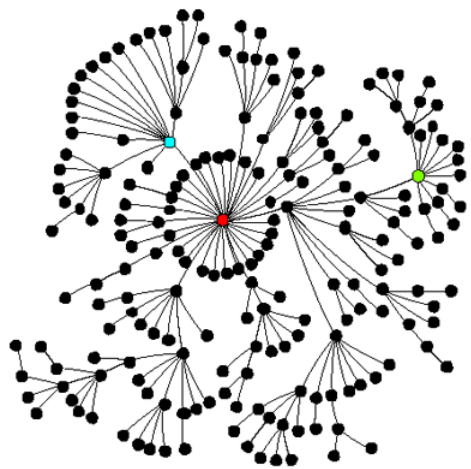
Evolving topologies in a streaming context



Richard G. Clegg (richard@richardclegg.org)
Dept. of Electronic and Electrical Engineering, UCL

Talk to Bournemouth 2014

(Prepared using LaTeX and beamer.)

Research interests at UCL EE

- Group is Networks and Services Research.
- Main interest is internet design at the upper "layers" – protocols and applications.
- Research interests include the statisics of network traffic and the evolution of internet topology.
- Long term interest in internet measurement, getting new data, finding new things in old data.
- Some publications at `http://www.richardclegg.org/`
- This talk will mainly detail a particular research thread about graph topologies.

Introduction

- Evolving networks (graphs/topologies) are an important topic for research.
- Want to describe and understand processes which govern evolution.

Problem statement (vague)

- 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.

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 AS graph, web links, friendship networks. (Only approximation).

Preferential attachment

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

However, not whole truth – misses many aspects of real networks. Other models followed (e.g. prob prop to degree raised to power).

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?

Topology modelling appears to be progressing in the following manner:

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

Refined problem statement

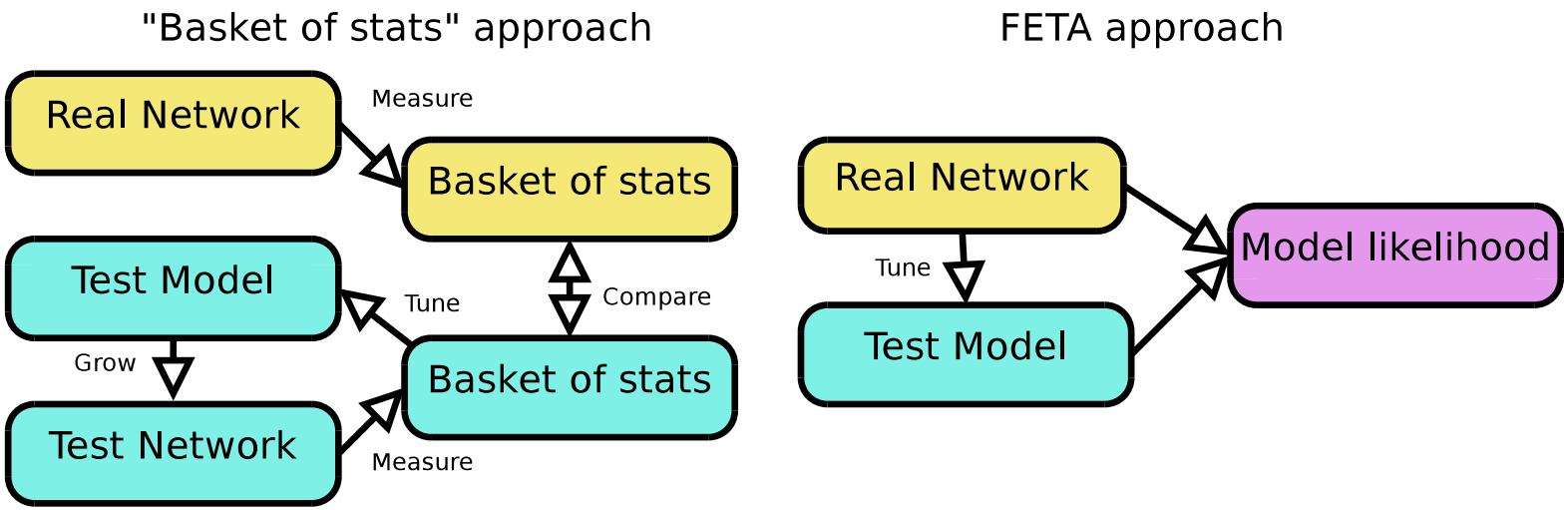
- Let $G(t)$ be a time evolving graph which evolves according to some probabilistic process.
- Let $\mathbf{G} = (G_i, G_{i+1}, \ldots, G_{i+n})$ be random variables representing this process observed at discrete times.
- Let $\mathbf{g} = (g_i, g_{i+1}, \ldots, g_{i+n})$ be a set of observations of $\mathbf{G}$.

Problem statement — more precise

Given observations of a graph $\mathbf{g}$ want to:

- Create models which formally specifies $\mathbb{P}\left[G_{t+1} = g_{t+1} | G_t = g_t, \ldots\right]$.
- Measure the likelihood of such a model producing $\mathbf{g}$.
- Automatically test many such models.

FETA approach



A probabilistic model of graph evolution

- Creating a model of $\mathbb{P}[G_{t+1} = g_{t+1} | G_t = g_t, \ldots]$. is not straightforward.
- This is not like normal stochastic process. The dimensionality of $G(t)$ changes over time.
- Could transform to some multi-dimensional process with dimension highest dimension graph will achieve (nasty solution).
- Also want a solution which is compatible with existing research in field (can test existing research methods).

The FETA model structure

Operation model

- Process to select an operation on the network.
- Could be: add node, add edge, remove node and so on.

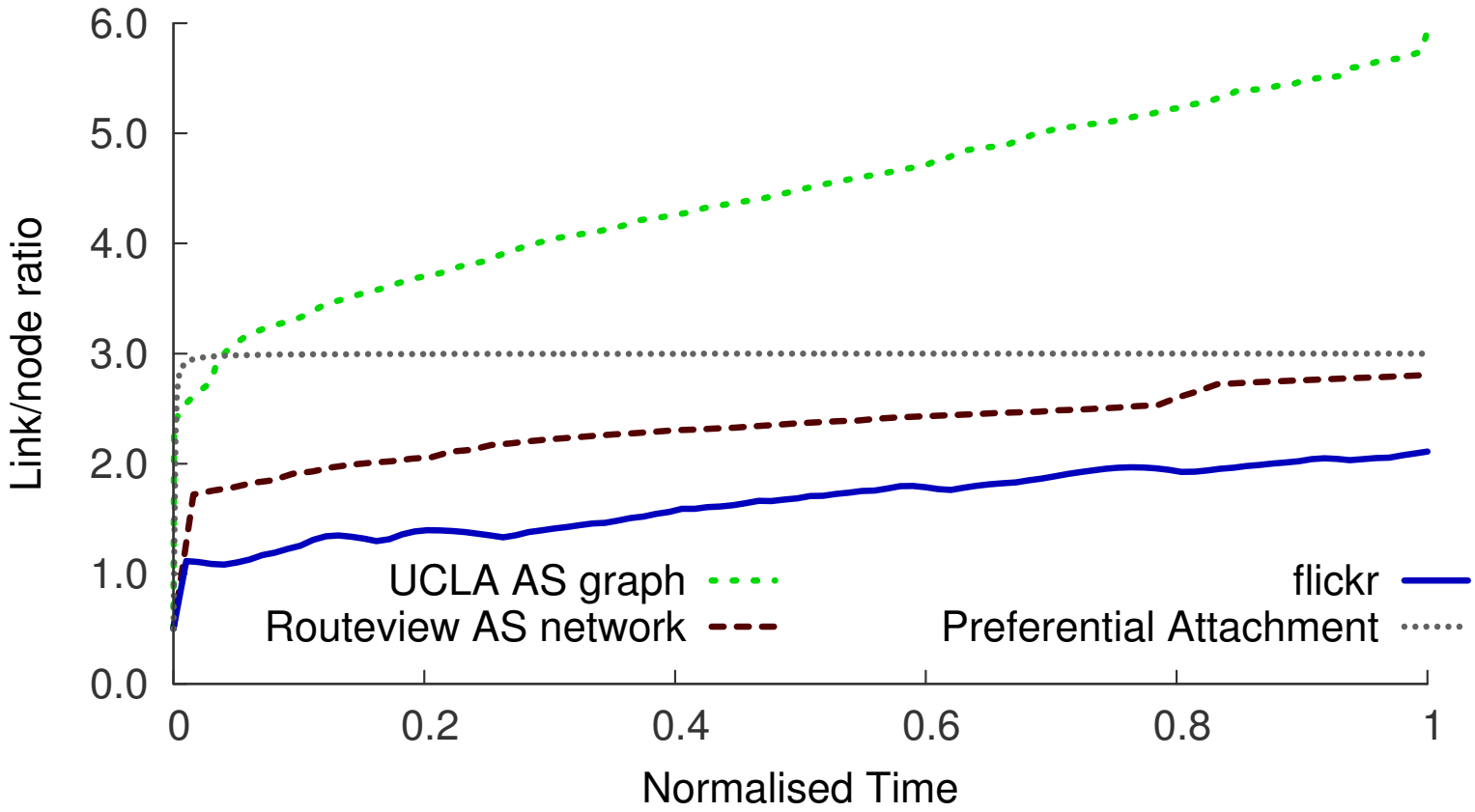
Object model

- Process selects which nodes/edges are involved in operation selected by operation model.
- Probabilities are assigned to nodes and potential edges for random selection.
- Edges selected by assigning probabilities to node pairs.
- Object model is main focus of this presentation.

FETA Model – operations model example

- In current version the operation model can select from:
  1. NewNodes(n, m) Create a new node and connect it to n new nodes and m existing nodes.
  2. NewLinks(n) Select an existing node and connect it to n existing nodes.
  3. NewClique(n, m) Create a clique between n new nodes and m existing nodes.
- Example: Original preferential attachment model is: NewNodes$(0, 3)$.
- Graph evolution is broken down into the addition of cliques, new nodes and links between existing nodes. (There is some ambiguity here).
- The full operations model gives the probability of each operation (with parameters) at each time step.
- More focus needed on the operations model. Here it is just "copied" for real data.

Importance of operations model



Object model examples

- For simplicity consider graphs which evolve using only the NewNode$(0, 1)$ operation – a new node is created and connects to one existing node.
- Let F designates some function which maps all possible choices (of node) to a probability.
- For example the Preferential Attachment model is $p_i = d_i/k$ where:
  - p_i is the probability of choosing node i.
  - d_i is the degree of node i.
  - k is a normalising constant such that $\sum_i p_i = 1$.
- The PFP model is $p_i = d_i^{1+\delta \log_{10}(d_i)}/k$ where δ is a parameter.

The likelihood of FETA model

- Let $F(\theta)$ be a parameterised FETA model which assigns probabilities to operations and object models with some parameters θ.
- Let F_i be the probabilities assigned at the ith step of evolution (θ dropped for convenience).
- Define $F_i(g_i) = \mathbb{P}\left[G_i = g_i | G_{i-1} = g_{i-1}, G_{i-2} = g_{i-2}, \ldots\right]$
- Then the likelihood of the observations $\mathbf{g}$ given $F(\theta)$ is $L(\mathbf{g}|F(\theta)) = \prod_{k=i+1}^{k=i+n} F_k(g_k)$.
- This likelihood defines how likely the observations were given the hypothesised model.
- It is the ability to assign a true likelihood to the graph evolution which is key to the FETA process.

Usable likelihood

- Define $l(\mathbf{g}|F(\theta)) = \log(L(\mathbf{g}|F(\theta))$.
- Because of normalisation problems standard log-likelihood maximisation techniques do not work.
- Likelihood can be split into operation model and object model components.
- Let F_0 that be the null hypothesis – all choices are equally likely. Let m be the number of choices.
- Human readable measure is c_0 the per choice likelihood ratio.

Per choice likelihood ratio c_0

$$c_0 = \left[\frac{L(C|F)}{L(C|F_0)}\right]^{1/m} = \exp\left[\frac{l(C|F) - l(C|F_0)}{m}\right].$$

Building object models from components

- Three possible object models have been introduced already.
  1. F_0 – all nodes equal.
  2. F_d – preferential attachment (nodes weighted by degree).
  3. $F_p(\delta)$ – PFP model δ is parameter.
- How about mixing models?
- $F = \beta_1 F_0 + \beta_2 F_d$ (nodes sometimes chosen randomly, sometimes by degree) – $0 < \beta_1 < 1$ and $\beta_1 + \beta_2 = 1$.
- On the positive site, a larger family of explanations, on the negative, more parameterisation.

Object model components

Throughout k is a normalising constant such that $\sum_i p_i = 1$ for all nodes considered. p_i is the probability of picking node i (at the stage being considered).

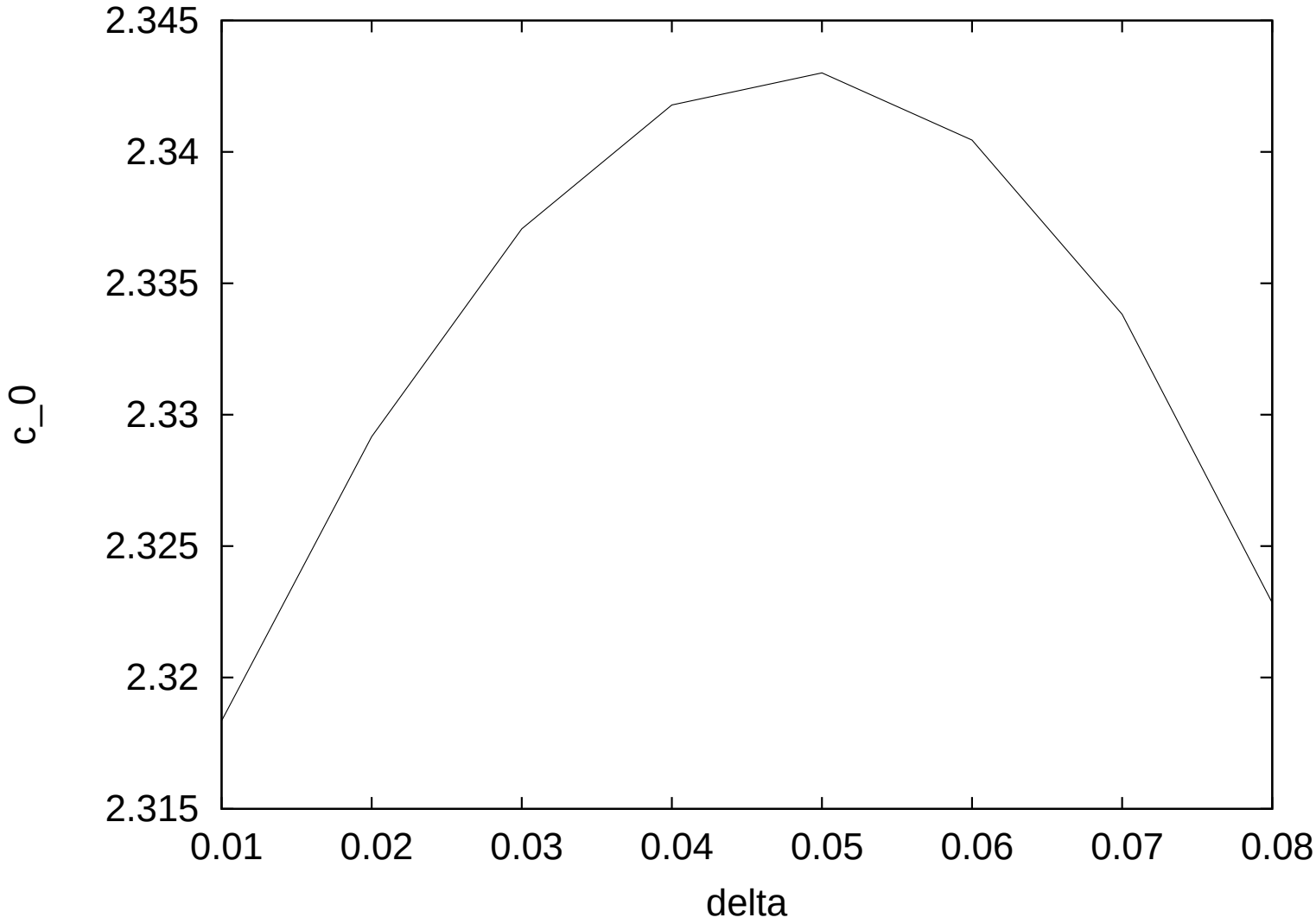
- Random model F_0 $\quad p_i = 1/k$.
- Preferential attachment F_d $\quad p_i = d_i/k$.
- PFP $F_p(\delta)$ $\quad p_i = d_i^{1+\delta \log_{10}(d_i)}/k$ where δ is a parameter.
- Degree power $F_d(\alpha)$ $\quad p_i = d_i^{\alpha}/k$ where α is a parameter.
- Triangle model F_t $\quad p_i = t_i/k$ where t_i is the triangle count of node i.
- Singleton model F_1 $\quad p_i = \begin{cases} 1/k & d_i = 1 \\ 0 & \text{otherwise} \end{cases}$.
- Doubleton model F_2 $\quad p_i = \begin{cases} 1/k & d_i = 2 \\ 0 & \text{otherwise} \end{cases}$.
- Hot model $F_h(n)$ $\quad p_i = \begin{cases} 1/k & \text{node chosen in last } n \text{ picks} \\ 0 & \text{otherwise} \end{cases}$

  where n is a parameter.

Artificial tests

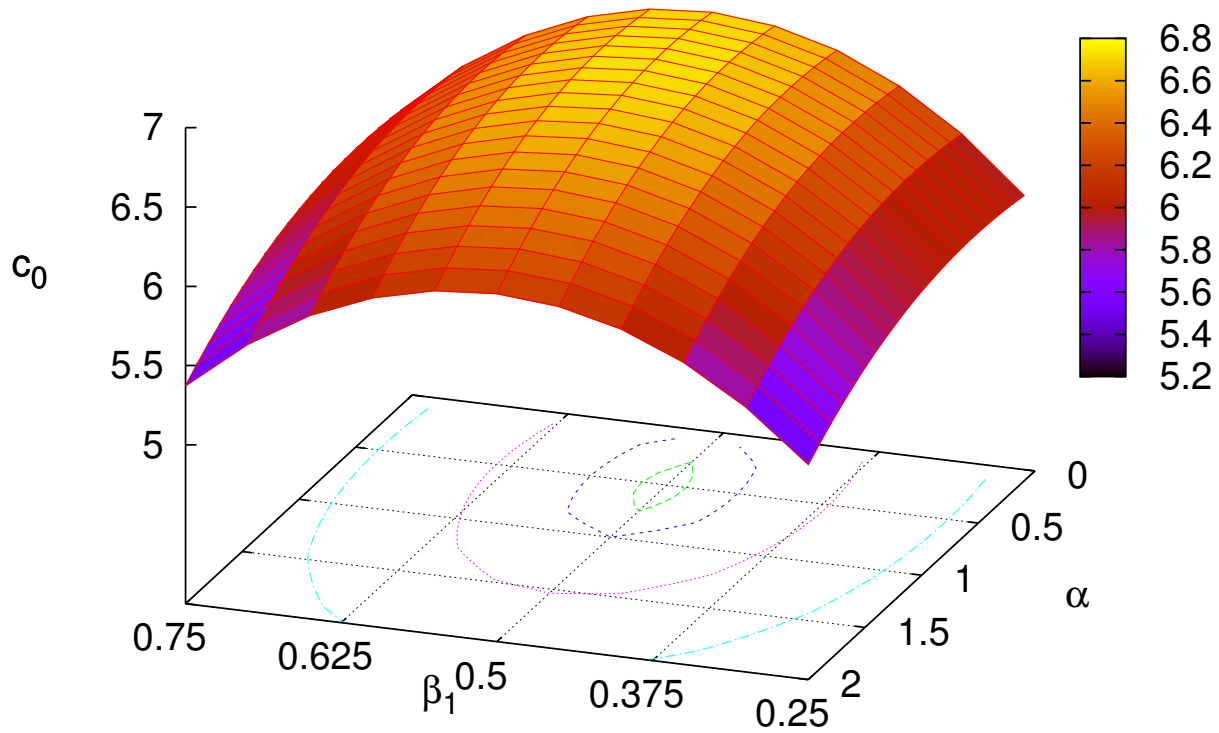
- Perhaps the most convincing test of such a model is its ability to recover parameters from a known model.

Sweep one parameter (10,000 link network)



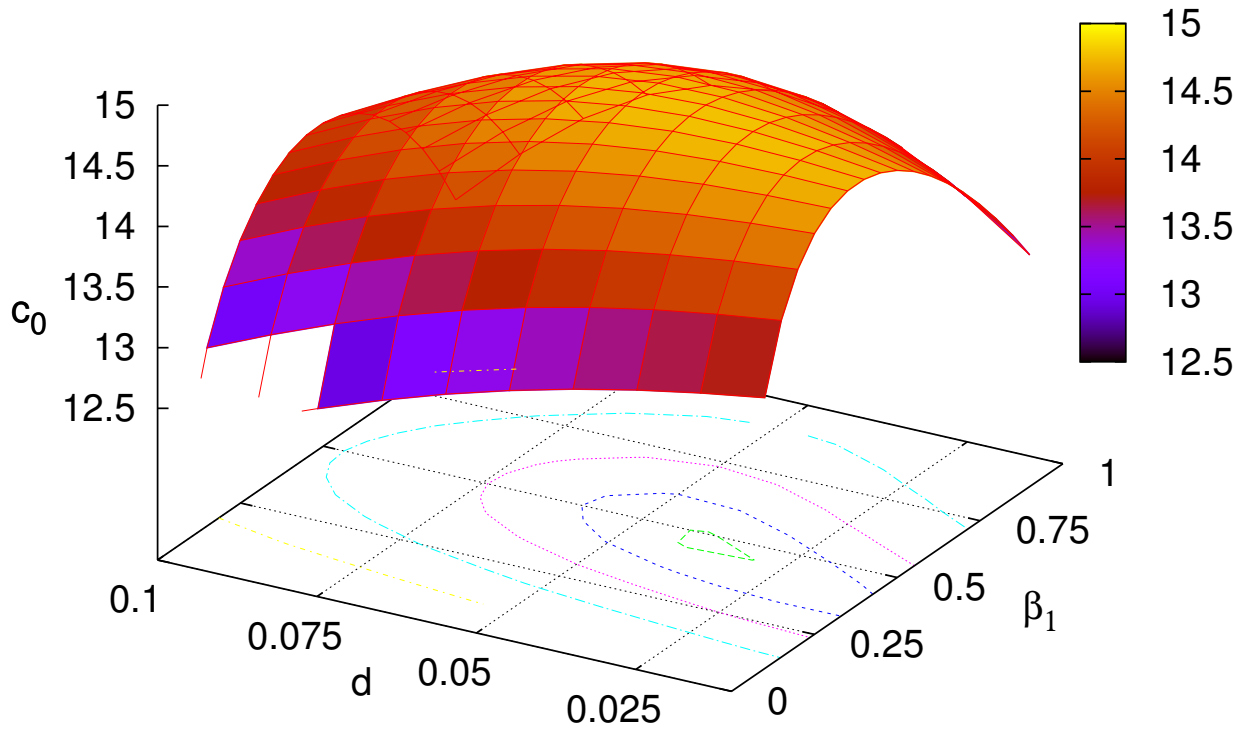
PFP model $F = F_d(0.05)$. Correct answer is $\delta = 0.05$.

Sweep two parameters (10,000 link network)



Correct model $F = 0.5F_2 + 0.5F_d(0.5)$ fitted $F = \beta_1 F_2 + (1 - \beta_1)F_d(\alpha)$.

Sweep two parameters (10,000 link network)



Correct model $F = 0.5F_p(0.05) + 0.5F_t$ fitted
$F = \beta_1 F_p(d) + (1 - \beta_1)F_t$

Parameter recovery using GLM procedure

- Test model $F = 0.25F_0 + 0.25F_t + 0.25F_1 + 0.25F_2$.
- 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_1	0.24 ± 0.016	0.1%
β_2	0.25 ± 0.020	0.1%

Work on real data

- Number of data sets used, enron emails, facebook wall posts, two views of the internet AS network, photo sharing websites.
- Models with higher likelihood are shown to be closer in the real statistics when artificial models grown.
- Likelihood is quicker to calculate than growing an artificial network and measuring results.

Moving to a streaming context

- Theoretically it should be trivial to move this work to a streaming context.
- Graphs can be analysed as they evolve, new links and nodes can be tested and the predicted parameters shifted.
- In reality it is likely that the parameters are not constant for the object or operations model.
- Need to find a compromise between quickly changing paramaters for the evolution of the graph “right now” and slowly changing parameters which are well estimated from many observations.

Conclusions and further work

- Likelihood method can be used to judge the best model from a number of hypotheses
- Parameterised models can have parameters estimated which maximise likelihood.
- In artificial data correct known parameters can be recovered and spurious parameters rejected.
- In real data higher likelihood models produce artificial networks “closer” to the real data set.
- Another step would be to parallelise, to measure likelihood from different subgraphs within larger graphs.
- This method is suitable for measurement of very large scale graph systems (e.g. twitter).

Other research interests

- Internet mining is an idea stemming from a previous EU proposal (rated well but unfunded).
- The idea is to combine measurements from various existing data streams which monitor the internet.
- Rapidly create informative summary statistics and analyse correlations between different streams of data.
- Stitch together large numbers of disparate rapidly changing data sources to provide a clearer overall picture.