J. Appl. Prob. **47,** 572–585 (2010)
Printed in England

PREFERENTIAL DUPLICATION GRAPHS

NETTA COHEN,* *University of Leeds*
JONATHAN JORDAN,** *University of Sheffield*
MARGARITIS VOLIOTIS,* *University of Leeds*

Abstract

We consider a preferential duplication model for growing random graphs, extending previous models of duplication graphs by selecting the vertex to be duplicated with probability proportional to its degree. We show that a special case of this model can be analysed using the same stochastic approximation as for vertex-reinforced random walks, and show that 'trapping' behaviour can occur, such that the descendants of a particular group of initial vertices come to dominate the graph.

Keywords: Preferential duplication graphs; stochastic approximation; vertex-reinforced random walk

2010 Mathematics Subject Classification: Primary 05C80
Secondary 60G99; 60K35

1. Introduction

Many naturally occurring networks ranging from subcellular biological networks to a variety of social networks are believed to grow by processes of vertex duplication. Indeed, graph growing models based on vertex duplications have been the subject of investigation over recent years (see, for example, [1], [5], [6], [7], and [10]). In most of these models vertices are chosen for duplication according to a uniform distribution, while in the model of [5] all vertices are duplicated simultaneously. By contrast, other graph growing algorithms rely on the preferential attachment of new vertices to highly connected existing ones to reproduce the broad, often scale-free, degree distributions (see [4] for a survey) found in many real-world networks. Here we present a generalisation of a duplication graph growing algorithm that is inspired by a 'friend-brings-a-friend' growth process, and reduces to the preferential attachment model in one limit. We call this growth algorithm preferential duplication.

Our model is defined as follows. Let G_0 be a finite (connected) graph with n_0 vertices (labelled with the integers $1, \dots, n_0$). There are two versions of the model, which we will call the 'false twins' version and the 'true twins' version.

In both models, we construct a sequence of graphs $(G_n)_{n\in\mathbb{N}}$ by a procedure which, to construct G_{n+1} from G_n, chooses a vertex v_{n+1} of G_n with probability proportional to its degree (that is, $v_{n+1} = v$ with probability

$$\frac{\deg_{G_n}(v)}{\sum_{w\in V(G_n)} \deg_{G_n}(w)},$$

Received 6 July 2009; revision received 17 March 2010.
* Postal address: School of Computing, University of Leeds, Leeds, LS2 9JT, UK.
** Postal address: Department of Probability and Statistics, University of Sheffield, Hicks Building, Sheffield, S3 7RH, UK. Email address: jonathan.jordan@shef.ac.uk

as in the preferential attachment graph), and duplicates v_{n+1} together with each of its edges with probability p, independently of each other. That is, if $v_{n+1} = v$, a new vertex v' is added to the graph. An edge exists in G_{n+1} between v' and w with probability p if an edge existed in G_n between v and w, and not otherwise; the existence of edges from v' to different neighbours of v_{n+1} is independent. Additionally, in the 'true twins' version of the model only, v' is connected to v with probability 1.

In the remainder of this paper we concentrate on the case $p = 1$; we intend to investigate the case $p < 1$ in a later paper. We note that the 'true twins' model with $p = 0$ becomes the preferential attachment model of [4] with the parameter $m = 1$.

The behaviour of preferential duplication graphs bears some resemblance to that of a completely different graph model—that of vertex-reinforced random walks (VRRWs; see [3], [8], and [11]). For a specific duplication event, with $p = 1$, the vertex and associated edges duplicated are simply a reproduction of the existing structure of the duplicated vertex. Therefore, it is convenient to collapse the new vertex and its associated edges onto the duplicated vertex and edges, while 'reinforcing' that vertex to indicate that such a collapse has taken place. This process would then be identical to the reinforcement in VRRWs. However, in the former, vertices are selected for duplication from the graph only as a function of their degree, whereas in the latter there is an additional constraint that any two successively reinforced vertices must be neighbours.

We show, in the special case of the preferential duplication model where $p = 1$, that a process representing the numbers of descendants of the original vertices becomes 'trapped' on certain subgraphs of the initial graph. In other words, given an initial graph G_0, we can find subgraphs of G_0 such that there is a positive probability that after a sufficient number of generations, all new vertices are descendants of vertices in G_0. Interestingly, such trapping is of the same nature as that which has been shown to occur in VRRWs; our proof method is to show that preferential duplication with $p = 1$ has a stochastic approximation equation linking it to the same dynamical system as is linked to VRRWs in [3] and [8]. We conjecture that, with probability 1, there exists one subgraph that traps the process.

In the 'false twins' case we show in Theorem 2.2 that the trapping subgraphs are of the form $S \cup B$, where S is a complete m-partite subgraph of G_0 satisfying certain conditions and B consists of those vertices with a neighbour in S. These trapping subgraphs are the same as those found in a VRRW on the initial graph G_0. In the 'true twins' case we show in Theorem 2.3 that the trapping subgraphs are of the form $S \cup B$, where S is a maximal clique of G_0 and B again consists of those vertices with a neighbour in S. In both cases any trapping subgraph has positive probability of trapping the process.

The trapping behaviour we find may give insight into the emergence of certain types of structure in a variety of complex systems with similar growth properties. In particular, the principle of a 'friend-brings-a-friend' is commonly used in real-world networks (though rarely with $p = 1$). Where such growth rules indeed lead to trapping behaviour, we would find that ancestry may be affected more by nuances such as the specific structure of G_0 than merely the distinctive features of a particular ancestor. Another interesting phenomenon is the symmetry breaking that can occur in this model, with an initial symmetric graph G_0 that has more than one trapping subgraph. More generally, the fact that trapping appears in two very different models of graph processes suggests that it may be a more universal property of particular classes of systems than previously known.

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

2. The case $p = 1$

We show that the case where $p = 1$ (so the new vertex is an exact copy of the vertex it was duplicated from) is closely related to a VRRW (see [3], [8], and [11] for more on VRRWs) on the initial graph.

2.1. Preliminaries and notation

Lemma 2.1. *In the 'false twins' case, all vertices of G_n will have the same set of neighbours in G_n as one of the initial vertices $1, \dots, n_0$.*

Proof. The new vertex added to G_n to form G_{n+1} has the same set of neighbouring vertices in G_{n+1} as the vertex it is a duplicate of. Hence, if two vertices have the same set of neighbours in G_n then they will continue to do so in G_m ($m > n$), and so all vertices of G_n will have the same set of neighbours in G_n as one of the initial vertices $1, \dots, n_0$.

This will also hold in the 'true twins' case if the set of vertices within distance 1 of each vertex (i.e. including the vertex itself) is considered.

We will describe a vertex as being descended from an initial vertex i if it was either duplicated directly from vertex i or duplicated from a vertex descended from vertex i. Hence, all vertices descended from vertex i have the same set of neighbours in G_n as vertex i.

For $n \geq 0$ and $1 \leq i \leq n_0$, let $d_i^{(n)}$ be the degree of vertex i in G_n, and let $c_i^{(n)}$ be the number of vertices of G_n which are descended from vertex i (including vertex i itself). Let X_n be a random variable taking values in $\{1, \dots, n_0\}$, whose value is the original vertex that v_n is descended from, so that $c_{X_{n+1}}^{(n+1)} = c_{X_{n+1}}^{(n)} + 1$ and $c_i^{(n+1)} = c_i^{(n)}$ for $i \neq X_{n+1}$.

Let $x_i^{(n)}$ be the proportion of vertices of G_n descended from vertex i,

$$x_i^{(n)} = \frac{c_i^{(n)}}{n + n_0},$$

and let $x^{(n)}$ be the vector of proportions, $x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \dots, x_{n_0}^{(n)})$, which can be regarded as an element of the $(n_0 - 1)$-dimensional simplex

$$\Delta^{n_0-1} = \left\{ x \in \mathbb{R}^{n_0-1}; x_i \geq 0 \text{ for all } i, \sum_{i=1}^{n_0-1} x_i \leq 1 \right\}.$$

Let $A = (a_{ij})_{i,j \in V(G_0)}$ be the adjacency matrix of G_0, and define a σ-algebra

$$\mathcal{F}_n = \sigma(G_0, G_1, \dots, G_n).$$

For the 'false twins' case, let $f(x) \colon \mathbb{R}^{n_0} \to \mathbb{R}^{n_0}$ be a function with coordinates given by

$$f_i(x) = \frac{x_i \sum_{j=1}^{n_0} a_{ij} x_j}{\sum_{k=1}^{n_0} x_k \sum_{j=1}^{n_0} a_{kj} x_j} = \frac{x_i (Ax)_i}{x^\top A x},$$

and let $F(x) = f(x) - x$. Following [3], for $x \in \Delta^{n_0-1}$, we write $N_i(x) = (Ax)_i$ and $H(x) = \sum_{i=1}^{n_0} x_i N_i(x) = x^\top A x$, so that

$$f_i(x) = \frac{x_i N_i(x)}{H(x)} \quad \text{and} \quad F_i(x) = \frac{x_i (N_i(x) - H(x))}{H(x)}.$$

Similarly, for the 'true twins' case, define $\hat{N}_i(x) = [(A + I)x]_i$ (where I is the $n_0 \times n_0$ identity matrix), $\hat{H}(x) = x^\top (A + I)x$, $\hat{f}_i(x) = x_i \hat{N}_i(x)/\hat{H}(x)$, and $\hat{F}(x) = \hat{f}(x) - x$.

For $i \in \{1, \dots, n_0\}$, let $\iota(i)$ be the unit vector in $\mathbb{R}^{n_0}$ with 1 in position i and 0s elsewhere.

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

2.2. Stochastic approximation and analysis of attractors

Theorem 2.1. *In the 'false twins' case, the sequence* $(x^{(n)})_{n\in\mathbb{N}}$ *satisfies the stochastic approximation equation*

$$x^{(n+1)} - x^{(n)} = \frac{1}{n+n_0+1} F(x^{(n)}) + \varepsilon^{(n+1)},$$

with $\mathrm{E}(\varepsilon^{(n+1)} \mid \mathcal{F}_n) = 0$ *and where* $\varepsilon^{(n+1)}(n+n_0+1)$ *is bounded.*

In the 'true twins' case, the sequence satisfies a different stochastic approximation equation,

$$x^{(n+1)} - x^{(n)} = \frac{1}{n+n_0+1} \hat{F}(x^{(n)}) + \varepsilon^{(n+1)} + R^{(n+1)},$$

with $\mathrm{E}(\varepsilon^{(n+1)} \mid \mathcal{F}_n) = 0$, *where* $\varepsilon^{(n+1)}(n+n_0+1)$ *is bounded and the remainder term* $R^{(n)} = O(n^{-2})$.

Proof. The graph G_n has $n+n_0$ vertices. The new vertex v_{n+1} will be descended from vertex i if and only if the selected vertex v_{n+1} is. Hence, in the 'false twins' case,

$$c_i^{(n+1)} = \begin{cases} c_i^{(n)} + 1 & \text{with probability } \dfrac{c_i^{(n)} d_i^{(n)}}{\sum_{k=1}^{n_0} c_k^{(n)} d_k^{(n)}}, \\ c_i^{(n)} & \text{with probability } 1 - \dfrac{c_i^{(n)} d_i^{(n)}}{\sum_{k=1}^{n_0} c_k^{(n)} d_k^{(n)}}. \end{cases}$$

Let a_{ij}, $1 \le i, j \le n_0$, be the entries of the adjacency matrix of G_0. Then $d_i^{(n)} = \sum_{j=1}^{n_0} a_{ij} c_j^{(n)}$, so we can rewrite the above as

$$c_i^{(n+1)} = \begin{cases} c_i^{(n)} + 1 & \text{with probability } \dfrac{c_i^{(n)} \sum_{j=1}^{n_0} a_{ij} c_j^{(n)}}{\sum_{k=1}^{n_0} c_k^{(n)} \sum_{j=1}^{n_0} a_{kj} c_j^{(n)}}, \\ c_i^{(n)} & \text{with probability } 1 - \dfrac{c_i^{(n)} \sum_{j=1}^{n_0} a_{ij} c_j^{(n)}}{\sum_{k=1}^{n_0} c_k^{(n)} \sum_{j=1}^{n_0} a_{kj} c_j^{(n)}}. \end{cases}$$

Hence, this can be treated as a generalisation of the Pólya urn model where category i is chosen with probability

$$\frac{c_i^{(n)} \sum_{j=1}^{n_0} a_{ij} c_j^{(n)}}{\sum_{k=1}^{n_0} c_k^{(n)} \sum_{j=1}^{n_0} a_{kj} c_j^{(n)}} = f_i(x^{(n)})$$

instead of with probability simply proportional to $c_i^{(n)}$.

We see that

$$x_i^{(n+1)} - x_i^{(n)} = \frac{1}{n+n_0+1}(1 - x_i^{(n)})\, \mathbf{1}_{\{X_{n+1}=i\}} - \frac{1}{n+n_0+1} x_i^{(n)} (1 - \mathbf{1}_{\{X_{n+1}=i\}}),$$

and, hence,

$$x^{(n+1)} - x^{(n)} = \frac{1}{n+n_0+1}(\iota(X_{n+1}) - x^{(n)}).$$

Taking conditional expectations,

$$\mathrm{E}(x^{(n+1)} - x^{(n)} \mid \mathcal{F}_n) = \frac{1}{n+n_0+1}(f(x^{(n)}) - x^{(n)}).$$

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

Letting

$$\varepsilon^{(n+1)} = \frac{1}{n+n_0+1}(\iota(X_{n+1}) - f(x^{(n)})),$$

we conclude that $\mathrm{E}(\varepsilon^{(n+1)} \mid \mathcal{F}_n) = 0$ and that $\varepsilon^{(n+1)}(n+n_0+1)$ is bounded, giving the result.

In the 'true twins' case,

$$d_i^{(n)} = \sum_{j=1}^{n_0} a_{ij} c_j^{(n)} + c_i^{(n)} - 1.$$

So the probability that some vertex descended from vertex i is chosen for duplication is

$$\begin{aligned} p_i^{(n)} &= \frac{c_i^{(n)}\left(\sum_{j=1}^{n_0} a_{ij} c_j^{(n)} + c_i^{(n)} - 1\right)}{\sum_{k=1}^{n_0} c_k^{(n)} \sum_{j=1}^{n_0} (a_{kj} c_j^{(n)} + c_k^{(n)} - 1)} \\ &= \frac{x_i^{(n)}\left(\sum_{j=1}^{n_0} a_{ij} x_j^{(n)} + x_i^{(n)} - 1/(n+n_0)\right)}{\sum_{k=1}^{n_0} x_k^{(n)} \sum_{j=1}^{n_0} (a_{kj} x_j^{(n)} + x_k^{(n)} - 1/(n+n_0))}, \end{aligned}$$

so

$$\begin{aligned} p_i^{(n)} &= \frac{x_i^{(n)}(\hat{N}_i(x^{(n)}) - 1/(n+n_0))}{\hat{H}(x^{(n)}) - 1/(n+n_0)} \\ &= \frac{x_i^{(n)} \hat{N}_i(x^{(n)})}{\hat{H}(x^{(n)})} + \frac{x_i^{(n)}(\hat{N}_i(x^{(n)}) - \hat{H}(x^{(n)}))}{\hat{H}(x^{(n)})(\hat{H}(x^{(n)})(n+n_0) - 1)} \\ &= \hat{f}_i(x^{(n)}) + O(n^{-1}). \end{aligned}$$

Hence, stochastic approximation theory [2], [9] relates the behaviour of $x^{(n)}$ to the behaviour of the continuous-time dynamical system given by the function $F(x)$. This is the same stochastic approximation as occurs for vertex-reinforced random walks in [3] and [8].

An *attractor* for a flow Φ on a metric space (M, d) is defined (e.g. in [2]) to be a subset $A \subseteq M$ which is invariant under Φ and has a neighbourhood W such that $d(\Phi_t x, A) \to 0$ as $t \to \infty$ uniformly for $x \in W$.

In [3], a *stable equilibrium* for the dynamical system of interest is defined as being an equilibrium where all eigenvalues of the Jacobian matrix are nonpositive. This does not necessarily imply that the stable equilibrium is in an attractor.

As H is a Lyapunov function for the stochastic approximation, and letting $L(x)$ be the limit set of the process $(x^{(n)})_{n\in\mathbb{N}}$, we can conclude the following.

Corollary 2.1. *If A is an attractor of the continuous-time dynamical system given by the function $F(x)$ (or $\hat{F}(x)$), then $\mathrm{P}(L(x) \subseteq A) > 0$. Furthermore, $L(x)$ consists of equilibria for the dynamical system, and $H(x^{(n)})$ (or $\hat{H}(x^{(n)})$) converges as $n \to \infty$.*

Proof. The first statement follows from Theorem 7.3 of [2] and the second statement follows from Proposition 6.4 of [2].

We now discuss the attractors and stable equilibria for the dynamical system in the 'false twins' case. First consider the case where G_0 is a complete m-partite graph. In this case vertices which belong to the same part have the same neighbours and so are indistinguishable, so we can just consider the case where G_0 is a complete graph on m vertices. The convergence of the stochastic approximation in this case is discussed in [8].

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

We now consider finding attractors for the dynamical system in more general graphs. Consider the case where G_0 contains a subgraph consisting of a 'core' S and its outer boundary B, S being a complete m-partite graph, $S = V_1 \cup V_2 \cup \cdots \cup V_m$, and B consisting of exactly those vertices which are outside S but have a neighbour in S. We have already dealt with the case where S is the whole graph above, so without loss of generality assume that $n_0 \notin S$. In [11] such a subgraph is defined to be a *trapping subgraph* if, for any vertex v in B, two criteria are met:

1. there is at least one part of S, V_i, such that v is not connected to V_i;
2. there is at least one vertex in $x' \in S \setminus V_i$ such that v is not connected to x'.

The following result shows that trapping subgraphs (under the definition in [11]) produce attractors of the dynamical system.

Proposition 2.1. *Let S be the core of a trapping subgraph, and let $y \in \Delta^{n_0-1}$ be such that $\sum_{i\in V_j} y_i = 1/m$ and that $y_i = 0$ if $i \notin S$, i.e. y represents a proportion $1/m$ of the vertices being in each part of the m-partite graph, and a proportion 0 outside the graph. The set of points of this form is an attractor for the dynamical system driven by F.*

Proof. It is fairly easy to see that y is a fixed point of F: the definition implies that $H(y) = (m-1)/m$ and that $N_i(y) = H(y)$ for all $i \in S$. It remains to prove that the set of fixed points of this form is an attractor for the dynamical system. To do this, we will evaluate the partial derivatives $d_{ij} = \partial F_i/\partial x_j$ at the fixed point y, and show that the eigenvalues of the resulting Jacobian are at most 0, and that the eigenspace corresponding to the eigenvalue 0 is contained within the set of fixed points.

Using the fact that $\sum_{i=1}^{n_0} x_i^{(n)} = 1$, we write $x_{n_0} = 1 - \sum_{i=1}^{n_0-1} x_i$ and treat f as a function from Δ^{n_0-1} to itself. Hence, we can rewrite the components of f:

$$f_i(x) = \frac{x_i \sum_{j=1}^{n_0-1}(a_{ij} - a_{i,n_0})x_j + x_i a_{i,n_0}}{\sum_{k=1}^{n_0-1}\left(\sum_{j=1}^{n_0-1}(a_{kj} - 2a_{k,n_0})x_k x_j + 2a_{k,n_0}x_k\right)}.$$

(We assume that G_0 has no loops, so $a_{ii} = 0$ for all i.)

Differentiating and substituting the above values for the x_i at the fixed point, we find that

$$d_{ij} = \left.\frac{\partial F_i}{\partial x_j}\right|_{x=y}$$
$$= \begin{cases} -\dfrac{ma_{i,n_0}y_i}{m-1} - 2y_i + \dfrac{2my_i}{m-1}\displaystyle\sum_{k\in S} a_{k,n_0}y_k, & i = j \in S, \\ \dfrac{ma_{ij}y_i}{m-1} - \dfrac{ma_{i,n_0}y_i}{m-1} - 2y_i + \dfrac{2my_i}{m-1}\displaystyle\sum_{k\in S} a_{k,n_0}y_k, & i \neq j,\ i, j \in S, \\ \dfrac{ma_{ij}y_i}{m-1} - \dfrac{ma_{i,n_0}y_i}{m-1} - \dfrac{2my_i}{m-1}\displaystyle\sum_{k\in S} a_{kj}y_k + \dfrac{2my_i}{(m-1)}\displaystyle\sum_{k\in S} a_{k,n_0}y_k, & i \in S,\ j \notin S, \\ \dfrac{m}{m-1}\displaystyle\sum_{k\in S} a_{ik}y_k - 1, & i = j \notin S, \\ 0, & i \notin S,\ j \neq i. \end{cases}$$

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

The off-diagonal zero entries where $i \notin S$ mean that the eigenvalues of this matrix are the eigenvalues of the matrix obtained by restricting to the rows and columns corresponding to S, together with the diagonal entries

$$\eta_i = \frac{m}{m-1}\sum_{k\in S} a_{ik}y_k - 1, \qquad i \notin S.$$

So we need to find the eigenvalues of the $|S| \times |S|$ matrix D with entries

$$d_{ij} = \left.\frac{\partial F_i}{\partial x_j}\right|_{x=y} = \begin{cases} -\dfrac{ma_{i,n_0}y_i}{m-1} - 2y_i + \dfrac{2my_i}{m-1}\displaystyle\sum_{k\in S} a_{k,n_0}y_k, & i = j \in S, \\ \dfrac{ma_{ij}y_i}{m-1} - \dfrac{ma_{i,n_0}y_i}{m-1} - 2y_i + \dfrac{2my_i}{(m-1)}\displaystyle\sum_{k\in S} a_{k,n_0}y_k, & i \neq j,\ i, j \in S. \end{cases}$$

Label the parts of the complete m-partite graph S as $1, \ldots, m$, and let $p(i)$ be the part containing vertex i. Then, given a set of constants $\alpha_1, \alpha_2, \ldots, \alpha_m$ with $\sum_{k=1}^m \alpha_k = 0$, define a vector $v \in \mathbb{R}^{|S|}$ by $v_i = y_i\alpha_{p(i)}$. Then, as a_{ij} is 1 if $p(i) \neq p(j)$ and 0 otherwise,

$$(Dv)_i = \frac{my_i}{m-1}\sum_{\{j:\ p(j)\neq p(i)\}} y_j\alpha_{p(j)} = \frac{y_i}{m-1}(-\alpha_{p(i)}) = -\frac{1}{m-1}v_i,$$

so this gives an eigenspace of dimension $m-1$ with eigenvalue $-1/(m-1)$.

Now let w be a vector with $w_j = 1$ for all j. Then

$$\begin{aligned}(w^\top S)_j &= \sum_{i\in S}\left(\frac{ma_{ij}y_i}{m-1} - \frac{ma_{i,n_0}y_i}{m-1} - 2y_i + \frac{2y_i}{m-1}\sum_{k\in S} ma_{k,n_0}y_k\right) \\ &= 1 - \frac{1}{m-1}\sum_{i\in S} ma_{i,n_0}y_i - 2 + \frac{2}{m-1}\sum_{i\in S} ma_{i,n_0}y_i \\ &= \frac{m}{m-1}\sum_{i\in S} a_{i,n_0}y_i - 1,\end{aligned}$$

so there is a dimension 1 eigenspace with eigenvalue

$$\eta_{n_0} = \frac{m}{m-1}\sum_{i\in S} a_{i,n_0}y_i - 1.$$

If $p(i) = p(j)$, $a_{ij} = 0$, so rows i and j of D are identical. Hence, each part k gives an eigenspace with eigenvalue 0 and dimension $|V_k| - 1$. The eigenspace of the zero eigenvalues is in the direction where F is constant.

We now need to consider the eigenvalues η_i for $i \notin S$. If our subgraph is a trapping subgraph then this ensures that, as i is outside S,

$$\sum_{k\in S} a_{ik}y_k < \frac{m-1}{m},$$

and so these eigenvalues are negative for all choices of y_k, $k \in S$. Hence, in this case all the eigenvalues are negative or 0, and the eigenspace of the zero eigenvalues is in the direction where F is constant, so the set of fixed points is an attractor.

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

A slightly weaker condition than that of a trapping subgraph in [11] arises if it is possible to find an equilibrium y with support S such that

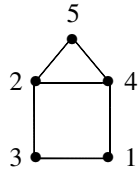
$$\sum_{k\in S} a_{ik}y_k < \frac{m-1}{m},$$

or, equivalently,

$$N_i(y) < H(y), \tag{2.1}$$

simultaneously for all $i \notin S$. In this case there is a region of the family of fixed points where all the eigenvalues are negative or 0, and the eigenspace of the zero eigenvalues is in the direction where F is constant, but this does not apply throughout the family of fixed points, so the family is not an attractor according to the standard definition. In what follows, we will extend the definition of a trapping subgraph from that in [11] by including those where it is possible to find such an equilibrium y.

For example, the simplest case where this arises is the graph with five vertices:



This graph contains a trapping subgraph according to the definition in [11], where S is the triangle formed by vertices $\{2, 4, 5\}$. However, if S is the bipartite graph $\{1, 2\} \cup \{3, 4\}$ and $y_2 + y_4 < \frac{1}{2}$, then the eigenvalues other than those within the family of fixed points are all negative.

In the context of VRRWs, it was shown in [3] that any stable equilibrium y of the dynamical system has support consisting of a complete m-partite graph S, and that the condition mentioned above that (2.1) is satisfied simultaneously for all $i \notin S$ implies that there is positive probability of VRRWs being trapped in a neighbourhood of y.

The results for VRRWs in [3] also show that stable equilibria of the dynamical system driven by $\hat{F}$, which appears in the 'true twins' case, are localised on cliques of the original graph G_0: if S is a clique of G_0 then any y with $\sum_{i\in S} y_i = 1$ is a stable equilibrium. The condition that $\hat{N}_i(y) < \hat{H}(y)$ simultaneously for all $i \notin S$ implies that S is not contained within a larger clique.

2.3. Convergence to stable equilibria

In this section we show that in the 'false twins' case any trapping subgraph (in the weaker sense described above) has a positive probability of trapping the process $x^{(n)}$. Throughout the proofs, similar arguments can be applied in the 'true twins' case to show that any clique of G_0 which is not contained within a larger clique can trap the process. The method, and the proofs of Lemmas 2.2 and 2.3, are based on those used for VRRWs in [3].

The following definitions and notation follow [3]. Let S be a complete k-partite subgraph of G_0 with outer boundary B. Let $\mathcal{S}$ consist of elements of Δ^{n_0-1} whose support is S, and let $\mathcal{S}'$ consist of elements of Δ^{n_0-1} which are nonzero at all elements of S. Let $\mathcal{L}(U)$ be the event that $x^{(n)}$ converges to a stable equilibrium $x^{(\infty)} \in \mathcal{S} \cap U$.

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

Given $q \in \mathcal{S}$, define the entropy function

$$V_q(y) = \begin{cases} -\sum_{i \in S} q_i \log\left(\frac{y_i}{q_i}\right) + 2\sum_{i \notin S} y_i, & y \in \mathcal{S}', \\ \infty, & \text{otherwise.} \end{cases}$$

We define two types of balls around q, one based on the entropy function,

$$B_{V_q}(r) = \{y \in \Delta^{n_0-1} \colon V_q(y) < r\},$$

and one based on the ∞-norm,

$$B_\infty(q, r) = \{y \in \Delta^{n_0-1} \colon \|y - q\|_\infty < r\}.$$

As stated in [3], there are increasing continuous functions $u_{1,q}$ and $u_{2,q} \colon \mathbb{R}^+ \to \mathbb{R}^+$ such that $u_{1,q}(0) = 0$ and $u_{2,q}(0) = 0$, and, for all $r > 0$, $B_\infty(q, u_{1,q}(r)) \subset B_{V_q}(r) \subset B_\infty(q, u_{2,q}(r))$.

Let

$$\zeta_i^{(n+1)} = \begin{cases} \dfrac{\varepsilon_i^{(n+1)}}{x_i^{(n)}}, & i \in S,\ x_j^{(n)} \neq 0 \text{ for all } j \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Still following [3], for $q, z \in \Delta^{n_0-1}$, let

$$\begin{aligned} I_q(z) &= -\sum_{i \in S} q_i (N_i(z) - H(z)) + 2\sum_{i \notin S} z_i (N_i(z) - H(z)) \\ &= -H(z)\left(-\sum_{i \in S} q_i \frac{F_i(z)}{z_i} + 2\sum_{i \notin S} F_i(z)\right). \end{aligned}$$

In the following lemma, this quantity will be related to the increment in entropy relative to q between $x^{(n)}$ and $x^{(n+1)}$.

The following lemma corresponds to Lemma 5 of [3], with a virtually identical proof.

Lemma 2.2. ([3, Lemma 5].) *Let $q \in \mathcal{S}$ be a stable equilibrium of the dynamical system with $N_i(q) < H(q)$ for all $i \in B$. There exists ε such that if n is large enough and $x^{(n)} \in B_{V_q}(\varepsilon)$, then*

$$V_q(x^{(n+1)}) - V_q(x^{(n)}) = \frac{I_q(x^{(n)})}{(n + n_0 + 1)H(x^{(n)})} - \langle q, \zeta^{(n+1)} \rangle + 2\sum_{i \notin S} \varepsilon_i^{(n+1)} + O\left(\frac{1}{(n + n_0)^2}\right),$$

and, furthermore,

$$I_q(x^{(n)}) \leq -\left(H(q) - H(x^{(n)}) + C_1 \sum_{i \notin S} x_i^{(n)}\right)$$

for a positive constant C_1.

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

Proof. We have

$$\begin{aligned}
&V_q(x^{(n+1)}) - V_q(x^{(n)}) \\
&\quad = -\sum_{i\in S} q_i \left(\log\left(\frac{x_i^{(n+1)}}{q_i}\right) - \log\left(\frac{x_i^{(n)}}{q_i}\right)\right) + 2\sum_{i\notin S}(x_i^{(n+1)} - x_i^{(n)}) \\
&\quad = -\sum_{i\in S} q_i \frac{x_i^{(n+1)} - x_i^{(n)}}{x_i^{(n)}} + 2\sum_{i\notin S}(x_i^{(n+1)} - x_i^{(n)}) + O\left(\frac{1}{(n+n_0)^2}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Taylor's expansion)} \\
&\quad = -\sum_{i\in S} q_i \frac{F_i(x^{(n)})}{(n+n_0+1)x_i^{(n)}} - \sum_{i\in S} q_i \zeta_i^{(n+1)} + 2\sum_{i\notin S}\left(\frac{F_i(x^{(n)})}{n+n_0+1} + \varepsilon_i^{(n+1)}\right) \\
&\qquad + O\left(\frac{1}{(n+n_0)^2}\right) \\
&\quad = \frac{I_q(x^{(n)})}{(n+n_0+1)H(x^{(n)})} - \langle q, \zeta^{(n+1)}\rangle + 2\sum_{i\notin S}\varepsilon_i^{(n+1)} + O\left(\frac{1}{(n+n_0)^2}\right).
\end{aligned}$$

For the inequality for $I_q(x^{(n)})$, observe that

$$\sum_{i\in S} q_i N_i(z) = \sum_{i\in G} q_i N_i(z) = \sum_{i\in G} z_i N_i(q) = H(q) + \sum_{i\in B} z_i (N_i(q) - H(q)),$$

by the definition of N_i and the fact that q is an equilibrium. So

$$I_q(z) = H(z) - H(q) + \sum_{i\in B} z_i (2(N_i(z) - H(z)) - (N_i(q) - H(q))),$$

so the inequality is satisfied if we choose ε small enough that if $z \in B_{V_q}(\varepsilon)$, $2(N_i(z) - H(z)) - (N_i(q) - H(q)) < -C_1$ for all $i \notin S$.

Lemma 2.3. ([3, Lemma 7].) *Let $q \in \mathcal{S}$ be a stable equilibrium of the dynamical system with $N_i(q) < H(q)$ for all $i \in B$. For sufficiently small ε and sufficiently large n, if $x^{(n)} \in B_{V_x}(\varepsilon/2)$,*

$$\mathrm{P}(\mathcal{L}(B_{V_q}(\varepsilon)) \mid \mathcal{F}_n) \geq 1 - \exp(-\varepsilon^2 C_2(n+n_0)).$$

Proof. The proof follows that of Lemma 7 of [3].

Fix ε small enough that, for all $x \in B_{V_q}(\varepsilon)$, $x_i \geq \alpha$ for all $i \in S$ and some positive α. Define martingales $(A_k)_{k\geq n}$, $(B_k)_{k\geq n}$, and $(\kappa_k)_{k\geq n}$ by $A_n = B_n = \kappa_n = 0$, and, for $k > n$,

$$\begin{aligned}
A_k &= \sum_{j=n+1}^{k} \zeta^{(j)} \mathbf{1}_{\{V_q(x^{(j-1)})<\varepsilon\}}, \\
B_k &= \sum_{j=n+1}^{k} \sum_{i\in B} \varepsilon_i^{(j)} \mathbf{1}_{\{V_q(x^{(j-1)})<\varepsilon\}}, \\
\kappa_k &= -\langle q, A_k\rangle + 2B_k.
\end{aligned}$$

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

By martingale convergence, all three converge almost surely and in L^2, and as increments of κ have moduli at most $C_3/(k+n_0)$ for some constant C_3,

$$\mathrm{E}(\exp(\theta(\kappa_k - \kappa_{k-1}) \mid \mathcal{F}_{k-1})) \le \exp\left(\frac{C_3^2}{2}\frac{\theta^2}{(k+n_0)^2}\right).$$

As $(\kappa_k)_{k\ge n}$ is a martingale, $(\exp(\kappa_k))_{k\ge n}$ is a submartingale, so Doob's submartingale inequality implies that

$$\mathrm{P}\left(\sup_{k\ge n}\kappa_k \ge c \;\middle|\; \mathcal{F}_n\right) \le \mathrm{e}^{-\theta c}\,\mathrm{E}(\mathrm{e}^{\theta\kappa_\infty} \mid \mathcal{F}_n) \le \exp\left(-\theta c + \frac{\theta^2 C_3^2}{2(n+n_0)}\right),$$

so if $\theta = c(n+n_0)/C_3^2$ then

$$\mathrm{P}\left(\sup_{k\ge n}\kappa_k \ge c \;\middle|\; \mathcal{F}_n\right) \le \exp\left(-\frac{c^2}{2C_3^2}(n+n_0)\right).$$

Let Υ be the event that $\sup_{k\ge n}\kappa_k < \varepsilon/4$; then

$$\mathrm{P}(\Upsilon \mid \mathcal{F}_n) \ge 1 - \exp(-\varepsilon^2 C_2(n+n_0))$$

for a new constant C_2.

Lemma 2.2 implies that

$$V_q(x^{(k)}) - V_q(x^{(n)}) \le \kappa_k + \frac{\varepsilon}{4}$$

if n is large enough (as in [3], we use the fact that Lemma 4 of [3] implies that $H(q) - H(x^{(n)}) \ge 0$ if $x^{(n)} \in B_{V_q}(\varepsilon)$ for small enough ε). Hence, on Υ, $V_q(x^{(k)}) < \varepsilon$ for all $k \ge n$.

Lemma 2.2 now implies that, as κ_k converges as $k \to \infty$, $\liminf_{k\to\infty}(-I_q(x^{(k)}) = 0$ (otherwise $V_q(x^{(k)}) \to -\infty$, but $V_q(x) > 0$ for all x) and so

$$\liminf_{k\to\infty}\left(H(q) - H(x^{(k)}) + C_1\sum_{i\notin S}x_i^{(k)}\right) = 0.$$

Hence, there exists a subsequence $(j_k)_{\{k\ge 0\}}$ with

$$\lim_{k\to\infty} H(x^{(j_k)}) = H(q)$$

and

$$\lim_{k\to\infty}\sum_{i\notin S}x_i^{(j_k)} = 0.$$

As in [3], we identify an accumulation point r of $(x^{(j_k)})_{\{k\ge 0\}}$, which will have $H(r) = H(q)$ and, hence (by the lemmas in [3]), be a stable equilibrium if ε is small enough.

Redefine the martingale $(\kappa_k)_{k\ge n}$ in terms of r instead of q, and let j_k be far enough along this subsequence that $V_r(x^{(j_k)}) < \varepsilon/2$ and $\sup_{k\ge j_k}|\kappa_k - \kappa_j| < \varepsilon/4$. Then Lemma 2.2 implies that, for $j' > j > j_k$,

$$V_r(x^{(j')}) \le V_r(x^{(j)}) + \sup_{k>j}|\kappa_k - \kappa_j| + \frac{C_4}{j}.$$

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

As $\liminf_{j\to\infty} V_r(x^{(j)}) = 0$ and $\lim_{j\to\infty}(\sup_{k>j} |\kappa_k - \kappa_j| + C_4/j) = 0$, we have

$$V_r(x^{(n)}) \to \lim_{k\to\infty} V_r(x^{(j_k)}) = 0$$

and so $x^{(n)} \to r$.

Lemma 2.4. *Assume that, for a given vertex i, $N_i(x^{(n)})/H(x^{(n)})$ converges to $\lambda_i \in (0,\infty)$. Then, for $i \in B$, $c_i^{(n)}/n^{\lambda_i}$ converges to a limit in $(0,\infty)$ almost surely.*

Proof. Let

$$Y_i^{(n)} = \sum_{k=1}^{n} \frac{\mathbf{1}_{\{X_k=i\}}}{c_i^{(k-1)}},$$

and let

$$M_i^{(n)} = Y_i^{(n)} - \sum_{k=1}^{n} \frac{N_i(x^{(k-1)})}{H(x^{(k-1)})(k+n_0)}.$$

Then

$$\mathrm{E}(Y_i^{(n+1)} - Y_i^{(n)} \mid \mathcal{F}_n) = \frac{N_i(x^{(n)})}{H(x^{(n)})(n+n_0)},$$

so $(M_i^{(n)})_{n\geq 1}$ is a martingale.

Now,

$$\begin{aligned}
\mathrm{E}\left(\sum_{n=1}^{\infty}(M_i^{(n)} - M_i^{(n-1)})^2\right) &= \mathrm{E}\left(\sum_{n=1}^{\infty}\left(\frac{\mathbf{1}_{\{X_n=i\}}}{c_i^{(n-1)}} - \frac{N_i(x^{(n-1)})}{H(x^{(n-1)})(n+n_0)}\right)^2\right) \\
&\leq \mathrm{E}\left(\sum_{n=1}^{\infty}\left(\frac{\mathbf{1}_{\{X_n=i\}}}{c_i^{(n-1)}}\right)^2 + \left(\frac{N_i(x^{(n-1)})}{H(x^{(n-1)})(n+n_0)}\right)^2\right) \\
&\leq \infty,
\end{aligned}$$

so martingale convergence implies that

$$\log c_i^{(n)} \equiv Y_i^{(n)} \equiv \lambda_i \log n,$$

giving the result.

Let $\mathcal{R}_{n,k}$ be the range of the process $(X_j)_{j\in\mathbb{N}}$ between times n and k, and let $\mathcal{R}_{n,\infty}$ be the range of the process $(X_j)_{j\in\mathbb{N}}$ for times $j \geq n$.

We now combine our results.

Theorem 2.2. *In the 'false twins' case, let G_0 contain a complete m-partite graph S with outer boundary B such that there exists a stable equilibrium q of the dynamical system driven by F with support S and with $N_i(q) < H(q)$ for all $i \in B$. Then, with positive probability, for some stable equilibrium $r \in B_{V_q}(\varepsilon)$ with support S,*

1. *$x^{(n)} \to r$;*
2. *for $i \in B$, $c_i^{(n)}/n^{N_i(r)/H(r)}$ converges to a limit in $(0,\infty)$ almost surely;*
3. *for some (random) time n, $\mathcal{R}_{n,\infty} = S \cup B$.*

https://doi.org/10.1239/jap/1276784910 Published online by Cambridge University Press

Furthermore, if there is a stable equilibrium r in the limit set $L(x)$ of $(x^{(n)})_{n\in\mathbb{N}}$ with support S and with $N_i(r) < H(r)$ for all $i \in B$, then, almost surely, $x^{(n)} \to r$ as $n \to \infty$.

Proof. That convergence occurs with positive probability follows from Lemma 2.3, and the behaviour of $c_i^{(n)}$ for $i \in B$ follows from Lemma 2.4.

For $i \notin S \cup B$, for which $N_i(r)/H(r) \to 0$ as $n \to \infty$ on $x^{(n)} \to r$, the same argument as in Lemma 2.4 shows that $c_i^{(n)}/n^\alpha \to 0$ for any $\alpha > 0$, from which it follows that $f_i(x^{(n)})/n^\alpha \to 0$ as $n \to \infty$ if $\alpha > \max_{j\in B}(N_j(r)/H(r)) - 2$, which implies that, almost surely, i is visited only finitely many times.

The last part is also a consequence of Lemma 2.3.

Theorem 2.3. *In the 'true twins' case, let G_0 contain a clique S with outer boundary B such that S is not contained in a larger clique, and let q be a stable equilibrium with support S. Then, with positive probability, for some stable equilibrium $r \in B_{V_q}(\varepsilon)$ with support S,*

1. *$x^{(n)} \to r$;*
2. *for $i \in B$, $c_i^{(n)}/n^{\hat{N}_i(r)/\hat{H}(r)}$ converges to a limit in $(0, \infty)$ almost surely;*
3. *for some (random) time n, $\mathcal{R}_{n,\infty} = S \cup B$.*

Furthermore, if there is a stable equilibrium r in the limit set $L(x)$ of $(x^{(n)})_{n\in\mathbb{N}}$ with support S, then, almost surely, $x^{(n)} \to r$ as $n \to \infty$.

Proof. The proofs of Lemmas 2.2 and 2.3 apply in this case as well, with N_i and H replaced by $\hat{N}_i$ and $\hat{H}$. For Lemma 2.4, if we define

$$Y_i^{(n)} = \sum_{k=1}^{n} \frac{\mathbf{1}_{\{X_k=i\}}}{c_i^{(k-1)}},$$

as before, then

$$\mathrm{E}(Y_i^{(n+1)} - Y_i^{(n)} \mid \mathcal{F}_n) = \frac{\hat{N}_i(x^{(n)})}{\hat{H}(x^{(n)})(n+n_0)} + \frac{\hat{N}_i(x^{(n)}) - \hat{H}(x^{(n)})}{[\hat{H}(x^{(n)})(n+n_0) - 1]\hat{H}(x^{(n)})(n+n_0)},$$

so we redefine the martingale $M_i^{(n)}$ by

$$M_i^{(n)} = Y_i^{(n)} - \sum_{k=1}^{n}\left(\frac{N_i(x^{(k-1)})}{H(x^{(k-1)})(k+n_0)} + \frac{\hat{N}_i(x^{(n)}) - \hat{H}(x^{(n)})}{[\hat{H}(x^{(n)})(n+n_0) - 1]\hat{H}(x^{(n)})(n+n_0)}\right).$$

The rest of the argument is the same as in the proof of Lemma 2.4.

Acknowledgement

The authors would like to acknowledge the support of the EPSRC funded Amorphous Computing, Random Graphs and Complex Biological Systems group (grant reference EP/D003105/1).

References

[1] Bebek, G. *et al.* (2006). The degree distribution of the generalized duplication model. *Theoret. Comput. Sci.* **369,** 239–249.

[2] Benaïm, M. (1997). Dynamics of stochastic approximation algorithms. In *Séminaires de Probabilités XXXIII* (Lecture Notes Math. **1709**), Springer, Berlin, pp. 1–68.

[3] Benaïm, M. and Tarrès, P. (2008). Dynamics of vertex-reinforced random walks. To appear in *Ann. Prob.*
[4] Bollobás, B. and Riordan, O. M. (2003). Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks*, eds S. Bornholdt and H. G. Schuster, Wiley-VCH, Weinheim, pp. 1–34.
[5] Bonato, A. *et al.* (2010). Models of on-line social networks. To appear in *Internet Math.*
[6] Chung, F., Lu, L., Dewey, T. G. and Galas, D. J. (2003). Duplication models for biological networks. *J. Comput. Biol.* **10,** 677–687.
[7] Kumar, R. *et al*. (2000). Stochastic models for the web graph. In *Proc. 41st Annual Symp. on Foundations of Comput. Sci.* (Redondo Beach, CA, 2000), IEEE Computer Society Press, Los Alamitos, CA, pp. 57–65.
[8] Pemantle, R. (1988). Random processes with reinforcement. Doctoral Thesis, Department of Mathematics, Massachusetts Institute of Technology.
[9] Pemantle, R. (2007). A survey of random processes with reinforcement. *Prob. Surveys* **4,** 1–79.
[10] Raval, A. (2003). Some asymptotic properties of duplication graphs. *Phys. Rev. E* **68,** 066119, 10 pp.
[11] Volkov, S. (2001). Vertex-reinforced random walk on arbitrary graphs. *Ann. Prob.* **29,** 66–91.