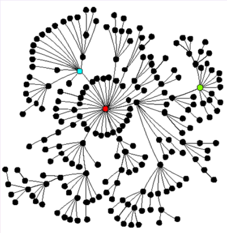
Probabilistic models for evolving network topologies



Richard G. Clegg (richard@richardclegg.org) – work developed with R. Landa and M. Rio, Dept. of Electronic and Electrical Engineering, UCL

Talk to QMUL 2009

(Prepared using LaTeX and beamer.)

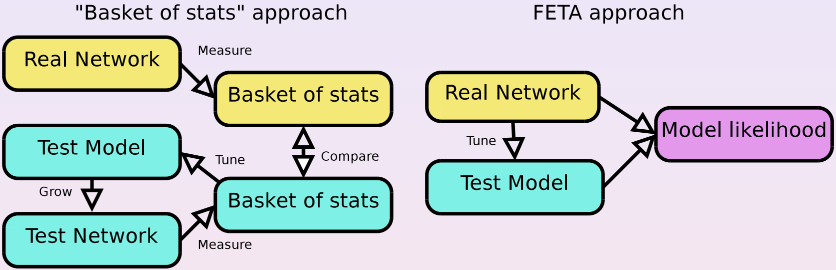
Introduction

- There is much work on creating models which “grow” artificial networks to match real ones.
- Existing models: Erdős–Rényi, Preferential attachment, Positive feedback preference (PFP) and General Linear Preference (GLP).
- How can new models be evaluated and compared?

FETA – a framework for evolving topology analysis

- Statistically rigorous approach to assessing models which generate artificial topologies to match real data.
- Comparison with a null model specific to network growth.
- Ability to automatically “optimise” some model parameters.
- Uses (requires) growth data about network.

FETA approach



The FETA general topology model

Outer model

- Process to select an operation on the network.
- Could be: add node, add edge, remove node and so on.
- Currently two: connect edge(s) to new node and add edge between existing nodes.

Inner model

- Process selects node or edge for operation.
- Probabilities are assigned to nodes and potential edges for random selection.
- Edges selected by assigning probabilities to node pairs.
- FETA focuses exclusively on the inner model.

Inner model evaluation

- For simplicity consider graphs which evolve using only the "connect to new node" operation.
- Let G_0 be some known starting graph and assume that $G_1, \ldots, G_t$ are also known.
- From G_{i-1} and G_i we can infer N_i the node selected at stage i of construction.
- Let θ be some candidate model – assigns node probabilities.
- Let θ_0 be the null model – all node probabilities equal.
- Probabilities assigned based on graph properties plus possible exogenous inputs.

Inner model evaluation (2)

- Let $p_j(i|\theta)$ be the probability that θ assigns to node i for choice j (based on G_{j-1}).
- At choice j node N_j was selected – the likelihood of this selecion given θ is $p_j(N_j|\theta)$.
- Want likelihood of observed choices $C = N_1, \ldots, N_t$.

Likelihood of observed choices C

The likelihood of the observed node choices C inferred from the graphs $G_0, G_1, \ldots, G_t$ is given by

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

Useful statistics

- Log likelihood – $l(C|\theta) = \log(L(C|\theta)) = \sum_{j=1}^{t} \log[p_j(N_j|\theta)]$.
- Per choice likelihood ratio c_A – ratio of likelihood versus model θ_A normalised by $|C| = t$,
$c_A = \left[\frac{L(C|\theta)}{L(C|\theta_A)}\right]^{1/t} = \exp\left[\frac{l(C|\theta) - l(C|\theta_A)}{t}\right]$.
- If a model has $c_A > 1$ is better explains the choice set C than model A.
- Particularly useful c_0 the per choice likelihood ratio relative to the null (random) model θ_0.

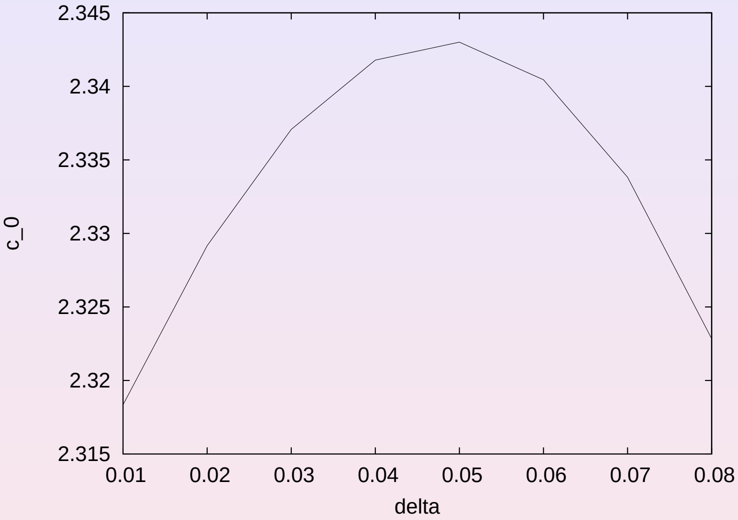
Combining and automatically fitting models

- A node choice model θ could be built from component models such as:
  1. θ_d Preferential attachment model (probability proportional to node degree).
  2. $\theta_p(\delta)$ the PFP model (with delta parameter).
  3. θ_T triangle model (probability proportional to no of triangles node is in).
  4. θ_1 singleton model (probability constant for nodes with degree 1 or 0 otherwise).
- $\theta = \beta_d\theta_d + \beta_p\theta_p(\delta) + \beta_T\theta_T$ is a valid model if $\beta \in (0, 1)$ and $\sum \beta = 1$.
- The β parameters can be tuned using generalised linear model (GLM) fitting techniques.
- Non linear parameters such as δ can be tuned using c_0 and state space search.

Artificial data tests

- Generate 10,000 link test network with $\theta = 0.5\theta_d + 0.5\theta_1$ (pref. attach. + singletons).
- Model $\theta = 0.5\theta_d + 0.5\theta_1$ has $c_0 = 7.40$.
- Fitting model $\beta_d\theta_d + \beta_1\theta_1$ using GLM gives $\beta_d = 0.47 \pm 0.03$ and $\beta_1 = 0.53 \pm 0.3$.
- This model has $c_0 = 7.39$ (almost indistinguishable).
- Fitting model $\beta_T\theta_T + \beta_d\theta_d$ (triangles + pref attach) gives $\beta_T = -0.00024 \pm 0.00050$ and $\beta_d = 1.0 \pm 0.042$ – essentially θ_d.
- The model θ_d has $c_0 = 0.727$ – worse than random model θ_0.
- Fitting model $\beta_d\theta_d + \beta_0\theta_0$ (pref. attach. + random) gives the illegal model $\beta_0 = 1.07 \pm 0.075$ and $\beta_d = -0.077$.
- The final model fit also says that θ_d has no statistical significance to the fit. This is because that model alone is a worse model than θ_0.

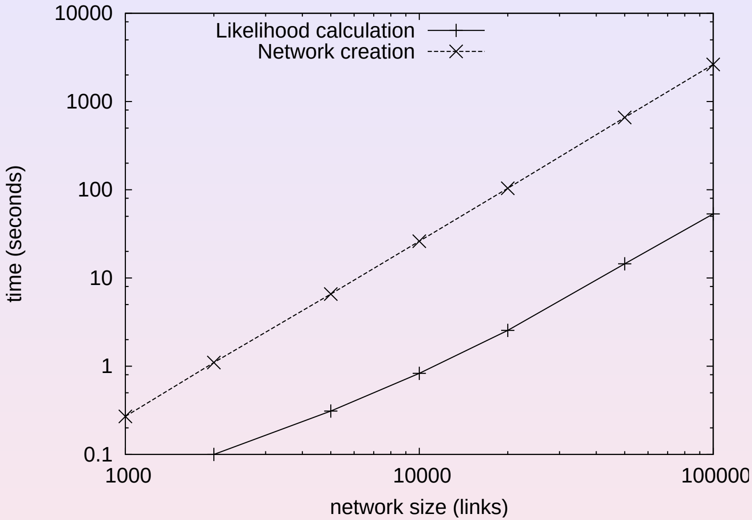
Delta sweep to recover known PFP δ parameter 0.05 (10,000 nodes)



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.
- Obviously for real networks we cannot know the true underlying model.
- Various hypothetical models were tested on the real network using a "basket of statistics".
- Those models with higher c_0 performed better when judged by the "basket of statistics".
- Interpreting which was the better from two models with close c_0 was often tricky.
- PFP was the most successful model component tried – δ close to zero for connection between inner nodes.

Runtime of likelihood estimate versus network creation



Conclusions and further work

- The likelihood parameters and the null model here provide a rigorous way to assess a potential dynamic model of network evolution.
- A GLM approach can be used to optimise parameters in linear combinations of models.
- In tests on artificial models the optimisation can recover parameters from linear combinations of models.
- Further work could improve the outer model (currently very simple).
- Multiplicative model combinations might have greater success: $\theta = K\theta_d^{\beta_d}\theta_T^{\beta_T}\cdots$.
- Software and data freely available – please email `richard@richardclegg.org`.