Statistical and computational methods for bioinformatics and social network analysis

or how did I learn to stop worrying and love the bomb

George G Vega Yon

University of Southern California, Department of Preventive Medicine

October 2, 2019

Paper 1: Exponential Random Graph Models for Small Networks

Paper 2: On the prediction of gene functions using phylogenetic trees

Future directions

Things that are very interesting but I most probably won't have any time to discuss with the attendees

References

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka ERGMs are:

What are Exponential Random Graph Models

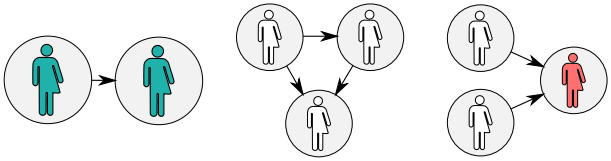
Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks
- In simple terms: statistical inference on what network patterns/structures/motifs govern the data-generating process



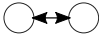
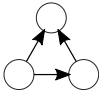
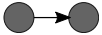
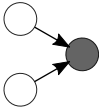
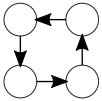
Representation	Description
	Mutual Ties (Reciprocity) $\sum_{i \neq j} y_{ij} y_{ji}$
	Transitive Triad (Balance) $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$
	Homophily $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$
	Covariate Effect for Incoming Ties $\sum_{i \neq j} y_{ij} x_j$
	Four Cycle $\sum_{i \neq j \neq k \neq l} y_{ij} y_{jk} y_{kl} y_{li}$

Figure: Besides of the common edge count statistic (number of ties in a graph), ERGMs allow measuring other more complex structures that can be captured as sufficient statistics.

What are Exponential Random Graph Models: State of the Art

What are Exponential Random Graph Models: State of the Art

Small-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic approximation.
- Maximum Pseudo Likelihood (MPLE)

What are Exponential Random Graph Models: State of the Art

Small-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic approximation.
- Maximum Pseudo Likelihood (MPLE)

large-huge networks (up to the millions of vertices)

- Semi-parametric bootstrap
- Conditional joint estimation (like snowball sampling, a.k.a. divide and conquer)
- Equilibrium Expectation Algorithm (millions of vertices)

What are Exponential Random Graph Models: State of the Art

Small-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic approximation.
- Maximum Pseudo Likelihood (MPLE)

large-huge networks (up to the millions of vertices)

- Semi-parametric bootstrap
- Conditional joint estimation (like snowball sampling, a.k.a. divide and conquer)
- Equilibrium Expectation Algorithm (millions of vertices)

But who cares about tiny to small networks?

What are Exponential Random Graph Models: The MC-MLE approach

One of the most popular methods for estimating ERGMs is the MC-MLE approach (citations here)

This consists on the folling steps

1. Start from a sensible guess on what should be the population parameters (usually done using pseudo-MLE esimtation)
2. While the algorithm doesn't converge, do:
   - 2.1 Simulate a stream of networks with the current state of the parameter, θ_t
   - 2.2 Using the law of large numbers, approximate the ratio of likelihoods based on the parameter θ_t, this is the objective function
   - 2.3 Update the parameter by a Newton-Raphson step
   - 2.4 Next iteration

What are Exponential Random Graph Models: The MC-MLE approach

MC-MLE works great (we have some simulations showing this), but it has some problems:

- While lots of advances have been made, there are restrictions on what can be done with it, after all, it is an approximation,
- In the case of small networks, issues regarding near-degeneracy during estimation are common (unstable MCMC process, bad sampling, problems)

What shall we do then?

Exponential Random Graph Models for Small Networks

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible[1]
- In the case of networks with 5 nodes, 1,048,576 different configurations, we can compute the likelihood exactly.

Using the exact likelihood opens a huge window of methodological-possibilities

[1] A thing mentioned in literature several times, although not much attention paid

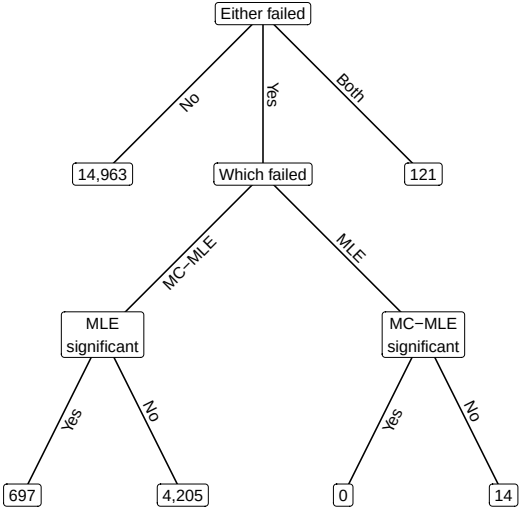
The ergmito

Paper 1 Simulation Studies

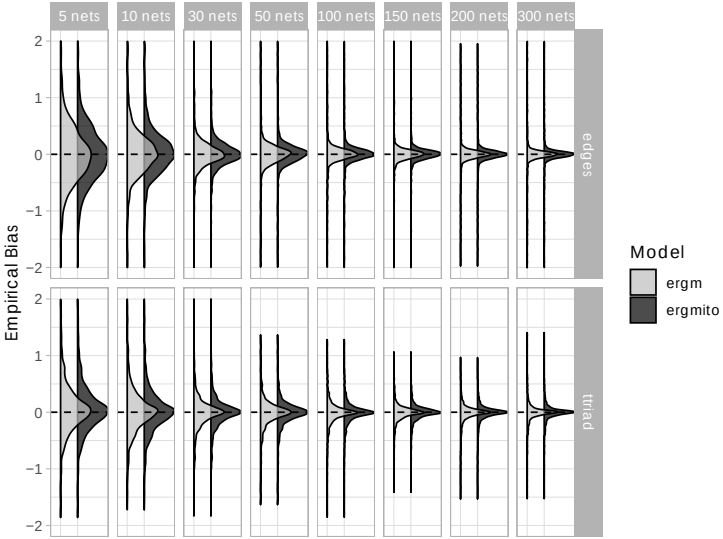
In order to compare the MLE with the MC-MLE estimation method, we performed a simulation study with the following features:

- Draw 20,000 samples of groups of small networks
- Each group had prescribed: (model parameters, number of networks, sizes of the networks)
- We estimated the models using the ergm and ergmito R packages

Paper 1 Simulation Studies: Error rate



Paper 1 Simulation Studies: Empirical Bias

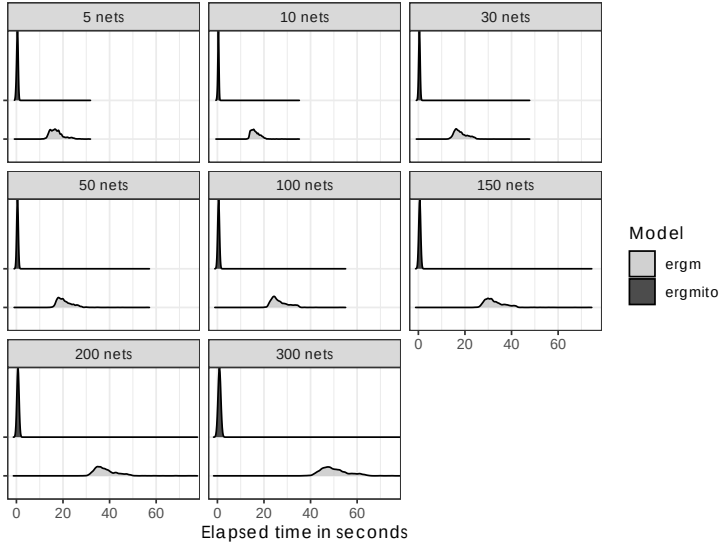


Paper 1 Simulation Studies: Empirical Type I error

		P(Type I error)		
Sample size	N. Simulations	MC-MLE	MLE	chi2
5	2,189	0.084	0.057	11.71 ***
10	2,330	0.070	0.045	12.46 ***
15	2,395	0.084	0.066	5.55 *
20	2,430	0.074	0.060	3.58
30	2,460	0.057	0.052	0.67
50	2,495	0.046	0.044	0.17
100	2,499	0.048	0.048	0.00

Table: Empirical Type I error rates. The χ^2 statistic is from a 2-sample test for equality of proportions, and the significance levels are given by *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$. The lack of fitted samples in some levels is due to failure of the estimation method.

Paper 1 Simulation Studies: Elapsed time



Paper 1: Exponential Random Graph Models for Small Networks

Paper 2: On the prediction of gene functions using phylogenetic trees

Future directions

Things that are very interesting but I most probably won't have any time to discuss with the attendees

References

Phylogenetic Trees

- It can be very general: think of the tree of life
- Nowadays, thanks to gene-sequencing techniques, we are building trees at the gene level (using sequence-alignment methods, i.e. comparing gene sequences to see how much similar/different two genes are between and within species (whattt!)).
- A single phylogenetic tree can host multiple species

A common phylogenetic tree

Gene Functional Annotations

The Gene Ontology Project

Accession	GO:0060047
Name	heart contraction
Ontology	biological_process
Synonyms	heart beating, cardiac contraction, hemolymph circulation
Alternate	IDs None
Definition	The multicellular organismal process in which the heart decreases in volume in a characteristic way to propel blood through the body. Source: GOC:dph

source: http://amigo.geneontology.org/amigo/term/GO:0060047

Speciation

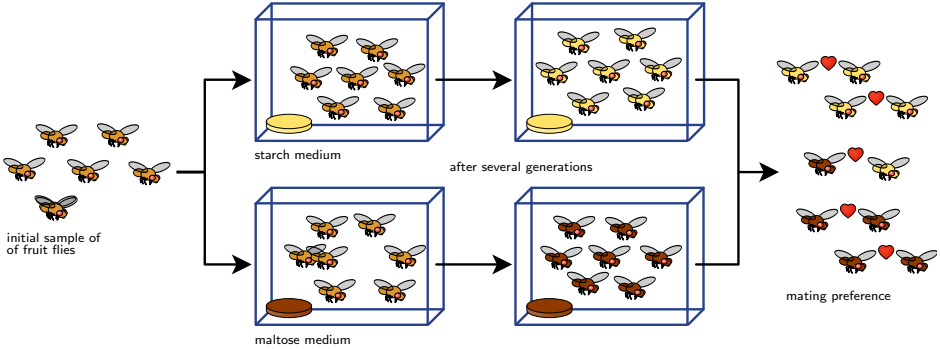


Figure: Dodd (1989): After one year of isolation, flies showed a significant level or assortativity in mating (wikimedia)

Duplication

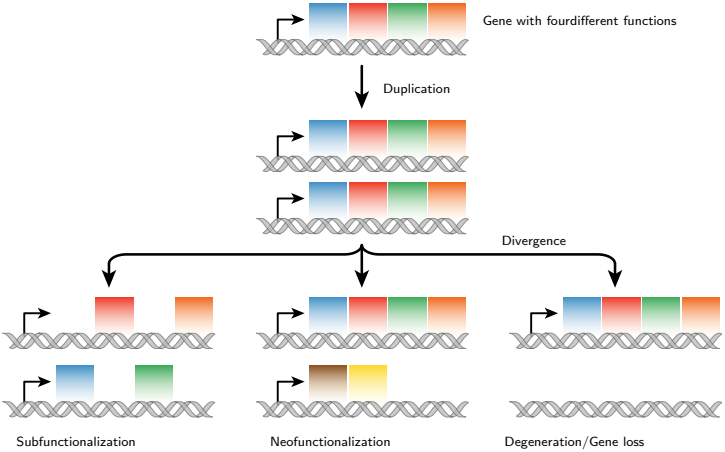


Figure: A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge (wikimedia)

An evolutionary model of gene functions

The general points of the model

- The rootnode in a phylogenetic tree is the best idea we have about the past, meaning, it could be that the tree has more behind, i.e. so functions may be gained since the beginning
- At each step in evolution (interior node), there is a probability that the gene may gain/loss the function
- Those probabilities vary depending on the type of the node: We belive that functional changes may happen at Duplication nodes
- That's it!

An evolutionary model of gene functions: Formal statement

The whole is based on the markov-assumption: The current state of the gene can be fully explained by its parent(s).

For this we use Felsensteins' pruning algorithm (also known as...)

Formally

$$P(x = 1) = P(x = 1|x_p = 0)P(\text{Gain}) + P(x = 1|x_p = 1)P(\text{No loss})$$

The aphylo

Future directions: ERGMitos

- Identify an adequate test for goodness-of-fit assesment
- Extend to estimation of large graphs by splitting the networks in induced-subgraphs

Future directions: Gene functional prediction

Possible venues to continue

- Incorporate more external information using leaf(and node?) level features.
- Adapt the model to incorporate joint estimation of functions using pseudo-likelihood.

$$P(a, b, c) \approx P(a, b)P(b, c)P(a, c)$$

- Make the model hierarchical when pooling trees: different mutation rates.

Here are some by-products of my research here at USC

- The slurmR R package
- The pruner C++ library
- The fmcmc R package

References I

Dodd, D. M. B. (1989). Reproductive isolation as a consequence of adaptive divergence in drosophila pseudoobscura. *Evolution*, *43*(6), 1308–1311. Retrieved from http://www.jstor.org/stable/2409365