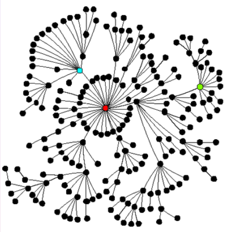
A statistically rigorous way to analyse network topology models



Richard G. Clegg (richard@richardclegg.org) (UCL)

Cosener's NGN 2009

(Prepared using LaTeX and beamer.)

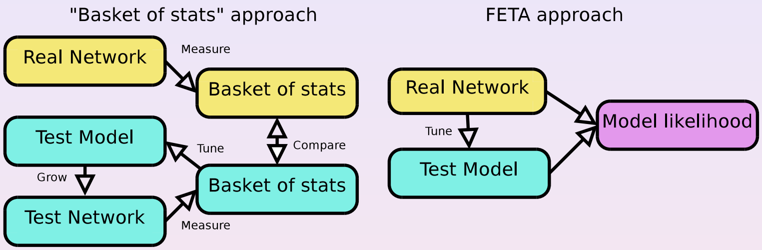
Introduction

Growing artificial networks

- Want to grow networks with same properties as real networks.
- Want to be able to describe evolution of the real network.
- Want to assess simple processes which explain the evolution of the network.
- Want to be able to compare rival theories about the evolution.

- Background: scale free networks, Preferential Attachment, PFP, GLP models.
- Use historic data on evolution.
- FETA – Framework for Evolving Topology Analysis.
- Framework for comparing models not to give best model.
- Single rigorous statistic not many indicative ones.

FETA approach



Inner model evaluation

- For simplicity consider graphs which evolve using only the "connect to new node" operation.
- Let θ be some candidate inner model – a map from node numbers to probability distribution.
- Model must explain observed node choices $C = N_1, N_2, \ldots, N_t$.
- Want to compare θ with rival model θ' or with null model θ_0.
- Let $p_j(k|\theta)$ be the probability node k is chosen at stage j (based on graph at this stage and possibly other factors).

Likelihood of observed choices C

The likelihood of the observed node choices C given model θ is

$$L(C|\theta) = \prod_{j=1}^{t} p_j(N_j|\theta).$$

Building models from components

- Inner model θ could be built from components:
  1. θ_d Preferential attachment model – prob. prop. to degree d.
  2. $\theta_p(\delta)$ the PFP model with δ parameter –prob. prop. to $d^{(1+\delta \log_{10}(d))}$.
  3. θ_S singleton model – prob. const. for degree $= 1$ or 0 otherwise.
  4. $\theta_r(N)$ the "recent" model – prob. const. for nodes picked in the last N choices or 0 otherwise.

Example model from components

$$\theta = \beta_S \theta_S + \beta_p \theta_p(\delta) + \beta_r \theta_r(N),$$

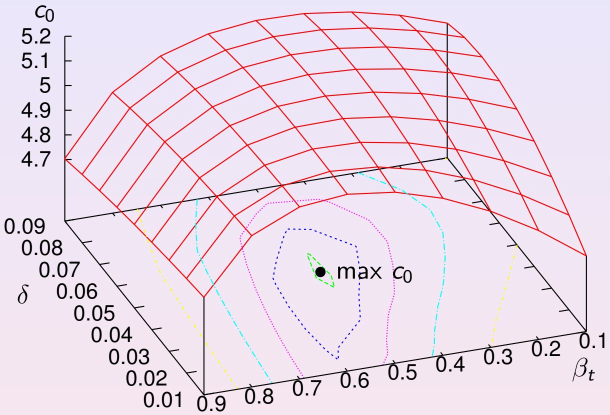
where $\beta_\bullet \in (0, 1)$ and $\beta_S + \beta_p + \beta_r = 1$.

Need to optimise β_S, β_p, β_r, δ and N!

Artificial tests – parameter sweep

- The most convincing test of such a model is its ability to recover parameters from a known model.
- Consider the inner model $\theta = 0.5\theta_p(0.05) + 0.5\theta_t$ (PFP + triangles).
- Remember for PFP prob. of connecting to node i is $p_i \sim d_i^{1+\delta \log_{10} d_i}$ for triangles prob is proportional to node triangle count.
- Outer model is simple – node connects to three nodes.
- Create a test network of 10,000 nodes .
- Now try to recover "unknown" δ and β parameters
- Measure c_0 – ratio of likelihood versus θ_0 normalised by $|C| = t$,
- Find δ and β_t to maximise c_0.

Two dimensional parameter sweep for $\beta_p\theta_p(\delta) + \beta_t\theta_t$



Max c_0 at $\delta = 0.0525$ and $\beta_t = 0.5$.

Artificial tests – General linear models

- Test model $\theta = 0.25\theta_0 + 0.25\theta_t + 0.25\theta_S + 0.25\theta_D$.
- Here the GLM is tested with an additional spurious model component θ_d (preferential attachment).
- The θ_d component is rejected.

Parameter	Estimate	Significance
β_0	0.33 ± 0.059	0.1%
β_t	0.29 ± 0.017	0.1%
β_S	0.24 ± 0.016	0.1%
β_D	0.23 ± 0.022	0.1%
β_d	-0.089 ± 0.059	5%

Real data tests

- Tests have been performed on five real networks – two from social networks (photo sharing), two models of the internet AS and one publication network (arxiv).
- Model sizes varied from 15,788 links to 98,931.
- Hypothetical models are created from components using GLM and their c_0 measured.
- Claim is that the c_0 is a good predictor of success at predicting network.
- Test three candidate models "random" (θ_0), "best PFP" (PFP model with optimised δ) and "best" (best combination of submodels found.
- Calculate "best model" using c_0 value.
- Grow artificial models and measure sample network statistics.

Real data results

- In all networks tested, c_0 was an excellent predictor of how well an artificial network would replicate statistics.
- It is much quicker to measure c_0 than to grow an artificial network and measure statistics.
- The sub models tested here did not perfectly replicate all network statistics (but then that was not the aim).
- In particular the sub models I use now do not capture clustering or assortativity well.
- If the data is available then this likelihood statistic is the way we should be assessing potential network models.
- The c_0s statistic is a single, fast and rigorous measure of network likelihood.

Further work

Take home messages

- Likelihood measures are the way to assess network models.
- New network models created from combining sub models.
- Standard statistics techniques (GLM) can optimise submodel weights.

-
- Software and data freely available – see website `http://www.richardclegg.org/software/FETA`
- I am very keen to collaborate – give me your network and I will analyse it for you.