AIP | Chaos
An Interdisciplinary Journal of Nonlinear Science
25TH ANNIVERSARY

Mean field theory for biology inspired duplication-divergence network model

Shuiming Cai, Zengrong Liu, and H. C. Lee

Citation: Chaos **25**, 083106 (2015); doi: 10.1063/1.4928212
View online: http://dx.doi.org/10.1063/1.4928212
View Table of Contents: http://scitation.aip.org/content/aip/journal/chaos/25/8?ver=pdfcov
Published by the AIP Publishing

Mean field theory for biology inspired duplication-divergence network model

Shuiming Cai,[1,a)] Zengrong Liu,[2,b)] and H. C. Lee[3,4,5,c)]
[1]*Faculty of Science, Jiangsu University, Zhenjiang 212013, China*
[2]*Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[3]*Institute of Systems Biology and Bioinformatics, National Central University, Zhongli, 32001 Taiwan*
[4]*Center for Dynamical Biomarkers and Translational Medicine, National Central University, Zhongli, 32001 Taiwan*
[5]*Department of Physics, Chung Yuan Christian University, Zhongli, 32023 Taiwan*

(Received 21 April 2015; accepted 24 July 2015; published online 11 August 2015)

The duplication-divergence network model is generally thought to incorporate key ingredients underlying the growth and evolution of protein-protein interaction networks. Properties of the model have been elucidated through numerous simulation studies. However, a comprehensive theoretical study of the model is lacking. Here, we derived analytic expressions for quantities describing key characteristics of the network—the average degree, the degree distribution, the clustering coefficient, and the neighbor connectivity—in the mean-field, large-N limit of an extended version of the model, duplication-divergence complemented with heterodimerization and addition. We carried out extensive simulations and verified excellent agreement between simulation and theory except for one partial case. All four quantities obeyed power-laws even at moderate network size ($N \sim 10^4$), except the degree distribution, which had an additional exponential factor observed to obey power-law. It is shown that our network model can lead to the emergence of scale-free property and hierarchical modularity simultaneously, reproducing the important topological properties of real protein-protein interaction networks. © *2015 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4928212]

Biological processes that make living cells function are wired by interaction networks of various cellular components such as proteins, DNA, RNA, metabolites, and small molecules. The structures of many such networks, including networks of protein-protein interactions and transcription-regulatory networks, have been revealed. Studies on the topological structure of these complex biological networks have indicated that they share a number of characteristic features such as sparseness, small-world pattern, scale-free connectivity, hierarchical modularity, and disassortativity. Recently, various network growth models invoking duplication and divergence (DD) have been constructed to recapture the topological properties of real protein-protein interaction networks. However, those network growth models invoking duplication and divergence have so far only been elucidated through numerous simulations; a comprehensive theoretical analysis is still lacking. The mechanism underlying the evolution of protein-protein interaction networks is therefore still not well understood. In this paper, we comprehensively explored the duplication-deletion-heterodimerization-addition (DDHA) model, derived from it analytical solutions for the average degree, the degree distribution, the clustering coefficient, and the neighbor connectivity in the mean-field and large-N approximation, and conducted extensive simulations for validation. Our results indicated that all four quantities obeyed power-laws even at moderate network size ($N \sim 10^4$), except the degree distribution, which had an additional exponential factor; the exponent of this factor also obeyed a power-law. Particularly, it is shown that our network model can lead to the emergence of scale-free property and hierarchical modularity simultaneously, reproducing the important topological properties of real protein-protein interaction networks.

I. INTRODUCTION

Biological processes that make living cells function are wired by interaction networks of various cellular components such as proteins, DNA, RNA, metabolites, and small molecules.[1] The structures of many such networks, including protein-protein interaction, metabolic, signaling, and transcription-regulatory networks, were revealed through the development of high-throughput data-collection methods and new technology platforms.[2] These networks are not independent; rather they form a "network of networks" that drive cell function. A major challenge of contemporary biology is to understand and model quantitatively the topological and dynamic properties of these complex biological networks by integrating theory with experimental data.[2]

Protein-protein interactions are central to biological processes, and the systematic identification of all protein-protein interactions is key to gain insight into the inner workings of a cell.[3] New developments in experimental and computational techniques have led to the systematic determination of putative and actual protein interactions in many model organisms. The information of protein-protein interaction networks at the whole-genome level is now available from several organisms, including *Saccharomyces cerevisiae*,[4–7] *Caenorhabditis*

[a)]Electronic mail: caishuiming2008@126.com
[b)]Electronic mail: zrongliu@126.com
[c)]Electronic mail: hclee12345@gmail.com

elegans,[8] *Drosophila melanogaster*,[9] *Homo sapiens*,[10,11] and *Plasmodium falciparum*.[12] Studies on the topological structure of these protein-protein interaction networks and other large-scale biological networks have revealed that they share a number of interesting characteristics: (1) They are sparse graphs, with a small average number of links.[2,3,13] (2) They are scale-free networks;[2,13,14] there is no typical number of links per node, rather the distribution of the number of links (k) per node (P) decays as a power law: $P(k) \sim k^{-\gamma_P}$. That is, there are many nodes with few links and a small but still significant number of nodes (hubs) with many links. (3) They have a small-world architecture;[2,8,9,13] they are highly clustered but the average shortest path length almost as low as that for random networks. (4) They exhibit hierarchical modularity structure,[2,3,15,16] with $C(k)$, the average cluster coefficient of k-degree nodes, obeying a power-law $C(k) \sim k^{-\gamma_C}$,[3,16] indicating that low-degree nodes tend to be more clustered than high-degree ones. (5) They show a disassortative structure, in which $K_{nn}(k)$, the average degree among the neighbors of all k-degree nodes, follows $K_{nn}(k) \sim k^{-\gamma_{nn}}$.[18–20] That is, connections between a hub and a low-degree node are favored, while those among hubs and among low-degree nodes are suppressed.[18–20]

Small world phenomena and power-law degree distributions have previously been observed in a number of naturally occurring graphs such as communication networks, web graphs, research citation networks, neural nets, among others.[21,22] It is possible to generate networks that satisfy these two properties by an iterative process that adds one new node to the graph at each step, with the new node preferentially attached to some of the existing high-degree nodes.[21,22] However, such a model of preferential attachment does not capture the essence of the genome evolution and is therefore not suitable for modeling biological networks.

Duplication and divergence have been widely recognized as the two dominant mechanisms driving the evolution of genome[23–28] and cellular network.[29] Duplication is the driving force for creating new genes in genomes: at least 50% of prokaryotic genes[30,31] and over 90% of eukaryotic genes[32] are products of gene duplication, while divergence generates function diversity.[13,23] Recent work has shown that interaction networks constructed on the principle of DD tend to exhibit scale-free and small-world properties.[17–19,33–39]

Studies have unveiled that biological networks from protein-protein interactions to metabolic and regulatory networks characteristically exhibit hierarchical modularity.[2,15,16] Despite networks constructed in DD-based models successfully predict the scale-free and small-world properties,[17–19,33–39] they failed to exhibit hierarchical modularity.[40,41] To overcome this difficulty, the duplication-divergence-heterodimerization (DDH) model, that is, duplication and divergence complemented with the heterodimerization process, has been proposed.[18,19,41,42] Simulation studies have shown that the DDH model could generate networks that exhibit hierarchical modularity and scale-free connectivity.[18,19,42] Heterodimerization, or the enhanced linkage of pairs of target and replica nodes, is essential for generating clustering in protein-protein interaction networks.[41,42]

Those network growth models invoking duplication and divergence do capture the topological properties of real protein-protein interaction networks that have so far only been elucidated through numerous simulations;[17–19,33–39,42] a comprehensive theoretical analysis is still lacking. The mechanism underlying the evolution of protein-protein interaction networks is therefore still not well understood. In this paper, we comprehensively explored the DDHA model, derived from it are analytical solutions for the average degree, the degree distribution, the clustering coefficient, and the neighbor connectivity in the mean-field and large-N approximation, and conducted extensive simulations for validation. The new ingredient in our model, addition, by linking a newly created duplicate to nodes not connected to its ancestor, reflected the process of mutation through which the duplicated one could develop a new, independent and original interaction pattern and function.[33–35] Our results indicated that all four quantities obeyed power-laws even at moderate network size ($N \sim 10^4$), except the degree distribution, which had an additional exponential factor; the exponent of this factor also obeyed a power-law. It is shown that our network model can lead to the emergence of scale-free property and modularity simultaneously, reproducing the important topological properties of real protein-protein interaction networks. We hope that the results derived in this study will provide some insight into the mechanisms underlying various topological properties of biological networks.

II. THE NETWORK MODEL

The model was a topologically based approximation intended to capture generic features of proteome evolution.[33,34] It translated the evolution of the protein-protein interaction networks into a growing network and did not include functionality or dynamics of the proteins involved. Protein-protein interaction networks in cells do not directly evolve as described in the model. Rather, they so evolve as a consequence of evolution of the genome, driven mainly by gene duplication and subfunctionalization (i.e., diversification).[26,27]

In the network model, each node was considered as the protein expressed by a gene, and the duplication of a protein was meant to represent the consequence of a gene duplication in the genome. We restricted the duplication to single-protein in the model because multiple-gene duplication or larger duplications in the genome are not universal or in any case are relatively rare events.[23,33,34] After protein duplication, the ancestor and its duplicate will have the same interactions.[23] In the course of subsequent evolution, in a majority of cases, one of the ancestro-duplicate pair will be lost through redundancy. In other cases, both proteins survive by divergence, in which one or both of the pair lose some old functions or acquire new ones.[23,33,34] In the model, these phenomena are emulated by letting the duplicate start by having links to all neighbors of the ancestor, followed by random removal of these links from the duplicate. When a self-interacting gene is duplicated, the ancestor-duplicate pair will interact with each other, and the link between the pair may survive after divergence.[13,38,41] In the model, this was mimicked by establishing a new link between the pair with some probability, forming a heterodimer.[38,41] To account for mutations, the model allowed limited

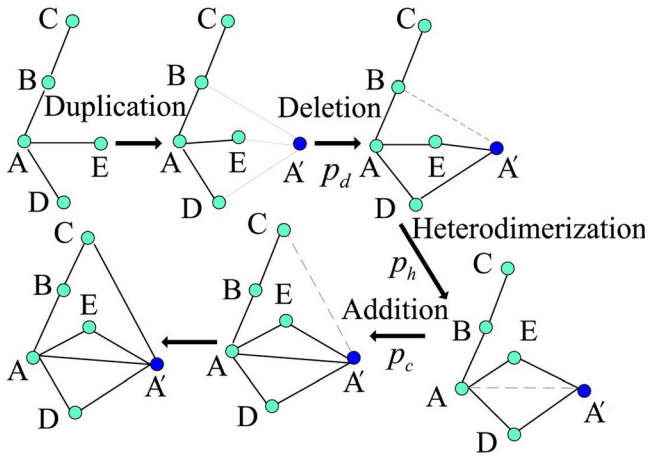

FIG. 1. Schematic representation of the four operations in the DDHA model: duplication, deletion (removal), heterodimerization, addition.

random attachments of new links between a newly created duplicate to nodes not connected to its ancestor.[33–35] These processes were formalized in our model as follows.

A connected network of relative small size (N_0) is given an initial configuration with random connections. At each subsequent time-step, it is made to evolve and grow through the following four actions until the network reaches a desired size (Fig. 1):

(I) Duplication: Randomly select a node A from the network in its current state, duplicate A by adding a new node A' to the network such that A' is connected to all nodes connected to A.
(II) Deletion: Delete with probability p_d each of the edges connected to A'.
(III) Heterodimerization: Connect with probability p_h A' to A.
(IV) Addition: Connect with probability p_c A' to all nodes except A and those connect to A.

Duplication is a key mechanism in all biological growth.[23–29] Divergence, actions (ii) and (iv), give rise to complexity in biological systems.[18,19,33–39,41–45] Heterodimerization is essential for clustering.[18,19,41,42] After this four-step procedure, if the new node A' is not connected to any other node, it is removed. This guarantees the connectedness of the network, as protein-protein interaction networks must be.[18,19,39,41,42]

III. MEAN FIELD THEORY

Mean field theory (or mean-field approximation), which is originated in statistical physics, has been frequently used in the investigation of complex networks for deriving analytical expressions of the quantities describing the characteristics of network evolution models, such as degree distribution, average path length, and clustering coefficient.[46–49] The main idea of mean field theory is to replace all interactions on any one body with an average or effective interaction, sometimes called a molecular field.[50] In this section, we give a mean-field analysis for the DDHA model, which allows us to derive analytic forms of power-law exponents in the DDHA model for the average degree, degree distribution, clustering coefficient, and neighbor connectivity as functions of parameters of the model.

We denote N as the number of nodes in network, or network size; t as the growth "time," or the number of event steps (for our purpose, we set $\Delta t \approx \Delta N = N(t+1) - N(t) = 1$, hence $N(t) \approx N(0) + t$). We set n_c, but not p_c, to be a constant small compared to N, and let $p_c = n_c/(N - k_A - 1)$, where k_A is the degree of A.[33–35,43] In order to provide a general analytical understanding of the DDHA model, we analyzed the statistical property of the network generated in the DDHA model by considering in the mean field theory the time evolution of the average degree K_N, the degree distribution $P(k)$, clustering coefficient $C(k)$, and neighbor connectivity $K_{nn}(k)$, where the variable k denotes degree. The derived mean-field approximations of power laws of these quantities are summarized in Table I.

A. Average degree

Let K_N be the average degree of the network when it has N nodes. After a duplication event $N \to N+1$, the degree of the network becomes $K_N N + (2K_N - 2p_d K_N) + (2n_c + 2p_h)$, where the first term indicates the degree of the network with N nodes, the second term corresponds to the duplication of one node and the average elimination of $p_d K_N$ links emanating from the new node, the last term accounts for the addition of $p_c(N - K_N - 1) \approx n_c$ new links pointing to the new node, and the addition of p_h new links via heterodimerization. Hence, after the duplication event $N \to N+1$, the change of the average degree is

$$K_{N+1} - K_N = \frac{K_N N + (2K_N - 2p_d K_N) + (2n_c + 2p_h)}{N+1} - K_N = \frac{K_N - 2p_d K_N + 2n_c + 2p_h}{N+1}.$$

For large N, using the continuous approximation, then the evolution equation of K_N is

$$\frac{dK_N}{dt} \approx \frac{dK_N}{dN} \approx K_{N+1} - K_N = \frac{1}{N}(K_N - 2p_d K_N + 2n_c + 2p_h). \quad (1)$$

The large-N solution for Eq. (1) is

$$K_N = \begin{cases} \xi + (K_0 - \xi)N^{1-2p_d}, & p_d \neq 1/2, \\ 2(n_c + p_h)\ln(N/N_0) + K_0, & p_d = 1/2, \end{cases} \quad (2)$$

where $\xi = 2(n_c + p_h)/(2p_d - 1)$, and K_0 and N_0 are integration constants. Therefore, for large N, K_N grows with power law with exponent $\beta_K = 1 - 2p_d$ when $p_d < 1/2$, is logarithmic when $p_d = 1/2$, and is a constant, $K_\infty = \lim_{N\to\infty} K_N = \xi$, independent of N when $p_d > 1/2$ (Fig. 2(a)). This indicates that $p_d = 1/2$ is a critical value for DDHA networks. A realistic finite average degree is recovered only when the deletion probability $p_d > 1/2$. In other words, a DDHA network may have the biological properties of sparseness and small-worldness only when more than half of the links on a newly duplicated node are removed immediately after duplication. Note that the power-law property of K_N does not depend on p_h or p_c.

TABLE I. Mean-field estimates of power-law exponents of characteristic network quantities in DDHA model. The variable k is the degree.

Power law	Exponent
Average degree $K_N \sim N^{\beta_K}$	$\beta_K = \begin{cases} 1-2p_d, & p_d < \frac{1}{2}; \\ 0\,(\text{logarithmic}), & p_d = \frac{1}{2}; \\ 0, & p_d > \frac{1}{2}. \end{cases}$
Degree distribution[a] $P(k) \sim k^{-\gamma_P}$	$\gamma_P = 1,\ p_d \le 0.4329$ $1 < \gamma_P \le 2,\ 0.4329 < p_d \le 0.5$ $2 < \gamma_P \le 3,\ 0.5 < p_d \le 0.5858$
Clustering coefficient $C(k) \sim k^{-\gamma_C}$	$\gamma_C = \begin{cases} 1, & \frac{1}{2} \le p_d < 1; \\ 2p_d, & 0 < p_d \le \frac{1}{2}. \end{cases} \quad (p_h > 0)$
Degree correlation $K_{nn}(k) \sim k^{-\gamma_{nn}}$	$\gamma_{nn} = \begin{cases} 0, & p_d \ge \frac{1}{2}; \\ 2p_d - 1, & p_d < \frac{1}{2}. \end{cases}$

[a]Also has an exponential factor for finite N; see text and Fig. 5.

B. Degree distribution

The degree distribution, $P(k)$, is an important statistical property of the characterization of a network, which is defined as the probability that a randomly selected node has exactly k links.[2] Denote k as the degree of a node and $f(k, t)$ be the number of nodes with degree k at time t, then the degree distribution at time t is $P(k) = f(k,t)/N(t)$. To describe the degree distribution of the network model, we need to establish a basic relationship between the number of nodes with degree k at successive time-steps. Recall that duplication starts adding one node per time-step ($t \to t+1$) at $t_0 = N_0$ (Fig. 1). Therefore, in the mean-field approximation, the expected value of the number of nodes with degree k at time $t+1$, $f(k, t+1)$, satisfies the following iterative equation for $t \ge t_0$:

$$
\begin{aligned}
f(k,t+1) = {} & \left[f(k,t) - p_h \frac{f(k,t)}{N(t)} - (1-p_d)\frac{kf(k,t)}{N(t)} \right. \\
& \left. - \left(1 - \frac{k_+ f(k,t)}{N(t)}\right)\frac{n_c f(k,t)}{N(t)} \right] \\
& + \left[p_h \frac{f(k_-,t)}{N(t)} + (1-p_d)\frac{k_- f(k_-,t)}{N(t)} \right. \\
& \left. + \left(1 - \frac{kf(k_-,t)}{N(t)}\right)\frac{n_c f(k_-,t)}{N(t)} \right] \\
& + (1-p_h) \sum_{m\ge 0} \sum_{n \ge k-m} \frac{f(n,t)}{N(t)} \\
& \times \mathcal{B}(n, 1-p_d; k-m)\mathcal{B}(N(t), p_c; m) \\
& + p_h \sum_{m\ge 0} \sum_{n \ge k_- - m} \frac{f(n,t)}{N(t)} \mathcal{B}(n, 1-p_d; k_- - m) \\
& \times \mathcal{B}(N(t), p_c; m), \qquad (3)
\end{aligned}
$$

where $k_\pm = k \pm 1$ and $\mathcal{B}(N(t), s; m)$ is given in the expansion $(s-(1-s))^{N(t)} \equiv \sum_{m=0}^{N(t)} \mathcal{B}(N(t), s; m) = \sum_m \cdots s^m \cdots$. In Eq. (3), the first term on the right-hand-side is the expected number of k-degree nodes at time t remaining so at $t+1$ through actions (i)–(iv); the second term is the expected number of $(k-1)$-degree nodes at t becoming k-degree at

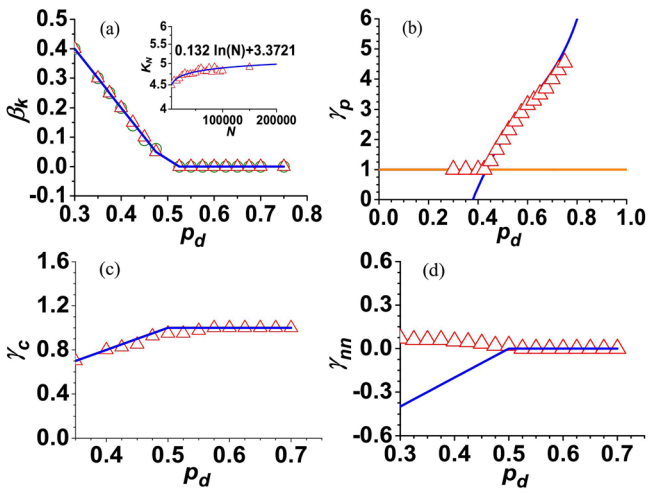

FIG. 2. (a) β_K for the average degree K_N; K_N is predicted to be logarithmic at $p_d = 1/2$ (inset). (b) γ_P for $P(k)$ after correction by an exponential factor (see Fig. 5); a non-trivial solution exists for $p_d > 0.4329$. (c) γ_C for $C(k)$. (d) γ_{nn} for $K_{nn}(k)$. Triangles (circles) are obtained in simulations with $p_h = 0.065$ ($p_h = 0$) and $n_c = 0.001$ for $N = 2 \times 10^6$ networks. Lines are theoretical large-N mean-field results (see Table I).

$t+1$ through actions (i)–(iv); the last two terms are the expected number of n-degree nodes at t becoming k-degree at $t+1$ through action (iv).

Recall that the degree distribution at t is $P(k) = f(k,t)/N(t)$, and $N(t+1) - N(t) = 1$. In the large-t limit, Eq. (3) can be reduced to the following form:

$$\begin{aligned}0 =& [-1-p_h-(1-p_d)k-n_c]P(k)\\ &+[p_h+(1-p_d)(k-1)+n_c]P(k-1)\\ &+(1-p_h)\sum_{m\geq 0}\sum_{n\geq k-m}P(n)\mathcal{B}(n,1-p_d;k-m)\mathcal{B}(N(t),p_c;m)\\ &+p_h\sum_{m\geq 0}\sum_{n\geq k_- -m}P(n)\mathcal{B}(n,1-p_d;k_- -m)\mathcal{B}(N(t),p_c;m).\end{aligned} \tag{4}$$

By assuming $P(k)$ obeys a power law[37,43] and writing $P(k) \propto k^{-\gamma_P}$, we have

$$\begin{aligned}0 =& [-1-p_h-(1-p_d)k-n_c]\\ &+[p_h+(1-p_d)(k-1)+n_c]\frac{k^{\gamma_P}}{(k-1)^{\gamma_P}}\\ &+(1-p_h)\sum_{m\geq 0}\sum_{n\geq k-m}\frac{k^{\gamma_P}}{n^{\gamma_P}}\mathcal{B}(n,1-p_d;k-m)\mathcal{B}(N(t),p_c;m)\\ &+p_h\sum_{m\geq 0}\sum_{n\geq k_- -m}P(n)\mathcal{B}(n,1-p_d;k_- -m)\mathcal{B}(N(t),p_c;m).\end{aligned} \tag{5}$$

Note that $\left(\frac{k}{k-1}\right)^{\gamma_p} = 1+\frac{\gamma_p}{k}+\mathcal{O}(k^{-2})$,[37,43] so that the first two terms on the right-hand side of Eq. (5) are $-1-(\gamma_p-1)(1-p_d)+\mathcal{O}(k^{-1})$. On the other hand, for any constant $\gamma > 0$, and any j, l one has[37,43]

$$\binom{j}{j-l}\left(\frac{l}{j}\right)^{\gamma} = \left(1+\mathcal{O}\left((l+1)^{-1}\right)\right)\binom{j-\gamma}{j-l},$$

thus,

$$\begin{aligned}&\sum_{n\geq k-m}\frac{k^{\gamma_P}}{n^{\gamma_P}}\mathcal{B}(n,1-p_d;k-m)\\ &=\sum_{n\geq k-m}\binom{n}{k-m}(1-p_d)^{k-m}(p_d)^{n-(k-m)}\left(\frac{k}{n}\right)^{\gamma_p}\\ &=\sum_{n\geq k-m}\binom{n}{n-(k-m)}\left(\frac{k-m}{n}\right)^{\gamma_p}(1-p_d)^{k-m}(p_d)^{n-(k-m)}\left(\frac{k}{k-m}\right)^{\gamma_p}\\ &=\left(1+\mathcal{O}\left(\frac{1}{(k-m)+1}\right)\right)\left(\frac{k}{k-m}\right)^{\gamma_p}\sum_{n\geq k-m}\binom{n-\gamma_p}{n-(k-m)}(1-p_d)^{k-m}(p_d)^{n-(k-m)}\\ &=\left(1+\mathcal{O}\left(\frac{1}{(k-m)+1}\right)\right)\left(\frac{k}{k-m}\right)^{\gamma_p}(1-p_d)^{k-m}\sum_{j\geq 0}\binom{j+(k-m)-\gamma_p}{j}(p_d)^j\\ &=\left(1+\mathcal{O}\left(\frac{1}{(k-m)+1}\right)\right)\left(\frac{k}{k-m}\right)^{\gamma_p}(1-p_d)^{k-m}\sum_{j\geq 0}\binom{\gamma_p-(k-m)-1}{j}(-1)^j(p_d)^j\\ &=\left(1+\mathcal{O}\left(\frac{1}{(k-m)+1}\right)\right)\left(\frac{k}{k-m}\right)^{\gamma_p}(1-p_d)^{k-m}(1-p_d)^{\gamma_p-(k-m)-1}\\ &=\left(1+\mathcal{O}\left(\frac{1}{(k-m)+1}\right)+\mathcal{O}\left(\frac{m}{k}\right)\right)(1-p_d)^{\gamma_p-1} = \left(1+\mathcal{O}\left(\frac{1}{k}\right)\right)(1-p_d)^{\gamma_p-1}.\end{aligned} \tag{6}$$

Since $\sum_{m\geq 0}\mathcal{B}(N(t),s;m) = 1$, the third term on the right-hand side of Eq. (5) is $(1-p_h)(1+\mathcal{O}(k^{-1}))(1-p_d)^{\gamma_p-1}$. Similarly, we can derive that the last term on the right-hand side of Eq. (5) is $p_h(1+\mathcal{O}(k^{-1}))(1-p_d)^{\gamma_p-1}$. Therefore, we obtain an equation for γ_P for large N

$$(1-\gamma_P)(1-p_d) = (1-p_d)^{\gamma_P-1} - 1. \tag{7}$$

Once again γ_P depends only on p_d, not on p_h or p_c. Similar results have been reported.[35,37,43] A numerical solution of Eq. (7) for γ_P as a function of p_d (Fig. 2(b)) shows that there is a p_d-independent trivial solution giving $\gamma_P = 1$ and a non-trivial solution giving $\gamma_P > 1$ for $p_d > 0.4329$. In particular, $1 < \gamma_P \leq 2$ when $0.4329 < p_d \leq 0.5$, and $2 \leq \gamma_P \leq 3$ when $0.5 \leq p_d \leq 0.5858$. The empirical values of γ_P extracted from biological networks mostly lie in the range 2 to 3, with a few between 1 and 2.[2,3,7–11,18,19]

C. Clustering coefficient

Hierarchical modularity is a feature shared by a large number of real biological networks.[2,3,15,16] The node-specific clustering coefficient, the cohesiveness of the neighborhood of a node that has k_i links, has been used to examine hierarchical modularity in scale-free networks.[2,3,15,16] The clustering coefficient of k_i-degree node is defined as $C(k_i) = 2g(k_i)/[k_i(k_i-1)]$, where $g(k_i)$ is the number of links between its neighbors.[51] This quantity measures how close the local

neighborhood of a node is to being part of a clique (module) in which every node is connected to all other nodes. In practice, the average, $C(k)$, of clustering coefficients of nodes having the same degree k is used to characterize the network hierarchical modularity. For many real biological networks, it has been observed that $C(k) \propto k^{-1}$, which is an indication of a network's hierarchical character.[2,15,16]

For convenience, we denote the set of all nodes linked to node A, or its (nearest) neighbors, by $\mathcal{S}_A$. In the following, we derive the analytic expression of clustering coefficient $C(k)$ in the DDHA model. First, we consider the change in A's degree, k_A, after a time step ($t \to t+1$). From Fig. 1, it is easy to see that there are three potential sources of change: (i) A is duplicated and heterodimerization does occur; (ii) a neighbors of A is duplicated, and the link between the new node and A is not deleted; (iii) a node that is neither A nor one of its neighbors is duplicated, and a new link between the new node and A is added. Therefore, the change of the degree k_A after a time step is $k_A(t+1) - k_A(t) = \frac{p_h}{N(t)} + \frac{k_A}{N(t)}(1-p_d) + (1 - \frac{k_A+1}{N(t)})p_c$. For large t, using $p_c \approx \frac{n_c}{N(t)}$ and the continuous approximation, we have

$$\frac{\mathrm{d}k_A}{\mathrm{d}t} = \frac{p_h}{N(t)} + \frac{k_A}{N(t)}(1-p_d) + \left(1 - \frac{k_A+1}{N(t)}\right)p_c \approx \frac{1-p_d}{N(t)}(k_A + \eta), \tag{8}$$

where $\eta \equiv (p_h + n_c)/(1-p_d)$.

Now, we consider the change of g_A, the number of links among the nodes in $\mathcal{S}_A$, after a time step ($t \to t+1$). Three events will cause g_A to increase in a cycle (i.e., one time step) of growth triggered by a duplication (Fig. 3): (i) A is duplicated (call it A'). A triangle (A, A',B), for any $B \in \mathcal{S}_A$, will form, provided A and A' dimerize and the new link between A' and B is not deleted (Fig. 3(a)). This adds $p_h(1-p_d)$ to g_A; (ii) a neighbor $B \in \mathcal{S}_A$ is duplicated (call it B'). A triangle (A, B, B') will form, provided B and B' dimerize and the new link between B' and A is not deleted, and a triangle (A, B', j'), $j' \in \mathcal{S}_A$ and not B, will form, provided B and j' are linked and both the links between A and B' and between j' and B' are not deleted (Fig. 3(b)). Since the clustering coefficient of a node is the probability that its two neighbors are linked,[2] the expected number of links between the neighbor B and other neighbors of node A is given by $(k_A - 1)C(k_A)$. This means that event (ii) adds $(1-p_d)(p_h + (1-p_d)(k_A-1)C(k_A))$ to g_A; and (iii) a next nearest neighbor E of A is duplicated (call it E'), namely, $E \in \mathcal{S}_B$, $B \in \mathcal{S}_A$, and $E \notin \mathcal{S}_A$. Then, a triangle (A, B, E') will form provided the E'–B link is not deleted and an E'–A link is added (Fig. 3(c)). This adds $m_{EA}(1-p_d)p_c$ to g_A, where m_{EA} is the number of nodes in the intersect $\mathcal{S}_A \cap \mathcal{S}_E$. Adding (i), (ii), and (iii) over all possible participating nodes, and replacing p_c by $n_c/N(t)$, leads to a rate of change in g_A

$$\frac{\mathrm{d}g_A}{\mathrm{d}t} = \frac{1-p_d}{N(t)}\left(k_A p_h + k_A(p_h + (1-p_d)(k_A-1)C(k_A)) + \sum_{E\neq A, E\notin \mathcal{S}_A} m_{EA}\frac{n_c}{N(t)}\right), \tag{9}$$

where $1/N(t)$ is a weigh factor and the factor $(\sum_{E\neq A, E\notin \mathcal{S}_A} m_{EA})$ is the just the number of unique two-link paths from node A to all its next nearest neighbors. The n_c-dependent term in Eq. (9) is of order $\mathcal{O}(N^{-2})$, which may not contribute to the leading term in g_A for large N. It is therefore ignored (Figs. 2(c) and 4).

Noting that $C(k_A) = 2g_A/(k_A(k_A-1))$ and using Eq. (8), we obtain for large N

$$\frac{\mathrm{d}g_A}{\mathrm{d}k_A} = \frac{\mathrm{d}g_A}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}k_A} \approx \frac{2(1-p_d)}{k_A+\eta}g_A - \frac{2p_h\eta}{k_A+\eta} + 2p_h. \tag{10}$$

The solution of Eq. (10) is

$$g_A = \begin{cases} \dfrac{2p_h}{2p_d-1}k_A + a_0(k_A+\eta)^{2-2p_d} + \dfrac{p_h\eta}{(2p_d-1)(1-p_d)}, & p_d \neq 1/2, \\ 2p_h(k_A+\eta) - 2p_h\eta\ln(k_A+\eta) + a_0, & p_d = 1/2, \end{cases}$$

where a_0 is a constant. For large k

$$C(k) \approx \frac{2g}{k(k-1)} \sim k^{-\gamma_C}, \qquad \gamma_C = \begin{cases} 1, & \frac{1}{2} \le p_d < 1,\ p_h > 0; \\ 2p_d, & 0 < p_d \le \frac{1}{2},\ p_h > 0. \end{cases} \tag{11}$$

This shows that $C(k)$ decays with a power law, implying that low-degree nodes tend to be more clustered than high-degree ones. The power-law exponent depends on the deletion probability p_d and heterodimerization probability p_h but not on the addition probability p_c ($\approx n_c/N$). When $p_h = 0$, the model loses its main mechanism for triangle formation, and a low value of $C(k)$ is expected.[33–35] Our simulations with $p_h = 0$ showed that $C(k)$ was of the order of 10^{-5} to 10^{-4}. It was concluded in Refs. 13 and 41 that links between recently duplicated pairs of protein are common, implying that heterodimerization of pairs of duplicates regularly occurs in real biological networks. Therefore, we focused on the more relevant case with $p_h > 0$. Then, p_d is again the deciding parameter and $p_d = 1/2$ is a critical value. If $p_d > 1/2$, then, with $C \propto k^{-1}$, the network will have hierarchical modularity structure, as seen in many real biological networks.[2,15,16] This shows that the deletion of links and heterodimerization are two key factors for the emergence of hierarchical modularity. Indeed, if the duplicated node is a self-interacting

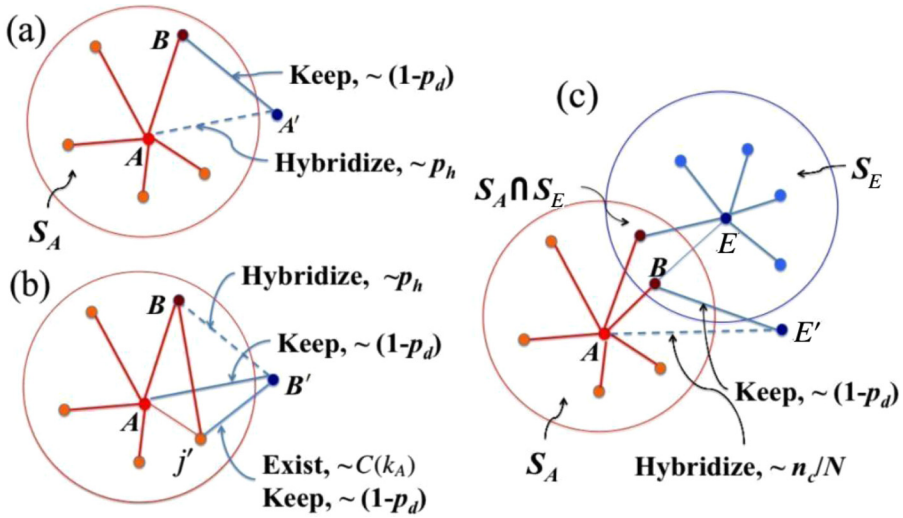

FIG. 3. Three ways, (a), (b), and (c), corresponding to (i), (ii), and (iii) discussed in text, a new duplication event generate triangles contributing to the change in g_A.

protein, it will interact with the newly generated node,[13,18,41] leading to heterodimerization.

D. Neighbor connectivity

Degree correlation is the correlation between the degrees of two connected nodes. When nodes of high degree preferentially connect with other nodes of high degree, the network is said to be assortative, whereas when nodes of high degree preferentially connect with nodes of low degree, the network is said to be disassortative.[52–54] It was reported that social networks such as coauthorships of scientific papers and collaborations in the film industry are assortative, whereas technological and biological networks including food web, neural network, and protein-protein interaction networks are disassortative.[18–20,52–54]

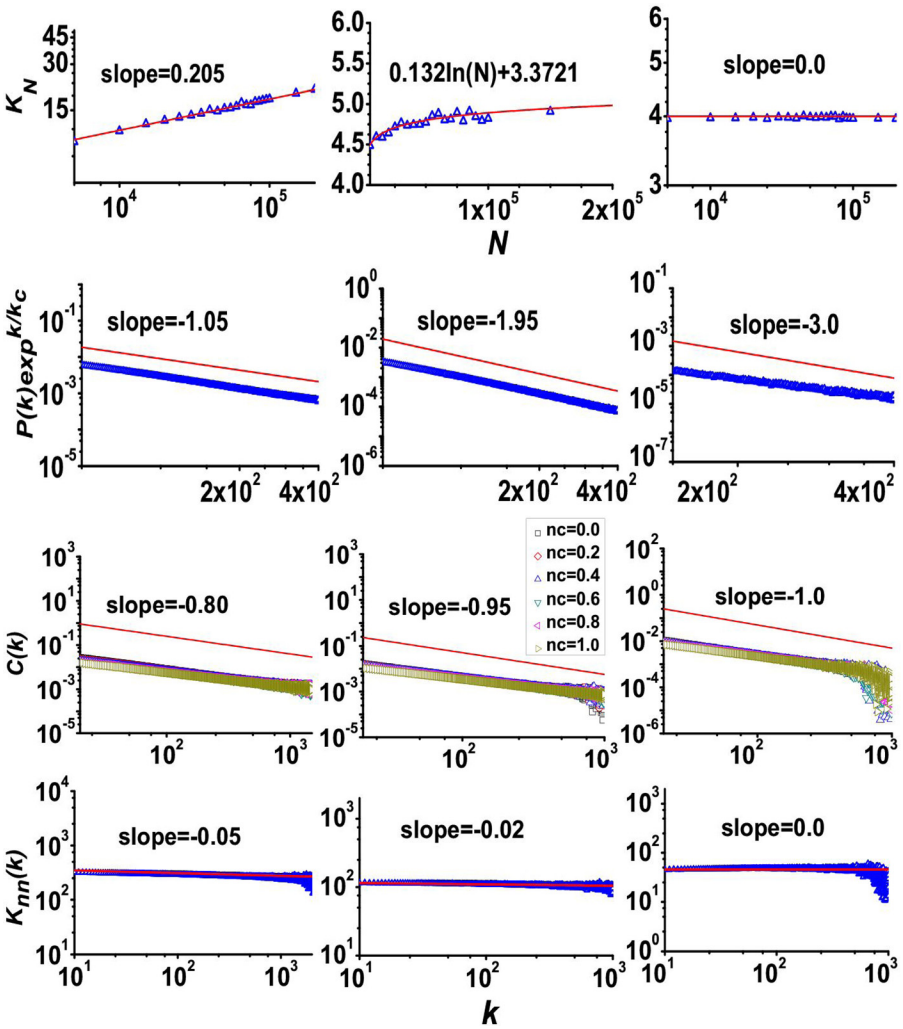

FIG. 4. Some simulation results with $p_h = 0.065$ unless otherwise specified. First row, K_N *vs.* N. Second to fourth rows, $P(k)\exp(k/k_c)$, $C(k)$, and $K_{nn}(k)$ *vs.* k, with $N = 2 \times 10^6$, and $p_d = 0.4$, 0.5, and 0.6, $n_c = 0.001$, 0.001, and 0.335 for the three columns, left to right, in the second row, $k_c = 225$, 400, and 600 for $p_d = 0.4$, 0.5, and 0.6, respectively. Lines indicate mean-field predictions (Table I), and values given for "slope" are from linear regression of simulation data.

The neighbor connectivity of node i, $K_{nn}(k_i)$, is the average of the number of links on its neighbors: $K_{nn}(k_i) = k_i^{-1} \sum_{B\in\mathcal{S}_i} k_B \equiv T_{nn}^{(i)}/k_i$, where $T_{nn}^{(i)}$ denotes the sum of the degrees of node i's neighbors.[52] The assortativity of a network is determined by the k-dependence of $K_{nn}(k)$, the average of all $K_{nn}(k_i)$ with $k_i = k$: the network is assortative or disassortative if $K_{nn}(k)$ is an increasing or decreasing functions of k, respectively.[52,54] When $K_{nn}(k)$ is independent of k, there are no degree correlations. Protein-protein interaction networks have been reported to be disassortative with $K_{nn}(k) \sim k^{-\gamma_{nn}}$.[11,18–20]

In order to derive the analytical solution for the neighbor connectivity of the DDHA model, we first compute $K_{nn}(k_A)$ of node A by considering the change in $T_{nn}^{(A)}$ after a time step ($t \to t+1$). This can happen in three ways: (i) A is duplicated (with duplicate A'). This adds $(1-p_d)p_h k_A + p_h((1-p_d)k_A + p_h + n_c)$ degrees to $T_{nn}^{(A)}$ provided A and A' dimerize, where the first term is the increased degrees coming from the k_A neighbors of A and the second terms coming from A' through actions (ii)–(iv); (ii) a neighbor $B \in \mathcal{S}_A$ is duplicated (call it B'). This generates (a maximum of) k_B neighbors for B', among which about $C(k_A)(k_A-1)$ are also neighbors of A. This adds $(1-p_d)((1-p_d)C(k_A)(k_A-1) + p_h) + (1-p_d)((1-p_d)(k_B-1)+p_h+n_c)$ degrees to $T_{nn}^{(A)}$ provided the new link between B' and A is preserved; and (iii) a node l that is neither A nor one of its neighbors is duplicated (call it $\hat{l}$). This adds $(1-p_d)m_{lA}n_c/N(t) + ((1-p_d)\ k_l + p_h + n_c)n_c/N(t)$ degrees to $T_{nn}^{(A)}$ provided a new $\hat{l}$–A link is added (with probability $\approx n_c/N(t)$), where m_{lA} is the number of nodes in the intersect $\mathcal{S}_l \cap \mathcal{S}_A$. Therefore, for large $N(t)$ and k_A, using the continuous approximation, the rate of change in $T_{nn}^{(A)}$ is

$$\begin{aligned}\frac{dT_{nn}^{(A)}}{dt} &= \frac{1-p_d}{N(t)}\left(2p_h k_A + (1-p_d)\sum_{B\in\mathcal{S}_A}\left((k_A-1)C(k_A) + (k_B-1) + \frac{2p_h+n_c}{1-p_d}\right) + \frac{p_h(p_h+n_c)}{1-p_d}\right.\\ &\quad \left. + \sum_{l\neq A, l\notin\mathcal{S}_A}\left(m_{lA} + k_l + \frac{p_h+n_c}{1-p_d}\right)\frac{n_c}{N(t)}\right)\\ &= \frac{1-p_d}{N(t)}\left(\zeta_0 + \zeta_1 k_A + \zeta_2 C_0 k_A^{2-\gamma_C} + (1-p_d)T_{nn}^{(A)}\right),\end{aligned}$$

where we used $k_l \approx k_A$, $C(k_A) \approx C_0 k_A^{-\gamma_C}$ (see Eq. (11)) and dropped $\mathcal{O}(1/N^2)$ terms, and $\zeta_0 = (p_h+n_c)^2/(1-p_d)$, $\zeta_1 = 4p_h + n_c + p_d - 1 + n_c/(1-p_d)$, $\zeta_2 = (1-p_d)$, and C_0 is a constant. For large $N(t)$ and k_A, using (8) and $\tilde{k}_A = k_A - \eta \approx k_A$, we obtain

$$\frac{dT_{nn}^{(A)}}{d\tilde{k}_A} = \tilde{\zeta}_0 \tilde{k}_A^{-1} + \zeta_1 + \zeta_2 C_0 \tilde{k}_A^{1-\gamma_C} + (1-p_d)T_{nn}^{(A)}\tilde{k}_A^{-1}, \quad (12)$$

where $\tilde{\zeta}_0 = \zeta_0 - \zeta_1\eta$. The solution is

$$T_{nn}^{(A)} = -\frac{\tilde{\zeta}_0}{1-p_d} + \frac{\zeta_1}{p_d}\tilde{k}_A + \frac{\zeta_2}{1+p_d-\gamma_C}C_0\tilde{k}_A^{2-\gamma_C} + T_0\tilde{k}_A^{1-p_d},$$

where T_0 is a constant depending on the initial condition of $T_{nn}^{(A)}$. We obtain in the large-k approximation

$$K_{nn}(k_A) \sim \tilde{T}_0 + \frac{\zeta_1}{p_d} + \frac{\zeta_2}{1+p_d-\gamma_C}C_0 k_A{}^{1-\gamma_C} + \mathcal{O}\left(k_A^{-p_d}\right), \quad (13)$$

where $\tilde{T}_0$ is a constant. Substituting the values for γ_C (Eq. (11)) for large k, we have

$$K_{nn}(k) \approx \frac{T_{nn}}{k} \sim k^{-\gamma_{nn}},\ \gamma_{nn} = \begin{cases} 0, & p_d \geq \frac{1}{2}; \\ 2p_d - 1, & p_d < \frac{1}{2}. \end{cases} \quad (14)$$

This suggests that a DDHA network is associative ($\gamma_{nn} > 0$) or neutral ($\gamma_{nn} = 0$), when p_d is $< \frac{1}{2}$ or $\geq \frac{1}{2}$, respectively, but is never dissociative ($\gamma_{nn} < 0$).

IV. NUMERICAL EXAMPLES

To verify the power-laws given in Table I, we carried out extensive *in silico* network construction following the four-step procedure stated earlier. Data from protein-protein interaction networks of yeast, fly, and human suggest $p_h \leq 0.1$.[39] Estimates of values for p_d and n_c from yeast protein-protein interaction network data give a ratio of $n_c/p_d \leq 1$.[13] Therefore, we used parameter values in the ranges $0.30 \leq p_d \leq 0.75$, $0.01 \leq p_h \leq 0.1$, and $0.001 \leq n_c \leq 1.0$ in our numerical simulations. Without loss of generality, a single connected random network containing 100 nodes with the average degree $K_{100} = 3.98$ was taken as an initial network. The sizes of the simulated networks were $N = 1\times10^5$, 5×10^5, 1×10^6, 2×10^6, and 3×10^6. For each set of parameters, 20 networks were generated and averaged measurements were taken.

Except for one case, the simulated results demonstrated clear power-law behavior (Fig. 4) and there was excellent agreement between the mean-field predictions and measured (by simulation) values of the power-law exponents (Fig. 2). The non-power-law case, in which K_N was predicted to have logarithmic dependence on N at $p_d = 1/2$, was also borne out by simulation (mid panel, top row, Fig. 4). The parameter n_c had little effect on the power-law exponent γ_C, verifying the effectiveness of the approximation method used to derive Eq. (10). In addition, our simulations revealed that the $P(k)$,

besides its power-law dependence on k, had a multiplicative exponential factor $\exp(-k/k_c)$. The exponential dependence was effective already when $k \sim 10$ (Fig. 5(a)) hence may not be attributed entirely to finite-size cut-off effect.[14,18,19,33] The general trend of the exponential dependence was that k_c increased with N (Fig. 5(b)), was independent of n_c (Fig. 5(c)), and increased with p_d (Fig. 5(d)). This corroborates the suggestions that $P(k)$ approaches its large-N asymptotic power-law slowly.[35,39]

Although mean-field theory predicted $K_{nn}(k)$ to grow with k when $p_d < 1/2$ (Table I), our simulation showed $K_{nn}(k)$ had no k dependence (Fig. 4), i.e., our result did not show a DDHA network to be assortative. In our DDHA model, nodes were randomly selected for duplication, and hence each node had equal probability of being duplicated at each time step. Similarly, links selected for deletion and for addition were both random. As a result, neither the connections between high-degree nodes and high-degree nodes nor those between high-degree nodes and low-degree nodes were favored. It is therefore no surprise that our model created networks that exhibited no assortativity.

We checked individual cases for further verification. When $p_d = 0.4$, it was measured that $C(k) \approx 0.1k^{-0.8}$ (Fig. 4). In this case, the third term on the right-hand-side of Eq. (13) $\zeta_2/(1+p_d-\gamma_C)C_0 {k_A}^{1-\gamma_C} \approx 0.1k_A^{0.2} \leq 0.5731$, for $k \leq 2000$. It was so small that the first two terms on the right-hand-side of Eq. (13) always dominated. In other cases for $p_d < 1/2$, the third term on the right-hand-side of Eq. (13) was also small and had no effect on the values of $K_{nn}(k)$. Thus, a DDHA network is not assortative.

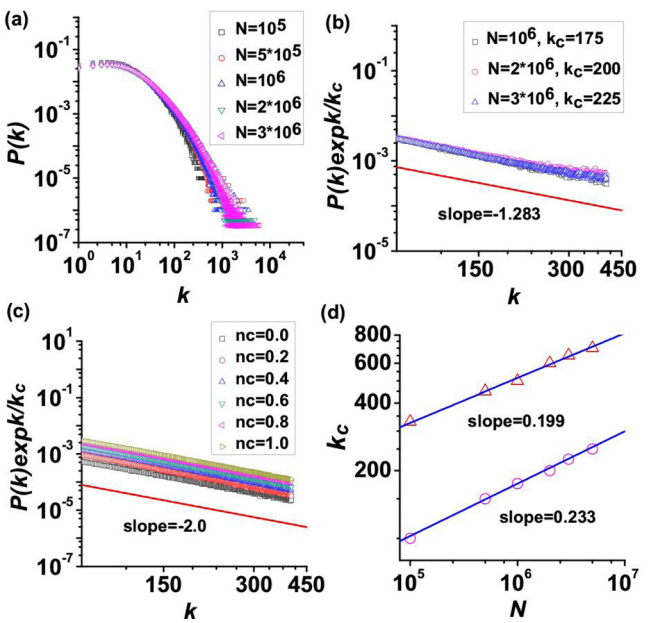

FIG. 5. Log-log plots for $P(k)$; in all cases $p_h = 0.08$. (a) (Simulated) $P(k)$ versus k for different network sizes; $p_d = 0.45$ and $n_c = 0.40$. (b) $P(k)\exp(k/k_c)$ versus k for different network sizes and best k_c; $p_d = 0.45$. (c) $P(k)\exp(k/k_c)$ versus k, with $k_c = 400$ for all n_c values; $N = 2 \times 10^6$ and $p_d = 0.50$. (d) Power-law relation between the parameter k_c and N; circles, $p_d = 0.45$; triangles, $p_d = 0.55$.

V. DISCUSSION AND CONCLUSION

We derived mean-field analytic forms of power-law exponents in the DDHA model for the average degree (β_K), degree distribution (γ_P), clustering coefficient (γ_C), and neighbor connectivity (γ_{nn}) as functions of parameters of the model: the deletion probability p_d, the heterodimerization probability p_h, and the addition probability p_c. All four exponents depended on the all important parameter p_d. Only when p_d is larger than the critical value 1/2 will a realistic finite average degree is recovered, i.e., networks may have the biological properties of sparseness and small-worldness. Only the exponent γ_C depended on p_h and none depended on p_c. Instead of having a constant value, here p_c was given for n_c/N, where n_c was a constant. As there was no other factors to help it counter the diminishing effect of the $1/N$ factor, terms dependent on p_c was destined to be small in the large N limit. Had p_c itself been made a constant, it would have impacted the role played by p_h in the large N behavior of γ_{nn} (Table I). However, a constant p_c would not be biological realistic.

The degree distribution was scale free only in the limit of large network size. For finite network size, the exponent γ_P had an exponential dependence on the degree, a dependence that weakens (as a power-law) with increasing network size. Except for one case—value of γ_{nn} when $p_d < 1/2$ (Fig. 2(d))—power-law exponents extracted from our large-scale simulations agreed extremely well with the mean-field results.

It has been reported that the number of links retained after gene duplication is considerably different between duplicates.[55] In the present work, we generated this difference between the duplicates by adopting the asymmetric divergence model, in which the removal of links might occur only in newly generated nodes. This allowed us to derive closed-form expressions for the power-laws. The model was adopted in several previous studies.[18,19,35,37,39,45] The other extreme is the symmetric divergence model, in which links are removed from both duplicates with equal probability.[19,41] Actual biological networks should lie somewhere between these two extremes but are expected to be heterogeneous in this aspect. Recent studies have indicated that the two models do not yield essential differences in network properties of the type considered here.[19,41]

Studies of the topological structure of protein-protein interaction networks in yeast, worm, fly, human, and malaria parasite have revealed that they tend to be disassortative ($\gamma_{nn} > 0$).[10,18–20] However, our model created networks that were neither assortative nor disassortative: $K_{nn}(k)$ was k-independent ($\gamma_{nn} \approx 0$). Theoretically, disassortative networks can be generated by favoring low-degree nodes in duplication,[19,42] for instance, by making the probability of duplicating a node *inversely proportional* to its degree.[42] Biologically, this bias is reasonable, because the cost for duplicating a node should be approximately proportionally to its degree. Owing to preferential duplication of low-degree nodes in this asymmetric model, links between a high-degree nodes and low-degree nodes are preferentially generated, resulting in a disassortative network. However,

we were not able to derive a close-form mean-field expression for $K_{nn}(k)$ for this asymmetric model.

ACKNOWLEDGMENTS

S.C. was supported by the National Science Foundation of China (Grant No. 11402100), the Tian Yuan Special Foundation (Grant No. 11326193), and the Research Foundation for Advanced Talents of Jiangsu University (Grant No. 13JDG027). Z.L. was supported by the National Science Foundation of China (Grant No. 11172158). H.C.L. was supported by the ROC National Science Council (Grant Nos. 101-2911-I-0008-001 and 102-2911-I-008-001) and the ROC National Center for Theoretical Sciences. The authors are grateful to the editor and anonymous reviewers for their constructive comments and suggestions that helped to improve the content and the quality of the paper.

[1]A. Prachumwat and W. H. Li, Mol. Biol. Evol. **23**, 30 (2006).
[2]A.-L. Barábasi and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[3]S. H. Yook, Z. N. Oltvai, and A.-L. Barábasi, Proteomics **4**, 928 (2004).
[4]P. Uetz *et al*., Nature **403**, 623 (2000).
[5]T. Ito *et al*., Proc. Natl. Acad. Sci. U. S. A. **98**, 4569 (2001).
[6]U. Güldener *et al*., Nucl. Acids Res. **34**, 436 (2006).
[7]H. Yu *et al*., Science **322**, 104 (2008).
[8]S. Li *et al*., Science **303**, 540 (2004).
[9]L. Giot *et al*., Science **302**, 1727 (2003).
[10]J. F. Rual *et al*., Nature **437**, 1173 (2005).
[11]U. Stelzl *et al*., Cell **122**, 957 (2005).
[12]D. J. LaCount *et al*., Nature **438**, 103 (2005).
[13]A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001).
[14]H. Jeong, S. P. Mason, A.-L. Barábasi, and Z. N. Oltvai, Nature **411**, 41 (2001).
[15]E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[16]E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).
[17]V. Noort, B. Snel, and M. A. Huynen, EMBO Rep. **5**, 280 (2004).
[18]T. Hase, Y. Niimura, T. Kaminuma, and H. Tanaka, PLoS ONE **3**, e1667 (2008).
[19]T. Hase, Y. Niimura, and H. Tanaka, BMC Evol. Biol. **10**, 358 (2010).
[20]S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[21]R. Albert and A.-L. Barábasi, Rev. Mod. Phys. **74**, 47 (2002).
[22]M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[23]S. Ohno, *Evolution by Gene Duplication* (Springer, Berlin, 1970).
[24]M. Lynch and J. S. Conery, Science **290**, 1151 (2000).
[25]S. D. Hooper and O. G. Berg, Mol. Biol. Evol. **20**, 945 (2003).
[26]M. Lynch and A. Force, Genetics **154**, 459 (2000).
[27]J. Zhang, Trends Ecol. Evol. **18**, 292 (2003).
[28]L. S. Hsieh, L. F. Luo, F. M. Ji, and H. C. Lee, Phys. Rev. Lett. **90**, 018101 (2003).
[29]S. A. Teichmann and M. M. Babu, Nat. Genet. **36**, 492 (2004).
[30]S. E. Brenner *et al*., Nature **378**, 140 (1995).
[31]S. A. Teichmann, J. Park, and C. Chothia, Proc. Natl. Acad. Sci. U. S. A. **95**, 14658 (1998).
[32]J. Gough *et al*., J. Mol. Biol. **313**, 903 (2001).
[33]R. V. Solé, R. Pastor-Satorras, E. Smith, and T. B. Kepler, Adv. Comput. Syst. **5**, 43 (2002).
[34]R. Pastor-Satorras, E. Smith, and R. V. Solé, J. Theor. Biol. **222**, 199 (2003).
[35]J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner, Phys. Rev. E **66**, 055101(R) (2002).
[36]A. Raval, Phys. Rev. E **68**, 066119 (2003).
[37]F. Chung, L. Lu, T. G. Dewey, and D. J. Galas, J. Comput. Biol. **10**, 677 (2003).
[38]A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, ComPlexUs **1**, 38 (2003).
[39]I. Ispolatov, P. L. Krapivsky, and A. Yuryev, Phys. Rev. E **71**, 061911 (2005).
[40]M. Middendorf, E. Ziv, and C. H. Wiggins, Proc. Natl. Acad. Sci. U. S. A. **102**, 3192 (2005).
[41]I. Ispolatov, P. L. Krapivsky, I. Mazo, and A. Yuryev, New J. Phys. **7**, 145 (2005).
[42]X. Wan, S. Cai, J. Zhou, and Z. Liu, Chaos **20**, 045113 (2010).
[43]G. Bebek *et al*., Theor. Comput. Sci. **369**, 239 (2006).
[44]K. Evlampiev and H. Isambert, Proc. Natl. Acad. Sci. U. S. A. **105**, 9863 (2008).
[45]A. Wagner, Proc. R. Soc. London, Ser. B **270**, 457 (2003).
[46]A.-L. Barábasi, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[47]M. E. J. Newman, C. Moore, and D. J. Watts, Phys. Rev. Lett. **84**, 3201 (2000).
[48]X. Cheng, H. Wang, and Q. Ouyang, Phys. Rev. E **65**, 066115 (2002).
[49]K. Takemoto and C. Oosawa, Math. Biosci. **208**, 454 (2007).
[50]P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*, 4th ed. (Cambridge University Press, Cambridge, 2007).
[51]D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[52]R. Pastor-Satorras, A. Vazquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[53]M. E. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[54]L. F. Costa, F. A. Rodrigues, G. Travieso, and V. Boas, Adv. Phys. **56**, 167 (2007).
[55]A. Wagner, Mol. Biol. Evol. **19**, 1760 (2002).