Preferential Attachment

Introduction to Network Science

Carlos Castillo

Topic 05

Universitat Pompeu Fabra Barcelona

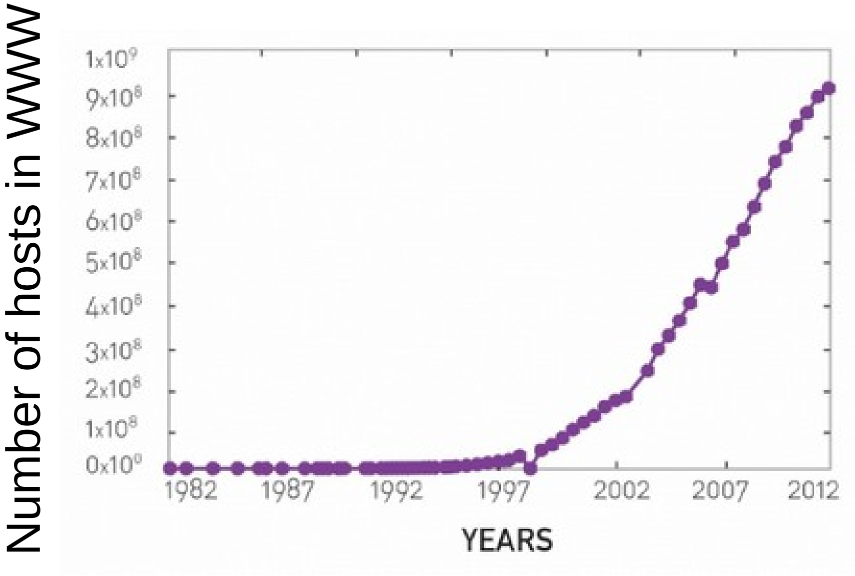
Contents

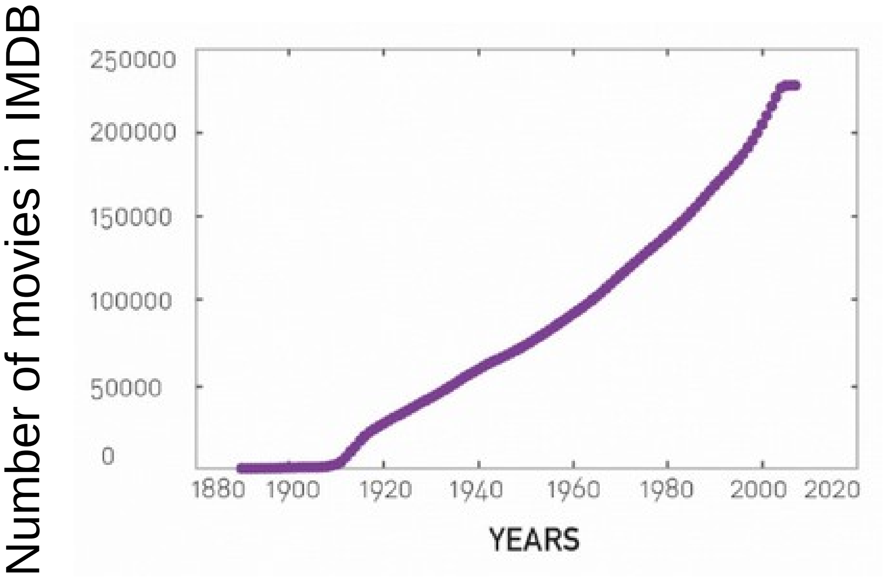
- The uniform random attachment model
- The BA or preferential attachment model
- Degree distribution under the BA model
- Distance distribution under the BA model
- Clustering coefficient under the BA model

Sources

- Albert László Barabási (2016) Network Science
  - Preferential attachment follows chapter 05
- Ravi Srinivasan 2013 Complex Networks Ch 12
- Networks, Crowds, and Markets Ch 18
- Data-Driven Social Analytics course by Vicenç Gómez and Andreas Kaltenbrunner

The number of nodes N increases: we need models of network growth





Preliminary:
Uniform Random Attachment

Growth in an ER network

- Two assumptions in ER networks:
  - There are N nodes that **pre-exist**
  - Nodes connect **at random**
- Let’s challenge the first assumption

Uniform Attachment

- Network starts with m fully-connected nodes
- Time starts at $t_0=m$
- At every time step we add 1 node
- This node will have m outlinks

Expected degree over time

- Probability of obtaining one link: m/t
  - Decreases over time
- Expected degree of node born at $m < i < t$

$$m + \frac{m}{i+1} + \frac{m}{i+2} + \frac{m}{i+3} + \cdots + \frac{m}{t} \approx m\left(1 + \log\left(\frac{t}{i}\right)\right)$$

Compute expected degree over time using a differential equation

$$\frac{dk_i(t)}{dt} = \frac{m}{t}$$

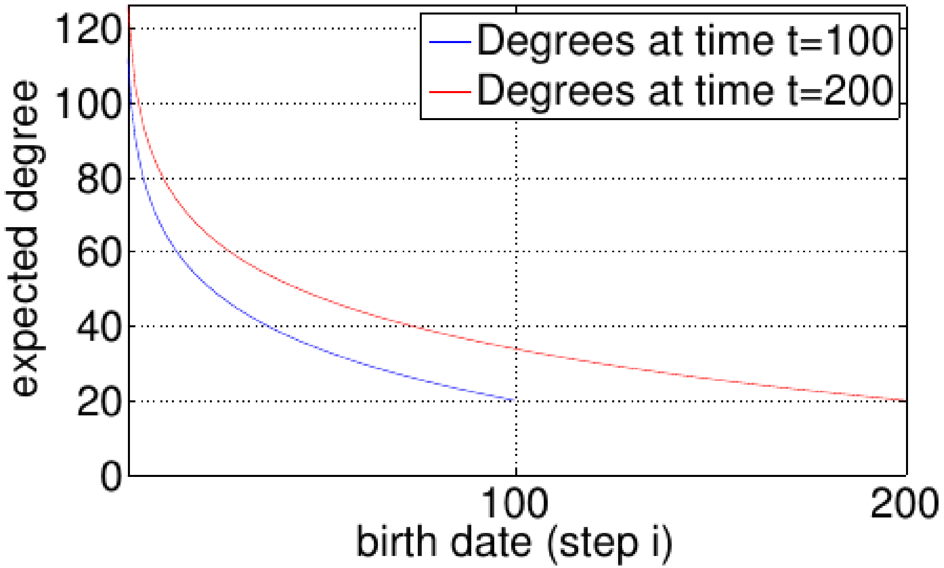
(1) Integrate between time i and time t

(2) Use initial condition $k_i(i) = m$

$$\int \frac{1}{t} = \log t + C$$

Degree distribution over time is not static

Degree of node born at time $m < i < t = m\left(1 + \log\left(\frac{t}{i}\right)\right)$



Tail of degree distribution

How many nodes of degree smaller than k are there at time t?

The fraction is $\dfrac{te^{-\frac{k-m}{m}}}{t} = e^{-\frac{k-m}{m}}$

Decreases exponentially with k: it's vanishingly rare to find high-degree nodes

$$m\left(1+\log\left(\frac{t}{i}\right)\right) < k$$

$$1+\log\left(\frac{t}{i}\right) < \frac{k}{m}$$

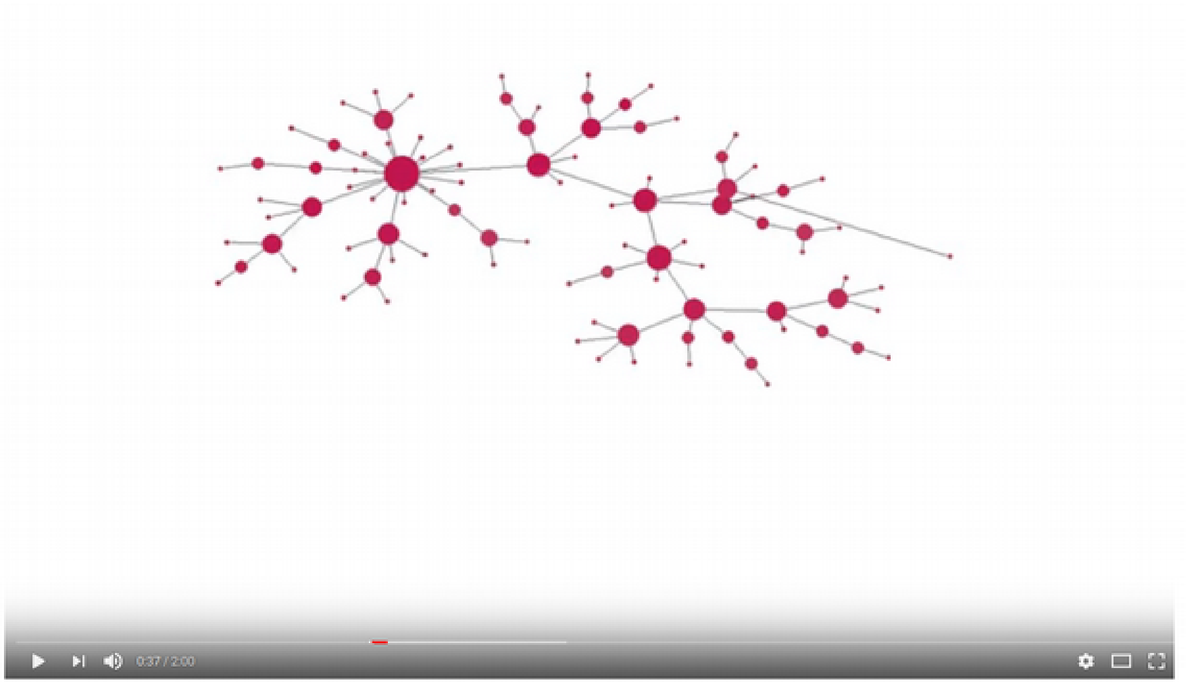
$$\log\left(\frac{t}{i}\right) < \frac{k-m}{m}$$

$$\frac{t}{i} < e^{\frac{k-m}{m}}$$

$$i > te^{-\frac{k-m}{m}}$$

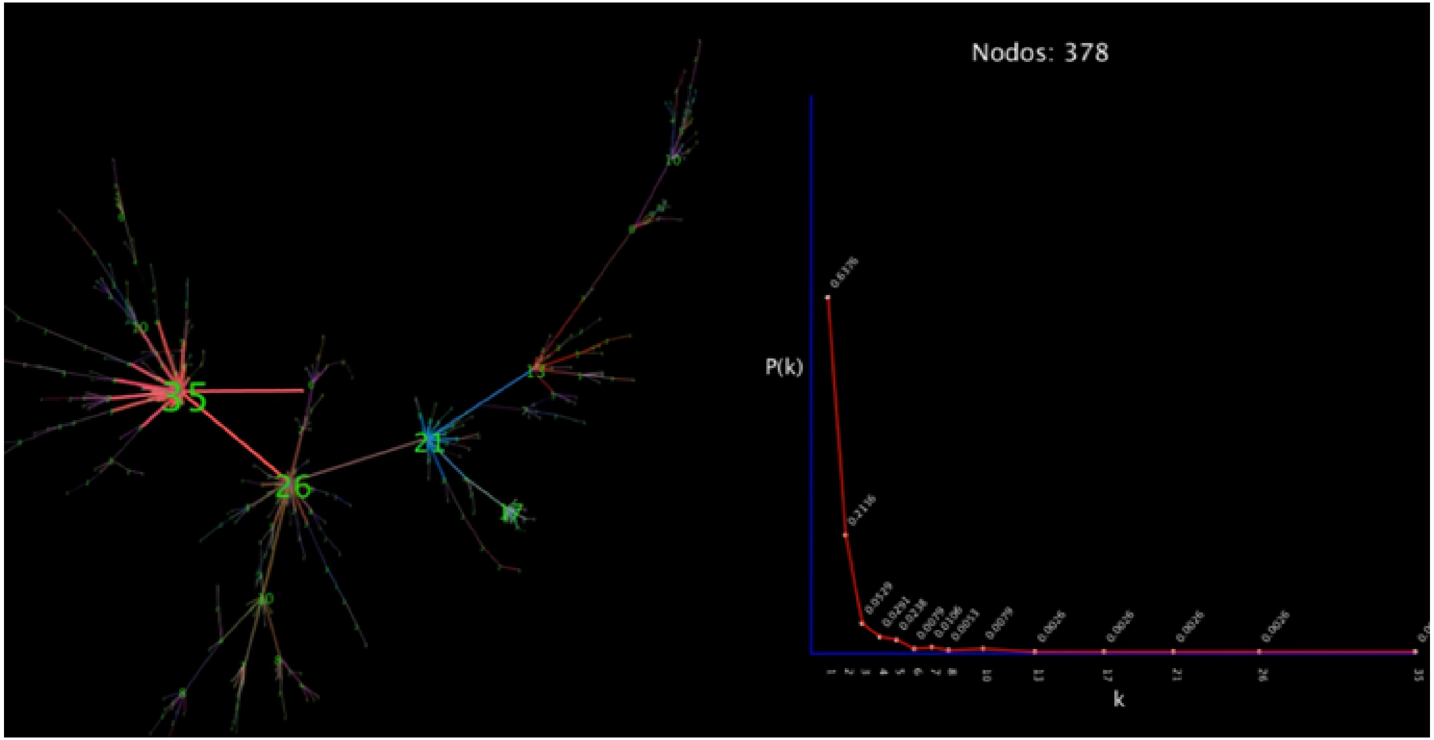
Preferential Attachment

Preferential attachment simulation



https://www.youtube.com/watch?v=4GDqJVtPEGg

Degree distribution in simulation



https://www.youtube.com/watch?v=5RIQweqPT6A

We have seen what but not why

- Power-law degree distributions are prevalent
  - Why?
- Two assumptions in ER networks:
  - There are N nodes that **pre-exist**
  - Nodes connect **at random**
- Let’s challenge both assumptions

Growth

- Suppose there are two web pages on a topic, one with many inlinks the other with few, which one am I most likely to link to?
- Which scientific papers are read? Which are cited?
- Which actors are more sought after for new movies?

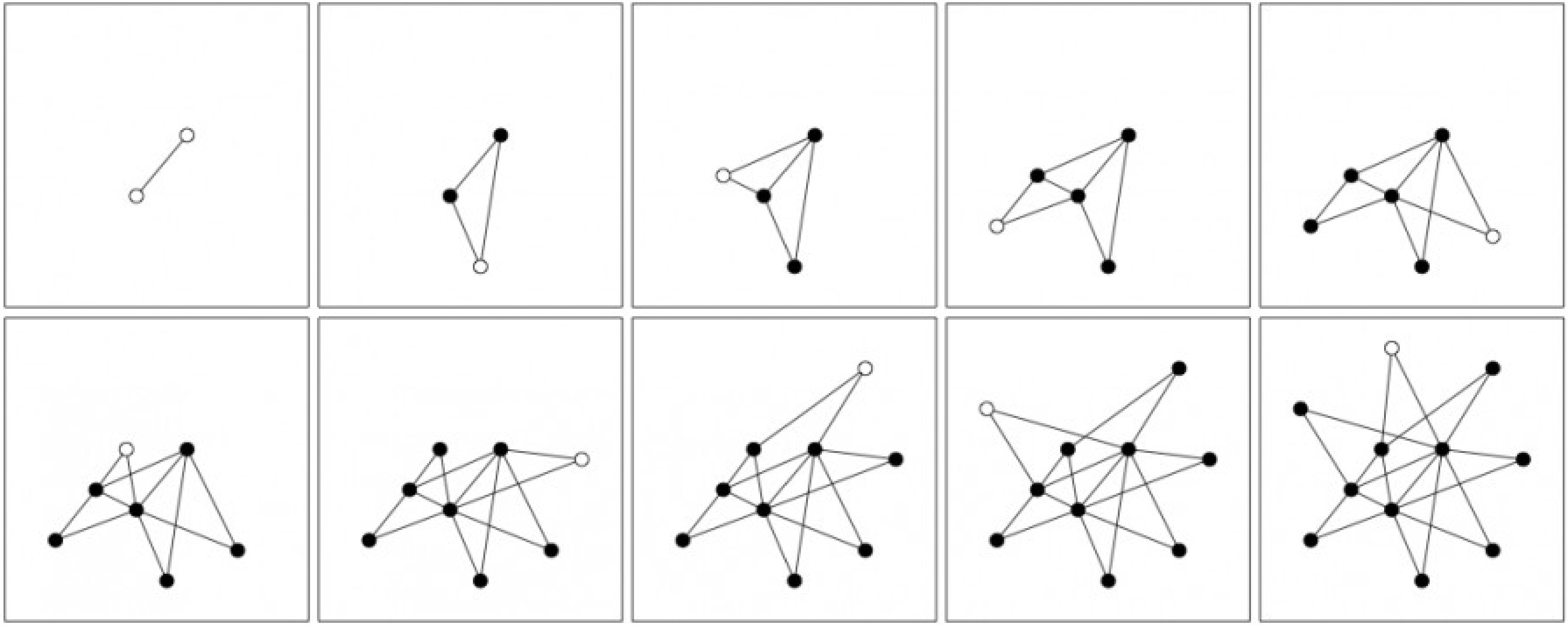

BURNS
UNIVERSITY
Ut
Uberiores
Divites
Our motto: *Ut uberiores divites.*
FOX 32

The Barabási-Albert (BA) model

- Network starts with m nodes
- At every time step we add 1 node
- This node will have $m < m_0$ outlinks
- The probability of an existing node to gain a link is

$$\Pi(k_i) = \frac{k_i}{\sum_{j=1}^{N-1} k_j}$$

Example (m_0 = 2; m=2)



The Barabási-Albert (BA) model

- Network starts with m nodes
- At every time step we add 1 node
- This node will have $m < m_0$ outlinks
- The probability of an existing node to gain a link is

$$\Pi(k_i) = \frac{k_i}{\sum_{j=1}^{N-1} k_j}$$

Write the formula for N(t) and L(t)

Degree $k_i(t)$ as a function of time

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j}$$

$$\sum_{j=1}^{N-1} k_j = 2m(t-1)$$ (All nodes minus the current)

$$\frac{dk_i}{dt} = \frac{k_i}{2t-2} \approx \frac{k_i}{2t}$$ (For large t)

Degree $k_i(t)$ … continued

$$\frac{dk_i(t)}{dt} = \frac{k_i(t)}{2t}$$

$$\frac{1}{k_i(t)} dk_i(t) = \frac{1}{2t} dt$$

$$\int_{t=t_i}^{t} \frac{1}{k_i(t)} dk_i(t) = \int_{t=t_i}^{t} \frac{1}{2t} dt$$

(t_i is the creation time of node i)

$$\log k_i(t) - \log k_i(t_i) = \frac{1}{2} \log t - \frac{1}{2} \log t_i$$

$$\log k_i(t) = \frac{1}{2} \log t - \frac{1}{2} \log t_i + \log m$$

Degree $k_i(t)$ ... continued

$$\log k_i(t) = \frac{1}{2}\log t - \frac{1}{2}\log t_i + \log m$$

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}$$

Is the degree growth linear, super-linear, or sub-linear? Intuitively, why?

Degree $k_i(t)$ … consequences

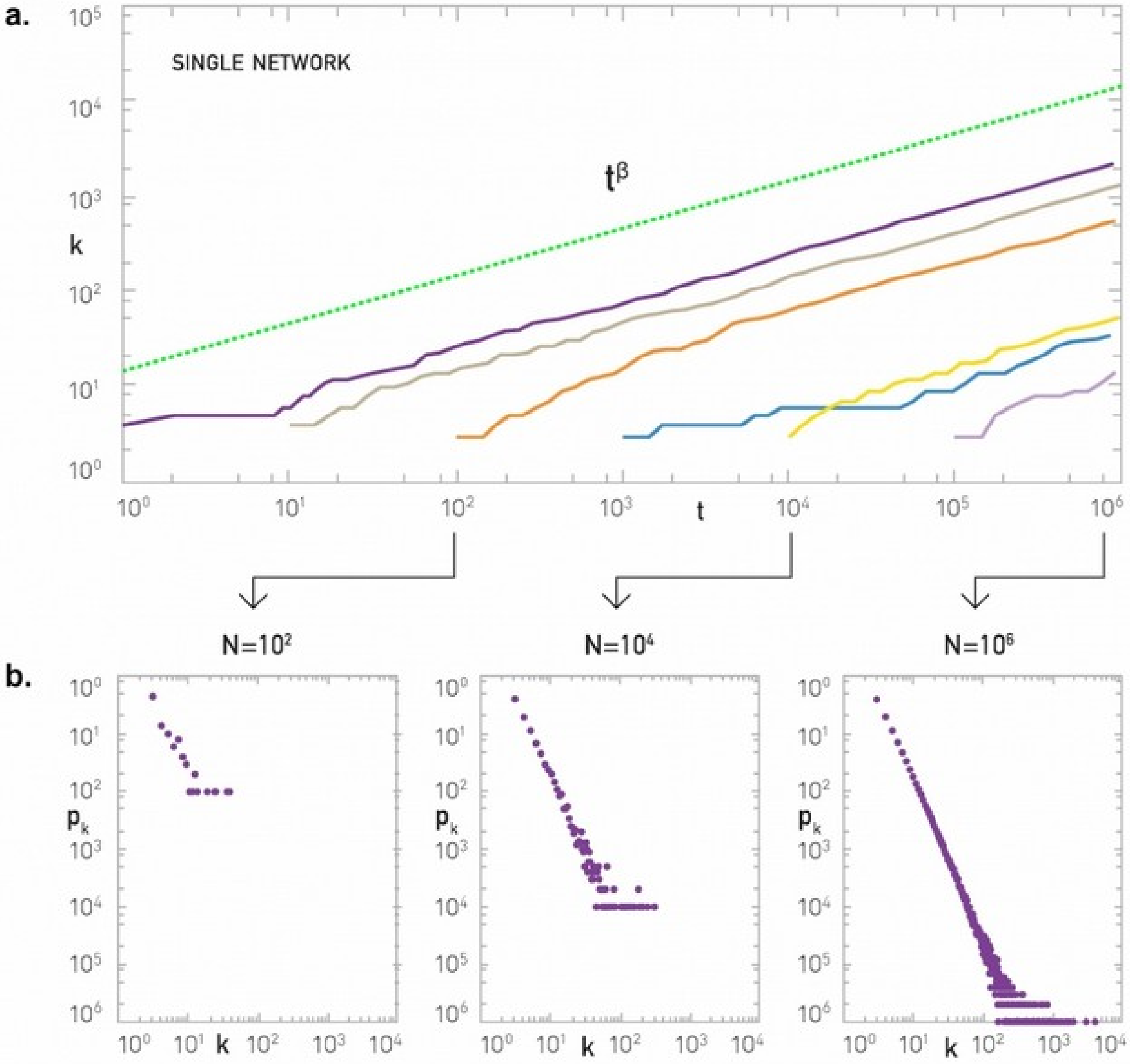
$$\log k_i(t) = \frac{1}{2}\log t - \frac{1}{2}\log t_i + \log m$$

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}$$

$$\frac{dk_i(t)}{dt} = \frac{k_i(t)}{2t} = \frac{m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}}{2t} = \frac{m}{2\left(t \cdot t_i\right)^{\frac{1}{2}}}$$

If t_i < t_j (node i is older than node j), what do we expect of k_i and k_j?

Simulation results



Model

Nodes with t_i = 1, 10, 100, 1000, 10000, ...

Degree distribution

- Let's calculate the CDF of the degree distribution

$$
\begin{aligned}
Pr(k_i \leq k) &= 1 - Pr(k_i > k) \\
&= 1 - Pr\left(m \left(\frac{t}{t_i} \right)^{\beta} > k \right) \\
&= 1 - Pr\left(\left(\frac{m}{k} \right)^{1/\beta} > \frac{t_i}{t} \right) \\
&= 1 - \left(\frac{m}{k} \right)^{1/\beta}
\end{aligned}
$$

$$\frac{ti}{t} \sim \mathrm{Uniform}(0, 1)$$

Degree distribution

Now let's take the derivative of the CDF to obtain the PDF

$$p_k = \frac{\partial}{\partial k} Pr(k_i \leq k) = \frac{\partial}{\partial k}\left(1 - \left(\frac{m}{k}\right)^{1/\beta}\right)$$

$$= -\frac{\partial}{\partial k}\left(\left(\frac{m}{k}\right)^{1/\beta}\right) = -m^{1/\beta}\frac{\partial}{\partial k}\left(\frac{1}{k^{1/\beta}}\right)$$

$$= \frac{1}{\beta}\frac{m^{1/\beta}}{k^{1/\beta+1}}$$

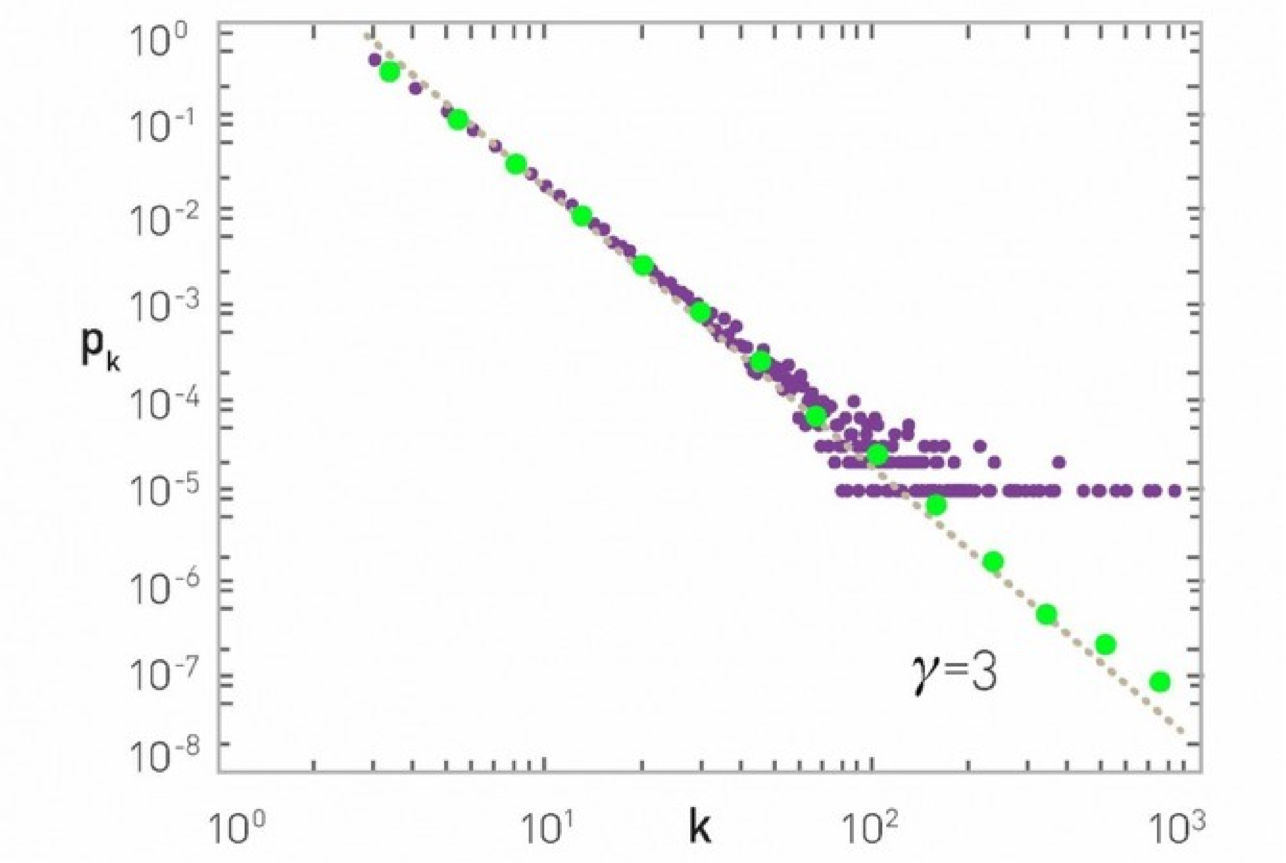
$$p(k) \propto k^{-(1/\beta+1)} = k^{-3}$$

Degree distribution

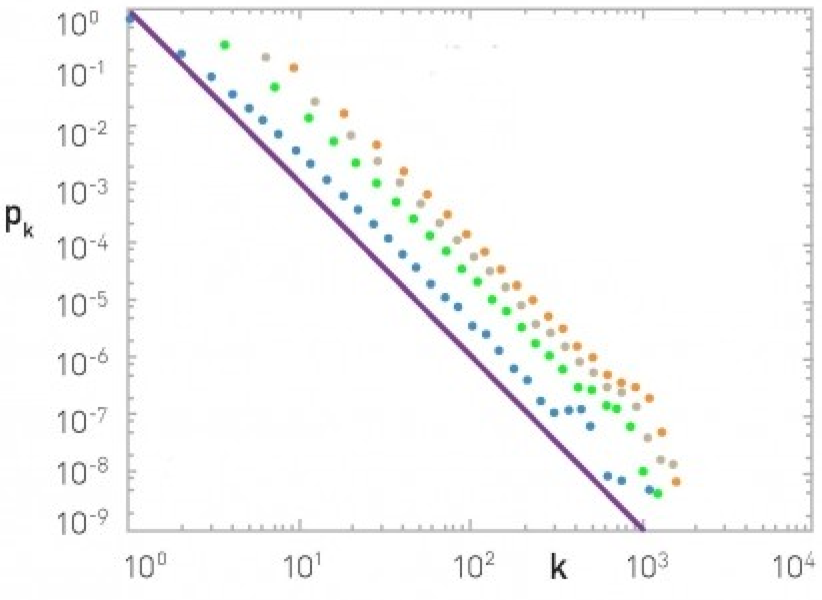
- $\beta = 1/2$ is called the dynamical exponent
- $\gamma = \frac{1}{\beta} + 1 = 3$ is the power-law exponent

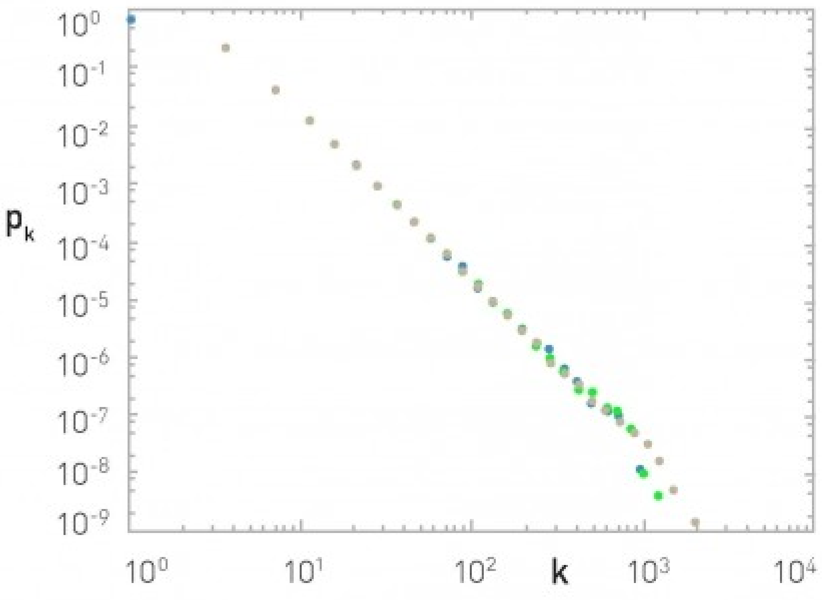
- Note that $p(k) \approx 2m^2 k^{-3}$
  does not depend on t
  hence, it describes a stationary network

Degree distribution, simulation results
N=100,000 m=3



More simulations





N = 100,000; m_0 = m =
1 (blue), 3 (green), 5 (gray), 7 (orange)

Observe γ is independent of m (and m_0)

The slope of the purple line is -3

m_0 = m = 3; N =
50K (blue), 100K (green), 200K (gray)

Observe p_k is independent of N

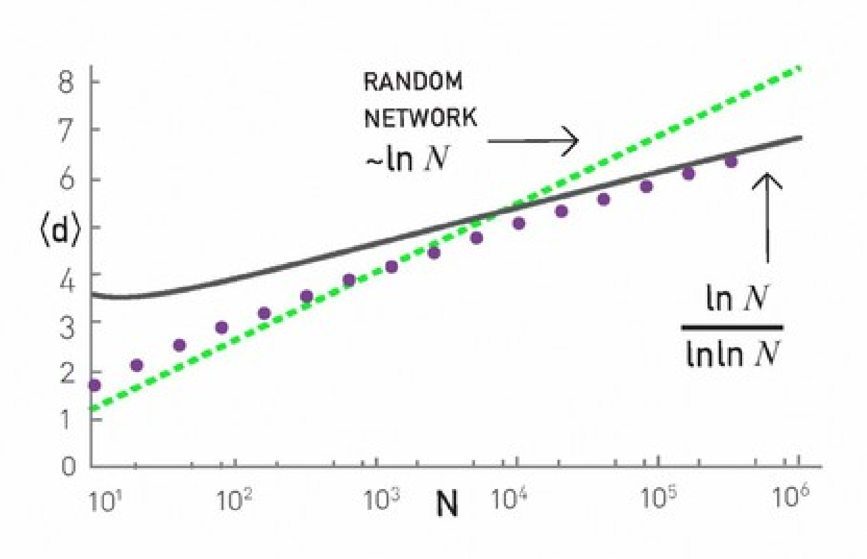
Processes that generate BA networks

- Link-selection model (step)
  - Add one new node v to the network
  - Select an existing link at random and connect v to one of the edges of that existing link
- Copy model (step)
  - Add one new node v to the network
  - Pick a random existing node u
  - With probability p link to u
  - With probability 1-p link to a neighbor of u

Average distance

- Distances grow slower than log N

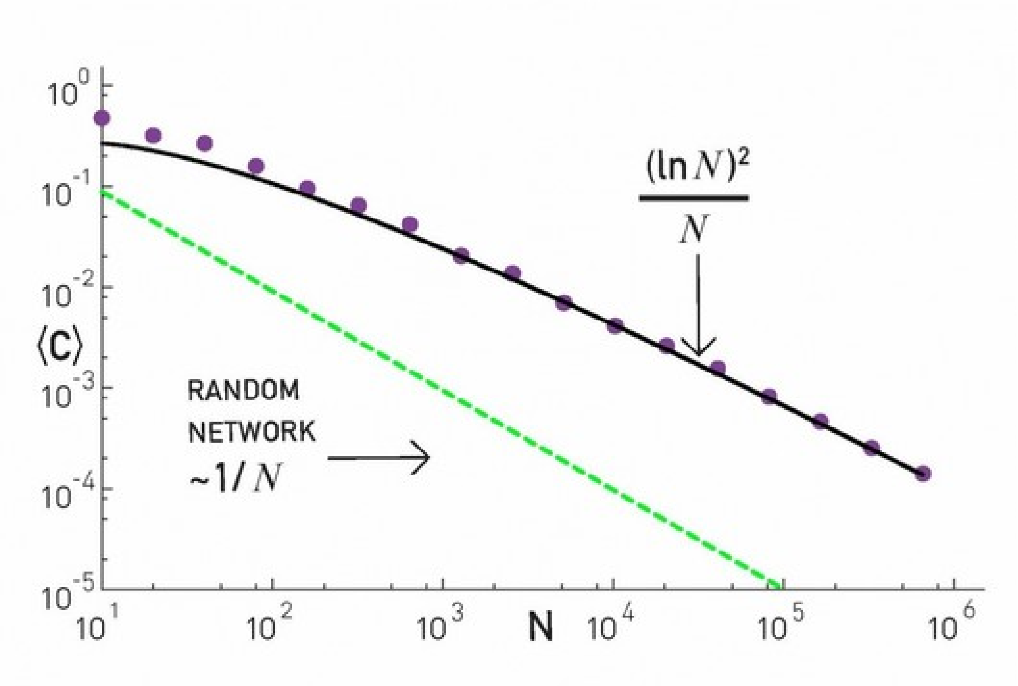
$$\langle d \rangle \sim \frac{\log N}{\log \log N}$$



Clustering coefficient

- BA networks are locally more clustered than ER networks

$$\langle C \rangle \sim \frac{(\log N)^2}{N}$$



Limitations of the BA model

- Predicts a fixed exponent of -3
- Assumes an undirected network, while many real complex networks are directed
- Does not consider node deletions or edge deletions which are common in practice
- Considers that all nodes are equal except for their arrival times

Exercise: the copy model

- In the copy model, at every step t:
  - Add one new node v to the network
  - Pick a random existing node u
  - With probability p link to u
  - With probability 1-p link to a neighbor of u
- Simulate it on paper for 5 nodes with p=0.5
  - Make sure you understand the model fully!
- What is N(t) and L(t)?

In the copy model, at every step t:
1) Add one new node v to the network
2) Pick a random existing node u
3) With probability p link to u
4) With probability 1-p link to a neighbor of u

- What is k_j^{out} ?
- We will compute $k_j^{\text{in}} = k_j$
- How many links on average gets node i at time t? In other words, what is …

$$\frac{d}{dt} k_i(t)$$

- Hint: it has a term with p and a term with 1-p

- Rearrange terms in $\frac{d}{dt}k_i(t)$
- … to do something similar to what we did for the BA model, then integrate between t_i and t to obtain an expression for $k_i(t_i)$
- Note that now

$$k_i(t_i) = k_i^{\mathrm{in}}(t_i) = 0$$

- Once you have a expression for $k_i(t_i)$
- Compute $Pr(k_i(t_i) > k)$
- Now write the cumulative distribution function of $k_i(t_i)$
- And compute its derivative to obtain

$$p_k = Pr(k_i(t) = k) = \frac{d}{dk} Pr(k_i(t) \leq k)$$

- It should show exponent $\gamma = \dfrac{2-p}{1-p}$

Practice on your own

- Try to reconstruct the derivations we have done in class, including the exercise
  - Try to understand every step
- Insert a small change in the model and try to recalculate what we have done