

A Likelihood Based Framework for Assessing Network Evolution Models Tested on Real Network Data

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

Raul Landa (UCL)

Uli Harder (Imperial), Miguel Rio (UCL)

SIMPLEX conference, Venice 2009

(Prepared using \LaTeX and beamer.)

Introduction

Growing artificial networks

- Want to grow networks with the **same properties** as real networks.
 - Want to be able to describe the **evolution** of the real network.
 - Want to test and compare the abilities of simple models hypothesised as explanations of complex network evolution.
-
- How do we know which properties are important?
 - If we have historic data about the network can this be used?
 - What if the growth process changes part way through?
 - Here we propose FETA – Framework for Evolving Topology Analysis.

The “basket of statistics” approach

- Current approach – call it the “basket of statistics” method.
 - ① Select several statistics which can be measured on network snapshot (mean degree, scaling parameter, clustering coeff, etc etc etc).
 - ② Use test model to grow test network (same size as real network).
 - ③ Compare the “basket of statistics” on real and test data.
- New statistics motivate new models – but what if not all stats match?

Problem to solve

Need a statistically sound framework to compare and test models. This should use growth information. The framework will also be able to tune parameters (automatically?). This framework will be a test-bed for future network models.

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, \dots, 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.
- An inner model is a map from some graph and node number to a probability. It assigns a probability (may be zero) to each node and this totals to one.
- Let θ be some hypothesised inner 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 – function may be time varying and may depend on previous choices.

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 selection given θ is $p_j(N_j|\theta)$.

Likelihood of observed choices C

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

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

Useful statistic c_0 (per choice likelihood ratio) – ratio of likelihood versus null normalised by $|C| = t$,

$$c_0 = \left[\frac{L(C|\theta)}{L(C|\theta_0)} \right]^{1/t} = \exp \left[\frac{l(C|\theta) - l(C|\theta_0)}{t} \right].$$

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_t triangle model – prob. prop. to Δ count.
 - ④ θ_S singleton model – prob. const. for degree = 1 or 0 otherwise.
 - ⑤ $\theta_r(N)$ 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 to maximise c_0 (per choice likelihood ratio).

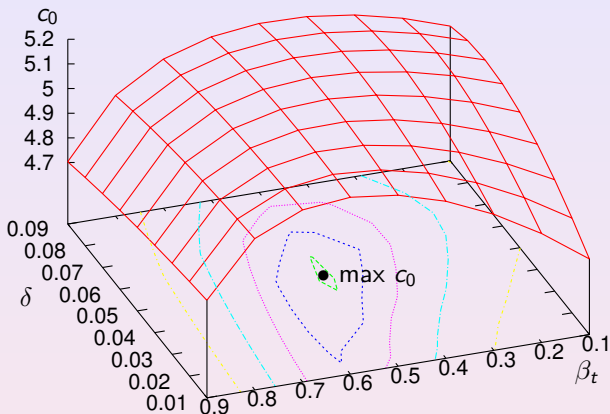
A GLM approach to optimise β parameters

- Want to fit $p_i = \beta_1\theta_1 + \beta_2\theta_2 + \dots + \varepsilon$ to data.
- p_i is not known, only whether the node was “picked”.
- Define I_i an indicator variable.
- For each node choice step:
 - 1 For each node record the relevant parameters at that step (degree, triangle coefficient, age of node and so on).
 - 2 Record a 1 for I_i if node i was “picked” at this step.
 - 3 Record a 0 for I_i if node i was not “picked” at this step.
- $E[I_i] = p_i$ – the expectation of I_i is the probability i would be chosen by the model underlying the graph evolution.
- Fitting $I_i = \beta_1\theta_1 + \beta_2\theta_2 + \dots$ for all possible nodes for a given choice and for all known choices optimises the β .

Artificial tests

- 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 trianle 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 for models of the form $\beta_t\theta_t + (1 - \beta_t)\theta_p(\delta)$ with various δ and β_t values.
- 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 $\beta_t = 0.5$.

GLM procedure with incorrect model

- 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%

General comments on GLM procedure

- Works well to recover parameters to known model.
- Can have issues when model components express “similar” things (e.g. PFP and preferential attachment in same model).
- Acts as a guide to the user as to which model components to include and which to reject.
- Does not allow testing of non-linear parameters (e.g. δ) but can be combined with “parameter sweep”.
- Ultimately though, the likelihood estimate c_0 is the arbiter of which model is correct.

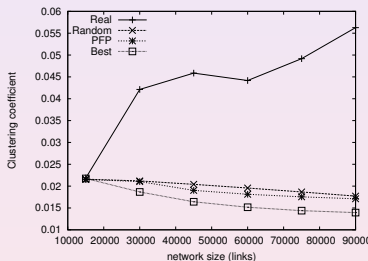
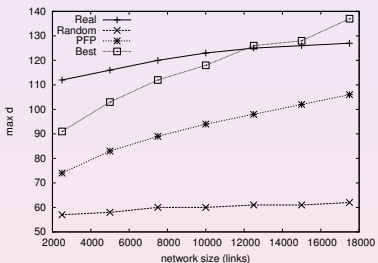
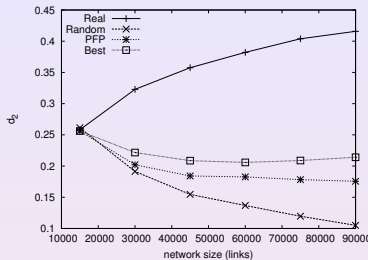
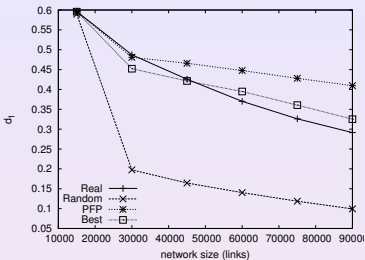
Real data tests

- 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.
- Inner models compared are “random”, PFP (pref attach is PFP with $\delta = 0$) and “best” combination of submodels.
- Outer model is “copy” of real outer actions to remove it from equation.
- Real network stats are compared with artificial.

Real data example – UCLA AS level internet

- UCLA AS level internet – partially complete evolution of the internet Autonomous System network.
- The best pure PFP model was $\theta_p(0.005) - c_0 = 4.81$.
- Best model had $0.82\theta_p(0.014) + 0.18\theta_R(1)$ (PFP + recent) for connections to new nodes.
- Best model had $0.71\theta_d + 0.22\theta_R(1) + 0.07\theta_S$ (preferential attachment + “recent” + singleton) for connections between existing nodes.
- Best model had $c_0 = 8.06$.
- Real networks grown with random, PFP and “best” and stats compared throughout evolution.
- Stats measured included prop. of nodes of degree one and two, maximum node degree, mean square node degree, clustering coefficient and assortativity.

UCLA results – network plots



Real data conclusions

- For all the networks the c_0 values were an excellent predictor of the order in which the models reproduced network statistics.
- The combinations of submodels considered in this paper were usually usually unable to reproduce assortativity and clustering coefficient.
- Other network statistics considered were reproduced well.
- Either new submodels or changes to the outer model (addition of cliques) to improve performance.
- Optimising among the submodels with the aid of GLM was relatively fast and all combinations could be tried in manageable time.
- The likelihood of many models could be tested in the time taken to “grow” and measure statistics on one artificial network.

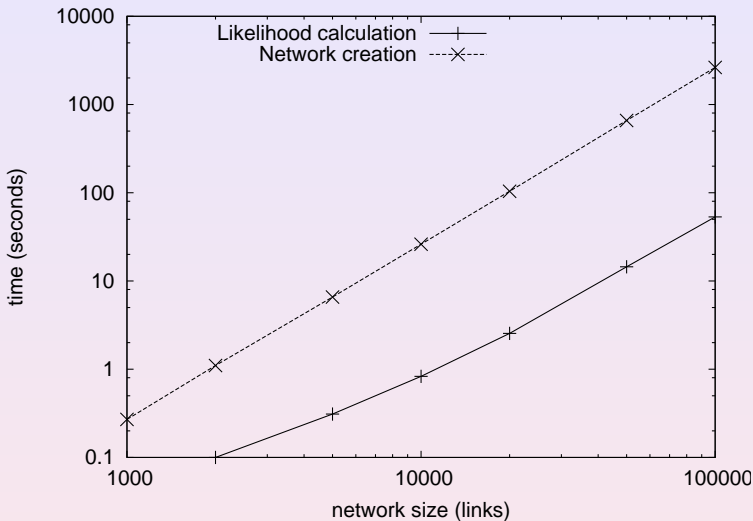
Conclusions

- The likelihood parameters and the null model here provide a rigorous way to assess a potential dynamic model of network evolution.
- Known model parameters can be recovered using sweeps of likelihood or GLM for linear parameters.
- The likelihood is reflected in improved performance on replicating network statistics.
- The advantages of this framework are several:
 - ① Assesses the dynamic history of the data not statistics of a snapshot.
 - ② Single statistically rigorous estimate of model likelihood.
 - ③ Quicker than growing a network and testing statistics (using same codebase).
- An exciting new way to test theories about topologies if you have the data for it.

Further work

- What model components can be added (particularly for assortativity and clustering).
- More data must be found – currently data from transport networks and biological systems is being investigated.
- Further work must be done on the outer model.
- Multiplicative model combinations might have greater success:
$$\theta = K\theta_d^{\beta_d}\theta_T^{\beta_T}\dots$$
- Software and data freely available – please email richard@richardclegg.org
- See also the website
<http://www.richardclegg.org/software/FETA>
- I am very keen to collaborate – give me your network and I will analyse it for you.

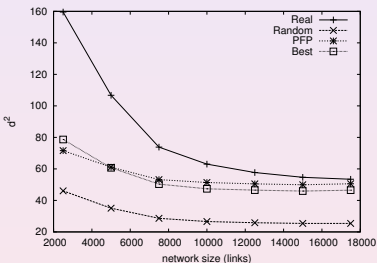
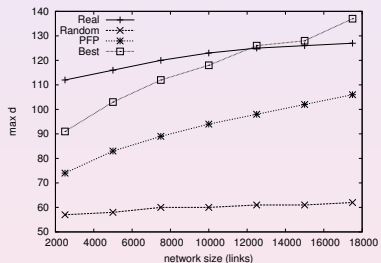
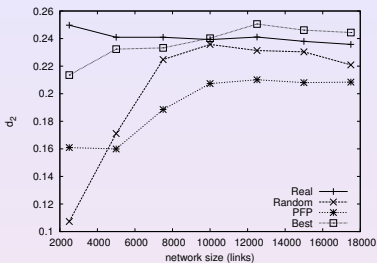
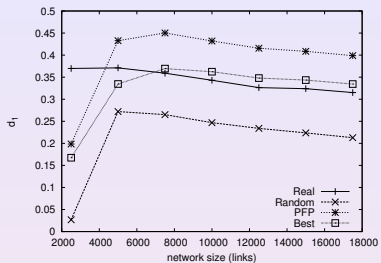
Runtime of likelihood estimate versus network creation



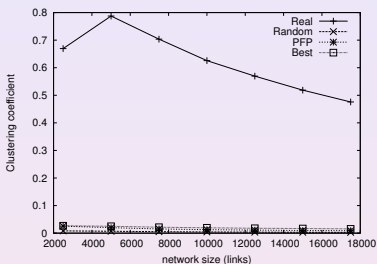
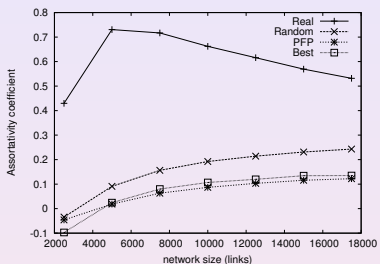
arXiv modelling

- arXiv co-authorship network for “math” library.
- Approx 17,500 links representing two authors on same paper.
- Outer model as before.
- Random model θ_0 – obviously has $c_0 = 1$.
- Best “pure PFP” model $\theta_p(-0.005)$ (negative δ parameter common in “human” networks) – has $c_0 = 1.31$.
- “Best” model found has $c_0 = 6.25$.
 - New node connections $0.56\theta_p(-0.29) + 0.28\theta_r(3) + 0.16\theta_S - \text{PFP} + \text{“recent”} + \text{singleton}$.
 - Inner edge connections $0.57\theta_p(-0.03) + 0.39\theta_r(3) + 0.04\theta_S - \text{PFP} + \text{“recent”} + \text{singleton}$.
- Expect “Best” better than PFP which is slightly better than random.

arXiv results – successful results



arXiv results – much less successful results



All models hopelessly wrong (cliques an issue?).