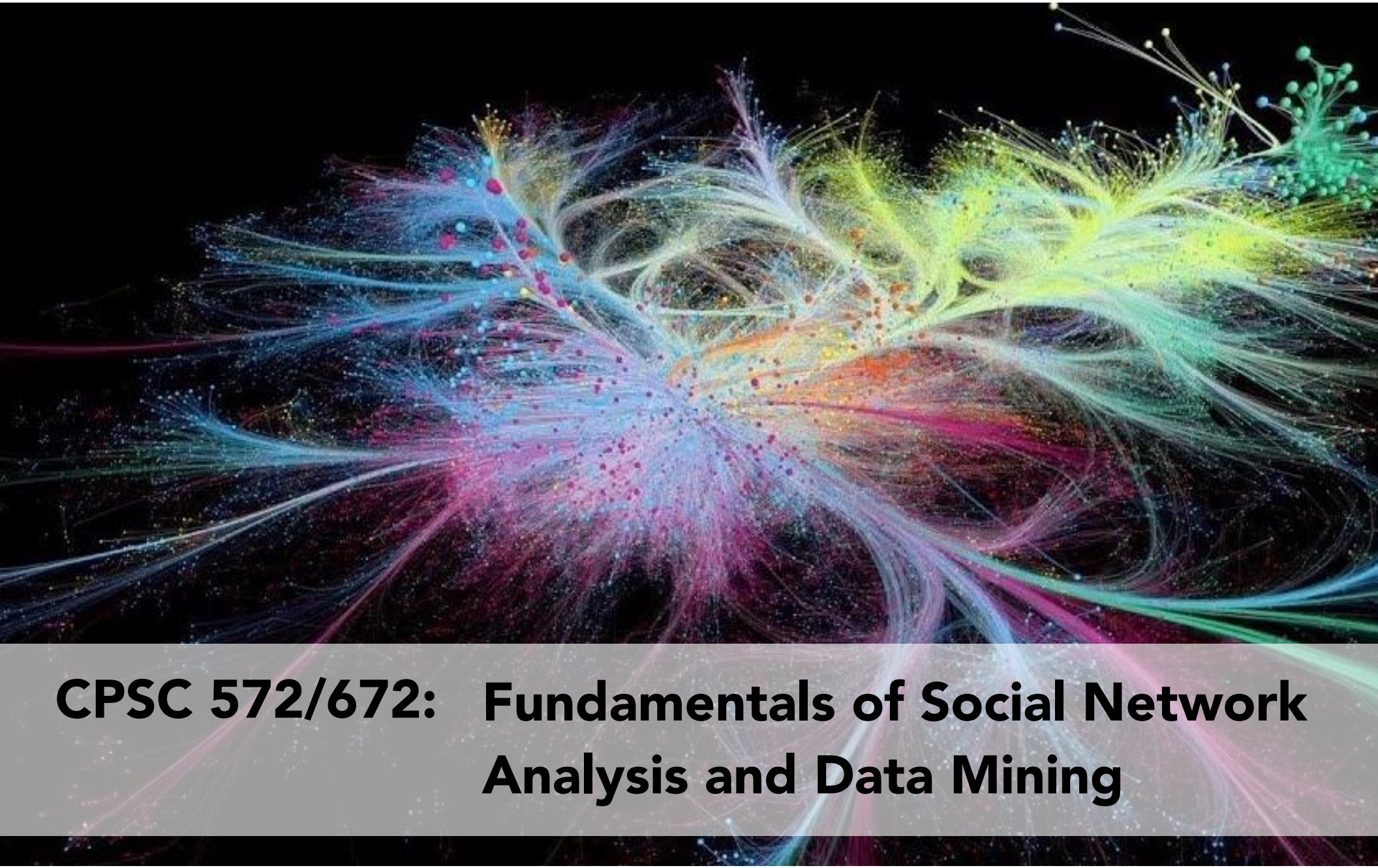


CPSC 572/672: Fundamentals of Social Network Analysis and Data Mining

Schedule

Project check-in presentations

Thursday 20th and 27th October

- Quick recap (what are your nodes and links? Why do we care about your network?)
- Screenshot of your data
- How did you construct your network from your data?
- Visualisation (can be a subset)
- Degree distribution
- Have your research questions changed?

4 mins - strictly enforced (2 mins feedback)
NOT graded, just for feedback and the benefit of the class
Upload a pdf by midnight the night before your check-in

Paper presentations

Tue.1st, Thu. 3rd November

- Work in your project groups
- Find a paper on the topic of social network analysis
    - Social network: same interpretation as for your projects
    - Towards social justice: First and/or last author must be from an underrepresented group in the sciences.
    - How to find papers?
- Review the paper
    - Main findings
    - What is good about the paper, what are the limitations?
    - How does it fit into what we are learning?
- Present your review - 6 mins, strictly enforced

Upload a pdf of your slides by midnight the night before your presentation.

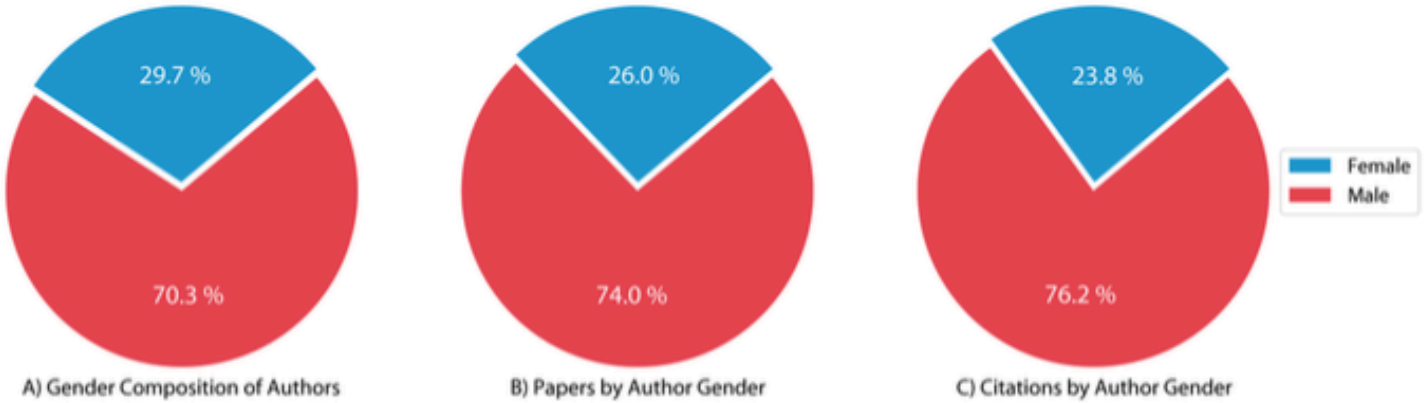


Figure 1: Gender breakdown of Network Science articles published in since 1998 for (A) unique authors (each unique author counted once) (B) authorships (authors with multiple papers counted multiple times) (C) citations (authors with multiple citations counted multiple times)

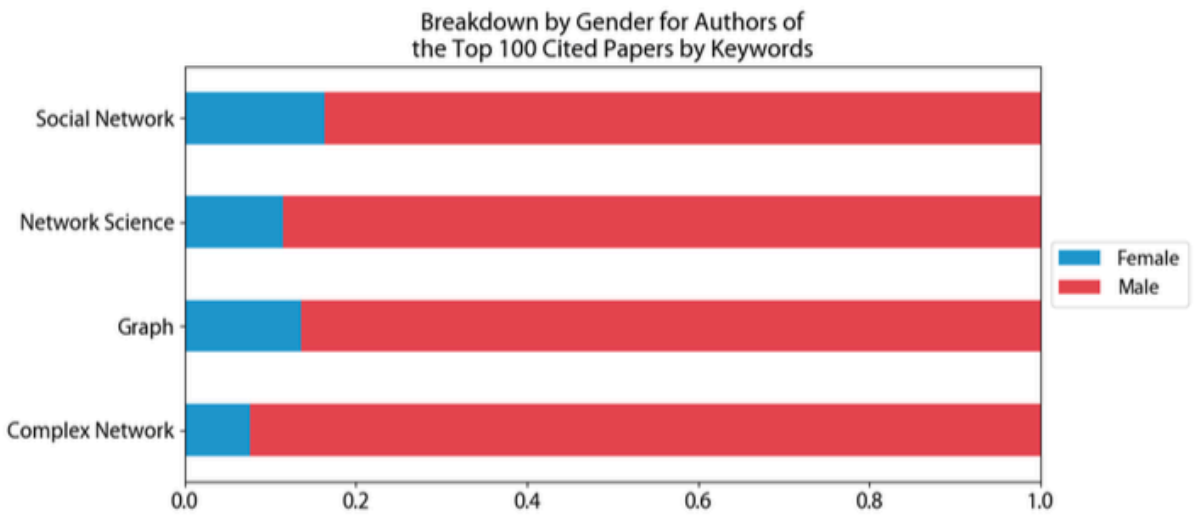


Figure 5: Gender breakdown of the 100 most highly cited authors in Network Science by topic keyword.

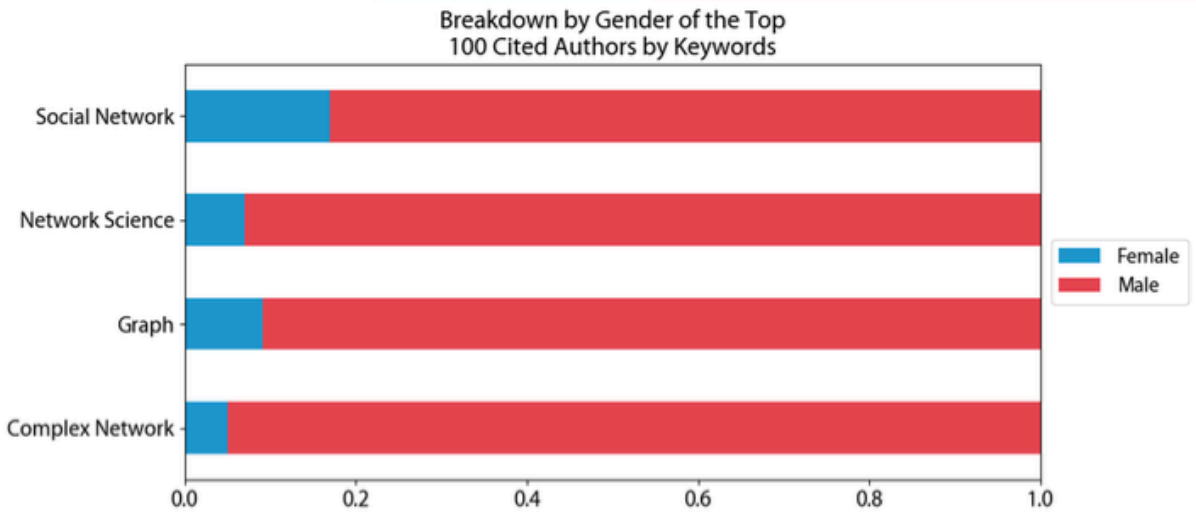


Figure 6: Gender breakdown of the 100 most highly cited authors in Network Science by topic keyword.

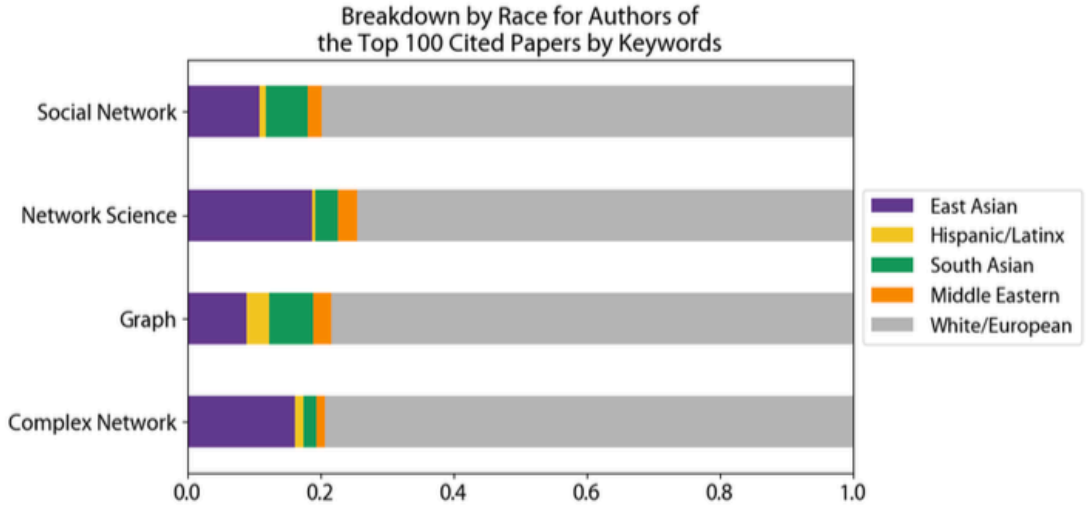


Figure 7: Breakdown by race of the authors of the 100 most highly cited papers in Network Science by topic keyword.

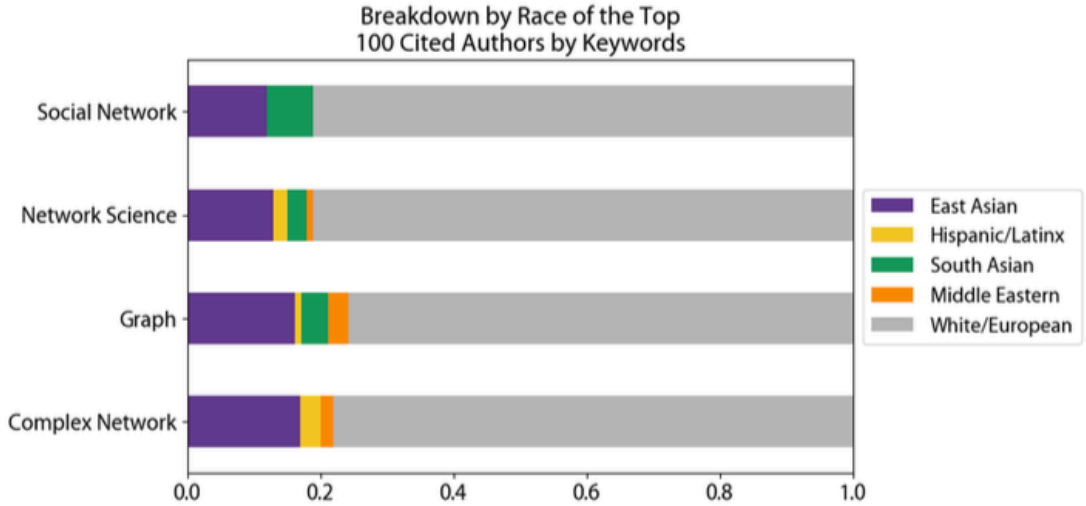


Figure 8: Breakdown by race of the 100 most highly cited authors in Network Science by topic keyword.

Barabási-Albert Model

Hubs represent the most striking difference between a random and a scale-free network. Their emergence in many real systems raises several fundamental questions:

•Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?

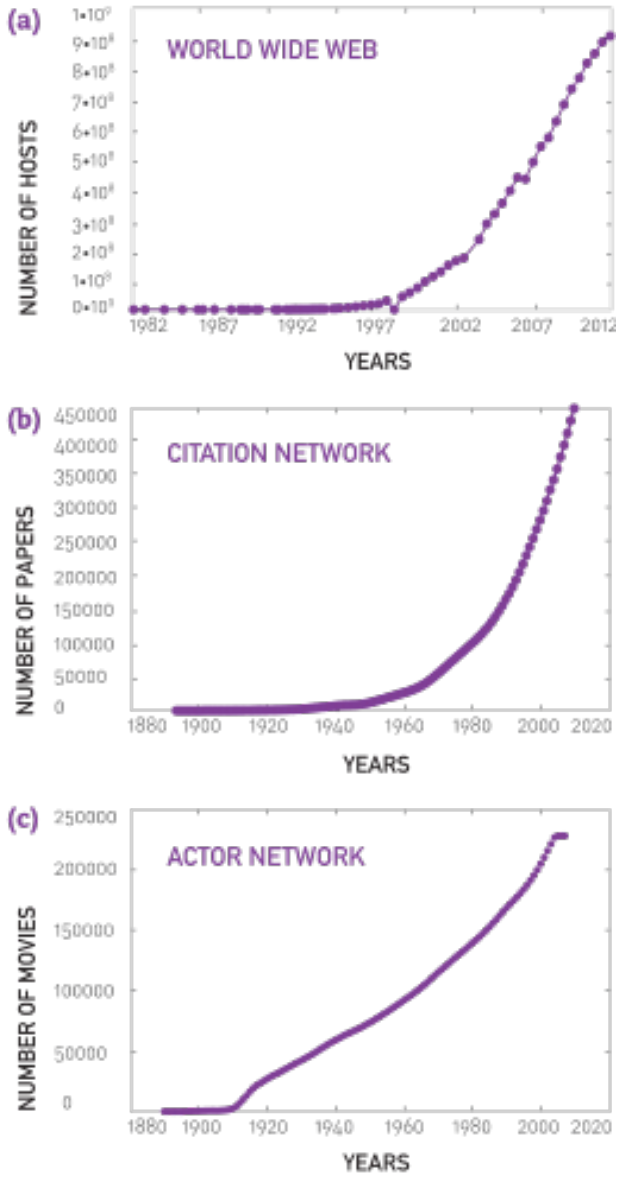
• Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

Growth and preferential attachment

BA MODEL: Growth

ER model:
the number of nodes, N, is fixed (static models)

networks expand through the addition of new nodes



Barabási & Albert, *Science* **286,** 509 (1999)

BA MODEL: Preferential attachment

ER model: links are added randomly to the network

New nodes prefer to connect to the more connected nodes

Barabási & Albert, *Science* **286,** 509 (1999)

Section 2: Growth and Preferential Sttachment

The random network model differs from real networks in two important characteristics:

Growth: While the random network model assumes that the number of nodes is fixed (time invariant), real networks are the result of a growth process that continuously increases.

Preferential Attachment: While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.

The Barabási-Albert model

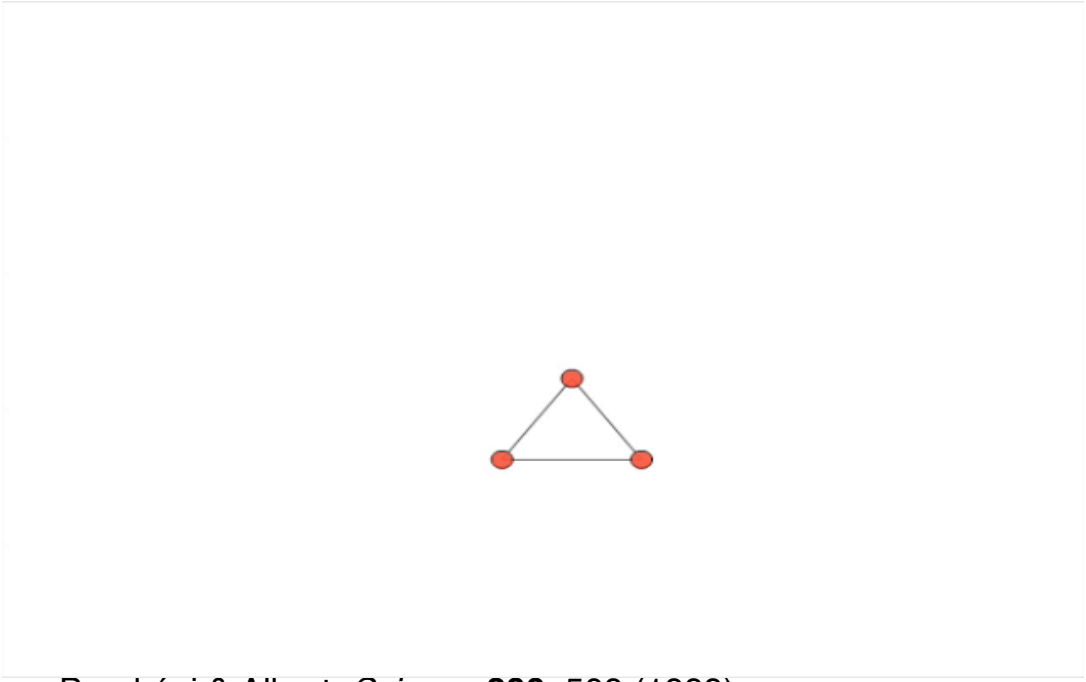
Origin of SF networks: Growth and preferential attachment

(1) Networks continuously expand by the addition of new nodes

WWW : addition of new documents

(2) New nodes prefer to link to highly connected nodes.

WWW : linking to well known sites



Barabási & Albert, *Science* **286,** 509 (1999)

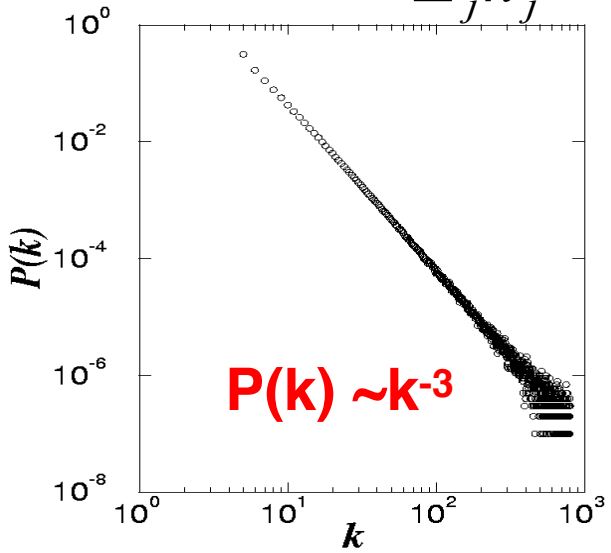
GROWTH:

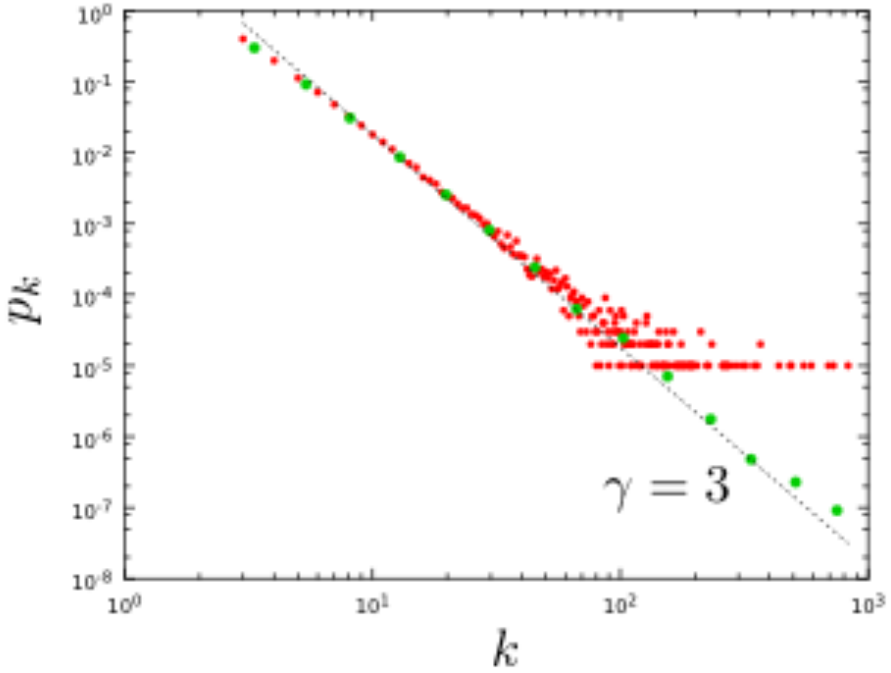
add a new node with m links

PREFERENTIAL ATTACHMENT:

the probability that a node connects to a node with *k* links is proportional to *k*.

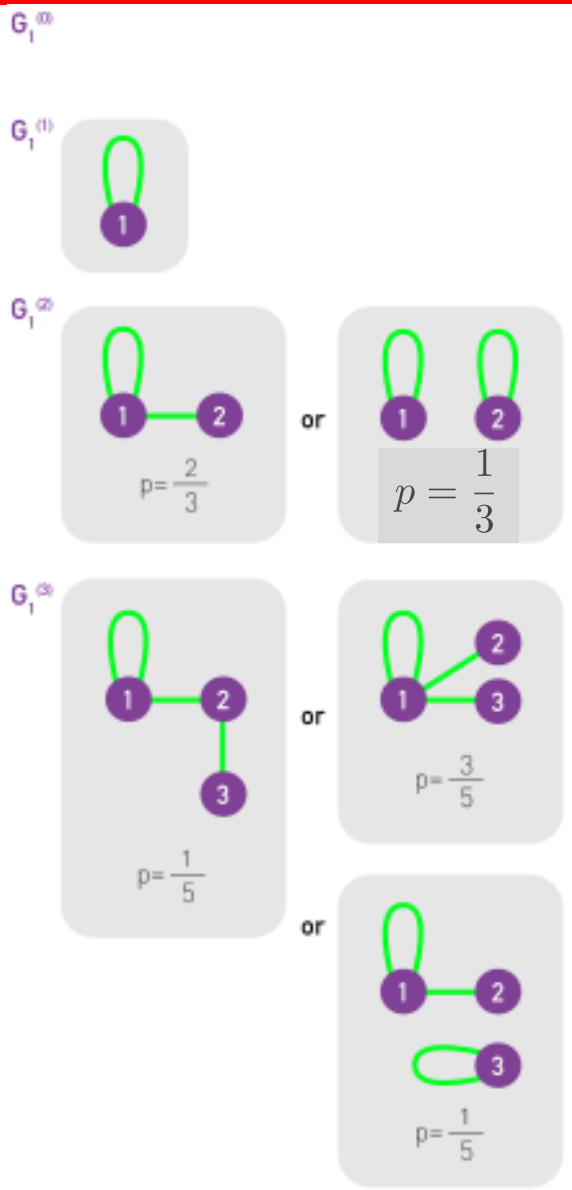
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$




10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
p_k
10^0
10^1
10^2
10^3
k
$\gamma = 3$

The definition of the Barabási-Albert model leaves many mathematical details open:

- It does not specify the precise initial configuration of the first m_0 nodes.
- It does not specify whether the m links assigned to a new node are added one by one, or simultaneously. This leads to potential mathematical conflicts: If the links are truly independent, they could connect to the same node i, leading to multi-links.



Open your iPython notebook

Degree dynamics

All nodes follow the same growth law

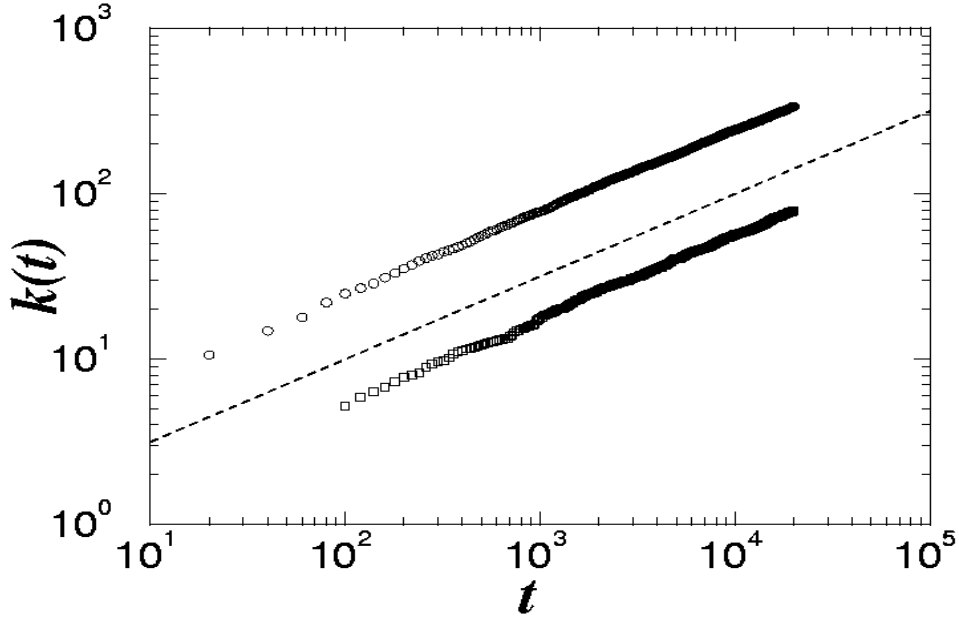
$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) = A\frac{k_i}{\sum_j k_j}$$

A.-L.Barabási, R. Albert and H. Jeong, *Physica* A **272,** 173 (1999)

All nodes follow the same growth law

$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) = A\frac{k_i}{\sum_j k_j}$$

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

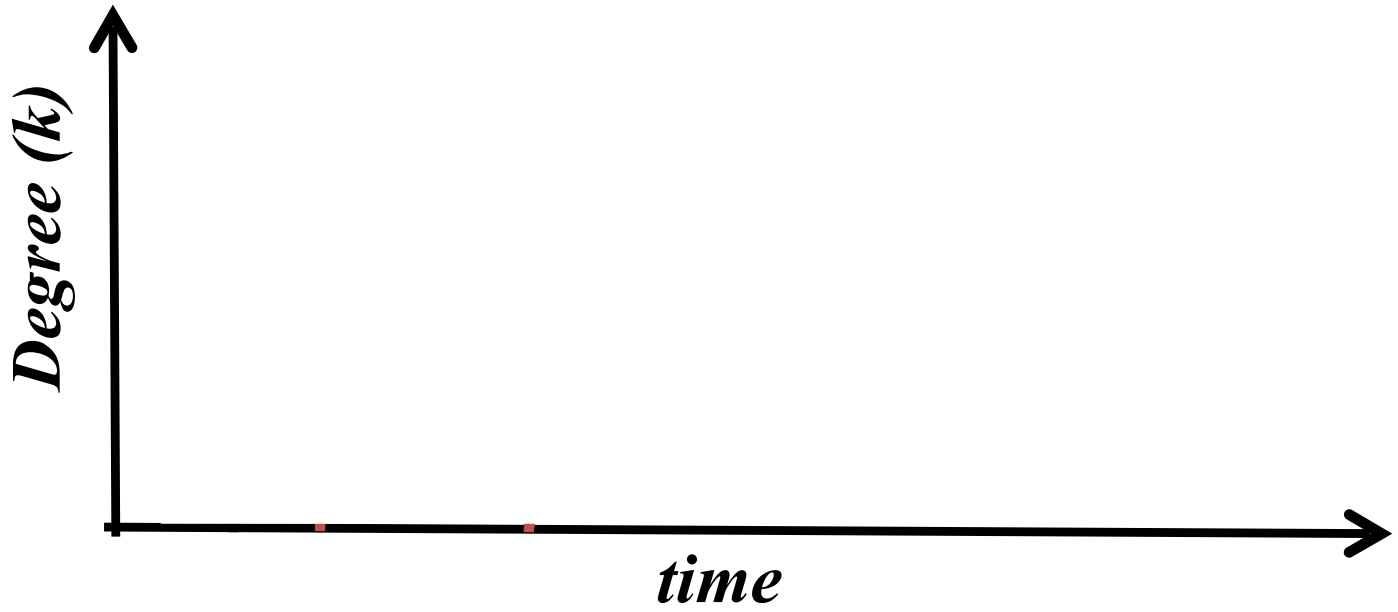


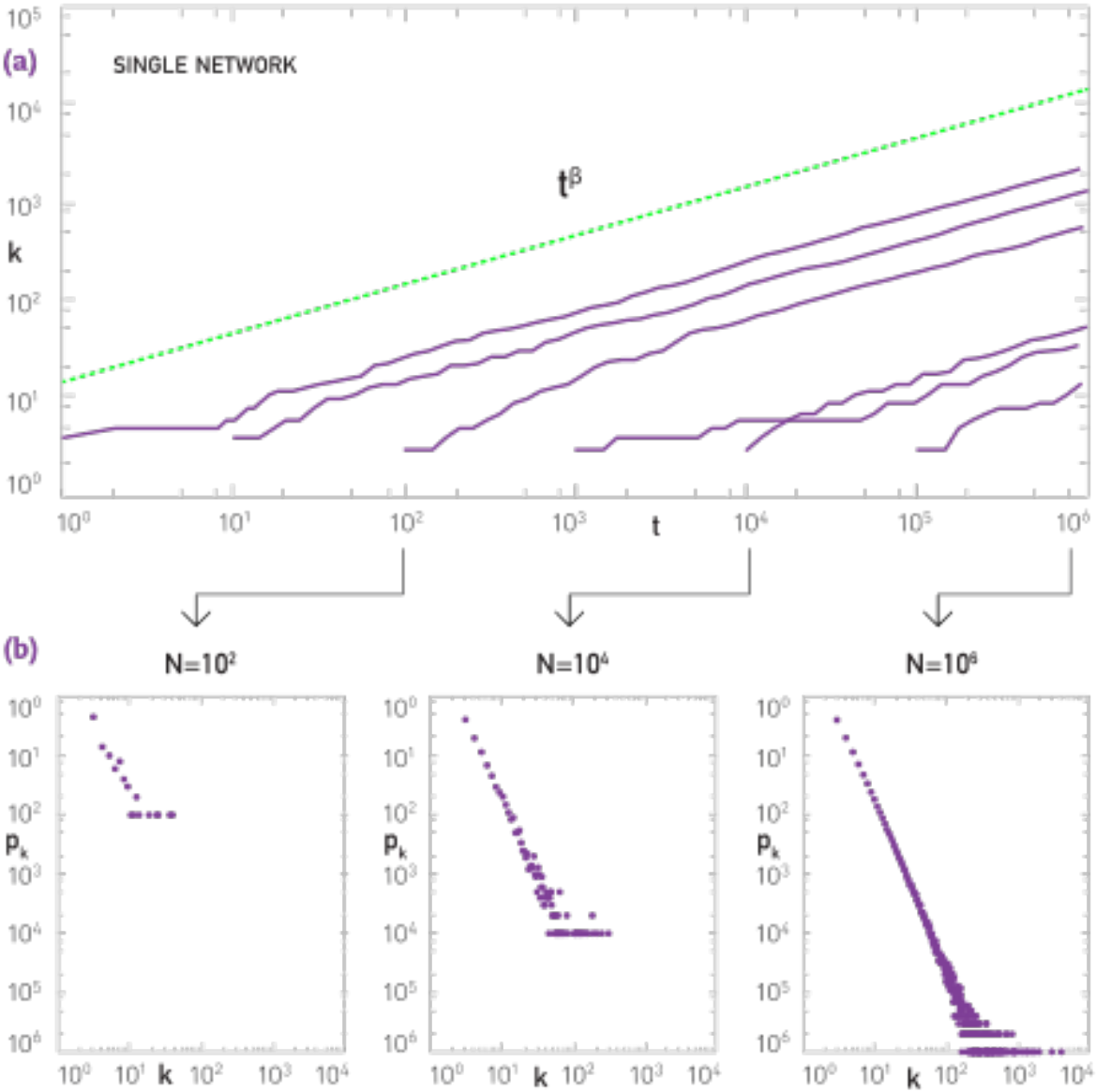
β: dynamical exponent

A.-L.Barabási, R. Albert and H. Jeong, *Physica* A **272,** 173 (1999)

All nodes follow the same growth law

SF model: $k(t) \sim t^{1/2}$ **(first mover advantage)**





- The degree of each node increases following a power-law with the same dynamical exponent β =1/2 (Figure 5.6a). Hence all nodes follow the same dynamical law.

- The growth in the degrees is sublinear (i.e. $\beta < 1$). This is a consequence of the growing nature of the Barabási-Albert model: Each new node has more nodes to link to than the previous node. Hence, with time the existing nodes compete for links with an increasing pool of other nodes.

- The earlier node i was added, the higher is its degree $k_i(t)$. Hence, hubs are large because they arrived earlier, a phenomenon called *first-mover advantage* in marketing and business.

- The rate at which the node i acquires new links is given by the derivative of (5.7)

$$\frac{dk_i(t)}{dt} = \frac{m}{2}\frac{1}{\sqrt{t_i t}}, \tag{5.8}$$

indicating that in each time frame older nodes acquire more links (as they have smaller t_i). Furthermore the rate at which a node acquires links decreases with time as $t^{-1/2}$. Hence, fewer and fewer links go to a node.

Degree distribution

Degree distribution

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

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \qquad \boxed{\gamma = 3}$$

$$P(k) \sim k^{-3} \qquad \text{for large } k$$

(i) The degree exponent is independent of m.

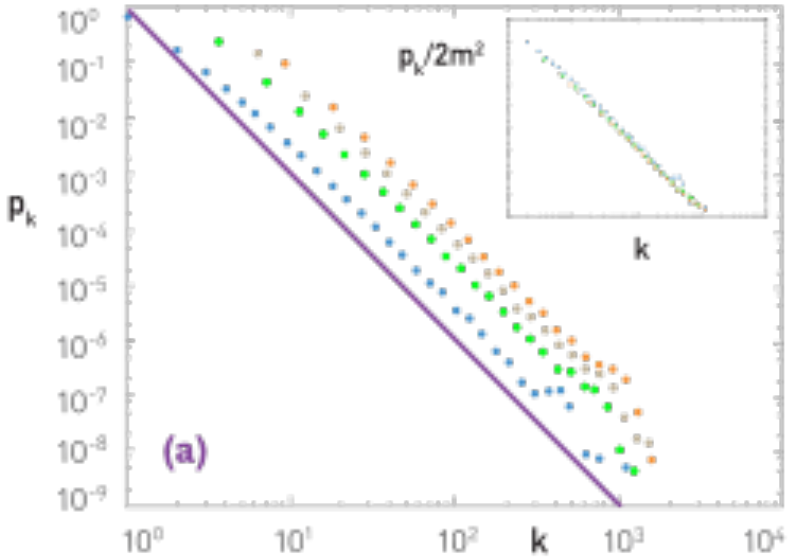
(ii) As the power-law describes systems of rather different ages and sizes, it is expected that a correct model should provide a time-independent degree distribution. Indeed, asymptotically the degree distribution of the BA model is independent of time (and of the system size N)

→ the network reaches a stationary scale-free state.

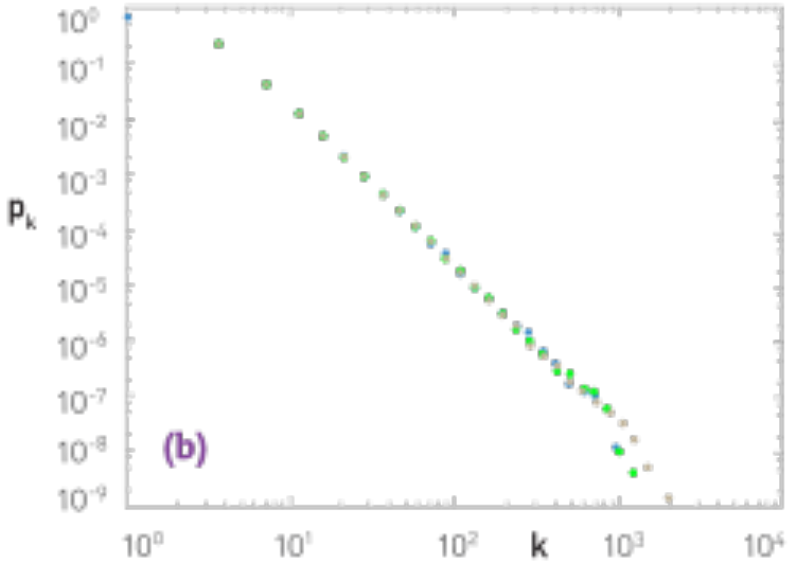
(iii) The coefficient of the power-law distribution is proportional to m^2.

NUMERICAL SIMULATION OF THE BA MODEL

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}$$



(a) We generated networks with N=100,000 and m_0=m=1 (blue), 3 (green), 5 (grey), and 7 (orange). The fact that the curves are parallel to each other indicates that γ is independent of m and m_0. The slope of the purple line is -3, corresponding to the predicted degree exponent γ=3. Inset: (5.11) predicts $p_k \sim 2m^2$, hence $p_k/2m^2$ should be independent of m. Indeed, by plotting $p_k/2m^2$ vs. k, the data points shown in the main plot collapse into a single curve.



(b) The Barabási-Albert model predicts that p_k is independent of N. To test this we plot p_k for N = 50,000 (blue), 100,000 (green), and 200,000 (grey), with m_0=m=3. The obtained p_k are practically indistinguishable, indicating that the degree distribution is stationary, i.e. independent of time and system size.

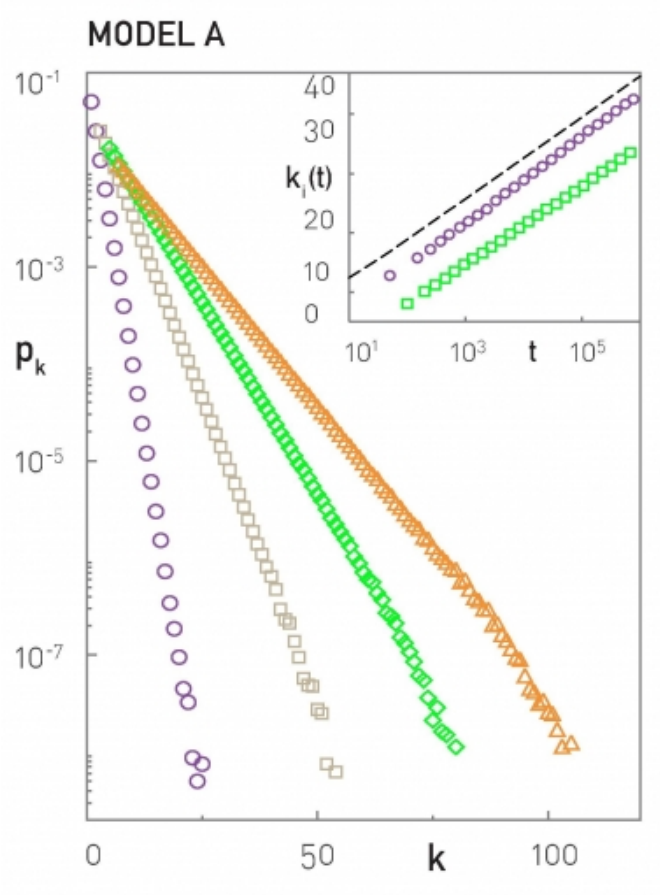
absence of growth and preferential attachment

MODEL A

growth ~~preferential attachment~~

$\Pi(k_i)$: uniform

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) = \frac{m}{m_0 + t - 1}$$



MODEL A

growth ~~preferential attachment~~

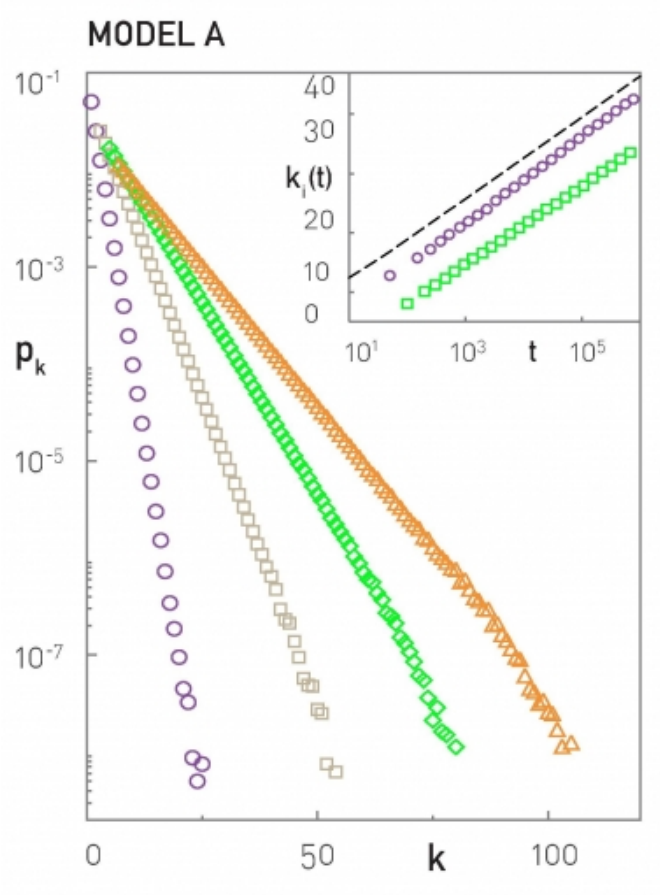
$\Pi(k_i)$: uniform

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) = \frac{m}{m_0 + t - 1}$$

$$k_i(t) = m\ln(\frac{m_0 + t - 1}{m + t_i - 1}) + m$$

$$P(k) = \frac{e}{m}\exp(-\frac{k}{m}) \sim e^{-k}$$

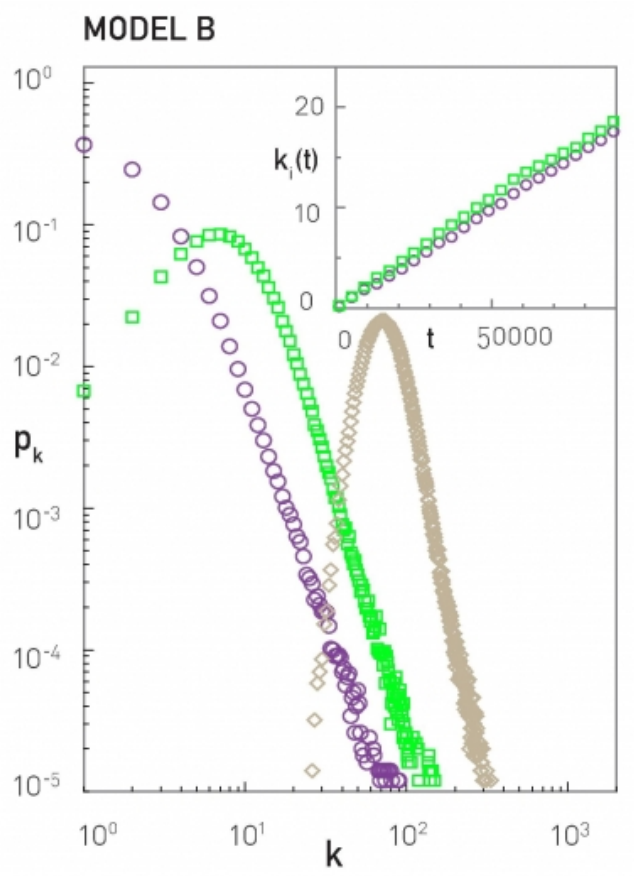
-> Exponential distribution -> no hubs



MODEL B

~~growth~~ preferential attachment

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) + \frac{1}{N} = \frac{N}{N-1}\frac{k_i}{2t} + \frac{1}{N}$$

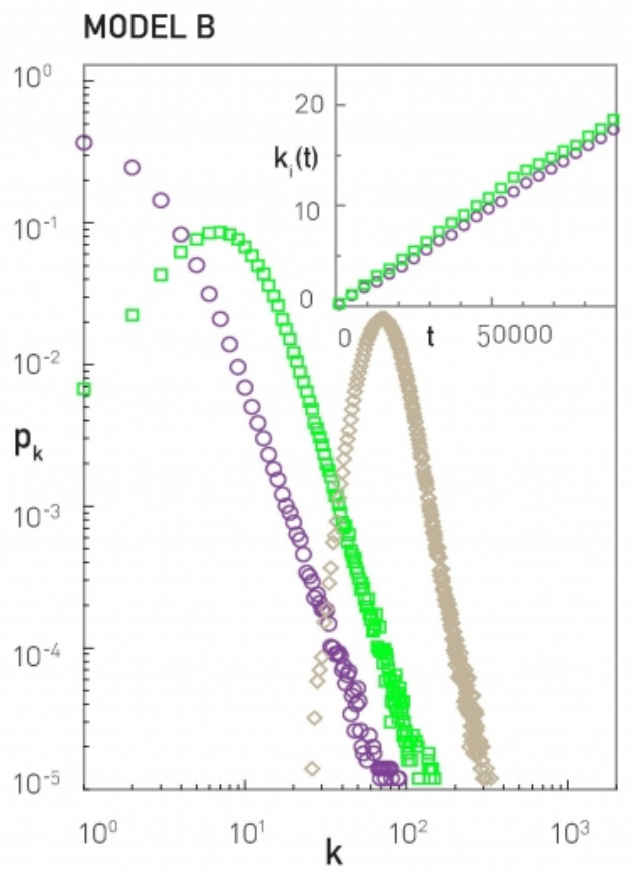


MODEL B

~~growth~~ preferential attachment

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) + \frac{1}{N} = \frac{N}{N-1}\frac{k_i}{2t} + \frac{1}{N}$$

$$k_i(t) = \frac{2(N-1)}{N(N-2)}t + Ct^{\frac{N}{2(N-1)}} \sim \frac{2}{N}t$$



p_k : not stationary. Power law (initially) →

→ Gaussian → Fully Connected

Do we need both growth and preferential attachment?

YEP.

Measuring preferential attachment

$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) \sim \frac{\Delta k_i}{\Delta t}$$

Plot the change in the degree Δk during a fixed time Δt for nodes with degree k.

To reduce noise, plot the integral of $\Pi(k)$ over

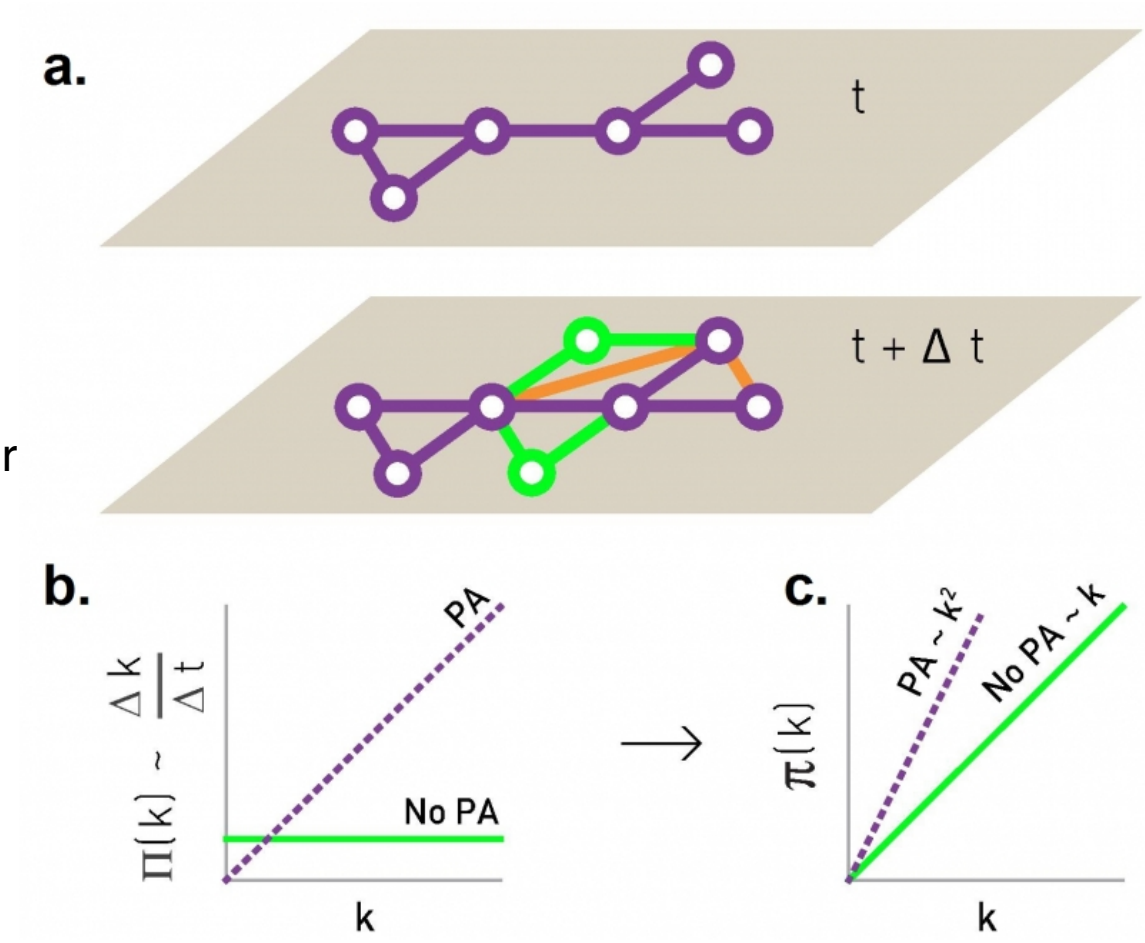
$$\kappa(k) = \sum_{K<k} \Pi(K)$$

No pref. attach:

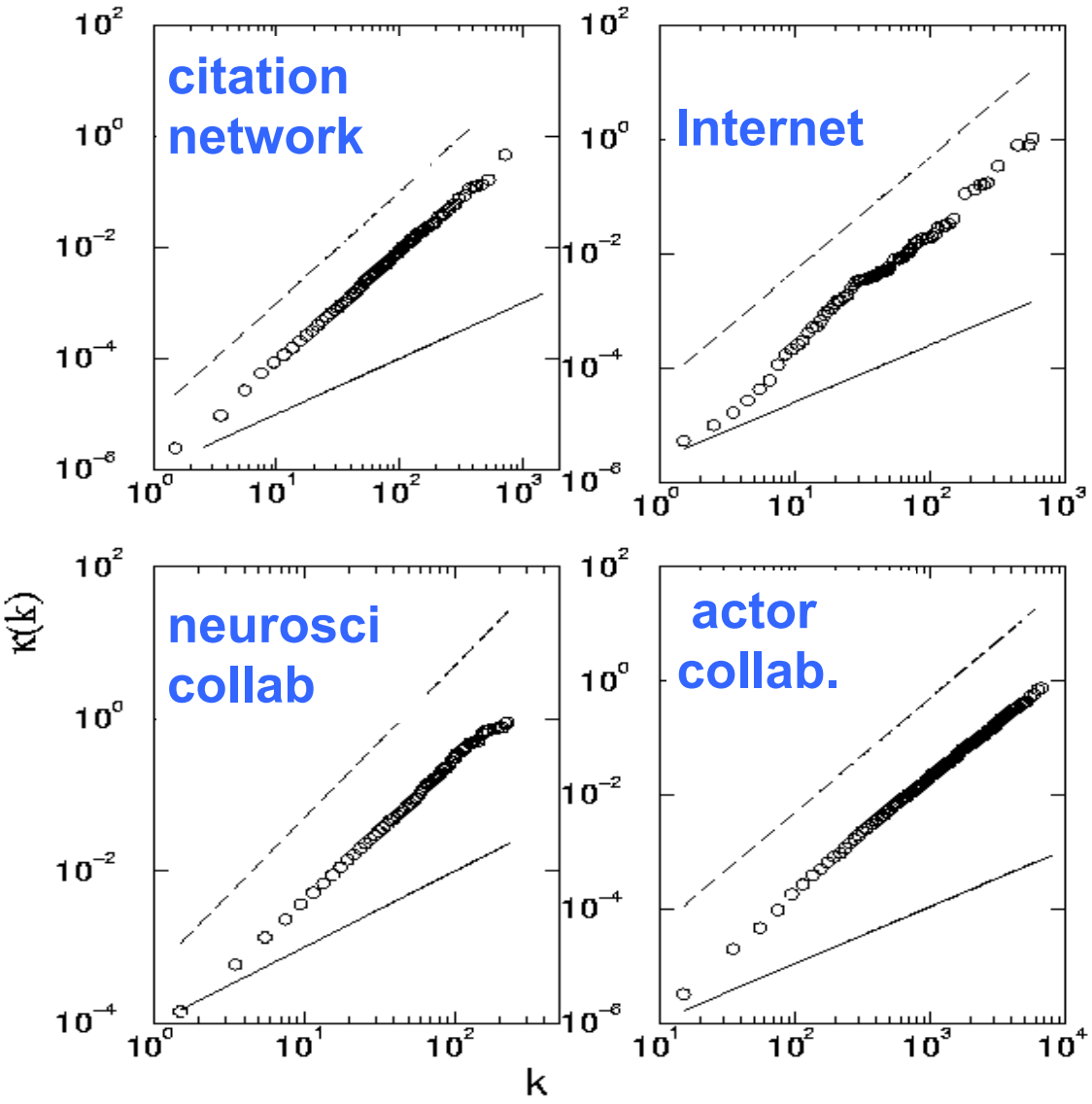
$\kappa \sim k$ ——

Linear pref. attach:

$\kappa \sim k^2$ - - - -



(Jeong, Neda, A.-L. B, Europhys Letter 2003; cond-mat/0104131)



Plots shows the integral of Π(k) over k:

$$\kappa(k) = \sum_{K<k} \Pi(K)$$

No pref. attach:

$\kappa \sim k$ ——

Linear pref. attach:

$\kappa \sim k^2$ - - - -