Stability in flux: Community structure in dynamic networks

By John Bryden[1†‡], Sebastian Funk[1,2‡], Nicholas Geard[3‡], Seth Bullock[3], Vincent A.A. Jansen[1]

[1]*School of Biological Sciences, Royal Holloway, University of London, Egham TW20 0EX, UK*

[2]*Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK*

[3]*School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK*

The structure of many biological, social and technological systems can usefully be described in terms of complex networks. Although often portrayed as fixed in time, such networks are inherently dynamic, as the edges that join nodes are cut and rewired, and nodes themselves update their states. Understanding the structure of these networks requires us to understand the dynamic processes that create, maintain and modify them. Here, we build upon existing models of coevolving networks to characterise how dynamic behaviour at the level of individual nodes generates stable aggregate behaviours. We focus particularly on the dynamics of groups of nodes formed endogenously by nodes that share similar properties (represented as node state) and demonstrate that, under certain conditions, network modularity based on state compares well to network modularity based on topology. We show that if nodes rewire their edges based on fixed node states, the network modularity reaches a stable equilibrium which we quantify analytically. Furthermore, if node state is not fixed, but can be adopted from neighbouring nodes, the distribution of group sizes reaches a dynamic equilibrium, which remains stable even as the composition and identity of the groups changes. These results show that dynamic

† Corresponding author - john.bryden@rhul.ac.uk
‡ John Bryden, Sebastian Funk and Nicholas Geard contributed equally to this work.

networks can maintain the stable community structure that has been observed in many social and biological systems.

Keywords: coevolutionary networks, opinion formation, modularity, dynamic equilibrium, protein-protein interaction

1. Introduction

Many scenarios exist in nature and society where individuals or entities bias their interactions to a limited subset of a population. Populations that split into subpopulations interacting strongly within themselves but much less strongly between themselves are said to exhibit community structure. In human and animal societies this means that they consist of partially independent groups, cliques and tribes (Brown, 2000; Schelling, 1971; Lusseau and Newman, 2004), which can be important for studying epidemic spread (Salathé and Jones, 2010). This notion can be extended to more abstract representations of interactions in natural systems, such as in genetic, protein-protein and metabolic interaction networks that are structured into dynamic and functionally, spatially or temporally separated modules (Bader and Hogue, 2003; Li et al., 2010; Przytycka et al., 2010); or in neural networks where neurons tend to cluster into groups based on activity patterns (Segev et al., 2003).

The analysis of networks using tools borrowed from graph theory has proven to be a useful approach for studying populations where individuals or entities within the population interact only with a certain subset of the remaining population, and significant effort has been put into developing methods to identify community structure in such networks (Girvan and Newman, 2002; Schaeffer, 2007; Porter et al., 2009; Fortunato, 2010). The networks are usually taken to be static – they are presented or measured as snapshots in time, which neglects the fact that both the properties of individuals and the interactions between individuals will usually change over time. For example, human social and communication networks display complex community structure despite individuals continually changing their patterns of association (Palla et al., 2007). Only recently have an increasing number of studies concentrated on the dynamical properties of networks (Gross and Blasius, 2008), as well as their relevance to the spread of infectious diseases (Gross et al.,

2006; Volz and Meyers, 2009; van Segbroeck et al., 2010; Funk et al., 2010; Bansal et al., 2010).

Previous models of dynamic networks have considered the coevolution of opinions and network connections under *homophily* – where edges are rewired to nodes of the same state (McPherson et al., 2001) – and *heterophily* – where edges are rewired to nodes of a different state (Kimura and Hayakawa, 2008). In these studies, homophilous processes are often contrasted with *state-spread* – where states are transferred (or equilibrated) between nodes (Deffuant et al., 2001; Holme and Newman, 2006; Kozma and Barrat, 2008; Kimura and Hayakawa, 2008; Fu and Wang, 2008; Vazquez et al., 2008). Existing work has tended to focus on exploring the probability of achieving consensus, or the time taken to do so, and pay less attention to the dynamics that occur when multiple groups or communities coexist stably in the population.

Here, we focus on a topic that so far has received little attention: the emergence of community structure in dynamic networks. We introduce a model where each node has a state – which is either a fixed or dynamic property – and the network stays dynamic under homophilous and random rewiring. In addition to propagating states between nodes, we also use an "innovation" process to continually introduce diversity into the population. With this model, we study the emergence and stability of community structure in the resulting dynamic networks, and how they relate to properties at either the level of individual nodes or at population level.

2. Methods

We first state our microscopic (individual-based) model as an algorithm. We will later study the corresponding macroscopic (population-level) model, which approximates the average behaviour of the microscopic model and allows for mathematical treatment of some aspects of the model behaviour.

We consider a network of n nodes and m undirected edges, where each node i is associated with a state $S_i \in \{s_1, s_2, s_3, \ldots\}$. We deliberately leave interpretations of the meaning of the state open at this point, as we will consider both scenarios where states are fixed properties of nodes and ones where they can spread over the edges of the network. Either way, what we deem states of nodes will form the basis

for our implementation of homophilous rewiring, where nodes change edges to be preferentially connected to nodes of the same state.

In the individual-based model exactly one of the possible processes below, chosen with probability proportional to the corresponding rate, is invoked at each timestep. The lengths of inter-event times are exponentially distributed, in line with Gillespie (1977), so that the timescale remains consistent across different parameter settings. Based largely on models of opinion flow (Kimura and Hayakawa, 2008) and of social group formation (Geard and Bullock, 2008), we analyse the effects of two classes of simple processes on the network, one containing *rewiring* events and the other *state change* events. Let us first consider the class of processes dealing with rewiring: edges may either be rewired to nodes of the same state (homophilous rewiring) or to random nodes (random rewiring).

- *homophilous rewiring* (rate p) – Choose a random edge (i, j). Choose a random node k where $k \neq i$, $S_i = S_k$ and there is no edge (i, k). Delete edge (i, j) and add edge (i, k). If there exists no suitable k, do nothing.
- *random rewiring* (rate q) – Choose a random edge (i, j). Choose a random node k such that there is no edge (i, k). Delete edge (i, j) and add edge (i, k). If there is no suitable k, do nothing.

The second class of processes changes the states of the nodes: nodes may copy the state of connected nodes or be updated with a random state.

- *symmetric state spread* (rate r) – Choose a random edge (i, j). Set $S_j = S_i$.
- *innovation* (rate w) – Choose a random node i and a random state s_k where $\forall j, S_j \neq s_k$, set $S_i = s_k$.

Note that our implementation of state spread is symmetric in the sense that once an edge is chosen, its endpoints are randomly designated to be source and target. Choosing a random node first which then spreads its state to a neighbouring node would make states with many nodes more likely to spread and invade other state groups; choosing a random node which then copies a neighbouring state, on the other hand, makes states with many nodes more likely to be invaded by other state groups. Our method attempts to avoid these biases.

The rates given for the four processes are to be understood as local (i.e., per-edge or per-node) rates. To obtain global rates, we multiply with the number of edges or nodes, respectively, depending on whether the events are node-based or edge-based. This yields the population-wide rates mp, mq, mr and nw.

In simulations run with the state-based processes, we initialise all our nodes with a *null state* to remove any biases from initial conditions. Nodes in that initial state do not actively rewire or spread their state to other nodes until they have been updated with another state. We then wait for a burn-in period until every node has a non-null state before we take measurements on the networks. The distribution of states thus emerges from the model dynamics.

3. Results

In the following, we will present our analysis of the dynamics that result from the interplay between the processes outlined above. We will first focus on a scenario of fixed states, where only the two rewiring processes occur, before turning to scenarios where all four processes can happen.

(*a*) *Fixed states*

When the state of each node is immutable, the only processes affecting the network are homophilous rewiring, with rate p, and random rewiring, with rate q. Here, state can be interpreted as a certain property in a simple biological network, or a relatively fixed property of individuals in a social network, such as relative age or a visible trait. We initialise the model by distributing a given number of states randomly among nodes.

When we run the model global network properties such as clustering coefficient, average shortest path length and modularity stabilise in spite of the ongoing dynamics. Generally, three different scenarios of network topology emerge (see Figure 1) depending on the distribution of states and the relative fraction of homophilous versus random rewiring events,

$$a = \frac{p}{q} \tag{3.1}$$

If a is small, or most rewiring events connect random nodes, the resulting dynamic networks are of Erdős-Rényi type at any point in time, with the usual characteristics of low clustering, short path lengths and low modularity. If a is large, or most rewiring events connect nodes of the same state, groups of nodes sharing the same state form tight communities with only transient connections to the rest of the network. These transient connections, when they come into place, are quickly rewired to again connect nodes of the same state. In that case, while the communities disconnect and reconnect over time, at any specific point in time the network fractures into components of nodes with the same state, with the size of these components depending on the abundance of the corresponding states. These network snapshots possess strong clustering, but since they are disconnected they cannot be associated with meaningful modularity and average path lengths.

Between these two extremes, an intermediate regime of values of a exists, where the networks are formed of tightly-connected groups of the same state, but there is still enough random rewiring to leave the networks connected at any point in time. In that case, the network snapshots have strong clustering, large modularity and average path lengths.

By considering a population-level analogue of the model described in the previous section, we can use mathematical analysis to predict the modularity of the resulting networks in the intermediate regime. Modularity is a measure of how well a network partition reflects the presence of communities, and is given by (Newman, 2006)

$$Q = x - \epsilon \tag{3.2}$$

where x is the proportion of all edges that are within-module edges – that is, those linking nodes in the same module. The factor $\epsilon = \sum_i (d_i/2m)^2$, where d_i is the total degree of nodes in the same module, corrects for the expected number of links between nodes of the same module if the links were placed at random. A simple algorithm for detecting modules then involves the identification of a network partition that maximises Q (Newman and Girvan, 2004). It is worth noting that modularity is not a perfect metric for community structure. It can fail to discriminate between structurally diverse partitions (Good et al., 2010), and using modularity to

detect communities in large graphs has been demonstrated to miss small communities (Fortunato and Barthelemy, 2007). These concerns do not preclude the use of modularity for our purposes: our networks are not so large that the resolution limit becomes a serious concern; also, our networks are artificial, and we are more interested in the level of modularity than the identity of modules.

We can take advantage of the fact that homophilous rewiring creates modules of tightly connected nodes of the same state if a is large enough. The partition that maximises Q will then be similar to one that identifies nodes of the same state in modules. Therefore, we can use the connections to nodes of the same state and to nodes of a different state as proxies for within-module and between-module connections. In other words, we can interpret x to mean the proportion of all edges that are within-state edges, or that link nodes of the same state.

If each node is initialised randomly with one of y states ($0 \ll y \ll n$), the value of ϵ is given by summing over a Poisson distribution,

$$\begin{aligned} \epsilon &= \sum_{i=0}^{\infty} y\mathrm{Pois}(i, n/y) \left(\frac{2im/n}{2m}\right)^2 \\ &= \frac{n+y}{ny} \,. \end{aligned} \tag{3.3}$$

In a similar way ϵ can be calculated for other state distributions. Over a period of time where every link is rewired at least once (which is in the order of $(p+q)^{-1}$), the proportion of within-state edges will converge to approximately $x \approx (p+\epsilon q)/(p+q)$, giving the modularity for the state partition as

$$Q_\mathrm{s} = \frac{p}{p+q}\left(1 - \frac{1}{n} - \frac{1}{y}\right). \tag{3.4}$$

The two processes can thus be used to generate a network that has a partition with modularity given by Q_s. This can be compared with the modularity Q_t of the partition of the same network that uses topological analysis to maximise modularity (e.g., Girvan and Newman, 2002). Since the community structure has been introduced by homophilously increasing the numbers of links between nodes of the same state, with all other links placed randomly, it is unlikely that any topological partition that splits up or combines groups of nodes of the same state could achieve

a greater level of modularity than that found in the state partition. This intuition is confirmed by Figure 2, which shows how the topologically generated partition corresponds to the state partition when the network has topological community structure ($Q_{\mathrm{t}} > 0.4$).

(*b*) *Dynamic states*

In many systems, such as social systems and neural networks, properties of the nodes in the network can be affected by those they interact with (Segev et al., 2003; Palla et al., 2007; Gautreau et al., 2009). For example, in human social systems we tend to form relationships based on an implicit set of criteria such as our interests, political affiliations, socioeconomic status or social norms (McPherson et al., 2001). However, human adaptability means that the criteria can change – either by copying others we interact with, or by innovating new sets of criteria. To reflect this, we relax the immutability of states and introduce the state spread and innovation processes described above. We may then apply our model to such a system by taking node state to represent a set of criteria shared by many people.

We find that, under appropriate parameters, the model shows community structure with several concurrent groups, many of which have relatively long lifetimes (Figure 3). The sizes of the groups, as well as their composition, are dynamic as nodes join and leave them in the close interplay of state changes and edge rewiring (Figure 4). Again, we see that, under a wide range of parameters, some global properties, such as clustering coefficient or network modularity, stabilise as the network keeps evolving. Mathematical analysis (see part(a) of the Appendix) also predicts stability of network modularity and gives a good approximation of the corresponding topological network modularity (as with Figure 2) when the state spread parameters maintain a moderate number of groups (between $n/50$ and $n/3$).

To capture the mutual feedback between state changes and network rewiring, we introduce two more quantities,

$$b = \frac{w}{r}, \tag{3.5}$$

the relative frequency of innovation versus state spread, and

$$c = \frac{p+q}{r+w} \tag{3.6}$$

the relative frequency of rewiring versus state update.

Depending on the model parameters, snapshots of the dynamic networks range from random-like graphs with a single or few dominant states to fragmentation into many small tight-knit communities, each of which share the same state (Figure 3). In an intermediate regime, highly connected communities emerge, each containing mostly the same state, with some interconnections between those communities, similar to what we observed for fixed states (Figure 1). As before, if most rewiring events are homophilous (large a), the network tends have high modularity or even break up into fragments. If, on the other hand, most rewiring events are random (small a), network snapshots resemble random graphs. If rewiring happens on timescales faster than state changes (large c), we tend to see more modular graphs, whereas if state changes are faster (small c), networks are more random. Lastly, the frequency of innovation (regulated by b) largely dictates the number of different states concurrently present in the network, with corresponding second-order effects on the distribution of state prevalence and modularity as communities in the network tend to be smaller if there are many concurrent states (see part (a) of the Appendix).

To characterise the distribution of states at a given moment in time (i.e., how many nodes are in each different state that coexists in a network) we exploit an analogy between our model and the canonical ensemble of statistical physics. This ensemble considers particles in a gas that exchange energy when they collide. In the case of our model, the analogue of particles are the different states, and the equivalent of their energy is the number of nodes that are in that state at a given moment in time. When a state spread event happens, a node in the network changes its state, therefore decreasing the number of nodes in its original state by one and increasing the number of nodes in its new state by one – a process equivalent to the exchange of energy between colliding particles.

If we assume such exchanges of nodes between groups of states to occur com-

pletely randomly, the probability distribution P_i of groups that have i nodes is given by the Boltzmann distribution

$$P_i = \frac{\exp(-iy/n)}{\sum_{i=1}^{n} \exp(-iy/n)} \tag{3.7}$$

Simulations confirm that the state distribution does indeed stabilise (Figure 5). However, while the shape of the distribution remains relatively constant, the identity of groups at a particular rank does not. The ongoing dynamics at the node level causes states to grow and shrink in abundance (Figure 6).

The state distribution we observe in simulations is steeper than that given by Eq. (3.7) (Figure 5). The most abundant state tends to contain a greater number of nodes than predicted, and the less abundant ones fewer. This is because large groups of the same state have more edges linking them to other states, and therefore more possibilities to acquire or lose nodes. If, on the other hand, there is only one node left of a given state it can stay in the network for a long time without being selected for an event, or anything happening to it.

In fact, every state that appears in the network via the innovation process will eventually go extinct due to the inherent stochasticity of the model. This becomes clear when we consider the lifetime distribution of states. In Figure 7, we compare the distribution of change of states in nodes (i.e., the time it takes until the state of a given node changes) with the distribution of lifetimes of states in all nodes (i.e., the time between a state is introduced through innovation and its extinction) where state spread and homophilous rewiring are much more frequent than the randomising processes of innovation and random rewiring. When state spread happens on timescales faster than homophilous rewiring, the changes in network structure resulting from rewiring will be too slow to create a modular structure – one dominant group forms and persists for a long time, while most newly innovated states go extinct quickly. Thus the distribution of node state changes and states largely coincide.

If homophilous rewiring and state spread happen with similar frequencies, both distributions are bimodal. The left mode is a reflection of the more than 50% chance of newly innovated states to go extinct before they are spread to a second node (50%

for the first spreading event involving the node plus a small chance that another innovation will happen in the same node). Some states, however, become established in the modular network, and the corresponding nodes will form a community and subsequently be protected from invasion as they are surrounded by nodes of the same state. This is why both distributions have another mode at longer lifetimes. Note that the curve representing the lifetime of states has a more pronounced tail because states can survive for a long time even if their composition of nodes change. If homophilous rewiring happens on a much faster timescale than state spread, the distributions again become unimodal. This is because innovations are immediately rewired away from, so that there cannot be rapid extinction.

4. Discussion

We have presented a model of dynamic networks in which, over a range of parameters, stable and connected community structure emerges. We have found the presence of such stable community structure to depend largely on the relative frequencies of random to homophilous rewiring. Furthermore, we have provided evidence that a partition of the network according to the state of nodes represents a partition of maximal modularity, and can therefore be used to predict topological modularity. This allowed us to treat modularity analytically, to predict convergent modularity and to quantify its value according to the ratio of random to homophilous rewiring.

The two simple processes of homophily and random rewiring alone can generate community structure reminiscent of that found in the topology of simple, but dynamic, biological networks with units (nodes) having fixed states but dynamic associations (edges). We consider two real-world examples where this is relevant. The first is protein-protein interaction networks where proteins (represented by nodes in our model) often interact (recent or frequent interactions are represented by edges) when they share similar amino-acid sequences (represented by states). This homophilous process can explain community structure found in such networks (Bader and Hogue, 2003; Li et al., 2010). Interestingly, yeast protein interaction data shows how communities in the network match well with actual protein complexes (Li et al., 2010). The second example is the Schelling segregation model, which suggests a mechanism for ghetto formation in humans of different eth-

nic groups (Schelling, 1971). With ethnicity represented by node states, Schelling's model features a rewiring process that only rewires individuals with a high enough proportion of different-state neighbours. This essentially homophilous process leads to a highly-modular network. In our model, the introduction of a random rewiring process means that modularity converges to an equilibrium.

When nodes have dynamic states we see how several groups of the same state can exist concurrently in a population with community structure. While the presence of these groups is relatively stable over time, their composition is not: individuals move between groups such that some groups grow, some groups shrink, and others have a roughly constant size, but a continual turnover in members. The behaviour of this model variant is reminiscent of data showing such dynamics in human social and communication networks (Palla et al., 2005; Newman et al., 2006) and so may eventually provide insights into how the dynamics on these networks are generated. We characterised the stable group size distribution by comparing it to the Boltzmann distribution, exploiting an analogy of our model to particle collisions in statistical physics. We also compared dynamics at different timescales – the relative timescales of state spread and innovation, as well as the relative timescales of processes affecting states and those affecting the network topology. We have characterised the influence of each of these relative timescales on the behaviour of the network dynamics over a wide range of parameters.

While our model can provide some insight into how endogenous processes produce community structure in real-world networks, there are some limitations to its generality. Communities in real systems can be overlapping (Palla et al., 2005), and the association between individual nodes and states may be non-exclusive (Geard and Bullock, 2010), increasing the complexity of both structure and dynamics. Moreover, our model dynamics are biased in that choosing a random edge in the symmetric state spread process increases the frequency with which more highly connected nodes update or spread their states. Other update rules could be argued for, such as degree-based preferential attachment and node-based state spread, each of which would result in different biases.

Future development and validation of our model will require stronger links with data, especially data that are resolved in time. Such data has traditionally been

difficult or costly to obtain, though new sources are becoming available, such as online social communities (Mislove et al., 2007; Lazer et al., 2009). In spite of its limitations, we believe this study to be useful as a baseline to which future models of more specific systems may be compared. We have shown how stable properties can emerge from a highly dynamical system, and focused on modularity, which is a known property of many social and biological systems.

This research was funded by the UK Engineering and Physical Sciences Research Council through standard research grant number EP/D002249/1.

Appendix A. Mathematical treatment

(*a*) *State-based modularity*

We can approximate the behaviour of x (the proportion of links that connect nodes of the same state) under the four processes in our model by making a few simplifying assumptions:

Fixed states

1. *homophilous rewiring* (rate mp) – In the random selection of edges, between-state links are selected with probability $1 - x$, and only in that case does homophilous rewiring take place. Assuming that there is always a node available for rewiring to, the between-state link is replaced with a within-state link. On average, this process thus increases x by $(1 - x)/m$.

2. *random rewiring* (rate mq) – If we assume that all edges created through random rewiring are between-state, we only need to consider events rewiring within-state links (as the ones rewiring between-state links do not change x). Picking within-state links happens with probability x, so this process will on average decrease x by x/m.

Dynamic states

3. *symmetric state spread* (rate mr) – Again, if a between-state link is selected (with probability $1 - x$), it becomes a within-state link. Assuming the node being updated does not have any other links to nodes with its new state, or

that the average degree $d = 2m/n$ is small with respect to the number of states currently in the network, it will on average have xd within-state links that become between-state links. Including the newly added within-state link, this process on average decreases x by $(1-x)(xd-1)/m$.

4. *innovation* (rate nw) – The updated node will have a new state so all its links will become between-state links. A typical node will have xd within-state links, so this process will on average decrease x by xd/m.

We can take all four processes together to give an equation for the temporal evolution of x:

$$\begin{aligned}\dot{x} &= mp(1-x)/m - mqx/m - mr(1-x)(xd-1)/m - nwxd/m \\ &= p(1-x) - qx - r(1-x)(xd-1) - 2wx\end{aligned} \tag{A 1}$$

Note that the process of state spread adds a nonlinearity because both the probability of selecting a between-state link, as well as the amount by that the fraction of between-state links is typically changed by state spread, depend on x itself.

We derive equilibrium values of x by solving $\dot{x} = 0$ in Equation (A 1); these are given by,

$$\tilde{x} = \frac{p+q+r(1+d)+2w \pm \sqrt{(p+q+r(1+d)+2w)^2 - 4rd(p+r)}}{2rd} \,. \tag{A 2}$$

Equilibria are stable if and only if,

$$\tilde{x} < \frac{p+q+r(1+d)+2w}{2rd} \,, \tag{A 3}$$

Substitution of Equation (A 3) into Equation (A 2) shows that unstable equilibria are only found when the $\pm$ term in Eq. (A 2) is positive, and stable equilibria are found when the $\pm$ term in Eq. (A 2) is negative. Algebraic manipulation can be used to show that unstable equilibria can only be found when $\tilde{x} > 1$. Similarly, stable equilibria are always in the region, $0 < \tilde{x} < 1$. This analysis thus shows that for all values of $p, q, r, w, d > 0$, there is always a stable equilibrium for x in the region $0 < \tilde{x} < 1$

Further manipulation can be done to show that $\tilde{x}$ will increase for increasing values of p (done in this case by comparing $\tilde{x}$ for p and $p+\delta$) and decrease for increasing values of q, w and d. When,

$$d > \frac{p+q+2w}{p}$$

$\tilde{x}$ will decrease for increasing values of r.

The prediction given in Equation (A 3) is compared with modularity generated from simulations over a range of parameters in the supplement to this article. Both the modularity of the state partition and the maximum modularity from topological analysis were calculated at several time steps (wide enough apart for the network to significantly change) after the burn in period. The prediction and mean modularities (with standard deviations) are plotted in Figures S1, S2 and S3. In the main, the mathematical prediction is good when there is community structure in the network – but there are small differences due to the correction for within-state links expected by a random rewiring of the network (ϵ) for the modularity measures. These will decrease as the number of nodes increases. The prediction is also good when the number of states is moderate (between $n/50$ and $n/3$).

(*b*) *State distribution*

To find the most likely distribution of states, we use an analogy with the distribution of particle energies in an ideal gas. Similarly to the way particles exchange energy in random collisions, the groups of states in our model exchange nodes. We conjecture that the most likely distribution of states y_i can be found by maximising the number of microstates yielding that distribution (equivalent to minimising the entropy) under the constraints of constant number of states

$$\sum_{i=1}^{n} y_i = y \tag{A 4}$$

and number of nodes

$$\sum_{i=1}^{n} i y_i = n. \tag{A 5}$$

The derivation of the most likely distribution follows the same steps as the derivation of the Maxwell-Boltzmann distribution of statistical physics. The number of microstates yielding a distribution y_1, y_2 $\dots$ y_n is the number of ways to distribute y states among these,

$$\Omega(n, y, \{y_i\}) = y! \prod_{i=1}^{n} \frac{1}{y_i!} \tag{A 6}$$

Maximising $\Omega(n, y, \{y_i\})$ is equivalent to maximising

$$\ln \Omega(n, y, \{y_i\}) = y \ln y - y + \sum_{i=1}^{n} (-y_i \ln y_i + y_i), \tag{A 7}$$

where we used Stirling's formula, $y! \approx y^y e^{-y}$.
We introduce Lagrange multipliers α, β to impose the constraints of constant number of states and particles.

$$\begin{aligned} f(y_i) &= \ln \Omega(n, y, \{y_i\}) + \alpha(y - \sum_{i=1}^{n} y_i) + \beta(n - \sum_{i=1}^{n} i y_i) \\ &= y \ln y - y + \alpha y + \beta n + \sum_{i=1}^{n} (-y_i \ln y_i + y_i - \alpha y_i - \beta i y_i), \end{aligned} \tag{A 8}$$

and maximise $f(y_i)$ by solving

$$\frac{\partial f}{\partial y_i} = -\ln y_i - \alpha - \beta i = 0, \tag{A 9}$$

yielding

$$y_i = e^{-\alpha - \beta i} \tag{A 10}$$

as the distribution that maximises $\Omega(n, y, \{y_i\})$. The first constraint, $\sum y_i = y$ yields

$$e^{-\alpha} = \frac{y}{\sum e^{-\beta i}}, \tag{A 11}$$

so that

$$y_i = y \frac{e^{-\beta i}}{\sum e^{-\beta i}}. \tag{A 12}$$

The second constraint, $\sum iy_i = n$, gives

$$\frac{\sum_{i=1}^{n} ie^{-\beta i}}{\sum_{i=1}^{n} e^{-\beta i}} = \frac{n}{y}. \tag{A 13}$$

To determine β analytically, we make a continuum approximation and replace the sums from 1 to n by integrals from 0 to infinity. This yields

$$\frac{\int_0^\infty ie^{-\beta i} di}{\int_0^\infty e^{-\beta i} di} = \frac{1}{\beta}, \tag{A 14}$$

and $\beta = y/n$ via Eq. (A 13). Putting this back into Eq. (A 12) and setting $P_i = y_i/y$ yields Eq. (3.7).

References

Bader, G. D. and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2.

Bansal, S., Read, J., Pourbohloul, B., and Meyers, L. A. (2010). The dynamic nature of contact networks in infectious disease epidemiology. *J Biol Dyn*, 4(5):478–489.

Brown, R. (2000). *Group Processes*. Oxford: Blackwell, 2nd edition.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. (2001). Mixing beliefs among interacting agents. *Adv. Complex Syst.*, 3:87–98.

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.*, 486:75–174.

Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA*, 104:36–41.

Fu, F. and Wang, L. (2008). Coevolutionary dynamics of opinions and networks: From diversity to uniformity. *Phys. Rev. E*, 78(1):016104.

Funk, S., Salathé, M., and Jansen, V. A. A. (2010). Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J R Soc Interface*, 7(50):1247–1256.

Gautreau, A., Barrat, A., and Barthélemy, M. (2009). Microdynamics in stationary complex networks. *Proc. Natl. Acad. Sci. USA*, 106(22):8847–8852.

Geard, N. and Bullock, S. (2008). Group formation and social evolution: a computational model. In *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems* (Ed. Bullock, S., Noble, J., Watson, R., and Bedau, M.), pp. 197–203. Cambridge: MIT Press.

Geard, N. and Bullock, S. (2010). Competition and the dynamics of group affiliation. *Adv. Complex Syst.*, 13(4):501–517.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826.

Good, B. H., Montjoye, Y.-A. D., and Clauset, A. (2010). The perfomance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106.

Gross, T. and Blasius, B. (2008). Adaptive coevolutionary networks: a review. *J. R. Soc. Interface*, 5(20):259–271.

Gross, T., D'Lima, C. J. D., and Blasius, B. (2006). Epidemic dynamics on an adaptive network. *Phys. Rev. Lett.*, 96(20):208701.

Holme, P. and Newman, M. E. J. (2006). Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E*, 74(5):056108.

Kimura, D. and Hayakawa, Y. (2008). Coevolutionary networks with homophily and heterophily. *Phys. Rev. E*, 78(1):016103.

Kozma, B. and Barrat, A. (2008). Consensus formation on adaptive networks. *Phys. Rev. E*, 77(1):016102.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M. V. (2009). Computational social science. *Science*, 323:721–723.

Li, X., Wu, M., Kwohl, C.-K., and Ng, S.-K. (2010). Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, 11(Suppl 1):S3.

Lusseau, D. and Newman, M. E. J. (2004). Identifying the role that animals play in their social networks. *Proc. Roy. Soc. B*, 271 Suppl 6:S477–S481.

McPherson, J. M., Smith-Lovin, L., and Cook, J. (2001). Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.*, 27:415–444.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07)*, pp. 29–42. Association for Computing Machinery.

Newman, M., Barabási, A.-L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446:664–667.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Not. Amer. Math. Soc*, 56:1082–1097.

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Brief Bioinform*, 11(1):15–29.

Salathé, M. and Jones, J. H. (2010). Dynamics and control of diseases in networks with community structure. *PLoS Computational Biology*, 6(4):e1000736.

Schaeffer, S. E. (2007). Graph clustering. *Comp. Sci. Rev.*, 1:27–64.

Schelling, T. C. (1971). Dynamic models of segregation. *J. Math. Sociol.*, 1:143–186.

Segev, R., Benveniste, M., Shapira, Y., and Ben-Jacob, E. (2003). Formation of electrically active clusterized neural networks. *Phys. Rev. Lett.*, 90(16):168101.

Stafford, R. (2006). Random vectors with fixed sum. [Online]. http://www.mathworks.com/matlabcentral/fileexchange/9700.

van Segbroeck, S., Santos, F. C., and Pacheco, J. M. (2010). Adaptive contact networks change effective disease infectiousness and dynamics. *PLoS Comput Biol*, 6(8).

Vazquez, F., Eguíluz, V. M., and Miguel, M. S. (2008). Generic absorbing transition in coevolution dynamics. *Phys. Rev. Lett.*, 100(10):108702.

Volz, E. and Meyers, L. A. (2009). Epidemic thresholds in dynamic contact networks. *J R Soc Interface*, 6(32):233–241.

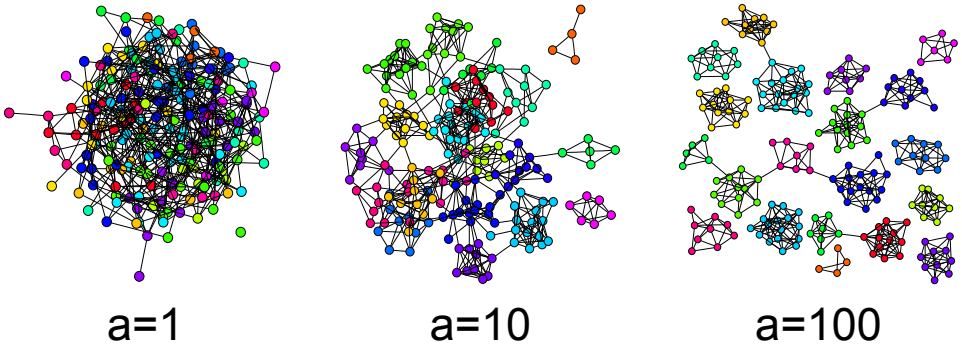


Figure 1: Network snapshots for different values of a (where $a = p/q$) when no state update occurs (i.e., $r = w = 0$). Different colours indicate different states. Three classes of stable system behaviour can be distinguished: (I) When the rate of random rewiring is high with respect to random rewiring (e.g., $a = 1$), network topology is random; (II) When the rate of random rewiring is low (e.g., $a = 0.01$), the network fractures into a set of disconnected, homogeneous components; (III) When homophilous and random rewiring are balanced (e.g., $a = 0.1$), densely connected homogeneous state groups are evident, but the network as a whole also remains connected.

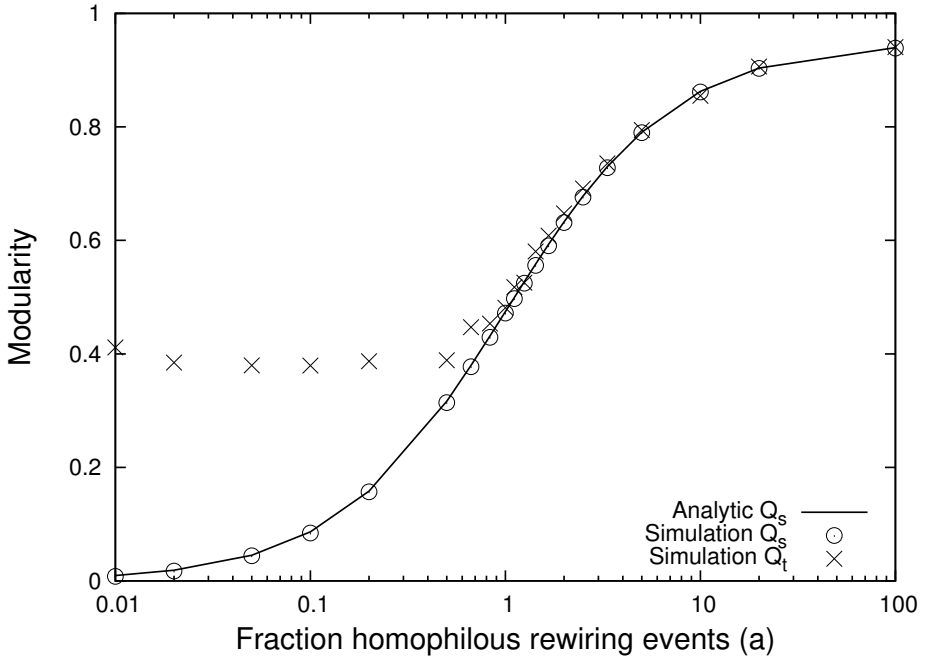


Figure 2: Modularity based on maximal topological modularity as given by the Girvan-Newman algorithm (Q_{t}) as measured in simulations (crosses), and as given by our algorithm identifying modules based on state (Q_{s}), as predicted analytically (line) and measured in simulations (circle), in terms of the fraction of rewiring events that are homophilous, $a = p/q$.

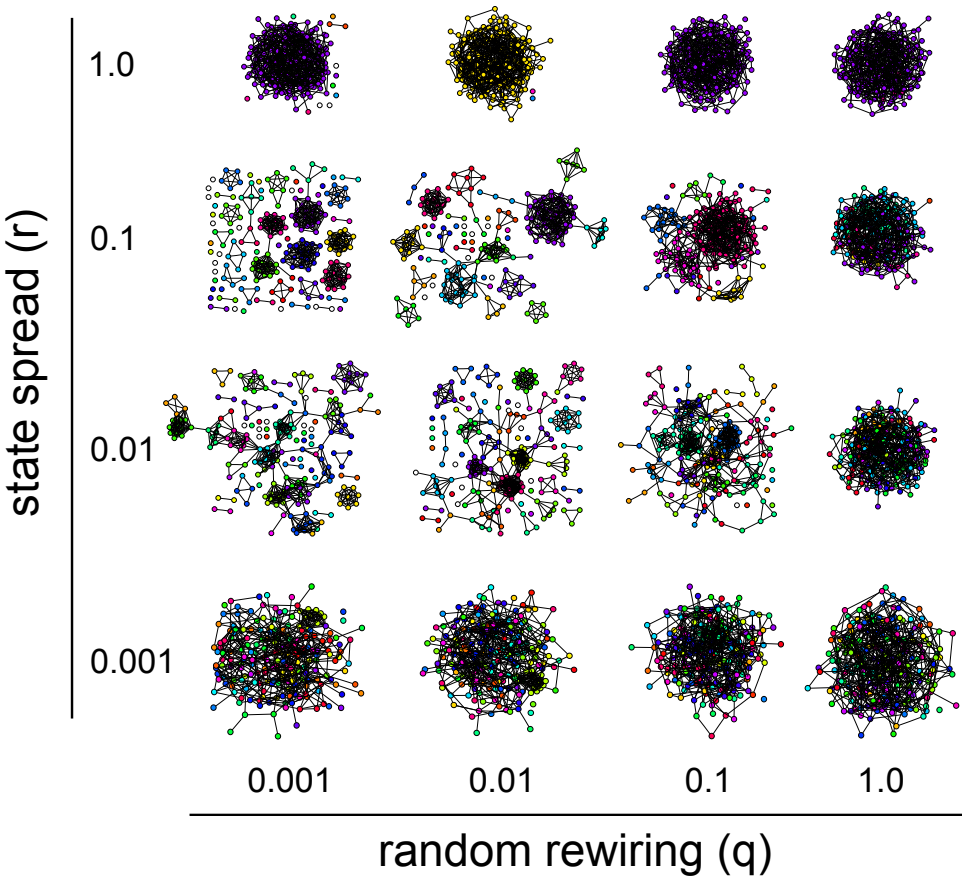


Figure 3: Network snapshots for different rates of state spread (r) and random rewiring (q) ($p = 1$ and $w = 0.001$). Snapshots were taken at $t = 5 \times 10^6$, to ensure that any transient dynamics had passed. Different colours indicate different states. Again, three classes of stable system behaviour can be distinguished: (I) Random network topologies result not only when the rate of random rewiring is high ($q = 1$), but also when the rate of state spread is either very low or very high. In the former case, the absence of state spread inhibits the organising tendencies of homophilous rewiring; in the latter case, a single group rapidly establishes itself and dominates the population, in which case homophilous rewiring becomes effectively equivalent to random rewiring. (II) When the rate of random rewiring is low and there is a moderate level of state spread (e.g., $r = 0.001; q = 0.1$), the network fractures into a set of disconnected, homogeneous components. (III) With intermediate levels of both state spread and random rewiring (e.g., $r = 0.01; q = 0.01$), densely connected homogeneous state groups are evident, but the network as a whole also remains connected.

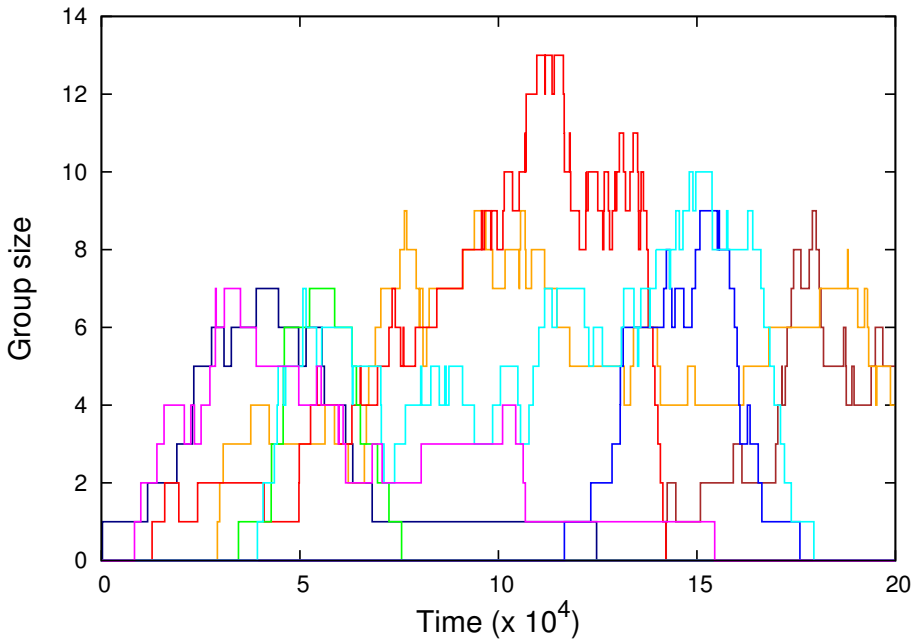


Figure 4: An illustration of the evolution of state groups. This figure plots the size of eight different state groups over 200,000 time steps ($p = 1; q = r = w = 0.01$). The eight state groups shown (of a total of 57 that existed at some point during the simulation run) were each the largest in the population at some point in time.

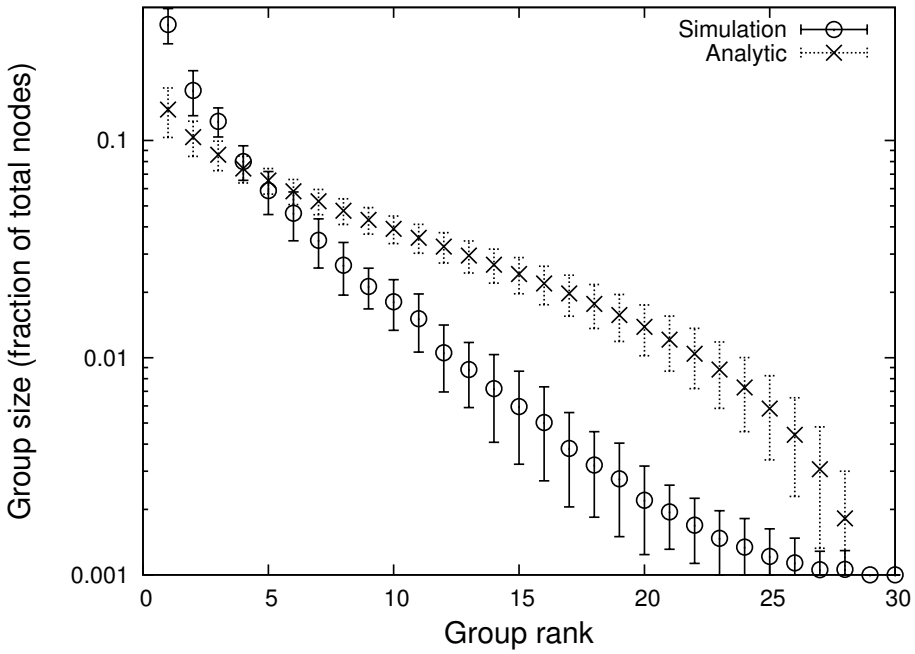


Figure 5: Size distribution of state groups. Shown is the mean size of the ith largest group across 20 snapshots from a simulation run (circles; $a = 100; b = 0.001; c = 0.3$), error bars indicating one standard deviation. Also shown is the distribution as predicted by Eq. (3.7) (crosses), obtained by sampling from $y = 28$ random numbers summing up to $n = 1000$, using the algorithm of Stafford (2006), until convergence was obtained. Despite the continually changing composition of state groups in a population (Figure 4), distribution of group sizes is relatively stable over time.

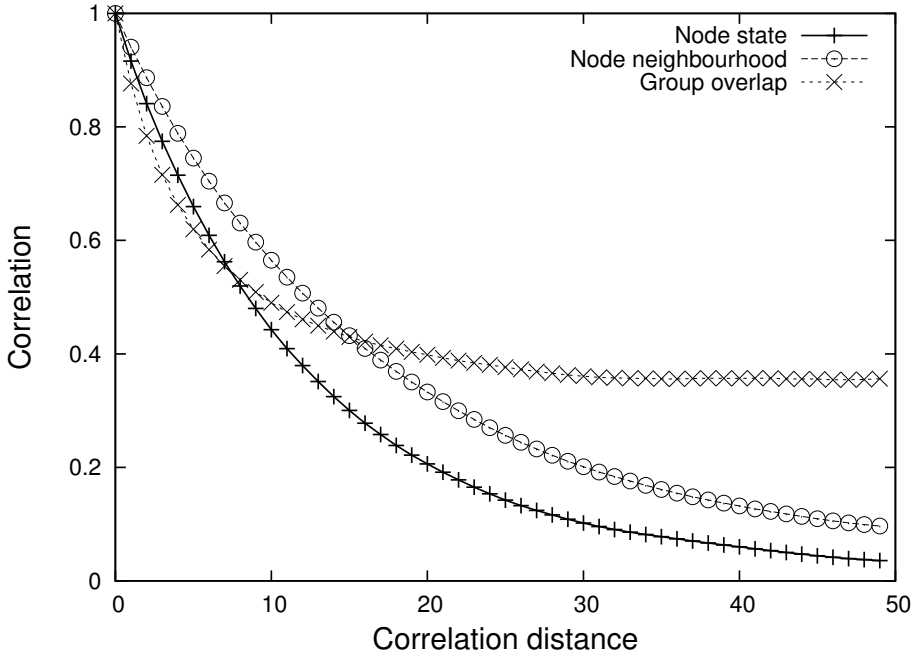
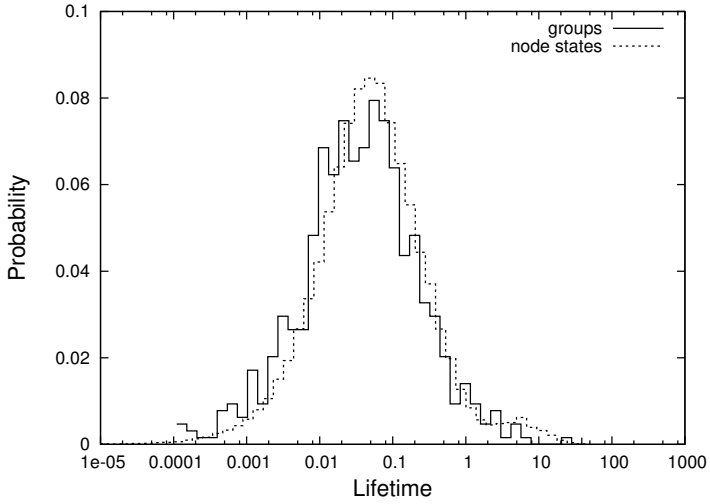
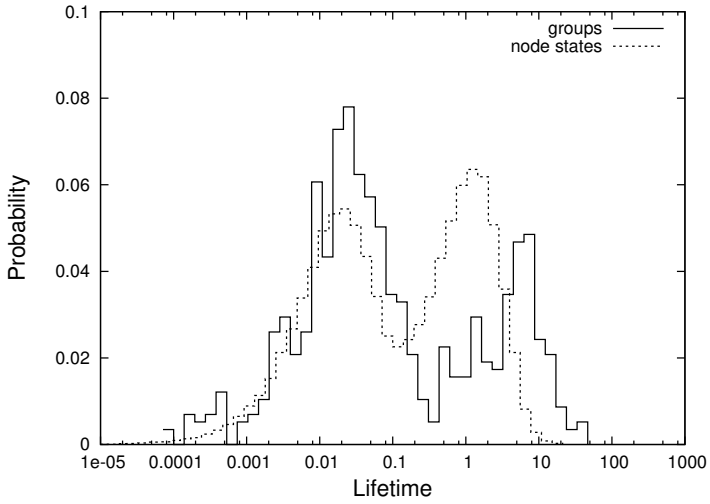


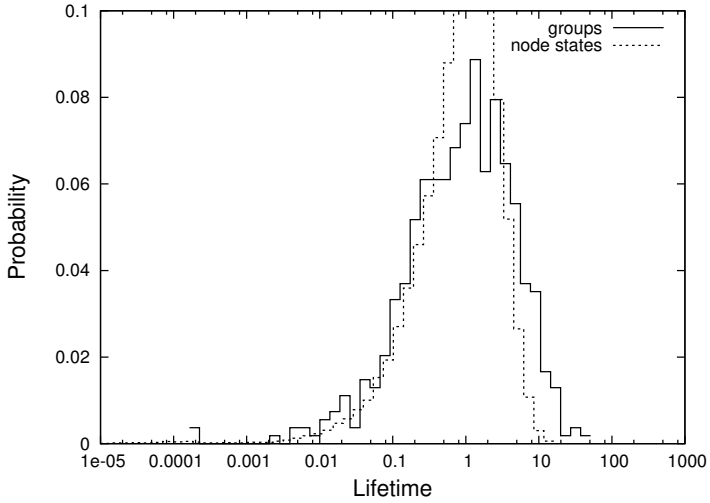
Figure 6: Autocorrelation measures for node and state group properties ($p = 1.0; q = r = w = 0.01$). Node state measures the fraction of nodes that are in the same state at time $t + d$ as they were at time t. Node neighbourhood measures the fraction of node pairs that are neighbours at time $t + d$ that were also neighbours at time t. Group overlap measures the relative overlap in group membership between time t and time $t+d$. Note that all three measures drop rapidly with initial increases in correlation distance; thereafter, some correlation remains at the group level, while node-level correlation drops close to zero.



(a) $a = 10^2; b = 10^{-3}; c = 10^{-3}$



(b) $a = 10^2; b = 10^{-3}; c = 10^{-1.5}$



(c) $a = 10^{1.5}; b = 10^{-3}; c = 10^3$

Figure 7: Distribution of the times it takes until a node changes its state (dashed line), and distribution of the total lifetimes of states from first innovation until they go extinct (solid line) for three different sets of parameters representing different relative timescales of state spread and homophilous rewiring: (a) fast state spread, (b) similar timescales, (c) fast rewiring.

Supplementary Information

J. Bryden, S. Funk, N. Geard, S. Bullock, V.A.A. Jansen

September 27, 2010

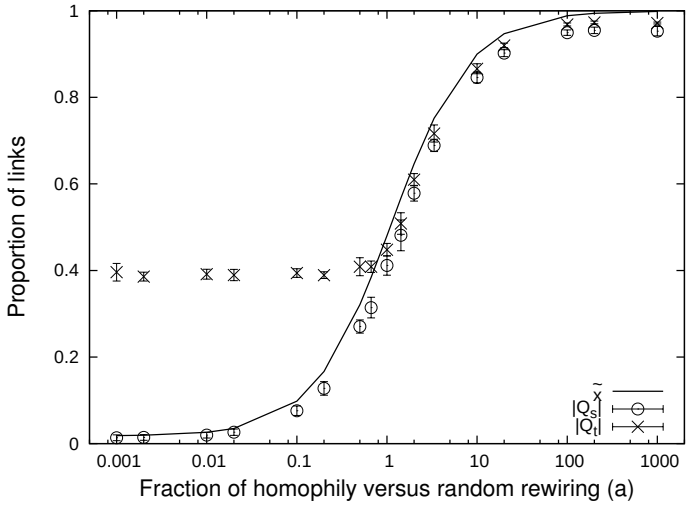


Figure S1: The relative frequency of homophilous rewiring to random rewiring, when state processes also happen ($b = 0.01$ and $c = 50$). The difference between the mathematical prediction $\tilde{x}$ of edges connecting nodes of the same state (line) and the modularities found in simulations based on node state (Q_s, circles) and topological analysis (Q_t, crosses) arises because in the mathematical analysis we do not account for within-state links created by random rewiring of the network (ϵ). Other parameters, $n = 1000$ and $m = 3000$.

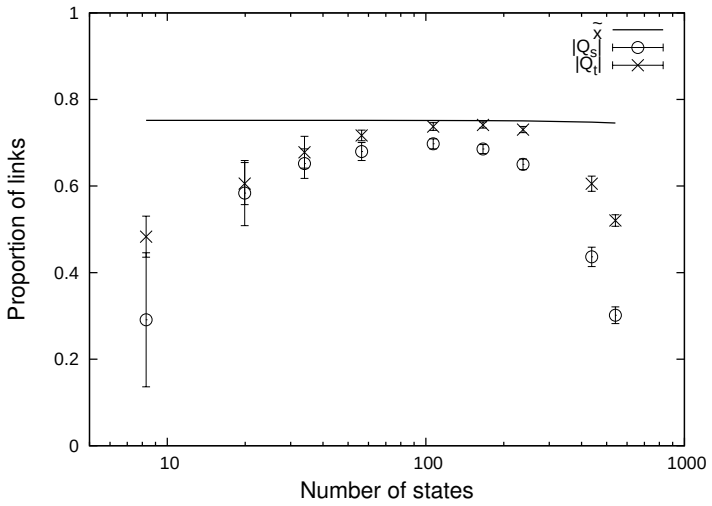


Figure S2: Changes to the relative frequency of innovation to state spread (increasing b), also changes the number of states existing contemporaneously. Shown is the mathematical prediction for the fraction $\tilde{x}$ of edges connecting nodes of the same state (line), as well as modularity found by simulations based on node state (Q_{s}, circles) or topological analysis (Q_{t}, crosses). When b is too large or too small (at the left and right of the graph), the network becomes to a random-like network at any given time. Other parameters, $n = 1000$, $m = 3000$, $a = 3.33$, $c = 50$, and b ranges from 0.001 on the left to 1 on the right of the figure.

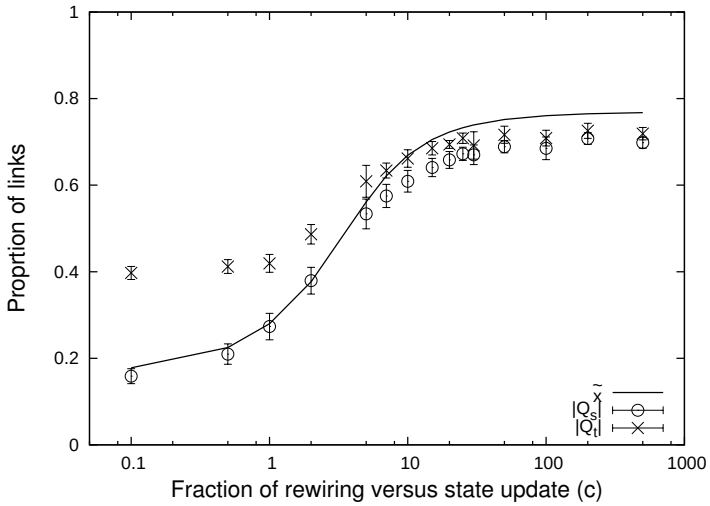


Figure S3: When varying the relative frequency of rewiring to state update, the mathematical prediction for the fraction $\tilde{x}$ of edges connecting nodes of the same state (line) is largely similar to the modularity found by simulations based on node state (Q_{s}, circles) or topological analysis (Q_{t}, crosses). When state spread is less frequent ($c > 1$), the difference between the mathematical prediction and the modularities found in simulations arises because in the mathematical analysis we do not account for within-state links created by random rewiring of the network (ϵ). When state spread is more frequent ($c < 1$) the network becomes a random-like network at any time, and the topological algorithm will find a partition with greater modularity than the state partition. Other parameters, $n = 1000$, $m = 3000$, $a = 3.33$ and $b = 0.01$.