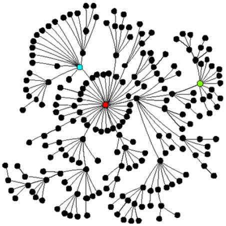
Mixed and time varying models for evolving complex networks



Richard G. Clegg (r.clegg@qmul.ac.uk)
Naomi Arnold (n.a.arnold@qmul.ac.uk)
Queen Mary University of London

3rd UK Network Science Workshop, Leeds 2019

(Prepared using LaTeX and beamer.)

Problem statement

- Evolving networks (graphs/topologies) are an important topic for research.
- Want to describe and understand processes which govern evolution.
- Motivating example models of similar form to Barabási–Albert (1999)

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.

Many many network models exist

- Preferential growth model [Barabási–Albert Science 1999]
  - Probability of attachment to i is proportional to degree d_i, $p_i \sim d_i$.
- Degree–power model [Krapivsky et al Phys Rev Lett 2000]
  - Probability of attachment to i is degree d_i modified by power $p_i \sim d_i^\alpha$ – α is tunable parameter.
- Positive Feedback Preference (PFP model) [Zhou–Mondragón Phys Rev E 2004] (targetted at the Internet)
  - Prob. of connecting to i is $p_i \sim d_i^{(1-\delta \log_{10} d_i)}$ where δ is a tunable parameter.
- Jackson–Rogers [American Econ Rev 2007] (targetted at social networks)
  - (Simplified explanation). Pick r nodes completely at random. Pick q nodes that neighbour these r nodes. q and r are tunable parameters.
- Many others: Stochastic block models, exponential random graphs, Waxman model, Subgraph generation model, Statistical exponential random graphs.

The "basket of statistics" approach

- Which model is "best"? How do we tune parameters?
- Define the current approach 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 specify $\mathbb{P}\left[G_{t+1} = g_{t+1} | G_t = g_t, \ldots\right]$.
- Likelihood: measures the probability a given data set arose if an explanatory model is true.
- Measure the likelihood of such a model producing $\mathbf{g}$.
- Automatically test many such models.

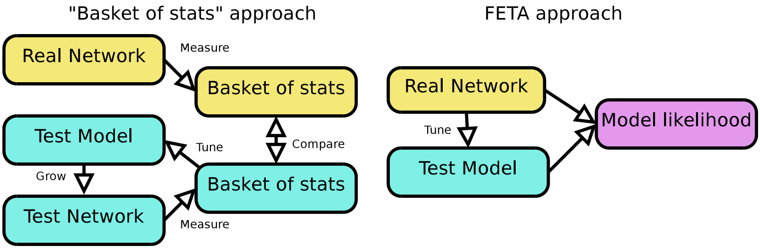
A probabilistic model of graph evolution

- Creating a parameterised model $M(\theta)$ of $\mathbb{P}\left[G_{t+1} = g_{t+1} | G_t = g_t, \ldots\right]$. 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).

FETA approach

FETA

Framework for Evolving Topology Analysis



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 (operation model, for now "copied").

The likelihood of FETA model

- Let $M(\theta)$ be a parameterised FETA model which assigns probabilities to operations and object models with some parameters θ.
- Define $f_{i,M(\theta)}(g_i) = \mathbb{P}[G_i = g_i | M(\theta), G_{i-1} = g_{i-1}, G_{i-2} = g_{i-2}, \ldots]$
- For convenience just write $f_i(g_i)$
- Then the likelihood of the model $M(\theta)$ given the observations $\mathbf{g}$ (from i to $i+n$) is $L(M(\theta)|\mathbf{g}) = \prod_{k=i+1}^{k=i+n} f_k(g_k)$.
- This likelihood defines how likely the model is given the observations (or conversely, how probable the observations given the model).
- It is the ability to assign a true likelihood to the graph evolution which is key to the FETA process.

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 $M_0 \quad p_i = 1/k$.
- Preferential attachment $M_d \quad p_i = d_i/k$.
- PFP $M_p(\delta) \quad p_i = d_i^{1+\delta \log_{10}(d_i)}/k$ where δ is a parameter.
- Degree power $M_d(\alpha) \quad p_i = d_i^{\alpha}/k$ where α is a parameter.
- Triangle model $M_t(N) = 1/k$ if node in neighbourhood of N most recently chosen.
- Singleton model $M_1(N) = 1/k$ if node has degree 1 0 otherwise.
- Doubleton model $M_2(N) = 1/k$ if node has degree 2 0 otherwise.

Insight: we can build models from components

Unlikely that any one model is the "whole" truth

Can we create new models by "mixing" existing model components? Take two (or more) models and create a model that mixes both in some proportion.

- How about mixture models?
- We can mix an arbitrary number of models. Let $p_i^{(n)}$ be the probability of choosing node i in model n

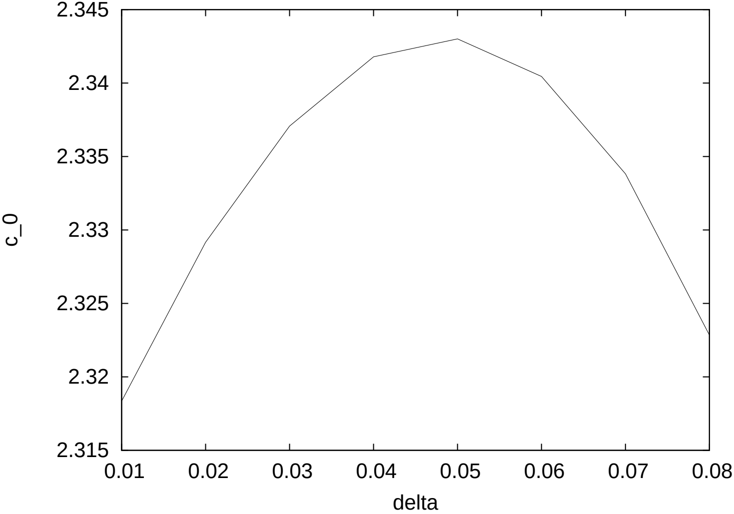
$$p_i = \beta_1 p_i^{(1)} + \beta_2 p_i^{(2)} + \cdots$$

where $0 \leq \beta_i \leq 1$ and $\sum_i \beta_i = 1$.

Artificial tests

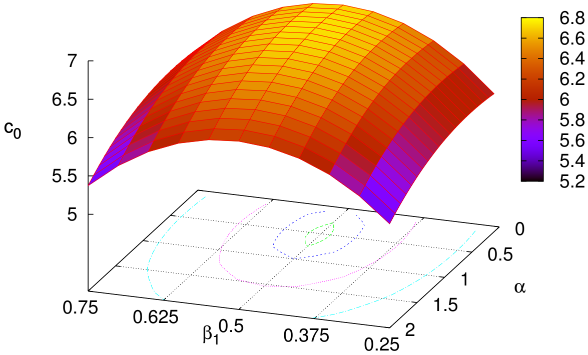
- Perhaps the most convincing test of such a model is its ability to recover parameters from a known model.
- Create a realisation of a graph with known $M(\theta)$.
- From this graph try to recover θ with maximum likelihood.
- Linear parameters of model can be assessed "in parallel" (model tracks likelihood for a range of answers simultaneously).
- Other (non-linear) parameters can be found from "parameter sweeps" (assessing the likelihood for many values of that parameter).
- Look at a measure c_0 which gives a "human readable" likelihood relative to a null model of "random" connections.

Sweep one parameter (10,000 link network)

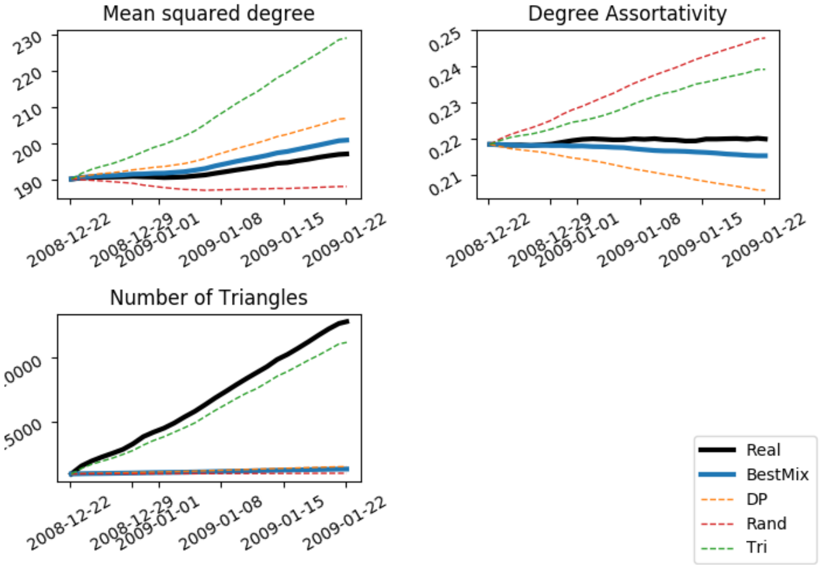


PFP model $M = M_p(0.05)$. Correct answer is $\delta = 0.05$.

Sweep two parameters (10,000 link network)

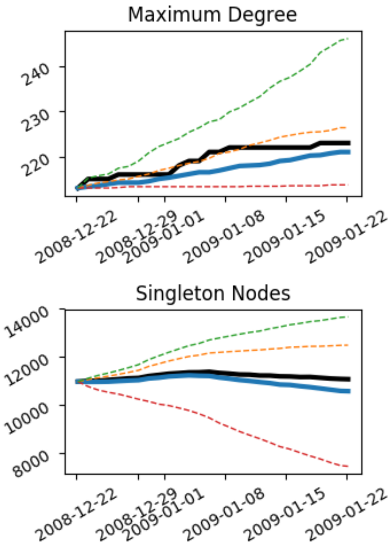


Correct model $M = 0.5M_2 + 0.5M_d(0.5)$ fitted
$M = \beta_1 M_2 + (1 - \beta_1)M_d(\alpha)$.



Mean squared degree
230
220
210
200
190
2008-12-22
2008-12-29
2009-01-01
2009-01-08
2009-01-15
2009-01-22
Degree Assortativity
0.25
0.24
0.23
0.22
0.21
2008-12-22
2008-12-29
2009-01-01
2009-01-08
2009-01-15
2009-01-22
Number of Triangles
2008-12-22
2008-12-29
2009-01-01
2009-01-08
2009-01-15
2009-01-22
Real
BestMix
DP
Rand
Tri

Real tests Facebook data (2)



Summary so far

- FETA can give an exact likelihood a model M underlies an observed dynamic graph.
- Given several models and a data set FETA can say which is the most likely to have produced that data.
- FETA can take several models and produce a “mixture” of all three.
- On artificial data sets produced from mixture models FETA can recover the parameters.
- On real data we can test mixtures of models improving replicated statistics.
- Next step, what if underlying model changes during lifetime of graph.

Models which change in time

- Motivating examples
  - A social network which changes its rules (default privacy assumptions, maximum number of followers).
  - A network associated with an organisation suffers a traumatic event (enron email network).
  - A network associated with a technological change (bitcoin network when chain forks, Internet as technology evolves).
- In this case it is likely the underlying model will change as time changes.
- How can we model this?

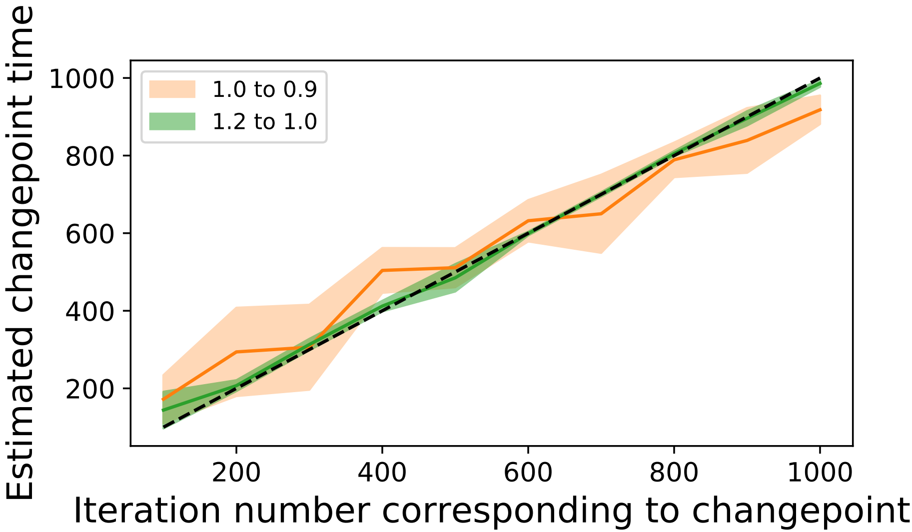
Models which change in time – artificial data

- Model changes once at time T. Use the "degree power" model and change the power.
- Let $p_i(t)$ be the probability of choosing i at time t.
- Let $d_i(t)$ be the degree of i at time t.

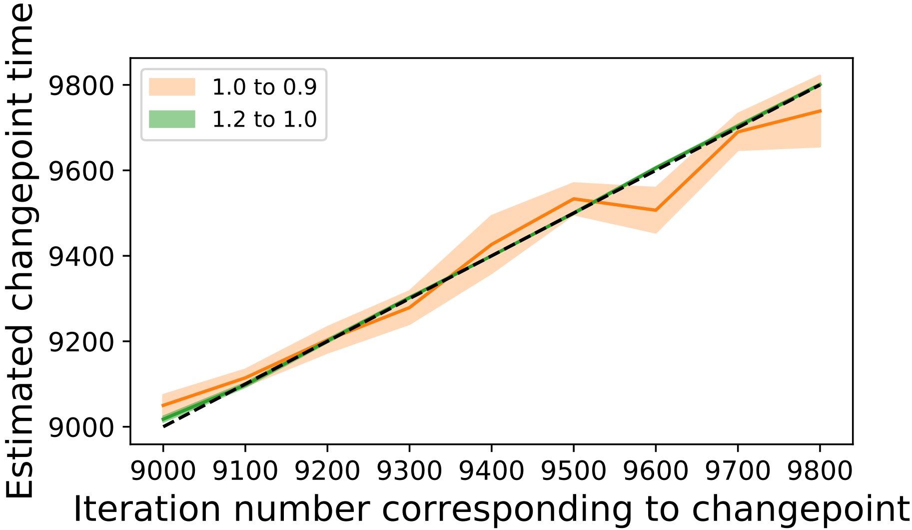
$$p_i(t) = \begin{cases} d_i(t)^\alpha / k & t < T \\ d_i(t)^\beta / k & t \geq T. \end{cases}$$

- Now we can grow a graph using this model for different α and β.
- Already shown for a long section of data we can recover the parameter.
- Can we locate the point at which the change occurs?

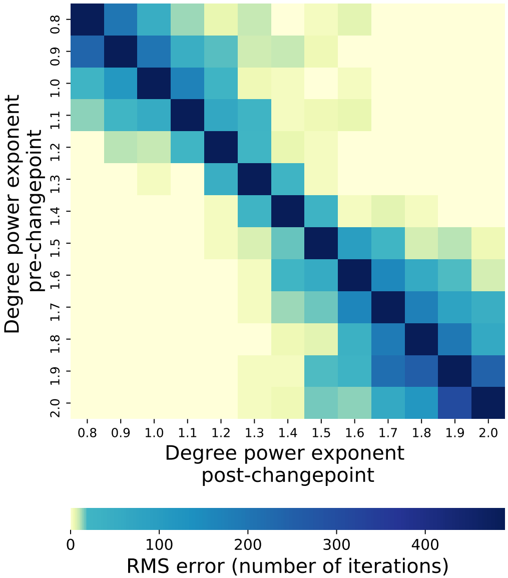
Finding the change point – 1000 node experiment



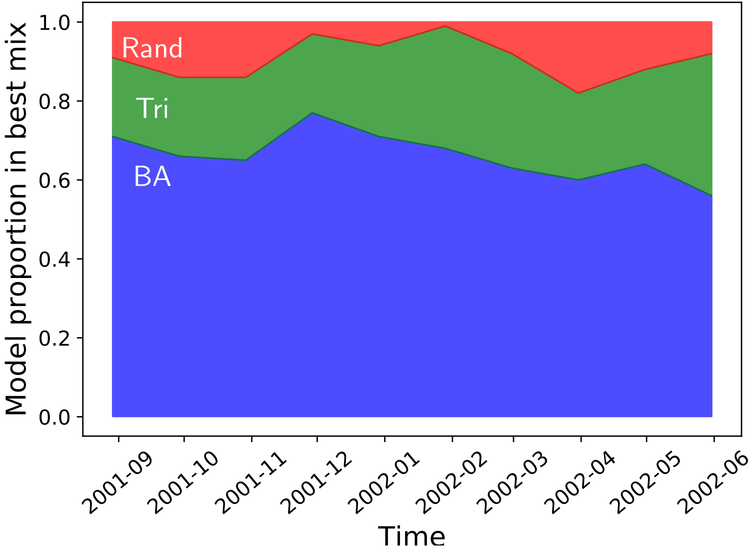
Finding the change point – 10000 node experiment



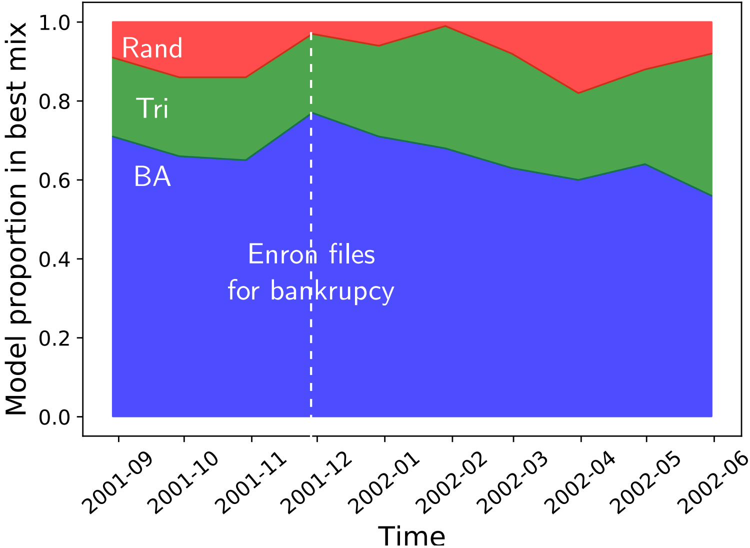
Which changes are easiest to detect?



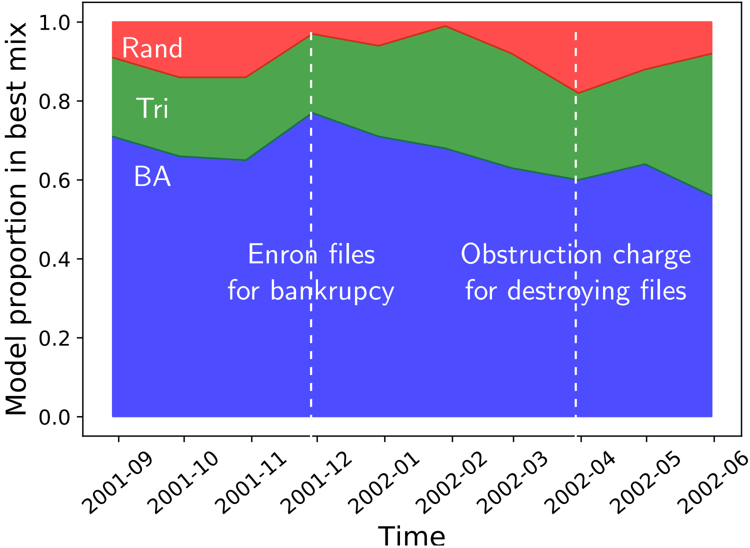
Changes in real data – Enron email network analysed as 3 model mixture 10 time periods



Changes in real data – Enron email network analysed as 3 model mixture 10 time periods



Changes in real data – Enron email network analysed as 3 model mixture 10 time periods



Conclusions

- The likelihood parameters and the null model here provide a rigorous way to assess a potential dynamic model of network evolution.
- 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.
  4. Models can be mixture of existing models and vary in time.
- Moving from network models based in physics to network models based in statistics.
- An exciting new way to test theories about topologies if you have the data for it.

Further work

- Ongoing work – lots still to do.
- Software and data freely available – please email `n.a.arnold@qmul.ac.uk`
- Code and data: `https://github.com/narnolddd/FETA3`