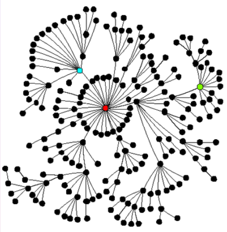
Likelihood methods for comparing network evolution models



Richard G. Clegg (richard@richardclegg.org)
Dept. of Electronic and Electrical Engineering, UCL
Help from Raul Landa and Miguel Rio (UCL), Uli Harder (Imperial)

Talk to Mathematics of Networks 2010

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 be able to compare rival theories about the evolution.

- How do we know which properties are important?
- If we have historic data about the network can this be used?
- Answer: FETA – Framework for Evolving Topology Analysis.

Topology modelling – the 1 minute history

Scale free networks

A scale free network is one where the degree distribution follows a power law – $\mathbb{P}[\deg = i] \sim i^{-\alpha}$.

Scale free networks said to include:

- Internet Autonomous System (AS) graph [Faloutsos x 3 INFCOM 1999],
- hyperlinks in web pages / wikipedia,
- co-authorship/citation networks, and other social networks,
- biological networks (protein networks).

Preferential attachment

Probability of attach to node prop to node degree. Leads to scale free network (Barabási–Albert [Science 1999]).

Other models

- Waxman model [Waxman IEEE Selected Areas in Communication 1988] – predates scale-free discovery.
- Generalised Linear Preference (GLP model) [Bu–Towsley, INFOCOM 2004] – uses non-linear connection probabilities.
- Positive Feedback Preference (PFP model) [Zhou–Mondragón Phys Rev E 2004]
    - Prob. of connecting to i is $p_i \sim d_i^{(1-\delta \log_{10} d_i)}$ where δ is a tunable parameter.
    - Combined with *interactive growth* model (how internal links connect).
    - δ tuned “by hand” to reproduce a number of statistics of interest.
    - Accounts for the fact that the fact that the internet is not pure power law.

The "basket of statistics" approach

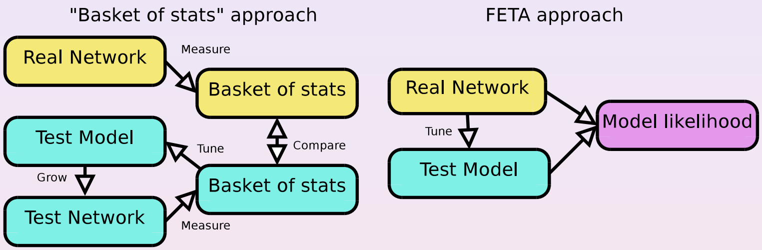
- Current approach – call it the "basket of statistics" method.
  1. Select several statistics which can be measured on net snapshot.
  2. Use test model to grow test network (same size as real network).
  3. Compare the "basket of statistics" on real and test.
- New statistics motivate new models – 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.

FETA approach

FETA – the Framework for Evolving Topology Analysis is a simpler approach if data is available.



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 selection 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.

In practice

- Hypothesise a model which "explains" some portion of the evolution of a graph G.
- The statistic c_0 measures how much "better" than random the model is (> 1 better than random and < 1 worse).
- For two models, the ratio of c_0 for each is the ratio of those models "per choice likelihood".
- An edge choice can be decomposed into two node choices.
- If a simple graph is desired the choice of the second node is made from a reduced choice set (to avoid repeated edges and self edges).

Building models from components

- A node choice model θ could be built from component models such as:
  1. θ_d Preferential attachment model.
  2. $\theta_p(\delta)$ the PFP model with δ parameter.
  3. θ_t triangle model (prob. prop. to Δ count).
  4. θ_S singleton model (prob. const. for degree $= 1$ 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!

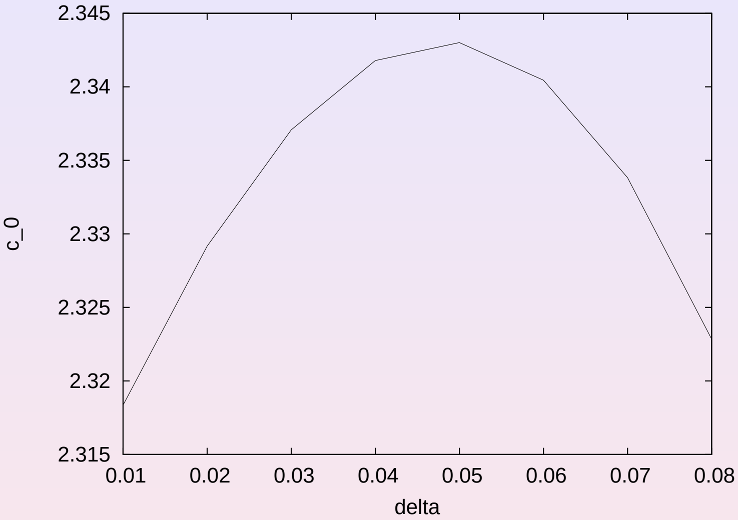
A GLM approach to optimise β parameters

- Want to fit $p_i = \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \varepsilon$ to data.
- This looks very like a Generalised Linear Model (GLM).
- Problem: 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.
- $\mathrm{E}[I_i] = p_i$ – the expectation of I_i is the probability i would be chosen by the model underlying the graph evolution.
- Fitting $I_i = \beta_1\theta_1 + \beta_2\theta_2 + \cdots$ for all possible nodes for a given choice and for all known choices optimises the β.

Artificial tests

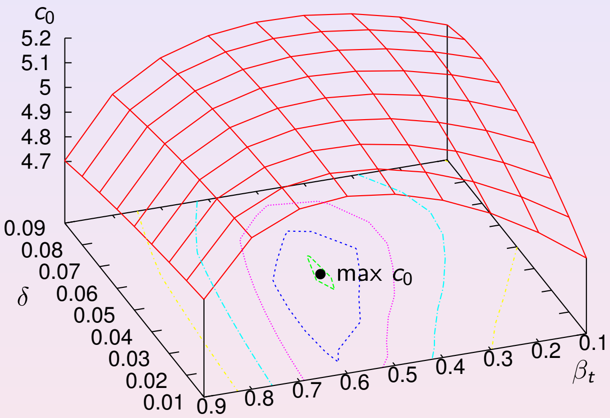
- The most convincing test of such a model is its ability to recover parameters from a known model.
- Consider the PFP model $\theta_p(\delta)$.
- Prob. of connecting to node i is $p_i \sim d_i^{1+\delta \log_1 0 d_i}$.
- Create a test network of 10,000 nodes with $\delta = 0.05$.
- Simple outer model adds one node and one link at each stage (start with one link).
- Now try to recover "unknown" δ.
- Measure c_0 for models of the form $\theta_p(\delta)$ with various δ values.
- Find δ to maximise c_0.

Parameter sweep to recover $\delta = 0.05$ (10,000 nodes)



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

Similar test on $\theta = 0.5\theta_p(0.05) + 0.5\theta_t$ (PFP + triangles) – new node connects to three nodes.



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

Parameter recovery using GLM procedure

- Test model $\theta = 0.25\theta_0 + 0.25\theta_t + 0.25\theta_S + 0.25\theta_D$.
- Random model + triangle model + singleton model + doubleton model.
- Generate 10,000 links and fit using GLM.

Parameter	Estimate	Significance
β_0	0.23 ± 0.021	0.1%
β_t	0.28 ± 0.017	0.1%
β_S	0.24 ± 0.016	0.1%
β_D	0.25 ± 0.020	0.1%

GLM procedure with incorrect model

- In reality we do not know which model components to use.
- 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.
- The c_0 is an accurate predictor of how well models replicated real network statistics.
- Note – claim is not that the models in this presentations are the best possible.
- Claim is that the c_0 is a good predictor of success at predicting network.

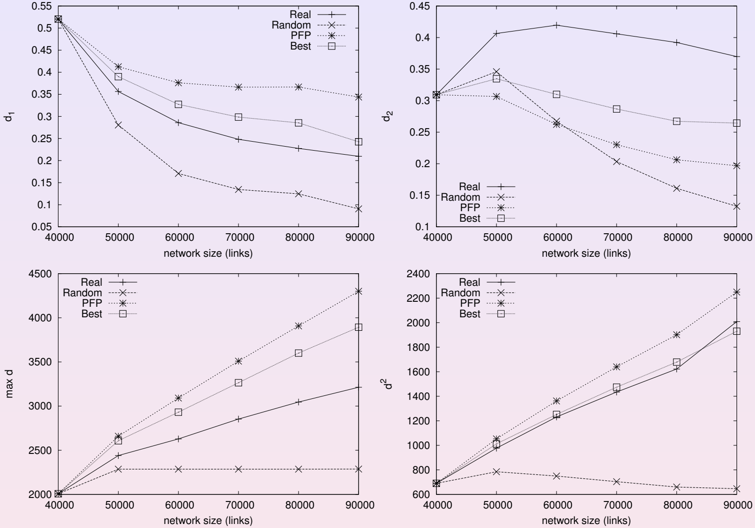
Routeviews AS data

- Network is internet Autonomous System graph.
- Daily measurements from April 11th, 2007 to January 16th, 2009.
- Nodes are always added to the model (even though in reality some die).
- Network grows from 42,000 edges to over 90,000.
- Fit the best inner model from components.
- Fit separate models for "new node" connections and for "inner edge" connections to get the best model.
- Compare with "random" and with "best pure PFP" – that is a PFP model with a single δ.

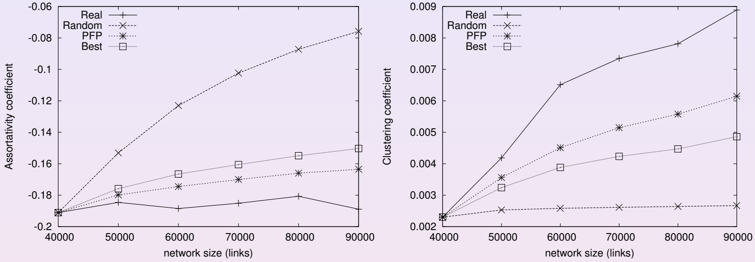
Routeviews models

- Outer model is always "copy" of real outer model (where real data added new node our model does).
- Random model θ_0 – obviously has $c_0 = 1$.
- Best "pure PFP" model $\theta_p(0.005)$ (very low δ parameter) – has $c_0 = 4.81$.
- Note this is not PFP as in [Zhou 2004] (no Interactive Growth part).
- "Best" model found has $c_0 = 8.06$.
  - New node connections $0.81\theta_p(0.014) + 0.17\theta_r(1)$ – PFP + "recent".
  - Inner edge connections $0.71\theta_d + 0.22\theta_r(1) + 0.07\theta_S$ – pref attach + "recent" + singleton.
- Expect "Best" better than PFP better than random.

Routeviews results – successful results



Routeviews results – less successful results



For assortativity and clustering coefficient PFP slightly beats "best".

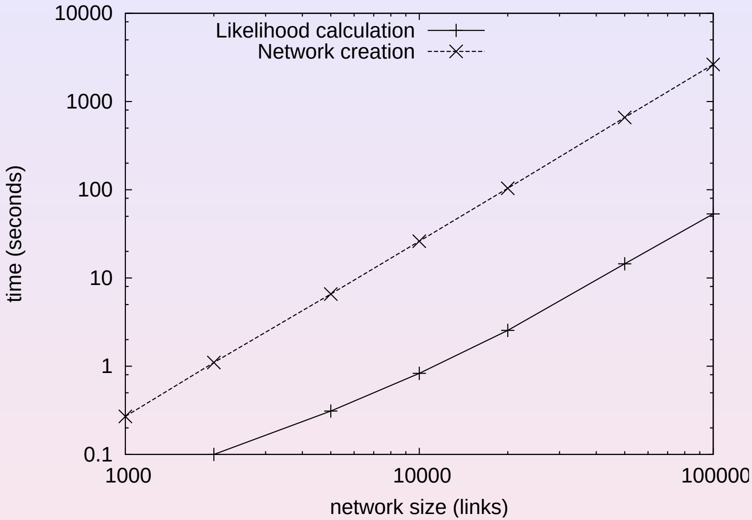
Conclusions

- 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: $\theta = K\theta_d^{\beta_d}\theta_T^{\beta_T}\cdots$.
- 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.

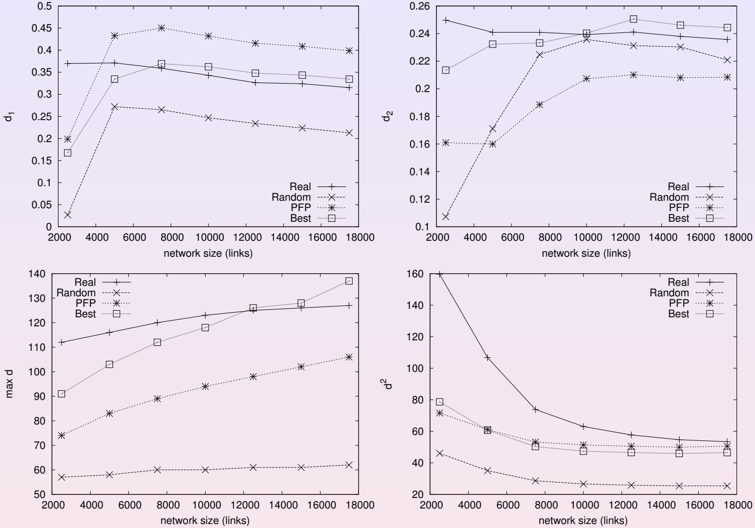
Runtime of likelihood estimate versus network creation



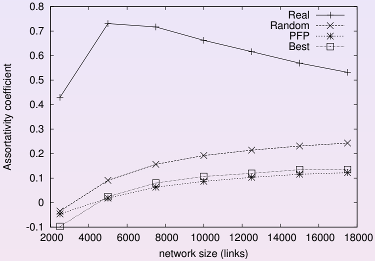
arXiv modelling

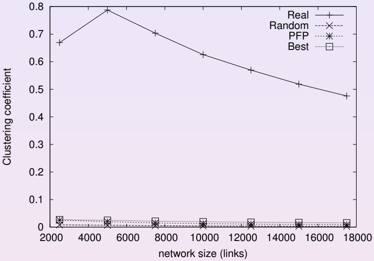
- arXiv co-authorship network for "math" library.
- Approx 17,500 links representing two authors on same paper.
- Outer model as before.
- Random model θ_0 – obviously has $c_0 = 1$.
- Best "pure PFP" model $\theta_p(-0.005)$ (negative δ parameter common in "human" networks) – has $c_0 = 1.31$.
- "Best"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