Modeling Biological Networks with Exponential Random Graph Models

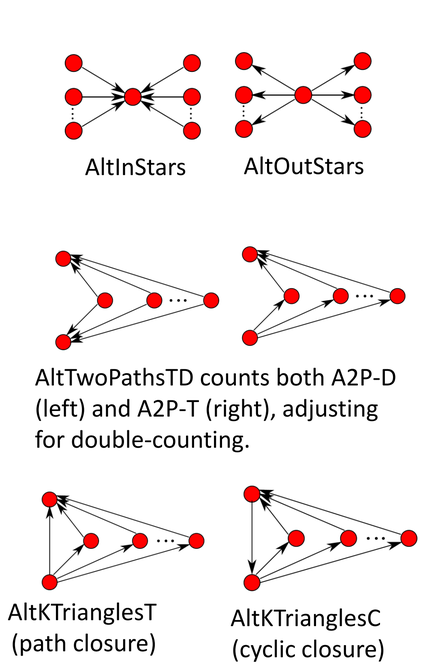
Alex Stivala[1,2], Maksym Byshkin[2], Antonietta Mira[2,3], Garry Robins[4], Alessandro Lomi[2,4]

[1]Swinburne University of Technology, Australia. [2]Università della Svizzera italiana, Lugano, Switzerland
[3]Università dell'insubria, Como, Italy [4]The University of Melbourne, Australia

Description	Nodes	Components	Size of largest component	Mean degree	Density	Clustering coefficient	Mean path length
A. thaliana PPI	2160	165	1632	3.70	0.00171	0.06645	6.72
Yeast PPI	2617	92	2375	9.06	0.00346	0.46862	5.10
Human PPI	4303	135	4100	6.24	0.00145	0.03326	4.06
C. elegans PPI	5038	87	4847	5.14	0.00102	0.05818	4.49
E. coli regulatory	418	29	328	2.48	0.00596	0.02382	4.82
Drosophila optic medulla	1781	6	1770	10.01	0.00562	0.06922	2.91

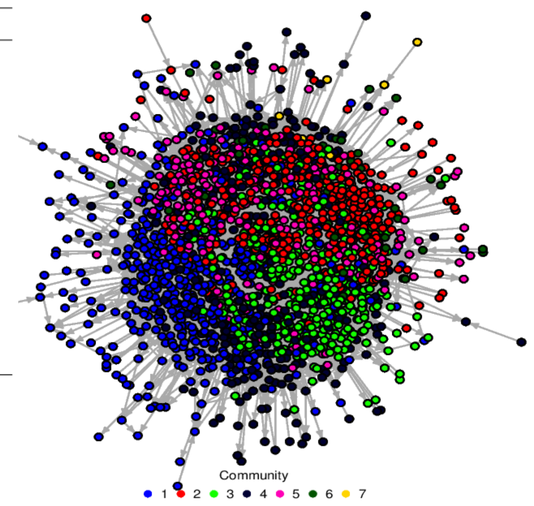
"Motifs" are often considered building blocks of complex biological networks. Exponential Random Graph Models (ERGMs) are a well-established class of statistical models widely used in social network analysis, which can determine the over or under-representation of such motifs (by significantly positive or negative estimated model parameters). Despite being introduced in the bioinformatics literature ten years ago, their use in biology has been limited due to networks being too large to be estimated with existing methods.

We used high performance computing to apply our recently developed new techniques (snowball sampling, improved fixed density (IFD) ERGM sampling, and the scalable Equilibrium Expectation (EE) algorithm) to several protein-protein interaction (PPI) networks, a gene regulatory network, and a neural network.



Effect	Model 1	Model 2
Arc	−10.040 (−10.076,−10.004)	−10.036 (−10.071,−10.002)
AltInStars	1.391 (1.375,1.407)	1.378 (1.365,1.392)
AltOutStars	0.868 (0.840,0.896)	0.870 (0.848,0.893)
Reciprocity	1.368 (1.319,1.416)	1.597 (1.555,1.639)
AltTwoPathsTD	−0.020 (−0.022,−0.019)	−0.020 (−0.021,−0.019)
AltKTrianglesT	1.316 (1.303,1.329)	1.368 (1.351,1.386)
AltKTrianglesC	—	−0.085 (−0.108,−0.063)

Drosophila optic medulla estimation results from EE. There is preferential attachment on both in- and out- degree. Path closure is over-represented but cyclic closure under-represented.



Drosophila optic medulla network with nodes coloured according to network community from Louvain algorithm.

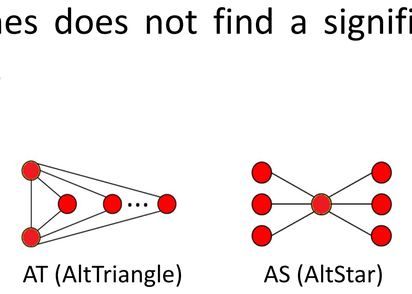
We gratefully acknowledge support from PASC, from Swiss NSF NRP 75 (project 167326) and from Melbourne Bioinformatics at the University of Melbourne (grant number VR0261).

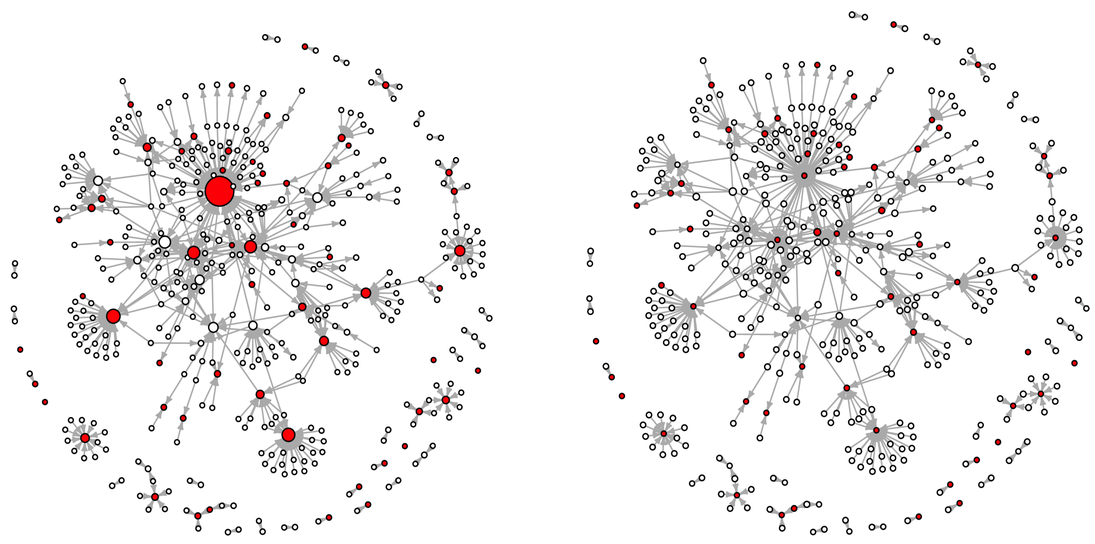
References

- Byshkin, M., Stivala, A., Mira, A., Krause, R., Robins, G., & Lomi, A. (2016). Auxiliary parameter MCMC for exponential random graph models. *Journal of Statistical Physics*, 165(4), 740-754.
- Byshkin, M., Stivala, A., Mira, A., & Lomi, A. (2018). Fast maximum likelihood estimation via Equilibrium Expectation for large network data. *arXiv preprint* arXiv:1802.10311.
- Saul, Z. M., & Filkov, V. (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19), 2604-2611.
- Stivala, A. D., Koskinen, J. H., Rolls, D. A., Wang, P., & Robins, G. L. (2016). Snowball sampling for estimating exponential random graph models for large networks. *Social Networks*, 47, 167-188.

Method	Network	Average sample size	N_c	Avg. estim. time (m)	Elapsed time
EE (IFD sampler)	*A. thaliana* PPI	2160	20	1.1	01 m 50 s
EE (IFD sampler)	Yeast PPI	2617	20	6.6	09 m 07 s
EE (IFD sampler)	Human PPI	4303	20	7.6	10 m 49 s
EE (IFD sampler)	*C. elegans* PPI	5038	20	6.8	09 m 35 s
EE (IFD sampler)	*E. coli* regulatory	418	20	0.6	00 m 43 s
EE (IFD sampler)	*Drosophila* optic medulla	1781	20	4.3	06 m 22 s
SA (IFD sampler)	*A. thaliana* PPI	2160	20	9.2	0 h 34 m 02 s
SA (IFD sampler)	Human PPI	4303	20	49.2	2 h 46 m 54 s
SA (IFD sampler)	Yeast PPI	2617	20	45.6	2 h 02 m 38 s
SA (IFD sampler)	*C. elegans* PPI	5038	20	766.5	25 h 07 m 44 s
SA (IFD sampler)	*E. coli* regulatory	418	20	0.0	0 h 00 m 06 s
SA (IFD sampler)	*Drosophila* optic medulla	1781	20	824.6	72 h 40 m 00 s
SA (basic sampler)	*A. thaliana* PPI	2160	0	—	(time limit)
SA (basic sampler)	Yeast PPI	2617	0	—	(time limit)
SA (basic sampler)	Human PPI	4303	0	—	(time limit)
SA (basic sampler)	*C. elegans* PPI	5038	3	204.5	7 h 40 m 20 s
SA (basic sampler)	*E. coli* regulatory	418	20	1.1	0 h 04 m 06 s
SA (basic sampler)	*Drosophila* optic medulla	1781	0	—	(time limit)
Snowball sampling	*A. thaliana* PPI	490.6	19	26.3	2 h 08 m 24 s
Snowball sampling	Yeast PPI	264.8	19	30.2	3 h 40 m 34 s
Snowball sampling	Human PPI	822.5	18	47.0	3 h 50 m 27 s
Snowball sampling	*C. elegans* PPI	496.4	16	270.7	40 h 00 m 33 s
Snowball sampling	*Drosophila* optic medulla	649.7	15	118.0	7 h 22 m 48 s

Estimation times for undirected network models on 20 Intel Haswell compute cores (2.3 GHz) on a Lenovo NeXtScale x86 cluster. The elapsed time limit was 99 hours. N_c is the number of converged runs. SA is "stochastic approximation", a widely used ERGM estimation algorithm. Although it is many times faster, the EE algorithm estimates are consistent with those from the MCMC MLE methods. Snowball sampling sometimes does not find a significant effect that other methods do.





E. coli regulatory network. Self-regulating operons in red. Node size is proportional to in-degree (left) and out-degree (right).

Effect	Model 1	Model 2	Model 3	Model 4
Arc	−8.390 (−8.655,−8.125)	−7.986 (−8.258,−7.715)	−7.846 (−8.133,−7.560)	−6.745 (−7.096,−6.394)
AltInStars	2.396 (2.282,2.510)	2.207 (2.091,2.324)	2.241 (2.120,2.362)	1.981 (1.842,2.120)
AltOutStars	−0.708 (−0.919,−0.498)	−0.858 (−1.076,−0.639)	−0.706 (−0.918,−0.495)	−0.370 (−0.627,−0.114)
AltTwoPathsTD	—	—	—	−0.988 (−1.155,−0.821)
AltKTrianglesT	1.252 (1.073,1.431)	1.477 (1.306,1.648)	1.250 (1.069,1.430)	2.299 (2.112,2.486)
Sender self	—	−1.494 (−1.863,−1.125)	—	—
Receiver self	—	0.425 (0.354,0.496)	—	—
Matching self	—	—	−0.515 (−0.625,−0.406)	−0.426 (−0.523,−0.330)

E. coli regulatory network estimation results from EE. There is centralization on in-degree but not out-degree. Transitive closure (feed-forward loop) is over-represented. Self-regulating operons are more likely to be regulated than to regulate others.