THE UNIVERSITY OF

WARWICK

Generalised network clustering and its dynamical implications

Thomas House

Abstract

A parameterisation of generalised network clustering, in the form of four-motif prevalences, is presented. This involves three real parameters that are conditional on one- two- and three-motif prevalences. Interpretations of these real parameters are presented that motivate a set of rewiring schemes to create appropriately clustered networks. Finally, the dynamical implications of higher order structure, as parameterised, for a contact process are considered.

1 Introduction

Networks have become one of the indispensible tools for the study of complex systems with many interacting components, as demonstrated by their ubiquity in the Proceedings of the recent European Conference on Complex Systems with which this journal issue is concerned. In particular, the combination of high clustering amongst nodes and short average path length, commonly known as the small world phenomenon [13], has been observed not only in social networks [11], but also in technological, metabolic and citation networks [9, 10, 12].

Small connected sub-graphs of complex networks, known as *motifs*, have also been observed to have significantly different prevalences from those expected in a random case, leading to scientific insight [8, 7]. This paper is concerned with an alternative approach to motif prevalence that conditions on standard, triangle-level clustering, as a guide to intuition for other applications of the concept of motifs. In particular, new wirings are presented that modify clustering without changing node degree, along the lines of [1, 6, 4, 2].

Networks, and population structure in general, have also become central to modern infectious disease epidemiology [5]. The impact of motif structure for *SIS* epidemics was considered in [3], and we combine the dynamical system developed in that work with the new parameterisation to gain insights into the impact of higher-order clustering on transmission / contact process dynamics.

2 Characterisation of motif structure

We start by considering the relatively simple structure of one- two- and three-motif prevalences. At orders one and two, there are only the number of nodes in a network and the number of links to consider. For simplicity, we consider networks with a single giant component of N nodes in which each individual has exactly n links connecting it to the rest of the network. This assumption

is not essential to the general thrust of analysis presented, but does simplify an already complex set of manipulations. In our notation, we use a diagramatic representation of a node and linked nodes enclosed in square brackets to denote prevalence of that motif in the network. This means that at order one and two:

$$\left[\,\bullet\,\right] = N \ , \quad \left[\,-\,\right] = nN \ . \tag{1}$$

So the motif structure at this level is given equivalently either by the raw motif prevalences $\left[\,\bullet\,\right], \left[\,-\,\right]$, or by the real numbers N, n. The benefit of the latter approach is that n tells us something about the number of links *per node*—i.e. two-motif structure conditional on one-motif structure.

Less trivially, there are two connected three-motifs: triangles and unclosed triples. Since every triple must be either closed or unclosed, the prevalences of three-motifs, notated using square brackets and diagrams as for other motifs, obey the identity

$$\left[\wedge\right] + \left[\triangle\right] = Nn(n-1) \ . \tag{2}$$

This means that a real parameter $\phi \in [0,1]$ can be introduced to partition this identity as below:

$$\left[\wedge\right] = Nn(n-1)(1-\phi) \ , \quad \left[\triangle\right] = Nn(n-1)\phi \ . \tag{3}$$

In network analysis, ϕ (the ratio of triangles to all triples, closed and unclosed) is often called the *clustering coefficient*. In the same way that n conditions on network size, ϕ conditions on network size and number of links to measure transitivity of the network in a different manner from raw counts of triangles.

We now attempt a similar parameterisation at order four. There are six connected graphs of size four, which can be represented pictorally using the following set of symbols:

$$\left\{\,\boxdiagup, \sqcap, \boxslash, \square, \boxbslash, \boxtimes\,\right\} \ .$$

A set of identities analogous to (2) was introduced in [3],

$$\left[\boxslash\right] + 2\left[\boxbslash\right] + \left[\boxtimes\right] = (n-2)\left[\triangle\right] \ ,$$
$$\left[\boxdiagup\right] + 2\left[\boxslash\right] + \left[\boxbslash\right] = (n-2)\left[\wedge\right] \ , \tag{4}$$
$$\left[\sqcap\right] + \left[\boxslash\right] + \left[\square\right] + \left[\boxbslash\right] = (n-1)\left[\wedge\right] \ .$$

Each of these identies is derived by starting with the three-motif appearing on the right-hand side of the identity (either a triangle or unclosed triple) and then joining a fourth node to one of the original three; the left-hand side of each identity can then be seen as an enumeration of the possible additional links between the new node and the two other nodes within the original three-motif. We now propose the main innovation of this work, a partition of these identities in terms of three real parameters, ψ, ζ, ξ.

We start this process by writing down the four-motif prevalences that would be expected if transitive closure of any given triple is a random event of constant probability ϕ. In the case where no triangles at all are present in the network, the only four-motif clustering structure possible is the closure of four-lines into squares, and the appropriate motifs obey $\left[\sqcap\right] + \left[\square\right] = Nn(n-1)^2$. This motivates the introduction of a square-level partition of these two motifs, ψ, analagous to ϕ in (3), but not equivalent in the case where some triangles are present in the network.

2

Finally, we introduce parameters ζ and ξ additively to the prevalences of the motifs ◺ and ⊠ respectively, and then use the identities (4) to carry through the consequences of this addition to other motif prevalences, yielding the form

$$\left[\diagup\!\!\!\!\diagdown \right] = Nn(n-1)(n-2)\left((1-\phi)^3 + 3\zeta\right) ,$$

$$\left[\sqcap \right] = Nn(n-1)\left((n-1)(1-\phi) - (n-2)\left(\phi(1-\phi) - \zeta - \frac{1}{3}\xi\right)\right)(1-\psi) ,$$

$$\left[\diagdown\!\!\!\!\diagup \right] = Nn(n-1)(n-2)\left(\phi(1-\phi)^2 - 2\zeta + \frac{1}{3}\xi\right) ,$$

$$\left[\square \right] = Nn(n-1)\left((n-1)(1-\phi) - (n-2)\left(\phi(1-\phi) - \zeta - \frac{1}{3}\xi\right)\right)\psi ,$$

$$\left[\diagdown\!\!\!\!\diagup \right] = Nn(n-1)(n-2)\left(\phi^2(1-\phi) + \zeta - \frac{2}{3}\xi\right) ,$$

$$\left[\boxtimes \right] = Nn(n-1)(n-2)\left(\phi^3 + \xi\right) .$$

$$(5)$$

Requiring that no motif prevalence be negative, the new parameters sit in the following ranges, provided each of the others is zero:

$$\psi \in [0,1] ,$$

$$\zeta \in \left[\max\left(-\frac{1}{3}(1-\phi)^3, -\phi^2(1-\phi)\right), \frac{1}{2}\phi(1-\phi)^2\right] ,$$

$$\xi \in \left[\max\left(-3\phi(1-\phi)^2, -\phi^3\right), \frac{3}{2}\phi^2(1-\phi)\right] .$$

$$(6)$$

A neighbourhood-based interpretation of these new parameters for certain limiting cases is considered in Figure 1. This figure shows a typical neighbourhood around an individual in a network with $n = 6$, and clustering parameter values varied. Plot (a) shows a completely unclustered graph—essentially a Cayley tree of degree 6. In (b), triangle-level clustering ϕ has been introduced, but in such a way that the triangles do not form highly connected fourth-order structures. (c) shows how the 'envelope' shape ◺ is more prevalent than would be expected on the basis of the three-motif structure: this means that ζ is positive, while for under-represented envelopes ζ would be negative. (d) shows that the 'four-clique' ⊠ is, in the same way, over-represented, implying positive ξ, while its under-representation would imply negative ξ. Finally, (e) shows that ψ represents the ratio of squares to all four-lines (closed and unclosed), and involves connections being made further away from the central node than other clustering parameters. In this plot, as with (b), the squares are shown maximally uncorrelated—obviously, at still higher orders of clustering, correlations between squares may be parameterised as for triangles.

Outside of these limiting cases, however, the interpretation of the new clustering parameters is more subtle, since the consistency conditions (4) are much more structured than (2). In this sense, the parameterisation of four-motif structure is not a straightforward extension of the methodology used at the three-motif level.

3 Rewiring schemes

Rewiring schemes that preserve the number of links attached to a node can play an important role in understanding, creating and manipulating networks. We now present a set of rewiring schemes that modify the clustering parameters we have introduced, two from existing work (together with applications) and three that are, to our knowledge, novel. These rewirings are an aid to intuition and also demonstrate that explicit networks of the kind considered here can be generated given sufficient computational resources. Nevertheless, their naïve implementation is highly computationally intensive, and does not scale well with network size, meaning that technical innovation beyond the scope of this paper is necessary to produce simulations equivalent to the results obtained below using moment closure.
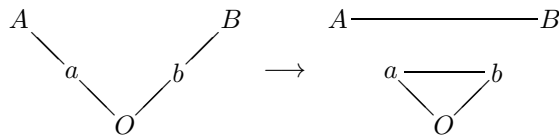
3.1 Randomiser

This rewiring was used recently in epidemiological applications [4, 2] to remove all forms of clustering without changing degree distribution.
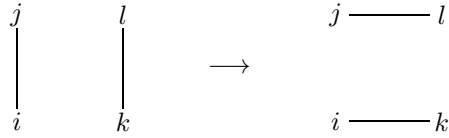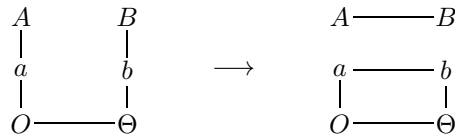
3.2 'Big V'

This rewiring was considered recently in [1, 6, 4] to increase ϕ without changing the degree distribution.

3.3 'Big U'

This novel rewiring increases ψ

3.4 'YV'

This novel rewiring increases ζ and ϕ.
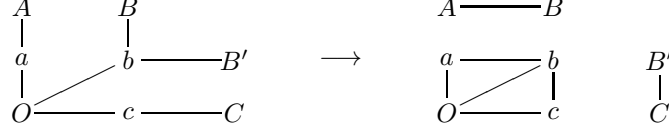
3.5 'YYY'

This novel rewiring increases ξ and ϕ.
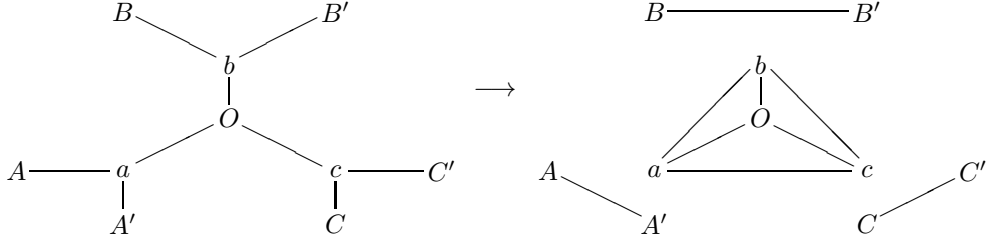
4 Contact-process dynamics

We now present the model of [3], used to investigate the impact of higher order-clustering on what epidemiologists call *SIS* dynamics. In these dynamics, often called a *contact process*, individuals are either susceptible (S) or infectious (I) with letters $A, B, C \ldots$ representing either of these states. Transmission of infection happens between infectious individuals I and susceptible individuals S linked on the network at a rate τ, while infectious individuals recover and become susceptible, since recovery is assumed not to offer lasting immunity, at a rate g. We use square brackets to denote the prevalence of certain structures in the network.

4.1 Exact dynamical equations

The model in question takes as its starting point a set of differential equations that are, in the $N \to \infty$ limit, exact but form an infite hierarchy. We present the first three orders of this:

$$\frac{d}{dt}[\,S\,] = -\tau[\,S\!-\!I\,] + g[\,I\,],$$

$$\frac{d}{dt}[\,I\,] = \tau[\,S\!-\!I\,] - g[\,I\,],$$

$$\frac{d}{dt}[\,S\!-\!S\,] = -2\tau[\,\underset{\cdots}{S\!-\!S\!-\!I}\,] + 2g[\,S\!-\!I\,],$$

$$\frac{d}{dt}[\,S\!-\!I\,] = \tau\left([\,\underset{\cdots}{S\!-\!S\!-\!I}\,] - [\,\underset{\cdots}{I\!-\!S\!-\!I}\,] - [\,S\!-\!I\,]\right) + g\left([\,I\!-\!I\,] - [\,S\!-\!I\,]\right),$$

$$\frac{d}{dt}[\,I\!-\!I\,] = 2\tau\left([\,\underset{\cdots}{I\!-\!S\!-\!I}\,] + [\,S\!-\!I\,]\right) - 2g[\,I\!-\!I\,],$$

$$\frac{d}{dt}[\,S\!-\!S\!-\!S\,] = -\tau\left(2[\,\overset{\cdots}{S\!-\!S\!-\!S\!-\!I}\,] + [\,\overline{S\!-\!S\!-\!S\cdots I}\,]\right) + g\left(2[\,S\!-\!S\!-\!I\,] + [\,S\!-\!I\!-\!S\,]\right),$$

$$\frac{d}{dt}[\,S\!-\!S\!-\!I\,] = \tau\left([\,\overset{\cdots}{S\!-\!S\!-\!S\!-\!I}\,] - [\,\overset{\cdots}{S\!-\!S\!-\!I\!-\!I}\,] - [\,\overline{S\!-\!S\!-\!I\cdots I}\,] - [\,S\!-\!S\!-\!I\,]\right)$$
$$+ g\left([\,S\!-\!I\!-\!I\,] + [\,I\!-\!S\!-\!I\,] - [\,S\!-\!S\!-\!I\,]\right),$$

$$\frac{d}{dt}[\,S\!-\!I\!-\!S\,] = +\tau\left([\,\overline{S\!-\!S\!-\!S\cdots I}\,] - 2[\,\overset{\cdots}{S\!-\!I\!-\!S\!-\!I}\,] - 2[\,S\!-\!I\!-\!S\,]\right)$$
$$+ g\left(2[\,S\!-\!I\!-\!I\,] - [\,S\!-\!I\!-\!S\,]\right),$$

$$\frac{d}{dt}[\,S\!-\!I\!-\!I\,] = \tau\left([\,\underset{\cdots}{S\!-\!I\!-\!S\!-\!I}\,] + [\,\overset{\cdots}{S\!-\!S\!-\!I\cdots I}\,] - [\,\underset{\cdots}{S\!-\!I\!-\!I\cdots I}\,]\right.$$
$$\left. + [\,S\!-\!I\!-\!S\,] + [\,S\!-\!S\!-\!I\,] - [\,S\!-\!I\!-\!I\,]\right) + g\left([\,I\!-\!I\!-\!I\,] - 2[\,S\!-\!I\!-\!I\,]\right),$$

$$\frac{d}{dt}[\,I\!-\!S\!-\!I\,] = \tau\left(2[\,\overset{\cdots}{S\!-\!S\!-\!I\cdots I}\,] - [\,\overline{I\!-\!S\!-\!I\cdots}\,] - 2[\,I\!-\!S\!-\!I\,]\right)$$
$$+ g\left([\,I\!-\!I\!-\!I\,] - 2[\,I\!-\!S\!-\!I\,]\right),$$

$$\frac{d}{dt}[\,I\!-\!I\!-\!I\,] = \tau\left(2[\,\overset{\cdots}{S\!-\!I\!-\!I\cdots I}\,] + [\,\overline{I\!-\!S\!-\!I\cdots}\,] + 2[\,S\!-\!I\!-\!I\,] + 2[\,I\!-\!S\!-\!I\,]\right)$$
$$- 3g[\,I\!-\!I\!-\!I\,],$$

$$\frac{d}{dt}[\,\underline{S\!-\!S\!-\!S}\,] = -3\tau[\,\underline{S\!-\!\overset{\cdots}{S\!-\!S}\!-\!I}\,] + 3g[\,\underline{S\!-\!S\!-\!I}\,],$$

$$\frac{d}{dt}[\,\underline{S\!-\!S\!-\!I}\,] = \tau\left([\,\underline{S\!-\!\overset{\cdots}{S\!-\!S}\!-\!I}\,] - 2[\,\underline{S\!-\!\overset{\cdots}{S\!-\!I\cdots I}}\,] - 2[\,\underline{S\!-\!S\!-\!I}\,]\right)$$
$$+ g\left(2[\,\underline{S\!-\!I\!-\!I}\,] - [\,\underline{S\!-\!S\!-\!I}\,]\right),$$

$$\frac{d}{dt}[\,\underline{S\!-\!I\!-\!I}\,] = \tau\left(2[\,\underline{S\!-\!\overset{\cdots}{S\!-\!I\cdots I}}\,] - [\,\underline{S\!-\!\overset{\cdots}{I\!-\!I\cdots I}}\,] + 2[\,\underline{S\!-\!S\!-\!I}\,] - 2[\,\underline{S\!-\!I\!-\!I}\,]\right)$$
$$+ g\left([\,\underline{I\!-\!I\!-\!I}\,] - 2[\,\underline{S\!-\!I\!-\!I}\,]\right),$$

$$\frac{d}{dt}[\,\underline{I\!-\!I\!-\!I}\,] = 3\tau\left([\,\underline{S\!-\!\overset{\cdots}{I\!-\!I\cdots I}}\,] + 2[\,\underline{S\!-\!I\!-\!I}\,]\right) - 3g[\,\underline{I\!-\!I\!-\!I}\,]. \tag{7}$$

Here and throughout this paper, we use dotted lines to imply expansion as below:

$$[\,\underset{\cdots}{A\!-\!B\!-\!C}\,] = [\,A\!-\!B\!-\!C\,] + [\,\underline{A\!-\!B\!-\!C}\,],$$

$$[\,\underset{\cdots}{A\!-\!B\!-\!C\!-\!D}\,] = [\,A\!-\!B\!-\!C\!-\!D\,] + [\,\underline{A\!-\!B\!-\!C}\!-\!D\,] + [\,A\!-\!\overline{B\!-\!C\!-\!D}\,] + [\,\underline{A\!-\!B\!-\!C\!-\!D}\,],$$

$$\vdots$$

$$\tag{8}$$

and similarly for other fourth-order terms.

4.2 Closure schemes

To integrate the system as presented so far, we need a closure scheme, previously introduced in [3], which is most easily expressed in terms of the raw motif prevalences.

$$[\,A\!-\!\overline{B\!-\!C}\ D\,] \approx [\,\text{motif}\,]\frac{\left[-\right]^3}{\left[\wedge\right]^3\left[\,\bullet\,\right]}\frac{[\,A\!-\!B\!-\!C\,]\,[\,A\!-\!B\!-\!D\,]\,[\,C\!-\!B\!-\!D\,][B]}{[\,A\!-\!B\,][\,B\!-\!C\,][\,B\!-\!D\,]}\,,$$

$$[\,A\!-\!B\!-\!C\!-\!D\,] \approx [\,\text{motif}\,]\frac{\left[-\right]}{\left[\wedge\right]^2}\frac{[\,B\!-\!C\!-\!D\,]\,[\,A\!-\!B\!-\!C\,]}{[\,B\!-\!C\,]}\,,$$

$$[\,\underline{A\!-\!B\!-\!C}\!-\!D\,] \approx [\,\text{motif}\,]\frac{\left[-\right]^3}{\left[\triangle\right]\left[\wedge\right]^2\left[\,\bullet\,\right]}\frac{[\,\underline{A\!-\!B\!-\!C}\,]\,[\,B\!-\!C\!-\!D\,]\,[\,A\!-\!C\!-\!D\,][C]}{[\,A\!-\!C\,]\,[\,B\!-\!C\,]\,[\,C\!-\!D\,]}\,,$$

$$[\,\underline{A\!-\!B\!-\!C\!-\!D}\,] \approx [\,\text{motif}\,]\frac{\left[-\right]^4}{\left[\wedge\right]^4}\frac{[\,A\!-\!B\!-\!C\,]\,[\,B\!-\!C\!-\!D\,]\,[\,C\!-\!D\!-\!A\,]\,[\,D\!-\!A\!-\!B\,]}{[\,B\!-\!C\,]\,[\,C\!-\!D\,]\,[\,D\!-\!A\,]\,[\,A\!-\!B\,]}\,,$$

$$[\,A\!-\!\overline{B\!-\!C}\!-\!\underline{D}\,] \approx [\,\text{motif}\,]\frac{\left[-\right]}{\left[\triangle\right]^2}\frac{[\,B\!-\!C\!-\!D\,]\,[\,\underline{A\!-\!B\!-\!C}\,]}{[\,B\!-\!C\,]}\,,$$

$$[\,\underline{A\!-\!\overline{B\!-\!C}\!-\!D}\,] \approx [\,\text{motif}\,]\frac{\left[-\right]^6}{\left[\triangle\right]^4\left[\,\bullet\,\right]^4}\times$$
$$\frac{[\,\underline{A\!-\!B\!-\!C}\,]\,[\,\underline{B\!-\!C\!-\!D}\,]\,[\,\underline{C\!-\!D\!-\!A}\,]\,[\,\underline{D\!-\!A\!-\!B}\,][A][B][C][D]}{[\,B\!-\!C\,]\,[\,C\!-\!D\,]\,[\,D\!-\!A\,]\,[\,A\!-\!B\,]\,[\,A\!-\!C\,]\,[\,B\!-\!D\,]}\tag{9}$$

Then (5), together with this closure and equations (7) create an integrable ODE system. Provided τ is sufficiently large compared to g, this system has a steady state with a non-trivial proportion I^* of the network infectious. Standardly, this equilibrium value is called the *endemic state*. We investigate this dynamically in Figure 2, where ϕ is increased for other parameters held constant in all plots, giving the common black line. We then modify either (a) ζ, (b) ξ or (c) ψ. This shows that, essentially, ψ has a significant but relatively stable effect in reducing the endemic state at each ϕ value, while ξ can have a significant effect in either direction at larger ϕ values. ζ, on the other hand, is relatively dynamically unimportant, except perhaps at moderate values of ϕ. We also note that positive values of ζ reduce the endemic state, and negative values increase it, while the opposite is true for ξ.

5 Discussion

This paper has presented a novel way of thinking about higher-order structure in networks, together with intuitive explanations of this, rewiring schemes and dynamical consequences. This opens up three main questions.

Firstly, what are reasonable parameter values for networks that are seen in nature, and which can be explicitly constructed? The exact values of clustering coefficients considered in the *SIS* model are perhaps slightly larger than are likely to be seen or constructed, although this should be mainly of quantitative importance since the qualitative dynamical implications found for higher order clustering are not modified at different coefficient values, and moment closure

(particularly in the three-motif case) has been extensively used in modelling *SIS* and *SIR* dynamics without producing qualitatively incorrect results [5]. Nevertheless, if sufficiently efficient methods were available to generate explicit networks to run stochastic simulations on, that would significantly increase the confidence in the results obtained here using moment closure.

Secondly, how can this analysis be generalised to networks with heterogeneous numbers of links, and (perhaps more problematically) preferential assortative connection between nodes of similar degree? While such analysis is doubtless possible, the large number of interacting quantities may make analytic results technically difficult. In particular, it it unlikely that arbitrary heterogeneity and clustering statistics are compatible with each other.

Finally, under what conditions is it necessary to consider k-motifs for a given k? Clearly, a high preponderance of triangles in a network would favour a pairwise model, but this answer is less clearly posed for four-motifs in general. However, the parameterisation suggested here goes some way towards answering this question: starting with a set of four-motif prevalences, are these significantly different from what would be predicted based on the values of three- two- and one-motif parameters ϕ, n and N? If the new generalised clustering parameters ϕ, ζ and ξ are significantly different from zero, then we would expect that at least four-motifs should be considered in analysis of the network.

Acknowledgments

References

[1] Bansal, S., Khandelwal, S., and Meyers, L. A., Exploring biological network structure with clustered random networks, *BMC Bioinformatics* **10** (2009) 405.

[2] Green, D. M. and Kiss, I. Z., Large-scale properties of clustered networks: Implications for disease dynamics, *Journal of Biological Dynamics* (2010).

[3] House, T., Davies, G., Danon, L., and Keeling, M. J., A motif-based approach to network epidemics, *Bulletin of Mathematical Biology* **71** (2009) 1693–1706.

[4] House, T. and Keeling, M. J., The impact of contact tracing in clustered populations, *PLoS Computational Biology* **6** (2010) e1000721.

[5] Keeling, M. J. and Rohani, P., *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2007).

[6] Kiss, I. Z. and Green, D. M., Comment on "properties of highly clustered networks", *Physical Review E* **78** (2008) 048101.

[7] Milo, R., Superfamilies of evolved and designed networks, *Science* **303** (2004) 1538–1542.

[8] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., Network motifs: simple building blocks of complex networks, *Science* **298** (2002) 824–7.

[9] Newman, M. E. J., The structure and function of complex networks, *SIAM Review* **45** (2003) 167–256.

[10] Strogatz, S. H., Exploring complex networks, *Nature* **410** (2001) 268–276.

[11] Wasserman, S. and Faust, K., *Social Network Analysis: Methods and Applications* (Cambridge University Press, 1994).

[12] Watts, D. J., *Six degrees: The science of a connected age* (WW Norton, 2003).

[13] Watts, D. J. and Strogatz, S. H., Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440–442.

(a) $\phi, \psi = 0$

(b) $\phi > 0, \psi = 0$, minimal ζ, ξ

(c) $\phi, \zeta > 0, \psi = 0$, minimal ξ

(d) $\phi, \xi > 0, \psi = 0$, minimal ζ
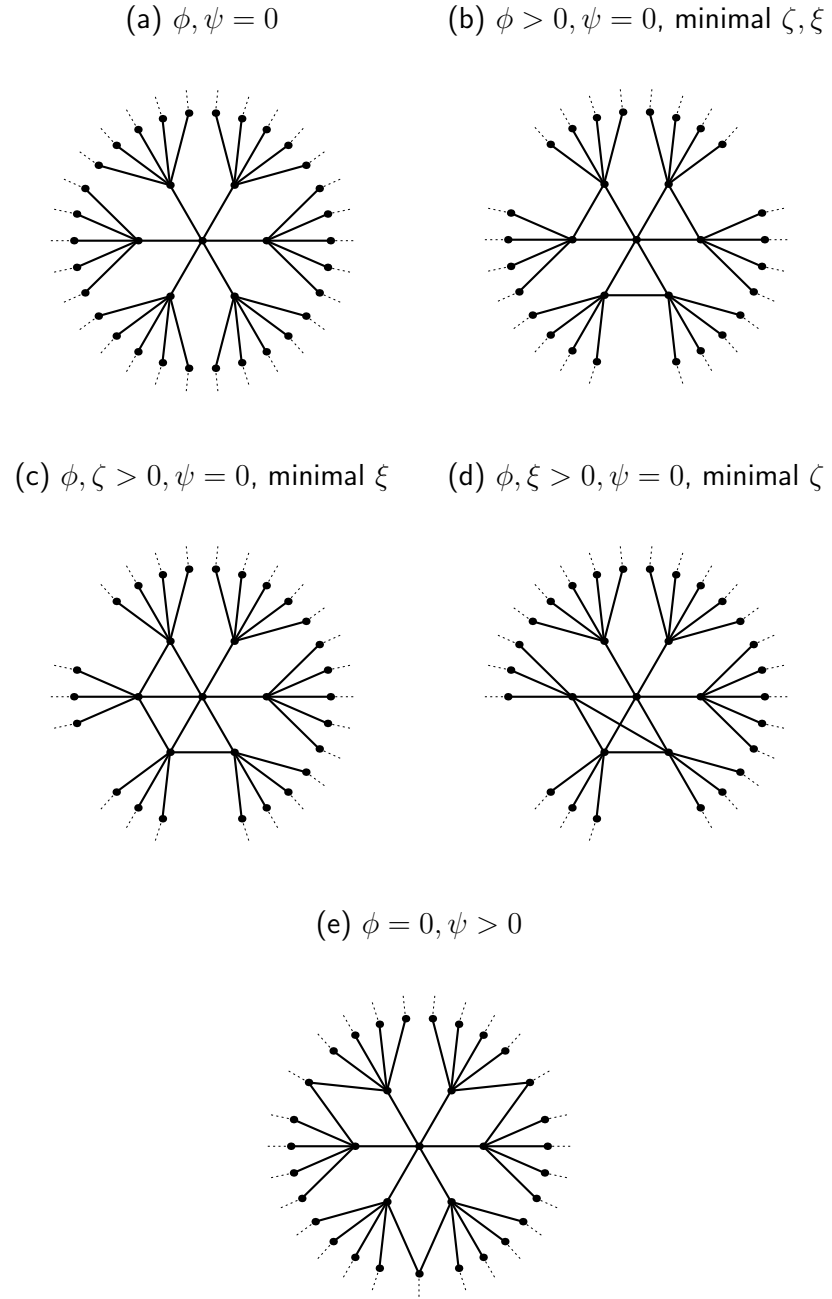
(e) $\phi = 0, \psi > 0$

Figure 1: Interpretation of the clustering parameters ϕ, ψ, ζ, ξ for a typical neighbourhood in a network with $n = 6$
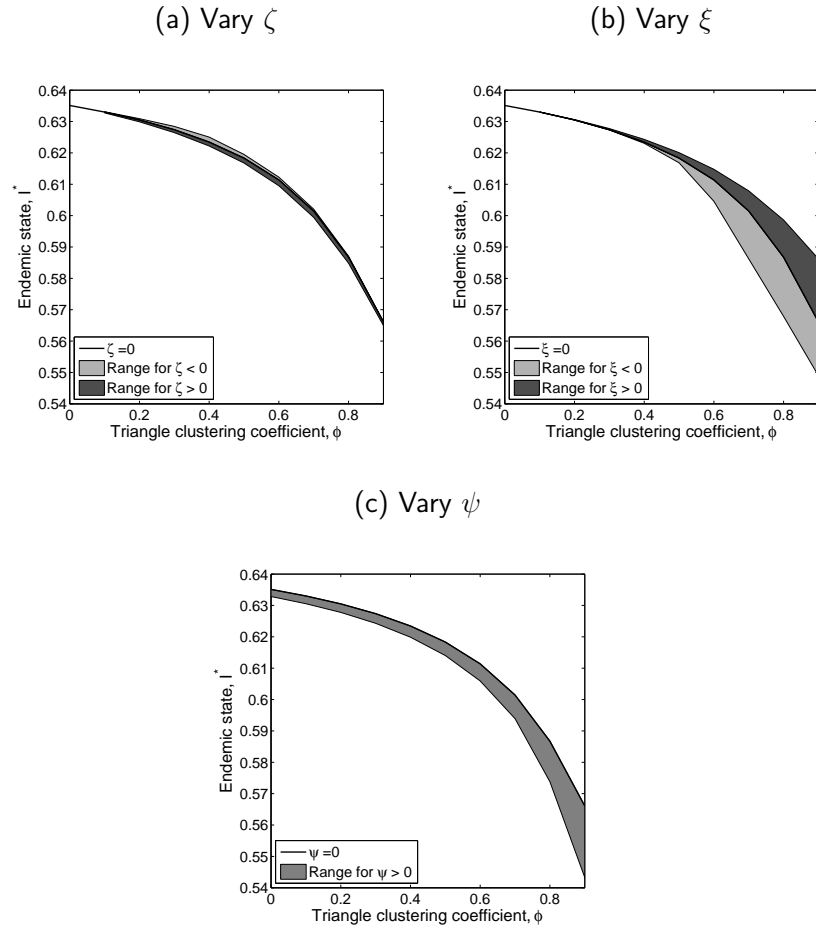
Figure 2: Dynamical results for the endemic state of the triplewise contact process model for $n = 6, g = 1, \tau = 3/5$ and other parameters varied.