Conservation and topology of protein interaction networks under duplication-divergence evolution

Kirill Evlampiev and Hervé Isambert*

Physico-chimie Curie, Centre National de la Recherche Scientifique Unité Mixte de Recherche 168, Institut Curie, Section de Recherche, 11 rue P. & M. Curie, 75005 Paris, France

Communicated by M. Gromov, Institut des Hautes Études Scientifiques, Bures-sur-Yvette, France, April 30, 2008 (received for review March 22, 2007)

Genomic duplication-divergence processes are the primary source of new protein functions and thereby contribute to the evolutionary expansion of functional molecular networks. Yet, it is still unclear to what extent such duplication-divergence processes also restrict by construction the emerging properties of molecular networks, regardless of any specific cellular functions. We address this question, here, focusing on the evolution of protein–protein interaction (PPI) networks. We solve a general duplication-divergence model, based on the statistically necessary deletions of protein–protein interactions arising from stochastic duplications at various genomic scales, from single-gene to whole-genome duplications. Major evolutionary scenarios are shown to depend on two global parameters only: (*i*) a protein conservation index (*M*), which controls the evolutionary history of PPI networks, and (*ii*) a distinct topology index (*M′*) controlling their resulting structure. We then demonstrate that conserved, nondense networks, which are of prime biological relevance, are also necessarily scale-free by construction, irrespective of any evolutionary variations or fluctuations of the model parameters. It is shown to result from a fundamental linkage between individual protein conservation and network topology under general duplication-divergence evolution. By contrast, we find that conservation of network motifs with two or more proteins cannot be indefinitely preserved under general duplication-divergence evolution (independently from any network rewiring dynamics), in broad agreement with empirical evidence between phylogenetically distant species. All in all, these evolutionary constraints, inherent to duplication-divergence processes, appear to have largely controlled the overall topology and scale-dependent conservation of PPI networks, regardless of any specific biological function.

evolutionary constraint | scale-free graph | functional motif | orthology | statistical model

The primary source of new protein functions is generally considered to originate from *duplication* of existing genes followed by functional *divergence* of their duplicate copies (1–3). In fact, duplication-divergence events have occurred and continue to occur at a wide range of genomic scales, from many independent duplications of individual genes[†] [10^{-3} fixed events per gene per million years (MY) (4)] to rare but evolutionary dramatic duplications of entire genomes [one fixed event per 100–200 MY (5)]. For instance, there have been between two and four *consecutive* whole-genome duplications in all major eukaryote kingdoms in the past 300–500 MY (5). This actually amounts to a more-or-less similar contribution of new genes from whole-genome duplication as from individual gene duplications [i.e., one fixed event per 100–200 MY $\simeq 10^{-3}$ fixed events per gene per MY, assuming a 10% fixation rate after a whole-genome duplication with ≈10,000 genes (5)].

This succession of whole-genome duplications, together with the accumulation of individual gene duplications, must have greatly contributed to shaping the global structure of large biological networks, such as protein–protein interaction (PPI) networks, that control cellular activities. In fact, concordant empirical evidence reveals the evolutionary persistence of duplication-derived protein–protein interactions. For instance, there are clear enrichments of recent protein duplicates around common protein partners compared with randomly picked pairs of proteins (5, 6), although the fraction of proteins identified as having undergone a (recent) duplication (<200 MY) remains typically small in absolute terms, for example, 10% (4). Similarly, protein residues implicated in protein–protein interaction are generally the most conserved at the surface of proteins (7), revealing their duplication-derived origin,[‡] with typically little more than one conserved binding interface per protein-binding domains.[§]

Ispolatov *et al.* (10) proposed an interesting local duplication-divergence model of PPI network evolution based on (*i*) the *statistical* deletion of individual, duplication-derived interactions and (*ii*) a *time-linear* increase in genome and PPI network sizes. Clearly, the deletion of redundant interactions arising from duplication is necessary to avoid the emergence of biologically irrelevant, densely connected PPI networks, lacking low-degree connectivities. Yet, we expect that *independent* local duplications and, *a forteriori*, partial- or whole-genome duplications all lead to *exponential*, not time-linear, evolutionary dynamics of PPI networks. In the long time limit, exponential dynamics should outweigh all time-linear processes that have been assumed in earlier PPI network evolution models (10–15). Models based on time-linear processes also assume that local evolutionary dynamics remain essentially frozen, as long as they are not directly affected by a local modification of the network. Yet, in reality, sequence mutations and environmental changes continue to affect the evolution of whole PPI networks, not just in the immediate surroundings of recently duplicated proteins.

In this article, we propose and asymptotically solve a general duplication-divergence model based on prevailing exponential dynamics[¶] of PPI network evolution under local, partial, or global genome duplications. The only interaction changes that are considered are *deletions* of duplication-derived interactions. In particular, the rewiring dynamics of PPI networks by *de novo* creation of protein-binding interfaces (4) is neglected (10), as suggested by the empirical evidence mentioned earlier (see also

Author contributions: H.I. designed research; K.E. and H.I. performed research; K.E. and H.I. analyzed data; and K.E. and H.I. wrote the paper.

The authors declare no conflict of interest.

*To whom correspondence should be addressed. E-mail: herve.isambert@curie.fr.

[†]Duplicated protein domains or subdomains are also quite common even within ancestral proteins, such as the ubiquitous aquaporin membrane proteins in eubacteria, archaea, and eukaryotes, or the TATA-binding protein from archaea and eukaryotes.

[‡]Except for a few interesting cases of protein-binding mimicry, typically found in virus–host protein–protein interactions (8).

[§]Except for domains that self-assemble into homo-oligomers, which *must* have at least two binding interfaces, see table 2 in ref. 9.

[¶]Results from the time-linear duplication-divergence model (10) are recovered as a special limit, see supporting information (SI) *Appendix*.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804119105/DCSupplemental.

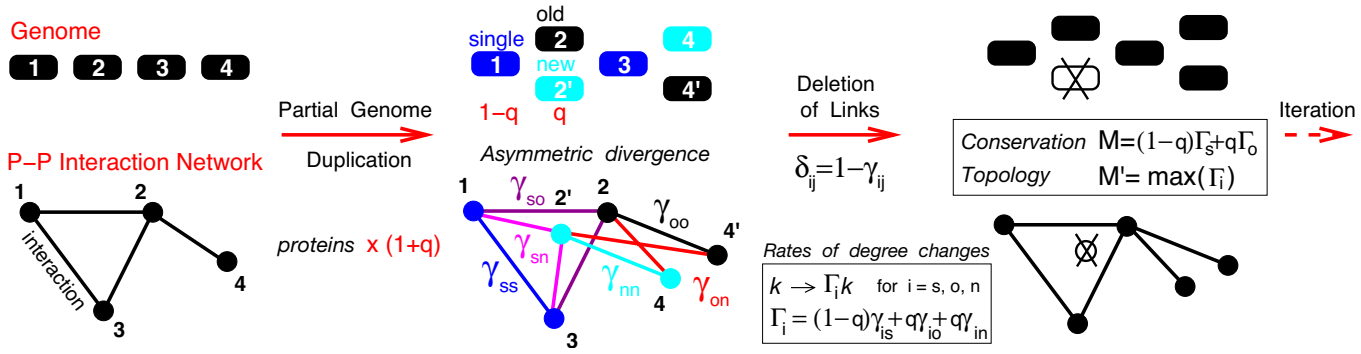

Fig. 1. General duplication-divergence model for protein–protein interaction network evolution. Successive duplications of a fraction q of genes are followed by an asymmetric divergence of gene duplicates (e.g., 2 vs. 2′). New duplicates (n) are left essentially free to accumulate neutral mutations with the likely outcome of becoming nonfunctional and eventually deleted unless some new, *duplication-derived* interactions are selected; old duplicates (o), however, are more constrained to conserve old interactions already present before duplication. Interactions on the locally ($q \ll 1$), partially ($q < 1$) or fully ($q = 1$) duplicated network are then preserved stochastically with different probabilities γ_{ij} ($0 \leq \gamma_{ij} \leq 1$, $i, j = s, o, n$) reflecting the recent history of each interacting partners, that are either singular, nonduplicated genes (s) or recently duplicated genes undergoing asymmetric divergence (o/n). Two effective parameters, M and M', that depend on the rates of connectivity change, Γ_i, and underlying parameters q and γ_{ij}, control the evolutionary history or conservation (M) and resulting structure or topology (M') of PPI networks (see text).

Evolution of PPI network motifs). Indeed, our aim here is to establish a theoretical baseline from which other evolutionary processes beyond strict gene duplication and interaction loss events, such as shuffling of protein domains (5) or horizontal gene transfers, can then be considered.

A visual overview of the model is shown in Fig. 1 including its two main effective parameters, M and M', that control, respectively, the evolutionary history or *conservation* (M) and resulting structure or *topology* (M') of PPI networks under duplication-divergence evolution. In this article, we demonstrate a fundamental relation between protein conservation (M) and network topology (M'), that is, $M \leq M'$, that is strictly independent from *any* evolutionary variations or fluctuations of the model parameters. We then discuss simple consequences in terms of evolutionary *linkage* between individual protein conservation and PPI network topology. The approach is also extended to outline the evolutionary statistics of small-network motifs including two or more proteins. In particular, we show that network motifs, unlike individual proteins, cannot be indefinitely conserved under general duplication-divergence evolution, regardless of any network-rewiring dynamics. Throughout the article, theoretical assumptions and results are commented on with brief discussions highlighting their biological relevance.

Results

General Duplication-Divergence Model. The general duplication-divergence (GDD) model is designed to capture PPI network properties caused by evolutionary constraints, inherent to duplication-divergence processes and independent of selective adaptation (3) or any specific biological function. Concretely, the GDD model analyzes the deletion statistics of protein–protein interactions that arise from stochastic duplications at various genomic scales, from single-gene to whole-genome duplications. This deletion statistics of duplication-derived interactions is indeed a necessary "background" dynamics of PPI network evolution to prevent the emergence of biologically irrelevant, densely connected PPI networks, lacking low-degree connectivities.

In practice, a fraction q of extant genes is randomly *duplicated* at each time step of the GDD model. The divergence of both duplicated and nonduplicated genes then leads to the stochastic deletion or conservation of their related interactions, before another round of duplication-divergence occurs (Fig. 1). In the following, we first solve the GDD model assuming that q is constant over evolutionary time scales. We then study more realistic scenarios combining, for instance, rare whole-genome duplications ($q = 1$) with more frequent local duplications of individual genes ($q \ll 1$), and including also stochastic fluctuations in *all* microscopic parameters of the GDD model (see Fig. 1 and below). To analyze the deletion statistics of duplication-derived interactions, we assume that ancient and recently duplicated interactions are stochastically conserved with distinct probabilities γ_{ij}'s, depending only on the recently duplicated or nonduplicated state of each protein partners, as well as on the *asymmetric* divergence between "old" and "new" (or more "conserved" and more "divergent") gene duplicates (5), see the Fig. 1 legend ("s" for "singular," nonduplicated genes and "o"/"n" for old/new asymmetrically divergent duplicates). Here, we consider nonoriented PPI networks, that is, $\gamma_{ij} = \gamma_{ji}$, for $i, j = s, o, n$.

The first effective parameters derived from these microscopic evolutionary parameters are the average rates of connectivity change Γ_i (i.e., $k \to k\Gamma_i$) for each type of node $i = s, o, n$, where $\Gamma_i = (1 - q)\gamma_{is} + q(\gamma_{io} + \gamma_{in})$ is independent from node connectivity k. In the following, we assume $\Gamma_o \geq \Gamma_n$ by definition of old and new duplicates caused by asymmetric divergence. Note that self-interacting proteins, corresponding to self-link loops, are not taken into account, for simplicity, in the main text, because they can be shown to have little effect on the asymptotic evolutionary regimes of the connectivity distribution (see *SI Appendix*, Fig. S3 and *SI Text*, for details).

We study the GDD evolutionary dynamics of PPI networks in terms of ensemble averages $\langle Q^{(n)} \rangle$ defined as the mean value of a feature Q over all realizations of the evolutionary dynamics after n successive duplications. This does not imply, of course, that all network realizations "coexist," but only that a random selection of them is reasonably well characterized by the theoretical ensemble average. Although it is generally not the case for exponentially growing systems, here, we can show that ensemble averages over all evolutionary dynamics indeed reflect the properties of typical network realizations for biologically relevant regimes (see *Statistical Properties of GDD Models* in *SI Appendix*).

In the following, we focus on the number of proteins (or "nodes") N_k of connectivity k in PPI networks, while postponing the analysis of GDD models for simple network motifs to the end of the article and the *SI Appendix*. The total number of nodes in the network is noted $N = \Sigma_{k \geq 0} N_k$ and the total number of interactions (or "links") $L = \Sigma_{k \geq 0} kN_k/2$. The dynamics of the ensemble averages $\langle N_k^{(n)} \rangle$ after n duplications is analyzed by using a generating function,

$$F^{(n)}(x) = \sum_{k\geq 0} \langle N_k^{(n)} \rangle x^k. \quad [1]$$

The evolutionary dynamics of $F^{(n)}(x)$ corresponds to the following recurrence deduced from the microscopic definition of the GDD model (see *SI Appendix*),

$$F^{(n+1)}(x) = (1-q)F^{(n)}(A_{s(x)}) + qF^{(n)}(A_{o(x)}) + qF^{(n)}(A_{n(x)}) \quad [2]$$

where we note for $i = s, o, n$,

$$A_i(x) = (1-q)(\gamma_{is}x + \delta_{is}) + q(\gamma_{io}x + \delta_{io})(\gamma_{in}x + \delta_{in}) \quad [3]$$

where $\delta_{ij} = 1 - \gamma_{ij}$ are deletion probabilities ($i, j = s, o, n$) and $A_i'(1) = (1-q)\gamma_{is} + q(\gamma_{io} + \gamma_{in}) = \Gamma_i$, average rates of connectivity change for each type of nodes i = s, o, n (Fig. 1).

Network Expansion (Γ) and Protein Conservation (*M*). The total number of nodes generated by the GDD model, $F^{(n)}(1)$, growths exponentially with the number of partial duplications, $F^{(n)}(1) = C\cdot(1+q)^n$, where C is the initial number of nodes, as a constant fraction of nodes q is duplicated at each time step. Yet, some nodes become completely disconnected from the rest of the graph during divergence and rejoin the disconnected component of size $F^{(n)}(0)$. From a biological point of view, these disconnected nodes represent genes that have presumably lost all biological functions and become pseudogenes before being simply eliminated from the genome. We neglect the possibility for nonfunctional genes to reconvert to functional genes again after suitable mutations, and remove them at each round of partial duplication[‖] focusing solely on the connected part of the graph.

In particular, the link growth rate $\Gamma = (1-q)\Gamma_s + q\Gamma_o + q\Gamma_n$ obtained by taking the first derivative of Eq. **2** at $x = 1$, controls whether the connected part of the graph is exponentially growing ($\Gamma > 1$) or shrinking ($\Gamma < 1$).

Let us now introduce another rate of *prime* biological interest, $M = (1-q)\Gamma_s + q\Gamma_o$. It is the *average rate of connectivity increase (M >1) of decrease (M < 1) for the most conserved duplicate lineage*, which corresponds to a stochastic alternance between singular (s) and most conserved (o) duplicate descents. In particular, we have by construction, $M < \Gamma = M + q\Gamma_n$, independently from any evolutionary parameters, q and $\gamma_{ij} > 0$. This implies three main evolutionary regimes from the perspective of network expansion (Γ) and protein conservation (M) (Fig. 2):

- If $M < \Gamma < 1$. PPI networks are vanishing in this regime with seemingly little biological relevance.
- If $M < 1 < \Gamma$. PPI networks are expanding, in this case, but their proteins are *not* conserved over long evolutionary time scales. This implies that the networks forget their evolutionary history exponentially fast, as most nodes eventually disappear and, with them, all traces of network evolution. These networks are *not* preserved over time, but instead are continuously renewed from duplication of the (few) most connected nodes (Fig. 2). Individual proteins of a given network realization are thus more similar to one another than to any protein of other network realizations, which can be seen, from a speciation perspective, as PPI networks of phylogenetically distant organisms. This is in sharp contrast to the widespread structural orthology observed across all extant life forms, even though functions of orthologs often differ (see *Evolution of PPI Network Motifs*).
- If $1 < M < \Gamma$. By contrast, PPI networks remember their past evolution from the very beginning, in this case, as proteins statistically keep on increasing their connectivity once they have emerged from a duplication-divergence event. This implies that most proteins are conserved *throughout* the evolution process and preserve some interaction partners. This is indeed in broad agreement with empirical evidence, because traces of protein conservation are even observed within the core transcriptional and translational machineries across all three major living kingdoms (16).

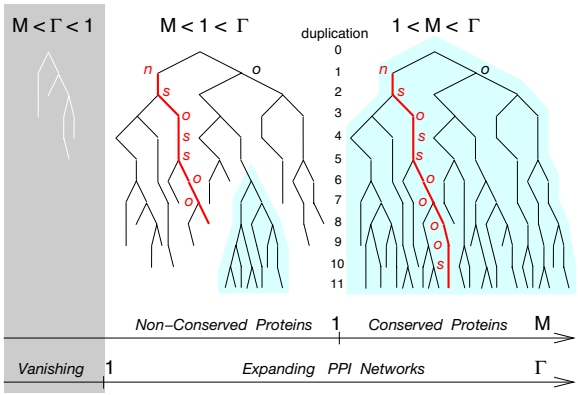

Fig. 2. Evolutionary growth (Γ) and protein conservation (M) of PPI networks. The constitutive constraint, $M < \Gamma$, defines three evolutionary regimes discussed in the text.

Evolution of PPI Network Degree Distribution. We now turn to the evolution of the degree distribution and other topological properties of PPI networks, which correspond to the technical core of the GDD model. To this end, we rescale the exponentially growing connected graph by introducing a normalized generating function for the average degree distribution,

$$p^{(n)}(x) = \sum_{k\geq 1} p_k^{(n)} x^k \quad \text{with} \quad p_k^{(n)} = \frac{\langle N_k^{(n)} \rangle}{\langle N^{(n)} \rangle}, \quad [4]$$

where $\langle N^{(n)} \rangle = \Sigma_{k\geq 1} \langle N_k^{(n)} \rangle$, that is, after removing $\langle N_0^{(n)} \rangle$. $F^{(n)}(x)$ can be reconstructed from the shifted degree distribution, $\tilde{p}^{(n)}(x) = p^{(n)}(x) - 1$, as

$$F^{(n)}(x) = \langle N^{(n)} \rangle \tilde{p}^{(n)}(x) + C\cdot(1+q)^n, \quad [5]$$

which yields the following recurrence for $\tilde{p}^{(n)}(x)$,

$$\tilde{p}^{(n+1)}(x) = \frac{(1-q)\tilde{p}^{(n)}(A_{s(x)}) + q\tilde{p}^{(n)}(A_{o(x)}) + q\tilde{p}^{(n)}(A_{n(x)})}{\Delta^{(n)}} \quad [6]$$

where $\Delta^{(n)}$ is the ratio between two consecutive graph sizes in terms of connected nodes, that is, $\Delta^{(n)} = \langle N^{(n+1)} \rangle / \langle N^{(n)} \rangle$,

$$\Delta^{(n)} = -(1-q)\tilde{p}^{(n)}(A_{s(0)}) - q\tilde{p}^{(n)}(A_{o(0)}) - q\tilde{p}^{(n)}(A_{n(0)}) > 0 \quad [7]$$

Although $\Delta^{(n)}$ is not known *a priori* and should, in general, be determined self-consistently with $\tilde{p}^{(n)}(x)$ itself, it is directly related to the evolution of the mean degree $\bar{k}^{(n)} = \Sigma_{k\geq 1} k p_k^{(n)}$ obtained by taking the first derivative of Eq. **6** at $x = 1$,

$$\frac{\bar{k}^{(n+1)}}{\bar{k}^{(n)}} = \frac{(1-q)\Gamma_s + q\Gamma_o + q\Gamma_n}{\Delta^{(n)}} = \frac{\Gamma}{\Delta^{(n)}}. \quad [8]$$

[‖]Note, however, that pseudogenes may still have a critical role in evolution by providing functional domains that can be fused to adjacent genes. This supports a view of PPI network evolution in terms of protein domains instead of entire proteins (*SI Appendix*, Fig. S6*B*, and ref. 5). Yet, we showed in ref. 5 that extensive domain shuffling does not change the resulting network topology from duplication-divergence models.

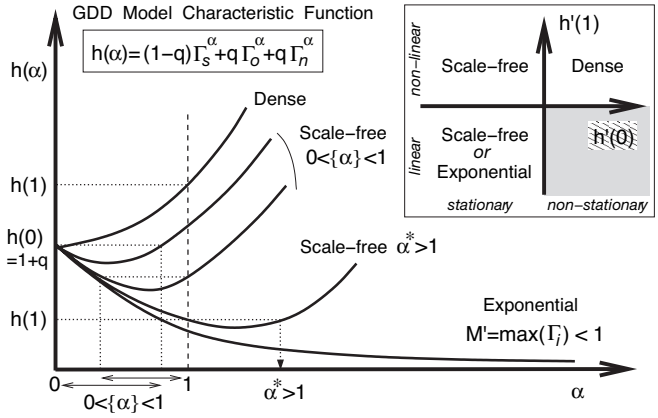

Fig. 3. Asymptotic degree distribution for GDD models. Asymptotic regimes are deduced from the *convex* characteristic function $h(\alpha)$ and its derivatives h′(0) and h′(1) (see text).

Hence, although connected networks grow exponentially both in terms of number of links (link growth rate Γ) and number of connected nodes (node growth rate $\Delta^{(n)}$), features normalized over these growing networks, such as node mean connectivity (Eq. **8**) or distributions of node degree (or simple network motifs, see below), exhibit richer evolutionary dynamics in the asymptotic limit $n \to \infty$, as we will now discuss.

Asymptotic Analysis of Node Degree Distribution (*M′*). The node degree distribution can be shown (see *SI Appendix*) to converge toward a limit function $p(x)$, with $\tilde{p}(x) = p(x) - 1$ solution of the functional Eq. **6**.

$$\tilde{p}(x) = \frac{(1-q)\tilde{p}(A_{s(x)}) + q\tilde{p}(A_{o(x)}) + q\tilde{p}(A_n(x))}{\Delta} \qquad [9]$$

where $\Delta = \lim_{n\to\infty}\Delta^{(n)}$ with both $\Delta \leq 1 + q$, the maximum node growth rate, and $\Delta \leq \Gamma$, the link growth rate, because the number of connected nodes cannot increase faster than the number of links. Asymptotic regimes with $\Delta = \Gamma$ correspond to the same exponential growth of the network in terms of connected nodes and links, and will be referred to as *linear* regimes, hereafter, whereas $\Delta < \Gamma$ corresponds to *nonlinear* asymptotic regimes, which imply a diverging mean connectivity $\bar{k}^{(n)} \to \infty$ in the asymptotic limit $n \to \infty$ (Eq. **8**).

To determine Δ and $p(x)$ self-consistently, we first express successive derivatives of $p(x)$ at $x = 1$ in terms of lower derivatives by using Eq. **9**,

$$\partial_x^k p(1)\left[1 - \frac{(1-q)\Gamma_s^k + q\Gamma_o^k + q\Gamma_n^k}{\Delta}\right] = \sum_{l=[k/2]}^{k} \alpha_{k,l}\partial_x^l p(1), \qquad [10]$$

where $\alpha_{k,l}$ are positive functions of the 1 + 6 parameters. Inspection of this expression readily defines two classes of asymptotic regimes, *regular* and *singular* regimes, depending on the value of a *topology index* $M' = \max_i(\Gamma_i)$, for $i = s, o, n$. The detailed analysis relies on the "characteristic function" $h(\alpha) = (1-q)\Gamma_s^\alpha + q\Gamma_o^\alpha + q\Gamma_n^\alpha$, as outlined below and in Fig. 3 (see *SI Appendix*, *Asymptotic Methods*, for proof details).

Regular regimes, if $M' = \max_i(\Gamma_i) < 1$, for $i = s, o, n$. In this case, the only possible solution is $\Delta = h(1)$ (i.e., linear regime). Hence, since $M' < 1$, $h(1) > h(k)$, and successive derivatives $\partial_x^k p(1)$ are thus finite and positive for all $k \geq 1$. This corresponds to an exponential decrease of the node degree distribution for $k \gg 1$, $p_k \propto e^{-\mu k}$ with a power law prefactor. The limit average connectivity (Eq. **8**) is finite in this case, $\bar{k} < \infty$.

Singular regimes, if $M' = \max_i(\Gamma_i) > 1$, for $i = s, o, n$. In this case, Eq. **10** suggests that there exists an integer $r \geq 1$ for which the rth derivative is negative, $\partial_x^r p(1) < 0$, which is impossible by definition. This simply means that neither this derivative nor any higher ones exist (for $k \geq r$). We thus look for self-consistent solutions of the "characteristic equation" $h(\alpha) = \Delta$ (with $r - 1 < \alpha \leq r$) corresponding to a singularity of $p(x)$ at $x = 1$ and a power law tail of p_k, for $k \gg 1$ (17),

$$p(x) = 1 - \cdots - A_\alpha(1-x)^\alpha + \cdots \text{ and } \quad p_k \propto k^{-\alpha-1} \qquad [11]$$

where the singular term $(1-x)^\alpha$ is replaced by $(1-x)^r \ln(1-x)$ for $\alpha = r$ exactly. Several asymptotic behaviors are predicted from the convex shape of $h(\alpha)$ ($\partial_\alpha^2 h \geq 0$), depending on the signs of its derivatives $h'(0)$ and $h'(1)$ (Fig. 3 *Inset*).

- If $h'(0) < 0$ and $h'(1) < 0$. There exists an $\alpha^\star > 1$ so that $h(\alpha^\star) = h(1)$ and the condition $\Delta \leq h(1)$ implies that $\alpha^\star \geq \alpha \geq 1$. The solution $\alpha = 1$ requires $h'(1) = 0$ and should be rejected in this case. Hence, because $\bar{k} < \infty$ for $\alpha > 1$, we must have $\Delta = h(1)$ (linear regime) and a scale-free limit degree distribution with a *unique* $\alpha = \alpha^\star > 1$, $p_k \propto k^{-\alpha^\star - 1}$ for $k \gg 1$.
- If $h'(0) < 0$ and $h'(1) = 0$. $\alpha = 1$, $\Delta = h(1)$, and $p_k \propto k^{-2}$ for $k \gg 1$ ($\bar{k}^{(n)} \to \infty$ as $n \to \infty$).
- If $h'(0) < 0$ and $h'(1) > 0$. The general condition $\Delta \leq \min(h(0), h(1))$ leads *a priori* to a whole range of possible $\alpha \in]0, 1]$ corresponding to stationary scale-free degree distributions with diverging mean degrees $\bar{k}^{(n)} \to \infty$. Yet, numerical simulations suggest that there might still be a unique asymptotic node growth rate Δ regardless of initial conditions or evolutionary trajectories, although convergence is extremely slow (see *SI Appendix*, *Numerical simulations*).
- If $h'(0) \geq 0$ and $h'(1) > 0$. $\Delta = h(0) = 1 + q$, implying that all duplicated nodes are selected in this case. No suitable α exists as the node degree distribution is exponentially shifted toward higher and higher connectivities. This is a dense, nonstationary regime with seemingly little relevance to biological networks.

Finally, note that the characteristic equation $\Delta = h(\alpha)$ can be recovered directly from the average change of connectivity $k \to k\Gamma_i$ and the following continuous approximation (by using $N^{(n)} = \sum_k N_k^{(n)} \simeq \int_u N_u^{(n)}\, du$ and $\langle N_k^{(n)}\rangle \propto k^{-\alpha-1}$),

$$\frac{\langle N^{(n+1)}\rangle}{\langle N^{(n)}\rangle} \simeq \frac{\int\langle (1-q)N_{k\Gamma_s}^{(n)}\Gamma_s + qN_{k\Gamma_o}^{(n)}\Gamma_o + qN_{k\Gamma_n}^{(n)}\Gamma_n\rangle dk}{\int_u \langle N_u^{(n)}\rangle du} = h(\alpha)$$

Local ($q \ll 1$) and Global ($q = 1$) Duplication Limits. The asymptotic degree distribution of the GDD model can be conveniently mapped into the (Γ, M) plane for two limit regimes of prime biological relevance: (*i*) for local duplication events ($q \ll 1$ and $\gamma_{ss} = 1$; Fig. 4*A*) and (*ii*) for whole-genome duplication events ($q = 1$; Fig. 4*B*). See *SI Appendix* for details.

The local duplication-divergence limit leads to scale-free limit degree distributions for both conserved and nonconserved networks, with power law exponents $1 < \alpha + 1 \leq 3$ if $\gamma_{so} \simeq 1$ (i.e., which ensures that most previous interactions are conserved in at least one copy after duplication). By contrast, the whole-genome duplication-divergence limit leads to a wide range of asymptotic behaviors from nonconserved, exponential regimes to conserved, scale-free regimes with arbitrary power law exponents. Conserved, nondense networks require, however, an asymmetric divergence between old and new duplicates ($\gamma_{oo} \neq \gamma_{nn}$) (5) and lead to scale-free limit degree distributions with the same range of exponents $1 < \alpha + 1 \leq 3$ for maximum divergence asymmetry ($\gamma_{oo} \simeq 1$ and $\gamma_{nn} \simeq 0$).

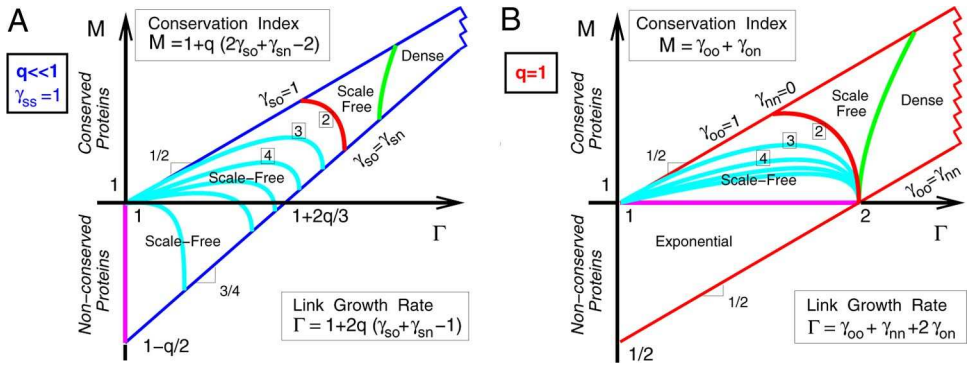

Fig. 4. Asymptotic phase diagram of PPI networks under the GDD model. (*A*) Local duplication-divergence limit ($q \ll 1$ and $\gamma_{ss} = 1$). (*B*) Whole-genome duplication-divergence limit ($q = 1$). Boxed figures are power law exponents ($\alpha + 1$) of scale-free regimes (Eq. **11**).

Evolutionary Variations of Model Parameters. The previous analysis with fixed parameters $\{q, \gamma_{ij}\}$ can be readily extended to combine local and global PPI network duplications (Fig. 4 *A* and *B*) or even include *any* evolutionary variations and stochastic fluctuations of the GDD model parameters with arbitrary series $\{q^{(n)}, \gamma_{ij}^{(n)}\}_R$ (see *SI Appendix*). Protein conservation is then found to be controlled by the cumulated product of connectivity growth/decrease rates following the most conserved, old duplicate lineage,

$$M = \left(\prod_n^R [(1 - q^{(n)})\Gamma_s^{(n)} + q^{(n)}\Gamma_o^{(n)}] \right)^{1/R} \quad [12]$$

with conserved (resp. nonconserved) protein evolutionary regimes corresponding to $M > 1$ (resp. $M < 1$).

A similar geometric average also controls the nature of the asymptotic degree distribution as the network topology index now reads,

$$M' = \left(\prod_n^R \max_i(\Gamma_i^{(n)}) \right)^{1/R} \quad [13]$$

with $M' < 1$ corresponding to exponential networks and $M' > 1$ to scale-free (or dense) networks with an effective node degree exponent α and effective node growth rate Δ that are self-consistent solutions of the generalized characteristic equation,

$$h(\alpha) = \left(\prod_n^R h^{(n)}(\alpha) \right)^{1/R} = \Delta, \quad [14]$$

where $h^{(n)}(\alpha) = (1 - q^{(n)})\Gamma_s^{(n)\alpha} + q^{(n)}\Gamma_o^{(n)\alpha} + q^{(n)}\Gamma_n^{(n)\alpha}$, as before. This leads to *exactly* the same discussion for singular regimes as with constant q and Γ_i (Fig. 3) because of the convexity of the generalized function $h(\alpha)$ ($\partial_\alpha^2 h(\alpha) \geq 0$; see *SI Appendix* for details and discussion on the $R \to \infty$ limit).

In particular, because $(1 - q^{(n)})\Gamma_s^{(n)} + q^{(n)}\Gamma_o^{(n)} \leq \max_i(\Gamma_i^{(n)})$ for *all* $q^{(n)}$ and $\Gamma_i^{(n)}$ ($i = s, o, n$), we *always* have $M \leq M'$. This relation implies a fundamental linkage between protein conservation and network topology under general duplication-divergence evolution, regardless of all possible evolutionary variations of the model parameters, $q^{(n)}$ and $\Gamma_i^{(n)}$. We expect, in particular, that *all conserved networks are necessarily scale-free* (or dense) ($1 < M \leq M'$), whereas *exponential networks can never be conserved* ($M \leq M' < 1$), under general duplication-divergence evolution.

Evolution of PPI Network Motifs. The generating function approach, introduced for the one-node degree distribution $p_k^{(n)}$ (Eqs. **1–6**), can be generalized to analyze the evolutionary statistics of multinode correlation functions and related clustering coefficient, distribution of first-neighbor average connectivity g_k (18) (see Fig. 6) and small-network motifs. Yet, although M' also controls transitions between major evolutionary regimes for multinode correlation functions, their analysis remains technically involved (*SI Appendix*).

By contrast, the conservation property of network motifs under general duplication-divergence evolution turns out to be remarkably simple, as outlined in Fig. 5. We derive conservation indices for specific network motifs by summing over all possible combinations of s nodes (with probability $\kappa_s = 1 - q$) or o nodes (with probability $\kappa_o = q$) and the corresponding γ_{ij} ($i, j = s, o$) (Fig. 5). Clearly, network motifs with a larger number of interactions, $p \geq 1$, have lower conservation indices, $M_p \simeq O(\gamma_{ij}^p)$ (Fig. 5). Moreover, because the probability to conserve a specific interaction γ_{ij} cannot be exactly 1, because of deleterious mutations (i.e., $\gamma_{ij} < 1$), motif conservation indices M_p must *all* be <1, regardless of any parameter variations, $q^{(n)}$ and $\gamma_{ij}^{(n)}$.

Hence, network motifs *cannot* be indefinitely conserved under duplication-divergence evolution, even though their individual proteins *are* typically conserved in the network (if $M > 1$) (Fig. 2). This implies that *structural* orthology between individual proteins from phylogenetically distant species *cannot* indefinitely coincide with *functional* orthology at the level of protein interactions and complexes, in broad agreement with empirical evidences (19). The resulting turnover toward more and more divergent interaction partners is a simple evolutionary consequence of the GDD model, regardless of any network-rewiring dynamics (that have been neglected here). In particular, even the most conserved orthologous proteins (*s*/*o* descents) must eventually perform different functions, but conserved ancestral functions are inevitably passed down to less conserved protein complexes (*s*/*o*/*n* descents) in phylogenetically distant species. This inherent evolutionary constraint of the GDD model sets

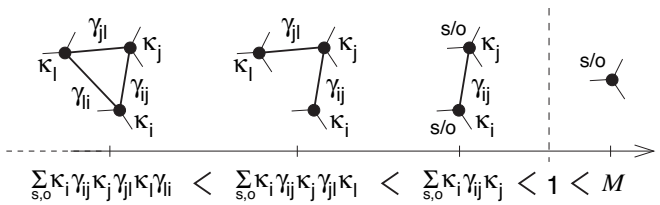

Fig. 5. Motif conservation indices. Although individual proteins are typically conserved (if $M > 1$), network motifs including two or more proteins *cannot* be indefinitely conserved under general duplication-divergence evolution ($\kappa_o = 1 - \kappa_s = q$; see text).

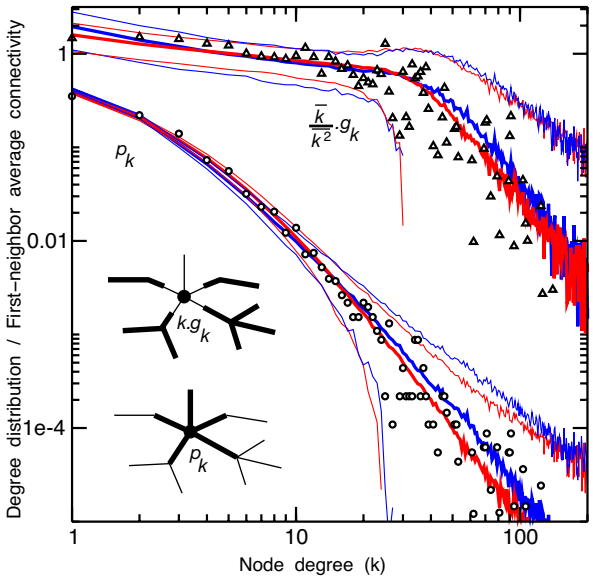

Fig. 6. Comparison between empirical PPI network data (20) and finite-size, duplication-divergence simulations. Protein physical interaction data for yeast are taken from the Biomolecular Interaction Network Database (20) (4,576 proteins, 9,133 physical interactions, $\bar{k} = 3.99$, $\bar{k}^2 = 106$). Both connectivity distribution p_k (open circles) and first-neighbor average connectivity g_k (18) (open triangles) are shown. Best one-parameter fit of the data (blue curves) ($\pm 2\sigma$ for expected deviations) for the most asymmetric genome duplication-divergence model with $\gamma_{on} = 0.26$ (and $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$) (5). Best two-parameter fit of the data (red curves) ($\pm 2\sigma$ for expected deviations) with the same most asymmetric whole-genome duplication limit but at the level of protein domains instead of entire proteins, with $\gamma_{on} = 0.1$ and $\lambda = 0.3$, which corresponds to $1/(1 - \lambda) = 1.5$ average number of binding domains per proteins. See *SI Appendix*, Fig. S6 and ref. 5 for details about this model including domain shuffling.

conceptual and practical limitations on functional annotation transfer between genomes (19).

Comparison with Empirical PPI Network Data. Although the GDD model can, in principle, accommodate $1q + 6\gamma_{ij}$ fitting parameters, and any evolutionary variations or fluctuations $q^{(n)}$ and $\gamma_{ij}^{(n)}$, we have restricted all quantitative comparisons with the empirical data (20) to biologically relevant regimes, limited to one or two effective fitting parameters, only.

The results, corresponding to 10^3 to 10^4 protein nodes and low network growth rates in the conserved, scale-free regime ($1 < M \leq M'$), are compared with Yeast PPI network data (20) (Fig. 6), both for the connectivity distribution p_k and the first-neighbor average connectivity g_k (18) (see Fig. 6 legend and *SI Appendix* for model and fitting details).

Extending these numerical simulations to larger PPI network sizes ($>10^5$ nodes) shows little changes in these distributions for small $k \leq 20$ (*SI Appendix*, Figs. S5–S7). This suggests that the available empirical PPI network for yeast (20) has essentially reached its asymptotic degree distribution, for $k \leq 20$. By contrast, global convergence of the GDD model is typically quite slow for large growth rate regimes with little biological relevance, as shown in *SI Appendix*, *Numerical simulations*, Figs. S5–S7.

Conclusions

Although conservation and network topology have *a priori* no reason to be related features of emerging PPI networks in the course of evolution, we demonstrate in this article that they are, in fact, linked properties under general duplication-divergence evolution, because of a fundamental linkage between protein conservation and network topology indices, that is, $M \leq M'$, regardless of any variations or fluctuations of the model parameters.

By contrast, we also showed that conservation of network motifs cannot be indefinitely preserved under general duplication-divergence evolution (independently from any network rewiring dynamics). This underlies intrinsic limitations on functional annotation transfer across phylogenetically distant genomes (19).

All in all, these evolutionary constraints, inherent to duplication-divergence processes and independent from selective adaptation (3), appear to have largely controlled the overall topology and scale-dependent conservation of PPI networks, regardless of any specific biological function.

ACKNOWLEDGMENTS. We thank U. Alon, R. Bruinsma, M. Cosentino-Lagomarsino, T. Fink, R. Monasson, and M. Vergassola for discussion. This work was supported by the Ministry of Higher Education and Research, Centre National de la Recherche Scientifique, and Institut Curie.

1. Ohno S (1970) *Evolution by Gene Duplication* (Springer, New York).
2. Li WH (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
3. Lynch M (2007) *The Origins of Genome Architecture* (Sinauer, Sunderland, MA).
4. Berg J, Lässig M, Wagner (2004) A Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol* 4:51.
5. Evlampiev K, Isambert H (2007) Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst Biol* 1:49.
6. Maslov S, Sneppen K, Eriksen KA, Yan KK (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* 4:9.
7. Jones S, Thornton JM, (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 93:13–20.
8. Henschel A, Kim WK, Schroeder M (2006) Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics* 22:550–555.
9. Kim WK, Henschel A, Winter C, Schroeder M, (2006) The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput Biol* 2:e124.
10. Ispolatov I, Krapivsky PL, Yuryev (2005) A Duplication-divergence model of protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys* 71:061911.
11. Albert R, Barabási A-L (2002) Statistical mechanisms of complex networks. *Rev Mod Phys* 74:47–97.
12. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5:101–113.
13. Raval A (2003) Some asymptotic properties of duplication graphs. *Phys Rev E Stat Nonlin Soft Matter Phys* 68:066119.
14. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. *ComPlexUs* 1:38–44.
15. Ispolatov I, Krapivsky PL, Mazo I, Yuryev A (2005) Cliques and duplication-divergence network growth. *N J Phys* 7:145.
16. Hartman H, Favaretto P, Smith TF (2006) The archaeal origins of the eukaryotic translational system. *Archaea* 2(1):1–9.
17. Flajolet P, Sedgewick R (2006) *Analytic Combinatorics.* Available at http://algo.inria.fr/flajolet/Publications/books.html. Accessed March 22, 2007.
18. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910.
19. Yu H, *et al.* (2004) Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. *Genome Res* 14:1107–1118.
20. Alfarano C, *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 33:D418–D424.

Supporting Information

1 Proof of the evolutionary recurrence for the node degree generating function (Eq. 2)

The generating function for node degrees $N_k^{(n)}$ after n duplications is defined as,

$$F^{(n)}(x) = \sum_{k\geq 0} \langle N_k^{(n)} \rangle x^k. \tag{S1}$$

where $\langle\cdot\rangle$ corresponds to the ensemble average over *all* possible trajectories of the evolutionary dynamics. The x^k term of $F^{(n)}(x)$ "counts" the statistical number of nodes with exactly k links (one x per link).

At each time step $n \to n+1$, each node can be either duplicated with probability q, giving rise to two node copies, or non-duplicated with probability $1-q$. Hence, in the general case with asymmetric divergence of duplicates (with a more conserved, "old" copy and a more divergent, "new" copy), there are 3 $F^{(n)}\big(A_i(x)\big)$ contributions to the updated $F^{(n+1)}(x)$ coming from each node type, $i = s, o, n$, for singular nodes, old and new duplicates,

$$F^{(n+1)}(x) = (1-q)F^{(n)}\big(A_s(x)\big) + qF^{(n)}\big(A_o(x)\big) + qF^{(n)}\big(A_n(x)\big) \tag{S2}$$

where the substitutions $x \to A_i(x)$ in each $F^{(n)}$ terms ($i = s, o, n$) should reflect the statistical fate of a particular link "x" between a node of type i and a neighbor node which is either singular (s) with probability $1-q$ or duplicated (o/n) with probability q. In practice, the duplication of a fraction q of (neighbor) nodes first leads to the replacement $x \to (1-q)x + qx^2$ corresponding to the maximum preservation of links for both singular (x) and duplicated o/n (x^2) neighbors, and then to the subtitution $x \to \gamma_{ij}x + \delta_{ij}$ for each type of neighbor nodes $j = s, o, n$ where γ_{ij} is the probability to preserve a link "x" (and $\delta_{ij} = 1-\gamma_{ij}$ the probability to erase it). Hence, the complete substitution correponding to the GDD model reads $x \to (1-q)(\gamma_{is}x + \delta_{is}) + q(\gamma_{io}x + \delta_{io})(\gamma_{in}x + \delta_{in}) = A_i(x)$ for $i = s, o, n$, leading to Eq.(S2).

2 Statistical properties of the model

The approach we use to study the evolution of PPI networks under general duplication-divergence processes is based on ensemble averages over all evolutionary trajectories. We characterize, in particular, PPI network evolution in terms of average number of nodes and links and average degree distribution. Yet, in order for these average features to be representative of typical network dynamics, statistical fluctuations around the mean trajectory should not be too large. In practice, it means that the relative variance $\chi_Q^2(n)$ for a feature $Q^{(n)}$ should not diverge in the limit $n \to \infty$,

$$\chi_Q^{2\,(n)} = \left(\frac{\langle Q^2\rangle - \langle Q\rangle^2}{\langle Q\rangle^2}\right)^{(n)} < \infty \quad \text{as} \quad n \to \infty$$

and more generally the pth moment of $Q^{(n)}$ should not diverge more rapidly than the pth power of the average. If it is not the case, successive moments exhibit a whole multifractal spectrum and ensemble averages do not represent typical realizations of the evolutionary dynamics. In order to check whether it is or not the case here for general duplication-divergence models, we proceed by analyzing the probability distributions for the number of links and nodes.

The number of link L has a probability distribution $\mathcal{P}(L)$ whose generating function $\mathcal{P}(x) = \sum_{L\geq 0} \mathcal{P}(L)x^L$ satisfies

$$\begin{aligned} &\mathcal{P}^{(n+1)}(x) = \mathcal{P}^{(n)}[a(x)], \\ &a(x) = (1-q)^2(\gamma_{ss}x+\delta_{ss}) + 2q(1-q)(\gamma_{so}x+\delta_{so})(\gamma_{sn}x+\delta_{sn}) + q^2(\gamma_{oo}x+\delta_{oo})(\gamma_{nn}x+\delta_{nn})(\gamma_{on}x+\delta_{on})^2. \end{aligned} \tag{S3}$$

This relation can be justified in a way similar to that of the fundamental evolutionary recurrence above: each node of the initial graph will be either duplicated d with probability q or kept singular s with probability $1-q$, leading to three possible node combinations for each link: $s-s$ link with probability $(1-q)^2$, $s-d$ or $d-s$ links with probability $2q(1-q)$ and $d-d$ link with probability q^2. Then each $s-s$ link is either kept with γ_{ss} and erased with δ_{ss} leading to the substitution $x \to \gamma_{ss}x+\delta_{ss}$ in the corresponding term; each $s-d$ or $d-s$ link can lead to two links between s and each o/n duplicate, *i.e.* $x \to (\gamma_{so}x+\delta_{so})(\gamma_{sn}x+\delta_{sn})$, while each $d-d$ link can lead up to 4 links after duplication, *i.e.* $x \to (\gamma_{oo}x+\delta_{oo})(\gamma_{nn}x+\delta_{nn})(\gamma_{on}x+\delta_{on})^2$. Combining all these operations eventually yields Eq.(S3).

Successive moments of this distribution are obtained taking successive derivatives of Eq.(S3),

$$A_k^{(n)} = \partial_x^k \mathcal{P}^{(n)}(x)\big|_{x=1}, \tag{S4}$$

and lead to the following recurrence relations

$$A_k^{(n+1)} = [h(1)]^k A_k^{(n)} + \frac{C}{2}k(k-1)[h(1)]^{k-2}A_{k-1}^{(n)} + \dots$$

where $h(1) = a'(1) = (1-q)\Gamma_s + q\Gamma_o + q\Gamma_n$ and $C = a''(1)$ are constants depending on microscopic parameters. These relations can be solved to get the leading order behavior of successive moments

$$A_k^{(n)} = \tilde{A}_k [h(1)]^{kn} \Big(1 + \mathcal{O}\big([h(1)]^{-n}\big) \Big), \tag{S5}$$

where $\tilde{A}_k$ are some functions of microscopic parameters.

The latter relation implies that the kth moment is equal (modulo some finite constant) to the kth power of the first moment in the leading order when $n \to \infty$. This suggests that in this limit the probability distribution should take a scaling form,

$$\mathcal{P}^{(n)}(L) \simeq \frac{1}{\langle L^{(n)} \rangle} \mathcal{F}\Big(\frac{L}{\langle L^{(n)} \rangle} \Big), \quad n \gg 1. \tag{S6}$$

This hypotesis can be verified directly from the explicit form of Eq.(S3)(see Appendix A for details).

Although we are not able to determine the scaling function $\mathcal{F}$ from previous considerations, we can derive some of its properties from the successive moments Eq.(S4): in particular for $n \gg 1$ the link distribution and the function $\mathcal{F}$ do not present a vanishing width around their mean value but instead a finite limit width corresponding to a finite relative variance,

$$\chi_L^{2\,(n)} = \left(\frac{\langle L^2 \rangle - \langle L \rangle^2}{\langle L \rangle^2} \right)^{(n)} \to \frac{1}{L^{(0)}} \left(\frac{a''(1)}{a'(1)(a'(1)-1)} - 1 \right) < \infty,$$

This relation is found solving explicitly Eq.(S4) for $k = 1$ and $k = 2$ given the initial number of links $L^{(0)}$. Hence, although fluctuations in the number of links are important, they remain of the same order of magnitude as the mean value, Fig. S1. This result is in fact rather surprising for a model which clearly exhibits a memory of its previous evolutionary states and might, in principle, develop diverging fluctuations in the asymptotic limit.

Fluctuations for the total number of nodes, $N^{(n)}$, and the number of nodes of degree $k \geq 1$, $N_k^{(n)}$, can also be evaluated using the previous result on link fluctuations and the double inequality $N_k \leq N \leq 2L$, valid for *any* graph realization. Indeed, we obtain the following relations between the pth moments and the pth power of the corresponding first moments,

$$\begin{aligned}
\langle (N^p)^{(n)} \rangle &\leq 2^p \langle (L^p)^{(n)} \rangle \propto 2^p \langle L^{(n)} \rangle^p = \big(\overline{k}^{(n)} \big)^p \langle N^{(n)} \rangle^p, \\
\langle (N_k^p)^{(n)} \rangle &\leq 2^p \langle (L^p)^{(n)} \rangle \propto 2^p \langle L^{(n)} \rangle^p = \big(\overline{k}^{(n)} \big)^p \langle N^{(n)} \rangle^p = \left(\frac{\overline{k}^{(n)}}{p_k^{(n)}} \right)^p \langle N_k^{(n)} \rangle^p.
\end{aligned}$$

using $\langle L^{(n)} \rangle = \overline{k}^{(n)} \langle N^{(n)} \rangle$ and $\langle N_k^{(n)} \rangle = p_k^{(n)} \langle N^{(n)} \rangle$, for all $n \geq 1$ and $k \geq 1$. Hence, we find that fluctuations for both N and N_k remain finite in the asymptotic limit for *linear* asymptotic regimes corresponding to exponential or scale-free degree distributions with *finite* limit values for both mean degree, $\overline{k}^{(n)} \to \overline{k} < \infty$ and degree distribution $p_k^{(n)} \to p_k > 0$, for all $k \geq 1$. This corresponds presumably to the most biologically relevant networks. On the other hand, for *non-linear* (scale-free or dense) asymptotic regimes previous arguments do not apply as $\overline{k}^{(n)} \to \infty$ (and $p_k^{(n)} \to 0$ for dense regime) when $n \to \infty$. The numbers of nodes $N^{(n)}$ and $N_k^{(n)}$ grow exponentially more slowly than the number of links $L^{(n)}$ in this case, and the growth process might develop, in principle, diverging fluctuations as compared to their averages, $\langle N^{(n)} \rangle$ and $\langle N_k^{(n)} \rangle$, respectively. Yet, numerical simultations (see section 8 below) tend to show that it is actually *not* the case, suggesting that the ensemble average approach we have used to study the GDD model is still valid for *non-linear* asymptotic regimes.

In summary, we found that the general duplication-divergence dynamics is not *stricto sensu* "self-averaging" [2] (*i.e.* $\chi^{(n)} \to 0$), however, fluctuations remain finite (*i.e.* $\chi^{(n)} < \infty$) in the asymptotic limit $n \to \infty$, which corresponds to a so-called 'fractal' regime. This implies that ensemble averages of PPI network evolution are actually good representatives of typical PPI network realizations in biologically relevant regimes. Conversely, an ensemble averaging approach would become inappropriate in the presence of diverging fluctuations ($\chi^{(n)} \to \infty$) correponding to a so-called 'multifractal' regime. Hence, our findings on the finite fluctuations of the model justifies its statistical foundations and, thereby, our theoretical approach based on functional recurrences between successive generating functions.

Interestingly, Fig. S1 shows that duplication-divergence evolution do *not* tend to be more "self-averaging" in the low growth rate limit, nor equivalently in the local duplication limit ($q \ll 1$). This means that the "self-averaging" property suggested in previous works on the basis of numerical simulations [2] might actually results as a consequence of the time-linear, single protein duplication dynamics adopted by these authors.

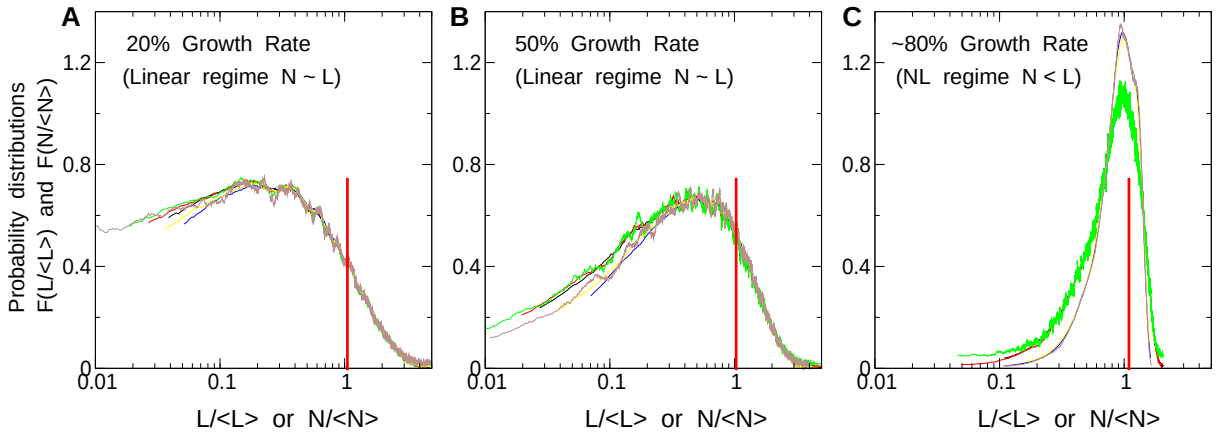

Figure S1: **Numerical simulations of link and node probability distributions.** Scaling functions $\mathcal{F}(x/\langle x\rangle)$, eq.(S6), obtained for link probability distribution (brown, $x = L$) and node probability distribution (green, $x = N$) under whole genome duplication dynamics with 20%, 50% and 80% network growth rates, respectively. These distributions are stationnary in all three cases and are very close for nodes and links in linear dynamical regimes, as expected (**A** and **B**), and distinct for non-linear regimes (**C**).

3 Asymptotic methods

In this section, we give more details about the asymptotic analysis of node degree distribution defined by the recurrence relation on its normalized generating function $p^{(n)}(x)$, Eq.(6).

First of all, the series of $p^{(n)}(x)$ can be shown to converge at each point at least for some initial conditions. Indeed, let us introduce a linear operator $\mathcal{M}$ defined on functions continous on $[0,1]$ and acting according to Eq.(6), *i.e.*, $p^{(n+1)} = \mathcal{M}p^{(n)}$. For two non-negative functions $f(x)$ and $g(x)$ so that $f(0) = 0$, $g(0) = 0$, $f(1) = 1$ and $g(1) = 1$, we have,

$$\forall x \in [0,1] \;\; f(x) \leq g(x) \Rightarrow \forall x \in [0,1] \;\; (\mathcal{M}f)(x) \leq (\mathcal{M}g)(x). \tag{S7}$$

It can be verified that if $p^{(0)}(x)=x$ (one simple link as initial condition), $\mathcal{M}p^{(0)}(x)\leq p^{(0)}(x)$ $\forall x \in [0,1]$ and by consequence, when applying $\mathcal{M}^n$ to this inequality, the following holds

$$0 \leq p^{(n+1)}(x) \leq p^{(n)}(x), \;\; \forall x \in [0,1]$$

which means that at each point the series of $p^{(n)}(x)$ is decreasing and converges to some non-negative value $p(x)$. Futhermore, numerical simulations show that for an arbitrary initial condition, there exists an $n_0 > 1$ suffisiently large so that $p^{(n)}(x)$ decreases for $n \geq n_0$. Hence, we can take the limit $n \to \infty$ on both sides of Eq.(6) to get the Eq.(9) for the limit function $p(x)$.

We analyze the properties of this generating function $p(x)$ for the limit degree distribution, using asymptotic methods. Indeed, we have no mean to solve analytically this functional equation to precisely obtain the corresponding limit degree distribution, but we have enough information to deduce its asymptotic behavior at large k, since it is directly related to the asymptotic properties of $p(x)$ for $x \to 1$. In the following, we note $h(\alpha) = (1-q)\Gamma_s^\alpha + q\Gamma_o^\alpha + q\Gamma_n^\alpha$, following the same notation as in the main text.

First, we consider the relation between successive derivatives of $p(x)$ at $x = 1$ deduced from Eq.(6) by taking the corresponding number of derivatives, Eq.(10),

$$\left[1 - \frac{h(k)}{\Delta}\right] \partial_x^k p(1) = \sum_{l=[k/2]}^{k} \alpha_{k,l} \, \partial_x^l p(1), \tag{S8}$$

with some positif coefficients $\alpha_{k,l}$. The value of Δ in this relation is still unknown and should be determined self-consistently with $p(x)$. Each of these derivatives can also be obtained as a limit of value $\partial_x^k p(1) = \lim_{n\to\infty} \partial_x^k p^{(n)}(1)$, with the following recurrence relation for $\partial_x^k p^{(n)}(1) = m_k^{(n)}$

$$m_k^{(n+1)} = \frac{h(k)}{\Delta^{(n)}} m_k^{(n)} + \frac{\tilde{C}}{2} k(k-1) \frac{h(k-2)}{\Delta^{(n)}} m_{k-1}^{(n)} + \dots \tag{S9}$$

directly derived from Eq.(6). Different regimes can be identified depending on the general convex shape of $h(\alpha)$ ($\partial_\alpha^2 h(\alpha) \geq 0$).

Regular regimes - $h(\alpha)$ strictly decreasing for $\alpha > 0$ iff $M' = \max_i(\Gamma_i) < 1$, for $i = s, o, n$.
In this case, if we suppose that $p'(1)$ is finite, all the derivatives of $p(x)$ at $x = 1$ are finite since $\Delta = h(1)$ and $h(k) < h(1)$ for $\forall k \geq 2$. In fact, the alternative situation $p'(1) = \infty$ and $\Delta < h(1)$ is not possible as it would imply that some first moments in Eq.(S9), at least $m_1^{(n)}$ and $m_2^{(n)}$, would diverge exponentially as $(h(1)/\Delta)^n$. However, since $h(k) < h(1)$ for $k \geq 2$, this would contradict the fact that the nth moment grows more rapidly than the nth power of the first one. Hence, we must have $\Delta = h(1)$ and the solution is not singular at $x = 1$ but may have a singularity at some $x_0 > 1$.

Taking an anzats for the asymptotic expansion in the form

$$p(x) = A_0 - A_1(x_0 - x) + A_2(x_0 - x)^2 + A_\alpha(x_0 - x)^\alpha + \mathcal{O}\big((x_0 - x)^{\alpha+1}\big). \tag{S10}$$

and inserting it in Eq.(9) we find that, in order to have the singularity at $x = x_0$ present on both sides of the equation, x_0 has to be chosen as the root closest to 1 in the following three equations,

$$A_s(x) = x,\ A_o(x) = x,\ A_n(x) = x, \tag{S11}$$

where, $A_i(x) = (1-q)(\gamma_{is}x + \delta_{is}) + q(\gamma_{io}x + \delta_{io})(\gamma_{in}x + \delta_{in})$ for $i = s, o, n$, or explicitly (since the second root is always 1)

$$x_0 = \min\left(\frac{(1-q)\delta_{ss} + q\delta_{so}\delta_{sn}}{q\gamma_{so}\gamma_{sn}}, \frac{(1-q)\delta_{so} + q\delta_{oo}\delta_{on}}{q\gamma_{oo}\gamma_{on}}, \frac{(1-q)\delta_{sn} + q\delta_{on}\delta_{nn}}{q\gamma_{on}\gamma_{nn}}\right).$$

Since $h(\alpha)$ is strictly decreasing when $\Gamma_s < 1$, $\Gamma_o < 1$ and $\Gamma_n < 1$, it is straightforward to prove that all three values are greater than one, and hence, $x_0 > 1$ for regular regimes.

The value of α is obtained from the same equation Eq.(6) by comparing the coefficients in front of the singular terms when developping each term near $x = x_0$

$$\alpha = \frac{\ln(\epsilon_i \Delta)}{\ln(2 - \Gamma_i)}, \tag{S12}$$

where $i = s$, o or n if x_0 is the solution of $A_i(x) = x$, $\epsilon_s = (1-q)^{-1}$, $\epsilon_o = \epsilon_n = q^{-1}$, and replacing also $\epsilon_i \to 1/2\epsilon_i$ or $1/3\epsilon_i$ if two or all three Γ_i's happen to be equal, respectively.

We recall that for $h(\alpha)$ under consideration $\overline{k} = p'(1)$ is finite and $\Delta = h(1)$. Therefore, in this regime the asymptotic growth of the graph is exponential with respect to the number of links and the number of nodes with a common growth rate $\Delta = h(1)$. We call this asymptotic behavior "*linear*" because $\langle L^{(n)} \rangle$ and $\langle N^{(n)} \rangle$ are asymptotically proportional.

The decrease of the limit degree distribution for $k \gg 1$ is given by [3]

$$p_k \propto k^{-\alpha-1} x_0^{-k} \left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right),\ \ k \gg 1 \tag{S13}$$

and is thus *exponential* with a power law prefactor. When one of the Γ_i's tends to one, simultaneously $x_0 \to 1$ and $\alpha \to \infty$ and, as we will see below, we meet the singular scale-free regime for the limit mean degree distribution.

The emergence of an exponential tail for p_k when $k \gg 1$ naturally comes from the fact that at each duplication step the probability for a node to duplicate one of its links (keeping both the original link and its copy), $q\gamma_{oo}\gamma_{on}$ for o nodes, $q\gamma_{so}\gamma_{sn}$ for s nodes and $q\gamma_{on}\gamma_{nn}$ for n nodes, is smaller than the corresponding probabilities to delete the initial link, $(1-q)\delta_{so} + q\delta_{oo}\delta_{on}$, $(1-q)\delta_{ss} + q\delta_{so}\delta_{sn}$ and $(1-q)\delta_{sn} + q\delta_{on}\delta_{nn}$ (it is in fact equivalent to $x_0 > 1$). For this reason at each duplication only few nodes are preserved and they keep only few of their links, the graph contains many small components and has no memory about previous states. In a different way, we can develop this argument in terms of a particular node degree evolution. When $\Gamma_o < 1$, $\Gamma_s < 1$ and $\Gamma_n < 1$, nodes o and s as well as their copies n loose links in proportion to their connectivities. It means that the number of nodes of a given connectivity is modified by a Poissonian prefactor, representing the overall tendency to follow an exponentially decreasing distribution for large number of duplications.

Singular regimes - $h(\alpha)$ has a minimum on $\alpha > 0$ iff $M' = \max_i(\Gamma_i) > 1$, for $i = s, o, n$.
In this case, from Eq.(S8) we can be sure to have a negative value for some derivative: since $h(\alpha)$ has a unique minimum, there exists an integer $r \geq 1$ so that $h(r) < \Delta < h(r+1)$ implying that $\partial_x^{r+1} p(1) < 0$ which is impossible by construction. In fact, this indicates the presence of an irregular term in the development of $p(x)$ in the vicinity of $x = 1$, and for this reason the function itself is r times differentiable at this point while its $(r+1)$th and following derivatives do not exist. Hence, we take an anzats for $p(x)$ in the neighborhood of $x = 1$ using the following form

$$p(x) = 1 - A_1(1 - x) + A_2(1 - x)^2 + A_\alpha(1 - x)^\alpha + \mathcal{O}\big((1 - x)^{\alpha+1}\big) \tag{S14}$$

A priori, we do not know the exact value of Δ, and it is to be determined self-consistently with $p(x)$. We then substitute Eq.(S14) into Eq.(9) to get a "characteristic" equation relating α and Δ,

$$h(\alpha) = (1-q)\Gamma_s^\alpha + q\Gamma_o^\alpha + q\Gamma_n^\alpha = \Delta. \tag{S15}$$

If we find a nontrivial value of $\alpha^* > 0$ that are solutions of this equation, it will give us an asymptotic expression for the coefficients of the generating function of the *scale free* form

$$p_k \propto k^{-\alpha^*-1}\left(1+\mathcal{O}\left(\frac{1}{k}\right)\right), \quad k \gg 1. \tag{S16}$$

Note that when the solution takes an integer value $\alpha^* = r \geq 1$ the form of the asymptotic expansion should differ from Eq.(S14) because formally it is not longer singular in this case. In fact, in the anzats a logarithmic prefactor should be added in the singular term

$$p(x) = 1 - A_1(1-x) + A_2(1-x)^2 + \ldots + A_r(1-x)^r + \tilde{A}_r(1-x)^r \ln(1-x) + \mathcal{O}\big((1-x)^{r+1}\big) \tag{S17}$$

In order for this asymptotic expansion to satisfy Eq.(9), we should have $h(r) = \Delta$, as before, as well as an additional condition for $r = 1$ namely $h'(1) = 0$.

Note also, that the characteristic equation $h(\alpha) = \Delta$ can be recovered directly (although less rigorously) using the connectivity change $k \to k\Gamma_i$ on average for i-type of nodes ($i = s, o, n$) at each duplication and the following continuous approximation, $N^{(n)} = \sum_k N_k^{(n)} \simeq \int_u N_u^{(n)} du$,

$$\Delta = \frac{\langle N^{(n+1)}\rangle}{\langle N^{(n)}\rangle} \simeq \frac{\int_k \big\langle (1-q)N^{(n)}_{k\Gamma_s}\Gamma_s + qN^{(n)}_{k\Gamma_o}\Gamma_o + qN^{(n)}_{k\Gamma_n}\Gamma_n\big\rangle dk}{\int_u \langle N_u^{(n)}\rangle du} = \frac{\big((1-q)\Gamma_s^\alpha + q\Gamma_o^\alpha + q\Gamma_n^\alpha\big)\int_u \langle N_u^{(n)}\rangle du}{\int_u \langle N_u^{(n)}\rangle du} = h(\alpha)$$

where we assumed that $\langle N_k^{(n)}\rangle \propto k^{-\alpha-1}$.

Three cases should now been distinguished depending on the signs of $h'(0)$ and $h'(1)$ (see Fig. 3 in main text):

1. $h'(0) < 0$ and $h'(1) < 0$.

Since $h(\alpha) > h(1)$ for $\alpha < 1$, any solution of Eq.(S15) has to be greater than one (as $\Delta \leq h(1)$) which implies, by vertue of Eq.(S16), $\overline{k} < \infty$ and consequently $\Delta = h(1)$ exactly (which is consistent with previous considerations). So, for the parameters satisfying $h'(1) < 0$ the value of α we are looking for is the unique solution, $\alpha^\star > 1$, of

$$h(\alpha^\star) = (1-q)\Gamma_s^{\alpha^\star} + q\Gamma_o^{\alpha^\star} + q\Gamma_n^{\alpha^\star} = (1-q)\Gamma_s + q\Gamma_o + q\Gamma_n = h(1) \tag{S18}$$

The other solution $\alpha = 1$ should be discarded here as it corresponds to a solution only if $h'(1) = 0$ (see proof for the most general dupication-divergence hybrid models, below).

Evidently, in this regime there exists an entier $k_0 \geq 1$ for which

$$h(k_0) < h(\alpha^*) \leq h(k_0+1),$$

and so all the derivatives of $p^{(k)}(1)$ are finite for $k \geq k_0$ while all following derivatives are infinite. Finally, when we fix Γ_i which are less than one and make other $\Gamma_i \to 1$ the value of α^* tends to infinity, the scale free regime Eq.(S16) meets the exponential one Eq.(S13).

2. $h'(0) < 0$ and $h'(1) \geq 0$.

The condition $\Delta \leq h(1)$ implies that only solutions with $0 < \alpha \leq 1$ are possible. Therefore, surely $\overline{k} = \infty$ in this case but there is no additional constraints *a posteriori* on Δ which might take, in principle, a whole range of possible values between $\min_\alpha(h(\alpha))$ and $\min(h(0), h(1))$. Yet, numerical simulations suggest that there might still be a unique asymptotic node growth rate Δ regardless of initial conditions or evolution trajectories, although convergence is extremely slow (See *Numerical simulations* below).

3. $h'(0) \geq 0$ (we always have $h'(1) > 0$ in this case).

The minimum of $h(\alpha)$ is achieved for $\alpha_0 < 0$ in this case, and $\Delta \leq 1 + q \leq h(1)$. Yet because solutions of Eq.(S15) cannot be negative by definition of $p(x)$, the only possibility is $\Delta = 1 + q$, implying that the graph grows at the maximum pace. From the point of view of the graph topology, it means that the mean degree distribution is not stationnary and for any fixed k the mean fraction of nodes with this connectivity k tends to zero when $n \to \infty$, the number of links grows too rapidly with respect to the number of nodes so that the graph gets more and more dense. For this reason, we refer at this regime as the *dense* one.

4 Whole genome duplication-divergence model ($q = 1$)

The case $q = 1$ describes the situation for which the entire genome is duplicated at each time step, corresponding to the evolution of PPI networks through whole genome duplications, as discussed in ref. [1]. All results obtained above in the general case remain valid although there are now no more "singular" genes (s) and thus no γ_{ij}'s involving them. We just summarize these results here adopting the notations of ref. [1] for the only 3 relevant γ_{ij}'s left: $\gamma_o \equiv \gamma_{oo}$, $\gamma_n \equiv \gamma_{nn}$ and $\gamma \equiv \gamma_{on}$, hence

$$\Gamma_o = \gamma_o + \gamma, \;\; \Gamma_n = \gamma_n + \gamma, \tag{S19}$$

The model analysis then yields three different regimes (we do not consider the case $\Gamma_o+\Gamma_n < 1$ for which graphs vanish)

1. Exponential regime $\Gamma_o+\Gamma_n > 1$**,** $\max(\Gamma_o, \Gamma_n) < 1$. The limit degree distribution is nontrivial and decreases like Eq.(S13) with

$$x_0 = \begin{cases} \delta_o\delta/\gamma_o\gamma, \; \Gamma_o < \Gamma_n \\ \delta_n\delta/\gamma_n\gamma, \; \Gamma_o \geq \Gamma_n \end{cases} \tag{S20}$$

and

$$\alpha = \frac{\ln(\Gamma_o + \Gamma_n)}{\ln\big(2 - \max(\Gamma_o, \Gamma_n)\big)}, \; \Gamma_o \neq \Gamma_n \tag{S21}$$

while

$$\alpha = \frac{\ln(\Gamma_o)}{\ln(2 - \Gamma_o)}, \;\; \text{for} \;\; \Gamma_o = \Gamma_n. \tag{S22}$$

The rate of graph growth in number of nodes as well as in number of links is $\Delta = \Gamma_o+\Gamma_n$

2. Scale free regime ($\Gamma_o > 1$**,** $\Gamma_n < 1$) **or** ($\Gamma_o < 1$**,** $\Gamma_n > 1$). The limit degree distribution is surely nontrivial for

$$h'(1) = \Gamma_o \ln \Gamma_o + \Gamma_n \ln \Gamma_n < 0, \tag{S23}$$

and described by an asymptotic formula Eq.(S16) with $\alpha^* > 1$ solution of

$$\Gamma_o^\alpha + \Gamma_n^\alpha = \Gamma_o + \Gamma_n, \tag{S24}$$

In this case the ratio of two consecutive sizes is also $\Delta = \Gamma_o+\Gamma_n$. When

$$h'(1) = \Gamma_o \ln \Gamma_o + \Gamma_n \ln \Gamma_n > 0, \;\; \Gamma_o\Gamma_n < 1 \;\; (h'(0) < 0),$$

the mean degree distribution is still expected to converge to a nontrivial asymptotically scale-free distribution with $0 \leq \alpha \leq 1$.

3. Dense regime $\Gamma_o\Gamma_n > 1$ (***i.e.*** $h'(0) > 0$). The mean degree distribution is not stationary: the growing graphs get more and more dense in the sense that the fraction of nodes with an arbitrary fixed connectivity tends to zero when $n \to \infty$. Almost all new nodes are kept in the duplicated graph $\Delta = 1 + q$.

Because all these regimes are defined in terms of two independent parameters (instead of three), the model phase diagram can be drawn in a plane (Γ_o, Γ_n), or equivalently in $(\Gamma_o+\Gamma_n, \Gamma_o)$ (See Fig. S2A & B). This last representation is adapted to show explicitly the domains of node conservation and graph growth, while the alternative choice $(\Gamma_o+\Gamma_n, \Gamma_o-\Gamma_n)$ used in ref. [1] is best suited to illustrate the asymmetric divergence requirement to obtained scale-free networks (see [1] for a detailed discussion).

5 Local duplication-divergence limit ($q \to 0$)

A different limit model is obtained for q going to zero when the mean size of the graph tends to infinity, Fig. S2C. In principle, the most general model of this kind is the one defined by a monotonous decreasing function $q(\langle N \rangle)$ with

$$\lim_{x\to\infty} q(x) = 0.$$

For any function of this type, the graph growth rate in terms of links depends essentially on γ_{ss} because

$$\frac{\langle L^{(n+1)} \rangle}{\langle L^{(n)} \rangle} = (1-q)\Gamma_s + q\Gamma_o + q\Gamma_n = \gamma_{ss} + 2q(\gamma_{so} + \gamma_{sn} - \gamma_{ss}) + \mathcal{O}(q^2),$$

and if $\gamma_{ss} < 1$ the ensemble average of graphs will never reach infinite size, it will have at most some finite dynamics. So, we will suppose that $\gamma_{ss} = 1$, to ensure an infinite growth. We remark also that γ_{oo}, γ_{on} and γ_{nn} appear only in the term of order q^2 in the last expression because two new nodes have to be kept in order to add any link of the type *oo*, *on* or *nn*.

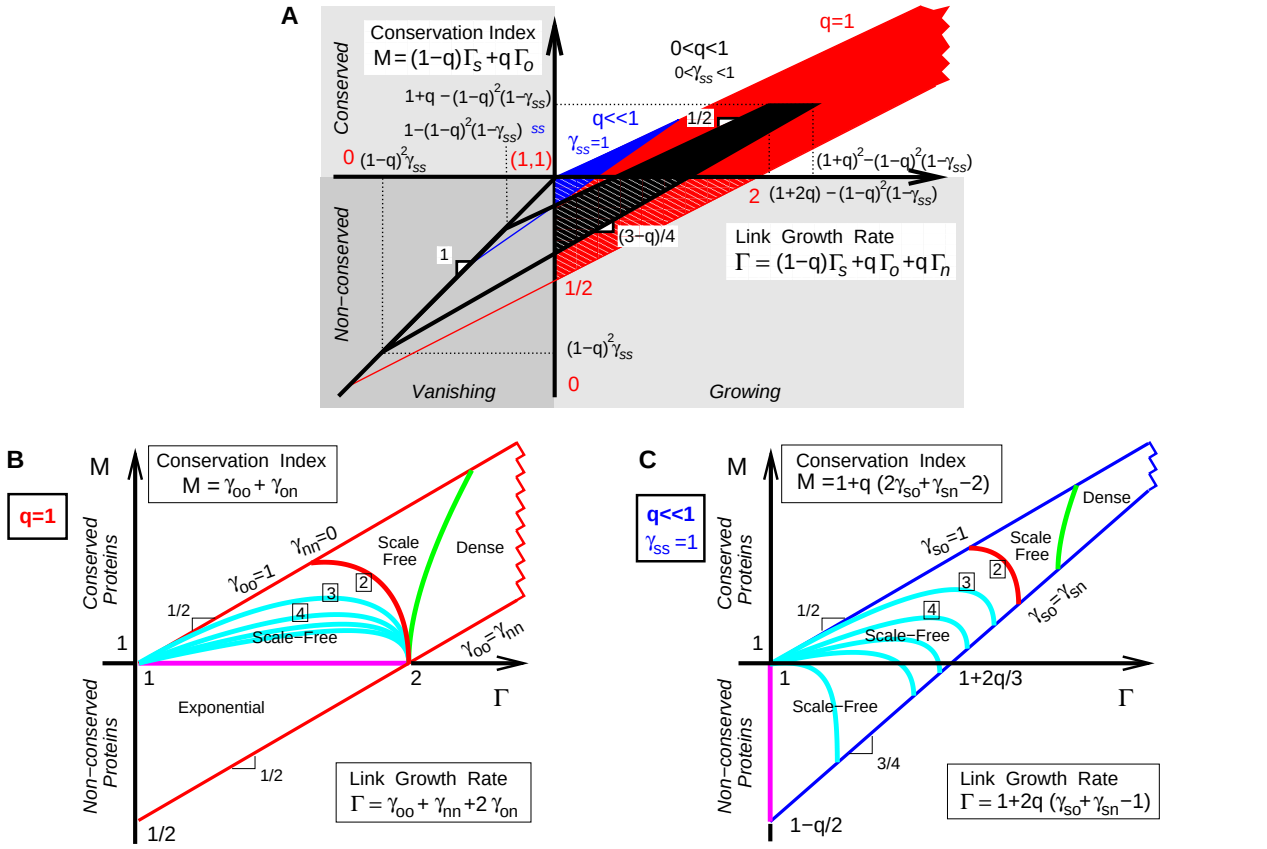

Figure S2: **Asymptotic phase diagram of PPI networks under the GDD model. A.** Phase diagram of GDD models for local (blue, $q \ll 1$, $\gamma_{ss} = 1$), partial (black, $q < 1$) and whole genome (red, $q = 1$) duplications, in the (Γ,M) plane. In particular, the condition $1+q-(1-q)^2(1-\gamma_{ss}) > 1$ should be enforced, implying $\gamma_{ss} > 1-q$, in the local duplication limit, $q \ll 1$. **B.** Whole genome duplication-divergence limit ($q = 1$). **C.** Local duplication-divergence limit ($q \ll 1$ and $\gamma_{ss} = 1$). Boxed figures are power law exponents ($\alpha + 1$) of scale-free regimes.

When q becomes small an approximate recursion relation for generating functions can be obtained by developping Eq.(6) (we set $\gamma_{ss} = 1$) with

$$\begin{aligned}
\tilde{p}^{(n)}\big(A_s(x)\big) &= \tilde{p}^{(n)}(x) + q\big((\delta_{so}+\gamma_{so}x)(\delta_{sn}+\gamma_{sn}x) - x\big)\partial_x\tilde{p}^{(n)}(x) + \mathcal{O}(q^2)\\
\tilde{p}^{(n)}\big(A_o(x)\big) &= \tilde{p}^{(n)}(\delta_{so}+\gamma_{so}x) + \mathcal{O}(q)\\
\tilde{p}^{(n)}\big(A_n(x)\big) &= \tilde{p}^{(n)}(\delta_{sn}+\gamma_{sn}x) + \mathcal{O}(q)
\end{aligned}
\quad , \tag{S25}$$

gives in linear order of q

$$\tilde{p}^{(n+1)}(x) = \frac{(1-q)\tilde{p}^{(n)}(x) + q\big((\delta_{so}+\gamma_{so}x)(\delta_{sn}+\gamma_{sn}x) - x\big)\partial_x\tilde{p}^{(n)}(x) + q\tilde{p}^{(n)}(\delta_{so}+\gamma_{so}x) + q\tilde{p}^{(n)}(\delta_{sn}+\gamma_{sn}x)}{\Delta^{(n)}} + \mathcal{O}(q^2)$$

with

$$\Delta^{(n)} = 1 - q\Big(\delta_{so}\delta_{sn}\partial_x\tilde{p}^{(n)}(0) + \tilde{p}^{(n)}(\delta_{so}) + \tilde{p}^{(n)}(\delta_{sn}) - 1\Big), \tag{S26}$$

an expression which does only depend on 3 of the 6 general γ_{ij}'s: γ_{ss}, γ_{so} and γ_{sn}. By neglecting terms in q^2 we obtain a model for which duplicated nodes are completely decorrelated in the sense that the probability for an o or s node to have two new neighbours is zero, and consequently any two new nodes do not have common neighbors. This model can be regarded as a generalization of the local duplication model proposed in [2] for which only one node is duplicated per time step and without modification of the connectivities between any other existing nodes, *i.e.* $\gamma_{ss} = 1$ and $\gamma_{so} = 1$. Indeed, when taking for q a decreasing law

$$q\big(\langle N\rangle\big) = \frac{A}{\langle N\rangle}, \quad A > 0$$

on average A nodes per step are duplicated. By setting $\gamma_{so} = 1$ in Eq.(S26) we first get the following form for the recurrence relation,

$$\tilde{p}^{(n+1)}(x) = \frac{\tilde{p}^{(n)}(x) + q\gamma_{sn}x(x-1)\partial_x\tilde{p}^{(n)}(x) + q\tilde{p}^{(n)}(\delta_{sn} + \gamma_{sn}x)}{\Delta^{(n)}}, \quad \Delta^{(n)} = 1 - q\tilde{p}^{(n)}(\delta_{sn}), \tag{S27}$$

and then using the definitions of $\Delta^{(n)}$ and $p^{(n)}(x)$ Eq.(4) to reexpress it as,

$$N_k^{(n+1)} \quad = \quad N_k^{(n)} + A\gamma_{sn}(k-1)p_{k-1}^{(n)} - A\gamma_{sn}kp_k^{(n)} + A\sum_{s\geq k} C_s^k \gamma_{sn}^k \delta_{sn}^{s-k} p_s^{(n)}, \tag{S28}$$

This expression is identical to the basic recurrence relation in the model of ref. [2] for $A = 1$. For an arbitrary A the asymptotic properties of the growing graph are essentially the same as in ref. [2], with only the growth rate modified by a factor proportional to A.

In the more general cases for which both γ_{sn} and γ_{so} may vary (with $\gamma_{ss} = 1$ remaining fix to ensure a non-vanishing graph), an asymptotic analysis can be carried out for the limit degree distribution with an asymptotic solution of the form

$$\tilde{p}(x) = -A_1(1-x) + A_2(1-x)^2 + A_\alpha(1-x)^\alpha + \mathcal{O}\big((1-x)^{\alpha+1}\big)$$

satisfying Eq.(S26) with $q \propto A/\langle N^{(n)}\rangle$. The characteristic equation thus becomes,

$$h_l(\alpha) = \gamma_{so}^\alpha + \gamma_{sn}^\alpha + \alpha(\gamma_{so} + \gamma_{sn} - 1) - 1 = \varphi,$$

where φ is defined as

$$\varphi = \lim_{n\to\infty} \frac{\big(\Delta^{(n)} - 1\big)}{q^{(n)}} \quad \Leftrightarrow \quad \Delta^{(n)} \simeq 1 + q^{(n)}\varphi, \; n \to \infty.$$

while the graph growth rate in terms of number of links is given by,

$$(1-q)\Gamma_s + q\Gamma_o + q\Gamma_n = 1 + q(2\gamma_{so} + 2\gamma_{sn} - 2) + \mathcal{O}(q^2),$$

at first order in q. Since the number of nodes can not grow more rapidly than the number of links, we can conclude that $\varphi \leq 2\gamma_{so} + 2\gamma_{sn} - 2$, in addition to, $\varphi \leq 1$, correponding to the maximum growth rate. Focussing the analysis on the case $\gamma_{so} + \gamma_{sn} > 1$ for which the graph does not vanish, one finds that the "characteristic" function $h_l(\alpha)$ is always convex, and the following results are obtained as in the asymptotic analysis of Sec. 3 in Supp. Information:

- When $h_l'(0) < 0$ and $h_l'(1) < 0$ the characteristic equation has a solution, $\alpha^* > 1$, and the limit degree distribution is asymptotically scale-free $p_k \propto k^{-\alpha^*-1}$ with α^* varying on the interval $[1, \infty)$ (depending on parameters γ_{so} and γ_{sn}) while $\varphi = h_l(1)$.
- For $h_l'(1) = 0$ precisely, the singular term of the asymptotic solution becomes $(1-x)\ln(1-x)$ and the limit degree distribution decreases as $p_k \propto k^{-2}$, for $k \gg 1$.
- When $h_l'(0) < 0$ and $h_l'(1) > 0$, scale-free regimes with slowly decreasing degree distributions are expected in general with $\varphi \leq \min(2\gamma_{so} + \gamma_{sn} - 2, 1)$ and the corresponding $0 < \alpha < 1$.
- For $h_l'(0) > 0$ the mean degree distribution is not stationary, $\varphi = 1$.

Fig. S2 summarizes these results for the limit degree distribution. More generally for

$$q\big(\langle N\rangle\big) = \frac{A}{\langle N\rangle^\beta}, \quad A > 0, \quad \beta > 0,$$

when $\beta > 1$, nodes a rarely duplicated so that the interval between two succesfull duplications in number of steps is approximately

$$n \propto \langle N^{(n)}\rangle^{\beta-1}.$$

Therefore $\beta > 1$ gives a model equivalent to $\beta = 1$ with a change of time scale. On the other hand, for $0 < \beta < 1$ a set of nontrivial models is obtained.

6 General Duplication-Divergence models including self-interacting proteins

One type of interactions, which is not included in the main GDD model, Fig. 1 and Eq. 2, corresponds to self-interacting loops, Fig. S3. Although self-link loops exist and might have played a critical role in the actual emergence of molecular interaction networks during early evolutionary times, they can be shown to have very little effects on the long time scale evolution of PPI networks under GDD model dynamics, which is why we have omitted them for simplicity in the main text.

Actually, the possibility of protein homo-oligomerization can be explicitly taken into account by introducing 2 types of nodes corresponding respectively to *i)* self-interacting proteins with self-link loops and *ii)* non-self-interacting proteins without self-link loops, Fig. S3. Available data on PPI networks reveals that about 10 to 15% of interacting proteins are self-interacting [5]. In principle, the detailed evolution of PPI network conservation and topology is affected by self-link loops which provide a source of duplication-derived *de novo* interactions between "old" and "new" copies of duplicated self-interacting proteins. We introduce four new evolutionary parameters, μ_s, μ_o, μ_n and μ_{on}, corresponding, respectively, to the probability to conserve a self-link interaction on a "singular" gene, "old" or "new" duplicated genes or to retain the duplication-derived *de novo* interaction between an old and new pair of gene copies from a duplicated self-interacting genes, Fig. S3. Together with previous "$1q+6\gamma$" evolutionary parameters this defines an *homogeneous* "$1q+6\gamma+4\mu$" GDD model where all γ's are independent from the self-interacting or non-self-interacting nature of each protein partner.

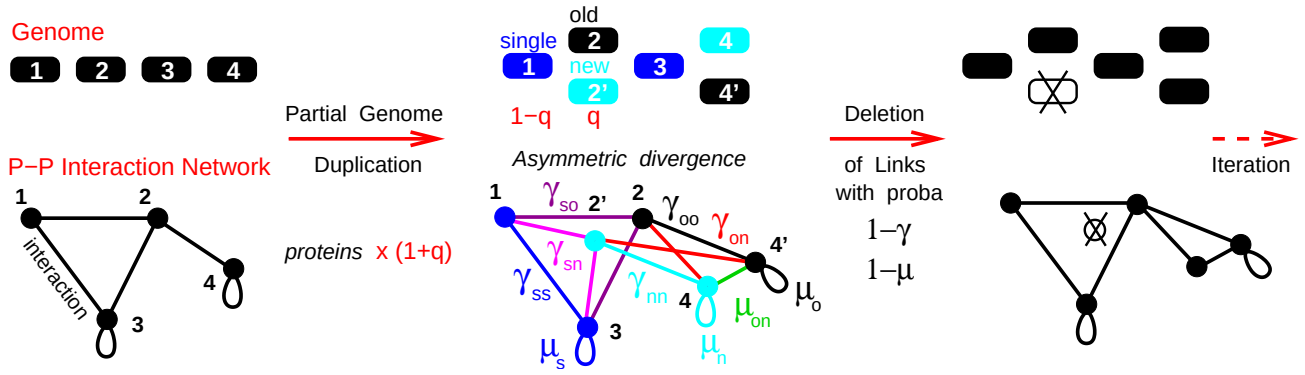

Figure S3: **GDD model of PPI network evolution including *self-link* interactions** ("$1q+6\gamma+4\mu$" parameters).

From a theoretical perspective, we now have to solve two coupled functional recurrences for the generating functions, $F_\ell^{(n)}(x)$ and $F_{n\ell}^{(n)}(x)$, respectively, with and without self-link loops. The global generation function including all network nodes is then simply $F^{(n)}(x) = F_\ell^{(n)}(x) + F_{n\ell}^{(n)}(x)$.

Hence, we readily obtain, using the same notation as before, in addition to the self-link "μ" parameters,
• Generating function WGD recurrence for the self-link loops, $F_\ell^{(n)}(x)$:

$$F_\ell^{(n+1)}(x) = (1-q)\mu_s F_\ell^{(n)}\big(A_s(x)\big) + q\big(\mu_{on}x + 1 - \mu_{on}\big)\Big[\mu_o F_\ell^{(n)}\big(A_o(x)\big) + \mu_n F_\ell^{(n)}\big(A_n(x)\big)\Big], \tag{S29}$$

• Generating function WGD recurrence without self-link loops, $F_{n\ell}^{(n)}(x)$:

$$\begin{aligned} F_{n\ell}^{(n+1)}(x) \;=\; & (1-q)\Big[F_{n\ell}^{(n)}\big(A_s(x)\big) + (1-\mu_s)F_\ell^{(n)}\big(A_s(x)\big)\Big] + q\Big[F_{n\ell}^{(n)}\big(A_o(x)\big) + F_{n\ell}^{(n)}\big(A_n(x)\big)\Big] \\ & + \; q\big(\mu_{on}x + 1 - \mu_{on}\big)\Big[(1-\mu_o)F_\ell^{(n)}\big(A_o(x)\big) + (1-\mu_n)F_\ell^{(n)}\big(A_n(x)\big)\Big] \end{aligned} \tag{S30}$$

• And the global generating function including all network nodes, $F^{(n)}(x) = F_\ell^{(n)}(x) + F_{n\ell}^{(n)}(x)$:

$$F^{(n+1)}(x) = (1-q)F^{(n)}\big(A_s(x)\big) + q\Big[F^{(n)}\big(A_o(x)\big) + F^{(n)}\big(A_n(x)\big)\Big] + q\mu_{on}(x-1)\Big[F_\ell^{(n)}\big(A_o(x)\big) + F_\ell^{(n)}\big(A_n(x)\big)\Big] \tag{S31}$$

Note, in particular, that,

- *i)* the evolution of self-link loops, $F_\ell^{(n)}(x)$, is not coupled to non-self-interacting proteins, $F_{n\ell}^{(n)}(x)$, while the global network evolution, $F^{(n)}(x)$, is formally different from Eq.(2) if and only if $\mu_{on} \neq 0$.

- *ii)* the existence of self-link loops in the PPI network does *not* affect the arguments of *any* generating functions, leading instead to self-link-dependent prefactors in all three generating function recurrences. This implies that the leading term of successive derivatives at $x = 1$ of these generating functions involve successive powers Γ_i^k, $k \geq 1$ where $\Gamma_i = \partial_x A_i|_{x=1}$, for $i = s, o, n$, as before.

Hence, applying the same asymptotic method approach as described in the main text readily yields the following asymptotic regimes for *i)* self-interacting proteins and *ii)* global PPI network,

• *i)* we always have $\Delta_\ell = (1-q)\mu_s + q(\mu_o + \mu_n)$, for the exponential growth rate of the number of self-link loops, and for scale-free regimes, $M' = \max_i(\Gamma_i) > 1$, $\Delta_\ell = (1-q)\mu_s + q(\mu_o + \mu_n) = (1-q)\mu_s\Gamma_s^{\alpha_\ell} + q[\mu_o\Gamma_o^{\alpha_\ell} + \mu_n\Gamma_n^{\alpha_\ell}]$, which defines the power law exponent, α_ℓ, for the limit degree distribution of self-interacting proteins, $p_{\ell k} \propto k^{-\alpha_\ell - 1}$, $k \gg 1$.

• *ii)* there are two cases for the global network topology in linear regimes, which we only consider here (*i.e.* same network growth rates in terms of node or link numbers, see main text):

1- either $\Delta_\ell = (1-q)\mu_s + q(\mu_o+\mu_n) < (1-q)\Gamma_s + q[\Gamma_o+\Gamma_n] = \Delta = (1-q)\Gamma_s^\alpha + q[\Gamma_o^\alpha+\Gamma_n^\alpha]$, then the network growth rate Δ is dominated by non-self-interacting proteins, which implies a negligeable effect of self-link loops and no changes from the paper conclusions, in particular $\alpha = \alpha_{n\ell}$, corresponding to the scale-free exponent without self-link loops defined as $(1-q)\Gamma_s + q[\Gamma_o+\Gamma_n] = (1-q)\Gamma_s^{\alpha_{n\ell}} + q[\Gamma_o^{\alpha_{n\ell}}+\Gamma_n^{\alpha_{n\ell}}]$ as in main text.

2- or $\Delta_\ell = (1-q)\mu_s + q(\mu_o+\mu_n) = \Delta = (1-q)\Gamma_s^\alpha + q[\Gamma_o^\alpha+\Gamma_n^\alpha] > (1-q)\Gamma_s + q[\Gamma_o+\Gamma_n] = (1-q)\Gamma_s^{\alpha_{n\ell}} + q[\Gamma_o^{\alpha_{n\ell}}+\Gamma_n^{\alpha_{n\ell}}]$, and the network growth rate Δ is then dominated by self-interacting proteins, which implies some non negligeable effects of self-link loops but actually *no changes* from the paper main conclusions on network conservation and topological regimes, except for the precise value of the power law exponent α in scale-free regimes, which increases from $\alpha = \alpha_{n\ell}$ to $\alpha_{n\ell} < \alpha < \alpha_\ell$. Note, however, that self-interacting proteins exhibit a larger connectivity exponent α_ℓ than the global PPI network, $\alpha < \alpha_\ell$.

Hence, overall, the general conservation and topological properties of PPI networks is actually little affected by the presence of self-link loops, in the asymptotic limits of large PPI networks and large node degrees. As can be seen from the above argument, this is because conservation and topological properties of PPI networks are controlled by the *exponential* increase of their node degrees, $k \to k\Gamma_i$, for $i = s, o, n$, while the contribution of *de novo* interactions arising from duplicated self-interacting proteins can at most lead to a *linear* increase of node degrees, with a maximum increment of +1 link per duplication event and protein. Thus, although an abundance of self-interacting proteins can significantly affect the evolution of low connectivity proteins, it cannot lead to a change of topological regimes for the highly connected nodes of the PPI networks (*e.g.* from exponential to scale-free node degree distribution or vice versa). Hence, to a first approximation, self-interacting proteins can be simply ignored to establish the asymptotic conservation and topology regimes of PPI network evolution, as we have done in the main text

Note, however, that self-link loops might still be important for the evolution of certain network motifs whose initial emergence might precisely depend on the presence of self-interacting proteins (*e.g.* the triangle motif unless one triangle at least is already present in the initial network).

7 General duplication-divergence hybrid models

We start the analysis of GDD hybrid models with the case of *two* duplication-divergence steps involving some fractions q_1 and q_2 of duplicated genes, introducing explicit dependencies in q and x for $A_i(q,x)$ and $\Gamma_i(q) = \partial_x A_i(q,1)$ functions $(i = s, o, n)$, $A_i(q,x) = (1-q)(\gamma_{is}x + \delta_{is}) + q(\gamma_{io}x + \delta_{io})(\gamma_{in}x + \delta_{in})$ and $\Gamma_i(q) = (1-q)\gamma_{is} + q(\gamma_{io} + \gamma_{in})$.

An evolutionary recurrence for the hybrid generating function can be found by introducing the intermediate step explicitly, $\tilde{p}^{(n)} \to \tilde{r}^{(n)} \to \tilde{p}^{(n+1)}$ where,

$$\tilde{r}^{(n)}(x) = \frac{(1-q_1)\tilde{p}^{(n)}\big(A_s(q_1,x)\big) + q_1\,\tilde{p}^{(n)}\big(A_o(q_1,x)\big) + q_1\,\tilde{p}^{(n)}\big(A_n(q_1,x)\big)}{\Delta_1^{(n)}}$$
$$\Delta_1^{(n)} = -(1-q_1)\tilde{p}^{(n)}\big(A_s(q_1,0)\big) - q_1\,\tilde{p}^{(n)}\big(A_o(q_1,0)\big) - q_1\,\tilde{p}^{(n)}\big(A_n(q_1,0)\big) > 0,$$

and then $\tilde{r}^{(n)} \to \tilde{p}^{(n+1)}$ with,

$$\tilde{p}^{(n+1)}(x) = \frac{(1-q_2)\tilde{r}^{(n)}\big(A_s(q_2,x)\big) + q_2\,\tilde{r}^{(n)}\big(A_o(q_2,x)\big) + q_2\,\tilde{r}^{(n)}\big(A_n(q_2,x)\big)}{\Delta_2^{(n)}}$$
$$\Delta_2^{(n)} = -(1-q_2)\tilde{r}^{(n)}\big(A_s(q_2,0)\big) - q_2\,\tilde{r}^{(n)}\big(A_o(q_2,0)\big) - q_2\,\tilde{r}^{(n)}\big(A_n(q_2,0)\big) > 0,$$

which finally yields for the effective $\tilde{p}^{(n)} \to \tilde{p}^{(n+1)}$ step,

$$\begin{aligned}
\tilde{p}^{(n+1)}(x) \;=\; & \frac{(1-q_2)\Big((1-q_1)\tilde{p}^{(n)}\big(A_s(q_1,A_s(q_2,x))\big) + q_1\tilde{p}^{(n)}\big(A_o(q_1,A_s(q_2,x))\big) + q_1\tilde{p}^{(n)}\big(A_n(q_1,A_s(q_2,x))\big)\Big)}{\Delta_1^{(n)}\Delta_2^{(n)}} \\
& + \frac{q_2\Big((1-q_1)\tilde{p}^{(n)}\big(A_s(q_1,A_o(q_2,x))\big) + q_1\tilde{p}^{(n)}\big(A_o(q_1,A_o(q_2,x))\big) + q_1\tilde{p}^{(n)}\big(A_n(q_1,A_o(q_2,x))\big)\Big)}{\Delta_1^{(n)}\Delta_2^{(n)}} \\
& + \frac{q_2\Big((1-q_1)\tilde{p}^{(n)}\big(A_s(q_1,A_n(q_2,x))\big) + q_1\tilde{p}^{(n)}\big(A_o(q_1,A_n(q_2,x))\big) + q_1\tilde{p}^{(n)}\big(A_n(q_1,A_n(q_2,x))\big)\Big)}{\Delta_1^{(n)}\Delta_2^{(n)}}
\end{aligned}$$

Expressing successive derivatives at $x=1$, $\partial_x^k p(1)$, for $k\geq 2$ in the asymptotic limit $p^{(n)}(x)\to p(x)$ and $\Delta_1^{(n)}\Delta_2^{(n)}\to\Delta^2$ for $n\to\infty$, yields, $\partial_x^k p(1)=(1-q_2)(1-q_1)\Gamma_s^k(q_2)\Gamma_s^k(q_1)\partial_x^k p(1)+(1-q_2)q_1\Gamma_s^k(q_2)\Gamma_o^k(q_1)\partial_x^k p(1)+\cdots$ and hence,

$$\partial_x^k p(1)\Big[1-\frac{\Big((1-q_1)\Gamma_s^k(q_1)+q_1\Gamma_o^k(q_1)+q_1\Gamma_n^k(q_1)\Big)\Big((1-q_2)\Gamma_s^k(q_2)+q_2\Gamma_o^k(q_2)+q_2\Gamma_n^k(q_2)\Big)}{\Delta^2}\Big]=\sum_{l=[k/2]}^{k}\alpha_{k,l}\,\partial_x^l p(1), \quad \text{(S32)}$$

In fact, this simple duplication-divergence combination can be generalized to *any* duplication-divergence hybrid models with arbitrary series of the 1+6 microscopic parameters $\{q^{(n)},\gamma_{ij}^{(n)}\}_R\in[0,1]$, for $i,j=s,o,n$ and $1\leq n\leq R$. Each duplication-divergence step then corresponds to a different linear operator $\mathcal{M}^{(n)}$ defined by $q^{(n)}$ and the functional arguments $A_i^{(n)}(q^{(n)},x)=(1-q^{(n)})(\gamma_{is}^{(n)}x+\delta_{is}^{(n)})+q^{(n)}(\gamma_{io}^{(n)}x+\delta_{io}^{(n)})(\gamma_{in}^{(n)}x+\delta_{in}^{(n)})$ and $\Gamma_i^{(n)}=\partial_x A_i^{(n)}(q^{(n)},1)$ for $i=s,o,n$ (with $A_i^{(n)}(q^{(n)},1)=1$). Hence, applying the same reasoning as in *Asymptotic methods* to the series of linear operators $\{\mathcal{M}^{(n)}\}_R$ implies that *any* duplication-divergence hybrid model converges in the asymptotic limit (at least for simple initial conditions).

In the following, we first assume that the evolutionary dynamics remains cyclic with a finite period R, before discussing at the end the $R\to\infty$ limit, which can ultimately include intrinsic stochastic fluctuations of the microscopic parameters.

In the cyclic case with a finite period R, successive derivatives at $x=1$, $\partial_x^k p(1)$, can be expressed in the asymptotic limit, $p^{(n)}(x)\to p(x)$ as,

$$\partial_x^k p(1)\left(1-\frac{\prod_n^R\Big[(1-q^{(n)})\Gamma_s^{(n)\,k}+q^{(n)}\Gamma_o^{(n)\,k}+q^{(n)}\Gamma_n^{(n)\,k}\Big]}{\Delta^R}\right)=\sum_{l=[k/2]}^{k}\alpha_{k,l}\,\partial_x^l p(1), \quad \text{(S33)}$$

Network conservation for such general duplication-divergence hybrid model corresponds to the condition $M>1$, where the *conservation index* M now reads

$$M=\left(\prod_n^R\Big[(1-q^{(n)})\Gamma_s^{(n)}+q^{(n)}\Gamma_o^{(n)}\Big]\right)^{1/R} \quad \text{(S34)}$$

while the nature of the asymptotic degree distribution is controlled by the network *topology index*,

$$M'=\left(\prod_n^R\max_i\big(\Gamma_i^{(n)}\big)\right)^{1/R} \quad \text{(S35)}$$

with $M'<1$ corresponding to exponential networks and $M'>1$ to scale-free (or dense) networks with an effective node degree exponent α and effective node growth rate Δ that are self-consistent solutions of the generalized characteristic equation,

$$h(\alpha)=\left(\prod_n^R h^{(n)}\big(\alpha,q^{(n)}\big)\right)^{1/R}=\Delta \quad \text{(S36)}$$

The resolution of this generalized characteristic equation is done following *exactly* the same discussion for singular regimes as with constant q and Γ_i (Fig. 3 and main text) due to the convexity of the generalized $h(\alpha)$ function, $\partial_\alpha^2 h(\alpha)\geq 0$. Indeed, the first two derivatives of $h(\alpha)$ yield (with implicit dependency in successive duplication-divergence steps, $q\equiv q^{(n)}$, $\Gamma_i\equiv\Gamma_i^{(n)}$, etc, for $i=s,o,n$),

$$\begin{aligned}
\partial_\alpha h(\alpha) &= \left(\prod_n^R h(\alpha,q)\right)^{1/R}\frac{1}{R}\sum_n^R\frac{\partial_\alpha h(\alpha,q)}{h(\alpha,q)}\\
&= \left(\prod_n^R h(\alpha,q)\right)^{1/R}\frac{1}{R}\sum_n^R\frac{(1-q)\Gamma_s^\alpha\ln\Gamma_s+q\Gamma_o^\alpha\ln\Gamma_o+q\Gamma_n^\alpha\ln\Gamma_n}{(1-q)\Gamma_s^\alpha+q\Gamma_o^\alpha+q\Gamma_n^\alpha}\\
\partial_\alpha^2 h(\alpha) &= \left[\frac{1}{R}\sum_n^R\frac{(1-q)q\Gamma_s^\alpha\Gamma_o^\alpha\big(\ln\Gamma_s-\ln\Gamma_o\big)^2+(1-q)q\Gamma_s^\alpha\Gamma_n^\alpha\big(\ln\Gamma_s-\ln\Gamma_n\big)^2+q^2\Gamma_o^\alpha\Gamma_n^\alpha\big(\ln\Gamma_o-\ln\Gamma_n\big)^2}{h^2(\alpha,q)}\right.\\
&\quad\left.+\left(\frac{1}{R}\sum_n^R\frac{\partial_\alpha h(\alpha,q)}{h(\alpha,q)}\right)^2\right]\left(\prod_n^R h(\alpha,q)\right)^{1/R} \;\geq\; 0
\end{aligned}$$

Let us now show that the solution of the generalized characteristic equation corresponding to $\alpha=1$ implies $h'(1)=0$, which is an essential condition to prove the existence of scale-free asymptotic regimes with a unique power law exponent, $p_k\propto k^{-\alpha^\star-1}$, with $\alpha^\star>1$ (see main text).

The generalized functional equation defining the limit degree distribution for a GDD hybrid model with an arbitrary sequence of duplications contains a sum over 3^R terms with R times nested functional arguments,

$$\tilde p(x)=\frac{1}{\Delta^R}\sum_I c_I\cdot\tilde p(\underbrace{A_{i_1}^{(1)}(q^{(1)},A_{i_2}^{(2)}(q^{(2)},A_{i_3}^{(3)}(q^{(3)},\dots A_{i_R}^{(R)}(q^{(R)},x))))))}_{R\text{ times}}$$

with all possible $i_j = s, o, n$ for $1 \leq j \leq R$, and a prefactor c_I for $I = \{i_1, \ldots, i_R\}$ equal to a product of $(1 - q^{(j)})$ or $q^{(j)}$ corresponding to each occurence of $A_s^{(j)}(q^{(j)}, \cdots)$ or $A_{o,n}^{(j)}(q^{(j)}, \cdots)$, respectively, within the nested functional argument. Inserting the expansion anzats for $\alpha = 1$ near $x = 1$,

$$\tilde{p}(x) = -a_1(1-x) - a'(1-x)\ln(1-x) + \mathcal{O}\big((1-x)^2 \ln(1-x)\big)$$

in the general functional equation yields the following form for each of the 3^R terms $\tilde{p}(\mathcal{A}_I(x))$ of the functional sum (where $\mathcal{A}_I(x)$ is the nested functional argument),

$$\begin{aligned}\tilde{p}(\mathcal{A}_I(x)) \quad \to \quad & -a_1(1-\mathcal{A}_I(x)) - a'(1-\mathcal{A}_I(x))\ln(1-\mathcal{A}_I(x)) = \\ & -a_1\mathcal{A}'_I(1)(1-x) - a'\mathcal{A}'_I(1)\ln\mathcal{A}'_I(1)(1-x) - a'\mathcal{A}'_I(1)(1-x)\ln(1-x) + \mathcal{O}\big((1-x)^2\ln(1-x)\big).\end{aligned}$$

where,

$$\mathcal{A}'_I(1) = \prod_n^R \partial_x A_{i_n}^{(n)}(q^{(n)}, 1) = \prod_n^R \Gamma_{i_n}^{(n)}, \quad \text{and} \qquad \sum_I c_I \mathcal{A}'_I(1) = [h(1)]^R$$

Hence, after collecting all 3^R terms together we get for the functional equation,

$$\tilde{p}(x) = -a'(1-x)\frac{1}{\Delta^R}\sum_I c_I \mathcal{A}'_I(1)\ln\mathcal{A}'_I(1) - a_1\Big(\frac{h(1)}{\Delta}\Big)^R(1-x) - a'\Big(\frac{h(1)}{\Delta}\Big)^R(1-x)\ln(1-x).$$

As the solution $\alpha = 1$ implies $\Delta = h(1)$, the last two terms on the right side of the functional equation correspond exactly to the expansion anzats of $\tilde{p}(x)$ near $x = 1$ for $\alpha = 1$, implying that the first term must vanish (with $a' \neq 0$). This imposes the supplementary condition,

$$\sum_I c_I \mathcal{A}'_I(1)\ln\mathcal{A}'_I(1) = 0$$

which is in fact equivalent to $h'(1) = 0$.

Finally, let us discuss the case of infinite, *non-cyclic series* of duplication-divergence events, which can include intrinsic stochastic fluctuations of all microscopic parameters. Formally, analyzing non-cyclic, instead of cyclic, infinite duplication-divergence series implies to exchange the orders for taking the two limits $p^{(n)}(x) \to p(x)$ and $R \to \infty$ (with $1 \leq n \leq R$). Although this cannot be done directly with the present approach, either double limit order should be equivalent, when there is a *unique* asymptotic form independent from the initial conditions (and convergence path). We know from the previous analysis that it is indeed the case for the *linear* evolutionary regimes (with $h(1) = \Delta$) leading to exponential or scale-free asymptotic distributions (with a unique $\alpha^\star \geq 1$). Hence, the main conclusions for biologically relevant regimes of the GDD model are insensitive to stochastic fluctuations of microscopic parameters.

On the other hand, when the asymptotic limit is *not* unique, as might be the case for *non-linear* evolutionary regimes, the order for taking the double limit $p^{(n)}(x) \to p(x)$ and $R \to \infty$ (with $1 \leq n \leq R$) might actually affect the asymptotic limit itself. Still, asymptotic convergence remains granted in both limit order cases (see above) and we do not expect that the general scale-free form of the asymptotic degree distribution radically changes. Moreover, numerical simulations seem in fact to indicate the existence of a unique limit form (at least in some *non-linear* evolutionary regimes) but after extremely slow convergence, see *Numerical simulations* below. Yet, the unicity of the asymptotic form of the GDD model for *general non-linear* evolutionary regimes remains an open question.

Combining local and global duplications

We now outline the predictions for a realistic GDD model combining $R-1 \gg 1$ local duplications ($q \ll 1$) for each whole genome duplication ($q = 1$). This hybrid model of PPI network evolution amounts to a simple extension of the initial GDD model with fixed q.

Network conservation is now controlled by the cummulated product of connectivity growth/decrease rates over one whole genome duplication and $R-1$ local duplications, following the most conserved, "old" duplicate lineage,

$$M = \Big(\Gamma_o(1) \cdot \big[(1-q)\Gamma_s(q) + q\Gamma_o(q)\big]^{R-1}\Big)^{1/R} \tag{S37}$$

where we note the explicit dependence of Γ_i in q ($i = s, o, n$): $\Gamma_i(q) = (1-q)\gamma_{is} + q(\gamma_{io} + \gamma_{in})$. Hence, conserved [resp. non-conserved] networks correspond to $M > 1$ [resp. $M < 1$].

A similar cummulated product also controls the effective node degree exponent α and node growth rate Δ which are self-consistent solutions of the characteristic equation,

$$\Big(h(\alpha, 1) \cdot \big[h(\alpha, q)\big]^{R-1}\Big)^{1/R} = \Delta \tag{S38}$$

where we note the explicit dependence of function $h(\cdot)$ for α and q: $h(\alpha, q) = (1-q)\Gamma_s^\alpha(q) + q\Gamma_o^\alpha(q) + q\Gamma_n^\alpha(q)$ as before.

Hence, the asymptotic degree distribution for the hybrid model is controlled by the averaged *topology index*,

$$M' = \Big(\Gamma_o(1) \cdot \max_i\big(\Gamma_i^{R-1}(q)\big)\Big)^{1/R} \tag{S39}$$

with $M' > 1$ [resp. $M' < 1$] for scale-free (or dense) [resp. exponential] limit degree distribution. In particular, assuming $\Gamma_s(q) \geq \Gamma_o(q)$, we find ${M'}^R = \Gamma_o(1)\Gamma_s^{R-1}(q)$ and thus,

$$\begin{aligned} {M'}^R &= \Gamma_o(1) \cdot \gamma_{ss}^{R-1}\Big(1 + q\big(\frac{\gamma_{so}+\gamma_{sn}}{\gamma_{ss}} - 1\big)\Big)^{R-1} \\ &\simeq \Gamma_o(1)\sqrt{[h(1,q)]^{R-1}} \quad \text{for} \quad \gamma_{ss} = 1, \;\; Rq^2 \ll 1 \end{aligned}$$

The square root dependency in terms of cummulated growth rate by $R-1$ local duplications, $[h(1,q)]^{R-1}$, implies that non-conserved, exponential regimes for whole genome duplications (if $\Gamma_o(1) < 1$) are not easily compensated by local duplications, suggesting that *asymmetric* divergence between duplicates is still required, in practice, to obtain (conserved) scale-free networks. In this case, the asymptotic exponent of the hybrid model α_h lies between those for purely local (α_ℓ) and purely global (α_g) duplications, that are solution of $h(\alpha_\ell, q) = \Delta_\ell$ and $h(\alpha_g, 1) = \Delta_g$, with typical scale-free exponents $\alpha_\ell + 1$, $\alpha_g + 1$ and, hence, $\alpha_h + 1 \in [2, 3]$, for $\overline{k} < \infty$. Analysis of available PPI data is discussed in [1].

8 Non-local properties of GDD Models

The approach, based on generating functions we have developped to study the evolution of the mean degree distribution can also be applied to study the evolution of simple non-local motifs in the networks. Here, we consider two types of motifs: the two-node motif, $N_{k,l}^{(n)}$ (Fig. S4B), that contains information about the correlations of connectivities between nearest neighbors, and the three-node motif, $T_{k,l,m}^{(n)}$ (Fig. S4C), describing connectivity correlations within a triangular motif. Two generating functions can be defined for the average numbers of each one of these simple motifs,

$$H^{(n)}(x,y) = \sum_{k\geq 0,\, l\geq 0} \langle N_{k,l}^{(n)}\rangle x^k y^l, \tag{S40}$$

$$T^{(n)}(x,y,z) = \sum_{k\geq 0,\, l\geq 0,\, m\geq 0} \langle T_{k,l,m}^{(n)}\rangle x^k y^l z^m. \tag{S41}$$

By construction these functions are symmetric with respect to circular permutations of their arguments.

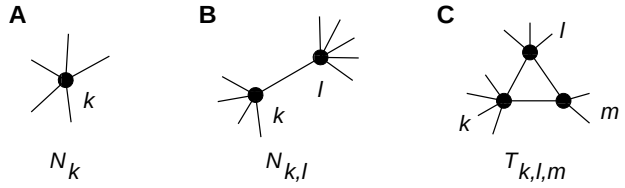

Figure S4: **Simple correlation motifs in PPI networks.**

By definition and symmetry properties of these generating functions, one obtains the mean number of links $\langle L^{(n)}\rangle$ or triangles $\langle T^{(n)}\rangle$, by setting all arguments to one,

$$\begin{aligned} H^{(n)}(x=1,y=1) &= 2\langle L^{(n)}\rangle, \\ T^{(n)}(x=1,y=1,z=1) &= 6\langle T^{(n)}\rangle, \end{aligned}$$

Hence, we can appropriately normalize these generating functions as,

$$h^{(n)}(x,y) = \sum_{k\geq 0,\, l\geq 0} \frac{\langle N_{k,l}^{(n)}\rangle}{2\langle L^{(n)}\rangle} x^k y^l, \tag{S42}$$

$$t^{(n)}(x,y,z) = \sum_{k\geq 0,\, l\geq 0,\, m\geq 0} \frac{\langle T_{k,l,m}^{(n)}\rangle}{6\langle T^{(n)}\rangle} x^k y^l z^m \tag{S43}$$

which yields two rescaled generating functions, varying from zero to one, for the two motif distributions.

Linear recurrence relations can then be written for these generating functions $H^{(n)}$, $T^{(n)}$, $h^{(n)}$ and $t^{(n)}$, using the same approach as for the evolutionary recurrence Eq.(S2) (see Appendix B for details). These relations which contain all information on 2- and 3-node motif correlations, can also be used to deduce simpler and more familiar quantities, such as the average connectivity of neighbors [4,6], $g(k)$, and the clustering coefficient [7,8], $C(k)$.

$g(k)$ is defined on a particular network realization as,

$$g(k) = \sum_{i:d_i=k} \sum_{j\in\langle i\rangle} d_j/kN_k,$$

where d_i denotes the connectivity of node i. This can be expressed in terms of the two-node motif of Fig. S4B and averaged over all trajectories of the stochastic network evolution after n duplications as,

$$g^{(n)}(k) = \left\langle \frac{\sum_{l\geq 0}(l+1)N^{(n)}_{k-1,l}}{kN^{(n)}_k} \right\rangle \simeq \frac{\sum_{l\geq 0}(l+1)\langle N^{(n)}_{k-1,l}\rangle}{k\langle N^{(n)}_k\rangle}, \tag{S44}$$

where the average of ratios can be replaced, in the asymptotic limit $n\to\infty$, by the ratio of averages for *linear* growth regimes, for which fluctuations of $N^{(n)}_k$ do not diverge (see section on *Statistical properties of GDD models*). Note, however, that this requires $N^{(n)}_k \gg 1$ which excludes by definition the few most connected nodes (or "hubs", $k\geq k_h$) for which $\langle N^{(n)}_k\rangle \leq 1$ (See section on *Numerical simulations*, below).

With this asymptotic approximation ($k\leq k_h$), $g^{(n)}(k)$ can then be expressed in terms of $h^{(n)}(x,y)$ and its derivatives,

$$g^{(n)}(k) = \frac{\partial_x^{k-1} h^{(n)}_1(x)|_{x=0}}{\partial_x^{k-1} h^{(n)}_0(x)|_{x=0}} + 1 \tag{S45}$$

where $h^{(n)}_1(x) = \partial_y h^{(n)}(x,y)|_{y=1}$, and $h^{(n)}_0(x) = h^{(n)}(x,y=1)$. Hence, we can reduce the recurrence relation on $h^{(n)}(x,y)$ to two recurrence relations on single variable functions $h^{(n)}_1(x)$ and $h^{(n)}_0(x)$ with,

$$h^{(n)}_0(x) = \left(\overline{k}^{(n)}\right)^{-1} \partial_x p^{(n)}(x)$$

using the mean distribution function defined in Eq.(4).

By construction $g(k)$ reflects correlations between connectivities of neighbor nodes and can actually be related to the conditional probability $p(k'|k)$ to find a node of connectivity k' as a nearest neighbor of a node with connectivity k

$$p(k'|k) = \frac{N_{k-1,k'-1}}{kN_k}, \;\; g(k) = \sum_{k'} p(k'|k)k'.$$

It is important to stress that $g^{(n)}(k)$ defined in this way might be non-stationary even though a stationary degree distribution may exists. Indeed, by definition $g^{(n)}(k)$ satisfies the following normalization condition,

$$\overline{k^2}^{(n)} = \sum_k k^2 p^{(n)}_k = \sum_k k g^{(n)}(k) p^{(n)}_k. \tag{S46}$$

which implies that $g^{(n)}(k)$ should diverge whenever $\overline{k^2}^{(n)}$ does so (and $\overline{k}^{(n)} \to \overline{k} < \infty$). This is in particular the case for actual PPI networks with scale-free degree distribution $p_k \propto k^{-\alpha-1}$ with $2<\alpha+1\leq 3$. When comparing actual PPI network data with GDD models (as discussed in ref. [1]), we have found that such divergence can be appropriately rescaled by the factor $\overline{k}^{(n)}/\overline{k^2}^{(n)}$, which yields quasi-stationary rescaled distributions $\overline{k}^{(n)} g^{(n)}(k)/\overline{k^2}^{(n)}$ (see *Numerical Simulations*).

The clustering coefficient, $C(k)$, is traditionally defined as the ratio between the mean number of triangles passing by a node of connectivity k and $k(k-1)/2$, the maximum possible number of triangles around this node. When replacing the mean of ratios by the ratio of means in the asymptotic limit, as above, we can express $C^{(n)}(k)$ as,

$$C^{(n)}(k) = \frac{\sum_{l\geq 0,\; m\geq 0}\langle T^{(n)}_{(k-2,l,m)}\rangle}{k(k-1)\langle N^{(n)}_k\rangle}. \tag{S47}$$

Hence, this distribution is entirely determined by the following two generating functions $p^{(n)}(x)$ and $t^{(n)}_0(x) = t^{(n)}(x,1,1)$

$$C^{(n)}(k) = \frac{6\langle T^{(n)}\rangle}{k(k-1)\langle N^{(n)}\rangle} \frac{\partial_x^{k-2} t^{(n)}_0(x)|_{x=0}}{\partial_x^k p^{(n)}(x)|_{x=0}}, \tag{S48}$$

where $6\langle T^{(n)}\rangle = t^{(n)}_0(1)$. A self-consistent recurrence relation on $t^{(n)}_0(x)$ can be deduce from the general recurrence relation on $T^{(n)}(x,y,z)$. We postpone the detailed analysis of these quantities to futur publications.

9 Numerical simulations – GDD model convergence *versus* Empirical network data

We present in this section some numerical results which illustrate the main predicted regimes of the GDD model. The most direct way to study numerically PPI network evolution according to the GDD model is to simulate the local evolutionary rules on a graph defined, for example, as a collection of links. This kind of simulation gives access to all observables associated with the graph, while requiring a memory space and a number of operations per duplication step roughly proportional to the number of links. On the other hand, if we are interessed in node degree distribution only, a simpler and faster numerical approach can be used: instead of detailing the set of links explicitly, one can solely monitor the information concerning the collection of connectivities of the graph, ignoring correlations between connected nodes. At each duplication-divergence step, a fraction q of nodes from the current node degree distribution is duplicated and yields two duplicate copies ("old" and "new") while the complementary $1-q$ fraction remains "singular". Duplication-derived interactions are then deleted assuming a random distribution of old/new vs singular neighbor nodes with probability q vs $1-q$. The evolution of the connectivity distribution derived in this way corresponds exactly to the evolution of the average degree distribution; even though particular realizations are different, we obtain on average the correct mean degree distribution. This simulation only requires a memory space proportional to the maximum connectivity and a number of operation that is still proportional to the number of links. Since the number of links grows exponentially more rapidly than the maximum connectivity, this numerical approach provides an efficient alternative to perform large numbers of duplications as compared with direct simulations. The numerical results presented below are obtained using either approach and correspond only to a few parameter choices of the GDD model in the whole genome duplication-divergence limit ($q=1$). These examples capture, however, the main features of every network evolution regimes.

From scale-free to dense regimes

We first present results for the most asymmetric whole genome duplication-divergence model [1] $q=1$, $\gamma_{oo}=1$ and $\gamma_{nn}=0$ for four values of the only remaining variable parameter $\gamma_{on}=\gamma=0.1$, 0.26, 0.5 and 0.7, Figs. S5A&B. As summarized on the general phase diagram for $q=1$, Fig. S2B, this model does not present any exponential regime, but a scale-free limit degree distribution $p_k \sim k^{-\alpha^\star-1}$ with a unique $\alpha^\star$ satisfying

$$(1+\gamma)^{\alpha^\star}+\gamma^{\alpha^\star}=1+2\gamma$$

for $\gamma<0.318$, and a non-stationary dense regime for $\gamma>(\sqrt{5}-1)/2\simeq 0.618$, while the intermediate range $0.318<\gamma<0.618$ corresponds to stationary scale-free degree distributions in the non-linear asymptotic regime (*i.e.* $(1+\gamma)^{\alpha}+\gamma^{\alpha}=\Delta\leq 1+2\gamma$) which we would like to investigate numerically in order to determine whether or not it corresponds to a unique pair (α,Δ), see discussion in *Asymptotic methods*.

As can be seen in Fig. S5A, for $\gamma=0.1$ the degree distribution becomes almost stationary with the predicted power law exponent ($\alpha^\star+1\simeq 2.75$) for more than a decade in k and typical PPI network sizes (about 10^4 nodes). Besides, this small value $\gamma=0.1$ appears to be within the most biologically relevant range of GDD parameters to fit the available PPI network data, Fig. S6A (orange curves), with a refined duplication-divergence model of PPI network evolution at the level of protein domains instead of entire proteins, Fig. S6B. Details about this model, which can also take into account *indirect* interactions within protein complexes and protein domain shuffling events, are discussed in ref. [1]. Fig. S5A also shows that the degree distribution and average connectivity of first neighbors are already essentially stationary at small $k\leq 20$ for PPI networks including 10^3-10^4 nodes corresponding to typical sizes of empiral PPI networks. The best one-parameter fit of this PPI network data (p_k and g_k) corresponds to $\gamma=0.26$, see Figs. S5 & S6 (cyan curves) and ref. [1]. On the other hand, numerical node degree distributions are still quite far from convergence at large $k\geq 20$ for $\gamma=0.26$ and even more so in the non-linear regime with $\gamma=0.5$ (green curves), even for very large PPI network sizes $>10^5$ connected nodes. This is related to the very slow global convergence for $\gamma=0.26$ or $\gamma=0.5$ by contrast with $\gamma=0.1$, as discussed below, Fig. S7.

Simulation results for the distributions of average connectivity of first neighbor proteins $g(k)$ [4, 6] are also shown in Fig. S5A. $g(k)$ is in fact normalized as $g(k)\cdot\overline{k}/\overline{k^2}$ to rescale its main divergence [1]. A slow decrease of $g(k)$ is followed by an abrupt fall at a threshold connectivity k_h beyond which nodes (with $k>k_h$) are rare and can be seen as "hubs" in individual graphs of size N (k_h corresponds to $N\cdot p_{k_h}\sim 1$). Degree distributions for large $k>k_h$ are governed by a "hub" statistics which is different, in general, from the predicted asymptotic statistics (although this is not so visible from the node degree distribution curves).

Fig. S5B shows the evolution of the node degree distribution for the same most asymmetric whole genome duplication-divergence model with $\gamma=0.7$, corresponding to the predicted non-stationnary dense regime. As can be seen, the numerical curves obtained for different graph sizes are clearly non-stationary in the regions of small and large k, with local slopes varying considerably with the number of duplications (and mean size). This was obtained using the efficient numerical approach ignoring connectivity correlations (see above), which cannot, however, be used to study the average connectivity of first neighbor proteins $g(k)$ (direct simulations can be performed though up to about $N=10^4$ nodes, as shown in Fig. S5B).

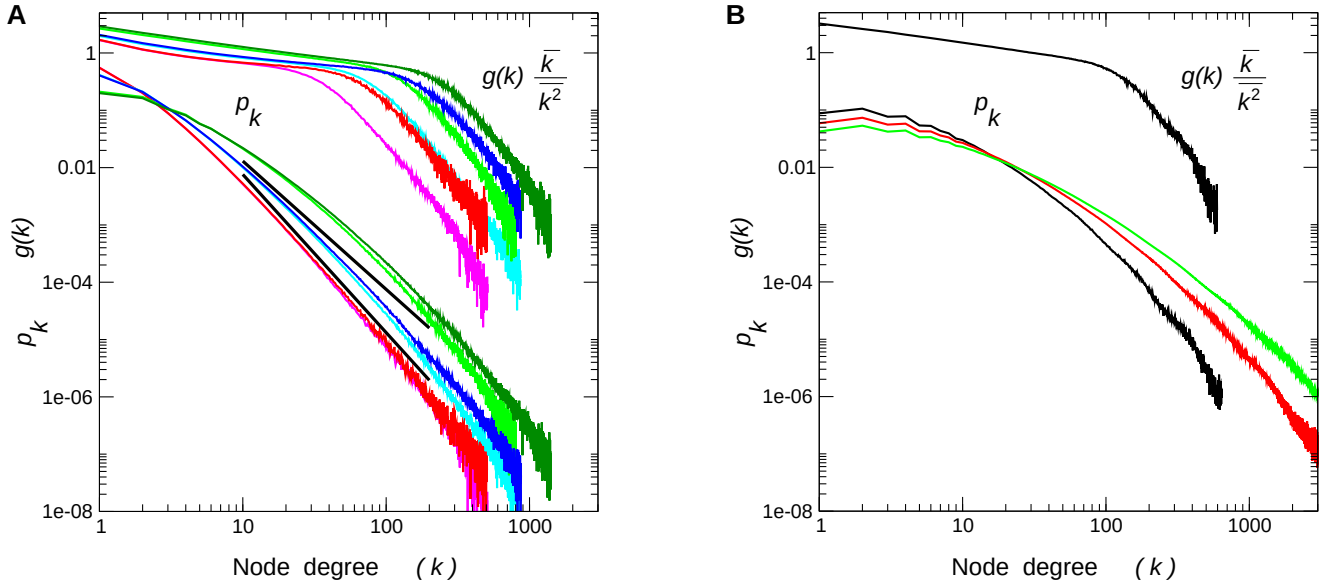

Figure S5: **Simulation results in the whole genome duplication-divergence limit with largest divergence asymmetry** ($q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$, $\gamma_{on} = \gamma = 0.1, 0.26, 0.5, 0.7$) **A.** Distribution p_k and $g(k)$ obtained for $\gamma = 0.1$ with $n = 50$ (magenta, $N = 7 \times 10^3$, $L = 9 \times 10^3$) and $n = 60$ (red, $N = 4 \times 10^4$, $L = 5.3 \times 10^4$); for $\gamma = 0.26$ with $n = 25$ (cyan, $N = 1.7 \times 10^4$, $L = 3.5 \times 10^4$) and $n = 30$ (blue, $N = 1.3 \times 10^5$ $L = 2.9 \times 10^5$); for $\gamma = 0.5$ with $n = 16$ (light green, $N = 1.3 \times 10^4$, $L = 6.4 \times 10^4$) and $n = 18$ (green, $N = 4.6 \times 10^4$, $L = 2.7 \times 10^5$); average curves are obtained for 1000 iterations. **B.** Distribution p_k obtained for $\gamma = 0.7$ with $n = 12$ (black, $N = 4 \times 10^3$, $L = 3.6 \times 10^4$, $g(k)$ is also shown in this case), $n = 16$ (red, $N = 6 \times 10^4$, $L = 1.2 \times 10^6$) and $n = 20$ (green, $N = 9 \times 10^5$, $L = 3.9 \times 10^7$). Distributions are averaged over 2000 iterations.

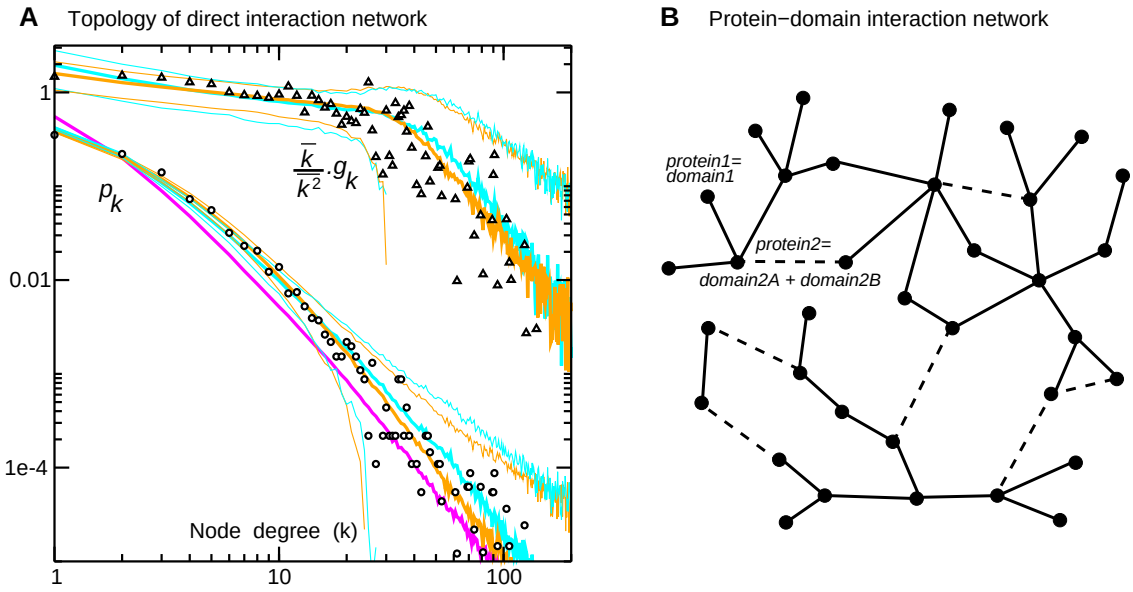

Figure S6: **Comparison between empirical PPI network data and finite-size, duplication-divergence simulations.** **A.** Comparison with protein direct interaction data for Yeast from BIND [9] database (4576 proteins, 9133 physical interactions, $\overline{k} = 3.99$, $\overline{k^2} = 106$). Both connectivity distribution p_k (**open circles**) and average connectivity of first neighbor proteins g_k [6] (**open triangles**) are shown. Data are statistically averaged to account for gaps in connectivities for large $k \geq 20$, due to the finite size of Yeast PPI network. (**cyan curves**) Best one-parameter fit of the data (p_k and g_k) with the previous genome duplication-divergence model with $\gamma = 0.26$, Fig. S5A (cyan curves). Numerical distributions are averaged over 10,000 network realizations (central cyan lines) and averaged distributions plus or minus two standard deviations ($\pm 2\sigma$) are also displayed to show the predicted dispersions (upper and lower cyan lines). (**orange curves**) Best two-parameter fit of the data with a genome duplication-divergence model at the level of protein domains (see **B**) with effective shuffling of protein domains (see ref. [1] for details about this model). The two adjusted parameters ($\gamma = 0.1$ and $\lambda = 0.3$) correspond to a network growth rate of 20% ($1 + 2\gamma$) and an average of 1.5 ($1/(1 - \lambda)$) protein-binding domains per protein, in good agreement with known estimates [1]. (**magenta curve**) The connectivity distribution of the underlying single-domain network (corresponding to $\gamma = 0.1$ and $\lambda = 0.0$ in Fig. S5A, magenta) is also shown to illustrate

the corresponding changes from the single-domain network (magenta curve) to the multi-domain protein network (orange curves with $\lambda = 0.3$). **B.** PPI network model based on domain-domain interactions and multi-domain proteins with $1/(1-\lambda)$ domains per protein.

Finally, we have studied numerically the convergence of the GDD model for these four parameter regimes, $\gamma = 0.1$, 0.26, 0.5 and 0.7. The results are presented in terms of $\Delta^{(n)}$ (Fig. S7A) and its distribution (Fig. S7B) as well as through the node variance $\chi_N{}^{(n)} = \left(\langle N^{(n)2}\rangle - \langle N^{(n)}\rangle^2\right)^{1/2}/\langle N^{(n)}\rangle$ (Fig. S7C). Fig. S7A confirms that the convergence is essentially achieved for $\gamma = 0.1$ while $\gamma = 0.26$, $\gamma = 0.7$ and above all $\gamma = 0.5$ are much further away from their asymptotic limits. For instance, we have $\Delta^{(n)} \simeq 1.86$ for $\gamma = 0.5$ when $\langle N^{(n)}\rangle \simeq 10^7$ nodes, while we know from the main asymptotic analysis detailed earlier that $1.9318 \leq \Delta \leq 2$ in the corresponding asymptotic limit. Yet, it is interesting to observed that these numerical simultations suggest that the asymptotic form for the non-linear regime $\gamma = 0.5$ might still be unique, as convergence appears to be fairly insensitive to topological details of the initial graphs (Fig. S7A) and stochastic dispersions of the evolutionary trajectories: distributions of $\Delta^{(n)}$ become even more narrow with successive duplications (Fig. S7B), while the dispersion in network size given by $\chi_N{}^{(n)}$ is typically smaller for non-linear than linear regimes with a very slow increase for large network size $\langle N^{(n)}\rangle > 10,000$ nodes (Fig. S7C). Still, a formal proof of such a unique asymptotic form (if correct) remains to be established, in general, for *non-linear* asymptotic regimes of the GDD model.

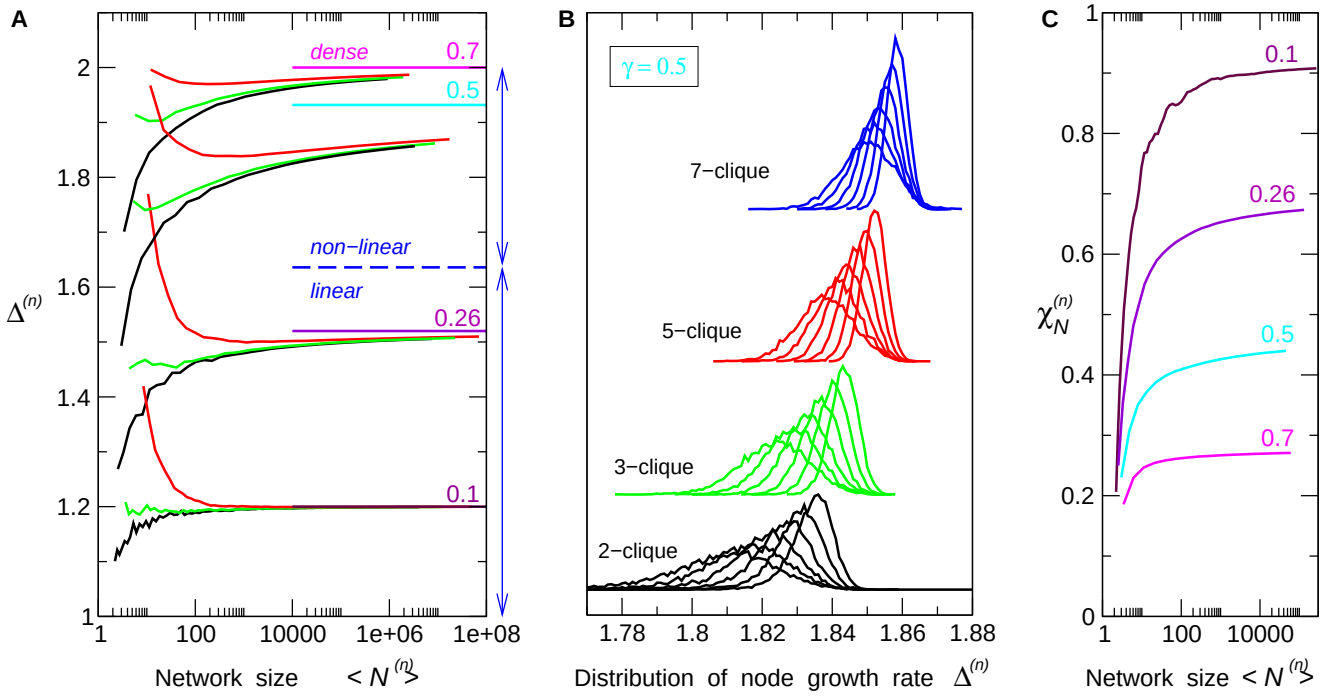

Figure S7: **Asymptotic convergence for the whole genome duplication-divergence limit with largest divergence asymmetry. A.** Asymptotic convergence of $\Delta^{(n)}$ from a simple initial link (black), triangle (green) or 6-clique (red) for the GDD model with $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$ and four values of $\gamma_{on} = \gamma = 0.1$, 0.26, 0.5 and 0.7. The corresponding asymptotic limits, $\Delta = 1.2$, 1.52, [1.9318;2] and 2, as well as the linear to non-linear regime threshold are shown on the right hand side of the plot. **B.** Distribution of $\Delta^{(n)}$ for successive duplications from different initial network topologies in the non-linear regime with $\gamma = 0.5$. **C.** Node variance $\chi_N{}^{(n)} = \left(\langle N^{(n)2}\rangle - \langle N^{(n)}\rangle^2\right)^{1/2}/\langle N^{(n)}\rangle$ for the GDD model with $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$ and four values of $\gamma_{on} = \gamma = 0.1$, 0.26, 0.5 and 0.7 and starting from a simple link (2-clique).

From exponential to dense regimes

An example of GDD model exhibiting an exponential asymptotic degree distribution can be illustrated with a perfectly symmetric whole duplication-divergence model $q = 1$, $\gamma_{oo} = \gamma_{on} = \gamma_{nn} = \gamma \leq 0.5$. The corresponding Fig. S8A shows a good agreement between theoretical prediction and the quasi exponential distribution obtained from simulations with $\gamma = 0.4 \leq 0.5$ (as $\gamma \geq 0.5$ correspond to non-stationary dense regimes, see below).

Finally, the same symmetric whole genome duplication-divergence model exhibits also a peculiar property due to the explicit form of its recurrence relation

$$p^{(n+1)}(x) = p^{(n)}\big((\gamma x + \delta)^2\big)$$

which happens to be precisely of the class of the link probability distribution Eq.(S49) studied in Appendix A. Hence, in the limit of large n the corresponding degree distribution should have a scaling form as defined by Eq.(S50). Indeed, the simulation results depicted in Fig. S8B show that the scaling functions $\overline{k}^{(n)} p_k^{(n)} = w(k/\overline{k}^{(n)})$ plotted for different graph sizes are perfectly close in the asymptotic limit, although the overall evolutionary dynamics is in the non-stationary dense regime, here, with $\gamma = 0.6 \geq 0.5$ (*i.e.* $\overline{k}^{(n)} \to \infty$ and $p_k^{(n)} \to 0$ when $n \to \infty$).

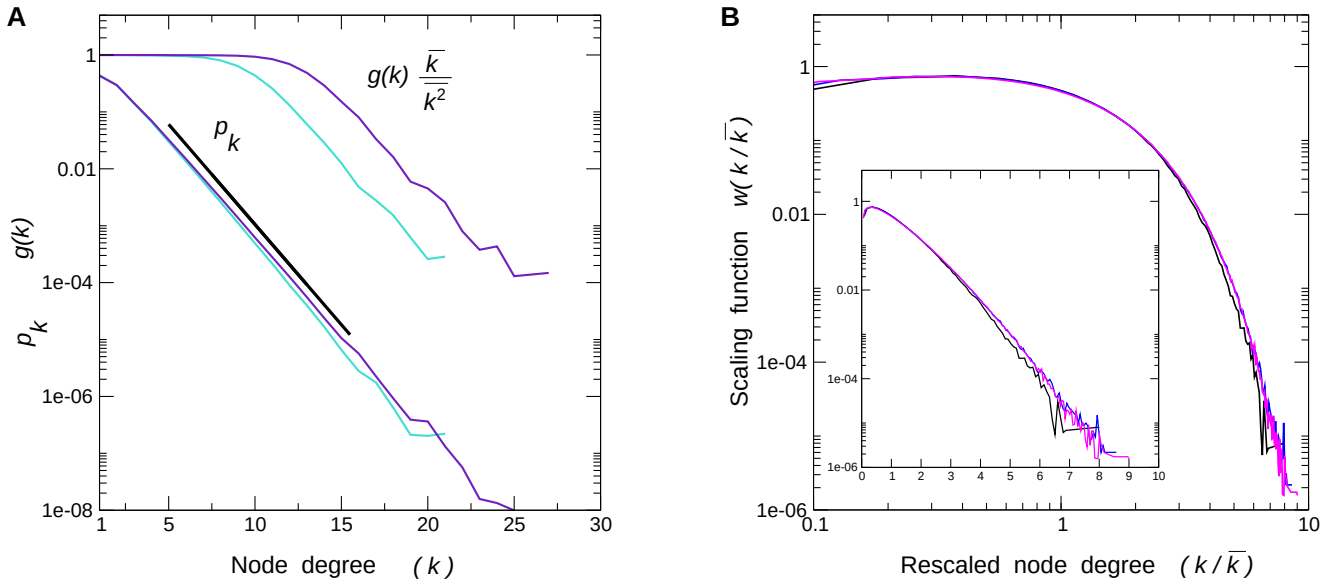

Figure S8: **Simulation results in the whole genome duplication-divergence limit with symmetric gene divergence.** **A.** Distribution p_k obtained for $\gamma = 0.4$ with $n = 15$ (black, $N = 1.2 \times 10^3$, $L = 1.1 \times 10^3$) and $n = 20$ (blue, $N = 1.2 \times 10^4$, $L = 1.2 \times 10^4$); **B.** Scaling function $w\big(k/\overline{k}^{(n)}\big)$ (see text) obtained for $\gamma = 0.6$ with $n = 10$ (black, $N = 1.2 \times 10^3$, $L = 6.3 \times 10^3$), $n = 12$ (blue, $N = 4.7 \times 10^3$, $L = 3.7 \times 10^4$) and $n = 13$ (magenta, $N = 9.3 \times 10^3$, $L = 8.8 \times 10^4$) ; $w\big(k/\overline{k}^{(n)}\big)$ is shown in both log-log and log-lin (inset) representations; average curves are obtained for 1000 iterations.

Appendices

A Scaling for Probability Distributions

Let $p_k^{(n)}$ be a probability distribution whose generating function $P^{(n)}(x) = \sum_k p_k^{(n)} x^k$ satisfies the following recurrence relation

$$P^{(n+1)}(x) = P^{(n)}\big[a(x)\big], \tag{S49}$$

with $a(x)$ a polynome with positive coefficients of degree $m > 1$ with $a(1) = 1$ and $a'(1) > 1$. This probability distribution can be shown to exhibit a scaling property

$$p_k^{(n)} = [a'(1)]^{-n} F\big(k/[a'(1)]^n\big), \quad n \gg 1. \tag{S50}$$

Indeed, we first remark that any polynome of this kind can be decomposed as

$$a(x) = \prod_{i=1}^{m_1} (\delta_i + \gamma_i x) \prod_{j=1}^{m_2} (a_j (x + c_j)^2 + b_j), \quad m_1 + m_2 = m,$$

where the first product collects the real roots of the polynome while the second product corresponds to all pairs of complex conjugate roots. Since all coefficients are positive, γ_i, δ_i, a_j, b_j and c_j are also positive. In addition, we can choose $\gamma_i + \delta_i = 1$ and $a_j(1 + c_j)^2 + b_j = 1$ for all i and j.

Then, the recurrence relation Eq.(S49) is equivalent to

$$p_s^{(n+1)} = \sum_{k=[s/m]}^{D_n} p_k^{(n)} \sum_{l_1=0}^{k} \cdots \sum_{l_{m_1}=0}^{k} \binom{k}{l_1} \cdots \binom{k}{l_{m_1}} \gamma_1^{l_1} \delta^{k-l_1} \cdots \gamma_{l_{m_1}} \delta^{k-l_m} \sum_{h_1=0}^{k} \sum_{s_1=0}^{2h_1} \cdots \sum_{h_{m_2}=0}^{k} \sum_{s_{m_2}=0}^{2h_{m_2}} \binom{k}{h_1} \binom{2h_1}{s_1} \cdots \binom{k}{h_{m_2}} \binom{2h_{m_2}}{s_{m_2}} a_1^{h_1} b_1^{k-h_1} c_1^{2h_1-s_1} \cdots a_{m_2}^{h_{m_2}} b_{m_2}^{k-h_{m_2}} c_{m_2}^{2h_{m_2}-s_{m_2}} \delta\Big(\sum_i l_i + \sum_j s_j - s\Big) \quad \text{(S51)}$$

where $D_n = nmD_0$ is the degree of $P^{(n)}(x)$. In the following, we fix $n \gg 1$ and suppose that the first moment is large $A = [a'(1)]^n \gg 1$, so that we can rescale all the variables as

$$x = s/A,\ y = k/A,\ y_i = l_i/A,\ w_j = h_j/A,\ z_j = s_j/A$$

and finally replace the sums by integrals over rescaled variables. We choose also n to be sufficently large to have $D_n/A \gg 1$. We then apply Stirling formula to get a continuous approximation for binomial coefficients and use the expected scaling form of $p_k^{(n)}$ from Eq.(S50), so that, when replacing sums by integrals in the continious approximation, we obtain,

$$p_s^{(n+1)} = A^{-n} A^{m_1/2+m_2-1} \int_{x/m}^{\infty} dy F(y) \int_0^y \cdots \int_0^y dy_1 \ldots dy_{m_1} \int_0^y dw_1 \int_0^{2w_1} dz_1 \ldots \int_0^y dw_{m_2} \int_0^{2w_{m_2}} dz_{m_2}$$
$$\delta\big(\sum_i y_i + \sum_j z_j - x\big) e^{Af} G(y, \ldots), \quad \text{(S52)}$$

with

$$\begin{aligned} f(y, \{y_i\}, \{w_j\}, \{z_j\}) = & \sum_i \Big(y \ln y - (y-y_i)\ln(y-y_i) - y_i \ln y_i + y_i \ln \gamma_i + (y-y_i)\ln \delta_i \Big) + \\ & \sum_j \Big(y \ln y - (y-w_j)\ln(y-w_j) - w_i \ln w_j + w_i \ln a_j + (y-w_j)\ln b_j + \\ & + 2w_i \ln 2w_j - (2w_j - z_j)\ln(2w_j - z_j) - z_j \ln z_j + (2w_j - z_j)\ln c_j \Big) \end{aligned}$$

and

$$G(y, z_1, \ldots, z_m) = \prod_{i=1}^{m_1} \Big(\frac{y}{2\pi y_i (y-y_i)} \Big)^{1/2} \prod_{j=1}^{m_2} \Big(\frac{2y}{(2\pi)^2 z_j (y-w_j)(2w_j - z_j)} \Big)^{1/2}.$$

Since A is large, we can apply the Laplace method first to the $m_1 + 2m_2$ internal integrals. We have to minimize f with respect to y_i, w_j and z_j given that $\sum_i y_i + \sum_j z_j = x$. This can be performed by the Lagrange multiplier method by looking for the minimum of

$$f(y, \{y_i\}, \{w_j\}, \{z_j\}) - \lambda(\sum_i y_i + \sum_j z_j - x)$$

and setting $\sum_i y_i + \sum_j z_j = x$ for the solution.

In this way we obtain a unique minimum at

$$y_i^0 = \frac{y}{a_i},\ w_j^0 = \frac{y}{h_j},\ z_j^0 = \frac{2y}{h_j g_j},$$

with

$$a_i = 1 + \frac{\delta_i}{\gamma_i} e^{\lambda},\ g_j = 1 + c_j e^{\lambda},\ h_j = 1 + \frac{b_j e^{2\lambda}}{a_j g_j^2}$$

and λ is determined implicitly as a function of x and y from the normalization condition

$$y \sum_i \frac{1}{a_i} + 2y \sum_j \frac{1}{h_j g_j} = x.$$

After some algebra, we find that the values of f in the minimum is given by

$$\begin{aligned} w(y,x) = f(y, \{y_i^0\}, \{w_j^0\}, \{z_j^0\}) = & y \sum_i \Big(-(1-a_i^{-1})\ln(1-a_i^{-1}) - a_i^{-1} \ln a_i^{-1} + a_i^{-1} \ln \gamma_i + (1-a_i^{-1})\ln \delta_i \Big) + \\ & y \sum_j \Big(-(1-h_j^{-1})\ln(1-h_j^{-1}) - h_j^{-1} \ln h_j^{-1} + h_j^{-1} \ln a_j + (1-h_j^{-1})\ln b_j + \\ & -2h_j^{-1}\big[-(1-g_j^{-1})\ln(1-g_j^{-1}) - g_j^{-1} \ln g_j^{-1} + (1-g_j^{-1})\ln c_j \big] \Big) \end{aligned}$$

Therefore we write the leading contribution from the $m_1 + 2m_2$ internal integrals in Eq.(S52) as,

$$e^{Aw(y,x)} g(y,x) A^{-(m_1+2m_2-1)/2}, \tag{S53}$$

with $g(y,x)$ collecting all the contributions of the integrals, while the power of A can just be determined by the number of integrations left after integrating the delta function.

The last integral to calculate in Eq.(S52) is on y

$$A^{-1/2} \int_{x/m}^{\infty} dy H(y,x) F(y) e^{Aw(y,x)}$$

where we have collected all slow varying terms and constants in $H(y,x)$. When applying the Laplace method we calculate the derivative of $w(y,x)$ with respect to y that turns out to have a simple expression

$$\partial_y w(y^0,x) = \sum_i \ln\big(\frac{\delta_i a_i}{a_i - 1}\big) + \sum_j \ln\big(\frac{b_j h_j}{h_j - 1}\big) = \sum_i \ln(\delta_i + \gamma_i e^{-\lambda}) + \sum_j \ln\big(a_j (e^{-\lambda} + c_j)^2 + b_j\big) = 0.$$

The last condition is equivalent to $\prod_i (\delta_i + \gamma_i e^{-\lambda}) \prod_j \big(a_j (e^{-\lambda} + c_j)^2 + b_j\big) = 1$ which has a unique solution $\lambda = 0$, and for the saddle point we get $y^0 = x/(\sum_i \gamma_i + 2\sum_j a_j(1 + c_j)) = x/a'(1)$.

Now it is just a matter of tedious calculations to prove that the prefactor shrinks to $1/a'(1)$ so that

$$p_k^{(n+1)} = [a'(1)]^{-n-1} F\Big(k/[a'(1)]^{n+1}\Big), \;\; [a'(1)]^{n+1} = A \cdot a'(1),$$

as anticipated from the scaling expression Eq.(S50). We were not able to determine the exact shape of the scaling function F which is strongly dependent on the initial probability distribution (an example is shown in Fig. S8B).

B Recurrence relations on $H^{(n)}$ and $T^{(n)}$

In order to relate $H^{(n)}$ and $H^{(n+1)}$ we remark that by partial duplication process one motif (k,l) of type Fig. S4B can generate up to three new motifs of this kind. If the middle link of this motif links two s nodes (probability $(1-q)^2$), the motif itself is kept with the probability γ_{ss} and its external connectivities are modified in the same way as the connectivities in the fundamental evolutionnary recurrence, *i.e.*,

$$x^k y^l \mapsto [A_s(x)]^k [A_s(y)]^l,$$

so that the contribution of ss links to the $H^{(n+1)}$ is given by

$$(1-q)^2 \gamma_{ss} F^{(n)}(A_s(x), A_s(y)).$$

If the middle link of the motif connects one s and one o nodes (proba $q(1-q)$), the link is presented with probability γ_{so}, and we have to substitute

$$x^k y^l \mapsto [A_s(x)]^k [A_o(y)]^l$$

for external links plus one new link sn which gives the factor $(\delta_{sn} + \gamma_{sn} x)$. By itself this link can create a new motif whose consecutive substitution is

$$x^k y^l \mapsto [A_s(x)]^k [A_n(y)]^l.$$

Therefore, the contribution of these two kinds of motifs is

$$q(1-q)\gamma_{so}(\delta_{sn} + \gamma_{sn} x) H^{(n)}(A_s(x), A_o(y)) + q(1-q)\gamma_{sn}(\delta_{so} + \gamma_{so} x) H^{(n)}(A_s(x), A_n(y)),$$

and the contribution from motifs with the middle link os is just obtained through the permutation $x \leftrightarrow y$

$$q(1-q)\gamma_{so}(\delta_{sn} + \gamma_{sn} y) H^{(n)}(A_o(x), A_s(y)) + q(1-q)\gamma_{sn}(\delta_{so} + \gamma_{so} y) H^{(n)}(A_n(x), A_s(y)).$$

Finally, motifs with the middle oo link can create 3 new motifs whose common contribution is obtained the same way as above

$$\begin{aligned} &q^2 \gamma_{oo}(\delta_{on} + \gamma_{on} x)(\delta_{on} + \gamma_{on} y) H^{(n)}(A_o(x), A_o(y)) + q^2 \gamma_{on}(\delta_{oo} + \gamma_{oo} x)(\delta_{nn} + \gamma_{nn} y) H^{(n)}(A_o(x), A_n(y)) + \\ &+ q^2 \gamma_{on}(\delta_{nn} + \gamma_{nn} x)(\delta_{oo} + \gamma_{oo} y) H^{(n)}(A_n(x), A_o(y)) + q^2 \gamma_{nn}(\delta_{on} + \gamma_{on} x)(\delta_{on} + \gamma_{on} y) H^{(n)}(A_n(x), A_n(y)) \end{aligned}$$

By consequence, when collecting all this contributions we get a recurrence relation on the generating function $H^{(n)}$

$$\begin{aligned}
&H^{(n+1)}(x,y) = (1-q)^2\gamma_{ss}F^{(n)}(A_s(x),A_s(y)) + \\
&+q(1-q)\gamma_{so}(\delta_{sn}+\gamma_{sn}x)H^{(n)}(A_s(x),A_o(y)) + q(1-q)\gamma_{sn}H^{(n)}(A_s(x),A_n(y)) + (x\leftrightarrow y) + \\
&+q^2\gamma_{oo}(\delta_{on}+\gamma_{on}x)(\delta_{on}+\gamma_{on}y)H^{(n)}(A_o(x),A_o(y)) + q^2\gamma_{on}(\delta_{oo}+\gamma_{oo}x)(\delta_{nn}+\gamma_{nn}y)H^{(n)}(A_o(x),A_n(y)) + \\
&+q^2\gamma_{on}(\delta_{nn}+\gamma_{nn}x)(\delta_{oo}+\gamma_{oo}y)H^{(n)}(A_n(x),A_o(y)) + q^2\gamma_{nn}(\delta_{on}+\gamma_{on}x)(\delta_{on}+\gamma_{on}y)H^{(n)}(A_n(x),A_n(y))
\end{aligned} \tag{S54}$$

This relation preserves explicitly the symmetry with respect to $x \leftrightarrow y$.

The recurrence relation on $T^{(n)}$ is derived using the same arguments as above. We remark first that a triangle already presented in the graph can generate at most 7 new triangles, or more precisely no new triangle if it has $3s$ nodes, one new triangle if it has $1o/2s$ nodes, up to 3 new triangles for $2o/1s$ nodes, and at most 7 new triangles when it consists of $3o$ nodes. As previously, for external links of the motif we just have to replace x, y or z by the respective functions A_s, A_o or A_n. The contribution of $3s$ triangles is

$$(1-q)^3\gamma_{ss}^3T^{(n)}(A_s(x),A_s(y),A_s(z)),$$

the contribution of $1o/2s$ triangles

$$q(1-q)^2\gamma_{so}^2\gamma_{ss}(\delta_{sn}+\gamma_{sn}y)(\delta_{sn}+\gamma_{sn}z)T^{(n)}(A_o(x),A_s(y),A_s(z)) + \tag{S55}$$

$$+\quad q(1-q)^2\gamma_{sn}^2\gamma_{ss}(\delta_{so}+\gamma_{so}y)(\delta_{so}+\gamma_{so}z)T^{(n)}(A_n(x),A_s(y),A_s(z)) + (x\to y\to z) \tag{S56}$$

where the last term stands for 4 terms obtained by circular permutations of 3 variables. The contribution of $2o/1s$ triangle will contain 4 terms plus 8 terms resulting from circular permutations of variables

$$\begin{aligned}
&q^2(1-q)\gamma_{so}^2\gamma_{oo}(\delta_{sn}+\gamma_{sn}x)^2(\delta_{on}+\gamma_{on}y)(\delta_{on}+\gamma_{on}z)T^{(n)}(A_s(x),A_o(y),A_o(z)) + \\
&q^2(1-q)\gamma_{so}\gamma_{on}\gamma_{sn}(\delta_{sn}+\gamma_{sn}x)(\delta_{so}+\gamma_{so}x)(\delta_{oo}+\gamma_{oo}y)(\delta_{nn}+\gamma_{nn}z)T^{(n)}(A_s(x),A_o(y),A_n(z)) + \\
&q^2(1-q)\gamma_{so}\gamma_{on}\gamma_{sn}(\delta_{sn}+\gamma_{sn}x)(\delta_{so}+\gamma_{so}x)(\delta_{nn}+\gamma_{nn}y)(\delta_{oo}+\gamma_{oo}z)T^{(n)}(A_s(x),A_n(y),A_o(z)) + \\
&q^2(1-q)\gamma_{sn}^2\gamma_{nn}(\delta_{so}+\gamma_{so}x)^2(\delta_{on}+\gamma_{on}y)(\delta_{on}+\gamma_{on}z)T^{(n)}(A_s(x),A_n(y),A_n(z)) + \\
&+(x\to y\to z).
\end{aligned} \tag{S57}$$

The contribution of $3o$ triangles contains 8 terms

$$\begin{aligned}
&q^3\gamma_{oo}^3(\delta_{on}+\gamma_{on}x)^2(\delta_{on}+\gamma_{on}y)^2(\delta_{on}+\gamma_{on}z)^2T^{(n)}(A_o(x),A_o(y),A_o(z)) + \\
&q^3\gamma_{nn}^3(\delta_{on}+\gamma_{on}x)^2(\delta_{on}+\gamma_{on}y)^2(\delta_{on}+\gamma_{on}z)^2T^{(n)}(A_n(x),A_n(y),A_n(z)) + \\
&q^3\gamma_{oo}\gamma_{on}^2(\delta_{nn}+\gamma_{nn}x)^2(\delta_{oo}+\gamma_{oo}y)^2(\delta_{oo}+\gamma_{oo}z)^2T^{(n)}(A_n(x),A_o(y),A_o(z)) + (x\to y\to z) + \\
&q^3\gamma_{nn}\gamma_{on}^2(\delta_{oo}+\gamma_{oo}x)^2(\delta_{nn}+\gamma_{nn}y)^2(\delta_{nn}+\gamma_{nn}z)^2T^{(n)}(A_o(x),A_n(y),A_n(z)) + (x\to y\to z).
\end{aligned}$$

When getting all these contributions together, the full recurrence relation on $T^{(n)}$ is obtained.

The mean number of triangles is evaluated from this relation by setting all variables to one, or directly when applying previous arguments to triangles irrespective of their external connectivities

$$\langle T^{(n+1)}\rangle \quad = \quad \big[(1-q)^3\gamma_{ss}^3 + q(1-q)^2\gamma_{ss}(\gamma_{so}^2+3\gamma_{sn}^2) + \tag{S58}$$

$$+\quad q^2(1-q)(\gamma_{oo}\gamma_{so}^2+3\gamma_{nn}\gamma_{sn}^2+6\gamma_{so}\gamma_{sn}\gamma_{on}) + \tag{S59}$$

$$+\quad q^3(\gamma_{oo}^3+3\gamma_{oo}\gamma_{on}^2+3\gamma_{nn}\gamma_{on}^2+\gamma_{nn}^3\big]\langle T^{(n)}\rangle. \tag{S60}$$

It evidently presents an exponential growth, that is common for many extensive quantities related to the graph dynamics.

References

1. Evlampiev, K. & Isambert, H. (2007)
Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst. Biol.* **1**:49.
2. Ispolatov, I, Krapivsky, P. L, & Yuryev, A. (2005)
Duplication-divergence model of protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys.* **71**, 061911.
3. Flajolet, P & Sedgewick, R. (2006)
Analytic Combinatorics. http://algo.inria.fr/flajolet/Publications/books.html.
4. Pastor-Satorras, R, Vázquez, A, & Vespignani, A. (2001)
Dynamical and Correlation Properties of the Internet. *Phys. Rev. Lett.* **87**, 258701.
5. Ispolatov, I, Yuryev, A, Mazo, I, & Maslov, S. (2005)
Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res.* **33**, 3629–3635.
6. Maslov, S & Sneppen, K. (2002) Specificity and stability in topology of protein networks *Science* **296**, 910.
7. Watts, D. J & Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature* **393**, 440.
8. Strogatz, S. H. (2001) Exploring complex networks. *Nature* **410**, 268.
9. Alfarano, C *et al.* (2005)
The Biomolecular Interaction Network Database and related tools 2005 update *Nucleic Acids Res.* **33**, D418–D424.