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 1, 2019

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

What are Exponential Random Graph Models: State of the Art

Currently, in the ERGMs world we have

- ▶ Estimation of small-large (dozens to a couple of thousand vertices) networks is done using MCMC based approaches.
- ▶ Estimation of large-huge networks (up to the millions of vertices) is done using approximation methods like

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

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

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

- ASDAS

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