lorenzo.dallamico@unito.it

di.unito.it

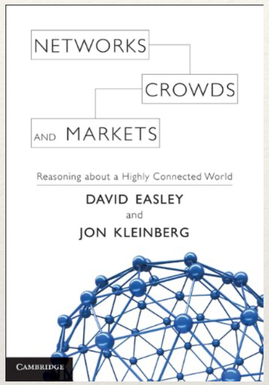
Lecture 10.ns07

Power Laws, Rich-Get-Richer Phenomena and preferential attachment

Course: **Complex Networks Analysis and Visualization**
Sub-Module: **NetSci**

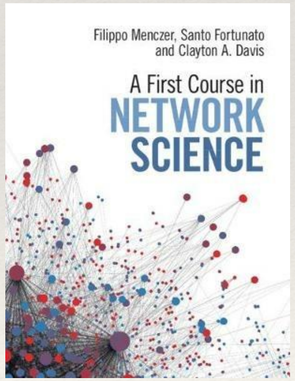
-

References



[ns2] Chapter 18 (18.1 - 18.6)

"Power Laws and Rich-Get-Richer Phenomena" —-> https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch18.pdf



Chapter 5: Network Models

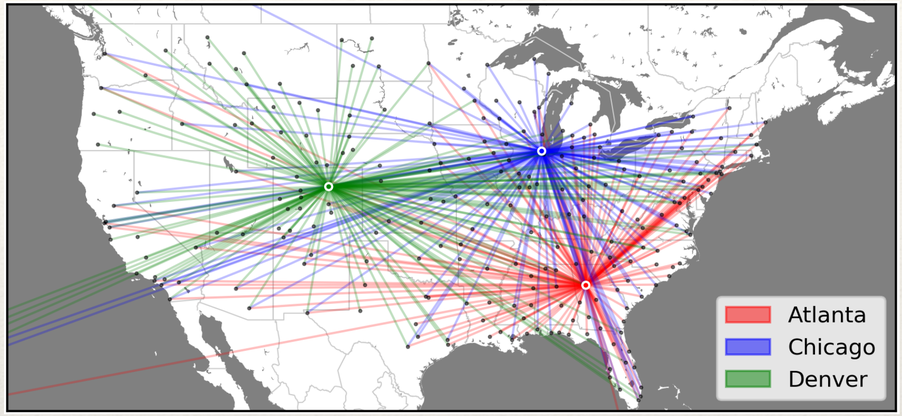
Please, check your general understanding with exercises at the end of the chapter!

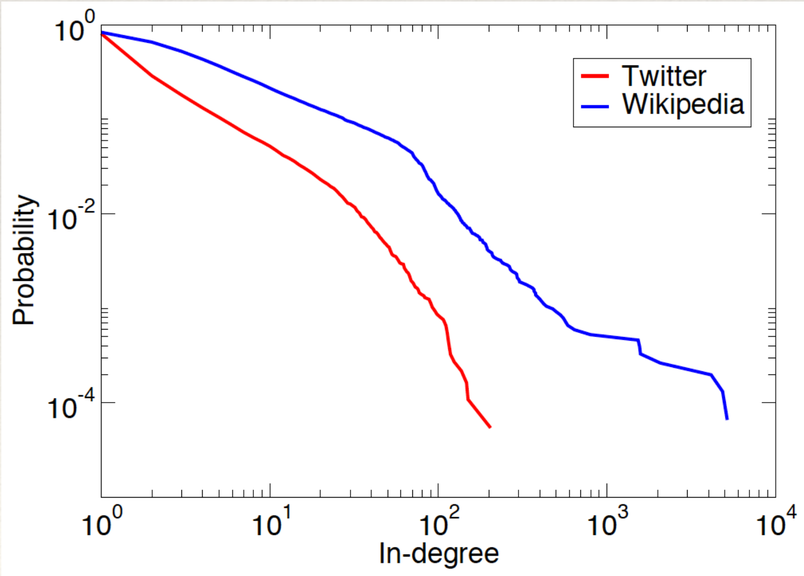
Popularity as a Network Phenomenon

Popularity, heterogeneity and networks

- recap:
  - real networks are heterogeneous
  - in-links as a measure of popularity
  - the heterogeneity parameter as a measure of distribution's broadness

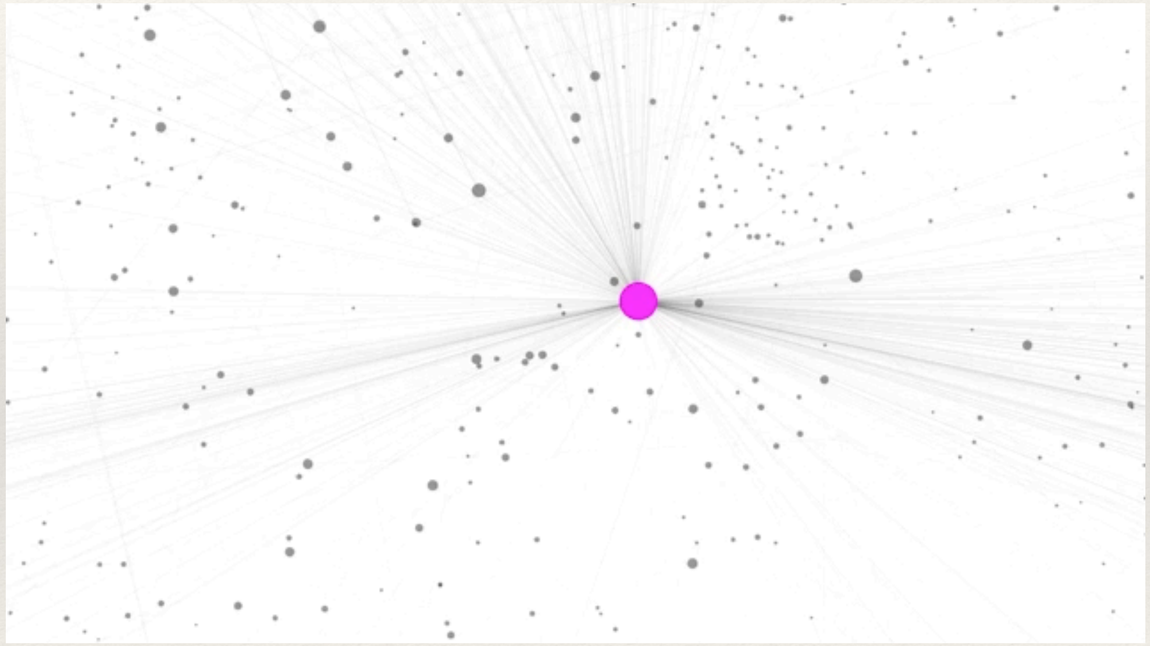
$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$$





Case study: the Web

- Characterizing popularity reveals imbalances (inequalities)
  - almost everyone is popular for very few people
  - very few people achieve high popularity
  - very very few people achieve global popularity
- Why? Is this phenomenon intrinsic to the whole idea of popularity itself?

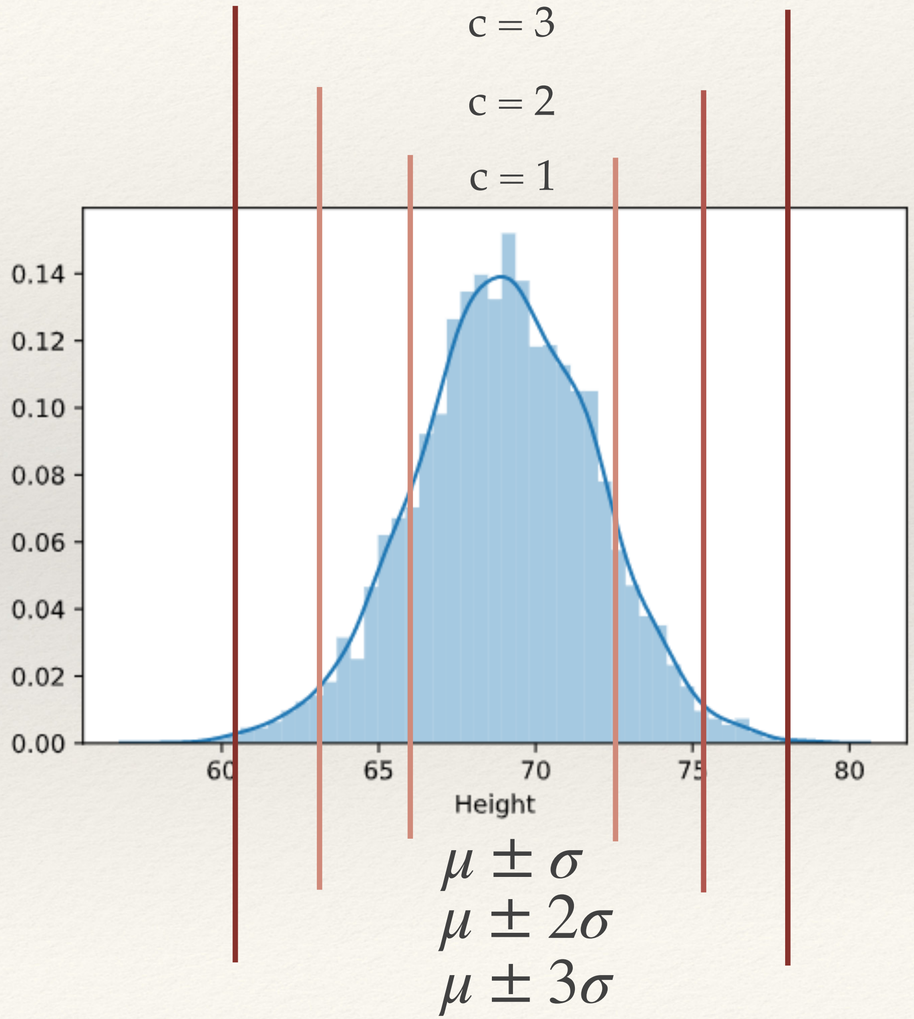


Looking for a popularity scale

- "As a function of *k*, what fraction of (web) pages have *k* links?"
- larger *k* corresponds to greater popularity
- First (and simple) hypothesis: **normal distribution**
- the mean defines a *scale* of the population: good for estimation/prediction

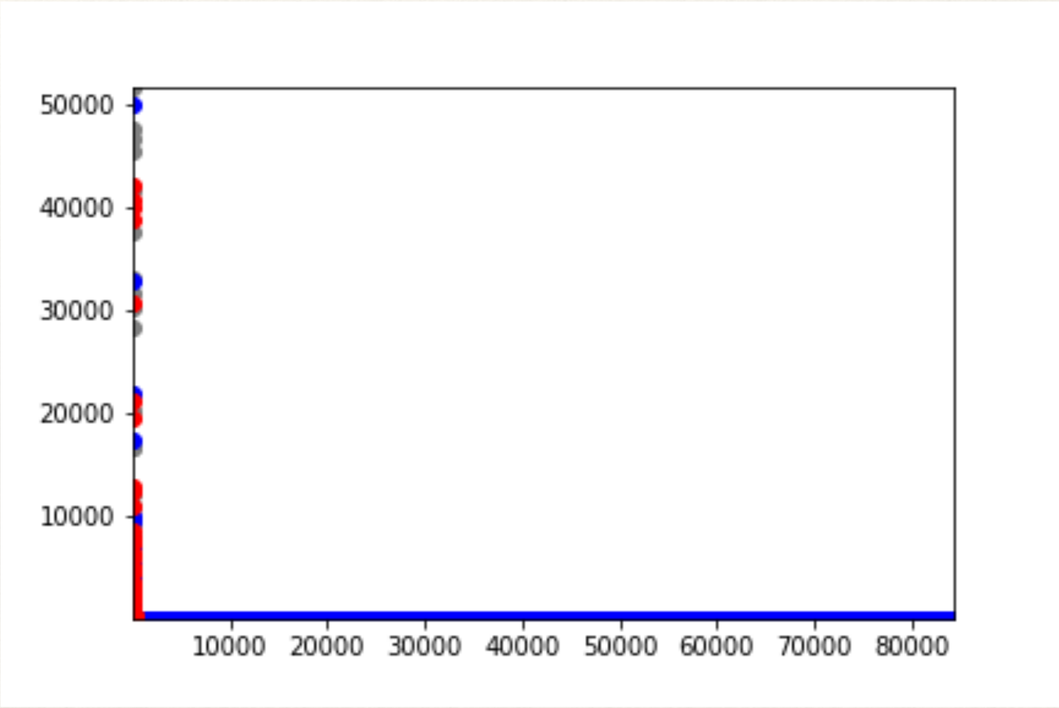
Ex. heights

- If we look at people's heights distributions
- mean: $\mu = 69.03$ (feet)
- std: $\sigma = 2.86$
- The prob. of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially in c
  - amazingly high persons are very unlikely



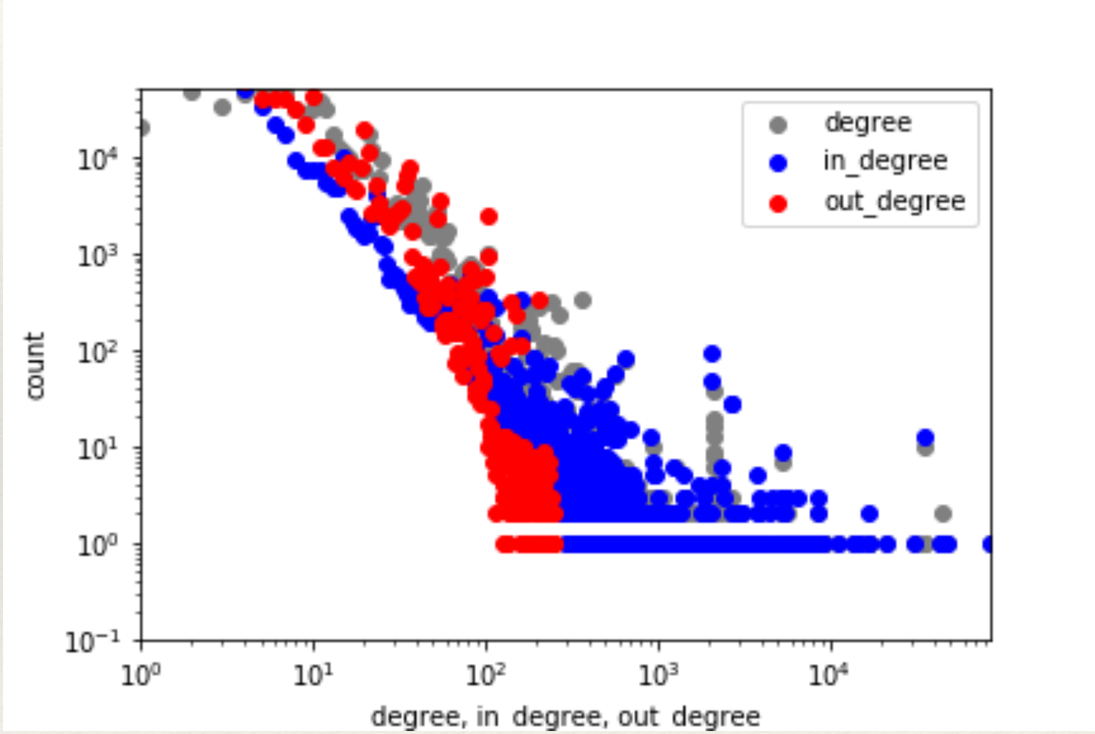
Ex. the Web (a sample)

- source dataset: https://snap.stanford.edu/data/web-BerkStan.html
- the "Berkeley-Stanford web graph"
  - Nodes: 68,5230, Edges: 7,600,595
- If we plot degree, in-degree, out-degree distribution, we find a different picture



log-log scale

- It turns out that the best way to plot heavy tailed distributions is to use log-log scales
- $\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2} = \frac{81782.51}{22.18^2} = 166.18$
- standard deviation: $\sigma = \sqrt{\langle k^2 \rangle} = 285.98$
- the degree of a randomly chosen node is $22.18 - 285.98 \leq k_{in} \leq 22.18 + 285.98$
  - not very informative…

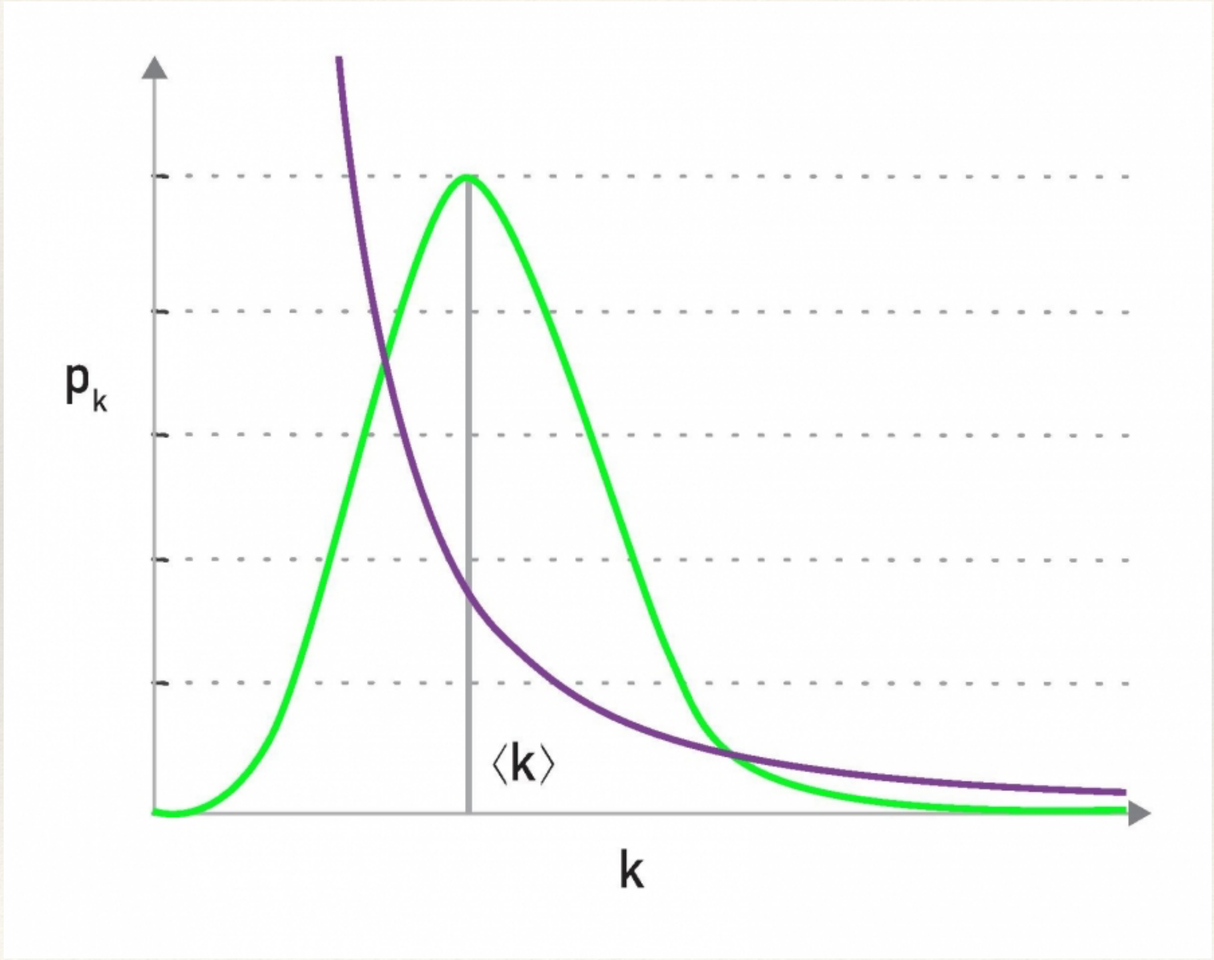


Power Laws

Empirical findings

- The fraction of web page that have k in-links is: $f(k) \approx \frac{1}{k^c} = k^{-c}$
  - c = 2.1
  - in other networks $2 < c < 3$ very often
- $f(k) = ak^{-c}$: **power law** distribution
- you can calculate that in the $2 < c < 3$ regime we have that when $N \to \infty$ then $\langle k^2 \rangle \to \infty$
  - **scale free networks**

The lack of a scale



- power laws in the scale free regime is an unbounded distribution when $N \to \infty$
- variance is infinite, so it is standard deviation
- main consequence: popular web pages are more common than we would expect with a normal distribution

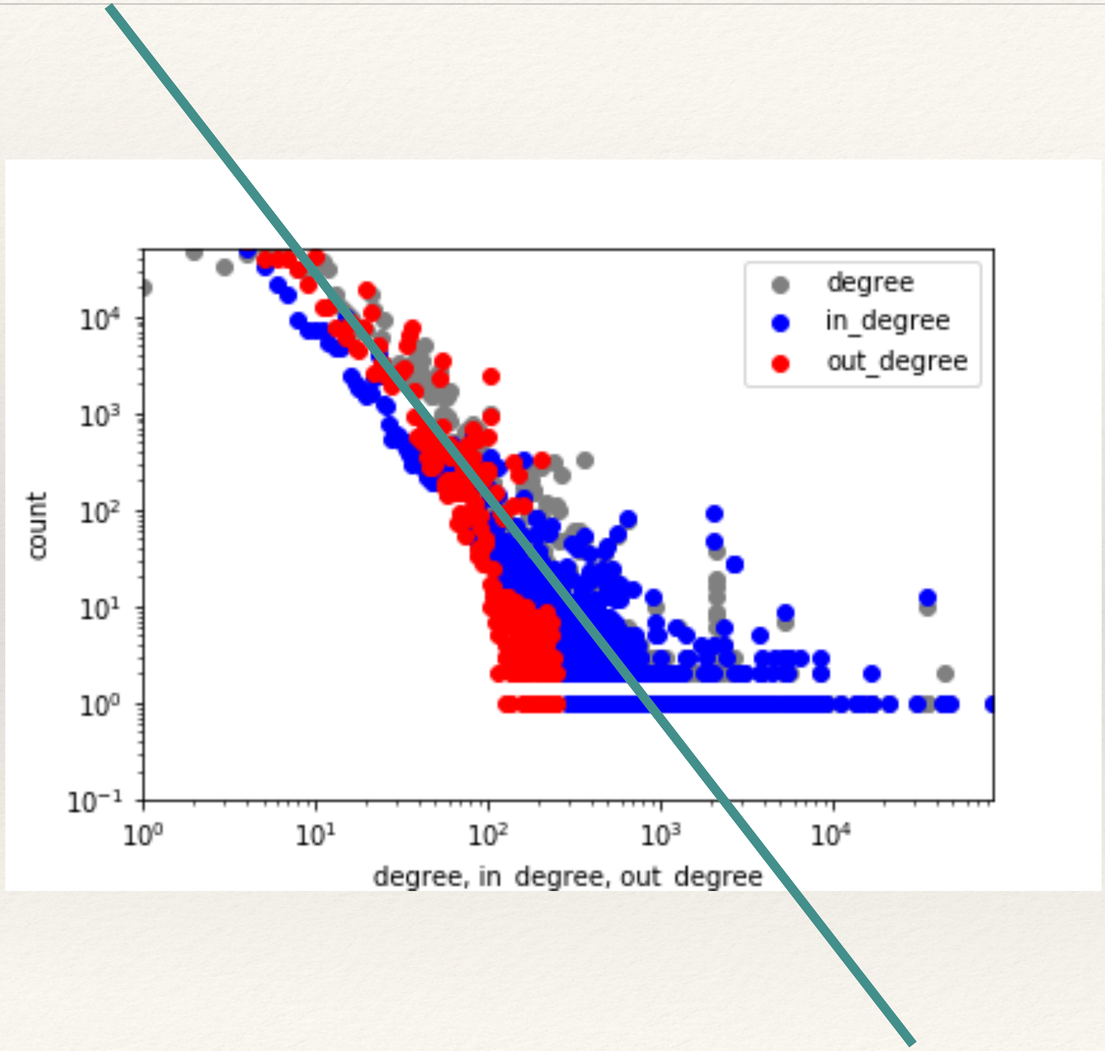
Disclaimer

- There is a long standing debate in the scientific communities if scale free networks are rare or frequent
- It depends on the severity of the test
- For many practical reasons: we do not care if the network is a *real* scale free network
  - checking heterogeneity is often enough to distrust the average degree as a good estimator (and all the other consequences: friendship paradox, robustness, and so on)
- however, understanding some characteristic of power laws is useful - as much as looking for an underlying process that explains the emergence of hubs!

power laws

- $f(k) \approx \frac{1}{k^c} = k^{-c}$ decreases much more slowly as k increases
  - pages with very large k are much more common than expected with the normal distribution
  - emergence of hubs is likely
- This can be observed empirically in many domains

Fitting with a straight line



- approximations of power laws are very common
- $f(k) = ak^{-c}$ for some constants a and c

$$\log f(k) = \log(ak^{-c})$$

$$\log f(k) = \log a - c \log k$$

$$y = \log a - cx$$

in a log-log plot:
$\log a$ is the **intercept**
$-c$ is the **slope**

Why do hubs emerge?

- Let's accept that power laws represent many phenomena
- Why?
- We are observing a kind of "order" emerging from chaos
- Is there an underlying process that keeps the line so straight?

Preferential models

Network growth

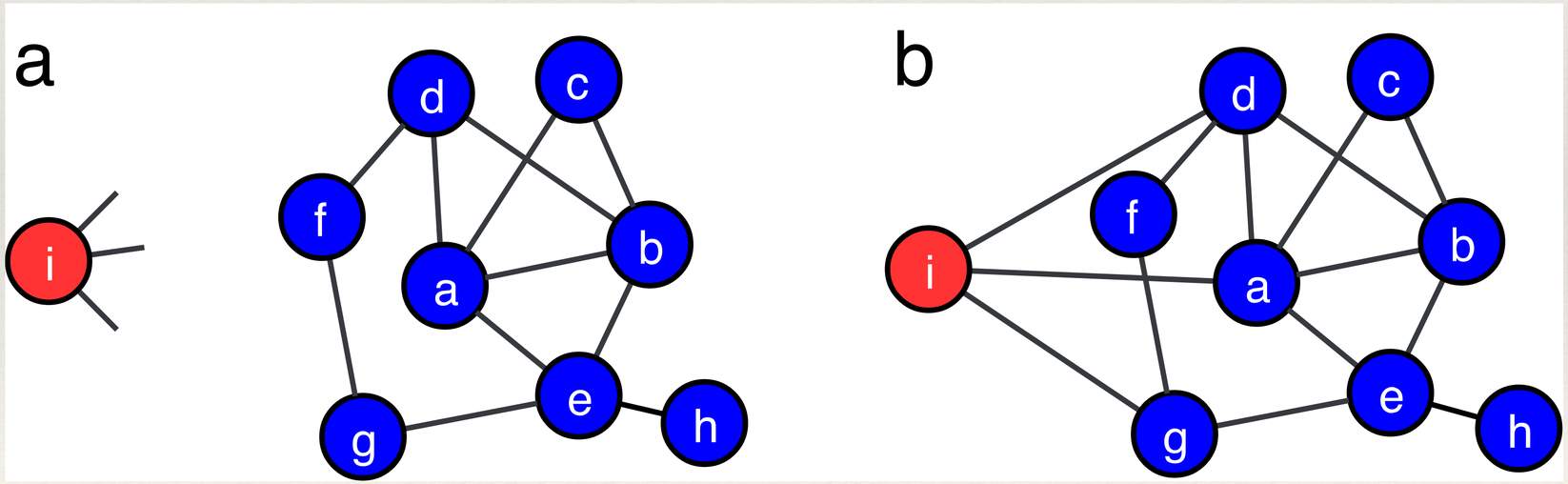
- **Note:** Real-world networks are growing
- **Examples:**
  - The Web in 1991 had a single node, today there are trillions
  - Citation networks of scientific articles and collaboration networks of scientists keep growing due to the publication of new papers
  - The collaboration network of actors keeps growing due to the release of new movies
  - The protein interaction network has been growing over the course of 4 billion years: from a few genes to over 20,000

Network growth

- **General procedure:**

1. A new node comes with a given number of stubs, indicating the number of future neighbors of the node (degree)

2. The stubs are attached to some of the old nodes, according to some rule



Preferential attachment

- **Note:** Nodes prefer to link to the more connected nodes
- **Examples:**
  - Our knowledge of the Web is biased towards popular pages, which are highly linked, so it is more likely that our website points to highly linked Web sites
  - Scientists are more familiar with highly cited papers (which are often the most important ones), so they will tend to cite them more often than poorly cited ones in their own papers
  - The more movies an actor makes, the more popular they get and the higher the chances of being cast in a new movie

Which model?

- Our network model should have the following features:
  - **Growth:** the number of nodes grows in time following the addition of new nodes. The models considered so far are **static**
  - **Preferential attachment:** new nodes tend to be connected to the more connected nodes. The models considered so far set links among pairs of random nodes, regardless of their degree

Polya's urn model

- **Start:** an urn contains X white and Y black balls
- **Process:** a ball is drawn from the urn and put back in with another ball of the same color
- **Example:** if we first pick a white ball, there will be X+1 white and Y black balls in the urn; **white will become more likely to be picked than black in the future**
- Preferential attachment used to **explain heavy-tail distributions of many quantities:** the number of species per genus of flowering plants, the number of (distinct) words in a text, the populations of cities, individual wealth, scientific production, citation statistics, firm size, etc.

The Barabási-Albert model

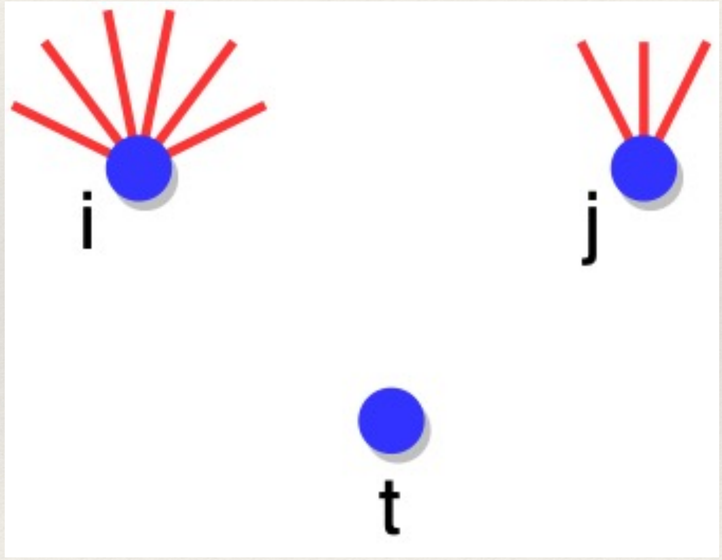
- **Procedure:**
  - Start with a group of m_0 nodes, usually fully connected (clique)
  - At each step a new node i is added to the system, and sets m links with some of the older nodes ($m \leq m_0$)
  - The probability that the new node i chooses an older node j as neighbor is **proportional to the degree k_j of j:**

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

  - The procedure ends when the given number N of nodes is reached

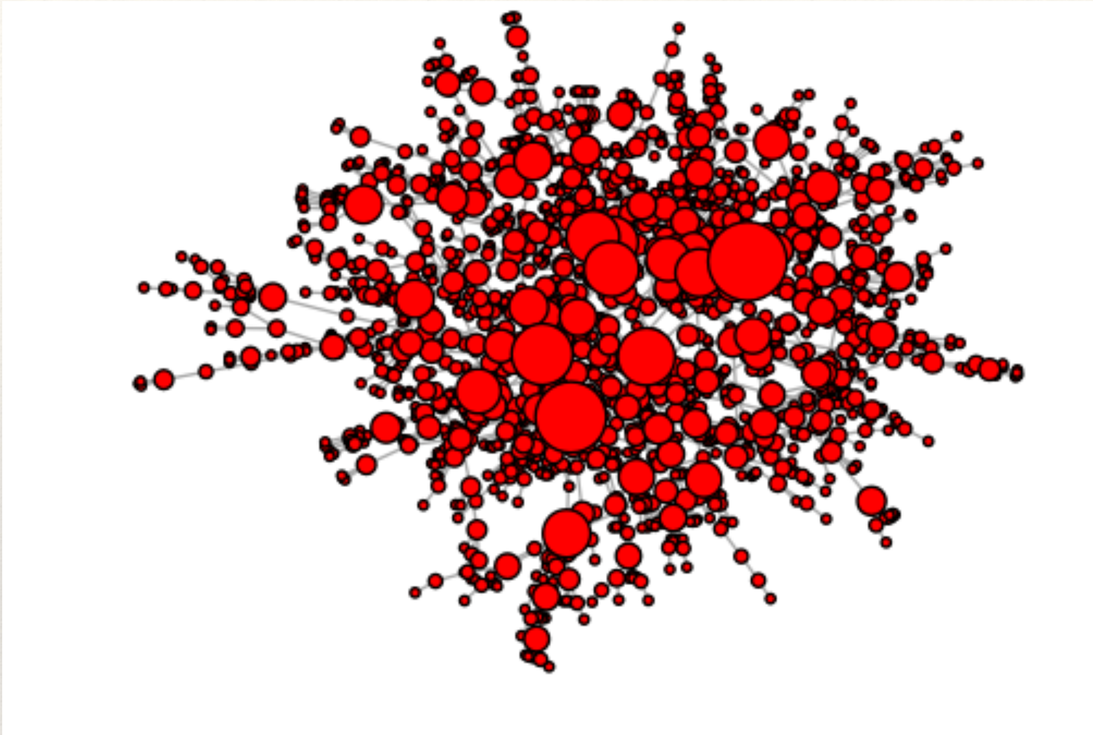
The Barabási-Albert model

Example: if *t* has to choose between node *i*, with degree 6, and node *j*, with degree 3, the probability of choosing *i* is twice the probability of choosing *j*



The Barabási-Albert model

- **Rich-gets-richer phenomenon:** due to preferential attachment, the more connected nodes have higher chances to acquire new links, which gives them a bigger and bigger advantage over the other nodes in the future!



- This is how **hubs** are generated

```
# BA model network
G = nx.barabasi_albert_graph(N,m)
```

The Barabási-Albert model

See http://networksciencebook.com/chapter/5

Why do we get a power law decay? Let $\bar{k}_i^{(t)} = \mathbb{E}[k_i^{(t)}] \quad \bar{p}_i^{(t)} = \mathbb{E}[p_i^{(t)}]$

$$\bar{k}_i^{(t+1)} = \bar{k}_i^{(t)}(1 - \bar{p}_i^{(t)}) + (\bar{k}_i^{(t)} + 1)\bar{p}_i^{(t)}$$

$$\bar{k}_i^{(t+1)} = \bar{k}_i^{(t)} + \bar{p}_i^{(t)}$$

$$\frac{d\bar{k}_i^{(t)}}{dt} = \bar{p}_i^{(t)} \approx \frac{\bar{k}_i^{(t)}}{2t}$$

The Barabási-Albert model

We thus obtain $\bar{k}_i^{(t)} = m \left(\frac{t}{t_i} \right)^{\frac{1}{2}}$

Letting t = n, the nodes with expected degree smaller than k are those for which

$$t_i > n \left(\frac{m}{k} \right)^2$$

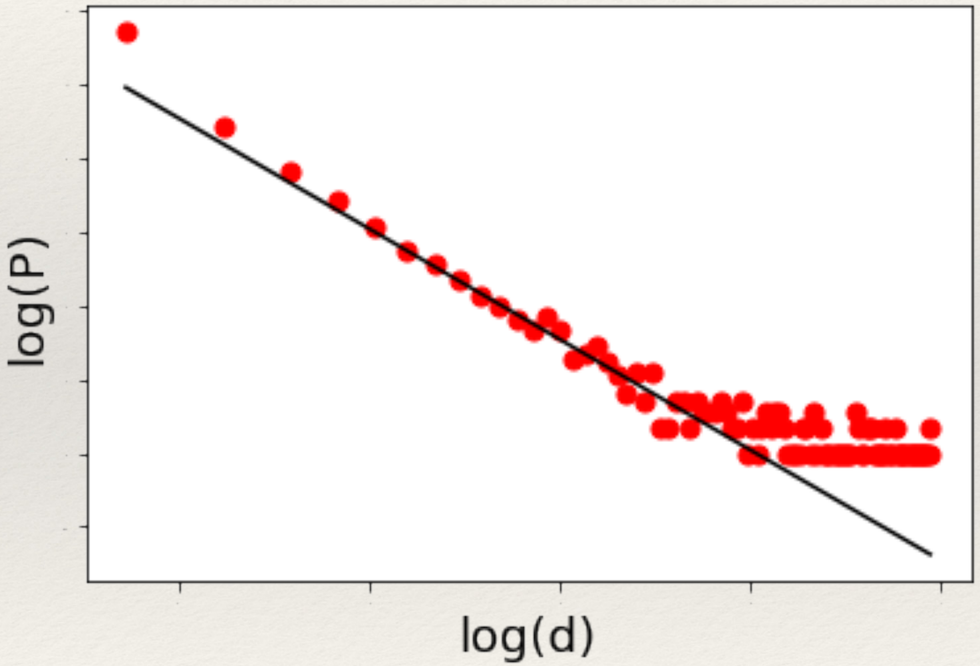
We thus obtain the degree distribution

$$\mathbb{P}(\bar{k}_i^{(t)} > k) = 1 - \left(\frac{m}{k} \right)^2$$

$$\frac{d\mathbb{P}(\bar{k}_i^{(t)} > k)}{dk} \sim k^{-3}$$

The Barabási-Albert model

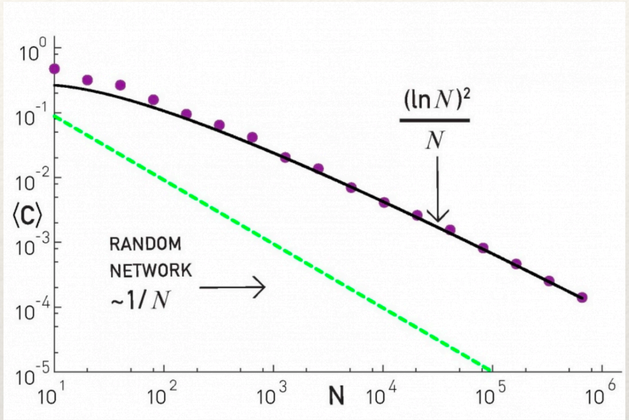
Empirical validation



The Barabási-Albert model

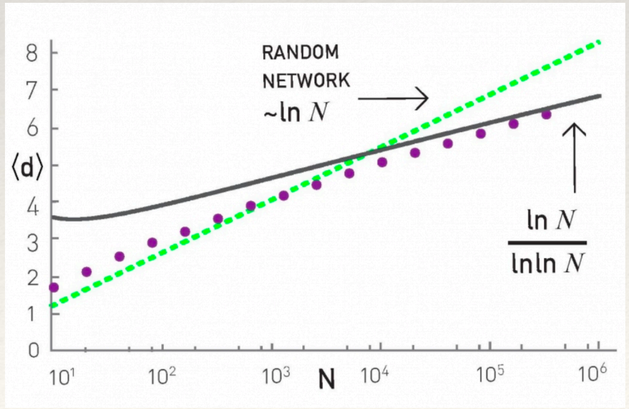
It can be proved that the clustering coefficient scales as

$$C_i \sim \frac{\log^2(n)}{n}$$



While the diameter as

$$\langle l \rangle \sim \frac{\log(n)}{\log\,\log(n)}$$



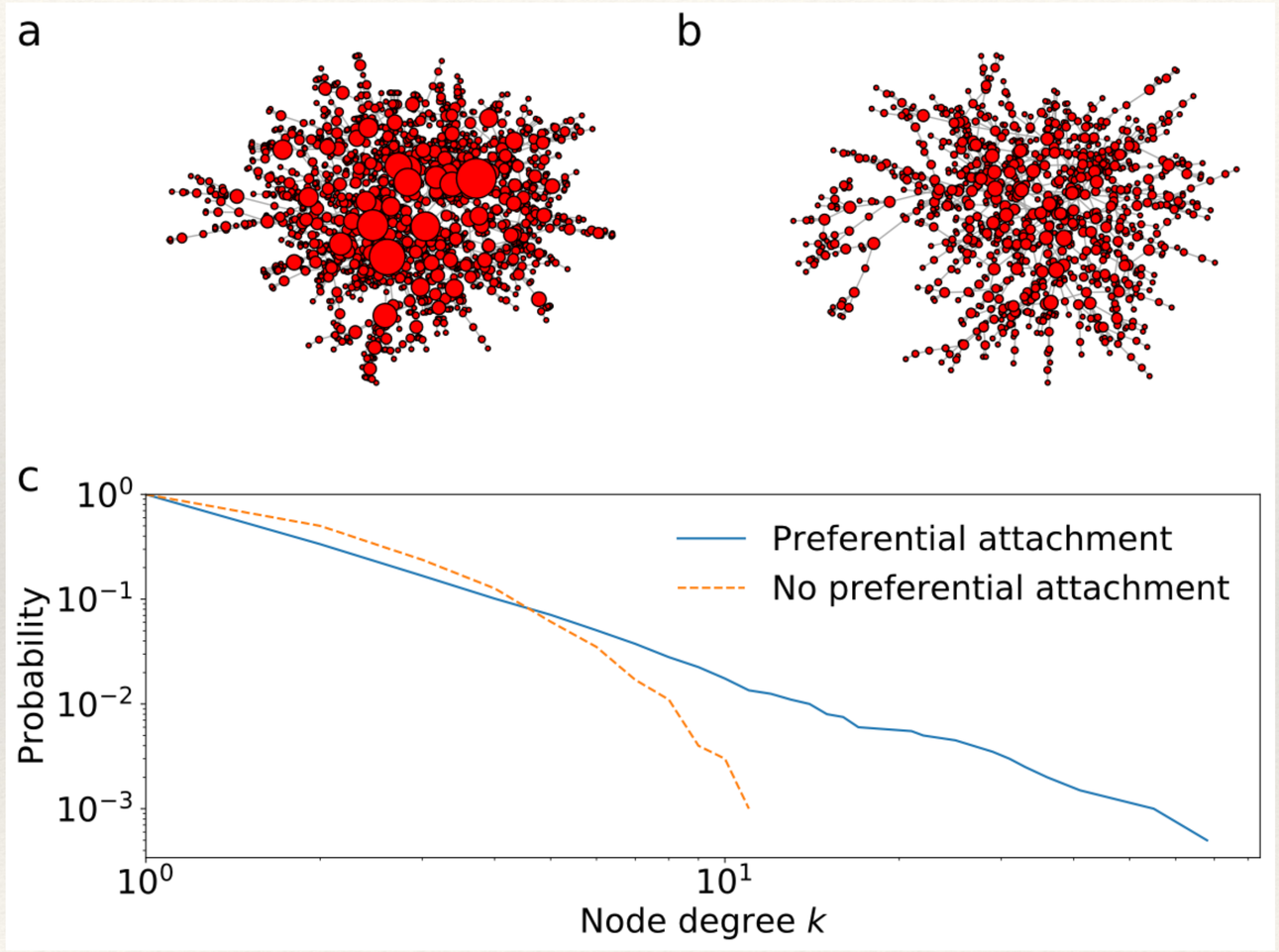
The Barabási-Albert model

- Hubs are the **oldest** nodes: they get the initial links and acquire an advantage over the other nodes, which increases via preferential attachment

- **Question:** if old nodes have an advantage over newer nodes anyway, do we need preferential attachment at all? Can we explain the existence of hubs just because of growth?

- **Alternative model:** each new node chooses its neighbors at random, not with probability proportional to their degree

The Barabási-Albert model



Conclusion: growth + random attachment does not generate hubs.

Preferential attachment is necessary!

$$\bar{k}_i^{(t)} \propto \log(n)$$

Other preferential models

- The BA model uses **linear preferential attachment:** the linking probability is **proportional** to the degree

- **Question:** what happens if the linking probability is **proportional to a power of the degree?**

- **Non-linear preferential attachment!**

Non-linear preferential attachment

- **Procedure:**
  - Start with a group of m_0 nodes, usually fully connected (clique)
  - At each step a new node i is added to the system, and sets m links with some of the older nodes ($m \leq m_0$)
  - The probability that the new node i chooses an older node j as neighbor is **proportional to the power α of the degree k_j of j:**

$$\Pi_\alpha(i \leftrightarrow j) = \frac{k_j^\alpha}{\sum_l k_l^\alpha}$$

  - The procedure ends when the given number N of nodes is reached

Non-linear preferential attachment

$$\Pi_\alpha(i \leftrightarrow j) = \frac{k_j^\alpha}{\sum_l k_l^\alpha}$$

- For $\alpha = 1$ we recover the **linear preferential attachment (BA model)**
- **Question:** what happens when $\alpha \neq 1$?
- **Answer:** it depends on whether $\alpha > 1$ or $\alpha < 1$

Non-linear preferential attachment

- For α <1, the link probability does not grow fast enough with degree, so the advantage of high-degree nodes over the others is not as big. As a result, **the degree distribution does not have a heavy tail: the hubs disappear!**
- If α >1, high-degree nodes accumulate new links much faster than low-degree nodes. As a consequence, one of the nodes will end up being connected to a fraction of all other nodes. For α >2, a single node may be connected to all other nodes (**winner-takes-all effect**), all other nodes having low degree
- **Conclusion:** Non-linear preferential attachment fails to generate hubs. Linear preferential attachment is the only way to go
- **Problem:** Strict proportionality of linking probability to degree appears unrealistic!

Limits of preferential attachment

- It yields a **fixed pattern for the degree distribution:** the slope is the same for any choice of the model parameters. Degree distributions in real-world networks could decay faster or more slowly
- The **hubs are the oldest nodes**: new nodes cannot overcome their degree
- It **does not create many triangles**: the average clustering coefficient is much lower than in many real-world networks
- **Nodes and links are only added:** in real networks they can also be deleted
- Since each node is attached to older nodes, the **network consists of a single connected component**. Many real-world networks have multiple components

Extensions of the BA model: Attractiveness model

- **Pitfall of preferential attachment:** What happens if a node has no neighbors (degree zero)? It will never get connections from other nodes!

- **No problem for standard initial condition:** the initial subgraph is complete (clique), so every node has nonzero degree

- **What if the network is directed and the linking probability is proportional to the in-degree? Bad**, as each new node has in-degree zero, so it will never be linked by future nodes!

Extensions of the BA model: Attractiveness model

- **Procedure:**
  - Start with a group of m_0 nodes, usually fully connected (clique)
  - At each step a new node i is added to the system, and sets m links with some of the older nodes ($m \leq m_0$)
  - The probability that the new node i chooses an older node j as neighbor is proportional to the sum of the degree k_j of j and an attractiveness A, indicating the intrinsic appeal:

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

Extensions of the BA model: Attractiveness model

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

- For $A = 0$ we recover the **BA model**
- For every value of A we get networks with heavy-tailed degree distributions
- The pattern of the distribution **changes with** A, so it is possible to match distributions of real-world networks, **unlike the BA model**

Extensions of the BA model: Fitness model

- **Pitfall of preferential attachment:** the hubs are the oldest nodes. **Unrealistic!**
- **Examples:**
  - In the Web, new pages can overrun old pages (*e.g.*, Google!)
  - In science, new papers can be more successful than (many) old papers
- **Reason:** each node has its own individual **appeal!**

Extensions of the BA model: Fitness model

- **Procedure:**
  - Start with a group of m_0 nodes, usually fully connected (clique)
  - At each step a new node i is added to the system, and sets m links with some of the older nodes ($m \leq m_0$)
  - The probability that the new node i chooses an older node j as neighbor is **proportional to the product of the degree k_j of j with a fitness η_j**, indicating the intrinsic appeal of j:

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

Extensions of the BA model: Fitness model

- The fitness values are extracted from a distribution $\rho(\eta)$ and assigned to each new node

- **Difference with attractiveness model**

  - The fitness enters as a **factor** in the link probability, not as a summand

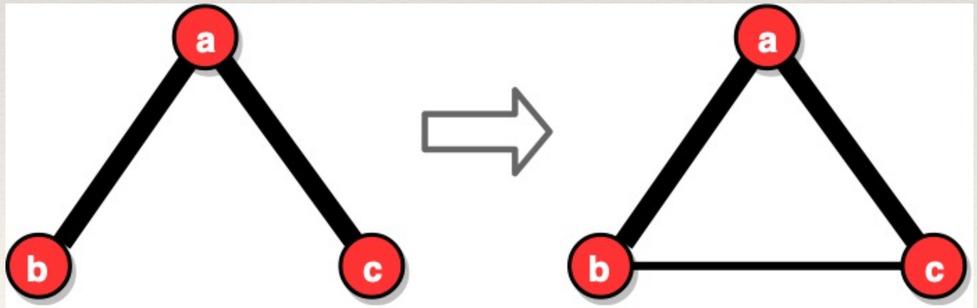
  - The fitness is characteristic of each node, it is not a constant

Extensions of the BA model: Fitness model

- **Results:**
  - If the fitness distribution $\rho(\eta)$ has finite support, *i.e.*, the fitness is distributed over a finite range of values, the degree distribution of the network is heavy-tailed
  - If the fitness distribution $\rho(\eta)$ has infinite support, *i.e.*, the fitness is distributed over an infinite range of values, the node with largest fitness attracts a fraction of all links (**monopoly**)
  - Nodes with large fitness can acquire a large degree even if they are introduced late in the system (**good!**)

Extensions of the BA model: Random walk model

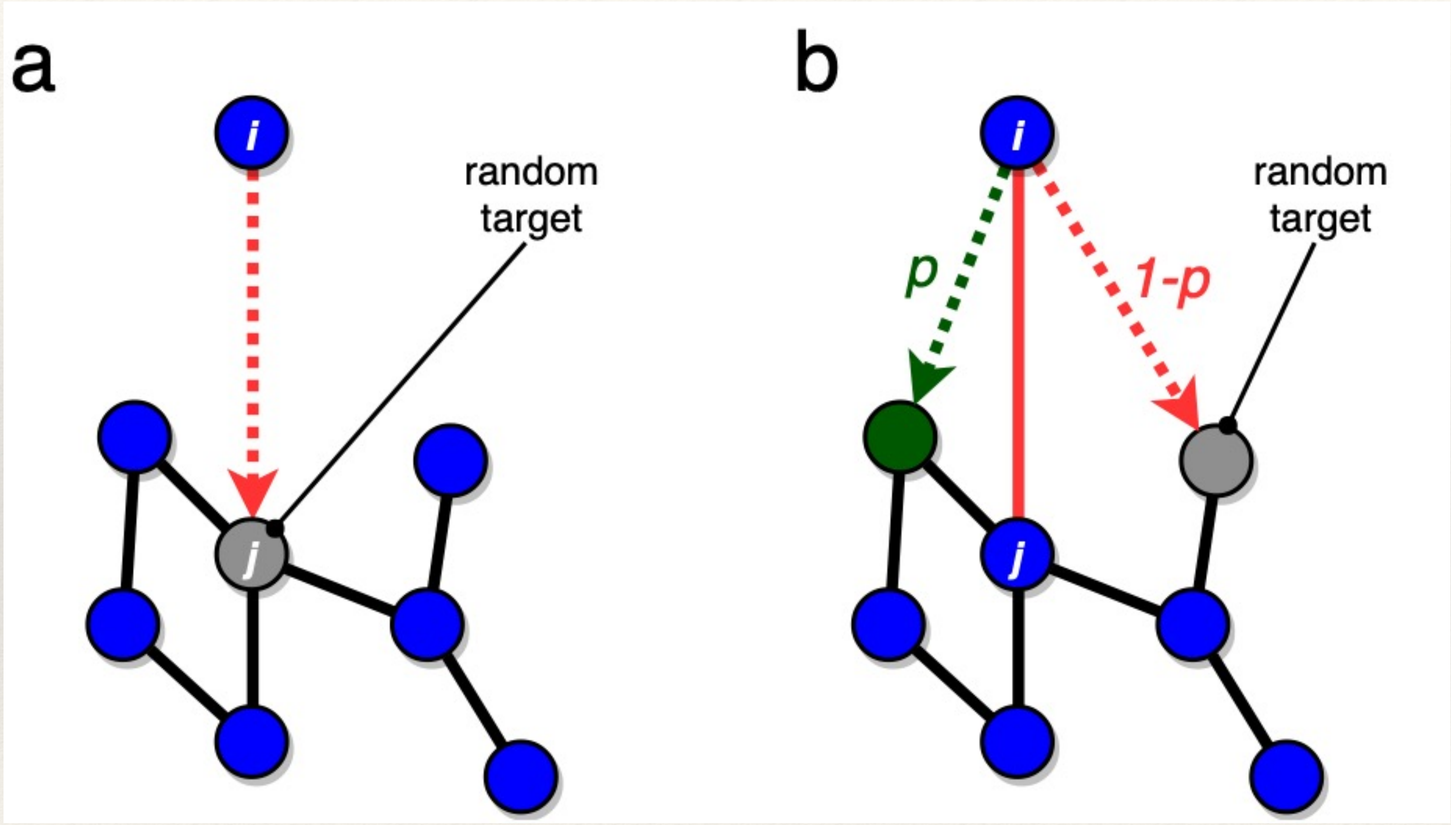
- **Pitfall of preferential attachment:** the BA model does not generate many triangles. **Why?**
- To close a triangle we need to set a link between two neighboring nodes, whereas in the BA model links are set based on degree, regardless of whether the future neighbors have common neighbors
- **Solution:** introduce a mechanism for **triadic closure** in the model!



Extensions of the BA model: Random walk model

- **Procedure:**
  - Start with a group of m_0 nodes, usually fully connected (clique)
  - At each step a new node i is added to the system, and sets m links with some of the older nodes ($1 < m \leq m_0$)
  - The first link targets a randomly chosen node j
  - From the second link onwards:
    - With probability p the link is set with a neighbor of j, chosen at random
    - With probability $1–p$ the link is set with a randomly chosen node

Extensions of the BA model: Random walk model



Extensions of the BA model: Random walk model

- **Results:**
  - The degree distribution is heavy-tailed
  - The average clustering coefficient is much higher than in BA networks (the larger, the greater the probability p of triadic closure)
  - When the triadic closure probability p is sufficiently high that many triangles are formed (p~1) the network has **community structure**, *i.e.*, it is made of cohesive groups of nodes, loosely connected to each other

Extensions of the BA model: Random walk model

- **Question:** if links are set at random, as it seems, how can the model generate hubs?

- **Answer:**

  - Choosing a random node and a random neighbor of the node amounts to choosing a link at random

  - The probability that the endpoint(s) of a randomly selected link have a given degree **is proportional to the degree**

Extensions of the BA model: Random walk model

- **Conclusion:** the triadic closure mechanism of the random walk model **induces effective preferential attachment!**

- **Take-home message:** preferential attachment can be induced by simple mechanisms based on random choices; **it is not necessary to require the knowledge of the degree of the nodes, nor a strict expression of the link probability!**

Extensions of the BA model: Rank model

- **Pitfall of preferential attachment:** BA model implies that nodes have a perception of how important other nodes are, *i.e.*, how large is their degree

- **Objection:** in the real world there is no such perception of the absolute value of things, **it is far easier to perceive the relative value!**

- **Solution:** ranking!

Extensions of the BA model: Rank model

- **Procedure:**
  - Nodes are ranked based on a property of interest (*e.g.*, age, degree). The rank of node i is R_i
  - Start with a group of m_0 nodes, usually fully connected (clique)
  - At each step a new node i is added to the system, and sets m links with some of the older nodes ($m \leq m_0$)
  - The probability that the new node i chooses an older node j as neighbor is **proportional to a power of the rank of j:**

$$\Pi(i \leftrightarrow j) = \frac{R_j^{-\alpha}}{\sum_l R_l^{-\alpha}}$$

Extensions of the BA model: Rank model

$$\Pi(i \leftrightarrow j) = \frac{R_j^{-\alpha}}{\sum_l R_l^{-\alpha}}$$

- **Remark:** highly-ranked nodes (those with low values of R) have high probabilities of being linked, much higher than poorly-ranked nodes

- **Result:** the model generates networks with hubs, for **any value** of the exponent α and **any property** used to rank the nodes!

The Unpredictability of Rich-Get-Richer Effects

The fragility of popularity

- Power laws are produced by feedback effects
- Initial stages of the process that gives rise to the popularity of a node is a relatively fragile thing
- Focus on "cultural market": can we predict the popularity of a song, a movie, a book, etc.?
- We can expect initial fluctuations: this brings unpredictability

predicting hubs emergence

- we can predict that a power law can emerge after a while, and so we can predict that we will have hubs
- but, which hubs?
- predicting the success of an individual item is not like predicting that some individual will have global success!

The MusicLab experiment

Science Contents News Careers Journals

SHARE

REPORT

Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market

Matthew J. Salganik[1,2,*], Peter Sheridan Dodds[2,*], Duncan J. Watts[1,2,3,*]

+ See all authors and affiliations

Science 10 Feb 2006:
Vol. 311, Issue 5762, pp. 854-856
DOI: 10.1126/science.1121066

https://science.sciencemag.org/content/311/5762/854

The MusicLab experiment

- MusicLab: a site where you could listen to songs, and download your favorites
  - unknown songs from unknown artists
  - different qualities
- Visitors: they were randomly assigned to different sessions
- Download counts: measure of popularity
- For every session's category we can produce a songs ranking

The MusicLab experiment

- Very good songs did not end up at the bottom
- Very bad songs did not end up at the top
- What about all the other songs?
  - they resulted in very mixed positions
- Observe that in some sessions the order was established by means of popularity
- Social influence was found relevant at the end of the process
- but **initial fluctuations are unpredictable**

The Long Tail

The Long Tail

- Popularity can be characterized by power laws
- That means that a small set of items is enormously popular
- If you could bet your money on "niches" or "hits", what would you do?
- Chris Anderson's idea:
  - do not focus on hits, but try to estimate the market sales of all the niches

Focus on hits

- stereotype of media business is to focus only on 'hits'

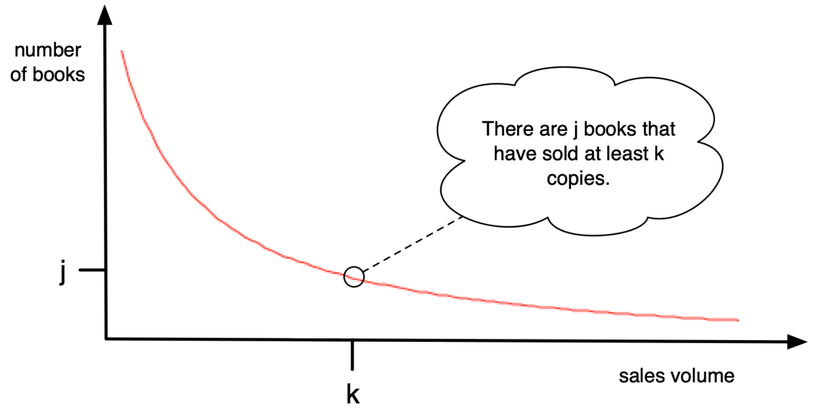


Figure 18.3: The distribution of popularity: how many items have sold at least k copies?

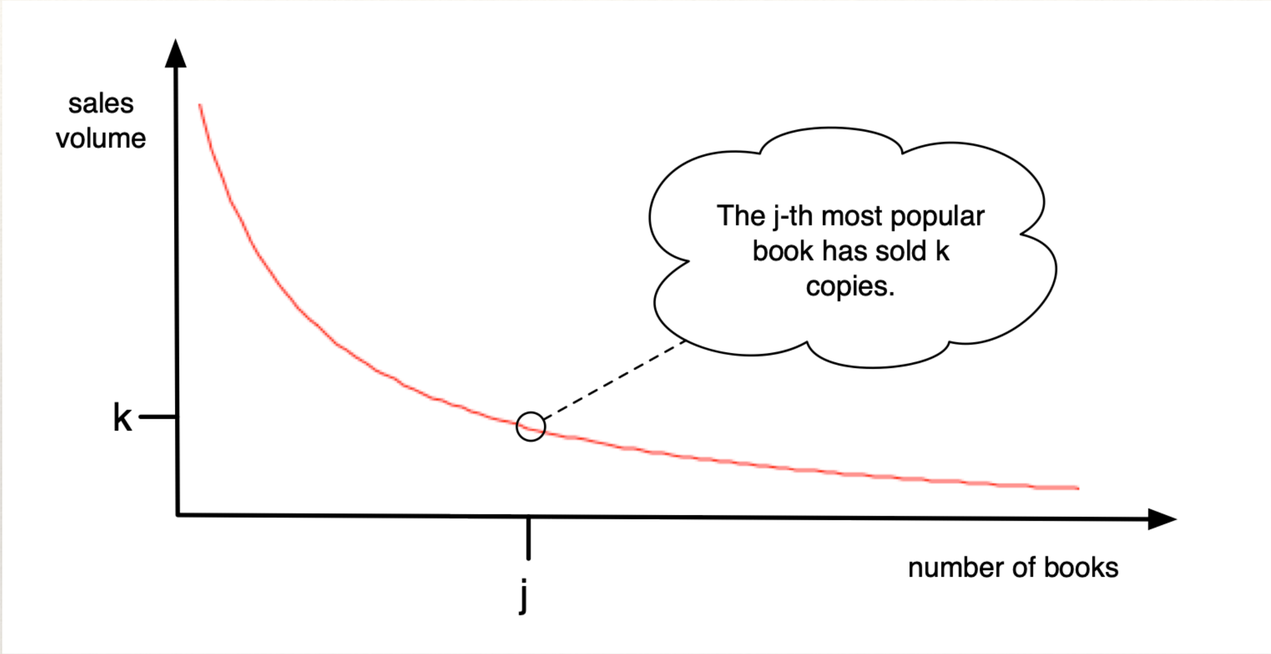
- Which area is bigger? unpopular vs popular items

As a function of k, what fraction of items have popularity exactly k?

Focus on niches

- Switch the axes:

As a function of j what number of items have popularity at least k?



- Focus on the "long tail": when we move to volumes of sales of many niche products. We need to compare if there is significantly more area under the left part of this curve (hits) or the right (niche products).

Zipf′s law

- For the record: the previous plot is known as a **Zipf's plot**
- Introduced by George Kinsley Zipf, a Harvard linguistic professor
- Zipf's law usually refers to the 'size' k of an occurrence of an event relative to it's rank j
  - it states hat the size of the j'th largest occurrence of the event is inversely proportional to it's rank
  - $k \approx j^{-b}$, with b usually close to 1

Pareto's law

- Many of you have probably found similarities with another famous law due to Pareto
- Wilfred Fritz Pareto (a former student of UniTo!) was interested in the distribution of the income
- Instead of asking which was the j-th largest income, he asked how many people have an income greater than j
- Pareto's law is a cumulative probability distribution (cdf):
  - $P(K > k) \approx k^{-\gamma}$

Three ′similar′ laws

- Zipf's law: $k \approx j^{-b}$
- Pareto's law: $P(K > k) \approx k^{-\gamma}$
- Power law: $f(k) \approx k^{-c}$
- They are all connected!
- It is possible to prove that $c = 1 + \gamma$ and that $\gamma = \frac{1}{b}$
- They are just three sides of the same coin!

for more information

Zipf, Power-laws, and Pareto - a ranking tutorial

Lada A. Adamic

Information Dynamics Lab
Information Dynamics Lab, HP Labs
Palo Alto, CA 94304

Abstract

The Effect of Search Tools and Recommendation Systems

Search tools

- Search tools make the rich-get-richer dynamics more evident
- Other aspects that make the effect less extreme:
  - different queries brings to different Search Engine results
  - targeted and personalized search makes unpopular items ranked first
  - recommendation system's serendipity exploits the long tail argument
- Complex effect in already complex systems