NETWORK SCIENCE OF ONLINE INTERACTIONS

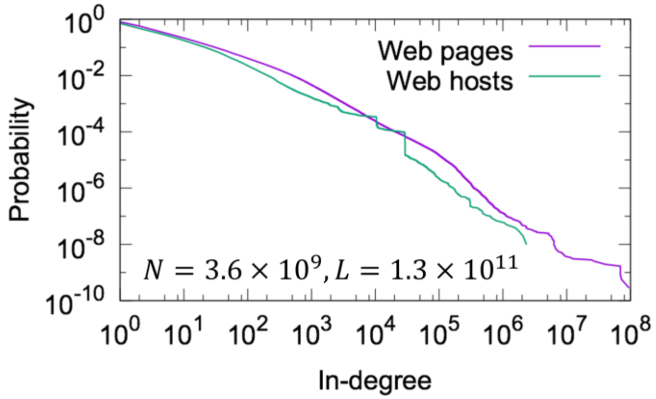
Chapter 5: Network Models

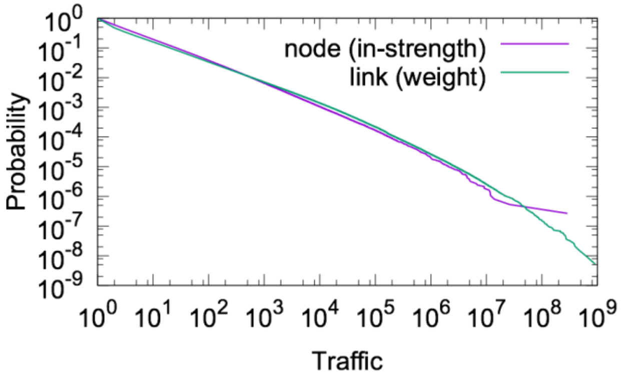
Joao Neto

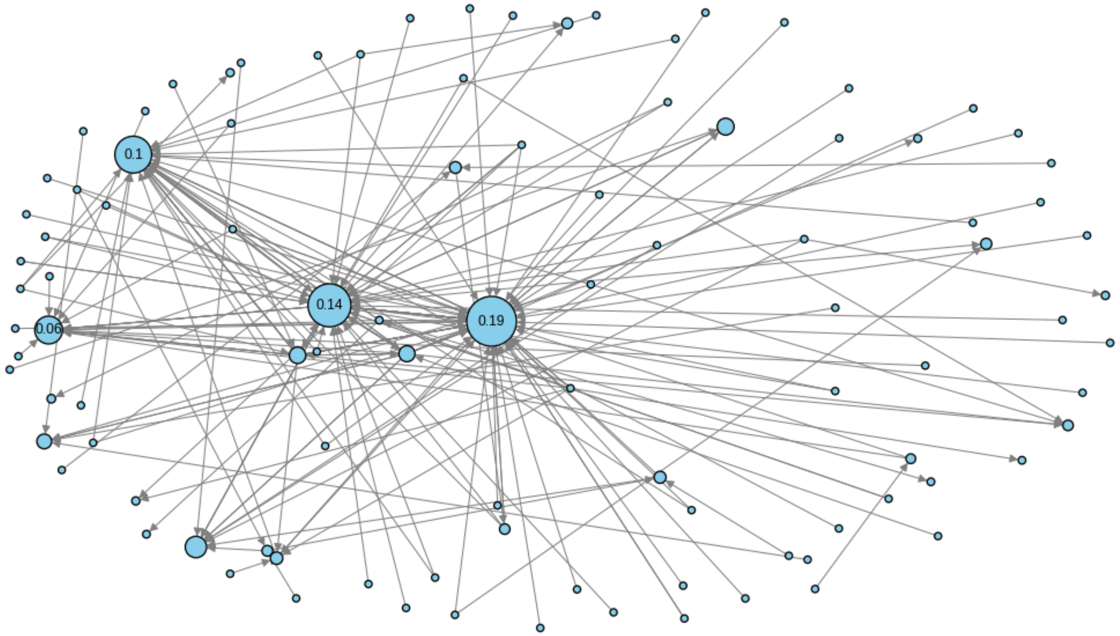
19/May/2023

SUMMARY

- Lots of directed networks are heavy-tailed both in degree and weights
- Word embedding is a powerful tool to study content networks (e.g. social media)
- Diffusion models are very useful
  - Metrics that model diffusion can excel (PageRank)
  - Can be used to study spread of misinformation in social media







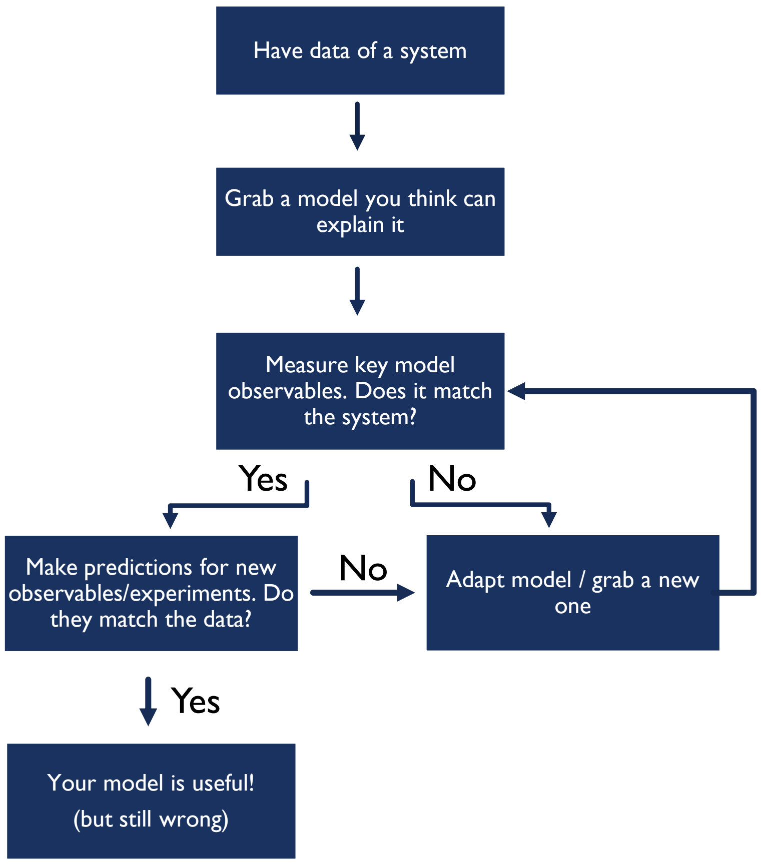
NETWORK MODELS

- Real-world networks tend to have certain properties
  - Short paths between nodes
    - Small-world property
  - Many triangles
    - High clustering coefficient
  - Heterogeneous distributions
    - Power-law degree distribution

- What simple mechanisms can create those properties?
  - “All models are wrong, some are useful”

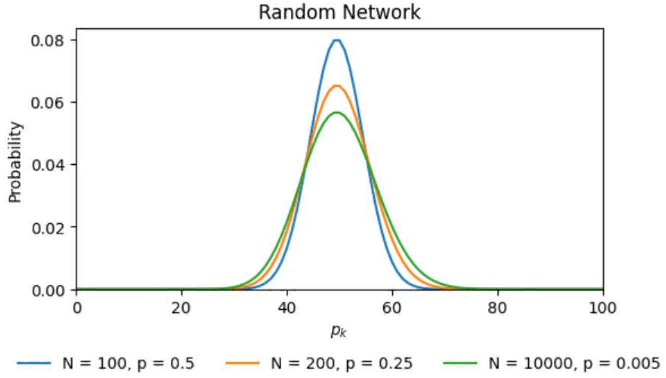
- Iterative process: change/explore model until it tells you things you didn’t know about the real system

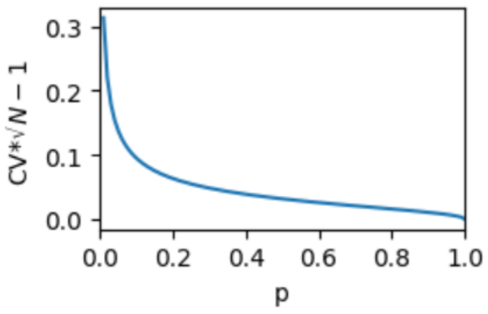
Modeling flowchart

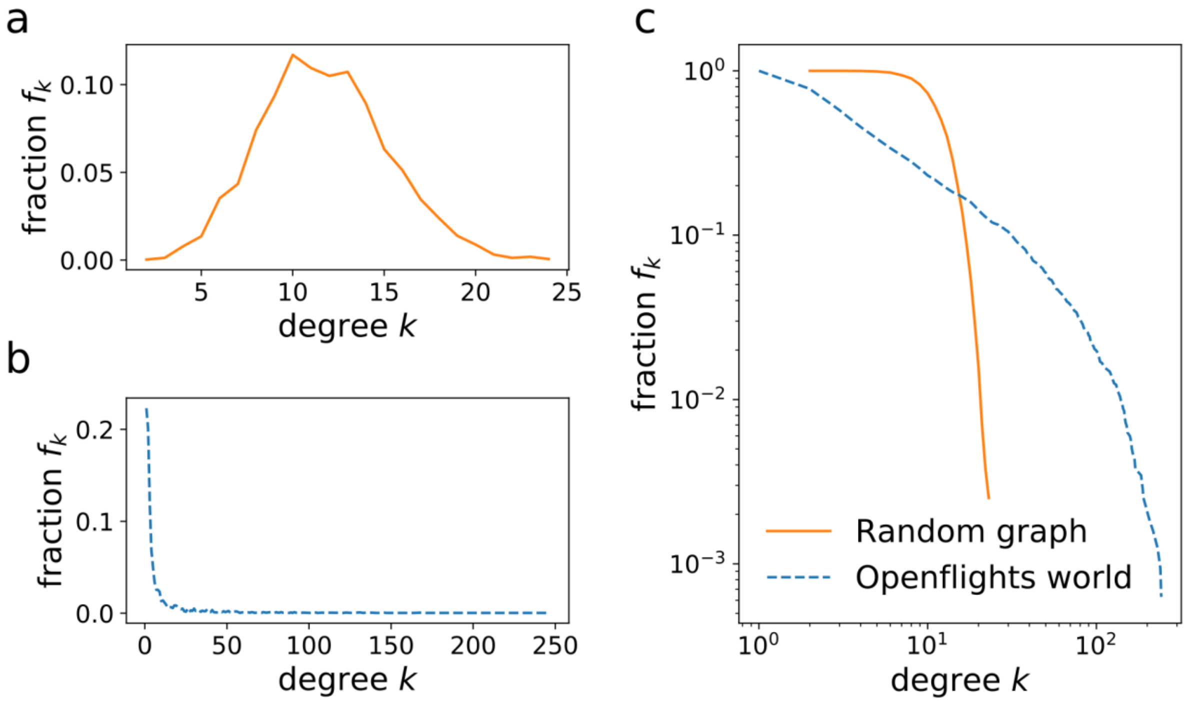


5.1 RANDOM NETWORKS

- What is the simplest, most assumption-free network we can create?
- Useful as a comparison baseline
- **Erdős-Rényi** (ER) network
  - Each possible link exists with probability p
  - Average degree $\langle k \rangle = (N-1)p$
  - Binomial degree distribution
    - $p_k \sim \binom{N-1}{k} p^k (1-p)^{N-1-k}$
    - $CV = \frac{\sqrt{(N-1)p(1-p)}}{(N-1)p} = \sqrt{\frac{1-p}{(N-1)p}} = \sqrt{\frac{1}{\langle k \rangle} - \frac{1}{N-1}}$
    - **Random networks do not have heavy tails**

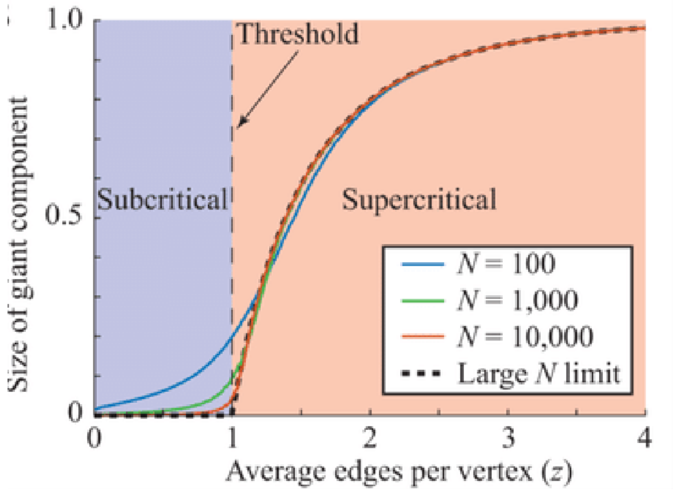


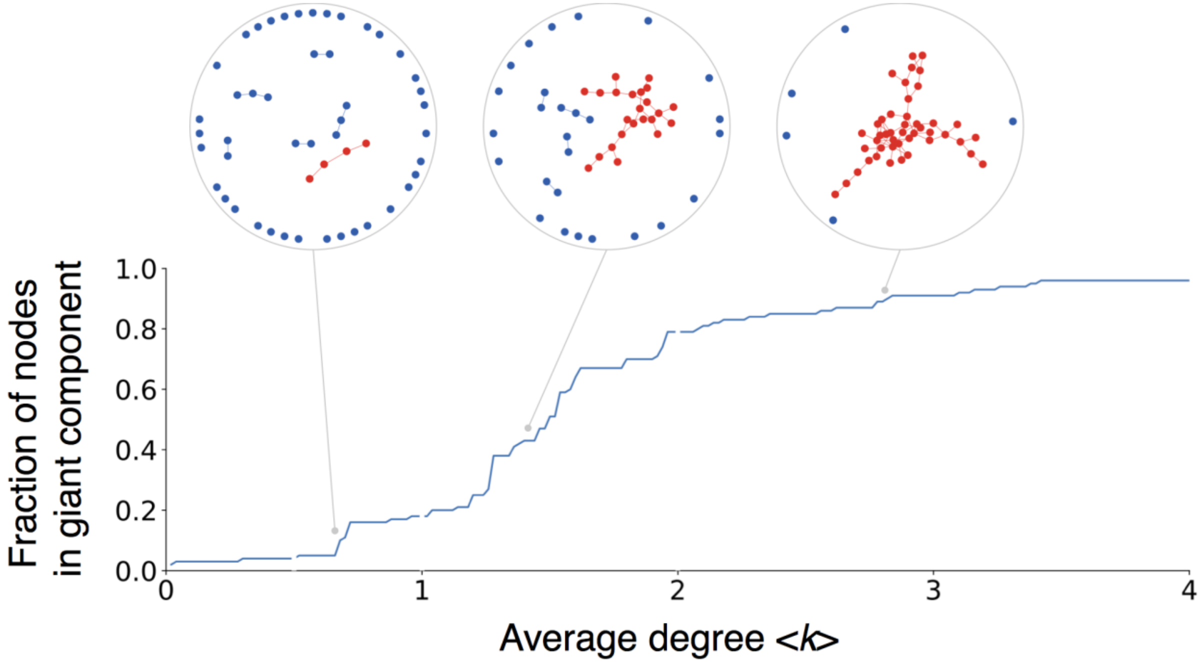




5.1 RANDOM NETWORKS

- Connectedness:
  - As $\langle k \rangle = (N-1)p$ increases, so does the size of the giant component
  - This is not smooth, and it is a phase transition





5.1 RANDOM NETWORKS

- Path length
  - (expected) max path length:
    - Assume $k_i = k$
    - Nodes 2 steps from i: $k(k-1)$
    - Nodes l steps from i: $k(k-1)^{l-1} \approx k^l$
    - $k^{l_{max}} = N \therefore l_{max} = \frac{\ln N}{\ln k}$
    - In reality [1]: $l_{max} = \frac{\ln N}{\ln k} +$ smaller terms
  - Average path length [2]: $\langle l \rangle = \frac{\ln N - \gamma}{\ln\langle k \rangle} + \frac{1}{2}, \gamma \approx 0.577$
  - Example:
    - $N = 8\ billion, \langle k \rangle = 150$
    - $\langle l \rangle = 4.93, l_{max} = 4.1 +$ (terms) ≈ 6
  - **Random networks have short path lengths**

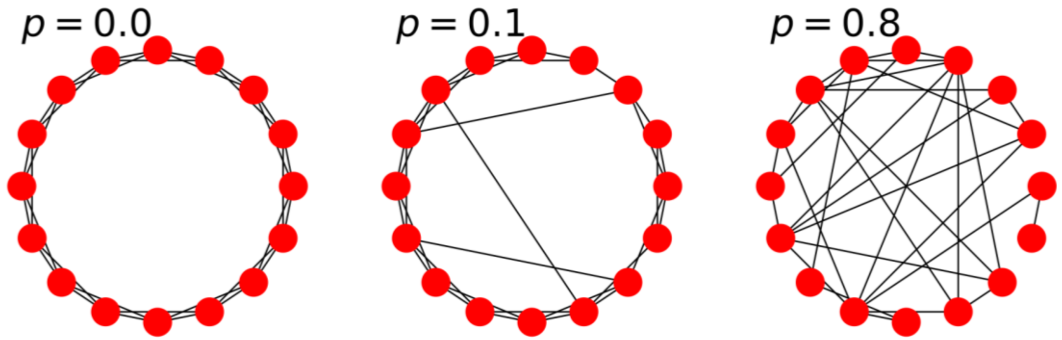
- Clustering coefficient
  - $p_{triangle} = p^3$
  - $C = C_i = p^3/p^2 = p$
  - In real-world networks p would be very small
    - $\langle k \rangle = (N-1)p$
  - **Random networks have very low clustering coefficient**
- Summary
  - Short paths
  - Low clustering
  - No heavy-tail
  - No hubs

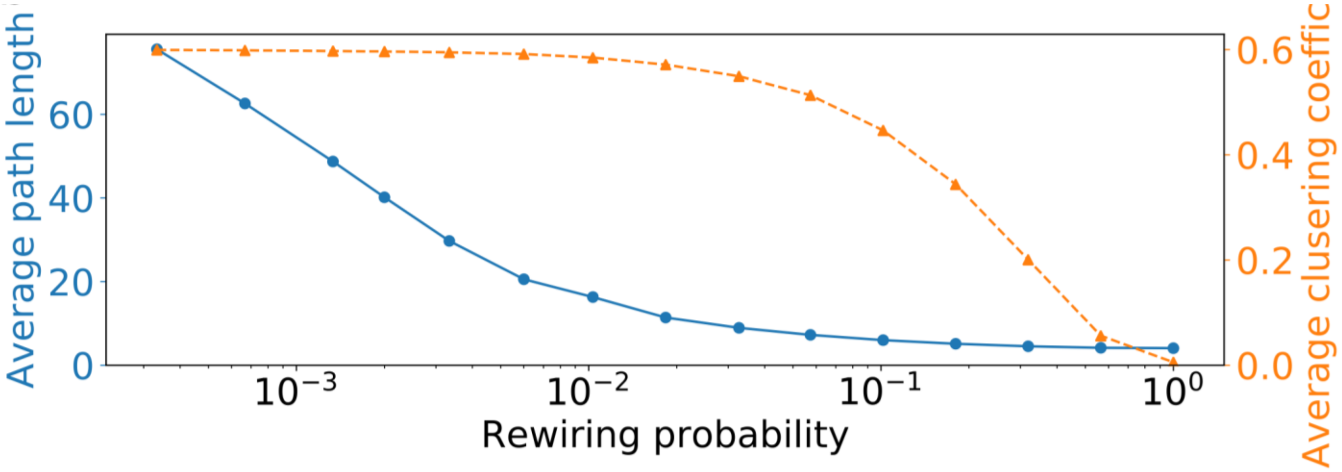
```
#small networks
G = nx.erdos_renyi_graph(N, p)

#large & sparse networks
G = nx.fast_gnp_random_graph(N, p)
```

[1] Riordan, O. & Wormald, Probability and Computing, 19(5-6), 835–926 (2010). https://doi.org/10.1017/s0963548310000325
[2] Fronczak et al, Physical Review E 70, 056110 (2014): https://doi.org/10.1103/PhysRevE.70.056110

5.2 SMALL WORLDS

- Can we get a minimal model with high clustering?
- **Watts-Strogatz** (WS) model
  1. Start from a high clustering ring network
  2. Rewrite nodes with probability p
- What are the new properties?
  - Short paths
  - High clustering
  - Nodes have mostly the same degree
    - No heavy-tails
    - No hubs



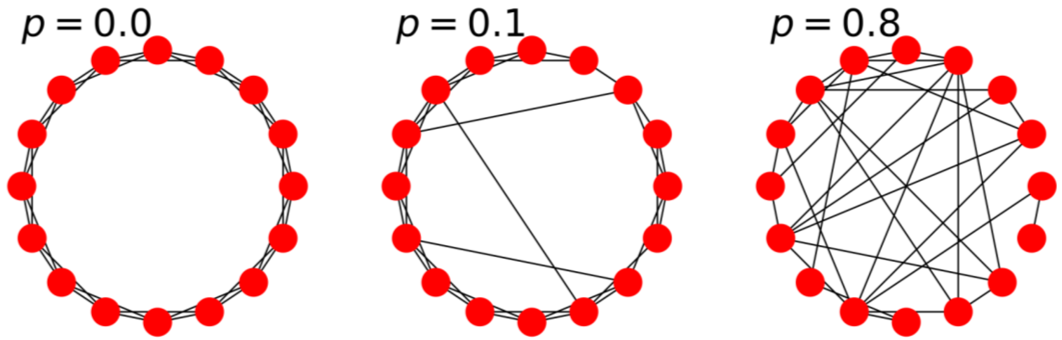


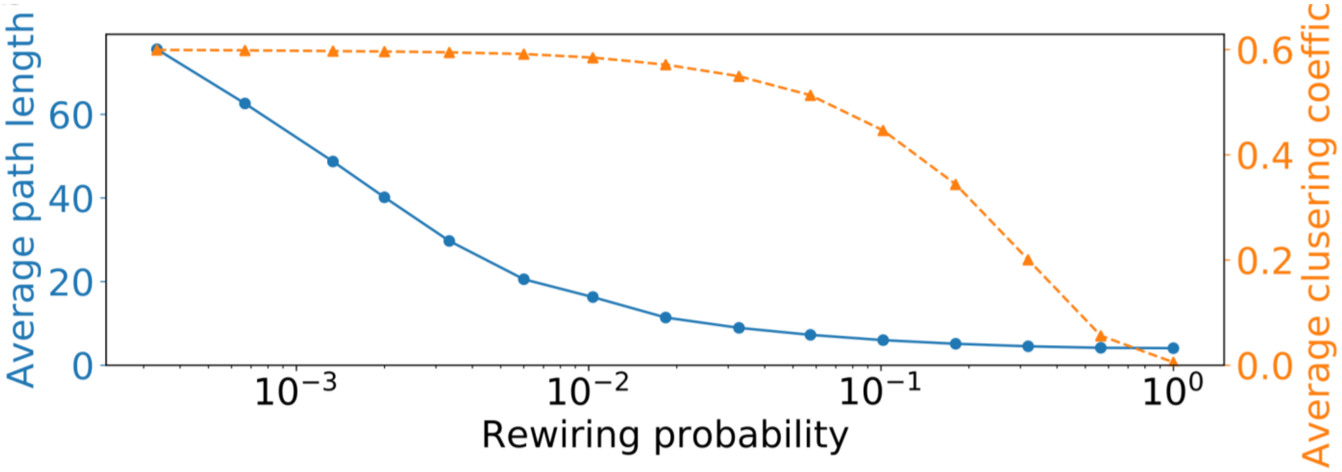
Watts, D., Strogatz, S. Nature 393, 440–442 (1998). https://doi.org/10.1038/30918 .

5.2 SMALL WORLDS

- Path length [1] $\langle l \rangle = (2k^2p)^{-1} \log 2Nkp$
  - Valid for $Nkp \gg 1$
- Clustering: $C(p) = \frac{3(K-2)}{4(K-1)}(1-p)^3$
- Small-world definition: $\langle l \rangle \sim \log N$
  - ER network: small-world with low clustering
  - WS network: small-world with high clustering

```
# Standard version
G = nx.watts_strogatz_graph(n,k,p)
# Only adds links
G = nx.newman_watts_strogatz_graph(n,k,p)
# Ensures connected network
G = nx.connected_watts_strogatz_graph(n,k,p)
```



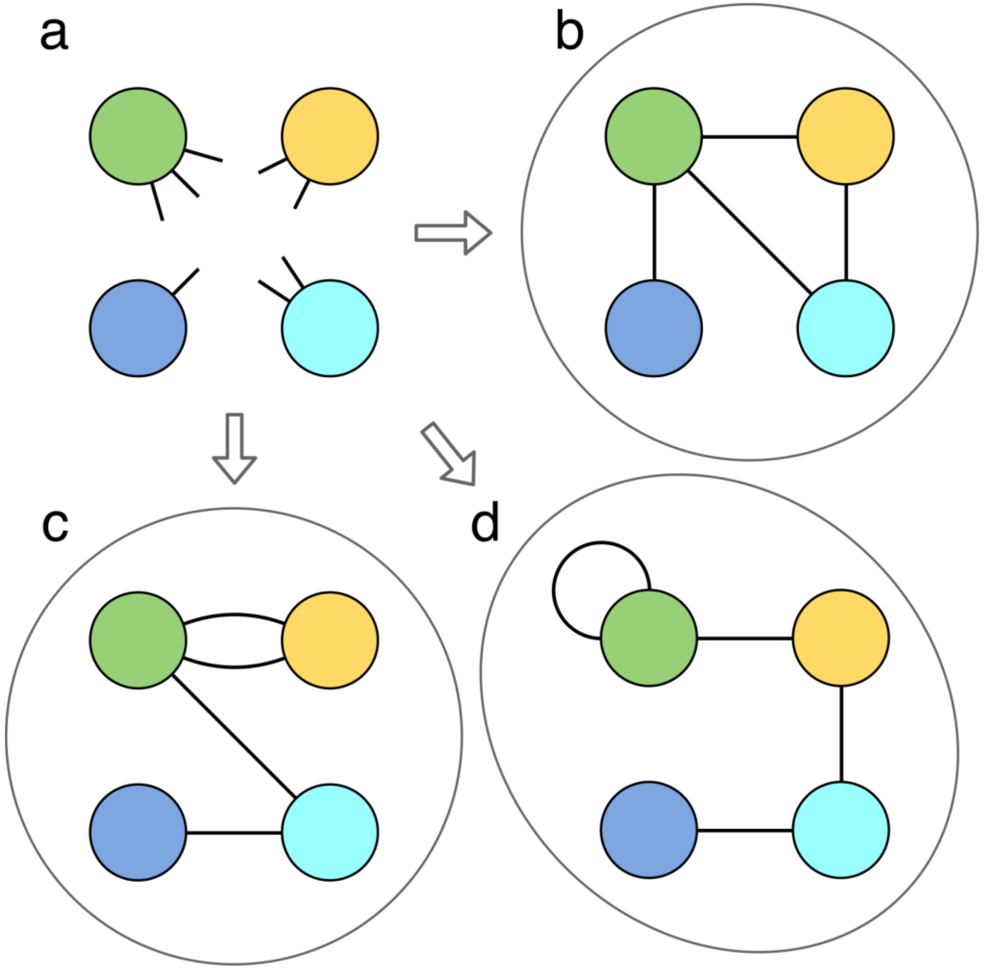


[1] Newman, M. E. J. et al. Physical Review Letters, 84(14), 3201–3204 (2000) http://doi.org/10.1103/physrevlett.84.3201

5.3 CONFIGURATION MODEL

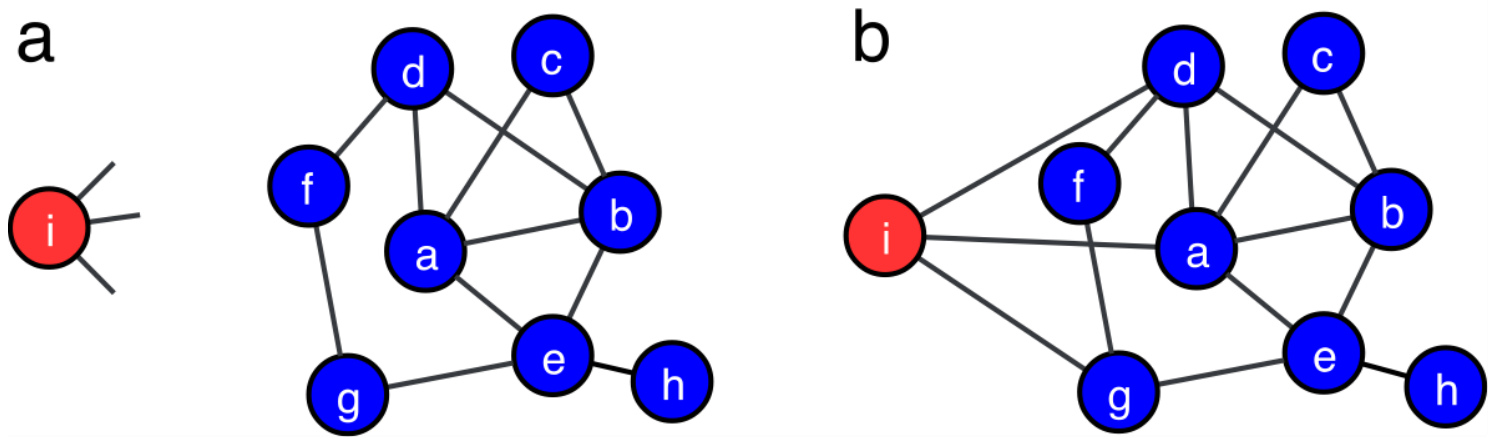
- What if we want an *exact* degree distribution?
- The idea: property of interest is encoded in the *degree*
- **Configuration model**
  - Nodes start with **stubs**
  - Stubs are randomly connected
- Operationalizing it
  1. Measure metrics of a network
  2. Measure the same metrics of many configuration models with the same degree sequence
  3. If the metrics are the same, the properties are encoded in the degree sequence. If not, something else is needed

```
G = nx.configuration_model(degree_sequence)
```



5.4 PREFERENTIAL ATTACHMENT

- Networks grow
  - New node comes with some stubs
  - Connects to other nodes following some rule
- **The idea: nodes prefer to connect to well-connected nodes**
  - Node *fitness* is based on degree

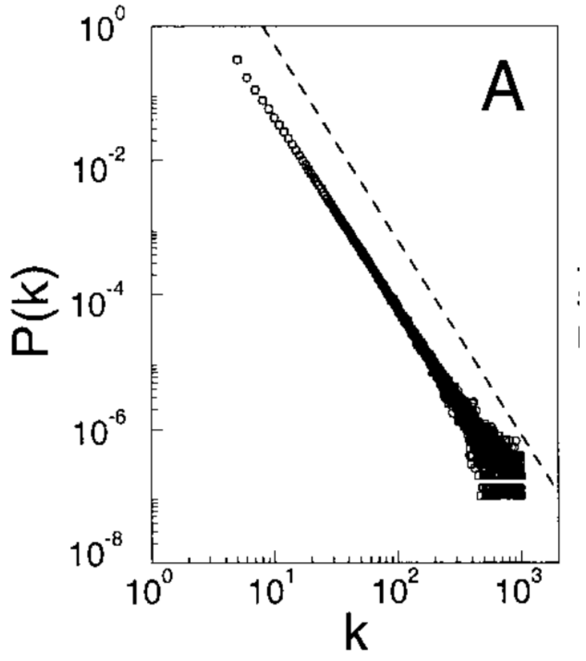


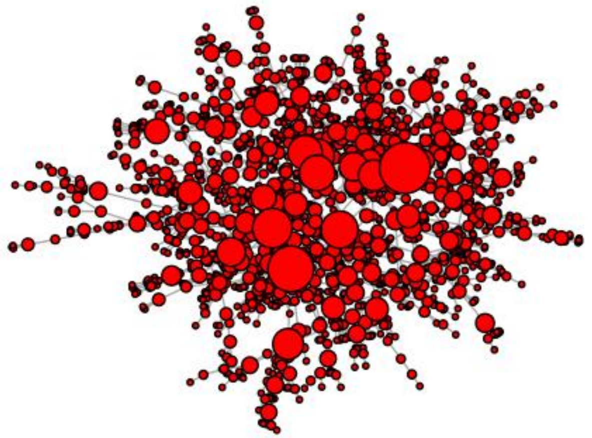
5.4 PREFERENTIAL ATTACHMENT

- **Barabasi-Albert** (BA) network
  1. Starts from a (usually fully) connected *core*
  2. Adds new nodes one by one, with m links each
  3. The probability of connecting to j is $\sim k_j$: **preferential attachment (PA)**

$$\Pi(i \leftrightarrow j) = \frac{k_j}{\sum_l k_l}$$

- Results in a power-law degree-distribution
  - $p_k = 2m^2/k^3$





```
G = nx.barabasi_albert_graph(N,m)
```

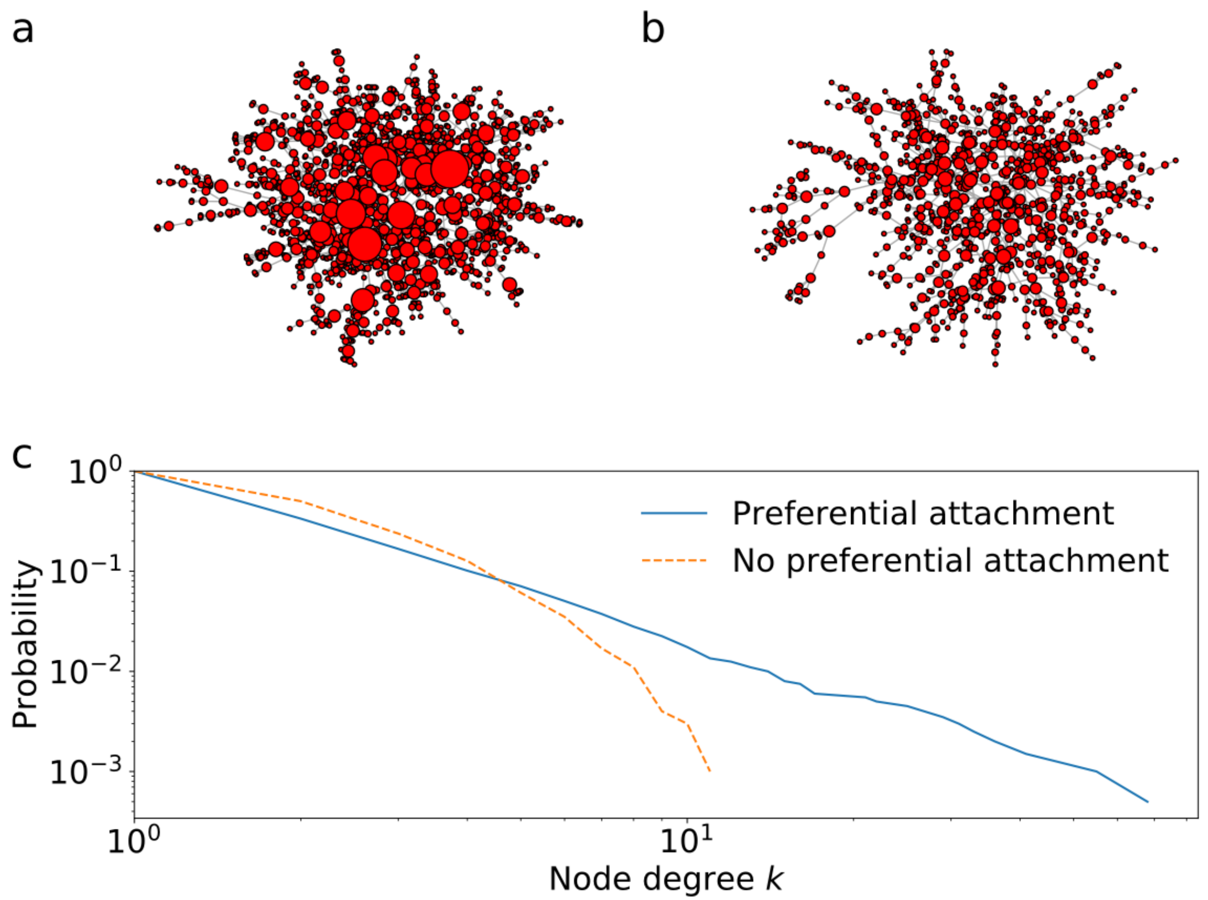
A. L. Barabási and R. Albert. Science 286, pp 509-512 (1999) https://doi.org/10.1126/science.286.5439.509

5.4 PREFERENTIAL ATTACHMENT

- Ingredients: growth and linear PA
  - Are both necessary?

- Network with only growth
  - Nodes are added over time, but links are random
  - Maybe just "time in the game" is enough

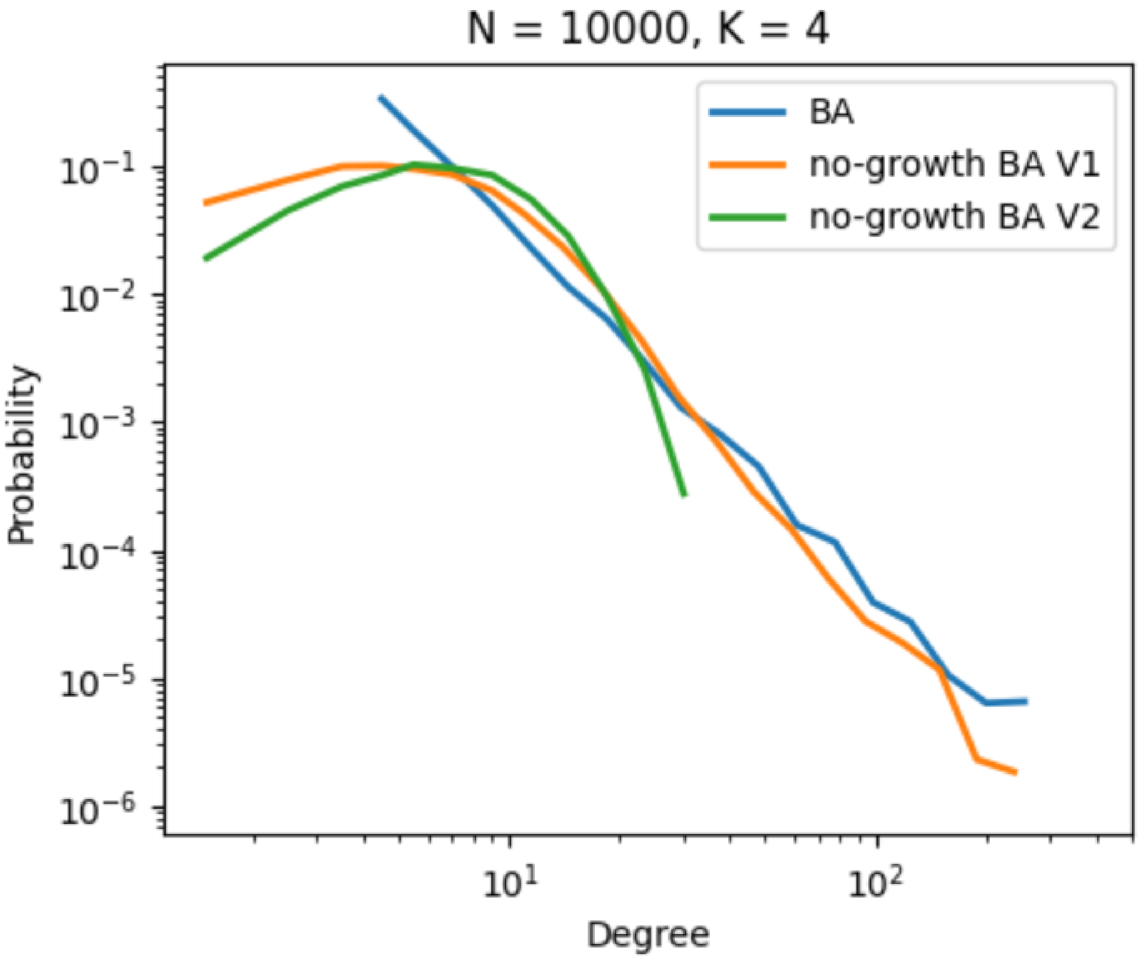
- Result
  - Exponential degree distribution: $p_k \sim e^{-\beta k}$
  - Preferential attachment is needed



5.4 PREFERENTIAL ATTACHMENT

- Network with only PA
  1. All nodes from the start
  2. Randomly select a node
  3. Link it to another with PA
     - V1: $\Pi \sim k_i$
     - V2: $\Pi \sim k_i + 1$
  4. Run until network as NK links

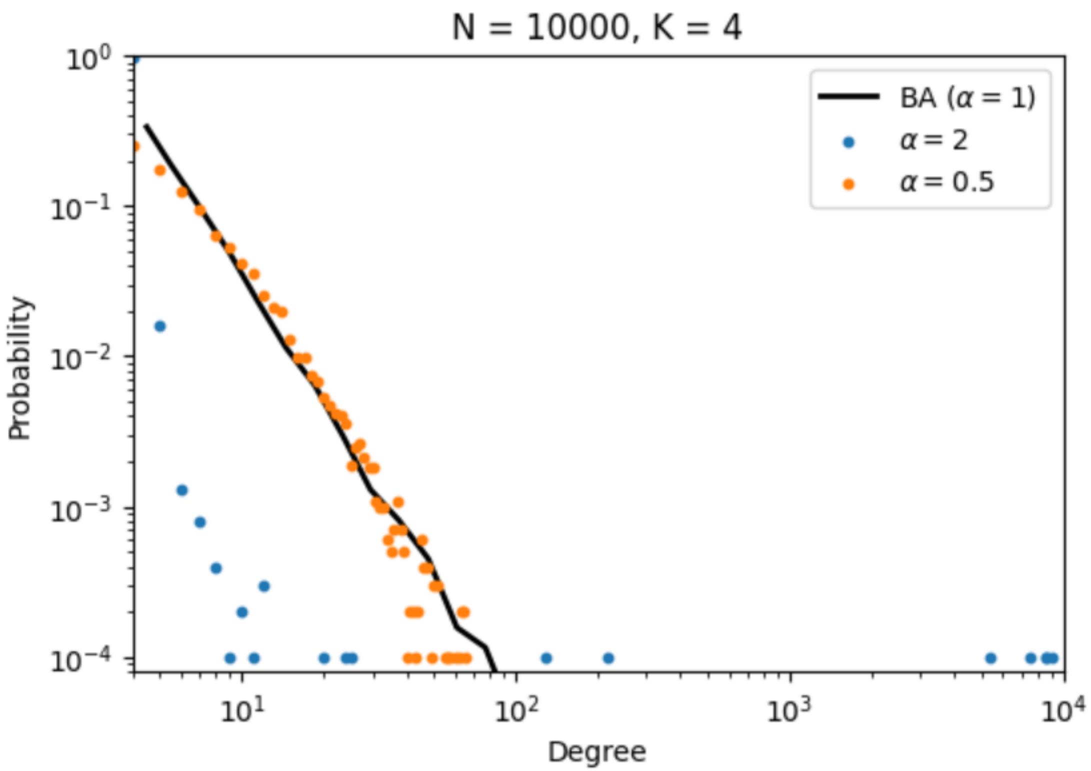
- Results
  - Power-law *tail* for V1, no power-law for V2
  - PA requires enough *imbalance* (temporal, fitness, etc)
  - Needs to justify stopping after NK links



5.5 OTHER PREFERENTIAL MODELS

- BA has linear preferential attachment ($\Pi \sim k_i$)
  - What if $\Pi \sim k_i^{\alpha}$?
- If $\alpha > 1$
  - Hubs hyper-concentrate the links
  - Hub-and-spoke structure
- If $\alpha < 1$
  - Effect of hubs is weakened
  - heavy-tail disappears as $\alpha \to 0$

- A natural model for heavy-tails includes
  - Growth
  - **Linear** preferential attachment



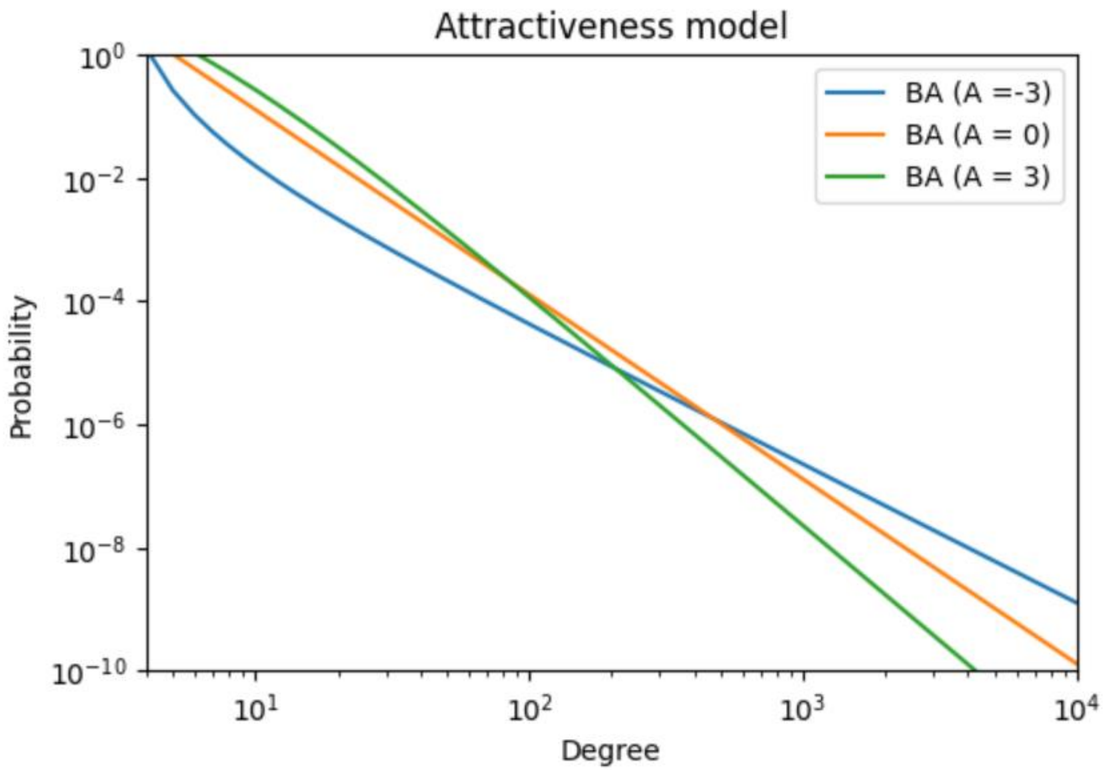
5.5 OTHER PREFERENTIAL MODELS

- BA model limitations
  - Fixed degree distribution $p_k \sim k^{-3}$
  - Hubs are always the oldest nodes (older gets richer)
  - Low clustering coefficients
  - No node deletion
  - Necessarily connected
  - Requires linear PA

- Plenty of modifications
  - Attractiveness model
  - Fitness model
  - Random Walk model
  - Copy model
  - Rank model

5.5 OTHER PREFERENTIAL MODELS

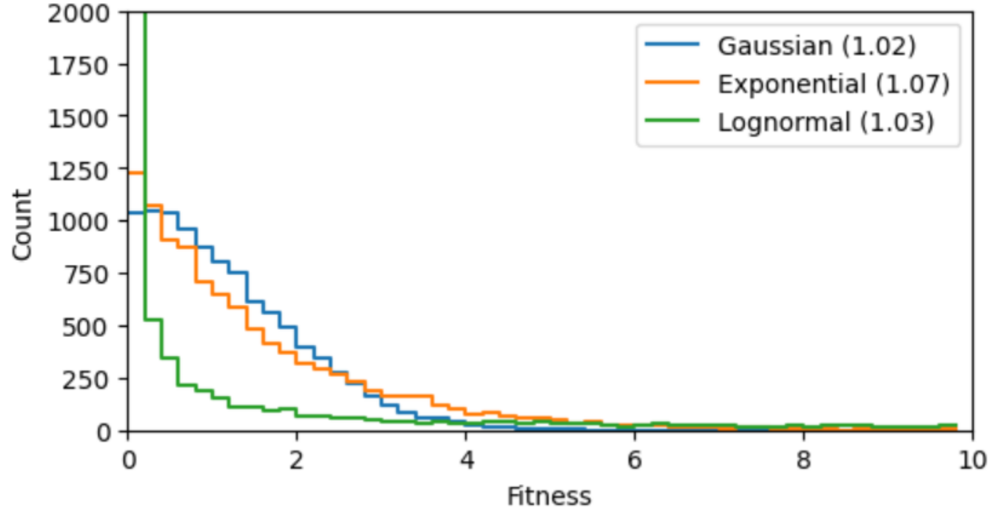
- **Attractiveness model**
  - Nodes have a baseline attractiveness A
  - $\Pi = \frac{A+k_j}{\sum_l (A+k_l)}$
  - Degree distribution [1]
    - $p_k \sim k^{-3-A/m}$
    - Valid for $A > -m$
    - **Exponents** $-2 < \alpha < -3$
  - Larger A → larger α
  - *Less statistical weight* in the tail

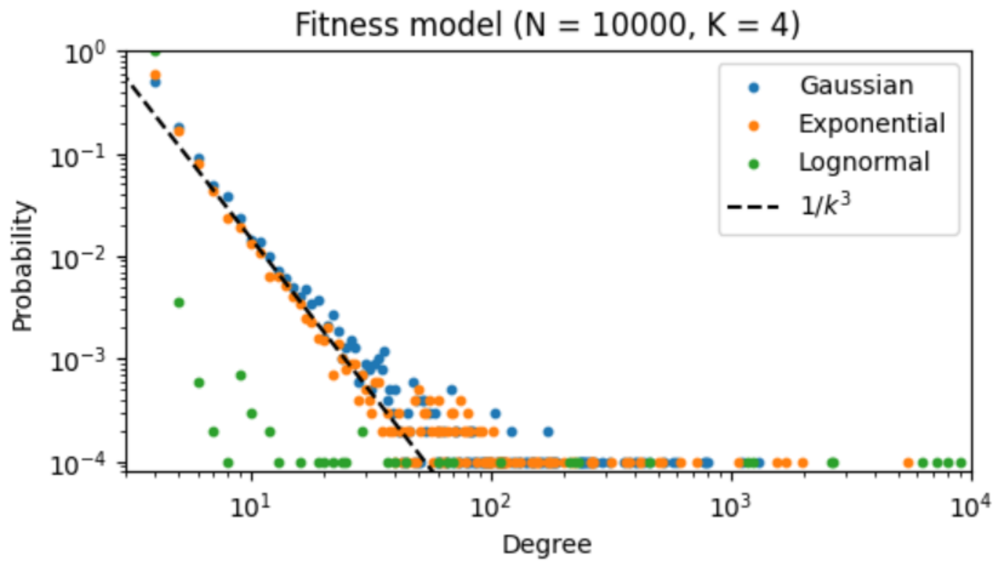


[1] Onody, R. N., & de Castro, P.A. Physica A, 336(3-4), 491–502 (2004). http://doi.org/10.1016/j.physa.2004.01.027

5.5 OTHER PREFERENTIAL MODELS

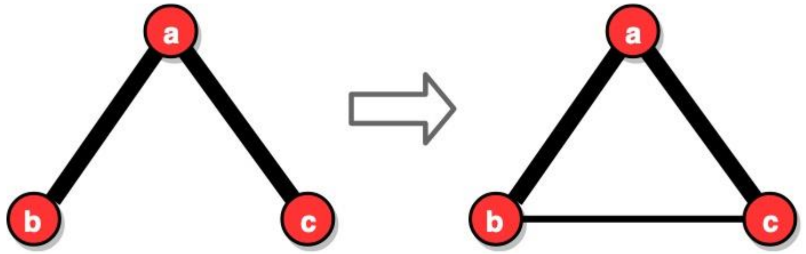
- **Fitness model**
  - Each node has an individual attractiveness (fitness) η
  - $\Pi = \eta_j k_k / \sum_l \eta_l k_l$
  - Drawn for $\eta \sim \rho(\eta)$
  - Distribution will be heavy-tailed if $\max(\rho)$ is finite
  - Late nodes can still attract links
- Comparing distributions
  - Gaussian, Exponential, Lognormal
  - Same median $\mathrm{med}\{X\} = 1$
  - Lognormal: heavy tail creates hyper-connected hubs, destroys distribution



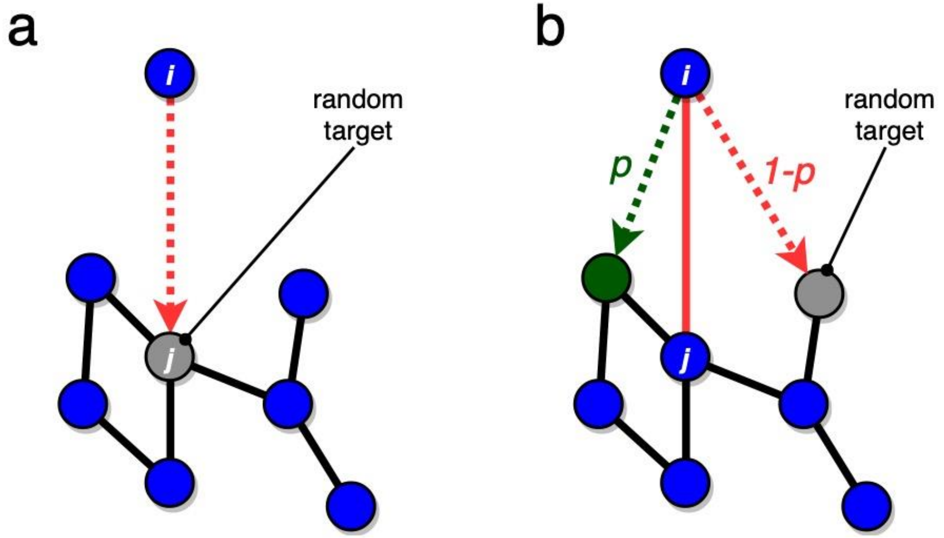


5.5 OTHER PREFERENTIAL MODELS

- BA networks have low clustering
- **Triadic closure:**
  - neighbours of a node are more likely to be connected



- **Random Walk model**
  1. Start with (fully-connected) core
  2. Grow network adding a node with m links
     - The first of the m links chooses a random node j
     - The other $m-1$ links choose a neighbour of j with probability p, and a random node with probability $1-p$
  - Can produce heavy-tailed distributions for large-enough p
    - Implicit preferential attachment from hubs being well-connected
  - Can produce both high clustering and community structure

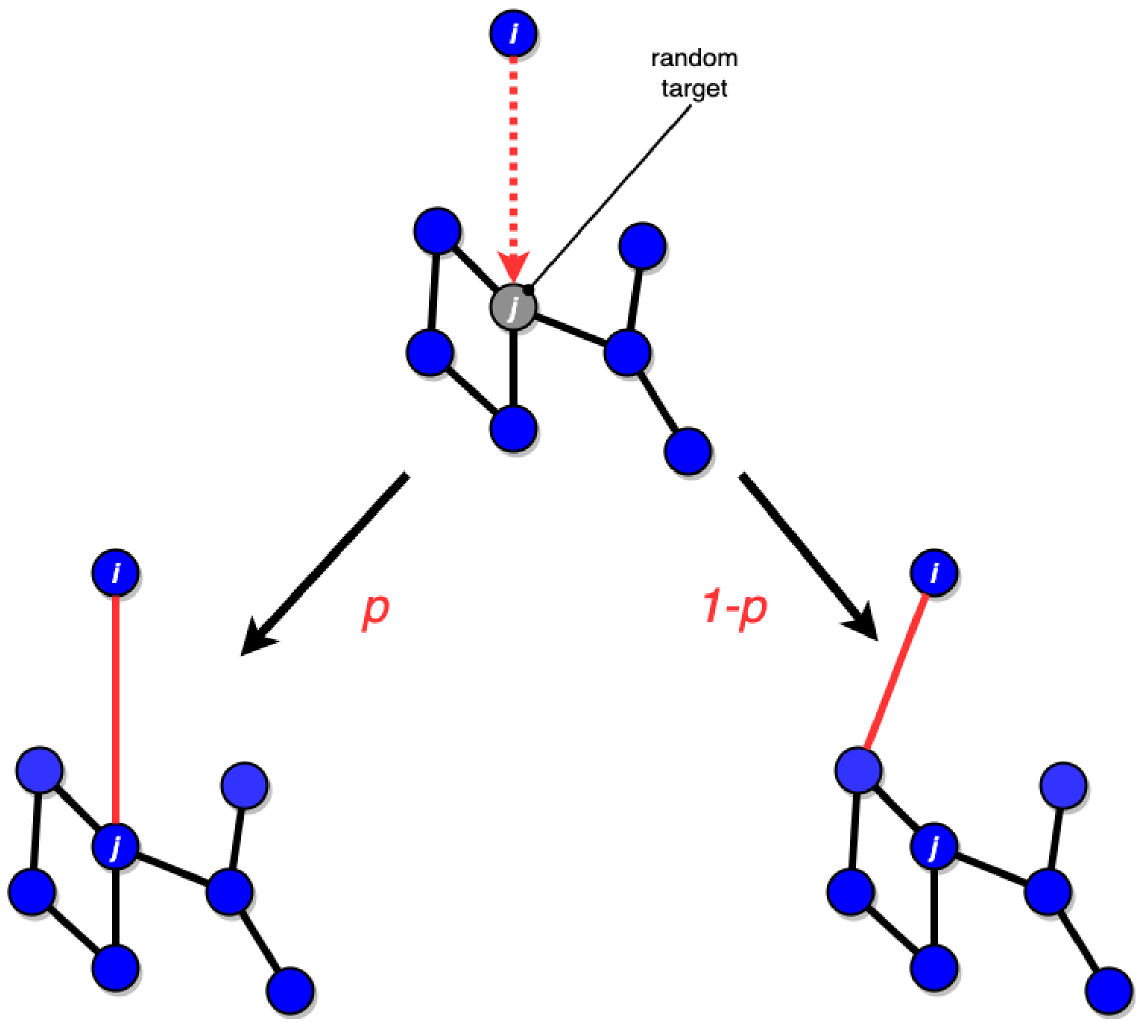


5.5 OTHER PREFERENTIAL MODELS

- **Copy model**
  1. Add a new node with m links
  2. Select a target j
  3. Either
     - link to j with probability p
     - Link to a *neighbour* of j with $1 - p$

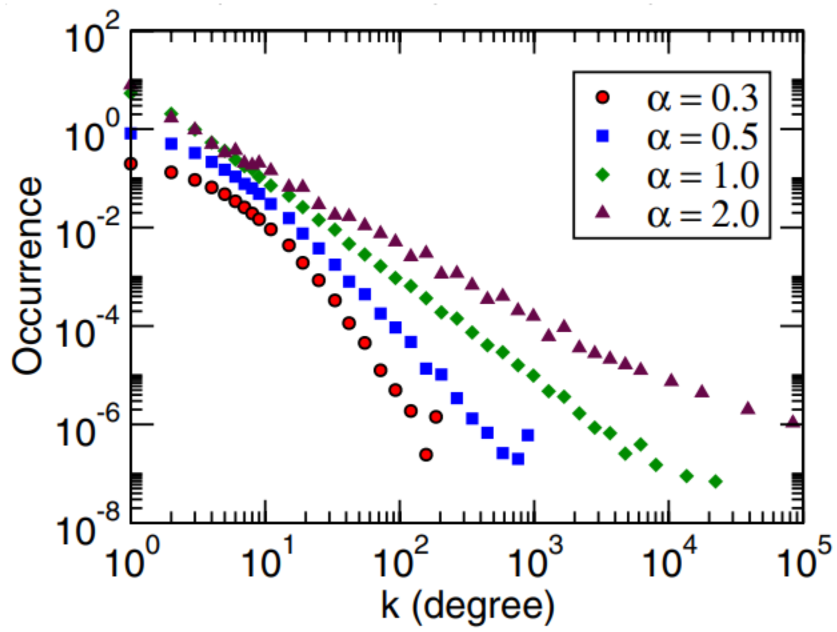
- In many processes, new nodes copy the connections of older ones
  - Citation networks, websites, etc

- Properties:
  - **No triadic closure**
  - Hubs



5.5 OTHER PREFERENTIAL MODELS

- BA model requires an absolute metric of worth (degree)
- Perceived value can be based on the relative *ranking* instead
- **Rank model**
  1. Nodes receive ranks $R = 1,2, \ldots$
  2. Add a node with m links
  3. Connection probability $\Pi \sim R_j^{-\alpha}$
  4. If R depends on e.g. degree, re-rank nodes



  - Robustly creates heavy-tailed distributions
  - For ranking by time: $p_k \sim 1/k^{1+1/\alpha}$

Fortunato, S. et al. Physical Review Letters, 96(21), 218701 (2006). http://doi.org/10.1103/PhysRevLett.96.218701

SUMMARY

- Many models focusing on emulating certain properties
  - Degree distribution, clustering, triadic closure, etc
- No single "best model"
- Preferential attachment is a key mechanism
  - Can create heavy-tailed distributions
  - If unbalanced, can create hyper-concentrated hubs
- Variations of PA models can create a variety of degree distributions