# Chapter 5: Growing Random Networks

Social and Economic Networks, Matthew O. Jackson

Jesse C. Chen

2023-02-11

# Roadmap

1. Uniform Randomness
   - Mean-Field Approximation
   - Continuous Time Approximation of Degree Distribution
2. Preferential Attachment
3. Hybrid Models
4. Other Aspects of a Growing Network
   - Diameter
   - Positive Assortativity and Degree Correlation
   - Clustering in Growing Random Networks
5. Summary

# Recap and Introduction

We considered "random graph-based models" last week, so why consider "growing networks"?

1. In many applications, it is more natural to consider *growing* networks.
   - web pages, scientific journals (citations)
   - people entering new environments (e.g., schools, workplaces, neighbourhoods, cities)
2. Growing models provide extra richness to the model that is not present in Poisson random networks.

   That is, with the **time dimension** now in consideration, we can model features such as *fat-tailedness* and *clustering* naturally.

# Uniform Randomness

# Notation

- Discrete time: $t \in \{0, 1, 2, ...\}$.
- One node is born at each time, and the node is named after its birth date.
- Let $d_i(t)$ denote the degree (undirected unless specified) of node $i$ at time $t$.

# Exponential Model (Uniform Randomness)

- At the birth of each node, it connects to $m$ other nodes randomly (uniformly).
- We want to derive the (asymptotic) *degree distribution* of this model setting.

# Degree Distribution of Exponential Model

- For the model to be well-defined, we assume that we are starting with a $m$-node clique. This assumption will not affect the degree distribution in the limit.
- Thus, the first new born node is $m + 1$.
- At time $t$, each node $i \in \{m + 1, ..., t\}$ has **expected** links

$$m + \frac{m}{i+1} + \cdots + \frac{m}{t} \approx m\left(1 + \ln(t) - \ln(i)\right)$$

- Given a degree $d$ at time $t$, nodes that have expected degree less than $d$ are those such that

$$m\left(1 + \ln(t) - \ln(i)\right) < d \implies i > t \exp\left(1 - \frac{d}{m}\right),$$

i.e., nodes that are born after $t \exp(1 - d/m)$.

# Degree Distribution of Exponential Model (cont'd)

- Hence, the fraction of nodes that have expected degree less than $d$ is

$$\frac{t \exp\left(1 - \frac{d}{m}\right)}{t} = \exp\left(1 - \frac{d}{m}\right)$$

- Therefore, the expected degree distribution is

$$F_t(d) = 1 - \exp\left(1 - \frac{d}{m}\right) \tag{1}$$

This is a variation of an exponential distribution.

## Note

- The distribution is time-independent. This is not true in general.
- We are using **expected** degree to obtain the distribution.

## Uniform Randomness

### Mean-Field Approximation

# Mean-Field Approximations

The technique we used (using expected degree) is called **mean-field approximation**.

- The burning question is: **Is this approximation good for “actual” degrees?**
- For simple models, such as the exponential model and the preferential attachment model, we know this approximation is correct.
- However, the answer is ¯_(*_*)_/¯ most of the time.
- The next-best thing to do is use simulation to check whether the approximation is good.

# Uniform Randomness

## Continuous Time Approximation of Degree Distribution

# Degree Distribution of Exponential Model (Conti. time)

- Now we consider a *continuous time view* to approximating the degree distribution.
- A node starts with $d_i(t = i) = m$; then, the node gains $m/t$ links in expectation at any time $t$ after $i$.
- This description yields an differential equation

$$\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = \frac{m}{t}.$$

- This differential equation has a solution

$$d_{i}(t) = m + m \ln\left(\frac{t}{i}\right)$$

- Note that $d_i(t)$ is decreasing in $i$ and increasing in $t$. This fits the intuition that older nodes have higher degrees.

# Degree Distribution of Exponential Model (Continued)

- Let $i_t(d)$ denote a node such that node $i_t(d)$ has degree $d$ at time $t$, i.e., $d_{i_t(d)}(t) = d$.
- Since $d_i(t)$ is decreasing in $i$, ...
  1. $i_t(d)$ is well-defined,
  2. only the nodes born after $i_t(d)$ have degrees less than $d$.
- In our case, $i_t(d)$ assumes the form

$$i_t(d) = t \exp\left(1 - \frac{d}{m}\right)$$

- Hence, the resulting degree distribution is

$$F_t(d) = 1 - \frac{i_t(d)}{t} = 1 - \exp\left(1 - \frac{d}{m}\right),$$

which is identical to what we have obtained earlier.

# Degree Distribution of Exponential Model (Continued)

## Remarks

Solving first order ODE's has several advantages:

1. It is often simpler than the direct method.
2. It can make model specification simpler, as we only have to specify (1) the initial condition of $d_i(t)$ at $t = i$ and (2) how $d_i(t)$ evolves through time.

- Continuous time approximation is not a big problem. It is relatively minor and it smooths things out.
- The main problem of approximating the degree distribution is still the discrepancy between “expected degrees” and “actual degrees”.

# Preferential Attachment

# Preferential Attachment

- **Motivation**: New comers do not form links with existing members *at random*. More likely, they form links with existing members that has a lot of connections already.
- The two main ingredients of this process is
  1. The system grows over time.
  2. The existing objects grow at rates proportional to their size.

  These two properties leads to **scale-free distributions**.
- Many big names studied this phenomenon:
  - Pareto (1896): wealth distribution (Pareto distribution)
  - Yule (1925): explain the distribution of city sizes
  - Zipf (1949): word frequency (Zipf's law)
  - Simon (1995): formalize processes that generates scale-free distributions
  - Price (1965): citation network

# Degree Distribution of Preferential Attachment

- Each newborn node still forms $m$ links, but not uniformly across existing nodes.
- Each new node links to a preexisting node with probabilities proportional to their degrees, i.e., an existing node $i$ is expected to get

$$m\frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)} = m\frac{d_i(t)}{2tm} = \frac{d_i(t)}{2t}$$

  links from the newborn node at time $t$.
- Thus, the mean-field, continuous-time approximation of this process is

$$\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = \frac{d_i(t)}{2t}$$

  with initial condition $d_i(t = i) = m$.

# Degree Distribution of Preferential Attachment (Cont'd)

- A solution to the ODE is

$$d_i(t) = m\left(\frac{t}{i}\right)^{1/2} \quad \rightsquigarrow \quad i_t(d) = t\left(\frac{m}{d}\right)^2 .$$

- Hence, the degree distribution is

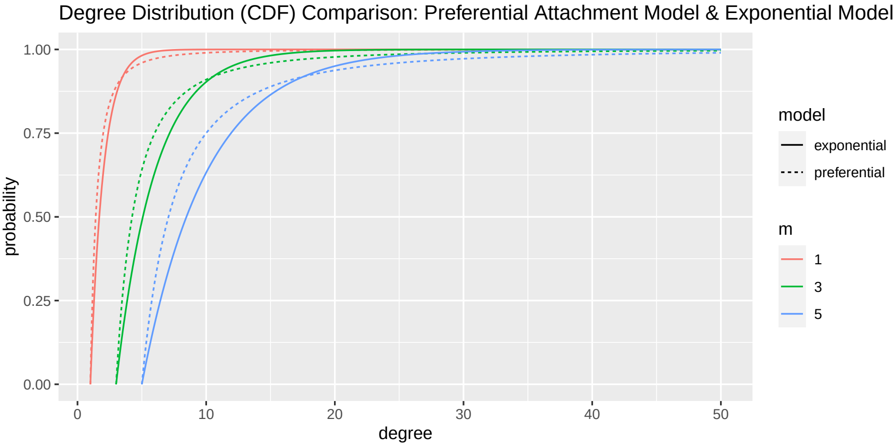
$$F_t(d) = 1 - \frac{i_t(d)}{t} = 1 - \left(\frac{m}{d}\right)^2$$

The corresponding density function is

$$f_t(d) = 2m^2 d^{-3}.$$

Thus, the preferential attachment process naturally motivates a **scale-free distribution** with exponent $3$. (scale-free distribution: $p(d) = cd^{-3}$)

# Comparison of Preferential Attachment and Exponential Model



# Motivating a Different Exponent

- The exponent $3$ comes from the fact that $\sum_{j=0}^{t} d_j(t) = 2tm$.
- Thus, if we specify the differential equation more generally thus

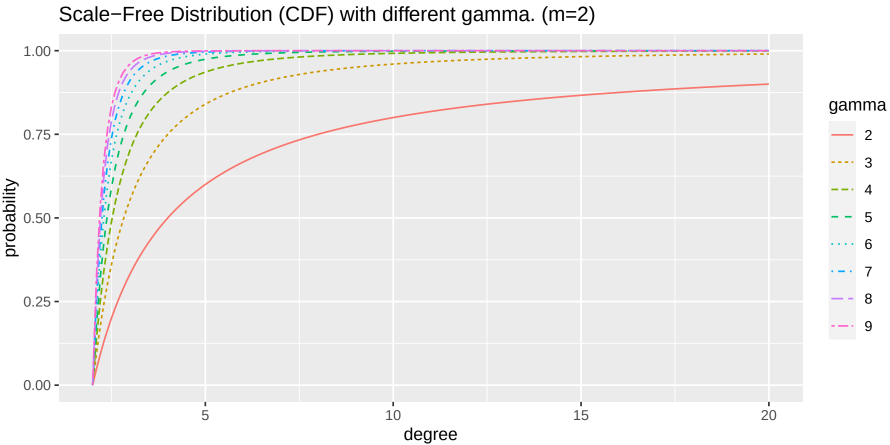
$$\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = \frac{d_i(t)}{\gamma t},$$

the corresponding degree distribution would be

$$f_t(d) = \gamma m^{\gamma} d^{-\gamma-1}.$$

- Intuitively, the growth of $d_i(t)$ is proportional to itself (the main feature of preferential attachment) and is scaled by a factor of $\gamma^{-1}$. The smaller $\gamma$ is, the faster $d_i(t)$ grows over time, which leads to a fatter tail.

# Motivating a Different Exponent (Cont'd)



# Motivating a Different Exponent (Cont'd)

- It is not very straight forward to justify/motivate a $\gamma$ that is not $2$.
- A possible motivation is thus:
    - At each time period, a group of nodes are born.
    - In the same period, $m$ new connects are created.
    - Of the $m$ connects, $\alpha m$ are made to existing nodes, and $(1-\alpha)m$ are made within the new nodes.
- In this setting, an existing node $i$ is expected to get

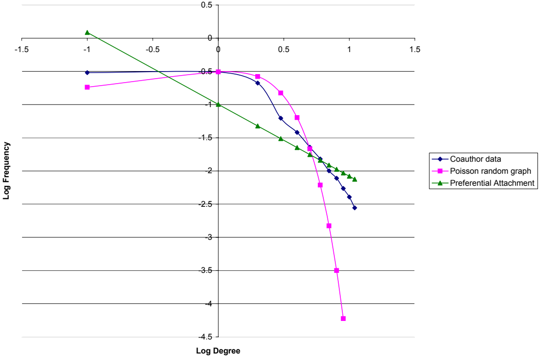
$$\alpha m \frac{d_i(t)}{2mt} = \alpha \frac{d_i(t)}{2t} = \frac{d_i(t)}{(2/\alpha)t}$$

new links. Here, $\gamma = 2/\alpha$.

# Hybrid Models

# Real Data

- Co-authorship data fits between a uniformly random network and a preferential attachment network. That is, the data has a fat tail, but less fat than preferential attachment network.



- Thus, we are motivated to build a hybrid model.

# Hybrid Model

- We can intuitively build a hybrid model:
- Each new node form $m$ connections, of which $m\alpha$ links randomly to existing nodes and of the rest $m(1-\alpha)$ links to existing node via preferential attachment:

$$\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = \alpha\frac{m}{t} + (1-\alpha)\frac{d_i(t)}{2t}$$

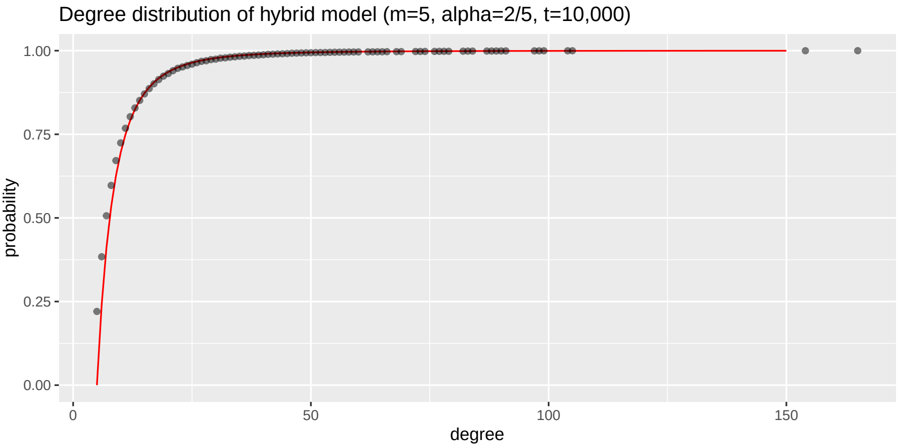
- The ODE has solution

$$d_{\color{red}i}({\color{blue}t}) = \left(m + \frac{2\alpha m}{1-\alpha}\right)\left(\frac{\color{blue}t}{\color{red}i}\right)^{(1-\alpha)/2} - \frac{2\alpha m}{1-\alpha}$$

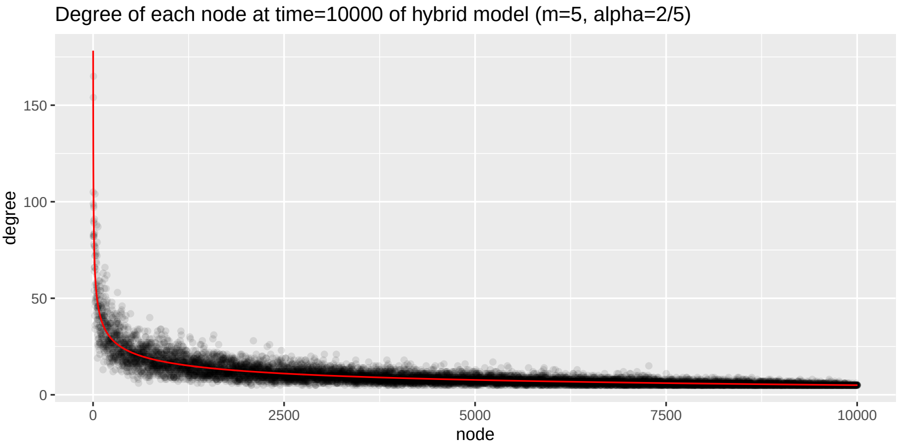
- The degree distribution is

$$F_t(d) = 1 - \frac{i_t(d)}{t} = 1 - \left(\frac{m + \frac{2\alpha m}{1-\alpha}}{d + \frac{2\alpha m}{1-\alpha}}\right)^{\frac{2}{1-\alpha}} \qquad (2)$$

# Hybrid Model (Simulation)



# Hybrid Model (Simulation, Cont'd)



# Fitting the Hybrid Model

- The parameter $\alpha$ is interesting, since it can be interpreted as the proportion that the network formation process is through uniform randomness/preferential attachment.
- Parameter $m$ can be observed directly by dividing the total degree by $2t$. Many methods can be used to estimate the CDF:

$$\ln(1 - F(d)) = \frac{2}{1-\alpha} \ln\left(m + \frac{2\alpha m}{1-\alpha}\right) - \frac{2}{1-\alpha} \ln\left(d + \frac{2\alpha m}{1-\alpha}\right)$$

- Fitting this model to co-authorship data listed on EconLit during 1990's yields an estimation of $\hat{\alpha} = 0.56$. (Goyal, van der Liej, Moraga-Gonzalez, 2006)

# Other Aspects of a Growing Network

Now we take a look at other aspects of a network.

- Diameter
- Positive Assortativity
- Clustering

# Other Aspects of a Growing Network

## Diameter

# Diameter

- In general, the diameter (or average path length) of a network is very difficult to calculate beyond Poisson random graph.
- Intuitively, a model with preferential attachment should have lower diameter (compared to a Poisson graph), since there high degree nodes serves as hubs.

**Theorem (Bollobás & Riordan, 2004)**

Consider a preferential attachment model where each newborn node forms $m \geq 2$ links. As $t$ increases, the resulting graph will consist of a single component with diameter proportional to $\frac{\log t}{\log \log t}$ almost surely.

- The diameter of a Poisson network is proportional to $\log t$ if the average degree is fixed.

## Other Aspects of a Growing Network

### Positive Assortativity and Degree Correlation

# Positive Assortativity and Degree Correlation

### Assortativity

a preference for a network's nodes to attach to others that are *(dis)similar* in some way.

- Many networks exhibits positive assortativity, a feature absent in Poisson graphs.

### Theorem (Jackson & Rogers, 2007b)

Consider a hybrid model. Under mean-field estimation, the estimated distribution of $i$'s neighbors' degree strictly first-order stochastically dominates that of $j$'s at each $t > j$ for all $j > i$. In particular, $F_i^t(d) < F_j^t(t)$ for all $d < d_i(t)$.

- This means, an older node has friends that have more connections.

# Other Aspects of a Growing Network

## Clustering in Growing Random Networks

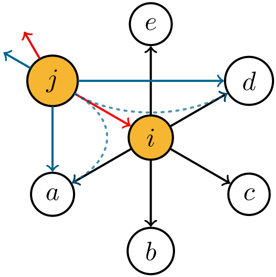
# Clustering

- Clustering is another idea that is observed in reality but not captured by Poisson graphs.
- Even in a exponential graph or hybrid models, clustering is not captured. That is, clustering converges to $0$ as $t \to \infty$.
- Intuitively, the probability of forming clusters is too low in exponential graphs. The probability of two new links creasting a cluster is

$$\frac{tm \text{ (all existing links)}}{\binom{t}{2} \text{ (all pairs of nodes)}} = \frac{2m}{t-1} \to 0 \quad \text{as} \quad t \to \infty.$$

- **Idea**: For a model to capture "clustering", the network formation process has to depend on "the existing graph structure" rather than only on the degree.

# A Meeting-Based Network Formation Model



$m_r = 2$

$m_n = 3$

$m_r + m_n = m = 5$

- A new node ($j$) is born. Let $m = m_r + m_n$.
- **Step 1**: Pick $m_r$ nodes randomly to link to. ($j$ picks $i$ in this step)
- **Step 2**: Randomly choose $m_n$ of the *out-links* of the $m_r$ nodes and link to the corresponding nodes, i.e., linking to "friends of friends." ($j$ picks $2$ of out-links of $i$ and links to corresponding node $a$ and node $d$)

- Thus, we have the following ODE characterizing the change in in-degree:

$$\frac{\mathrm{d}d_i^{\text{in}}(t)}{\mathrm{d}t} = \underbrace{\frac{m_r}{t}}_{\textbf{Step 1}} + \underbrace{\frac{m_r d_i^{\text{in}}(t)}{t}\frac{m_n}{m_r m}}_{\textbf{Step 2}} = \frac{m_r}{t} + \frac{m_n d_i^{\text{in}}(t)}{mt}$$

$$= \underbrace{\frac{m_r}{m}}_{\alpha}\frac{m}{t} + \underbrace{\frac{m_n}{m}}_{1-\alpha}\frac{d_i^{\text{in}}(t)}{t}$$

with initial condition $d_i^{\text{in}}(i) = 0$.

- The resulting in-degree distribution is

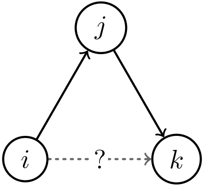
$$F(d^{\text{in}}) = 1 - \left(\frac{rm}{d^{\text{in}} + rm}\right)^{1+r}$$

where $r = m_r/m_n$. ◂ compare

# Discussion of Meeting-Based Network

1. The reason we consider a directed network here is for the tractability of **Step 2**.
   - In our case, $d_t^{\text{out}}(t) = m \; \forall t$ makes the calculation easy.
   - In an undirected network, it is hard to count the total degree of the $m_r$ chosen nodes.
2. A directed network with constant out-degree might be a good model for webpages or scientific articles, but not for human interactions.

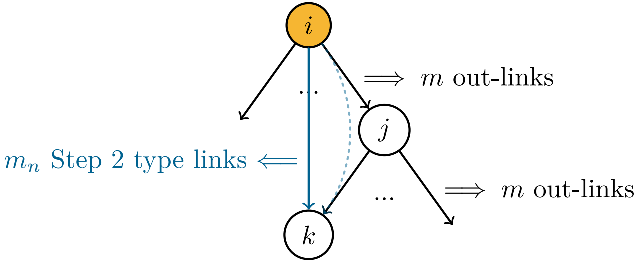
# Clustering



- Now we consider the property of interest: *clustering*.
- Recall that **transitive triple clustering measure** is

$$\mathrm{Cl}^{\mathrm{TT}}(g) = \frac{\sum_{i;j\neq i;k\neq j} g_{ij} g_{jk} g_{ik}}{\sum_{i;j\neq i;k\neq j} g_{ij} g_{jk}}$$

- Clearly, this model is designed in a way such that transitive triples are common. But how common exactly?



$i$
$j$
$k$
...
...
$\Longrightarrow$ $m$ out-links
$m_n$ Step 2 type links $\Longleftarrow$
$\Longrightarrow$ $m$ out-links

- The denominator of $\mathrm{CI}^{\mathsf{TT}}(g)$ is $tm^2$.
- The nominator of $\mathrm{CI}^{\mathsf{TT}}(g)$ is “at least” $tm_n$.
- Thus, the lower-bound of $\mathrm{CI}^{\mathsf{TT}}(g)$ is

$$\frac{tm_n}{tm^2} = \frac{m_n}{m^2} = \frac{1}{(1+r)m}$$

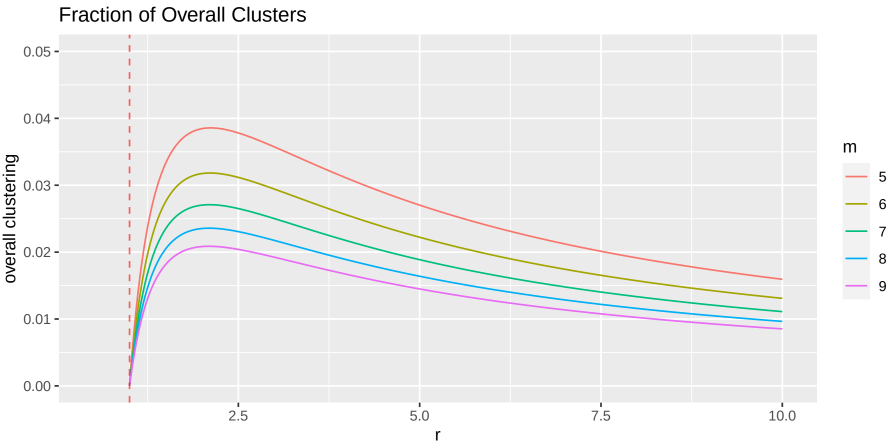
  where $r = m_r/m_n$.
- This lower-bound turn out to be the correct $\mathrm{CI}^{\mathsf{TT}}$ when $r \geq 1$.

- In order to simplify the calculation, consider a special process:
    - when $r \geq 1$, then at most one link is formed in each node found in **Step 1**.
    - when $r < 1$, then exactly $m_n/m_r$ are formed in each node found in **Step 1**.
- The parameter $r < 1$ means that more than half of the friends made by a new node are “friends of friends”. This means any link created in **Step 2** is very likely to generate more than one transitive triple.

### Proposition (Jackson & Rogers, 2007b)

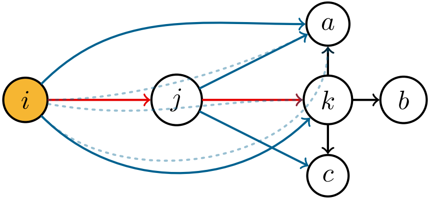
Under a mean-field approximation specified above, the fraction of transitive triples, $\mathrm{CI}^{\mathrm{TT}}$, tends to

$$\begin{cases} \frac{1}{(1+r)m} & \text{if } r \geq 1, \\ \frac{r(m-1)}{m(m-1)(1+r)r-m(1-r)} & \text{if } r < 1. \end{cases}$$



- There is actually a kink at $r = 1$.

# Example: Case $r < 1$



- Consider the case where $m_r = 1$ and $m_n = 2$.
- Previously, node $j$ connects to $k$ in **Step 1** and connects to $a$ and $c$ in **Step 2**.
- Now, a new node $i$ connects to $j$ in **Step 1**, then it connects to $a$ and $k$ in **Step 2**.
- In this case, $3$ transitive triples are generated instead of $2$.
  - $i \to j \to a \implies i \to a$.
  - $i \to j \to k \implies i \to k$.
  - $i \to k \to a \implies i \to a$. (This triple we did not expect)

# Example: Case $r < 1$ (Cont'd)

- To account for the unexpected triples, we can do the following calculation

$$\mathsf{Cl}^{\mathsf{TT}} = \frac{m - 1 + \binom{m-1}{2}\frac{\mathsf{Cl}^{\mathsf{TT}} m^2}{\binom{m}{2}}}{m^2} \implies \mathsf{Cl}^{\mathsf{TT}} = \frac{m-1}{2m}.$$

  where the $m_n = m - 1$ is the number of triples we originally expect and the $\mathsf{Cl}^{\mathsf{TT}} m^2 / \binom{m}{2}$ is the probability that any two neighbours of $j$ are already linked.

- The result of the previous proposition can be obtained similarly.

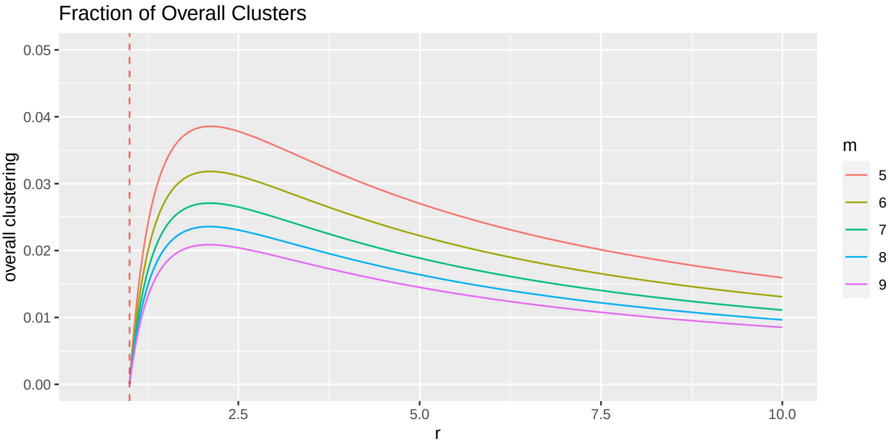
# Clustering, Undirected

- We can also measure overall clusters, i.e., ignoring the directions of links.
- It turns out the result is quite different:

## Proposition (Jackson & Rogers, 2007b)

Under a mean-field approximation, the overall clustering tends to

$$\begin{cases} 0 & \text{if } r \leq 1, \\ \frac{6(r-1)}{(1+r)(3(m-1)(r-1)+4mr)} & \text{if } r > 1. \end{cases}$$



- In the case where $r \leq 1$, nodes with very high degree starts to appear and dominates the calculation.
- So if we want overall clustering, we must have $r > 1$, but not too large.

◂ compare to transitive triples

# Discussion of Clustering

- The choice of which measure of clustering to use is crucial.
- The meeting-based model can explain clustering, there are also other reasons that clustering might emerge:

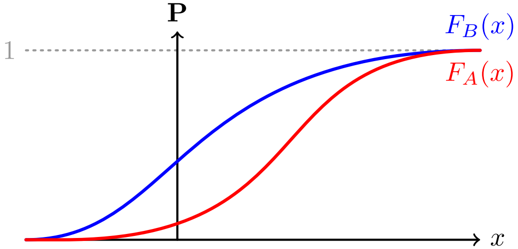
1. Common characteristics among nodes can lead to clustering, e.g., a low connection cost due to geographical reasons.
2. Specifying "active" and "inactive" nodes can lead to clustering (Klemm & Eguíluz, 2002):
   - When a new node is born, it is "active," while one other node turns "inactive." (with probability proportional to inverse degree)
   - A new born node first links to all active nodes, then with probability $\mu$, each link is rewired to a random node according to preferential attachment.

# Summary

# Summary

- We consider growing networks for two main reasons:
  1. It is very natural.
  2. It motivates many properties of real networks.
- **Exponential model** is the natural extension of the Poisson model.
- **Preferential attachment model** motives the scale-free distribution.
- We can model a mix of "uniformness" and "fat-tailedness" with a **hybrid model**.
- Other characteristics such as diameter, assortativity, and clustering can also be motivated by growing networks.
- The main technical challenge is that even the simplest properties of these networks become increasingly difficult to calculate. **Mean-field approximation** and **continuous time approximation** are our friends.

# First-Order Stochastic Dominance



- Let $A, B$ be two random variables with CDF $F_A$ and $F_B$ respectively.
- $A$ is said to **first-order stochastically dominate** $B$ if $F_A(x) \leq F_B(x)\ \forall x$.
- It "dominates" in the sense that.

$$F_A(x) \leq F_B(x) \iff \mathbf{P}\left\{A > x\right\} \geq \mathbf{P}\left\{B > x\right\}$$

◂ back