Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries

Jithin Sreedharan, Krzysztof Turowski and Wojciech Szpankowski

NSF Center for Science of Information and Dept. of Computer Science, Purdue University

{jithinks, kturowsk, szpan}@purdue.edu

paper code

PURDUE UNIVERSITY

Center for Science of Information

NSF Science and Technology Center

Motivation & Contributions

- **Symmetries of the graph.** 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

  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

- **Graph parameter recurrences.** Existing methods heavily depend upon *steady-state assumption and asymptotic properties* of the graph model.

  We derive exact non-asymptotic recurrence relations of degree, no. of wedges and no. of triangles

- **Maximum likelihood method.** MLE via importance sampling requires $\Theta(n^3/\varepsilon^2)$ computations for the DD-model (n and ε being the number of nodes and the required resolution)

  Our approach based on recurrence relations requires only $\Theta(n/\varepsilon \log(1/\varepsilon))$ steps

- **Seed graph choice.** It is well known that seed graphs play an important role in biological networks

  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

Duplication-Divergence Graph Model (DD-model)

Start with seed graph G_{n_0}. At time step k:

- Duplication: Select a node u from G_k uniformly at random. New node v copies all connections of u
- Divergence: Each of the new made connections of v are randomly deleted with probability $1-p$. For all other nodes, create a connection randomly with v with probability r/k

Duplication step

Divergence step

Datasets

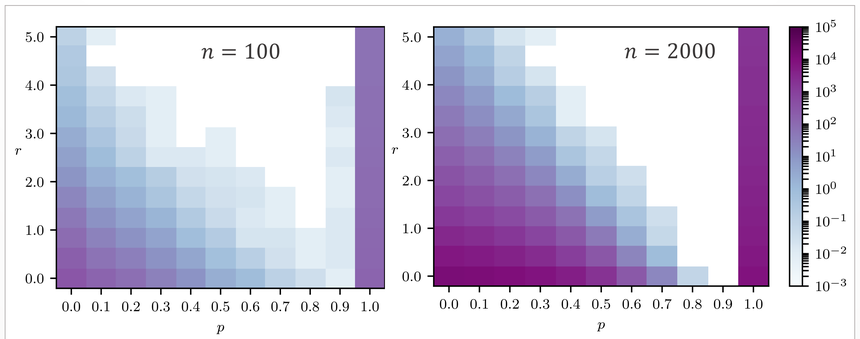
Protein-protein interaction (PPI) networks of 7 species

Organism	Scientific name	Original graph G_{obs} # Nodes	# Edges	$\log \lvert\text{Aut}(G)\rvert$	Seed graph G_{n_0} # 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

Selection of seed graph

Select the seed graph as the graph induced in the PPI networks by the oldest proteins, with the largest phylogenetic age (taxon age). 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.

Influence of Parameters on Symmetries of the Model



$\mathbb{E}[\log|\text{Aut}(G_n)|]$ generated from the DD-model. The seed graph $G_{n_0} = K_{20}$

- Other graph models like the preferential-attachment and Erdős–Rényi models are asymmetric with high probability [1,2]
- Presence of large number of symmetries in the DD-model for certain parameter range makes it suitable for fitting PPI networks and other biological networks

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

Let $G_n^{(1)}, \ldots, G_n^{(m)}$ be m graphs generated from the DD-model with the estimated parameters using any fitting method

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

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

$$p\text{-value} = 2\min\{p_u, p_l\}.$$

Why existing parameter estimation methods fail in practice?

Organism	$\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log\lvert\text{Aut}(G_n)\rvert]$	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

Mismatch in the number of symmetries and graph statistics with the mean-field approach [3]

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

Dependence on power-law behavior in the estimation techniques

Our Method: Parameter Estimation Using Recurrence-Relations

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

$D(G_n) = n^{-1}\sum_{i=1}^{n} \deg_n(i)$ is the mean degree

$S_2(G_n)$ is the number of *wedges*

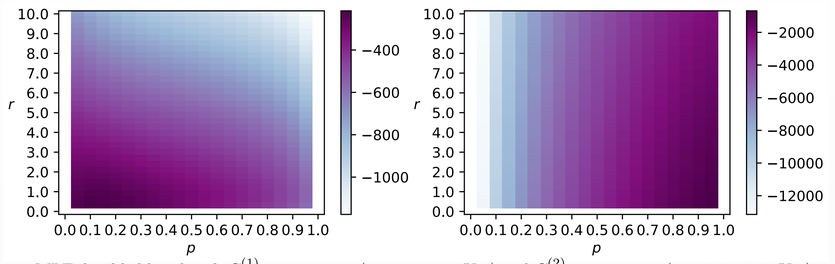
$$\mathbb{E}[D(G_{n+1})|G_n] = D(G_n)\left(1 + \frac{2p-1}{n+1} - \frac{2r}{n(n+1)}\right) + \frac{2r}{n+1}$$

$$\mathbb{E}[S_2(G_{n+1})|G_n] = S_2(G_n)\left(1 + \frac{2p+p^2}{n} - \frac{2(p+1)r}{n^2} + \frac{r^2}{n^3}\right) + D(G_n)\left(pr + p + r - \frac{pr+r+r^2}{n} + \frac{r^2}{n^2}\right) + \frac{r^2}{2}$$

- Similar expressions derived for mean squared degree and number of triangles
- Find solution set $\{(\widehat{p}, \widehat{r})\}$ with recurrence-relations of each graph properties
- If we find a concurrence in their solutions, a necessary condition for the presence of duplication-divergence model has been satisfied
- Output the converging point as the fitted parameter set

Numerical Results with Recurrence-Relation method

Results on synthetic graphs

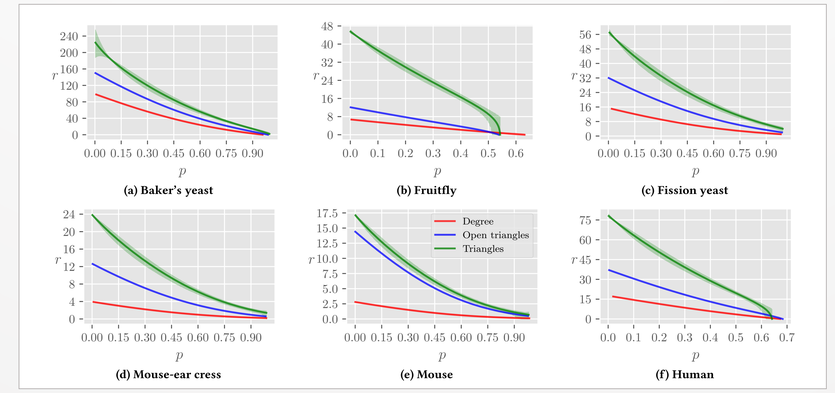


MLE: log-likelihood with $G_n^{(1)} \sim \texttt{DD-model}(100, 0.1, 0.3, K_{20})$ and $G_n^{(2)} \sim \texttt{DD-model}(100, 0.99, 3.0, K_{20})$

- Log-likelihood function of MLE is nearly flat for large values of p, and thus MLE returns less reliable estimates

Model parameters	$\log\lvert\text{Aut}(G_{\text{obs}})\rvert$	Recurrence-Relation $\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log\lvert\text{Aut}(G_n)\rvert]$	p-value	MLE $\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log\lvert\text{Aut}(G_n)\rvert]$	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

Results on real-world PPI networks



Organism	$\widehat{p}$	$\widehat{r}$	$\mathbb{E}[\log\lvert\text{Aut}(G_n)\rvert]$	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

Discussion

- We focus 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. We combine the number of automorphisms with a faster method of recurrence relations to allows us to narrow down the parameter search space
- 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
- The methods introduced in this work is applicable to a variety of dynamic network models, as for many models one can derive recurrence relations similar to the ones presented here

Please see the paper for references