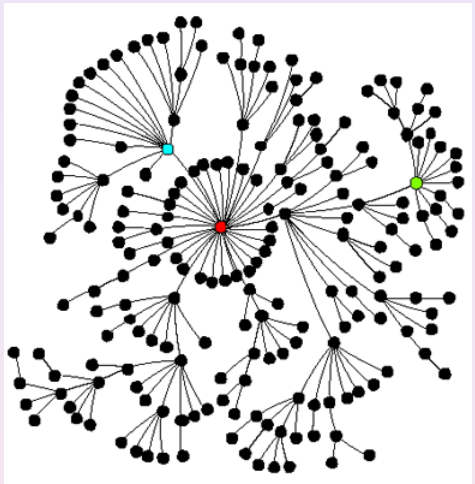
A new method for generating network topologies



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

Talk to UCL 2008

(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 *growth process* for the real network.
- Want to be able to compare rival theories about the growth process.

- 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?

Talk structure

1. A quick history of network topology modelling.
   1. Scale-free networks
   2. Other models and statistics
   3. Problem statement
2. A way forward for network modelling
   1. A general class of network growth models
   2. A Framework for Evolving Topology Analysis (FETA)
3. Testing the framework
   1. Tests on artificial data
   2. Tests on real data
4. Conclusions and future work

Scale-free networks

A scale-free (power-law) network

Let X be the degree of a node in a network. A network is said to be scale free if

$$\mathbb{P}[X = k] \sim k^{-\alpha},$$

where $\sim$ means *asymptotically proportional to* and $\alpha \in (0, 2)$.

Examples include:

- internet AS network [Faloutsos $\times$ 3, INFOCOM 1999],
- hyperlinks in web pages and in wikipedia,
- science citation networks,
- human social networks,
- some biological networks (protein networks).

The preferential attachment model

- Model due to Barabási–Albert [Science 1999] gives a scale free network.
- A new node connects to three existing nodes with selected probabilities.
- This can be shown to give a scale-free network as size grows to infinity.

Preferential attachment probabilities

Let X be a random variable representing the node to connect to. The probability of selecting node k is given by

$$\mathbb{P}[X = k] = Cd_k,$$

where d_k is the degree of node k and $C = 1/\sum_k d_k$.

Network statistics

Various network statistics can be considered.

- Average node degree.
- Maximum node degree – largest d_i where d_i is the degree of node i.
- Assortativity (r) – assortative networks have $r > 0$ (nodes of similar degree likely to connect), disassortative have $r < 0$ (nodes of similar degree unlikely to connect – correlation coefficient of node degrees of connected nodes.
- Top clique size – the size of the largest set of nodes which all connect to each other.
- Clustering coefficient – number of triangles divided by the possible number. For node i the coefficient is $\frac{T_i}{d_i(d_i-1)/2}$ with $d_i \geq 2$, where T_i is the number of triangles involving node i.

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]
  - Probability of connection proportional to $d_i^{(1-\delta \log_{10} d_i)}$ where δ is a tunable parameter.
  - δ tuned "by hand".
  - Reproduces 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

The current way of assessing a test model can be characatured as the "basket of statistics" method.

1. Select a "basket of statistics" for comparison.
2. Use test model to grow test network (same size as real network).
3. Compare the "basket of statistics" on real and test.

Network modelling appears to be progressing in the following manner:

1. Find some statistic the current model does not replicate (add this to "basket").
2. Create a new model which replicates the new statistic without affecting old ones.
3. Test using the above procedure.

Problem statement

- Models aim to replicate several statistics.
- Not certain which statistics are correlated and which are most important.
- Important aspects may be missed by the statistics used.
- Growth information about network often available but not used.

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.

A general model for network topology creation

A general framework for the type of models analysed here.

1. Outer model
   - This describes how nodes and edges are added.
   - Example 1: A new node is added at every stage which connects to three existing nodes.
   - Example 2: At every step, with probability $\mathbb{P} = 0.25$ a new node is added and connects to one existing node. With probability $\mathbb{P} = 0.75$ two existing nodes are connected.
   - The outer model can be of arbitrary complexity, it is not the focus of this work.
2. Inner model
   - This describes which nodes are chosen for connection.
   - Nodes are assigned probabilities based upon their properties.
   - Separate inner models can be created for connecting to new nodes and for joining existing nodes.

Calculating model likelihood

- Let $P_i(t)$ be the probability that node i was chosen at step t (according to some inner model).
- Let n_t be the node which was actually chosen at step t (assume n_t is known for steps s to e).
- Given node choices $n_s, \ldots, n_e$ and knowing the inner model(s), the likelihood can be calculated.

Likelihood of given network evolution

The likelihood L for the hypothesised model is the product of the probability of each individual node choice, $L = \prod_{t=s}^{e} P_{n_t}(t)$. The higher the value of $L \in [0, 1]$ the better the fit to the data – 1 being "perfect" and 0 "impossible".

For some parameterised models, it is possible to calculate which parameters minimise L (or $\log(L)$).

Clarification of Likelihood

- The probability is some function of the graph $P_i(t) = f(G(t), i)$.
- Taking logs makes the maths easier $\log(L) = \sum_{t=s}^{e} \log(P_{n_t}(t)$.
- Now we want to find a function for $P_i(t)$ which maximises L (or $\log(L)$).
- A perfect fit would be $P_i(t) = 1$ iff $i = n_t$ but this model is over specified (as many parameters as data).
- Conversely, if a model ever says $P_{n_i}(i) = 0$ then $L = 0$ – the model is impossible.
- Look for a *parsimonious* model (few parameters) which maximises L.

Generalized Linear Models (GLM)

- Motivating example: y is weight of person (kg), x is height (m)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

  where β_i are parameters to fit and ε are errors.
- This approach is well known, the likelihood can be maximised to fit the model. (If the errors are iid normal mean zero this minimises the RMSE too).
- The assumption is that the errors are mean zero – a known "structure" of errors can be assumed (normal, binomial, Poisson etc).
- Standard statistical packages can be used to fit this model, get optimal β parameters.
- The models also produce confidence intervals for the parameters and statistical significance for each.
- How can this be applied to the situation here?

A Framework for Evolving Topology Analysis (FETA)

- We want a model for probabilities based on node parameters, for example $p_i(t) = \beta_1/N(t) + \beta_2 C_d(t) d_i(t)$ where $N(t)$ is no. of nodes, $C_d(t)$ is normalising const.
- However, $p_i(t)$ cannot be directly measured.
- Instead all we know is whether node i was actually selected or not at stage t.
- Let $I_i(t)$ be an indicator variable such that $I_i(t)$ is one if node i was chosen for choice t and zero otherwise.
- By definition $\mathrm{E}[I_i(t)] = p_i(t)$.
- Therefore, we fit models of the form $I_i(t) = \beta_1/N(t) + \beta_2 C_d(t) d_i(t) + \varepsilon_{i,t}$.
- Obviously many models of this form can be tried. Statistical significance will reject unnecessary variables.

FETA in practice

- 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 $I_i(t)$ if node i was picked at step t.
  3. Record a 0 for $I_i(t)$ if node i was not picket at step t.
- The amount of data recorded scales as $O(n^2)$ with the number of choices made.
- Sampling may therefore be required to fit a model.
- The data set can be fitted using standard statistical packages (like R).
- The errors ε are assumed to have a binomial structure (since $I_i(t)$ is 1 or 0).
- We can think of it as letting rival models "fight" to explain the data.

Tests on artificial data

Why artificial data?

We can test the method by recovering known parameter from networks constructed with a known algorithm.

- Given an inner and outer model build a realistic sized network.
- Construct the FETA input data.
- This data must be sampled to get it down to a usable size.
- The GLM fitting procedure in R can be used to recover the parameters of the inner model.
- If the correct parameters are recovered then we know it works.
- Correctly specified and misspecified models can be tried.

Test model 1 – Random + Preferential attachment

- Outer model – new node every step connects to one existing node.
- Inner model – $p_i(t) = \beta_0 + \beta_1/N(t) + \beta_2 C_d(t) d_i(t)$ where $N(t)$ is no. of nodes, $C_d(t)$ is normalising const.
- Choose $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = 0.5$.

Param	True Value	Estimated Value	Significance Level
β_0	0	$(-2.3 \pm 2.4) \times 10^{-6}$	none
β_1	0.5	0.47 ± 0.17	1%
β_2	0.5	0.54 ± 0.11	0.1%
β_0	0	$(-1.3 \pm 2.2) \times 10^{-5}$	none
β_1	0.5	0.46 ± 0.16	1%
β_2	0.5	0.60 ± 0.11	0.1%
β_0	0	$(-3.0 \pm 2.3) \times 10^{-5}$	none
β_1	0.5	0.54 ± 0.16	0.1%
β_2	0.5	0.47 ± 0.11	0.1%

Test model 2 – Random + PFP (known δ)

- Outer model – new node every step connects to one existing node.
- Inner model – $p_i(t) = \beta_1/N(t) + \beta_2 C_p(t) d_i(t)^{1+\delta \log_{10}(d_i(t))}$, with $\delta = 0.048$ and $C_p(t)$ as normalising const.
- Choose $\beta_1 = 0.5$, $\beta_2 = 0.5$.

Param	True Value	Estimated Value	Significance Level
β_1	0.5	0.53 ± 0.11	0.1%
β_2	0.5	0.46 ± 0.11	0.1%
β_1	0.5	0.44 ± 0.097	0.1%
β_2	0.5	0.56 ± 0.10	0.1%
β_1	0.5	0.58 ± 0.10	0.1%
β_2	0.5	0.42 ± 0.10	0.1%

Test model 3 – pure PFP unknown δ

- Outer model – new node every step connects to one existing node.
- Inner model – $p_i(t) = \beta_1 C_p(t) d_i(t)^{1+\delta \log_{10}(d_i(t))}$, with $\delta = 0.048$ and $\beta_1 = 1$.
- Assume delta unknown and calculate deviance and RMSE for various δ.
- "Correct" δ should minimise deviance (maximise Likelihood) but not necessarily RMSE.

Test model 3 – Deviance versus δ

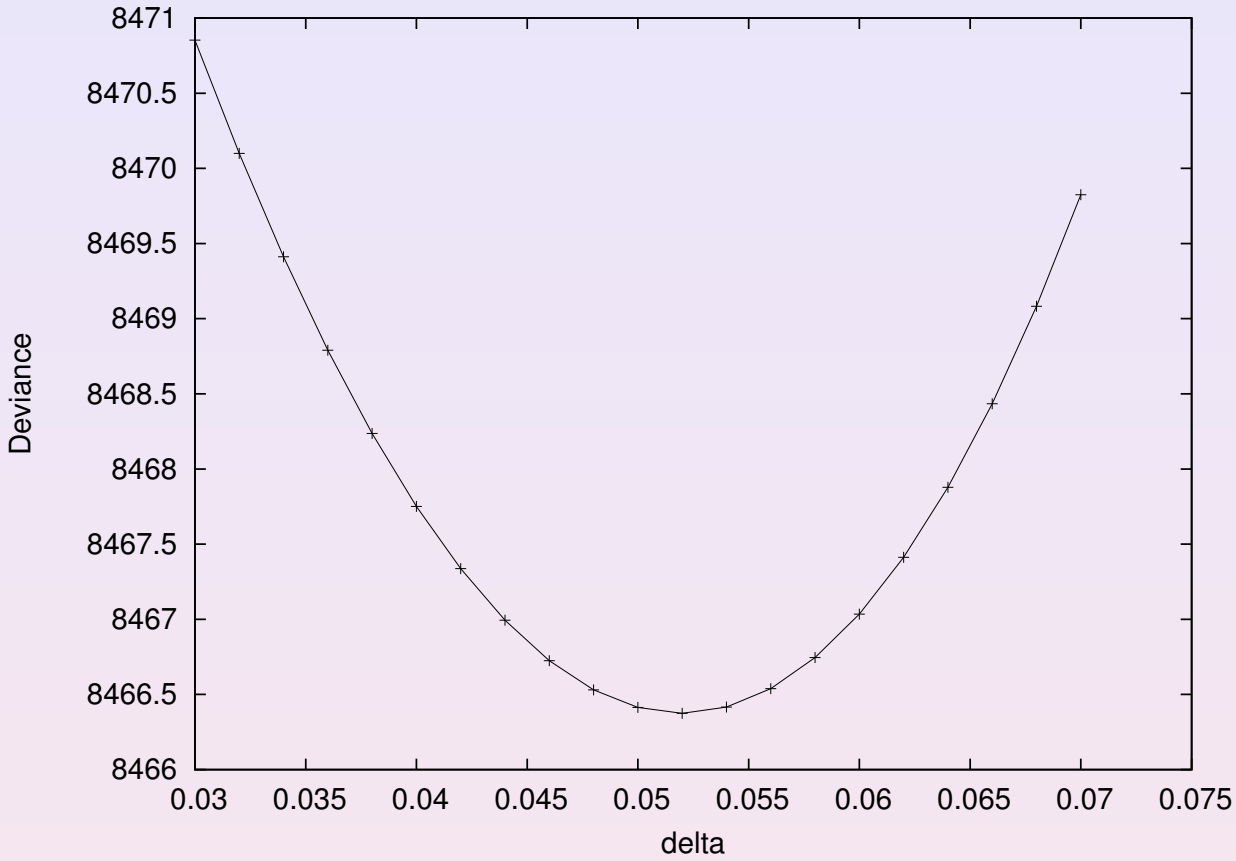


Figure: PFP deviance with various δ.

Test model 3 – RMSE versus δ

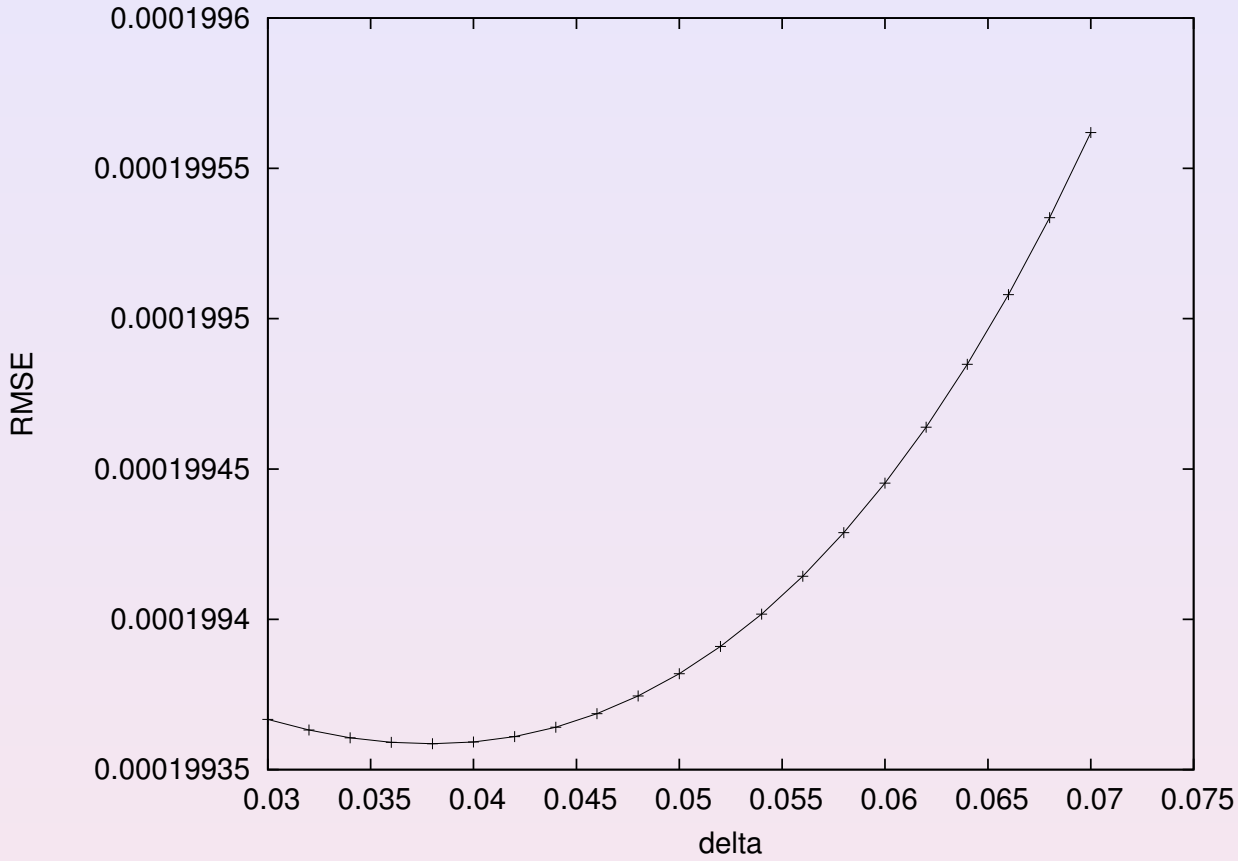


Figure: PFP RMSE with various δ.

Misspecified models 1

- Outer model – After each new node $\mathbb{P} = 0.9$ add one internal link.
- Inner model (new nodes) – $p_i(t) = \beta_1/N(t) + \beta_2 C_d(t) d_i(t)$, with $\beta_1 = \beta_2 = 0.5$.
- Inner model (internal edges) – $p_i(t) = \beta_1/N(t) + \beta_2 C_p(t) d_i(t)^{1+\delta \log_{10}(d_i(t))}$, with $\delta = 0.048$ and $\beta_1 = 0.2$ and $\beta_2 = 0.8$.
- Fitted model (both) – $p_i(t) = \beta_1/N(t) + \beta_2 C_d(t) d_i(t) + \beta_3 C_T(t) T_i(t)$, where $T_i(t)$ is number of triangles.
- Data for new nodes can be put in a separate file to internal edges and fitted separately in R.

Misspecified models 1

Param	True Value	Estimated Value	Significance Level
β_1^N	0.5	0.51 ± 0.049	0.1%
β_2^N	0.5	0.49 ± 0.048	0.1%
β_3^N	0	-0.0028 ± 0.0019	none
β_1^E	0.2	0.17 ± 0.032	0.1%
β_2^E	0.8	0.83 ± 0.042	0.1%
β_3^E	0	0.0022 ± 0.00086	0.5%

While β_3^E is claimed to have stat. sig. the low value may lead to an experimenter removing it anyway.

Misspecified models 2 – pure PFP

- Outer – new node every iter. edge to one internal node.
- Inner – $p_i(t) = C_p(t) d_i(t)^{1+\delta \log_{10}(d_i(t))}$ with $\delta = 0.048$.
- Fitted model is

$$p_i(t) = \beta_1 C_r(t)/N(t) + \beta_2 C_d(t) d_i(t) + \beta_3 C_p(t) d_i(t)^{1+\delta \log_{10}(d_i(t))},$$

Param	True Value	Estimated Value	Significance Level
	Random + BA + PFP($\delta = 0.048$)		
β_1	0	-0.11 ± 0.11	none
β_2	0	0.21 ± 0.61	none
β_3	1.0	0.90 ± 0.54	10%
	Random + BA + PFP($\delta = 0.06$)		
β_1	0	-0.11 ± 0.11	none
β_2	0	0.43 ± 0.47	none
β_3	1.0	0.68 ± 0.41	10%

Misspecified model 2 – pure PFP

Param	True Value	Estimated Value	Significance Level
	Random + BA + PFP($\delta = 0.01$)		
β_1	0	-0.10 ± 0.11	none
β_2	0	-4.1 ± 3.3	none
β_3	1.0	5.2 ± 3.2	none
	BA + PFP($\delta = 0.048$)		
β_2	0	-0.16 ± 0.47	none
β_3	1.0	1.16 ± 0.47	5%
	BA + PFP($\delta = 0.06$)		
β_2	0	0.12 ± 0.35	none
β_3	1.0	0.88 ± 0.36	5%
	BA + PFP($\delta = 0.01$)		
β_2	0	-5.62 ± 2.7	5%
β_3	1.0	6.6 ± 2.7	5%

UCLA data set

- Set of AS topology links, from UCLA Jan 2004 – Aug 2008 `http://irl.cs.ucla.edu/topology/`.
- Updated daily using data sources such as BGP routing tables and updates from RouteViews, RIPE, Abilene and LookingGlass servers.
- Times of oberservations noted.
- Some preprocessing necessary:
  1. Add links in order which they were first observed.
  2. If neither node is yet in network, delay addition (keep network connected).
  3. Two datasets Permanent has all links observed Transient has only links which have been seen in last two months.
- Note: this is a controversial data set. FETA is general and does not depend on correctness of this data.

Real network evolution data

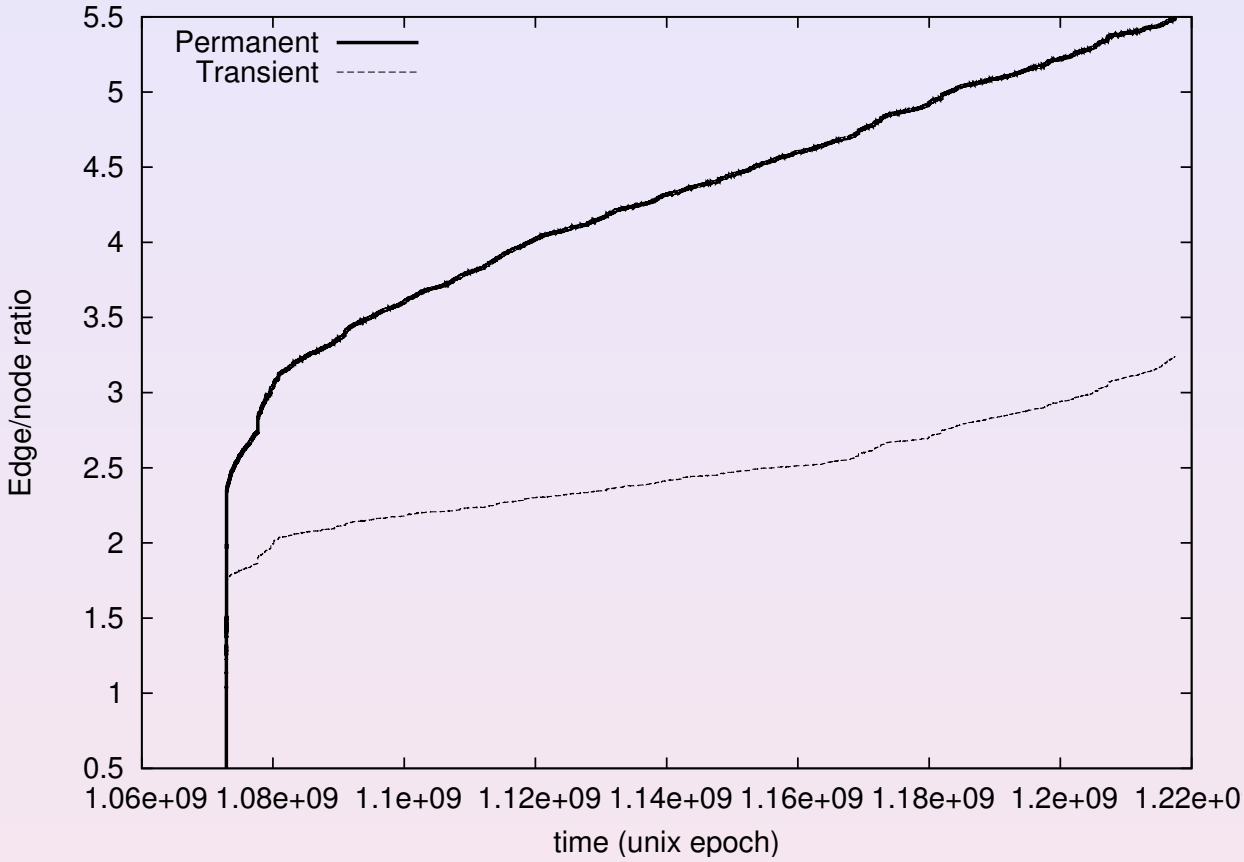


Figure: Edge/node ratio for the two topologies as time progresses.

Real network data stats

This table shows some summary statistics for the UCLA AS data.

Statistic	Permanent	Transient
Nodes	35,723	29,702
Links	196,040	96,237
Mean degree	11.0	6.5
Max degree	5,497	3,157
Top clique	11	7
Assort. coeff.	-0.15	-0.19
Clust. coeff	0.076	0.058

Initial test model

- Separate inner models were fitted for new nodes and for internal edges.
- The assumed outer model was simple: Follow every new node with two ($\mathbb{P} = 0.75$) or three ($\mathbb{P} = 0.25$) internal edges.
- The model had a random component, a degree based component and a PFP component with $\delta = 0.048$.

Param	Estimated Value	Significance Level
β_r^N	-0.18 ± 0.035	0.1%
β_d^N	1.0 ± 0.42	5%
β_p^N	0.14 ± 0.38	none
β_r^E	-0.014 ± 0.048	none
β_d^E	1.28 ± 0.20	0.1%
β_p^E	-0.23 ± 0.16	none

Final fitted model

- The components with least significance were removed and the models rerun.
- The PFP parameter δ was tuned by maximising the MLE (with repeated runs).
- Note that this invalidates the significance figure for the internal edge model (but the p-value is extremely low anyway).

Param	Estimated Value	Significance Level
β_r^N	-0.18 ± 0.026	0.1%
β_d^N	1.18 ± 0.16	0.1%
β_d^E	1.1 ± 0.067	0.1%
β_p^E	-0.059 ± 0.0043	0.1%

Statistical comparison

Of course the traditional "basket of statistics" method can now be considered.

Statistic	Real AS Transient	PFP $\delta = 0.048$	PFP $\delta = -0.091$	FETA
Nodes	29,702	29,634	29,616	29,575
Links	96,237	96,234	96,233	96,232
Mean degree	6.48	6.49	6.50	6.51
Max degree	3,157	7,211	1,244	1,416
Top clique	7	31	27	27
Assort. coeff.	-0.19	-0.31	-0.15	-0.20
Clust. coeff	0.058	0.015	0.062	0.064

Note that the first three stats come from statistical variations in the (identical) outer model. These aside, FETA performs best or equal best in three of the four statistics and is very close to the best answer in the fourth statistic.

Conclusions

- Big idea 1: Likelihood methods provide a rigorous statistical underpinning to the subject as opposed to an ad hoc "basket of statistics".
- Big idea 2: By incorporating the generalised linear model we can test which components of a model are important and get significance levels.
- FETA provides a method for assessing and optimising topology creation algorithms in a variety of fields.
- In tests on artificial networks, the model could recover parameters in most cases.
- There are issues with scalability but sampling can help with this.
- FETA, or something like it, is the way we should be investigating network topology algorithms in future.

Future work (1)

- It seems that further investigation (possibly with larger computers) could quickly improve model of AS network.
- What other networks does the data exist for?
  1. Social networking sites?
  2. Web connections (can we get evolution data?)
  3. Academic publications data?
- A number of hypotheses could be tested in this framework:
  1. nodes have a "likely to acquire partners" phase when first joining,
  2. the inner and/or outer model changes as the network grows,
  3. other properties of a node may be important to growth (triangles, age, star sign?).

Future work (2)

While FETA works well, there are some obvious routes for improvement.

- The data requirements are large $O(n^2)$ – can the unchosen nodes somehow be "pooled"?
- Can the MLE problem be solved directly for interesting general cases?
- Additive probability equations are unlikely, multiplicative is more likely

$$p_i = C x_{i,1}^{\beta_1} x_{i,2}^{\beta_2} x_{i,3}^{\beta_3} \cdots .$$

- A possible answer is logistic regression (part of GLM) but normalisation is a big problem (PFP looks very natural in logistic form).
- There is lots of work to do in this area but it could change network modelling in many fields – an exciting new area for anyone who wants to get involved.