

## Research Article

# A Markov Chain Approach to Randomly Grown Graphs

**Michael Knudsen and Carsten Wiuf**

*Bioinformatics Research Center, University of Aarhus, Høegh-Guldbergs Gade 10, Building 1090, 8000 Århus C, Denmark*

Correspondence should be addressed to Michael Knudsen, micknudsen@gmail.com

Received 29 June 2007; Accepted 3 January 2008

Recommended by Rahul Roy

A Markov chain approach to the study of randomly grown graphs is proposed and applied to some popular models that have found use in biology and elsewhere. For most randomly grown graphs used in biology, it is not known whether the graph or properties of the graph converge (in some sense) as the number of vertices becomes large. Particularly, we study the behaviour of the degree sequence, that is, the number of vertices with degree  $0, 1, \dots$ , in large graphs, and apply our results to the partial duplication model. We further illustrate the results by application to real data.

Copyright © 2008 M. Knudsen and C. Wiuf. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1. Introduction

Over the past decade, networks have played a prominent role in many different disciplines including theoretical physics, technology, biology, and sociology [1–9]. Particularly in biology, networks have become fundamental for the description of complex data structures. The appeal of networks may, at least partly, be due to the fact that in addition to being based on a rigorous mathematical base [10–14], they also provide a convenient graphical representation of the data which allows for visual interpretation. Examples of complex data structures that can be described by networks include food webs in ecology, sexual partner networks in sociology, and protein interaction networks in biology.

The canonical model in random graph theory has been Erdős-Renyi random graphs, where each of a fixed number of vertices has a Poisson-distributed number of links to other vertices. A Poisson number of links have turned out barely to be realistic for many empirically observed networks, and other models have been suggested to accomodate the discrepancies between theory and observation. Barabási and Albert [2] proposed a simple stochastic model, the preferential attachment (PA) model, whereby the network gradually is built up by adding one vertex at a time until the network reaches the desired size. This model is able to account for

the scale-free degree distribution that is observed in some empirical networks, but not many of the other features and motifs that are found in real networks (e.g., [15–18]). Therefore, for mathematical and statistical analysis of network data, many other stochastic models have been proposed, in particular models that fall in the class of randomly grown graphs (RGGs; see next section for a definition) which share the property of the PA model of gradual growth. Overviews of different models and their properties can be found in [13, 16, 19, 20].

While the PA model has been under close mathematical scrutiny (e.g., [20]), other RGGs have been treated less extensively (e.g., [19, 21]) and mostly in the context of considering a continuous time approximation to the original discrete time Markov process (e.g., [13, 22, 23]). In this paper, we specifically address the question of the behavior of the vertex degrees as the number of vertices grows large. For a class of RGGs (including the PA model), the existence of a limiting degree distribution has been proven and its analytical form has been derived [21]. However, for most RGGs applied in biology, it is not known whether a limiting distribution exists, let alone its form.

Biologically, it is of great interest to know whether the network stabilizes as it grows, or whether the degree distribution is a function of the size of the network, even for large network sizes. It relates to the question whether the network in an evolutionary perspective reaches an equilibrium, such that adding new vertices does not change the overall connectivity of the network. For example, in relation to protein interaction networks where vertices represent proteins and edges represent physical interactions between proteins, both scenarios seem a priori possible. Proteins may be able to engage in an unlimited number of interactions, or the number of interactions may be restricted by a number of factors such as space, time, and protein production rates. With the increasing statistical interest in analyzing complex biological networks with respect to evolutionary and functional properties [1, 5, 9, 13, 14, 24], it is also becoming of interest to understand the mathematical properties of the models.

We study a large class of RGGs that allows the construction of a simple, but time-inhomogeneous, Markov chain. For a given RGG, the corresponding Markov chain can be used to address questions about the RGG, for example, questions about the degree distribution. In particular, we focus on a special RGG, the partial duplication model, which has recently been used in the study of biological protein interaction networks [16, 18, 25, 26] and has formed the basis for new and more biologically realistic models (e.g., [16, 27]). The partial duplication model has two parameters ( $p$  and  $q$ ) and we give conditions under which the chain is ergodic or transient. Further, based on the time-inhomogeneous Markov chain, we define a time-homogeneous Markov chain and a continuous time, time-homogeneous Markov process, and demonstrate that these, in general, are easier to study and apply than the original chain. Proofs rely on general theory of discrete Markov processes, which makes it easy to prove similar results for other RGGs.

Finally, we apply our results to a collection of real protein interaction data.

## 2. RGGs

An RGG is a Markov chain  $\{G_t\}_{t \geq s}$  on undirected graphs such that  $G_t$  has exactly  $t$  vertices, and the set of vertices of  $G_t$  is a subset of the set of vertices of  $G_{t+1}$  for all  $t \geq s$ . Note that we do *not* require the set of edges of  $G_t$  to be a subset of the set of edges of  $G_{t+1}$ .

Denote by  $n_t(k)$  the expected number of vertices of degree  $k$  at time  $t$ . Since, by assumption, the graph  $G_t$  has exactly  $t$  vertices, the expected relative frequency of vertices of degree

$k$  at time  $t$  is  $f_t(k) = n_t(k)/t$ . For many RGGs, one can derive a recursive formula for  $n_t(k)$ , often referred to as the *master equation* [13]. Here, we consider a very general class of master equations given by

$$n_{t+1}(k) = \sum_{j \geq 0} A(t)_{k,j} n_t(j), \quad (2.1)$$

where  $A(t)$  for all  $t \geq s$  is an infinite real matrix with  $A(t)_{k,j} = 0$  for  $k > j + 1$ , and such that all columns sum to the same number  $a(t)$ . Furthermore, assume for suitable real numbers  $b_{k,j}$  that

$$A(t)_{k,j} = \begin{cases} 1 - \frac{b_{k,k}}{t} & \text{for } k = j, \\ \frac{b_{k,j}}{t} & \text{for } k \neq j \end{cases} \quad (2.2)$$

with  $b_{k,j} = 0$  for  $k > j + 1$ . The latter condition guarantees that the vertex degree can increase by at most one. By construction,  $n_t(k) = 0$  for  $k + 1 > t$ , and hence  $A(t)$  is effectively a  $t \times (t - 1)$  matrix. We assume that the entries (2.2) in this submatrix are positive.

One particular example of a model fulfilling the conditions above is the partial duplication model (details are found in Section 4). The master equation is given by

$$\begin{aligned} n_{t+1}(k) = & \left(1 - \frac{q + kp}{t}\right) n_t(k) + (1 - q) \sum_{j \geq k} \binom{j}{k} p^k (1 - p)^{j-k} \frac{n_t(j)}{t} \\ & + \frac{q + (k-1)p}{t} n_t(k-1) + q \sum_{j \geq k-1} \binom{j}{k-1} p^{k-1} (1 - p)^{j-k+1} \frac{n_t(j)}{t}. \end{aligned} \quad (2.3)$$

For several other models, the master equation takes a similar form. Among these models are the duplication-divergence model [16], an approximation to the duplication-mutation model [22, 23], and the models discussed in [21] after a suitable modification (see Section 5.2). Generally, (2.1) is fulfilled whenever the expected degree change in a vertex depends on the degree only, and not on the degrees of the other vertices.

It follows immediately from (2.1) that

$$f_{t+1}(k) = \sum_{j \geq 0} B(t)_{j,k} f_t(j), \quad (2.4)$$

where  $B(t)$  is the transpose of  $(t/(t+1))A(t)$ , and by assumption all rows of  $B(t)$  sum to  $b(t) = (t/(t+1))a(t)$ . It follows that

$$1 = \sum_{k \geq 0} f_{t+1}(k) = \sum_{k \geq 0} \sum_{j \geq 0} B(t)_{j,k} f_t(j) = b(t) \sum_{j \geq 0} f_t(j) = b(t), \quad (2.5)$$

that is,  $b(t) = 1$  and the matrices  $\{B(t)\}_{t \geq s}$  describe a Markov chain with time-dependent transition probabilities.

**Proposition 2.1.** *Assume that  $\sum_{k \geq 0} b_{j,k} < \infty$  for all  $j \geq 0$ . If  $f_t(j) \rightarrow f(j)$  pointwise for all  $j \geq 0$ , then  $\{f(j)\}_{j \geq 0}$  satisfies*

$$0 = \sum_{k \neq j} b_{j,k} f(k) - (1 + b_{j,j}) f(j), \quad \sum_{j \geq 0} f(j) \leq 1. \quad (2.6)$$

*Proof.* The second part of the proposition is a simple application of Fatou's lemma. By using (2.4), the definition of  $B(t)$ , and  $\sum_{k \geq 0} b_{j,k} < \infty$ , it follows that

$$(t+1)(f_{t+1}(j) - f_t(j)) = \sum_{k \neq j} b_{j,k} f_t(k) - (1 + b_{j,j}) f_t(j) \longrightarrow d_j \quad \text{for } t \longrightarrow \infty \quad (2.7)$$

for some real number  $d_j$ , and it remains to prove that  $d_j = 0$ . Note that

$$\sum_{n=s}^t d_j(n) = (t+1)f_{t+1}(j) - s f_s(j) - \sum_{n=s}^t f_n(j), \quad (2.8)$$

and by using Cesaro's lemma, we get

$$d_j = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=s}^t d_j(n) = f(j) - f(j) = 0. \quad (2.9)$$

□

Consider the *jump chain* corresponding to the Markov chain  $\{B(t)\}_{t \geq s}$ , that is, the Markov chain with transition probabilities  $B(t)_{j,k} / (1 - B(t)_{j,j})$  for  $j \neq k$ , unless  $B(t)_{j,j} = 1$  in which case the probability is put to 0. The jump chain has time-independent transition probabilities given by

$$p_{j,k} = \frac{b_{k,j}}{1 + b_{j,j}} \quad \text{for } k \neq j, \quad (2.10)$$

and  $p_{j,j} = 0$  for all  $j \geq 0$ . If  $1 + b_{j,j} = 0$ , then  $p_{j,k} = 0$ . Occasionally, we consider a slightly modified jump chain (still with time-independent transition probabilities) which is allowed to stay in the same state with positive probability.

If a stationary distribution  $\{\pi_j\}_{j \geq 0}$  for the jump chain exists, it fulfills

$$\pi_j = \sum_{k \neq j} \pi_k \frac{b_{j,k}}{1 + b_{k,k}} \quad \forall j \geq 0. \quad (2.11)$$

Assume that  $\inf_{j \geq 0} (1 + b_{j,j}) > 0$  and put  $\pi'_j = \pi_j / (1 + b_{j,j})$ . Then we obtain that

$$0 = \sum_{k \neq j} b_{j,k} \pi'_k - (1 + b_{j,j}) \pi'_j \quad \forall j \geq 0, \quad (2.12)$$

and hence  $\{\pi'_j\}_{j \geq 0}$  is a solution to the equation in Proposition 2.1. Furthermore, we may normalize  $\{\pi'_j\}_{j \geq 0}$  to get a distribution, and hence (2.11) and (2.12) may be used to transfer a stationary distribution for the jump chain to the limit of the time-inhomogeneous Markov chain and vice versa.

In our main example, the partial duplication model (see Section 4 for details), we have  $b_{0,0} = 2q - 1$  and

$$b_{j,j} = q + jp - (1 - q)p^j - qp^{j-1}(1 - p) \quad \text{for } j \geq 1, \quad (2.13)$$

and hence the assumption  $\inf_{j \geq 0} (1 + b_{j,j}) > 0$  is fulfilled if  $q > 0$ .

### 3. A continuous time approximation

In this section, we show that the time-inhomogeneous Markov chain converges to a continuous time, time-homogeneous Markov process after a suitable time transformation.

Denote by  $T_i$  the time of the  $i$ th jump in the time-inhomogeneous chain after a given time  $t_0$ , and let  $J_i$  be the state to which it jumps. Set  $T_0 = t_0$  and  $J_0 = j_0$ , the state of the chain at time  $t_0$ . To simplify notation further, introduce  $s_i = (t_i, j_i)$  and  $S_i = (T_i, J_i)$ .

Note that at time  $t$ , the probability of staying in state  $j$  is  $B(t)_{j,j} = 1 - (b_{j,j} + 1)/(t + 1)$ . In particular, if we let  $\alpha_i = b_{j_{i-1},j_{i-1}} + 1$ , then

$$P(T_i > t_i \mid S_{i-1} = s_{i-1}) = \prod_{u=t_{i-1}+1}^{t_i} \left(1 - \frac{\alpha_i}{u+1}\right) \approx \left(\frac{t_i}{t_{i-1}}\right)^{-\alpha_i} \quad (3.1)$$

for large  $t_{i-1}$  and  $t_i$ . Now consider the transformation  $Z_i = \log T_i - \log T_{i-1} = \log T_i/T_{i-1}$ . It follows that

$$P(Z_i > z \mid S_{i-1} = s_{i-1}) = P(T_i > t_{i-1}e^z \mid S_{i-1} = s_{i-1}) \longrightarrow e^{-\alpha_i z} \quad (3.2)$$

as  $t_{i-1} \rightarrow \infty$ . That is, in the limit, the transformed waiting time is exponentially distributed with parameter  $\alpha_i$ .

**Proposition 3.1.** *Let  $X_{s_0}(z)$ ,  $z \geq 0$ , take the value of the time-inhomogeneous Markov chain at time  $t$ , where  $t = \lfloor t_0 e^z \rfloor$  and  $\lfloor x \rfloor$  denotes the integer part of  $x$ . At time 0,  $X_{s_0}(0) = j_0$ . For fixed  $j_0$ , the process  $X_{s_0}(z)$  converges to a continuous time, time-homogeneous Markov process as  $t_0 \rightarrow \infty$ .*

*Proof.* Clearly, the process  $X_{s_0}(z)$ ,  $z \geq 0$ , is Markovian by definition. Let  $Z_i^*$  be the time of the  $i$ th jump, that is,  $T_i = \lfloor t_0 e^{Z_i^*} \rfloor = t_0 e^{Z_i^*}$  and  $Z_i = Z_i^* - Z_{i-1}^*$  in the notation above. It follows from (3.2) that

$$P_{s_0}(Z_i^* > z \mid Z_{i-1}^* = z_{i-1}^*, J_{i-1} = j_{i-1}) \longrightarrow e^{-\alpha_i(z - z_{i-1}^*)} \quad (3.3)$$

for  $t_0 \rightarrow \infty$ . (Subscript  $s_0$  in  $P_{s_0}$  is used to underline the implicit dependency of  $s_0 = (t_0, j_0)$ .) Recall the transition probabilities (2.10) in the original jump chain. It follows immediately that

$$P_{s_0}(J_i = j_i \mid Z_{i-1}^* = z_{i-1}^*, J_{i-1} = j_{i-1}) = P_{s_0}(J_i = j_i \mid J_{i-1} = j_{i-1}) = \frac{\beta_i}{\alpha_i}, \quad (3.4)$$

where  $\beta_i = b_{j_{i-1},j_i}$ . Combined with (3.3) this shows that, in the limit as  $t_0 \rightarrow \infty$ , the rate of jumping to  $j_i$  from  $j_{i-1}$  is  $\beta_i$ . More precisely, it demonstrates that  $X_{s_0}(z)$ ,  $z \geq 0$ , converges to a continuous time, time-homogeneous Markov process with transition rate matrix  $Q = \{q_{j,k}\}_{j,k \geq 0}$  given by  $q_{j,k} = b_{k,j}$  for  $j \neq k$ , and  $q_{j,j} = -q_j = \sum_{k \neq j} q_{j,k}$ . This sum is indeed finite because by assumption  $b_{k,j} = 0$  for  $k > j + 1$  (see Section 2).  $\square$

Note that a stationary equation  $\{\pi_j\}_{j \geq 0}$  for the continuous-time Markov chain fulfills the equation in Proposition 2.1 with  $f(j)$  replaced by  $\pi_j$ .

#### 4. The partial duplication model

Consider the model  $\{G_t\}_{t \geq s}$ , where  $G_s$  is a simple graph with  $s$  vertices, and where  $G_{t+1}$  is obtained from  $G_t$  in the following way: introduce a new vertex  $v$  and choose  $u \in G_t$  uniformly. With probability  $q$ , connect  $v$  and  $u$ . Independently of each other, connect each neighbor of  $u$  to  $v$  with probability  $p$ .

In this section, we follow the path outlined in the previous section. That is, we first find the jump chain corresponding to the partial duplication model. As already stated in Section 1, the master equation is given by

$$\begin{aligned} n_{t+1}(k) = & \left(1 - \frac{q + kp}{t}\right) n_t(k) + (1 - q) \sum_{j \geq k} \binom{j}{k} p^k (1 - p)^{j-k} \frac{n_t(j)}{t} \\ & + \frac{q + (k-1)p}{t} n_t(k-1) + q \sum_{j \geq k-1} \binom{j}{k-1} p^{k-1} (1 - p)^{j-k+1} \frac{n_t(j)}{t}. \end{aligned} \quad (4.1)$$

It can be seen in the following way: the first term corresponds to the case where a vertex of degree  $k$  keeps its degree, and this is the case unless one of two things happens: (i) the vertex is copied and receives a link to the new vertex, or (ii) it receives a link because one of its  $k$  neighbors is copied. The probabilities of these two events are  $q/t$  and  $kp/t$ , respectively. Similarly, the third term corresponds to the case where a vertex of degree  $k-1$  gets a new link in one of the above-mentioned ways. The two remaining terms correspond to the cases where the new vertex has degree  $k$ . The new vertex has degree  $k$  when a vertex of degree  $\geq k$  is copied and receives exactly  $k$  links to the neighbors of the copied vertex and no link to the copied vertex, or if a vertex of degree  $\geq k-1$  is copied and receives a link to the copied vertex and exactly  $k-1$  links to the neighbors of the copied vertex.

The cases  $q = 0$  and  $q = 1$  have been studied in [19, 26], respectively. Note, however, that the master equation given in [26] is incorrect. For general  $q$ , the model has been discussed in [18]. It follows immediately that

$$\begin{aligned} f_{t+1}(k) = & \left(1 - \frac{q + kp}{t}\right) \frac{t}{t+1} f_t(k) + \frac{1-q}{t+1} \sum_{j \geq k} b_j(k) f_t(j) \\ & + \frac{q + (k-1)p}{t+1} f_t(k-1) + \frac{q}{t+1} \sum_{j \geq k-1} b_j(k-1) f_t(j), \end{aligned} \quad (4.2)$$

where we, in order to simplify notation, define

$$b_j(k) = \binom{j}{k} p^k (1 - p)^{j-k}. \quad (4.3)$$

From (4.2), we may read off the description of the matrix  $B(t)$ . Its entries satisfy that

$$(t+1)B(t)_{j,k} = \begin{cases} (1-q)b_j(k) + qb_j(k-1) & \text{for } k < j, \\ t - (q + jp) + (1-q)b_j(j) + qb_j(j-1) & \text{for } k = j, \\ q + jp + qb_j(j) & \text{for } k = j + 1, \end{cases} \quad (4.4)$$

and  $B(t)_{j,k} = 0$  otherwise. An easy calculation shows that

$$(t+1) \sum_{k \neq j} B(t)_{j,k} = 1 + q + jp - [(1-q)b_j(j) + qb_j(j-1)] \quad (4.5)$$

from which it follows that the probability of jumping from state  $j$  is

$$1 - B(t)_{j,j} = \frac{1 + q + jp}{t+1} + \frac{(1-q)b_j(j) + qb_j(j-1)}{t+1}. \quad (4.6)$$

Motivated by this formula, we allow the jump chain to stay in state  $j$  with probability  $(1-q)b_j(j) + qb_j(j-1)$ , and it follows that the transition probabilities  $p_{j,k}$  in the modified jump chain satisfy that

$$(1 + q + jp)p_{j,k} = \begin{cases} (1-q)b_j(k) + qb_j(k-1) & \text{for } k \leq j, \\ q + jp + qb_j(j) & \text{for } k = j+1, \end{cases} \quad (4.7)$$

and  $p_{j,k} = 0$  otherwise.

In particular, the chain is irreducible if and only if  $0 < q < 1$ . If  $q = 0$ , the state 0 is absorbing, and if  $q = 1$ , the state 0 is not reachable from any other state. If state 0 is ignored, the resulting chain is irreducible for  $q = 1$ .

#### 4.1. Classification of states

We first recall a theorem from [28]. The theorem is reformulated in [29], and we will use that formulation. If  $q = 1$ , then we ignore the state 0, and since in this case all  $p_{j,0}$  are zero, the conditions stated in theorems below stay the same.

**Theorem 4.1.** *Let  $\{p_{j,k}\}_{j,k \geq 0}$  be a Markov chain. If there exist a sequence of non-negative real numbers  $\{x_j\}_{j \geq 0}$  and an integer  $N \geq 1$  such that*

$$\sum_{k=0}^{\infty} p_{j,k} x_k \leq x_j \quad \forall j \geq N, \quad x_j \rightarrow \infty \text{ for } j \rightarrow \infty, \quad (4.8)$$

then the chain is ultimately recurrent.

Applied to the partial duplication model the theorem states that if there is a sequence  $\{x_j\}_{j \geq 0}$  of nonnegative real numbers with  $x_j \rightarrow \infty$  such that

$$\sum_{k=0}^{j+1} x_k p_{j,k} \leq x_j \quad \forall j \gg 0, \quad (4.9)$$

then, if  $q = 0$ , the probability of ultimate absorption in 0 is 1. If  $q \neq 0$ , the conclusion of the theorem is that all states are persistent.

The solution  $p$  of  $\log(p) + p = 0$ , where  $\log$  denotes the natural logarithm, is known as the omega constant, and we denote it by  $\Omega$ . We have  $\Omega \approx 0.5671$ .

**Proposition 4.2.** *Let  $p < \Omega$  in the partial duplication model. If  $q = 0$ , the probability of ultimate absorption in 0 is 1, and if  $q > 0$ , the Markov chain is persistent.*

In [26] it is claimed that for  $q = 0$  there exists a limiting distribution different from the one we find.

*Proof.* Let  $x_j = \log(j + 1)$ . Then  $\{x_j\}_{j \geq 0}$  is a nonnegative sequence of real numbers with  $x_j \rightarrow \infty$ , and hence it suffices to show that, for the choices of  $p$  and  $q$  in the proposition, the sequence satisfies (4.9). Since  $\log$  is a concave function, Jensen's inequality implies that  $E(\log(X + 1)) \leq \log(E(X) + 1)$  for a positive random variable  $X$ . In particular, using this for binomially distributed random variables, we get

$$\sum_{k=0}^{j+1} x_k p_{j,k} \leq \frac{(1-q) \log(jp+1) + q \log(jp+2) + (q+jp) \log(j+2)}{1+q+jp}, \quad (4.10)$$

and hence we need only prove that the right-hand side of this inequality is less than or equal to  $\log(j+1)$  for  $j \gg 0$ . This may, for  $j \gg 0$ , be rewritten as

$$q \log\left(\frac{jp+2}{jp+1}\right) + q \log\left(\frac{j+2}{j+1}\right) + \log\left(\frac{jp+1}{j+1}\right) + jp \log\left(\frac{j+2}{j+1}\right) \leq 0, \quad (4.11)$$

and here the first two  $\log$ -terms converge to 0, while the two remaining terms converge to  $\log(p)$  and  $p$ , respectively. Here we have used that

$$jp \log\left(\frac{j+2}{j+1}\right) = p \frac{j}{j+1} \frac{\log(1+1/(j+1))}{1/(j+1)} \rightarrow p \quad \text{for } j \rightarrow \infty. \quad (4.12)$$

Note that since  $p < \Omega$  by assumption, we have  $\log(p) + p < 0$ , and hence the inequality in (4.11) holds for all  $j \gg 0$ .  $\square$

Since zero is the only absorbing state, it follows that for  $p \geq \Omega$ , a limiting distribution takes the form  $(a_0, 0, 0, \dots)$ , with  $a_0 \leq 1$ . To infer the behaviour of the Markov chain for other values of  $q$ , we first recall a result proved in [30].

**Theorem 4.3.** *Let  $\{p_{j,k}\}_{j,k \geq 0}$  be an irreducible, aperiodic Markov chain. If there exist a sequence of positive real numbers  $\{x_j\}_{j \geq 0}$  and an integer  $N \geq 1$  with*

$$\sum_{k=0}^{\infty} p_{j,k} x_k \leq x_j \quad \forall j \geq N, \quad x_j \rightarrow 0 \text{ for } j \rightarrow \infty, \quad (4.13)$$

*then the chain is transient.*

Let  $\Phi$  denote the golden ratio conjugate. That is,  $\Phi$  is the unique positive real number  $p$  satisfying that  $1/p = p + 1$ . We have  $\Phi \approx 0.6180$ .

**Proposition 4.4.** *Let  $q > 0$  in the partial duplication model. Then the Markov chain is transient for all  $p > \Phi$ .*

*Proof.* Put  $x_j = 1/(j+1)$  for all  $j \geq 0$ . Then  $x_j > 0$  for all  $j \geq 0$ , and  $x_j \rightarrow 0$ . Thus, in order to apply Theorem 4.3, we only need to verify that  $\{x_j\}_{j \geq 0}$  is a solution to the inequalities in (4.9).

It follows from a straightforward calculation that

$$(1+q+jp) \sum_{k=0}^{j+1} p_{j,k} x_k \leq \frac{1}{(j+1)p} + \frac{q+jp}{j+2} \quad (4.14)$$

such that  $\{x_j\}_{j \geq 0}$  is a solution if the right-hand side of this inequality is less than or equal to  $(1+q+jp)/(j+1)$  for  $j \gg 0$ . This is equivalent to

$$\frac{1}{p} - \frac{q+jp}{j+2} \leq 1 \quad \text{for } j \gg 0, \quad (4.15)$$

and the left-hand side converges to  $1/p - p$  as  $j \rightarrow \infty$ . Since  $p > \Phi$ , it follows that  $1/p - p < 1$ , and hence the inequality in (4.15) holds for all  $j \gg 0$ .  $\square$

Let  $q > 0$  such that the chain is irreducible. One may ask for which  $p$  the chain is ergodic. By Proposition 4.4, a necessary condition is  $p < \Phi$ . However, as we will see, this may not be sufficient. To see this, we first recall another theorem from [28].

**Theorem 4.5.** *Let  $\{p_{j,k}\}_{j,k \geq 0}$  be an irreducible, aperiodic Markov chain. If there exist an  $N \geq 1$  and a nonnegative sequence  $\{x_j\}_{j \geq 0}$  of real numbers such that*

$$\sum_{k=0}^{\infty} p_{j,k} x_k \leq x_j - 1 \quad \text{for } j \geq N, \quad \sum_{k=0}^{\infty} p_{j,k} x_k < \infty \quad \text{for } j < N, \quad (4.16)$$

*then the chain is ergodic.*

In the partial duplication model, the second condition in the theorem is always fulfilled since  $p_{j,k} = 0$  for  $k > j+1$ . Let  $X_t$  denote the state of the chain at time  $t$ . If there exists  $N \geq 0$  and  $\varepsilon > 0$  such that

$$E[X_t | X_{t-1} = j] \leq j - \varepsilon \quad \forall j \geq N, \quad (4.17)$$

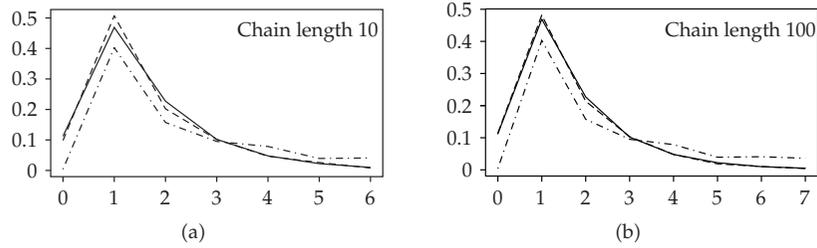
then this  $N$ , together with  $x_j = j/\varepsilon$ , will work in Theorem 4.5. This is pointed out in [28].

**Proposition 4.6.** *Let  $q > 0$ . Then the Markov chain is ergodic for all  $p < 1/2$ .*

*Proof.* We find

$$E[X_t | X_{t-1} = j] - j = \frac{j(2p-1) + 2q}{1+q+jp} \rightarrow 2 - \frac{1}{p} \quad \text{for } j \rightarrow \infty. \quad (4.18)$$

Note that  $p < 1/2$  implies  $2 - 1/p < 0$ , and hence  $2 - 1/p \leq -\varepsilon$  for all sufficiently small  $\varepsilon > 0$ . That is, for a large  $N$ , (4.17) is fulfilled.  $\square$



**Figure 1:** Shown is the distribution of vertex degrees of 50 simulated networks (solid) and that of numbers simulated from the corresponding Markov chain (dashed), using parameters estimated for the *P. falciparum* dataset. In addition, the observed degree distribution for *P. falciparum* is shown (dot-dashed).

In general, it is not an easy task to actually find the stationary distribution of the jump chain or the time-inhomogeneous Markov chain. For  $q = 1$ , an attempt to solve (2.4) has been made in [19]. They assume that  $\{f_t(j)\}_{j \geq 0}$  converges and show that, under this assumption, the limit (for  $p > 0$ ) has a power-law tail. However, this does not establish the existence of a stationary distribution. Further, the power-law they provide for  $p > \Omega$  is in fact not a distribution. In the special case  $p = 0$ , the stationary distribution is  $\pi_j = (1/2)^j$  for  $j \geq 1$ .

It is natural to ask what happens for the values of  $p$  not covered in the propositions above. In general, this is difficult. However, if  $\Omega$  is not the maximal upper bound in Proposition 4.2, the culprit must be the particular choice of  $\{x_j\}_{j \geq 0}$ . Indeed, the damage provided by the use of Jensen's inequality is not severe. This may be seen in the following way: denote by  $\mu_k(j)$  the  $k$ th central moment of a binomially distributed random variable  $X$  with parameters  $j$  and  $p$ . From [31], we get  $\mu_k(j) = O(j^{-k/2})$ , and by expanding  $\log(X + 1)$  as a Taylor series around  $jp$ , it follows that  $E[\log(X + 1)] = \log(jp + 1) + O(j^{-1})$ .

#### 4.2. Application to protein interaction networks

We used the computer program developed for [18] to estimate the parameters under the partial duplication model for different protein interaction networks. The *Plasmodium falciparum* (*P. falciparum*) dataset is obtained from [32], and the remaining datasets are downloaded from the Database of Interacting Proteins (<http://dip.doe-mbi.ucla.edu>). Curiously, we note that according to Proposition 4.6, all pairs of  $p$  and  $q$  correspond to ergodic Markov chains, indicating that the networks stabilize as the number of vertices becomes large.

For one of the networks, *P. falciparum*, we conducted some further experiments where 50 networks were simulated with the same number of vertices as in *P. falciparum* (1271) and the degree distribution was computed. All simulations were started from an initial network of two vertices connected by an edge. Furthermore, 1271 runs of the corresponding Markov chain were performed, and the degree distribution was calculated and compared to the degree distribution obtained from the simulated networks. Here, the initial state of the Markov chain is 1. The length of the runs was varied, as shown in Figure 1.

The simulations indicate that the Markov chain approach may be used to approximate the degree distribution. This is particularly useful for simulation of large networks in terms of memory usage; storing the connections between vertices requires memory capacity proportional to the number of vertices times the average number of connections. Simulation of the

**Table 1:** Parameters estimated from protein interaction data.

Species	Vertices	Edges	$p$	$q$
<i>H. pylori</i>	675	1291	0.263	0.052
<i>P. falciparum</i>	1271	2642	0.026	0.789
<i>C. elegans</i>	2368	3767	0.315	0.105
<i>S. cerevisiae</i>	4968	17530	0.131	0.263

corresponding Markov chain requires memory capacity proportional to the current value of the chain.

The empirical degree distribution for *P. falciparum* shows that the partial duplication model does not provide a perfect fit. For example, no zero-degree vertices are included in the dataset (experimenter's choice), and this needs to be incorporated into the model.

## 5. Other models

We have applied the Markov chain approach to other models, and in this section we briefly present some of the results.

### 5.1. The duplication-divergence model

The *duplication-divergence model* is an extension of the partial duplication model, and it has been used for analysis of protein interaction networks as well [15, 16, 27, 33]. However, the model is slightly more complicated than the partial duplication model, and it has three parameters  $p$ ,  $q$ , and  $r$ . A step in this model is as follows: pick a vertex  $u$  in the graph uniformly, and add a new vertex  $v$ . Connect  $u$  and  $v$  with probability  $q$ , and create an edge  $e_w$  between  $v$  and  $w$  whenever there is an edge  $e'_w$  between the vertices  $u$  and  $w$ . Now modify the pairs  $(e_w, e'_w)$  independently of each other in the following way: with probability  $p$ , keep both edges; otherwise, with probability  $r$ , keep  $e_w$  and delete  $e'_w$ , and with probability  $1 - r$ , keep  $e'_w$  and delete  $e_w$ .

One can derive a master equation and go through the construction of the modified jump chain. In this case, the transition probabilities  $p_{j,k}$  satisfy that

$$(jp + 2)p_{j,k} = \begin{cases} (1 - q)b_j(k, 1 - \psi) + qb_j(k - 1, 1 - \psi) \\ \quad + (1 - q)b_j(k, p + \psi) + qb_j(k - 1, p + \psi) & \text{for } k \leq j, \\ jp + qb_j(j, 1 - \psi) + qb_j(j, p + \psi) & \text{for } k = j + 1, \end{cases} \quad (5.1)$$

and  $p_{j,k} = 0$  otherwise. Here  $\psi = (1 - p)(1 - r)$ , and  $b_j(k, s)$  is the binomial probability from (4.3) with  $p$  replaced by  $s$ .

In order to apply Theorem 4.1, we put  $x_j = \log(j + 1)$ . It follows from simple calculations, again using Jensen's inequality, that  $\{x_j\}_{j \geq 0}$  is a solution to (4.9) if  $p$  and  $r$  satisfy that

$$\log(1 - \psi) + \log(p + \psi) + p < 0. \quad (5.2)$$

Note that in the special cases  $r = 0$  and  $r = 1$ , the left-hand side of the inequality reduces to  $\log(p) + p$ , the same inequality as seen earlier. Actually, for  $r = 0$  the model is the partial duplication model. It follows that if  $r = 0$  or  $r = 1$ , a solution  $p$  of (5.2) must satisfy that  $p < \Omega$ . For  $0 < r < 1$  an exact upper bound on  $p$  is harder to derive. For these values of  $r$ , the solution  $p$  is less than  $\Omega$  and attains a minimum  $p \approx 0.5235$  for  $r = 1/2$ .

### 5.2. Another class of models

We believe that the Markov chain approach presented in this paper may be used to infer the behavior of other classes of models. In [21], simple models with master equations on the form

$$n_{t+1}(k) = \left(1 - \frac{a_k}{t}\right)n_t(k) + \frac{a_{k-1}}{t}n_t(k-1) + c_k, \quad (5.3)$$

where  $a_k$  and  $c_k$  are nonnegative numbers, are studied. The resulting master equation for the relative frequencies  $f_t(k)$  may be written in matrix form as

$$\frac{1}{t+1} \left( \begin{array}{c|c} t & \mathbf{1} \\ \hline \mathbf{c} & A(t) \end{array} \right) \begin{pmatrix} 1 \\ \mathbf{f}_t \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{f}_{t+1} \end{pmatrix}, \quad (5.4)$$

where  $\mathbf{1} = (1 \ 1 \ 1 \ \dots)$ , and  $\mathbf{c}$  and  $\mathbf{f}_t$  are the column vectors consisting of all the numbers  $c_k$  and  $f_t(k)$ , respectively. The matrix  $A(t)$  is given by

$$A(t) = \begin{pmatrix} t - a_0 & 0 & 0 & 0 & \cdots \\ a_0 & t - a_1 & 0 & 0 & \cdots \\ 0 & a_1 & t - a_2 & 0 & \cdots \\ 0 & 0 & a_2 & t - a_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (5.5)$$

Note that columns of the partitioned matrix in (5.4) sum to  $t + 1$ . That is, when divided by  $t + 1$ , the transpose of this matrix represents a Markov chain with time-dependent transition probabilities. We identify the countable set of states with  $\mathbf{N} \cup \{-\infty\}$  where the artificial state  $-\infty$  accounts for the first row and the first column in the partitioned matrix.

We may compute the corresponding jump process, and again it turns out that its transition probabilities  $p_{j,k}$  are time-independent. We may get rid of the state  $-\infty$  by simply forgetting the time we spend there. That is, for  $j, k \geq 0$ , we replace  $p_{j,k}$  by the sum  $p_{j,k} + p_{j,-\infty}p_{-\infty,k}$ , and this leads to a Markov chain with transition probabilities given by

$$p_{j,k} = \begin{cases} \frac{a_j + c_{j+1}}{1 + a_j} & \text{for } k = j + 1, \\ \frac{c_k}{1 + a_j} & \text{otherwise.} \end{cases} \quad (5.6)$$

These jump chains are in fact all ergodic, and the stationary distribution of the time-inhomogeneous Markov chains has been derived in [21].

### 5.3. Other extentions

Still other models do not fall under the conditions and assumptions introduced in this paper. For example, the master equation of the most general form of the duplication-mutation model [22, 23] depends on terms  $O(1/t^2)$ , and the columns of  $A(t)$  do not sum to the same number  $a(t)$  because of  $O(1/t^2)$  terms, and because the requirement  $A(t)_{k,j} = 0$  for  $k > j + 1$  is not fulfilled.

Some of these problems may be circumvented at the cost of a more technical and elaborate exposition, but often the results need to be stated as limiting results. For example, if the columns of  $A(t)$  do not sum to the same number, the jump chain in (2.10) should be considered as emerging in the limit  $t \rightarrow \infty$ .

Furthermore, one may choose to ignore terms of order  $O(1/t^2)$  in the master equation. As  $t \rightarrow \infty$ , the influence from higher-order terms often becomes insignificant, justifying such an approximation. This is, for example, the case for the duplication-mutation model.

## Acknowledgments

M. Knudsen is supported by the Centre for Theory in the Natural Sciences, University of Aarhus. C. Wiuf is supported by the Danish Cancer Society and the Danish Research Councils. They would like to thank an anonymous reviewer for valuable suggestions that improved the clarity of the paper.

## References

- [1] E. Alm and A. P. Arkin, "Biological networks," *Current Opinion in Structural Biology*, vol. 12, no. 2, pp. 193–202, 2003.
- [2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [3] Z. Burda, J. D. Correia, and A. Krzywicki, "Statistical ensemble of scale-free random graphs," *Physical Review E*, vol. 64, no. 4, Article ID 046118, 9 pages, 2001.
- [4] J. Cork and M. Purugganan, "The evolution of molecular genetic pathways and networks," *Bioessay*, vol. 26, no. 5, pp. 479–484, 2004.
- [5] T. Evans, "Complex networks," *Contemporary Physics*, vol. 45, no. 6, pp. 455–474, 2004.
- [6] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, no. 3, Article ID 036122, 8 pages, 2003.
- [7] J. Padgett, "Robust action and the rise of the medici," *American Journal of Sociology*, vol. 98, no. 6, pp. 1259–1319, 1993.
- [8] J. Scott, *Social Network Analysis*, Sage, Beverly Hills, Calif, USA, 2000.
- [9] E. de Silva and M. Stumpf, "Complex networks and simple models in biology," *Journal of the Royal Society Interface*, vol. 2, no. 5, pp. 419–430, 2005.
- [10] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.
- [11] B. Bollobas, *Random Graphs*, Academic Press, New York, NY, USA, 1998.
- [12] B. Bollobas and O. Riordan, "Mathematical results on scale-free graphs," in *Handbook of Graphs and Networks*, S. Bornholdt and H. Schuster, Eds., pp. 1–34, Wiley & Sons, New York, NY, USA, 2003.
- [13] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," in *From Biological Nets to the Internet and WWW*, Oxford University Press, Oxford, UK, 2003.
- [14] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
- [15] M. Middendorf, E. Ziv, C. Adams, et al., "Discriminative topological features reveal biological network mechanisms," *BMC Bioinformatics*, vol. 5, p. 181, 2004.
- [16] M. Middendorf, E. Ziv, and C. H. Wiggins, "Inferring network mechanisms: the *drosophila melanogaster* protein interaction network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 9, pp. 3192–3197, 2005.
- [17] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.
- [18] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf, "A likelihood approach to analysis of network data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 20, pp. 7566–7570, 2006.

- [19] F. Chung and L. Lu, *Complex Graphs and Networks*, vol. 107 of *CBMS Regional Conference Series in Mathematics*, American Mathematical Society, Providence, RI, USA, 2006.
- [20] R. Durrett, *Random Graph Dynamics*, vol. 20 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, New York, NY, USA, 2006.
- [21] O. Hagberg and C. Wiuf, "Convergence properties of the degree distribution of some growing network models," *Bulletin of Mathematical Biology*, vol. 68, no. 6, pp. 1275–1291, 2006.
- [22] A. Raval, "Some asymptotic properties of duplication graphs," *Physical Review E*, vol. 68, no. 6, Article ID 066119, 10 pages, 2003.
- [23] R. V. Solé, R. Pastor-Satorras, E. D. Smith, and T. Kepler, "A model of large-scale proteome evolution," *Advances in Complex Systems*, vol. 5, no. 1, pp. 43–54, 2002.
- [24] M. P. H. Stumpf, W. Kelly, T. Thorne, and C. Wiuf, "Evolution at the system level: the natural history of protein interaction networks," *Trends in Ecology & Evolution*, vol. 22, no. 7, pp. 366–373, 2007.
- [25] A. B. Bhan, D. J. Galas, and T. G. Dewey, "A duplication growth model of gene expression networks," *Bioinformatics*, vol. 18, no. 11, pp. 1486–1493, 2002.
- [26] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas, "Duplication models for biological networks," *Journal of Computational Biology*, vol. 10, no. 5, pp. 677–688, 2003.
- [27] O. Ratmann, O. Jørgensen, T. Hinkley, M. P. H. Stumpf, S. Richardson, and C. Wiuf, "Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*," *PLoS Computational Biology*, vol. 3, no. 11, p. e230, 2007.
- [28] R. L. Tweedie, "Sufficient conditions for regularity, recurrence and ergodicity of Markov processes," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 78, pp. 125–136, 1975.
- [29] E. Samuel-Cahn and S. Zamir, "Algebraic characterization of infinite Markov chains where movement to the right is limited to one step," *Journal of Applied Probability*, vol. 14, no. 4, pp. 740–747, 1977.
- [30] C. M. Harris and P. G. Marlin, "A note on feedback queues with bulk service," *Journal of the Association for Computing Machinery*, vol. 19, no. 4, pp. 727–733, 1972.
- [31] V. Romanovsky, "Note on the moments of a binomial  $(p + q)^n$  about its mean," *Biometrika*, vol. 15, no. 3-4, pp. 410–412, 1923.
- [32] D. J. LaCount, M. Vignali, R. Chettier, et al., "A protein interaction network of the malaria parasite *plasmodium falciparum*," *Nature*, vol. 438, no. 7064, pp. 103–107, 2005.
- [33] A. Wagner, "How the global structure of protein interaction networks evolves," *Proceedings of the Royal Society B*, vol. 270, no. 1514, pp. 457–466, 2003.