Impact of network clustering and assortativity on epidemic behaviour

Jennifer Badham[a,*], Rob Stocker[a]

[a]*Artificial Life and Adaptive Robotics Laboratory, School of ITEE, Australian Defence Force Academy, Northcott Drive, Canberra 2600*

Abstract

Epidemic models have successfully included many aspects of the complex contact structure apparent in real world populations. However, it is difficult to accommodate variation in number of contacts, clustering coefficient and assortativity. Investigations of the relationship between these properties and epidemic behaviour have led to inconsistent conclusions and have not accounted for their interrelationship. In this study, simulation is used to estimate the impact of social network structure on the probability of an SIR epidemic occurring and, if it does, the final size. Increases in assortativity and clustering coefficient are associated with smaller epidemics and the impact is cumulative. Derived values of R_0 over networks with the highest property values are more than 20% lower than those derived from simulations with zero values of these network properties.

Keywords: Disease spread, Transmission networks, Clustering coefficient, Assortativity, Social networks
PACS: 89.75.Hc, 87.23.Ge, 87.19.Xx

* Corresponding author.
Email addresses: research@criticalconnections.com.au (Jennifer Badham), r.stocker@adfa.edu.au (Rob Stocker)

1. Introduction

Epidemiologists have developed mathematical models to estimate the size and other key features of an epidemic. The simplest of these models assume homogeneity in social structure (Kermack and McKendrick, 1927). Among other simplifications, all people are assumed to have the same number of contacts per unit time and these contacts are between people selected uniformly at random.

The final size of the epidemic (denoted f) is the proportion of the population ever infected. The basic reproduction ratio (denoted R_0) is the number of infections generated directly from an average infectious individual in an otherwise susceptible population, and is related to the number of contacts and transmission probability. Under the assumptions of an equal number of randomly selected contacts, these epidemic properties are related by (Brauer, 2008, equation (2.3))

$$\log_e(1 - f) = -fR_0. \tag{1}$$

These simplifications are potentially more problematic for sexually transmitted diseases or where direct contact is required for transmission, and models for these diseases introduced more realistic assumptions concerning social structure (Gupta et al., 1989). A recent review article (Ferguson et al., 2003) identified several ways in which social structure has been introduced to smallpox models. A common approach is to define subpopulations based on age or socio-economic factors or, in patch models, geographic factors. Structure is introduced by varying contact probabilities for the pairs of subpopulations. Within a subpopulation, individuals are treated identically and the contacts are selected uniformly at random. The highest resolution models simulate individuals in the population. For example, the EpiSims project (Eubank et al., 2004) synthesises a population and their activities to generate a detailed set of contacts.

An intermediate approach uses an algorithm to generate a static network with specific characteristics. The network structure establishes a fixed set of contacts for the population and this set of contacts determines the way in which the disease is transmitted. There are several reviews of network models that focus on the implications for epidemic behaviour (Newman, 2002b, 2003; Watts, 2004; Keeling and Eames, 2005). However, these models have not examined the specific combination of social network properties described in this study.

2. Social network properties

In a social network, nodes (or vertices or actors) represent individual people and edges (or links) represent the relationship of interest. For epidemic behaviour, that relationship is sufficient contact to transmit the disease in some undefined period of time. The networks used in this study are static, simple, undirected and unweighted. This means: the edges do not change over time; there is only one edge between connected nodes; the disease can be transmitted in either direction; and the probability of transmission is identical for all pairs of nodes connected by an edge. Apart from the number of edges (contacts), nodes are homogenous.

The degree of a node is the number of edges that it shares with other nodes. The degree distribution is then the frequency distribution of different degrees across the nodes in the network. For clarity, the term degree distribution is used for both the density and the cumulative distribution by different authors, and is being used in the former sense here. Real-world network degree distributions are typically positively skewed (Newman and Park, 2003), with some nodes having very high degree nodes. See Figure 1 for an example distribution.

Clustering is a measure of network transitivity, the extent to which neighbours of a node are neighbours of each other. It formalises the concept that two people who have a friend in common are more likely to be friends themselves than two people selected at random. Clustering is defined as (Watts and Strogatz, 1998):

> Suppose that a vertex v has k_v neighbours [that is, nodes with which it shares an edge]; then at most $k_v(k_v - 1)/2$ edges can exist between them . . . . 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.

By convention, a node with degree of 0 or 1 is assigned clustering of 0. Each node contributes equally to the mean, regardless of its degree. The theoretical range is the interval [0,1].

Degree assortativity of a network (denoted r) is defined as (Newman, 2002a): *simply the Pearson correlation coefficient of the degrees at either ends of an edge*. The theoretical range is the interval [-1,1].

Social networks have higher assortativity and clustering coefficient values than would be expected in a network with randomly created edges constrained by the degree distribution (Newman and Park, 2003, and references therein). For example, a collaboration network of mathematicians has C of 0.15 instead of the expected 0.00015, and a similar network for physicists has r of 0.154. A further example of shared Board membership for company directors has C of 0.59 (expected 0.0035) and r of 0.276.

Analytical solutions and simulation over networks have been used to include these social factors in epidemic analysis to some extent. In particular, variation in infectivity and susceptibility arising from variation in number of contacts (referred to as degree heterogeneity in network theory), has been well studied (Becker, 1973; Lajmanovich and Yorke, 1976; Nold, 1980). Variation in number of contacts increases R_0 and the probability of an epidemic, but R_0 may appear to be reduced because equation (1) is invalid (Ball and Clancy, 1993). For positively skewed degree distributions, final size of the epidemic is lower than would arise with the same probability of infection but uniform degree at the mean, because low degree nodes have low susceptibility and comprise a high proportion of the nodes. With knowledge of the degree distribution, more complex aspects of epidemic behaviour can be described, such as the probability of outbreaks of different sizes or the rate of incidence early in the outbreak (Yan, 2008).

Assortativity has been introduced into the social structure underlying epidemic models in two ways, and analysed both theoretically and by simulation

for each. The first method divides the intended contacts into two groups: one group is used to connect people with the same number of contacts, and the other is allocated randomly (referred to as preferential and proportional mixing respectively) (Nold, 1980; Moreno et al., 2003). Assortativity is higher where the proportion in the former group is larger. The second method uses the (predefined) joint degree distribution to constrain the way in which connections are made (Newman, 2002a; Boguñá and Pastor-Satorras, 2002). These studies have consistently found that epidemics are smaller in the presence of positive assortativity. However, the results for the probability of an epidemic occurring are inconsistent. Most (Newman, 2002a; Nold, 1980; Boguñá and Pastor-Satorras, 2002) have found that increased assortativity increases the probability of an epidemic (or equivalently, lowers the epidemic threshold), but one (Moreno et al., 2003) finding a reduced probability of an epidemic with positive assortativity.

The impact of clustering on epidemic behaviour has had only limited study and simulations have only been conducted over networks with socially unrealistic degree distributions (Keeling, 1999, 2005). These studies have found that epidemics are less likely to occur with higher levels of clustering and, if they do occur, are smaller.

The authors are unaware of any published studies concerning epidemic behaviour where clustering and assortativity interact. Further, each of the studies concerning the separate impact of clustering or assortativity do not measure the other property and, as there is some evidence that these properties are related (Soffer and Vázquez, 2005; Holme and Zhao, 2007; Smith et al., 2008; Badham et al., 2008), the conclusions may have attributed the identified behaviour to the incorrect network property.

3. Experimental design

Use of a static network to describe the transmission opportunities inherently establishes a set of assumptions. In particular, the set of people available to be infected by each susceptible person is fixed and, further, the effective contact rate is reduced because a newly infected node is connected to the node that infected it, which is usually not susceptible. Instead of calculating theoretical values for comparison that incorporate these assumptions, the experiment also simulates epidemics on networks with uncorrelated degree (zero assortativity) and minimum clustering coefficient.

To isolate the effect of clustering coefficient and assortativity, all networks are generated with the same algorithm. The algorithm creates edges locally so as to induce clustering, but first moves nodes with similar target degrees close to each other to ensure those local connections also lead to positive degree assortativity (Badham and Stocker, 2010). Briefly, the steps are as follows:

1. Assign target degrees for each node, which are randomly located on a (notional) ring.
2. Uniformly randomly select pairs of nodes and calculate the mean target degree near each node in the ring. Swap the nodes if the higher degree

node is in the lower degree region of the ring. Repeat such selections and swaps some arbitrary number of times.

3. For each node in random order, create edges to the nearest nodes first. An edge is created with fixed probability and possible edges are tested moving further away until the target degree is reached.
4. Measure the assortativity of the generated network. If is too low, destroy the edges and return to the node swapping step.
5. Once a suitable network is generated, the ring locations are forgotten.

This network generation algorithm has input parameters of the target degree distribution, clustering coefficient and (positive) assortativity. A specific degree distribution is used and the target clustering coefficient and assortativity are varied for this study.

To incorporate a real-world shape, the degree distribution used is derived from the number of nominated friends in a study of the friendship network of young children (Rapoport and Horvath, 1961). Each child was nominated by between 0 and 29 of the children in the social group, with mean 6.84. The cumulative probability distribution is rescaled to increase mean degree to 8, to simplify future comparison with other degree distribution shapes. This rescaled distribution is sampled until the required number of nodes has been assigned a nonzero degree. Figure 1 shows the expected degree distribution for a 1 000 node network.

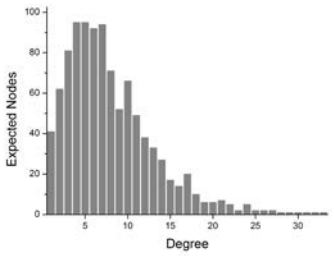

Figure 1: Expected degree frequency distribution for 1 000 node network. Note the positive skew that is characteristic of social networks.

Networks are generated that differ over three parameters:

- Size N: 500, 1 000, 2 500 nodes
- Assortativity r: 0.0 to 0.3, by 0.1
- Clustering C: 0.0 to 0.4, by 0.1

For each network size, 5 instances of the degree distribution are extracted. For each of these, up to 10 attempts are made to generate a network based on that degree sequence with each combination of assortativity and clustering coefficient. The algorithm does not guarantee that a network is successfully

generated. A successful attempt is defined as both assortativity and clustering coefficient within 0.05 of target.

The generated network may have some nodes with a final degree lower than its target. This occurs where target degree could only be achieved by creating an edge between a node and itself, or duplicating an existing edge. The degree sequences realised are more stable for the larger size experimental series than for the smaller. The 500 node networks have mean degree between 7.6 and 8.3, and the 2 500 node networks have mean degree between 7.9 and 8.1.

For each generated network, 30 SIR potential epidemics are simulated starting from a single infected node selected uniformly at random. An infected node has a 1/3 probability of recovery and becoming immune. That is, the mean duration of infection is 3 timesteps. While infected, a node has a given probability each timestep (1/12 or 1/8) of infecting each susceptible node with which it shares an edge. These infection probability values were chosen so that epidemics occur for many, but not all, simulations and their particular values are based on trial runs .

Thus, there are 6 series of experiments (3 network sizes and 2 infectivity rates) over which clustering coefficient and assortativity are varied. Up to 150 data points are obtained for each parameter combination. Table 1 displays the number of simulations in each experimental series (which is the same for each infectivity rate).

Table 1: Number of simulations by network size (N), assortativity (r) and clustering coefficient (0.0 to 0.4).

N	r	0.0	0.1	0.2	0.3	0.4
500	0.0	150	150	150	90	30
	0.1	120	150	150	150	120
	0.2	0	150	150	90	150
	0.3	0	150	150	150	150
1 000	0.0	150	120	90	120	120
	0.1	150	150	120	120	150
	0.2	0	150	150	150	150
	0.3	0	150	150	150	150
2 500	0.0	150	150	60	0	0
	0.1	150	150	150	150	150
	0.2	150	150	150	150	150
	0.3	150	150	150	150	150

Each simulation is run until there are no infected nodes. Results obtained from the simulation are final size (as a proportion of nodes) and whether an epidemic occurred. For this study, an epidemic occurs where at least 25 nodes ever become infected, which represents at least some secondary infections. In addition, equation (1) is used to estimate an "apparent" basic reproduction ratio. While the equation is not valid in a population where number of contacts varies, final size by degree is not available in real world situations and antibody

prevalence is used to estimate R_0 in the absence of better information (Anderson and May, 1992, Table 4.1).

4. Results

To simplify the results presentation, full details are presented for only one series: 1 000 nodes with infectivity of 1/8. Summary results are presented for all series and any differences in patterns in the detailed results are described.

The network generation algorithm does not generate networks uniformly at random from the ensemble of possible networks constrained by the degree distribution, assortativity and clustering coefficient. Thus, the analysis does not include measures of statistical significance, nor construct regression models of the relationship between network structure and epidemic features. However, the pattern of results by network property can suggest the magnitude of effects despite the lack of a statistical underpinning.

From Table 2, there is no consistent trend in epidemic proportion related to either property of interest. However, low epidemic occurrence is associated with a high value of either property (see Table 3). Further, the impact of these properties can be substantial, with a large difference in proportion of epidemics within each set of networks.

Table 2: Proportion of simulations with epidemics (at least 25 nodes infected) by network assortativity (r) and clustering coefficient: series 1 000 nodes and infectivity rate of 0.125.

r	0.0	0.1	0.2	0.3	0.4
0.0	0.65	0.64	0.67	0.53	0.65
0.1	0.75	0.73	0.72	0.73	0.63
0.2		0.73	0.63	0.73	0.71
0.3		0.61	0.62	0.65	0.67

Table 3: Epidemic proportion - Minimum and maximum across assortativity / clustering coefficient pairs, all series (size, infectivity).

Series	Min	r/C	Max	r/C
500, 0.083	0.29	0.1/0.3	0.56	0.1/0.0
500, 0.125	0.42	0.2/0.3	0.70	0.1/0.0
1000, 0.083	0.34	0.2/0.4	0.61	0.1/0.0
1000, 0.125	0.53	0.0/0.3	0.75	0.1/0.0
2500, 0.083	0.39	0.3/0.3	0.64	0.1/0.4
2500, 0.125	0.53	0.3/0.3	0.78	0.0/0.0

Unlike epidemic proportion, epidemic size does demonstrate a clear pattern over the range of structural values. Final size, and hence epidemic derived R_0, decreases as either assortativity or clustering coefficient increases (see Table 4 and Figure 2). For networks with 1 000 nodes and infectivity rate of 1/8, mean final size decreases from 0.80 to 0.59 over the range of network properties tested,

with a consequent decrease in apparent R_0 from 2.02 (r=0.0, C=0.0) to 1.55 (r =0.3, C=0.4).

Table 4: Epidemic final size f by network assortativity (r) and clustering coefficient, mean over (up to 150) simulations where epidemic occurred: series 1 000 nodes and infectivity rate of 0.125.

r	0.0	0.1	0.2	0.3	0.4
0.0	0.80	0.79	0.79	0.77	0.73
0.1	0.79	0.79	0.78	0.76	0.71
0.2		0.76	0.74	0.68	0.69
0.3		0.73	0.70	0.64	0.59

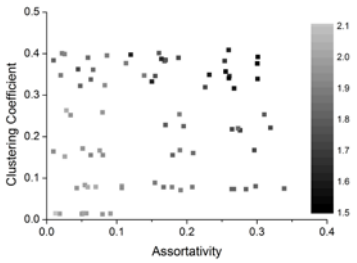

Figure 2: Epidemic derived R_0 by network properties: series 1 000 nodes and infectivity rate of 0.125. The mean R_0 of the (up to 30) epidemics simulated over each network is indicated by the colour of the point with coordinates determined by the assortativity and clustering coefficient of the network. Points are darker (indicating smaller values of mean R_0) for higher values of either network property.

Similar trends are displayed in all experimental series (Table 5), which displays the mean epidemic derived R_0 over the (up to 150) simulations where epidemics occurred. Epidemics over the networks with the highest values of assortativity and clustering coefficient are smaller, with derived values of R_0 about 75% of those derived from simulations with zero values of these network properties.

5. Discussion

This paper highlights a relatively neglected aspect of social structure in epidemic models. Sociologists studying friendship patterns and other relationships potentially relevant to epidemiology have developed several standard measures of structure, including clustering coefficient (C) and degree assortativity (r). Social networks are known to have positively skewed degree distributions, higher C than expected from degree distribution and positive r. These properties are difficult to jointly incorporate into epidemic models. However, simulation provides a viable alternative research methodology, with algorithms able to generate sets

Table 5: Minimum and maximum mean of epidemic derived R_0 across assortativity / clustering coefficient pairs, all series (size, infectivity).

Series	Min	r/C	Max	r/C
500, 0.083	1.22	0.3/0.4	1.57	0.0/0.0
500, 0.125	1.54	0.3/0.4	2.00	0.0/0.0
1000, 0.083	1.24	0.2/0.3	1.65	0.0/0.0
1000, 0.125	1.55	0.3/0.4	2.02	0.0/0.0
2500, 0.083	1.15	0.3/0.4	1.63	0.0/0.0
2500, 0.125	1.52	0.3/0.4	2.03	0.0/0.0

of networks that vary on specific properties. At a minimum, studies that investigate the effect of clustering coefficient or assortativity should also measure the other structural property, to ensure effects are attributed appropriately.

The study presented here examines only a limited set of networks using a specific generation algorithm, and there is no attempt to calibrate the simulations to a specific disease or the network to a specific population. Over the experimental sets, epidemics are smaller on networks with increased C or r and the effects of these properties are additive. For socially realistic values of these structural properties, the potential variation on estimated R_0 is greater than 20% when compared to networks where these properties are zero. The impact has a similar magnitude for all three network sizes and both infectivity rates examined.

For epidemic occurrence, the results are less consistent. Generally, a higher proportion of simulations resulted in epidemics over networks with lower values for C or r. That is, a high value of clustering coefficient or assortativity can lead to the area of infection becoming trapped. However, at lower property values, an increase in the property value is not consistently associated with a decrease in epidemic proportion.

These results suggest that structural properties identified by social network researchers are relevant for epidemiology and systematic research is warranted due to the potential size of the effect. If social structures are very different between the societies where an infection occurs and R_0 is derived, and the societies where the derived R_0 is applied, these results suggest the error could be in the order of 20%.

In future work, a broader selection of networks is necessary that include degree sequences with larger and smaller skewness, and different values for mean degree. Further, infectivity and recovery rates should be set to simulate real diseases.

6. Acknowledgement

The authors would like to thank Professor Hussein Abbass for many useful discussions about this project, and Katie Glass, David Philp and the anonymous

referees for several useful comments on an earlier draft. This research is primarily funded by a University College Postgraduate Research Scholarship and supported by the ARC Centre for Complex Systems grant number CEO0348249.

Anderson, R., May, R., 1992. Infectious diseases of humans: Dynamics and control.

Badham, J., Abbass, H., Stocker, R., 2008. Parameterisation of Keeling's network generation algorithm.

Badham, J., Stocker, R., 2010. A spatial approach to network generation for three properties: Degree distribution, clustering coefficient and degree assortativity.

Ball, F., Clancy, D., 1993. The final size and severity of a generalised stochastic multitype epidemic model.

Becker, N., 1973. Carrier-borne epidemics in a community consisting of different groups.

Boguñá, M., Pastor-Satorras, R., 2002. Epidemic spreading in correlated complex networks.

Brauer, F., 2008. Compartmental models in epidemiology.

Eubank, S., Guclu, H., Kumar, V., Marathe, M., Srinivasan, A., Toroczkai, Z., Wang, N., 2004. Modelling disease outbreaks in realistic urban social networks.

Ferguson, N., Keeling, M., Edmunds, W., Gani, R., Grenfell, B., Anderson, R., Leach, S., 2003. Planning for smallpox outbreaks.

Gupta, S., Anderson, R., May, R., 1989. Networks of sexual contacts: implications for the pattern of spread of HIV.

Holme, P., Zhao, J., 2007. Exploring the assortativity-clustering space of a network's degree sequence.

Keeling, M., 1999. The effects of local spatial structure on epidemiological invasions.

Keeling, M., 2005. The implications of network structure for epidemic dynamics.

Keeling, M., Eames, K., 2005. Networks and epidemic models.

Kermack, W., McKendrick, A., 1927. Contributions to the mathematical theory of epidemics - I.

Lajmanovich, A., Yorke, J., 1976. A deterministic model for gonorrhea in a nonhomogeneous population.

Moreno, Y., Gómez, J., Pacheco, A., 2003. Epidemic incidence in correlated complex networks.

Newman, M., 2002a. Assortative mixing in networks. Newman, M.,

2002b. Spread of epidemic disease on networks. Newman, M., 2003.

The structure and function of complex networks.

Newman, M., Park, J., 2003. Why social networks are different from other types of networks.

Nold, A., 1980. Heterogeneity in disease transmission modelling.

Rapoport, A., Horvath, W., 1961. A study of a large sociogram.

Smith, D., Lee, C., Onnela, J.-P., Johnson, N., 2008. Link-space formalism for network analysis.

Soffer, S., Vázquez, A., 2005. Network clustering coefficient without degree-correlation biases.

Watts, D., 2004. The 'new' science of networks.

Watts, D., Strogatz, S., 1998. Collective dynamics of 'small-world' networks.

Yan, P., 2008. Distribution theory, stochastic processes and infectious disease modelling.