Uncovering the evolution of dynamic networks using temporal data

N. A. Arnold R. J. Mondragón R. G. Clegg

School of Electronic Engineering and Computer Science
Queen Mary University of London

Statistical Inference for Network Models, NetSci 2018

Dynamics of network formation

Looking at how local processes

- how individuals in a social network make new connections
- how scientists choose papers to cite

influence the eventual global structure of a network

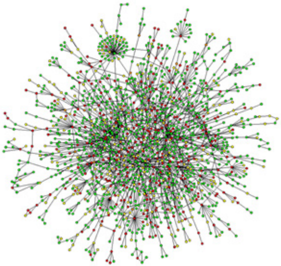


Figure: Saccharomyces cerevisiae protein-protein interaction network

Figure: Visualisation of Facebook graph

We use explanatory models to identify these mechanisms

How should we validate explanatory models?

Traditionally, based on their ability to reproduce networks with similar descriptive statistics on a to the network of interest such as: degree distribution $P(k)$, clustering coefficient, maximum degree.
Shortfalls of this approach:

- What if two possible models each perform better on different statistics?
- Which statistics should carry more weight?
- What if two different explanations give extremely similar end statistics?

I present an example of this last bullet point and a method to distinguish such models using temporal data.

An evolving network model template

Start with small connected network of m_0 nodes.

Label nodes $1, 2, \ldots, N(t)$ according to the order of their arrival.

At each iteration, add a node and connect to m existing nodes in the network.

Nodes are chosen without replacement from a distribution

$$\mathbb{P}(\text{choose node } i) = p_i, \quad \sum_{i=1}^{N} p_i = 1$$

Queen Mary
University of London

Two examples

The Barabási-Albert (BA) preferential attachment model sets $p_i \propto k_i$, the degree of node i.

- Nodes of higher degree have greater chance of attracting new links
- Dependent on network structure
- Theoretical scale-free degree distribution $P(k) \sim k^{-3}$

Two examples

The Barabási-Albert (BA) preferential attachment model sets $p_i \propto k_i$, the degree of node i.

- Nodes of higher degree have greater chance of attracting new links
- Dependent on network structure
- Theoretical scale-free degree distribution $P(k) \sim k^{-3}$

The rank preference (RP) model sets $p_i \propto i^{-\alpha}$.

- Longest established nodes have greater chance of attracting new links
- Independent of network structure
- Theoretical degree distribution $P(k) \sim k^{-\gamma}$ with $\gamma = 1 + 1/\alpha$

Henceforth let $\alpha = \frac{1}{2}$

Queen Mary
University of London

Degree distribution of realisation

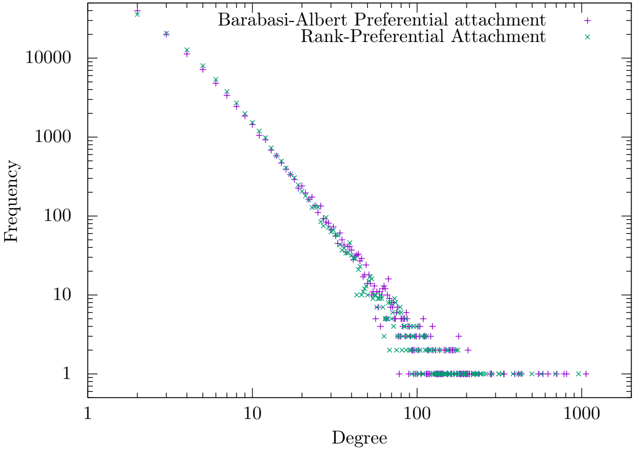
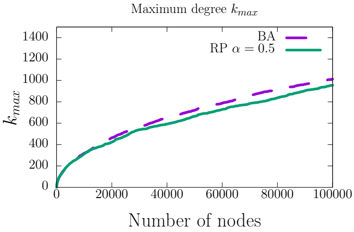
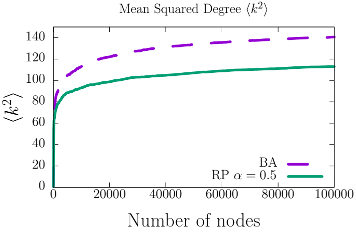


Figure: Degree distribution of realisation of BA (purple) and RP (green).

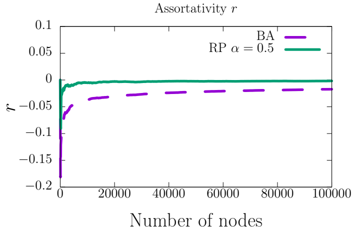
Evolution of other statistics



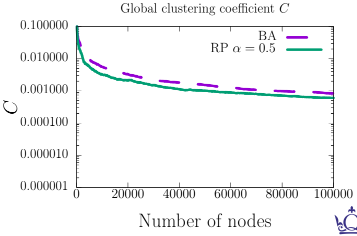
Maximum degree k_{max}



Mean squared degree $\langle k^2 \rangle$

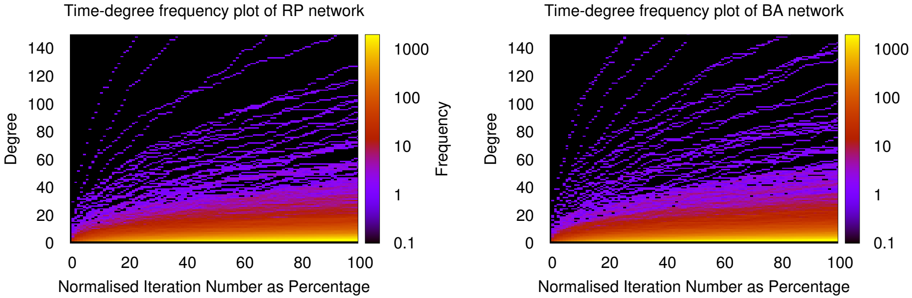


Assortativity



Clustering coefficient

Degree distributions over time



Rank Preference

Barabási-Albert

Can we distinguish the two models?

Introduce the mixture model $M(\beta)$, which gives probabilities of choosing a node as:

$$\mathbb{P}(\text{choose node } i) =$$

Can we distinguish the two models?

Introduce the mixture model $M(\beta)$, which gives probabilities of choosing a node as:

$$\mathbb{P}(\text{choose node } i) = \beta p_i^{\text{RP}} + (1-\beta) p_i^{\text{BA}}$$

where $\beta \in [0, 1]$, ie, a model that is part RP and part BA.

Can we distinguish the two models?

Introduce the mixture model $M(\beta)$, which gives probabilities of choosing a node as:

$$\begin{aligned}\mathbb{P}(\text{choose node } i) &= \beta p_i^{\text{RP}} + (1-\beta) p_i^{\text{BA}} \\ &= \beta \frac{i^{-\alpha}}{\sum_{j=1}^{N} j^{-\alpha}} + (1-\beta) \frac{k_i}{\sum_{j=1}^{N} k_j}\end{aligned}$$

where $\beta \in [0,1]$, ie, a model that is part RP and part BA.

Can we distinguish the two models?

Introduce the mixture model $M(\beta)$, which gives probabilities of choosing a node as:

$$\begin{aligned}\mathbb{P}(\text{choose node } i) &= \beta p_i^{\text{RP}} + (1-\beta) p_i^{\text{BA}} \\ &= \beta \frac{i^{-\alpha}}{\sum_{j=1}^{N} j^{-\alpha}} + (1-\beta) \frac{k_i}{\sum_{j=1}^{N} k_j}\end{aligned}$$

where $\beta \in [0,1]$, ie, a model that is part RP and part BA.
Given a synthetic network grown using model $M(\beta)$, can we reliably recover the parameter β?

Method: Model likelihood

[R. Clegg, B. Parker, M. Rio *Likelihood based assessment of network models*]

Definition

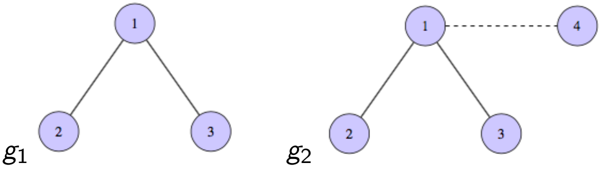
Let $G = G_t$ be an evolving network and g_t an observed snapshot, and let $M(\theta)$ be a probabilistic model. Then the likelihood of model $M(\theta)$ given the evolution sequence $\vec{g} = (g_1, g_2, \dots)$ of G is

$$L(M(\theta)|\vec{g}) = \mathbb{P}(G = \vec{g}|M(\theta))$$

Assuming we can calculate this likelihood, can fit model parameters by finding estimators which maximise the likelihood.
How do we calculate this?

Calculation of likelihood

Conditional probability of single observation:

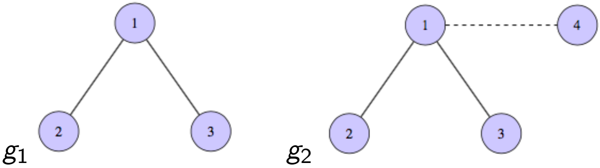


Example

Model adding node and one link at each timestep.

Calculation of likelihood

Conditional probability of single observation:

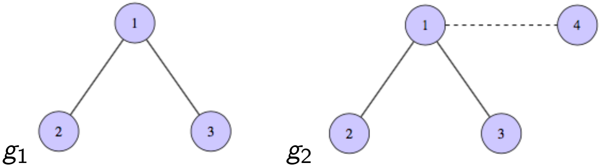


Example

Model adding node and one link at each timestep.
$L(\text{BA}|G_2 = g_2, G_1 = g_1) = \mathbb{P}_{\text{BA}}(\text{choose node 1}) = \frac{2}{1+2+1} = \frac{1}{2}$

Calculation of likelihood

Conditional probability of single observation:



Example

Model adding node and one link at each timestep.

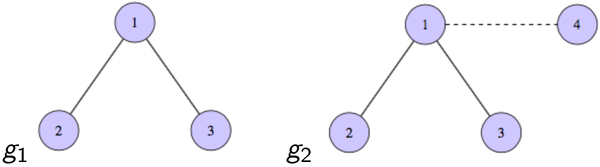
$L(\mathrm{BA}|G_2 = g_2, G_1 = g_1) = \mathbb{P}_{\mathrm{BA}}(\text{choose node 1}) = \frac{2}{1+2+1} = \frac{1}{2}$

$L(\mathrm{RP}|G_2 = g_2, G_1 = g_1) = \mathbb{P}_{\mathrm{RP}}(\text{choose node 1}) = \frac{1}{1+2^{-\frac{1}{2}}+3^{-\frac{1}{2}}} < \frac{1}{2}$

Queen Mary University of London

Calculation of likelihood

Conditional probability of single observation:



Example

Model adding node and one link at each timestep.

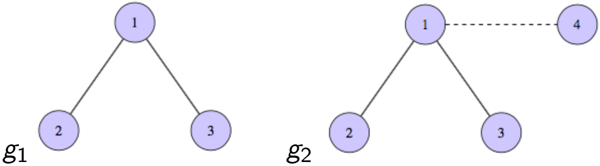
$L(\mathrm{BA}|G_2 = g_2, G_1 = g_1) = \mathbb{P}_{\mathrm{BA}}(\text{choose node 1}) = \frac{2}{1+2+1} = \frac{1}{2}$

$L(\mathrm{RP}|G_2 = g_2, G_1 = g_1) = \mathbb{P}_{\mathrm{RP}}(\text{choose node 1}) = \frac{1}{1+2^{-\frac{1}{2}}+3^{-\frac{1}{2}}} < \frac{1}{2}$

BA higher likelihood.

Calculation of likelihood

Conditional probability of single observation:



Theorem

Let $f_t(g_t|M(\theta)) = \mathbb{P}(G_t = g_t|g_{t-1}, g_{t-2}, \ldots, M(\theta))$. *Then*

$$L(M(\theta)|\vec{g}) = \prod_t f_t(g_t|M(\theta))$$

Experiment and Result

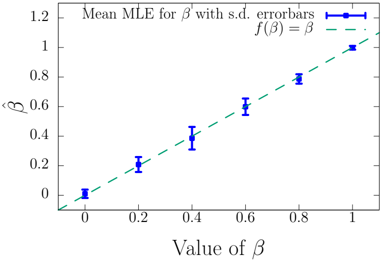
For $\beta = 0, 0.2, \ldots, 1$ we

1. Grew artificial networks to 10,000 nodes, adding a node at each timestep and connecting to m existing nodes with probabilities defined by $M(\beta)$.
2. Calculated maximum likelihood estimators $\hat{\beta}$ for β.
3. Repeated 10 times and obtained mean/sd.

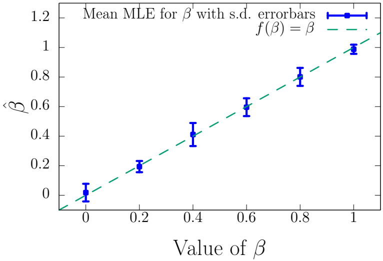
Experiment and Result

For $\beta = 0, 0.2, \ldots, 1$ we

1. Grew artificial networks to 10,000 nodes, adding a node at each timestep and connecting to m existing nodes with probabilities defined by $M(\beta)$.
2. Calculated maximum likelihood estimators $\hat{\beta}$ for β.
3. Repeated 10 times and obtained mean/sd.



m=1



m=3

Example: StackExchange MathOverflow Dataset

[A. Paranjape, A. R. Benson, and J. Leskovec: *Motifs in temporal networks*]

Online mathematics based Q & A forum.
Nodes are users and an edge can represent any interaction between two users:

- Answering a user's question
- Commenting on a question or user's answer to a question

mathoverflow

Queen Mary
University of London

Example: StackExchange MathOverflow Dataset

[A. Paranjape, A. R. Benson, and J. Leskovec: *Motifs in temporal networks*]

Online mathematics based Q & A forum.
Nodes are users and an edge can represent any interaction between two users:

- Answering a user's question
- Commenting on a question or user's answer to a question

We tested the model $M(\beta)$ where node probabilities are given as

$$\mathbb{P}(\text{choose node } i) = \beta p_i^{\text{RP}} + (1-\beta) p_i^{\text{BA}}$$

Example: StackExchange MathOverflow Dataset

[A. Paranjape, A. R. Benson, and J. Leskovec: *Motifs in temporal networks*]

Online mathematics based Q & A forum.
Nodes are users and an edge can represent any interaction between two users:

- Answering a user's question
- Commenting on a question or user's answer to a question

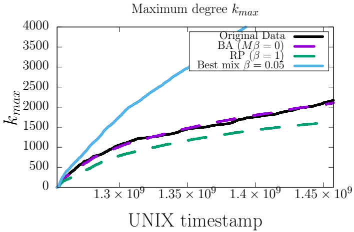
We tested the model $M(\beta)$ where node probabilities are given as

$$\mathbb{P}(\text{choose node } i) = \beta p_i^{\text{RP}} + (1-\beta) p_i^{\text{BA}}$$

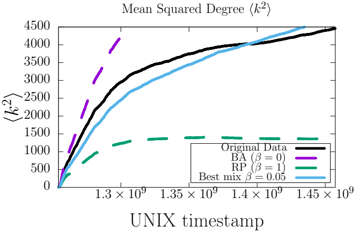
and found that $\beta = 0.05$ gives the maximum likelihood.

Queen Mary University of London

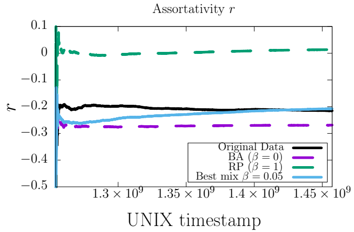
Best mixture model compared to non-mixed models



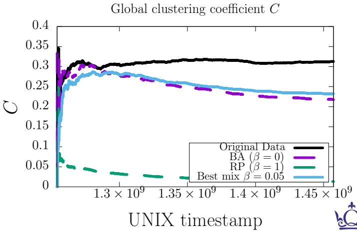
Maximum degree k_{max}



Second moment $\langle k^2 \rangle$



Assortativity



Clustering coefficient

Queen Mary
University of London

Conclusions & future directions

- Temporal data allows deeper understanding of mechanisms governing network evolution and opportunity to go beyond comparisons of snapshots.
- Micro-scale information about individual node and link arrivals can be used to find model likelihoods and validate explanations.
- We have a way of distinguishing very similar explanatory models when temporal data is available.
- Idea of model mixtures may be useful for modelling networks arising from a mixture of mechanisms.

Final

Thanks for listening!
Code available at https://github.com/narnolddd/FETA2
Dataset available at SNAP:
http://snap.stanford.edu/data/sx-mathoverflow.html

Questions?

if(timeleft$>\epsilon$): Degree Trichotomy vs TPA

The degree trichotomy model sets $p_i \propto \hat{k}_i$ where $\hat{k}_i = \begin{cases} L & k_i \leq L \\ k_i & L < k_i \leq U \\ U & k_i > U \end{cases}$

where L and U constants.

The temporal preferential attachment model batches nodes into time intervals $I_1, I_2, \ldots$ of equal size according to their arrival time. A new node arriving in the most recent time period I_t will choose m nodes to connect to by repeatedly:

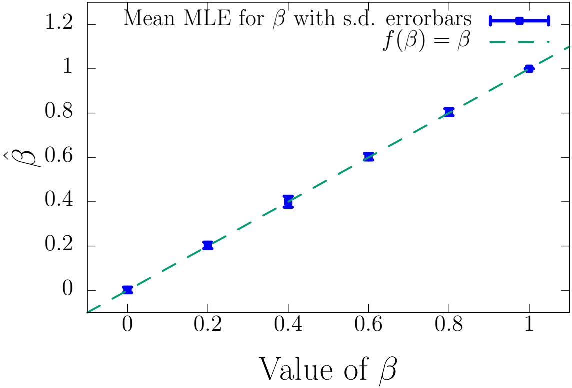
1. picking a time period with $\mathbb{P}(\text{choose } I_T) = f(t - T)$ where f is a decaying function (preferring more recent time intervals)
2. picking a node within that time interval according to Barabási-Albert preferential attachment.

Queen Mary
University of London

Result

Use a mixture model $M(\beta)$ assigning node probabilities

$$p_i = \beta p_i^{\mathrm{TPA}} + (1-\beta) p_i^{\mathrm{DT}}$$



Appendix: Copying network transformations

To grow the networks in Stack Exchange figure, we extracted from the edgelist the sequence of operations of the network's evolution, e.g.:

Time	Operation
1	New node added with 3 links
2	New link between existing nodes
3	New link between existing nodes
4	New node added with 5 links
⋮	⋮

and grew networks with the corresponding sequence, with node probabilities provided by choice of model M

Queen Mary
University of London