Influence of homology and node age on the growth of protein-protein interaction networks

Arianna Bottinelli,[1] Bruno Bassetti,[2,3] Marco Cosentino Lagomarsino,[4,5,6] and Marco Gherardi[2,3,*]

[1]*Department of Mathematics, Uppsala University, Uppsala, Sweden*
[2]*Dipartimento di Fisica, Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy*
[3]*Istituto Nazionale di Fisica Nucleare, Sezione di Milano, Via Celoria 16, 20133 Milano, Italy*
[4]*Genomic Physics Group, Unité Mixte de Recherche No. 7238 associée au Centre National de la Recherche Scientifique,*
"Microorganism Genomics," Paris, France
[5]*Université Pierre et Marie Curie, 15 rue de l'École de Médecine, 75005 Paris, France*
[6]*Dipartimento di Fisica, Università di Torino, via P. Giuria 1, Torino, Italy*
(Received 15 June 2012; published 31 October 2012)

Proteins participating in a protein-protein interaction network can be grouped into homology classes following their common ancestry. Proteins added to the network correspond to genes added to the classes, so the dynamics of the two objects are intrinsically linked. Here we first introduce a statistical model describing the joint growth of the network and the partitioning of nodes into classes, which is studied through a combined mean-field and simulation approach. We then employ this unified framework to address the specific issue of the age dependence of protein interactions through the definition of three different node wiring or divergence schemes. A comparison with empirical data indicates that an age-dependent divergence move is necessary in order to reproduce the basic topological observables together with the age correlation between interacting nodes visible in empirical data. We also discuss the possibility of nontrivial joint partition and topology observables.

PACS number(s): 87.15.km, 64.60.aq, 89.75.Fb, 87.23.Kg

I. INTRODUCTION

The protein-protein interaction (PPI) network represents the physical interactions between proteins in a cell [1]. The topological properties of this complex network provide an effective overview of the protein-protein interactions coded by a genome, with implications for the analysis of signaling and metabolic pathways [2].

In the course of evolution, a genome acquires new genes, and thus new proteins, by different evolutionary processes [3,4], which include gene duplication and horizontal transfers. These processes define groups of proteins with the same common ancestor, termed homology classes. Notably, homology classes follow well-defined quantitative laws with specific mathematical properties [4–6], dependent only on genome size and not on further details of a genome's evolutionary history [7,8].

Following gene duplications [9], proteins belonging to the same homology class can modify their binding interfaces to conserve ancient interactions, lose them, or evolve new ones. This process generates new PPI network configurations, which are subject to selective pressures of different kinds [10–12], and allow one to construct increasingly complex biomolecular machinery [13–15]. This mechanism of duplication divergence has inspired a thread of graph-growth modeling work within the physics and computational biology communities [16–22]. Generally speaking, these models generate random graph ensembles by iteratively adding new nodes that are initially copies of existing ones (and thus interact with all their binding partners) and subsequently lose and/or rewire interactions by a set of simplified prescription rules. This basic mechanism produces graph topologies resembling empirical PPI networks in many aspects. Comparison of model predictions and

empirical data leads to the hypothesis that duplication divergence can (at least in part) explain PPI network topologies [21,23,24], starting from the basic observation that duplicate proteins are often involved in similar protein-protein interactions [13,15].

While it appears that gene duplication plays a role in shaping PPI networks through evolutionary time [25], many questions remain open. For example, it has been pointed out that the duplication-age profiles naturally emerging from duplication-divergence models do not resemble empirical data and that the availability of binding interfaces could quite reasonably impose additional relevant constraints [26–28]. Accordingly, alternative models have been proposed, where the wiring rules account for these constraints [26]. Additionally, according to most of these models, collapsing multiple homologous neighbors of a protein into one neighbor should make the broad degree distribution considerably narrower, which does not seem to be the case in empirical data [29]. Thus the actual growth mechanisms of PPI networks is still under debate and it is unclear how much duplication divergence versus other constraints can account for the topology of empirical PPI networks [25,26,30]. Additionally, duplication-divergence models typically neglect the process of homology classes expanding and being formed within a genome and thus cannot describe how PPI network links are distributed among homology classes. However, the subdivision of genes into homology classes could constitute another relevant constraint for the PPI network's structure and should not be neglected *a priori*.

This work addresses the above issues through a modeling approach. We consider a (null) statistical graph-growth model describing the joint growth of the PPI network and homology classes structure. The output of the model is a growing graph whose nodes are partitioned into equivalence classes following the empirical size distributions of protein classes. The model defines a framework for testing alternative mechanisms of network growth, where duplication

*marco.gherardi@mi.infn.it

divergence can have different weight during the process and thus different consequences on the final properties of the network. Within this setting, we ask about the ingredients that can account for the joint growth of homology classes and network, as well as reproducing the main empirical observables such as degree distribution, degree correlation, and correlation between interacting duplication-age groups. In our analysis we find in particular that reproducing the empirical age correlation between interacting nodes requires a heavy bias on the duplication-divergence process, which must correspond to additional constraints of functional or physical origin.

II. BACKGROUND

A. Network growth by duplication divergence

Perhaps the simplest PPI network growth model incorporating the basic moves of duplication and divergence (DD) was introduced and studied in Ref. [19]. In this model the network grows by node duplication and subsequent deletion of some of the duplicate links with a prescribed probability (divergence). More precisely, at each step a randomly chosen network node is copied, initially inheriting all the interactions of the original node, and in a second substep the new node's links are deleted independently with probability $1 - \sigma$. If no link is left after divergence, the duplicate node itself is deleted, so that the network remains connected throughout its evolution. This process is completely asymmetric, meaning that the parent node (the one chosen for duplication) does not lose any connection, and the divergence process affects only the daughter. More general variants have been proposed, for instance, by relaxing the requirements of complete asymmetry and single-gene duplication [21] or by introducing rewiring between existing nodes (which can even become dominant in shaping the network [29]). For simplicity, we will restrict the discussion to the one-parameter model in the following.

One of the main features of this model is that the described mechanism leads to an effective preferential attachment principle since high-degree nodes are more likely to have a neighbor being duplicated by random choice. Specifically, the probability of a new link being attached to a node of degree k is proportional to k. As a consequence, the degree distribution of the growing network develops power-law tails $\sim k^{-\gamma}$ for large degrees [19]. Exponents in the range $\gamma \in [2,3]$ are realized by choices of $\sigma \in (0,1/2)$. Comparison with available subsets of empirical PPI networks yields values of the link-retention probability σ around $0.40(\pm 0.05)$ for *S. cerevisiae*, *D. melanogaster*, and *H. sapiens* [19]. The average total number of links $L(N)$ as a function of the network size N can also be predicted by mean-field calculations (see Sec. IV A).

B. Homology class partitioning by the Chinese restaurant process

Duplication plays a fundamental role in the evolution of homology classes as well [7], as it constitutes the main drive for class expansion, at least in eukaryotes. Equally, a genome innovation move (for instance, by horizontal transfer) causes the creation of new homology classes.

A simple class of partitioning processes incorporating the basic moves of class expansion and innovation is capable of explaining the scaling laws observed in domain-class partitioning [8]. The paradigm of these models is the so-called Chinese restaurant process (CRP) [6,31–33], which is the one that will be used here. In this process, at each iteration the genome goes from having n to $n + 1$ genes and either a new class is created (with probability p_{new}) or a domain is added to an existing class (with probability $p_{\text{old}} = 1 - p_{\text{new}}$). A crucial ingredient of the CRP is the dependence of p_{new} and p_{old} on the size of the growing proteome, whose effect is to reproduce in the model the observed sublinear scaling of the number of domain classes $F(N)$ with genome size N:

$$p_{\text{new}} = \frac{\alpha F(N) + \theta}{N + \theta}, \quad p_{\text{old}} = \frac{N - \alpha F(N)}{N + \theta}, \tag{1}$$

where $\alpha \in (0,1)$ and $\theta \geqslant 0$ are parameters of the model. (The extreme cases $\alpha = 0,1$ could be included, but we will neglect them here for clarity.) The per-class probability of duplication is defined as

$$p_{\text{old}}^{(i)} = \frac{j_i - \alpha}{N + \theta}, \tag{2}$$

where j_i is the size of the ith class. This corresponds to an asymptotically uniform extraction, which realizes an effective preferential attachment principle. The parameter α describes the dominance of innovation over duplication, while θ is a fixed size scale at which preferential attachment sets in. Mean-field calculations, supported by simulations, show [8] that the asymptotic behaviors of the class-size distribution $f(j,N)$ and of the total number of classes $F(N)$ are

$$f(j,N) \sim j^{-(1+\alpha)}, \quad F(N) \sim N^{\alpha} \tag{3}$$

for large N and j. As a consequence, p_{new} and p_{old} scale as

$$\begin{aligned} p_{\text{new}} &\sim \alpha N^{\alpha-1}, \\ p_{\text{old}} &\sim 1 - \alpha N^{\alpha-1}. \end{aligned} \tag{4}$$

These predictions are in good qualitative agreement with empirical data for prokaryotic proteomes [7,8].

III. MODEL AND METHODS

A. Definition of a statistical model combining genome partitioning and network growth

As we discussed, from a simplifying perspective, the growth of PPI networks and genome partitioning in homology classes are produced by essentially the same basic evolutionary moves of innovation and duplication on the genes. For this reason, the model proposed here is defined by abstract realizations of these basic moves on the level of both the network and the homology classes. This is achieved by a simple coupling between the duplication-divergence model of network growth and the CRP partitioning, as reviewed in Sec. II. In particular, a class expansion move is associated with a network duplication move
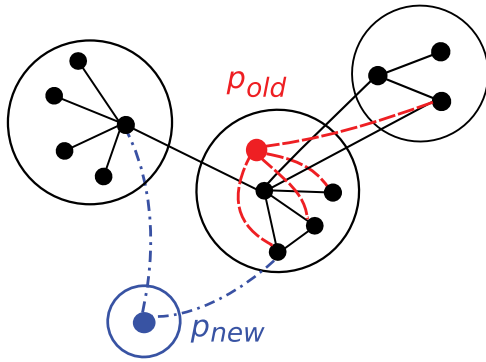
FIG. 1. (Color online) Illustration of the moves in the DDIW model. At each step either a new class containing one node is added and the new node is linked to one or more existing nodes (innovation wiring, done with probability p_{new}) or a randomly chosen node is duplicated inside a class and the replica's links activated independently with probability σ (duplication divergence, done with probability p_{old}). Solid circles are nodes and lines are links; large circles are homology classes; the red node and its dashed links are the results of a duplication-divergence move; the blue node and its dot-dashed links are the results of an innovation-wiring move.

and a proteome innovation move is associated with a network move wiring the new node to the existing network. Thus the model could be termed duplication-divergence innovation wiring (DDIW) and describes the growth of homology classes and the PPI network jointly.

Let p_{new}, $p_{\text{old}}^{(i)}$, and $p_{\text{old}} = \sum_i p_{\text{old}}^{(i)}$ be defined as in Eqs. (1) and (2) in terms of the number of classes $F(N)$ and the size of the ith class j_i. The basic data structure of the model includes the topology of the PPI network and the information on the partitioning of its nodes (see Fig. 1). Given a proteome or network of size N, the growth process is defined by the following two rules acting on the classes and the graph topology.

(I a) Duplication (classes). Choose a class i with probability $p_{\text{old}}^{(i)}$ and duplicate a randomly chosen target node inside class i.

(I b) Divergence (network). Attach the new node to each of the target's neighbors independently with probability σ.

(II a) Innovation (classes). Otherwise (i.e., with probability p_{new}), create a new node in a new class.

(II b) Wiring (network). Attach the new node to one or more nodes in the existing network, independently of their classes. (The additional rules describing this step are listed in Sec. III B.)

Altogether, there are three parameters governing the dynamics: $\alpha \in (0,1)$, $\theta \geqslant 0$, and $\sigma \in (0,1]$. Notice that while the network dynamics is dependent on the configuration of the partitioning, the evolution of the latter is not affected by what happens at the network level. Therefore, partitioning is assured by definition to reproduce the CRP predictions for all choices of the parameters. Notice that class expansion can also occur by horizontal transfer of members of an existing homology class [34], but we will disregard this process here. In fact, while this mechanism is widespread in bacteria, we

found that there was no need to incorporate it explicitly in the model in order to have a good fit with data for both networks and homology classes.

Technically, we choose a slightly different divergence rule from the model of Ref. [19]. In order for duplication to always be successful (i.e., no node being left without any links) we impose a randomly chosen link to be conserved and divergence to be performed on the remaining ones, i.e., the model assumes that each duplicated node is preserved by selection and cannot be disconnected from the existing network. The same hypothesis holds for the original model, but is implemented by removing the disconnected nodes. The different implementation implies that the divergence rule explained in Sec. II yields a degree-dependent probability of duplication since less connected nodes are more prone to have all their links disconnected; the rule used here instead assigns the same probability of duplication to every node. Despite this bias, the modified model incorporates the same basic mechanisms as the previous one and we verified that it leads to the same qualitative results (some features match also quantitatively; see Sec. IV A). The main rationale behind this choice is a simplification of the mean-field equations, as it makes it unnecessary to estimate the number of deleted nodes.

The initial condition will be chosen as the complete three-graph, which is the smallest nonbipartite network. Results do not change appreciably by starting with different small networks (we did not study systematically the dependence of the results from initial conditions built as large networks). We choose to exclude self-interactions from the model as they play a biologically distinct role in the network and they probably deserve to be considered separately [14].

B. Model variants allowing study of the effect of different growth mechanisms on the topology

The wiring rule is not completely specified by the definitions above. Its implementation will be given in the following. At the network level, the rules concerning the topology can be modified without affecting the basic structure of the model. Here we study a minimal version and consider different variants for such rules, which allow us to address the recently formulated problem of the age dependence of empirical interactions [26].

We start by focusing on the wiring move. Once introduced, the new node can be attached to a single node chosen in the existing network by a preferential-attachment (PA) or antipreferential-attachment (APA) principle with respect to the old node's degree. The former alternative describes the tendency of new, specialized proteins to interact more likely with old proteins that perform basic tasks, the latter reflects the relationship between the binding probability and the available interaction surface of existing nodes [26]. Alternatively, the new node can be wired to a size-dependent or configuration-dependent number l of existing nodes.

Other modifications are possible for the divergence move, for example, by making the link-retention probability σ depend on the current configuration of the network or on the age difference between the two nodes that are connected by the
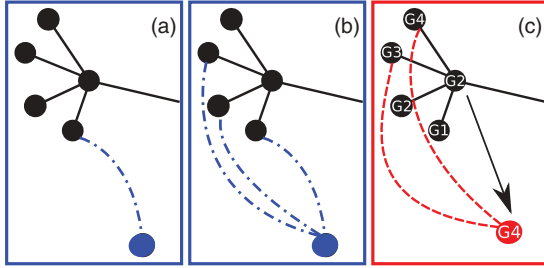
FIG. 2. (Color online) Variants of the model. Symbols and colors have the same meaning as in Fig. 1. (a) APA DDIW (antipreferential-attachment innovation with a single link). During innovation, the new node carries one new link whose target node is chosen with probability inversely proportional to its degree. (b) EAPA DDIW (antipreferential-attachment innovation, with multiple links). During innovation, the new node carries a number of links proportional to the current average degree. (c) AD DDIW (age-dependent divergence). During divergence of a duplicated node, the probability of keeping a link depends on the difference in age between the two nodes linked (higher age differences corresponding to lower probabilities).

link considered by divergence. Here we consider three main variants[1] (see Fig. 2).

(A) *APA DDIW.* The wiring move establishes a single new link between the new node and an existing node i of degree k_i, chosen with probability proportional to $1/k_i$. This antipreferential rule reflects the growing of the binding probability with the interaction surface available.

(B) *Extensive APA (EAPA) DDIW.* The wiring move attaches the new node to $l = [\gamma \langle k \rangle]$ existing nodes, chosen with antipreferential attachment; $[\gamma \langle k \rangle]$ is the closest integer to a fraction $\gamma \in (0,1)$ of the mean degree in the present configuration.

(C) *Age-dependent (AD) DDIW.* The wiring move is the same as for variant A. The divergence step implements a kind of preferential attachment that takes into account the node's age in the following way. Let a_i be the age of node i, i.e., the number of iterations the process underwent since the node was born. A link to node i inherited from the target node is kept with probability 1 if $a_i < \sigma N$, where N is the size of the network, and with probability 0 otherwise. This rule implements non-neutral selective pressure towards maintaining ancient well-established basic cellular machinery.

C. Empirical data sets and data analysis methods

Data for protein binding is obtained from the most recent (October 2011) Database of Interacting Proteins (DIP) [35]. We filter out self-interactions between single proteins and interactions between proteins expressed by different genomes; different strains are considered as different organisms. Moreover, we exclude all virus data and all networks with less than 10 nodes. We end up with 1 *archaeon*, 14 *bacteria*, and 7 *eukaryotes*; a list of all organisms considered in the study of

---

[1] We have also studied the PA DDIW variant, where the new node is wired to the old network according to a preferential-attachment principle, but the results did not show a significant difference from the APA DDIW variant described here and thus will not be reported.

TABLE I. Genomes from DIP [35] and corresponding values of the number of nodes N and number of links L. Note that the bacteria with small number of nodes are heavily undersampled in the data set, so that the number of effectively significant points is low (see Sec. V).

Organism	No. of nodes	No. of links
archaea		
Sulfolobus solfataricus	14	9
prokaryotes		
Arabidopsis thaliana	136	153
Bos taurus	30	23
Caenorhabditis elegans	2647	3985
Chlamydomonas reinhardtii	14	17
Danio rerio	13	9
Drosophila melanogaster	7500	22 737
Gallus gallus	11	6
Homo sapiens	1850	2370
Mus musculus	524	457
Pisum sativum	10	12
Rattus norvegicus	147	112
Saccharomyces cerevisiae	4998	21 881
Schizosaccharomyces pombe	80	160
Xenopus laevis	20	14
bacteria		
Bacillus subtilis	34	24
Caulobacter crescentus	18	11
Escherichia coli	2640	11 545
Helicobacter pylori	700	1354
Mycobacterium tuberculosis	13	9
Synechocystis sp.	32	29
Xanthomonas campestris	11	10

network topology is presented in Table I, together with the observed number of proteins N and interactions L. Notice that the networks we can construct from DIP include only subsets of the full proteomes. For example, the *C. elegans* network in our data set is smaller than that of *S. cerveisiae*, despite its genome being much larger, possibly creating significant undersampling problems in the data. See Sec. V for a discussion of this issue.

Homology classes are built starting from the SUPERFAM-ILY database for domain assignment [36]. We reconstruct the domain architectures as ordered lists of domains and gaps; a gap is defined as a subsequence of 100 or more AA not scored for domain [37]. Two proteins are in the same homology class if their architectures are exactly matching. We also tested a more relaxed criterion (allowing for repetitions of domain architectures) and obtained the same results as those presented in the following for the stricter criterion. Moreover, we also considered data restricted to longest transcripts in eukaryotes, finding no difference in the scaling (we remark that longest-transcript data in the data set are very incomplete, so we will not include them in the forthcoming analysis). We filter out genomes with more than 19 000 assignments; altogether, we work with data for 1384 organisms—87 *archaea*, 1077 *bacteria*, and 220 *eukaryotes*—for the homology classes, but only 22 networks with sufficiently large sampling of the interactions.

Beside network topology and homology classes, we are interested in evolutionary ages of proteins. For the proteome of *S. cerevisae*, we use data from Wapinski *et al.* [38], where

duplication events for a number of genes of *S. cerevisiae* are divided into ten classes, labeled A, B, C, D, E, WGD, G, H, I, and J, depending on when in the evolutionary history of *Ascomycota* they occurred (class A being the more recent). We further group these classes into four superclasses (labeled G1–4), keeping the whole-genome duplication (WGD) alone, due to the abundance of its elements:

$$G1 = I + J, \quad G2 = G + H,$$
$$G3 = WGD, \quad G4 = A + B + C + D + E.$$

By this procedure, we assign 210 genes to age group G1, 85 to G2, 691 to G3 (WGD), and 91 to G4. The age of a protein is defined as the superclass of the oldest duplication event in which it is reported to be involved. It should be noted that the WGD has a different phenomenology than the single-gene duplication events considered here; we do not exclude it from our data, but its modelization is beyond the scope of the present work (see Ref. [22]). In order to evaluate the history dependence of protein interactions, we use the interaction density $D_{m,n}$ between two age groups m and n as an indicator of age correlation. It is defined, following Ref. [26], as

$$D_{m,n} = \log_2 \left[\frac{L_{m,n}}{E_{m,n}} \frac{N(N-1)}{2L} \right], \tag{5}$$

where $L_{m,n}$ is the number of links between the age groups m and n and $E_{m,n}$ is the number of possible links between nodes of the two groups, which depends on only the number of nodes in m and n. The average interaction density gradient, defined as [26]

$$\Delta D = \sum_{n=2}^{4} \sum_{m<n} (D_{m+1,n} - D_{m,n}), \tag{6}$$

measures the overall correlation present between the ages of proteins; a positive value indicates that newer nodes preferentially link with newer nodes. We will use the sign of ΔD as a marker of correlation or anticorrelation between ages. Fits of data against nonlinear analytic expressions are performed by minimization of the squared residuals through the standard Levenberg-Marquardt method and are systematically checked for stableness under the introduction of a cutoff on small-size data.

IV. RESULTS

We ask under which conditions the model or its variants fulfill the following requirements. First, it should qualitatively reproduce the features of both the duplication-divergence and CRP "pure" models. Second, it describes the enriched data structure of network plus homology classes and it should predict the behavior of joint topology-partition observables, including the history dependence of interactions.

All variants of the DDIW model reproduce the same homology-class scaling as the pure CRP essentially because the class partitioning is not affected by the network dynamics by definition. A simple scaling argument suggests that the duplication-divergence predictions are expected to be recovered for large N since the scaling of p_{new} and p_{old} [Eq. (4)] shows that duplication becomes dominant in this regime. Therefore, the model is expected to behave as pure

duplication divergence in the large-N limit; it remains to clarify what happens at intermediate values of N. In the following sections we address some of these questions; the large-N behavior is clarified by mean-field techniques, while finite values of N are studied by means of numerical simulations. The analysis of how the partitioning into homology classes correlates with the network structure will be briefly addressed in Sec. V, but its systematic study will be left to future work.

A. Mean-field theory accurately predicts scaling of the total number of links

Mean-field calculations give reliable estimates for the behavior of the duplication-divergence network growth model and for the class-expansion innovation model separately [8,19], therefore it makes sense to apply the same procedure to the joint model. The mean-field approach essentially consists in neglecting the fluctuations due to the statistical nature of the models and writing "macroscopic" differential equations for the average quantities, which can be treated analytically. In this section we will use this tool to study the average total number of links $L(N)$ as a function of the number of nodes N for the variants of the joint-evolution model described in the preceding section. In principle, other characteristics of the network may be accessible through mean-field calculations, such as the degree distribution, but we will not treat them here.

For the duplication-divergence model alone (in the variant defined in Sec. III A), the simplification we introduced allows us to write a slightly more general expression for $L(N)$ than that obtained in Ref. [19]. Let N_k be the average number of nodes with k links in a network of size N (the average is intended on all realizations of the stochastic process up to size N). Clearly,

$$\sum_k N_k = N \tag{7}$$

and

$$\sum_k k N_k = 2L(N), \tag{8}$$

where the sums are extended to all possible values of the degree k (say, from 1 to ∞); $L(N)$ varies at each duplication following the mean-field equation

$$\Delta L(N) \simeq \sum_k \frac{N_k}{N} [1 + (k-1)\sigma], \tag{9}$$

where $\Delta L(N) \equiv L(N+1) - L(N)$. The summand takes into account the duplication of a node of degree k, which is performed with probability N_k/N. The term in square brackets reflects the fact that by definition at least one of the links is maintained, while the other $k-1$ links are kept independently with probability σ. Performing the sum by applying identities (7) and (8) yields

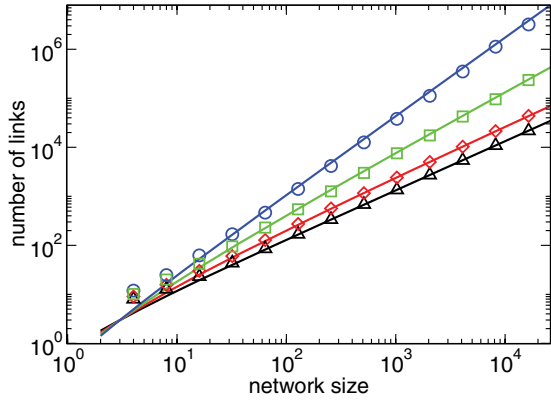$$\Delta L(N) \simeq (1-\sigma) + 2\sigma \frac{L(N)}{N}. \tag{10}$$

FIG. 3. (Color online) Average total number of links $L(N)$ as a function of network size for the pure duplication-divergence model. Solid lines show the mean-field prediction, while symbols are the results of numerical simulations (100 realizations); error bars are smaller than symbols. Triangles correspond to $\sigma = 0.2$, diamonds to 0.4, squares to 0.6, and circles to 0.8.

This can be approximated by the (large-N) differential equation

$$\frac{dL}{dN} \simeq (1 - \sigma) + 2\sigma \frac{L}{N}. \quad (11)$$

Solving this equation with a formal initial condition $L(N_0) = L_0$ gives the solution

$$L(N) \simeq \frac{1-\sigma}{1-2\sigma}N + \left(L_0 - \frac{1-\sigma}{1-2\sigma}N_0\right)\left(\frac{N}{N_0}\right)^{2\sigma}. \quad (12)$$

In the following we will fix the initial condition to the complete three-graph [$L(3) = 3$] in order to avoid the proliferation of irrelevant parameters. The presence of two regimes is apparent, where the first or the second term dominates, corresponding to $\sigma < 1/2$ and $\sigma > 1/2$, respectively. Notice the alternating-sign pattern of the corrections to scaling, which can cause the observation of a small-size effective exponent higher than both 1 and 2σ (see Sec. IV B). By taking the limit $\sigma \to 1/2$ one has $L(N) = 1/2(N \ln N) + O(N)$, thus recovering the three different regimes of the original DD model [19]. Figure 3 shows that mean-field predictions correctly reproduce the results of simulations, even for fairly small values of N; small deviations from mean field appear only for large values of σ, which are not very relevant empirically, as the link density would be too high compared to empirical data.

We now consider the different variants of the joint DDIW model. The increase in the total number of links at each step is given by either l (if the innovation move is chosen) or the same sum as in Eq. (9) (if the duplication move is chosen). We will not consider variant AD DDIW since in this case solving the mean-field equation for the number of links requires knowledge of the node-age distribution in the network. Thus, for the first two variants we have

$$\Delta L(N) \simeq p_{\text{new}}l(N) + p_{\text{old}}\sum_k \frac{N_k}{N}[1 + (k-1)\sigma],$$

where $l(N)$ is the average of l over realizations of the process up to size N. By plugging in the asymptotic forms (4) and

taking the continuum approximation as in Eq. (10) we obtain

$$\frac{dL}{dN} \simeq \alpha N^{\alpha-1}l(N) + (1 - \alpha N^{\alpha-1})\left(1 - \sigma + 2\sigma\frac{L}{N}\right), \quad (13)$$

which has to be solved separately for the two cases $l(N) = 1$ (APA DDIW) and $l(N) = \gamma 2L/N$ (EAPA DDIW). The solution is presented in some detail in the Appendix; we concentrate here on the asymptotic behavior. Up to exponential corrections of the form $\exp(x^{-\eta})$ with $\eta > 0$, the number of links scales as

$$L(N) \sim aN^{2\sigma} + bN \quad (14)$$

for the APA DDIW variant and as

$$L(N) \sim cN^{2\sigma} + dN^{\alpha} + eN \quad (15)$$

for the EAPA DDIW variant; a and b are functions of σ and α, while c, d, and e are functions of σ, α, and γ. The exponential corrections are proportional to $\exp(p_{\text{new}})$, which indicates the influence the partitioning process has on the early stages of the growth process. Figure 4 shows a comparison between mean-field results and numerical simulations. Deviations are apparent for $(\sigma, \alpha) = (0.6, 0.6)$ and $(0.2, 0.6)$ in the EAPA DDIW variant, but theoretical predictions are accurate for other values and the APA DDIW variant. The structure of the power-law corrections to scaling is similar to that of the pure DD model and as long as $\alpha < 2\sigma$ (which is the case for the universal fits to empirical data presented in Sec. IV B) the asymptotic behavior depends only on σ, up to the subleading order. This suggests that the scaling behavior of the hybrid DDIW model is to a certain extent robust with respect to the details of the innovation dynamics.

Concerning the scaling of the number of links in the AD DDIW variant (Fig. 4), note that in this case the definition of σ does not allow one to interpret this parameter as the average fraction of links retained after node duplication. This is due to a nontrivial correlation between node age and node degree, which is not straightforward to include in the mean-field calculation. Nevertheless, numerical simulations indicate that the asymptotic behavior of $L(N)$ derived for the APA DDIW variant also holds for the AD DDIW variant, up to a rescaling of σ. This can be seen in Fig. 4, where the mean-field predictions are compared with numerical results for the rescaled values $\hat{\sigma}$.

B. Scaling of the number of links and classes as functions of genome size is reproduced by universal parameters independent of the model variant

Having established that the scaling for the number of links is captured by simple mean-field estimates and indicates well-defined parameter regimes, we constrain the parameters by comparing them to the available empirical data. Specifically, we fix the three parameters α, θ, and σ by fitting the mean-field expressions against data for homology classes and PPI networks.

The calculations presented in the preceding section and in the Appendix are not easily extendable to finite values of θ; they are valid in the asymptotic limit or when $\theta = 0$. Nonetheless, a corrected expression of $F(N)$ for the case $\theta > 0$ and $F(1) = 1$ can be obtained (see Ref. [8]) and it is the one
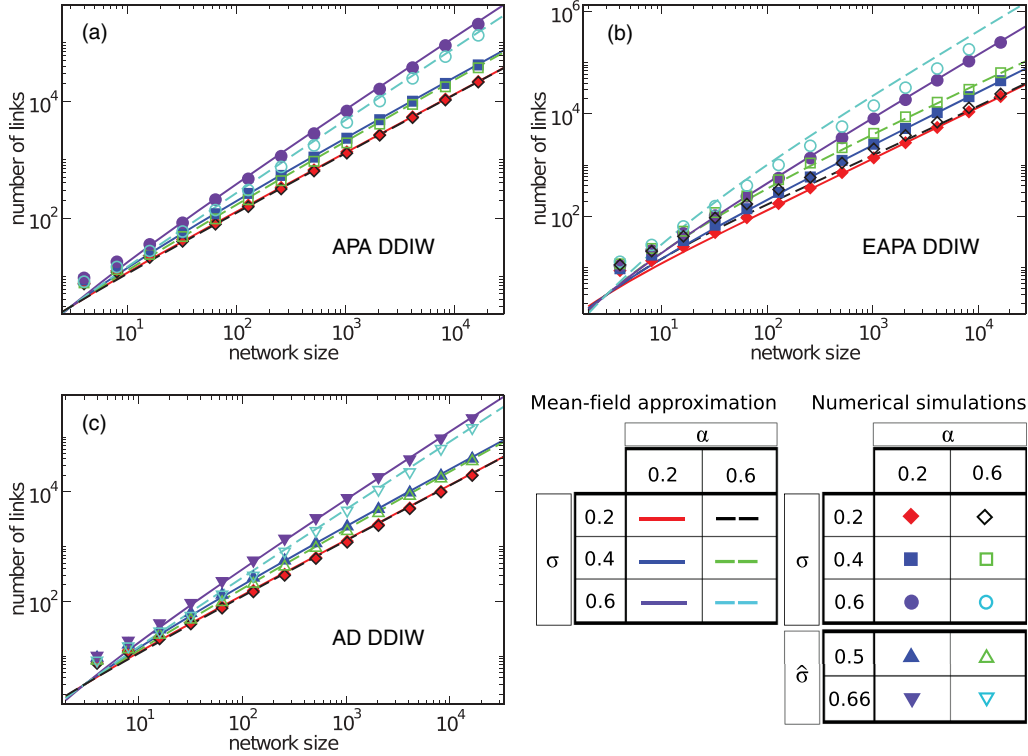
FIG. 4. (Color online) Scaling of the average total number of links $L(N)$ as a function of network size, captured by simple mean-field estimates for all model variants. Solid curves show the mean-field prediction, while symbols are numerical results averaged over 100 realizations. (a) APA DDIW (antipreferential attachment innovation move) variant: The mean-field estimate agrees with the simulation results. (b) EAPA DDIW (antipreferential attachment innovation move with an extensive number of links) variant: Deviations are present for the larger values of α, but there is good agreement between mean-field estimate and simulations. (c) AD DDIW (age-dependent duplication divergence): In this case, the mean-field estimates with the same slope as in (a) are valid for simulations with a rescaled value $\hat{\sigma}$ of the parameter σ, related to link retention (note that σ is not the link retention probability in this variant; see the text).

we use here to fit the number of homology classes as a function of genome size,

$$\alpha F(N) + \theta \sim \frac{\alpha + \theta}{(1 + \theta^\alpha)}(N + \theta)^\alpha. \tag{16}$$

We perform the fits on the empirical data set for homology classes defined by protein domain architectures, described in Sec. III C. By taking into account all data, we obtain $\alpha \simeq 0.42$ and $\theta \simeq 124$. Estimates change slightly by imposing a cutoff, since after $N \approx 1000$ data show a clearer power law. By including only data with $N \geqslant 1000$ we obtain $\alpha \simeq 0.43$ and $\theta \simeq 118$, which are compatible with the results obtained from the whole available range of genome sizes (without any cutoff). We will use the following estimates for all forthcoming computations:

$$\alpha = 0.43, \quad \theta = 121.$$

The theoretical mean-field curve for $F(N)$ is plotted against data in Fig. 5(a).

Turning to the network data and the fit for $L(N)$, a non-null value of θ is not expected to modify the asymptotic behavior, but to act only on the prefactors. Therefore we use the mean-field $L(N)$, even if the homology-class fits give a non-negligible value of θ. Note that the same scaling seems to apply to the prokaryotic genomes as well, despite their network dynamics not being dominated by duplication divergence;

indeed, homology classes prevalently expand by horizontal gene transfers [34]. A more precise analysis of this behavior can only be carried out with more reliable and abundant data; here we use both prokaryotic and eukaryotic data, as described in Sec. III C. Fits against the mean-field predictions for $L(N)$ given in the Appendix (with $\alpha = 0.43$ fixed) yield

$$\sigma = 0.457(10)$$

for the APA DDIW variant and

$$\sigma = 0.421(9) \quad (\gamma = 1),$$
$$\sigma = 0.460(10) \quad (\gamma \to 0)$$

for the EAPA DDIW variant; values of γ between 0 and 1 give estimates between the two extremes. In contrast, a fit against the pure DD prediction (12) gives

$$\sigma = 0.446(10).$$

We tested the stability of the foregoing fits by increasing the cutoff on the network size N from 10 to 100. The values do not change appreciably; errors increase by approximately 50%. A comparison between DIP data [35] and simulations of the AD DDIW variant, whose exact behavior cannot be calculated via mean-field derivation, gives approximately $\sigma \simeq 0.5$, which corresponds to an effective link-retention probability around 0.4 [see Fig. 4(c)].
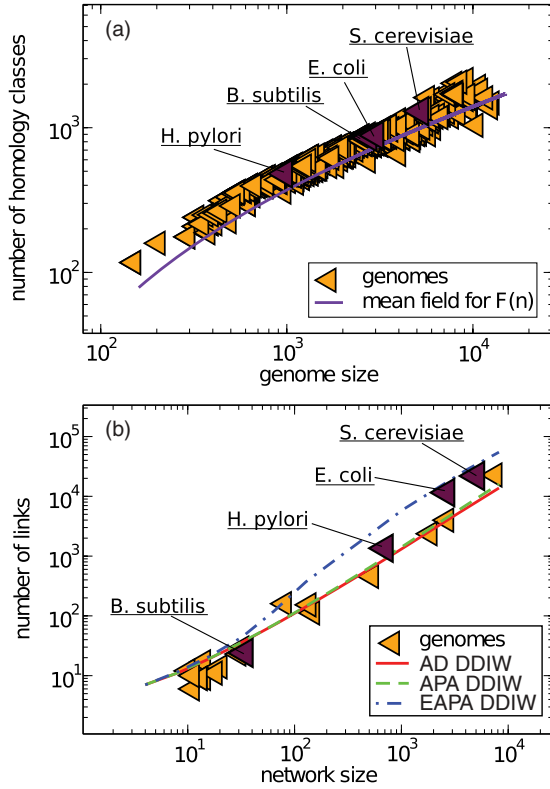
FIG. 5. (Color online) Universal behavior for the number of homology classes and the number of network links. (a) Number of homology classes versus total number of proteins. Symbols are data from the SUPERFAMILY database [36] and the line is a two-parameter (α, θ) fit from the model [Eq. (16)]. (b) Number of links versus size of sampled network. Symbols are data from the DIP dataset [35] and lines are the results of simulations for the three variants of the model, with all parameters fixed by fits. Darker triangles point out some examples of well-known genomes. Note that many networks (e.g., *B. subtilis* and all the triangles of smaller size) are heavily undersampled in the data set (see Sec. V).

Figure 5(b) shows numerical results for the three variants of our model (with α, θ, and σ fixed by the above fits) superimposed on the data from DIP. The initial network was chosen as the complete three-graph (see Secs. III A and IV A). Finite-size effects can be seen, especially for the EAPA DDIW variant, but the trend is consistent. The results for all parameters are compatible with each other, therefore we regard this as a model variant-independent fit: The two parameters α and σ can then be seen as universal (model variant-independent) quantities governing the scaling laws observed in genomes. Very similar values of σ (≈ 0.4) were also found in Ref. [19] with a more detailed analysis of the degree distribution of PPI networks and comparison to the model. Note that a simple fit of the form $L(N) \sim N^{2\sigma}$ on the empirical data would yield $\sigma = 0.52$, i.e., it would suggest a crossover regime. According to our analysis, such a higher exponent appears instead to be an artifact due to the cooperation of two terms ($N^{2\sigma}$ and N) with smaller exponents but with alternating signs.

Note that in principle the mean-field derivation is valid in the large-N limit. Figure 5(b) shows that differences in the fit

results can be noticed for small cutoffs. We chose a low cutoff to genomes with less than 10 nodes in order to show this. It must be noted, however, that many small networks are actually quite large in reality, but extremely undersampled in the data set.

C. Comparison with the empirical network of yeast reveals the necessity of age-dependent divergence

We now turn to the question of the topological properties and the age dependence of interactions. In order to perform a qualitative comparison between properties of an empirical PPI network and the results of computer simulations for the three model variants described above we choose the case of baker yeast *S. cerevisiae*, where reliable estimates of the age of nodes can be obtained from the literature (see Sec. III C). As pointed out in Refs. [26,39], while standard duplication-divergence network growth models well reproduce topological features of protein-protein interaction graphs, such as degree distribution and clustering coefficient, they fail to capture the empirically observed correlation between the evolutionary ages of interacting proteins. As they discuss, this might be obtained from an antipreferential-attachment principle if it becomes a dominant mechanism in defining the network topology.

In order to monitor the topology and history dependence of interactions we considered the following observables. (i) We measured two relevant topology-related quantities: the degree distribution n_k, defined as the fraction of nodes of degree k, and a measure of the degree-degree correlation, called d_k, defined as the average over all nodes of degree k of the mean degree of their neighbors. (ii) To check for age-age correlations we employed the interaction density $D_{m,n}$ and the interaction density gradient ΔD introduced in Sec. III C.

The behavior of the observables considered is shown in Fig. 6 for both the empirical PPI network of yeast and numerical simulations of the DDIW model variants. The model parameters are those obtained in Sec. IV B. As we pointed out before, results for the age class corresponding to the WGD in Fig. 6 should be taken carefully since homologs in that class were duplicated in a phenomenologically different event. For assessing how successful each variant is in reproducing the degree distribution and the degree correlation we adopt a qualitative criterion. Specifically, we consider a monotonically decreasing behavior of the two topological quantities to be compatible with empirical data since this is the behavior observed in yeast. Concerning node-age correlations, we measure the interaction density gradient and verify whether it is positive or negative; the reference data for yeast give a positive ΔD. The whole comparison is carried out in the same spirit as in Ref. [26].

The APA DDIW variant successfully reproduces the empirical degree correlation and degree distribution, but not the pattern of correlation between age groups ($\Delta D < 0$). In this model the innovation move gives a negligible contribution to network topology because the corresponding number of links is always subdominant. In fact, we verified that changing the antipreferential-attachment innovation move into preferential attachment has little or no effect on the main topological observables. As expected from this argument, this model
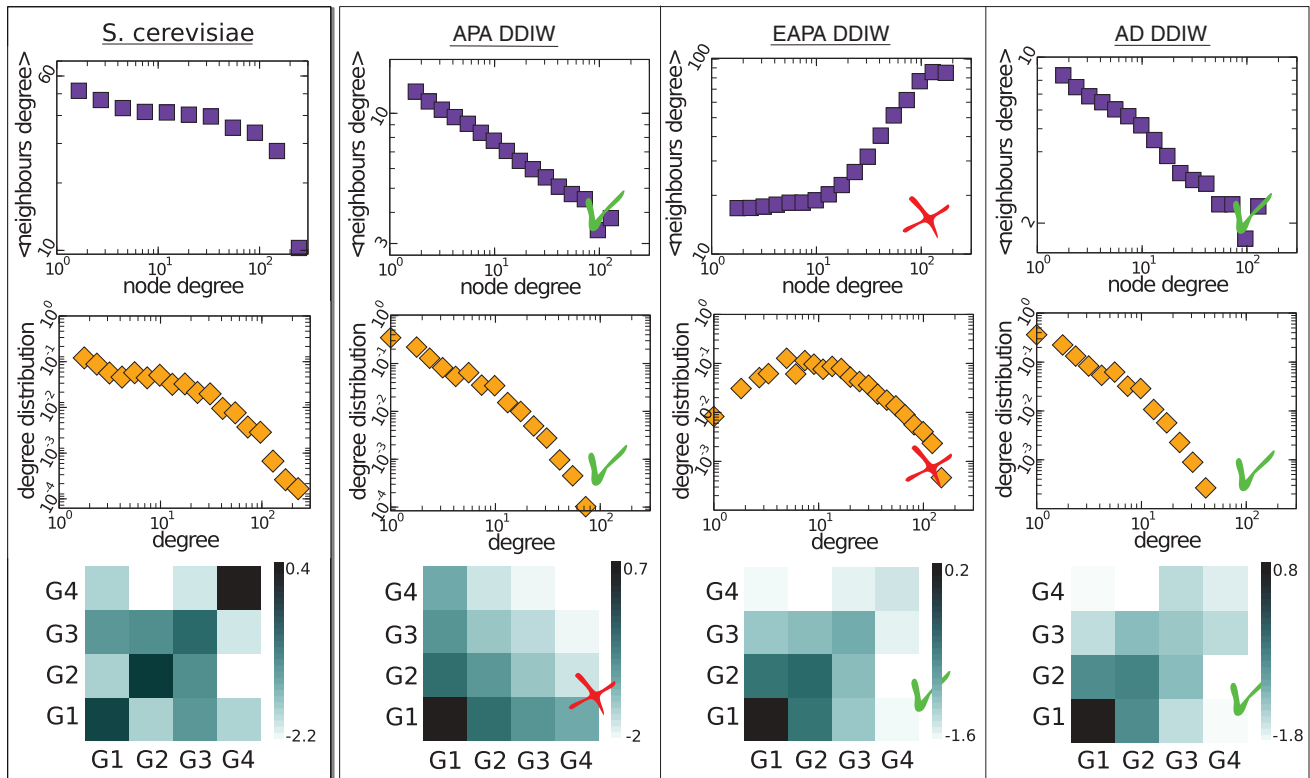
FIG. 6. (Color online) Qualitative comparison between model variants and empirical data. The average degree of nearest neighbors (top row), the degree distribution (middle row), and the interaction density [Eq. (5)] between age groups (bottom row) are measured for *S. cerevisiae* (left panel) and for simulations of the three variants of the DDIW model (right panel). The APA DDIW variant successfully reproduces the empirical degree correlation and degree distribution, but the wiring mechanism does not provide enough links to reproduce the empirical age correlation; the EAPA DDIW variant correctly shows correlation between protein ages, due to the increased number of links introduced by innovation, but it strongly distorts the topological features of the network; the AD DDIW variant effectively reproduces both topological and age-correlation features observed in the empirical network.

generates a network where new nodes are preferentially connected to old nodes, contrary to the pattern that emerges in yeast and equivalently to a pure duplication-divergence network growth. However, the antipreferential mechanism is capable of generating a qualitatively correct age correlation if it can build a large number of links, i.e., if the EAPA DDIW variant is considered. In this case, due to the progressive increasing in the number of links attached in the innovation move, one obtains the correct empirical age dependence ($\Delta D > 0$), but at the expense of completely disrupting the topology. For this variant, a scatterplot of the degree-degree correlation (not shown here) presents a slight bimodality in a small range of degrees; nonetheless, we chose to group the data in histograms in order to highlight how the overall behavior is different from the empirical one. Finally, the age-dependent DDIW variant is able to account for both the topological features and the age correlation.

As mentioned above, we have also tested the robustness of the results under further modifications of the innovation move. No relevant change in the results for the APA DDIW variant is detectable by applying a preferential-attachment principle instead of an antipreferential one nor by attaching the new node to a fixed number (greater than 1) of existing nodes. Moreover, the EAPA DDIW variant yields very similar results for all values of γ in (0,1] and therefore the actual value of this

parameter should not be regarded as an essential quantity. As far as the age-preference is concerned, we remark that an antipreferential wiring move gives the clearest results, but age-age correlations can be seen also in networks obtained by means of preferential-attachment wiring, as long as this does not dominate over the duplication-divergence move.

V. DISCUSSION AND CONCLUSION

The model presented here can be seen as the prototype of a rather general modeling framework where a graph grows by the addition of nodes and links within the constraint of a class structure. Indeed, new nodes are added to a new class or to an existing one with prescribed probabilities, their wiring rules being different in the two cases. Here we explored variants where nodes added together with a new class are wired to the old network according to an antipreferential-attachment principle, while nodes introduced into an existing class follow a duplication-divergence prescription. The goals of our work were twofold. First, we studied the joint evolution of the network by duplication and divergence and by class expansion and innovation. Second, as a case study and proof-of-principle application, we applied the unified framework to the study of age dependence, where some interesting questions are open. The two objectives are connected, as the scenarios we

explored would be ill defined outside this unified framework. For example, assigning antipreferential attachment to the innovation move requires one to be able to distinguish it from a duplication move, i.e., to separate new families from existing ones. To carry out both objectives, we stayed as close as possible to empirical data.

We considered probabilities of the addition of new nodes that vanish with $N \to \infty$ in order to reproduce the observed empirical scaling of homology classes [7]. As a consequence, unless it is imposed that new nodes (i.e., new nodes belonging to a new homology class) carry an extensive number of links, the wiring rule for innovation is of secondary importance with respect to the duplication-divergence move in determining the asymptotic features of the resulting graph ensemble. This is in accordance with the empirical observations indicating that duplication divergence is relevant in shaping the appearance of the PPI network [11,12,15,24]. The finite-size behavior, nonetheless, is sensitive to the innovation process, suggesting the existence of nontrivial features of the topology related to the dynamics of homology classes.

Following these indications, the framework considered here can in principle make more detailed predictions for observables that involve network and homology classes jointly. We analyzed the behavior of one such observable, namely, the correlation between the total number of links originating from a given class and the size of the class. While we find good agreement between data for the *E. coli* PPI network and simulations of the DDIW model (at least for the two nonextensive variants), they both agree with the null expectation that this scaling is linear (see Fig. 7). Indeed, in the random case, i.e., when the members of homology classes (of prescribed sizes) are chosen randomly among network nodes, the total degree
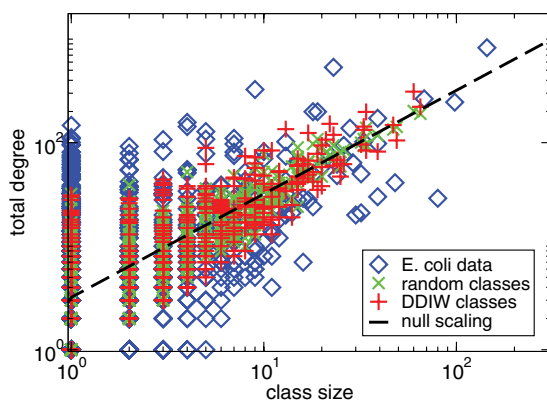

FIG. 7. (Color online) Linear scaling of the correlation between the total number of links originating from a given class and the size of the class: scatterplot of class degree (sum of the degrees of all nodes in a class) versus class size (number of nodes). Red pluses (+) represent results from the typical realization of the DDIW model with $N = 2640$ nodes as in the PPI network of *E. coli* (for the APA DDIW variant; AD DDIW yields a similar plot); green crosses (×) are obtained from the same DDIW realization by randomly permuting nodes between classes; blue diamonds are obtained by combining data for the network structure and homology classes for *E. coli*; the dashed line is the prediction for the average of the total class degree in the randomized case, i.e., the mean node degree times the class size (here $\langle k \rangle = 2.3$).

of a class will be, on average, equal to the number of nodes in the class times the mean node degree. Thus we were unable to find such an effect in the data available to us.

Despite the relation between class size and total degree not being discriminating, the DDIW model does generate nontrivial correlations from the joint evolution of the network and the partitioning into classes. The fact that currently available empirical data do not allow one to discriminate should not discourage the analysis of joint models until more abundant or precise data will be available. To give an example, let us focus on the number $F_N(1,1)$ of classes containing a single node with degree 1 in a network of size N. In the null model where nodes are shuffled randomly between classes (in a single realization of the network), this number is distributed following a hypergeometric distribution centered in $\langle F_N(1,1) \rangle = MC/N$, where M is the number of degree-one nodes in the whole network and C is the number of size-one classes. Simulations of the DDIW model for several realizations (in the two nonextensive variants) consistently yield values of $F_N(1,1)$ that lie several standard deviations above the mean of the null-model distribution. We have measured $M, C, N,$ and j for the *E. coli* PPI network, using both Ensembl [40] and SUPERFAMILY [36] homology data; the actual value of $F_N(1,1)$ is larger than the null average, in both data sets, by approximately 4–6 standard deviations, thus confirming the qualitative non-null prediction of the model. Future work could be directed towards a more detailed study of joint laws such as this one. As an example, the full numbers $F_N(i,k)$ of classes containing i nodes of total degree k are a class of interesting observables that are probably accessible by standard mean-field techniques.

The model variants can be approached by analytical estimates and direct simulation and matched with empirical data on both homology classes and PPI networks. This fitting procedure constitutes a proof of principle of the general applicability of the framework defined here. It also allows one to fix the few parameters of the model and produces well-defined comparisons of the model's predictions with data.

In order to explicitly carry out this comparison in a specific case study we considered the problem of reproducing the empirical age dependence of PPI network interactions through different variants of the model. We tested the predictions obtained against data from yeast, where both PPI network and gene duplications are well characterized and the duplication age of individual proteins is also available. We were able to show that the empirical duplication-age patterns of interacting protein pairs can be reproduced in two alternative ways: first, by an antipreferential-attachment prescription in the innovation move, associated with a heavy (extensive) contribution of this move to the number of links, and second, by inserting a strong negative bias towards forming protein-protein interactions with old nodes. However, the first choice leads to networks whose degree distribution and neighbor degree correlations do not resemble the empirical ones. Conversely, the bias imposed in the second case could be rationalized by biological arguments concerning the available binding interfaces (older proteins are more likely fully engaged with the interactions they participate in) and the conservation of basic biological functions (new interactions interfering with older ones could be detrimental). Thus an age-dependent

duplication-divergence move seems more satisfactory. Once established that such an age dependence in the divergence process is in qualitative agreement with data, one can ask whether the same features can be reproduced without considering the full partition and topology dynamics. We have performed additional numerical simulations and found that the qualitative patterns in Fig. 6 can be reproduced also by a simple duplication-divergence model with age bias and no innovation nor class dynamics. This is not in contrast with the importance of considering the problem in the more general framework since, in principle, as we have explained in the preceding section, other mechanisms, related to the innovation or wiring move, could have been responsible for the age-correlation patterns observed.

Overall, our analysis tends to support the hypothesis that duplication divergence alone does not account for the observed history dependence of the existing protein-protein interactions [26]. Note, however, that in the age-dependent DDIW model, as well as in the previous models of this kind, duplication divergence turns out to be a necessary ingredient in shaping biologically resembling degree distributions and degree correlations of nearest neighbors. This suggests that the mechanism of duplication and divergence might play a role in determining PPI network topologies [25]. Conversely, in the previous model of Kim and Marcotte, the age dependence is associated with model moves that, roughly speaking, are more similar to an antipreferential-attachment innovation move than to a duplication-divergence one [26]. We should also remark that the models we have explored here are based on totally asymmetric duplication divergence. We cannot exclude that the age-correlation patterns could be biased also by using general duplication-divergence schemes [21], where different values of σ are assigned to the connections between pairs of new nodes with respect to new-old node pairs. In this case, the introduction of an additional parameter could produce the age-correlation kernel in a natural way.

One important caveat is that the PPI data available to us are affected by strong subsampling problems since presumably for most organisms only a fraction of the protein-protein interactions are available in the DIP database [35]. Having small samples of large networks makes it problematic to estimate model parameters. For example, it is likely that the exponent for $L(N)$ is overestimated. We performed a numerical test by growing networks up to size N (and a fluctuating number of links L') and subsampling them to a fixed number of links L. In general one obtains networks with many more nodes N' compared to networks that are grown with the model at L' edges and not subsampled. For parameter values that match the available data, this error could be as large as 100%; in *C. elegans*, for instance, for which approximately 4000 interactions are known involving around 2600 proteins (out of approximately 20 000 genes), we obtain $N' \approx 5100$. On the positive side, restricting the parameter-matching analysis (Fig. 5) of the model to the few highly sampled genomes does not change our results. Nevertheless, it seems quite possible that a larger cross-genomic knowledge of PPI networks could change the quantitative picture emerging from these data and possibly also the qualitative one.

To conclude, despite the current open questions, we believe that this general framework might be important in posing questions about the growth of PPI networks as the network structure is intimately related to the partitioning in homology classes and, quite importantly, to the class of biological functions that a specific homology class can perform [41].

APPENDIX: MEAN-FIELD CALCULATION OF $L(N)$

We give here the solutions to the mean-field equation (13). Let us call $L_A(N)$ the solution with the choice $l(N) = 1$ (APA DDIW variant) and $L_B(N)$ the solution with the choice $l(N) = \gamma 2L/N$ (EAPA DDIW variant). For both choices Eq. (13) is a standard first-order ordinary differential equation, whose solution can be readily computed with the help of MATHEMATICA. One obtains

$$
L_A(N) = N^{2\sigma} e^{2\sigma P_\alpha(N)} \Bigg\{ \text{const} + \frac{1}{2} N^{1-2\sigma} e^{-2\sigma P_\alpha(N)}
$$
$$
- \frac{1}{2(1-\alpha)} N^{1-2\sigma} [2\sigma P_\alpha(N)]^{(1-2\sigma)(1-\alpha)}
$$
$$
\times \Gamma\left(-\frac{1-2\sigma}{1-\alpha}, 2\sigma P_\alpha(N) \right) \Bigg\}, \tag{A1}
$$

where $P_\alpha(N)$ is defined as

$$
P_\alpha(N) = \frac{\alpha}{1-\alpha} N^{\alpha-1} \tag{A2}
$$

[which is proportional to the asymptotic form of the innovation probability; see Eq. (4)] and $\Gamma(a,z)$ is the upper incomplete Gamma function

$$
\Gamma(a,z) = \int_z^\infty t^{a-1} e^t \, dt. \tag{A3}
$$

The constant term depends only on α, σ, and the initial condition $L(N_0) = L_0$. Note that $P_\alpha(N) \to 0$ when $N \to \infty$ since $\alpha \in (0,1)$. By substituting the asymptotic expansion for the incomplete Gamma function to leading order around $z = 0$,

$$
\Gamma(a,z) \sim \Gamma(a) - \frac{z^a}{a}, \tag{A4}
$$

into Eq. (A1) one sees that the first term in large curly brackets gives a contribution proportional to $N^{2\sigma}$ to the asymptotic form, while the second and third terms have a linear behavior proportional to N, thus recovering expression (14).

An expression similar to but more complicated than Eq. (A1) is found for $L_B(N)$. We do not quote it here because it is very large without being particularly instructive; the same analysis gives the corresponding asymptotic behavior (15).

[1] P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee, and E. M. Marcotte, Curr. Opin. Struct. Biol. **14**, 292 (2004).

[2] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[3] M. G. Sun and P. M. Kim, Genome Biol. **12**, 235 (2011).

[4] E. V. Koonin, Y. I. Wolf, and G. P. Karev, Nature (London) **420**, 218 (2002).

[5] E. Koonin, PLoS Comput. Biol. **7**, e1002173 (2011).

[6] J. Qian, N. M. Luscombe, and M. Gerstein, J. Mol. Biol. **313**, 673 (2001).

[7] M. Cosentino Lagomarsino, A. Sellerio, P. Heijning, and B. Bassetti, Genome Biol. **10**, R12 (2009).

[8] A. Angelini, A. Amato, G. Bianconi, B. Bassetti, and M. Cosentino Lagomarsino, Phys. Rev. E **81**, 021919 (2010).

[9] J. Zhang, Trends Ecol. Evol. **18**, 292 (2003).

[10] M. Heo, S. Maslov, and E. Shakhnovich, Proc. Natl. Acad. Sci. USA **108**, 4258 (2011).

[11] G.-Z. Wang and M. J. Lercher, PLoS One **6**, e18288 (2011).

[12] S. A. Teichmann, J. Mol. Biol. **324**, 399 (2002).

[13] J. B. Pereira-Leal, E. D. Levy, C. Kamp, and S. A. Teichmann, Genome Biol. **8**, R51 (2007).

[14] A. J. Venkatakrishnan, E. D. Levy, and S. A. Teichmann, Biochem. Soc. Trans. **38**, 879 (2010).

[15] J. B. Pereira-Leal and S. A. Teichmann, Genome Res. **15**, 552 (2005).

[16] R. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler, Adv. Complex Syst. **5**, 43 (2002).

[17] R. Pastor-Satorras, E. Smith, and R. V. Solé, J. Theor. Biol. **222**, 199 (2003).

[18] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, Complexus **1**, 38 (2003).

[19] I. Ispolatov, P. L. Krapivsky, and A. Yuryev, Phys. Rev. E **71**, 061911 (2005).

[20] R. R. Stein and H. Isambert, Phys. Rev. E **84**, 051904 (2011).

[21] K. Evlampiev and H. Isambert, Proc. Natl. Acad. Sci. USA **105**, 9863 (2008).

[22] K. Evlampiev and H. Isambert, BMC Syst. Biol. **1**, 49 (2007).

[23] I. Ispolatov, P. Krapivsky, I. Mazo, and A. Yuryev, New J. Phys. **7**, 145 (2005).

[24] A. Presser, M. B. Elowitz, M. Kellis, and R. Kishony, Proc. Natl. Acad. Sci. USA **105**, 950 (2008).

[25] S. Navlakha and C. Kingsford, PLoS Comput. Biol. **7**, e1001119 (2011).

[26] W. K. Kim and E. M. Marcotte, PLoS Comput. Biol. **4**, e1000232 (2008).

[27] M. G. F. Sun, M. Sikora, M. Costanzo, C. Boone, and P. M. Kim, PLoS Comput. Biol. **8**, e1002411 (2012).

[28] D. Talavera, S. G. Williams, M. G. S. Norris, D. L. Robertson, and S. C. Lovell, J. Mol. Biol. **419**, 387 (2012).

[29] J. Berg, M. Lassig, and A. Wagner, BMC Evol. Biol. **4**, 51 (2004).

[30] D. L. Robertson and S. C. Lovell, Biochem. Soc. Trans. **37**, 768 (2009).

[31] B. Bassetti, M. Zarei, M. Cosentino Lagomarsino, and G. Bianconi, Phys. Rev. E **80**, 066118 (2009).

[32] R. Durrett and J. Schweinsberg, Ann. Probab. **33**, 2094 (2005).

[33] J. Pitman, *Combinatorial Stochastic Processes*, Lecture Notes in Mathematics, Vol. 1875 (Springer, Berlin, 2006).

[34] T. J. Treangen and E. P. C. Rocha, PLoS Genet. **7**, e1001284 (2011).

[35] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, Nucleic Acids Res. **28**, 289 (2000).

[36] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, and J. Gough, Nucleic Acids Res. **37**, D380 (2009).

[37] L. Grassi, D. Fusco, A. Sellerio, D. Cora, B. Bassetti, M. Caselle, and M. Cosentino Lagomarsino, Mol. BioSyst. **6**, 2305 (2010).

[38] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev, Bioinformatics **23**, i549 (2007).

[39] Z. Liu, Q. Liu, H. Sun, L. Hou, H. Guo, Y. Zhu, D. Li, and F. He, BMC Evol. Biol. **11**, 133 (2011).

[40] P. Flicek *et al.*, Nucleic Acids Res. **40**, 84 (2012).

[41] J. Grilli, B. Bassetti, S. Maslov, and M. Cosentino Lagomarsino, Nucleic Acids Res. **40**, 530 (2012).