Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries

Jithin K. Sreedharan*, Krzysztof Turowski, and Wojciech Szpankowski, *Fellow, IEEE*

Abstract—Graph models often give us a deeper understanding of real-world networks. In the case of biological networks they help in predicting the evolution and history of biomolecule interactions, provided we map properly real networks into the corresponding graph models. In this paper, we show that for biological graph models many of the existing parameter estimation techniques overlook the critical property of graph symmetry (also known formally as graph automorphisms), thus the estimated parameters give statistically insignificant results concerning the observed network. To demonstrate it and to develop accurate estimation procedures, we focus on the biologically inspired duplication-divergence model, and the up-to-date data of protein-protein interactions of seven species including human and yeast. Using exact recurrence relations of some prominent graph statistics, we devise a parameter estimation technique that provides the right order of symmetries and uses phylogenetically old proteins as the choice of seed graph nodes. We also find that our results are consistent with the ones obtained from maximum likelihood estimation (MLE). However, the MLE approach is significantly slower than our methods in practice.

Index Terms—biological networks, protein-protein interaction, parameter estimation, duplication-divergence, random graphs.

1 Introduction

Many biological processes are regulated at the level of interactions between protein molecules. In recent decades, the development of experimental and bioinformatic methods allow researchers to obtain and make publicly available growing amount of data of various biological mechanisms involving different proteins. The whole network of these events that can be found and reconstructed using various techniques is customarily summarized as protein-protein interaction (PPI) networks. The PPI networks are often described as undirected graphs with nodes representing proteins and edges corresponding to protein-protein interactions.

The proliferation of biological data, in turn, encouraged a study of a series of theoretical models to develop a deeper understanding of the evolution and structural properties of the network representation. However, proper fitting of the biological data and development of statistical tests to check the validity of the models are critical to take advantage of the theoretical developments in graph models. The parameter estimation remains as a challenging problem in biological networks like PPIs, mainly because most of the classical estimation procedures were developed for static networks, and not tailored to specific properties of biological data and its underlying dynamics. In this paper, we propose a parameter estimation scheme for biological data with a new perspective of symmetries and recurrence relations, and point out many fallacies in the previous estimation procedures.

In this paper, we assume that the networks evolve according to the following duplication-divergence stochastic graph model.

Duplication-divergence model (DD-model)

There is a wide agreement [1], originally stemming from Ohno's hypothesis on genome growth [2], that the main mechanism driving evolution of PPI networks is the duplication mechanism, in which new proteins appear as copies of some already existing proteins in the network. This is supplemented by a certain amount of divergence of random mutations that lead to some differences between patterns of interaction for the source protein and the duplicated protein.

There are many variations of duplication-divergence models in the literature, although they have not yet been studied or compared systematically. In this work, we use the model suggested by Pastor-Satorras et al. in [3], recommended in several surveys [4], [5] as a good possible theoretical match for PPI networks.

The model of Pastor-Satorras et al. constructs the graph as described below. Let $\mathcal{N}_k(u)$ be the the set of neighbors of vertex u at time k. Given an undirected, simple, seed graph G_{n_0} on n_0 nodes and target number of nodes n, the graph G_{k+1} with $k+1$ nodes evolves from the graph G_k as follows (the subscript k in G_k can also be interpreted as time instant k). First, a new vertex v is added to G_k. Then the following steps are carried out:

- Duplication: Select an node u from G_k uniformly at random. The node v then makes connections to $\mathcal{N}_k(u)$.
- Divergence: Each of the newly made connections from v to $\mathcal{N}_k(u)$ are deleted with probability $1-p$. Furthermore, for all the nodes in G_k except $\mathcal{N}_k(u)$, create

- J. K. Sreedharan and W. Szpankowski are with NSF Center for Science and Information, Purdue University, West Lafayette, IN 47907. E-mail: {jithinks, szpan}@purdue.edu.
- K. Turowski is with the Theoterical Computer Science Department, Jagiellonian University, Krakow, Poland. E-mail: krzysztof.szymon.turowski@gmail.com.

The first two authors contributed equally to this work. Asterisk indicates corresponding author.
This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, NIH Grant 1U01CA198941-01, Polish National Science Centre grant 2018/31/B/ST6/01294.
Manuscript received ??? ??, 20??; revised ??? ??, 20??.

an edge from each of them to v independently with probability $\frac{r}{k}$.

The above process is repeated until the number of nodes in the graph reaches n. We denote the graph G_n generated from the DD-model with parameters p and r, starting from seed graph G_{n_0}, by $G_n \sim \texttt{DD-model}(n, p, r, G_{n_0})$. Note that the above model generalizes pure duplication model when $p = 1, r = 0$ [6], [7]. In some variations of the model (e.g. [8]), the nodes v and u will make a connection independently with a probability q that is much larger than r/k. However, the addition of q does not introduce significant changes in the properties of the graph, therefore we do not consider such a variation in this work.

Motivation and contributions

In this work, we rigorously study the problem of parameter estimation in the duplication-divergence model using PPI datasets of seven species. The following points motivate this work and we present our key results with them.

Symmetries of the graph

One important feature of the networks that was neglected in the previous studies is the distribution of the number of symmetries generated by the fitted models with fine-tuned parameters. The symmetries of a graph H are formally called automorphisms which is defined as the set of all permutations π of the vertex set of H with the property that, for any vertices u and v, we have $\pi(u) \sim \pi(v)$ if and only if $u \sim v$ where $\sim$ represents an edge (i.e., an automorphism is a an adjacency preserving permutation). For the real-world PPI networks it turns out that the number of symmetries is considerably high, which is in stark contradiction with properties of many random graph models. For example, it is known that graphs generated from Erdős-Renyi model [9] and from preferential attachment model [10] are asymmetric with high probability. Therefore they cannot be reasonably justified as underlying generation schemes for PPI networks. We shall also see that the same phenomena may occur for the DD-model with some ranges of the parameters p and r.

Automorphisms are rarely studied in the context of biological networks and graph models. So far there are no theoretical results on automorphisms in the duplication-divergence model except the work in [11] for the limiting case of $r = 0$, where it was discovered that when both $p = 0$ and $p = 1$ the model produces graphs with a significant amount of symmetry.

Our main focus in this work is to take into account the number of automorphisms of the observed network to restrict the parameter search to a more meaningful range. Moreover, we show that most parameters outputted by previous estimation techniques fail to produce graphs having an order of automorphisms close to that of the PPI networks and therefore they are, in this regard, do not fit the DD-model well. We also note that cross-checking with the number of automorphisms of the real-world network forms a null hypothesis test for the model under consideration.

Moreover, there are close relations between automorphism group of a graph and eigenvalues of the associated graph matrices (commonly called spectrum of a graph) [12]. The graph spectrum concisely abstracts many key characteristics of the graph like the number of triangles and other subgraphs, number of walks of a specified length, number of spanning trees etc. If the existing parameter estimation methods fit the given graph into a model that is not in agreement in terms of the number of automorphisms, the spectrum and many characteristics of the fitted graph do not get matched to the given graph.

Graph parameter recurrences

It is widely recognized that the asymptotics of structural properties such as the degree distribution and number of edges of the DD-model are crucial parameters, upon which judgment about the fitness of the model could be made. From the theoretical point of view, the analysis for the DD-model was presented in [13], [14], supporting the case that graphs derived from this model exhibit (under certain assumptions) power-law-like behavior. Moreover, the frequency of appearance of certain graph structures called graphlets (small subgraphs such as triangles, open triangles, etc.) can be viewed as another criterion for model fitting (see [4], [15] and the references therein). The triangles and *wedges* (paths of length 2 or star with two nodes) are particularly crucial as they are directly related to the network clustering coefficient. The high value of this coefficient is recognized in general as a significant characteristic of some biological networks including PPI networks [16], differentiating these networks from those which can be obtained, for example, from Erdős-Renyi or preferential attachment models.

Our approach is based on recreating graph evolution from a single snapshot of the observed network. We apply, for the first time, rigorous analyses to estimate parameters with the recurrence relations of degree and the number of wedges and triangles, recreating the dynamic process of DD-model construction. The advantages of this approach are twofold: first, the use of accurate iterative formulas allow us to achieve more realism and precision for finite graphs, which is in contrast to most of the previous studies that derive parameter estimates exclusively in terms of steady-state behavior. Second, it is not proven whether the steady state for such models even exists and whether the whole random process converges asymptotically (see Section 4 for more details).

Maximum likelihood method

To substantiate the accuracy of our estimation technique, we apply the maximum likelihood method (MLE) with the importance sampling to the parameter inference problem, adapting the work of Wiuf et al. [17] to the DD-model (see Section 5.2 for the implementation and details). It turns out that the results of the MLE method are very similar to those that derived from our estimation method, and the estimated parameters in both the techniques generate data that is consistent with the observed network in terms of the number of automorphisms.

However, the MLE algorithm has much larger computational complexity, $\Theta\left(n^3 \frac{1}{\varepsilon^2}\right)$ compared to $\Theta\left(n \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)$ needed by our approach based on the recurrences (n and ϵ being the number of nodes and required resolution). Therefore, the analysis of graphs using MLE method is significantly slower for networks in practice and its application is impractical for networks exceeding 1000 nodes, which includes most of the real-world PPI networks. Furthermore, in the case of MLE, a large set of parameter values maximizes the likelihood function when the true p value is close to one, thus making it less reliable (see Section 6.1).

Seed graph choice

It is well known that seed graphs play an important role in biological networks, and its improper selection will affect the estimated parameters of the fitted model [5], [18]. In particular, the task of determining the suitable range of parameters for the duplication-divergence model is always done under assumptions concerning the seed graph. In this work, we improve on the existing solutions by choosing the seed graph on the basis of phylogenetic ages of the proteins in the PPI data – the oldest proteins forms the seed graph. Although such a choice of seed graph is completely absent in prior literature, it is a natural pick as the seed graph itself is defined as the network that is comprised of the oldest entities in the given network.

Outline of the paper

In Section 2, we describe the PPI datasets that are considered in this paper. The influence of parameters of the `DD-model` on graph symmetries and the p-value calculation for comparing with the observed graph are given in Section 3. In Section 4, we provide a critique of various deficiencies of the previous approaches to PPI networks parameter estimation, like lack of symmetries and overemphasis on power-law behavior. Section 5 describes our approach based on both automorphisms counting and exact iterative formulas for certain graph statistics. In this section, we also present an MLE algorithm for parameter inference and compare it with our approach in terms of their complexity and practical usage. Section 6 contains numerical results for both synthetic data generated from the `DD-model` and real-world PPI networks. Section 7 reports the conclusions of the paper with a discussion of obtained results and their significance. At the end, in Section 8 and 9, we provide, as a supplement, an implementation of the MLE algorithm and the proof of our main theorem.

Table 1 provides the list of main notations used in the paper.

Notation	Meaning
G_{obs}	Observed real-world network
G_{n_0}	Seed graph (initial graph) with n_0 nodes
G_n	Realization of the `DD-model` with fixed parameters and n number of nodes
p, r	Parameters of the `DD-model`
γ	Power law exponent
$\text{Aut}(G)$	Automorphism group of graph G
$\lvert\text{Aut}(G)\rvert$	Number of automorphisms of graph G
$\mathcal{N}_s(t)$	Set of neighbors of node t at time s
$\deg_s(t)$	Degree of a node t at time s
$D(G_n)$	$n^{-1}\sum_{i=1}^{n}\deg_n(i)$
$D_2(G_n)$	$n^{-1}\sum_{i=1}^{n}\deg_n^2(i)$
$S_2(G_n)$	Number of wedges (stars with two nodes) in G_n
$C_3(G_n)$	Number of triangles in G_n

TABLE 1: List of main notations

2 Datasets

We use protein-protein interaction networks (PPI) to verify the estimation techniques proposed in this paper. The data is collected from the BioGRID, a popular curated biological database of protein-protein interactions. The networks formed from protein-protein interaction data are further cleaned by removing self-interactions (self-loops), multiple interactions (multiple edges), and interspecies (organisms) interactions of proteins. Thus the considered PPI networks only have physical and intra-species interactions. Unlike some of the previous studies that consider only the largest connected component, the `DD-model` we focus in this work incorporates disconnected subgraphs and isolated nodes.

Table 2 shows the different PPI datasets considered in this paper. We have also listed the logarithm of the number of automorphisms in the original graph, obtained using a publicly available program `nauty` [19]. We note here that the PPI dataset is growing as new interactions getting added on every new release of the dataset. Many previous studies were using older and less complete versions of the data, and therefore it is important to repeat the estimation procedures from those studies and compare them to our methods.

2.1 Selection of seed graph G_{n_0}

Previous studies typically assume the seed graph G_{n_0} as the maximal clique (or the largest two cliques) in the graph G_n [4], [5]. Here we consider a novel formulation for the seed graph. We select the seed graph as the graph induced in the PPI networks by the oldest proteins. That is, the proteins in the observed PPI data that are known to have the largest phylogenetic age (taxon age). It is reasonable to expect that the same protein which appeared over different species also appears in their common ancestor. Hence proteins shared across many different, distant species are supposed to be older than others.

More precisely, the age of a protein is based on a family's appearance on a species tree, and it is estimated via protein family databases and ancestral history reconstruction algorithms. We use Princeton Protein Orthology Database (PPOD) [20] along with OrthoMCL [21] and PANTHER [22] for the protein family database and asymmetric Wagner parsimony as the ancestral history reconstruction algorithm. These algorithms can be accessed via ProteinHistorian software [23].

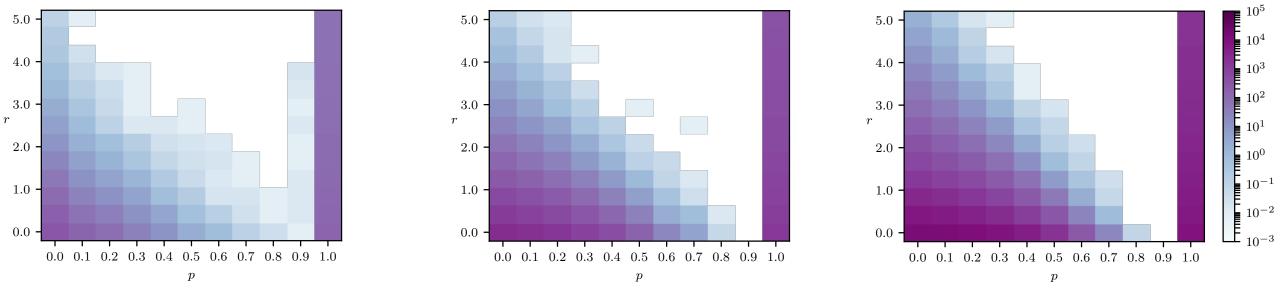
Table 2 also lists the statistics of seed graphs G_{n_0} for different PPI networks. Even if the original PPI network is connected, the `DD-model` under consideration allows a disconnected graph to be the seed graph. Thus, similar to the formation of the PPI network, we consider the graph induced by oldest proteins including isolated nodes and disconnected subgraphs, not restricting ourselves to a connected component that introduces biases in the results.

3 Influence of parameters on symmetries of the model

For certain range of values of the parameters p and r of the DD-model, given n and G_{n_0}, we show in this section that the model generates virtually only asymmetric graphs. However, we can put forward a question: are there any

		Original graph G_{obs}			Seed graph G_{n_0}	
Organism	Scientific name	# Nodes	# Edges	$\log \lvert\text{Aut}(G)\rvert$	# Nodes	# Edges
Baker's yeast	*Saccharomyces cerevisiae*	6,152	531,400	267	548	5,194
Human	*Homo sapiens*	17,295	296,637	3026	546	2,822
Fruitfly	*Drosophila melanogaster*	9,205	60,355	1026	416	1,210
Fission yeast	*Schizosaccharomyces pombe*	4,177	58,084	675	412	226
Mouse-ear cress	*Arabidopsis thaliana Columbia*	9,388	34,885	6696	613	41
Mouse	*Mus musculus*	6,849	18,380	7827	305	7
Worm	*Caenorhabditis elegans*	3,869	7,815	3348	185	15

TABLE 2: Statistics of PPI networks used in this paper and the generated seed graph G_{n_0} with nodes of the largest phylogenetic ages.



(a) $n = 100$ (b) $n = 500$ (c) $n = 2000$

Fig. 1: Logarithm of the expected number of automorphisms of graphs generated from the DD-model. The seed graph $G_{n_0} = K_{20}$.

values of parameters that will yield graphs with the number of automorphisms (i.e., the number of adjacency preserving permutations of the vertex set) close to the real-world PPIs?

In Figure 1, we present the average number of symmetries in the logarithmic scale for graphs with different sizes generated from the `DD-model` with a fixed set of parameters. As $p, r \to 0$ or when p becomes very close to 1 we observe significantly larger values for the average number of automorphisms (since the generated graphs tend to have numerous isolated nodes or they become closer to a complete graph). For instance in Figure 1a, $p = 1, r = 0.4$ has $\mathbb{E}[\log \text{Aut}(G_n)] = 1114$, and $p = 0, r = 0.4$ has $\mathbb{E}[\log \text{Aut}(G_n)] = 1253$. But for large ranges of p and r, it is practically impossible to generate a graph exhibiting any noticeable symmetries. For example, $p = 0.2, r = 2.4$ has $\mathbb{E}[\log \text{Aut}(G_n)] = 3.2$; $p = 0.6, r = 0$ has $\mathbb{E}[\log \text{Aut}(G_n)] = 1.3$; and $p = 0.4, r = 2.4$ has $\mathbb{E}[\log \text{Aut}(G_n)] = 0.12$. These observations are consistent for different n and G_{n_0} too, though the specific range of values of parameters will obviously change.

The number of automorphisms in the `DD-model` behaves differently as in many other graph models. The preferential-attachment graphs are asymmetric (no nontrivial symmetries) with high probability when the number of edges a new node brings into the graph exceeds 2 [10], and almost every graph from the Eŕdos-Rényi model is asymmetric [9]. On the other hand, the `DD-model` exhibits a large number of symmetries and it grows with the number of nodes, as shown in Figure 1.

These findings allow us to argue that only certain subsets of (p, r) pairs correspond to the expected number of automorphisms in the order of the required value. This means that it can be reasonably used as a falsification tool to discard certain parameter ranges and to verify parameter estimation methods. We provide a simple statistical test for checking the possibility of generating the required number of symmetries with the estimated parameters.

Statistical test for significance of the number of symmetries with the estimated parameters

Given the real-world network G_{obs}, seed graph G_{n_0}, and the estimated parameters $(\widehat{p}, \widehat{r})$ of the DD-model, we can estimate the statistical significance of the estimates with respect to the number of symmetries in G_{obs} as follows. Let $G_n^{(1)}, \ldots, G_n^{(m)}$ be m graphs generated from `DD-model`$(n, \widehat{p}, \widehat{r}, G_{n_0})$. Then the p-value is now calculated as follows:

$$p_u = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{\log |\text{Aut}(G_n^{(i)})| \geq \log |\text{Aut}(G_{\text{obs}})|\}$$

$$p_l = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{\log |\text{Aut}(G_n^{(i)})| \leq \log |\text{Aut}(G_{\text{obs}})|\},$$

with $\mathbf{1}\{A\}$ as the indicator function of the event A. Then p-value $= 2\min\{p_u, p_l\}$. As an example, for a fixed parameter set, the empirical distribution of $\log |\text{Aut}(G)|$ is shown in Figure 2. The distribution is symmetrical and this justifies use of the symmetrical definition of p-value. A lower p-value indicates that the estimated parameters do not fit the observed network, and a higher value gives an argument for the estimated parameters being in agreement with the number of symmetries in G_{obs}.

4 Parameter Estimation and Why Existing Methods Fail in Practice?

Previous methods for the parameter estimation problem in the DD-model was first sketched in [3] and then considered more rigorously in [13]. Later, [4], [5] provided some extensions to the estimation procedures using the mean-field approximation of the average degree $D(G_{\text{obs}})$ together with the steady-state expression of the power-law exponent

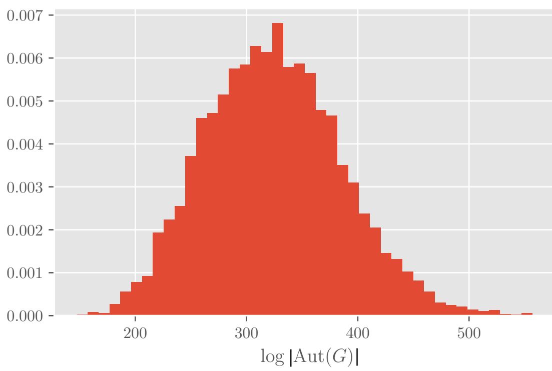


Fig. 2: Normalized histogram of logarithm of number of automorphisms when $G_n \sim \texttt{DD-model}(500, 0.3, 0.4, K_{20})$.

γ of the degree distribution. Then, the values of p and r are computed, respectively, from the formulas:

$$\gamma = 1 + \frac{1}{p} - p^{\gamma-2} \text{ and } r = \left(\frac{1}{2} - p\right) D(G_{\mathrm{obs}}), \text{ for } p < \frac{1}{2}.$$

Table 3 presents the estimates of parameters p and r using the above method. Additionally, we present the average logarithm of the number of automorphisms computed from 10,000 graphs generated from the `DD-model` with the estimated parameters.

Organism	$\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log\lvert\mathrm{Aut}(G_n)\rvert]$	p-value
Baker's yeast	0.28	38.25	0	0
Human	0.43	2.39	10.81	0
Fruitfly	0.44	0.75	3771.99	0
Fission yeast	0.46	1.02	897.48	0
Mouse-ear cress	0.44	0.43	18596.72	0
Mouse	0.48	0.12	34961.69	0
Worm	0.47	0.14	15700.26	0

TABLE 3: Estimated parameters of the `DD-model` and average number of symmetries using mean-field approach.

Mismatch in the number of symmetries and graph statistics

Comparing Tables 2 and 3, we observe that the number of symmetries of the graphs which are generated by the `DD-model` with parameters estimated via the mean-field approach differs significantly with that of the real-world PPI networks. Moreover, the estimated p-values are consistently zero for all the species because the observed values of the parameters under investigation fall far outside the range of the empirical distribution of the parameters for synthetic graphs generated with estimated p and r. This shows that the previously established estimation methods of the `DD-model` fail to capture the critical graph property of automorphisms, and thus do not fit the PPI networks accurately.

As shown in Table 2, the PPI networks exhibit some significant amount of symmetry, but far less than the maximum possible value (equal to $n \log n$), which is attained when every node can be interchanged with every other node. This observation, along with the p-value test in Section 3, allow us to discard not only many models which produce only asymmetric graphs with high probability (such as Erdős-Renyi or preferential attachment model), but also effectively stands as a hypothesis test to verify that the fitting obtained by an estimation procedure can be safely assumed to match the model underlying real-world structures.

Similarly, for certain graph statistics $D(G_n), S_2(G_n)$ and $C_3(G_n)$ (see Table 1 for notation), which are considered later in Section 5.1 for deriving our methods, we observe from Table 4 that the estimated parameters do not yield graphs that have the considered statistics close to the observed graph. Here the p-values are calculated in an equivalent way of number symmetries, just that now it is computed with respect to the graph statistics.

Next, we point out several other deficiencies in the known estimation procedures, which could be the reasons behind such a divergence between the number of symmetries of the PPI networks and its proposed theoretical model.

Power-law behavior

The parameter estimation of the `DD-model` introduced in prior works, such as the one that was presented at the beginning of this section, assumes that the PPI networks are scale-free. This property, stating that the degree distribution of the PPI networks is heavy-tailed or, more precisely, that the number of vertices of degree k is proportional to $k^{-\gamma}$ for some constant $\gamma > 0$ [24], [25]. With this assumption, some (see [8] for example) argue that the estimated value of the exponent for the PPI networks satisfies $2 < \gamma < 3$. However, there are counterarguments to this claim, and it is challenged on statistical grounds that the PPI graphs do not fall into the power-law degree distribution category [26], [27].

To each of the PPI networks in Table 2, we fit the coefficients of power-law distribution with the cut-off following the methodology of Clauset et al. [28]. We note here that cutoff is required in all the cases since the power-law behavior mostly happens in the tails of the degree distribution. However, we find that the cutoff neglects a huge percentage of the data. For example, for a fitting of baker's yeast PPI network, as shown in Figure 3, the cutoff is 582, which is at 94.98 percentile of the degree data, i.e., the power-law fitting does not take into account 94.98% of the data. With the cutoff and the percentiles of all the species listed in Table 5, we remark that any method to estimate the parameters p and r involving power-law exponent do not give reliable approximations since it discards the vast majority of the data.

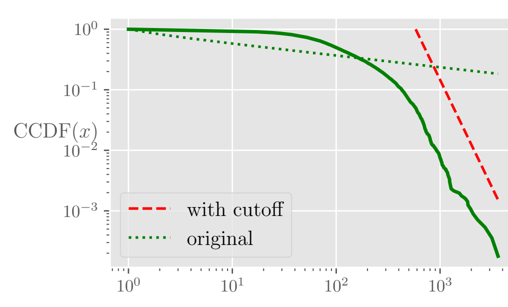


Fig. 3: Complementary cumulative distribution function (CCDF) of baker's yeast and power law fitting.

Steady state assumption

Previous research on the DD-model, both on the level of theoretical analysis of the model properties and the level

Organism	$D(G_{\text{obs}})$	$\mathbb{E}[D(G_n)]$	p-value	$S_2(G_{\text{obs}})$	$\mathbb{E}[S_2(G_n)]$	p-value	$C_3(G_{\text{obs}})$	$\mathbb{E}[C_3(G_n)]$	p-value
Baker's yeast	172.76	115.10	0	220.35M	45.33M	0	9.77M	370.49K	0
Human	34.30	19.39	0	52.25M	7.02M	0	1.07M	105K	0
Fruitfly	13.11	7.87	0	2.94M	1.45M	0	195.96K	77.61K	0
Fission yeast	27.64	6.72	0	7.42M	215.84K	0	223.61K	1.14K	0
Mouse-ear cress	7.39	2.23	0	2.98M	44.46K	0	23.34K	23.27	0
Mouse	5.35	0.82	0	2.95M	9.33K	0	10.22K	0.79	0
Worm	4.04	0.90	0	346.13K	5.32K	0	2.41K	0.49	0

TABLE 4: Comparison of certain graph statistics of the observed graph and that of the synthetic data with parameters estimated via the mean-field approach.

Organism	$\widehat{\gamma}$	Cutoff percentile
Baker's yeast	4.55	94.98
Human	2.85	92.33
Fruitfly	2.71	88.00
Fission yeast	2.43	88.31
Mouse-ear cress	2.68	93.89
Mouse	2.29	78.58
Worm	2.41	88.23

TABLE 5: Estimated power law exponent and required cutoff percentile with the mean-field approach

of parameter estimation of real-world PPI networks, focus heavily on the asymptotic and steady-state behavior [3]. Most of the previous results on the functional form of certain graph statistics in the `DD-model` are under the strong assumption of steady-state [8], [29]. But they do not provide any theoretical proof for convergence to steady-state. Moreover, these steady-state asymptotic results, even when achievable, do not give any bounds on the rate of convergence. This, in turn, raises questions about the straightforward applicability of such theoretical results to parameter estimation.

The previously used methods of parameter estimation also suffer from another issue: for simplicity, they assume that the average degree of the network does not change during the whole evolution from G_{n_0} to G_n. This is not only highly implausible in practice, but also impose direct relation between p and r, and hides any dependency that might be discovered from various properties of the networks.

Seed graph choice

As shown by previous studies (most notably in Hormozidari et al. [5]), choice of the seed graph plays a significant role in graph evolution, directly contributing to the order of growth of many important graph statistics.

The seed graph is typically assumed to be the largest clique (or a connected graph of the largest two cliques) of the observed graph. Then random vertices and edges are gradually added to the network, preserving the average degree of the final network, to make the size of the network to a fixed value of n_0. This method is motivated by the infinitesimally small probability with which there could appear a clique of a greater size during graph evolution. Such a procedure has no formal theoretical guarantees and does not have any clear justification from a biological perspective [5].

Our natural approach to select the seed graph is based on the extra-network information about the estimated age of proteins, described in Section 2.1.

5 Main Results

Our main constructive results concern the relation between the parameters of the model and the number of symmetries exhibited by graphs generated from it. Additionally, we present two parameter estimation algorithms, one based on recurrences characteristic for certain graph statistics, the other based on the well-known maximum likelihood approach.

5.1 Our method: parameter estimation using recurrence relations

Our basic tool to infer the parameters of `DD-model` for a given the PPI network is a set of the exact recurrence relations for basic graph statistics, which relate their values at time k and $k+1$ of graph evolution. Such recurrence relations are sufficient to estimate model parameters, as the whole sequence of graphs from the initial graph G_{n_0} to the final graph G_n can be split into steps consisting of the addition of a single vertex and the changes introduced by the added vertex.

Typically, five statistics of the random graph G_n are studied in literature: number of edges $E(G_n)$, mean degree of the network $D(G_n) = n^{-1}\sum_{i=1}^{n} \deg_n(i)$, mean squared degree $D_2(G_n) = n^{-1}\sum_{i=1}^{n} \deg_n^2(i)$, number of triangles (3-cliques) $C_3(G_n)$, and number of *wedges* $S_2(G_n)$ (wedges are also called 2-stars or paths of length 2 in prior literature, and number of wedges includes counts of triangles and open triangles).

However, for every graph H on n vertices it is true that $E(H) = \frac{n}{2}D(H)$ and $S_2(n) = \frac{n}{2}(D_2(H) - D(H))$. Therefore, it is sufficient to analyze only the three of above-mentioned graph statistics: $D(n)$, $S_2(n)$ and $C_3(n)$.

As a first step, we derive the following recurrence relations for the chosen statistics.

Theorem 1. *If* $G_{n+1} \sim \texttt{DD-model}(n+1, p, r, G_n)$, *then*

$$\mathbb{E}[D(G_{n+1})|G_n] = D(G_n)\left(1 + \frac{2p-1}{n+1} - \frac{2r}{n(n+1)}\right) + \frac{2r}{n+1}$$

$$\begin{aligned}\mathbb{E}[D_2(G_{n+1})|G_n] &= D_2(G_n)\left(1 + \frac{2p+p^2-1}{n+1} - \frac{2r(1+p)}{n(n+1)} + \frac{r^2}{n^2(n+1)}\right) \\ &+ D(G_n)\left(\frac{2p-p^2+2pr+2r}{n+1} - \frac{2r+2r^2}{n(n+1)} + \frac{r^2}{n^2(n+1)}\right) \\ &+ \frac{2r^2+2r}{n+1} - \frac{r^2}{n(n+1)}\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[C_3(G_{n+1})|G_n] \\
&\quad = C_3(G_n)\left(1+\frac{3p^2}{n}-\frac{6pr}{n^2}+\frac{3r^2}{n^3}\right) \\
&\quad + D_2(G_n)\left(\frac{pr}{n}-\frac{r^2}{n^2}\right)+D(G_n)\frac{r^2}{2n} \\
&\mathbb{E}[S_2(G_{n+1})|G_n] \\
&\quad = S_2(G_n)\left(1+\frac{2p+p^2}{n}-\frac{2(p+1)r}{n^2}+\frac{r^2}{n^3}\right) \\
&\quad + D(G_n)\left(pr+p+r-\frac{pr+r+r^2}{n}+\frac{r^2}{n^2}\right)+\frac{r^2}{2}-\frac{r^2}{2n}.
\end{aligned}$$

Proof. See Section 9. □

In Figure 4, we verify Theorem 1 by comparing $\mathbb{E}[D_n]$, for various n, computed using theory and experiments.

The expressions given in Theorem 1 implicitly define a function $\mathbb{E}[D(G_n)|G_{n_0}] = F_D(n,p,r,G_{n_0})$, which is a cornerstone of our algorithm. Similar functions exist for recurrences based on other statistics of G_{n_0} and G_n. Now we claim that the result of Theorem 1 in terms of expectation can be used for the graph statistics with high probability too. Figure 5 shows the concentration of empirical distribution of different graph statistics.

Although we don't need an explicit formula for F_D in our algorithm, we may derive one from the recurrences:

$$\begin{aligned}
\mathbb{E}[D(G_n)|G_{n_0}] &= D(G_{n_0})\prod_{k=n_0}^{n-1}\left(1+\frac{2p-1}{k+1}-\frac{2r}{k(k+1)}\right) \\
&+\sum_{k=n_0}^{n-1}\frac{2r}{k+1}\prod_{l=k+1}^{n-1}\left(1+\frac{2p-1}{l+1}-\frac{2r}{l(l+1)}\right).
\end{aligned}$$

Although this is outside of the scope of this article, we note that such an expression allows us to find, for example, the asymptotic order of growth for $D(G_n)$ and for other statistics.

Though closed form solution of recurrences with G_n and G_{n_0} could be difficult to obtain, Theorem 1 is sufficient to formulate an efficient algorithm for finding the parameters of the model. The crucial feature is that all parameters are monotonic, that is, the larger the parameters p and r, the larger the values of $D(G_n)$ and other statistics.

Algorithm 1 presents our estimation technique for finding $\widehat{p}$ with the recurrence relation for $D(G_n)$ (which will be $D(G_{\text{obs}})$ when we consider real-world network), assuming $\widehat{r}$ is known beforehand. However, sufficient number of samples of $\widehat{r}$ from the interval $[0, n_0]$ is adequate to get a feasible solution set of $\{(\widehat{p},\widehat{r})\}$ with a desired resolution. The algorithm also works for recurrence relations of $S_2(G_n)$ and $C_3(G_n)$ with evident modifications.

We note here that for each graph property under consideration, D, S_2 or C_3, the estimation algorithm returns a curve (more precisely, a set of feasible points). Now, if we find a concurrence in the solutions to the recurrence relations of various graph statistics, we know that a necessary condition for the presence of duplication-divergence model has been satisfied. On the other hand, if the curves were not having a common crossing point, it suggests that the DD-model may not be the appropriate fit for the observed network. We denote the above estimation procedure using the recurrence relations of all three graph statistics as the RECURRENCE-RELATION method.

Algorithm 1 Estimation of p via recurrence relation of $D(G_n)$.

```
1: function RECURRENCE-RELATION(n, r, G_{n_0}, D(G_n), ε)
2:     D_min ← F_D(n, 0, r, G_{n_0}), D_max ← F_D(n, 1, r, G_{n_0})
3:     if D_min > D(G_n) or D_max < D(G_n) then
4:         return "no suitable solution for p"
5:     p_min ← 0, p_max ← 1
6:     while p_max − p_min > ε do
7:         p' ← (p_min + p_max)/2, D' ← F_D(n, p', r, G_{n_0})
8:         if D' < D(G_n) then p_min ← p' else p_max ← p'
9:     return p_min
```

5.2 Parameter estimation via maximum likelihood method

An alternative way of estimating parameters of the DD-model is the maximum likelihood estimation (MLE). With MLE, the estimated parameters $\widehat{p}$ and $\widehat{r}$ are given by $\max_{\theta=(p,r)} L(\theta, G_n)$, where the likelihood function is $L(\theta, G_n)$ is the probability of generating G_n from G_{n_0} for fixed parameters θ, i.e.

$$\begin{aligned}
L(\theta, G_n) &:= \Pr(G_n|G_{n_0};\theta) \\
&= \sum_{G_{n_0+1},\ldots,G_{n-1},G_n\in\mathcal{G}(G_{n_0},G_n)}\ \prod_{k=n_0+1}^{n}\Pr(G_k|G_{k-1};\theta),
\end{aligned}$$

where $\mathcal{G}(G_{n_0}, G_n)$ is the set of all sequences of graphs that starts with G_{n_0} and ends at G_n. Given a fixed sequence of graph evolution history $(G_{n_0},\ldots,G_{n-1},G_n)$, it is straightforward to calculate the likelihood, but $L(\theta, G_n)$ requires summation over all histories, which has exponential number of possibilities. In [17], the authors present an importance sampling strategy to approximate the likelihood and thereby estimate the parameters. It is based on the idea of traversing backwards in history (G_n to G_{n-1} and G_{n-1} to G_{n-2} likewise) on one sample path of graph evolution sequence via Markov chain. We adapt their algorithm to our DD-model and the complete algorithm is presented later in Section 8.

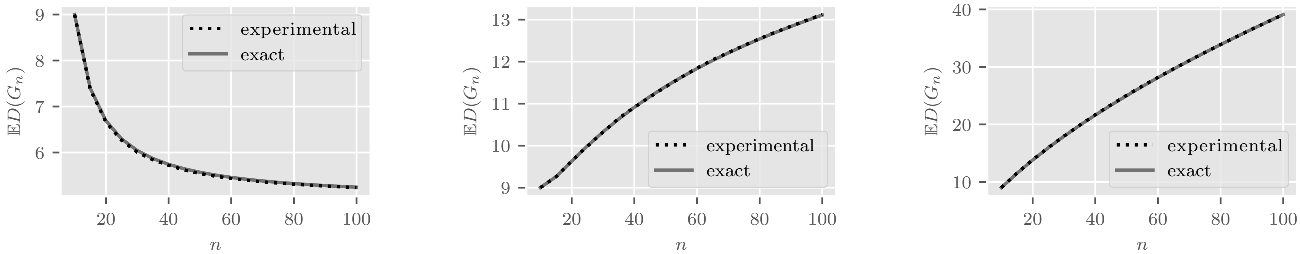
We now provide a brief description of the importance sampling procedure. The idea is to express likelihood in terms of a known reference parameter θ_0 instead of unknown θ. Now, the likelihood can be rewritten as an expectation with respect to θ_0 and can be estimated via Monte Carlo simulations (see [17] for more details):

$$L(\theta, G_n) = \mathbb{E}_{\theta_0}\left[\prod_{k=n_0}^{n} S(\theta_0,\theta,G_k,v)\right],$$

where

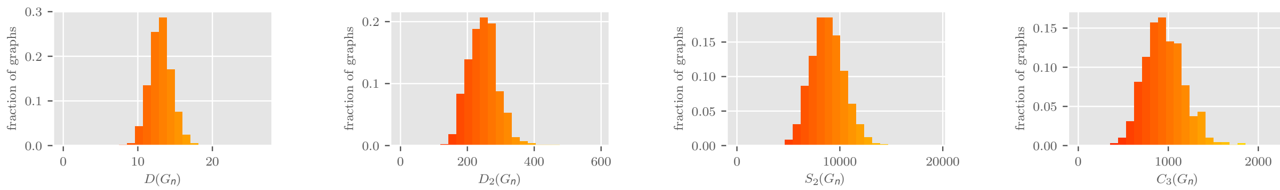
$$S(\theta_0,\theta,G_k,v) = \frac{1}{k}\omega(G_k,\theta,v)\frac{\omega(G_k,\theta_0)}{\omega(G_k,\theta_0,v)}.$$

Here $\omega(G_k,\theta,v)$ is the probability of creating the graph G_k from G_{k-1} through the addition of a node v, with parameter as θ. v can be chosen as any node in G_k such that its removal would result in a positive probability G_{k-1} under the DD-model. The variable $\omega(G_k,\theta)$ is the transition probability $\omega(G_k,\theta,v)$, summed over all possible v. In fact, $\omega(G_k,\theta,v)$ itself is the normalized sum of $\omega(G_k,\theta,v,w)$, over all possible nodes w, which is the probability of producing a graph



(a) $G_n \sim$ `DD-model`$(100, 0.2, 1.5, K_{10})$ (b) $G_n \sim$ `DD-model`$(100, 0.5, 1.5, K_{10})$ (c) $G_n \sim$ `DD-model`$(100, 0.8, 1.5, K_{10})$

Fig. 4: Comparison of $\mathbb{E}[D(G_n)]$ computed via Theorem 1 and via experiments.



(a) Distribution of $D(G_n)$: $CV = 10.57\%$ (b) Distribution of $D_2(G_n)$: $CV = 17.99\%$ (c) Distribution of $S_2(G_n)$: $CV = 18.01\%$ (d) Distribution of $C_3(G_n)$: $CV = 22.90\%$

Fig. 5: Empirical distribution of graph statistics: $G_n \sim$ `DD-model`$(100, 0.5, 1.5, K_{10})$. Coefficient of variation CV is defined as the ratio of empirical standard deviation and empirical mean. The lower values of CV in the sub-figures show the concentration of the considered graph statistics.

G_k from G_{k-1} by adding a node v that is duplicated from node w.

5.3 Computational complexity of parameter estimation methods

Let us now assume that the n_0 is fixed and we are interested in results up to a resolution ε, that is, the values of p and r are stored in such a way that two numbers within a distance less than ε are indistinguishable.

For our algorithm 1, RECURRENCE-RELATION, a single pass requires $\Theta\left(n \log \frac{1}{\varepsilon}\right)$, as it uses a binary search for p and for every intermediate value of p it executes exactly $n - n_0$ steps of *for* loop, each requiring constant time. Now it is sufficient to sample $\frac{n_0}{\varepsilon}$ different values of r, therefore the total running time to find suitable $(\widehat{p}, \widehat{r})$ pairs is $\Theta\left(n \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)$.

On the other hand, the MLE algorithm needs to compute at every step values of the ω function for all possible pairs of v and w for each graph G_k. This is the case because in `DD-model` every vertex v could be a duplicate of every other vertex w always with some non-zero probability at every stage of the algorithm. This means that we require $\Theta(k^2)$ steps at each iteration of the algorithm; therefore $\Theta\left(\sum_{k=n_0}^{n} k^2\right) = \Theta(n^3)$ steps in total. Unfortunately, even clever bookkeeping and amortization is not much of a help here.

Additionally, we need to estimate the likelihood for each pair (p, r) independently, as maximum likelihood function does not have the monotonicity property, so it requires in total $\Theta\left(n^3 \frac{1}{\varepsilon^2}\right)$ steps to find all feasible pairs up to a desired resolution of ϵ.

Moreover, as it was suggested by Wiuf et al. in [17], importance sampling provides good quality results only for at least 1000 independent trials. This adds up a constant factor not visible in the big-Θ notation, but significant in practice, effectively making the algorithm infeasible for the real-world data without using supercomputer power.

6 NUMERICAL EXPERIMENTS

In this section, we evaluate our methods on synthetic graphs and real-world PPI networks.

We made publicly available all the code and data of this project at https://github.com/krzysztof-turowski/duplication-divergence. The code supports random graph models and real-world networks.

Estimation of tolerance interval

We find the tolerance interval of the estimated p and r values for the fitted `DD-model` as follows. For a given network G_{obs} and a seed graph G_{n_0}, first the RECURRENCE-RELATION algorithm outputs a set of solutions $\{(\widehat{p}, \widehat{r})\}$. For each of the feasible pairs, we then estimate the confidence interval of the graph property with which recurrence was calculated. For instance, if the property used was the empirical mean D, graph samples generated from `DD-model`$(n, \widehat{p}, \widehat{r}, G_{n_0})$ are used to estimate expectation $\mathbb{E}[D(G_n)]$ and variance $\mathrm{Var}[D(G_n))]$. A 95% confidence interval of $D(G_n)$ is then given by the values

$$\mathbb{E}[D(G_n)] \pm 1.96 \,\mathrm{Var}[D(G_n))].$$

The Gaussian distribution assumption used in the above expression is indeed a good approximation for the distribution of D for large n. Now by fixing $\widehat{p}$, we can calculate a tolerance interval $(\widehat{r}_{\min}, \widehat{r}_{\max})$ for the estimated parameter $\widehat{r}$. In the following experiments, for demonstrating the above approach, we focus on two graph statistics D and C_3 (parameter estimation will include S_2 too).

Parameter estimation procedure for the experiments

Our parameter estimation procedure can be summarized follows:

1) We employ the RECURRENCE-RELATION algorithm for solving graph recurrences of the three graph statistics D, S_2 and C_3, and we identify a set of solutions for p and r.
2) With $G_n \sim$ `DD-model`$(n, \widehat{p}, \widehat{r}, G_{n_0})$, we find the tolerance interval of $\widehat{r}$ using the confidence interval of $D(G_n)$ and $C_3(G_n)$, as explained in the above subsection.
3) We look for crossing points of the plots in the figure, and the range of values of p and r where the confidence intervals meet around the crossing point. We call such a range of values as *feasible-box*.
4) Though any point in the feasible-box is a good estimate of p and r, to improve the accuracy, we uniformly sample a fixed number of points from the box and choose the pair that gives maximum p-value with respect to the number of automorphisms of the given graph G_{obs}.

The Theorem 1 provides theoretical guarantees (in the expected sense) for the RECURRENCE-RELATION algorithm and the idea of convergence of the three curves, the solution set of D, S_2 and C_3 statistics, in the above estimation procedure. But in practice, we allow some discrepancy in the convergence of the three curves. In the following experiments, we declare that the `DD-model` fits the given dataset when at least two curves converge and there is an intersection among the confidence intervals of the three curves or when one curve and the confidence interval of another curve intersect.

6.1 Synthetic graphs

In this section, we derive preliminary insights by studying our method and maximum likelihood estimation (MLE) empirically on synthetic data. See Sections 5.2 and 8 for the MLE implementation. We note here that the infeasibility of the MLE method for large graphs (e.g. with tens of thousands of vertices) is already explained in Section 5.3 and the comparison, in this section, between our method and the MLE on small graphs is provided only to prove the competitiveness of our approach as to the MLE method.

We generate two random graph samples $G_n^{(1)}$ and $G_n^{(2)}$ from the `DD-model` with the following parameters:

$$G_n^{(1)} \sim \texttt{DD-model}(n = 100, p = 0.1, r = 0.3, G_{n_0} = K_{20}),$$
$$G_n^{(2)} \sim \texttt{DD-model}(n = 100, p = 0.99, r = 3.0, G_{n_0} = K_{20}).$$

The choice of parameters in $G_n^{(1)}$ and $G_n^{(2)}$ show different regimes in the following studies. Moreover the parameters are chosen in such a way that the generated graphs have non-trivial symmetries.

Figures 6a and 6b plot the sets of feasible points identified by the recurrence relations using RECURRENCE-RELATION method. The light shaded bands show the tolerance intervals of r. We observe that the crossing points and the tolerance intervals are fairly close to the original parameters. Figures 6c and 6d display the heat-plot of log-likelihood function of the MLE for different values of the parameters and the maximum value of the log-likelihood in the heat-plot will give the parameters outputted by the MLE. The log-likelihood function maximizes at (p, r) pairs close to the original parameters, but not up to the resolution of RECURRENCE-RELATION method.

In Table 6 we produce the statistical significance of the best estimated parameter pairs via both the RECURRENCE-RELATION and the MLE. The best pair is found in the RECURRENCE-RELATION method from 1000 uniform samples in the feasible-box centered at the point where the three curves are in agreement, and for the MLE, it is found from 1000 uniform samples in the maximum log-likelihood area if no unique maximizer exists. The estimates from both the techniques demonstrate the presence of the `DD-model` in the given graphs $G_n^{(1)}$ and $G_n^{(2)}$ (p-value > 0.1), the best pair of RECURRENCE-RELATION estimator has much higher p-value and certainly outperforms MLE.

We note that for the first graph $G_n^{(1)}$ the results obtained by both methods are almost identical, in terms of $\mathbb{E}[\log|\mathrm{Aut}(G_n)|]$ and p-values. For the second graph $G_n^{(2)}$, *the log-likelihood function of MLE is nearly flat for large values of p, and thus MLE returns less reliable estimates*. This in turn results in a larger deviation of the number of automorphisms from the observed graph. Our algorithm on the other hand provides a better estimate even when p is close to 1. To sum up, we find that our algorithm does not perform worse than MLE in terms of quality and achieves better performance than MLE when p is high. It also has much lower computational complexity than the MLE. The MLE requires more than 1 million computations for this particular example, where as RECURRENCE-RELATION needs only at most 100 computations (scaling factors excluded). Detailed complexity calculations are provided in Section 5.3.

6.2 Real-world PPIs

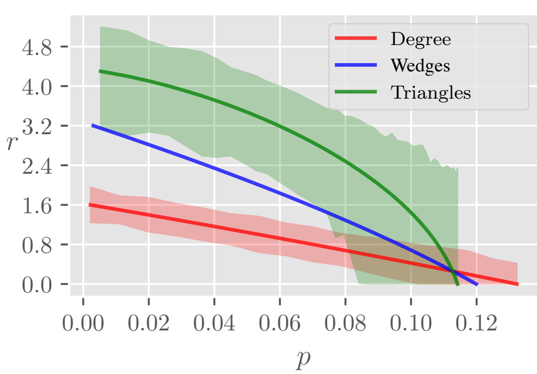
We apply recurrence-based estimator to PPI networks of seven species listed in Table 2. As mentioned in Section 2.1, the seed graph G_{n_0} is assumed as the graph induced by the nodes having the largest phylogenetic age.

The MLE solution is almost impossible to calculate for the PPI networks. Even for the smallest PPI network (Worm), MLE requires around 58 billion computations to obtain a result for a single (p, r) parameter set.

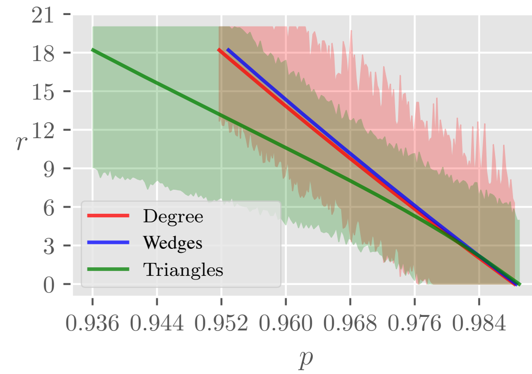
Figure 7 presents plots of RECURRENCE-RELATION estimator for seven species. In all the figures, the plots meet or come very close at a specific point. This illustrates the presence of the `DD-model` in all the considered species. Furthermore, Table 7 calculates the statistical significance of the fitted `DD-model` with respect to the number of automorphisms in the observed PPI networks. The estimated p-values are remarkably high and most often much larger than 0.4 (except in one case), demonstrating that the fitted DD-models exhibit symmetries closer to the real-world PPIs.

7 DISCUSSION

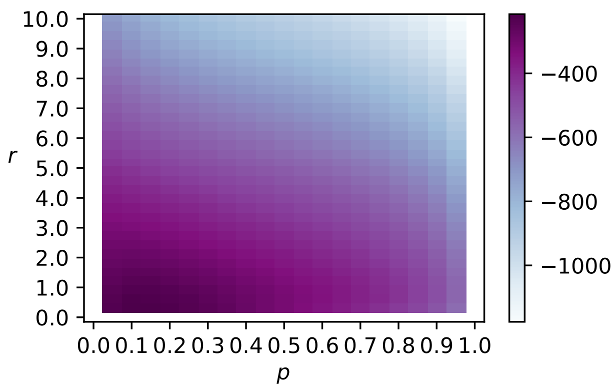
We focus in this work on fitting dynamic biological networks to a probabilistic graph model, from a single snapshot of the networks. Our attention here is on a key characteristic of the networks – the number of automorphisms – that is often neglected in modeling. Using the number of automorphisms as a measure to sample parameters from the parameter space may raise serious questions about its practicality



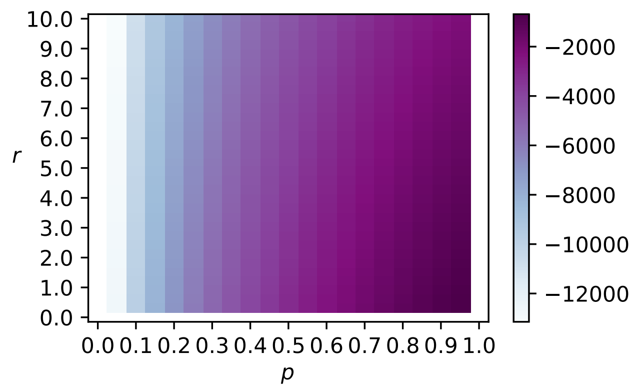
(a) RECURRENCE-RELATION: $G_n^{(1)} \sim \texttt{DD-model}(100, 0.1, 0.3, K_{20})$.



(b) RECURRENCE-RELATION: $G_n^{(2)} \sim \texttt{DD-model}(100, 0.99, 3.0, K_{20})$



(c) MLE: log-likelihood with $G_n^{(1)} \sim \texttt{DD-model}(100, 0.1, 0.3, K_{20})$.



(d) MLE: log-likelihood with $G_n^{(2)} \sim \texttt{DD-model}(100, 0.99, 3.0, K_{20})$

Fig. 6: Results on synthetic networks: RECURRENCE-RELATION and maximum likelihood estimation (MLE) methods

		RECURRENCE-RELATION				MLE			
Model parameters	$\log \lvert \mathrm{Aut}(G_{\mathrm{obs}}) \rvert$	$\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log \lvert \mathrm{Aut}(G_n) \rvert]$	p-value	$\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log \lvert \mathrm{Aut}(G_n) \rvert]$	p-value
$p = 0.1$, $r = 0.3$	81.963	0.09	0.3	81.974	0.980	0.1	0.3	78.794	0.820
$p = 0.99$, $r = 3.0$	16.178	0.99	2.5	16.588	0.980	0.95	0.3	0.368	0

TABLE 6: Results on synthetic networks: average number of automorphisms and p-value

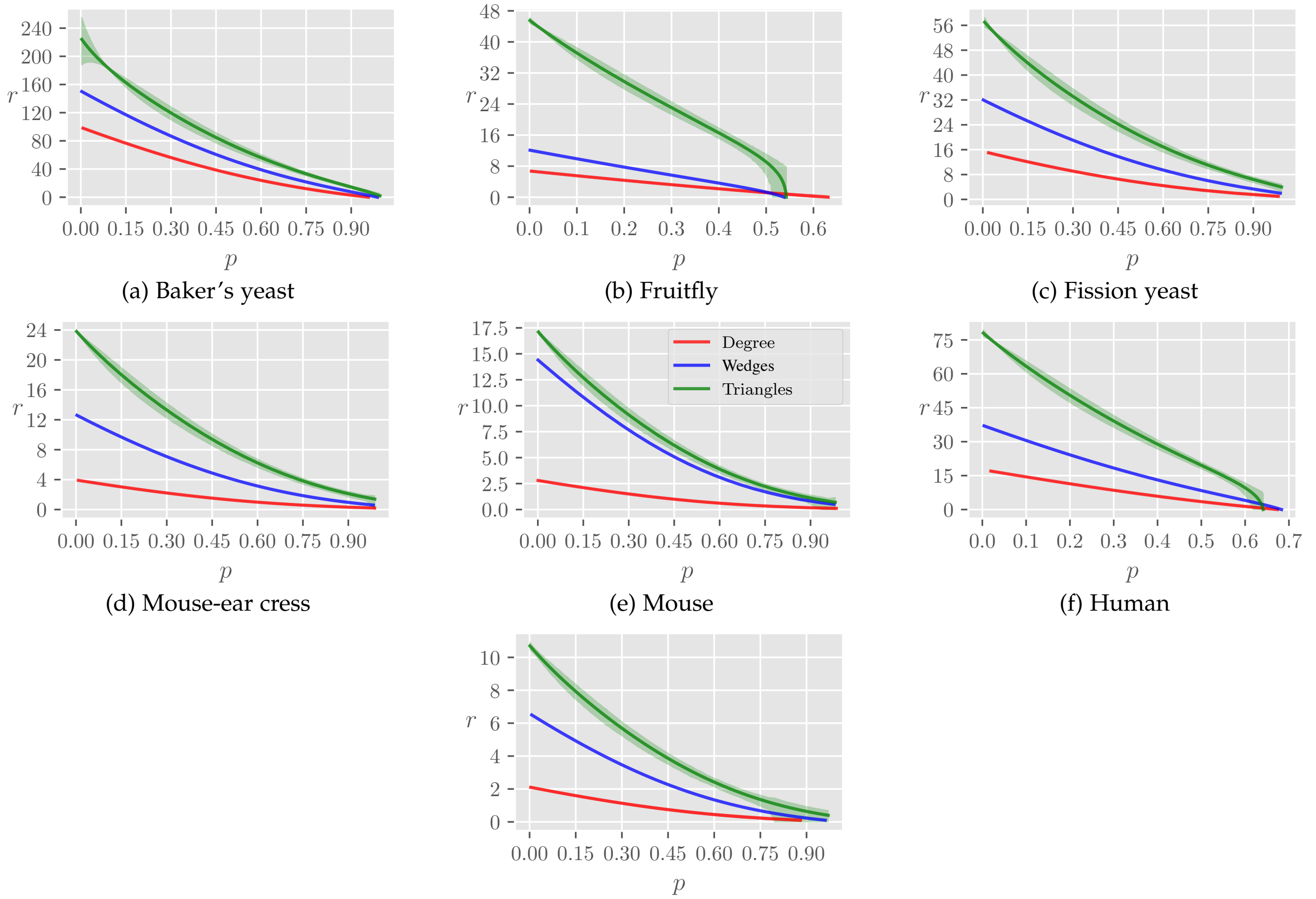
Organism	$\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log \lvert \mathrm{Aut}(G_n) \rvert]$	p-value
Baker's yeast	0.98	0.35	293.27	0.71
Human	0.64	0.49	2998.81	0.51
Fruitfly	0.53	0.92	1073.83	0.64
Fission yeast	0.983	0.85	705.278	0.74
Mouse-ear cress	0.98	0.49	6210.36	0.13
Mouse	0.96	0.32	8067.56	0.67
Worm	0.85	0.35	3352.91	0.48

TABLE 7: Parameters of the real-world PPI networks estimated using RECURRENCE-RELATION method

(like some slower maximum likelihood estimation methods for graph fitting). To address this, our approach in this paper to combine the number of symmetries with a faster method of recurrence relations, which allows us to narrow down the parameter search, finds high relevance in practice.

We argue that many existing parameter estimation techniques fail to take into account the number of symmetries of real-world networks, leading to serious concerns in the fitting methodology. Previous studies made unrealistic assumptions like the steady-state behavior of the model, and it could be the reason behind erroneous estimates. Our proposed fitting method based on exact recurrence relations, derived from rigorous theory, with minimal assumptions works well on synthetic data and real-world protein-protein interaction (PPI) networks of seven species. We also formulate a simple statistical test in terms of the number of symmetries. Since the PPI networks are expanding with new protein-protein interactions getting discovered, we make sure to use up-to-date data so that the fitted parameters in this paper can serve as a benchmark for future studies.

We note here that the method introduced in this work is applicable to a variety of dynamic network models, as for many models there exist recurrence relations similar to the ones presented here. A systematic way of parameter estimation can also be seen as an introductory work to other important problems in biological networks. One example of such a problem is the *temporal order problem* [30]: given a network, the task is to recover the chronology of the node arrivals in the network. Parameter estimation provides us with better knowledge about the specific characteristics of



(a) Baker's yeast (b) Fruitfly (c) Fission yeast (d) Mouse-ear cress (e) Mouse (f) Human (g) Worm

Fig. 7: Results on PPI networks: RECURRENCE-RELATION method

the model that retains temporal information in its structure.

8 MAXIMUM LIKELIHOOD ALGORITHM

Function MAXIMUMLIKELIHOODVALUE, as shown below in Algorithm 2, presents a single pass of the MLE technique. Here $\theta_0 = (p_0, r_0)$ is the initial parameter set which can be chosen with some extra knowledge of the given network or even arbitrarily; but with a proper choice of θ_0, a faster convergence to the true value of likelihood is guaranteed.

We note here that the MLE procedure has to be run multiple times, and the average of the L's (see algorithm) from the multiple runs gives an estimate of the likelihood at the inputted (p, r). The procedure needs to be repeated for all the relevant (p, r) pairs. The estimated parameters are the points for which the likelihood function is maximized. See Section 5.2 for details.

Algorithm 2 Single run for the likelihood value computation.

function MAXIMUMLIKELIHOODVALUE(G, n, n_0, p, r, p_0, r_0)
  $L \leftarrow 1$
  for $k = n, n-1, \ldots, n_0 + 1$ **do**
    Pick v at random with $\Pr[v] \sim \omega(G, p_0, r_0, v)$
    $L \leftarrow L \cdot \frac{1}{k} \frac{\sum_{u \in G} \omega(G, p_0, r_0, u)}{\omega(G, p_0, r_0, v)} \omega(G, p, r, v)$
    Remove v from G
  return L
function ω(G, p, r, v)
  $sum \leftarrow 0$, $n \leftarrow |V(G)|$
  for $u \in V(G)$, $u \neq v$ **do**
    $both \leftarrow |\mathcal{N}_n(u) \cap \mathcal{N}_n(v)|$
    $only_v \leftarrow |\mathcal{N}_n(v) \setminus \mathcal{N}_n(u)|$
    $only_u \leftarrow |\mathcal{N}_n(u) \setminus \mathcal{N}_n(v)|$
    $none \leftarrow n - |\mathcal{N}_n(u) \cup \mathcal{N}_n(v)|$
    $sum \mathrel{+}= p^{both} \left(\frac{r}{n}\right)^{only_v} (1-p)^{only_u} \left(1 - \frac{r}{n}\right)^{none}$
  return sum

9 PROOF OF THEOREM 1

Let $\deg_t(s)$ be the degree at time t of a vertex added at time s, which is same as node label s, and parent(t) be a vertex which was chosen from G_{t-1} for the duplication step at time t.

It follows from the definition of the model that degree of the new vertex $n+1$ is the total number of edges from the vertex $n+1$ to $\mathcal{N}_n(\text{parent}(n+1))$ (each of which is formed from choosing nodes independently from $\mathcal{N}_n(\text{parent}(n+1))$ with probability p) and to all other vertices (each of which is formed from nodes chosen independently from a set $V(G_n) \backslash \mathcal{N}_n(\text{parent}(n+1))$ with probability $\frac{r}{n}$).

It can be then expressed as a sum of two independent binomial variables:

$$\deg_{n+1}(n+1) \sim \mathrm{Bin}\left(\deg_n(\mathrm{parent}(n+1)), p\right) + \mathrm{Bin}\left(n - \deg_n(\mathrm{parent}(n+1)), \frac{r}{n}\right).$$

1. Recurrence for $D(G_n)$.

$$\begin{aligned}
&\mathbb{E}[\deg_{n+1}(n+1) \mid G_n] \\
&= \sum_{k=0}^{n-1} \Pr(\deg_n(\mathrm{parent}(n+1)) = k \mid G_n) \\
&\quad \times \sum_{a=0}^{k} \binom{k}{a} p^a (1-p)^{k-a} \\
&\quad \times \sum_{b=0}^{n-k} \binom{n-k}{b} \left(\frac{r}{n}\right)^b \left(1-\frac{r}{n}\right)^{n-k-b} (a+b) \\
&= \sum_{k=0}^{n-1} \Pr(\deg_n(\mathrm{parent}(n+1)) = k \mid G_n) \left(pk + \frac{r}{n}(n-k)\right) \\
&= \left(p - \frac{r}{n}\right) \sum_{k=0}^{n-1} k \Pr(\deg_n(\mathrm{parent}(n+1)) = k \mid G_n) + r.
\end{aligned}$$

Since in the definition of the model it is stated that the parent is selected uniformly at random, we know that $\Pr(\mathrm{parent}(n+1) = i \mid G_n) = \frac{1}{n}$ and therefore

$$\begin{aligned}
D(G_n) &= \sum_{i=1}^{n} \Pr(\mathrm{parent}(n+1) = i \mid G_n) \deg_n(i) \\
&= \sum_{k=0}^{n-1} k \Pr(\deg_n(\mathrm{parent}(n+1) \mid G_n) = k).
\end{aligned}$$

Combining the last two equations, we get

$$\mathbb{E}[\deg_{n+1}(n+1) \mid G_n] = \left(p - \frac{r}{n}\right) D(G_n) + r.$$

Using the above, we find the following recurrence for the mean degree of G_{n+1}:

$$\begin{aligned}
&\mathbb{E}[D(G_{n+1}) \mid G_n] \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \deg_{n+1}(i) \mid G_n\right] \\
&= \frac{1}{n+1} \left(\sum_{i=1}^{n} \deg_n(i) + 2\mathbb{E}\left[\deg_{n+1}(n+1) \mid G_n\right]\right) \\
&= \frac{1}{n+1} \left(nD(G_n) + 2\mathbb{E}[\deg_{n+1}(n+1) \mid G_n]\right) \\
&= D(G_n) \left(1 + \frac{2p-1}{n+1} - \frac{2r}{n(n+1)}\right) + \frac{2r}{n+1}.
\end{aligned}$$

Now from the law of total expectation:

$$\begin{aligned}
&\mathbb{E}[D(G_{n+1})] \\
&= \mathbb{E}[D(G_n)] \left(1 + \frac{2p-1}{n+1} - \frac{2r}{n(n+1)}\right) + \frac{2r}{n+1}.
\end{aligned}$$

2. Recurrence for $D_2(G_n)$.

$$\begin{aligned}
&\mathbb{E}[\deg^2_{n+1}(n+1) \mid G_n] \\
&= \sum_{k=0}^{n-1} \Pr(\deg_n(\mathrm{parent}(n+1)) = k \mid G_n) \\
&\quad \times \sum_{a=0}^{k} \binom{k}{a} p^a (1-p)^{k-a} \\
&\quad \times \sum_{b=0}^{n-k} \binom{n-k}{b} \left(\frac{r}{n}\right)^b \left(1-\frac{r}{n}\right)^{n-k-b} (a+b)^2 \\
&= \sum_{k=0}^{n-1} \Pr(\deg_n(\mathrm{parent}(n+1)) = k \mid G_n) \\
&\quad \times \left(k^2 \left(p^2 - \frac{2pr}{n} + \frac{r^2}{n^2}\right) + k\left(p - p^2 + 2pr - \frac{r + 2r^2}{n}\right)\right. \\
&\qquad \left. + r^2 + r - \frac{r^2}{n}\right) \\
&= D_2(G_n) \left(p^2 - \frac{2pr}{n} + \frac{r^2}{n^2}\right) \\
&\quad + D(G_n) \left(p - p^2 + 2pr - \frac{r + 2r^2}{n} + \frac{r^2}{n^2}\right) \\
&\quad + r^2 + r - \frac{r^2}{n},
\end{aligned}$$

since we have, as before,

$$\begin{aligned}
D_2(G_n) &= \sum_{i=1}^{n} \Pr(\mathrm{parent}(n+1) = i \mid G_n) \deg^2_t(i) \\
&= \sum_{k=0}^{n-1} \Pr(\deg_n(\mathrm{parent}(n+1)) = k \mid G_n) k^2.
\end{aligned}$$

Now we proceed with the second moment of degree distribution of G_n. Let $I_{n+1}(i)$ be an indicator variable whether there is an edge between $n+1$ and i. Then the following basic results follows:

$$\sum_{i=1}^{n} I^2_{n+1}(i) = \sum_{i=1}^{n} I_{n+1}(i) = \deg_{n+1}(n+1)$$

and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} \deg_n(i) I_{n+1}(i) \mid G_n\right] = \sum_{i=1}^{n} \deg_n(i) \mathbb{E}[I_{n+1}(i) \mid G_n] \\
&= \sum_{i=1}^{n} \deg_n(i) \left(\frac{\deg_n(i)}{n} p + \frac{n - \deg_n(i)}{n} \frac{r}{n}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \deg^2_t(i) \left(p - \frac{r}{n}\right) + \frac{1}{n} \sum_{i=1}^{n} \deg_n(i) r \\
&= \left(p - \frac{r}{n}\right) D_2(G_n) + r D(G_n).
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}[D_2(G_{n+1}) \mid G_n] = \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \deg^2_{n+1}(i) \mid G_n\right] \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n} \deg^2_t(i) + 2\sum_{i=1}^{n} \deg_n(i) I_{n+1}(i)\right. \\
&\qquad \left. + \deg_{n+1}(n+1) + \deg^2_{n+1}(n+1) \mid G_n\right] \\
&= D_2(G_n) \left(1 + \frac{2p + p^2 - 1}{n+1} - \frac{2r(1+p)}{n(n+1)} + \frac{r^2}{n^2(n+1)}\right) \\
&\quad + D(G_n) \\
&\quad \times \left(\frac{2p - p^2 + 2pr + 2r}{n+1} - \frac{2r + 2r^2}{n(n+1)} + \frac{r^2}{n^2(n+1)}\right) \\
&\quad + \frac{2r^2 + 2r}{n+1} - \frac{r^2}{n(n+1)}.
\end{aligned}$$

Then from the law of total expectation we obtained the desired formula.

3. Recurrence for $S_2(G_n)$. The recurrence for $S_2(G_n)$ is straightforward from the following *deterministic* relation for every graph:

$$S_2(G_n) = \frac{n}{2}\left(D_2(G_n) - D(G_n)\right).$$

Alternatively, it can be computed from the following relation using similar methods as before:

$$\mathbb{E}[S_2(G_{n+1}) \mid G_n] = \mathbb{E}\left[\sum_{i=1}^{n+1}\binom{\deg_{n+1}(i)}{2} \mid G_n\right].$$

4. Recurrence for $C_3(G_n)$. Finally, we find the expected number of triangles in the following way. Let us denote by $e_t(A,B)$ the number of edges with one endpoint in A and other in B, and by $e_t(A)$ the number of edges with both edges in A for some fixed $A, B \subseteq V(G_t)$, at time t.

For brevity, let us also introduce the following notations.

- $NP(n+1) = \mathcal{N}_n(\text{parent}(n+1))$ – the set of neighbors of the parent of $n+1$ in G_n,
- $X_1(G_n) := e_n(NP(n+1))$ – the number of edges within (open) neighborhood of the parent of $n+1$ in G_n,
- $X_2(G_n) := e_n\left(NP(n+1), V(G_n) \setminus NP(n+1)\right)$ – the number of edges between (open) neighborhood of the parent of $n+1$ and other vertices in G_n,
- $X_3(G_n) := e_n\left(V(G_n) \setminus NP(n+1)\right)$ – the number of edges between vertices not connected to the parent of $n+1$ in G_n.

It can be easily verified that

$$\begin{aligned} X_1(G_n) &= \frac{3C(G_n)}{n}, \\ X_1(G_n) + X_2(G_n) &= D_2(G_n), \\ X_1(G_n) + X_2(G_n) + X_3(G_n) &= \frac{n}{2}D(G_n). \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{E}[C_3(G_{n+1}) \mid G_n] \\ &\quad = C_3(G_n) + p^2\mathbb{E}[X_1(G_n) \mid G_n] \\ &\quad\quad + \frac{pr}{n}\mathbb{E}[X_2(G_n) \mid G_n] + \frac{r^2}{n^2}\mathbb{E}[X_3(G_n) \mid G_n] \\ &\quad = C_3(G_n)\left(1 + \frac{3p^2}{n} - \frac{6pr}{n^2} + \frac{3r^2}{n^3}\right) \\ &\quad\quad + D_2(G_n)\left(\frac{pr}{n} - \frac{r^2}{n^2}\right) + D(G_n)\frac{r^2}{2n}. \end{aligned}$$

We can now apply the law of total expectation to get the final result. □

References

[1] J. Zhang, "Evolution by gene duplication: an update," *Trends in Ecology & Evolution*, vol. 18, no. 6, pp. 292–298, 2003.

[2] S. Ohno, *Evolution by gene duplication*. Berlin–Heidelberg: Springer-Verlag, 1970.

[3] R. Pastor-Satorras, E. Smith, and R. V. Solé, "Evolving protein interaction networks through gene duplication," *Journal of Theoretical Biology*, vol. 222, no. 2, pp. 199–210, 2003.

[4] M. Shao, Y. Yang, J. Guan, and S. Zhou, "Choosing appropriate models for protein–protein interaction networks: a comparison study," *Briefings in Bioinformatics*, vol. 15, no. 5, pp. 823–838, 2013.

[5] F. Hormozdiari, P. Berenbrink, N. Pržulj, and S. C. Sahinalp, "Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution," *PLoS Computational Biology*, vol. 3, no. 7, p. e118, 2007.

[6] I. Ispolatov, P. Krapivsky, and A. Yuryev, "Duplication-divergence model of protein interaction network," *Physical Review E*, vol. 71, no. 6, p. 061911, 2005.

[7] A. Raval, "Some asymptotic properties of duplication graphs," *Physical Review E*, vol. 68, no. 6, p. 066119, 2003.

[8] F. Chung, L. Lu, T. G. Dewey, and D. Galas, "Duplication models for biological networks," *Journal of Computational Biology*, vol. 10, no. 5, pp. 677–687, 2003.

[9] J. H. Kim, B. Sudakov, and V. H. Vu, "On the asymmetry of random regular graphs and random graphs," *Random Structures & Algorithms*, vol. 21, no. 3-4, pp. 216–224, 2002.

[10] T. Łuczak, A. Magner, and W. Szpankowski, "Asymmetry and structural information in preferential attachment graphs," *Random Structures & Algorithms*, 2019.

[11] K. Turowski, A. Magner, and W. Szpankowski, "Compression of Dynamic Graphs Generated by a Duplication Model," in *56th Annual Allerton Conference on Communication, Control, and Computing*. Monticello, IL, US: IEEE, 2018, pp. 1089–1096.

[12] C. Godsil and G. F. Royle, *Algebraic graph theory*. Springer Science & Business Media, 2013.

[13] G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, and S. C. Sahinalp, "The degree distribution of the generalized duplication model," *Theoretical Computer Science*, vol. 369, no. 1-3, pp. 239–249, 2006.

[14] G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. H. Nadeau, and S. C. Sahinalp, "Improved duplication models for proteome network evolution," in *Systems Biology and Regulatory Genomics*. Berlin, Heidelberg: Springer, 2007, pp. 119–137.

[15] R. Colak, F. Hormozdiari, F. Moser, A. Schönhuth, J. Holman, M. Ester, and S. C. Sahinalp, "Dense graphlet statistics of protein interaction and random networks," in *Biocomputing 2009*. Singapore: World Scientific Publishing, 2009, pp. 178–189.

[16] A. Bhan, D. Galas, and T. G. Dewey, "A duplication growth model of gene expression networks," *Bioinformatics*, vol. 18, no. 11, pp. 1486–1493, 2002.

[17] C. Wiuf, M. Brameier, O. Hagberg, and M. Stumpf, "A likelihood approach to analysis of network data," *Proceedings of the National Academy of Sciences*, vol. 103, no. 20, pp. 7566–7570, 2006.

[18] S. Bubeck, E. Mossel, and M. Z. Rácz, "On the influence of the seed graph in the preferential attachment model," *IEEE Transactions on Network Science and Engineering*, vol. 2, no. 1, pp. 30–39, 2015.

[19] B. McKay and A. Piperno, "Practical graph isomorphism," *Journal of Symbolic Computation*, vol. 60, pp. 94–112, 2013.

[20] S. Heinicke, M. Livstone, C. Lu, R. Oughtred, F. Kang, S. Angiuoli, O. White, D. Botstein, and K. Dolinski, "The princeton protein orthology database (p-pod): a comparative genomics analysis tool for biologists," *PloS one*, vol. 2, no. 8, p. e766, 2007.

[21] L. Li, C. Stoeckert, and D. Roos, "Orthomcl: identification of ortholog groups for eukaryotic genomes," *Genome research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[22] P. Thomas, M. Campbell, A. Kejariwal, H. Mi, B. Karlak, Daverman *et al.*, "Panther: a library of protein families and subfamilies indexed by function," *Genome research*, vol. 13, no. 9, pp. 2129–2141, 2003.

[23] J. Capra, A. Williams, and K. Pollard, "Proteinhistorian: tools for the comparative analysis of eukaryote protein origin," *PLoS Computational Biology*, vol. 8, no. 6, p. e1002567, 2012.

[24] K. Brown, C. Hill, G. Calero, C. Myers, K. Lee, J. Sethna, and R. Cerione, "The statistical mechanics of complex signaling networks: nerve growth factor signaling," *Physical Biology*, vol. 1, no. 3, p. 184, 2004.

[25] J.-D. Han, N. Bertin, T. Hao, D. Goldberg, G. Berriz, L. Zhang, D. Dupuy, A. Walhout, M. Cusick, F. Roth *et al.*, "Evidence for dynamically organized modularity in the yeast protein–protein interaction network," *Nature*, vol. 430, no. 6995, p. 88, 2004.

[26] R. Tanaka, T.-M. Yi, and J. Doyle, "Some protein interaction data do not exhibit power law statistics," *FEBS Letters*, vol. 579, no. 23, pp. 5140–5144, 2005.

[27] R. Khanin and E. Wit, "How scale-free are biological networks," *Journal of Computational Biology*, vol. 13, no. 3, pp. 810–818, 2006.

[28] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[29] R. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler, "A model of large-scale proteome evolution," *Advances in Complex Systems*, vol. 5, no. 01, pp. 43–54, 2002.

[30] J. Sreedharan, A. Magner, A. Grama, and W. Szpankowski, "Inferring temporal information from a snapshot of a dynamic network," *Nature Scientific Reports*, vol. 9, no. 1, p. 3057, 2019.

Jithin K. Sreedharan is a Postdoctoral Research Associate at the NSF Center for Science of Information and Dept. of Computer Science in Purdue University. He received his Ph.D. in computer science from INRIA, France, in 2017 with a fellowship from INRIA-Bell Labs joint lab. Before that, he finished M.S. from Indian Institute of Science (IISc), Bangalore, in 2013, and received the best thesis award. His central research interest is in solving real-world problems in data science with a network perspective. His current works focus on data mining algorithms for large networks with probabilistic guarantees, statistical modeling, and inference on networks, and distributed techniques for analyzing big matrices.

Krzysztof Turowski is currently assistant professor at the Theoretical Computer Science Department at the Jagiellonian University, Krakow, Poland. He received his MS and PhD degrees from Gdansk University of Technology, Poland in 2011 and 2015, respectively, both in computer science. From 2010 to 2016 he was employed at the Department of Algorithms and System Modelling at Gdansk University of Technology and from 2016 to 2018 he worked at Google as a software developer for Google Compute Engine. From 2018 to 2019 he was a Postdoctoral Research Scholar in the NSF Center for Science of Information at Purdue University. His research interests include graph theory (especially various models of graph coloring), analysis of algorithms and information theory.

Wojciech Szpankowski is Saul Rosen Distinguished Professor of Computer Science at Purdue University where he teaches and conducts research in analysis of algorithms, information theory, analytic combinatorics, data science, random structures, and stability problems of distributed systems. He held several Visiting Professor/Scholar positions, including McGill University, INRIA, France, Stanford, Hewlett-Packard Labs, Universite de Versailles, University of Canterbury, New Zealand, Ecole Polytechnique, France, the Newton Institute, Cambridge, UK, ETH, Zurich, and Gdansk University of Technology, Poland. He is a Fellow of IEEE, and the Erskine Fellow. In 2010 he received the Humboldt Research Award and in 2015 the Inaugural Arden L. Bement Jr. Award. He is also the recipient of 2020 Flajolet Lecture Prize. He published two books: "Average Case Analysis of Algorithms on Sequences", John Wiley & Sons, 2001, and "Analytic Pattern Matching: From DNA to Twitter", Cambridge, 2015. In 2008 he launched the interdisciplinary Institute for Science of Information, and in 2010 he became the Director of the newly established NSF Science and Technology Center for Science of Information.