New Journal of Physics
The open-access journal for physics

Evolving networks through deletion and duplication

Nadia Farid[1] **and Kim Christensen**

Blackett Laboratory, Imperial College London, Prince Consort Road, London SW7 2BW, UK
E-mail: nadia.farid@imperial.ac.uk and k.christensen@imperial.ac.uk

New Journal of Physics **8** (2006) 212
Received 30 August 2006
Published 27 September 2006
Online at http://www.njp.org/
doi:10.1088/1367-2630/8/9/212

Abstract. We introduce a minimalistic model based on dynamic node deletion and node duplication with heterodimerization. The model is intended to capture the essential features of the evolution of protein interaction networks. We derive an exact two-step rate equation to describe the evolution of the degree distribution. We present results for the case of a fixed-size network. The results are based on the exact numerical solution to the rate equation which are consistent with Monte Carlo simulations of the model's dynamics. Power-law degree distributions with apparent exponents <1 were observed for generic parameter choices. However, a proper finite-size scaling analysis revealed that the actual critical exponent in such cases is equal to one. We present a mean-field argument to determine the asymptotic value of the average degree, illustrating the existence of an attractive fixed point, and corroborate this result with numerical simulations of the first moment of the degree distribution as described by the two-step rate equation. Using the above results, we show that the apparent exponent is determined by the heterodimerization probability. Our preliminary results are consistent with empirical data for a wide range of organisms, and we believe that through implementing some of the suggested modifications, the model could be well-suited to other types of biological and non-biological networks.

[1] Author to whom any correspondence should be addressed.

PII: S1367-2630(06)25317-5

Contents

1. Introduction **2**
2. DDDH model **4**
3. DDDH model: rate equation **5**
4. DDDH model: results **7**
4.1. Comparison between exact numerical solution of the rate equation and MC simulations . . . 7
4.2. Scaling for $p_{\mathrm{del}} = p_{\mathrm{dupl}} = p = 1, 0 < \theta \leqslant 1/2$ . . . 8
4.3. Mean-field equation for the average degree . . . 11
4.4. Effects of varying θ . . . 14
4.5. Effects of varying p . . . 14
4.6. Comment on relation to biological data . . . 15
5. Discussion **15**
5.1. Extensions . . . 16
6. Conclusions **17**
Acknowledgments **17**
References **17**

1. Introduction

In recent years, models of networks growing via single-node duplication, divergence and mutation of links, considered in isolation and combination, have assumed prominence in the literature on complex networks. In a series of independent studies, it was suggested that these duplication–divergence–mutation models (hereafter called 'duplication models' for brevity) are good candidates to describe the evolution and large-scale topological features of real protein–protein interaction networks (PINs) in several organisms such as *S. cerevisiae* and *H. pylori* [1]–[6].

In duplication models, proteins are represented by nodes, and a pairwise interaction between any two proteins is represented by an undirected link between the associated nodes, assumed to be fully operative at all times. In a duplication event, a mother node is chosen uniformly at random (u.a.r.) and each of its links are copied to a newly created daughter node. Divergence refers to the subsequent loss of links from the daughter node [2], [4]–[7], and/or the mother node [8], or a shared neighbour of both the daughter and mother node [3]. Often, for simplicity, it is assumed that only the daughter node diverges. In a mutation event, new links are added between the daughter node and all other nodes in the network which are not already connected to the mother node. Typically, one duplication–divergence event occurs at each update step, and mutations, when considered, are modelled at a rate much less than the divergence rate—typically, one new link per update step is added [4]–[6].

The idea of evolution through gene duplication is taken from biology [9]–[12], where it was popularized in the 1970s by Ohno [13] who conjectured that single and whole genome duplications could provide the raw material for evolutionary diversification. While there is mounting evidence that duplicate genes do occur in genomes [1, 12], [14]–[17], it is widely acknowledged that little is known about the details of the process of duplication itself, such as

the frequency of duplication events, the fate of duplicate genes, and the frequency with which duplicate genes acquire novel functions [1, 9, 11, 18]. In fact, while the microscopic parameters are not yet known exactly, it has been suggested by Berg *et al* [19] that the rate of gain and loss of interactions through mutations is at least an order of magnitude higher than the growth rate due to duplications; as a result, the link dynamics act as the dominant evolutionary force shaping the large-scale statistical structure of the network, and not the gene duplications; the slower gene duplication events only really affect the size of the network. Thus, it is still unclear if, and to what extent, gene duplication is the dominant mechanism responsible for the observed statistical features of PINs. Moreover, gene deletion and rearrangement are known to play important roles in the long-term evolution of genomes [10, 17, 20] but have received comparatively little attention in the literature on networks addressing PINs [20, 21].

In this paper, in a move to go beyond the duplication models and to expand upon the emerging literature on network models of PINs including gene deletion, we present a four parameter model addressing the scenario of network evolution through dynamic total node removal in conjunction with growth by node duplication and heterodimerization. We refer to our network model as a *deletion–duplication–divergence–heterodimerization* (DDDH) model (we differentiate between total node removal from the network and removal via the merging of nodes which has been considered elsewhere, see for example [22, 23]). Although the model is primarily aimed at describing the evolution of PINs, its description is kept general enough so as to be applicable to a diverse range of complex systems where components are added and removed throughout the system's evolution.

In the wider literature on complex networks, previous studies which have considered node deletion have generally regarded it as a perturbation effect, used to test the tolerance of a network to random and targeted attack [3], [24]–[26]. More recently, the mechanism of dynamic node removal in conjunction with growth by preferential attachment has been explored in independent studies by Chung and Lu [27], Cooper *et al* [28], and Wang [29]. They refer to such models as *growth-deletion* models. Since we consider growth by duplication as opposed to growth by preferential attachment, our model therefore also contributes to the literature on growth-deletion models.

We focus our analysis on the degree distribution, $P(k)$, characterizing the probability for a node to have exactly k links [30]–[32]. The degree distribution is the simplest topological feature to measure and as a result, it has attracted and received the most attention in the literature on complex networks. It has been shown [2]–[4], [6]–[8] that the degree distribution of networks generated by duplication models exhibits a power-law tail, $P(k) \sim k^{-\gamma}$, for $k \gg 1$, where the critical exponent γ can be tuned such that it is in agreement with observed exponents which are found to be in the range $1 < \gamma < 3$ [2, 33]. Importantly, it has also been shown that the degree distribution is a robust and generic property of PINs common across different data sets—an important consideration given that current experimental techniques are notorious for suffering from a high rate of false positives and false negatives [19, 33].

The paper is organized as follows. In section 2, we present the formulation of the dynamics which describe the rules of evolution of the network, defining the parameters and interpreting the rates. In section 3, we present and discuss the exact two-step rate equation for the evolution of the degree distribution. In section 4, we present results obtained from Monte Carlo (MC) simulations of the model and the exact numerical solution of the rate equation for a generic choice of parameter values. We discuss our results in section 5 and end with a conclusion in section 6.

2. DDDH model

We consider undirected networks where loops and multiple links are forbidden. We start with a network of known size, N, and degree distribution, $P(k, t = 0)$, and allow it to evolve under the following rules (see figure 1):

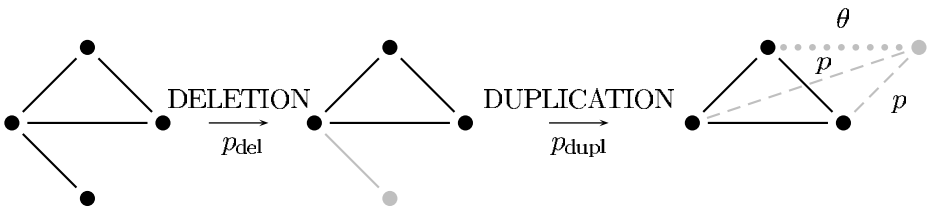

Figure 1. A schematic representation of a network evolving through deletion and duplication. (a) A node is chosen u.a.r. and deleted with probability p_{del} (grey). (b) A mother node is chosen u.a.r. and duplicated with probability p_{dupl} (green $\rightarrow$ red). The links are retained with probability p (dashed red line), and a further link (c) is established between the daughter and mother node with probability θ (dotted blue line).

1. *Deletion.* With probability p_{del}, a node is chosen u.a.r. This node and all of its links are deleted from the network.
2. *Duplication–Divergence–Heterodimerization.* With probability p_{dupl}, a mother node is chosen u.a.r. and duplicated. This entails a new daughter node being added to the network and linking to each of the neighbours of the mother node with probability p. A further link is established between the daughter and mother node with heterodimerization probability θ.

The evolution of the network is thus governed by four parameters: p_{del}, p_{dupl}, p, θ. In section 3, we cast these rules into a concrete mathematical framework. First, however, it is instructive to discuss the motivation behind the choice of dynamics and its implications.

The mechanism of growth by duplication is preferred to the mechanism of growth by preferential attachment in PIN network models as well as many other network models since it reproduces the effects of preferential attachment without having to artificially put the mechanism in. In other words, it arises naturally from the dynamics: nodes that have a large number of links are more likely to be neighbours of a duplicating node, and hence are more likely to gain a link to the newly created node. The DDDH model preserves this effect and introduces another one: implicit preferential *detachment*. Nodes that have a large number of links are more likely to be the neighbour of a node chosen for deletion, and therefore will be more likely to lose a link each time a node is deleted. It is interesting to note that these two effects do not cancel out each other, as one might intuitively expect. The inclusion of heterodimerization, $\theta > 0$, means that we do not consider mutations in the traditional sense (as described above), but rather we restrict the addition of new links to only occur between the mother and the daughter node. Heterodimerization is preferred over mutations as it has been noted that the former increases the likelihood of clique-formation—a feature observed in real PINs—while the latter, in order to form the observed

number of triads and higher cliques, would require a prohibitively/physically unrealistic high rate [8]. Moreover, in duplication models, whilst a lack of random linking (mutations) has been found to destroy fine structure such as the self-averaging and existence of a smooth degree distribution [4], the large-scale statistical features of the final network do not depend on the existence of mutations [3]–[5].

From the general definition of the model's dynamics given above, one can already determine specific features of the network, namely its overall size, by focusing on particular choices of the model's parameter values. For example, for a network that remains, on average, fixed in size, $p_{\mathrm{del}} = p_{\mathrm{dupl}}$. We will refer to this case as a 'fixed-size' network. For a network that is monotonically growing, on average, $p_{\mathrm{dupl}} > p_{\mathrm{del}}$; if $p_{\mathrm{del}} = 0$ no nodes are removed from the network. Moreover, if we fix $\theta = 0$ and $0 < p < 1$ and $p_{\mathrm{dupl}} = 1$ for this case, the DDDH model is equivalent to the duplication–divergence model [5, 6, 8, 34]. For $p_{\mathrm{del}} = 0$, $p_{\mathrm{dupl}} = 1$, $\theta = 0$ and $p = 1$ the duplication–divergence model is equivalent to Polya's Urn [2, 8, 35]. For an on average monotonically shrinking network, $p_{\mathrm{dupl}} < p_{\mathrm{del}}$; if $p_{\mathrm{dupl}} = 0$ no new nodes or links are added to the network (for an appropriately chosen initial network size one could imagine that this regime could be used to compare the results of perturbation effects with the results of gene knock-out experiments). For a network fluctuating in size the values of p_{del} and p_{dupl} would be stochastic variables chosen anew at each update step from a suitably chosen distribution (a similar mechanism has been considered by Slanina *et al* [36]). If $p = 1$, the daughter node inherits all of the links; this is the case of perfect, or full duplication. If $0 < p < 1$, the daughter node inherits only some of the links; this represents imperfect, or partial duplication. If $p = 0$, the daughter node inherits none of the links, that is, an isolated node is added to the network.

In this paper, we study the case of fixed-size networks, $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$, with heterodimerization $\theta > 0$ and perfect duplication $p = 1$, unless otherwise stated (in which case, similar to [2, 4, 6] we assume that only the daughter node diverges) (it is worth pointing out that originally, we set out to present the case of a growing network, evolving under dynamic node deletion and node duplication with heterodimerization, and compare these results to those obtained from duplication models. However, since we found the results for a fixed size network particularly striking and unusual, we have restricted our results in this paper to this special case; we defer the results for the case of a growing network to a further publication).

In section 3, we apply the rate equation approach [4, 37] to study the evolution of the expected number of nodes with k links at time t, $f(k, t) = N_t P(k, t)$ where N_t is the total number of nodes at time t.

3. DDDH model: rate equation

In this section, we present a two-step rate equation for the general DDDH model, and then explain in detail the origin of each term.

The two-step rate equation for the DDDH model is given by,

$$f(k, t+1) = f(k, t) - p_{\mathrm{del}} \frac{f(k, t)}{N_t} - p_{\mathrm{del}} \frac{k f(k, t)}{N_t} + p_{\mathrm{del}} \frac{(k+1) f(k+1, t)}{N_t}, \qquad (1a)$$

$$f(k,t+2) = f(k,t+1) - p_{\rm dupl}p\frac{kf(k,t+1)}{N_{t+1}} - p_{\rm dupl}\theta\frac{f(k,t+1)}{N_{t+1}} + p_{\rm dupl}p\frac{(k-1)f(k-1,t+1)}{N_{t+1}}$$
$$+p_{\rm dupl}\theta\frac{f(k-1,t+1)}{N_{t+1}} + p_{\rm dupl}\theta\sum_{j\geqslant(k-1)}\binom{j}{k-1}p^{(k-1)}$$
$$\times(1-p)^{j-(k-1)}\frac{f(j,t+1)}{N_{t+1}} + p_{\rm dupl}(1-\theta)\sum_{j\geqslant k}\binom{j}{k}p^k(1-p)^{j-k}\frac{f(j,t+1)}{N_{t+1}}. \quad (1b)$$

Equation (1) is *exact* and there are no approximations in its derivation. It describes the DDDH process for all parameters, $p_{\rm del}$, $p_{\rm dupl}$, p, θ. It holds for all $k \geqslant 0$, with $f(-1,t) = 0$ and $f(k > k_{\max}, t) = 0$ for all t. Moreover, $f(k,t)$ satisfies the following normalization conditions:

$$\sum_{k=0}^{\infty} f(k,t) = N_t, \quad (2a)$$

$$\sum_{k=0}^{\infty} kf(k,t) = 2L_t, \quad (2b)$$

where L_t is the total number of links at time t in the network.

The distinct feature of the DDDH rate equation compared to rate equations for duplication models is that it is defined in two steps, and not one, reflecting the fact that we now include a node deletion in addition to node duplication. To keep the notation simple, we have written $t+1$ and $t+2$ in equation (1) but one might have equally well written $t+1/2$ and $t+1$ to indicate that we *only* observe the network after *both* the deletion and the duplication steps have been completed.

Moreover, each of these actions is executed sequentially, highlighting the fact that we clearly also make the distinction between this process and the process of adding nodes at an 'effective' duplication rate, $p_{\rm dupl} - p_{\rm del}$, or merging nodes as is [22, 23], or substituting nodes where the duplication of a node automatically implies the deletion of some other node as in [20]. We note that the exact correspondence between time, t, and real biological timescales is unclear, however, at this stage.

In equation (1*a*) we consider the effects on the network of the deletion of a node and the removal of its links. The probability a deletion event occurs is given by $p_{\rm del}$. The terms on the right-hand side (RHS) are interpreted as follows. A *loss* in the number of nodes with degree k at time $t+1$ from deletion will occur either if the node deleted is of degree k at time t, or if a neighbour (of arbitrary degree) of a k-node is chosen for deletion, as the k-node will lose a link and become a node of degree $k-1$. Since every node has an equal probability of being deleted in a given time step, a node of degree k is chosen with probability $f(k,t)/N_t$ (second term); the probability that a neighbour of a k-node is chosen for deletion is $kf(k,t)/N_t$ (third term). The final term on the RHS represents a *gain* in the number of nodes with degree k at time t. This can occur if the neighbour of a node with degree $k+1$ is deleted. Given that a node of degree $k+1$ has $k+1$ neighbours, the probability a neighbour is chosen for deletion is $(k+1)f(k+1,t)/N_t$.

In equation (1*b*), we consider the effects of node duplication and subsequent hetero-dimerization. The probability a duplication event occurs is given by $p_{\rm dupl}$, and the probability

that subsequent heterodimerization occurs is given by θ. All but the two final terms on the RHS represent changes a duplication–heterodimerization event has on the existing nodes in the network; the last two terms represent the daughter node's contribution.

A *loss* in the number of nodes with degree k at time $t+2$ from duplication and heterodimerization can occur in one of two ways. If one of the neighbours (of arbitrary degree) of a node of degree k at time $t+1$ is duplicated, the k-node will, with probability p gain a link from duplication thus becoming a $k+1$ node at time $t+2$ (second term). Alternatively, if the mother node to be duplicated is already of degree k at time $t+1$, it will become a node with degree $k+1$ at time $t+2$ by gaining a link to the daughter node via heterodimerization. The probability a node of degree k is chosen for duplication is given by $f(k, t+1)/N_{t+1}$, and the probability of heterodimerization is θ (third term).

We arrive at the fourth and fifth terms which describe the *gain* in the number of nodes with degree k at time $t+2$ by similar considerations. If one of the neighbours of a node of degree $k-1$ at time $t+1$ is chosen for duplication, the $k-1$ node will, with probability p, gain a link from duplication, thus becoming a k node at time $t+2$ (fourth term). Alternatively, if the mother node to be duplicated is of degree $k-1$ at time $t+1$, it will become a node with degree k at time $t+2$ by gaining a link to the daughter node via heterodimerization (fifth term).

The final two terms on the RHS of equation (1*b*) account for the daughter node's contribution. The sixth term is to account for the contribution of the daughter node in the event it does establish a link, with probability θ, to the mother node. The seventh term is to account for the case where it does not, which happens with probability $(1-\theta)$. Note the lower limits on these sums are not identical. This is because if a link is established via heterodimerization, in order to become a node of degree k, the daughter node is restricted to copying $k-1$ out of j links, each with probability p. However, if such a link is not established, the daughter node is restricted to copying k out of j links of the mother node.

Since the exact analytical solution to the rate equation in equation (1) is not tractable at present, in section 4, we present results obtained from the numerical solution of the exact two-step rate equation and compare them to MC simulations of the model. Unless otherwise stated, our analysis is based on the case of a fixed-size network, $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$, evolving through perfect duplication, $p = 1$, with heterodimerization, $\theta > 0$.

4. DDDH model: results

4.1. Comparison between exact numerical solution of the rate equation and MC simulations

Figure 2 displays the evolution of the degree distribution obtained from the exact numerical solution of the rate equation compared to MC simulations of the model. The stationary regime is defined by $P(k, t) = P(k, t+2)$.

We start with an initial network of $N = 400$ nodes, with a random degree distribution centred around $k_{\mathrm{init}} = 100$, and iterate the rules with the following parameter settings: $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$ (fixed-size network), $p = 1$ (perfect duplication), with heterodimerization $\theta = 0.1$. The value for θ was chosen as such as it is believed that heterodimerization occurs at a rate not greater than 0.1 [8]. In figure 2, we show two snapshots of the network in the transient regime when $t = 1000$, 5000, respectively, and one in the stationary regime for $t = 10^6$. The MC simulations are averaged over 10^5 realizations for $t = 1000,\ 5000$ and 3×10^3 realizations for $t = 10^6$.

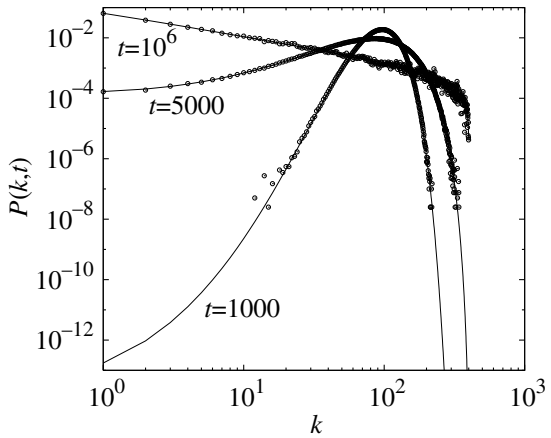

Figure 2. Exact numerical (—) and MC (○) results of the degree distribution of a fixed-size network, $p_{\rm del} = p_{\rm dupl} = 1$, with $N = 400$ nodes, evolving under perfect duplication, $p = 1$, with heterodimerization $\theta = 0.1$. Snapshots were taken at times $t = 1000$, 5000 (transient regime) and $t = 10^6$ (stationary regime). The MC simulations are averaged over 10^5 realizations for $t = 1000$, 5000, and 3×10^3 realizations for $t = 10^6$. The exact numerical results show excellent agreement with the MC simulations. The distribution in the stationary regime is well approximated by a power-law decay, $P(k) \sim k^{-\gamma}$ with an apparent exponent $\gamma = 0.8$ and sharp cutoff at $k_{\rm cutoff} = 399$. Note that the maximum degree a node can attain in networks of the type we consider is $N - 1$.

There is excellent agreement between the exact numerical solution and MC results, lending support to our statement earlier that there are no approximations involved in our derivation of the rate equation. We have verified through extensive simulations that the agreement holds true over a range of $p_{\rm del}$, $p_{\rm dupl}$, p, and θ values. Hence, all remaining figures are generated using data obtained from the exact numerical solution of the rate equation only, hereafter referred to as 'exact numerical results'. The interesting feature to note is that even for the simplest realization of the DDDH model which we have presented in figure 2, fat-tailed degree distributions are obtained in the stationary regime ($t = 10^6$ curve). This is in stark contrast to the duplication models where only the case of imperfect duplication leads to a power law [2]. In the following section, we describe quantitatively the exact form of the stationary degree distribution.

4.2. Scaling for $p_{\rm del} = p_{\rm dupl} = p = 1, 0 < \theta \leqslant 1/2$

We are interested in quantifying the form of the degree distribution, in the stationary regime, as a function of the model's parameters. Given that we are, for the moment, investigating fixed-size networks evolving under perfect duplication, $p_{\rm del}$, $p_{\rm dupl}$ and p are all fixed. This reduces the number of variables to just one: θ. However, given that we are observing the degree distribution for specific fixed network sizes, we have N as another variable in the problem. Hence, we would like to know how $P(k)$ depends on θ and N. In order to investigate this, we have performed numerical simulations for the following two cases: (i) fixed θ, varying N, and (ii) fixed N, varying θ. We discuss (i) in this subsection, and (ii) in subsection 4.3.

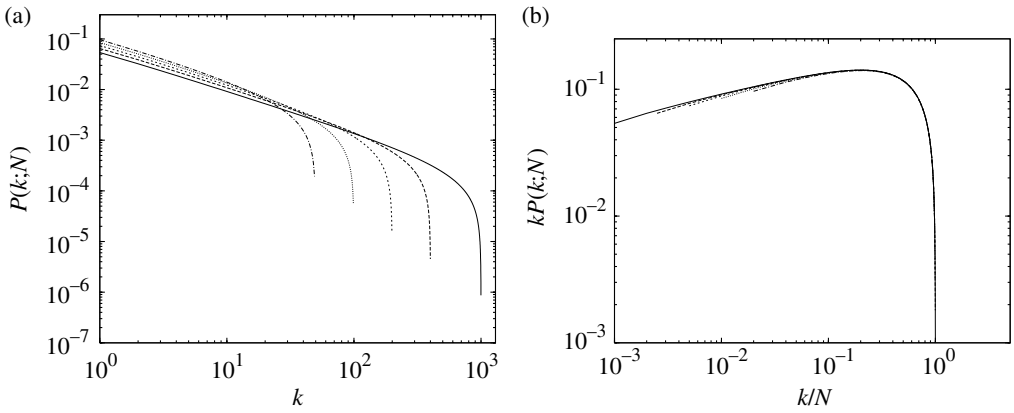

Figure 3. (a) Exact numerical results of the degree distribution in the stationary regime for fixed-size networks, $p_{\rm del} = p_{\rm dupl} = 1$, where $N = 50, 100, 200, 400, 1000$ (marked with lines of increasing dash-length) each having evolved through perfect duplication $p = 1$ with heterodimerization $\theta = 0.1$. There is no typical node-degree. For large node-degrees, the degree distribution is well-approximated by a power-law decay, $P(k; N) \sim k^{-\gamma}$, with an apparent exponent $\gamma = 0.8$. The power-law is characterized by a sharp cutoff at $k_{\rm max} = N - 1$, which increases with increasing system size. Note that the maximum degree a node can attain in networks of the type we consider is $N - 1$. (b) Data collapse of the exact numerical results of the degree distribution is obtained by plotting the transformed probability density $kP(k; N)$ versus the rescaled degree k/N. The curves collapse on to the graph of the scaling function, $\tilde{\mathcal{G}}(k/N) = \frac{1}{\Gamma(1-\gamma)\Gamma(1+\gamma)}(k/N)^{1-\gamma}(1 - k/N)^{\gamma}$, see equation (10).

Figure 3(a) shows the exact numerical results of the degree distribution, $P(k; N)$ versus k on a double logarithmic plot in the stationary regime for $\theta = 0.1$ and networks of increasing N, specifically, $N = 50, 100, 200, 400, 1000$ nodes. There are clear power-law fluctuations in the node degrees present in the network, implying an appreciable probability of finding a node with degree, $1 \leqslant k \ll k_{\rm max}$ in the network. $k_{\rm max}$ marks the cross-over between a power-law decay and a rapid decay in $P(k; N)$. In particular, $k_{\rm max}$ represents a characteristic scale in the node degree resulting from the finite size of the networks we can study numerically. Hence, we can say that $P(k; N)$ decays as a power-law for $1 \leqslant k \ll k_{\rm max}$ and has a sharp cutoff for $k \gg k_{\rm max}$, which can be expressed informally as,

$$P(k; N) \propto \begin{cases} k^{-\gamma} & 1 \leqslant k \ll k_{\rm max} \\ \text{sharp cutoff}, & k \gg k_{\rm max}. \end{cases} \tag{3}$$

From simulations, we find that $k_{\rm max} = N - 1$ which is equivalent to the maximum possible degree that a node in the network can acquire. This implies that $k_{\rm max}$ increases linearly with increasing network size, N, hence, in the limit of $N \to \infty$, the characteristic scale diverges and a pure power-law is recovered, as expected. In the region, $1 \leqslant k \ll k_{\rm max}$, we find that the gradient of the lines in figure 3(a) are well-approximated by an 'apparent' exponent, $\gamma = 0.8$ and we will shortly demonstrate that $\gamma = 1 - 2\theta$. Generally speaking, an exponent less than 1 is unusual,

and seems to contradict certain known results about scaling functions. However, we will be able to resolve this apparent contradiction in subsection 4.3.

With the above discussion in mind, we propose the following general ansatz for $P(k; N)$,

$$P(k; N) = a(N)k^{-\gamma}\mathcal{G}(k/N), \tag{4}$$

and, assuming for simplicity that $k_{\max}$ is approximated by N, the equation is valid for $1 \leqslant k \leqslant N$, where $N \gg 1$ and $\gamma < 1$. In equation (4), $a(N)$ is a prefactor, dependent on the network size, and $\mathcal{G}(x)$ is the cutoff function, dependent on the rescaled variable k/N. $\mathcal{G}(x)$ is required to fall-off fast enough to ensure $P(k; N)$ is finite and integrable. From figure 3(a), it seems reasonable to assume that $\mathcal{G}(x)$ is constant for $x \ll 1$ and decays abruptly for $x \gg 1$.

Note that equation (4) is *not* a finite-size scaling (FSS) ansatz since the prefactor, rather than being a constant (as is typical), is N-dependent. This difference turns out to be important as it leads to an interesting result about the critical exponent which we demonstrate below.

In theory, we can use the fact that the probability density function must be properly normalized,

$$\int_1^\infty P(k; N)\,\mathrm{d}k \equiv 1, \tag{5}$$

to derive an expression for $a(N)$. However, without knowing *a priori* the correct scaling for $P(k; N)$, we can only guess at the form of $\mathcal{G}(x)$. For simplicity, we assume that the cutoff function, $\mathcal{G}(x)$, is of the form,

$$\mathcal{G}(x) = \begin{cases} (1 - k/N)^\gamma & \text{for } 1 \leqslant k \leqslant N \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

(which is in-keeping with our stated requirements for the form of $\mathcal{G}(x)$ and is an excellent fit to the numerics). Substituting this into equation (4), we find that normalization requires,

$$\int_1^N a(N)k^{-\gamma}(1 - k/N)^\gamma\,\mathrm{d}k \equiv 1. \tag{7}$$

Evaluating the left-hand side (LHS) of equation (7) we find for $N \gg 1$,

$$a(N)N^{1-\gamma}\Gamma(1 - \gamma)\Gamma(1 + \gamma) = 1, \tag{8}$$

and it immediately follows that,

$$a(N) = \frac{N^{\gamma-1}}{\Gamma(1 - \gamma)\Gamma(1 + \gamma)}. \tag{9}$$

Using this result for $a(N)$, we can recast equation (4) into a scaling ansatz such that an *actual* critical exponent equal to 1 is obtained, as follows,

$$\begin{aligned} P(k; N) &= \frac{N^{\gamma-1}}{\Gamma(1 - \gamma)\Gamma(1 + \gamma)}k^{-\gamma}\mathcal{G}(k/N) \\ &= k^{-1}\frac{1}{\Gamma(1 - \gamma)\Gamma(1 + \gamma)}\frac{k^{1-\gamma}}{N^{1-\gamma}}\mathcal{G}(k/N) \\ &= k^{-1}\tilde{\mathcal{G}}(k/N), \end{aligned} \tag{10}$$

where $\tilde{\mathcal{G}}(x) = \frac{1}{\Gamma(1-\gamma)\Gamma(1+\gamma)} x^{1-\gamma}\mathcal{G}(x)$. Equation (10) is our 'proper' FSS ansatz. We have shown using consistent arguments that equation (4) can be recast into equation (10) assuming the cutoff function, $\mathcal{G}(x)$ is of the form given in equation (6), and using the requirement that the probability density function must be properly normalized to derive an expression for the N-dependent prefactor, $a(N)$. The point of interest is that on the LHS of equation (10) the leading power-law term has attained a fixed value equal to 1, independent of γ. Thus, even if the apparent measured exponent, γ, is in the range (0, 1) the actual critical exponent is always equal to 1.

To test the validity of the scaling ansatz given in equation (10), we have, in figure 3(b), plotted the transformed probability density $kP(k; N)$ versus the rescaled variable k/N using the same data as in figure 3(a). Multiplying both sides of equation (10) by k we get,

$$kP(k; N) = \tilde{\mathcal{G}}(k/N). \tag{11}$$

Therefore, we expect the curves to collapse on to the curve $\tilde{\mathcal{G}}(k/N) = \frac{1}{\Gamma(1-\gamma)\Gamma(1+\gamma)}(k/N)^{1-\gamma}(1 - k/N)^{\gamma}$, with the gradient of the slope to be equal to $1 - \gamma$. As shown in figure 3(b), a convincing data collapse is obtained, with all curves collapsing on to the scaling function described by $\tilde{\mathcal{G}}(x)$ with $\gamma = 1 - 2\theta = 0.8$. We have repeated the data collapse for different values of θ, and observed a convincing data collapse in all cases, with all curves collapsing on to the scaling function described by, $\tilde{\mathcal{G}}(x)$ with $\gamma = 1 - 2\theta$.

Together, equation (4) and the success of the data collapse in figure 3(b) demonstrate that the degree distribution for any network size, N, is determined by the scaling function $\tilde{\mathcal{G}}(x)$. This means that we can deduce the degree distribution for any network size, N, without having to actually perform the numerical simulation itself. Hence, our results are applicable to networks larger than those which we have demonstrated directly, that is for $N > 10^3$. This is in contrast to the duplication models considered in [8] where the networks generated do not attain power-law degree distributions even for very large networks.

Thus far, we have not yet justified the relation given between the apparent exponent and the parameter θ. In the following section, we derive a result for the average degree, $\langle k \rangle$. We then demonstrate how we can use this result to find an expression for the apparent exponent γ in terms of the parameter, θ.

4.3. Mean-field equation for the average degree

The average degree, $\langle k \rangle$, can be determined in various different ways. Ideally, one would be able to calculate it directly from the two-step rate equation for the evolution of the degree distribution in equation (1*a*). This would be achieved by taking the first moment of the normalized degree distribution, $P(k, t) = f(k, t)/N_t$, according to,

$$\langle k \rangle_t = \int_1^{\infty} kP(k, t)\, \mathrm{d}k. \tag{12}$$

Strictly speaking, where we write integration signs we should have sums, as we are dealing with a discrete probability distribution. Either way, the solution we are interested in is $lim_{t\to\infty}\langle k \rangle_t = \langle k \rangle$, which is analytically intractable. In order to get around this problem we have calculated this quantity numerically, and compared this result to the asymptotic value of $\langle k \rangle$ obtained as a solution to a mean-field rate equation approach (described below).

Figure 4 illustrates the exact numerical results for the time-evolution of $\langle k \rangle$ for $\theta \in [0.01, 0.5]$ and clearly illustrates the existence of a stationary asymptotic $\langle k \rangle$. We now describe

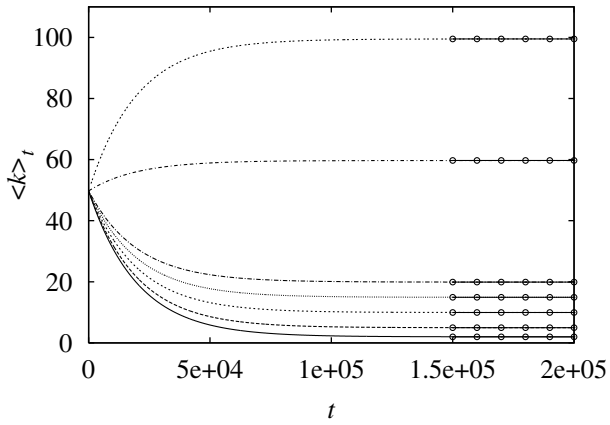

Figure 4. Exact numerical results for the average degree, $\langle k\rangle_t$, versus time for a fixed-size network, $p_{\text{del}} = p_{\text{dupl}} = 1$, with $N = 200$ nodes, and $k_{\text{initial}} = 50$. The network evolved through perfect duplication and increasing $\theta = 0.01$, 0.025, 0.05, 0.075, 0.1, 0.3, 0.5 (marked with lines of decreasing dash-length). In each case, $\langle k\rangle_t$ reaches a stationary value whose value is identical to that predicted analytically ('$\circ$'), $\langle k\rangle_{\text{MF}} = \theta(N-1)$, see equation (13).

our mean-field argument to determine the asymptotic value of the average degree, $\langle k\rangle$. At each time step, for each node we delete we lose on average, $\langle k\rangle_t$ links, and for each node we duplicate we gain on average, $p\langle k\rangle_{t+1} + \theta$ links. Therefore, the net change in the number of links, $\Delta L = -\langle k\rangle_t + p\langle k\rangle_{t+1} + \theta$. For the case of $p = 1$, we can rewrite this as, $\Delta L = -2L_t/N_t + 2L_{t+1}/N_{t+1} + \theta$, where we have used the standard relation, $\langle k\rangle = 2L/N$. Imposing the condition $\Delta L = 0$, which is valid in the stationary regime, $t \to \infty$, we find,

$$\langle k\rangle_{\text{MF}} = \theta(N-1) \tag{13}$$

for a fixed-size network evolving under perfect duplication for arbitrary θ. So, for a network of size $N = 200$, for example, we predict, using equation (13), $\langle k\rangle = 1.99, 19.9, 99.5$ for $\theta = 0.01, 0.1, 0.5$ respectively. We can compare this prediction with the exact numerical results for the first moment of the degree distribution. As shown in figure 4, there is exact agreement between the asymptotic value of the average degree determined from the mean-field calculation and the exact numerical results (for $t > 10^5$). Thus, both the mean-field calculation and the exact numerical results, as illustrated in figure 4, demonstrate the existence of an attractive fixed point in the average degree, $\langle k\rangle$. This is clearly related to the existence of a stationary degree distribution, that is, $P(k, t) = P(k, t+2)$.

Equation (13) highlights the importance of accounting for the mechanism of hetero-dimerization for finite-sized networks, as it shows that the network self-organizes to a stationary state where the average degree $\langle k\rangle$ is constant, and determined by the system size, N *and* the probability for heterodimerization, θ. Since real PINs are of finite size, typically with no more than 10^4 nodes (see table 1), the point of information regarding the role of θ in finite-sized networks is of significance.

We have also verified through further simulations varying p such that $p < 1$ for constant θ and N clearly dramatically reduces $\langle k\rangle$, as one would expect, although the precise nature of this

Table 1. Comparison between empirical data [8, 38] and the fixed-size, perfect duplication DDDH model. Values of θ and γ are quoted to two decimal places.

Data	N	$\langle k \rangle$	θ	γ
Yeast (I)	4873	6.6	0.0014	1.0
Yeast (II)	5397	29.2	0.0054	1.0
Fly	6954	5.9	0.00085	1.0
Human	5275	5.7	0.0011	1.0

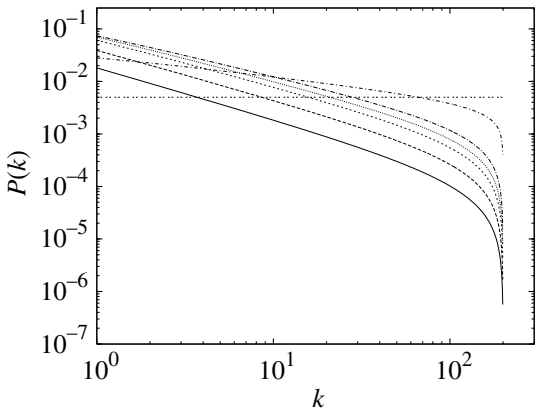

Figure 5. Exact numerical results for the degree distribution in the stationary regime for a fixed-size network, $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$, with $N = 200$ nodes, and $k_{\mathrm{initial}} = 50$. The network evolved through perfect duplication and increasing $\theta = 0.01, 0.025, 0.05, 0.07, 0.1, 0.3, 0.5$ (marked with lines of decreasing dash-length). The value $\theta = 0.5$ marks a change in behaviour from a power-law decay with a negative exponent to a uniform distribution.

effect has not yet been quantified and seems to be non-trivial. Thus, equation (13) actually gives an upper bound for $\langle k \rangle$.

An alternative method for calculating $\langle k \rangle$ analytically, is to calculate the first moment of the degree distribution as expressed in equation (10), $\langle k \rangle_{\mathrm{SF}}$. This turns out to be very useful as far as determining an expression for the apparent exponent, γ, in terms of θ. We find that,

$$\begin{aligned}
\langle k \rangle_{\mathrm{SF}} &= \int_1^N k P(k;\, N)\, \mathrm{d}k \\
&= \int_1^N \frac{1}{\Gamma(1-\gamma)\Gamma(1+\gamma)} \left(\frac{k}{N}\right)^{1-\gamma} \left(1 - \frac{k}{N}\right)^{\gamma} \mathrm{d}k \\
&= \frac{N}{2} \frac{\Gamma(2-\gamma)}{\Gamma(1-\gamma)} \quad \text{for } N \to \infty \\
&= \frac{N}{2}(1-\gamma). \qquad (14)
\end{aligned}$$

Since we already know that $\langle k \rangle_{\mathrm{MF}} = \theta(N-1)$ from equation (13), the RHS of equation (14) must be equivalent to equation (13), hence,

$$\gamma = 1 - 2\theta. \tag{15}$$

This justifies our previous finding in subsection 4.2 that the apparent exponent is $\gamma = 1 - 2\theta$. In the following section, we investigate the effect on the degree distribution of varying $\theta \in (0, 0.5)$ for fixed N, completing the analysis of the two scenarios outlined at the beginning of subsection 4.2.

4.4. Effects of varying θ

Figure 5 illustrates the topological effect of varying the probability to heterodimerize in the range $0 < \theta \leqslant 0.5$, in a fixed-size network ($N = 200$), evolving through perfect duplication.

We find that the degree distribution exhibits a power-law with a slope that varies with θ according to $\gamma = 1$–2θ, reaching a uniform distribution at a value of $\theta = 0.5$. We have repeated the above simulations for a range of network sizes, up to $N = 1000$, and confirmed this behaviour. Hence, the result for γ is consistent with our previous findings in subsection 4.2.

Current estimates from empirical data for yeast, fly and human PINs indicate that θ never exceeds 0.1 [8]. Since we observe power-law degree distributions in the range $0 < \theta < 0.5$, a value of $\theta < 0.1$ in our model is consistent with empirical data. The fact that $\gamma = 1 - 2\theta$ might go some way towards explaining why in the duplication model considered in [8], for realistic values of $\theta < 0.1$ their results were not affected.

4.5. Effects of varying p

Up until now, we have been investigating the effects of varying θ and N, for fixed $p_{\mathrm{del}} = p_{\mathrm{dupl}} = p = 1$, on the degree distribution. We now report our findings for a third possible scenario: the effect of varying p, for fixed θ and N (keeping $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$, as before).

Figure 6 illustrates the topological effect of varying p in a fixed-size network with heterodimerization $\theta = 0.1$. There is clearly a marked difference between the curve for $p = 1$ and the family of curves for $p < 1$. We established in subsections 4.2–4.4 that for $p = 1$ and $\theta \in (0, 0.5)$, the degree distribution is well approximated by a power-law decay. We now see that for $p < 1$, the power-law behaviour is no longer observed and a characteristic degree is present. We have confirmed this result for a range of network sizes, specifically, $N = 50, 100, 400, 1000$. Moreover, we have found that the second moment $\langle k^2 \rangle$ does not diverge with increasing network size as one would expect if, in the limit of $N \to \infty$, the degree distribution were indeed described by a power-law decay.

We can offer a simple heuristic argument to account for the difference between the case $p = 1$ and $p < 1$. We believe that it is directly related to the choices we have made for the remaining parameters of the model, namely $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$ and $\theta > 0$. For $p < 1$ at each duplication event only some of the links of the mother node are copied by the daughter node, whereas for $p = 1$, all of the mother node's links are copied. Given that $p_{\mathrm{del}} = 1$ we delete a node and all of its links at each time step, and thus at each duplication event, if $p < 1$, we do not compensate for the loss of links incurred through the deletion process. Hence, the repeated application of duplication events with $p < 1$, given that $p_{\mathrm{del}} = 1$ accounts for the fact that the observed degree distribution does not follow a power-law decay but rather has a characteristic degree present.

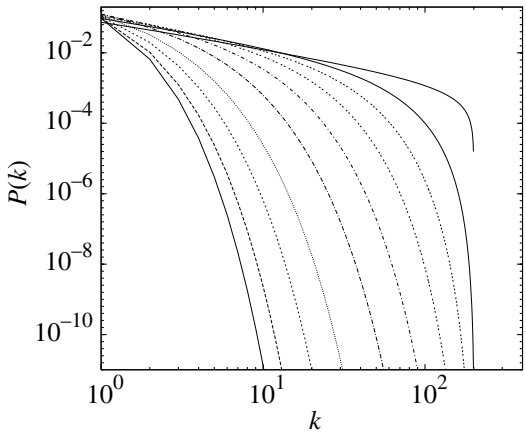

Figure 6. Exact numerical results for the degree distribution in the stationary regime for a fixed-size network, $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$, with $N = 200$ nodes, for increasing duplication probabilities $p = 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.975, 0.99, 1$ (marked with lines of decreasing dash-length) and heterodimerization $\theta = 0.1$. There is a marked difference between the curves generated with $p < 1$ and with $p = 1$. Whereas the latter is described by a power-law decay, the former are not and a characteristic degree is present.

4.6. Comment on relation to biological data

We can test the suitability of the DDDH model as a representation of the evolution of PINs by comparing our results against empirical data. Using data cited in [8], we can obtain values of the number of proteins in the network and estimates of the average degree of the PINs for yeast, fly, and human. From these, we can derive estimates for θ using equation (13) and the apparent scaling exponent, $\gamma = 1 - 2\theta$. The results are tabulated in table 1.

For all data sets, the calculated value of θ is of the order of 10^{-3} which agrees well with the fact that it is believed heterodimerization occurs at a rate not greater than 0.1 [8]. Moreover, the corresponding apparent scaling exponent extracted in all cases is found to be $\gamma = 1.0$. This is consistent with our FSS analysis and corroborates [8] where a power-law with degree exponent $\gamma \approx 1.1$ is given for the data for yeast derived from [38].

5. Discussion

The results in section 4 indicate that the degree distribution of fixed-size networks where the rate of node deletion is equal to the rate of node duplication, $p_{\mathrm{del}} = p_{\mathrm{dupl}} = 1$, is dependent on several features. We summarize these as follows. (i) In order to observe a power-law degree distribution it is necessary to have $p = 1$, that is, perfect duplication, and $0 < \theta < 0.5$. (ii) No power-law is observed in the degree distribution for $p < 1$, or for $p = 1$ and $\theta \geqslant 0.5$. (iii) In all cases, it is necessary for θ to be non-zero in order for a positive average degree to be obtained in the stationary regime since for $\theta = 0$, $\langle k \rangle = 0$. (iv) In cases where a positive average degree is obtained, it is notable that the network self-organizes into a stationary state. We see this as an advantageous feature of this model and comment that this is in contrast to some other network

models, such as in [19], where the average degree is a fixed parameter in the model, or the duplication models where the average degree scales with the network size. (v) Our FSS analysis indicates that for fixed $\theta \in (0, 0.5)$, the scaling exponent of the associated (proper) FSS ansatz is fixed and equal to one. The new result is obtained through the inclusion of a system-size dependency in the prefactor, $a(N)$, necessary when the apparent exponent $\gamma < 1$, which results in the scaling function being recast in such a way that the only relevant scale parameterizing the system is determined by the cutoff, given by N in our case. This example illustrates that it is necessary to be cautious when doing a FSS ansatz in systems where the apparent exponent appears to be less than 1 [39]. (vi) Estimates of θ and γ for networks for yeast, fly and human are consistent with estimates from empirical data and our FSS analysis.

If we accept the fact that the fixed-size version of the DDDH model in spite of its simplicity is able to reproduce certain observed topological features of PINs, this in turn would require us to revise the idea that the protein repertoire has evolved over millions of years from a small set of genes to the genomes we observed today in multi-cellular organisms which are typically composed of tens of thousands of genes since the two are not compatible. Clearly, this is a rather drastic measure. Rather than accept such a state of affairs, perhaps all that the results of the fixed-size DDDH model indicate thus far is that we should exercise caution when interpreting minimalistic network models, as attractive as they are. Since we can conjure up many varied and simple network models, with and without growth, which are capable of reproducing observed features of complex systems perhaps the only recourse when trying to pick one network model over the other is to carefully use our knowledge of the essence of the original real system [40].

5.1. Extensions

As a first step in investigating the behaviour of the DDDH model, it seems reasonable to keep things as simple as possible, as we have done here. We have reported on the case of fixed-size networks, and are currently investigating how the topology is affected by varying the relative rates of node deletion and duplication. Moreover, sensitivity to initial conditions is being probed further; we believe that for the fixed-size case, the network features are independent of initial conditions.

However, beyond the steps we have mentioned, there are obvious extension of this model which further investigations could benefit from including. For example, the model is based on an undirected network—it would be interesting to see how best to incorporate dynamics based on directed links and what affect this would have on the in-degree distribution and out-degree distribution. Incorporating this feature would make the model well-suited to describing genetic regulatory networks, for example.

Moreover, the model assumes that the rate of node deletion and duplication are independent of one other, and independent of any feature of the network such as the size; one could imagine the scenario where this is not the case. Moreover, we consider single-node deletion and single-node duplication—an interesting variation would be to consider multiple-node deletion or duplication, or even duplication of whole modules (motifs) as in [41], for example.

Finally, the only cause of an increase or decrease in the number of links is either due to a node deletion or node duplication event: links are not added or deleted through any other mechanism. The scenario where (directed) links are stochastically added or removed between already existing nodes would be an interesting amendment to investigate, particularly with regards to the resulting effect on the degree distribution and its corresponding exponent [27, 29, 42].

A final example is that there is no fitness parameter in the model, nor any rule based on selection—our results are independent of both of these features at the gene/protein level, and at the network level, yet it is widely believed that both features are driving forces in the evolution of most, if not all, biological systems. Including these features in a meaningful way would be a highly relevant step towards understanding some of the thornier questions in modern biology today.

6. Conclusions

We have introduced and discussed a minimalistic model governed by four parameters, based on dynamic node deletion and node duplication with heterodimerization. The model is intended to capture some basic features in the evolution of PINs but we believe that it is also suited to other types of networks in light of the suggested modifications.

Power-law degree distributions were observed for generic parameter values, and a novel FSS effect was observed for the case of fixed-size networks evolving through perfect duplication and $\theta \in (0, 0.5)$. The existence of an attractive fixed point in the average degree was derived based on mean-field arguments, and corroborated with numerical simulations of the first moment of the degree distribution as described by the two-step rate equation. The above results were then used to derive a relation for the apparent exponent, $\gamma = 1 - 2\theta$.

Our results thus far indicate consistency with empirical data. Further investigations are required to fully explore and understand the wider phase-space inhabited by this model, and several suggestions have been made to this end.

Acknowledgments

We thank the referees for very useful comments on a previous version of this paper. The authors are grateful for helpful discussions with Henrik J Jensen and Simon Laird. We are also grateful for Matthew Stapleton for insightful discussions on FSS. NF is thankful to EPSRC DTG for funding.

References

[1] Wagner A 2001 *Mol. Biol. Evol.* **18** 1283
[2] Chung F, Lu L, Dewey T G and Galas D J 2003 *J. Comput. Biol.* **10** 677
[3] Vázquez A, Flammini A, Maritan A and Vespigniani A 2003 *ComplexUs* **1** 38
[4] Kim J, Krapivsky P L, Kahng B and Redner S 2002 *Phys. Rev.* E **66** 055101(R)
[5] Pastor-Satorras R, Smith E and Sole R V 2003 *J. Theor. Biol.* **222** 199
[6] Sole R V, Pastor-Satorras R, Smith E D and Kepler T 2002 *Adv. Complex Sys.* **5** 43
[7] Raval A 2003 *Phys. Rev.* E **68** 0066119
[8] Ispolatov I, Krapivsky P L and Yuryev A 2005 *New J. Phys.* **7** 145
Ispolatov I, Krapivsky P L and Yuryev A 2005 *Phys. Rev.* E **71** 061911
[9] Hughes A L 1994 *Proc. Biol. Sci.* **256** 119
[10] Kent W J, Baertsch R, Hinrichs A, Miller W and Haussler D 2003 *Proc. Natl Acad. Sci. USA* **100** 11484
[11] Lynch M and Conery J S 2000 *Science* **290** 1151
[12] Zhang J 2003 *Trends Ecol. Evol.* **18** 292
[13] Ohno S 1970 *Evolution by Gene Duplication* (New York: Springer)

[14] Friedman R and Hughes A L 2001 *Genome Res.* **11** 373
[15] Gu Z, Cacalcanti A, Chen F-C, Bouman P and Li W-H 2002 *Mol. Biol. Evol.* **19** 256
[16] Wolfe K H and Li W-H 2003 *Nature Genet.* **33** 255
[17] Kellis M, Birren B W and Lander E S 2004 *Nature* **428** 617
[18] Prince V E and Pickett F B 2003 *Nature Rev. Genet.* **3** 827
[19] Berg J, Lässig M and Wagner A 2004 *BMC Evol. Biol.* **4** 51
[20] Axelsen J B, Yan K-K and Maslov S 2005 *Preprint* q-bio.GN/050702
[21] van Noort V, Snel B and Huynen M 2004 *EMBO Rep.* **5** 280
[22] Minnhagen P, Rosvall M, Sneppen K and Trusina A 2004 *Physica* A **340** 725
[23] Dorogvtsev S N, Mendes J F F and Samukhin A N 2002 *Europhys. Lett.* **57** 334
[24] Bollobás B and Riordan O 2002 Mathematical results on scale-free graphs *Handbook of Graphs and Networks* ed S Bornholdt and H Schuster (Berlin: Wiley-VCH)
[25] Jeong H, Mason S P, Barabási A-L and Oltvai Z W 2001 *Nature* **411** 41
[26] Gu Z, Steinmetz L M, Gu X, Scharfe C, Davis R W and Li W-H 2003 *Nature* **421** 63
[27] Chung F and Lu L 2004 *Internet Math.* **1** 409
[28] Cooper C, Frieze A and Vera J 2004 *Internet Math.* **1** 464
[29] Wang C 2005 *Proceedings of Combinatorial and Algorithm Aspects of Networking* to be published in a special issue of *Internet. Math.*
[30] Albert R and Barabási A 2002 *Rev. Mod. Phys.* **74** 47
[31] Newman M E J 2003 *SIAM Rev.* **45** 167
[32] Dorogovtsev S N and Mendes J F F 2002 *Adv. Phys.* **51** 1079
[33] Yook S-H, Oltvai Z N and Barabási A-L 2004 *Proteomics* **4** 928
[34] Sole R V and Fernandez P 2003 *Preprint* q-bio.GN/0312032v1
[35] Chung F, Handjani S and Jungreis D 2003 *Ann. Comb.* **7** 141
[36] Slanina F and Kotrla M 1999 *Phys. Rev. Lett.* **83** 5587
Slanina F and Kotrla M 2000 *Phys. Rev.* E **62** 6170
Slanina F, Kotrla M and Steiner J 2002 *Europhys. Lett.* **60** 14
[37] Krapivsky P L and Redner S 2001 *Phys. Rev.* E **63** 066123-1
[38] von Mering C, Krause R, Snel B, Cornell M, Oliver S G, Fields S and Boork P 2002 *Nature* **417** 399
[39] Christensen K, Farid N, Pruessner G and Stapleton M private communication
[40] Watts D J 2003 *Six Degrees: The Science of a Connected Age* (New York: Norton)
[41] Ravasz E and Barabási A-L 2003 *Phys. Rev.* E **67** 026112
[42] Tadić B 2001 *Physica* A **293** 273