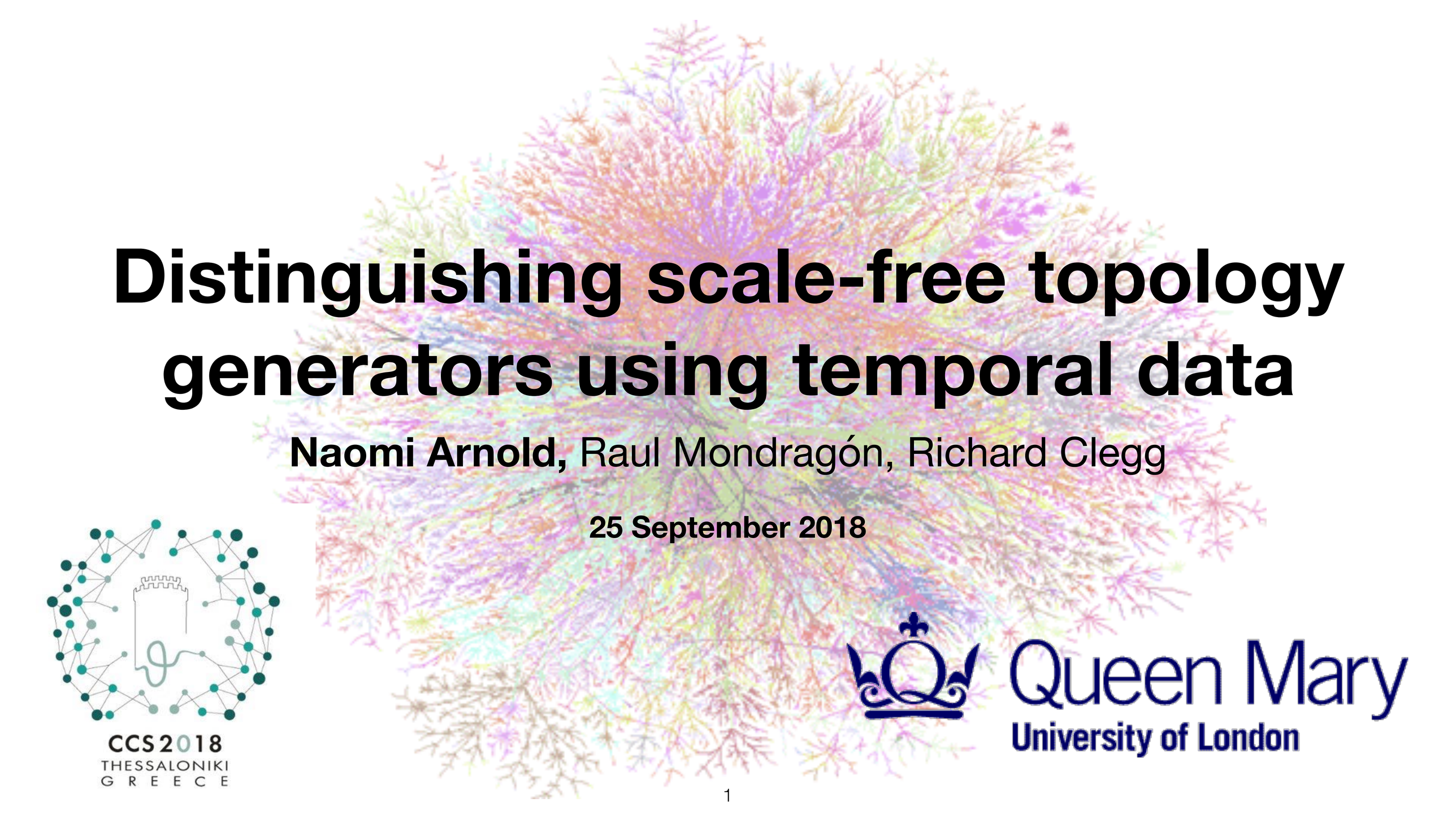


Distinguishing scale-free topology generators using temporal data

Naomi Arnold, Raul Mondragón, Richard Clegg

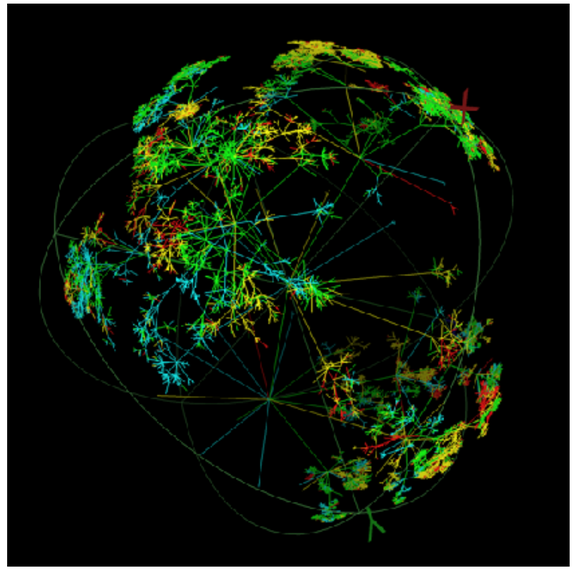
25 September 2018

CCS2018
THESSALONIKI
GREECE

Queen Mary
University of London

Motivation

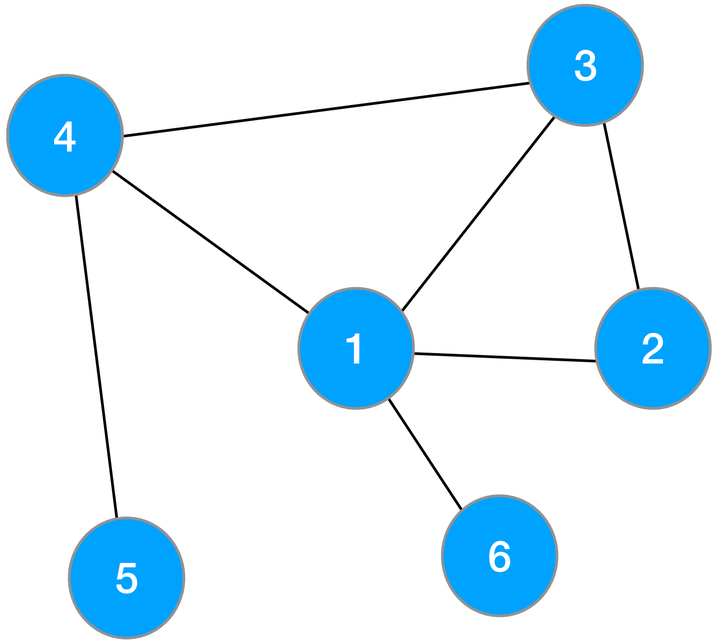
- The structure of many networks of interest is often dynamic in nature
- Rich temporal data for network topologies is becoming more prevalent
- This allows us to better investigate models for network growth over time



Model for evolving networks

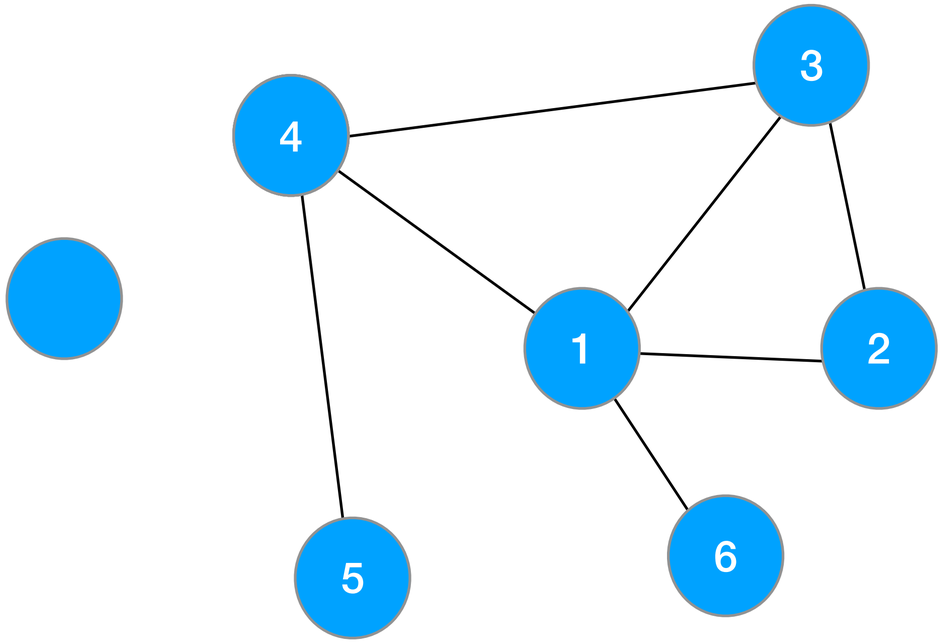
Model for evolving networks

Seed network



Model for evolving networks

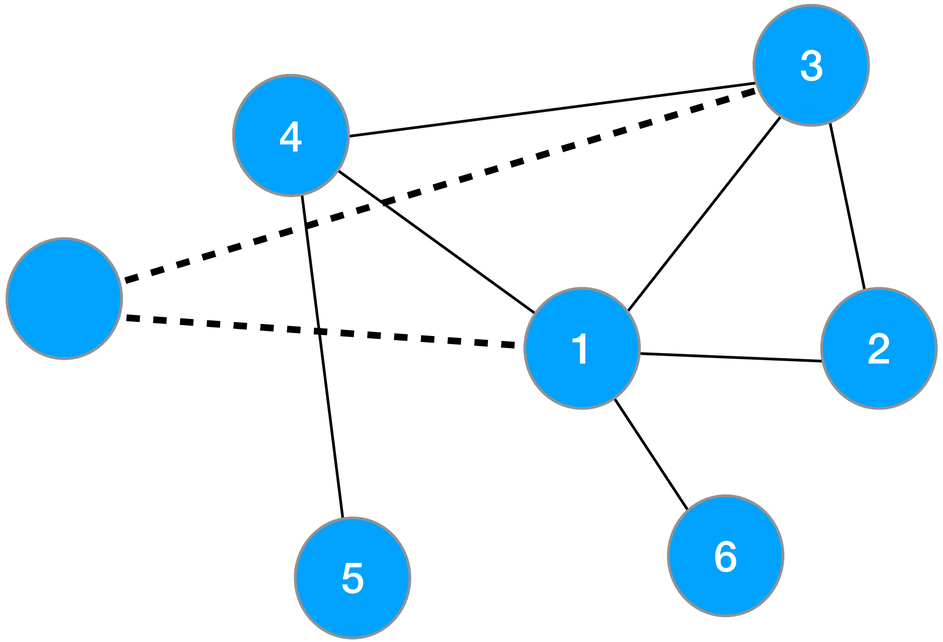
Seed network



Model for evolving networks

Action (new node/internal link)

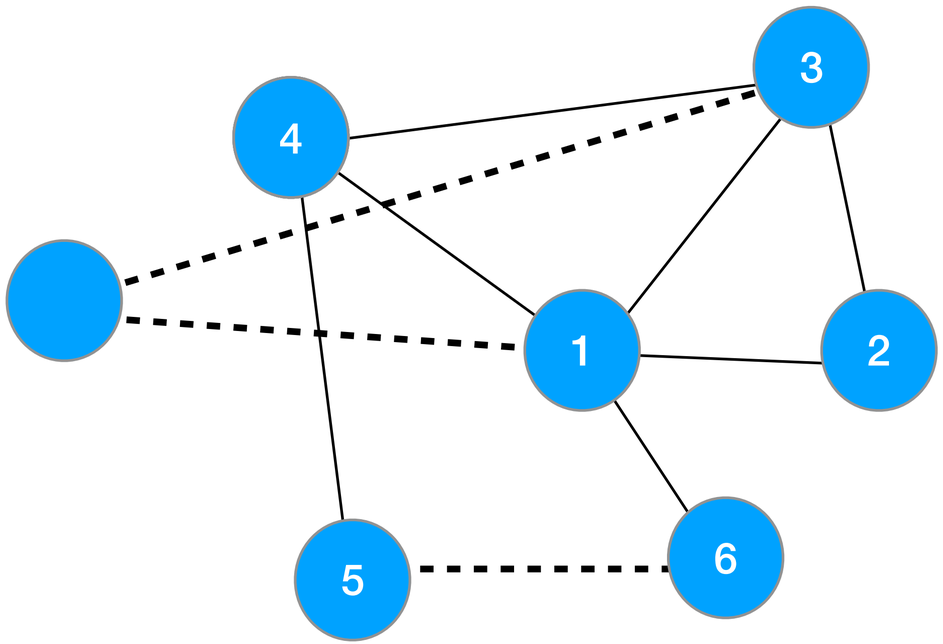
Seed network



Model for evolving networks

Action (new node/internal link)

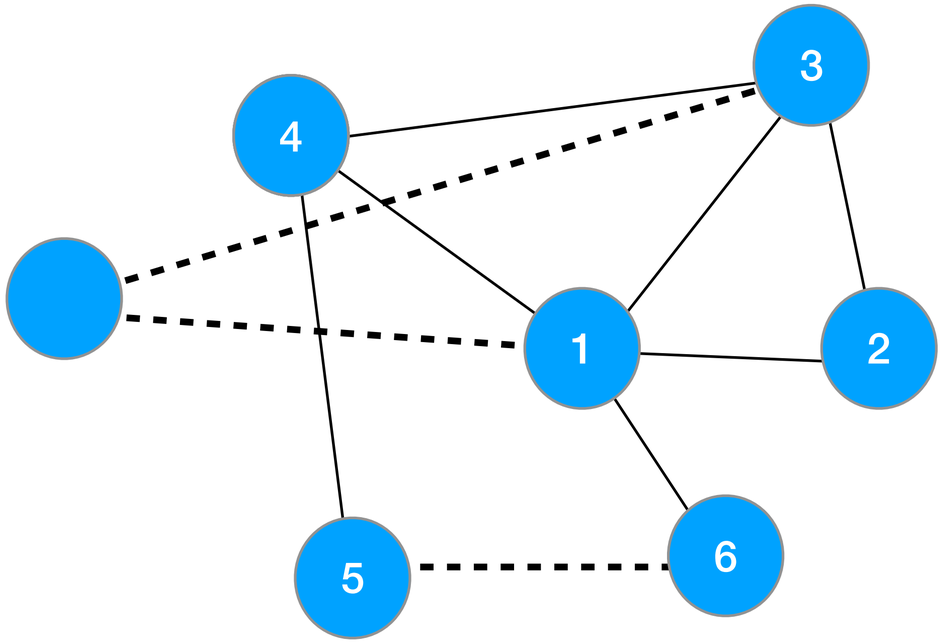
Seed network



Model for evolving networks

Action (new node/internal link)

Seed network



Attachment probabilities

$$p_i$$

corresponding to

$\mathbb{P}$(**choose node** i)

Attachment probabilities

Attachment probabilities

$$p_i \propto 1$$

Random/neutral model. All nodes equally likely

Attachment probabilities

$$p_i \propto 1$$

Random/neutral model. All nodes equally likely

$$p_i \propto f(k_i)$$

Function of node i's degree, e.g. BA model

Attachment probabilities

$$p_i \propto 1$$

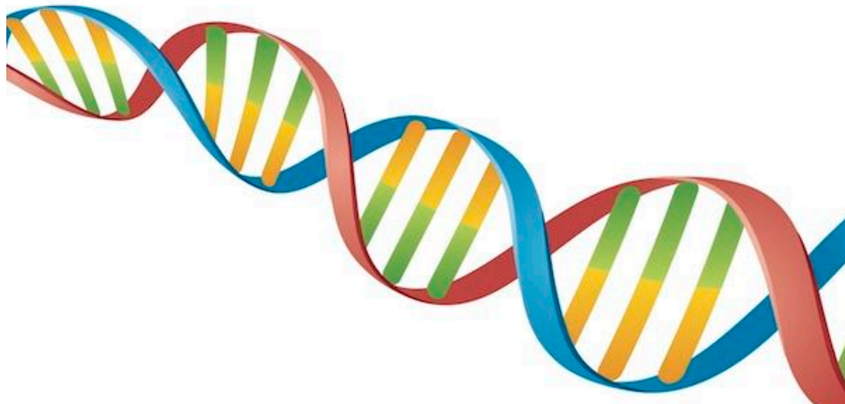
Random/neutral model. All nodes equally likely

$$p_i \propto f(k_i)$$

Function of node i's degree, e.g. BA model

$$p_i \propto f(\eta_i)$$

Function of some other intrinsic node property



How do we quantify how good a fit a model is to real data?

How do we quantify how good a fit a model is to real data?

- **Traditional approach:** generate a network from model of the same size as target network, and compare on different statistics

How do we quantify how good a fit a model is to real data?

- **Traditional approach:** generate a network from model of the same size as target network, and compare on different statistics
- **Problem:** generating large networks is time consuming

How do we quantify how good a fit a model is to real data?

- **Traditional approach:** generate a network from model of the same size as target network, and compare on different statistics
- **Problem:** generating large networks is time consuming
- **Problem:** two networks having similar properties doesn’t necessarily mean their evolution was governed in the same way

How do we quantify how good a fit a model is to real data?

- **Traditional approach:** generate a network from model of the same size as target network, and compare on different statistics
- **Problem:** generating large networks is time consuming
- **Problem:** two networks having similar properties doesn't necessarily mean their evolution was governed in the same way ***Why?***

Two different models for scale-free networks

[A.-L. Barabasi, R. Albert 1999: Emergence of scaling in random networks]

[S. Fortunato, A. Flammini, F. Menczer 2006: Scale-free network growth by ranking]

Two different models for scale-free networks

Barabási Albert model

$$p_i \propto k_i$$

Higher degree nodes likelier to attract new links

[A.-L. Barabasi, R. Albert 1999: Emergence of scaling in random networks]

[S. Fortunato, A. Flammini, F. Menczer 2006: Scale-free network growth by ranking]

Two different models for scale-free networks

Barabási Albert model

$$p_i \propto k_i$$

Higher degree nodes likelier to attract new links

Mean field prediction:

$$P(k) \sim k^{-3}$$

[A.-L. Barabasi, R. Albert 1999: Emergence of scaling in random networks]

[S. Fortunato, A. Flammini, F. Menczer 2006: Scale-free network growth by ranking]

Two different models for scale-free networks

Barabási Albert model

$$p_i \propto k_i$$

Higher degree nodes likelier to attract new links

Mean field prediction:

$$P(k) \sim k^{-3}$$

[A.-L. Barabasi, R. Albert 1999: Emergence of scaling in random networks]

Rank-preference model

$$p_i \propto i^{-\alpha}$$

Older nodes likelier to attract new links

[S. Fortunato, A. Flammini, F. Menczer 2006: Scale-free network growth by ranking]

Two different models for scale-free networks

Barabási Albert model

$$p_i \propto k_i$$

Higher degree nodes likelier to attract new links

Mean field prediction:

$$P(k) \sim k^{-3}$$

[A.-L. Barabasi, R. Albert 1999: Emergence of scaling in random networks]

Rank-preference model

$$p_i \propto i^{-\alpha}$$

Older nodes likelier to attract new links

Mean field prediction:

$$P(k) \sim k^{-(1+1/\alpha)}$$

[S. Fortunato, A. Flammini, F. Menczer 2006: Scale-free network growth by ranking]

Two different models for scale-free networks

Barabási Albert model

$$p_i \propto k_i$$

Higher degree nodes likelier to attract new links

Mean field prediction:

$$P(k) \sim k^{-3}$$

[A.-L. Barabasi, R. Albert 1999: Emergence of scaling in random networks]

Rank-preference model

$$p_i \propto i^{-\alpha}$$

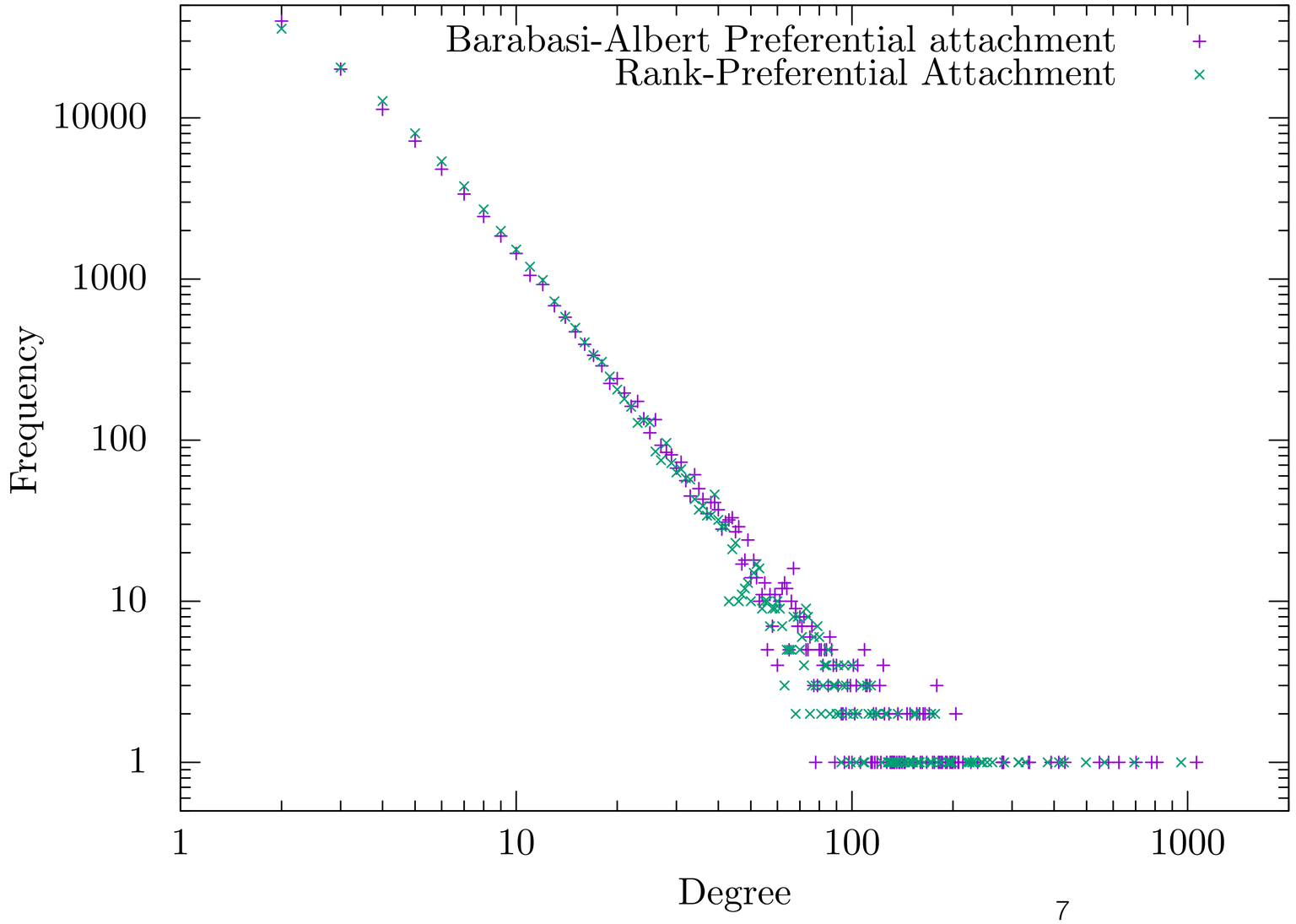
Older nodes likelier to attract new links

Mean field prediction:

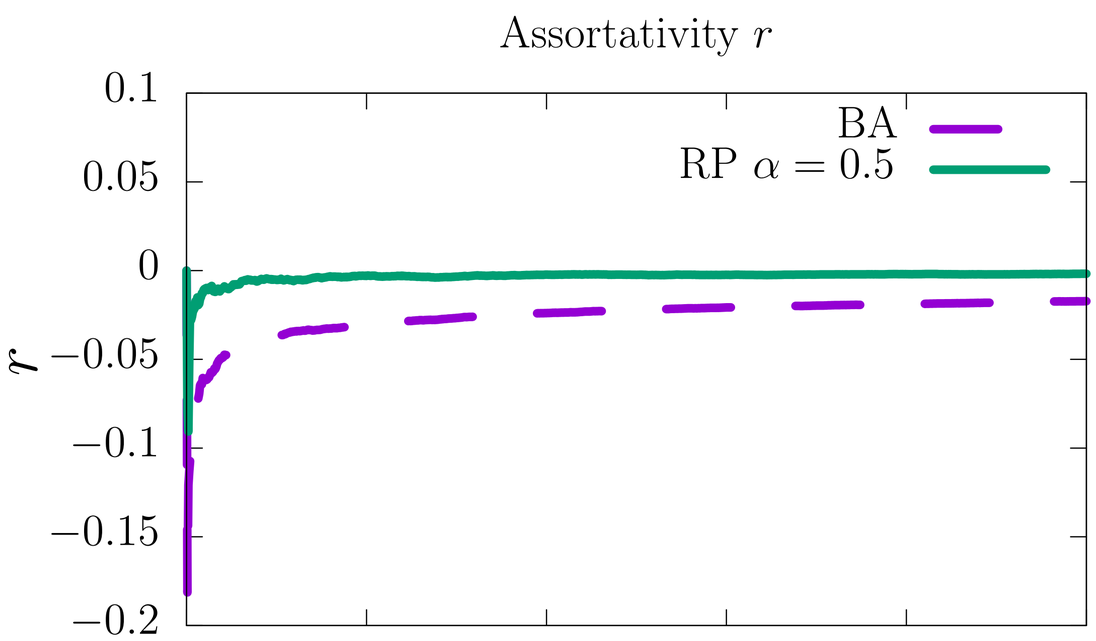
$$P(k) \sim k^{-(1+1/\alpha)}$$

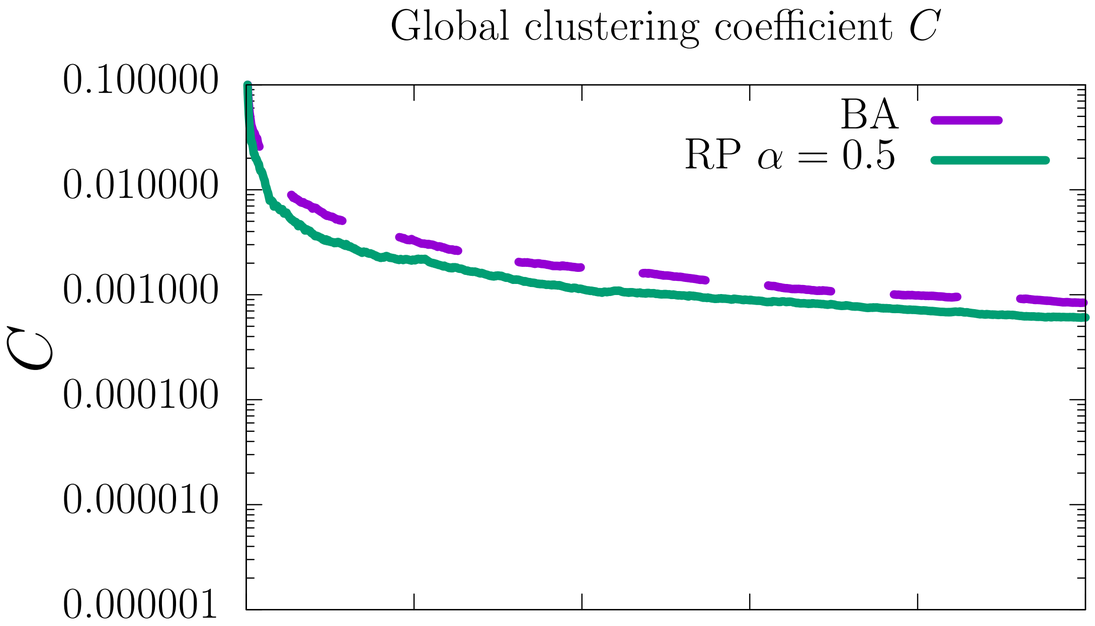
[S. Fortunato, A. Flammini, F. Menczer 2006: Scale-free network growth by ranking]

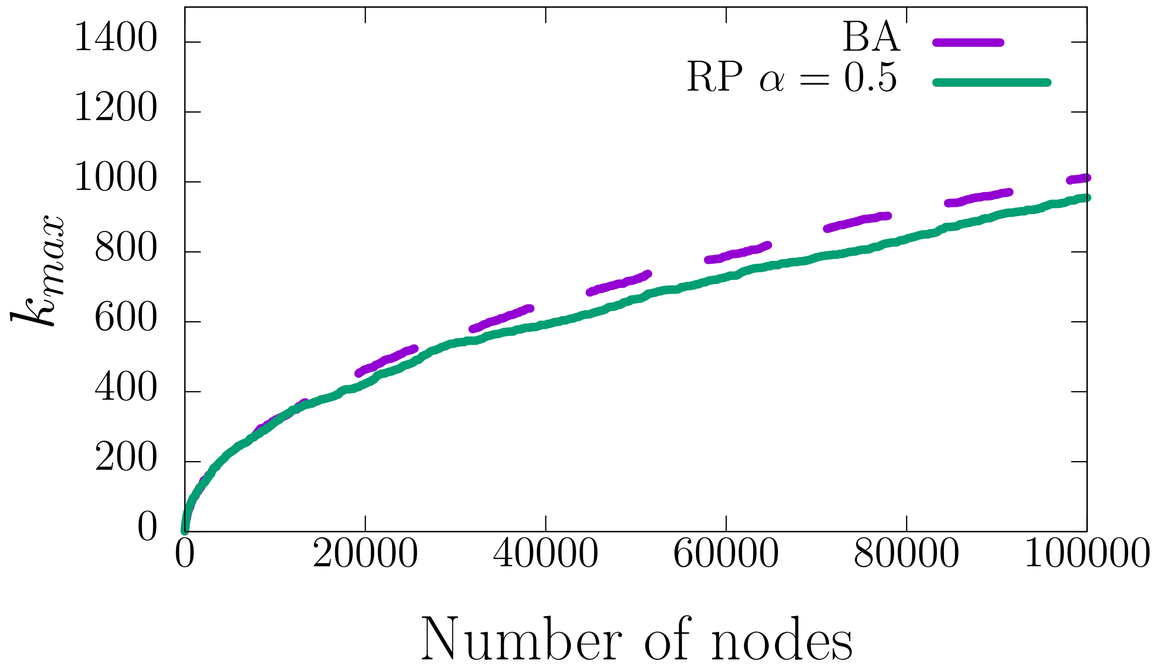
Degree distribution

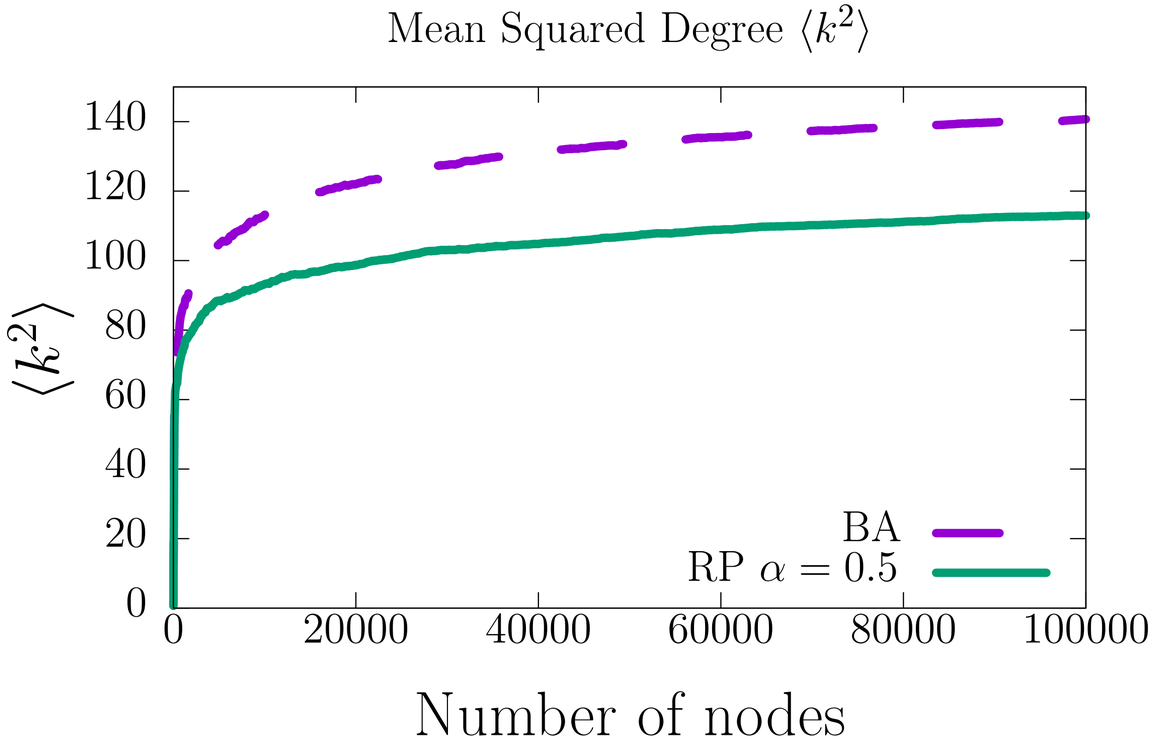


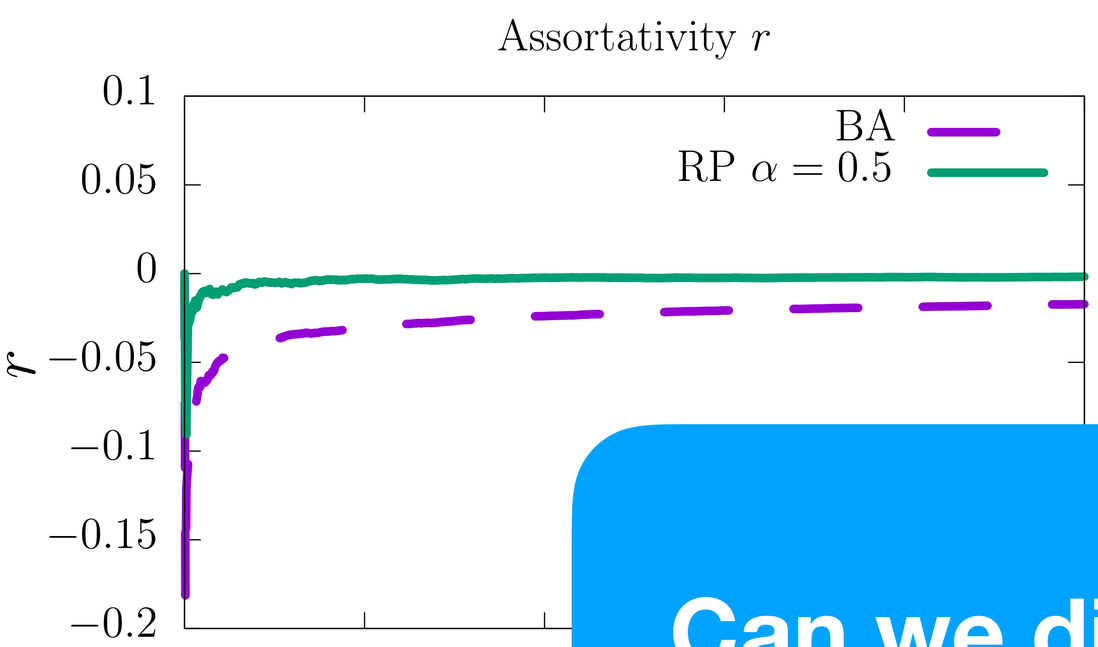
Network of 100,000 nodes generated from the Barabási-Albert and Rank Preference models, each new node having initially 3 neighbours

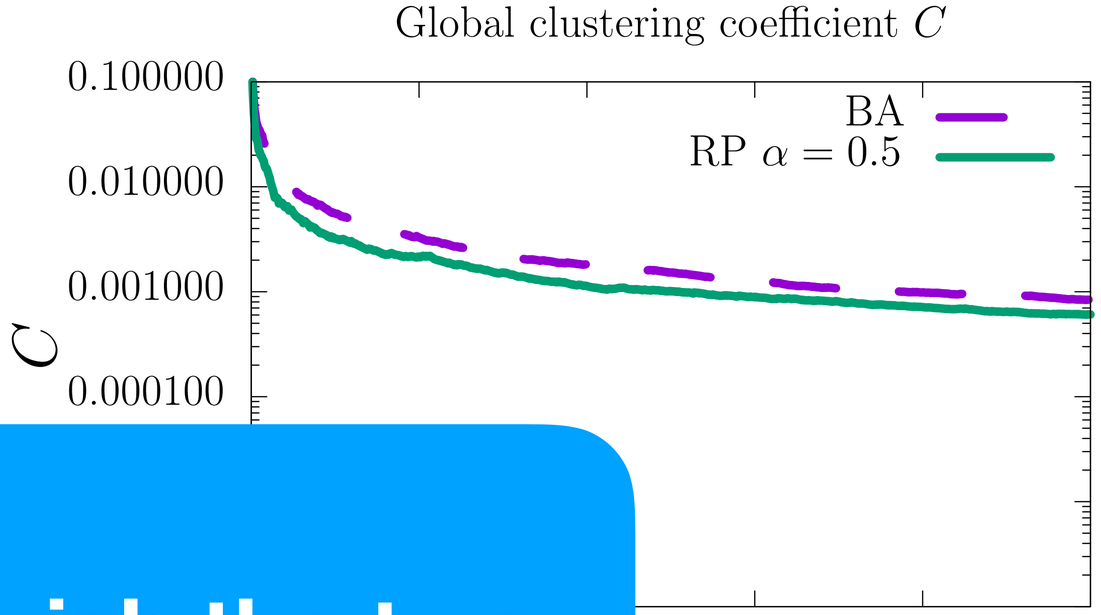

Assortativity r
BA
RP $\alpha = 0.5$
r
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2


Global clustering coefficient C
BA
RP $\alpha = 0.5$
C
0.100000
0.010000
0.001000
0.000100
0.000010
0.000001

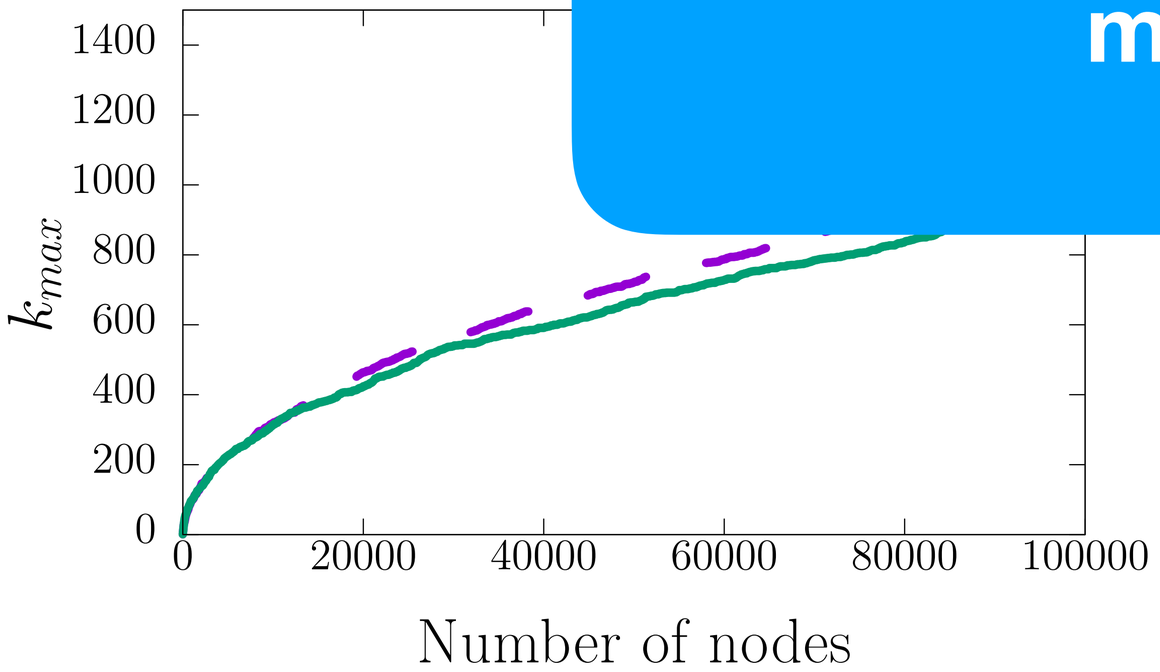

Maximum degree k_{max}
BA
RP $\alpha = 0.5$
k_{max}
1400
1200
1000
800
600
400
200
0
0
20000
40000
60000
80000
100000
Number of nodes

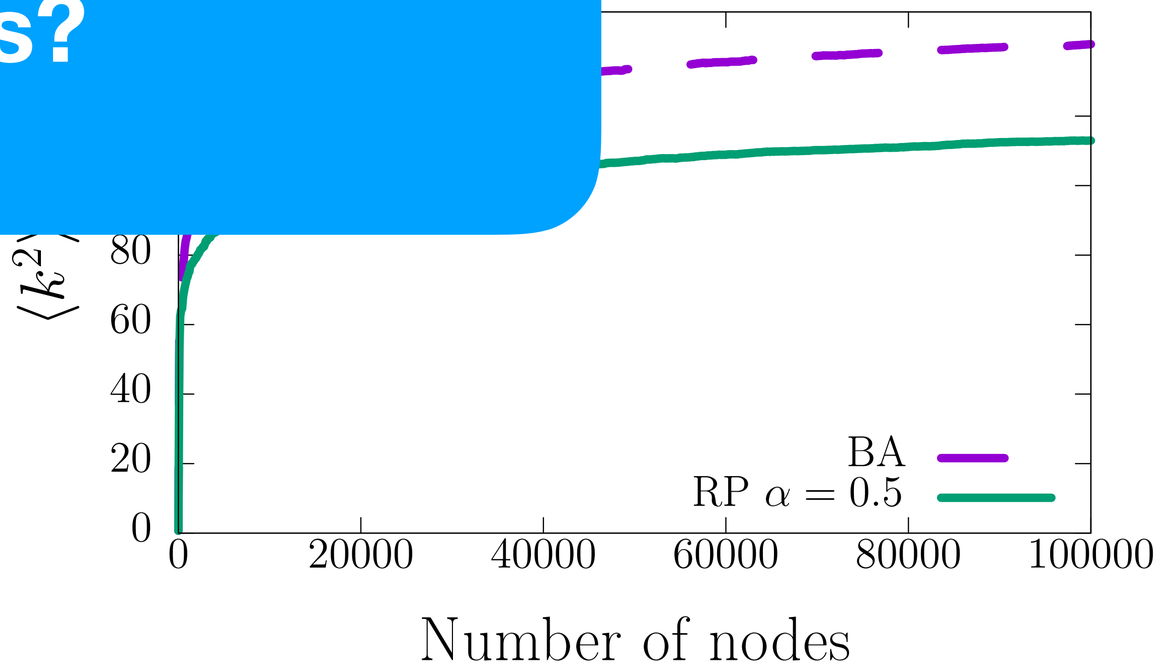

Mean Squared Degree $\langle k^2 \rangle$
BA
RP $\alpha = 0.5$
$\langle k^2 \rangle$
140
120
100
80
60
40
20
0
0
20000
40000
60000
80000
100000
Number of nodes





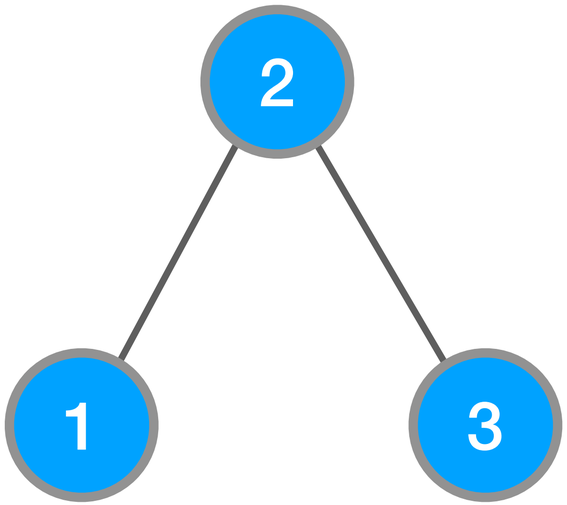
Can we distinguish the two models?





Different approach: Model likelihood

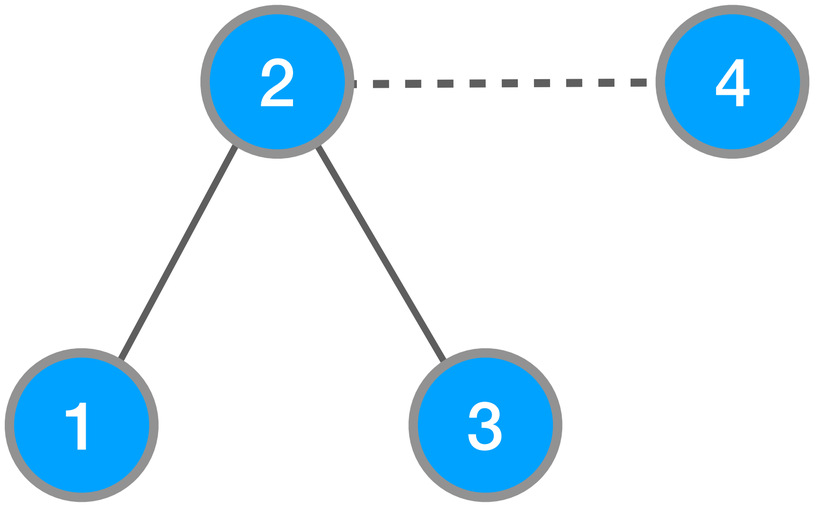
Likelihood of model given observation = probability of seeing observation given model



[R. Clegg, B. Parker, M. Rio 2016: Likelihood-based assessment of dynamic network models]

Different approach: Model likelihood

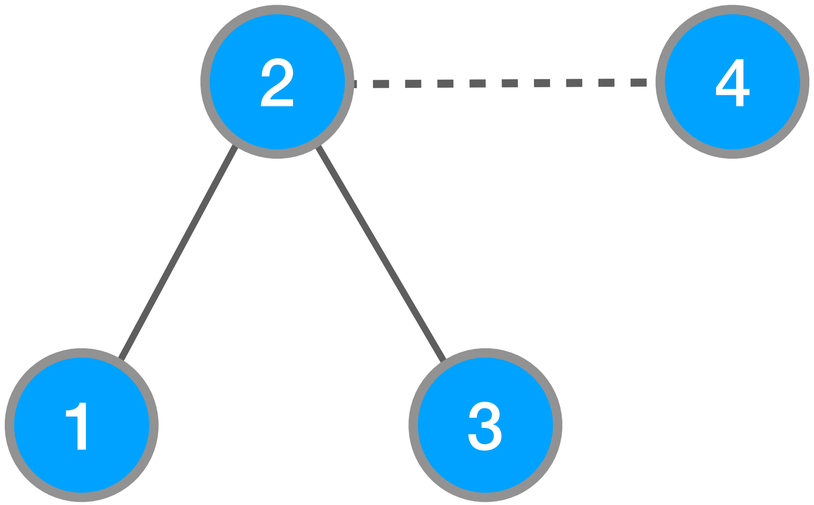
Likelihood of model given observation = probability of seeing observation given model



[R. Clegg, B. Parker, M. Rio 2016: Likelihood-based assessment of dynamic network models]

Different approach: Model likelihood

Likelihood of model given observation = probability of seeing observation given model



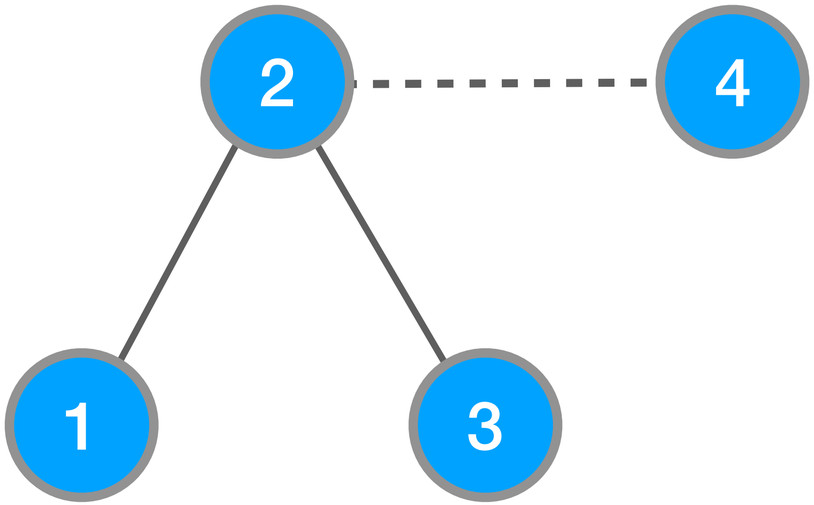
Likelihood of random/uniform model given by:

$$\mathbb{P}_{rand}(\textbf{Choose node } 2) = \frac{1}{3}$$

[R. Clegg, B. Parker, M. Rio 2016: Likelihood-based assessment of dynamic network models]

Different approach: Model likelihood

Likelihood of model given observation = probability of seeing observation given model



Likelihood of random/uniform model given by:

$$\mathbb{P}_{rand}(\textbf{Choose node } 2) = \frac{1}{3}$$

Likelihood of BA preferential attachment model given by:

$$\mathbb{P}_{BA}(\textbf{Choose node } 2) = \frac{2}{1+2+1} = \frac{1}{2}$$

[R. Clegg, B. Parker, M. Rio 2016: Likelihood-based assessment of dynamic network models]

Likelihood: Remarks

Likelihood: Remarks

- Quickly calculated, compared to generating networks

Likelihood: Remarks

- Quickly calculated, compared to generating networks
- Given a number of models, can define the ‘best’ as that which has the highest likelihood

Likelihood: Remarks

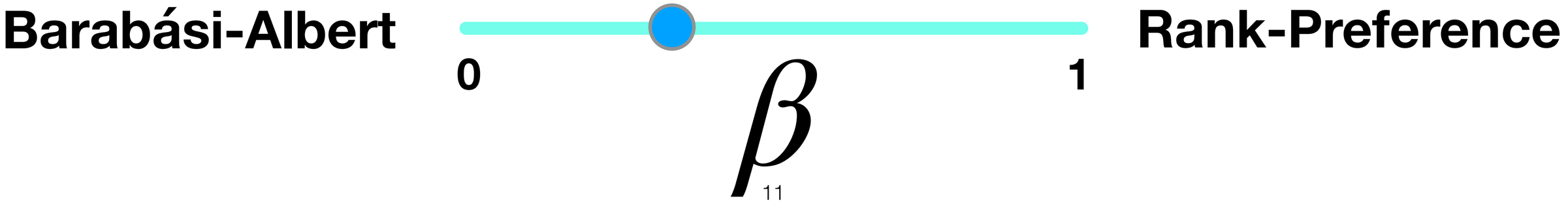
- Quickly calculated, compared to generating networks
- Given a number of models, can define the ‘best’ as that which has the highest likelihood
- For models with parameters, can find maximum likelihood estimators for params

Model Mixtures

Hypothesis: network growth likely to be governed by a mixture of mechanisms, not just one

Example: Mixture of BA and rank preference model

$$p_i = (1 - \beta)p_i^{BA} + \beta p_i^{RP}$$

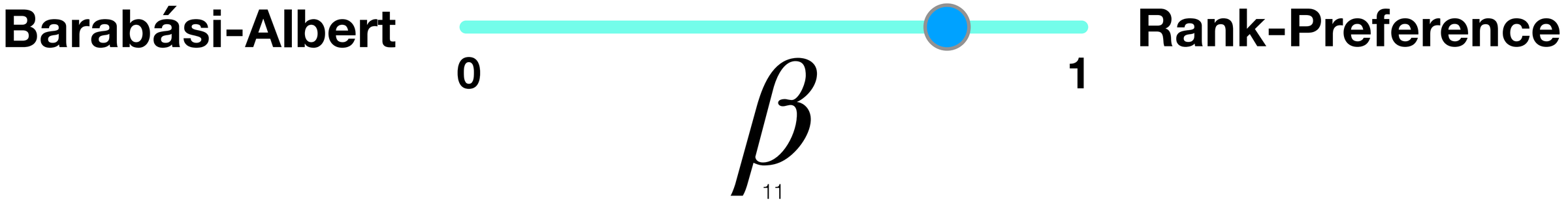


Model Mixtures

Hypothesis: network growth likely to be governed by a mixture of mechanisms, not just one

Example: Mixture of BA and rank preference model

$$p_i = (1 - \beta)p_i^{BA} + \beta p_i^{RP}$$

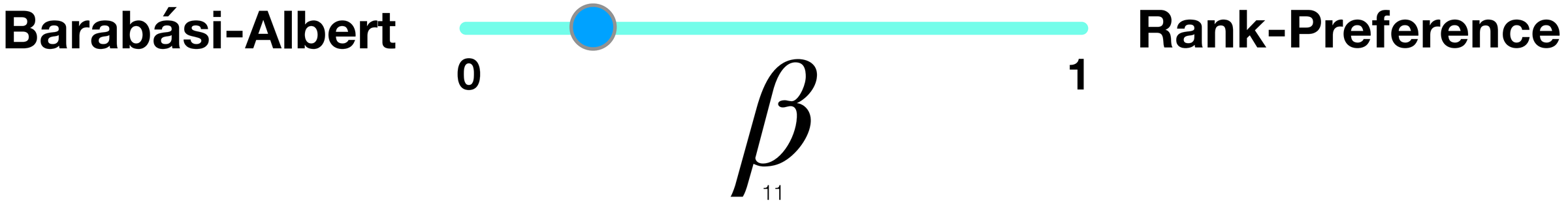


Model Mixtures

Hypothesis: network growth likely to be governed by a mixture of mechanisms, not just one

Example: Mixture of BA and rank preference model

$$p_i = (1-\beta)p_i^{BA} + \beta p_i^{RP}$$



Distinguishing using maximum likelihood estimation

Having generated artificial networks using:

$$p_i = \beta p_i^{RP} + (1 - \beta) p_i^{BA}$$

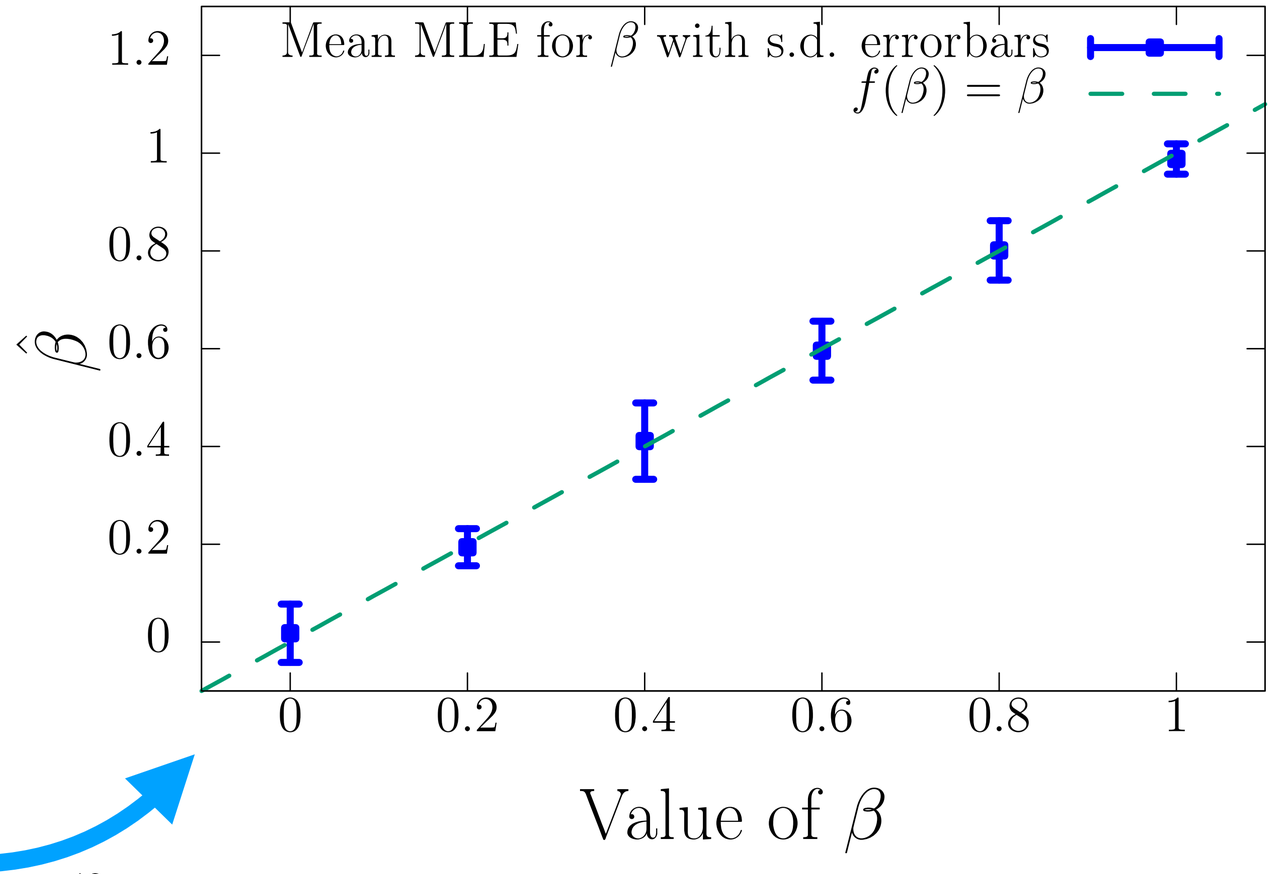
We can accurately recover the proportion β as an MLE!

Distinguishing using maximum likelihood estimation

Having generated artificial networks using:

$$p_i = \beta p_i^{RP} + (1 - \beta) p_i^{BA}$$

We can accurately recover the proportion β as an MLE!



Real data example: Math Overflow Social Network

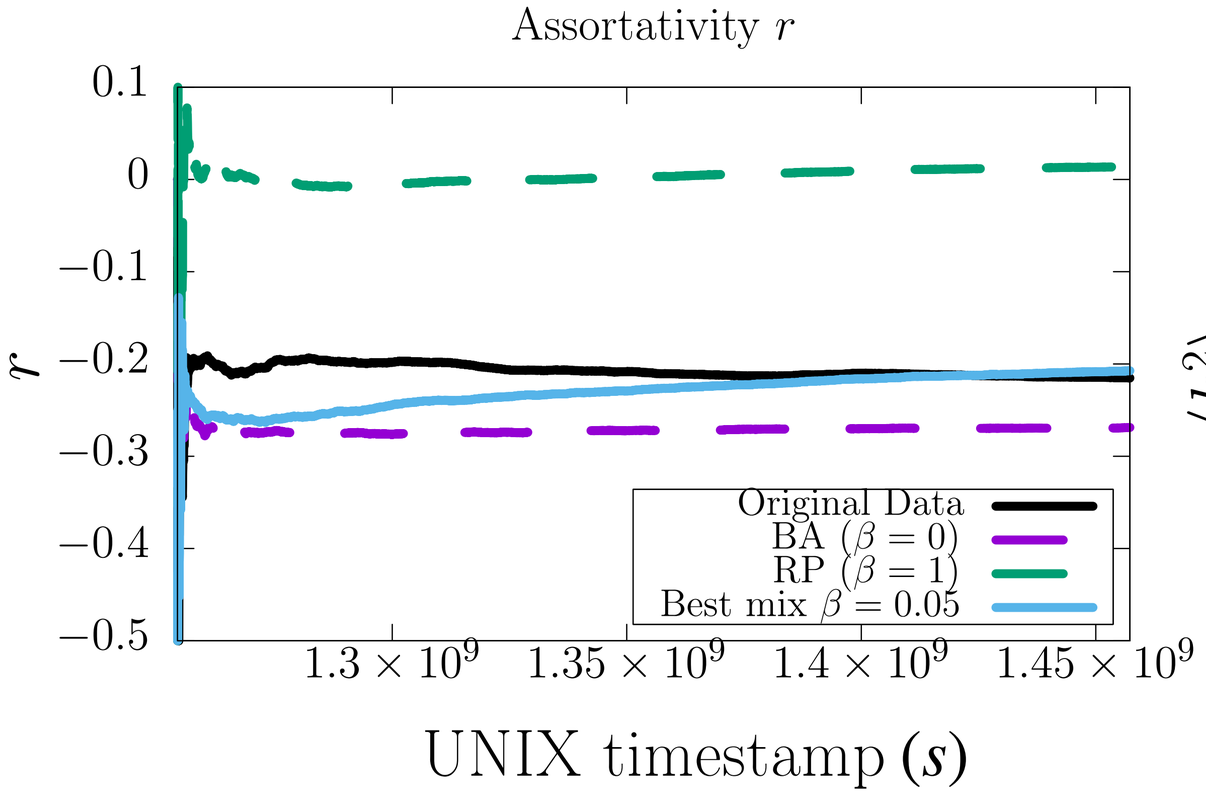
Dataset from [A. Paranjape, A. Benson, J. Leskovec 2017: Motifs in temporal networks]

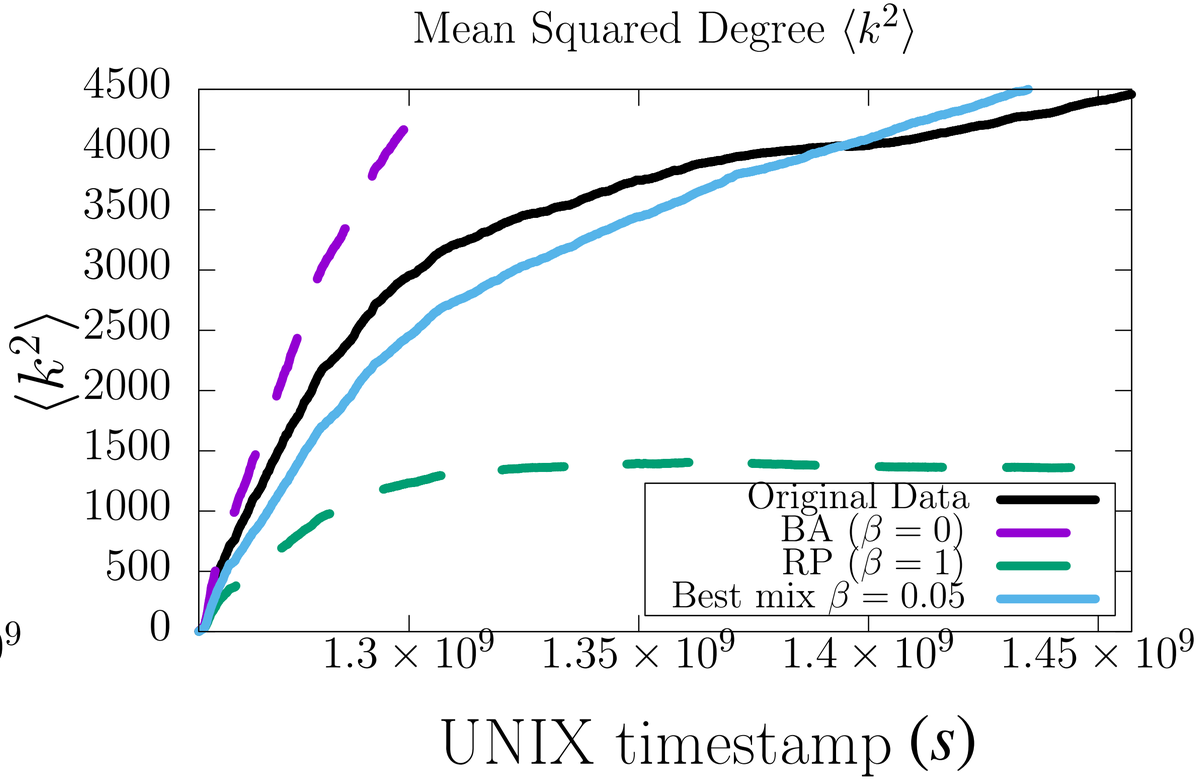
Real data example: Math Overflow Social Network

- Q&A site for mathematical problems
- Nodes are users
- An undirected edge between node A and B if A answers a question by B, A comments on B's answer or question
- Multiple edges collapsed
- Models components tested: BA, static rank preference

Dataset from [A. Paranjape, A. Benson, J. Leskovec 2017: Motifs in temporal networks]

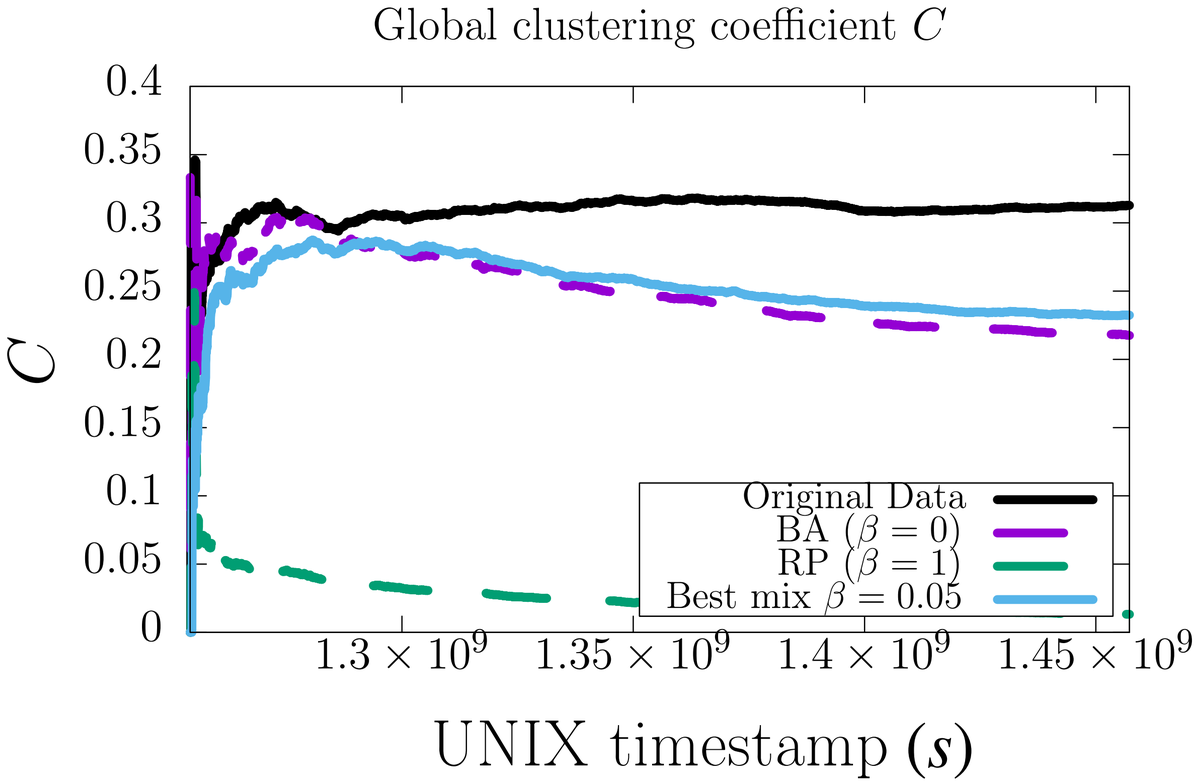
Real data example: Math Overflow Social Network

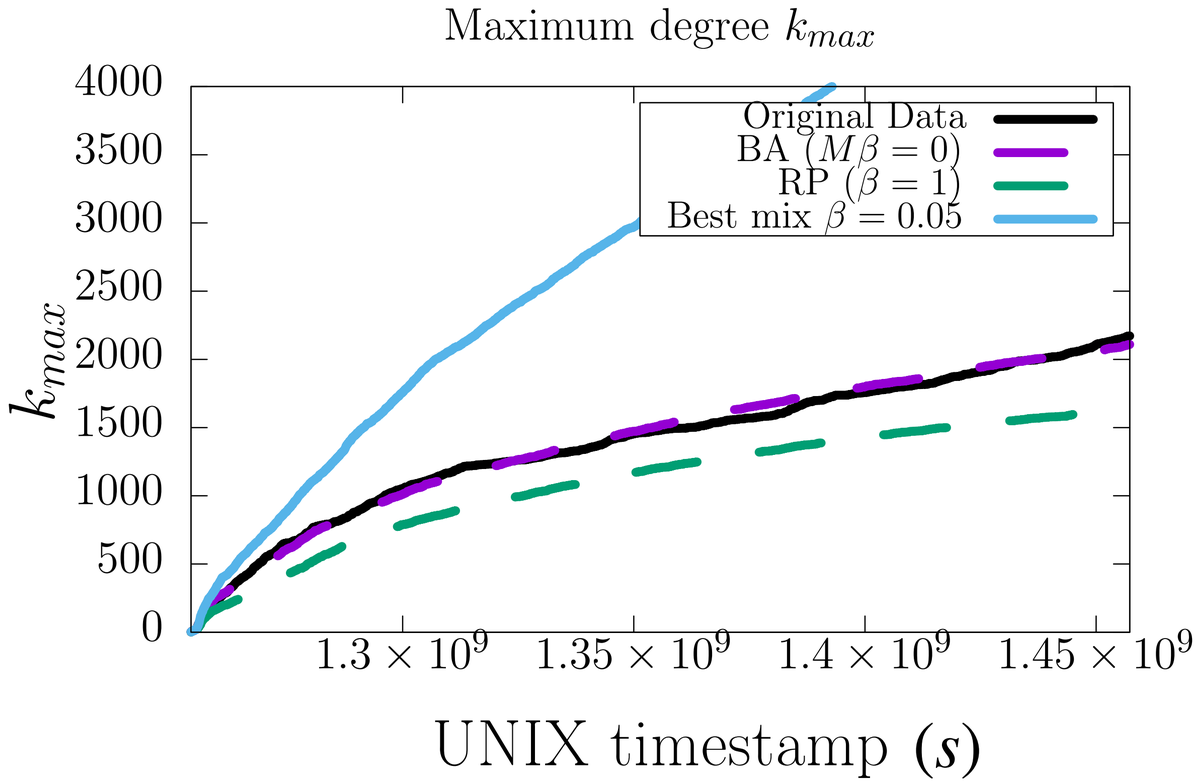




Dataset from [A. Paranjape, A. Benson, J. Leskovec 2017: Motifs in temporal networks]

Real data example: Math Overflow Social Network





Dataset from [A. Paranjape, A. Benson, J. Leskovec 2017: Motifs in temporal networks]

Real data example: Routeviews AS topology dataset

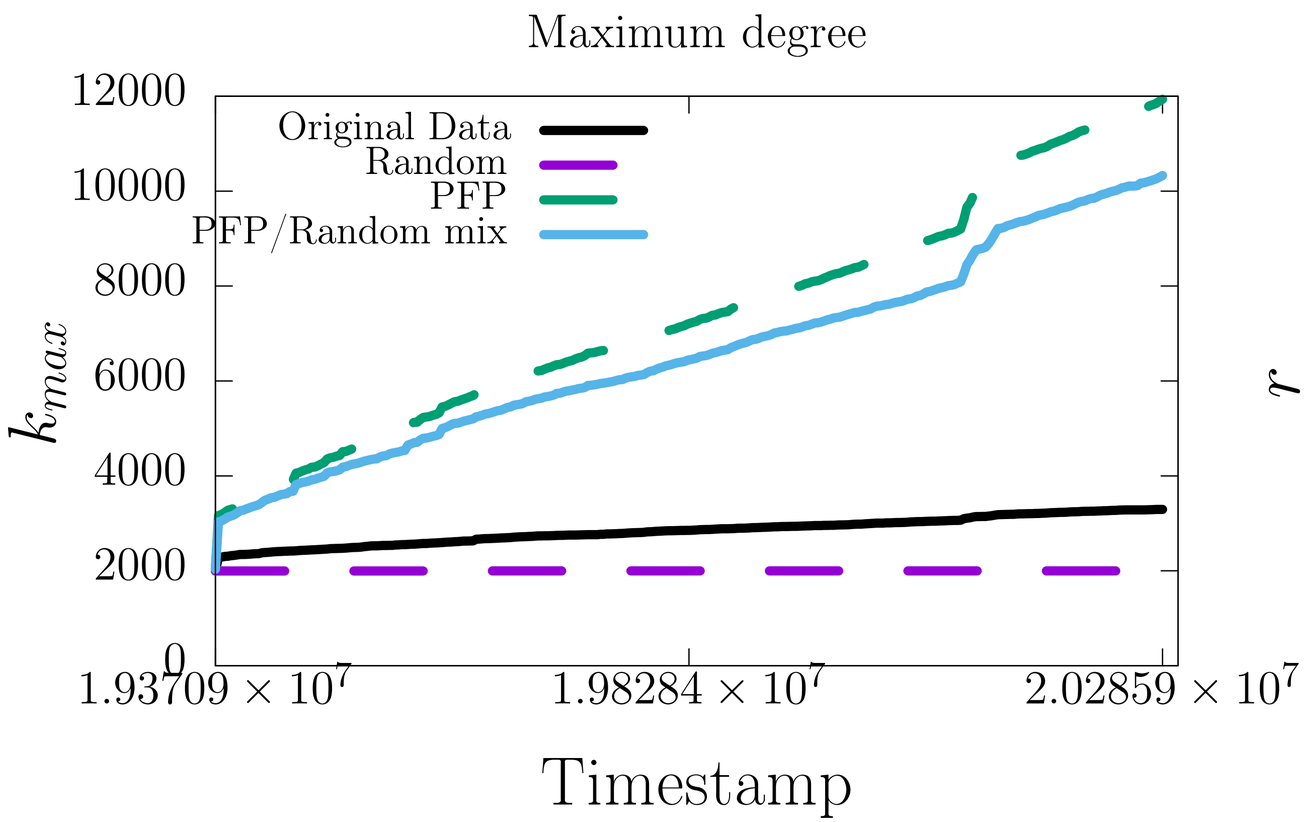
- Timestamped dataset: nodes autonomous systems and links represent peering relationship
- Model components tested: random/uniform model, positive-feedback preference model

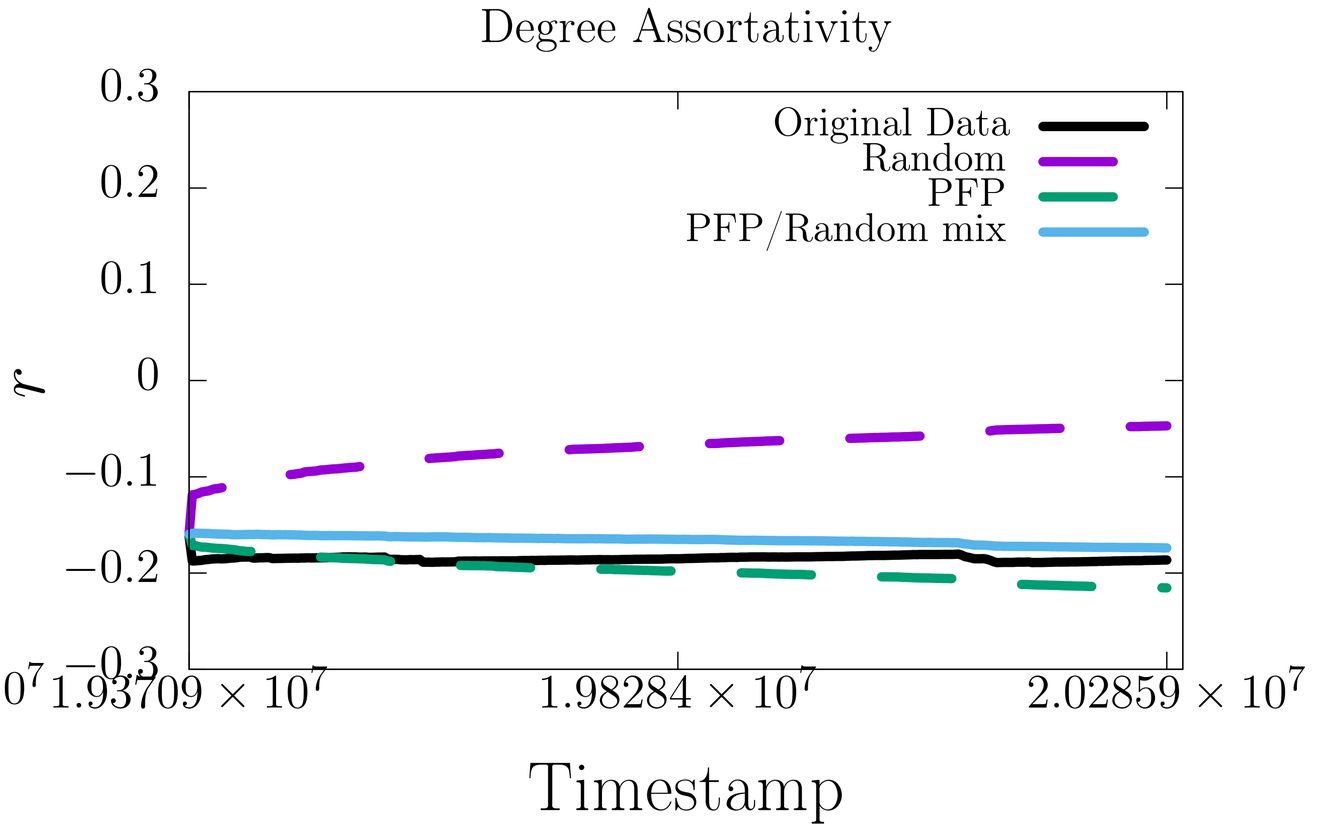
Note - positive-feedback preference model gives:

$$p_i \propto k_i^{1+\delta \log_{10} k_i}$$

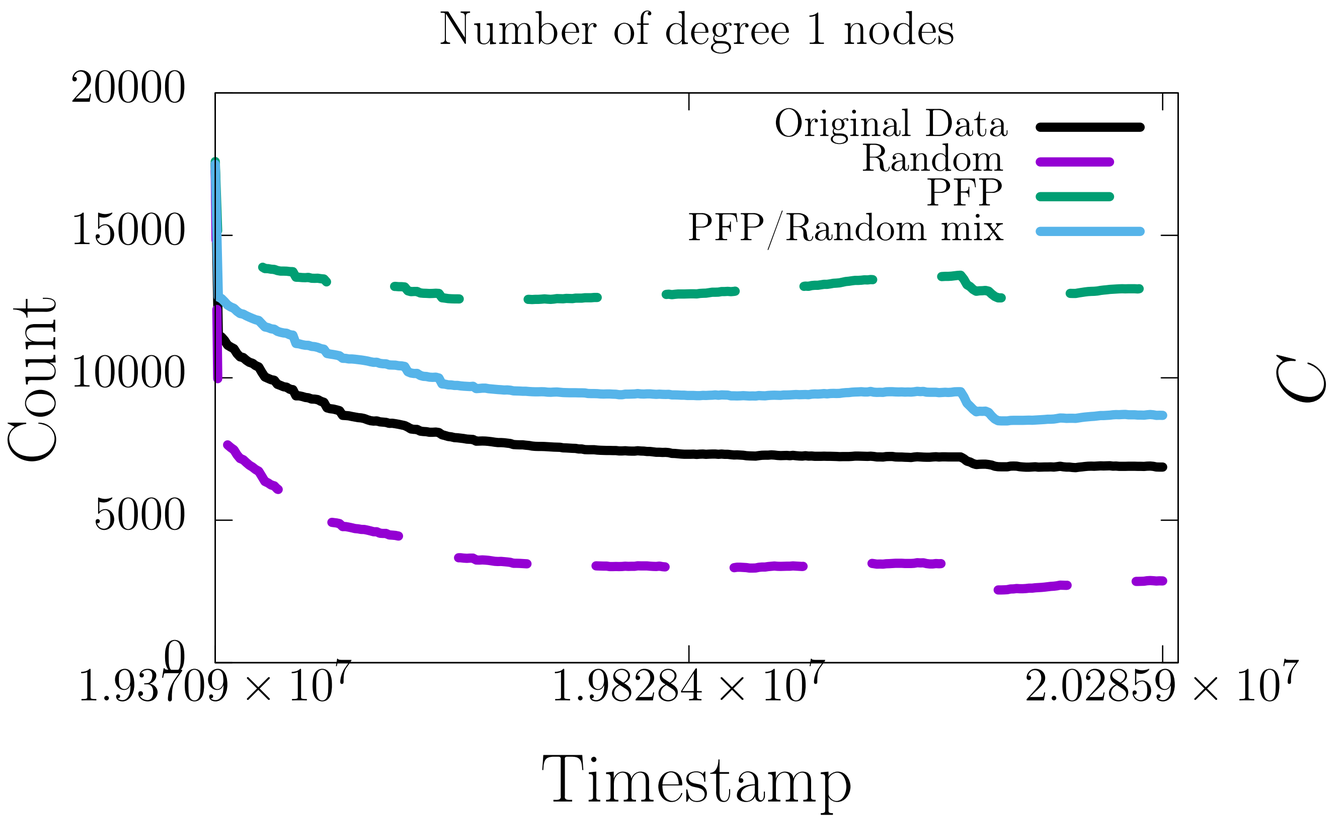
PFP model: [S. Zhou, R. Mondragón 2004: Accurately modelling the Internet Topology]

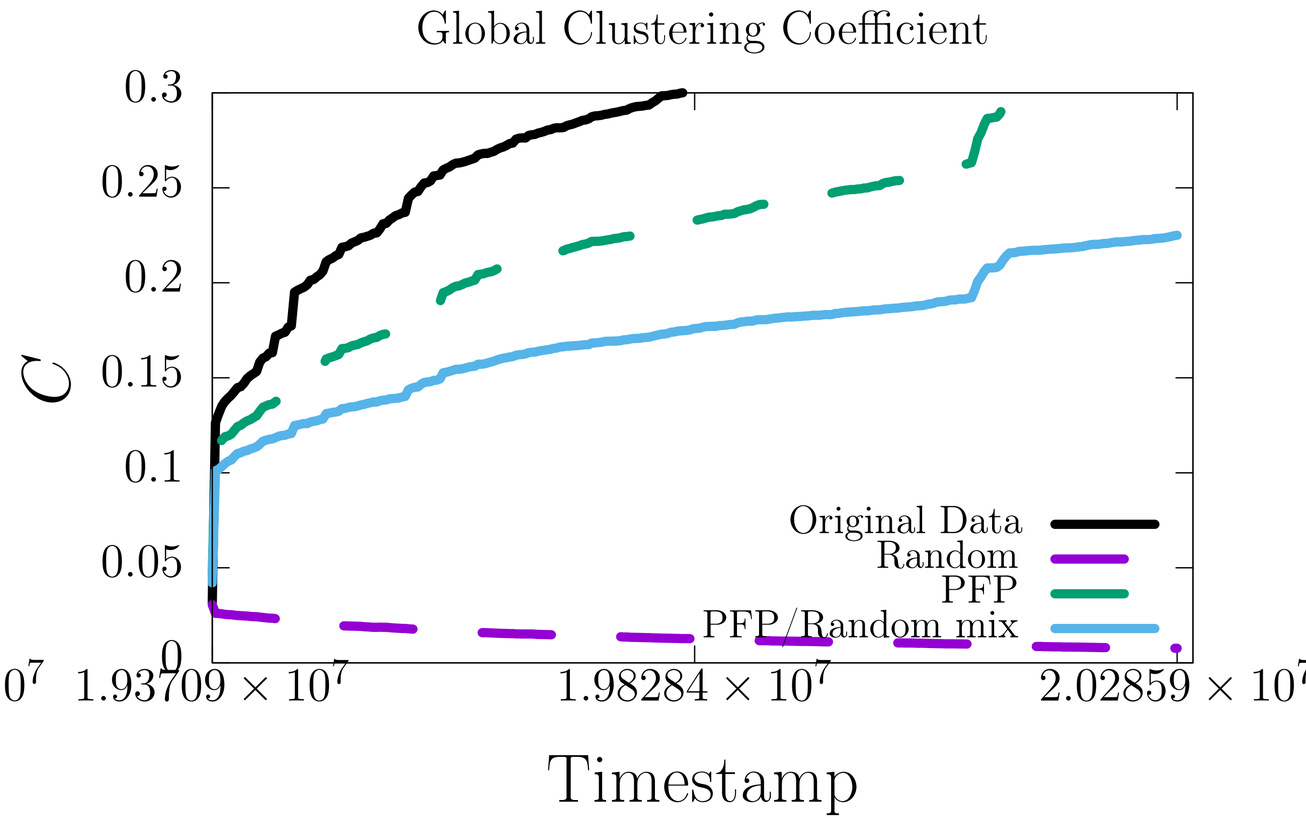
Real data example: Routeviews AS topology dataset





Real data example: Routeviews AS topology dataset





Remarks

- Often (**not always**) the highest likelihood mixture of model components may generate networks with better matching statistics than any single component alone

- But this highest likelihood mix is not guaranteed to be a good model - still need to look at network statistics

- Finding maximum likelihood mix of more than two model components may become expensive - candidate for ML techniques

Conclusions and future directions

- Temporal data allows calculation of likelihoods of dynamic models given observed network evolution
- With this measure we can distinguish models which generate networks that are similar in structure
- Fitting mixed models helps uncover the roles of different processes governing network growth
- Work in progress: using likelihood measure to analyse how such processes may change over time

Thank you for listening!
What are your questions?

github.com/narnolddd

n.a.arnold@qmul.ac.uk

@narnolddd