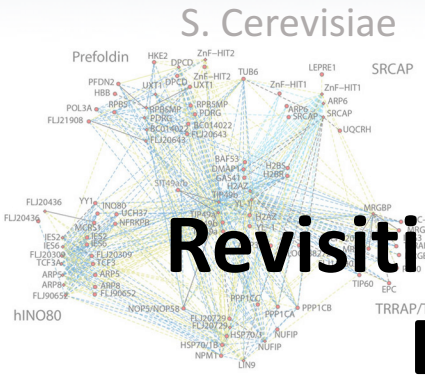
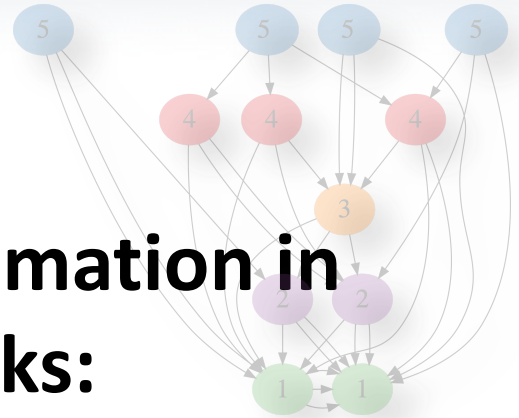




Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries

Jithin K. Sreedharan
Purdue University

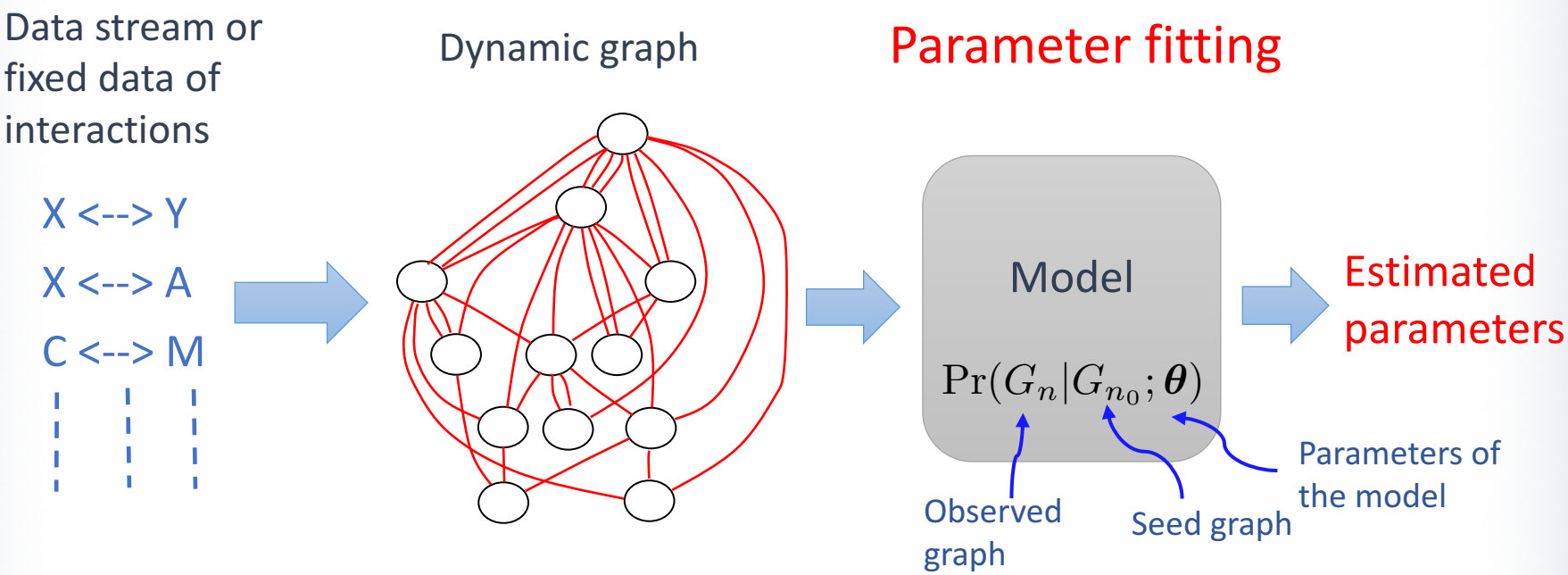
Krzysztof Turowski (Purdue)

Wojceich Szpankowski (Purdue)

Center for Science of Information
NSF Science and Technology Center

BioKDD August 5, 2019

The Problem: Fitting a model



- Data usually represents a single snapshot $G_{\text{obs}} := G_n$ of the graph of dynamic evolution $G_n, G_{n-1}, \ldots, G_{n_0}$
- Random graph models tailored to specific applications provide deep insights unlike general learning models
- Examples: asymptotic behavior, clustering properties, properties of motifs (subgraphs or lower/higher order structures), diffusion over the graph etc

Why need to revisit the estimation methods?

Symmetries of the graph

- Most of the existing parameter estimation techniques overlook the critical property of graph symmetry (also known formally as graph automorphisms).
- The estimated parameters give statistically insignificant results concerning the observed network

Goal-1: Take into account the number of automorphisms of the observed network to restrict the parameter search to a more meaningful range

Parameter estimation methods

- Existing methods heavily depend upon *stead-state assumption and asymptotic properties* of the graph model
- Many of these assumptions has been proven not to exist or exist with strong conditions

Goal-2: Use exact non-asymptotic relations

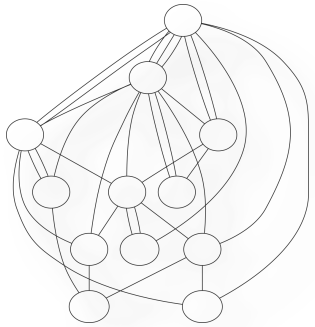
Why need to revisit the estimation methods?

Maximum likelihood method

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \Pr(G_n | G_{n_0}; \boldsymbol{\theta})$$

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

set of all sequences of graphs that starts with G_{n_0} and ends at G_n



Given one snapshot of the graph, $(n - n_0)!$ ways to arrange the order of arrival of nodes

- Direct computation of likelihood of a dynamic graph model requires $O(n!)$ computations
- Clever techniques with importance sampling or expectation-maximization still requires huge complexity
- For e.g., for Duplication-Divergence graph model, it is $\Theta(n^3/\varepsilon^2)$ with a large hidden constant factor (n: no. of nodes, ε: required resolution)

Goal-3: Achieve $\Theta(n)$ complexity

Seed graph choice

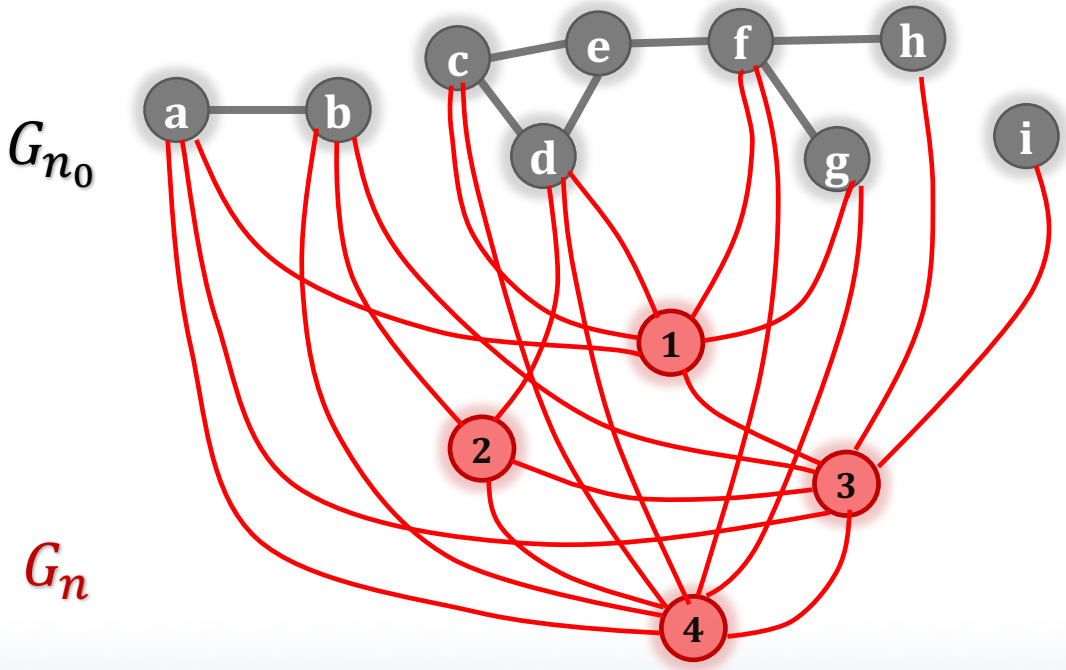
- Seed graphs play an important role in biological networks
- Previous solutions form seed graphs as cliques

Goal-4: Form a seed graph with biological relevance

Duplication-Divergence model (vertex-copying model)

Start with seed graph G_{n_0}. A 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



Datasets Used

Protein-protein interaction (PPI) networks of 7 species

Data collected from BioGRID. Removed self-interactions (self-loops), multiple interactions (multiple edges), and interspecies (organisms) interactions of proteins.

		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

Selection of seed graph

- As the graph induced in the PPI network by the oldest proteins, those with the largest phylogenetic age (taxon age)
- The age of a protein is based on its family's appearance on a species tree, and is estimated via protein family databases and ancestral history reconstruction algorithms

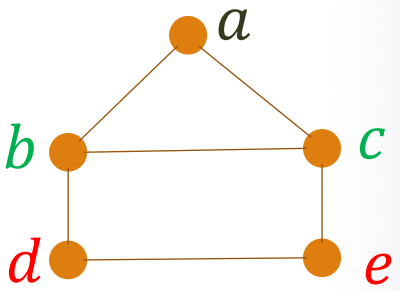
  Princeton Protein Orthology Database (PPOD) along with OrthoMCL and PANTHER for the protein family database and asymmetric Wagner parsimony as the ancestral history reconstruction algorithm

Influence of Parameters on Symmetries of the Model

Symmetries of the graph (Graph Automorphism):

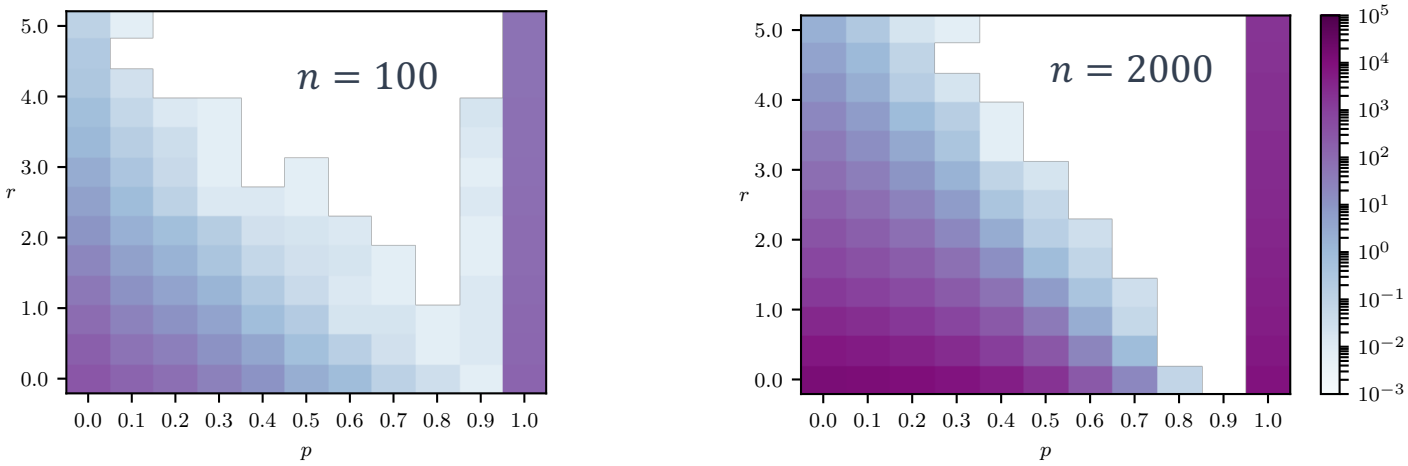
An automorphism of G is adjacency preserving permutation of vertices of G (i.e., a form of symmetry)

The collection Aut(G) of automorphisms of G is called autmorphism group of G



- Neglected in most of the prior works
- Real-world PPI networks exhibit large number of symmetries
- Erdős–Rényi and preferential attachment models are asymmetric with high probability
- Cross-checking with the number of automorphisms of the real-world network forms a null hypothesis test for the model under consideration

Influence of Parameters on Symmetries of the Model



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

For large ranges of p and r, it is impossible to generate graphs with large number of automorphisms

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 $\texttt{DD-model}(n, \widehat{p}, \widehat{r}, G_{n_0})$ with the estimated parameters using any fitting method

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

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

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

Why existing parameter estimation methods fail in practice (contd.)?

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

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

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

Power-law exponent

Average degree

References

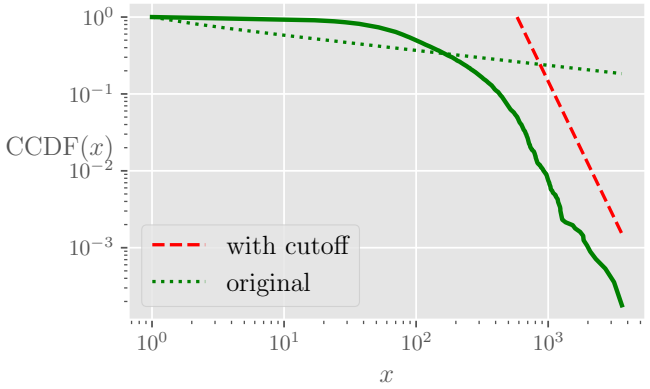
- R. Pastor-Satorras, Eric Smith, and Ricard V Solé. Journal of Theoretical Biology, 2003
- Fereydoun Hormozdiari, Petra Berenbrink, Nataša Pržulj, and Süleyman Cenk Sahinalp. PLoS Computational Biology, 2007
- Mingyu Shao, Yi Yang, Jihong Guan, and Shuigeng Zhou. Briefings in Bioinformatics, 2013

Why existing parameter estimation methods fail in practice (contd.)?

Power-law behavior

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

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



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

Cutoff neglects a huge percentage of the data

Asymptotic and steady-state assumption

- No theoretical proof for convergence to steady-state.
- Moreover, steady-state asymptotic results, even when achievable, do not give any bounds on the rate of convergence
- Assumes the average degree of the network does not change during the whole evolution.

Our method based on recurrence relations of graph statistics

A set of the exact recurrence relations for basic graph statistics, which relate their values at time k and $k + 1$ of graph evolution.

Theorem

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

Mean degree

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

Mean squared degree

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

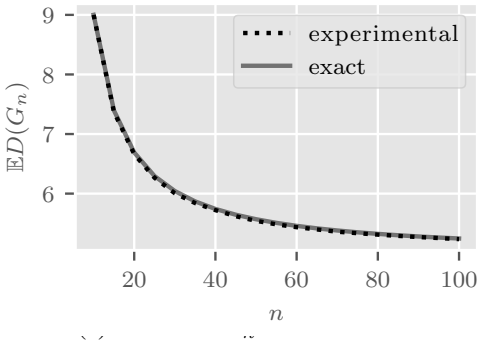
No. of triangles

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

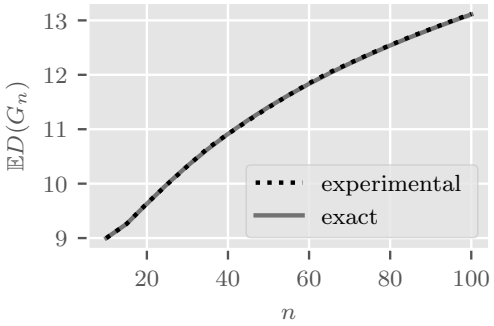
No. of wedges (paths of length 2)

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

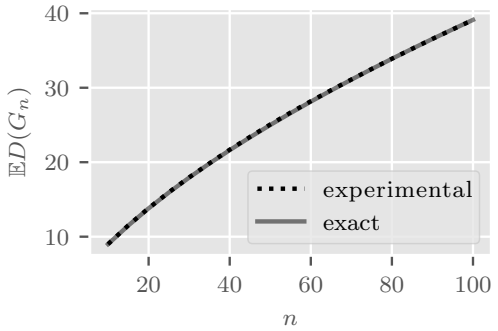
Our method based on recurrence relations of graph statistics (contd.)



DD-model$(100, 0.2, 1.5, K_{10})$



DD-model$(100, 0.5, 1.5, K_{10})$



DD-model$(100, 0.8, 1.5, K_{10})$

- 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
- Computational complexity is $\Theta(n/\varepsilon \log(1/\varepsilon))$

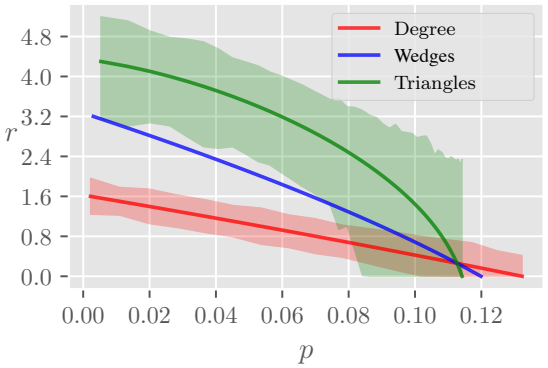
For theoretical results, see the paper

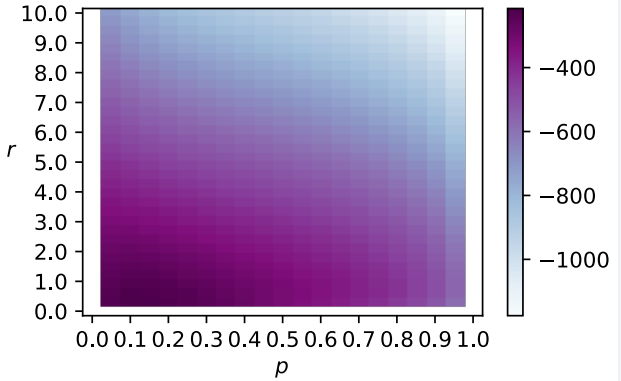
Results on synthetic networks: **Recurrence -Relation method**

Recurrence-Relation method

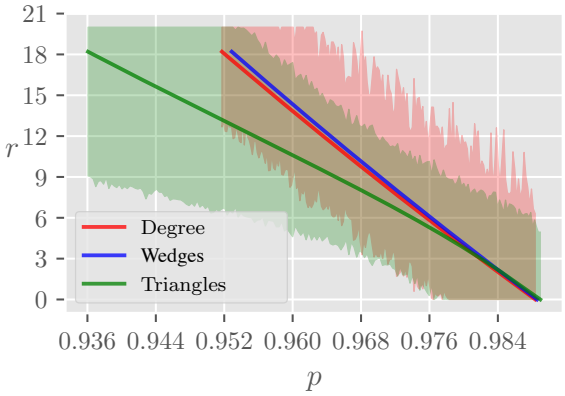
Log-likelihood

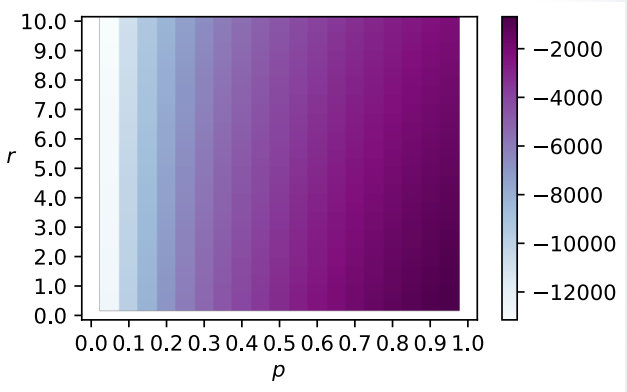
$G_n^{(1)} \sim \texttt{DD-model}(n = 100, p = 0.1, r = 0.3, G_{n_0} = K_{20})$,





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



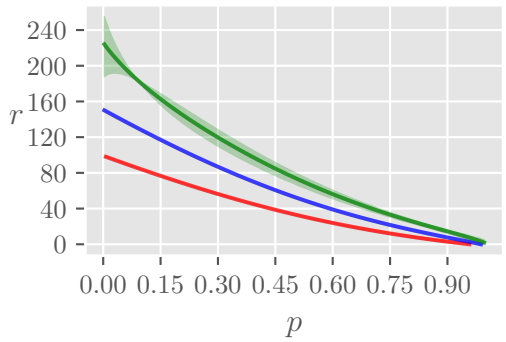


For confidence interval calculation, see the paper

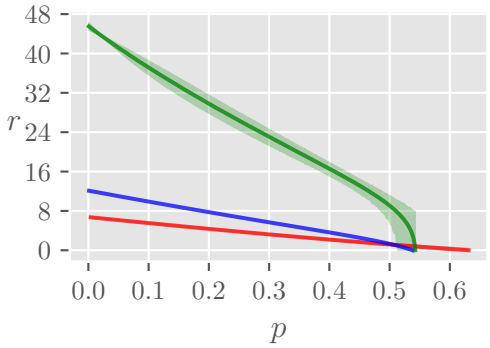
Model parameters	$\log \lvert\mathrm{Aut}(G_{\mathrm{obs}})\rvert$	Recurrence-Relation $\widehat{p}$	Recurrence-Relation $\widehat{r}$	Recurrence-Relation $\mathbb{E}[\log \lvert\mathrm{Aut}(G_n)\rvert]$	Recurrence-Relation p-value	MLE $\widehat{p}$	MLE $\widehat{r}$	MLE $\mathbb{E}[\log \lvert\mathrm{Aut}(G_n)\rvert]$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

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

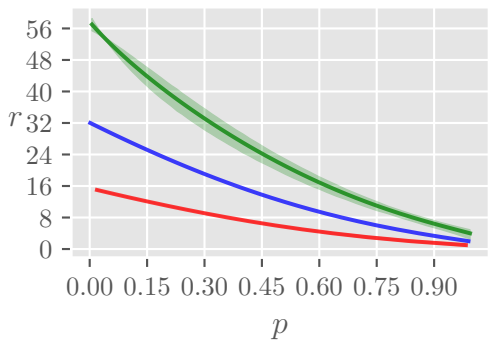
Results on protein-protein networks: Recurrence -Relation method



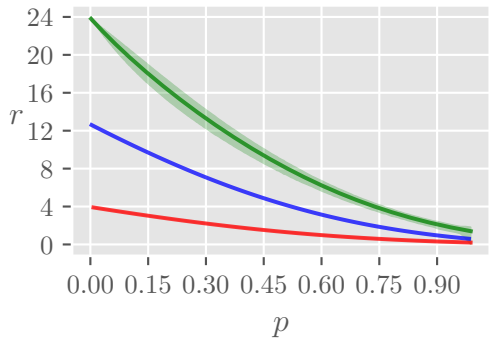
Baker's yeast



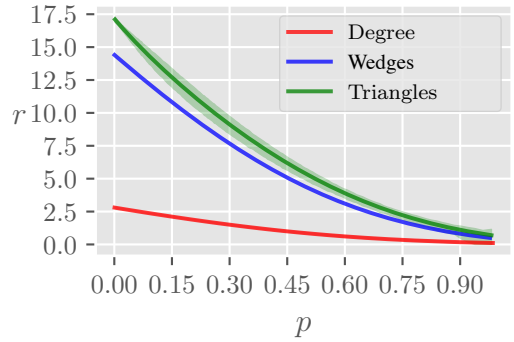
Fruitfly



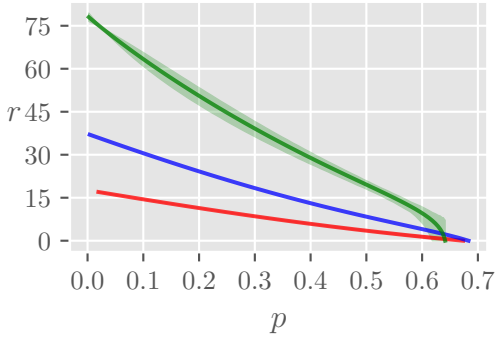
Fission yeast



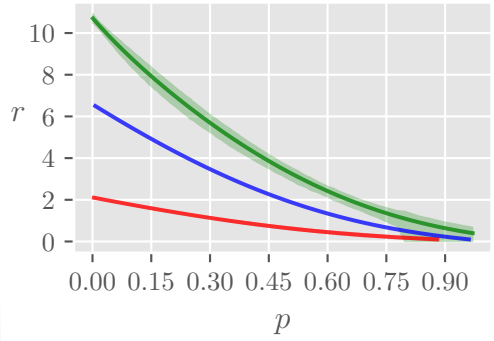
Mouse-ear cress



Mouse



Human



Worm

Results on protein-protein networks (contd.)

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

Parameters of the real-world PPI networks estimated using Recurrence -Relation method

Conclusions

- Fitting dynamic biological networks to a probabilistic graph model from a single snapshot of the evolution with stress on a key characteristic of the networks – the number of automorphisms – that is often neglected in modeling.
- Combined the number of automorphisms with a faster method of recurrence relations to to narrow down the parameter search space
- Much lower computational complexity
- Tested on protein-protein interaction data of 7 species
- Be extra careful when applying mean-field approach without strong theoretical guarantees
- Used up-to-date PPI data so that the fitted parameters in this paper can serve as a benchmark for future studies

Slides, paper, code, and data are available at cs.purdue.edu/homes/jithinks/

Thank You!

Extra Slides

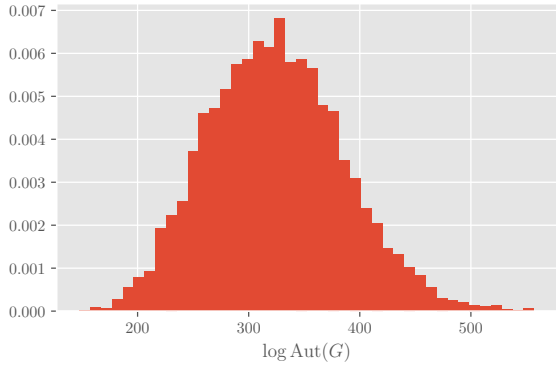


Figure 2: Normalized histogram of logarithm of number of automorphisms when $G_n \sim \texttt{DD-model}(500, 0.3, 0.4, K_{20})$.

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.

Why existing parameter estimation methods fail in practice?

Seed graph choice

- The seed graph is typically assumed to be the largest clique 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
- No formal theoretical guarantees and does not have clear justification

Algorithm 1 Parameter estimation via recurrence relation of $D(G_n)$.

function RECURRENCE-RELATION(n, r, G_{n_0}, $D(G_n)$, ε)
    $D_{\min} \leftarrow F_D(n, 0, r, G_{n_0})$, $D_{\max} \leftarrow F_D(n, 1, r, G_{n_0})$
    if $D_{\min} > D(G_n)$ or $D_{\max} < D(G_n)$ **then**
        return "no suitable solution for p"
    $p_{\min} \leftarrow 0$, $p_{\max} \leftarrow 1$
    while $p_{\max} - p_{\min} > \varepsilon$ **do**
        $p' \leftarrow \frac{p_{\min}+p_{\max}}{2}$, $D' \leftarrow F_D(n, p', r, G_{n_0})$
        if $D' < D(G_n)$ **then** $p_{\min} \leftarrow p'$ **else** $p_{\max} \leftarrow p'$
    return $p_{\min}$

Our estimation procedure can be summarized follows:

- 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.
- With $G_n \sim \texttt{DD-model}(n, \widehat{p}, \widehat{r}, G_{n_0})$, we find the tolerance interval of $\widehat{r}$ using the confidence interval of $D(G_n)$ and $C_3(G_n)$.
- 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*.
- 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}.