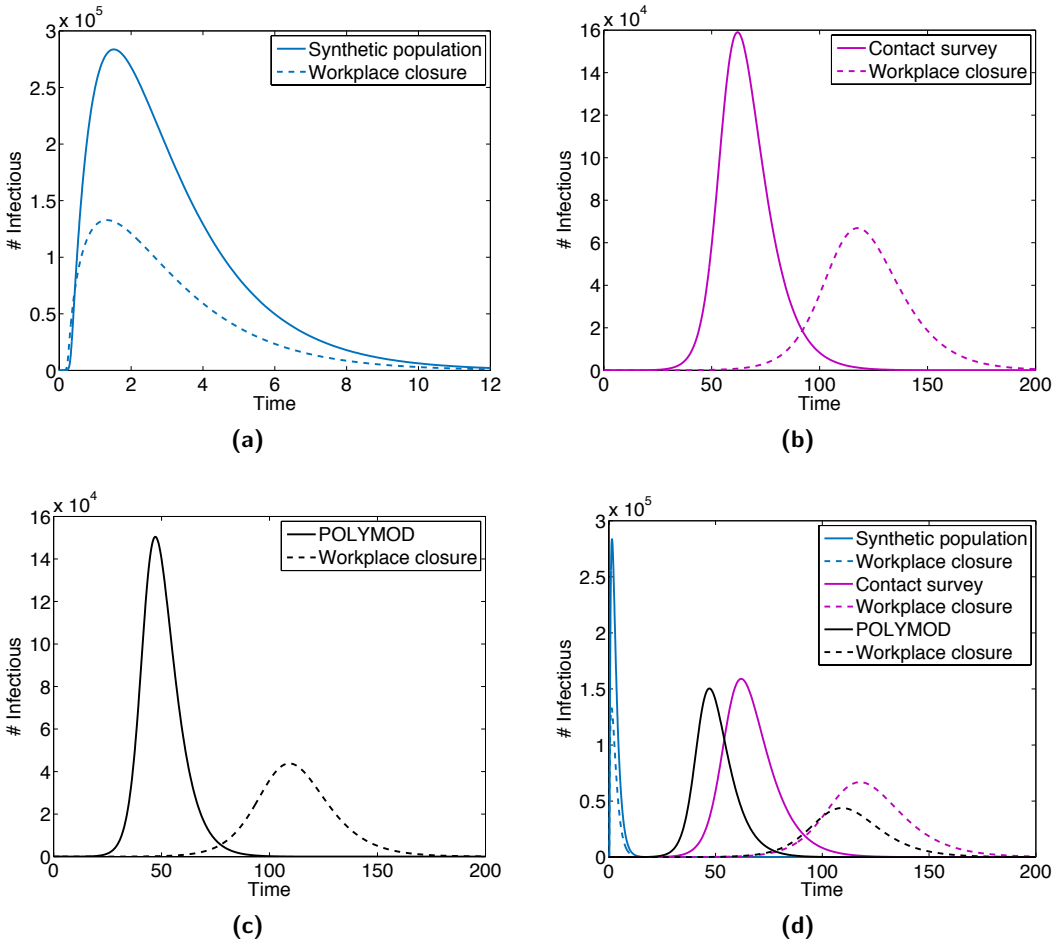


Figure 6.6: Figure showing the number of infected individuals over the course of an epidemic, as given by the pairwise approximation (3.28), inputting the degree distribution of the synthetic population model, the UK contact survey [Danon et al., 2012] and the POLYMOD survey for the UK [Mossong et al., 2008] respectively in (a), (b) and (c). The final size is always 0.6 of the population when the full degree distribution is considered. In (d) these are shown side by side where the differences in the spreads are more obvious. The dashed lines in each figure give the spread if degree distributions where workplace contacts are ignored is used.

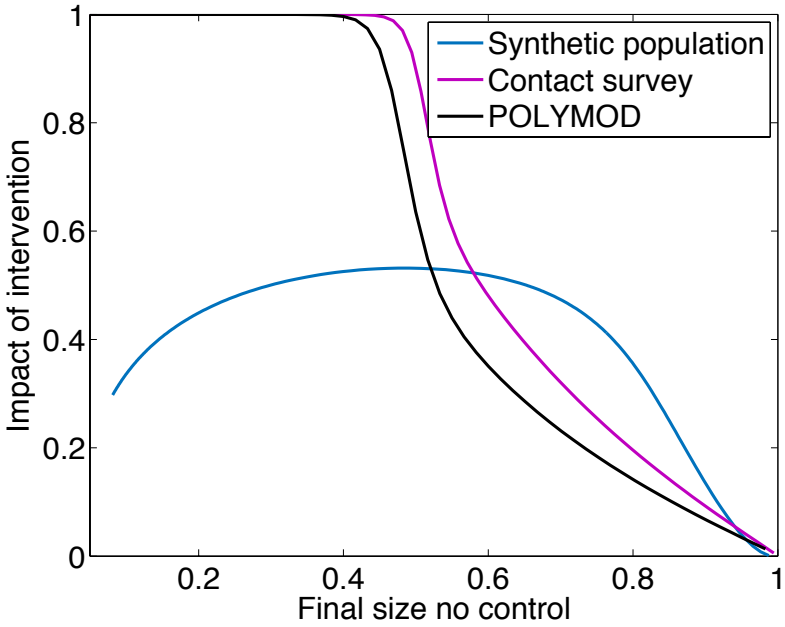


Figure 6.7: Figure showing the proportion of infections that are averted by closing workplaces in a population, calculated using the pairwise approximation to the degree distributions of the synthetic population, POLYMOD [Mossong et al., 2008] and the UK Social contact survey [Danon et al., 2012]. The x-axis gives the proportion of individuals infected in the situation where no controls are applied.

6.3.1 Pairwise approximation to degree distributions - summary & limitations

We have seen that the two degree distributions produced by the POLYMOD and UK contact surveys [Mossong et al., 2008; Danon et al., 2012], give qualitatively similar behaviour to each other both in the epidemics trails that are predicted using the pairwise approximation ODEs, but also in the impact of closing workplaces. This implies that these two contact surveys (not unexpectedly) are capturing the information about numbers of contacts in a way which is consistent with each other, and also that workplaces are captured in much the same way in each survey.

The behaviour of the epidemic using the degree distribution of the synthetic population is much removed from that seen in the other two populations. A postulated reason for this is the impact of the "other" locations on the spread on a configuration model type network with this degree distribution. The comparison is somewhat unfair, as in the true synthetic population, the contacts are weighted, which is incompatible with the methods used in the section. However, the contrast between these survey based model populations and the synthetic population are stark, and the fact that the predicted impact of workplace closure is similar using this approximation as in the few cases available to compare with the full synthetic population is likely coincidental.

An interesting next step from here would be to construct weighted networks in accordance with those given by the synthetic population, and compare these with results from the full synthetic population along with weighted networks for the UK contact survey and POLYMOD.

The limitations of this approach, include the fact that we are considering unclustered networks and using unweighted networks as discussed in §3.1.1. The fact that the POLYMOD and UK contact survey include biases in terms of the population represented in them means that there will be a bias in the degree distributions produced. Additionally, there is no attempt to investigate the knock-on effects of closing workplaces, such as the altering of the contact structure of these people who will now not be in work, and therefore altering other peoples contact network through this effect.

6.4 Who acquires infection from whom matrix method

Along with comparing simple configuration networks, we can also compare the spread predicted by the WAIFW matrices which are visualised in figure 5.4. We consider three different datasets to construct the WAIFW matrices for this section. The first is the synthetic population, where the numbers of contacts between people are taken by simply counting the number of contacts in the synthetic population, seen in 5.4a. The second is the same as this but weighted by duration of contact, seen in 5.4b, and the third is from the POLYMOD study displayed in 5.4c.

To model the spread of an epidemic through a population, with contacts following a given WAIFW matrix $\mathbf{W}$, we follow methods set out in Mossong et al. [2008] with notation similar to Metcalf et al. [2012]. We use a discrete time SIR model, with automatic recovery after one day. We use data from 2001 census to define the age distribution of people in England and Wales. The model used is indexed by age, therefore the state space of the model, $\mathbf{n}(t)$, is given by

$$\mathbf{n}(t) = (S_0, I_0, R_0, S_1, \dots R_z) \ , \tag{6.1}$$

where z is the maximum age in the population data. There is no demographic modelling, meaning that all people who begin in a specific age class will end the epidemic in the same age class, and there are no deaths or births.

We define the WAIFW matrix we are considering, $\mathbf{W}$, as follows:

$$\mathbf{W} = W_{z+1,z+1} = \begin{pmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,z} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,z} \\ \vdots & \vdots & \ddots & \vdots \\ w_{z,0} & w_{z,1} & \cdots & w_{z,z} \end{pmatrix} \ , \tag{6.2}$$

where $w_{i,j}$ is the average number of people age j that a person of age i will meet in a day. Along with this matrix and the population age distribution we define β to be the transmission rate between two contacts.

To calculate the value of β required for a given value of R_0 the following method is followed. One definition of R_0 is the average number of infections which will be caused by introducing a single 'typical' infected individual into a fully susceptible population. Therefore, for each age i up to the maximum age in the population we introduce a single infected of that age and calculate the number of people who this person will meet on an average day. This is given by the sum column of $\mathbf{W}$ corresponding to age i and denoted by π_i say. To satisfy the fact that this must be a 'typical' individual, we weight the mean number of individuals infected by a person of a given age by the proportion of the population who is that age.

Therefore the mean number of infecteds, denoted by η, is given by $\eta = \sum_i N_i \pi_i / N$ where N_i is the number of people aged i and N is the total population. Once we have calculated this we set $\beta = R_0/\eta$.

We denote the probability of infection of an individual of age a in a given time step by $\phi(t)$, which is given by

$$\phi_a(t) = \beta \sum_j \frac{I_j w_{j,a}}{N_a} \ , \tag{6.3}$$

where N_a is the number of individuals of age a and the term $I_j w_{j,a}$ gives the number of contacts of age a that infecteds of age j make in a day.

To calculate the probability of infection across all ages we take the row sums of the matrix produced by the following expression,

$$\phi(t) = \beta \begin{pmatrix} I_0 & I_1 & \cdots & I_z \\ I_0 & I_1 & \cdots & I_z \\ \vdots & \vdots & \ddots & \vdots \\ I_0 & I_1 & \cdots & I_z \end{pmatrix} \odot \begin{pmatrix} w_{0,0}/N_0 & w_{1,0}/N_0 & \cdots & w_{z,0}/N_0 \\ w_{0,1}/N_1 & w_{1,1}/N_1 & \cdots & w_{z,1}/N_1 \\ \vdots & \vdots & \ddots & \vdots \\ w_{0,z}/N_z & w_{1,z}/N_z & \cdots & w_{z,z}/N_z \end{pmatrix} \ , \tag{6.4}$$

where $\odot$ denotes element-wise multiplication.

To calculate the number of people who move from state to state, for each state in $\mathbf{n}(t)$, we draw from a multinomial distribution, with probabilities dependent on the compartment in question along with the probability of infection by age. Define $\mathbf{K}_{S_a}(t) = (P_{S_a}, P_{I_a}, P_{R_a})$ to be the probabilities of moving to and of the three states of the model for susceptible individuals of age a. These probabilities are given by,

$$\mathbf{K}_{S_a}(t) = (1 - \phi_a(t), \phi_a(t), 0) \ . \tag{6.5}$$

Infected and removed individuals are assigned to the removed class with probability 1, as the length of infection is assumed to be 1 day (or whatever length the timestep we choose is).

In figure 6.8 we consider the WAIFW matrices from the three populations mentioned above, along with uniform mixing, where the number of contacts in each age group is proportional to its population. Figure 6.8a shows the age specific R_0 for each age in the different populations in question. To generate this figure, a transmission rate was calculated that gave a value of $R_0 = 1$ for a typical individual in the population. By this we mean that for a randomly selected individual, where the probability of an individual of age a being selected is proportional to the number of people aged a, the expected value of R_0 is 1. We see that for the uniform mixing population all ages have $R_0 = 1$. For the other populations there is a large swing in the values of R_0 over all ages, with the largest deviations arriving in the weighted synthetic population and the smallest ones from the POLYMOD data. This implies that the heterogeneity is larger in the synthetic populations than is generated by the questionnaires used to inform the POLYMOD data.

To generate figures 6.8b, 6.8c and 6.8d, 1000 stochastic simulations were performed, where a single infection from an individual with an age chosen according to the number of people in each age bracket was introduced into an otherwise fully susceptible population. These were then averaged to give the figures shown.

In figure 6.8b, we can see the average size of an epidemic according to each of these

WAIFW's for values of R_0 from 0 to 3. We see that the mean final size is 0 across all populations for $R_0 < 1$ and at $R_0 = 1$ it becomes non-zero for the WAIFW defined by the weighted synthetic population along with the POLYMOD dataset. For values of R_0 below 1.5, going from smallest to largest in terms of final size in the four populations we have the uniform population, un-weighted synthetic population, POLYMOD and then weighted synthetic population. At $R_0 = 1.5$, this order becomes less clear, but by $R_0 = 2.2$, the order of final size has reversed. The mechanism for this behaviour is discussed below.

In 6.8c we see the probability that the outbreak 'took off' in the population for a given value of R_0. To be classified as taking off, there must have been at least 1,000 infections in the population. As the population is approximately 52,000,000, this constitutes a small proportion of the population becoming infected, but means that we do not include any times that the epidemic died out immediately. We see a similar pattern here in that for low values of R_0, in this case until $R_0 = 0.9$, the values are all 0. Once the values are non-zero, the order is initially synthetic weighted, POLYMOD, synthetic non-weighted and then uniform mixing. Again the order changes at around $R_0 = 1.5$ and at the end the order is uniform mixing, POLYMOD, synthetic non-weighted and then synthetic weighted. Again, we discuss the reasons for this below.

Finally 6.8d shows the expected final size if the epidemic takes off for each population. The weighted synthetic population and POLYMOD again have non-zero values of this for lower values than uniform or non-weighted synthetic population, and again for large values of R_0 the uniform population has the largest expected size, with the non-weighted synthetic population being very close, followed by the weighted synthetic population and then POLYMOD.

To explain the observed probability that the epidemic will take off, given in figure 6.8d, we refer to the age-specific R_0 given in figure 6.8a. The fact that the weighted synthetic population and the POLYMOD are always the first two populations to have non-zero values can be explained by the fact that it is these two populations which have the greatest maximum values for the age-specific values of R_0. This implies that when the epidemic reaches anyone in these age brackets, the average number of people who are infected by them increases, making it more likely that the epidemic will take off. For large values of R_0, the fact that the minimum values of age-specific R_0 are smaller for both synthetic populations implies that there is a larger probability that there will be a small or no outbreak at all, as if the ages with very low values of R_0 are initially infected, then it is more likely that the epidemic will not take off, or only a handful of people will become infected.

For figures relating to the expected final size of the epidemic (figures 6.8b and 6.8d) when R_0 is low, we again only need to consider the age-specific R_0, as the magnitude of the largest age-specific R_0's in each population implies the order in the final sizes. As R_0 gets larger however it is not simply a case that we can explain the order in final sizes by considering figure 6.8a. For example, the fact that the expected final size, given that the epidemic has take off, for POLYMOD is well below the other populations for high values of R_0 is not explainable by the age-specific R_0.

To explain this, we need to also consider the the age-age mixing for each population, which is shown in figure 5.4. When we look at this figure, we can see that the amount of mixing between ages for the POLYMOD population is extremely heterogeneous, with the mixing between people of different ages being very small in comparison with the mixing between people of the same age, especially for individuals between the ages of 5

and 20. Note that these are the same individuals who have increased age-specific values of R_0. As the population value of R_0 increases and large portions of the population are becoming infected, the fact that the members of the population who are driving the epidemic with their high age-specific R_0's are mainly infecting each other implies that the effective R_0 in these age groups, along with the population as a whole will be decreasing faster than expected as the available susceptibles are quickly diminishing. We can apply similar arguments to the age-age mixing for the synthetic populations, which shows that the mixing is more heterogeneous between ages for the weighted population than for the non-weighted one, which again implies the rank of these two populations in terms of final size for an epidemic which has taken off.

Along with the final size of the epidemic, of interest is the time that the epidemic will last, along with the peak level of infection. This is interesting from a control perspective as if we have two epidemics with the same final size, one of which ends after 20 days and the other after 100 days, then the time available to react and put controls in place is significantly different. To consider this we can simulate an epidemic on each of the populations, all or which have the same final size and see what the size of the epidemic is at each time point. This can be seen in figure 6.9.

Figure 6.9a shows the spread using the different mixing estimates in the population for 20 separate simulations, all of which have a final size of 0.42 times the population size. The pattern of spread seems to indicate that the POLYMOD structure gives the largest peak along with the quickest time to the end of the epidemic. However, this is not clear due to the variation in times to reach the peak when we consider the same mixing estimate. To solve this problem we relabel time 0 to be the point at which 500 infections were made in the population, and then plot the spread of the epidemic from that point. This can be seen for the same 20 simulations in figure 6.9b, which confirms the fact that POLYMOD gives the largest peak and quickest epidemic, followed by the weighted synthetic population, the non-weighted synthetic population and then uniform mixing. Figure 6.9c shows that this pattern is also seen at higher values of R_0, as for this figure the final size of the epidemic is 68% of the population.

The impact of vaccination in the WAIFW scheme is simple to investigate, as this is equivalent to putting a given proportion of the population into the removed class at the point at which vaccination is applied to the population. We consider the impact of vaccinating 20% of the population at time zero, and compare the spread of the epidemic through time for the different mixing patterns. Figures 6.9d and 6.9e show the spread in the population for the case where the final size is 42% and 68% of the population respectively. We can see that the inclusion of vaccination does not alter the predicted order among the models in terms of the peak epidemic height or the time to the end of the epidemic, but decreases the peak height more significantly for the mixing patterns which give a smaller height, and also increases the length of the epidemic more significantly for these populations (most obviously for the uniform mixing population in figure 6.9d).

Finally figure 6.9f shows the proportion that the size of the epidemic is reduced by if a randomly selected 20% of the population is vaccinated.The x-axis gives the final size of the epidemic in the population when no controls are used. We see that for the majority of values of final size (and therefore R_0) the population which this impacts most significantly is the uniform mixing population. This is due to the fact that the spread is driven evenly by all members of the population here, but in the other cases, the spread will be driven by a minority of the population, meaning that randomly vaccinating people will not be an

optimally efficient method for controlling the epidemic. We see that it is more effective in the POLYMOD population than in the weighted synthetic population, and is again more effective for the non-weighted synthetic population, which again comes down to the level of heterogeneity in the age-specific R_0's for these populations.

We can compare this to the decrease in the spread of infection in the population when we consider the full synthetic population and vaccinate 20% of the population as is seen in figures 6.2 and 6.3. For the first figure, the final size is approximately 40% of the population and when we vaccinate 20%, the number of infections decrease by 52%. For the second the final size is 56% and 36% of these are prevented with 20% vaccination. For these final sizes in the non vaccinated population, figure 6.9f gives a decrease of approximately 70% and 50% respectively for the weighted synthetic population and even larger for the non-weighted synthetic population. This predicted impact of vaccination is far higher than that seen in the full synthetic population. An important factor in terms of explaining this is that to create the WAIFW for the synthetic population we are already averaging behaviour and therefore missing out on the variability in the population. Therefore a similar argument can be applied to the true synthetic population when compared to the population defined by the synthetic WAIFW as was applied to this WAIFW in comparison to uniform mixing in that the people who are driving the epidemic are in a minority (and are more extreme than for the WAIFW) and therefore will be less likely to be removed from the dynamics than people with less extreme behaviour.

When comparing the impact of vaccination on the spread of the epidemic in a population with heterogeneous levels of mixing, it is possible to consider what changes when we use targeted vaccination instead. In this scenario we would vaccinate a given percentage of the population using some statistic of the population to choose which individuals this percentage includes. Here we would target those individuals who have the highest age-specific R_0's, meaning that for the mixing defined by the weighted synthetic population and POLYMOD we would get the greatest increase in efficacy of vaccination and for the uniform mixing there would be no difference to random vaccination. This has not been done here, but the impact of targeted vaccination on populations with weighted contacts can be seen in Eames et al. [2009].

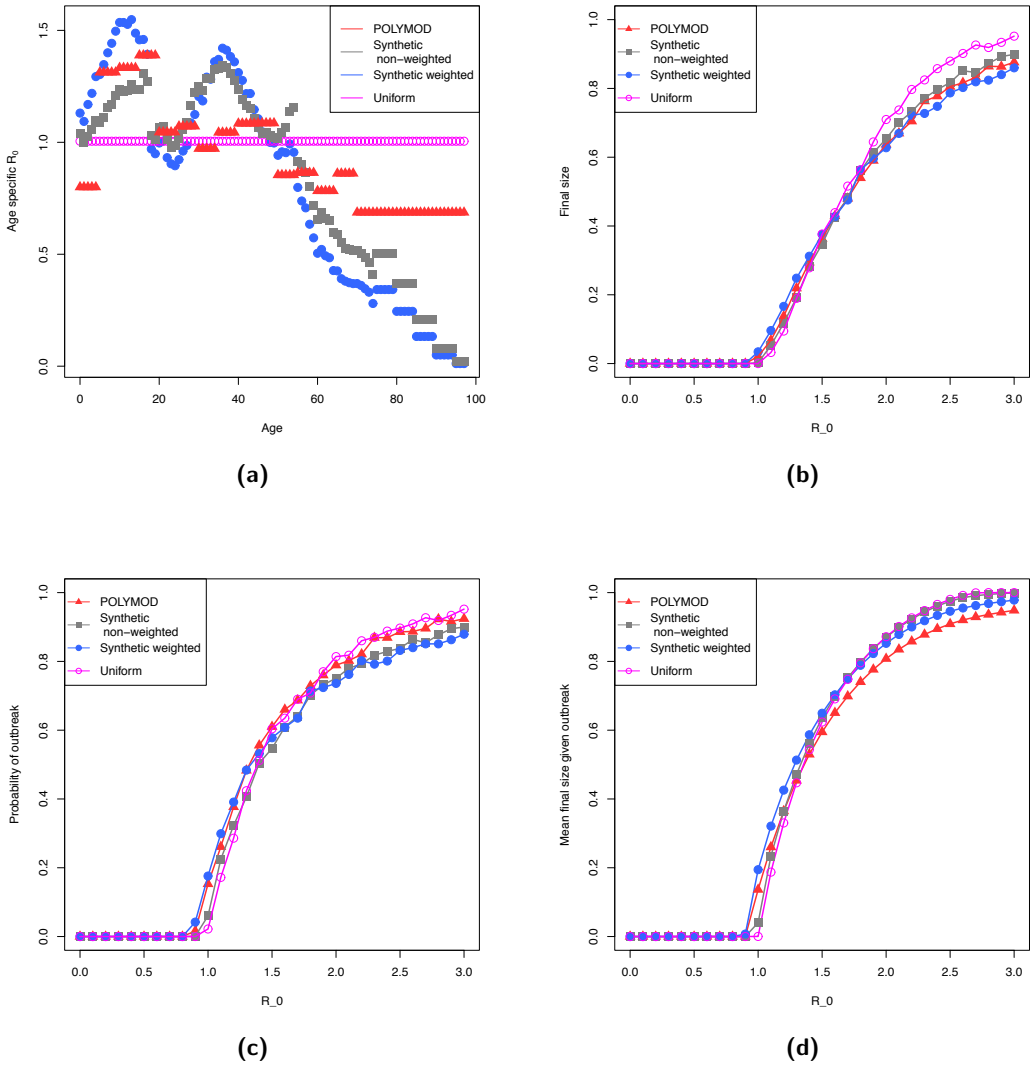


Figure 6.8: Characteristics of epidemics on a population whose contact structure is modelled by the WAIFW matrices for the synthetic population, both weighted by contact duration and non-weighted, POLYMOD and uniform mixing. (a) shows the age specific value of R_0 and its dependence on the population in question when the overall R_0 for the populations is equal to 1. (b) shows the final size of the epidemic for different values of R_0 if we use the method described above to simulate an epidemic. We consider the synthetic population where the WAIFW is given by counting contacts between age groups and also when this is weighted by contact duration along with the POLYMOD WAIFW for Great Britain and uniform mixing. (c) gives the probability that the epidemic will 'take off' in these four population for the same range of R_0's. Taking off here implies that at least 1,000 infections took place. (d) gives the mean outbreak size if we condition on the fact that the outbreak took off.

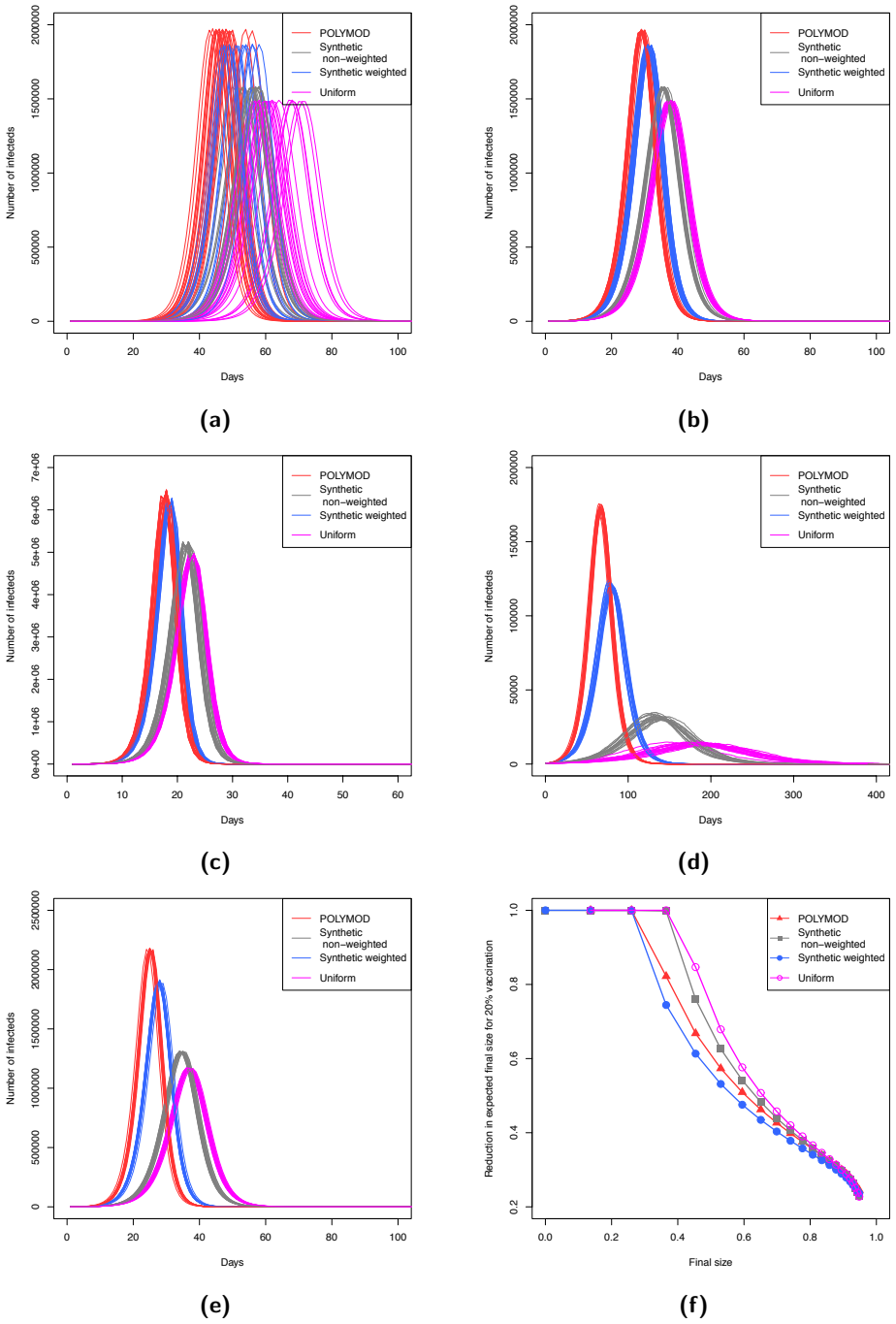


Figure 6.9: Simulations performed on population whose contact structure is modelled by the WAIFW matrices for the synthetic population, both weighted by contact duration and non-weighted, POLYMOD and uniform mixing. (a) shows the trails of 20 epidemic simulations which all took off and reached a final proportion of 0.42 for each of the four populations in question. In (b) we see the same 20 epidemics as in (a), but here time 0 is taken to be the point at which there were 500 infections in the population. (c) is similar to (b), but the final size is 0.68 in this case. Note that the x-axis goes from 0 - 60 in this figure, but from 0 - 100 in (a) and (b). (d) shows the outbreaks in the four populations if 20% of the population is vaccinated for the parameters which produce (b), and (e) does similarly for the parameters of (c). Note the x-axis on (d) is from 0 - 400 instead of 0 - 100 in (b) and for (e) goes from 0 - 100 rather than 0 - 60 in (c). (f) shows the reduction in final size if the population is 20% vaccinated.

6.4.1 WAIFW summary and limitations

We have seen that for epidemics which are spread according to the WAIFW's produced by the synthetic population along with the POLYMOD dataset and uniform mixing, we get epidemics which behave relatively similarly at the aggregate level, that is once we strip away the epidemic trajectories and focus solely on the probability that an epidemic will take place, or on the expected final size of an epidemic in these populations. However when we look at the actual spread through time, we see that there are significant differences in terms of the timing of the epidemics along with the peak of the epidemic. This has repercussions for how effective plausible intervention strategies may be predicted to be, as by the time that control strategies are prepared the stage in which the epidemic is in can be vastly different. This combines with the fact that the peak of the epidemic has consequences in terms of how effectively the treatment of potential diseases is, due to the increased burden on healthcare systems caused by an increase in peak incidence.

In terms of the use of the WAIFW to model the spread of the epidemic, this is in some ways closer to the actual synthetic population than the previously examined pairwise approximation model, as this is based entirely on the contact network that is produced for the synthetic population, and does not model simply one aspect of the population as the degree distribution does.

We now look at the spread when we include more heterogeneity and clustering in the population by constructing meta-populations for these model populations, using similar principles as used to construct the full synthetic population.

The main limitation of this approach is that we remove any heterogeneity from the populations, as all people of a given age are assumed to be equivalent to each other. Again we don't consider the knock-on effects of closing workplaces, which would have a large impact on the contact structure.

There is also underlying uncertainty in the WAIFW matrices produced by either the synthetic population, due to the use of a large number of assumptions to construct the population, or the survey based techniques used to derive POLYMOD and the UK contact survey, which make assumptions regarding the required contact to spread an infection.

6.5 Meta-population description

To add additional structure in order to compare the synthetic population against POLYMOD [Mossong et al., 2008] and the UK contact survey [Danon et al., 2012], we can consider a meta-population which is constructed using similar principles to the synthetic population itself. This is done for each of the populations that have been compared to each other previously, so the following process is used to construct representations of the synthetic population, POLYMOD and UK contact survey.

We consider population sizes on the order of 10^5 rather than the size of the synthetic population ($\approx 50 \times 10^6$), as this means that we can consider more parameter values and compare it with other population models. To do this we begin with the degree distributions for the home, workplace and other contacts for each of these populations. Each person is assigned a number of contacts in these three circumstances according to the distributions in question. Once this is done, people are grouped into households,

workplaces and other groups by the given size. For example, a person who has household size 3 is randomly assigned to a household, where the two other household contacts are chosen at random from the people in the population with household size 3. This is repeated for workplace and other contacts. In the remainder of this section, when referring to the synthetic population, POLYMOD or UK contact survey population or model, this refers to a population constructed using this method.

Once this population is assembled a disease can be spread on it. The model that we use here is the 'chain binomial' model, with non-overlapping generation times. This is a discrete time model, where each time point describes the spread of the disease through either houses, workplaces or other locations. At each time point, we first check for any infected person who has been infected for the given number of time steps needed to recover. These infecteds are then placed in the recovered compartment.

For the remaining infectious individuals, we then choose from a binomial distribution the number of people who it infects at this time point, where the transmission rate, and hence the probability of infecting a given susceptible contact, is dependent on the size of the location in question. We then move onto the next time point when all infectious individuals have drawn from a binomial distribution, the number of infections that they will perform. We continue looping through the three locations in turn until the epidemic ends.

The parameters used for the binomial are n = number of susceptibles remaining in given location, and $p = 1 - \exp(\nu(k-1)^{-\delta})$, where the exponential power is the transmission rate. To construct this transmission rate, ν is some constant dependent on R_0 and the location type, k is the location size (so $k-1$ is the maximum number of contacts in the location) and δ is a number between 0 and 1, which can be used to tune between frequency and density dependent transmission (δ is equivalent to $1-\epsilon$ introduced in § 4.2).

For workplaces and other locations we multiply ν by 0.5, to indicate that more time is spent in the household than in any other location [UK Data Service]. This is a somewhat arbitrary choice. It can obviously be argued that this value should be lower in all cases other than households and that it should be higher in workplaces than for other locations, but the sensitivity of results to this has not been investigated here.

This construction method has traits in common with the construction of the full synthetic population in that people are assigned activities to take part in each day, and they interact with different people when doing these different activities. Here however we are considering all locations to be fully mixed, so this also means that the level of clustering in the population is likely to be higher than that seen in the synthetic population. Additionally people are assigned a maximum of one location for each type of contact, which is not true for the synthetic population, where people can be assigned multiple work and other locations. As we are interested in comparing the synthetic population with smaller, less complex models, the population size is also far smaller here, typically between 100,000 and 200,000. People are also limited to workplaces of size 50 in the meta-population and can only make contacts with up to 100 people in other locations, which is far lower than the synthetic population.

6.5.1 Meta-population simulations

Once the populations are constructed, we simulate on them as described above. Figure 6.10a shows the value of R_0 given the transmission rate of the epidemic. Also examined is the impact of altering δ to tune between frequency dependent and non-frequency dependent transmission. The values looked at here are $\delta = 1$ (which is frequency dependent transmission) and $\delta = 0.85$. For $\delta = 1$, we see that the population with the largest value of R_0 is the POLYMOD dataset, followed by the synthetic population and then the contact survey. Altering δ, we can see that the values of R_0 for POLYMOD and the synthetic population are virtually identical. The fact that the synthetic population's reproductive number has increased so much relative to the other populations shows clearly that there are more locations with larger numbers of people in them, as it is these locations which will increase the value of R_0 as δ is decreased.

In figure 6.10b we consider the final size of the outbreak in the three populations for given values of R_0. Again we consider $\delta = 1$ and $\delta = 0.85$. We can see that the final size of the epidemic in the populations at a given value of R_0 is largest in the contact survey, followed by POLYMOD and finally by the synthetic population. This can be explained by the level of clustering in the populations. As mentioned previously increasing clustering decreases the number of total infections, due to the fact that the number of susceptibles that an individual has contact with is diminished by infections not only caused by itself, but also by other infected people who they are connected to. This results in a must faster depletion of susceptibles than is seen in configuration type networks.

Note that as the workplaces and other locations are fully connected in all of the populations, this implies that the level of clustering is likely to be larger than in reality in all cases. For the synthetic population, the level of clustering will be higher than for the other two populations, as this is a measure of the number of triangles in the population in comparison to the number of total triples in the population. As there are more places with a large number of connections in them in the synthetic population, this implies that there will be proportionally more triangles in the synthetic population than the other two.

This can be explained by considering an example. Consider the two networks shown in 6.11. We take 6.11a to be in the case where there is 1 home contact, 1 other contact and 2 work contacts and 6.11b to be similar but with 3 work contacts. We wish to calculate the clustering around the central node using the method given in Watts and Strogatz [1998]. To do this, we calculate the size of the neighbourhood of this node in each case, where the neighbourhood is defined to be the number of nodes which are connected to it. This can be written as,

$$N_i = \{v_j \in V : e_{ij} \in E \wedge e_{ji} \in E\} \ , \tag{6.6}$$

where V is the set of vertices and E is the set of edges in the network. For 6.11a this is 4 and for 6.11b it is 5. The level of clustering is then given by the proportion of the total number of possible connections between members of the neighbourhood that are actually in the network. Again as it is an undirected network, all connections are double counted to reflect the fact that it is equivalent to a directed network where if two nodes are connected in one direction then they are also connected in the other. The number of unique connections that are able to exist between k nodes (not allowing multiple connections between nodes)

in an undirected network is $k(k-1)/2$, and so clustering is given by the following,

$$C_i = \frac{2|\{e_{jk} \in E : v_j \in N_i \wedge v_k \in N_i\}|}{|N_i|(|N_i|-1)} \ . \tag{6.7}$$

Performing this calculation for 6.11a we get $C_i = 1/3$ and for 6.11b, we get $C_i = 3/5$, meaning that it is higher in 6.11b, due to the inflation of the number of work contacts.

As mentioned above, due to the assumption of fully connected workplaces and other locations, the level of clustering is likely to be higher than in reality. However, as can be seen in 5.6, the level of clustering is larger in the synthetic population than for the contact survey, and so even though both will be exaggerated, the rank of clustering will be maintained. Additionally we can see that altering the transmission rates away from being frequency dependent has no impact in the contact survey and POLYMOD meta-populations and a relatively small impact in the synthetic population.

Finally 6.10c gives the expected infection profile for outbreaks with a final size of 40% of the population for all three populations for $\delta = 1$ and $\delta = 0.85$. We see that the synthetic population has by far the largest peak and the fastest time to the end of the epidemic whilst the POLYMOD and contact surveys are far more similar to each other, with the peak being marginally faster and higher in the POLYMOD population. It is interesting to note that for the pairwise approximation predictions this is also observed (figure 6.6), but for simulations involving the WAIFW, we get that the peak is higher and time to the peak is faster for POLYMOD when compared to the synthetic population; especially in the case of the non-weighted synthetic population, which comparing only degree distributions is equivalent to (figure 6.9). For the pairwise approximations is due to the large variance in number of contacts which drives the early growth period of the epidemic. For the meta-population, this can be explained by the network structure and the level of clustering in the population.

As seen in 6.10b, the value of R_0 needed to give the same value of final size in the synthetic population is considerably larger than for the other two populations. This implies that the epidemic will initially spread more quickly through the model population, as everyone will initially infect more people. This leads to a faster and higher peak in populations with the same level of clustering, due to the quicker depletion of susceptibles in the population when the transmission rate is increased. Though the level of clustering is higher in the synthetic population, this increased value of R_0 compensates sufficiently to still have a higher and faster peak whilst infecting the same number of people. This is similarly why the peak is higher and faster for the POLYMOD meta-population, but the difference is far less pronounced here. Again the difference between the epidemic spreads for different values of δ is negligible, hence the impact of altering this will not be considered from here on.

In 6.12a and 6.12b, we again consider the time to the peak and the size of the peak for the epidemics on the different populations for a range of values of the final size. It can be seen that the results discussed relating to the peak from 6.10c are not limited to a specific value of the final size, and are true in general for these meta-populations. It is again clear that the POLYMOD and contact survey populations are far more similar to each other than the synthetic population.

In 6.12c, we consider the effectiveness of different intervention strategies on the spread of the epidemic on these meta-populations. We consider the impact of discounting any con-

tacts from workplaces against vaccinating a randomly selected 20% of the population. For the synthetic population, closing workplaces is more effective in all cases than vaccinating 20% of the population, whilst this is not the case for the other two meta-populations as vaccination becomes more efficacious once the final size is approximately 80% of the population.

Considering the impact of closing workplaces and vaccinating 20% of the population on the spread of the epidemic, we see that this method is much more effective in the meta-population construction of the synthetic population than it is in the full synthetic population. Comparing the impact of these interventions in 6.2 and 6.3 for vaccination we get an impact of 52% and 36% for final sizes of 40% and 56%, whilst for the meta-population the impact is slightly above 60% in the first final size and slightly below 60% in the second. For closing of workplaces, in the full synthetic population, the impact is approximately the same for both final sizes at 53%, but for the meta-population, the impact is 100% for these final sizes.

Again this shows that removing the heterogeneity of the full synthetic population to model the populations results in us getting large differences in the behaviour of the epidemics, especially when we consider the impact of interventions.

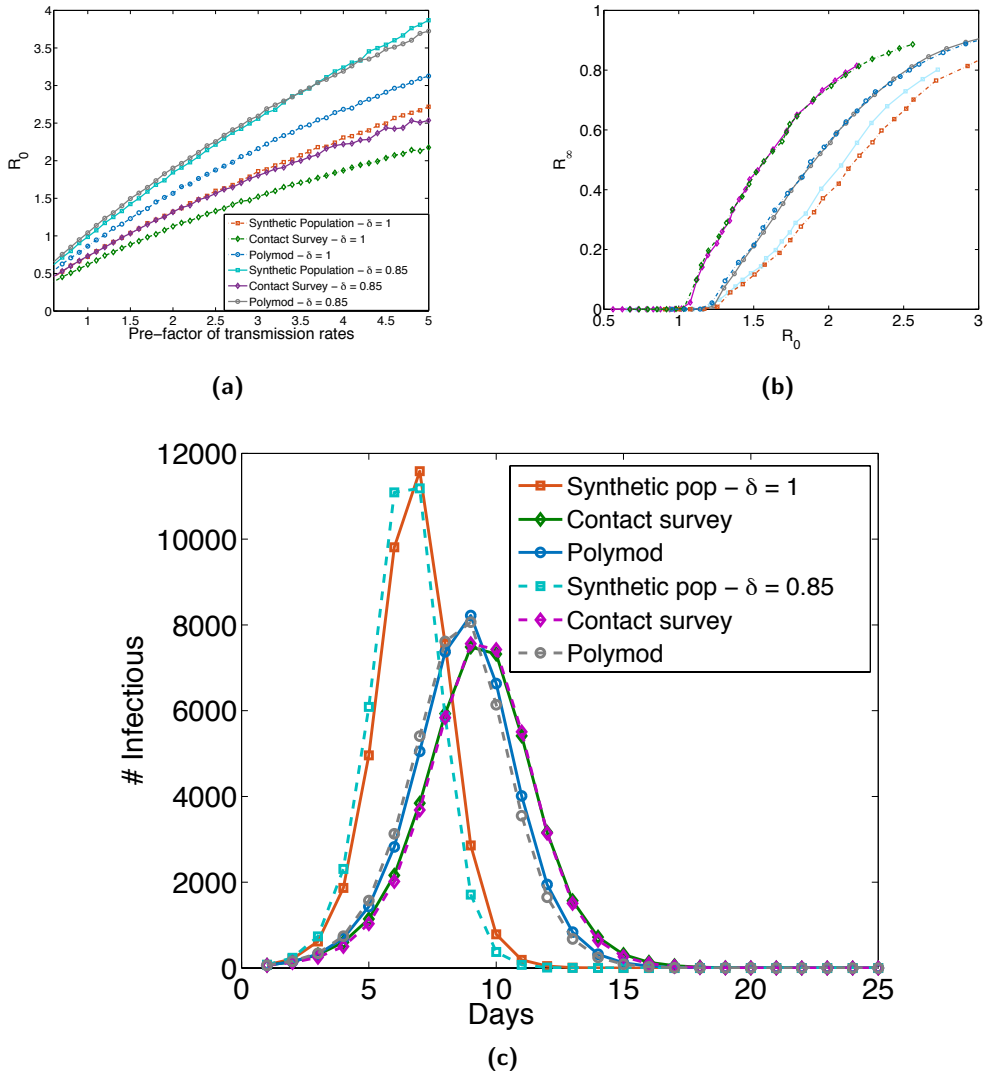
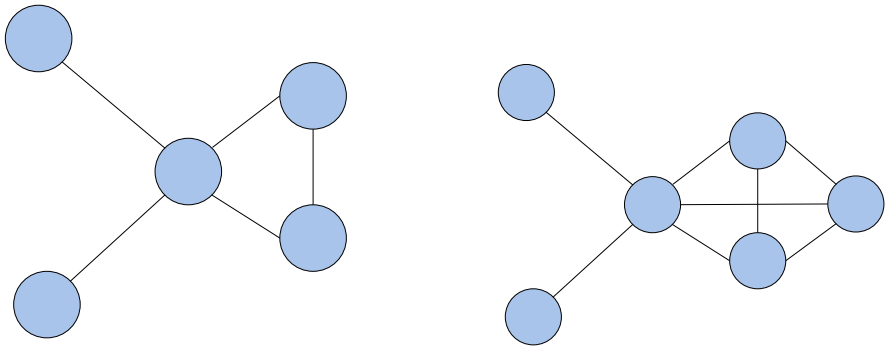
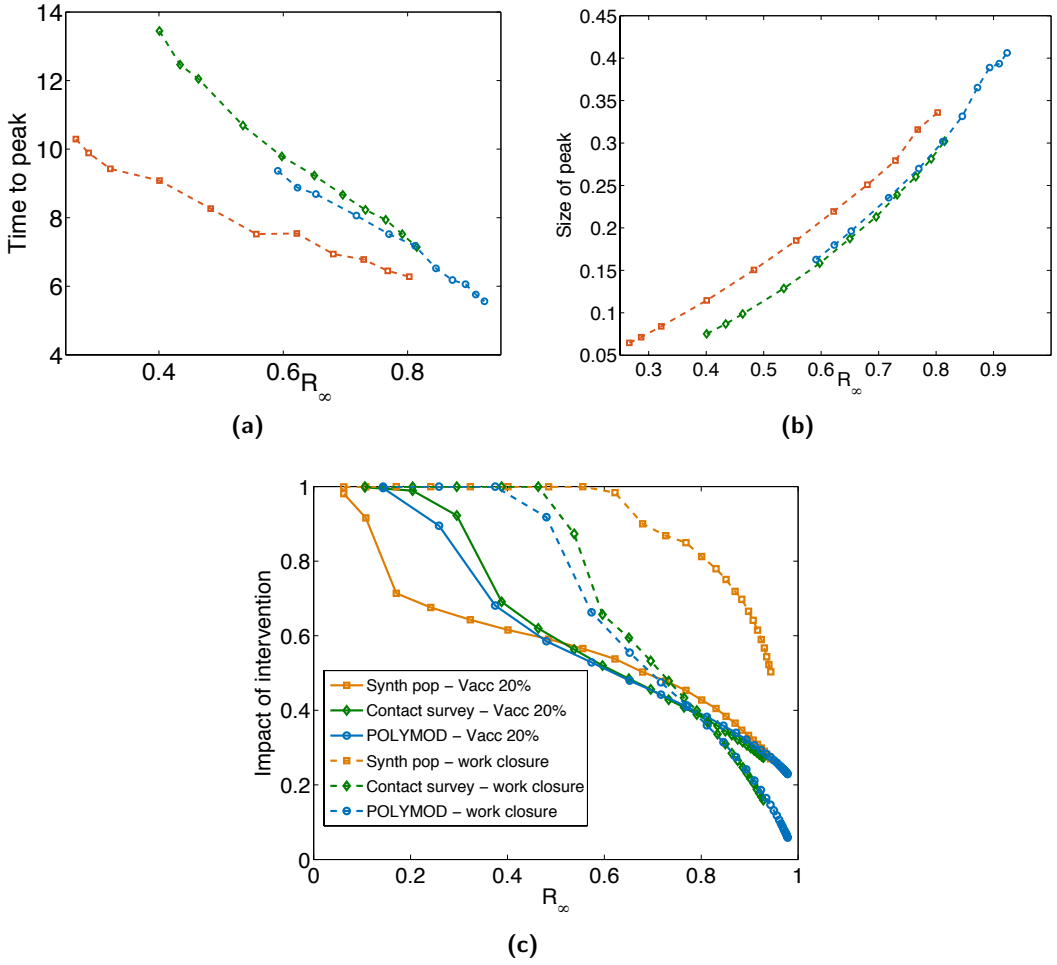


Figure 6.10: Data from simulations performed on meta-populations for a population modelled according to the synthetic population, POLYMOD and UK Social contact survey. In (a) we can see the change in R_0 which is given by multiplying transmission rates by a given pre-factor. Transmission rates are given by $A(k-1)^{-\delta}$, where A is the transmission pre-factor and δ is given by either 1 or 0.85 in this figure. In (b) we have the accompanying value of R_∞ for given values of R_0 across the three populations that we are considering. (c) gives the average spread by day of the epidemic in the scenario where the final size is 0.4 in all populations.



(a) (b)

Figure 6.11: Two configurations of neighbours in the home, work and other context.



(a) (b) (c)

Figure 6.12: Data from simulations performed on meta-populations constructed as described in 6.5. (a) shows the time to the peak for various values of R_∞. In all cases here $\delta = 1$ and the colour scheme follows that given in 6.10, where orange is synthetic population, green is contact survey and blue is POLYMOD. (b) gives the size of the peak for different values of R_∞. (c) shows the proportion of infections that are averted by either closing workplaces of vaccinating 20% of the population at $t = 0$.

6.5.2 Meta-populations summary & limitations

By constructing meta-populations based on the degree distributions of the synthetic population, UK and POLYMOD contact surveys, we have been able to consider the impact of clustering in the population on the spread of an epidemic. The level of clustering that this method produced in the synthetic population and POLYMOD in comparison to the UK contact survey meant that we needed to increase the value of R_0 in these population models considerably to observe a similar final size as that observed for the UK contact survey. This led to a faster and more highly peaked epidemic for the synthetic population, and to some extent for the POLYMOD model too.

We also observed that for a given final size, in general the POLYMOD and UK contact survey derived population models behave similarly to each other, whilst the synthetic population behaves much differently.

In terms of intervention, it was possible to consider the impact of vaccination and workplace closure here. For the majority of values of final size, the impact of closing workplaces was greater than vaccination. For the synthetic population model, this was true for all values of R_∞, whilst for the other two, around a final size of 0.8, vaccination became a better option. This is unlike what was observed in the full synthetic population, where as the final size increased, the impact of workplace closure improved in terms of its comparison to vaccination at 20%. However if we increased the size of the epidemic in the full synthetic population, at some point, we would definitely see that vaccination was better in terms of reducing the final size, as this would always reduce infections by 20%, whilst closing workplaces may ultimately prevent no infections if transmission is great enough. Whether this occurs at a final size of around 80% is unknown currently.

Limitations of this approach include reliance on the degree distributions to define the meta-populations, which are base on many assumptions, especially in the synthetic population case. We assume that the very large degree distributions seen in the synthetic population are not artefacts, and by assuming that transmission in 'other' locations is as strong in workplaces and half as strong as homes, it is likely that the impact of other locations is overstated. This will also have an impact on the effectiveness of closing workplaces as an intervention, as a large proportion of the spread will be produced by the 'other' locations.

6.6 Summary of chapter

In this chapter, we have compared the predicted spread in several different models of the same population, which have been modelled in different ways, which include different levels of heterogeneity.

Firstly, the full synthetic population, whose construction was described in the previous chapter was examined. Due to the size and complexity of this population, relatively few simulations were done, with a relatively small range of parameters investigated. Here we compared vaccination to closing workplaces and saw that as final size increases, the difference in effectiveness of the closure of workplaces in comparison to vaccination increased.

Next we considered the pairwise approximation and saw that the degree distribution de-

fined by the synthetic population gave huge growth during the early period of the epidemic when compared to the other UK contact survey and the POLYMOD survey model populations. The behaviour of the two model populations derived from the surveys was quite similar to each other, whilst the synthetic population showed much different behaviour, both in terms of the spread of the epidemic and the efficacy of closing workplaces. We did not consider vaccination of the population here.

The who-acquires-infection-from-whom (WAIFW) matrices for the synthetic population, weighted and un-weighted by time were then compared to POLYMOD and uniform mixing. Here the peak was greater and the time of the peak and length of the epidemic was shorter for the POLYMOD defined contact structure than for the synthetic population based models, which was not in agreement with the pairwise approximations.

Finally, meta-populations were considered and again we saw that the UK contact survey and POLYMOD survey were in much closer agreement to each other than they were with the synthetic population.

We have seen that the predicted impact of an intervention, along with which intervention is predicted to be better is keenly dependent on both the assumed contact structure of the population and the predicted size of the epidemic in the case where no control measures have been enforced.

We note that the use of full workplace closure as an intervention is unrealistic, as this would be impossible in reality. There are also knock-on effects, as this will likely increase the number of household contacts and significantly alter the overall dynamics of the epidemic. This was used purely as a way to get insight on the different models used and the populations considered, as the interaction of the epidemic with this intervention, along with the impact of this intervention in comparison to vaccination is a useful way of seeing the differences caused by modelling technique and the contact structure of the population.

In conclusion, this divide between what is predicted for the POLYMOD and UK contact surveys and for the synthetic population is an issue that needs to be addressed. The surveys conducted are seemingly sensible ways to get at the contact structure, and the method used to construct the synthetic population is also (arguably) sensible, as several large datasets are combined in ways which satisfy many marginal statistics, along with the use of statistical techniques designed to use time use data in an informative way (CART). By considering the WAIFW matrices, along with the degree distribution, it is clear that there is much more heterogeneity included in the synthetic population, and whether this is an artefact of the methods used, or it is something which is in the population, but not sampled by the contact surveys [Mossong et al., 2008; Danon et al., 2012] is unclear, and merits additional investigation.

Chapter 7

Final Discussion

Throughout this thesis, the spread of epidemics on networks and in heterogeneous populations has been considered. In general, the main investigation has been on the impact of heterogeneity in a population on the spread of an epidemic, be that in populations that can be described by a contact network, as in §3, §5, §6.2, §6.3 and §6.5, or in populations where the spread of the epidemic is similar to that seen using mean-field dynamics, as in §4 and §6.4.

In §3, the focus was on the early exponential growth period of a stochastic epidemic on a heterogeneous network. The level of variance in the number of infected individuals which can be seen during this early period was calculated analytically. This has been done for the SIS model previously [Dangerfield et al., 2009], but has not been considered for the SIR model before. It was shown that the variance was dependent on the first three moments of the degree distribution, along with the transmission and recovery parameters. This used a density-dependent diffusion approximation to the pairwise equations describing the disease dynamics which exploited an assumption about the neighbourhood of susceptible individuals. In addition an argument was made about how to correctly think of the neighbourhood of infected individuals in this period, which was shown to be consistent with two constraints that were enforced by the exact but unclosed differential equations describing this spread.

Next, in §4, the distribution of workplace sizes for the UK was considered, which included the analysis of a novel dataset [Bluesheep data source]. The fit of three heavy-tailed distributions to this dataset was considered using numerous criteria, likelihood and distance-based, along with the size of predicted epidemics over these distributions when compared to the true distribution. It was shown that, of these distributions, the one which fit the best, has previously been put forward as a good fit to workplaces in the US. This implies that this may be a likely candidate to the distribution of workplace sizes in other countries too, at least in similarly developed countries. It was also shown that the impact of changing the workplace size distribution can be to increase the expected number of infected people considerably, but via consideration of the secondary attack rate in the locations, that the time to do so was long enough that effective control strategies could be implemented to limit the spread.

In §6.2, the construction of a synthetic population describing England and Wales was discussed. This has been done for other countries, such as the USA [Barrett et al., 2005; Eubank et al., 2004] and Italy [Iozzi et al., 2010], but this is the first time that a synthetic

population for England and Wales base been constructed. Several datasets from the Office for National Statistics, and data detailing the time use of individuals, were used in an attempt to approximate the contact network of this large area in a realistic and accurate way. Once constructed, this model was used to simulate the spread of epidemics on, along with the impact of vaccination and the closure of workplaces on this spread. The aim here was to model the spread of an epidemic through a realistic population for England and Wales, to gain insight into the effectiveness of control methods, as has been done previously in synthetic populations in the USA and Thailand [Eubank et al., 2004; Ferguson et al., 2006]. However, the size and complexity of this model meant that it was difficult to truly understand the results that were seen, and there was no way to check whether our results were sensible.

We therefore used simpler models to compare existing survey based contact structures for the UK which were given by POLYMOD and the UK contact surveys [Mossong et al., 2008; Danon et al., 2012]. The degree distribution and WAIFW matrix which made up the synthetic population were compared to the POLYMOD and UK contact surveys using simpler models; pairwise approximation to a network, spread using the WAIFW and a meta-population type model. This demonstrated the difference resulting from deriving contact structures by asking people about their interactions, to using available datasets in order to construct a contact network. In general, we observe that the populations derived from surveys were much more similar to each other than the synthetic population. This implies that there is some qualitative difference in the contact structures created by the synthetic population than for the structures we get from the surveys.

Also of consideration was the necessity of using the extremely large and complicated synthetic population model in order to model the spread of infection in this population. We saw that using these simpler models, we produce results which are qualitatively different, especially in the impact of intervention of the spread of epidemics. This implies that there is some gain in information from using the full synthetic population (not unexpectedly), though the use of simpler models allows us to explain results that we observe, whilst for the synthetic population, the results are less transparent.

Additionally, the POLYMOD and UK contact surveys have never been compared to each other in terms of the disease dynamics that are implied by their structures, and this thesis was able to show that they are qualitatively similar to each other. There has also been little comparison between contact structures of synthetic populations and structures from other data sources. The Little Italy study included a comparison between the Little Italy model and POLYMOD, but this synthetic population is very small and non-representative of the population. Seeing that there is a large difference in these contact structures gives strong support to further investigation of the differences between these types of contact structures. Additionally, it implies that more thought and study is needed in order to construct synthetic populations which are representative of the population that they purport to represent.

In summation, the level of heterogeneity assumed to be present in the contact structures of our population, along with the amount of detail included in our model has a great impact on the predicted spread of an epidemic, and the predicted efficacy of potential control strategies.

It has been shown that the different models that we consider and the different assumptions that we make, can have a large impact on the spread of an epidemic through a population. In addition to this, there are many other factors in the spread of an infectious disease

which are unknown and haven't been considered here. These include how different levels of interaction, such as touching compared with talking face-to-face with someone will have on the likelihood of spreading a disease, or how likely it is that someone can become infected from touching a door handle which has previously been used by someone with a disease.

A significant problem with trying to consider the impact of the contact structure of a population on the spread of the epidemic, is that there are these unknown factors taking place continuously and they are very difficult to include in model, let alone parameterise accurately. However, it may be that the impact of these un-modelled interactions, will average out over a large enough population, and therefore a stochastic model will be able to model the overall spread 'correctly' when we average the results. Whether this is the case is unknown, but there is currently little that can be included in a model that can target these questions, which wouldn't just be adding arbitrarily derived noise to the model. Additionally, whether these will have an impact which is as large as the errors contained in the contact structures used to represent the true contact structure is possible, though unlikely.

Given this, we can ask the question of what we are gaining by even considering the impact of the contact structure, and why we don't use a simple compartment model, with few divisions between different types of people (such as including ages)? This is compounded by the difficulty of being sure that what we do include in the model is accurate, as can be seen by the disparity between the synthetic population and the POLYMOD and UK contact surveys. The most obvious and potentially motivating answer to this question, is that these simple compartmental models are not telling us the whole story in terms of what is occurring during the spread of an infectious disease, as people are treated as being identical in these models, and it is clear that heterogeneity exists within populations in many forms. Therefore, we must try to investigate factors that have an impact on the epidemic, and that we can at least get a handle on, in an attempt to get closer to the truth. We can also attempt to target intervention strategies more sensibly with the increase in knowledge that we gain from these type of models, especially as they become more accurate due to increased attempts to survey populations, which isn't possible in standard compartmental models.

Continuing this line of reasoning, one can argue that the use of individual-based models should be encouraged. However, given our current understanding of the interactions that people have with each other and the transmission rates between people, this is likely to cause us to make many explicit and implicit assumptions about our population. We have seen in this thesis that the interpretation of such models can also be difficult, and that it is unlikely that we will have agreements between these type of models and those based on (arguably) more accurate data. Until there is a large increase in the availability of micro-level data relating to peoples interactions (if access to this type of data ever arrives), along with computing power which can handle the large amounts of data needed to perform many simulations on this type of model, then one can argue that there is too much in these models which is unknown to make them trustworthy enough to be truly useful to policy makers.

In order to improve our ability to predict and control diseases, the increase in our knowledge relating to the contact structures involved is of key importance. The use of phylogenetic data, in order to identify infections pathways [Volz and Frost, 2013], will hopefully help us to achieve this, though a lot of work is needed in order to make this a realistic

and helpful tool for modeling [Frost and Volz, 2013]. Along with this, increasing the use of surveys to assess peoples contact structure, will help to increase the accuracy of these data-rich models.

This thesis adds to the understanding of heterogeneities in transmission processes and the subsequent tailoring of intervention methods to combat the spread of infectious diseases.

Bibliography

Center for disease control and prevention. `http://www.cdc.gov/excite/library/glossary.htm`. Accessed: 23/04/2014.

2001 census. `http://www.ons.gov.uk/ons/guide-method/census/census-2001/index.html`. Accessed: 05/05/2014.

H Abbey. An examination of the reed-frost theory of epidemics. *Hum Biol.*, 3(24):201–233, 1952.

B Adams and D D Kapan. Man bites mosquito: Understanding the contribution of human movement to vector-borne disease dynamics. *PLoS ONE*, 4(8), 08 2009.

H Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

L J S Allen, F Brauer, P Van den Driessche, and J Wu. *Mathematical epidemiology*, volume 1945. Springer, 2008.

D Alonso, A J McKane, and M Pascual. Stochastic amplification in epidemics. *Journal of the Royal Society Interface*, 4(14):575–582, 2007.

R M Anderson and R M May. Epidemiological parameters of hiv transmission. *Nature*, (333), 1988.

R M Anderson and R M May. Infectious Diseases of Humans; Dynamics and Control. *Epidemiology and Infection*, 108(1), 1992.

H Andersson and T Britton. *Stochastic Epidemic Models and Their Statistical Analysis*, volume 151 of *Springer Lectures Notes in Statistics*. Springer, Berlin, 2000.

J Arino and P van den Driessche. A multi-city epidemic model. *Math. Popul. Stud.*, 10: 175–193, 2003.

M Baguelin, A J Van Hoek, M Jit, S Flasche, P J White, and W J Edmunds. Vaccination against pandemic influenza A/H1N1v in England: A real-time economic evaluation. *Vaccine*, 28(12):2370–2384, Mar 2010.

N T J Bailey. *The mathematical theory of epidemics*. London, UK: Griffin, 1957.

N T J Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. London, UK: Griffin, 1975.

F Ball and P Neal. Network epidemic models with two levels of mixing. *Mathematical Biosciences*, 212(1):69–87, March 2008. ISSN 0025-5564. doi: 10.1016/j.mbs.2008.01.001.

S Bansal, B T Grenfell, and L A Meyers. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface*, 4 (16):879–91, October 2007. ISSN 1742-5689. doi: 10.1098/rsif.2007.1100.

C Barrett, K Bisset, S Eubank, X Feng, and M Marathe. Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, 2008.

C L Barrett, S G Eubank, and J P Smith. If smallpox strikes portland ... *Scientific American*, 292, 2005.

M Barthélemy, A Barrat, R Pastor-Satorras, and A Vespignani. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Phys. Rev. Lett*, 92(17), 2004.

M. S. Bartlett. Measles periodicity and community size. *Journal of the Royal Statistical Society: Series A (General)*, 120, 1957.

D Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole. *Mémoires de Mathématiques et de Physique de l'Académie Royale des Sciences, Paris*, 1, 1766.

K Bisset and M Marathe. A cyber-environment to support pandemic planning and response. *DOE SciDAC Magazine*, 13:36–47, 2009.

O N Bjørnstad, B F Finkenstädt, and B T Grenfell. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series sir model. *Ecological Monographs*, (2):169–184, May 2002.

A Black, A McKane, A Nunes, and A Parisi. Stochastic fluctuations in the susceptible-infective-recovered model with distributed infectious periods. *Physical Review E*, 80(2): 21922, 2009. ISSN 15393755. doi: 10.1103/PhysRevE.80.021922.

Bluesheep data source. `http://bluesheep.bluegroupinc.com/products/ukbu`. Accessed: 05/05/2014.

M Boguñá, R Pastor-Satorras, and A Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.*, 90:028701, Jan 2003. doi: 10.1103/PhysRevLett.90.028701. URL `http://link.aps.org/doi/10.1103/PhysRevLett.90.028701`.

T Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

J. H. Brown and A. Kodric-Brown. Turnover rates in insular biogeography: effect of immigration and extinction. 1997.

C Cattuto, W Van den Broeck, A Barrat, V Colizza, J-F Pinton, and A Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS One*, 5(7):e11596, 07 2010.

S Chatterjee and R Durrett. Contact processes on random graphs with power law degree distributions have critical value 0. *The Annals of Probability*, 37(6):2332–2356, 11 2009. doi: 10.1214/09-AOP471.

CINET: Cyber-Infrastructure for Network Science. GaLib. http://cinet.vbi.vt.edu/cinet_new/content/galib.

D Clancy, P D O'Neill, and P K Pollett. Approximations for the long-term behavior of an open-population epidemic model. *Methodology And Computing In Applied Probability*, 3:75–95, 2001.

A Clauset and Y Virkar. Power-law distributions in binned empirical data, 2012. URL http://tuvalu.santafe.edu/~aaronc/powerlaws/bins/.

A Clauset, C Shalizi, and M Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. doi: 10.1137/070710111.

F Clementi and M Gallegati. Power law tails in the italian personal income distribution. *Physica A: Statistical Mechanics and its Applications*, 350(2–4):427 – 438, 2005. doi: http://dx.doi.org/10.1016/j.physa.2004.11.038.

Code-Point Open. http://www.ordnancesurvey.co.uk/business-and-government/products/code-point-open.html. Accessed: 20/11/2014.

V Colizza and A Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *J. T. Bio.*, 251:450–467, 2008.

V Colizza, A Barrat, M Barthélemy, and A Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS*, 103(14), 2006.

P Crépey, F P Alvarez, and M Barthélemy. Epidemic variability in complex networks. *Phys. Rev. E*, 73, 2006.

M E Crovella, M S Taqqu, and A Bestavros. Heavy-Tailed Probability Distributions in the World Wide Web. In *In A Practical Guide To Heavy Tails, chapter 1*, volume 1, pages 3–26, 1998.

C E Dangerfield, J V Ross, and M J Keeling. Integrating stochasticity and network structure into an epidemic model. *Journal of the Royal Society Interface*, 6(38):761–74, September 2009. ISSN 1742-5662. doi: 10.1098/rsif.2008.0410.

L Danon, A P Ford, T House, C P Jewell, M J Keeling, G O Roberts, J V Ross, and M C Vernon. Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on Infectious Diseases*, 2011:1–28, 2011.

L Danon, T A House, J M Read, and M J Keeling. Social encounter networks: collective properties and disease transmission. *J. R. Soc. Interface*, (9):2826–2833, 2012. doi: 10.1098/rsif.2012.0357.

L Danon, J M Read, T House, M C Vernon, and M J Keeling. Social encounter networks: characterizing Great Britain. *Proceedings of the Royal Society B*, 280(1765):20131037, 2013.

L Decreusefond, J-S Dhersin, P Moyal, and V Chi Tran. Large graph limit for a SIR process in random network with heterogeneous connectivity. *Annals of Applied probability*, 22 (2):541–575, 2012.

O Diekmann and J A Heesterbeek. *Mathematical epidemiology of infectious diseases*. John Wiley & Sons Ltd., 2000. ISBN 0471492418.

K Dietz. Epidemics and rumours: A survey. *Journal of the Royal Statistical Society. Series A*, 130(4), 1967.

R Durrett. *Random Graph Dynamics*. Cambridge University Press, 2007.

K T D Eames and M J Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *PNAS*, 99(20):13330–13335, Jan 2002.

K T D Eames and M J Keeling. Contact tracing and disease control. *Proc. R. Soc. B*, (270), 2003.

K T D Eames, J M Read, and W J Edmunds. Epidemic prediction and control in weighted networks. *Epidemics*, 1(1):70–76, 2009. doi: 10.1098/rsif.2008.0013.

B Erens, S McManus, and J Field. *National Survey of Sexual Attitudes and Lifestyles. II: Technical Report*. London: National Centre for Social Research,, 2001.

S N Ethier and T G Kurtz. *Markov Processes: Characterization and Convergence (Wiley Series in Probability and Statistics)*. Wiley, 1986.

S Eubank, H Guclu, V S A Kumar, M V Marathe, A Srinivasan, Z Toroczkai, and N Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

S Eubank, V S A Kumar, M V Marathe, A Srinivasan, and N Wang. Structure of social contact networks and their impact on epidemics. *AMS-DIMACS Special Volume on Epidemiology*, 4(8), 08 2006.

N M Ferguson, C A Donnelly, and R M Anderson. The foot-and-mouth epidemic in great britain: pattern of spread and impact of interventions. *Science*, 5519, 2001a.

N M Ferguson, C A Donnelly, and R M Anderson. Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain. *Nature*, 413, 2001b.

N M Ferguson, M J Keeling, W J Edmunds, R Gant, B T Grenfell, R M Amderson, and S Leach. Planning for smallpox outbreaks. *Nature*, 425(6959):681–685, Jan 2003.

N M Ferguson, D A T Cummings, S Cauchemez, C Fraser, S Riley, A Meeyai, S Iamsirithaworn, and D S Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437, 2005.

N M Ferguson, D A T Cummings, C Fraser, J C Cajka, P C Cooley, and D S Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442:448–452, 2006.

A N J Fish, D V I Fairweather, J D Oriel, and G L Ridgway. Chlamydia trachomatis infection in a gynaecology clinic population: identification of high-risk groups and the value of contact tracing. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 31, 1989.

S D W Frost and Erik M Volz. Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 2013.

G R Fulford, M G Roberts, and J A P Heesterbeek. The metapopulation dynamics of an infectious disease: Tuberculosis in possums. *Theoretical Population Biology*, 61(1):15 – 29, 2002.

A P Galvani and R M May. Epidemiology: Dimensions of superspreading. *Nature*, 438 (1207):293–295, 2005.

Geoportal. `https://geoportal.statistics.gov.uk/geoportal/catalog/main/home.page`. Accessed: 20/11/2014.

L Gilbert, R Norman, K M Laurenson, H W Reid, and P J Hudson. Disease persistence and apparent competition in a three-host community: an empirical and analytical study of large-scale, wild populations. *Journal of Animal Ecology*, 70(6):1053–1061, 2001.

D T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. ISSN 0022-3654. doi: 10.1021/j100540a008.

D T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, 115(1716), 2001.

M C Gonzalez, C A Hidalgo, and A-L Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. doi: 10.1038/nature06958.

N J Gotelli. Metapopulation models: The rescue effect, the propagule rain, and the core-satellite hypothesis. *American Naturalist*, 138(3), 1991.

M Graham and T A House. Dynamics of stochastic epidemics on heterogeneous networks. *J Math. Biol.*, 68(7), 2013.

F Granath, J Giesecke, G Scalia-Tomba, K Ramstedt, and L Forssman. Estimation of a preference matrix for women's choice of male sexual partner according to rate of partner change, using partner notification data. *Mathematical Biosciences*, 107(2):341 – 348, 1991.

J Gómez-Gardeñes, V Latora, Y Moreno, and E Profumo. Spreading of sexually transmitted diseases in heterosexual populations. *PNAS*, 5(105), 2008.

H M Götz, G van Doornum, H G M Niesters, J G den Hollander, H B Thio, and O de Zwart. A cluster of acute hepatitis c virus infection among men who have sex with men – results from contact tracing and public health implications. *Aids*, 19, 2005.

C B Hall, R G Douglas Jr, J M Geiman, and M P Meagher. Viral shedding patterns of children with influenza b infection. *J Infect Dis*, 4(140):610–613, 1979.

H H Handsfield, R J Rice, M C Roberts, and K K Holmes. Localized outbreak of penicillinase-producing neisseria gonorrhoeae. paradigm for introduction and spread of gonorrhea in a community. *JAMA*, 16(261), 1989.

I Hanski. Spatially realistic theory of metapopulation ecology. *Naturwissenschaften*, 88 (9):372–381, 2001.

D D Heckathorn. Respondent-driven sampling: A new approach to the study of hidden population. *Social Problems*, 44(2), 1997.

J Herbert and V Isham. Stochastic host-parasite interaction models. *Journal of Mathematical Biology*, 40(4):343–371, 2000. doi: 10.1007/s002850050184.

H W Hethcote and D W Tudor. Integral equation models for endemic infectious diseases. *Journal of Mathematical Biology*, 9(1):37–47, 1980. ISSN 0303-6812. doi: 10.1007/BF00276034.

H W Hethcote and J A Yorke. *Gonorrhea Transmission Dynamics and Control*. Lecture notes in biomathematics. Springer-Verlag, 1984. ISBN 9780387138701. URL `http://books.google.co.id/books?id=YfHVMQEACAAJ`.

T House and M J Keeling. Household structure and infectious disease transmission. *Epidemiology and Infection*, 137:654–661, 2009. doi: 10.1017/S0950268808001416.

T House and M J Keeling. The impact of contact tracing in clustered populations. *PLoS Comput Biol*, 6(3), 2010. doi: 10.1371/journal.pcbi.1000721.

T House and M J Keeling. Epidemic prediction and control in clustered populations. *J. T. Bio*, 272(1), 2011a. doi: 10.1016/j.jtbi.2010.12.009.

T House and M J Keeling. Insights from unifying modern approximations to infections on networks. *Journal of the Royal Society Interface*, 8(54):67–73, January 2011b. ISSN 1742-5662. doi: 10.1098/rsif.2010.0179.

T House, J V Ross, and D Sirl. How big is an outbreak likely to be? methods for epidemic final-size calculation. *Proc. R. Soc. A*, 469(2150), 2012. doi: http://dx.doi.org/10.1098/rspa.2012.0436.

InFuse. `http://infuse.mimas.ac.uk/`. Accessed: 20/11/2014.

F Iozzi, F Trusiano, M Chinazzi, F C Billari, E Zagheni, S Merler, M Ajelli, E D Fava, and P Manfredi. Little italy: An agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological data. *PLoS Comput Biol*, 6(12):700–721, 2010.

V Isham. Stochastic models of host-macroparasite interaction. *The Annals of Applied Probability*, 5(3):720–740, 1995.

ISIS VBI. `http://ndssl.vbi.vt.edu/apps/isis/`. Accessed: 04/08/2014.

ISTAT homepage. `http://www.istat.it`. Accessed: 05/05/2014.

C Kamp. Untangling the interplay between epidemic spread and transmission network dynamics. *PLoS Computational Biology*, 6(11):e1000984, Nov 2010.

M Keeling. The implications of network structure for epidemic dynamics. *Theoretical population biology*, 67(1):1–8, 2005.

M J Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B*, 266(1421):859–67, 1999. doi: 10.1098/rspb.1999.0716.

M J Keeling. J. theo. biol. *Bulletin of Mathematical Biology*, 205, 2000.

M J Keeling and C A Gilligan. Metapopulation dynamics of bubonic plague. *Nature*, 407, 2000.

M J Keeling and P Rohani. *Modeling Infectious Diseases*. Princeton University Press, 2008.

M J Keeling and J V Ross. On methods for studying stochastic disease dynamics. *J. R. Soc. Interface*, 5, 2008.

W O Kermack and A G McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115(772):700–721, 1927.

B Killingley, J E Enstone, J Greatorex, A S Gilbert, R Lambkin-Williams, S Cauchemez, J M Katz, R Booy, A Hayward, J Oxford, C B Bridges, N M Ferguson, and J S N Van-Tam. Use of a human influenza challenge model to assess person-to-person transmission: Proof-of-concept study. *Journal of Infectious Diseases*, 205(1):35–43, 2012.

I Z Kiss, D M Green, and R R Kao. Disease contact tracing in random and clustered networks. *Proc. R. Soc. B*, (272), 2005.

I Z Kiss, D M Green, and R R Kao. The network of sheep movements within great britain: network properties and their implications for infectious disease spread. *J. R. Soc. Interface*, (3), 2006a.

I Z Kiss, D M Green, and R R Kao. The effect of contact heterogeneity and multiple routes of transmission on final epidemic size. *Mathematical Biosciences*, 203(1):124–36, September 2006b. ISSN 0025-5564. doi: 10.1016/j.mbs.2006.03.002.

A S Klovdahl. Social networks and the spread of infectious diseases: the aids example. *Social science & medicine*, 21:1203–1216, 1985.

I Krishnarajah, A Cook, G Marion, and G Gibson. Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology*, 67, 2005.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

T G Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970. ISSN 00219002.

T G Kurtz. Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes. *Journal of Applied Probability*, 8(2):344–356, 1971. ISSN 00219002.

E O Laumann and Y M Youm. Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the united states: A network explanation. *Sexually Transmitted Diseases*, 26(5), 1999.

N Lee, P K Chan, D S Hui, T H Rainer, E Wong, K W Choi, G C Lui, B C Wong BC, R Y Wong, W Y Lam, I M Chu, R W Lai, C S Cockram, and J J Sung. Viral loads and duration of viral shedding in adult patients hospitalized with influenza. *J Infect Dis*, (200):492–500, 2009.

J Lessler, D A T Cummings, J M Read, S Wang, H Zhu, G J Smith, Y Guan, C Q Jiang, and S Riley. Location-specific patterns of exposure to recent pre-pandemic strains of influenza a in southern china. *Nat Commun.*, 2(423), 2011. doi: 10.1038/ncomms1432.

G E Leventhal, R Kouyos, T Stadler, V Von Wyl, S Yerly, J Böni, C Cellerai, T Klimkait, H F Günthard, and S Bonhoeffer. Inferring epidemic contact structure from phylogenetic trees. *PLoS computational biology*, 8(3), 2012.

B Lewis, S Eubank, A M Abrams, and K P Kleinman. In silico surveillance: evaluating outbreak detection with simulation models. *BMC Med. Inf. & Decision Making*, 13, 2013.

M Y Li and J S Muldowney. Global stability for the seir model in epidemiology. *Mathematical Biosciences*, 125, 1995.

M Y Li, J R Graef, L Wang, and J Karsai. Global dynamics of a seir model with varying total population size. *Mathematical biosciences*, 160(2):191–213, 1999.

F Liljeros, C R Edling, L A Amaral, H E Stanley, and Y Aberg. The web of human sexual contacts. *Nature*, 411(6959):907–908, June 2001.

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

AL Lloyd. Estimating variability in models for recurrent epidemics: assessing the use of moment closure techniques. *Theor. Pop. Biol.*, 65:49?65, 2004.

J O Lloyd-Smith, S J Schreiber, P E Kopp, and W M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438, 2005.

I M Longini. The generalized discrete-time epidemic model with immunity: a synthesis. *Mathematical Biosciences*, 82(1):19 – 41, 1986. doi: http://dx.doi.org/10.1016/0025-5564(86)90003-9.

R H MacArthur and E O Wilson. *The Theory of Island Biogeography*. Princeton: Princeton University Press, 1967.

A Machens, F Gesualdo, C Rizzo, A E Tozzi, A Barrat, and C Cattuto. A contribution to the mathematical theory of epidemics. *BMC Infectious Diseases*, 13(185):700–721, 2013.

G Magiorkinis, V Sypsa, E Magiorkinis D Paraskevis, A Katsoulidou andR Belshaw, C Fraser, O G Pybus, and A Hatzakis. Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. *PLoS Comput Biol*, 9(1), 2013.

N D Martinez. Artifacts or attributes? effects of resolution on the little rock lake food web. *Ecological Monographs*, 61(4), 1991.

R M May and R M Anderson. Transmission dynamics of hiv infection. *Nature*, 326: 137–142, 1987.

R M May and A L Lloyd. Infection dynamics on scale-free networks. *Physical Review E*, 64(066112), 2001.

N McCreesh, S D Frost, J Seeley, J Katongole, M N Tarsh, R Ndunguse, F Jichi, N L Lunel, D Maher, L G Johnston, P Sonnenberg, A J Copas, R J Hayes, and R G White. valuation of respondent-driven sampling. *Epidemiology*, 23(1), 2012.

L McKusick, W Horstman, and T J Coates. Aids and sexual behavior reported by gay men in san francisco. *Am J Public Health*, 5(75), 1985.

C J E Metcalf, J Lessler, P Klepac, F Cutts, and B T Grenfell. Impact of birth rate, seasonality and transmission rate on minimum levels of coverage needed for rubella vaccination. *Epidemiol. Infect.*, 140(12):2290–301, 2012.

C J E Metcalf, K Hampson, Tatem A J, B T Grenfell, and O N Bjørnstad. Persistence in epidemic metapopulations: Quantifying the rescue effects for measles, mumps, rubella and whooping cough. *PLoS ONE*, 8(9), 2013.

L A Meyers, B Pourbohloul, M E Newman, D M Skowronski, and R C Brunham. Network theory and sars: predicting outbreak diversity. *J T Bio*, 1(232):71–817, 2005.

S Milgram. The Small World Problem. *Psychology Today*, 1:61–67, 1967.

J C Miller. A note on a paper by Erik Volz: SIR dynamics in random networks. *J Math Biol.*, 62(3):349–358, Mar 2010.

J C Miller and E M Volz. Model hierarchies in edge-based compartmental modeling for infectious disease spread. *J Math Biol.*, 67(4), 2012.

H L Mills, C Colijn, P Vickerman, D Leslie, V Hope, and M Hickman. Respondent driven sampling and community structure in a population of injecting drug users, Bristol, UK. *Drug and Alcohol Dependence*, 126(3):324 – 332, 2012.

M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

M Molloy and B Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–179, 1995.

J Mossong, N Hens, M Jit, P Beutels, K Auranen, R Mikolajczyk, M Massari, S Salmaso, G Scalia Tomba, J Wallinga, J Heijne, M Sadkowska-Todys, M Rosinska, and J Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):381–391, Jan 2008.

D Moulay and Y Pigné. A metapopulation model for chikungunya including populations mobility on a large-scale network. *J. Theor. Biol.*, 318, 2013.

I Nåsell. Stochastic models of some endemic infections. *Mathematical Biosciences*, 2002.

I Nåsell. Moment closure and the stochastic logistic model. *Theor. Popul. Biol.*, 63, 2003.

M Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1), July 2002a. ISSN 1063-651X. doi: 10.1103/PhysRevE.66.016128.

M E J Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89, 2002b.

M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45 (2):167–256, 2003.

NOMIS. `https://www.nomisweb.co.uk/census/2011/postcode_headcounts_and_household_estimates`. Accessed: 20/11/2014.

N G van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 1992.

NPD. https://www.gov.uk/government/publications/national-pupil-database-user-guide-and-supporting-information. Accessed: 20/11/2014.

Office for National Statistics. 2001 Census: Commissioned Table C0844, a. ESRC/JISC Census Programme.

Office for National Statistics. Statistical wards, cas wards and st wards. http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/electoral-wards-divisions/statistical-wards--cas-wards-and-st-wards/index.html, b. Accessed: .

Office for National Statistics. 2001 Census: Special Licence Household Sample of Anonymised Records (SL-HSAR). http://discover.ukdataservice.ac.uk/catalogue?sn=5278, c.

R S Ostfeld and F Keesing. Biodiversity and disease risk: the case of lyme disease. *Conservation Biology*, 14:722?728, 2000.

W Otten, D J Bailey, and C A Gilligan. Empirical evidence of spatial thresholds to control invasion of fungal parasites and saprotrophs. *New Phytologist*, (163):125–132, 2004.

N Parikh, S Swarup, P E Stretz, C M Rivers, B L Lewis, M V Marathe, S G Eubank, C L Barrett, K Lum, and Y Chungbaek. Modeling human behavior in the aftermath of a hypothetical improvised nuclear detonation. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 949–956, 2013a.

N Parikh, M Youssef, S Swarup, and S Eubank. Modeling the effect of transient populations on epidemics in Washington DC. *Sci. Rep.*, 3, 2013b. doi: doi:10.1038/srep03152.

R Pastor-Satorras and A Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63(066117), 2001.

M Paunio, H Peltola, M Valle, I Davidkin, M Virtanen, and O P Heinonen. Explosive school-based measles outbreak. *American Journal of Epidemiology*, 148:1103–1110, 1998.

L Pellis, N M Ferguson, and C Fraser. Threshold parameters for a model of epidemic spread among households and workplaces. *J. R. Soc. Interface*, (6):979–987, 2008.

J J Potterat, R B Rothenberg, D E Woodhouse, J B Muth, C I Pratts, and J S Fogle. Gonorrhea as a social disease. *Sex Transm Dis*, 12(1), 1985.

D A Rand. Correlation Equations and Pair Approximations for Spatial Ecologies. *Advanced Ecological Theory*, 12(3-4):100–142, 1999.

J M Read, K T Eames, and W J Edmunds. Dynamic social networks and the implications for the spread of infectious disease. *J R Soc Interface*, 26(5):1001–1007, 2008. doi: 10.1098/rsif.2008.0013.

J M Read, J Lessler, S Riley, S Wang, L J Tan, K O Kwok, Y Guan, C Q Jiang, and D A T Cummings. Social mixing patterns in rural and urban areas of southern china. *Nat Commun.*, 9(2), 2011. doi: 10.1038/ncomms1432.

J M Read, J Lessler, S Riley, S Wang, L J Tan, K O Kwok, Y Guan, C Q Jiang, and D A T Cummings. Social mixing patterns in rural and urban areas of southern china. *Proc Biol Sci*, 281(1785):20140268, 2014. doi: 10.1098/rspb.2014.0268.

S Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4(2):131–134, 1998. ISSN 1434-6028. doi: 10.1007/s100510050359.

S Riley and N M Ferguson. Smallpox transmission and control: Spatial dynamics in great britain. *PNAS*, 103:12637–12642, 2006. doi: 10.1073/pnas.0510873103.

S Riley, C Fraser, C A Donnelly, A C Ghani, L J Abu-Raddad, A J Hedley, G M Leung, L-M Ho, T-H Lam, T Q Thach, P Chau, K-P Chan, S-V Lo, P-Y Leung, T Tsang, W Ho, K-H Lee, E M C Lau, N M Ferguson, and R M Anderson. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–6, Jun 2003.

T Rogers. Maximum-entropy moment-closure for stochastic systems on networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011.

P Rohani, X Zhong, and A A King. Contact network structure explains the changing epidemiology of pertussis. *Science*, 330:982–985, 2010.

J V Ross. A stochastic metapopulation model accounting for habitat dynamics. *J. M. Bio*, 52(6):788–806, 2006.

F C Santos, J F Rodrigues, and J M Pacheco. Epidemic spreading and cooperation dynamics on homogeneous small-world networks. *Physical Review E*, 72(5):056128, 2005.

K Sato, H Matsuda, and A Sasaki. Pathogen invasion and host extinction in lattice structured populations. *J. Math. Biol.*, (32):251–268, 1994.

L Sattenspiel and K Dietz. A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128(1):71 – 91, 1995.

A Schneeberger, C H Mercer, S A J Gregson, N M Ferguson, C A Nyamukapa, R M Anderson, AM Johnson, and G P Garnett. Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sexually Transmitted Diseases*, 31(6):380–7, Jun 2004.

T Sellke. On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Prob.*, (20):390–394, 1983.

P L Simon, M Taylor, and I Z Kiss. Exact epidemic models on graphs using graph-automorphism driven lumping. *Journal of Mathematical Biology*, 62(4):479–508, 2011. doi: 10.1007/s00285-010-0344-x.

S N Soffer and A Vázquez. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E*, 71:057101, 2005. doi: 10.1103/PhysRevE.71.057101.

R A Stein. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*, 15(8):510–513, 2011.

M Taylor, P L Simon, D M Green, T House, and I Z Kiss. From markovian to pairwise epidemic models and the performance of moment closure approximations. *Journal of mathematical biology*, 64(6):1021–1042, 2012.

M J Tildesley, N J Savill, D J Shaw, R Deardon, S P Brooks, M E J Woolhouse, B T Grenfell, and MJ Keeling. Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature*, 440(7080):83–6, Mar 2006.

Twins UK. `http://www.twinsuk.co.uk/twinstips/18/9934166/multiple-birth-statistics,-facts-&-trivia/multiple-birth-statistics-2010-(rel-2012)/`. Accessed: 30/01/2013.

UK Business. `http://www.ons.gov.uk/ons/rel/bus-register/uk-business/index.html`. Accessed: 20/11/2014.

UK Data Service. United Kingdom Time Use Survey, 2000. `http://discover.ukdataservice.ac.uk/series/?sn=2000054`.

UKDS. `http://discover.ukdataservice.ac.uk`. Accessed: 20/11/2014.

Y Virkar and A Clauset. Power-law distributions in binned empirical data. *arXiv*, 2012.

E Volz. SIR dynamics in random networks with heterogeneous connectivity. *J. M. Bio*, 56(3):293–310, 2008.

E M Volz and S D W Frost. Inferring the Source of Transmission with Phylogenetic Data. *PLoS Comput Biol*, 12(9), 2013.

E M Volz, S L K Pond, M J Ward, A J L Brown, and S D W Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.

J Wadsworth, J Field, A M Johnson, S Bradshaw, and K Wellings. Methodology of the national survey of sexual attitudes and lifestyles. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 156(3):pp. 407–421, 1993.

D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.

W Willinger and V Paxson. Where mathematics meets the internet. *Notices of the American Mathematical Society*, 45:961–970, 1998.

E B Wilson and M H Burke. The epidemic curve. *PNAS*, 28(9), 1942.

G-Q Zhang, S-Q Cheng, and G-Q Zhang. A universal assortativity measure for network analysis. *CoRR*, 2012.