Role of social network properties on the impact of direct contact epidemics

Author:
Badham, Jennifer Marette

Publication Date:
2008

DOI:
https://doi.org/10.26190/unsworks/18032

Role of social network properties on the impact of direct contact epidemics

Jennifer Marette Badham

A thesis submitted for the degree of
Doctor of Philosophy

School of Information Technology and Electrical Engineering
Australian Defence Force Academy

2008

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed _____

Date _____

Abstract

Epidemiological models are used to inform health policy on issues such as target vaccination levels, comparing quarantine options and estimating the eventual size of an epidemic. Models that incorporate some elements of the social network structure are used for diseases where close contact is required for transmission.

The motivation of this research is to extend epidemic models to include the relationship with a broader set of relevant real world network properties. The impact of degree distribution by itself is reasonably well understood, but studies with assortativity or clustering are limited and none examine their interaction.

To evaluate the impact of these properties, I simulate epidemics on networks with a range of property values. However, a suitable algorithm to generate the networks is not available in the literature. There are thus two research aspects: generating networks with relevant properties, and estimating the impact of social network structure on epidemic behaviour.

Firstly, I introduce a flexible network generation algorithm that can independently control degree distribution, clustering coefficient and degree assortativity. Results show that the algorithm is able to generate networks with properties that are close to those targeted.

Secondly, I fit models that account for the relationship between network properties and epidemic behaviour. Using results from a large number of epidemic simulations over networks with a range of properties, regression models are fitted to estimate the separate and joint effect of the identified social network properties on the probability of an epidemic occurring and the basic reproduction ratio. The latter is a key epidemic parameter that represents the number of people infected by a typical initial infected person in a population.

Results show that social network properties have a significant influence on epidemic behaviour within the property space investigated. Ignoring the

differences between social networks can lead to substantial errors when estimating the basic reproduction ratio from an epidemic and then applying the estimate to a different social network. In turn, these errors could lead to failure in public health programs that rely on such estimates.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

Most important in guiding me through the varied intellectual problems I encountered in this study have been my supervisors, Prof Hussein Abbass and Dr Rob Stocker. Prof Abbass has been working with me for the entire project and I am grateful for his intellectual curiosity and broad ranging interests, as well as his ability to help me focus on the key issues when required. Dr Stocker became involved partway through the project and brought with him a great deal of enthusiasm, as well as the capacity to ask the right question to help me understand issues at a deeper level.

I am also grateful to many colleagues for providing a range of practical assistance. Dr Alden Klovdahl, Dr Geoff Aldis and Dr Ruhul Sarker helped me access broader fields of study, making the journey easier. Many members of the Artificial Life and Adaptive Robotics Lab helped me to resolve programming problems, and I would particularly like to recognise Dr Yin Shan, Dr Ang Yang and Mr Kamran Shafi in this regard.

Finally, I would like to thank the Australian Defence Force Academy for providing scholarship funding to enable me to undertake this study.

Chapter 1: Introduction

An epidemic is "a sudden outbreak of infectious disease that spreads rapidly through the population, affecting a large proportion of people" (Martin 1994). Historians have recorded epidemics since at least 400 BCE, where a contemporaneous record noted the contribution of an epidemic in the fall of Athens to Sparta (Thucydides 431 BCE). Modern historians and anthropologists have studied the role of epidemics not only in conquest, but also in shaping society, for example through their role in labour shortages (popular accounts include McNeill 1976; Diamond 1998; Sherman 2006).

Quantitative methods were introduced to epidemiology with Bernoulli's comparison of smallpox immunisation techniques (1766). Snow's proof of the cause of cholera through identification of an infected water pump (Snow 1855) was the first use of systematic analysis for contagion control.

General mathematical models of the way in which diseases spread through the population were not developed until after 1900 (Hamer 1906; Kermack and McKendrick 1927). These were developed to examine key epidemiological questions such as:

- Under what conditions does an epidemic occur?
- What is the eventual size of a specific epidemic?
- How quickly is a specific epidemic expected to spread?

An understanding of epidemic behaviour sufficient to answer these questions is necessary in order to inform health policy on issues such as target vaccination levels and comparing quarantine options.

These early models made strong assumptions about social interaction that simplified the mathematics. The strongest is that all people have the same level of social contact and that contacts form uniformly at random. Notwithstanding this simplification, these models are very successful in

modelling airborne diseases such as measles, where only limited or incidental contact is required to transmit the infection.

More realistic assumptions about social interaction were introduced for models of sexually transmitted diseases and other diseases where transmission depends on close contact. In particular, variation in number of contacts was explicitly included in models developed for gonorrhoea (Lajmanovich and Yorke 1976) and HIV (Gupta et al. 1989). These studies demonstrated that modelling of social structure is necessary to explain some aspects of epidemic behaviour, such as a disease remaining present in a population despite very low average prevalence.

1.1 Research questions

Network theorists have defined a variety of properties in addition to the variation in number of contacts already described (referred to as the degree distribution), and calculated their values for many different real world networks (Newman 2003c).

Real world networks typically have a positively skewed degree distribution with a long tail of nodes with high degree, rather than the symmetric Poisson distribution that would be expected from random edge creation (Newman 2003c, Section III.C and Figure 6).

Further, social networks display some characteristic differences from other types of real world networks, such as links in web pages and other technological networks (Newman and Park 2003). In particular, they have positive degree assortativity and higher clustering coefficient than would be expected from the degree sequence (see Section 2.2.1 and glossary at Appendix A). There has been only limited investigation of the impact of assortativity and clustering on epidemic behaviour.

This observation sets the context for the **primary research question**:

> What is the relationship between epidemic behaviour and three key features of social networks: positively skewed degree

distribution, positive clustering coefficient and positive assortativity?

A key parameter in epidemic models is the basic reproduction ratio, denoted by R_0, which represents the expected number of infections directly generated by the first infection. It is related to both whether an epidemic occurs and, if so, the size of the epidemic.

Focussing the primary research question to concentrate on these aspects of epidemic behaviour leads to the more specific **secondary research questions 1 to 3**:

1) How does each of these properties affect epidemic occurrence?

2) How does each of these properties affect the basic reproduction ratio R_0?

3) Do these social network properties influence epidemic behaviour separately or jointly and, if the latter, how do they interact?

Simulation provides the greatest potential to evaluate the impact of the full range of property values and interactions. However, this method requires networks to be generated with positively skewed degree distribution, positive assortativity and positive clustering coefficient.

An algorithm to generate networks that control these three properties is not available in the literature. This leads to **secondary research question 4**:

4) How can networks be generated for simulations with various values of degree sequence, assortativity and clustering coefficient, separately and jointly?

1.2 Structure of the thesis

To enable investigation of these questions, Chapter 2 first sets out the relevant background from epidemiology and social network analysis (with a glossary at Appendix A). It also describes commonly used network generation algorithms and those that generate networks with some of the social properties sought. Several studies have investigated aspects of the relationship between social structure and epidemic behaviour with mathematical models from epidemiological and network perspectives, and with simulations. Chapter 2 concludes with a description of these studies and those results relevant to the research questions.

Chapter 3 assumes suitable networks can be generated and describes the experimental design. There are three aspects to the design: the properties of the networks, the epidemic properties of the simulation (infectivity, recovery and immunity probabilities), and practical aspects of the simulation such as the update process.

I introduce an algorithm in Chapter 4 that is able to generate networks, controlling the three properties of interest. This chapter identifies how the algorithm inputs relate to the properties of the generated networks and the algorithm is then used to generate the networks for simulation required by the experimental design.

These networks are used to simulate up to 100 epidemics for each combination of network property values and epidemic parameters. For each simulation, the results of interest are whether an epidemic occurred and, if so, the basic reproduction ratio implied by the size of the epidemic. Chapter 5 includes a detailed analysis of a selected simulation parameter set and summary results for all simulation sets.

This analysis includes fitting of regression models, which estimate the separate and joint effects on epidemic behaviour of the degree distribution, assortativity and clustering of the social networks over which the epidemic occurs. The computer code and other files used for the analysis are described

in Appendix B and included on the supplementary DVD. The DVD also contains the detailed results for all simulation sets, indexed at Appendix C.

The conclusions and implications of the research are presented in Chapter 6.

1.3 Summary of contributions

The first group of contributions concerns network generation:

- The major contribution in this group is design of a novel three phase approach to generate networks with separate control of degree sequence, degree assortativity and clustering coefficient (Section 4.2.1), responding to secondary research question 4.
- I implement this approach with a specific algorithm: one dimensional ring with stochastic edge swaps (Section 4.2.2). The algorithm is validated, and the relationship between input parameters and properties of generated networks is described (Section 4.3).
- In addition, an existing network generation algorithm is modified to generate networks with a given degree distribution that are connected (Section 4.1), to enable equivalent network implementation of the basic epidemiological model with its simple social structure assumptions.

The second group of contributions describes aspects of the relationship between social network properties and the behaviour of epidemics for the property space examined, based on simulation over networks generated with the new algorithm:

- The major contribution in this group is the first analysis of the effect of each of three network properties (positively skewed degree distribution, positive degree assortativity and positive clustering coefficient) in the presence of the other properties (Chapter 5), responding to the primary research question.
- An operational definition of epidemic for simulation studies is proposed (Section 5.1).

- I study the relationship between degree heterogeneity and epidemic occurrence in the presence of assortativity and clustering (Section 5.2.1), responding to secondary research question 1.
    - For low infectivity rates, the results provide further support for published studies that suggest epidemics are more likely to occur as degree variation increases in the absence of assortativity and clustering. However, this relationship interacts with infectivity, and for some higher infectivity rates, epidemics are less likely to occur as degree variation increases.
    - The results also provide evidence that the relationship between epidemic occurrence and degree heterogeneity in networks with positive assortativity and clustering has the same direction as in networks with zero values of those properties.
- I investigate the relationship between degree heterogeneity and epidemic size in the presence of assortativity and clustering (Section 5.2.2).
    - The established view that epidemics are smaller as degree variation increases is supported.
    - This same pattern generally occurs in networks that have positive assortativity or clustering, but the opposite pattern occurs for highly clustered networks where the epidemic has low infectivity.
- I study the relationship between epidemic occurrence, assortativity and clustering, while controlling for degree heterogeneity (Section 5.3), responding to secondary research question 1. The likelihood of an epidemic decreases for higher values of one or both of these properties, but the significant property differs across parameter sets.
- While controlling for degree heterogeneity, I study the relationship between assortativity and clustering, and basic reproduction ratio (R_0) as derived from epidemic final size where immunity is conferred (SIR

epidemics) (Section 5.4) or from equilibrium prevalence where there is no immunity (SIS epidemics) (Section 5.5.1), responding to secondary research questions 2 and 3. Results show:

- ♦ assortativity and clustering affect R_0 independently and linearly, with only limited evidence of interaction or nonlinearity;
- ♦ increases in either assortativity or clustering lead to lower values of R_0;
- ♦ the relative importance of assortativity and clustering differ between degree distribution types, infectivity levels and SIR or SIS epidemic type;
- ♦ the impact of assortativity and clustering on R_0 is substantial, with real world values reducing R_0 (as compared to its value for zero values of the network properties) by between 9% and 45% over the various simulation sets (Section 5.5.3);
- ♦ some evidence that SIR epidemics are more strongly affected by clustering and assortativity in the social network than SIS epidemics (Section 5.5.2).

- Based on the concept of secondary reproduction number proposed in (Eguíluz and Klemm 2002), an alternative method to include the impact of network social structure into epidemic models is proposed, accessible proportion of network in a specified number of steps (Section 5.6). Using this measure:
  - ♦ by itself - models are generally able to account for less of the variability in the value of epidemic derived R_0 than models based on assortativity and clustering coefficient;
  - ♦ in addition to assortativity and clustering coefficient – models have only limited additional explanatory power than models based on assortativity and clustering coefficient only.

Chapter 2: Literature Analysis

The role of social structure in epidemic behaviour can be studied from the perspective of three overlapping fields of study. Sociologists studying social networks have defined a variety of properties and calculated their values for many different real world social networks. Mathematicians and other physical scientists have studied dynamic processes, including epidemic spread, on idealised networks through mathematical techniques and by simulation. Finally, epidemiologists have incorporated elements of social structure in models of disease spread.

This chapter initially presents basic concepts and results from epidemiology and from social network theory. This is followed by a description of several published network generation algorithms, selected because they are commonly used for simulation of processes on networks, or they generate networks with some control over the network properties of interest. These basic concepts and algorithms each contribute to published results concerning the effect of the selected social properties on epidemic behaviour. These results are then described and gaps identified, particularly as they relate to the research questions.

2.1 Fundamentals of epidemic modelling

Simple epidemic models fall into two broad categories described by the available states for the population. A person who is available to be infected is referred to as susceptible (*S*). Once the disease is successfully transmitted to a person and that person is able to transmit the disease to other members of the population, that person is referred to as infected (*I*). At the end of the infectious period, the person either returns to the susceptible state or is removed (*R*) from the relevant population through immunity (or death). The model where immunity is conferred is referred to as an SIR epidemic. If there is no immunity, it is an SIS epidemic.

Other states are used in more complex models. For example, if there is a period between the time at which a person becomes infected and the time at which they can infect others, they are referred to as exposed (*E*) during that period. Models including this state are referred to as SEIR or SEIS. Only SIR and SIS epidemics are modelled in this study.

2.1.1 Basic epidemic model

The first mathematical models of person to person epidemic transmission included two key concepts, the mass action principle and threshold theory. The mass action principle (Hamer 1906) states that the spread of a disease is proportional to the rate of contact between infected and susceptible individuals. Implicit in this principle are two assumptions:

- The probability of transmission of infection is the same for all pairs of susceptible and infected persons; and
- Every susceptible individual has an equal probability of coming into contact with every infected individual.

Threshold theory arises from the system of differential equations now referred to as the Kermack-McKendrick model or basic epidemiological model. This model was developed to examine the reasons why epidemics die out.

> Two of the reasons commonly put forward … are (1) that the susceptible individuals have all been removed, and (2) that during the course of the epidemic the virulence of the causative organism has gradually decreased (Kermack and McKendrick 1927, pg 34).

As well as the mass action principle, the model was based on several assumptions:

- One (or more) infected persons introduced to community;
- Disease spread by contact;
- All members of the community are equally susceptible;

- Infectivity is a function of infection age (that is, time since infection);
- Removal (immunity or death) is a function of infection age;
- Permanent immunity is conferred by a single infection;
- Population is constant (with the exception of deaths caused by the epidemic) as the disease timeframe is short compared to demographic timeframe.

While Kermack and McKendrick were unable to develop a general solution to the equation system in the form of a function for the number of infected persons over time, they provided solutions for several special cases. One of these special cases is where the infection and removal functions are independent of infection age. By the assumptions of the model, infectivity and susceptibility are also independent of seasonality, characteristics of the person infected or exposed, or any other potentially relevant factor. This simplifies the equation system considerably and is the usual presentation of the model (for example, in Bailey 1975; Anderson 1991; Diekmann & Heesterbeek 2000).

The model is given by:

$$\begin{aligned} S + I + R &= N \\ \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \qquad (2.1)$$

where: N is total population
S is number susceptible
I is number infected
R is number removed (dead or immune)
β is infection transmission parameter
unit: per person per unit time
γ is removal (or recovery) rate
unit: per unit time

The infection transmission parameter (or contact rate or infectivity rate) is defined as:

> The proportion of total possible contacts between infectious cases and susceptibles that lead to new infections (Last 2001, pg 94).

The parameter incorporates two separate elements. The reference to "possible" contacts is important in this definition as one element is the mean contact rate between susceptible and infected persons in the population, or stochastically the mean probability that there is contact between any specified infective and any specified susceptible. The other element is the proportion of contacts that transmit the infection.

Under these assumptions, an epidemic occurs only when the population exceeds a critical ratio of the two rates:

$$N > \frac{\gamma}{\beta} \tag{2.2}$$

Further, an epidemic dies out because the number of susceptible people decreases, thereby decreasing the contact rate component of the infection transmission parameter. The critical threshold increases and, from the point of time where N is insufficient, the number of people being removed is higher than the number of new infections.

Population size is the somewhat unintuitive critical factor because the assumptions of the model have the number of contacts increasing as population size increases. One interpretation is that the population is contained in a limited space and the relevant disease is airborne, so changes in population numbers do lead to proportional changes in contact rates.

An alternative formulation of the model use S, I and R to denote the density of susceptible, infected and recovered persons instead of the number of people in each of these states. Similar notation is used, which can lead to confusion about the interpretation of parameters in equation system (2.1).

See the discussion in (Hethcote and van Ark 1987, pp 90-91) about the differences between the two formulations.

From the basic epidemiological model, neither reason previously put forward is required to explain the fact that an epidemic dies out before infecting an entire population. Kermack and McKendrick (1927) also found an approximate solution for the final size of the epidemic (total number of people infected), valid only near the threshold. They were also able to demonstrate that similar results exist in the more general situations of nonconstant infectivity and removal rates, and where transmission is through contact with an intermediate host.

As pointed out by Anderson (1991), while Kermack and McKendrick did not present their result in this way, it is fundamental to the modern concept of basic reproduction ratio (R_0). R_0 is defined (Diekmann et al. 1990, pp 365-6) as:

> the expected number of secondary cases produced, in a completely susceptible population, by a typical infected individual during its entire period of infectiousness.

Calculation of R_0 for a disease is of key concern to epidemiologists because the threshold theorem, in its modern form, states that an epidemic can occur if $R_0 > 1$. That is, an epidemic occurs if an infected person who has contact only with susceptible people is able to produce at least one other infected person on average. This is equivalent to equation (2.2) as $R_0 = \beta N/\gamma$ in this model.

2.1.2 Key results from the basic model

As well as the threshold theorem, R_0 is related to other key features of epidemic behaviour. For an SIR epidemic, the total proportion of the population ever infected is referred to as final size. It is given by the nonzero root f satisfying (Kendall 1956; Diekmann & Heesterbeek 2000, equation 1.11):

$$\log_e(1-f) = -f R_0 \tag{2.3}$$

For an SIS epidemic, final size is not relevant. Instead, equilibrium is reached, where a constant proportion of the population remains infected. This proportion is at the level where new infections are matched by infected persons recovering and becoming susceptible. Thus, ongoing prevalence p is given by (Anderson & May 1992, equation 2.1):

$$p = 1 - \frac{1}{R_0} \tag{2.4}$$

Also, by definition, R_0 provides the growth rate per generation in the initial stages of an epidemic, where the proportion of infected (or removed) persons is negligible. The growth rate per unit time will depend also on the recovery rate. In each generation, an infected person infects R_0 other persons and recovers themselves. Thus:

$$I_G = I_0^{G(R_0-1)} \tag{2.5}$$

where: G is number of generations
I_0 is number initially infected
I_G is number infected at generation G

and:

$$g_0 = R_0 - 1 \tag{2.6}$$

where: g_0 is initial growth rate per generation

These results are all based on a deterministic consideration of epidemic behaviour. Later researchers were also able to examine epidemics from a stochastic viewpoint. The key difference is that the number of people infected from each infection is treated as a probability distribution instead of the exact value R_0.

Where a significant proportion of the population is infected, the stochastic impact is small and the deterministic results apply. However, in the early stages of an epidemic, stochastic considerations give rise to a nonzero probability that an epidemic will not occur, despite a basic reproduction ratio of greater than one.

Following (Diekmann & Heesterbeek 2000, section 1.2.2), branching theory can be used to show that the probability of an epidemic going extinct before infecting a significant proportion of the population is given by the smallest root z satisfying:

$$z = \sum_{k=0}^{\infty} q_k z^k \tag{2.7}$$

where: q_k is the probability of an infected person infecting k persons

Under the assumptions of a fixed number of contacts c with probability of infection p given contact, and a completely susceptible population, the probability of an epidemic going extinct is given by z satisfying:

$$z = \sum_{k=0}^{c} \binom{c}{k} p^k (1-p)^{c-k} z^k \tag{2.8}$$

2.2 Fundamentals of social network theory

Graph Theory is a mathematical abstraction used to describe relationships between entities. More generally, a graph is a set of nodes and a set of edges between those nodes. The edges denote the existence and, in some cases, direction of the relationship of interest between the pair of nodes that it

connects. For a given set of nodes many different relationships can exist. Each relationship defines a different set of edges, and hence a different graph.

For example, if the nodes represent people, relationships could include living in the same residence, older than, likes, wearing the same coloured shirt etc. This study concerns the behaviour of human epidemics. In this situation, nodes represent people and the relationship of interest is social contact sufficient to transmit the disease in some undefined period of time. Clearly, diseases with different transmission modes (such as airborne, blood, sexual activity) would define different graphs over the same population.

The term ‘graph’ tends to be used in mathematical studies of theoretical results that apply to ensembles of graphs with specified properties (for example, in Bollobás 2001) and the term ‘network’ is used by sociologists investigating a real instance (for example, in Wasserman & Faust 1994) where nodes and edges have additional properties (such as age). This study will use the term 'network'.

Networks have been studied from various perspectives. For example, sociologists may be interested in fully mapping a specific example of a network and identifying a small set of behavioural rules that allow networks to evolve that are similar to the network of interest, or investigating properties of the network that influence the current social behaviour. Such a set of rules or properties can then be used to provide insight into how society operates (for a general reference, see Wasserman & Faust 1994). In contrast, the mathematical field of Graph Theory investigates general classes of abstract networks regardless of whether there are specific existing examples of the network class (for example, Bollobás 2001). Researchers from many other fields have also used networks as a way to represent and analyse various relationships. These include biological relationships such as food webs or the proteins involve in gene regulation, and constructed relationships such as power transmission grids and links between world wide web pages (for a comprehensive review, see Newman 2003c).

This study considers only social networks, as they provide the contact for person to person disease transmission in humans. Thus, only those aspects of network theory relevant to social networks will be considered.

2.2.1 Relevant network properties

This section defines the network properties that are to be compared for their impact on epidemic behaviour. There are many other properties defined, measured and interpreted in social network literature, some of which are likely relevant to epidemic behaviour. For example, the concept of "betweenness" emphasises that a person may be involved in greater or fewer of the paths between randomly selected people in the network. Excellent descriptions of network properties are included in (Wasserman & Faust 1994) and (Newman 2003c).

2.2.1.1 General network properties

For this study, the relationships of interest are between network properties and epidemic behaviour. In order to illuminate these relationships, several simplifying assumptions are made about the networks (discussed at the start of Chapter 4); they are undirected, static and unweighted. That is:

- the relationship denoted by an edge between arbitrary nodes *i* and *j* exists from node *i* to node *j* and from node *j* to node *i* (undirected);
- the set of edges (and nodes) does not change once the network is created (static); and
- all edges in the network transmit infection with equal probability (unweighted).

A network is referred to as connected if any node can be reached from any other node by following edges. For networks that are not connected, each section of the network in which all nodes can be reached from each other is referred to as a component. Thus, a connected network has only one component. In a disconnected network, the largest component is referred to as the giant component if it includes more than half the nodes.

In this study, edges denote epidemic transmission opportunities. Therefore, if a network is not connected, the disease cannot spread through the entire population and the size of the giant component constrains the potential size of the epidemic.

A network is referred to as simple if there are no multiple edges or self edges. That is, each edge connects two different nodes (rather than connects a node to itself) and each pair of nodes either does not have an edge between them or has exactly one edge between them. This is in contrast to a multigraph, which may contain multiple edges between the same pair of nodes. Self edges and multiple edges have no epidemic interpretation if an edge is defined by the opportunity to infect, so the networks of interest are simple.

2.2.1.2 Degree distribution

The degree (sometimes referred to as connectivity) of a node is the number of edges connected to it or, equivalently, the number of neighbours it has. The degree distribution is then the frequency distribution of different degrees across the nodes in the network.

There are several aspects of the degree distribution that are potentially relevant to epidemic behaviour. The most basic of these is mean degree, the average degree across all nodes in the network. It is given by:

$$\hat{k} = M/2N \tag{2.9}$$

where: M is number of (undirected) edges
N is number of nodes

The basic epidemiological model (Section 2.1.1) explicitly assumes that all members of the community are equally infectious and equally susceptible. This assumption requires all individuals to have the same rate of contact with all other individuals in the community. That is, all nodes have the same degree.

An alternative interpretation is that any variation in number of contacts is exactly balanced by a variation in intrinsic infectivity or susceptibility. While

this is mathematically equivalent, it is artificial and the previous interpretation is the standard.

Other features of the degree distribution that may be practically useful include the general form of the degree probability function, variance in the degree, and various measures of bias in the function such as skewness, entropy of the data sequence and indices concerning concentration of edges amongst particular nodes. Such summary measures provide a convenient way to compare degree distributions with a small set of parameters.

2.2.1.3 Clustering coefficient

Clustering is a measure of network transitivity, the extent to which neighbours of a node are neighbours of each other. There are two alternative approaches to calculating the clustering coefficient for the network.

Unless specifically noted, this study will use the definition of clustering coefficient given in (Watts and Strogatz 1998):

> Suppose that a vertex v has k_v neighbours; then at most $k_v(k_v - 1)/2$ edges can exist between them (this occurs when every neighbour of v is connected to every other neighbour of v). Let C_v denote the fraction of these allowable edges that actually exist. Define C (the clustering coefficient) as the average of C_v over all v.

That is, the local clustering coefficient is calculated for each vertex and the mean of these coefficients is used. By convention, a node with degree of 0 or 1 is considered to have local clustering coefficient of 0. Formally:

$$\begin{aligned} C &= \frac{1}{N}\sum_{v=1}^{N} C_v \\ C_v &= \begin{cases} 0 & \text{if } k_v = 0,1 \\ \dfrac{2\left|e_{ij} \text{ given } e_{vi} \text{ and } e_{vj} \text{ exist}\right|}{k_v\left(k_v - 1\right)} & \text{if } k_v \geq 2 \end{cases} \end{aligned} \tag{2.10}$$

where: C is clustering coefficient
C_v is local clustering coefficient for node v
v, i, j are nodes
N is the total number of nodes
e_{ij} indicates that an edge exists between nodes i and j
k_v is degree of node v

For the example at Figure 2-1, the network clustering coefficient would be calculated as the mean of 0.33 (for V), 1 (for a), 0 (for b), 1 (for c) and 1 (for d), for a network clustering coefficient of 0.67.

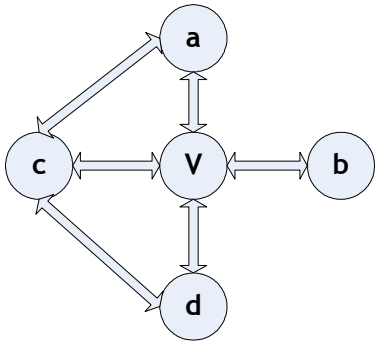

***Figure 2-1: Clustering coefficient example** : Vertex V has four neighbours, vertices a, b, c and d. Thus, there are 6 possible undirected edges between these pairs of neighbours: ab, ac, ad, bc, bd and cd. Of these, only two edges (ac and cd) exist, so the clustering coefficient for vertex V is 2/6.*

The other approach is to calculate the proportion of realised neighbour pair edges in the network (Holland and Leinhardt 1970), referred to in the social network literature as the fraction of transitive triples. For the example at Figure 2-1, this method would give a clustering coefficient of 0.55 (from (2+1+0+1+1)/(6+1+0+1+1))

2.2.1.4 Geodesic path lengths

In a connected graph, any node can be reached from any other node through a series of edges and nodes. The shortest path (or geodesic) between a pair of nodes is the smallest number of edges required to get from one node in the pair to the other. Two properties of interest are the mean and maximum

(diameter) of the shortest paths across all pairs of nodes, as they provide indicators of the potential speed of an epidemic across the network.

$$\begin{aligned} \text{Mean geodesic} &= \frac{1}{M}\sum_{i=1}^{N}\sum_{j=i+1}^{N} G_{ij} \\ \text{Diameter} &= Max\left(G_{ij}\right) \quad i,j \in \left[1,N\right] \end{aligned} \tag{2.11}$$

where: N is the total number of nodes
M is the total number of (undirected) edges
G_{ij} is the length of the shortest path (geodesic) between node i and node j

2.2.1.5 Assortativity

The term assortativity is used broadly in social network theory to describe preferential relationships. For example, the 'friendship' relation is more likely to exist between persons of similar age and/or similar interests than would be expected purely from the age and interest proportions of people in the population. However, unless specifically stated, assortativity is used in this study in the narrow sense of assortativity with respect to degree.

Newman (2002a) defines the assortativity coefficient of a network as:

> … simply the Pearson correlation coefficient of the degrees at either ends of an edge.

For an undirected network, each edge must be included twice in the correlation calculation, once in each direction. The assortativity of an undirected network can be calculated as follows:

$$r = \frac{M\sum_i j_i k_i - \left[\sum_i j_i\right]^2}{M\sum_i j_i^2 - \left[\sum_i j_i\right]^2} \tag{2.12}$$

where: j_i and k_i are the degrees of the nodes at the ends of edge i with $i = 1...M$
M is the number of edges

The network literature also uses the term degree correlation, but this obscures the fact that the correlation of interest is between the degrees at each end of the edges. As for correlation coefficients generally, assortativity has a value in the range [-1,1], with positive assortativity indicating that nodes preferentially link to nodes with similar degree.

2.2.2 Properties of social networks

Newman, together with occasional colleagues, has pursued a research thread collating and comparing published properties of social and other networks, classified as information, technological or biological (Newman 2002; Newman 2003a; Newman 2003c; Newman and Park 2003a). This research thread identifies properties relevant to social networks and investigates potential factors giving rise to these properties.

Table 2-1 displays the network properties for social networks where all information is available, extracted from (Newman 2003, Table IIc). All networks are undirected except for email address books.

Table 2-1: Summary of network properties for published social networks

Network	Nodes	Mean degree	Clustering*	Assortativity
Film actors	449 913	113.43	0.78	0.208
Company directors	7 673	14.44	0.88	0.276
Maths co-authorship	253 339	3.92	0.34	0.120
Physics co-authorship	52 909	9.27	0.56	0.363
Biology co-authorship	1 520 251	15.53	0.60	0.127
Email address books	16 881	3.38	0.13	0.092
Student relationships	573	1.66	0.001	-0.029

* *Values for both definitions of clustering coefficient are given in the original table. Only the value for C(2) is displayed here as that corresponds to the definition being used.*

From this table, it is clear that a broad range of values can exist for key properties of real world social networks. In general, social networks show positive assortativity and relatively high clustering. These conclusions should, however, be qualified by the observation that the social networks presented

are not the type of networks that would give rise to epidemic transmission. However, they are consistent with studies of social networks in which diseases are transmitted through sexual contact or needle sharing (Rothenberg 2003).

In contrast, other types of networks generally show negative degree assortativity and clustering that is consistent with the value expected if edges occurred between pairs of nodes selected uniformly at random but constrained by the degree distribution.

2.2.3 Property relationships

The requirement that the network be simple has been shown empirically to cause negative assortativity (Maslov et al. 2002; Park and Newman 2003) in networks with degree distributions that are highly positively skewed. The popular but unquantified explanation is that, if the highest degree nodes have degrees in the order of $\sqrt{N}$, the expected number of edges between some high degree nodes given random connections is greater than one. Because only one edge is formed, the excess edge(s) must instead be made with a lower degree node and the created network has negative assortativity.

For any specific degree sequence, a network constructed by randomly making edges until the target degree is met (Molloy and Reed 1995, described in Section 2.3.2.1) has an expected assortativity of zero (Newman 2002a). The same type of network has an expected (transitive triples) clustering coefficient given by (Newman and Park 2003):

$$C = \frac{1}{N}\frac{\left[\text{var}(k) - \hat{k}^2 - \hat{k}\right]^2}{\hat{k}^3} \qquad (2.13)$$

where: C is the transitive triples version of clustering coefficient
N is the number of nodes
var(k) is the variance of the degree distribution
$\hat{k}$ is mean degree

While there has been no systematic study of the relationship between assortativity and clustering in networks, some relevant results have been

reported. The internet shows much lower clustering than random networks with the same degree distribution, but significantly higher clustering than random networks constructed with the same degree distribution and assortativity (Maslov et al. 2002). The potential assortativity / clustering space was mapped for a specific degree sequence in (Holme and Zhao 2006, Figure 1a) and higher values for one property are strongly linked to higher values of the other.

2.3 Commonly used network generation algorithms

There are two broad approaches to network generation. One method uses detailed data to synthesise the population of interest and their activities so as to derive contact information. For example, the EpiSims project uses population census and traffic survey data (Eubank et al. 2004) to generate a large urban population for simulating the effect of different diseases and public health techniques.

Alternatively, there are several algorithms available that enable a network to be generated with specific values of some properties. This is the method to be used in this study as it enables variation in the network properties, which is essential to develop relationships between these properties and epidemic behaviour.

Several network generation algorithms have been widely used in the published literature. Networks generated with different algorithms have different properties.

Algorithms that generate graphs that are not simple are included, as post-hoc corrections can be made by simply deleting self-edges and duplicate edges. However, such corrections may change the characteristics of the algorithm. For example, nodes with higher degree are more likely to have self edges, so deleting these may reduce the range of achieved degrees.

It is important to recognise that algorithms to generate networks have been developed for many different purposes and that their suitability for epidemic

simulation does not necessarily impact on their suitability for analysis of other types of networks. This is because the properties of social networks can be very different from the properties of other types of networks (Newman and Park 2003).

2.3.1 Random graphs (Erdös-Rényi algorithms)

In their classic paper on random graphs, Erdös and Rényi (1960) presented two network generation processes. For both of these algorithms, the graphs generated are simple because of the construction rules, but not necessarily connected.

Their primary method creates graphs with exactly the specified number of nodes and edges (and hence mean degree) with equal probability from the set of all possible simple graphs with that number of nodes and edges. The algorithm is as follows:

1) Take a set of *N* nodes;

2) Form an edge between any pair of nodes, selecting with equal probability from all pairs that do not have an edge already between them;

3) Repeat edge creation until the desired number of edges (*M*) has been created.

The other method described, but not used, generates graphs with the required number of nodes, but mean degree is stochastic. The generation algorithm is as follows:

1) Take a set of *N* nodes;

2) For each pair of nodes (different from each other), create an edge between them with a given probability (*p*).

Various authors (including Erdös and Rényi 1960; Bollobás 2001) have demonstrated that large graphs generated using the two construction methods have similar properties. One objective of random graph theory is

> … to determine at what stage of the evolution a particular property of the graph is likely to arise (Bollobás 2001, pg xiii).

Consistent with this approach, Erdös and Rényi investigated the mathematical properties of graphs constructed using probabilistic techniques and the real world correspondence of their networks was not relevant. However, they anticipated future work by noting (pg 19):

> … if one aims at describing such a real situation, one should replace the hypothesis of equiprobability of all connections by some more realistic hypothesis.

For both methods, each node can have a possible *(N-1)* edges so the constant probability for an edge between any pair of nodes necessary to generate an expected *M* edges is given by:

$$p = \frac{M}{N(N-1)} \tag{2.14}$$

Thus, degree distribution is a binomial distribution with mean degree:

$$\hat{k} = 2p(N-1) \tag{2.15}$$

and probability of degree *k* given by:

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.16}$$

For large *N*, this can be approximated by the Poisson (for $p<<1$) or Gaussian (for *p* not close to 0 or 1) distributions.

As each edge is constructed independently of all other edges, the probability of an edge occurring between two nodes with a common neighbour is the same as any other edge. That is, the clustering coefficient is given by *p*. The expected assortativity for networks constructed by these methods is 0.

2.3.2 Given degree sequence – configuration and Markov chain models

There are two major algorithms in the literature for producing a random network with a specified degree sequence. The algorithms are the configuration model, which uses a matching process, and the Markov model, which uses a swapping process.

2.3.2.1 Configuration model (Molloy-Reed algorithm)

Although implicit in earlier studies (see Bender and Canfield 1978; Wormald 1981) or presented in a different form (Békéssy et al. 1972), the configuration model is generally attributed to Molloy and Reed, who gave the first explicit algorithmic presentation (Molloy and Reed 1995, pg 166). This algorithm constructs a network by matching pairs of imaginary stubs (edges not yet formed). Each component is fully constructed and, if there are nodes not yet in the network when the algorithm runs out of available stubs, a new component is started.

The method is as follows:

1) Form a set that contains d_i copies of node i for all i, where d_i is the degree of node i

2) Choose any two members of the set uniformly at random and remove the pair from the set

3) Choose any member of the set for which at least one of its copies has already been removed and choose its partner at random from the whole of the set

4) Repeat step 3 until there are no members of the set which meet the conditions, in which case return to step 2, or until the set is empty

5) Construct the network by including an edge for each pair of nodes removed from the set

A network constructed with this algorithm is not necessarily simple (as the two members of the pair may be copies of the same node) nor connected. If a simple graph is required, the common practice is to reject any proposed pair where the nodes are the same or the pair has already been chosen and select another or, if there are no suitable pairs available, restart the algorithm. While this introduces bias (King 2004), there is some evidence that the bias is small (Milo et al. 2004).

All random choices are made uniformly. In practice, this means that a node will be chosen from the available nodes (all in step 2, or only those already chosen in step 3) in proportion to residual degree (that is, degree minus pairs already formed).

The Molloy-Reed algorithm generates uniformly any graph with a given degree distribution, where the nodes are labelled and therefore isomorphic networks are considered different. If, however, the graph is unlabelled (as occurs when working from the degree distribution rather than the degree sequence), this algorithm generates networks with probability proportional to the number of isomorphic networks with the given degree sequence. In principle, if a uniform selection is required from the degree distribution, a specific degree sequence from the given degree distribution could be selected with probability inversely proportional to the number of isomorphisms of that sequence. This is similar to the approach taken in (Goldberg and Jerrum 1996) to uniformly sample connected multigraphs.

2.3.2.2 Given degree sequence – Markov chain model

The other standard algorithm for generating networks with a specific degree sequence uses a Markov chain edge swap process to generate a random network from a starting network with the required degree sequence (Rao et al. 1996; Roberts Jr 2000; Snijders 1991).

Typically, the initial network is created using the Havel-Hakimi algorithm (Havel 1955; Hakimi 1962). This algorithm repeatedly selects the node with the highest residual degree (degree *d*) and makes edges with the *d* nodes with the next highest residual degrees. A simple graph is created.

The Markov Chain transition process performs an edge swap. Two edges are selected that do not share a node (say edge from node A to B and edge from node C to D). The potential swap involves breaking the existing edges and making new edges across pairs (so AC and BD, or AD and BC). This potential switch is carried out only if the new edges do not duplicate any existing edge (so the graph remains simple). This transition process maintains the degree distribution.

Because any simple graph with the given degree sequence can be generated with sufficient edge swaps from any other (which provides irreducibility) and the probability of generating a particular network depends only on the current network state, Markov Chain theory states that a uniformly random network is generated following sufficient edge swaps.

Several modifications to this algorithm have been proposed to generate connected graphs (Gkantsidis et al. 2003; Viger and Latapy 2005; Stauffer and Barbosa 2005). As well as the edge duplication test, these algorithms perform a connectedness test and the swap is only performed if the resultant graph is connected. However, to increase efficiency because connectedness tests are much more expensive than edge swaps, the test is only performed after some number *w* of potential edge swaps, and all edge swaps since the last test are undone if connectedness fails. The algorithms differ with respect to the size of *w* but they each increase *w* whenever there is a successful connectedness test and decrease *w* if connectedness fails.

Regardless of which of these two methods are used to construct a random network with specified degree sequence, the network will have a positive clustering coefficient arising from its degree sequence (mean probability of edge), and expected assortativity of 0.

2.3.3 Motif models – including exponential random graph (Frank-Strauss p*) models

A motif is a small subgraph, usually comprising 3 or 4 nodes and the edges between them (see Figure 2-2 for examples). There are several strands of

research considering how motifs may be used to understand and model real world networks. Motif models count the number of times each particular motif appears in the full network and uses this information to identify motifs that appear more or less frequently than expected.

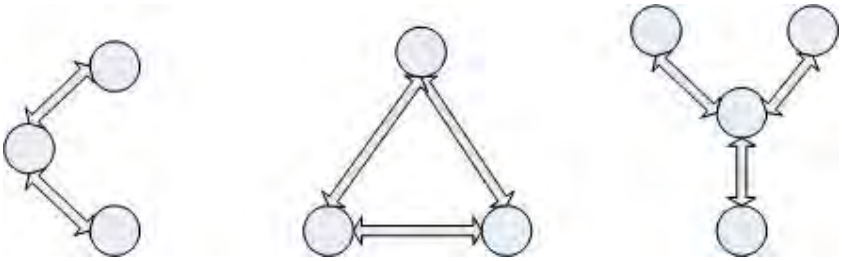

Figure 2-2: Motif examples*: Each of these subgraphs is a motif. They are referred to as (from left to right) the 2-star, triangle and 3-star.*

The most fully developed motif approach is exponential random graph (sometimes referred to as *p**) models. These models (Wasserman and Pattison 1996; Robins et al. 2005) generalise Markov graphs originally elaborated in (Holland and Leinhardt 1981; Frank and Strauss 1986). The key assumption is that the presence of an edge is dependent only on the presence of edges that are incident (have a node in common).

The model is then of the form:

$$P(\text{edge}) = \frac{1}{\kappa}\exp\left(\sum_{i=1}^{M}\theta_i z_i\right) \qquad (2.17)$$

where: i represents the particular Markov configuration (motif), such as edge, triangle, star
θ_i is the model parameter for the motif
z_i is the number of times the motif appears in the real network being modelled
κ is a normalisation parameter to ensure a probability function

Fitting models is a complicated process, with degeneracy problems and a high computation cost associated with the normalisation parameter. Several different approaches have been used including maximum likelihood

estimation, pseudo likelihood estimation and simple regression of approximate logit models (Anderson et al. 1999; Snijders et al. 2006). Once fitted, however, the model can be used to generate networks with the same conditional probability distribution of motifs.

Related work (Milo et al. 2002) has used a simulation approach to identify those motifs with 3 or 4 nodes that appear substantially more often in selected real world networks than would be expected in a randomly generated network with the same degree distribution.

2.3.4 Preferential growth (Barabási-Albert algorithm)

A highly studied network generation algorithm is that developed by Barabási and Albert (1999) with the intention of explaining and reproducing the scale-free degree distribution shown by many large networks, particularly the network formed by world wide web site links. The key parameter is the number of edges added per node. The original description of this algorithm has the following steps:

1) Start with a small number of nodes

2) Add remaining nodes one at a time

3) For each new node, add a fixed number of edges (m) to the existing network from that node. Each edge is attached to a node already in the network, selected in proportion to its existing degree (with no multiple edges).

The paper demonstrated that the two features of growth over time and preferential attachment (steps 2 and 3 respectively) lead to a network with a power law degree distribution. That is, the probability of degree k for a node is given by:

$$P(k) = c\, k^{-\alpha} \tag{2.18}$$

with c a constant to ensure total probability is 1 and $a = 3$. Note that the degree distribution does not truly satisfy the power law, as the power law

does not extend to very small degrees. In particular, the minimum degree is generally *m* and occurs for nodes added toward the end of the process. Lower degree nodes can only arise in the unlikely event that at least one of the initial nodes attracts fewer than *m* edges in the initial construction plus the growth phase.

As noted in (Bollobás and Riordan 2003), this description is not a full specification of an algorithm. Most importantly, the initial nodes have no edges and therefore the selection probability is undefined. The method used in this study refines the initialisation step of the algorithm in the following ways:

1) The initial number of nodes is set equal to the edges per node *m*

2) The initial network is complete. That is, each initial node has an edge with each of the other initial nodes, and hence degree of *m-1*

By ensuring connectivity in the initial network, this form of the algorithm generates a connected network. Also, the network after the first additional node will have *m+1* nodes, each of which has a degree of *m*, so this version has a minimum degree of *m*.

By construction, there is an age effect in this algorithm, where the high degree nodes are the early nodes and, further, these nodes are very close to each other. Consequently, preferential growth networks have a shorter average path length than random graphs with the same degree distribution (Albert and Barabási 2002, pp 74-75 and references therein).

Other models exist for generating power law distribution networks. The most direct is to select the desired degree sequence and use the configuration model described above. This approach was taken in (Aiello et al. 2001) to analyse the relationship between component size and the powers (*α*, *β* below) where the number of nodes of degree *k* (denoted n_k) follows the more general power law:

$$n_k = e^{\alpha} / d^{\beta} \qquad (2.19)$$

An alternative explanatory model was analysed in (Caldarelli et al. 2002; Servedio and Caldarelli 2004). It relies on an intrinsic 'fitness' of each vertex and a general linking function based on the fitness of the nodes at the ends of the potential edge. This method generates networks with a power law degree distribution for a broad range of fitness probability distributions and linking functions.

Two extensions of the Barabási-Albert algorithm are particularly important for modelling social networks. The first of these (Dorogovtsev et al. 2000) considers the case where nodes have some identical initial attractiveness (A) and selection is then proportional to attractiveness plus degree. This model avoids the problem of undefined probabilities at the start of the growth phase for $A > 0$. It also provides a more general power law degree distribution, with the power a able to vary from 2 (when A is 0) through 3 (when A is 1) to ∞ (when A is ∞).

The other extension of interest modifies the algorithm to allow the clustering coefficient to be set (Holme and Kim 2002) instead of accepting the default value close to zero. In the original algorithm, the new node is connected to existing nodes selected proportional to their degree. In the modified model, this selection process is used only for the first edge from the new node and, for an appropriate probability, for later edges. An alternative selection process is used for the other cases, where the existing node is selected from the neighbours of the node selected for the first edge. This alternative selection process ensures clustering occurs.

For networks of finite size, the skewness of the degree distribution in networks generated with the preferential growth algorithms and the fact that the networks are simple, leads to negative assortativity values (see Section 2.2.3 for discussion).

2.3.5 Lattice structures

Lattice networks are often generated in one (ring) or two (lattice) dimensions, but higher dimension structures could be used. Nodes are placed

at regular intervals and each node is connected to all of the nodes physically located next to it (see Figure 2-3). The lattice of nodes is generally considered to 'wrap' so that those on one edge of the lattice directly connect to those on the other side.

In lattice networks, each node has the same degree. Mean shortest paths are relatively high because there are no edges that connect physically distant sections of the network and many short hops are needed to get between nodes.

Clustering coefficients depend on the specific algorithm used. For example, in the 1D ring shown in Figure 2-3, each node has 4 (network) neighbours, with 3 of the possible 6 edges between them in place so the clustering coefficient is 0.5. However the 2D square lattice with degree 4 has a clustering coefficient of 0. Assortativity is meaningless for lattices as all nodes have the same degree.

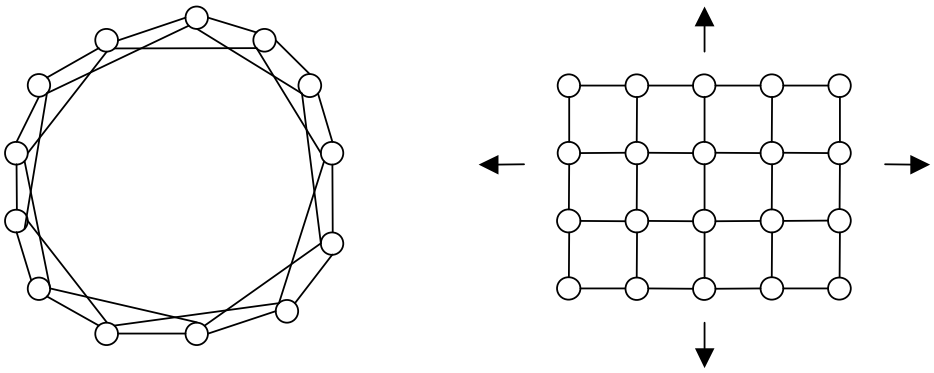

Figure 2-3: Lattice network examples *: The network on the left shows a one dimensional (1D) ring layout with each node connected to the two nearest nodes on either side (degree 4). The right hand side network shows a 2D rectangular layout with each node connected to the four nodes that are next to it, vertically or horizontally (also degree 4).*

Cellular automata models use a lattice as the underlying network structure. Lattices are particularly useful for modelling systems where physical location is important in the system behaviour, such as the spread of crop epidemics.

2.3.6 Small-world networks (Watts-Strogatz algorithm)

Real world social networks display the small-world property. That is, they have similar mean geodesic as random graphs of the same size and edge density, but with a very high clustering coefficient. This is in contrast to random graphs, which have a low clustering coefficient, and also to lattices, which have a high mean geodesic.

Watts and Strogatz (1998) developed an algorithm to generate networks with this property, by starting from a ring lattice and addind a rewiring procedure. For each edge, with given fixed probability, the edge is removed and a new edge created that joins one of the nodes previously connected with the original edge to a uniform randomly selected node in the network. The original implementation considered the edges in a particular order and this ordering also chose which node from the removed edge would be joined with the new edge.

Other implementations are possible. For example, the underlying lattice could be of any form. The edge creation process could have both end nodes selected randomly instead of remaining connected to one of the original nodes.

For any of these implementations, the principle of providing short cuts across the lattice structure is the same:

> … the rewired edges must typically connect vertices that would otherwise be much further apart than [the mean shortest path in a random graph with the same number of nodes] (Watts and Strogatz 1998, pg 441).

By varying the probability of edge rewiring, the algorithm provides intermediate networks between a regular lattice (probability of 0) and a

random graph (probability of 1). Note that the particular implementation used by Watts and Strogatz, with only one end of the edge rewired, does not generate a true random graph as defined in Section 2.3.1. In particular, the minimum degree in the network will be at least half the mean degree. However, implementation with both ends rewired does generate true random graphs where the edge rewiring probability is 1.

For a broad range of rewiring probability values, the mean geodesic of the networks generated by the algorithm increase logarithmically with the size of the network while clustering coefficient remained stable, thus demonstrating the required small-world property. Mean geodesic is lower for higher rewiring probability.

For networks generated with the small-world algorithm of rewiring a lattice, the clustering coefficient is close to the clustering coefficient of the original lattice and expected assortativity is 0.

2.4 Network generation with community structure

Several network generation algorithms are less widely used, but have been specifically developed to incorporate community structure in some way. Thus, they are potentially more relevant to epidemic simulation because the generated networks are expected to have more socially realistic properties.

2.4.1 Keeling's focal points algorithm

In spatially constructed network generation algorithms, physical location is used to define the network structure but, after the network is generated, the physical location has no relevance and is discarded. Simple examples have been explored in (Waxman 1988; Hong et al. 2005). Waxman's RG1 model (pp 1619-20) starts with N nodes distributed in rectangular 2D space at uniformly random selected integer coordinate points. For each pair of nodes, an edge is made between them with probability based on the Euclidean distance between them:

$$P(\text{edge}) = \beta \exp\left(\frac{-d}{\alpha L}\right) \tag{2.20}$$

where: d is Euclidean distance
L is maximum Euclidean distance
$\alpha \in (0,1]$ is a parameter for ratio of short or long edges
$\beta \in (0,1]$ is a parameter for edge density

A variation of this model has been used by Keeling and his colleagues (Keeling 1999; Eames and Keeling 2002; Keeling 2005) to examine the impact of some aspects of social network structure on epidemic behaviour, particularly mean degree, variance of degree and clustering. Keeling's algorithm has some additional steps and parameters:

1) The space used is a periodic square of size such that the mean node density is 1 node per unit area (so side length is $\sqrt{N}$)

2) Nodes are located anywhere in the space (coordinates uniformly random selected) rather than integer coordinates

3) Focal points are uniform randomly located in space (number or density of these is a model parameter)

4) Each node moves toward the closest focal point a fixed proportion (model parameter) of the distance to that focal point

5) For each pair of nodes, an edge is made between them with probability based on the Euclidean distance between them:

$$P(\text{edge}) = H \exp\left(\frac{-d^2}{2V}\right) \tag{2.21}$$

where: d is Euclidean distance (wrapped surface)
H is a parameter for edge density
V is a parameter for ratio of short or long edges (clustering)

6) Only the giant component is used, and only if it includes at least 90% of the available nodes.

This model is particularly suited to epidemic consideration because of the introduction of community structure through focal points. Nodes that move toward the same focal point also move toward each other, thereby increasing the probability of an edge being created between them. Further, the higher the proportion of distance moved, the stronger this effect. Thus, focal points play the role of

> … places where people congregate, and could represent schools and workplaces, or family groups, depending on their number (Keeling 2005, pg 3).

The edge creation probability is inspired by the Gaussian normal distribution, but there is no requirement that $H = \sqrt{2\pi V}$. Instead, H (height) and V (variance) are able to be set entirely independently.

H represents the maximum probability of an edge, as it is the probability that two nodes identically 'located' have an edge between them. Thus, $H < 1$ ensures that a proportion of node pairs do not have edges, providing an upper limit to edge density. Also, $H > 1$ establishes a distance where any two nodes that are closer than that distance always have an edge between them in the created network.

V provides a distance over which edge creation occurs. If V is very small, probability of an edge is low for all but very small distances. In contrast, if V

is very large, the distance has minimal impact on the probability and the probability of an edge approaches *H* for all node pairs.

Using a standardised version of the *V* parameter, the relationship between algorithm parameters and network properties has been examined in (Badham et al. 2007). This paper shows that the impacts of the *H* and (standardised) *V* parameters on the network properties are consistent and predictable. In contrast, the focal point parameters have inconsistent effect (see summary at Table 2-2).

Table 2-2: Impact of parameters in Keeling network generation

Parameter		Mean Degree	Degree Variance	Clustering Coefficient	Assortativity
Nodes	N	Parallel	Parallel	Minimal	Minimal
Edge creation	H	Parallel	Parallel	Parallel	Parallel
Edge creation	V	Parallel	Parallel	Minimal	Opposing
Focal point density	f	Minimal	Minimal	Various	Various
Move proportion	m	Minimal	Minimal	Various	Various

There are three difficulties with using this algorithm to generate networks for the epidemic simulations required. First, the degree distribution is approximately binomial. There is no capacity to generate the highly skewed degree sequences that can arise in social networks. Second, the relationships between parameters and properties break down where very small values of *V* are required; that is, for sparse networks with low mean degree but large size. Third, values for clustering coefficient and assortativity tend to be similar, with scope for separation only in networks with large mean degree.

2.4.2 Newman's community structure algorithm

Another network generation algorithm that explicitly includes communities also provides some control over clustering and assortativity in the generated network (Newman and Park 2003; Newman and Park 2007). While this model was developed to estimate the value of network properties arising naturally

from the distribution of nodes across groups, it can also be used to generate networks.

In this model, nodes are distributed to multiple groups, with the number of groups for the node randomly selected from some arbitrary distribution. The size of each group is also selected from some (different) arbitrary distribution. Edges are created between all pairs of nodes within groups with a fixed probability. The final parameter is either the total number of nodes or the total number of groups. These are related to each other through the means of each distribution and the edge probability.

Where the groups are of different size, there is assortativity in the generated network. Further, the value of the assortativity coefficient can be calculated theoretically, using the moments of the two distributions. The clustering coefficient is related to the edge probability (with reasoning similar to Newman 2003b).

This algorithm is unsuitable for generating the required networks for epidemic simulation because the properties of interest are generally not independent. In particular, mean degree, shape of degree distribution and assortativity all depend on the two distributions.

Some control exists for the specific case where the number of groups for each node is an integer constant C and the group size is taken from a Poisson distribution. In this case, assortativity (ρ) depends only on the probability of connection (p) and the number of groups, and is given by:

$$\rho = p/C \tag{2.22}$$

For this case, clustering coefficient depends on p and mean degree can be adjusted through group size without impacting on the other properties. However, there is no way to control the degree distribution shape.

2.4.3 Using motif algorithms for network properties

Researchers using the motif algorithms have focussed on fitting models to real networks and quantifying the extent to which the motif frequencies differ

from that expected (see Section 2.3.3). While there has been no research explicitly examining the relationship between model parameters and network properties, there has been some recent work examining the goodness of fit of models (Robins et al. 2007; Hunter et al. 2005), which is relevant to generation of networks with specific properties.

The goodness of fit research uses the fitted model to generate a number of networks and then examines whether the original network differs from the simulated sample. This comparison examines both the distribution of motifs (which is the input to the model) and broader network properties such as moments of distributions of degree, geodesics and local clustering coefficient. At least for some sets of motifs, this research shows that the properties of networks simulated from a model fitted to those motifs are reasonably stable and consistent with the real world network properties.

2.5 Incorporating social structure in epidemiological models

The basic epidemiological model excludes several significant biological factors that could be important in modelling the behaviour of an epidemic. These include such issues as births and natural deaths (that is, not caused by the epidemic), differential impact of a disease on population subgroups (such as age groups), impact of maternal disease status on infants, and loss of immunity over time (for a discussion of these, refer to Anderson & May 1992).

Some of these factors were incorporated by Kermack and McKendrick in their later work (1932; 1933). Amongst other things, this work demonstrated that a disease that conferred immunity could nevertheless maintain an endemic state in a population where new susceptible individuals were added to the population through birth or immigration. That is, the disease does not die out, instead maintaining a stable prevalence.

The key issue to be considered further in this study is the impact on epidemic behaviour of relaxing the assumption of the mass action principle. Under this

principle and the consequent assumption of equal infection transmission parameter, patterns of contact within the community are ignored and every susceptible individual has an equal probability of coming into contact with every infected individual. All other assumptions of the basic epidemiological model will be maintained, including homogeneity of disease impact and use of a fixed population.

Social network structure has been incorporated into epidemiological models analytically and by simulation. There are several reviews (including sections of Newman 2002b; Newman 2003c; Watts 2004; Keeling and Eames 2005) for the specific consideration of epidemic behaviour on networks.

2.5.1 State model

There are many potential sources of heterogeneity that may affect epidemic dynamics and therefore need to be incorporated into models. These include (Anderson & May 1992; Diekmann & Heesterbeek 2000):

- Demographic structure: infectivity, recovery and other rates may differ for different age and/or gender groups;
- Genetic and comorbidity variation: different people of the same 'type' (age, gender, infection status) may have different infectivity, susceptibility or recovery rates due to factors specific to the individual, including the presence of other disease;
- Social structure: individuals with the same demographic and genetic factors may have different social activity levels that impact on their infectivity or susceptibility, particularly in the number of contacts and the frequency of partner change;
- Spatial structure: this is a specific type of social structure impact where the contact rate is entirely dependent on the local population density.

Although different authors may use different presentations, terminology and/or levels of rigour and formality, these structures are each incorporated into models in the same way.

The key steps are (Nold 1980; May and Anderson 1984; Hethcote and van Ark 1987; Diekmann et al. 1990; and others):

1) Define a set of states (or subpopulations), with each state incorporating a particular value or range of values for each factor (for example, a state may be age 25-34, female, 1-2 sexual partners per year);

   - The set of states must be exhaustive and mutually exclusive (that is, each individual can be assigned to exactly one state);

2) Calculate population counts for each state;

3) Assume that individuals with the same state are homogenous in terms of any characteristics that affect epidemic behaviour;

4) Define a 'who acquires infection from whom' (WAIFW) transmission matrix K with elements k_{ij} set by the number of new cases of state i caused by a single infected individual of state j;

5) The matrix and initial population counts are then used to simulate epidemic behaviour.

The transmission matrix can also be formulated as the probability that a member of class i will infect a member of class j (instead of the number of infections), with consequential changes in the presentation of results.

There has been considerable progress in determining analytical results for state based models. In particular, Diekmann et al (1990) demonstrates that the basic reproduction ratio is the spectral radius of the transmission matrix:

$$R_0 = \lim_{n\to\infty} \left\| K^n \right\|^{1/n} \tag{2.23}$$

The interpretation of this is that the transmission matrix K operates on one generation of the epidemic to describe the state specific number of infections in the next generation of the epidemic. The spectral radius is then the average over many generations of the generational infection change.

Further, R_0 is the dominant eigenvalue of the transmission matrix as all elements are positive. Results for initial growth rate, probability of a minor outbreak, final size and prevalence for an endemic are presented in (Diekmann & Heesterbeek 2000). However, the mathematics is intractable in most situations.

Averaged across possible states for the initial infection, the transmission matrix is similar to the next generation operator (Diekmann et al. 1990) or secondary reproduction ratio (Eguíluz and Klemm 2002), providing the expected number of infections arising directly from a single infection. Under the assumption that the probability of infection is constant (given a contact between an infected and a susceptible node), this is also equivalent to the proportion of the network accessible by travelling along a single edge from the starting position, averaged over all nodes (potential starting positions).

While the state approach is able to incorporate many observable sources of variation, there are many other less clearly defined aspects that may also impact on epidemic behaviour. These include environmental factors (such as the impact of the weather on the capacity of the infective agent to survive the transfer process), presence of other diseases that affect immunity, and nutrition and hygiene behaviour.

2.5.2 Analytical approaches to contact variation

One specific application of the general state model is to define states by number of contacts. That is, all members of the population (or nodes in the social network) are considered homogenous except for contact rate. This introduces the network theory concept of degree distribution to the epidemic model. In general, degree heterogeneity increases R_0 (Becker 1973; Adler 1992).

Note that an inherent assumption of the state model is that all individuals in the same state are equivalent and any may be uniform randomly chosen for the purposes of the transmission matrix. This is equivalent to an assumption

that the clustering coefficient is equal to the expected clustering coefficient given the degree distribution and that assortativity is zero.

As an aside, many results from state model epidemiology require that the transmission matrix be irreducible. This is satisfied by any network that is connected, because a node in any state can be reached from a node in any other state.

One assumption often applied to epidemic state models, which simplifies the mathematics considerably, is that of separable mixing. This assumption requires that the states of the infective and susceptible individuals involved in any transmission are independent. Equivalently, the social network defined by potential epidemic transmission has assortativity of zero with respect to the combination of properties used to define states. Where states are defined solely by contact rates, separable mixing is equivalent to assuming zero (degree) assortativity. Zero assortativity social networks with degree heterogeneity but otherwise homogenous nodes also satisfy the stronger requirements for proportionate mixing, where the relative infectivity of a group is equal to its relative susceptibility, because a node's infectivity and susceptibility are both proportional to its degree.

Under the assumption of separable mixing, R_0 is given by (Nold 1980; Diekmann et al. 1990):

$$R_0 = R_{\hat{k}}\left[1 + \frac{\text{var}(k)}{\hat{k}^2}\right] \tag{2.24}$$

where: $R_{\hat{k}}$ is the basic reproduction ratio with uniform degree $\hat{k}$
var(k) is the variance of the degree distribution
$\hat{k}$ is mean degree

The correction factor reflects the fact that, regardless of how the initial infected node is selected, nodes infected later are selected proportional to their susceptibility, in this case degree, and this in turn means that the infected nodes also have relatively high infectivity.

The inclusion of the degree variance term can allow epidemics to exist in a population with much lower average infectivity per node. This is because a core group with high contact rates has a disproportionate effect arising from the high potential of core group members to become infected and also the opportunity to infect many others. At the extreme, a scale free network of sufficient size can maintain an epidemic for an arbitrarily low mean prevalence (Pastor-Satorras and Vespignani 2001).

A separate research thread (Becker 1973; Ball 1985; Andersson and Britton 1998; Lefevre and Malice 1988) has examined the size of epidemics in the presence of infectivity and susceptibility heterogeneity. While this research is not specifically considering the contact rate component of infectivity and susceptibility, the results can be applied to degree heterogeneity. In general, heterogeneity leads to a smaller epidemic than would be expected from the mean infectivity and susceptibility, weighted by number of nodes. However, the opposite is true when infectivity is low.

Thus, the higher basic reproduction ratio does not translate into larger epidemics as would be expected from equations (2.3) and (2.4). Extended equations have been developed under separable mixing for both SIS (Nold 1980; May and Anderson 1984; Hethcote and van Ark 1987a) and SIR (Ball and Clancy 1993; Britton 2001) epidemics. These equations show a similar form, but the contribution of subpopulations is weighted, with the weights incorporating the degree of the subpopulation and the correction factor for degree variation. Because of the nonlinearity of the relationship between R_0 and final size (SIR) or endemic prevalence (SIS), the degree variation can lead to a substantial difference between the R_0 derived from observed epidemic behaviour and the actual R_0.

The counterintuitive result of degree heterogeneity leading to higher R_0 but smaller epidemics is related to the shape of the degree distribution. The highest degree nodes are the most susceptible and are able to infect the most additional nodes, but also represent the smallest proportion of nodes. The low degree nodes suffer the least impact of the epidemic but form a much greater

proportion of the nodes. This issue is discussed in detail with an example in Section 5.7.1.

2.5.3 Dependency in number of contacts (assortativity)

In the absence of separable mixing, there are two approaches that have been taken to introduce assortativity to the transmission matrix. Note that networks with uniform degree must have assortativity of 1 by definition. Hence, any analysis of the impact of assortativity on epidemic behaviour must also have degree heterogeneity.

The first approach considers the special case where there are two components to the mixing: proportional mixing across states plus preferential mixing within states.

> A proportion $1 - s$ of contacts [edges] are distributed in proportion to the fractional activity levels [in this case, degree] of the groups contacted (Nold 1980, pg 237).

The remaining edges connect nodes within groups. The parameter s allows assortativity to range between 0 (at $s = 0$) and 1 (at $s = 1$). Note that the fully assortative network is disconnected, with each component containing only nodes of the same degree. Under this scheme,

> … there exists an interval of s values such that the disease persists, as long as the infectious contact number of some group exceeds 1 (Nold 1980, pg 238).

That is, an assortative network will have subnetworks with relatively high mean degree and other subnetworks with relatively low mean degree. The high degree subnetworks have a reproduction ratio within that network section of greater than one, and are able to maintain the epidemic. This is different from the degree variability case discussed in Section 2.5.2. For high assortativity networks, the epidemic is maintained by direct transfer between high degree nodes. In the case of high degree variability without assortativity, the epidemic is transferred from high to low to high degree nodes.

An expression equivalent to the epidemic threshold is derived for similar mixing schemes by (Diekmann et al. 1990, section 3.2 and references therein). The required level of vaccination is derived in (Anderson & May 1992, section 9.6 and Appendix D).

The same mixing scheme is applied to power law degree distribution networks in (Moreno et al. 2003). SIR epidemic behaviour is analysed using mean field methods and by simulation. This paper finds the probability of an epidemic is reduced for the assortative networks compared to random networks, but no difference in the limit of large network size. In addition, the epidemic is smaller in the presence of assortativity.

The second approach is based on the joint distribution of the degrees at each end of the edges. This approach is not restricted to the analytically tractable proportional plus preferential mixing. Given the joint degree distribution, the assortativity can be directly calculated. Newman (2002a) derives an expression for the size of the giant component of such a network, which is related to the SIR final size.

In the same paper, Newman describes an algorithm to generate a network with the specified joint distribution by successive edge swaps from a randomly generated network with the appropriate degree sequence. This algorithm and the analytical results are tested for the specific distribution where the degree at each end of the edge is from a binomial distribution with the same probability and that probability then controls the assortativity. As assortativity increases, the giant component forms more easily and is smaller. Newman concludes that positive assortativity (as occurs in social networks) could lead to a:

> core group … [that sustains] an epidemic even in cases in which the network is not sufficiently dense on average for the disease to persist (Newman 2002a, pg 4).

Similarly, (Boguña and Pastor-Satorras 2002) uses the complete joint degree distribution (referred to as the connectivity matrix) to derive a relationship between the required infectivity to reach the epidemic threshold and the

largest eigenvalue of the WAIFW matrix. Epidemics are simulated on networks of various sizes (and, due to the generating algorithm, different assortativity) to provide empirical support for the relationship. The epidemic threshold is found to be lower on assortative networks than for epidemics on networks with the same degree distribution but zero assortativity.

The two approaches thus lead to consistent conclusions that epidemics are smaller for assortative networks than for randomly mixed networks with the same degree distribution. Only Moreno et al (2003) quantified the relationship, noting that assortativity can decrease size by 15-20% for moderate (in terms of their simulation parameters) infectivity rates. Further, it is suggested by the reported simulation results but never explicitly stated that a larger assortativity leads to a smaller epidemic, though the methods used make such comparisons difficult as other network features may also change for some studies.

There are inconsistent results for epidemic occurrence. The paper by Moreno et al (2003) found that assortativity decreases the probability of an epidemic for finite network size. Other studies found the epidemic threshold to be lower, which suggests the probability of an epidemic is higher.

2.5.4 Clustering within the social network

The network phenomenon of clustering cannot be examined from the perspective of a WAIFW transmission matrix as the nodes to connect are not randomly selected from those nodes with a given state.

An alternative analytical approach was taken in (Keeling 1999) where the effect of (fraction of transitive triples) clustering was considered in a fixed degree distribution network. The edges were classified by the epidemic status of the nodes at each end. Because the proportion of infected-infected edges is non-zero in clustered networks, the basic reproduction ratio is reduced compared to equivalent random networks, with consequently slower initial growth and smaller final size.

Simulated SIR epidemics supported the analytical results and also found that clustered networks have a higher probability of failure to achieve an epidemic. These networks were generated using an algorithm that assigned locations to the nodes and weighted the edge creation probability by distance.

A more sophisticated network generation algorithm was used in (Keeling 2005) to also allow comparison of the impact of degree variance and interaction with mean degree and clustering. In this paper the simulated SIR epidemics were compared to basic epidemic models, fitted to the initial growth. Clustering was again found to inhibit early growth, but the final size was higher than for the best fit model. This is because the best fit model without clustering version actually has a lower underlying infection rate and the inhibition in the clustered network is reduced once the epidemic escapes from the initial infection area.

2.5.5 Incorporation of spatial structure

Spatial location is not considered in the network theory and epidemiological models already described. However, it clearly influences network structure, as contact requires some form of collocation.

Lattice based cellular automata models emphasise spatial structure but are unrealistic in other social network aspects. Various methods have been used to introduce social structure into epidemic simulations on lattice networks.

One method (Eidelson and Lustick 2004) assigns identities to each node and these identities are able to change over time. Probability of transmission is higher if the infected and susceptible neighbours share an identity. Another model assigns one or more mirror identities (reflecting home, school, transport etc) to nodes (Huang et al. 2004; Huang et al. 2005). Thus, each node appears in multiple locations and its neighbourhood is composed of the neighbourhoods of all the locations.

Lattice models can also be extended by allowing movement of agents within the lattice structure. For example, one model of human epidemics (Rhodes

and Anderson 1996) has examined the impact on epidemic behaviour of different speeds of movement.

This research emphasises that spatial structure impacts on epidemic behaviour, but no general rules are developed about the form of the impact.

An alternative approach to spatial structure is taken by metapopulation or patch models. In these models (Lajmanovich and Yorke 1976; Lloyd and May 1996; Grenfell and Harwood 1997), the population is divided into an arbitrary number of subpopulations of arbitrary size that reflect households, towns or other spatially defined communities. The simplest model uses two different rates for contact rate; a relatively high rate within a subpopulation and a lower rate between different subpopulations. However, this approach is simply another form of a general state model (Section 2.5.1), with the state defined by subpopulation and arbitrary contact rates between any two subpopulations. The contact rates are then the basis for the WAIFW transmission matrix.

A more sophisticated version of the patches model (Morris 1995) uses socially meaningful characteristics such as age to define the patches instead of spatial structure. As for clustering or degree variation, relatively small subpopulations with high contact rates are able to maintain an epidemic and transmit it to other groups. In this form, the patches model is very similar to Newman's community structure network generation algorithm (Section 2.4.2). With Newman's algorithm, there is a contact rate (probability of edge) within a group, the same contact rate between selected pairs of groups (those pairs where there is a node in common) and a zero contact rate between any other pairs of groups.

2.6 Relevance to research questions

The literature is clear that social structure has an impact on epidemic behaviour, including whether an epidemic occurs and the size of an epidemic (Section 2.5). The primary research question is therefore a legitimate field for research.

> **Primary research question**: What is the relationship between epidemic behaviour and three key features of social networks: positively skewed degree distribution, positive clustering coefficient and positive assortativity?

Other network properties may also influence epidemic behaviour. However, the properties selected for inclusion are key characteristics of social networks (Newman and Park 2003). A key parameter in epidemic models is the basic reproduction ratio (Diekmann et al. 1990). Thus, the term 'epidemic behaviour' in the primary research question can be focussed to create two secondary research questions.

> **Secondary research question 1:** How does each of these properties affect epidemic occurrence?

> **Secondary research question 2:** How does each of these properties affect the basic reproduction ratio R_0?

The relationship between degree heterogeneity and epidemic behaviour has been extensively studied (Section 2.5.2). Increased degree variation leads to higher epidemic occurrence but smaller epidemics.

In contrast, the impacts of assortativity and clustering are poorly understood (Sections 2.5.3 and 2.5.4). Assortativity increases the probability of an epidemic, while clustering decreases it. Both properties reduce the size of epidemics that do occur. Quantitative relationships exist only for specific combinations of artificial degree distributions and a narrow range of property values.

It is also clear that real world social networks simultaneously exhibit degree heterogeneity, clustering and positive assortativity. The degree heterogeneity results assume that clustering and assortativity are absent. Also, there is no information about the joint effect of assortativity and clustering. Thus, there are no existing results relevant to secondary research question 3.

> **Secondary research question 3:** Do these social network properties influence epidemic behaviour separately or jointly and, if the latter, how do they interact?

To investigate these questions, the experimental design requires simulation of epidemics on networks with a range of property values. Further, each property must be able to be considered separately and in combination with various values of the other properties. Thus, simulation requires a network generation algorithm that allows each of the three social network properties to be controlled separately.

Section 2.4 describes published algorithms that are able to generate networks that control any two of the three sought properties: degree distribution, assortativity and clustering coefficient. However, the remaining property either has an expected value arising naturally from the degree distribution (such as zero for assortativity), or the inherent structure of the algorithm leads to a limited range of values for the 'free' property. A suitable algorithm to generate the networks with all three properties is not available in the literature and is the subject of secondary research question 4.

> **Secondary research question 4:** How can networks be generated for simulations with various values of degree sequence, assortativity and clustering coefficient, separately and jointly?

Chapter 3: Experimental Design

There are three network properties of interest: degree distribution, clustering and assortativity. In principle, there are at least two methodologies possible to consider the impact of these network properties on epidemic behaviour, mathematical modelling and simulation.

Mathematical modelling has the advantage that the resultant model can be used to calculate the expected behaviour from any set of inputs. However, models of epidemic spread must incorporate the degree of the infected node and the expected degree of the susceptible nodes to which it is connected and average this over all infected nodes. Existing models have incorporated some elements of degree distribution and either assortativity or clustering (see Sections 2.5.2 to 2.5.4). The degree distribution impacts on the degree of the infected node and the averaging process. Clustering has a complex effect on the probability of a node previously being exposed and hence on the susceptibility of neighbour nodes. Assortativity modifies the expected degree distribution of the neighbour nodes to depend on the degree of the instant node. The existing models are already complicated and extending them to incorporate a third property appears infeasible.

Simulation is a suitable approach for modelling complex social systems but has its own limitations (Gilbert & Troitzsch 1999; Marney and Tarbert 2000; Goldspink 2002). In particular, simulations are specific to the system parameters actually simulated and results cannot necessarily be extrapolated to more general results. There are also more subtle issues where the implementation decisions can affect the results (Agar 2003; Polhill et al. 2005). Also, there is no published algorithm that can generate networks with the range of property values required.

However, generating the networks required is more feasible than extending the mathematical models. Thus, simulation was selected as the methodology

for the study. The experimental design described in this chapter assumes the problem of generating networks with relevant properties can be solved.

Clearly, simulated networks will need various combinations of values of the social properties. Also, the simulations will be run with different infectivity levels to enable examination of the consistency of derived relationships. In addition, the relationships must be examined for the two basic categories of epidemic models, SIR (infection provides full immunity) and SIS (no immunity conferred).

The properties to be varied in this study fall into two groups. The network properties are degree distribution (3 types), clustering coefficient (all available values) and assortativity (all available values). The epidemic properties are infectivity (3 values) and immunity (2 values).

In general, 100 simulations are run for each specific combination of network and epidemic parameters. These are comprised of 10 epidemic runs on each of 10 networks.

To minimise differences between networks with different clustering and assortativity levels, the networks with different values of these properties are generated with the same target degree sequence. That is, 10 target degree sequences are used for each of the three distribution types and these are used to generate the 10 networks for each set of network properties. Some networks are more difficult than others to generate and, in these cases, fewer than 10 networks may be generated for some property combinations and, hence, fewer than 100 epidemic simulations are available for those networks.

Additional networks are generated to implement the assumptions of the basic epidemiological model in a network context. Epidemics are simulated on these networks with the same set of infectivity and immunity parameters to allow a comparison.

3.1 Networks for epidemic simulation

A novel algorithm, referred to in this thesis as the neighbour algorithm (see Section 4.2 for details) is used to construct the networks with specific values for target degree sequence (3 broad shapes), clustering coefficient and assortativity. Up to 10 networks are generated for each combination of network property values.

Other network properties are held constant for all simulations. In particular, all networks have the same number of nodes (1 000) and target mean degree (8). These values are arbitrary, chosen for sufficient size but moderate enough for repeated computation.

The three types of degree distribution shapes used are an exemplar real world degree distribution, power law, and normal distribution. Each network is generated with a different seed.

For each degree distribution type, degree sequence instances are extracted from networks constructed with generation algorithms common in the literature (normal, power law) or directly from the degree distribution (real world). Using the neighbour algorithm, a network is generated with each degree sequence and a range of specific clustering and assortativity values.

Target assortativity and clustering coefficient values are 0, 0.1, … to the maximum possible. A single attempt is made with a range of input parameters to generate networks, and any property pairs not obtained are then specifically targeted with up to 10 attempts each. This process is shown in Figure 3-1.

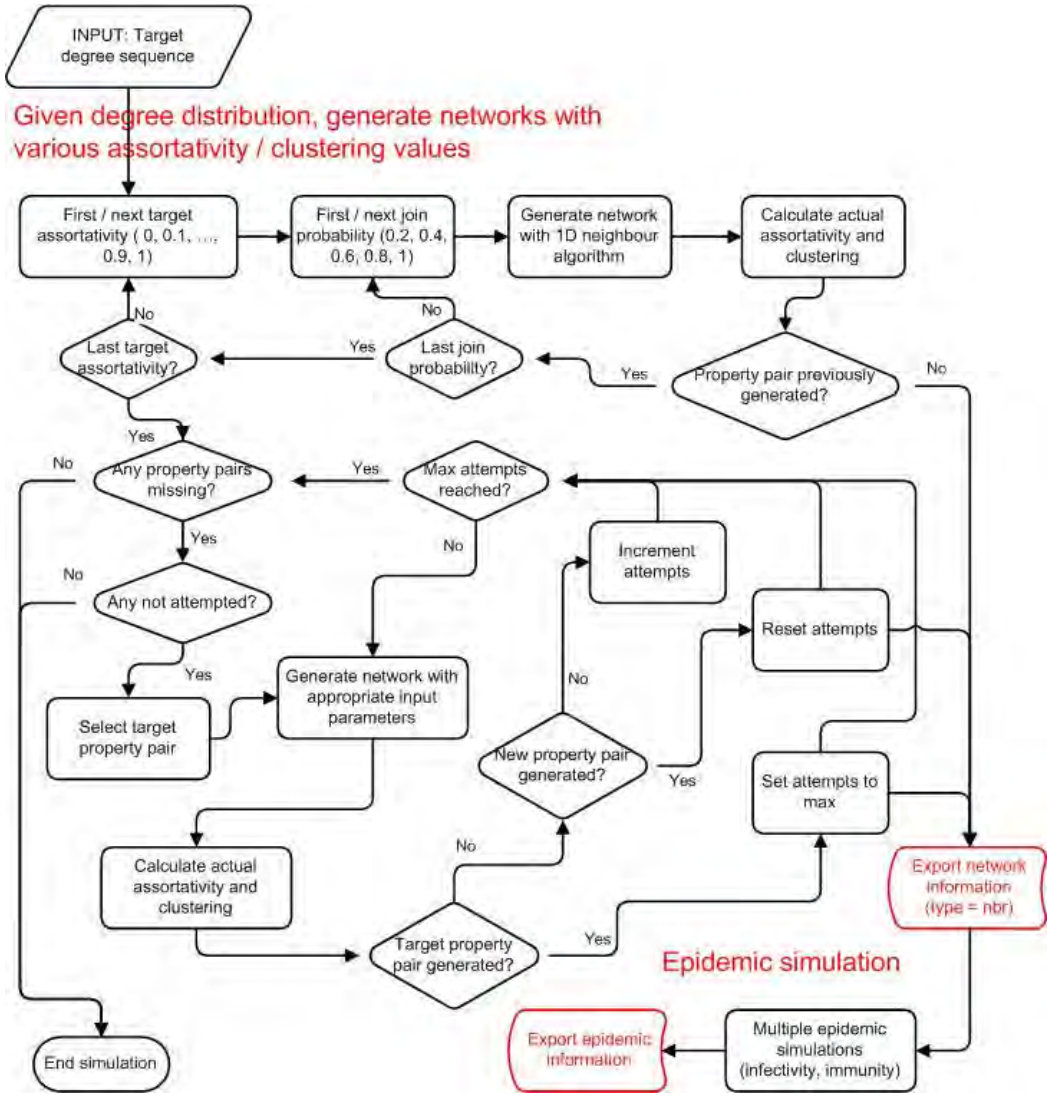

***Figure 3-1: Experimental process: Simulation of epidemics over networks with specific properties**: From the degree sequence, a network is generated with specific assortativity and clustering coefficient values and all epidemic simulations are run before generating the next network.*

3.1.1 Normal degree distribution

For the normal degree distribution, the fixed number of edges variant of the Erdös-Rényi algorithm (Erdös and Rényi 1960, and Section 2.3.1) is used to generate ten network instances, each with 1 000 nodes and 4 000 edges.

For each network, the degree sequence is extracted and used with the neighbour algorithm to generate a network for each accessible pair of assortativity and clustering coefficient values (refer to Figure 3-1).

3.1.2 Real world degree distribution

The real world distribution exemplar is taken from an early study of a friendship network between children (Rapoport and Horvath 1961). Each child nominated up to 8 friends. The number of nominations received is then used as the basis of the undirected real world degree distribution.

The full distribution was published (Rapoport and Horvath 1961, Table 5). Each child was nominated by between 0 and 29 of the children in the social group, with mean 6.84. For use in the experiments, the cumulative probability distribution is rescaled by multiplying each degree point by 8/6.84 (to increase mean degree to 8), and linear piecewise interpolation is used to create the corrected cumulative probability distribution. Rescaled for mean degree of 8, the probability density function is displayed at Figure 3-2.

As can be seen from this figure, the distribution displays positive skewness, characteristic of real world networks. The degree distribution is the only aspect of the exemplar network used in this study.

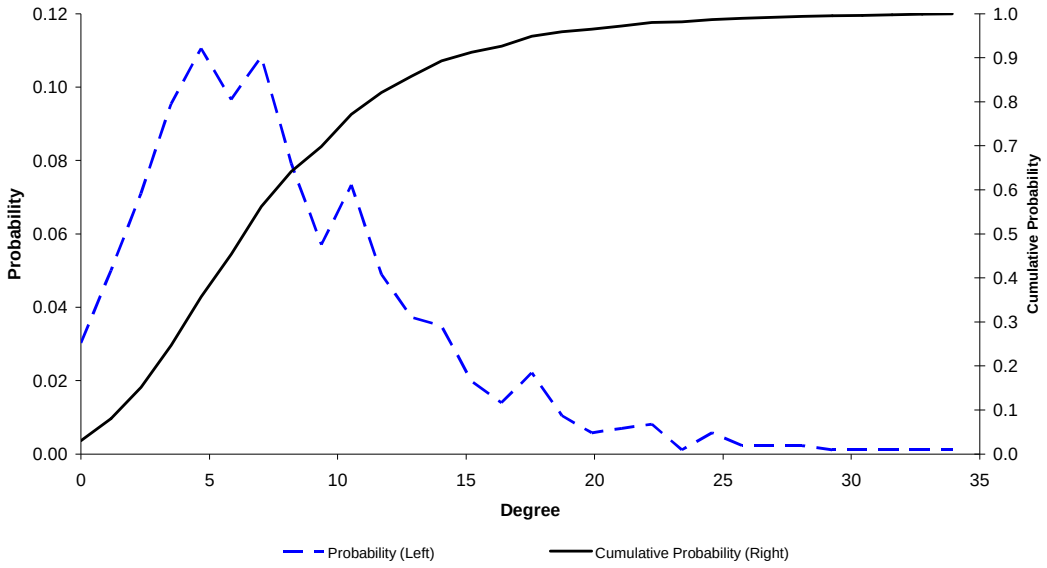

Figure 3-2: Real world degree distribution

Ten instances of a degree sequence are generated, each by sampling from the rescaled degree distribution until 1 000 non-zero degree values are selected. This degree sequence instance is used with the neighbour algorithm to generate a network for each accessible pair of assortativity and clustering coefficient values (refer to Figure 3-1).

3.1.3 Power law degree distribution

For the power degree distribution, the Barabási-Albert algorithm (Barabási and Albert 1999, and Section 2.3.4) is used to generate ten network instances. The initial network has 4 nodes, each connected to all other nodes. Each additional node also adds 4 edges. The generated networks hence have 1 000 nodes and 3 990 edges.

For each network, the degree sequence is extracted and used with the neighbour algorithm to generate a network for each accessible pair of assortativity and clustering coefficient values (refer to Figure 3-1).

3.1.4 Basic epidemiological model (uniform degree)

In addition to the neighbour networks used for analysis of property impact, networks are constructed to implement the basic epidemiological model. This network has uniform degree and is generated with the modified Molloy-Reed algorithm described in Section 4.1. That is, all nodes have the same degree lin this case, 8). The modification is to ensure the network is connected.

By definition, networks with uniform degree have assortativity of 1. Hence, this distribution type is not used to generate networks with specific assortativity and clustering values for the relationship analysis, but instead provides a comparison point.

3.1.5 Multiple degree sequences or multiple networks?

There are two broad choices for sampling of the ten networks with any parameter set: generate 10 networks from the same degree sequence (or

instance of the degree distribution), or generate 1 network from each of 10 degree sequences.

The advantage of a single degree sequence with multiple networks approach is that one source of variation is removed, making it easier to identify relationships. On the other hand, a single degree sequence introduces a risk that the relationship identified depends on the specific degree sequence rather than more general characteristics of the degree distribution such as shape.

To assist with this decision, networks with normal and power law degree sequences were generated using the neighbour algorithm with 1 000 nodes, mean degree of 8, target assortativity of 0.2 and maximum achievable clustering coefficient (0.5 for normal and 0.4 for power law). For each distribution, 30 networks were generated from a single distribution instance and a single network was generated for each of 30 distribution instances.

An SIS epidemic (that is, no immunity) was simulated on each network with 3 nodes initially infected, automatic transmission from an infected node to an adjacent susceptible node and probability of recovery of 0.3333 in each timestep (hence, mean infection period of 3 timesteps). The number of infected nodes was counted at timesteps 5, 10 and 20.

Table 3-1: Variation in number of infected nodes – degree sequence or network instance? *Number of nodes currently infected at timestep, mean and coefficient of variation over 30 networks*

	Timestep 5		Timestep 10		Timestep 20	
	Mean	**CV**	**Mean**	**CV**	**Mean**	**CV**
Normal – vary degree	414	24.0%	1065	5.9%	1412	2.2%
Normal – same degree	476	31.9%	1058	11.8%	1419	3.0%
Power law – vary degree	671	25.8%	1102	5.6%	1445	2.4%
Power law – same degree	708	18.2%	1120	4.1%	1457	2.0%

From Table 3-1, the two types of networks give different results. For normal degree distribution, generating multiple networks from the same degree

sequence led to greater variation in the number of infected nodes at all three timesteps. For power law degree distribution, greater variation occurred with multiple degree sequences. For both types of degree distribution, variation reduced over time.

The relatively low variation in the power law single instance simulations emphasised the high risk that any relationship identified depends on the specific degree sequence. Further, both versions of the normal degree distribution design showed relatively high variation, suggesting either would be suitable for the experimental simulations.

The multiple degree sequence with single network option was selected (as described in the sections for each degree distribution, Sections 3.1.2 to 3.1.3). That is, the approach used was to generate 10 different degree sequences for each type of distribution. For each instance, one network was generated with the required property values for assortativity and clustering coefficient.

3.2 Epidemic simulation design

Once the networks are generated, 60 epidemic simulations are run for 100 timesteps. The 60 simulations comprise 10 runs with each of 3 infectivity levels and 2 immunity types. Each simulation is started with a single infected node, selected uniformly at random, and uses a different seed. For each timestep, the numbers of infected, susceptible and immune nodes are recorded.

3.2.1 Model update process

The simulations are updated synchronously. Each timestep, the infected nodes are checked for potential transmission of infection to susceptible neighbours and then checked for potential recovery.

The full process is (also see Figure 3-3):

1) Infect initial node (at timestep 0)

2) Add 1 to timestep counter

3) FOR EACH infected node

    4) Identify all susceptible neighbours

    5) FOR EACH susceptible neighbour

        6) With probability of infectivity rate * susceptibility, mark for infection in next timestep

    7) END for susceptible neighbours

    8) With probability of recovery rate, mark the infected node for recovery in next timestep

9) END for infected nodes

10) Infect all susceptible nodes listed for infection

11) For each infected node listed to recover, make immune (SIR) or susceptible (SIS)

12) Count nodes in each state (susceptible, infected or immune)

13) Move to next timestep (that is, return to step 2)

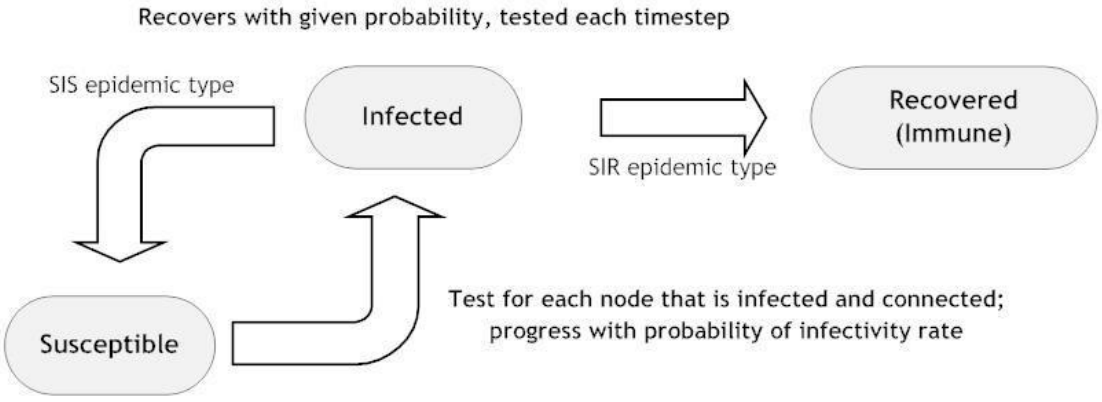

Figure 3-3: Epidemic state progression for nodes in simulation

The initial infected node is infected at timestep 0, so secondary infections start at timestep 1. The initial node is included in the cumulative count of infected nodes.

3.2.2 Epidemic parameters

Three infectivity rates are used. The values are arbitrary but are intended to provide R_0 values near but exceeding 1. At this value, epidemics are likely to occur but also fail. At higher infectivity levels, an epidemic would occur for almost all simulations and epidemic sizes would be similar, so the effect of network properties on epidemic behaviour would be less clear. For example, R_0 values of 4 and 6 for an SIR epidemic would lead to 98.1% and 99.7% of nodes becoming infected respectively.

The probability of an infection occurring over an edge between an infected and a susceptible node during an entire infectious period is given by:

$$
\begin{aligned}
P(\text{infection}) &= p + (1-r)(1-p)p + (1-r)^2(1-p)^2 p + \ldots \\
&= p\sum_{i=0}^{\infty}(1-r)^i(1-p)^i \\
&= p\sum_{i=0}^{\infty}(1-r-p+rp)^i \\
&\quad \textit{Sum of infinite geometric series (Dwight 1961, pg 6 item 26.1)} \\
&= p\frac{1}{1-(1-r-p+rp)} \qquad \text{for } |1-r-p+rp| < 1 \\
&= \frac{p}{r+p-rp}
\end{aligned}
\tag{3.1}
$$

where: $p =$ probability of infection in timestep
$r =$ probability of recovery in timestep

For all simulations, susceptibility is set to 1 so that infection transmission depends only on infectivity rate. Infectivity levels (p in equation (3.1)) are 1/24, 2/24 and 3/24.

Recovery rate (r in equation (3.1)) is 1/3. That is, the probability of an infected node becoming immune (SIR) or susceptible (SIS) in a timestep is 1/3.

The distribution of infectious periods follows an exponential distribution with mean period of infection of 3 timesteps. This value is arbitrary, but must be long enough for nodes in clustered networks to be infected by a node directly and still be infected when a mutual neighbour becomes infected.

Thus, given that a newly infected node is connected to a susceptible node (and ignoring the potential for the susceptible node to become infected from some other node), the probabilities of infection across the edge are given by equation (3.1) and displayed in Table 3-2. The expected values for R_0 are also displayed, assuming the degree for all nodes in all networks is 8, each potential infection is independent, and ignoring the effects of clustering and assortativity in the network.

Table 3-2: Simulated infectivity rates and R_0 for degree 8

Infectivity rate	Probability of infection	R_0 ignoring network structure
1/24	3/26	0.9231
2/24	6/28	1.7143
3/24	9/30	2.4000

Immunity probability takes two values, 0 for SIS epidemics and 1 for SIR epidemics.

3.3 Summary of experimental design

Epidemic behaviour on networks is potentially analysed over several dimensions. These include network size, network degree features, other network properties and epidemic parameters such as infectivity. The actual data items for analysis also add dimensions.

These dimensions are summarised at Table 3-3, together with comments concerning how they are considered in the experimental design.

Table 3-3: Dimensions of simulated networks and epidemics

Aspect	Approach
Network features	
Network size (nodes)	Not varied (set at 1 000)
Mean degree	Not varied (target is 8)
Degree distribution shape	Normal, Real world, Power law
Variation in degree	Not analysed: averaged over networks
Degree assortativity coefficient	Varied, target values 0.0, 0.1, …, max
Clustering coefficient	Varied, target values 0.0, 0.1, …, max
Epidemic parameters	
Nodes infected at timestep 0	Not varied (set at 1)
Initial node(s) selection method	Random uniform used
Infectivity rate	3 values, all nodes: 1/24, 2/24, 3/24
Susceptibility rate	Not varied, all nodes 1
Recovery rate	Not varied, all nodes 1/3
Immunity response	2 values, all nodes: all (SIR) or none (SIS)
Epidemic behaviour	
Timestep	First 100 timesteps recorded
Prevalence (current status infected)	Relevant for SIS only
Epidemic size (Σ new infected)	Relevant for SIR only
Whether epidemic occurs	Calculated from simulation results
Epidemic derived R_0	Calculated from simulation results

Chapter 4: Network Generation

The motivation for this study is epidemic behaviour, where the relationship is defined by contact sufficient to transmit a specific disease. Several simplifying assumptions are made about the contact process:

- if two people are in contact, the disease can be transmitted from either person to the other;
- probability of transmission is independent of type and duration of contact;
- probability of transmission is equal between any pair of infected and susceptible persons; and
- contact patterns do not change.

The network consequences of these assumptions are that the networks of interest are undirected, static and unweighted. That is, the defining relationship has the following characteristics:

- the relationship is between two nodes, rather than from one node to the other;
- the relationship does not change over time, or at least the relationship changes in a much longer timeframe than the issues being investigated so the network can be considered to have a fixed structure; and
- the relationship either exists or does not exist, there is no consideration of strength of the relationship.

To use simulation to study the impact of social network properties on epidemic behaviour, many networks are required. Networks (both real and simulated) have many different properties. Furthermore, different networks can be similar in some ways and quite different in others. In order to investigate epidemic behaviour using simulated networks, it is important that

the relevant properties of real world social networks are replicated. The relevant properties selected for this study are:

- degree distribution;
- clustering coefficient; and
- (degree) assortativity coefficient.

As discussed in Section 2.4, existing network generation algorithms are not able to generate networks controlling for all three of these properties. Thus, a new algorithm is presented that enables targeting of these properties (the neighbour algorithm, described at section 4.2). The algorithm's performance is analysed to assess the success in satisfying property targets (section 4.3). This algorithm is then used to generate networks that allow examination of the separate and joint impact of these properties on epidemic behaviour (section 4.4).

Separately, an algorithm is required to generate uniform degree networks that implement the basic epidemiological model and provide an additional comparison point. This algorithm is presented first.

4.1 Generation of arbitrary degree simple connected networks

For the uniform degree basic epidemiological model implementation, an algorithm is developed as a modification to the Molloy-Reed algorithm described in Section 2.3.2.1. The modification is to ensure the generated network is both simple and connected. The network must be simple because multiple edges and self edges have no epidemiological interpretation. The network must be connected because the basic epidemiological model assumes there is a fixed probability of transmission between any infected and susceptible persons. No parameters are required apart from the degree sequence, and all nodes have the same degree for the required networks.

4.1.1 Description of algorithm

The simple connected algorithm is as follows (flowchart at Figure 4-1). The term residual degree means the difference between the intended degree of the node and the number of edges already established for that node, and any node with a residual degree of at least 1 is referred to as open.

1) Group the nodes into those that have degree at least 2 (multiple degree group) and those that have degree of 1;

2) Randomly order the nodes in the multiple degree group;

3) Start network with first node in the multiple degree group;

4) FOR EACH multiple degree node, attach the specified node to a node selected from the open nodes already in the network with probability proportional to residual degree (ensures connectivity by creating a spanning tree);

5) FOR EACH single degree node, attach the specified node to a node already in the network randomly selected in proportion to residual degree;

6) FOR EACH multiple degree node, close the network by attaching the specified node to as many open nodes as required (excluding the instant node and nodes to which it is already connected) with probability proportional to residual degree.

The algorithm may fail in the final step because the only open nodes would result in self edges or multiple edges. In this case, it restarts.

As an alternative to restarting, a rewiring process could be added whereby an existing edge is broken to provide access to other nodes. However, a rewiring process introduces a bias, as those simple networks with many 'close' self edge or multiple edge networks would be more likely to be generated. A rewiring process is not used in this study.

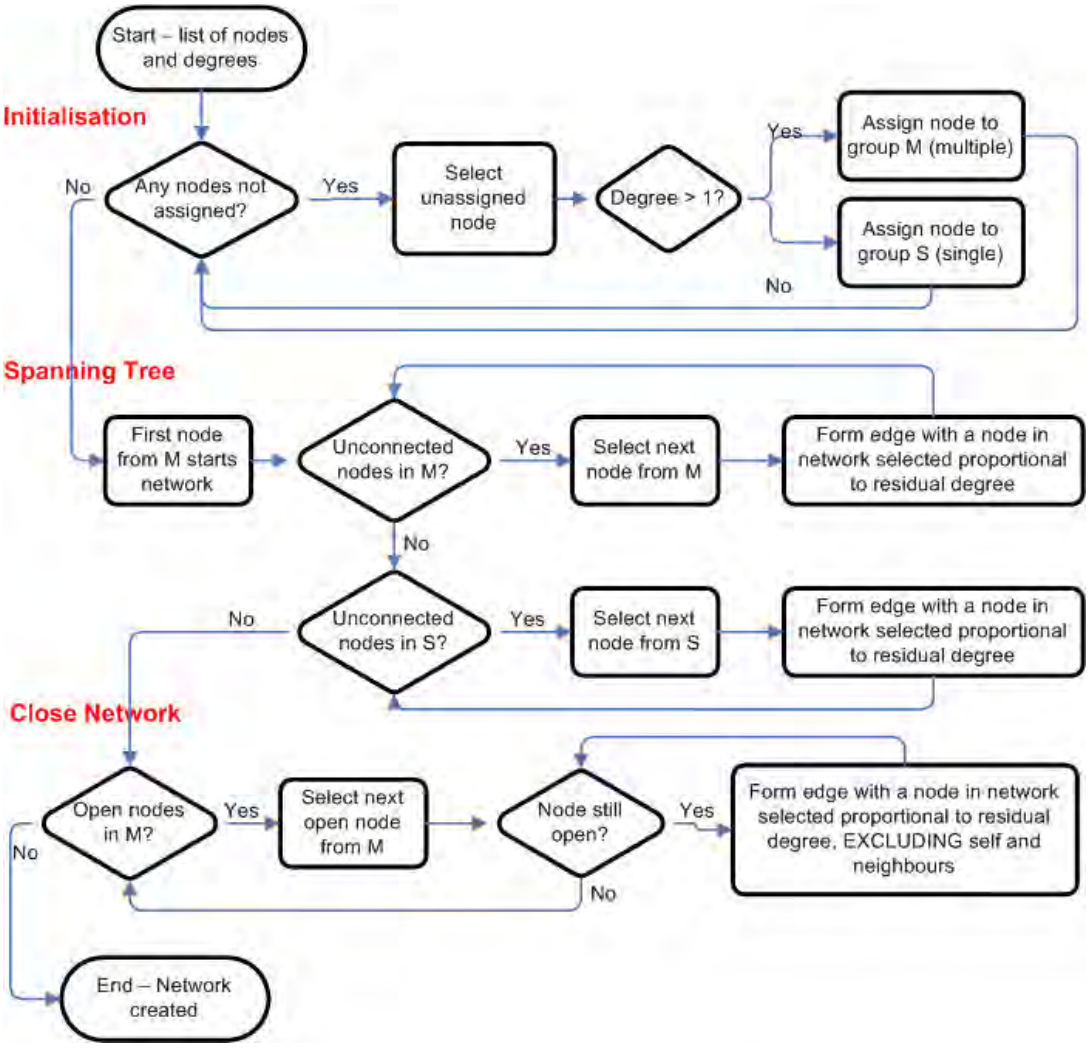

Figure 4-1: SC algorithm: Generating an arbitrary degree network that is simple and connected

4.1.2 Bias assessment

The use of the spanning tree introduces a potential bias to the networks generated by the simple connected (SC) modified Molloy-Reed algorithm. Thus, an experiment was conducted to compare the relevant properties of networks generated by the Erdös-Rényi (ER) random graphs algorithm (Erdös and Rényi 1960, described in Section 2.3.1) and the SC algorithm.

Networks were constructed by the ER algorithm with various edge probability settings (p = 0.002, 0.005, 0.01) and a size of 5 000 nodes. The degree

sequence was extracted and used as the input to the SC algorithm. Forty pairs of networks were randomly generated for each edge probability.

For each network property of interest, a paired t-test is conducted for each parameter set. The properties are:

- Clustering coefficient;
- Shortest paths: mean and maximum (diameter);
- Assortativity.

If the SC algorithm generates a truly random instance of a graph with the specified degree distribution, the properties of the ER and SC graphs should be the same.

Table 4-1 summarises the property values for each set of experiments and suggests there is little difference between the networks created by each algorithm. This conclusion is supported by the paired t-tests, with none rejecting the hypothesis that the property values are the same at the 90% significance level (that is, $p > 0.1$).

It is therefore not unreasonable to conclude that the SC algorithm generates a random simple connected network with the given degree distribution. As the full degree distribution is preserved, this method is better than merely retaining the giant component where these characteristics are important, for example in modelling epidemics.

For the uniform input degree distribution, the SC algorithm provides a network implementation of the basic epidemiological model. That is:

- all nodes have the same degree;
- an infection can access any node in the network because it is connected; and
- the edges are formed between nodes selected randomly in proportion to the difference between target degree and the number of edges already created that involve that node.

Table 4-1: Properties of generated networks – ER vs SC *(mean for parameter set, with standard deviation in ())*

Nodes	**5 000**	**5 000**	**5 000**
Probability of edge	**0.002**	**0.005**	**0.01**
Clustering coefficient ER	*0.0020*	0.0050	0.0100
	(0.0002)	(0.0001)	(0.0001)
Clustering coefficient SC	0.0020	0.0050	0.0100
	(0.0002)	(0.0001)	(0.0001)
Mean shortest path ER	3.95 (0.009)	2.93 (0.002)	2.59 (0.002)
Mean shortest path SC	3.95 (0.008)	2.93 (0.002)	2.59 (0.002)
Diameter ER	6.6 (0.5)	4.0 (0.0)	3.3 (0.4)
Diameter SC	6.7 (0.5)	4.0 (0.0)	3.2 (0.4)
Assortativity coefficient ER	0.001	0.000	0.000
	(0.007)	(0.004)	(0.003)
Assortativity coefficient SC	-0.003	-0.001	-0.001
	(0.007)	(0.004)	(0.003)

4.2 Neighbour algorithm: Generating networks with specific properties of interest

To generate networks for investigating the impact of network properties on epidemic behaviour, the relevant network properties must be directly entered as inputs to the algorithm, or related to input parameters in some predictable way. For this study, the properties of interest are degree distribution (either fully specified or defined by a probability function), assortativity and clustering coefficient.

Ideally, the generation algorithm would be able to create a network with a neutral impact on other network properties. That is, it would generate a uniform randomly selected instance of a network with the given degree distribution, degree assortativity and clustering coefficient. For practical reasons, the algorithm should also be efficient.

4.2.1 General approach

The proposed approach is inspired by the spatial models of Waxman (1988) and Keeling (2005) discussed in Section 2.4.1. Like these models, nodes are connected with some parameterisable probability that depends on distance. That is, nodes that are closer in the space are more likely to have an edge created between them. The space is then discarded.

However, in the proposed algorithm, the positioning of the nodes in that space depends on the desired assortativity. This is implemented through a layout modification stage so that nodes with similar degree are closer together and therefore more likely to have an edge created.

The neighbour algorithm therefore has three phases:

1) Initialisation (degree distribution): uniform randomly locating the nodes in space and assigning a target degree to each.

2) Layout modification (assortativity): moving nodes so that those with similar target degree are relatively close.

3) Edge creation (clustering coefficient): For each pair of nodes, create an edge with some probability that depends on the degree of each and the distance between them.

These phases map the network generation problem to (initialisation) and from (edge creation) physical space. In physical space, layout modification to target assortativity is much more straightforward than in network space. Hence, the mapping allows each property to be dealt with individually.

4.2.2 Implementation: One dimension with node swap

This general approach can be implemented in different ways. The method used for this study uses a one dimensional wrapped space (ring) as the notional space. Layout modification is implemented with stochastic conditional node swaps. For edge creation, edges for each node are created with fixed probability, tested from nearest to furthest nodes, until the

desired target degree is achieved. A flowchart of this implementation is at Figure 4-2.

4.2.2.1 Initialisation

Required information for the initialisation is the degree distribution, which also provides the number of nodes *N*. This can be fully specified with a count of the number of nodes with each degree or with a probability distribution.

Each node is assigned an identifier and a target degree randomly selected from the degree distribution or degree sequence. Each node is randomly assigned a position between 1 and *N* (inclusive).

The position is used to identify which nodes are in each neighbourhood in the later phases. Position *N* is next to position *N*-1 and position 1, so the position space is wrapped (the locations can be conceptualised on a ring).

4.2.2.2 Layout modification

The principle for the layout modification stage is that, because edge creation is more probable for closer node pairs, nodes are moved so that the nodes that should be connected to attain the desired assortativity are closer together.

Each layout update iteration compares the mean degree of the neighbourhoods of two uniform randomly selected nodes and swaps the location of the nodes only if such a swap would place the higher degree node in the higher degree neighbourhood. The neighbourhood is of size k/p, where k is the higher degree of the two nodes and p is the probability to be used in the edge creation phase.

The layout modification stage consists of some number of layout update iterations.

4.2.2.3 Edge creation

For each node, the edge creation stage creates all required edges to reach the node's target degree before moving to the next node, with nodes

considered in random order. For the randomly selected node, each node that does not already have its required number of edges and does not have an edge with the instant node is considered in location order: location of instant node +1 slot, then location of instant node -1 slot, then location of instant node +2 slots and so on. For the node being considered, an edge is made with the instant node with a set probability. If all nodes have been considered before the target degree is reached for the instant node, the nodes are considered again in the same order. If there are insufficient nodes that do not have edges with the instant node and have not reached their target degree, the algorithm fails.

4.2.2.4 Connecting the phases

While the edge creation phase uses the target degree, there is no natural stopping criterion for the layout update phase. The algorithm does not know when it has reached the target assortativity except by generating a network and measuring its assortativity.

Thus, there are three arbitrary evaluation points: how often to generate a test network, and after how many iterations to abandon the algorithm, and assortativity tolerance to accept the network. As implemented (for a 1 000 node network, Figure 4-2), a test network is generated every 500 iterations and accepted if actual assortativity is at least target assortativity minus 0.05. That is, the tolerance for assortativity is 0.05, but the layout iterations are also stopped if actual assortativity is too high, as there is no mechanism to reduce assortativity. The network is accepted regardless of its properties at 50 000 iterations, with an error message.

Thus, the implementation of the three phase approach is:

INITIALISE:

1) INPUTS: number of nodes, target degree for each node, target assortativity and edge creation probability

2) Randomly locate each node with an assigned target degree on a 1D ring

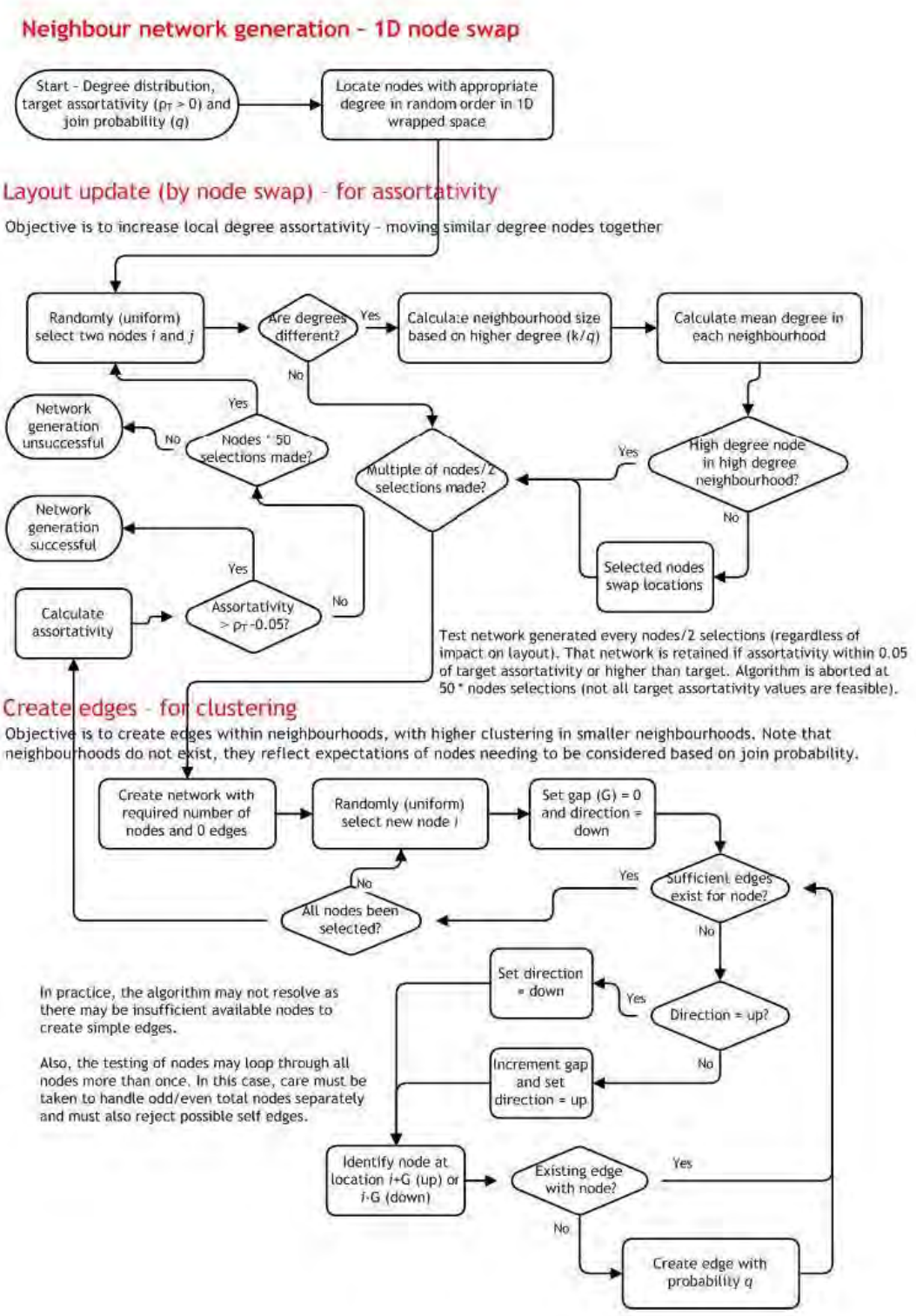

Figure 4-2: Neighbour network algorithm: 1D wrapped space with layout by swap

LAYOUT MODIFICATION:

3) FOR EACH of 500 pairs of nodes, if the higher degree node is in the lower degree neighbourhood, swap the locations of the nodes in the pair

EDGE CREATION:

4) FOR EACH node, with given probability, create edges with nodes until target degree is reached, starting with nearest nodes

TEST FOR COMPLETION:

5) Calculate assortativity of network

6) IF assortativity is less than target assortativity minus 0.5 and step 3 has been performed fewer than 10 times, return to step 3

7) ELSE network is retained and algorithm stops

Although arbitrary, these evaluation points are selected based on an initial investigation of the algorithm convergence.

4.3 Evaluation of neighbour algorithm

For the purposes of this study, the algorithm must generate networks with degree sequences from varied distribution families, a range of clustering coefficients and a range of assortativity values. Furthermore, these properties must be able to be analysed both jointly and independently, so multiple combinations of values are required.

In this context, the algorithm is valid if the implementation is able to control each of the target properties independently of the other two properties. It is reliable if the output network properties can be estimated from the input parameters.

Thus, some analysis of the networks generated by the algorithm is required to investigate issues such as the relationship between input parameters and network properties and the feasibility of generating networks with the desired

properties, to assess the suitability of the algorithm to generate the required networks.

4.3.1 Targeting of network properties

A single instance of a normal degree distribution and a power law degree distribution were generated. As for the networks to be used for epidemic simulation, these were generated using the Erdös-Rényi (Section 2.3.1) and Barabási-Albert (Section 2.3.4) algorithms respectively, with 1 000 nodes and mean degree 8. For each of these distributions, the neighbour network generation algorithm was run with a target assortativity of 1 and various values of edge probability (0.25, 0.5 and 1). Each 100 iterations of the layout update phase, 10 networks were generated and their properties measured. In addition, the number of nodes that are actually in a different location compared to the previous test point was recorded.

4.3.1.1 Assortativity: layout update phase

If all node pair tests led to a swap and no node was selected twice, there is a maximum of 200 node changes in each 100 iterations. As there is a nonzero probability of two nodes being selected having identical degree, fewer than half of the considerations could be expected to lead to swaps initially, with a decreasing number over time. This is consistent with the results at Figure 4-3. Fewer than 5% of nodes are being moved by iteration 20 000 for both the normal (ER) and power law (BA) distributions. That is, the algorithm is finding it increasingly difficult to identify potential increases in assortativity.

The question then is whether these location changes are having the intended impact on achieved assortativity. To this end, the mean assortativity for the 10 generated networks at each test point are shown at Figure 4-4. For iterations up to approximately 5 000, assortativity is steadily increasing as expected with the relative ease of node location changes. As swaps are reduced, the increase in assortativity also slows.

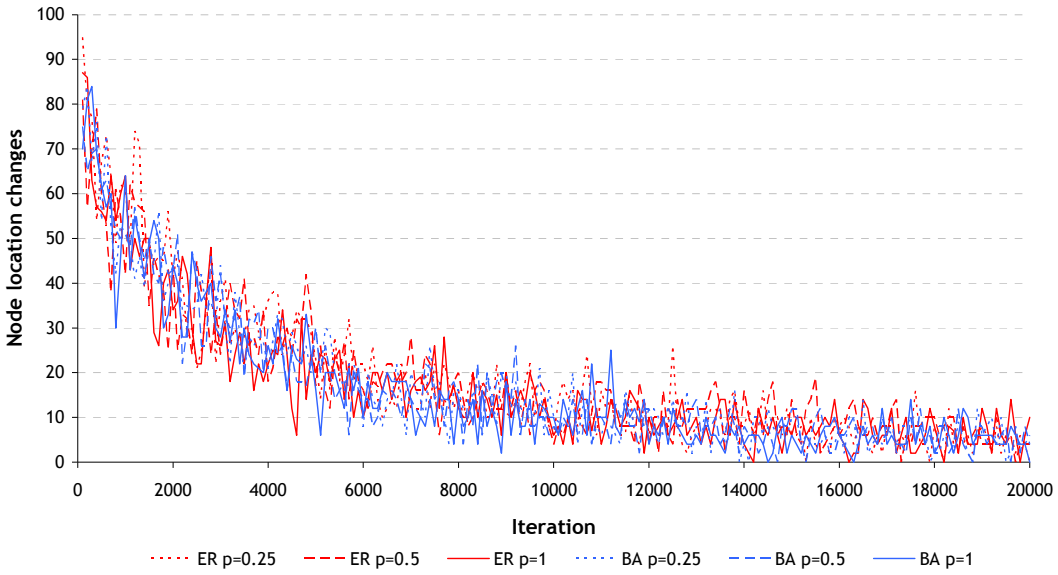

Figure 4-3: Layout update iterations: number of nodes in different locations *Individual plots are not important here, just the general shape. For both degree distribution types and all three edge creation probabilities, up to half the 200 considered nodes moved in the initial 100 layout update iterations and this decreases throughout the layout update phase.*

For the ER distribution, the maximum possible assortativity is 0.93 and this value is being approached by iteration 20 000. However, for the BA distribution, the maximum is 0.44 but additional iterations are having little impact reaching this level. Unlike the ER distribution, the higher edge creation probability also apparently impacts on the assortativity achieved for the BA distribution. If this is not a small sample artifact, one possible explanation is that the algorithm relies on the higher degree node to set the size of the neighbourhood for mean degree calculation to determine node swaps. For the BA distribution, the potentially extreme degrees could lead to big differences in the neighbourhood used for layout changes and the neighbourhood in which the actual edges are created.

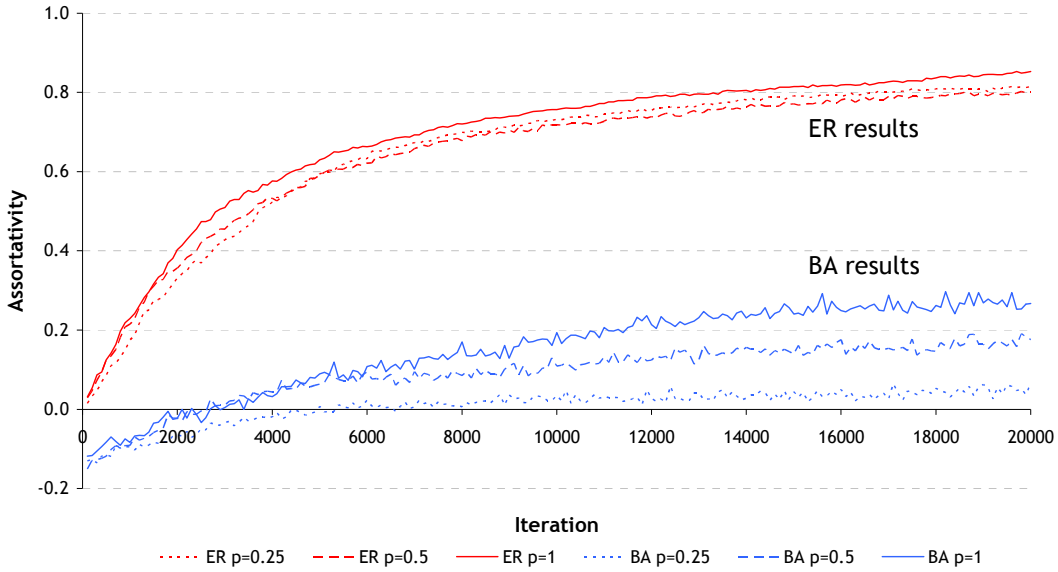

Figure 4-4: Layout update iterations: impact on assortativity

This issue of how much of the property space is actually accessible to the neighbour algorithm is revisited when the actual networks used for epidemic simulation are examined (Section 4.4.3).

4.3.1.2 Clustering coefficient: edge creation phase

The design of the algorithm intends assortativity and clustering to be independently controlled, with the layout updates influencing only assortativity and the edge creation probability influencing clustering. This is supported by Figure 4-5, with the mean clustering coefficient of the test networks maintaining a constant value regardless of layout update iterations.

The clustering coefficient value is slightly less than half the edge creation probability for both distribution types and all probability values. This highlights a weakness of the neighbour algorithm as implemented; an upper limit to the achievable clustering coefficient, achieved with edge creation probability of 1. From Table 2.1, this is too low to cover the full range of real world social networks.

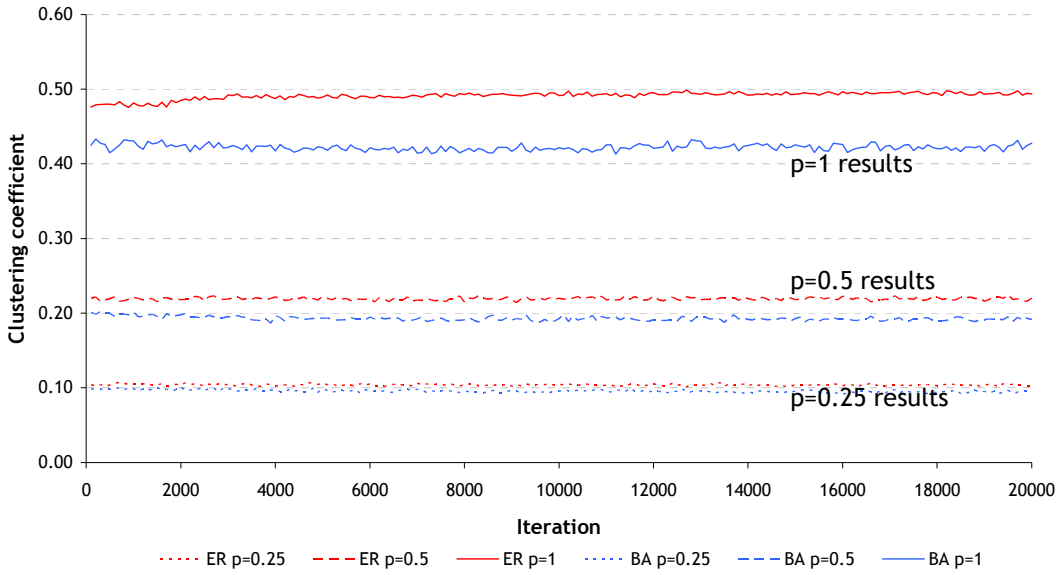

Figure 4-5: Layout update iterations: impact on clustering coefficient

Consider a ring lattice of some number at least $2m+1$ of identical nodes, each with degree $2m$. Select an arbitrary node and label it as location 0. That node will have edges with the nodes from location $-m$ to $+m$. The node at location $-m$ will have edges with nodes at $-(m-1)$ to 0 (as well as others that do not have edges with the node at location 0). Similarly, the node at location $-j$ (for $0 < j < m$) will have edges with nodes from location $-m$ to $+(m-j)$, and the node at location j will have edges with nodes from location $-(m-j)$ to m.

Since each edge is counted twice by this analysis of edges from one node to the other, the total number of edges between the neighbours of the node at 0 is given by:

$$\begin{aligned}\sum_{j=1}^{m}\left[(m-j)+m-1\right] &= m(2m-1)-\sum_{j=1}^{m} j \\ &= m(2m-1)-\frac{m(m+1)}{2}\end{aligned} \tag{4.1}$$

The number of possible edges between neighbours is:

$$\frac{\hat{k}\left(\hat{k}-1\right)}{2} \tag{4.2}$$

where: $\hat{k} = 2m$ is degree

Hence, clustering coefficient for a ring lattice with uniform degree is given by:

$$1-\frac{\hat{k}+2}{4\left(\hat{k}-1\right)} \tag{4.3}$$

where: $\hat{k}$ is (uniform) mean degree

This is the maximum clustering coefficient achievable for a network generated by the neighbour algorithm. Decreasing the edge creation probability or increasing the degree variation reduces the clustering coefficient achievable.

4.3.2 Stability of properties of generated networks

Having established that the neighbour algorithm does work as expected and set the stopping criteria, the next area of investigation is the stability of the properties of networks generated from fixed input parameters. That is, how similar are multiple networks generated from a single set of input parameters? Thirty neighbour networks were generated for each combination of the following input parameters:

1) Target mean degree 8 and 100 nodes

2) Distribution type: normal, real world, power law (using the same algorithms as are to be used for the epidemic simulation networks, defined in Section 3.1)

3) Target assortativity: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6

4) Edge creation probability: 0.2, 0.4, 0.6, 0.8, 1.0

4.3.2.1 Stability of degree sequence

At first examination, the neighbour algorithm performs very poorly in obtaining a network with the desired degree sequence for the skewed distributions (Table 4-2). For example, for the power law distribution, 36% of the generated networks have a mean degree less than 90% of the target mean degree. The variation in the degree is also reduced, with achieved standard deviation in degree of 4.9 instead of the target 6.1 for power law networks. Note that, by construction, the number of edges cannot exceed the edges in the target degree sequence.

Table 4-2: Stability of networks - Degree sequence

	Normal	Real world	Power law
Mean degree - target	8.00	7.81	7.80
Mean degree – achieved	7.93	7.44	7.09
Minimum edges	366	285	277
Maximum edges	400	497	390
Proportion <95% target edges	0.00	0.39	0.80
Proportion <90% target edges	0.00	0.08	0.36
Degree CV – target	0.34	0.66	0.78
Degree CV – achieved	0.34	0.63	0.69
Degree SD – target	2.74	5.19	6.06
Degree SD – achieved	2.70	4.68	4.90
Edges in top 5% nodes - target	8.8%	13.9%	18.1%
Edges in top 5% nodes – achieved	8.7%	13.2%	16.7%
Edges in top 5% nodes - <95% edges	9.1%	13.4%	16.5%
Edges in top 5% nodes - <90% edges	n/a	14.1%	16.3%

However, more detailed analysis suggests the flaws are not so substantial as to interfere with the capacity of the algorithm to generate networks suitable for investigation of epidemic behaviour. The general shape of the degree sequence is maintained, with the proportion of edges connected to the 5% highest degree nodes consistent even for those networks with fewer edges.

The degree truncation problem for the power law networks is also evident in examination of the edge representation of the top 5% of nodes, but the truncation is no worse for the lower mean degree networks than for the higher mean degree networks.

4.3.2.2 Stability of clustering coefficient

The networks generated by the neighbour algorithm show consistent values of the clustering coefficient for a given edge creation probability. The mean clustering coefficient increases linearly with edge creation probability and has a consistent standard deviation across degree distribution types of up to 0.03 (Table 4-3).

The highest coefficient of variation is 19% for real world with 0.2 edge creation probability. This high value reflects the low mean rather than being an indication of unusually high variation. The largest range is for real world networks with 1.0 edge creation probability, with a minimum clustering coefficient of 0.36 and a maximum of 0.52.

Table 4-3: Stability of networks – Clustering coefficient

		Edge creation probability				
		0.2	**0.4**	**0.6**	**0.8**	**1.0**
Normal	Mean	0.10	0.18	0.27	0.38	0.49
	SD	0.01	0.02	0.02	0.02	0.02
Real world	Mean	0.12	0.17	0.25	0.34	0.44
	SD	0.02	0.02	0.03	0.03	0.03
Power law	Mean	0.14	0.18	0.25	0.35	0.43
	SD	0.02	0.02	0.02	0.03	0.03

Overall, the algorithm generates networks with stable clustering coefficients. If targeting a particular value, a suitable starting edge creation probability would be twice the target clustering coefficient.

4.3.2.3 Stability of assortativity

For assortativity, the target value is the relevant input parameter to the network generation algorithm. A network is accepted if the achieved assortativity is 0.05 below the target assortativity or higher. Thus, at the end of the algorithm, as well as assortativity close to target, a generated network can have an assortativity higher than intended or that is too low but the algorithm was abandoned.

For the purposes of Table 4-4, target assortativity is achieved if the assortativity of the generated network is within 0.05, or correct to one decimal place. Except for the higher target values for the power law networks (which may not be feasible), the algorithm is able to successfully generate a network with the required assortativity in the majority of attempts.

Table 4-4: Stability of networks – Target assortativity achieved

	Target assortativity						
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
Normal	52%	71%	83%	87%	91%	85%	78%
Real world	66%	71%	73%	73%	70%	59%	55%
Power law	72%	61%	59%	45%	16%	3%	5%

For the normal distribution networks, for target assortativity values up to 0.4, the majority of failures arise because the achieved assortativity is too high, with the algorithm terminating before reaching the target assortativity being the larger contributor to failure for target assortativity of 0.5 and 0.6. For power law and real world distributions, positive assortativity values are more difficult to attain and termination is the major contributor for much lower target assortativity values of 0.1 and above (except 0.3 for real world).

4.3.2.4 Suitability for epidemic simulation

The neighbour algorithm generates networks with predictable clustering and assortativity properties. Further, broad ranges of values for these properties are able to be achieved. While generated networks have a reduced variation

of degree compared to the target for highly skewed degree distributions, the general shape is maintained and there remains substantial variation so the networks generated would allow results to incorporate degree variation.

Thus, the algorithm is suitable as a tool for generating networks with a broad range of socially relevant properties to enable investigation of epidemic behaviour by simulation.

4.3.3 Small-world property

Social networks exhibit the small-world property (Watts and Strogatz 1998), where the mean geodesic of networks is logarithmically related to the number of nodes (as compared to a linear relationship for random graphs) while maintaining high clustering coefficients. While the network property of mean geodesic is specifically excluded from examination in this study, the small-world property requires some investigation because it is a fundamental characteristic of social networks. The neighbour algorithm creates edges between nodes that are physically close, in the same way as lattice networks which do not exhibit this property.

To investigate this aspect, 10 neighbour networks were generated for each combination of the following input parameters:

1) Target mean degree 4, 8 and 12

2) Nodes 100, 500, 1 000, 5 000 and 10 000

3) Distribution type: uniform

4) Target assortativity: no layout updates (irrelevant as uniform degree)

5) Edge creation probability: 0.2, 0.4, 0.6, 0,8, 1.0

Uniform degree distribution was selected as any degree variation enables higher degree nodes to make relatively long edges and reduce mean geodesics. Uniform degree is thus the type of distribution for which the small-world property would be least evident for neighbour networks.

Since the edge creation probability defines the tightness of the lattice structure, mean geodesic can be expected to be directly related to this probability.

To satisfy the small-world property, networks generated by the neighbour algorithm must display a logarithmic relationship between number of nodes and mean geodesic within a fixed degree and edge creation probability parameter pair. For the networks with degree of 8, Figure 4-6 displays the number of nodes and mean geodesic of the generated networks. As the number of nodes is plotted on a logarithmic scale, the small world property is displayed as a linear plot.

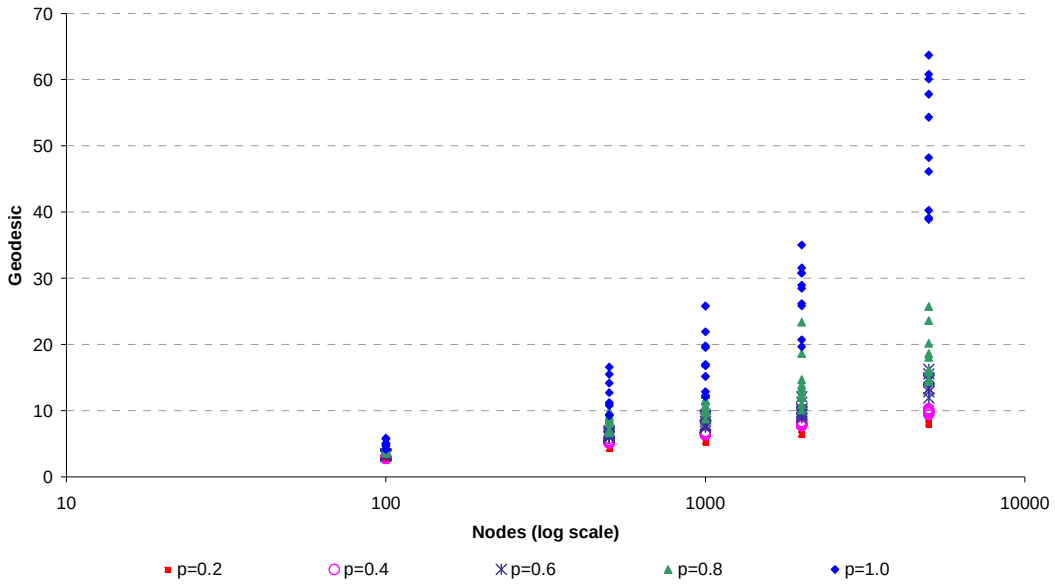

***Figure 4-6: Relationship between nodes and mean geodesic**: Uniform degree distribution, Neighbour algorithm, degree of 8*

From this figure, the relationship is clearly logarithmic for probability values of 0.2, 0.4 and 0.6 (R^2 > 90% for all three). It starts to break down for p = 0.8 (though logarithmic is a better fit than linear) and the small world property fails for p = 1. A similar pattern is also observed for the networks of other degree (Table 4-5). That is, the neighbour networks satisfy the small-world property for moderate edge probabilities, but the relationship breaks down for very high edge creation probabilities.

Table 4-5: Nodes versus mean geodesic – Logarithmic or linear?

	Edge creation probability				
	0.2	**0.4**	**0.6**	**0.8**	**1.0**
Degree 4	log	log	log	linear	linear
Degree 8	log	log	log	log	linear
Degree 12	log	log	log	log	linear

While the small-world property is able to be maintained for moderate clustering coefficients, some rewiring would be necessary where both small-world and very high clustering are required for generated networks. Watts and Strogatz's paper suggests a small rewiring process (1% to 10% of nodes) has a substantial impact on mean geodesic, but very little on clustering. Any impact on assortativity could be limited by rewiring to the same degree nodes where possible.

4.4 Networks generated for epidemic simulation

As a reminder of the key points from the experimental design, 10 instances of each of three degree distribution types are generated for 1 000 nodes and mean degree of 8. For each instance, one neighbour network is generated for each feasible combination of 11 clustering coefficient values and 11 assortativity values. Table 4-6 summarises key properties of the networks used for the epidemic simulations.

In addition, 10 networks are generated with uniform degree using the SC algorithm. These are to allow comparison with a network implementation of the basic epidemiological model.

Table 4-6: Properties of generated networks

Property	*Normal*	*Real world*	*Power law*
Mean proportion in giant component	1.00	0.98	1.00
Mean of mean degree[a]	7.99	7.58	7.67
Mean of degree CV	0.35	0.67	0.97
Mean of degree in top 5% nodes	8.9%	14.4%	21.2%
Max potential assortativity	0.955	0.926	0.527
Min achieved assortativity	-0.066	-0.098	-0.168
Mean achieved assortativity	0.384	0.346	0.086
Max achieved assortativity	0.864	0.863	0.490
Min achieved clustering	0.008	0.012	0.019
Mean achieved clustering	0.246	0.240	0.249
Max achieved clustering	0.510	0.446	0.463

a The neighbour algorithm is not guaranteed to create all desired edges. Thus, there are some differences between target and actual degree sequences.

4.4.1 Giant component or entire network?

While an epidemic can only be transmitted within a component of a network, the whole network was retained for the simulations rather than only the giant component. This was to maintain, to the extent possible, the experimental design, which calls for the same degree sequence to be used to generate networks with different values of assortativity and clustering coefficient.

As the only distribution with a considerable number of very low degree nodes, the real world distribution networks were the only group to have significant numbers of networks where the giant component included less than 95% of nodes (11.5%). The proportion increased for higher clustering and/or assortativity values, but the mean proportion of nodes in the giant component exceeded 90% for all property values.

Two aspects of epidemic behaviour are examined in this study (see Chapter 5), occurrence of an epidemic and basic reproduction ratio (R_0) derived from epidemic size.

Components that are not the giant component are generally too small to allow an epidemic that meets the definition developed in Section 5.1. For epidemic occurrence, retaining the whole network increases the likelihood that an epidemic does not occur because there is a non-zero probability that the initial node for infection will be in a small component. However, this is also true for real world epidemics and the full network is the more valid equivalent for simulation than just the giant component.

As the only successful epidemics will occur in the giant component, the number of nodes infected during the epidemic is not affected by whether the full network or giant component is used. However, epidemic derived R_0 is based on infected proportion. Using the nodes in the giant component for this conversion would more faithfully preserve the theoretical relationship between R_0 and epidemic size. However, epidemiologists do not have access to information about the size of the giant component and using the full network, or population size, more faithfully reflects the practice of deriving R_0 from an epidemic.

4.4.2 Degree sequence: target versus generated

Networks were used even if the algorithm failed to resolve. That is, if a node had insufficient edges to meet its target degree, but the only nodes available for connection would lead to a self edge or a multiple edge, edge creation was abandoned and the network was retained with the reduced number of edges.

Like the networks generated to investigate property stability (see Section 4.3.2.1), the simulation networks for the skewed degree distributions have lower degree, reduced degree variation and reduced impact of the highest degree nodes than targeted (Table 4-7). These effects are more pronounced in the power law than real world degree distribution types.

While the effect of degree variation has been blunted by the neighbour network generation algorithm, there remains a substantial skewing even in those networks where the fewest edges were able to be created. Thus, the

networks are sufficient to investigate differences in simulated epidemic behaviour arising from differences in the shape of the degree distribution.

Table 4-7: Properties of experimental networks - Degree sequence

	Normal	Real world	Power law
Mean degree - target	8.00	7.81	7.80
Mean degree – achieved	7.93	7.44	7.09
Minimum edges	366	285	277
Maximum edges	400	497	390
Proportion <95% target edges	0.00	0.39	0.80
Proportion <90% target edges	0.00	0.08	0.36
Degree CV – target	0.34	0.66	0.78
Degree CV – achieved	0.34	0.63	0.69
Degree SD – target	2.74	5.19	6.06
Degree SD – achieved	2.70	4.68	4.90
Edges in top 5% nodes - target	8.8%	13.9%	18.1%
Edges in top 5% nodes – achieved	8.7%	13.2%	16.7%
Edges in top 5% nodes - <95% edges	9.1%	13.4%	16.5%
Edges in top 5% nodes - <90% edges	n/a	14.1%	16.3%

4.4.3 Feasible assortativity and clustering coefficients

Table 4-8 displays the number of networks generated by the neighbour algorithm and used for the epidemic simulations, by approximate assortativity and clustering coefficient. Note that these property values are the achieved, not target, values. There are 1 111 neighbour networks in total distributed between the various degree distribution types, assortativity and clustering coefficients.

From Table 4-8, there are some combinations of network properties for which at least some of the possible 10 networks could not be generated. As expected from the considerations described in Section 4.3.1.2, the more skewed distributions are unable to achieve the same clustering coefficients as the networks with normal degree distribution.

Table 4-8: Number of neighbour networks for epidemic simulation, by network properties

		Clustering Coefficient					
Distribution	**Assortativity**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
normal	0.0	10	10	6	3	1	1
	0.1	10	10	10	10	10	10
	0.2	10	10	10	10	10	10
	0.3	10	10	10	10	10	10
	0.4	10	10	10	10	10	10
	0.5	-	10	10	10	10	10
	0.6	-	10	10	10	10	10
	0.7	-	10	10	10	10	10
	0.8	-	10	10	10	10	10
	0.9	-	9	10	10	10	10
real world	0.0	10	10	10	10	9	-
	0.1	10	10	10	10	10	-
	0.2	1	10	10	10	10	-
	0.3	-	10	10	10	10	-
	0.4	-	10	10	10	10	-
	0.5	-	10	10	10	10	-
	0.6	-	10	10	10	10	-
	0.7	-	10	10	10	10	-
	0.8	-	10	10	10	10	-
	0.9	-	-	8	10	10	-
power law	0.0	7	10	10	10	10	1
	0.1	-	10	10	10	10	-
	0.2	-	8	10	10	10	-
	0.3	-	3	10	10	10	-
	0.4	-	1	6	7	7	-
	0.5	-	-	-	2	1	-

There is also a weak relationship between achievable values of the two properties. For normal degree distribution, the neighbour algorithm has least success in generating networks with high values of one property and low

values of the other. For real world and power law degree distribution the maximum assortativity achieved is lower for a lower clustering coefficient.

Power law networks exhibit the greatest restriction on achievable properties. The maximum assortativity is approximately 0.5 instead of approximately 0.9 for the other degree distribution types. However, this is more a reflection of the degree distribution than the neighbour algorithm. Given a specific degree sequence, the simple network with the maximum possible assortativity is that constructed by the Havel-Hakimi algorithm. In this algorithm, the nodes are ordered by descending degree and the highest degree node is connected to the appropriate number of nodes from the next highest degree down (Havel 1955; Hakimi 1962). This process continues through the nodes. Table 4-9 shows the assortativity of the Havel-Hakimi network and the maximum assortativity of networks constructed with the neighbour algorithm for various approximate clustering coefficients.

Table 4-9: Maximum assortativity achieved

	Normal	**Real world**	**Power law**
Clustering ~0.0	0.369	0.154	-0.026
Clustering ~0.1	0.857	0.819	0.374
Clustering ~0.2	0.859	0.858	0.429
Clustering ~0.3	0.864	0.861	0.490
Clustering ~0.4	0.863	0.863	0.475
Clustering ~0.5	0.862	n/a	-0.044
Maximum feasible	0.955	0.926	0.527

Referring again to Table 4-8, for almost all values of assortativity or clustering, there are 10 networks for each of several values of the other property. Thus, the networks provide a suitable range of independent property values to build a model of impact on epidemic behaviour.

4.5 Discussion

Previously published algorithms have been able to generate networks with only specific ranges of interactions between degree distribution, assortativity and clustering coefficient (Section 2.4). However, independent control of each of these properties is required to examine the interaction of these properties.

The network generation neighbour algorithm developed in this chapter has a three phase approach. Each of these phases influences one of the target social network properties. The initialisation phase uniform randomly locates nodes in a notional space and assigns a target degree to each node. The layout modification controls assortativity by moving nodes with similar target degrees nearer to each other. The edge creation phase creates edges with given probability between nodes that are located near each other in the notional space provided they have not reached their target degree. The search for potential edge partners starts with the closest nodes and moves further away. Thus, a lower probability leads to a larger search space and smaller clustering coefficient.

In developing this algorithm, the objective was to generate networks with a range of values for the properties of interest. As such, within the general approach described, a suitable implementation was developed and validated. The implementation used a one dimensional ring as the notional space and stochastic node pair swaps for layout update.

Validation analysis confirmed the independent control over the three targeted social network properties. Test networks generated at regular intervals during layout modification displayed increasing levels of assortativity and a constant level of clustering.

The neighbour algorithm was also able to generate networks with property values close to those targeted for each property.

For degree distribution, if the algorithm successfully resolves, the generated network will have the same degree sequence as the target degree sequence.

However, the validation analyses demonstrated that the algorithm only rarely resolves and only where degree heterogeneity is low. If networks are accepted without resolution, the generated networks show the same skewed shape as the target degree sequence but mean degree and heterogeneity are both reduced. The reduction is greatest for the most highly skewed distribution types.

The difficulty is that the highest degree nodes are unable to find sufficient free nodes with which to make edges. This problem could be alleviated by ordering the edge creation phase so that higher degree nodes are connected first. However, consider the networks with edge creation probability of 1. The higher degree nodes would connect to all their immediate physical neighbours. Thus, a low degree node near several high degree nodes would not have the opportunity to connect to other low degree nodes, potentially introducing a bias in both assortativity and clustering coefficient of the generated networks. The random ordering implemented avoids this bias but at the cost of lower rates of resolution.

For assortativity, the algorithm is inelegant in its convergence as implemented. In particular, the edge creation phase is implemented periodically throughout the layout modification phase to generate test networks. If the test network has an assortativity that is near to the target assortativity, that network is retained as the generated network. If the assortativity of the test network is too low, the layout modification phase continues and a further test network is generated. However, if the test network assortativity is too high, the algorithm is abandoned because there is no mechanism to reduce assortativity. In general, over half the networks generated had assortativity values within the tolerance of 0.05 of the target values. This could be improved by introducing a mechanism to reduce assortativity, such as swapping a certain number of randomly selected pairs of nodes.

The algorithm is successful in the scope of achievable assortativity values. For all distribution types, it was able to generate networks with over 90% of the maximum value feasible given the degree sequence.

For clustering coefficient, the neighbour algorithm is able to consistently generate networks with very similar properties using the same input edge creation probability parameter. Further, there is a linear relationship between the edge creation probability and the clustering coefficient of the generated network, which simplifies the use of the neighbour algorithm to generate networks with specific target properties. If targeting a particular value, a suitable starting edge creation probability would be twice the target clustering coefficient. This could be increased or decreased if the networks being generated had clustering coefficients that were too low or high respectively.

However, there is a significant limitation in the capacity of the neighbour algorithm to generate networks with specific clustering coefficients. The maximum feasible clustering coefficient is approximately 0.5 and is reduced in the presence of degree variation. As some real world social networks have clustering coefficients higher than this value, this restriction does impact on the study of epidemic behaviour.

There is also some evidence that the neighbour algorithm is restricted in the combination of assortativity and clustering coefficient values able to be achieved. The algorithm was unable to generate higher assortativity networks in combination with a clustering coefficient near zero.

This could be a structural limitation rather than an artifact of the algorithm, particularly for smaller networks. Higher assortativity requires nodes of similar degree to have edges between them. If there are only a small number of nodes with similar degree for parts of the degree sequence, this will also lead to clustering. There is some evidence supporting this interpretation from the Keeling algorithm (Keeling 2005; Badham et al. 2007), which also has difficulty in separating the values of assortativity and clustering coefficient,

and also from a study of feasible property space for a specific degree sequence (Holme and Zhao 2006).

In addition to the properties for which the neighbour algorithm was developed, the test networks were also assessed for their compliance with the small world property. Social networks are known to exhibit the small-world property (Watts and Strogatz 1998), but lattice networks do not and the neighbour algorithm uses lattice like construction methods.

Neighbour networks satisfy the small-world property for moderate edge probabilities, but for very high edge creation probabilities, mean geodesic increases linearly with network size.

This suggests that some rewiring would be necessary where both small-world and very high clustering are required for generated networks. Watts and Strogatz's (1998) paper suggests a small rewiring process (1% to 10% of nodes) has a substantial impact on mean geodesic, but very little on clustering. Any impact on assortativity could be limited by rewiring to the same degree nodes where possible.

The neighbour algorithm demonstrated its suitability for generating networks for simulation studies of the impact of degree distribution, assortativity and clustering coefficient by generating a suitable sample for this study of epidemic behaviour. Multiple networks were obtained for a range of assortativity and clustering values for each type of degree distribution, which enables the impact of these properties to be assessed separately and jointly.

Chapter 5: Epidemic Simulation

This chapter investigates the way in which epidemic behaviour is affected by differences in the properties of the network: positively skewed degree distribution, (degree) assortativity and clustering coefficient. As described in Chapter 3, up to 100 SIR and SIS epidemic simulations are run for each simulation set defined by network property (degree distribution type, assortativity and clustering coefficient) and epidemic parameter (infectivity and immunity). In total, 66 720 simulations are run on neighbour networks generated by the neighbour algorithm developed in Chapter 4, with another 600 simulations to implement the basic epidemiological model in a network context.

Two aspects of epidemic behaviour are examined. The first of these is the proportion of simulations in which an epidemic occurs. This will be based on the definition of an epidemic as developed in Section 5.1.1.

The second examination concerns the intensity of the epidemic, given that an epidemic occurs. This uses the standard epidemiological measure of basic reproduction ratio R_0 derived from the size of the epidemic. For the SIR simulations, an epidemic will eventually die out because available nodes are immune. Thus, the cumulative infections at the end of the epidemic is an appropriate measure of size (referred to as final size). For the SIS process, the epidemic will eventually stabilise so that there is a relatively constant number of infected nodes. Thus, prevalence after stability is achieved is an appropriate measure of size. Both of these size measures can be used to estimate the underlying value of R_0 for the epidemic.

5.1 Analytical framework

For the simulations, the probability of infection transmission given an edge between an infected and a susceptible node is shown in Table 3.2. Note that

this value ignores lower probability events such as the infected node transmitting the infection more than once to the same node. It also ignores the possibility of the susceptible node becoming infected from some other source.

From equations (2.3) and (2.4), there is a relationship between R_0 and the observable epidemic property of final size (for SIR) or prevalence (for SIS), expressed as the proportion of nodes. Rewriting so that R_0 is a function of observable properties:

For SIR epidemics (Diekmann & Heesterbeek 2000,pg 13 eq 1.11):

$$R_0 = {-\log_e(1-f)}\big/{f} \tag{5.1}$$

where: f is final size, the proportion of the population ever infected

For SIS epidemics (Anderson & May 1992, pg 17 eq 2.1):

$$R_0 = {1}\big/{1-p} \tag{5.2}$$

where: p is the prevalence at equilibrium as a proportion of the population

While these relationships are invalid once degree variation is introduced (see Section 5.7.1 for discussion), prevalence and final size information is not available by degree in real world epidemics. Thus, to measure the impact of network properties, the observed value of R_0 is calculated from the simulations with known assortativity and clustering, and compared to the value over networks with the same degree distribution but zero values of those properties.

The experimental dataset is summarised at Table 5-1. The main analysis dataset is the neighbour networks with specific assortativity and clustering coefficient values for each of three degree distribution types. This is described more fully in Section 5.1.1. The properties of the networks have already been described in Section 4.4. There are also 10 networks with

uniform degree, which are used to implement the basic epidemiological model for selected comparisons.

Table 5-1: Number of records used for analysis

	Uniform	Real	Power	Normal	Total
Networks generated	10	409	183	520	1 122
60 simulations per network (3 infectivity rates x SIR/SIS x 10 instances)					
Simulations performed	600	24 540	10 980	31 200	67 320
Timestep records (million)	0.06	2.45	1.10	3.12	6.673

Note: The actual number of records in the dataset is 8.24 million as it also includes epidemic simulations on some additional comparison networks.

There are properties not controlled by the experimental design that are also changed as assortativity and clustering vary, at least some of which are likely to impact on epidemic behaviour (such as mean geodesic). The impact of these factors is not specifically considered in this analysis.

The supplementary DVD includes the C++, Matlab and SPSS code used for the major analyses, datasets and analysis output in Excel, Matlab or SPSS. Where important results were produced in SPSS, the output is also available in HTML. A description of the analysis process and files used is at Appendix B.

5.1.1 Analysis dataset

To examine the impact of network structure, the analysis dataset comprises epidemic simulations over neighbour networks with a range of assortativity and clustering values. The analysis will consider each type of degree distribution and infectivity value separately. The number of simulations and the number of those simulations that are epidemics (defined in Section 5.1.3), are displayed in Table 5-2.

For simplicity, the detailed analysis is generally presented for only a single simulation set or property combination, though summary results are presented for all cases. Where only selected results are presented, the results for all simulations are included on the supplementary DVD and indexed in Appendix C. Any conflicting results are noted in the text.

Table 5-2: Number of simulations and epidemics in dataset

	Infectivity rate		
	0.0417	**0.0833**	**0.1250**
Simulations			
Normal	5 200	5 200	5 200
Real world	4 090	4 090	4 090
Power law	1 830	1 830	1 830
SIR Epidemics			
Normal	326	2 152	3 506
Real world	480	1 573	2 331
Power law	276	627	900
SIS Epidemics			
Normal	711	2 797	3 830
Real world	719	1 863	2 554
Power law	322	761	1 083

The simulation set described fully is that for the real world distribution with highest infectivity rate. The real world distribution was selected as the normal distribution is socially unrealistic and the power law distribution has a smaller range of property values available for analysis. The highest infectivity value was selected as it has the highest proportion of successful epidemics, and therefore the richest dataset when considering the impact on basic reproduction ratio.

For analyses where assortativity and clustering coefficient are held constant, the values selected for presentation will be 0.2 and 0.4 respectively as these are moderate values for published social networks (see Table 2.1). The analysis dataset includes networks with assortativity of up to 0.9, so 0.2 is within the property space. For clustering coefficient, 0.4 is a very high value with respect to the networks able to be generated but is included in the simulations.

The dataset also includes simulations on 10 uniform degree networks. These networks implement the basic epidemiological model within the same

network modelling framework as the simulations to be used for the analysis of relationships. The number of simulations and epidemics in the dataset for these networks is at Table 5-3.

Table 5-3: Number of uniform degree simulations and epidemics in dataset

	Infectivity rate		
	0.0417	**0.0833**	**0.1250**
Simulations	100	100	100
SIR Epidemics	9	54	82
SIS Epidemics	4	59	74

5.1.2 Epidemic definition: Empirical reproduction ratio

In order to assess whether an epidemic occurs, some definition is required for whether the epidemic 'occurred' or 'failed'.

The generally accepted definition of epidemic amongst epidemiologists is disease specific:

> The occurrence in a community or region of cases of an illness … clearly in excess of normal expectancy (Last 2001, pg 60).

Thus, relevant organisations such as the United States' Centre for Disease Control set epidemic thresholds that are specific to the disease, available information and location. For example, the (US) influenza epidemic threshold is that the proportion of deaths attributable to pneumonia and influenza is in excess of 1.645 standard deviations above the seasonal baseline percentage (Center for Disease Control 2007).

The epidemic threshold theorem (Kermack and McKendrick 1927; Diekmann et al. 1990) suggests a more relevant potential definition. An epidemic occurs when an infected node in a susceptible population is able to infect, on average, at least one node during the period of its infection or equivalently, that the basic reproduction ratio R_0 is greater than or equal to 1. However, this measure is not observable from a single simulation, so cannot be used to determine whether an epidemic occurred.

Instead, define timestep specific empirical reproduction ratio (E_t) as the number of new infections arising in the next infectious period per current infection. That is:

$$E_t = \frac{\sum_{i=t+1}^{t+1/r} J_i}{I_t} \tag{5.3}$$

where: E_t is the empirical reproduction ratio at timestep t
J_i is the new infections at timestep i
r is recovery rate
I_t is the number of nodes infected at timestep t

Note that this is not a measure of reproduction per generation; such a measure would require tracking of the number of nodes that each node directly infects. In contrast, the presented ratio excludes infections already achieved by the currently infected nodes, excludes infections achieved by the currently infected nodes more than an average generation time in the future and includes infections achieved by nodes infected by the currently infected nodes, provided the infections occur quickly. In addition, timestep is a discrete measure so E_t will incorrectly estimate the actual reproduction ratio unless $1/r$ is an integer. However, E_t is an observable measure of epidemic behaviour.

What is the appropriate timestep at which to test for reproduction at least one and hence epidemic success?

If the test is too early, false negatives could arise where a later generation 'kickstarts' the epidemic. Alternatively, false positives could occur where the initial success is not able to develop into a full epidemic. This latter error may be a particular problem on highly clustered networks, where the infected nodes may be trapped in areas where all their neighbours are immune (for SIR) or already infected (SIS and SIR).

Another problem arises because the basic reproduction ratio is a theoretical measure based on a completely susceptible population. Once a simulation has started, the population is no longer completely susceptible and reproduction per infection decreases. Thus, testing too late could lead to false negatives, where an epidemic occurred but is no longer producing sufficient new infections to appear to be an epidemic.

For this study, the preliminary definition of an epidemic is that an epidemic occurs if the empirical reproduction ratio is at least 1 for some timestep:

$$\exists t \text{ such that } E_t \geq 1 \tag{5.4}$$

This is implemented by calculating the achieved reproduction ratio (E_t) for all timesteps after a short run in time (15 timesteps ≈ 5 generations). The run in time is intended to avoid the value being determined by a single high degree or otherwise successful node in the period where a single node can dominate the ratio. The maximum over these timesteps is then used to assess whether an epidemic succeeds. An epidemic is considered to have occurred where $E \geq 1$.

$$E = \max_t E_t \Big|_{t=15}^{t=98} \tag{5.5}$$

5.1.3 Epidemic definition: Assessment and refinement

While the operational epidemic definition has a theoretical justification, it must also meet conditions of operational utility and validity.

The first question of validity is whether the definition produces expected results in relation to infectivity. That is, is increased infectivity associated with a higher proportion of epidemic occurrence? This threshold test is satisfied for both SIR and SIS simulations for each of the four types of degree distribution (Figure 5-1 and Figure 5-2 respectively).

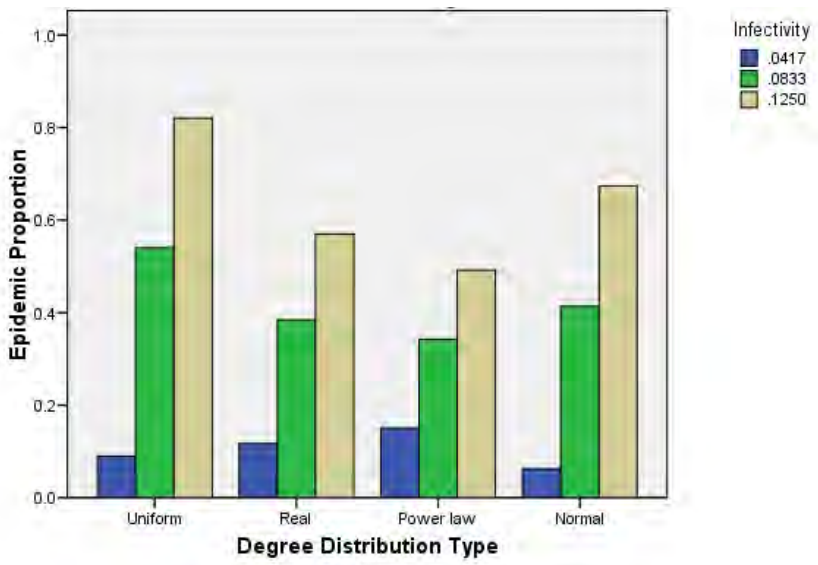

***Figure 5-1: Epidemic occurrence by infectivity, SIR**: Comparisons across degree distribution types are not valid as each has a different proportion of networks with higher values for assortativity and clustering.*

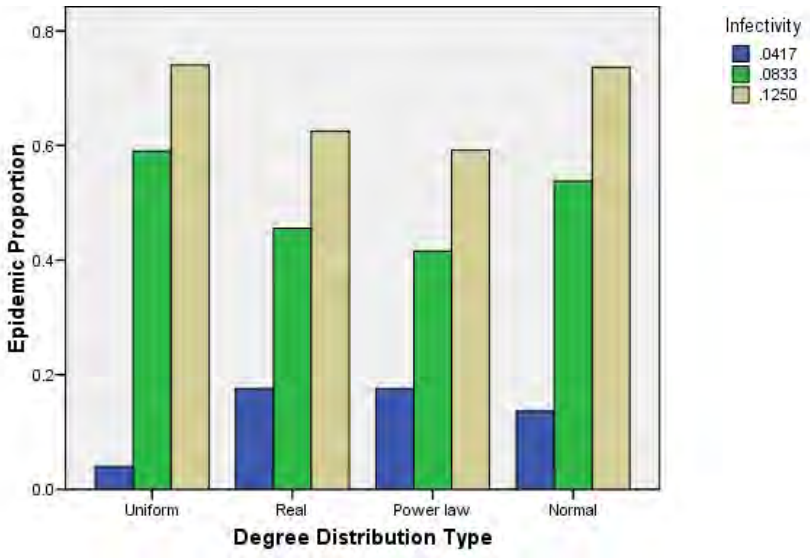

***Figure 5-2: Epidemic occurrence by infectivity, SIS**: Note that comparisons across degree distribution types are not valid as each has a different proportion of networks with higher values for assortativity and clustering.*

The next question concerns the extent to which the definition distinguishes between successful and failed epidemics. Figure 5-3 displays the distribution

of the maximum reproduction ratio (in the simulation period of 100 timesteps) over the 33 660 SIR epidemic simulations, with the equivalent information for SIS at Figure 5-4. Both of these figures show a substantial difference (note the logarithmic scale) in the number of simulations with reproduction ratios in the range [0,1) compared to [1,2), which provides support for the epidemic definition.

Table 5-4 provides limited further support for the definition of epidemic. It shows general agreement between the proposed theoretically based definition and whether more than 33 total infections were achieved in the first 33 generations. However, these classifications are not independent and the table also raises concerns about false positives and negatives.

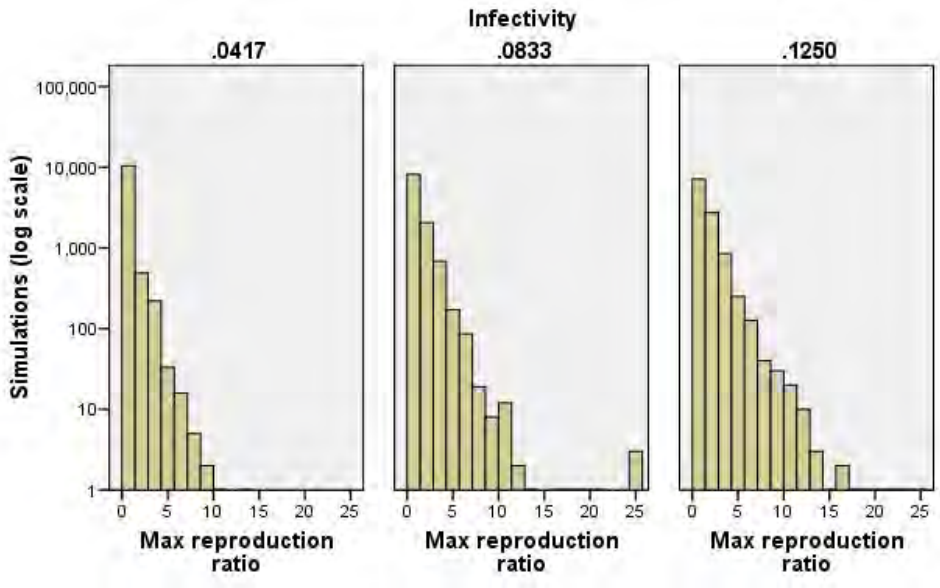

Figure 5-3: Number of simulations by reproduction ratio, SIR *The reproduction ratio scale has been truncated at 27. The maximum value that occurs in the dataset for SIR simulations is 92 and there are 8 simulations excluded.*

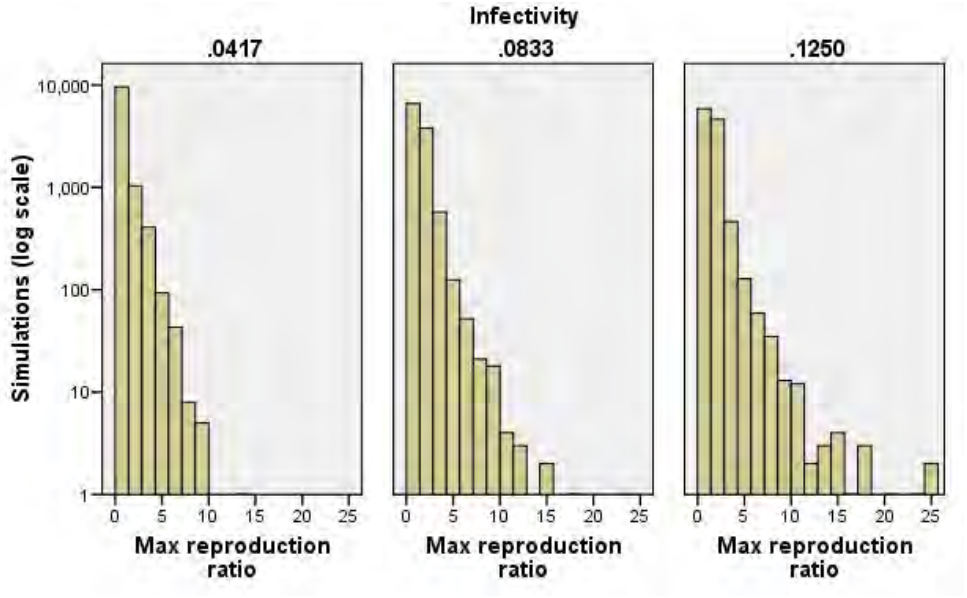

Figure 5-4: Number of simulations by reproduction ratio, SIS *The reproduction ratio scale has been truncated at 27. The maximum value that occurs in the dataset for SIR simulations is 97 and there are 5 simulations excluded from the chart.*

For SIR simulations, 1.6% of those simulations meeting the definition of an epidemic do not achieve 34 total infections in the first 33 generations, so the reproduction is not maintained. For SIS epidemics, only 1.0% are not maintained. That is, the epidemic reproduces but then becomes trapped by already infected (and immune, for SIR) nodes and dies out.

False negatives for SIR simulations are a much greater problem with the definition. Of those simulations identified as not epidemics, 8.6% of SIR but only 0.1% of SIS runs have achieved 34 infections by timestep 100. Furthermore, many of these simulations infect a substantial proportion of the population, with at least 500 infections achieved by 1 396 SIR simulations not defined as epidemics. This situation arises where significant growth occurs during the run in period (first 5 generations).

Table 5-4: Potential incorrect epidemic classification

	SIR		SIS	
	Number	**Percent**	**Number**	**Percent**
Not epidemic, infections > 33	2 897	8.6%	18	0.1%
Not epidemic, infections ≤ 33	21 282	63.2%	18 876	56.1%
Epidemic, infections > 33	8 951	26.6%	14 431	42.9%
Epidemic, infections ≤ 33	530	1.6%	335	1.0%
Total	33 660	100.0%	33 660	100.0%

To reduce this problem, maximum growth rate was considered as an alternative epidemic indicator. Define the timestep specific growth rate of the epidemic as the average growth rate per generation up to that timestep:

$$G_t = \left(\frac{\sum_{i=0}^{t} J_i}{I_0} \right)^{\frac{1}{tr}} - 1 \qquad (5.6)$$

where: G_t is the growth rate at timestep t
J_i is new infections at timestep i
tr is the number of mean infection periods
I_0 is the number of nodes initially infected

As for the reproduction ratio, define the maximum growth rate G as:

$$G = \max_t G_t \Big|_{t=15}^{t=100} \qquad (5.7)$$

Figure 5-5 displays the distribution of the maximum growth rate (in the initial 100 timesteps) over the 33 660 SIR epidemic simulations, with the equivalent information for SIS at Figure 5-6.

The minimum value of the (maximum) growth rate is 0 as there is at least one infection (the initial node) for all simulations. Thus, unlike reproduction ratio, there is no suitable value for growth rate to be used to define an epidemic. However, it can be used to supplement the theoretically sound reproduction

ratio definition. In particular, *G* can be used to identify those epidemics where there is significant initial growth in infections.

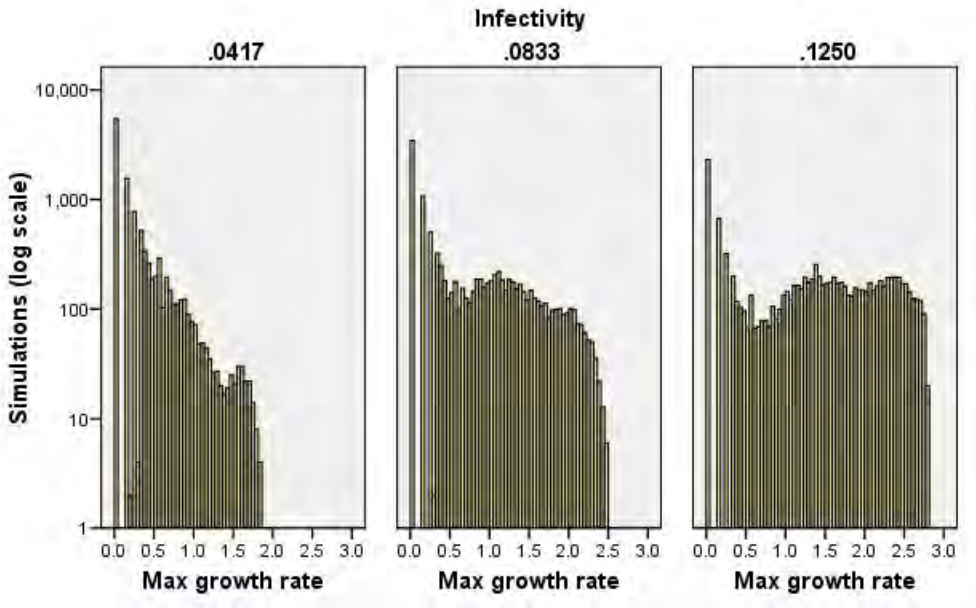

Figure 5-5: Number of simulations by growth rate, SIR *Note that if the initial node results in only one additional node becoming infected,* G *will have the value 2^(1/5)-1 = 0.15, which produces the gap in the histogram.*

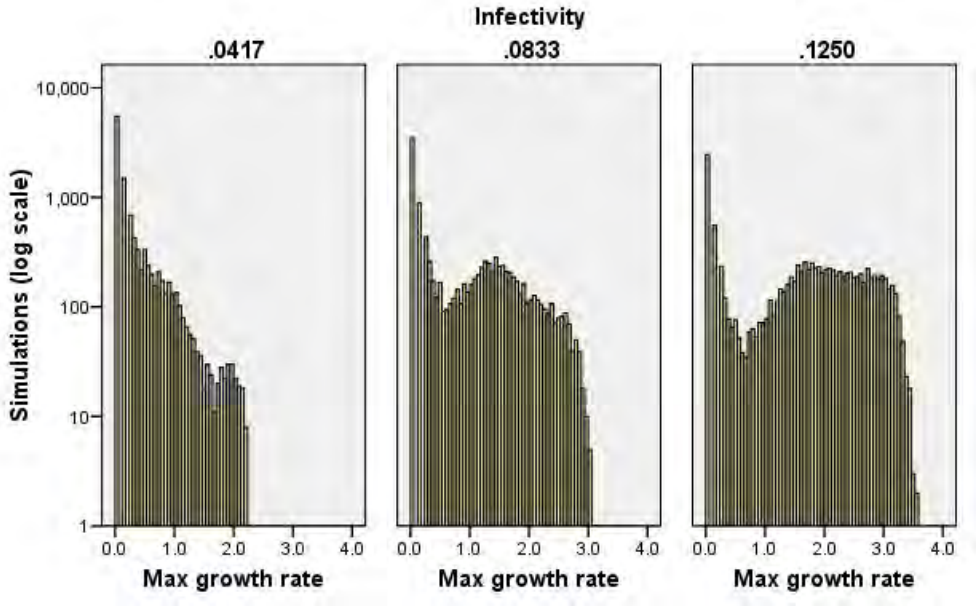

Figure 5-6: Number of simulations by growth rate, SIS

An epidemic is defined to occur when either the reproduction rate is at least 1 or *G* is greater than some threshold value. The threshold value is arbitrary, but should be sufficiently small so as to increase the number of epidemics by minimising false negatives. However, a very low value of *G* will lead to additional false positives, where a simulation is considered an epidemic because of overall growth and the reproduction ratio test will have no practical impact.

Several potential threshold values are examined at Table 5-5. From this table, the threshold value of 1 was chosen for the definition of an epidemic. Assuming that the growth occurs in the first 5 generations (run in period, where reproduction ratio not calculated) and then slows, this threshold is equivalent to the simulation achieving 32 infections by timestep 15. The consistency with the maintenance target of 34 infections by timestep 100 contributes to the strong performance of this threshold value in removing false negatives. Note that 6 infections by timestep 15 (that is, 1 replacement infection per generation) then no further infections would provide a *G* value of 0.43, but such a low value substantially increases the proportion of simulations that are defined as an epidemic but have minimal impact on the population.

Table 5-5: Impact of various values of G threshold on epidemic definition, SIR

G tested	Consistent	False positives	False negatives
> 0.0	68.5%	31.5%	0.0%
> 0.5	91.0%	9.0%	0.0%
> 0.7	95.6%	4.4%	0.0%
> 0.9	97.7%	2.3%	0.0%
> 1.0	98.2%	1.6%	0.2%
> 1.1	97.7%	1.6%	0.8%
> 1.2	97.1%	1.6%	1.3%
> 1.3	96.6%	1.6%	1.8%

Table 5-6 provides the same information as Table 5-4 with the revised definition of epidemic.

Table 5-6: Potential incorrect epidemic classification, with G>1 in epidemic definition

	SIR		SIS	
	Number	**Percent**	**Number**	**Percent**
Not epidemic, infections > 33	73	0.2%	7	0.0%
Not epidemic, infections ≤ 33	21 271	63.2%	18 876	56.1%
Epidemic, infections > 33	11 775	35.0%	14 442	42.9%
Epidemic, infections ≤ 33	541	1.6%	335	1.0%
Total	33 660	100.0%	33 660	100.0%

In summary, an epidemic is defined to occur if either of two conditions is met. The first condition is that, after an initial run-in period of approximately five generations, there is at least one timestep where the number of infected nodes is at least the same size as the number of infected nodes one generation (three timesteps) previously. This is an operational approximation of the epidemic threshold theorem, which states that an epidemic occurs if the basic reproduction ratio is at least 1.

The alternative condition is that there is at least one timestep for which the average growth per generation since the start of the simulation is greater than 1. That is, each generation has infected enough nodes to replace itself (as the nodes recover) and reproduce, so each node must infect more than two other nodes. This condition is to recognise epidemics that grow strongly during the run-in period of the simulation.

5.2 Epidemic impact of degree variation

The role of degree variation in epidemic behaviour has been extensively studied (Becker 1973; Adler 1992; and others, see Section 2.5.2), with a greater probability of an epidemic occurring as degree variance increases. The result arises from the fact that higher degree nodes have greater exposure and hence susceptibility, leading to over-representation in the average degree of the potential epidemic path. Thus, the effective contact

rate is higher than mean degree and this increases the basic reproduction ratio R_0. In turn, this leads to a greater probability of an epidemic occurring.

Despite the higher value of R_0, in positively skewed degree distributions, epidemic size generally decreases with an increase in degree variation (Becker 1973; and Section 2.5.2). This is due to the high proportion of low degree nodes where probability of infection is relatively low and can be seen in the example given in Section 5.7.1. The opposite pattern occurs when infectivity is low (Andersson and Britton 1998).

As there is substantial literature on this topic, I undertook only basic analysis on the independent impact of degree distribution. This was to confirm that the relationship was consistent regardless of the network structure imposed by different levels of assortativity and clustering, as these published studies did not consider these network properties.

The experimental design has three types of degree distribution for which neighbour networks are constructed with target assortativity and clustering coefficients. These are normal, real world and power law. The normal degree distribution networks have substantially lower degree variation than those with real world degree distribution, which are in turn lower than the values for the power law networks.

Using the degree sequence, networks are constructed with various values of assortativity and clustering coefficient. To isolate the effect of degree heterogeneity, the simulations can be compared between degree distribution types for specific values of assortativity and clustering coefficient. Table 5-7 displays aspects of the degree distribution and epidemic behaviour, comparing the degree distribution types for three selected sets of network property values.

***Table 5-7: Network and epidemic properties by degree distribution**: selected assortativity and clustering values, infectivity=0.1250*

	Normal	Real World	Power Law
Clustering~0.0, Assortativity~0.0			
Degree – coefficient of variation	33% - 36%	65% - 69%	86% - 103%
Degree – top 5% nodes	9% - 9%	14% - 15%	19% - 22%
Number of simulations	100	100	70
SIR proportion epidemics	0.70	0.73	0.66
SIR mean final size	0.880	0.780	0.786
SIS proportion epidemics	0.78	0.69	0.77
SIS mean endemic prevalence	0.554	0.503	0.507
Clustering~0.2, Assortativity~0.1			
Degree – coefficient of variation	33% - 36%	66% - 69%	92% - 113%
Degree – top 5% nodes	9% - 9%	14% - 15%	21% - 24%
Number of simulations	100	100	100
SIR proportion epidemics	0.67	0.54	0.51
SIR mean final size	0.838	0.751	0.595
SIS proportion epidemics	0.77	0.64	0.54
SIS mean endemic prevalence	0.553	0.497	0.455
Clustering~0.4, Assortativity~0.2			
Degree – coefficient of variation	33% - 36%	66% - 69%	90% - 105%
Degree – top 5% nodes	9% - 9%	14% - 15%	20% - 22%
Number of simulations	100	100	100
SIR proportion epidemics	0.77	0.51	0.32
SIR mean final size	0.464	0.617	0.430
SIS proportion epidemics	0.83	0.56	0.51
SIS mean endemic prevalence	0.548	0.476	0.407

The first set (clustering ~ 0.0 and assortativity ~ 0.0) provides unstructured networks similar to those assumed in theoretical analysis of the impact of degree heterogeneity. The third set (clustering ~ 0.4 and assortativity ~ 0.2) uses the example values selected in Section 5.1.1 for their real world relevance. The other set (clustering ~ 0.2 and assortativity ~ 0.1) provides an intermediate comparison value.

5.2.1 Relationship with epidemic occurrence

While the literature suggests the normal degree distribution simulations should have the lowest proportion of epidemics through the impact of degree heterogeneity on R_0 (Nold 1980), this pattern is not evident from the simulations shown in Table 5-7. However, the literature does assume that the network is created randomly, without assortativity or clustering.

To investigate this further, Table 5-8 displays the epidemic occurrence for all neighbour network simulation sets with assortativity and clustering coefficient of zero, and also the basic epidemiological model as implemented through the uniform degree network. There is no evident trend of increase or decrease with degree heterogeneity (which increases from Uniform to Power law networks).

Table 5-8: Proportion of simulations where epidemic occurs, zero structure

Infectivity	Immunity	Uniform	Normal	Real World	Power Law
0.0417	SIR	0.09	0.06	0.10	0.16
	SIS	0.04	0.13	0.16	0.09
0.0833	SIR	0.54	0.58	0.54	0.46
	SIS	0.59	0.52	0.59	0.61
0.1250	SIR	0.82	0.70	0.73	0.66
	SIS	0.74	0.78	0.69	0.77

However, there is some evidence of a pattern across all simulations. Table 5-9 compares the proportion of simulations in which epidemics occur for normal (N), real world (R) and power law (P) networks, restricted to those values of assortativity and clustering coefficient for which more than 50 simulations were performed for all three distribution types.

Table 5-9: Distribution type comparison of epidemic proportion *The number of specific approximate assortativity and clustering coefficient pairs for which the relative proportions of epidemics in each of the three degree distribution types show the relationship in the column heading. N denotes normal, R denotes real world and P denotes power law.*

Immunity	Infectivity	N<P<R	N<R<P	P<R<N	P<N<R	R<P<N	R<N<P
SIR	0.0417	7	11	0	1	0	1
	0.0833	0	1	13	4	1	0
	0.1250	0	0	17	2	2	0
SIS	0.0417	4	7	4	2	0	2
	0.0833	0	1	17	1	1	0
	0.1250	0	0	12	0	6	1

This table shows that for the lowest infectivity level, the normal degree distribution simulations generally show the lowest probability of achieving an epidemic for all assortativity and clustering coefficient combinations. However, for the two higher infectivity levels, the smallest epidemic proportion occurs in the power law degree distribution simulations.

One potential explanation for this is related to the selection method for the initial node to infect. This node is selected uniformly from all nodes. For the lowest infectivity rate, all simulation sets have low proportions achieving an epidemic. For the normal distribution simulations, no more than 20% achieve epidemics across all assortativity and clustering combinations where all three distributions are represented. For those networks with higher degree variation, there is a reasonable chance for the initial node to infect a very high degree node that is then able to stimulate an epidemic.

In contrast, for the higher infectivity rates, epidemic occurrence is not as reliant on infecting a very high degree node, moderate degree nodes are sufficient. For the normal distribution simulations, 30% to 90% achieve an epidemic for infectivity of 0.0833 with higher rates for infectivity of 0.1250. Instead, the skewed degree distribution is a disadvantage, because there is a very high proportion of very low degree nodes and there continues to be a

significant probability that the low degree initial node is unable to infect the necessary high degree node.

5.2.2 Relationship with epidemic size

For epidemic size, the literature suggests the power law networks to be smallest except where infectivity is low (Andersson and Britton 1998) and this is true for most of simulations shown in Table 5-7.

Table 5-10 displays the mean size of epidemics for all neighbour network simulation sets with assortativity and clustering coefficient of zero. In addition, the basic epidemiological model is implemented through the uniform degree network.

Unlike epidemic occurrence, there is an evident trend of increase or decrease with degree heterogeneity. For the lowest infectivity rate, size appears to increase with degree heterogeneity and it decreases for the higher infectivity rates. This is supported by an ANOVA (Tabachnick & Fidell 2006), which found the differences between distribution types to be significant ($p<0.001$) for all immunity and infectivity combinations.

Table 5-10: Mean epidemic size, zero structure

Infectivity	Immunity	Uniform	Normal	Real World	Power Law
0.0417	SIR final size	0.035	0.040	0.093	0.231
	SIS prevalence	-	0.15	0.106	0.136
0.0833	SIR final size	0.707	0.681	0.602	0.601
	SIS prevalence	0.411	0.411	0.386	0.366
0.1250	SIR final size	0.920	0.880	0.780	0.786
	SIS prevalence	0.572	0.554	0.503	0.507

Table 5-11 compares the epidemic size for normal (N), real world (R) and power law (P) networks, restricted to those network structure values for which more than 10 epidemics were achieved for all three distribution types.

Table 5-11: Distribution type comparison of epidemic size *The number of specific approximate assortativity and clustering coefficient pairs for which the relative epidemic size in each of the three degree distribution types show the relationship in the column heading. N denotes normal, R denotes real world and P denotes power law.*

Immunity	Infectivity	N<P<R	N<R<P	P<R<N	P<N<R	R<P<N	R<N<P
SIR	0.0417	0	1	0	0	0	0
	0.0833	9	1	5	4	0	0
	0.1250	2	0	11	4	2	0
SIS	0.0417	0	0	0	0	0	0
	0.0833	0	0	16	3	0	0
	0.1250	0	0	19	0	1	0

For the lowest infectivity rate, there were insufficient epidemics to compare the distribution types. For the highest infectivity rate and the moderate infectivity rate for SIS epidemics, the pattern expected from the literature (P<R<N) was achieved for almost all clustering and assortativity combinations.

For the moderate infectivity rate for SIR epidemics, however, there was no consistent pattern, with the normal and the power law distributions each the smallest for approximately half the assortativity and clustering pairs. With detailed investigation, the pattern crystallises into two groups: the power law distribution epidemics are smallest for the relatively high assortativity and low clustering, and the normal distribution epidemics are smallest for relatively low assortativity and high clustering.

Thus, the infectivity role in the impact of degree heterogeneity on epidemic size is more complex than shown in (Andersson and Britton 1998); it also interacts with assortativity and clustering coefficient. This suggests a possible explanation.

For the normal distribution network simulations, the high clustering level is apparently inhibiting the ability of the epidemic to expand and the infectivity rate is not sufficient to overcome this barrier. However, for the networks with more variable degree, the high degree nodes provide surges to break out of the clustered areas.

Where overall infectivity is higher, clustering poses less of a problem and the variation in the degree leads to the expected result of relatively low probability of infection for the high proportion low degree nodes and hence a smaller epidemic.

5.2.3 Discussion

The simulations support the study that has previously found degree heterogeneity to interact with infectivity in its impact on epidemic size (Andersson and Britton 1998). Further, they suggest that a similar complex dependency exists for epidemic occurrence.

While networks with various levels of assortativity and clustering generally show the same relationship between degree heterogeneity and epidemic behaviour as networks with zero values of these properties, there is some suggestion that assortativity and clustering affect the relationship between infectivity, degree heterogeneity and epidemic size.

The role of infectivity introduces an extra dimension to the study and is outside the scope of the research question, which is focussed on the role of social network properties on epidemic behaviour. Further, the impact of degree heterogeneity on epidemic behaviour is complex, with reversals of patterns at different infectivity levels. A general model that includes degree heterogeneity would therefore need to include infectivity in a nonlinear way.

Hence, further analysis will examine the impact of assortativity and clustering coefficient on epidemic behaviour within simulation sets defined by degree distribution type and infectivity. Trends across degree distribution types or infectivity will not be examined because of the potential interference from

the interaction of degree heterogeneity and infectivity. General rules will be sought by identifying common features from the various simulation sets.

Since assortativity and clustering coefficient concern the selection of nodes to connect to each other, these properties will be loosely referred to as network structure.

5.3 Impact of network structure on epidemic occurrence

In contrast with the impact of degree heterogeneity on epidemic occurrence, the effect of network structure has had only limited study. From (Boguña and Pastor-Satorras 2002; and others, see Section 2.5.3), the literature suggests that the proportion of simulations in which epidemics occur should increase as assortativity increases, although one study found the opposite relationship. From (Keeling 2005, see Section 2.5.4) the probability of an epidemic occurring decreases as clustering increases. However, this result is from a single study, albeit with both theoretical and simulation elements.

From the simulations, there is an apparent relationship between network structure and whether an epidemic occurs. However, the relationship is subtle and somewhat inconsistent.

Consider the results in Table 5-12, which shows the proportion of SIR simulations meeting the definition of epidemic on the real world degree distributions with infectivity of 0.1250 (or probability of transmission of infection of 0.30). Generally, within a specific assortativity value, the proportion of epidemics decreases as clustering increases. This is consistent with the results in the literature. However, there are substantial aberrations, particularly for the 0.0 and 0.1 values of assortativity, where there is an increase in the proportion of epidemics for the most clustered networks.

***Table 5-12: Proportion of simulations satisfying epidemic definition: SIR**, real world degree distribution, infectivity=0.1250*

Approximate Assortativity	Approximate Clustering Coefficient					
	0.0	0.1	0.2	0.3	0.4	Total
0.0	73%	59%	56%	73%	59%	64%
0.1	68%	68%	54%	59%	61%	62%
0.2	70%	63%	60%	66%	51%	60%
0.3		58%	61%	57%	57%	58%
0.4		57%	71%	51%	56%	59%
0.5		63%	51%	54%	51%	55%
0.6		58%	57%	55%	49%	55%
0.7		55%	58%	52%	50%	54%
0.8		57%	54%	51%	41%	51%
0.9			53%	46%	42%	46%
Total	70%	60%	58%	56%	52%	57%

The trend across assortativity values within specific clustering values is less clear. With the lower assortativity values, there are both increases and decreases in epidemic proportion as assortativity increases. However, for very high assortativity values (at least 0.6), epidemic proportion decreases as assortativity increases, contrary to the results from the literature.

For the SIS simulations for the same network type and infectivity level (Table 5-13) there is no apparent relationship between epidemic proportion and clustering. For assortativity, the same pattern is observed as for SIR simulations; for high values of assortativity, epidemic proportion decreases as assortativity increases.

One possible explanation for inconsistent results is that the number of simulations differs between property groups, leading to a false result from the smaller samples. However, from Table 5-14, this explanation is false.

***Table 5-13: Proportion of simulations satisfying epidemic definition: SIS**, real world degree distribution, infectivity=0.1250*

Approximate Assortativity	Approximate Clustering Coefficient					
	0.0	0.1	0.2	0.3	0.4	Total
0.0	69%	63%	61%	58%	66%	63%
0.1	73%	59%	64%	72%	64%	66%
0.2	60%	70%	55%	62%	56%	61%
0.3		63%	71%	61%	63%	65%
0.4		60%	67%	60%	67%	64%
0.5		67%	60%	65%	70%	66%
0.6		57%	64%	65%	67%	63%
0.7		53%	63%	56%	64%	59%
0.8		59%	61%	52%	58%	57%
0.9			58%	61%	56%	58%
Total	70%	61%	62%	61%	63%	62%

***Table 5-14: Number of contributing simulations**: Real world, infectivity=0.125*

Approximate Assortativity	Approximate Clustering Coefficient					
	0.0	0.1	0.2	0.3	0.4	Total
0.0	100	100	100	100	100	500
0.1	100	100	100	100	100	500
0.2	10	100	100	100	100	410
0.3		100	100	100	100	400
0.4		100	100	100	100	400
0.5		100	100	100	100	400
0.6		100	100	100	100	400
0.7		100	100	100	100	400
0.8		100	100	100	100	400
0.9			80	100	100	280
Total	210	900	980	1,000	1,000	4,090

5.3.1 Is the relationship significant?

A logistic regression (Tabachnick & Fidell 2006) was used to more rigorously examine the effect of the two properties and their interaction on epidemic occurrence. As assortativity and clustering are of the same order of magnitude, no rescaling is necessary.

To fit the model, the independent variables were added progressively (Elliott & Woodward 2007) with the most significant added in each step (that is, smallest value of significance and provided $p<0.05$). After adding the new variable, the model was recalculated with each of the variables removed individually to assess whether a variable is no longer required. Finally, the Hosmer-Lemeshow test (Tabachnick & Fidell 2006) was used to assess whether the model is a good fit.

For the SIR simulations on the real world degree distribution networks and infectivity of 0.1250, the results of this logistic regression process were as follows. The significant coefficients of the model were 0.519 for the constant ($p<0.001$) and -2.266 for the interaction term ($p<0.001$). In addition, the full model was significant when compared to the constant only model (model chi-square test, $p<0.001$) and the Hosmer-Lemeshow statistic found a good fit ($p=0.561$, nonsignificant chi-square is interpreted as a good fit).

Thus, for a given assortativity A and clustering coefficient C, the probability of an epidemic is estimated by:

$$\text{Prob(epidemic)} = \frac{e^{0.519-2.266AC}}{1+e^{0.519-2.266AC}} \tag{5.8}$$

For example, for a random network with A=0 and C=0, the estimated probability of an epidemic is 62.7%. This is reduced to 57.8% with A=0.3 and C=0.3. Despite the good statistical fit of the model, these results show a much smaller range than the actual values of 73% and 57% respectively.

Instead of using the model to estimate the probability of an epidemic with specific values of assortativity and clustering, the influence of the network

properties can be seen from the coefficient of the odds ratio (Tabachnick & Fidell 2006). The odds ratio is the probability of an epidemic divided by the probability of no epidemic. While the specific coefficient values are difficult to interpret because they depend on both the size of the effect and the starting value, a coefficient of less than 1 means that the probability of an epidemic is reduced as the value of that property increases. Further, a value close to 1 indicates that that epidemic occurrence is almost independent of the property.

The results of the logistic regression for all degree distribution types and infectivity rates are at Table 5-15 (SIR) and Table 5-16 (SIS). These tables show the coefficients of the odds ratio for each significant ($p<0.05$) property or interaction and the value of the Hosmer-Lemeshow goodness of fit statistic.

Table 5-15: Influence of network properties on epidemic occurrence (SIR): *logistic regression odds ratio coefficients ("-" indicates not significant)*

Distribution	Infectivity	Assortativity	Clustering	Interaction	Fit (>0.05)*
Real world	0.0417	-	0.240	-	0.298
	0.0833	0.382	-	-	0.160
	0.1250	-	-	0.104	0.561
Power law	0.0417	-	-	-	-
	0.0833	0.047	-	-	0.277
	0.1250	-	0.429	0.005	0.949
Normal	0.0417	-	0.154	-	0.212
	0.0833	0.481	0.128	-	0.238
	0.1250	-	-	0.076	0.554

* *The Hosmer-Lemeshow goodness of fit statistic is report in the column headed Fit(>0.05) and the fit is considered good if the value is greater than 0.05.*

From Table 5-15, except for the simulations with a power law degree distribution and infectivity of 0.0417, the social network structural properties have a significant effect on the probability of an epidemic. Unfortunately, different properties are significant for different simulation sets.

One area where there is some consistency is which properties are significant when comparing across simulation sets with the same infectivity level. That is, assortativity has a significant influence for all simulation sets with infectivity of 0.0833 and the interaction term has a significant influence for all simulation sets with infectivity of 0.1250.

The other area of consistency is that all odds ratio coefficients are smaller than one. That is, an increase in the property (or interaction) value decreases the probability of an epidemic occurring. This is consistent with the literature for clustering, but conflicts with previous studies concerning the effect of assortativity.

The results for SIS simulations (Table 5-16) also show insufficient consistency to allow generalisation. Overall, the network properties have a lesser impact on the occurrence on an epidemic for SIS as compared to SIR, with three simulation sets unable to find a model, and one model with a poor fit.

For specific property effects, clustering is not a significant variable in the regression for any simulation set, unlike the SIR simulations. However, it does have an impact through the interaction term.

***Table 5-16: Influence of network properties on epidemic occurrence (SIS)**: logistic regression odds ratio coefficients*

Distribution	Infectivity	Assortativity	Clustering	Interaction	Fit (>0.05)*
Normal	0.0417	-	-	0.185	0.516
	0.0833	-	-	0.362	0.792
	0.1250	-	-	-	-
Real world	0.0417	-	-	-	-
	0.0833	0.380	-	3.632	0.518
	0.1250	0.731	-	-	0.868
Power law	0.0417	-	-	-	-
	0.0833	0.167	-	-	0.616
	0.1250	0.187	-	-	No fit 0.004

* *The Hosmer-Lemeshow goodness of fit statistic is report in the column headed Fit(>0.05) and the fit is considered good if the value is greater than 0.05.*

The other difference is that there is an odds ratio coefficient that is more than one. However, that coefficient is on the interaction term and assortativity is also significant. As a result, if the network has a clustering value of no more than 0.7245, the decrease in epidemic occurrence arising from a difference only in assortativity is larger than the increase arising from the interaction term. For a given value of assortativity, simulations on networks with a higher clustering coefficient are more likely to result in an epidemic than on networks with a lower clustering coefficient.

5.3.2 Discussion

For those simulation sets (degree distribution type by immunity type by infectivity rate) in this study where a relationship was found, the presence of network structure consistently decreases the occurrence of epidemics. However, the structural property with the strongest influence differed between the simulation sets for both SIR and SIS epidemics.

For SIR simulations, the highest infectivity rate logistic models had the best fit and these found the interaction of assortativity and clustering to be significant. Logistic models were fitted to other simulation sets that found only assortativity or only clustering to impact on epidemic occurrence, albeit with lower significance.

For SIS simulations, four of the nine simulation sets found no significant relationship between network structure and epidemic occurrence (as compared to one for SIR). Where a relationship was found, assortativity or the interaction term was the relevant structural variable (with one set using both in the model). Clustering coefficient did not appear in any model outside of its influence through the interaction with assortativity.

The SIR simulations provide some evidence to support published results that higher clustering is associated with a reduction in the probability of an epidemic occurring, with six of the nine simulation sets finding a significant relationship with either clustering or the interaction term. However, the SIS simulations found only limited evidence with no simulation sets identifying

clustering as a significant factor. Of the three sets that found the interaction term to be significant, one has increased probability of epidemics with an increase in clustering.

This suggests that clustering may restrict SIR epidemics more than it does SIS epidemics. Such an interpretation is reasonable, because loops of infection that return to previously infected nodes are permanently blocked for SIR epidemics, but the node may have become susceptible again in SIS epidemics.

For assortativity, all simulation sets either found no significant relationship with epidemic occurrence or found that an increase in assortativity is associated with a decrease in epidemic occurrence. This is in conflict with published results.

In summary, over the property space simulated, the effect of network structure on epidemic occurrence varies between simulation sets (degree distribution type by infectivity level), and the results differ between SIR and SIS epidemics. There is some consistency in that, where an effect is observed, increasing assortativity or clustering coefficient decreases epidemic occurrence, but the relevant property is different for different sets and may impact through the interaction term.

5.4 Basic reproduction ratio for SIR epidemics in the presence of network structure

Three papers (Nold 1980; Newman 2002a; Moreno et al. 2003) have considered the impact of assortativity on epidemic size. Nold's study considered networks with the specific mixing scheme of part proportional and part within the same degree, and the other two papers used the full specification of the joint probability distribution. The three papers all considered SIR epidemics and agreed that the epidemic is smaller in the presence of assortativity. Due to the relationship between epidemic size and basic reproduction ratio, this would also lead to a reduction in epidemic derived R_0.

Only Moreno et al (2003) quantified the relationship, noting that assortativity can decrease size by 15-20% for moderate (in terms of their simulation parameters) infectivity rates. The results also suggest that a higher assortativity coefficient leads to a smaller epidemic in the absence of other changes.

The published literature on the impact of clustering on epidemic size is very limited. The only study to have specifically compared SIR epidemics on clustered networks compared to similar networks without clustering (Keeling 1999) found that increasing clustering decreases epidemic size and R_0.

To examine the relationship between social network properties and SIR epidemic behaviour, only those SIR simulations on neighbour networks where an epidemic occurred will be included in the analysis. The size of the dataset for each of the nine simulation sets is at Table 5-17 (extracted from Table 5-2 on page 101). For the real world degree distribution with infectivity of 0.1250, the network properties for the 2 331 epidemics are distributed as shown in Table 5-18.

Table 5-17: Number of epidemics in dataset (SIR)

	Infectivity rate		
	0.0417	0.0833	0.1250
Normal	326	2 152	3 506
Real world	480	1 573	2 331
Power law	276	627	900

From the experimental design, up to 100 epidemics were simulated for each property combination (at each infectivity value), with 10 simulations for each of up to 10 networks. Further, the set of 10 target degree instances was identical for each property combination, replicating network structure to the extent possible.

***Table 5-18: Number of contributing simulations: SIR**, real world, infectivity=0.1250*

Approximate Assortativity	Approximate Clustering Coefficient					
	0.0	0.1	0.2	0.3	0.4	Total
0.0	73	59	56	73	59	320
0.1	68	68	54	59	61	310
0.2	7	63	60	66	51	247
0.3	-	58	61	57	57	233
0.4	-	57	71	51	56	235
0.5	-	63	51	54	51	219
0.6	-	58	57	55	49	219
0.7	-	55	58	52	50	215
0.8	-	57	54	51	41	203
0.9	-	-	42	46	42	130
Total	148	538	564	564	517	2 331

The values of the network properties and R_0 as derived from epidemic final size are shown at Table 5-19. For the lowest infectivity rate of 0.0417, the small number of epidemics occurring in the simulations is consistent with the low values of R_0 for all epidemics. As infectivity increases, the maximum of the derived value of R_0 increases accordingly, but there continue to be epidemics that almost failed.

The differences in the ranges of assortativity and clustering across different infectivity levels within the same degree distribution type indicate that epidemics did not occur on some networks.

Table 5-19: Range of property and SIR derived R_0 values in dataset

Distribution	Infectivity	Assortativity	Clustering	R_0
Normal	0.0417	0.00 to 0.86	0.01 to 0.50	1.00 to 1.13
	0.0833	-0.02 to 0.86	0.01 to 0.51	1.00 to 1.85
	0.1250	-0.02 to 0.86	0.01 to 0.51	1.01 to 2.66
Real world	0.0417	-0.05 to 0.86	0.01 to 0.45	1.00 to 1.17
	0.0833	-0.05 to 0.86	0.01 to 0.45	1.00 to 1.67
	0.1250	-0.05 to 0.86	0.01 to 0.45	1.00 to 2.13
Power law	0.0417	-0.05 to 0.48	0.02 to 0.46	1.00 to 1.20
	0.0833	-0.05 to 0.49	0.02 to 0.46	1.00 to 1.63
	0.1250	-0.05 to 0.49	0.02 to 0.46	1.01 to 2.09

5.4.1 Is there a relationship?

Before undertaking a regression analysis to model the relationship between the network properties of assortativity and clustering, and epidemic behaviour as measured by epidemic size and the basic reproduction ratio R_0, exploratory techniques were used to assess whether a relationship is apparent.

5.4.1.1 Epidemic behaviour over time, varying assortativity or clustering

The first exploratory analysis examined the number of infections over time for specific assortativity values, comparing different levels of clustering, and for specific clustering coefficients, comparing different levels of assortativity.

Again using the real world degree distribution with infectivity of 0.1250 as an example, Figure 5-7 shows the mean cumulative infections of the simulations over networks with assortativity of approximately 0.2 where an epidemic occurred, by clustering coefficient of the network.

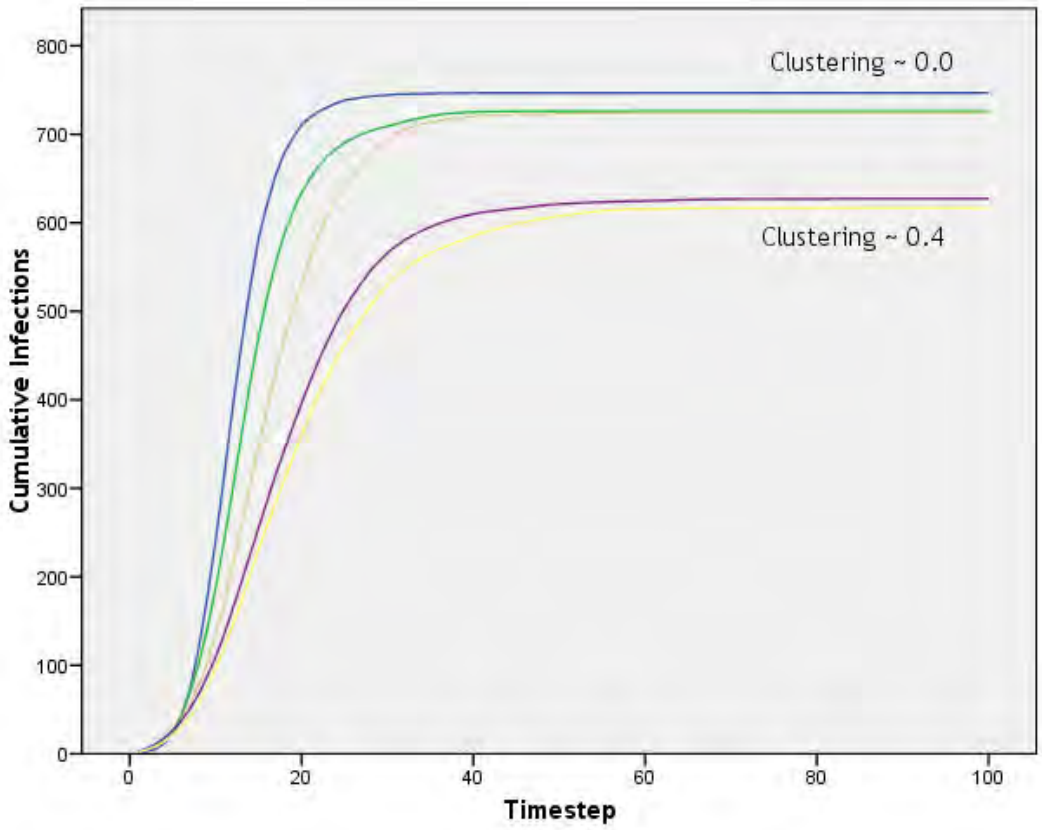

Figure 5-7: Epidemic size over time (cumulative infections) by clustering coefficient: Real world, infectivity=0.1250, Assortativity~0.2

While assortativity of 0.2 was selected for the example because it is a moderate and socially realistic level (see Table 2.1), it is coincidentally reasonably well behaved. For each increase in clustering coefficient, cumulative infections increase more slowly and reach a smaller final size. However, there is little difference between the results for clustering coefficients of 0.1 and 0.2, and only slightly more between 0.3 and 0.4.

Similarly, Figure 5-8 shows the mean cumulative infections of the simulations over networks with clustering coefficient of approximately 0.4 where an epidemic occurred, by assortativity of the network.

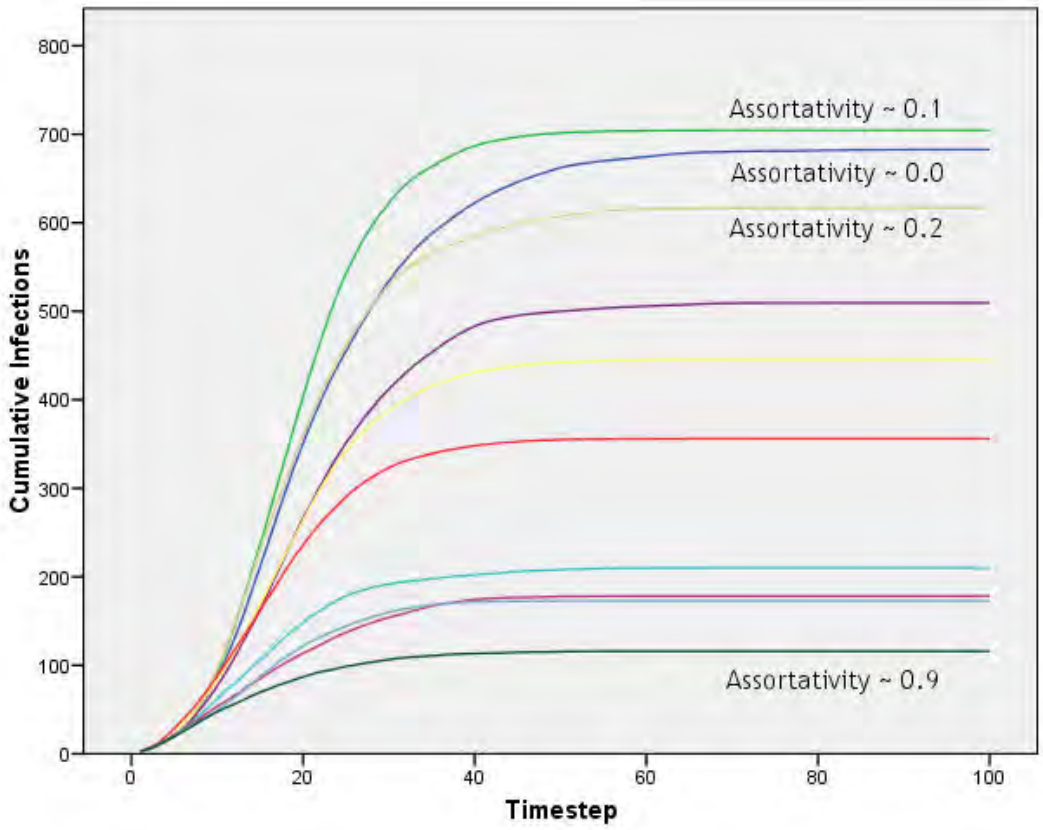

Figure 5-8: Epidemic size over time (cumulative infections) by assortativity: *Real world, infectivity=0.1250, Clustering~0.4*

Again, 0.4 was selected because it is a moderate and socially realistic level, but it is again reasonably well behaved. For each increase in assortativity, cumulative infections increase more slowly and reach a smaller final size. The only aberration is that the cumulative infection curves for the two lowest values of assortativity are reversed. That is, the epidemics on networks with assortativity of 0.0 show slower growth and a smaller final size than for epidemics on networks with assortativity of 0.1.

While not all of the specific clustering or assortativity values show the consistency of the relationships in these examples, there is a clear pattern. That pattern is that for each of the degree distribution types and infectivity levels, epidemics are smaller as assortativity or clustering increases while holding the other property fixed.

There is a notable group of exceptions. The fixed clustering charts for infectivity of 0.0417 for all three degree distribution types show no clear pattern between assortativity values and cumulative infections over time. This lack of a pattern is consistent with the complex interactions between degree heterogeneity, infectivity and network structure identified in Section 5.2.

5.4.1.2 Epidemic final size, with assortativity and clustering

Having established that there is a difference in epidemic behaviour over time, the second exploratory analysis considers how the epidemic final size varies with both assortativity and clustering coefficient simultaneously. Final size is important because it is the epidemic parameter used to estimate R_0.

Table 5-20 displays the average of the final size for all simulations in which an epidemic occurred (for the specified simulation set). That is, up to 100 simulations contribute to each average, 10 epidemic simulations on each of 10 networks. The actual number of contributing epidemics is given at Table 5-18.

While there are some aberrations (such as 705 for assortativity of 0.1 and clustering of 0.4), there is a clear trend whereby final size decreases as the value of either structural property increases. Also immediately noticeable is that these differences are very large, with the epidemic infecting 116 nodes in the most highly structured networks on average, compared to 780 nodes in the unstructured networks. In networks of 1,000 nodes, this is a reduction of final size to 0.116 from 0.780.

With three continuous variables and dependency between assortativity and clustering coefficient, there is no simple test for a trend in the final size values. However, Table 5-21 provides the standard error for each mean and these are small enough to suggest that the observed relationship is statistically valid.

***Table 5-20: Mean epidemic final size: SIR**, real world, infectivity=0.1250*

Approximate Assortativity	Approximate Clustering Coefficient					
	0.0	0.1	0.2	0.3	0.4	Total
0.0	780	774	760	744	683	749
0.1	768	764	751	737	705	746
0.2	747	726	724	627	617	677
0.3		693	666	598	509	618
0.4		692	633	543	446	583
0.5		665	600	480	356	533
0.6		616	491	373	210	432
0.7		602	449	314	178	393
0.8		517	379	271	173	349
0.9			307	185	116	202
Total	773	675	585	507	422	563

***Table 5-21: Standard error of mean epidemic final size: SIR**, real world, infectivity=0.1250*

Approximate Assortativity	Approximate Clustering Coefficient					
	0.0	0.1	0.2	0.3	0.4	Total
0.0	2	4	3	10	22	5
0.1	3	3	5	5	12	3
0.2	10	4	5	20	18	7
0.3		12	13	22	26	10
0.4		4	16	24	26	11
0.5		10	14	20	23	12
0.6		18	26	24	24	15
0.7		12	22	24	19	14
0.8		22	25	22	21	14
0.9			28	19	11	14
Total	2	5	8	10	11	5

Again, there is no clear relationship between assortativity and final size in the simulation sets where infectivity is 0.0417.

5.4.1.3 Correlation between R_0 and assortativity or clustering

The third exploratory analysis calculates the partial correlation coefficient between the basic reproduction ratio R_0 as derived from the epidemic final size, and each of assortativity and clustering coefficient, removing the influence of the other.

From Table 5-22, this analysis indicates that there is a significant negative linear relationship between R_0 and assortativity for all distribution types with infectivity of 0.0833 or 0.1250 (correlation of -0.38 to -0.79, $p<0.001$). That is, increases in assortativity are associated with a decrease in the R_0 calculated from epidemic final size for both these infectivity rates.

Table 5-22: Correlation between epidemic derived R_0 and network properties (SIR): *Partial correlation correcting for other network property (significant with $p<0.001$)*

Distribution	Infectivity	Assortativity	Clustering coefficient
Normal	0.0417	0.21	-0.43
	0.0833	-0.38	-0.78
	0.1250	-0.50	-0.84
Real world	0.0417	-	-0.50
	0.0833	-0.65	-0.67
	0.1250	-0.82	-0.58
Power law	0.0417	-0.31	-0.46
	0.0833	-0.69	-0.44
	0.1250	-0.71	-0.38

The three simulation sets with infectivity of 0.0417 show a different result for each of the three degree distribution types. For real world distribution networks there is no significant linear relationship. For normal distribution networks, there is a positive significant linear relationship ($p<0.001$), opposite to that shown in simulations with higher infectivity. For power law distribution networks, there is a significant linear relationship ($p<0.001$)

consistent with the higher infectivity simulations, where an increase in assortativity is associated with a decrease in R_0.

For all degree distribution types and infectivity levels, there is a significant negative linear relationship between R_0 and clustering coefficient (correlation of -0.40 to -0.83, $p<0.001$). That is, increases in clustering are associated with a decrease in the R_0 calculated from epidemic final size.

5.4.1.4 Relationship between R_0 and assortativity and clustering

The final exploratory analysis examined the specific values of assortativity and clustering coefficient, instead of comparing within groups of similar values. A scatter plot was constructed for each simulation set with the structural property values as the axes, using colour to indicate the mean R_0 value (Figure 5-9). For each point, 10 simulations and therefore up to 10 epidemics contributed to the mean.

From this figure, R_0 values range from approximately 1.0 to 2.1. The grey pales as assortativity or clustering increases, indicating a decrease in the mean value of R_0, consistent with the earlier analyses.

For infectivity of 0.0417, the scatter plots highlight the conflicting results from those simulation sets with. Each of these shows a small range of R_0 values with only a few points with colours that are different from the majority. While the relationship between clustering coefficient and R_0 is obvious, there is no apparent relationship between assortativity and R_0. This is despite the significant partial correlation found for two of the degree distribution types.

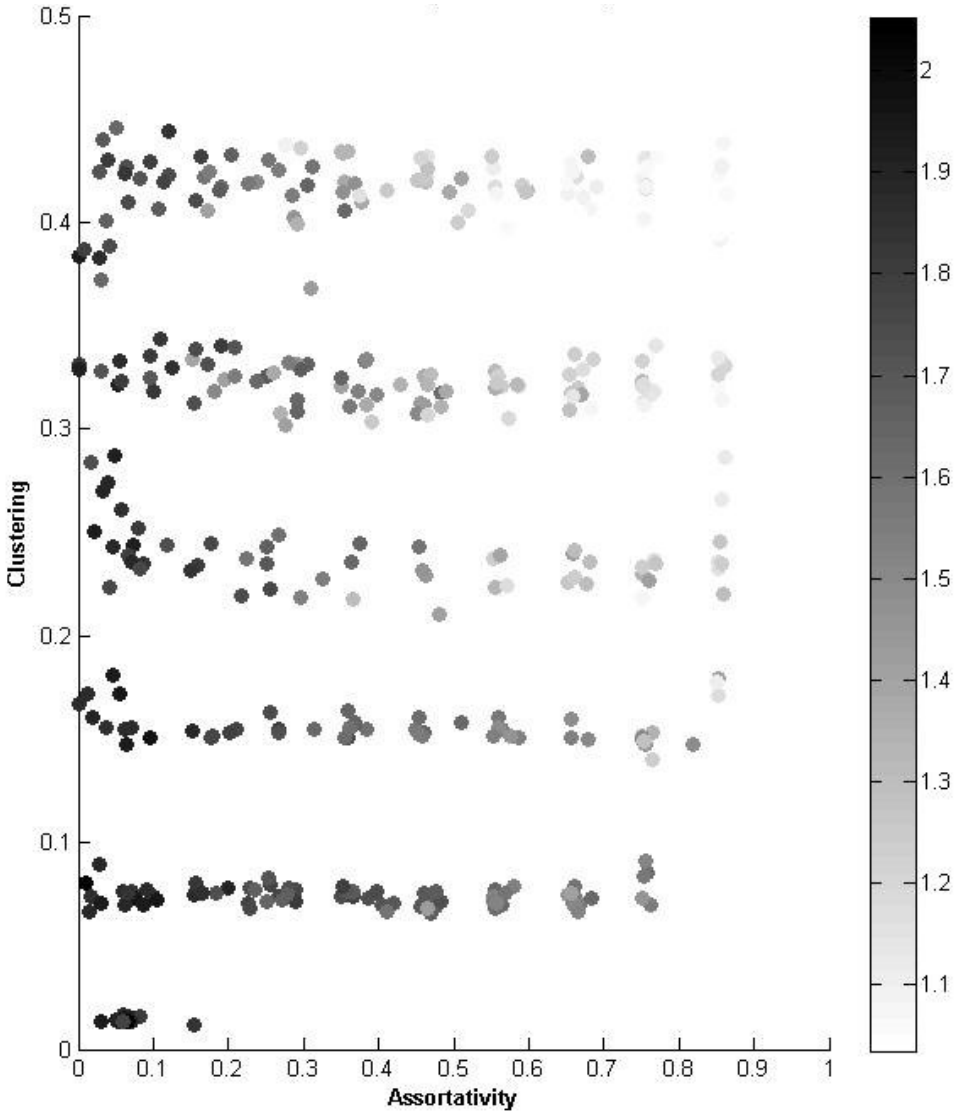

***Figure 5-9**: Assortativity, clustering coefficient and* R_0 *derived from epidemic final size: SIR, Real world, infectivity=0.1250, colour indicates* R_0

5.4.2 Relationship between network structure and the derived basic reproduction ratio (SIR)

While the preliminary analysis suggests there is a relationship between assortativity, clustering and R_0, there is no indication whether the relationship

is linear and whether the two network properties operate independently or have some additional interacting influence.

Multiple linear regression (Tabachnick & Fidell 2006) was used to model the relationship between the basic reproduction ratio R_0, as derived from epidemic final size, and the network structure properties of assortativity and clustering coefficient. A separate model is fitted for each degree distribution type and infectivity level.

5.4.2.1 Adequacy of linear model without interaction terms

To assist in assessing whether a linear model is adequate, three sets of regression models are fitted. The first is the multiple regression model containing only assortativity and clustering coefficient as independent variables. This models the linear relationship without interaction.

Two broader regression models add independent variables progressively (Elliott & Woodward 2007) from a pool that includes nonlinear and interaction transformations of assortativity and clustering coefficient using stepwise selection. This method selects the variable from the pool that is most significant (that is, lowest p value) and then checks the model to determine whether any of the variables is of insufficient significance and should be removed.

The variables available for selection are listed at Table 5-23. Where the nonlinear transformation involves $\log A$ or $\sqrt{A}$, some data points are lost because there are networks with negative values of assortativity. Thus, expanded models use a smaller dataset to fit the model and there is the possibility that the model will have lesser explanatory power than a linear model fitted to the complete data.

One set of expanded models starts with the linear model based on assortativity and clustering before the stepwise selection of additional variables. The other set of models has no variables initially included.

Table 5-23: Linear, nonlinear and interaction variables tested for model fitting

	Linear	Nonlinear, no interaction	Nonlinear, interaction
Assortativity (A)	X		
Clustering (C)	X		
log A		X	
log C		X	
A log A		X	
C log C		X	
1 / (1+e^(5-10A))		X	
1 / (1+e^(5-10C))		X	
1 / A		X	
1 / C		X	
√A		X	
√C		X	
A √A		X	
C √C		X	
A^2		X	
C^2		X	
AC			X
A √C			X
C √A			X
A^2C			X
AC^2			X
A log C			X
C log A			X
C / (1+e^(5-10A))			X
A / (1+e^(5-10A))			X

The purpose of fitting the expanded models is not to determine the final form of the regression model. Such a model would have substantial collinearity between the independent variables. Rather, it is intended to assess the potential additional explanatory power from adding nonlinear or interaction terms and to help identify potential variables should they be required.

Table 5-24 reports the adjusted R^2 for the linear model and the best model achieved by either set of expanded regression models. The adjusted R^2 is used as it includes a correction for the number of variables in the model and therefore provides a more useful comparison the fit of models with different numbers of independent variables. As can be seen from this table, there is little gain in adding the nonlinear and interaction terms and, in some cases, there is a reduction in the fit of the model.

Based on this analysis, only a multiple linear regression model will be fitted to the simulation results. Confirmation of the adequacy of the linear model will be sought through an analysis of the residuals.

Table 5-24: Adjusted R^2 for linear and nonlinear regressions, SIR

Distribution	Infectivity	Adj R^2 linear	Adj R^2 expanded
Normal	0.0417	0.194	0.221
	0.0833	0.668	0.800
	0.1250	0.769	0.806
Real world	0.0417	0.248	0.315
	0.0833	0.659	0.654
	0.1250	0.753	0.766
Power law	0.0417	0.296	0.324
	0.0833	0.582	0.405
	0.1250	0.586	0.357

5.4.2.2 Model of relationship

For all simulation sets, a linear model was able to be fitted that is significant ($p<0.001$), so there is a linear relationship between R_0 and at least one of assortativity and clustering coefficient. However, the explanatory power of the models varies substantially (see Table 5-25). For simulations with infectivity of 0.0417, only 20% to 30% of the variability of R_0 is accounted for by the linear model. Within degree distribution type, the explanatory power of the models for simulations with infectivity of 0.0833 and 0.1250 are of

similar magnitude, but slightly greater for the models of epidemics with higher infectivity.

Table 5-25: Performance of multiple linear regression models, SIR

Distribution	Infectivity	R^2	Significance
Normal	0.0417	0.198	p<0.001
	0.0833	0.668	p<0.001
	0.1250	0.769	p<0.001
Real world	0.0417	0.251	p<0.001
	0.0833	0.659	p<0.001
	0.1250	0.753	p<0.001
Power law	0.0417	0.301	p<0.001
	0.0833	0.584	p<0.001
	0.1250	0.587	p<0.001

Consider the simulations over real world degree distribution networks with infectivity of 0.1250. The regression equation is given by (Table 5-26 on page 145):

$$\text{Predicted } R_0 = 2.036 - 0.781 \text{ Assortativity} - 0.789 \text{ Clustering} \qquad (5.9)$$

The first analysis of the adequacy of this model considers the distribution of the residuals. The residual is calculated as:

$$\text{Residual} = \text{actual value of } R_0 - \text{model prediction for } R_0 \qquad (5.10)$$

In Figure 5-10, the residuals have been standardised. While the histogram is similar to the expected normal distribution, there is a tail with several residuals more than 3 standard deviations below the mean. That is, the model prediction is very high for more epidemics than expected.

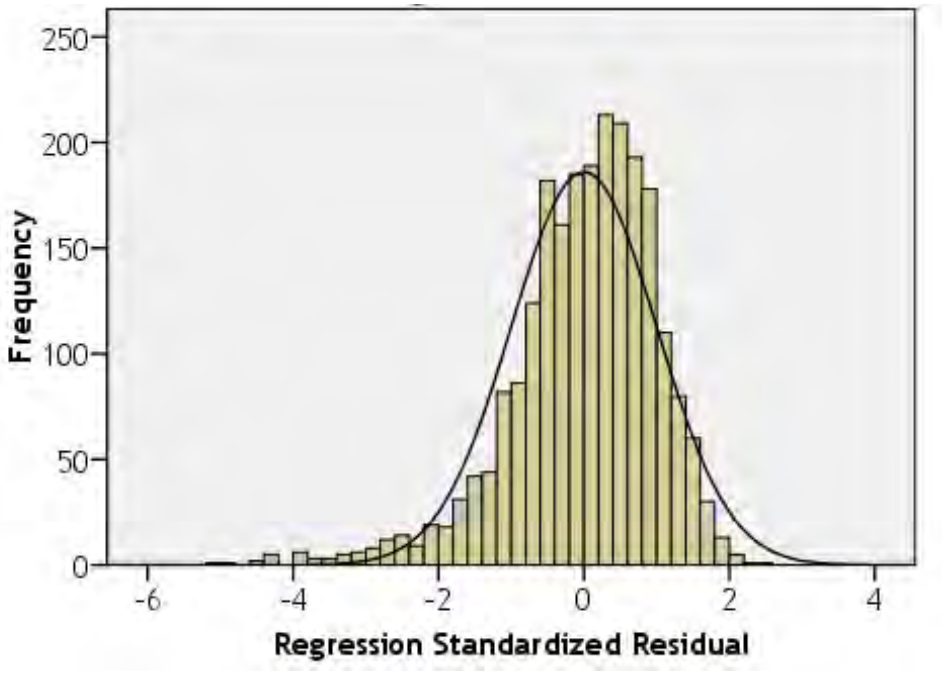

***Figure 5-10: Histogram of standardised regression residuals**: SIR, Real world, infectivity=0.1250*

To analyse whether these very high predictions indicates a systematic error in the regression model, the residual is plotted separately against each variable: predicted R_0 (Figure 5-11), assortativity (Figure 5-12) and clustering coefficient (Figure 5-13). Such a systematic error could, for example, indicate nonlinearity in the actual relationship that is not accounted for in the model.

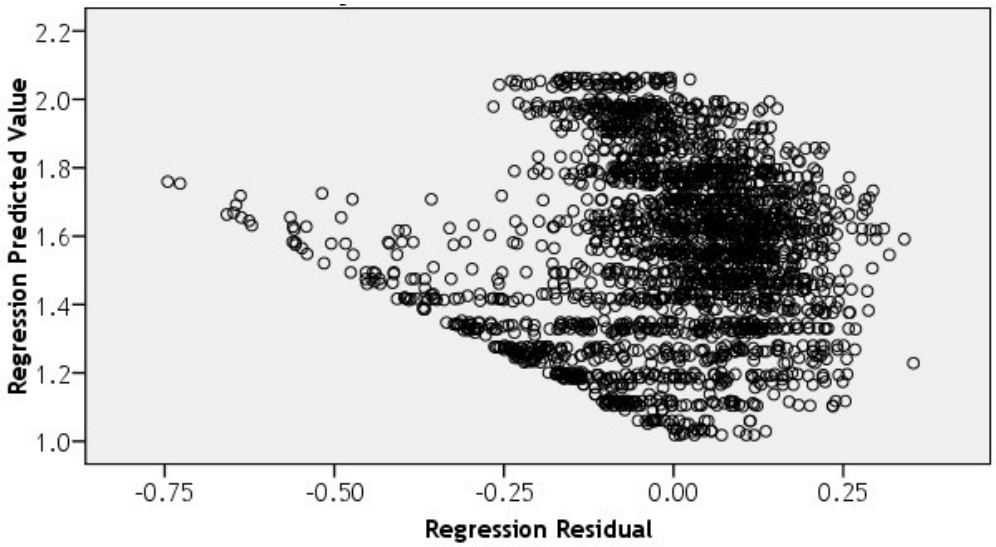

Figure 5-11: Residual plotted against regression prediction for R_0***:*** *SIR, Real world, infectivity=0.1250*

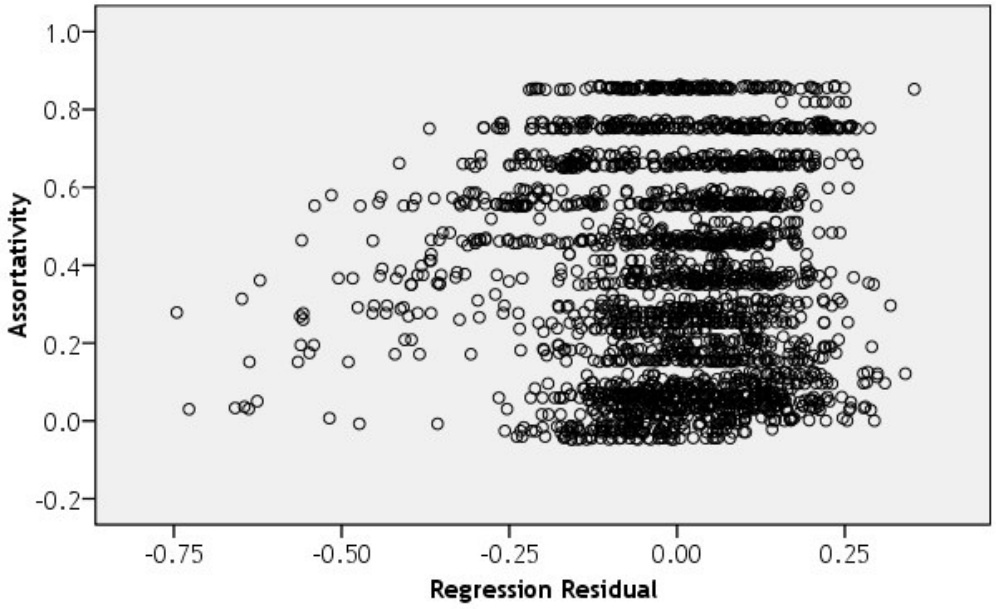

Figure 5-12: Residual plotted against assortativity: *SIR, Real world, infectivity=0.1250*

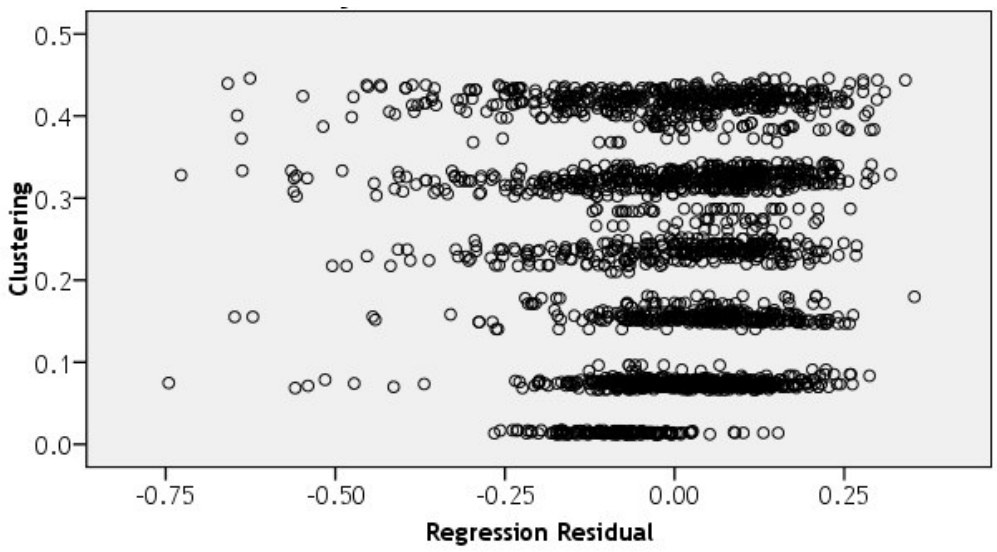

***Figure 5-13: Residual plotted against clustering coefficient**: SIR, Real world, infectivity=0.1250*

The first of these plots suggests that any bias is occurring in the midrange of the predicted R_0 values. This could indicate nonlinearity, with the predicted values not decreasing quickly enough through the middle values of assortativity and clustering, but then 'catching up' at the higher values. The plots against each network property show large negative residual values throughout the range of clustering coefficient values, with some suggestion that the overestimates of R_0 occur where assortativity is lower. However, for these assortativity values, the majority of the residual values are close to zero.

Instead, the apparent bias in the model predictions actually reflects a property of R_0. Basic reproduction ratio for an epidemic that occurs, derived from the epidemic final size, must be at least 1. Thus, the residual for a predicted value of 1+x is at least –x. This provides the boundary evident in Figure 5-11.

Overall, the linear regression model is a good fit for epidemic simulations over real world networks with infectivity of 0.1250. Explanatory power is high, and there is no evidence of a systematic bias that could indicate the presence of nonlinearity or interaction terms that are not included in the model.

Consistent with the results of the preliminary analysis, the regression models for simulation sets with infectivity of 0.0417 have insufficient explanatory power to be of any practical use (from Table 5-25). However, the models for the other simulation sets with infectivity of 0.0833 or 0.1250 all support conclusions similar to the model detailed above. Regression coefficients for these models are shown in Table 5-26 and all are significant ($p<0.001$).

Table 5-26: Regression coefficients (SIR), influence of assortativity and clustering on basic reproduction ratio

Distribution	Infectivity	Intercept	Assortativity	Clustering
Normal	0.0417[a]	1.026	0.013	-0.049
	0.0833	1.608	-0.236	-1.200
	0.1250	2.564	-0.575	-2.580
Real world	0.0417[a]	1.080	not significant	-0.144
	0.0833	1.565	-0.339	-0.692
	0.1250	2.036	-0.781	-0.789
Power law	0.0417[a]	1.139	-0.080	-0.135
	0.0833	1.410	-0.566	-0.301
	0.1250	1.748	-1.114	-0.496

a Values for infectivity 0.0417 are reported but are not of practical use

From the regression coefficients, basic reproduction ratio is reduced as assortativity or clustering increases for the exploitable models. There is, however, insufficient evidence to support any other general rules as the relative impact of the two properties differ between degree distribution types and infectivity levels.

5.5 Comparison between SIS and SIR results

Studies of the impact of assortativity and clustering have focused on SIR epidemic behaviour. However, except for the mean field approach, the results have not relied on the fact that the epidemic is SIR. Thus, the SIR results can be used as a first estimate of the effect of network structure on SIS epidemics and any behaviour differs between epidemic types are of particular interest.

The SIR and SIS epidemics in this study are simulated over identical networks with the same sets of infectivity rates. Hence, it is reasonable to expect some consistency of epidemic behaviour and observed relationships between the two immunity groups.

The same analyses of relationships between network properties and R_0 as described in Section 5.4 for SIR epidemics were also conducted for SIS epidemics. In addition, values of R_0 from equivalent SIR and SIS simulations are directly compared in Section 5.5.2.

5.5.1 Summary of SIS simulation results

To examine the relationship between social network properties and SIS epidemic behaviour, only those SIS simulations on neighbour networks where an epidemic occurred will be included in the analysis. In addition, at least one node must be infected at timestep 100, the end of the period for which epidemic information is recorded in the simulations. This additional requirement is because a simulation can meet the definition of an epidemic, but not maintain endemic equilibrium. Such equilibrium is needed to calculate the basic reproduction ratio R_0 from prevalence using equation (5.2) .

The size of the dataset for each of the nine simulation sets is at Table 5-27. Even some of the high infectivity rate epidemics are unable to sustain an endemic, but it is a particular problem for normal degree distribution epidemics with the lowest infectivity rate. For these simulations, despite the epidemic occurring, there is neither sufficient probability of infection nor a core group of high degree nodes to sustain the infections.

The values of the network properties and R_0 as derived from epidemic prevalence are shown at Table 5-28. Prevalence is calculated as the mean prevalence over the final 9 timesteps (3 average generations). This is to minimise the impact of the skewed degree distribution on variability of prevalence and hence on variability of the derived R_0.

Table 5-27: Number of epidemics in dataset (SIS)

	Infectivity rate		
	0.0417	0.0833	0.1250
Normal	711	2 797	3 830
with prevalence > 0	220	2 745	3 821
Real world	719	1 863	2 554
with prevalence > 0	559	1 804	2 533
Power law	322	761	1 083
with prevalence > 0	307	716	1 054

Table 5-28: Range of property and SIS derived R_0 values in dataset

Distribution	Infectivity	Assortativity	Clustering	R_0
Normal	0.0417	-0.01 to 0.86	0.01 to 0.50	1.00 to 1.09
	0.0833	-0.02 to 0.86	0.01 to 0.51	1.00 to 1.78
	0.1250	-0.02 to 0.86	0.01 to 0.51	1.01 to 2.37
Real world	0.0417	-0.05 to 0.86	0.01 to 0.45	1.00 to 1.20
	0.0833	-0.05 to 0.86	0.01 to 0.45	1.01 to 1.72
	0.1250	-0.05 to 0.86	0.01 to 0.45	1.00 to 2.14
Power law	0.0417	-0.05 to 0.49	0.02 to 0.44	1.04 to 1.20
	0.0833	-0.05 to 0.49	0.02 to 0.46	1.00 to 1.65
	0.1250	-0.05 to 0.49	0.02 to 0.46	1.00 to 2.11

5.5.1.1 Is there a relationship?

Before undertaking the regression analysis, exploratory techniques were used to assess whether there appears to be a relationship between the network properties of assortativity and clustering, and epidemic behaviour as measured by endemic prevalence and the basic reproduction ratio R_0. The four exploratory analyses conducted were the same as for the SIR epidemic analysis, except for the substitution of prevalence (current infections) instead of final size (cumulative infections).

The first exploratory analysis examined charts of current infections over time (SIR analysis described at Section 5.4.1.1), with either assortativity or

clustering held constant and separate plots for each value of the other property.

SIS epidemics show an increase in prevalence over time until the equilibrium prevalence level is achieved. With fixed assortativity, as for SIR epidemics, the epidemic grows more slowly for each increase in clustering coefficient. However, the final prevalence levels achieved do not show the same separation as SIR final size values except for those networks with very high assortativity levels, where higher clustering is associated with a lower prevalence.

With fixed clustering coefficients, the epidemic grows more slowly and the final prevalence is at a lower level for each increase in assortativity. This is consistent with the SIR results.

However, there is a complication for higher clustering coefficients. Consider Figure 5-14, the mean prevalence for epidemics on networks with assortativity of approximately 0.9 has apparently not reached stability. This problem exists for several of the simulation sets where both properties have high values, particularly for the lowest infectivity level simulations.

More detailed examination reveals that the simulations defined as epidemics on these highly structured networks separate into three groups (example numbers are from the 56 epidemics in the Assortativity~0.9 group in Figure 5-14):

- those that grow early and reach stability at a relatively high prevalence level (approximately 33, depending on definition of 'early');
- those that maintain a low prevalence throughout the simulation period (7); and
- those that maintain a low prevalence early in the simulation and then grow relatively suddenly to reach a higher stable prevalence (approximately 15, depending on definition of 'early').

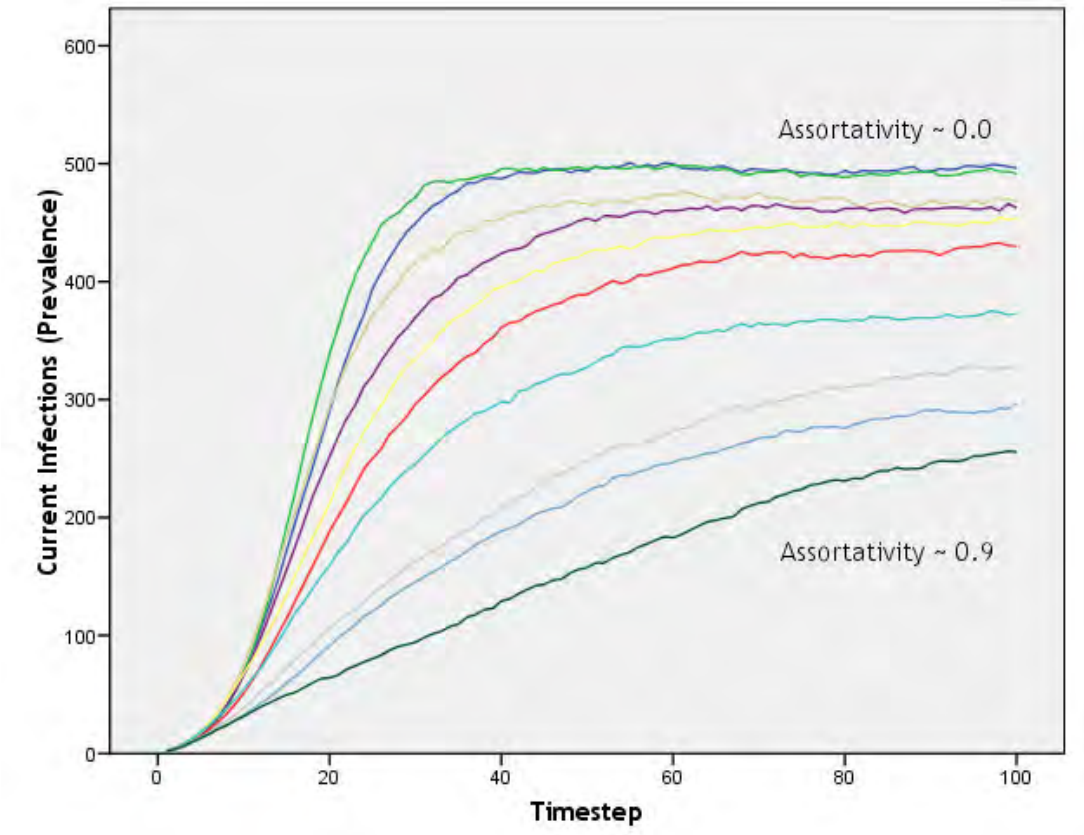

***Figure 5-14: Epidemic size over time (current infections) by assortativity (SIS)** : Real world, infectivity=0.1250, Clustering~0.4*

There is a fourth group that is not included in Figure 5-14; those epidemics that maintain a low prevalence early in the simulation period and then fail. These simulations meet the definition of epidemic but are not used in the analysis because non-zero prevalence is needed to estimate R_0 (see Table 5-27).

The apparent failure to reach instability hence reflects changes in the mean prevalence as individual epidemics jump from low to high prevalence. Given a longer time period, some of the other low prevalence epidemics may also jump to a higher prevalence level, but they may also fail. That is, prevalence and R_0 are bi-modal. Both modal values arise from valid epidemics with an impact on the population, but the lower value is removed from the dataset as simulation period increases. This issue is discussed further in Section 5.5.1.3.

Overall, for each of the degree distribution types and infectivity levels, there is some evidence that epidemics are smaller as assortativity or clustering increases while holding the other property fixed. However, the apparent relationship is not as strong or consistent as that shown by the SIR epidemics.

The second exploratory analysis considers how prevalence varies with both assortativity and clustering coefficient simultaneously (SIR analysis at Section 5.4.1.2) by considering the mean values of prevalence achieved by the end of the simulation.

Consistent with the relationship identified for the SIR epidemics, prevalence decreases with increases in assortativity. For higher assortativity levels only, prevalence also decreases with increases in clustering, whereas clustering had an impact for all assortativity levels in SIR epidemics. Unlike the SIR epidemics, this pattern also occurs for the SIS simulation sets with infectivity of 0.0417, except for the normal degree distribution simulations. However, the simulations with infectivity of 0.0417 show a slower epidemic growth overall and there is no evidence that this relationship would exist if sufficient time had elapsed to ensure all epidemics had achieved equilibrium.

The third exploratory analysis calculates the partial correlation coefficient between epidemic prevalence derived R_0 and each of assortativity and clustering coefficient, removing the influence of the other (equivalent SIR analysis at Section 5.4.1.3).

From Table 5-29, this analysis indicates that there is a significant negative linear relationship between R_0 and assortativity for all distribution types with infectivity of 0.0833 or 0.1250 (correlation of -0.48 to -0.82, $p<0.001$). That is, increases in assortativity are associated with a decrease in the R_0 calculated from endemic prevalence for both these infectivity rates.

***Table 5-29: Correlation between epidemic derived R_0 and network properties (SIS)**: Partial correlation correcting for other network property (significant with p<0.001)*

Distribution	Infectivity	Assortativity	Clustering coefficient
Normal	0.0417	0.48	-0.50
	0.0833	-0.51	-0.74
	0.1250	-0.48	-0.45
Real world	0.0417	-0.19	-0.56
	0.0833	-0.82	-0.37
	0.1250	-0.81	-0.30
Power law	0.0417	-0.75	-0.18
	0.0833	-0.81	-0.19
	0.1250	-0.76	-0.17

The three simulation sets with infectivity of 0.0417 show conflicting results. For real world distribution networks there is a smaller but no less significant negative linear relationship. For normal distribution networks, there is a positive significant linear relationship ($p<0.001$), opposite to that shown in simulations with higher infectivity. For power law distribution networks, the relationship is consistent with the higher infectivity simulations.

For all degree distribution types and infectivity levels, there is a significant linear relationship between R_0 and clustering coefficient (correlation of -0.17 to -0.75, $p<0.001$), despite the lack of any apparent relationship with prevalence from the earlier analysis. That is, increases in clustering are associated with a decrease in the epidemic derived R_0.

One potential explanation of a significant correlation despite an apparent lack of a relationship is that the relationship exists and is consistent, but that clustering makes only very small differences in R_0 as derived from endemic prevalence. Such a relationship would lead to small coefficients for clustering in the regression model.

The final exploratory analysis examined the specific values of assortativity and clustering coefficient, instead of comparing within groups of similar values (SIR analysis at Section 5.4.1.4) with a scatter plot coloured to indicate the mean R_0 value.

Results were consistent with other exploratory analysis results for infectivity levels of 0.0833 and 0.1250. That is, R_0 values decrease as assortativity increases and, for higher assortativity values, as clustering increases.

5.5.1.2 Adequacy of linear regression model without interaction terms

Multiple linear regression was used to model the relationship between the basic reproduction ratio R_0, as derived from prevalence, and the network structure properties of assortativity and clustering coefficient. A separate model is fitted for each degree distribution type and infectivity level.

The modelling process used is the same as for the SIR multiple linear regression. Three sets of regression models are fitted. The first is the multiple regression model containing only assortativity and clustering coefficient as independent variables to model the linear relationship without interaction. Two broader regression models add independent variables progressively using stepwise selection from a pool that includes nonlinear and interaction transformations of assortativity and clustering coefficient (variables listed at Table 5-23 on page 139).

Table 5-30 reports the adjusted R^2 for the linear model and the best model achieved by either set of expanded regression models. As can be seen from this table, there is little gain in adding the nonlinear and interaction terms and, in some cases, there is a reduction in the explanatory power of the model (due to the removal of information from networks with negative assortativity).

However, the normal degree distribution simulations with infectivity of 0.1250 do show a substantial improvement in the model fit with the addition of the nonlinear and interaction variables. Most of this improvement arose with the

addition of a single interaction variable that combines assortativity and a logistic transformation of clustering coefficient.

The same variable was also the first selected for addition to the model starting with assortativity and clustering for the real world degree distribution simulations with infectivity of 0.1250.

Table 5-30: Adjusted R^2 for linear and nonlinear regressions, SIS

Distribution	Infectivity	Adj R^2 linear	Adj R^2 expanded
Normal	0.0417	0.332	0.394
	0.0833	0.662	0.737
	0.1250	0.420	0.626
Real world	0.0417	0.359	0.412
	0.0833	0.717	0.735
	0.1250	0.692	0.755
Power law	0.0417	0.595	0.505
	0.0833	0.679	0.456
	0.1250	0.613	0.372

Based on this analysis, a multiple linear regression model will be fitted to the simulation results. Confirmation of the adequacy of the linear model will be sought through an analysis of the residuals. A nonlinear model is also fitted to the real world and normal degree distributions, infectivity of 0.1250 results (see Section 5.5.1.4). Those models also include the additional independent variable.

5.5.1.3 Linear model of relationship

For all simulation sets, a significant linear model was able to be fitted ($p<0.001$), so there is a linear relationship between R_0 and at least one of assortativity and clustering coefficient. However, the explanatory power of the models varies substantially (see Table 5-31) with between 34% and 72% of the variability of R_0 accounted for by the linear model.

Table 5-31: Performance of multiple linear regression models, SIS

Distribution	Infectivity	R^2	Significance
Normal	0.0417	0.338	p<0.001
	0.0833	0.662	p<0.001
	0.1250	0.421	p<0.001
Real world	0.0417	0.361	p<0.001
	0.0833	0.717	p<0.001
	0.1250	0.693	p<0.001
Power law	0.0417	0.597	p<0.001
	0.0833	0.680	p<0.001
	0.1250	0.613	p<0.001

As for the SIR epidemics, consider the simulations over real world degree distribution networks with infectivity of 0.1250. The regression equation is given by (Table 5-32 on page 157):

Predicted R_0 = 2.069 – 0.502 Assortativity – 0.233 Clustering

As for the SIR results, there are several simulations for which the model substantially overestimates R_0, by up to 7 times the standard deviation in the predicted values of R_0. However, unlike the SIR results, there is some evidence that these very high predictions indicate a systematic error in the regression model.

The plot of residuals against predicted R_0 (Figure 5-15) suggests that any bias is occurring in the lowest of the predicted R_0 values. The plots or residuals against assortativity (Figure 5-16) and clustering coefficient (Figure 5-17) show that the model underestimates the reduction in R_0 associated with large increases in either property.

In combination, these residuals plots suggest there is a systematic bias in the model where both network property values are high, and that this bias can lead to substantial overestimates of the value of R_0; that is, there are simulations for which the actual value is much lower than the value predicted from the model.

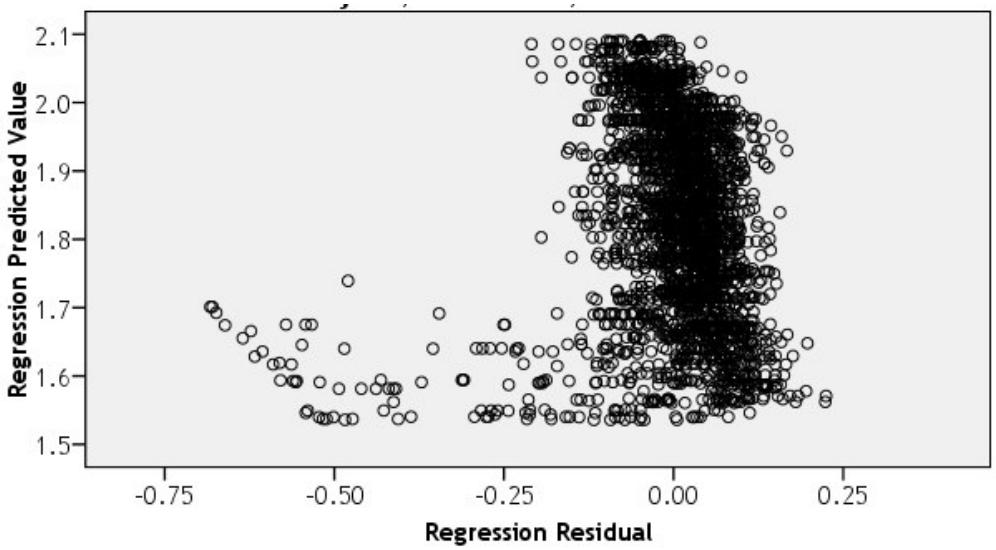

Figure 5-15: Residual plotted against regression prediction for R_0***: SIS, Real world, infectivity=0.1250***

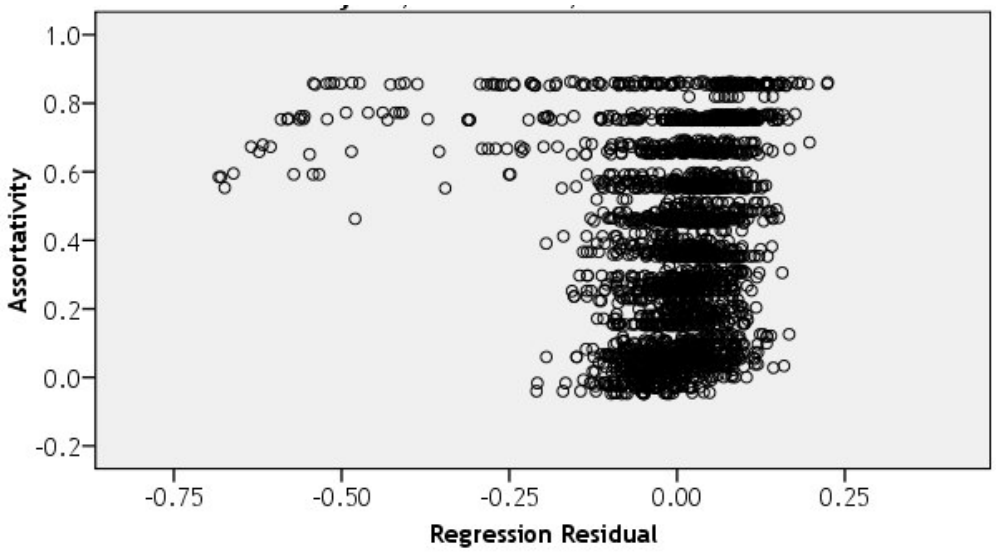

Figure 5-16: Residual plotted against assortativity: SIS, Real world, infectivity=0.1250

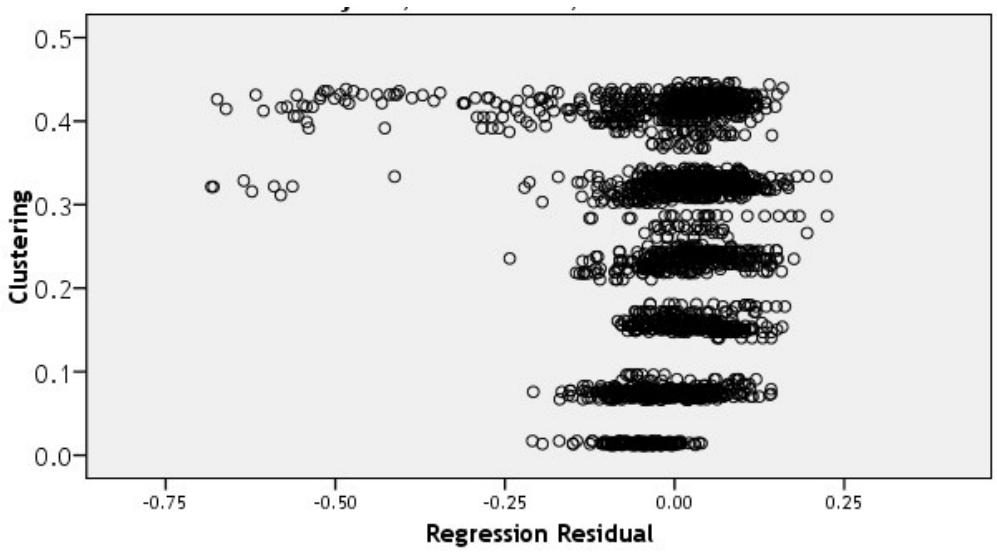

***Figure 5-17: Residual plotted against clustering coefficient**: SIS, Real world, infectivity=0.1250*

A detailed examination of the individual simulations suggests an alternative explanation than a systematic bias. Even for the networks with high values of both assortativity and clustering, the majority of the residuals are close to 0 and only a minority have very large (negative) residuals. Thus, there is not a systematic bias in the model; these results reflect the two different prevalence levels described in Section 5.5.1.1. The model estimates the higher R_0 value, where the majority of the results fall.

Overall, the linear regression model is a good fit for the simulation sets with infectivity of 0.0833 or 0.1250. Explanatory power is high, and there is no evidence of a systematic bias that could indicate the presence of nonlinearity or interaction terms that are not included in the model (though a nonlinear interaction model is also fitted for two simulation sets in Section 5.5.1.4). Regression coefficients for these models are shown in Table 5-32 and all are significant ($p<0.002$).

***Table 5-32: Regression coefficients (SIS), influence of assortativity and clustering on basic reproduction ratio**: Values for infectivity 0.0417 are reported but are only of practical use for the power law degree distribution.*

Distribution	Infectivity	Intercept	Assortativity	Clustering
Normal	0.0417	1.022	0.035	-0.073
	0.0833	1.819	-0.271	-0.829
	0.1250	2.382	-0.344	-0.513
Real world	0.0417	1.150	-0.027	-0.180
	0.0833	1.658	-0.371	-0.207
	0.1250	2.069	-0.502	-0.233
Power law	0.0417	1.158	-0.134	-0.024
	0.0833	1.505	-0.567	-0.082
	0.1250	1.926	-0.833	-0.134

The regression models for real world and normal degree distribution simulation sets with infectivity of 0.0417 have insufficient explanatory power to be of any practical use (Table 5-31), consistent with the situation for SIR simulation sets. However, the model for power law degree distribution and infectivity of 0.0417 is of comparable explanatory power to the models for higher infectivity simulations.

From the regression coefficients, basic reproduction ratio is reduced as assortativity or clustering increases for the exploitable models. As for SIR results, there is insufficient evidence to support any other general rules. The relative contribution of each network property differs between the degree distribution types.

5.5.1.4 Nonlinear model of relationship

As foreshadowed in Section 5.5.1.2, an additional model was fitted to the real world and normal degree distribution simulation sets with infectivity of 0.1250. As well as the assortativity and clustering coefficient variables, these models also include ASigmoidC, calculated as:

$$\text{ASigmoidC} = \frac{\text{Assortativity}}{1 + e^{(5 - 10\ \text{Clustering})}} \tag{5.11}$$

The regression coefficients for the linear and the expanded models are at Table 5-33. The additional term increases the explanatory power of the real world degree distribution model from 69.3% to 75.5%, and the normal degree distribution model from 42.1% to 62.2%.

The ASigmoidC variable has the property that it is very small for low to moderate values of clustering coefficient and approximately equal to the product of assortativity and clustering coefficient for values of clustering coefficient near 0.5. Thus, for the parameter space of the simulations, it has a stronger effect for the higher values of assortativity and clustering, correcting for the weakness in the linear model.

Table 5-33: Regression coefficients, linear and nonlinear models (SIS): *infectivity=0.1250*

Distribution	Model	Intercept	Assortativity	Clustering	ASigmoidC
Normal	linear	2.382	-0.344	-0.513	na
	nonlinear	2.236	-0.138	0.198	-1.663
Real world	linear	2.069	-0.502	-0.233	na
	nonlinear	1.996	-0.365	0.107	-1.184

Despite this correction, the residuals show a similar pattern as arose for the linear model. The histogram for the normal degree distribution simulations is at Figure 5-18. As for the linear model, the large residuals occur where the predicted value of R_0 is much higher than the actual value for some of the epidemics on networks with high assortativity and clustering coefficients.

While the additional nonlinear interaction term increases the proportion of variability in R_0 accounted for by the model, the residuals distribution is essentially unchanged. Thus, there is only limited evidence for a joint and nonlinear effect of assortativity and clustering on R_0, and that evidence is limited to the single simulation set of SIS simulations with normal degree distribution and the highest infectivity rate used.

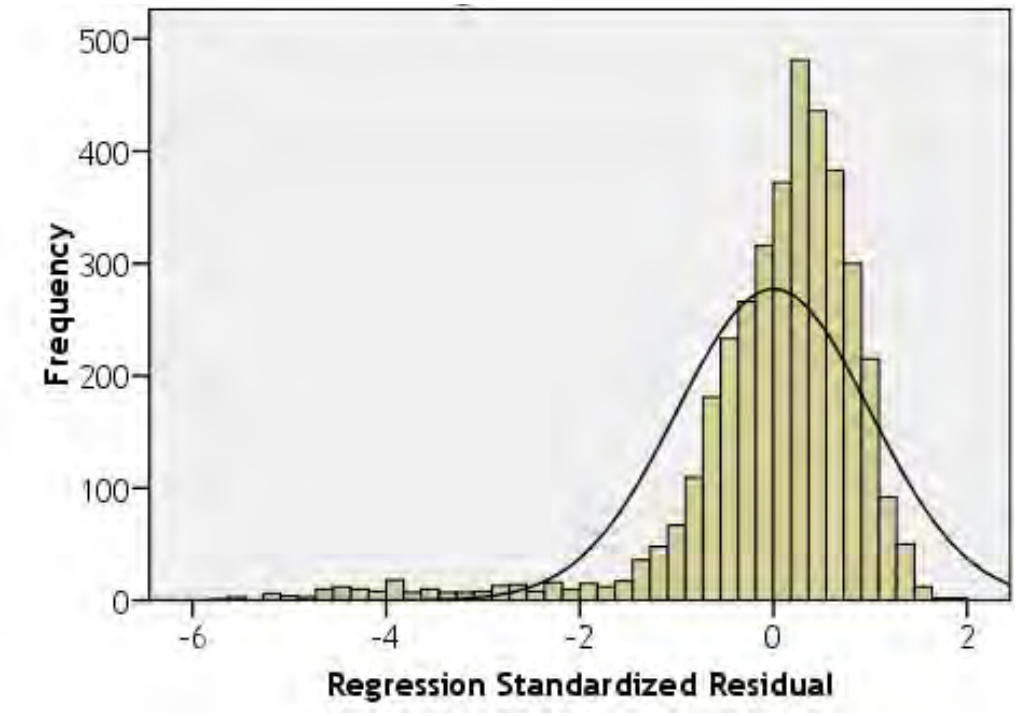

***Figure 5-18: Histogram of standardised regression residuals**: SIS, Normal, infectivity=0.1250, with nonlinear interaction term*

5.5.2 Consistency in epidemic derived basic reproduction ratio

The consistency between the values of R_0 derived from epidemic size in equivalent SIR and SIS simulations was assessed with both exploratory analyses and, for the observed values, a two sample (independent groups) t-test (Tabachnick & Fidell 2006).

The analysis grouped simulations by approximate assortativity and clustering coefficient, in addition to degree distribution and infectivity level. This allows up to 100 SIR and 100 SIS epidemics to contribute to the comparison. As the simulations with infectivity of 0.0417 had few occurrences of epidemics, and inconsistent behaviour of those epidemics (as shown by poor model fitting), they are excluded from the analysis.

5.5.2.1 Consistency in range of observed and predicted values

The first analysis considered the range of epidemic derived R_0 and linear model predicted R_0. Selected ranges are reported at Table 5-34. These suggest that the prevalence of the SIS epidemics generally leads to a higher R_0 than the final size of the equivalent SIR epidemics, but there is a great deal of overlap. This pattern occurs across all degree distribution types, infectivity levels and network property values.

In addition, the minimum R_0 for SIR epidemics falls to approximately 1.00 (the minimum possible) for simulations with lower levels of assortativity and clustering than occurs for SIS epidemics. One explanation for this phenomenon is that epidemics that 'only just' qualify under the definition of epidemic are retained in the SIR dataset, but are likely excluded from the SIS dataset because they fail before the end of the simulation period and prevalence is zero. That is, the effect is an artifact of the experimental design rather than a difference between SIS and SIR epidemic behaviour.

Table 5-34: Range of epidemic derived and predicted* R_0*, SIR and SIS*: Real world, infectivity=0.1250***

		Epidemic R_0		Predicted R_0	
Assortativity	**Clustering**	**SIR**	**SIS**	**SIR**	**SIS**
Approx 0.0	Approx 0.0	1.79-2.08	1.88-2.13	2.00-2.06	2.05-2.09
Approx 0.1	Approx 0.2	1.62-2.08	1.84-2.10	1.75-1.87	1.95-2.00
Approx 0.2	Approx 0.4	1.02-1.80	1.77-2.00	1.52-1.59	1.85-1.89
Approx 0.4	Approx 0.4	1.02-1.72	1.70-1.93	1.39-1.44	1.77-1.80
Approx 0.4	Approx 0.6	1.01-1.50	1.01-1.81	1.24-1.28	1.67-1.69
Approx 0.4	Approx 0.8	1.01-1.37	1.01-1.74	1.09-1.13	1.58-1.60

The R_0 values predicted from the SIR and SIS linear regression models have a small overlap for the networks with approximately zero for both assortativity and clustering coefficient. However, as the structure increases through either property, the range of predicted R_0 values separate, with the prediction for SIS consistently higher than from SIR. The same pattern occurs for normal

degree distribution simulations with infectivity of 0.1250. The simulations for real world and normal degree distributions with infectivity of 0.0833 and both infectivity levels for power law degree distributions show the same separation as network structure increases. However, even for unstructured networks, there is no overlap.

There are two aspects to this pattern that are of potential interest. The higher epidemic derived value of R_0 from SIS simulations as compared to SIR simulations suggests that SIS epidemics are less affected by the impact of degree variation in reducing R_0 derived from mean epidemic behaviour (see Section 5.7.1). The other aspect is that the separation of ranges suggests SIR epidemics are more strongly affected by assortativity and clustering structure in the network or, alternatively, that such structure exaggerates the influence of degree variation.

Alternatively, epidemics with relatively low R_0 values in SIS simulations fail before the end of the simulation period, increasing the mean value of R_0. While this interpretation can explain the high values at the bottom of the R_0 range for SIS simulations, it does not explain why the top of the range values are also high.

5.5.2.2 Individual values of epidemic derived R_0

The second exploratory analysis examined the values of R_0 derived from individual simulations. Figure 5-19 displays the frequency of these values in the absence of network structure. There is substantial overlap, but there are more high values from SIS simulations and low values from SIR simulations. Compare these results to Figure 5-20 for more structured networks. Overlap is considerably reduced and the higher SIS values are more obvious. This pattern can be observed in all degree distribution types and both infectivity levels.

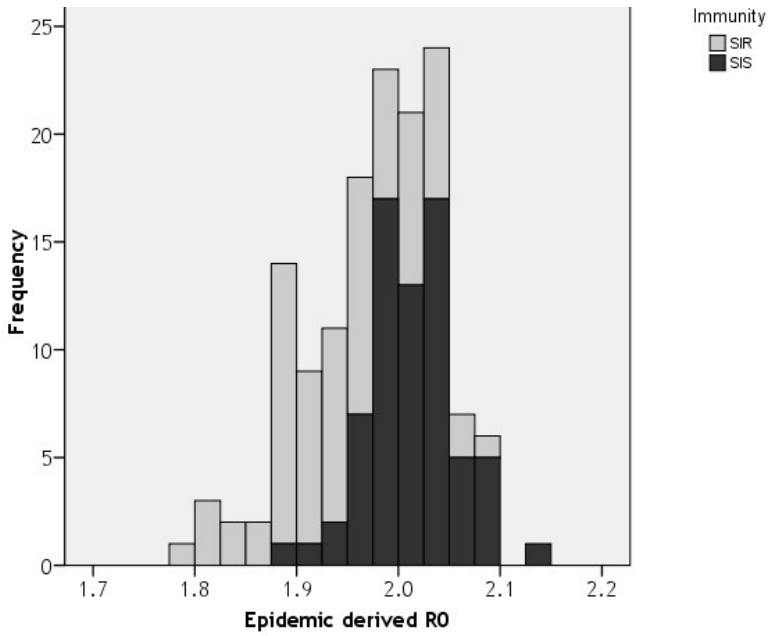

***Figure 5-19: SIR and SIS epidemic derived R0 values, frequencies**: Real world, infectivity=0.1250, assortativity~0.0, clustering~0.0*

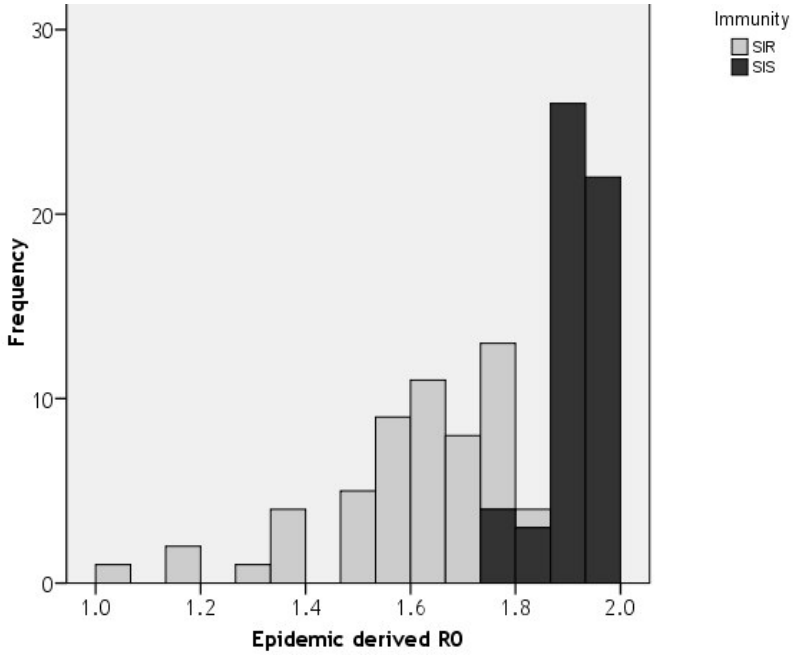

***Figure 5-20: SIR and SIS epidemic derived R_0 values, frequencies**: Real world, infectivity=0.1250, assortativity~0.2, clustering~0.4*

5.5.2.3 Consistency in range of observed and predicted values

A two sample (independent groups) t-test (Tabachnick & Fidell 2006) confirms that the apparent separation is statistically significant. That is, the hypothesis that there is no difference between mean R_0 derived from the final size of SIR epidemics and mean R_0 derived from the prevalence of SIS epidemics is rejected for almost all degree distribution and infectivity combinations ($p<0.01$).

Table 5-35 provides the results of the t-test for the standard examples from real world distribution with infectivity of 0.1250, as well as those simulation sets for which the general hypothesis rejection was not true. Each simulation set is determined by degree distribution type, infectivity level (inf), assortativity (A) and clustering coefficient (C). This table displays the mean value of R_0 for the SIR and SIS simulations and the significance for Levene's test of homogeneity of variance. If that test is significant ($p<0.05$), the t-test reported does not assume equal variances in the distribution of R_0.

Table 5-35: t-test to compare SIR and SIS epidemic derived R_0 values: *selected simulation groups are from standard analysis set, or because results are unusual*

Group	SIR R_0	SIS R_0	Levene's	Reject same?
Real world inf=0.0833 A~0.2 C~0.0	1.46	1.58	p=0.031	no (p=0.195)
Real world inf=0.1250 A~0.0 C~0.0	1.94	2.02	p=0.001	yes (p<0.001)
Real world inf=0.1250 A~0.2 C~0.4	1.59	1.91	p<0.001	yes (p<0.001)
Power law inf=0.0833 A~0.5 C~0.3	1.11	1.29	p=0.071	yes (p=0.002)
Power law inf=0.1250 A~0.5 C~0.4	1.14	1.42	p=0.091	no (p=0.144)
Normal inf=0.0833 A~0.0 C~0.0	1.68	1.70	p<0.001	no (p=0.129)
Normal inf=0.0833 A~0.1 C~0.0	1.68	1.69	p<0.001	no (p=0.271)
Normal inf=0.1250 A~0.1 C~0.2	2.24	2.24	p<0.001	no (p=0.902)
Normal inf=0.1250 A~0.2 C~0.2	2.17	2.24	p<0.001	yes (p=0.025)
Normal inf=0.1250 A~0.3 C~0.1	2.25	2.21	p<0.001	yes (p=0.015)
Normal inf=0.1250 A~0.4 C~0.1	2.19	2.20	p<0.001	no (p=0.470)
Normal inf=0.1250 A~0.5 C~0.1	2.20	2.18	p<0.001	no (p=0.488)

In almost all simulation sets, the mean value of R_0 from SIS simulations is higher than the mean value from equivalent SIR simulations (with equivalence determined by degree distribution, infectivity, assortativity and clustering). The exceptions are the simulations over normal degree distribution networks with infectivity of 0.1250 and limited social structure: assortativity of up to approximately 0.5 with clustering up to approximately 0.1 (except assortativity~0.4 with clustering~0.1).

5.5.2.4 R_0 values from SIR and SIS regression models

The SIR and SIS models can also be compared directly. Excluding the models for the lowest infectivity rate, Table 5-36 displays the linear model regression coefficients (SIR from Table 5-26 and SIS from Table 5-32). The standard errors for these coefficients are at Table 5-37.

Table 5-36: Regression coefficients, influence of assortativity and clustering on basic reproduction ratio

Distribution	Infectivity	Model	Intercept	Assortativity	Clustering
Normal	0.0833	SIR	1.608	-0.236	-1.200
		SIS	1.819	-0.271	-0.829
	0.1250	SIR	2.564	-0.575	-2.580
		SIS	2.382	-0.344	-0.513
Real world	0.0833	SIR	1.565	-0.339	-0.692
		SIS	1.658	-0.371	-0.207
	0.1250	SIR	2.036	-0.781	-0.789
		SIS	2.069	-0.502	-0.233
Power law	0.0833	SIR	1.410	-0.566	-0.301
		SIS	1.505	-0.567	-0.082
	0.1250	SIR	1.748	-1.114	-0.496
		SIS	1.926	-0.833	-0.134

From these tables, the 99% confidence intervals for the regression coefficient values from the SIR and matching SIS models only overlap for the assortativity regression coefficient for infectivity of 0.0833, and do so for all three degree

distribution types. The confidence intervals do not overlap for any other pair of coefficients.

Table 5-37: Standard error of regression coefficients

Distribution	Infectivity	Model	Intercept	Assortativity	Clustering
Normal	0.0833	SIR	0.007	0.013	0.021
		SIS	0.005	0.009	0.014
	0.1250	SIR	0.009	0.017	0.028
		SIS	0.006	0.010	0.017
Real world	0.0833	SIR	0.006	0.010	0.019
		SIS	0.004	0.006	0.012
	0.1250	SIR	0.007	0.011	0.023
		SIS	0.004	0.007	0.014
Power law	0.0833	SIR	0.007	0.024	0.024
		SIS	0.005	0.016	0.016
	0.1250	SIR	0.011	0.037	0.040
		SIS	0.007	0.022	0.024

5.5.3 Discussion

For the two higher infectivity rates, significant linear models were successfully fitted that were able to account for substantial variability in the value of epidemic derived R_0, based only on assortativity and clustering coefficient. For the SIR simulations, the models account for between 58% and 77% of the variation of R_0, and between 42% and 72% for the SIS simulations. With the exception of the SIS simulations on power law degree distribution networks, the models for the lowest infectivity rate had insufficient explanatory power for any practical use.

Additional nonlinear and interaction terms increase the explanatory power of some of the models, but the increases are generally not large. The exception is the SIS simulations over the normal degree distribution networks with highest infectivity, where variability of R_0 accounted for in the model increased from 42.1% to 62.2% with the addition of the nonlinear interaction term.

Residuals analysis suggests that the linear models are appropriate. While there are some epidemics for which the model substantially overestimates R_0, there is no apparent systematic bias. Even the model where the nonlinear interaction term adds explanatory power does not substantially affect the residuals distribution. Thus, there is only limited evidence for a joint or nonlinear effect of assortativity and clustering on R_0, and that evidence is limited to the single simulation set of SIS simulations with normal degree distribution and the highest infectivity rate used.

Hence, it is reasonable to conclude that the structural properties of assortativity and clustering operate separately and with no interaction over the property space investigated.

Ignoring the models with low explanatory power, the regression coefficients for assortativity and clustering coefficient are negative for all simulation sets. That is, R_0 is reduced (and the epidemic is smaller) as either assortativity or clustering increases. This is consistent with the literature. However, the relative contribution of each property differs between degree distribution types and infectivity levels, suggesting that further work is required to develop a general model.

The regression coefficients for the individual degree distribution / infectivity models suggest that network structure can substantially alter R_0. Consider the example of assortativity of 0.2 and clustering coefficient of 0.4. From the models, this level of assortativity reduces R_0 by up to 12.7% and the clustering coefficient reduces R_0 by up to 40.2%. The combined effect is a reduction in R_0 of between 9.4% and 44.7%. On the other hand, if R_0 is being estimated from the behaviour of an epidemic occurring in a highly structured social network, the estimates could substantially underestimate the relevant R_0 for other social networks. This has implications for public health policy, for example in setting target vaccination levels or in calculating health resource needs from epidemic size estimates using R_0.

Finally, the comparison of derived and predicted R_0 values from the SIR and SIS models applying to the same simulation sets show either similarity or a

higher value for SIS simulations. Furthermore, as the networks are more structured, the values separate.

There are two aspects to this pattern that are of potential interest. The higher value of R_0 from SIS simulations suggests that SIS epidemics are less affected by the impact of degree variation in reducing R_0 derived from mean epidemic behaviour. This interpretation could be analysed theoretically. The other aspect is that the separation of ranges suggests SIR epidemics are more strongly affected by assortativity and clustering structure in the network or, alternatively, that such structure exaggerates the influence of degree variation.

As SIS simulations are restricted in their maximum size through the timestep process in a way that does not affect SIR simulations (see Section 5.7.3), the SIS simulations could be expected to show a smaller, rather than larger, epidemic derived R_0. Thus, the difference would be expected to be larger in a study without this restriction.

5.6 Accessible network proportion

Instead of trying to build models that include network properties, some authors have instead considered the impact of those network properties through the secondary reproduction number (Eguíluz and Klemm 2002). This is defined as the mean degree of the network neighbours of the highest degree nodes multiplied by the average transmission probability. That is, how many nodes can a high degree node be expected to infect?

Extend this concept to a higher numbers of steps. A node can infect its neighbours and each can infect their neighbours and so on. However, clustering increases the likelihood that the nodes that are reached by the neighbours have already been exposed by the original infected node. Similarly, degree distribution and assortativity both impact on the number of neighbour nodes at various distances from the original node.

Define the *h*-step extended neighbourhood of a node as the number of unique nodes that can be reached in *h* steps from that node as a proportion of the total number of nodes. For example, the 3-step extended neighbourhood of node *i* is the proportion of the network that is accessible from node *i* in 3 steps. The *h*-step accessible proportion is then the mean *h*-step extended neighbourhood over all nodes.

For each network, this was recorded for *h* values of 1 to 5. Proportion accessible directly measures the effect of network structure as an alternative to using the network structure property values. Models based on this measure were fitted for comparison to models using the property values.

Across the three degree distribution types and three infectivity levels, the epidemic derived R_0 has the highest correlation with the proportion accessible in either 2 or 5 steps. The highest correlation coefficient for each simulation set is at Table 5-38.

The proportion accessible in 2 steps is highly correlated with the proportion accessible in 5 steps, so either but not both are assessed for possible addition to models to predict R_0. The explanatory power of three models are compared in Table 5-39, the assortativity and clustering model already developed, a model using only the proportion accessible, and a model with all three variables.

Table 5-38: Correlation between epidemic derived R_0 ***and accessible proportion of network***

Immunity	Distribution	Infectivity		
		0.0417	**0.0833**	**0.1250**
SIR	Normal	0.379	0.882	0.828
	Power law	0.582	0.788	0.800
	Real world	0.511	0.830	0.813
SIS	Normal	0.282	0.696	0.466
	Real world	0.620	0.741	0.698
	Power law	0.749	0.790	0.748

Models with proportion accessible have similar explanatory power as models using assortativity and clustering coefficient for SIR simulations, but do not perform as well for SIS simulations. Further, proportion accessible has only limited additional explanatory power where assortativity and clustering coefficient are known.

Table 5-39: Adjusted R^2 for network structure models of R_0: *Infectivity=0.1250*

Immunity	Distribution	Assortativity / Clustering	Proportion accessible	Combined
SIR	Normal	0.769	0.685	0.770
	Real world	0.753	0.661	0.754
	Power law	0.586	0.640	0.672
SIS	Normal	0.420	0.217	0.491
	Real world	0.692	0.487	0.695
	Power law	0.613	0.558	0.664

Conceptually, models using proportion accessible incorporate the same network elements as models using values of network structure properties. This table suggests that proportion accessible is worth further investigation as an alternative predictor for R_0 for situations where proportion accessible is able to be estimated but not assortativity or clustering.

5.7 Methodological limitations

Based solely on the basic epidemiological model with uniform degree, it would be expected that the regression intercept would reflect the relationship between R_0 and final size or prevalence as described by equations (5.1) and (5.2). That is, R_0 calculated from the epidemic behaviour in the absence of social structure (assortativity and clustering both zero) should reflect the basic model. From Table 5-40, it is clear that the R_0 values calculated from the epidemic behaviour are quite different from those expected from theory.

Table 5-40: Expected and derived R_0 values based on epidemic behaviour in the absence of network structure

Epidemic	Distribution	Infectivity 0.0417	0.0833	0.1250
Expected	Uniform	0.92	1.71	2.40
Expected	Normal	1.03	1.92	2.69
Expected	Real world	1.26	2.34	3.29
Expected	Power law	1.65	3.04	4.25
SIR	Uniform	1.02	1.74	2.76
SIR	Normal	1.02	1.68	2.41
SIR	Real world	1.05	1.54	1.94
SIR	Power law	1.14	1.53	1.97
SIS	Uniform	na[a]	1.70	2.34
SIS	Normal	1.02	1.70	2.25
SIS	Real world	1.12	1.63	2.02
SIS	Power law	1.16	1.58	2.03

a No epidemics had a nonzero prevalence to enable calculation of R_0

There are several confounding factors that impact on the theoretical R_0 value, the relationship between R_0 and expected epidemic behaviour, and the actual behaviour of the simulated epidemics. Degree variation increases the theoretical value of R_0, as discussed in Section 2.5.2. Other factors are discussed in this section operate so as to decrease the capacity of an infected node to transmit infection, reducing the size of the epidemic. Thus, the calculated R_0 is lower than expected from theory.

5.7.1 Impact of degree variation

In the presence of degree variation, for SIR epidemics, the relationship between final size and R_0 is given by (Hethcote and van Ark 1987, equation 8.7; Britton 2001, equation (3) observation (c)):

$$R_0 = \frac{-\sum_k n_k \log(1-f_k)}{\sum_k n_k f_k} \tag{5.12}$$

where: k is degree
n_k is the proportion of the population with degree k
f_k is the proportion of the population with degree k that became infected (final size)

Similarly, for SIS epidemics, the relationship between prevalence and R_0 is given by (Hethcote and van Ark 1987, equation 7.8; Nold 1980):

$$R_0 = \frac{\sum_k n_k \left(\frac{p_k}{1-p_k}\right)^2}{\sum_k n_k \frac{p_k^2}{1-p_k}} \tag{5.13}$$

where: k is degree
n_k is the proportion of the population with degree k
p_k is the proportion of the population with degree k that is infected at equilibrium (prevalence)

In the absence of degree variation, these relationships simplify to equations (5.1) and (5.2) respectively. However, using the simpler relationship to estimate R_0 from epidemic behaviour can lead to substantial error arising from the skewness in the degree distribution.

From (Nold 1980), the proportion of a specific subpopulation ever infected (SIR) or infected at equilibrium (SIS) is the same proportion as would be infected if that subpopulation made up the whole population, but with the inflation factor also included that recognises the natural weighting of nodes that become infected as discussed at Section 2.5.2.

That is, for an SIR epidemic, the final size (f_k) for the subpopulation with degree k is given by:

$$\log_e(1-f_k) = -f_k R_k \tag{5.14}$$

Similarly, for an SIS epidemic, endemic prevalence (p_k) for the subpopulation with degree k is given by:

$$p_k = 1 - \frac{1}{R_k} \tag{5.15}$$

where the reproduction ratio (R_k) for the subpopulation with degree k is given by:

$$R_k = \beta k \left[1 + \frac{\text{var}(k)}{\hat{k}^2} \right] \tag{5.16}$$

where: β is the probability of transmission of infection
k is the degree
var(k) is the variance of the degree distribution
$\hat{k}$ is mean degree

However, this correction factor is for the reproduction ratio in a completely susceptible population (by definition of R_0). As the higher degree nodes become infected first, the average degree of the susceptible population reduces and the impact of the degree variation also reduces.

To see the effect of these varying influences, consider a specific instance of a degree distribution generated by the Barabási-Albert algorithm (Section 2.3.4) with 1 000 nodes and 4 edges per node. From that degree sequence, a network was generated with the modified Molloy-Reed algorithm (Section 4.1). The constructed network had a maximum degree of 83 and almost no community structure, with assortativity of -0.01 and a clustering coefficient of 0.03.

Five hundred SIR epidemic simulations were run on this network until no infected nodes remained. With infectivity of 0.2 and recovery of 1/3, probability of infection transmission was 0.43. Of the 500 simulations, 438 had

at least 51 nodes ever infected and were retained for the analysis. Number of nodes and the mean proportion infected by degree for these simulations is at Table 5-41. For all epidemic simulations, all nodes with degree 16 or greater became infected except for 1 node with degree 17 (of 4) in one simulation and 1 node with degree 20 (of 4) in another.

Using the degree specific final size and equation (5.14), the empirical values of R_k can be determined. Dividing each R_k by the degree provides the effective transmission probability, which incorporates the underlying transmission probability, the degree variation inflation factor and the various confounding factors related to simulation over a network discussed at Section 5.7.2. This value is shown in the final column of Table 5-41. Note that it is not calculated for higher degree nodes, where all nodes become infected and thus equation (5.14) is undefined.

Table 5-41: Epidemic impact by degree *(nodes with degree not specifically displayed had final size of 1 except as noted in text)*

Degree	Nodes	Infected	Final size	Calculated R_k	R_k per degree
4	344	300.3	0.873	2.36	0.59
5	183	169.2	0.925	2.80	0.56
6	124	118.4	0.954	3.24	0.54
7	81	78.8	0.973	3.72	0.53
8	46	45.2	0.984	4.18	0.52
9	35	34.7	0.990	4.69	0.52
10	24	23.9	0.994	5.21	0.52
11	20	19.9	0.996	5.66	0.51
12	24	24.0	0.999	6.70	0.56
13	15	15.0	0.998	6.50	0.50
14	22	22.0	1.000	8.08	0.58
15	12	12.0	0.999	7.47	0.50
16-82	68	all	1.000	undefined	undefined
83	2	2.0	1.000	undefined	undefined
7.98	1 000	933.4	0.933		

Taking the mean of these values weighted by the number of nodes (only up to degree 15), gives an effective transmission probability of 0.56 per degree. This is higher than the underlying transmission probability of 0.43, but much lower than the value implied by the correction factor in equation (5.16), which is 0.94 (network degree coefficient of variation is 1.09).

Applying the effective transmission probability to each degree individually, converting to final size for that degree and adding across degrees, the expected number of nodes infected is 921.0. This is lower than the mean simulation result of 933.4 because of the high proportion of nodes with degree 4, which had a relatively high value of R_k by degree. Alternatively, using the mean degree of 7.98, R_0 is 4.464 and the expected number of nodes infected is 987.8.

The difference between expected final size of 921 nodes and 988 nodes (from 1 000) is substantial and is entirely created by the positively skewed degree distribution. The low degree nodes are both the highest number and the least likely to become infected.

This difference is exacerbated by the nonlinearity of the relationship between final size and R_0 (see Figure 5-21). Converting each to R_0, the values are 2.76 for the degree specific expected final size, 2.90 for the mean simulation final size and 4.64 for the mean degree expected final size. Thus, a 7.3% overestimate in the final size converts to a 62.0% overestimate of R_0.

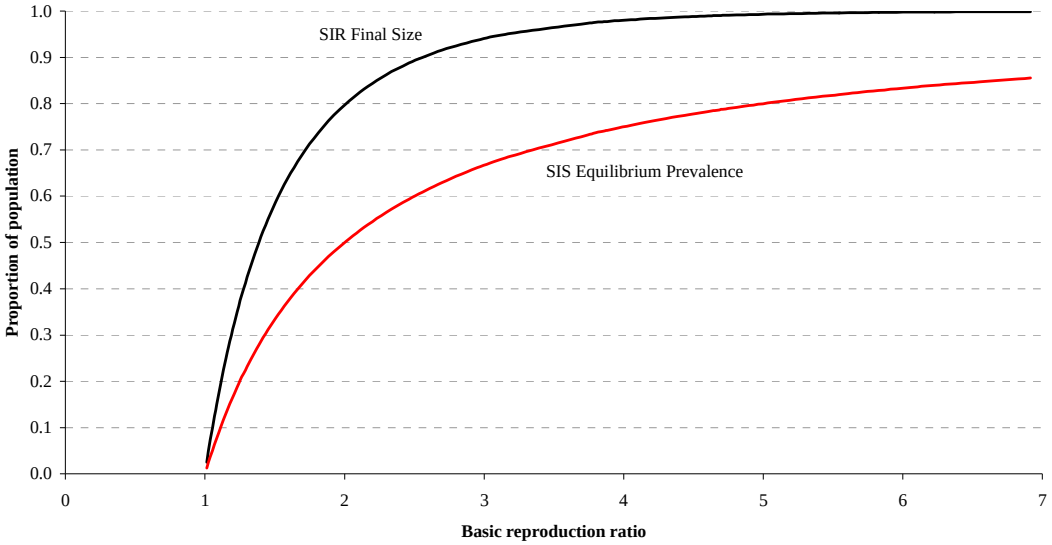

Figure 5-21: Relationship between epidemic behaviour and derived R_0

A similar analysis can be performed with SIS epidemics with similar results, but SIS epidemic simulations are much more affected by the simulation update process as described in Section 5.7.3, confounding the analysis.

Within a degree distribution type, the impact of degree variation will be different for each distribution instance despite generation by the same algorithm. The specific instance is used as the target degree sequence for the neighbour networks with different social parameters so the experimental design preserves the subpopulation structure and hence impact on epidemic size. However, resolution failure of the neighbour network introduces differences in the achieved subpopulation structure.

5.7.2 Representing contacts as a static network

For both SIR and SIS epidemics, the relationship between epidemic size and R_0 assumes the mass action principle (see Section 2.1.1). Specifically, this assumes that the number of infections produced by an infected node is proportional to the probability that the node at the other end of the edge is susceptible.

However, when using a static network to describe the contact process, the degree variation correction factor in equation (5.16) overstates the capacity of a node to reproduce in a network. This is because the node that infected a node is not susceptible and the available neighbours must be reduced by 1 (Diekmann & Heesterbeek 2000, section 10.5.2).

5.7.3 Simulation update process

The simulation model has synchronous timing (see Section 3.2.1). Consider an infected node in an SIS simulation that becomes infected immediately after recovering from the current infection. The state counter is updated at the end of each timestep. That is, the node is counted as susceptible for the timestep in which it recovers. For this study, the mean duration of infection is 3 timesteps. Thus, even if the node is immediately infected whenever it recovers, it is susceptible for one timestep in four on average. Across all nodes, this translates into a maximum equilibrium prevalence of 0.75.

One way to reduce the impact of this effect is to split the timesteps into smaller slices. For example, if a model was structured by days with a mean infectious period of 3 days, the internal slices could represent one hour. Probability of infection and recovery would need to be suitably rescaled. Results can still be kept at the longer scale simply by only counting states every 24 slices. The impact of this scale change is that a node immediately becoming reinfected is only susceptible for one 'hour' instead of one 'day' and the maximum prevalence increases from 1/4 to 1/73.

Note that the final size of an SIR epidemic does not have a similar artificial maximum. However, for both epidemic types, time rescaling could have a more subtle impact on epidemic behaviour, especially in networks with high clustering. This is because the distribution of infectious periods is changed, despite maintaining the mean with appropriate changes in probabilities.

The synchronous timing of the simulation update also leads to competition for susceptible nodes in the same timestep, and any epidemic is smaller than expected from the underlying value of R_0. Once a significant proportion of

the population is infected, several infected nodes may be connected to the same susceptible node and the potential to generate new infections is reduced.

This effect is easiest to see with an SIS simulation on a uniform network. Consider a network with 1 000 nodes and degree of 8 for each node, randomly connected. With infectivity of 0.1 and recovery of 1/3, the infection transmission probability is 0.25 so theoretical R_0 is 2 and equilibrium prevalence is 50%.

Consider a susceptible node within this network with prevalence of 50%. It has edges with 8 nodes of which 4 are infected. Each infected node has probability 0.25 of successfully infecting the susceptible node. Thus, it has probability 0.684 of becoming infected. However, from the perspective of the four infected nodes, each had the chance of infecting the susceptible with probability of 0.25. Thus, four contacts were used with a total expectation of 1 new infection to generate a new infection with probability of only 0.684. Thus, effective reproductive ratio is less than 1 and equilibrium prevalence must be higher than the theoretical value to accommodate the infected nodes competing for the same susceptible.

5.8 Discussion

This chapter examines the impact of three network properties (degree distribution, assortativity and clustering) on epidemic occurrence and epidemic derived basic reproduction ratio R_0.

Degree heterogeneity was found to interact with infectivity in its impact on epidemic behaviour. For low infectivity levels, degree heterogeneity increases epidemic occurrence or epidemic size, and the opposite occurs for higher infectivity levels. For epidemic size, the simulations suggest that clustering may impact on the infectivity level at which this reversal occurs. In general, however, the presence of network structure through positive assortativity or clustering did not affect these relationships.

To better focus on network structure, the study separately considered each of the simulation sets defined by degree distribution type and infectivity. Within degree distribution type, the experimental design uses identical target degree sequences for the networks with varied structural properties (assortativity and clustering coefficient).

For epidemic occurrence, the simulations found that the presence of network structure decreases the likelihood of an epidemic occurring. However, the property or interaction contributing most significantly to that decrease varied between simulation sets.

For epidemic derived R_0, clustering and assortativity were found to independently and linearly decrease its value. Again, the relative contribution of each property was inconsistent between simulation sets.

The potential impact of network structure was found to be substantial. With reasonable real world values of network properties, R_0 can be reduced by up to 45% in comparison with R_0 based on simulations with zero values for assortativity and clustering.

Further, due to the nonlinearity in the relationship between epidemic behaviour and derived R_0, small differences in R_0 can lead to large differences in the estimated impact of an epidemic for values up to about 2.5 (see Figure 5-21). For example, $R_0 = 1.5$ leads to an SIR epidemic affecting 58.3% of the population, but the final size for $R_0 = 1.7$ is 69.1%.

The simulations also suggest that the impact of all three network properties considered may impact more strongly on SIR epidemics than SIS epidemics.

Chapter 6: Conclusions

Historians and anthropologists recognise the potentially substantial role of epidemics in social and cultural development through a variety of mechanisms, such as affecting the outcome of wars or creating labour shortages. The basic reproduction ratio R_0 is a key parameter in epidemiological models. It incorporates information about the disease itself but also about the society in which it is embedded. The objective of this study is to draw out the implications of that embedding.

The role of social structure in epidemic behaviour can be studied from the perspective of three overlapping fields of study. Sociologists studying social networks have defined a variety of properties and calculated their values for many different real world social networks. Mathematicians and other physical scientists have studied dynamic processes, including epidemic spread, on idealised networks through mathematical techniques and by simulation. Finally, epidemiologists have incorporated elements of social structure in models of disease spread.

Social networks are well studied and there are three properties (amongst others) that social networks display that are in conflict with the assumptions about social structure used to develop the epidemiological models based on R_0: degree heterogeneity, positive degree assortativity and clustering.

The literature (summarised in Chapter 2) verifies the importance of the three selected network properties for social networks. It also suggests that the impact of degree heterogeneity on epidemic behaviour is well understood. However, the implications of assortativity and clustering coefficient have received only limited attention and the joint effect has not been studied.

Each of these three social network properties influences the number of susceptible nodes available to any infected node, either through direct connection, or through the connections of the neighbour nodes. Thus, any

mathematical model would be very complex and simulation appeared the more viable analysis technique.

Simulation, however, requires an algorithm to generate networks with a range of values for the properties of interest.

The experimental design (described in Chapter 3) assumes such an algorithm is available. To focus on the specific study question, only a single network size (1 000 nodes) and target mean degree (8) is used for all networks. Up to 10 networks are generated for each combination of degree distribution type, assortativity value and clustering coefficient value. Ten simulations are carried out on each network for each of 3 infectivity levels with both SIR and SIS immunity settings. Thus, there are up to 100 simulations for each network property and epidemic parameter combination.

In total, 66 720 simulations are conducted with various network and epidemic properties. These simulations allow the relationship between epidemic behaviour and network properties to be analysed. Two aspects of epidemic behaviour are examined: whether an epidemic occurred and, if so, the basic reproduction ratio as derived from epidemic size.

6.1 Generating networks with specific properties

The first major contribution of this thesis is the development of a network generation algorithm that is able to generate networks with independent control of degree distribution, assortativity and clustering coefficient. This algorithm responds to **secondary research question 4**:

> How can networks be generated for simulations with various values of degree sequence, assortativity and clustering coefficient, separately and jointly?

Such an algorithm strengthens the link between the fields of social networks as studied by sociologists and network analysis as studied by mathematical physicists, by enabling generation of networks with more realistic properties. This algorithm is developed in Chapter 4.

The general approach has three phases. The nodes are assigned target degrees and uniform randomly located in a notional space. The nodes are then moved so that nodes with similar target degrees are closer together. Edges are created taking into account the target degrees but favouring nodes in the local area or (physical) neighbourhood.

I implement this approach with a specific algorithm: one dimensional ring with stochastic node pair swaps (Section 4.2.2).

Testing demonstrates that the algorithm is valid: the three phase neighbour approach is able to target degree sequence, assortativity and clustering coefficient separately. Previously published algorithms have, at best, been able to generate networks with a target degree sequence and only one structural property. The capacity for two structural properties with independent control makes the algorithm very flexible, with much broader potential applications than epidemiology.

I validate the implementation and describe the relationship between input parameters and properties of generated networks in Section 4.3. The algorithm is reliable with some limitations.

The target degree sequence can take any shape and the exact degree sequence is achieved if the algorithm resolves (as compared to the generated degree sequence being random with an expected value that matches the target). For the target degree sequences with the greatest variation, however, the algorithm is unable to resolve and the generated networks have slightly reduced variation and mean degree.

For assortativity, the target value is also directly specified as an input to the algorithm. While the algorithm does not exactly match the target assortativity, tolerance is arbitrary. For the small (100 node) networks tested, over 50% of the generated networks had assortativity within 0.05 of the target value. Further, the algorithm is able to generate networks across the entire feasible space. Generated networks were able to achieve assortativity values of over 90% of the maximum possible given the degree sequence.

For clustering coefficient, the relevant input to the algorithm is the edge creation probability. Achieved clustering coefficient is approximately half the probability value. As there are real world networks with clustering coefficient greater than 0.5, this limitation makes the algorithm unsuitable for investigating social networks with very high clustering coefficients.

The algorithm also has difficulty generating networks with higher assortativity in combination with a clustering coefficient near zero. However, both assortativity and clustering coefficient are properties arising from the characteristics of specific node pairs that have edges between them, and feasible values of these properties may be interdependent as well as dependent on the degree sequence. Further work is warranted to examine the relationship between the two structural properties to determine whether the difficulty in achieving certain combinations of properties is a characteristic of the algorithm or such networks are not feasible.

Clustering was also related to the small world property. Compliance with the small world property degenerates for the highest probability values.

For this study of the relationship between epidemic behaviour and social network properties, the one dimensional ring with stochastic node pair swaps implementation of the neighbour algorithm was sufficient, and Chapter 4 also details the properties of the networks generated for the study.

In addition to the neighbour network generation algorithm, I also developed a modification to the Molloy-Reed algorithm (see Section 4.1) to implement the basic epidemiological model in the network context.

6.2 Impact of network properties on epidemic behaviour

The second major contribution of this thesis is the use of simulated epidemics to investigate the relationship between properties of the social network over which an epidemic occurs, and epidemic occurrence and basic reproduction ratio. These simulations enable epidemiological models to take greater

advantage of knowledge about social networks and processes that occur on networks.

This analysis responds to the **primary research question**:

> What is the relationship between epidemic behaviour and three key features of social networks: positively skewed degree distribution, positive clustering coefficient and positive (degree) assortativity?

This can be separated into **three secondary research questions** that focus on specific aspects of the relationship:

> 1 How does each of these properties affect epidemic occurrence?
>
> 2 How does each of these properties affect the basic reproduction ratio R_0?
>
> 3 Do these social network properties influence epidemic behaviour separately or jointly and, if the latter, how do they interact?

Chapter 5 details the results of the simulations conducted. The full analysis is presented for the simulation set with real world degree distribution and the highest infectivity rate, with summary results for all simulation sets. The detailed results for other simulation sets are included on the supplementary DVD, and indexed at Appendix C.

The first aspect of epidemic behaviour investigated is the relationship between network properties and epidemic occurrence (secondary research question 1). I first propose an operational definition of epidemic for simulation studies (Section 5.1).

The relationship with degree heterogeneity has been well studied in the literature, but previous studies have assumed the absence of network structure. I find that the presence of assortativity and clustering has no apparent impact on the relationship between epidemic occurrence and degree

heterogeneity (Section 5.2.1). However, the relationship is not as clear as the literature would suggest. For the lowest infectivity rate, the networks with greater degree variation had a higher occurrence of epidemics as expected. For the higher infectivity rates, the results were the opposite, with the normal degree distribution simulations having the highest occurrence of epidemics. Further work is required to draw out where this reversal occurs, particularly concerning the relationship to both infectivity and mean degree.

The relationship between epidemic occurrence and network structure has attracted only limited attention, with the literature suggesting that assortativity increases occurrence and clustering decreases occurrence. This relationship is examined in (Section 5.3), controlling for degree heterogeneity.

There are nine simulation sets defined by degree distribution and infectivity for both SIR and SIS epidemics. A significant relationship was found between network structure and epidemic occurrence for SIR epidemics in eight simulation sets and for SIS epidemics in five simulation sets. While the particular property differed, all the relationships found that the proportion of epidemics decreased as assortativity or clustering coefficient or their interaction increased.

For assortativity, the result from this study is in conflict with the majority of previous studies. However, it is a more general result, not relying on a particular mixing scheme or joint degree distribution. One previous simulation study using power law degree distribution networks also found increases in assortativity decrease epidemic occurrence.

For clustering coefficient, the result from this study confirms the results of the single previous study.

In addition, this study provides the first indication that assortativity and clustering coefficient have a joint effect in decreasing epidemic occurrence, separate to their individual effects.

The second aspect of epidemic behaviour analysed is the basic reproduction ratio R_0, derived from the epidemic final size (SIR) or equilibrium prevalence (SIS), responding to secondary research question 2. Epidemic size has a strictly monotonic but nonlinear relationship with R_0.

For degree heterogeneity, this study generally confirms previously published results that the mean epidemic size (and hence derived R_0) decreases as the variance of degree increases (Section 5.2.2). However, the opposite relationship is found for SIR simulations with relatively low infectivity combined with high clustering, where the normal degree simulations have the smallest epidemic size. That is, the epidemic becomes trapped in a small region of the network under specific conditions.

To estimate the quantitative impact of network structure on R_0, regression models were fitted with assortativity and clustering coefficient as the independent variables (Sections 5.4 and 5.5). For the lowest infectivity level, R_0 was very close to its minimum of 1 and, while the fitted models were significant, they were able to account for only a small proportion of the variation in R_0.

For the higher infectivity rates, assortativity and clustering coefficient were each found to have a significant linear relationship with R_0 for all simulation sets for both SIR and SIS epidemics. Increases in either property reduce the value of R_0 derived from epidemic size. This result is consistent with previous qualitative studies that have identified the direction of the relationship between these properties and epidemic size.

The potential size of this impact is critical in determining whether network structure can be ignored for health planning and other applications of epidemiological models. Analysis of real world social networks suggests that relatively high but reasonable values of these properties are 0.2 for assortativity and 0.4 for clustering coefficient. From the regression models (Section 5.5.3), R_0 in such a network would be reduced by 9.4% to 44.7% compared to R_0 for a network with the same degree sequence but no structure. As such, calculating R_0 from an epidemic over a less structured

social network could substantially underestimate the potential impact of the same epidemic in a more structured social network.

For almost all simulation sets, including nonlinear and interaction terms did not substantially improve the explanatory power of the regression model. For one simulation set, a nonlinear interaction term did provide substantial additional explanatory power for the regression model. That simulation set is SIS epidemics with the highest infectivity level over normal degree distribution networks. Responding to secondary research question 3, as this occurred for only one simulation set from twelve (Section 5.5.1.4), it is reasonable to conclude that any joint effect from the two network structure properties can generally be excluded from models of R_0.

As well as models using the network properties of assortativity and clustering coefficient, I also propose a method to include the effect of network social structure that directly measures the effect of network structure; accessible proportion of network in a specified number of steps. I find (Section 5.6) that this measure is generally less successful in accounting for the variability in the value of epidemic derived R_0 and provides only limited additional explanatory power.

6.3 Future work

Clearly there are applications other than epidemic behaviour for which similar network properties are likely to be important. Other dynamic processes on social networks are already studied using the model of contagion, such as information transfer, development of fads and opinion exchange. The network generation algorithm allows suitable networks to be constructed for simulation of any of these dynamic processes.

The three phase network generation approach is very flexible. The one dimensional ring with edge swapping used in this study to generate the networks for epidemic simulation is only one possible implementation. There are several potential research threads in developing the algorithm for other simulation studies.

The first research thread concerns improving the performance of the specific implementation used in this study (Section 4.2.2). For example, there is no procedure to rewire existing edges when a node with high target degree is attempting to make edges and there are insufficient nodes available. Breaking an existing edge potentially makes two new nodes available and may improve resolution and hence convergence to the target degree sequence. A rewiring process could also be developed to improve compliance with the small-world property. Finally, the assortativity overshoot can be reversed by randomly swapping the locations of pairs of nodes.

The second research thread concerns alternative implementations of the three phase neighbour network generation approach, to improve reliability in generating specific properties or to extend the algorithm for other network properties. For example, using a higher dimensional space is likely to weaken the control over clustering coefficient, but may allow the algorithm to control mean geodesic for the generated networks through the edge creation phase. Other types of assortativity could also be targeted, such as age or gender association.

Processes other than edge swapping are also available for the layout update phase. For example, a combination of attractive and repulsive forces could be used, with attractive forces based on the properties of the nodes (such as target degree) and the network property of interest. Repulsive forces, based on distance between nodes, would be required to maintain separation.

The third research thread concerns the feasible property space for generated networks. Degree distribution restricts feasible assortativity. This study suggests that there may also be a relationship between feasible or likely assortativity and clustering coefficient (Section 4.4.2). Such a relationship has ramifications when examining real world social networks, because a particular value of one property may affect the probability distribution of the values for the other property in random networks. Thus, the other property value would need to be compared to its expected value rather than considered at its absolute level.

For the specific issue of the impact of social network properties on epidemic behaviour, there are also several threads for future work.

To focus on the social properties of interest, the experimental design fixed other network properties and simplified epidemic parameters. These design decisions limited the scope of simulations and hence the scope of the results. While it is reasonable to expect any identified relationships between network properties and epidemic behaviour to also exist more generally, further studies would be needed to verify such relationships. In particular, it is likely that the values of regression coefficients would depend on factors such as:

- network size
- mean degree of network
- infectivity rate (probability of transmission of infection)
- recovery rate (or mean period of infection)

Some results may also be affected by the selection method for the initial infected node, particularly for network property relationships with epidemic occurrence. For example, if the initial selection is weighted by degree, the proportion of epidemics that occur is likely to increase for the skewed distribution types.

Further studies could also consider more complex epidemic states, including latency periods, partial immunity and time dependent infectivity. Such studies are possible with the neighbour algorithm to generate networks, paired with suitable implementation of epidemic parameters through simple models or more complex multi-agent systems.

The study also identified some issues that would benefit from targeted research. The results suggest that the influence of degree heterogeneity on epidemic behaviour has a complex interaction with infectivity but finer discrimination in both properties is required to draw out that relationship (Section 5.2). Some of the results suggest that network structure may restrict SIR epidemics more than it does SIS epidemics (Section 5.5.2).

The long term objective of these studies is to develop more general models that relate epidemic behaviour (or other dynamic processes) to the properties of the societies in which they occur. If such models can be developed, published values for different types of social networks could eventually allow models to be calibrated from a real world process on one social network, then applied to other social networks with corrections for the differences in social structure.

Separately, high resolution models are required to compare vaccination, quarantine and other public health strategies, to identify which strategies are appropriate in which circumstance. To provide all the information required for such models, the relationship between social structure and epidemic behaviour would need to be consider many aspects not explored in this study. This includes the additional parameters identified above (such as network size and complex epidemic states). However, it also potentially includes relaxing some of the assumptions inherent in the network model if the diseases to be considered did not require close or direct contact for transmission. In particular, the network assumptions of a fixed set of contacts (static) and equal and constant probability of transmission to each contact (unweighted) are likely to be valid for only specific diseases. Nevertheless, the approach used in this study through the neighbour algorithm could potentially be extended to generate networks where these assumptions are not made.

General models could be used to identify those aspects of network structure and of most importance. Future research could be targeted to improving the measurement of real world values of those network and epidemic parameters known to be important and hence contribute to the development of more effective high resolution simulations to improve epidemic management.

Chapter 7: References

Adler, F.R. 1992, "The effects of averaging on the basic reproduction ratio", *Mathematical Biosciences*, vol. 111, pp. 89-98.

Agar, M. 2003, "My kingdom for a function: Modeling misadventures of the innumerate", *Journal of Artificial Societies and Social Simulation*, [Online], vol. 6, no. 3, Available from: <http://jasss.soc.surrey.ac.uk/6/3/8.html>.

Aiello, W., Chung, F. & Lu, L. 2001, "A random graph model for power law graphs", *Experimental Mathematics*, [Online], vol. 10, pp. 53-66, Available from: <http://www.expmath.org/expmath/volumes/10/10.html>.

Albert, R. & Barabási, A.-L. 2002, "Statistical mechanics of complex networks", *Reviews of Modern Physics*, vol. 74, pp. 47-97.

Anderson, C.J., Wasserman, S. & Crouch, B. 1999, "A p* primer: logit models for social networks", *Social Networks*, vol. 21, pp. 37-66.

Anderson, R.M. 1991, "Discussion: The Kermack-McKendrick epidemic threshold theorem", *Bulletin of Mathematical Biology*, vol. 53, no. 1/2, pp. 3-32.

Anderson, R.M. & May, R.M. 1992, *Infectious Diseases of Humans: Dynamics and Control*, Oxford Science Publications, Oxford University Press, Oxford.

Andersson, H. & Britton, T. 1998, "Heterogeneity in epidemic models and its effect on the spread of infection", *Journal of Applied Probability*, vol. 35, no. 3, pp. 651-61.

Badham, J.M., Abbass, H.A. & Stocker, R., *Standardisation and parameterisation of Keeling"s network generation algorithm*, ALAR Technical Report Series, Artificial Life and Adaptive Robotics Laboratory, University of NSW (ADFA).

Bailey, N.T.J. 1975, *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd edn, Charles Griffin & Company, High Wycombe.

Ball, F. 1985, "Deterministic and stochastic epidemics with several kinds of susceptibles", *Advances in Applied Probability*, vol. 17, pp. 1-22.

Ball, F. & Clancy, D. 1993, "The final size and severity of a generalised stochastic multitype epidemic model", *Advances in Applied Probability*, vol. 25, no. 4, pp. 721-36.

Barabási, A.-L. & Albert, R. 1999, "Emergence of scaling in random networks", *Science*, vol. 286, no. 15 Oct 1999, pp. 509-12.

Becker, N.G. 1973, "Carrier-borne epidemics in a community consisting of different groups", *Journal of Applied Probability*, vol. 10, no. 3, pp. 491-501.

Békéssy, A., Békéssy, P. & Komlós, J. 1972, "Asymptotic enumeration of regular matrices", *Studia Scientiarum Mathematicarum Hungarica*, vol. 7, pp. 343-53.

Bender, E.A. & Canfield, E.R. 1978, "The asymptotic number of labeled graphs with given degree sequences", *Journal of Combinatorial Theory*, vol. 24, pp. 296-307.

Bernoulli, D. 1766, "Essai d"une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l"inoculation pour la prévenir", *Mémoires de Mathématique et de Physique, Académie Royale des Sciences*, pp. 1-45.

Boguña, M. & Pastor-Satorras, R. 2002, "Epidemic spreading in correlated complex networks", *Physical Review E*, vol. 66, no. 047104.

Bollobás, B. 2001, *Random Graphs*, Cambridge Studies in Advanced Mathematics 73, 2nd edn, Cambridge University Press, Cambridge.

Bollobás, B. & Riordan, O.M. 2003, "Mathematical results on scale-free random graphs" in *Handbook of Graphs and Networks: From the Genome to the Internet*, ed S. Bornholdt & G. Schuster, Wiley-VCH, Weinheim, pp. 1-34.

Britton, T. 2001, "Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 4, pp. 705-15.

Caldarelli, G., Capocci, A., de los Rios, P. & Muñoz, M.A. 2002, "Scale-free networks from varying vertex intrinsic fitness", *Physical Review Letters*, [Online], vol. 89, no. 258702, Available from: <http://link.aps.org/abstract/PRL/v89/e258702>.

Center for Disease Control, (10 November 2007), *Overview of Influenza Surveillance in the United States*, [Online], Available from: <http://www.cdc.gov/flu/weekly/fluactivity.htm>.

Diamond, J. 1998, *Guns, Germs and Steel: A Short History of Everybody for the Last 13,000 Years*, Random House.

Diekmann, O. & Heesterbeek, J.A.P. 2000, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, Wiley Series on Mathematical and Computational Biology, John Wiley & Sons.

Diekmann, O., Heesterbeek, J.A.P. & Metz, J.A.J. 1990, "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations", *Journal of Mathematical Biology*, vol. 28, pp. 365-82.

Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. 2000, 'Structure of growing networks with preferential linking', *Physical Review Letters*, vol. 85, pp. 4633-4636.

Dwight, H.B. 1961, *Tables of Integrals and Other Mathematical Data*, 4th edn, MacMillan Publishing Co Inc, New York.

Eames, K.T.D. & Keeling, M.J. 2002, "Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 13330-5.

Eguíluz, V.M. & Klemm, K. 2002, "Epidemic threshold in structured scale-free networks", *Physical Review Letters*, vol. 89, no. 10, p. 108701.

Eidelson, B.M. & Lustick, I. 2004, "VIR-POX: An agent based analysis of smallpox preparedness and response policy", *Journal of Artificial Societies and Social Simulation*, [Online], vol. 7, no. 3, Available from: <http://jasss.soc.surrey.ac.uk/7/3/6.html>.

Elliott, A.C. & Woodward, W.A. 2007, *Statistical Analysis Quick Reference Guidebook: with SPSS Examples*, Sage Publications, Thousand Oaks, California.

Erdös, P. & Rényi, A. 1960, "On the evolution of random graphs", *Publications of the Institute of Mathematics, Hungarian Academy of Science*, vol. 5, pp. 17-60.

Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z. & Wang, N. 2004, "Modelling disease outbreaks in realistic urban social networks", *Nature*, vol. 429, pp. 180-184.

Frank, O. & Strauss, D. 1986, "Markov graphs", *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 832-42.

Gilbert, N. & Troitzsch, K.G. 1999, *Simulation for the social scientist*, Open University Press, Buckingham.

Gkantsidis, C., Mihail, M. & Zegura, E. 2003, "The Markov Chain simulation method for generating connected power law random graphs", *Proceedings of SIAM Alenex 03* [Online], Available from: <http://www.cc.gatech.edu/~gantsich/TopologyGenerators/TheMCSimulationMethodForGeneratingConnectedPLRG.pdf>.

Goldberg, L.A. & Jerrum, M. 1996, "Randomly sampling molecules", *Proceedings of Eighth SIAM Conference on Discrete Mathematics*, Jun 17-20, 1996, Baltimore, Maryland [Online], Available from <http://citeseer.ist.psu.edu/ goldberg96randomly.html>.

Goldspink, C. 2002, "Methodological implications of complex systems approaches to sociality: Simulation as a foundation for knowledge", *Journal of Artificial Societies and Social Simulation*, [Online], vol. 5, no. 1, Available from: <http://www.soc.surrey.ac.uk/JASSS/5/1/3.html>.

Grenfell, B.T. & Harwood, J. 1997, "(Meta)population dynamics of infectious diseases", *Trends in Ecology & Evolution*, vol. 12, no. 10, pp. 395-9.

Gupta, S., Anderson, R.M. & May, R.M. 1989, "Networks of sexual contacts: implications for the pattern of spread of HIV", *AIDS*, vol. 3, no. 12, pp. 807-17.

Hakimi, S.L. 1962, "On realizability of a set of integers as degrees of the vertices of a linear graph", *Journal of the Society for Industrial and Applied Mathematics*, vol. 10, no. 3, pp. 496-506.

Hamer, W.H. 1906, "Epidemic disease in England: the evidence of variability and persistency of type", *The Lancet*, pp. 733-9.

Havel, V. 1955, "A remark on the existence of finite graphs", *Casopis pro Pestováni Matematiky [Czech]*, vol. 80, pp. 477-80.

Hethcote, H.W. & van Ark, J.W. 1987, "Epidemiological models for heterogeneous populations: Proportionate mixing, parameter estimation, and immunization programs", *Mathematical Biosciences*, vol. 84, pp. 85-118.

Holland, P.W. & Leinhardt, S. 1970, "A method for detecting structure in sociometric data", *American Journal of Sociology*, vol. 76, no. 3, pp. 492-513.

Holland, P.W. & Leinhardt, S. 1981, "An exponential family of probability distributions for directed graphs", *Journal of the American Statistical Association*, vol. 76, no. 1, pp. 33-50.

Holme, P. & Kim, B.J. 2002, "Growing scale-free networks with tunable clustering", *Physical Review E*, vol. 65, no. 026107.

Holme, P. & Zhao, J. (2006), Exploring the assortativity-clustering space of a network"s degree sequence, Preprint cond-mat/0611020, Available from: <http://arxiv.org/PS_cache/q-bio/pdf/0611/0611020.pdf>

Hong, L.H., Pattison, P. & Robins, G. 2005, "A spatial model for social networks", *Physica A*, vol. 360, pp. 99-120.

Huang, C.-Y. , Sun, C.-T. , Chen, Y.A. & Lin, H. 2005, "A novel small-world model: Using social mirror identities for epidemic simulations", *Simulation*, vol. 81, no. 10, pp. 671-99.

Huang, C.-Y. , Sun, C.-T. , Hsieh, J.-L. & Lin, H. 2004, "Simulating SARS: Small-world epidemiological modeling and public health policy assessments", *Journal of Artificial Societies and Social Simulation*, [Online], vol. 7, no. 4, Available from: <http://jasss.soc.surrey.ac.uk/7/4/2.html>.

Hunter, D.R., Goodreau, S.M. & Handcock, M.S. 2005, *Goodness of fit of social network models*, PennState Department of Statistics Technical Reports 05-02.

Keeling, M.J. 1999, "The effects of local spatial structure on epidemiological invasions", *Proceedings of the Royal Society London B*, vol. 266, pp. 859-69.

Keeling, M.J. 2005, "The implications of network structure for epidemic dynamics", *Theoretical Population Biology*, vol. 67, no. 1, pp. 1-8.

Keeling, M.J. & Eames, K.T.D. 2005, "Networks and epidemic models", *Journal of the Royal Society Interface*, [Online], vol. 2, no. 4, pp. 295-307, Available from: <http://www.journals.royalsoc.ac.uk/openurl.asp?genre=article&id= doi:10.1098/rsif.2005.0051>.

Kendall, D.G. 1956, "Deterministic and stochastic epidemics in closed populations", ed J. Neyman, *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press.

Kermack, W.O. & McKendrick, A.G. 1927, "Contributions to the mathematical theory of epidemics - I", *Proceedings of the Royal Society*, vol. 115A, pp. 711-21.

Kermack, W.O. & McKendrick, A.G. 1932, "Contributions to the mathematical theory of epidemics - II: The problem of endemicity", *Proceedings of the Royal Society*, vol. 138A, pp. 55-83.

Kermack, W.O. & McKendrick, A.G. 1933, "Contributions to the mathematical theory of epidemics - III: Further studies of the problem of endemicity", *Proceedings of the Royal Society*, vol. 141A, pp. 94-122.

King, O.D. 2004, "Comment on "Subgraphs in random networks"", *Physical Review E*, vol. 70, no. 058101.

Lajmanovich, A. & Yorke, J.A. 1976, "A deterministic model for gonorrhea in a nonhomogeneous population", *Mathematical Biosciences*, vol. 28, pp. 221-36.

Last, J.M. (eds) 2001, *A Dictionary of Epidemiology*, 4th edn, Oxford University Press.

Lefevre, C. & Malice, M.-P. 1988, "Comparisons for carrier-borne epidemics in heterogeneous and homogenous populations", *Journal of Applied Probability*, vol. 25, no. 4, pp. 663-78.

Lloyd, A.L. & May, R.M. 1996, "Spatial heterogeneity in epidemic models", *Journal of Theoretical Biology*, vol. 179, no. 1, pp. 1-11.

Marney, J.P. & Tarbert, H. 2000, "Why do simulation? Towards a working epistomology for practitioners of the dark arts", *Journal of Artificial Societies and Social Simulation*, [Online], vol. 3, no. 4, Available from: <http://www.soc.surrey.ac.uk/JASSS/3/4/4.html>.

Martin, E.A. (eds) 1994, *Concise Medical Dictionary: New Edition*, 4th edn, Oxford University Press, Oxford.

Maslov, S., Sneppen, K. & Zaliznyak, A. (2002), Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet, Preprint cond-mat/0205379 v2

May, R.M. & Anderson, R.M. 1984a, "Spatial heterogeneity and the design of immunization programs", *Mathematical Biosciences*, vol. 72, pp. 83-111.

May, R.M. & Anderson, R.M. 1984b, "Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes", *IMA Journal of Mathematical Applications in Medicine and Biology*, vol. 1, pp. 233-66.

McNeill, W.H. 1976, *Plagues and Peoples*, Anchor Press, New York.

Milo, R., Kashtan, N., Itkkovitz, S., Newman, M.E.J. & Alon, U. (2004), On the uniform generation of random graphs with prescribed degree sequences, Preprint cond-mat/0312028, Available from: <http://aps.arxiv.org/PS_cache/cond-mat/pdf/0312/0312028.pdf>

Milo, R., Shen-Orr, S., Itkkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. 2002, "Network motifs: Simple building blocks of complex networks", *Science*, vol. 298, pp. 824-7.

Molloy, M. & Reed, B. 1995, "A critical point for random graphs with a given degree sequence", *Random Structures and Algorithms*, vol. 6, pp. 161-79.

Moreno, Y., Gómez, J.B. & Pacheco, A.F. 2003, "Epidemic incidence in correlated complex networks", *Physical Review E*, vol. 68, no. 035103.

Morris, M. 1995, "Data driven network models for the spread of infectious disease" in *Epidemic Models: Their Structure and Relation to Data*, ed D. Mollison, Cambridge University Press, Cambridge, pp. 302-22.

Newman, M.E.J. 2002a, "Assortative mixing in networks", *Physical Review Letters*, vol. 89, no. 208701.

Newman, M.E.J. 2002b, "Spread of epidemic disease on networks", *Physical Review E*, vol. 66, no. 016128.

Newman, M.E.J. 2003a, "Mixing patterns in networks", *Physical Review E*, vol. 67, no. 026126.

Newman, M.E.J. 2003b, "Properties of highly clustered networks", *Physical Review E*, vol. 68, no. 026121.

Newman, M.E.J. 2003c, "The structure and function of complex networks", *SIAM Review*, [Online], vol. 45, no. 2, pp. 167-256, Available from: <http://arxiv.org/abs/cond-mat/0303516>.

Newman, M.E.J. & Park, J. 2003, "Why social networks are different from other types of networks", *Physical Review E*, vol. 68, no. 036122.

Newman, M.E.J. & Park, J. 2007, personal communication.

Nold, A. 1980, "Heterogeneity in disease transmission modelling", *Mathematical Biosciences*, vol. 52, pp. 227-40.

Park, J. & Newman, M.E.J. 2003, "Origin of degree correlations in the Internet and other networks", *Physical Review E*, vol. 68, no. 026112.

Pastor-Satorras, R. & Vespignani, A. 2001, "Epidemic spreading in scale-free networks", *Physical Review Letters*, vol. 86, no. 14, pp. 3200-3.

Polhill, J.G., Izquierdo, L.R. & Gotts, N.M. 2005, "The ghost in the model (and other effects of floating point arithmetic)", *Journal of Artificial Societies and Social Simulation*, [Online], vol. 8, no. 1, Available from: <http://jasss.soc.surrey.ac.uk/8/1/5.html>.

Rao, A.R., Jana, R. & Bandyopadhyay, S. 1996, "A Markov chain Monte Carlo method for generating random (0,1) matrices with given marginals", *Indian Journal of Statistics*, vol. 58, pp. 225-42.

Rapoport, A. & Horvath, W.J. 1961, "A study of a large sociogram", *Behavioral Science*, vol. 6, no. 4, pp. 279-91.

Rhodes, C.J. & Anderson, R.M. 1996, "Persistence and dynamics in lattice models of epidemic spread", *Journal of Theoretical Biology*, vol. 180, pp. 125-33.

Roberts Jr, J.M. 2000, "Simple methods for simulating sociomatrice with given marginal totals", *Social Networks*, vol. 22, no. 3, pp. 273-83.

Robins, G., Pattison, P., Kalish, Y. & Lusher, D. 2005, *A workshop on exponential random graph (p*) models for social networks*, Social Networks working paper No. 1/05, Psychology Department, University of Melbourne [Online], Available from: <http://www.psych.unimelb.edu.au/staff/gr/ergm.pdf>.

Robins, G., Snijders, T.A.B., Wang, P., Handcock, M.S. & Pattison, P. 2007, "Recent developments in exponential random graph (p*) models for social networks", *Social Networks*, vol. 29, no. 2, pp. 192-215.

Rothenberg, R. B. 2003, 'Large network concepts and small network characteristics', in *Networks and the Population Dynamics of Disease Transmission*, Institute for Mathematics and Its Applications, University of Minnesota (unpublished).

Servedio, V.D.P. & Caldarelli, G. 2004, "Vertex intrinsic fitness: How to produce arbitrary scale-free networks", *Physical Review E*, vol. 70, no. 056126.

Sherman, I.W. 2006, *The Power of Plagues*, ASM Press.

Snijders, T.A.B. 1991, "Enumeration and simulation methods for 0-1 matrices with given marginals", *Psychometrika*, vol. 56, pp. 397-417.

Snijders, T.A.B., Pattison, P., Robins, G. & Handcock, M.S. 2006, "New specifications for exponential random graph models", *Sociological Methodology*, vol. 36, pp. 99-153.

Snow, J. 1855, *On the Mode of Communication of Cholera*, 2nd edn, John Churchill, London.

Stauffer, A.O. & Barbosa, V.C. (2005), A study of the edge switching Markov-Chain method for the generation of random graphs, Preprint cond-mat/0512015, Available from: <http://arxiv.org/abs/cs/0512105>

Tabachnick, B.G. & Fidell, L.S. 2006, *Using Multivariate Statistics*, 5th edn, Allyn & Bacon, Boston.

Thucydides 431 BCE, *History of the Pelopponnesian War*, [Online], Available from: <http://classics.mit.edu/Thucydides/pelopwar.html>.

Viger, F. & Latapy, M. (2005), Fast generation of random graphs with prescribed degrees, Preprint cond-mat/0502085, Available from: <http://arxiv.org/PS_cache/cs/pdf/0502/0502085.pdf>

Wasserman, S. & Faust, K. 1994, *Social Network Analysis: Methods and Applications*, Structural Analysis in the Social Sciences, Cambridge University Press, Cambridge.

Wasserman, S. & Pattison, P. 1996, "Logit models and logistic regressions for social networks: I. An introduction to Markov random graphs and p*", *Psychometrika*, vol. 61, no. 3, pp. 401-26.

Watts, D.J. 2004, "The "new" science of networks", *Annual Review of Sociology*, [Online], vol. 30, pp. 243-70, Available from: <www.annualreviews.org (doi:10.1146/annurev.soc.30.020404.104342)>.

Watts, D.J. & Strogatz, S.H. 1998, "Collective dynamics of "small-world" networks", *Nature*, vol. 393, pp. 440-2.

Waxman, B.M. 1988, "Routing of multipoint connections", *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617-22.

Wormald, N.C. 1981, "The asymptotic connectivity of labelled regular graphs", *Journal of Combinatorial Theory, Series B*, vol. 31, pp. 156-67.

Appendix A: Glossary

Network analysis is a multidisciplinary field, with researchers in the overlapping fields using different terminology for related concepts. This glossary summarises key terms and identifies synonym terms across fields.

Actor	See node.
Adjacent	Two nodes are adjacent if an edge connects them. See neighbour.
Arc	A connection between two nodes that indicates that they are related. Usually used to indicate a directed relationship.
Assortativity	Pearson correlation coefficient of the degrees at the ends of each edge.
Basic reproduction ratio	The expected number of secondary cases produced, in a completely susceptible population, by a typical infected individual during its entire period of infectiousness.
Clustering coefficient	Mean local clustering coefficient over all nodes.
Complete	A network is complete if each pair of nodes has an edge between them. Thus, there are n(n-1)/2 edges in an undirected complete network with n nodes.
Component	If a network is not connected, each connected subnetwork is referred to as a component.
Connected	A network is connected if and only if any node can be reached from any other node by travelling along edges. Also see component.
Connectivity	Alternative term for degree
Contact	Epidemiological term where the relationship between two (or more) individuals allows a nonzero probability of disease transmission. From the perspective of network theory, a contact (for a node) can be interpreted as either the edges that connect to the node, or the nodes that have edges that connect to the specific node.

Degree	A node characteristic, the number of edges in which the node participates.
Degree correlation	Alternative term for (degree) assortativity.
Degree distribution	The frequency distribution for the degree of a node.
Degree sequence	A specific instance of a degree distribution. That is, a listing of the nodes with the degree for that node.
Diameter	Mean geodesic across all pairs of nodes in the network.
Directed	Indicates that a relationship from A to B is different from a relationship from B to A. For example, the relationship 'trusts' is undirected because A trusts B is not the same as B trusts A.
Edge	A connection between two nodes that indicates that they are related. Usually used to indicate a mutual or undirected relationship. Also see arc.
Epidemic	The occurrence in a community or region of cases of an illness … clearly in excess of normal expectancy.
Geodesic	Number of edges in the shortest path between a pair of nodes.
Graph	Comprises the nodes and the edges that join pairs of nodes
Host	Person who is infected, also see node.
Isomorphic	Two networks are isomorphic if one could be produced from the other by only changing any labels on the nodes. That is, the node and edge configuration is identical.
Local clustering coefficient	Characteristic of a node, the proportion of potential edges between neighbour nodes that are actual edges in the network.
Motif	Small subgraph that potentially appears many times within a larger graph.
Mutual	Alternative term for undirected.
Multigraph	A graph where pairs of nodes may have more than a single edge between them. Compare with simple graph.
Neighbour	If two nodes have an edge between them, each node is referred to as a neighbour of the other node. See adjacent.

Neighbour algorithm	Algorithm developed and described in thesis to generate networks with specific values of assortativity and clustering coefficient from a degree sequence.
Neighbour network	Network generated with neighbour algorithm
Network	Set of nodes and edges that describe a specific relationship.
Node	An individual who may be involved in the relevant relationship with another node. Also actor, vertex and, in the specific case of disease transmission, host.
Original network	Network generated with an algorithm from the literature
Reproduction ratio	(Also reproduction rate)
Proportionate mixing	For all people, the infectivity and susceptibility are proportional. Note that early research uses the term proportionate mixing for the mixing relationship later referred to as separable mixing.
Rewiring	Modification to a network through selection of two existing edges, removing those edges and constructing new edges for two different pairs of the four nodes involved.
Self edge	A self edge connects a node to itself.
Separable mixing	For any pair of infective and susceptible people involved in a potential transmission, the relevant characteristics of the two people are independent.
Simple graph	A graph where each pair of nodes has 0 or 1 edges between them (no multiple) and no edge joins a node to itself. Compare with multigraph.
SIR	An epidemic where the only states are susceptible, infected or removed. The disease confers either permanent immunity or death.
SIS	An epidemic where the only states are susceptible or infected. Once a person recovers from the disease, they are immediately susceptible to a new infection.

Undirected	Indicates that the network, relationship or edge is mutual. That is, A is related to B if and only if B is related to A. For example, the relationship 'is a sibling of' is undirected. Formally, the relation is reflexive.
Unweighted	In certain types of networks, nodes or edges may be weighted to reflect specific characteristics (such as a stronger relationship). In an unweighted network, neither nodes nor edges are weighted.
Vertex	Alternative term for node.

Appendix B: Files Used in Analysis

This appendix describes the programs and settings used to generate networks, simulate epidemics, and perform the major analyses for the study. It also identifies the output files. All the files are included on the supplementary results DVD.

In general, the CODE folder contains the C++ code to generate networks and simulate epidemics and SPSS and Matlab scripts used for analysis. The DATASETS folder contains the raw datasets generated by the simulation, the SPSS scripts to summarise the data into various views of the data and the SPSS format summary datasets. The outputs of the various analyses are at the top level of the DVD (with html format versions of the SPSS output files in the WEBOUTPUT folder).

B.1 C++ libraries

The C++ programs described in this appendix rely on several code libraries, included in the CODE folder on the DVD. In addition, the standard template library (STL) is used extensively.

The *Network* library is used to generate networks with different algorithms and measure network properties. To generate neighbour networks, the *MakeNeighbourNetwork1D* method is used. There are two related libraries: *DegDist* enables property calculation from the extracted degree distribution, and *NetBasic* is a subset of Network, to allow Network to call *DegDist* and *DegDist* to call *NetBasic* without circular referencing.

The *Epidemic* library runs the epidemic simulation over a specified network with method *Simulate*. It also contains methods to summarise simulation results files in different ways.

There are also three utility libraries. The *Array2D* class is an extension of the STL vector class to enable two dimensional index notation. The *Random* class implements a random number generator. Other utility functions, such as printing of STL data types, are included in the *Utilities* library.

B.2 Neighbour algorithm validation

Section 4.3.1 describes the evaluation of the neighbour algorithm with respect to its ability to target network properties separately. The relevant files are:

- C++ program: *Miscellanous.cpp* (switch 6), in the CODE folder
- Output files: *Convergence.xls*, in the RESULTS folder

Section 4.3.2 describes the way in which algorithm inputs are related to the generated neighbour networks. The relevant files are:

- C++ program: *NetParameters.cpp*, in the CODE folder
- Output files: *Networks Parameters.xls*, in the RESULTS folder

Conformity with the small-world property is assessed in section 4.3.3. The relevant files are:

- C++ program: *Miscellanous.cpp* (switch 4), in the CODE folder
- Output files: *Small World.xls*, in the RESULTS folder

B.3 Generate simulation data

The experimental design is described in Chapter 3. Up to 10 neighbour networks are generated for each of three degree distribution types, with each of various values of assortativity and clustering coefficient. For each network, up to 10 epidemics are simulated with a specific infectivity rate (3 values) and immunity setting (SIR or SIS).

Thus, there are potentially 100 SIS and 100 SIR epidemics for each network and epidemic property combination (10 networks by 10 epidemics). For each simulation, the number of nodes in each disease state (such as infected or susceptible) is recorded for 100 timesteps.

The relevant file to generate the networks and simulate the epidemics is the C++ program *EpiRunSample.cpp*. This program is run separately for each degree distribution type (distType: 0 for uniform, 1 for real world, 2 for power law and 3 for normal), with manual changing of the distType parameter and output file names.

```
// *****************************************************
// Parameters for experiment are set here
// *****************************************************

// experiment design
int     runDist =10;            // distribution instances
int     runEpi = 10;            // number of epidemic instances
int     maxAttempts = 10;       // number of attempts before fail for property pair
int     seed = 123;             // seed for randomisation process (-1 uses clock)

// network parameters
int     distType = 1;           // 0 uniform; 1 real world; 2 power law; 3 normal
int     netType;                // 0 is generating algorithm, 1 is neighbour
int     nodes = 1000;           // number of nodes
int     degree = 8;             // mean degree

seed += distType * 100;         // so seed doesn't need to be changed for each distribution run

// epidemic parameters
vector<double>  inf;            // vector for infectivity rate for all nodes
inf.push_back(0.5 * 1/24);
inf.push_back(1.0 * 1/24);
inf.push_back(1.5 * 1/24);
inf.push_back(2.0 * 1/24);
double  sus = 1;                // susceptibility rate for all nodes
double  rec = 0.3333;           // recovery rate for all nodes
double  imm = 0.5;              // proportion recover to immune (0 SIS, 1 SIR)
int     infStart = 1;           // nodes infected at timestep 0
bool    weighted = false;       // initial nodes selected weighted by degree or uniformly
int     timeSteps = 100;        // number of timesteps for epidemic run
double  threshold = 0.1;        // proportion nodes infected to be epidemic

// filenames with quotes, eg "c:\\outdata.csv" (remove path if using supercomputer)
string  dataFilename = "outdata.csv";           // detailed epidemic data
string  netsFilename = "outnetsused.csv";       // networks used
string  allnetsFilename = "outnetsmade.csv";    // info on all networks generated
```

Figure B-1: Parameter setup for main simulation program

The output files are included on the DVD in the RESULTS folder. They contain headings and the following information:

- *DistType Networks.csv*: Each record has network identification fields and network property values (see Table B-1).

Table B-1: Fields in Networks datasets

Field	Comment
Distribution type	Uniform, Real world, Power law (BA) or Normal (ER)
Distribution run	Counter for instance of distribution type
Network type	Algorithm used – original or neighbour
Network run / ID	Counter for instance of network
Assortativity target	Assortativity parameter for neighbour algorithm
Clustering probability	Edge parameter for neighbour algorithm
Distribution seed	Random number generator seed for distribution generation
Network seed	Random number generator seed for network generation
Max assortativity	Maximum possible assortativity given degree distribution
Valid	Status code for whether network generation resolved
Nodes	Number of nodes in generated network
Edges	Number of edges in generated network. This is important for determining whether to accept networks that are invalid.
Degree	Mean degree of generated network
CV Degree	Coefficient of variation for degree of nodes
Assortativity	Actual assortativity of generated network
Clustering	Mean clustering coefficient of nodes in generated network
Giant	Proportion of network in giant component
Components	Number of components in network
Entropy	Measure of skewness of degree distribution
Gini index	Measure of skewness of degree distribution
HHI	Herfindahl-Hirschman Index of degree concentration (power 2)
Nodes 5%	Proportion of edges accounted for by the highest degree 5% of nodes

- *DistType Data.csv*: Each record has network identification fields, epidemic parameters settings, and epidemic status information for a specific timestep (see Table B-2).
- *DistType Made.csv*: A table displaying the total number of networks generated by property combinations in order to obtain the required sample size.

For each degree distribution type, the detailed epidemic data is summarised with *Miscellaneous.cpp* (code switch 3). Created summary datasets are:

- *DistType Epidemic Stability.csv*: Identifies if, and at what timestep, epidemic stability is reached. A record is created for each simulation that contains network / epidemic identification information and epidemic status information averaged after stability.

The four datasets for each degree distribution type are combined with the SPSS syntax program *CreateDatasets.sps*. This program also adds derived variables including the empirical reproduction rate, whether an epidemic occurred (see section 5.1) and basic reproduction ratio derived from epidemic size, and creates three detailed and three summary SPSS datasets with various perspectives of the simulation data.

Note that the summary datasets were recreated several times to enable some of the included variables (such as whether an epidemic occurred) to be included after the relevant analysis was performed. That is, derivations of the new variables were added to the syntax file and the script rerun.

Table B-2: Fields in Data datasets

Field	Comment
Distribution type	Key to match with network
Distribution run	Key to match with network (redundant)
Network type	Key to match with network (redundant)
Network run	Key to match with network
Assortativity target	Key to match with network (redundant)
Clustering probability	Key to match with network (redundant)
Infectivity	Infectivity probability – applied to all nodes
Susceptibility	Susceptibility probability – applied to all nodes
Recovery	Recovery probability – applied to all nodes
Immunity	Probability of immunity when recovered
Start infected	Number of nodes initially selected as infected
Selection type	Initial infection uniform or proportional to degree
Epidemic run	Counter for instance of epidemic
Epidemic seed	Seed for random number generator for epidemic run
Timestep	Simulation step counter
New infections	Incidence: number of nodes that became infected in timestep
Cumulative infections	Impact: Number of nodes infected up to this timestep
Infected	Prevalence: number of nodes in infectious state
Susceptible	Number of nodes in susceptible state
Immune	Number of nodes in immune state
New infections degree	Mean degree of new infection nodes
Cum infections degree	Mean degree of all nodes that have been infected
Infected degree	Mean degree of infected nodes
Susceptible degree	Mean degree of susceptible nodes
Immune degree	Mean degree of immune nodes

The summary datasets are:

- *EpidemicSummary.sav*: Simulations (10 epidemic runs on 10 network instances) were aggregated by distribution type, network type, network properties (approximate clustering and assortativity) into two records, one for epidemics and one for simulations where an epidemic did not occur. Variables included the mean and standard deviation of the number of nodes in each epidemic state for each timestep.
- *EpidemicTimelines.sav*: The same information as *EpidemicSummary.sav* but there is a separate record for each timestep, rather than information for multiple timesteps in the same record (or case).
- *EpidemicNetworkMinimal.sav*: This dataset has a record for each simulation but contains only summary information instead of data for each timestep.

All datasets and code to create the SPSS datasets are included on the DVD in the DATASETS folder.

In addition to matching onto the epidemic datasets, the *Network.csv* files are used to create the Microsoft Excel workbook *Networks.xls*. This workbook describes the number and properties of the networks used for epidemic simulations.

B.4 Analysis of relationship between network properties and epidemic behaviour

Note that much of the analysis was performed with SPSS software. The syntax files (suffix sps) can be accessed with a text editor. However, the output files (suffix spo) are in a SPSS proprietary format that is only accessible with SPSS software. These output files have also been produced in a web browser accessible format.

A hyperlinked index to the html versions of the SPSS output files is at the top level of the DVD (*WebOutputIndex.htm*). The html version is complete but does not have any internal hyperlinks to find individual sections. As the comments and code from the syntax file is reproduced in the output file, specific analyses can be found by searching for the relevant text. Furthermore, the utility to produce the html file generates a separate file for each chart and many of the analyses contain large numbers of charts. Hence, the web format versions (same filename, suffix htm) are in a separate folder (WebOutput) with subfolders for those with large numbers of files.

Two SPSS syntax files are used to summarise the number of simulations and epidemics in the dataset (Section 5.1.1):

- For the neighbour networks with varying network properties: *SimulationCounts.sps* (CODE folder), with output at *SimulationCounts.spo*; and
- For the networks implementing the basic epidemiological model: *ZeroStructure.sps* (CODE folder), with output at *ZeroStructure.spo*.

For the definition of an epidemic (Sections 5.1.2 and 5.1.3), the syntax file to perform the analysis is *EpidemicDefn.sps* (CODE folder), with output at *EpidemicDefn.spo*. After this analysis was performed, the summary dataset creation syntax file was rerun with some additional code to implement the definition.

The analysis of impact of degree heterogeneity (Section 5.2) on epidemic occurrence and epidemic size used *DegreeImpact.sps* (CODE folder), with output at *DegreeImpact.spo*. Some of the results were transferred to the spreadsheet *DegreeImpact.xls* for additional analysis concerning the number of property combinations with different patterns of relativity between distribution types.

The analysis of the relationship between epidemic behaviour and the network properties of assortativity and clustering was performed with a series of SPSS syntax and related files:

- relationship with epidemic occurrence (Section 5.3) summary tables and logistic regressions used *PropertyImpactSeverity.sps* (CODE folder), with output at *PropertyImpactSeverity.spo*;
- relationship with basic reproduction ratio (Sections 5.4 and 5.5) exploratory analysis used *PropertyImpactBehaviour.sps* (CODE folder), with output at *PropertyImpactBehaviour.spo*;
- *PropertyImpactBehaviour.sps* also creates the text file *SizeAssClusData.csv*, which is used by the Matlab script *R0AssClusScatter.m* (CODE folder) to generate scatter plots (Section 5.4.1.4) of the relationship between epidemic derived R_0 and the structure properties, with the output contained in *SIR EpiR0 Properties.pdf* and *SIS EpiR0 Properties.pdf*; and
- regressions are fitted (Sections 5.4.2 and 5.5.1) using *PropertyImpactRegression.sps* (CODE folder), with output at *PropertyImpactRegression.spo*; and
- some additional analyses to compare predictions from the SIR and SIS regression models (Section 5.5.2) use *ConsistencySIRvSIS.sps* (CODE folder), with output at *ConsistencySIRvSIS.spo* and the Excel file *Consistency Predictions.xls*.

A separate analysis used regression models to examine the performance of an alternative way to incorporate network structure in epidemic models, accessible network proportion in a given number of steps (Section 5.6). The analysis used *ProportionAccessible.sps*, with output at *ProportionAccessible.spo*.

Appendix B: Files Used in Analysis

Appendix C: Index to Additional Results

Several of the tables and figures in Chapter 5 presented results for a single simulation set (for example, real world degree distribution with infectivity of 0.1250). This appendix identifies the file in the supplementary DVD where the results are available for all simulation sets for these tables (Table C-1) and figures (Table C-2).

The results files are in the proprietary SPSS output format, but have also been produced in a web browser accessible format. An index to the html versions of the SPSS output files is at the top level of the DVD (*WebOutputIndex.htm*).

Table C-1: Filename for supplementary results, Tables

Table number and caption	Results file
Table 5-7: Network and epidemic properties by degree distribution	*DegreeImpact.spo*
Table 5-12: Proportion of simulations satisfying epidemic definition: SIR	*PropertyImpactSeverity.spo*
Table 5-13: Proportion of simulations satisfying epidemic definition: SIS	*PropertyImpactSeverity.spo*
Table 5-14: Number of contributing simulations	*PropertyImpactSeverity.spo*
Table 5-18: Number of contributing simulations: SIR	*PropertyImpactSeverity.spo*
Table 5-20: Mean epidemic final size: SIR	*PropertyImpactBehaviour.spo*
Table 5-21: Standard error of mean epidemic final size: SIR	*PropertyImpactBehaviour.spo*
Table 5-35: t-test to compare SIR and SIS epidemic derived R_0 values	*ConsistencySIRvSIS.spo*
Table 5-39: Adjusted R^2 for network structure models of R_0	*ProportionAccessible.spo*

Table C-2: Filename for supplementary results, Figures

Figure number and caption	Results file
Figure 5-7: Epidemic size over time (cumulative infections) by clustering coefficient	*PropertyImpactBehaviour.spo*
Figure 5-8: Epidemic size over time (cumulative infections) by assortativity	*PropertyImpactBehaviour.spo*
Figure 5-9: Assortativity, clustering coefficient and R_0 derived from epidemic final size	*PropertyImpactBehaviour.spo*
Figure 5-10: Histogram of standardised regression residuals	*SIR R0 Properties.pdf* *SIS R0 Properties.pdf*
Figure 5-11: Residual plotted against regression prediction for R_0	*PropertyImpactRegression.spo*
Figure 5-12: Residual plotted against assortativity	*PropertyImpactRegression.spo*
Figure 5-13: Residual plotted against clustering coefficient	*PropertyImpactRegression.spo*
Figure 5-14: Epidemic size over time (current infections) by assortativity (SIS)	*PropertyImpactBehaviour.spo*
Figure 5-15: Residual plotted against regression prediction for R_0	*PropertyImpactRegression.spo*
Figure 5-16: Residual plotted against assortativity	*PropertyImpactRegression.spo*
Figure 5-17: Residual plotted against clustering coefficient	*PropertyImpactRegression.spo*
Figure 5-18: Histogram of standardised regression residuals	*PropertyImpactRegression.spo*
Figure 5-19: SIR and SIS epidemic derived R0 values, frequencies	*ConsistencySIRvSIS.spo*
Figure 5-20: SIR and SIS epidemic derived R0 values, frequencies	*ConsistencySIRvSIS.spo*