Social and Economic Networks

Growing Random Networks

Xiang Sun

October 18, 2017

Outline

1. Dynamic variation of Poisson random network
   - Discrete model
   - Continuous model
2. Power law distribution
   - History of power laws
3. Preferential attachment
   - Preferential attachment degree distribution
   - Barabasi–Albert model
4. Hybrid models

Growing random networks

- So far, we have focused on static random graph models in which edges among fixed n nodes are formed via random rules in a static manner.
  - Poisson random network has small distances, but low clustering and a rapidly falling degree distribution.
  - Small-world model provides a tractable model that has small distances and high clustering.
  - Configuration model generates arbitrary degree distributions.
- Most networks form dynamically whereby new nodes are born over time and form attachments to existing nodes when they are born.

Growing random networks (Cont.)

- Most networks form dynamically whereby new nodes are born over time and form attachments to existing nodes when they are born.
- Example: Consider the creation of web pages.
  - When a new web page is designed, it includes links to existing web pages.
  - Over time, an existing page will be linked to by new web pages.
- The same phenomenon is true in many other networks:
  - Networks of friendships, citations, professional relationships.
- Evolution over time introduces a natural heterogeneity to nodes based on their age in a growing network.
- These considerations motivate dynamic or generative models of networks.

Growing random networks (Cont.)

- When nodes are born we can consider different ways in which they attach to existing nodes.
- At one extreme, where newborn nodes pick nodes to link to uniformly at random.
  - We will just have a growing variation on a Poisson random network.
- At the other extreme, where they pick nodes in proportion to the current degrees of the existing nodes.
  - The older nodes have more chances to grow in degree and grow faster than younger nodes who have lower degrees.

Section 1

Dynamic variation of Poisson random network

Subsection 1

Discrete model

Discrete model

- As nodes are born over time, index them by the order of their birth.
- Node i is born at date i, where $i = 0, 1, 2, \ldots$.
- A node forms links to existing nodes when the new node is born.
- Let $d_i(t)$ be the degree of node i (born at time i) at a time t.

Discrete model

Consider a special case:

- Start the network with $m + 1$ nodes born at times $\{0, 1, \ldots, m\}$, each connected to each other.
- The first newborn node that we consider is the one born at time $m + 1$.
- Each newborn node randomly selects m of the existing nodes and links to them.
- The specifics of this will not be of great consequence when we look at limiting properties of the system, but it is helpful in order to be able to properly analyze the system.

Discrete model

- At the end of time $m+1$,
  - the newest node will have m links,
  - m of the $m+1$ pre-existing nodes will have new links and 1 of them will not.
  - Each of the pre-existing nodes expects to gain $\frac{m}{m+1}$ links.
- At the end of time $m+2$,
  - the newest node will have m links,
  - m of the $m+2$ pre-existing nodes will have new links and 2 of them will not.
  - Each of the pre-existing nodes expects to gain $\frac{m}{m+2}$ links.
- And so on.
- Probability: No longer binomial, as probabilities vary with time.

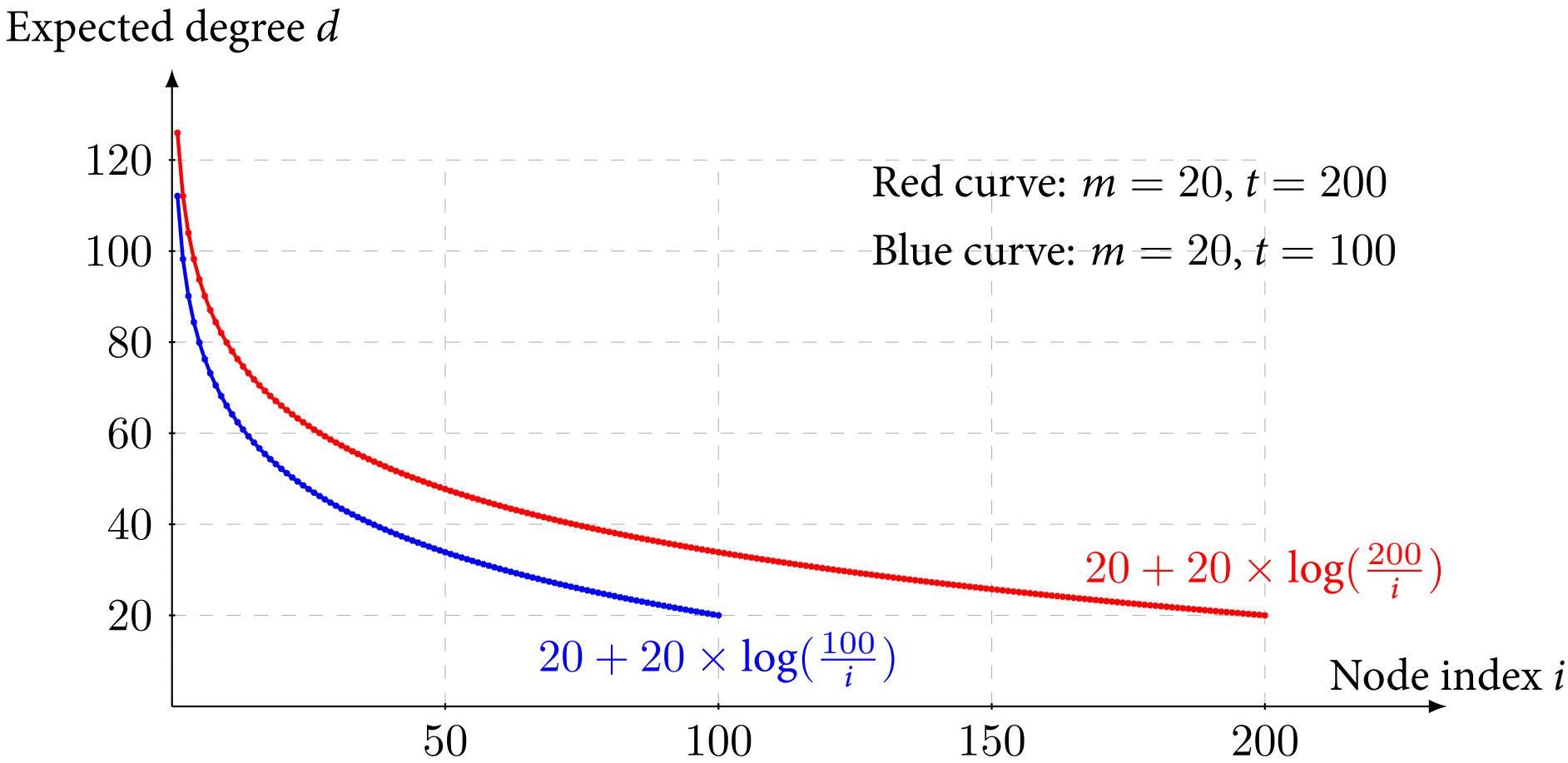
Realized network

- Depending on which 2 do not gain a link we have different possibilities for degree distributions that could be realized.
- As we continue, the number of possible realizations of the degree distribution grows.
- While it is hard to keep track of the potential realizations and their relative probabilities, we can do some more direct calculations.

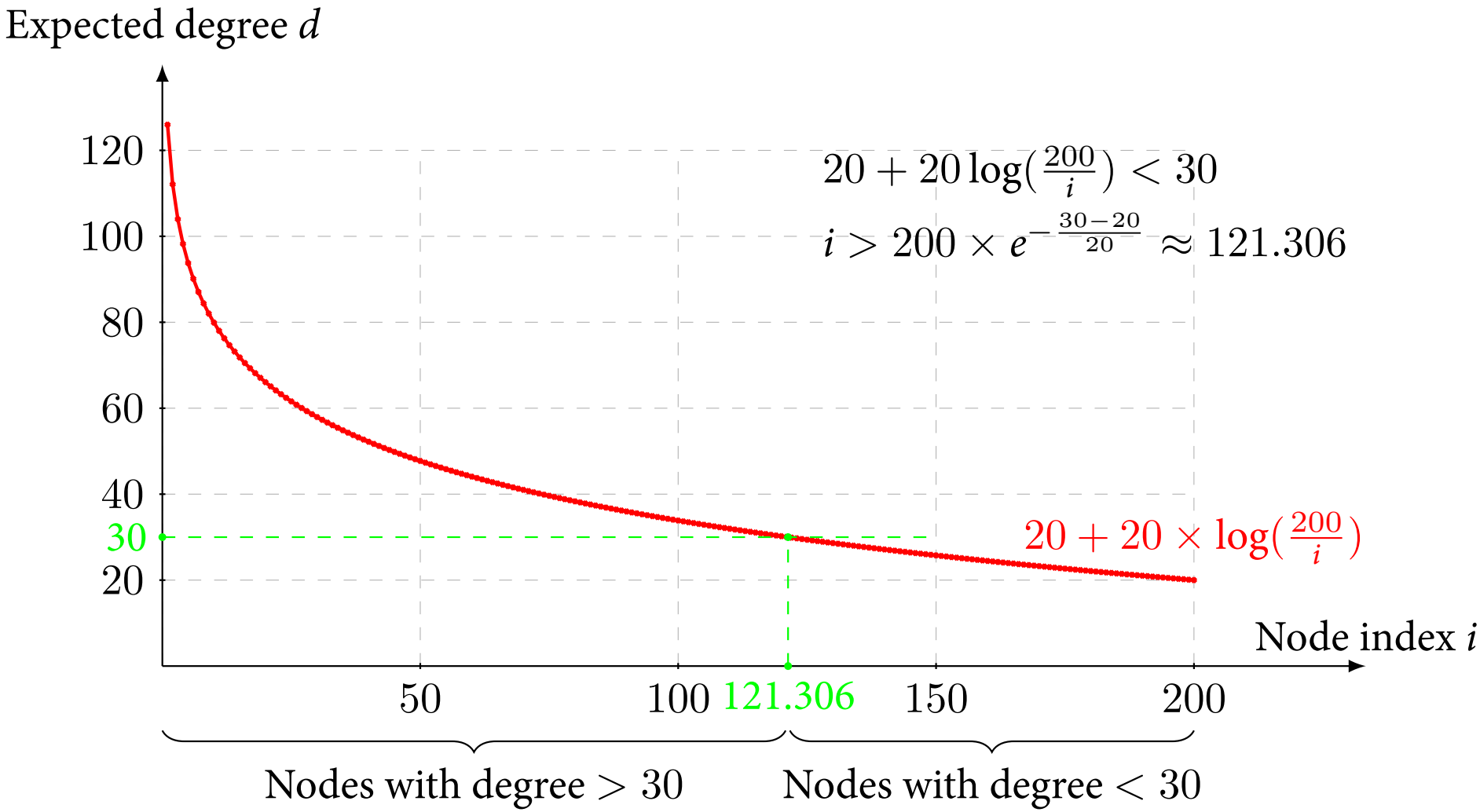
Discrete model: Expected degree

- For $m \leq i < t$, a node i born at time i has an expected degree at time t of
$$m + \frac{m}{i+1} + \frac{m}{i+2} + \cdots + \frac{m}{t}.$$
- For a large t, this is approximately
$$m + m \log \frac{t}{i}.$$
- For a large t, the nodes that have expected degree less than d are (using the approximation) those such that
$$m + m \log \frac{t}{i} < d.$$
- We require that $d < m + m \log \frac{t}{m}$—the largest possible degree.

Discrete model: Expected degree (Cont.)



Discrete model: Expected degree (Cont.)

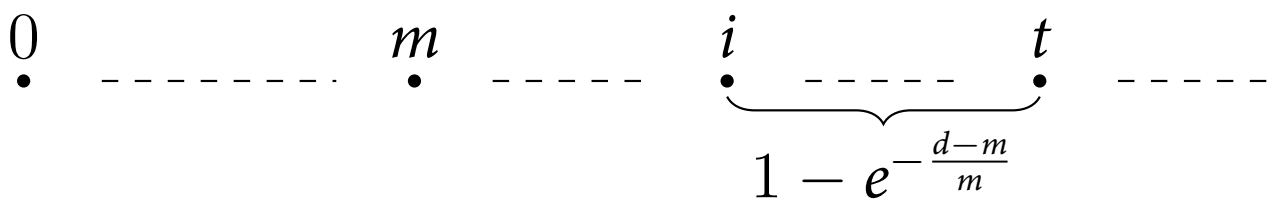


Discrete model: Expected degree distribution

- We rewrite this as the nodes i such that

$$i > te^{-\frac{d-m}{m}}.$$

- Thus, the nodes with expected degree less than d are those born after time $te^{-\frac{d-m}{m}}$.



- The fraction of nodes with expected degrees less than d is

$$F_t(d) = 1 - e^{-\frac{d-m}{m}}.$$

Discrete model: Expected degree distribution (Cont.)

- This is a variation of an exponential distribution.
- The faction of nodes with no more than some degree d is actually the same over time.
- The distribution is in fact independent of time t.

Expected vs. actual degree distribution

- The expected degree distribution is a good approximation for the actual degree distribution in this particular model.
- Issues:
  - The differences between the first m nodes and other nodes.
  - The rounding the sum of the harmonic series.
- The fraction of nodes whose ratios of realized to expected degrees are off by more than a given amount is going to 0.

Dynamic vs. static Poisson random networks

- Each node starts with a given number m of links. Then it is only the additional links that are random.

$\Rightarrow$ An appropriate benchmark random network would not be the Poisson random network, but instead a variation where there are t nodes, and each picks m others at random to link to.

$\Rightarrow$ There each node would approximately have a degree of m plus a Poisson random variable with expectation m.

- The exponential distribution has more of a spread to it:

  The older nodes tend to have higher degrees and the younger nodes have lower degrees.

Subsection 2

Continuous model

Continuous model

- A new node is born at time t. It forms m links by uniformly randomly picking m out of the t existing nodes.
- Node i's degree is thus described by
  - a starting condition of $d_i(i) = m$,
  - an approximate change over time of

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

for each $t > i$.

Continuous model (Cont.)

- This differential equation has a solution

$$d_i(t) = m + m \log \tfrac{t}{i}.$$

- If a node born at $i = i(d)$ has degree of exactly d, then

nodes whose expected degree $\leq d =$ nodes born on or after i.

- For any d and time t, we find the node $i(d)$ such that $d_{i(d)}(t) = d$.
- We solve for $i(d)$ such that

$$d = m + m \log \tfrac{t}{i(d)} \Rightarrow \tfrac{i(d)}{t} = e^{-\frac{d-m}{m}}.$$

Continuous model (Cont.)

- We have $\frac{i(d)}{t} = e^{-\frac{d-m}{m}}$.
- The nodes that have degree of less than d are then those born after $i(d)$.
- The fraction of nodes with expected degrees less than d is

$$F_t(d) = 1 - \frac{i(d)}{t} = 1 - e^{-\frac{d-m}{m}}.$$

- This is a negative exponential distribution with support from m to infinity and a mean degree of $2m$.

Section 2

Power law distribution

Power law distribution

- In studies over many different Web snapshots taken at different points in time, it has been observed that the degree distribution obeys a power law distribution.
  - The fraction of web pages with k in-links (or out-links) is approximately proportional to $k^{-2.1}$ (or $k^{-2.7}$).
- Many social and biological phenomena also governed by power laws.
  - Population sizes of cities observed to follow a power law distribution.
  - Number of copies of a gene in a genome follows a power law distribution.

Power law distribution (Cont.)

- A nonnegative random variable X is said to have a power law distribution if

$$\text{Prob}(X \geq x) \sim cx^{-\alpha},$$

for constants $c > 0$ and $\alpha > 0$. Here $f(x) \sim g(x)$ represents $\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1$.

- Roughly speaking, in a power law distribution, asymptotically, the tails fall of polynomially with power α.

Power law distribution (Cont.)

- Such a distribution leads to much heavier tails than other common models, such as normal and exponential distributions.
- One specific commonly used power law distribution is the Pareto distribution, which satisfies

$$\text{Prob}(X \geq x) = \left(\frac{x}{t}\right)^{-\alpha},$$

  for some $\alpha > 0$ and $t > 0$.
- The Pareto distribution requires $X \geq t$.

Subsection 1

History of power laws

History of power laws

- The earliest apparent reference is to the work by Pareto in 1897, who introduced the Pareto distribution to describe income distributions.
  - When studying wealth distributions, Pareto observed power law features, where there were many more individuals who had large amounts of wealth than would appear in normal or other distributions.
- Power laws also appeared in the work of Zipf in 1916, in describing word frequencies in documents and city sizes.
  - The empirical principle, known as Zipf's Law, states that the frequency of the j-th most common word in English (or other common languages) is proportional to j^{-1}.

History of power laws (Cont.)

- These ideas were further developed in the work of Simon in 1955, who showed that power laws arise when "the rich get richer", when the amount you get goes up with the amount you already have.
  - A city grows in proportion to its current size as a result of people having children.
  - Gene copies arise in large part due mutational events in which a random segment of the DNA is accidentally duplicated (a gene which already has many copies more likely to be in a random stretch of DNA).
- All of these examples exhibit rich get richer effects.

History of power laws (Cont.)

- Rich get richer effects quite fragile, there is great sensitivity to unpredictable initial fluctuations.
  - Empirically studied by Salganik, Dodds and Watts (2006): They created a music download site with 48 obscure songs. A visitor to the site can listen to the songs and also is shown the "current" download count for each song.
  - Each visitor at random is assigned to 8 "parallel copies" of the site, which started out identically.
  - Market share of different songs varied considerably across different copies.

History of power laws (Cont.)

- In 1965, Price applied these ideas to networks, with a particular focus on citation networks.
- Price studied the network of citations between scientific papers and found that the in degrees (number of times a paper has been cited) have power law distributions.
- His idea was that an article would gain citations over time in a manner proportional to the number of citations the paper already had.

History of power laws (Cont.)

- This is consistent with the idea that researchers find some article (e.g. via searching for keywords on the Internet) and then search for additional papers by tracing through the references of the first article.
- The more citations an article has, the higher the likelihood that it will be found and cited again.
- Price called this dynamic link formation process cumulative advantage.
- Today it is known under the name preferential attachment (偏好连结) after the influential work of Barabasi and Albert in 1999.

Section 3

Preferential attachment

Preferential attachment model

- Nodes are born over time and indexed by their date of birth.
- Assume that the system starts with a group of $m + 1$ nodes all connected to one another.
- Each node upon birth forms m edges with pre-existing nodes.

Preferential attachment model (Cont.)

- Instead of selecting m nodes uniformly at random, it attaches to nodes with probabilities proportional to their degrees.
  - For example, if an existing node has 3 times as many links as some other existing node, then it is 3 times as likely to be linked to by the newborn node.
- Thus, the probability that an existing node i receives a new link to the newborn node at time t is m times i's degree relative to the overall degree of all existing nodes at time t:

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

Preferential attachment model (Cont.)

- Since there are roughly tm total links at time t in the system (if t is large), it follows that

$$\sum_{j=1}^{t} d_j(t) = 2tm.$$

- Therefore, the probability that node i gets a new link in time t is

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

Preferential attachment model: Degree

- Hence, we can write down the evolution of expected degrees in continuous time as

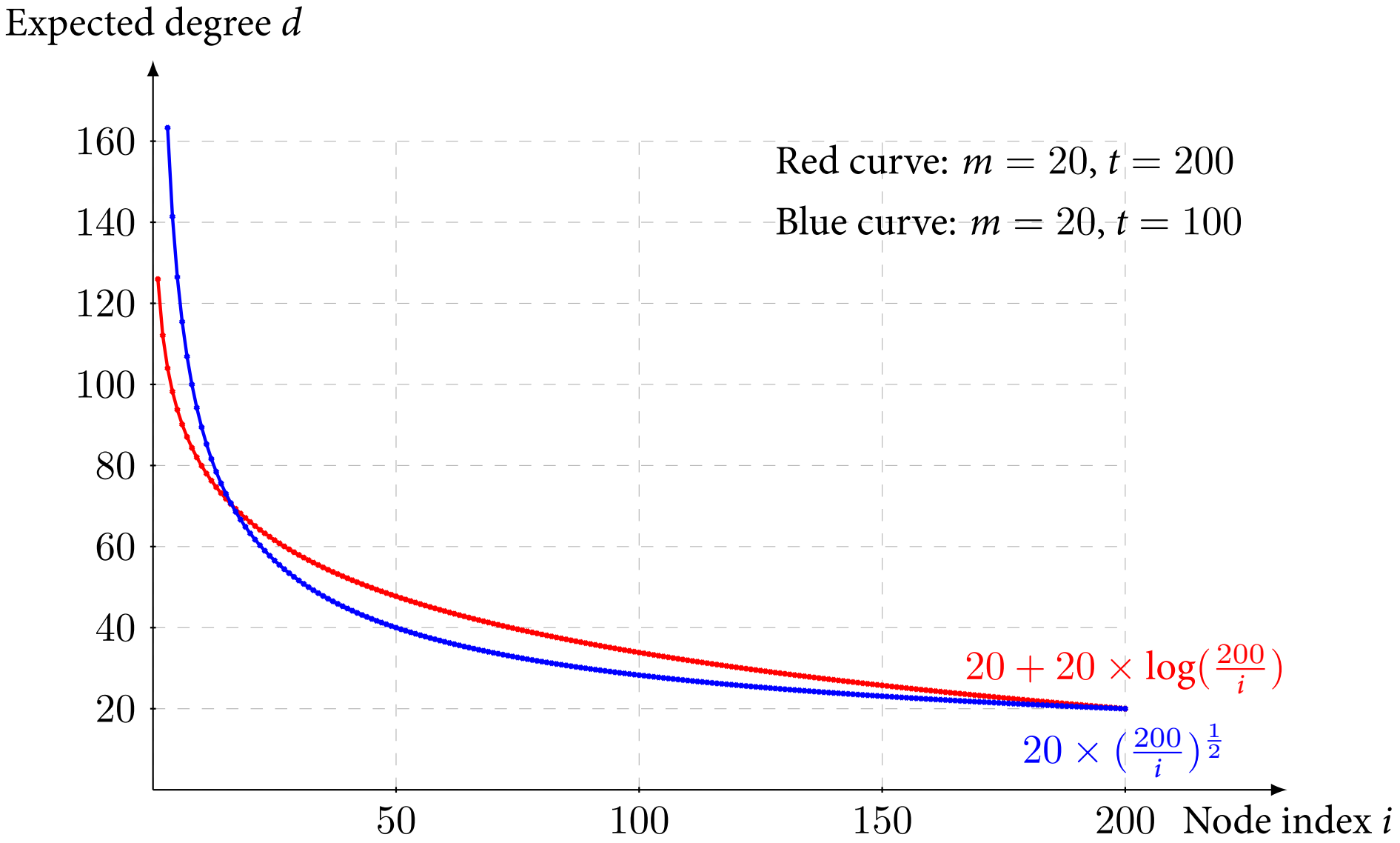
$$\frac{\mathrm{d}}{\mathrm{d}t} d_i(t) = \frac{d_i(t)}{2t},$$

with initial condition $d_i(m) = m$ (assuming degree is a continuous variable).

- This equation has a solution

$$d_i(t) = m\left(\frac{t}{i}\right)^{\frac{1}{2}}.$$

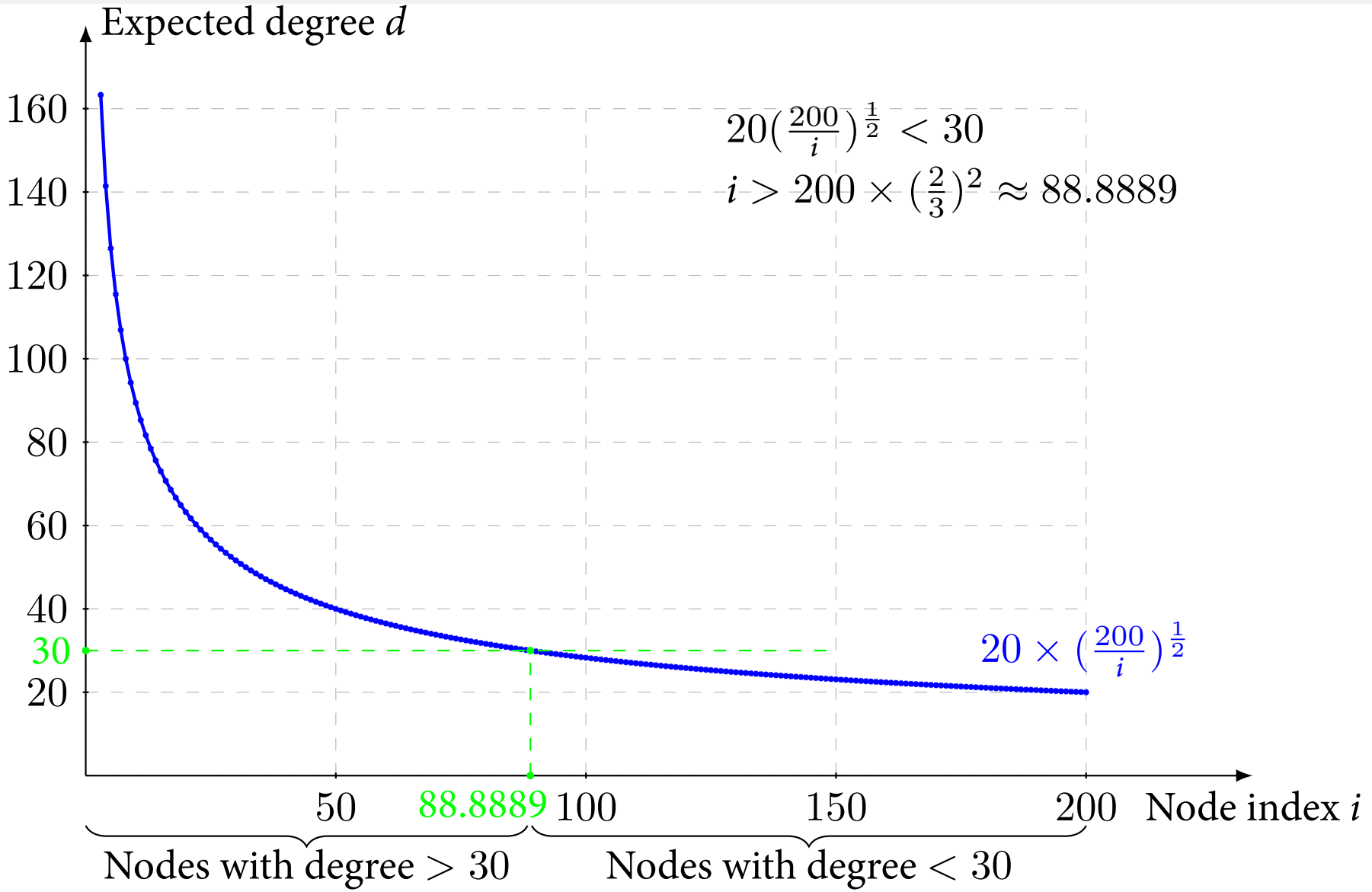
Preferential attachment vs. uniform at random



Subsection 1

Preferential attachment degree distribution

Preferential attachment degree distribution



Preferential attachment degree distribution

- As before, expected degrees of nodes are increasing over time.
- Hence, to find the fraction of nodes with degrees below a certain level d at time t, we need to identify which node is exactly at level d at time t.
- Let $i(d)$ be the node that has degree d at time t, or $d_{i(d)}(t) = d$.

Preferential attachment degree distribution (Cont.)

- From the degree expression, this yields

$$\frac{i(d)}{t} = \left(\frac{m}{d}\right)^2.$$

- The distribution function:

$$F(d) = 1 - \frac{i(d)}{t} = 1 - \left(\frac{m}{d}\right)^2.$$

- The density function:

$$\text{Prob}(d) = f(d) = 2m^2 d^{-3}.$$

- Networks generated by preferential attachment look very different from earlier models with similar average degree.

Preferential attachment degree distribution (Cont.)

- The density function:

$$\text{Prob}(d) = f(d) = 2m^2 d^{-3}.$$

- Log-log: $\log(\text{Prob}(d)) = \log(2m^2) - 3\log(d)$.

Subsection 2

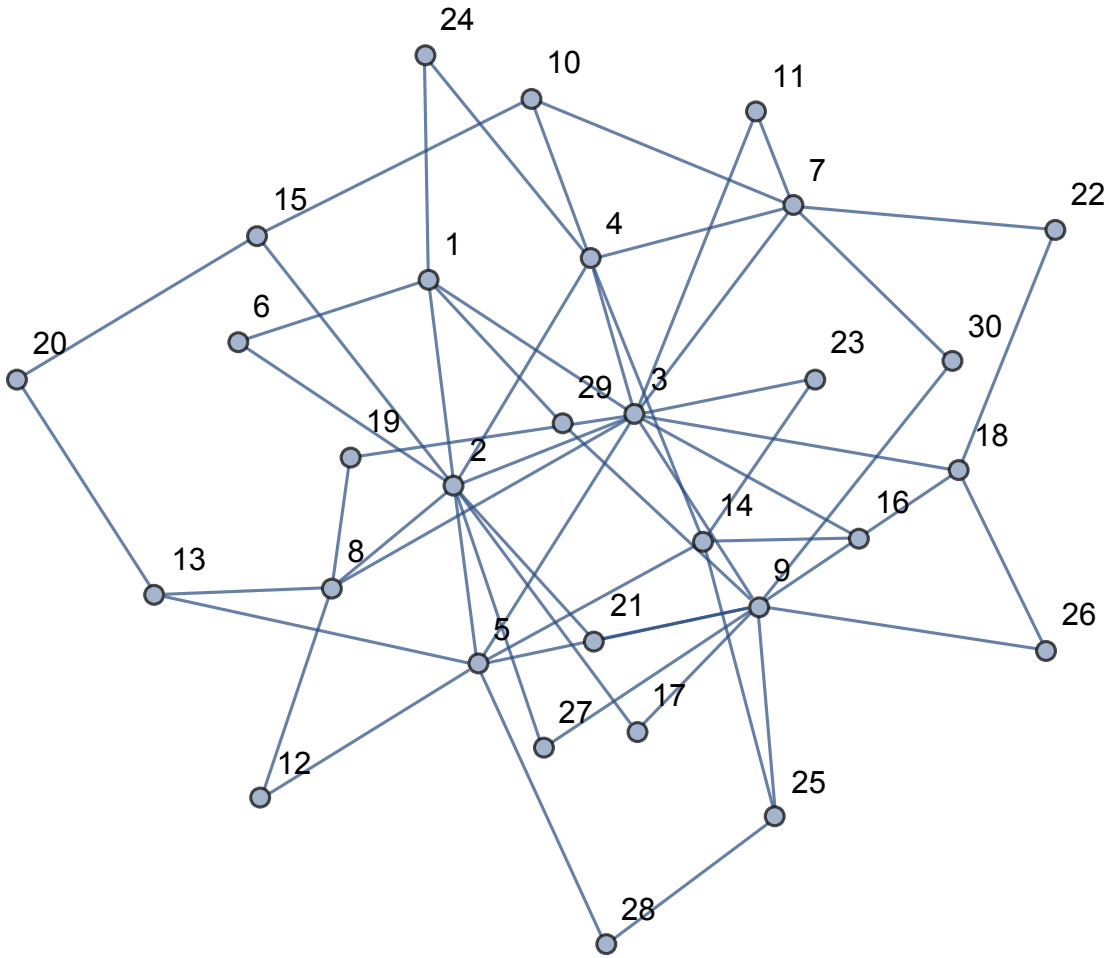
Barabasi–Albert model

Barabasi–Albert model

- The Barabasi-Albert model is an algorithm for generating random scale-free networks using a preferential attachment mechanism.
- `BarabasiAlbertGraphDistribution[n,m]` in Mathematica.
- A Barabasi–Albert graph distribution for n-vertex graphs where a new vertex with m edges is added at each step.

Barabasi–Albert model: Example

```
RandomGraph[BarabasiAlbertGraphDistribution[30, 2]]
```



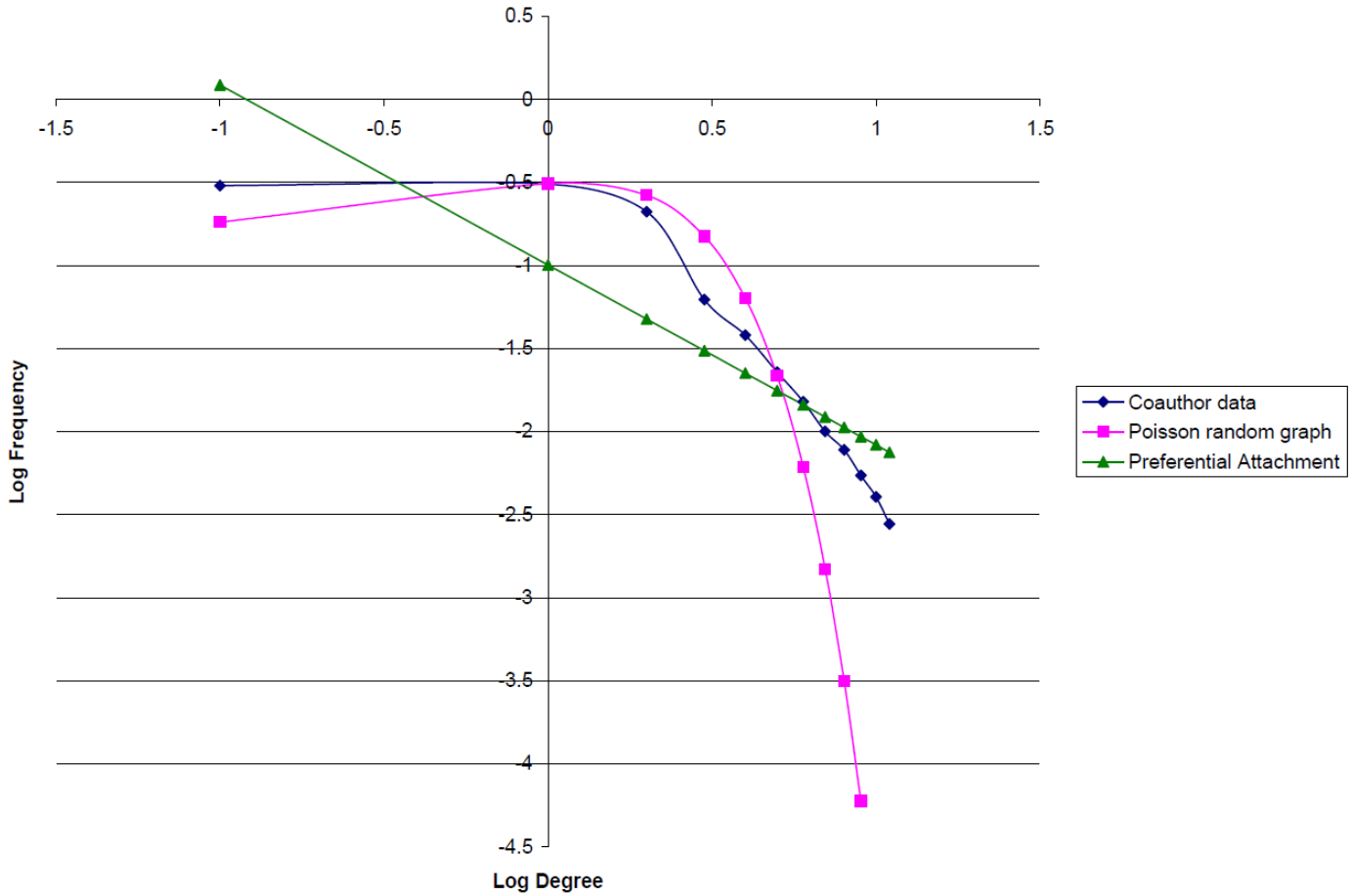
Section 4

Hybrid models

Hybrid models

- Many observed degree distributions match neither the exponential process nor the preferential attachment process.
- For example, consider the following degree distribution from the co-authorship network.

Hybrid models (Cont.)



Hybrid models (Cont.)

- Here we see a degree distribution that lies somewhere between the two extremes of uniformly random link formation and preferential attachment.
- This suggests that a more general network formation model is needed to match observed degree distributions.

Simple hybrid model

- A newborn node meets existing nodes via two different processes, where we combine the formation of links uniformly at random with preferential attachment.
- Each newborn node forms m links, with a fraction of $\alpha < 1$ of them formed to existing nodes selected uniformly at random, and a fraction $1 - \alpha$ of them formed to existing nodes via preferential attachment.

Simple hybrid model (Cont.)

- The change in the degree of a node over time can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t} d_i(t) = \alpha \frac{m}{t} + (1-\alpha) m \frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)} = \alpha \frac{m}{t} + (1-\alpha) \frac{d_i(t)}{2t}.$$

  - The first expression representing the chance of receiving one of the αm links being formed by picking uniformly at random from the t existing nodes.
  - The second expression has $(1-\alpha)m$ links being formed via preferential attachment and node i having a probability of $\frac{d_i(t)}{2mt}$ of receiving any one of them.

Simple hybrid model: Degree

- Solution:

$$d_i(t) = \left(d_0 + \frac{2\alpha m}{1-\alpha}\right)\left(\frac{t}{i}\right)^{\frac{1-\alpha}{2}} - \frac{2\alpha m}{1-\alpha},$$

where d_0 is the initial number of links that a node has when it is born.

- Setting $d_0 = m$, we have

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

Simple hybrid model: Degree (Cont.)

- When $\alpha = 0$, this is the degree distribution $1 - (\frac{m}{d})^2$, which is the power distribution that we found in the case of pure preferential attachment.
- When $\alpha \to 1$, the limit is harder to see directly, but it approaches the exponential distribution of $F(d) = 1 - e^{-\frac{d-m}{m}}$ for the model where links were formed uniformly at random.

Homework 3

- Suppose that newborn nodes come in groups of n in each period. Suppose that they attach a fraction f of their links uniformly at random to other newborn nodes, and a fraction $1-f$ to older nodes via preferential attachment. Using a continuous time approximation, develop an expression for the degree distribution.
- Answer: $F_t(d) = 1 - \left(\frac{m(1+f)}{d}\right)^{\frac{2}{1-f}}$.