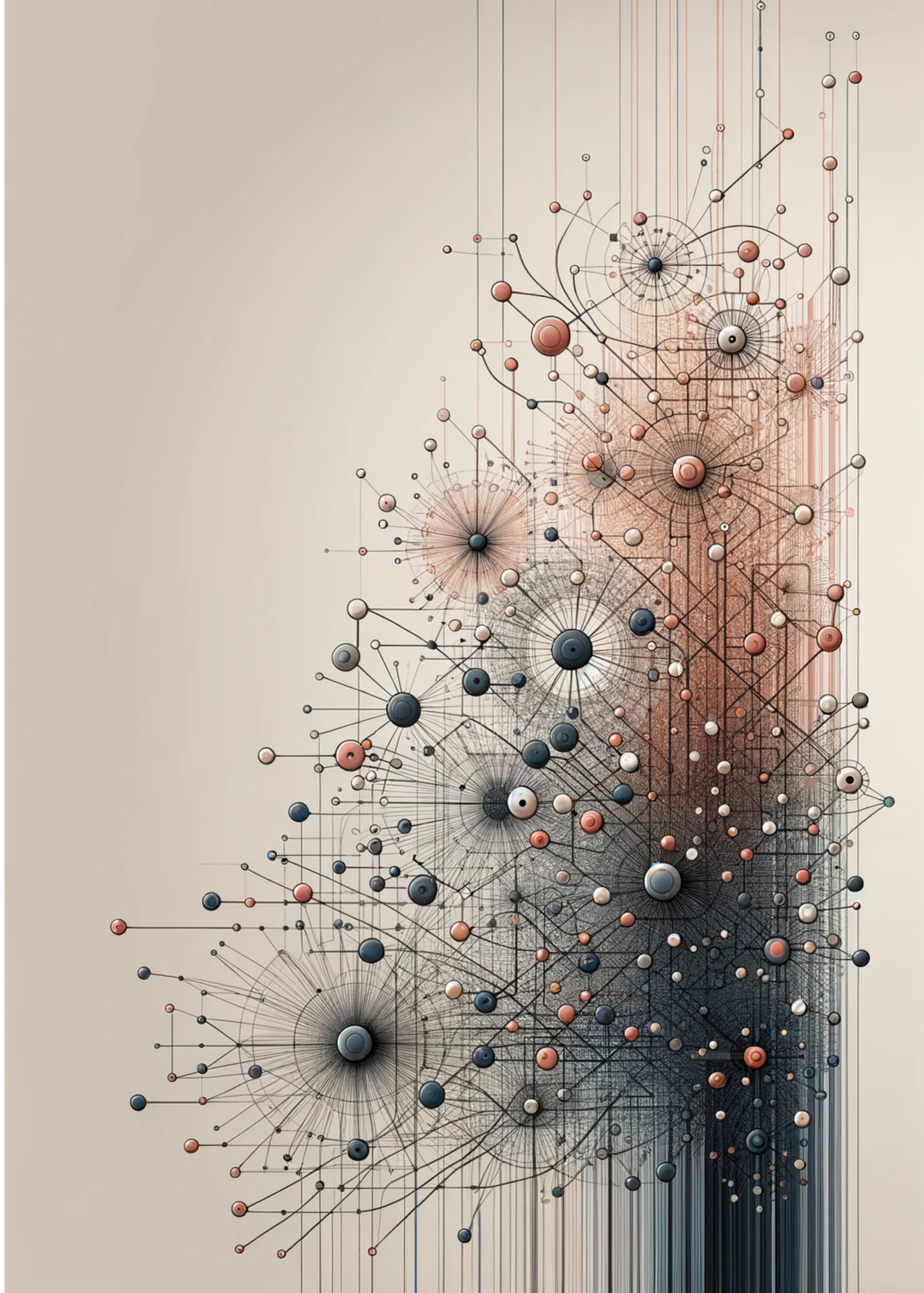
UNIVERSITÀ
DI TORINO

Analisi e Visualizzazione delle Reti Complesse

NS09 - Power Laws and Rich-Get-Richer Phenomena

Prof. Rossano Schifanella



Agenda

- Popularity as a Network Phenomenon
- Power Laws
- Rich-Get-Richer Models
- The Unpredictability of Rich-Get-Richer Effects
- The Long Tail
- The Effect of Search Tools and Recommendation Systems

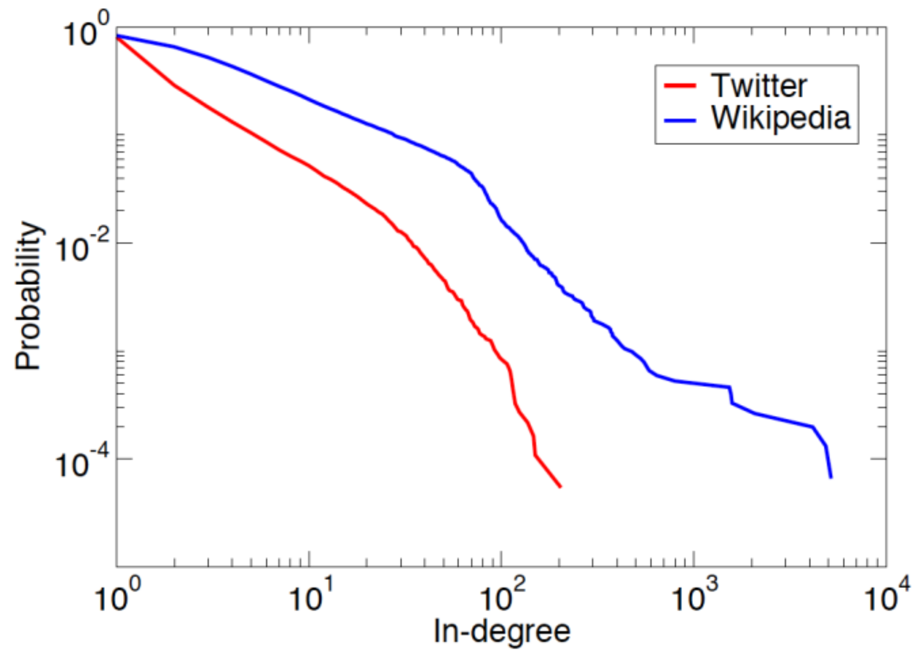
Popularity as a Network Phenomenon

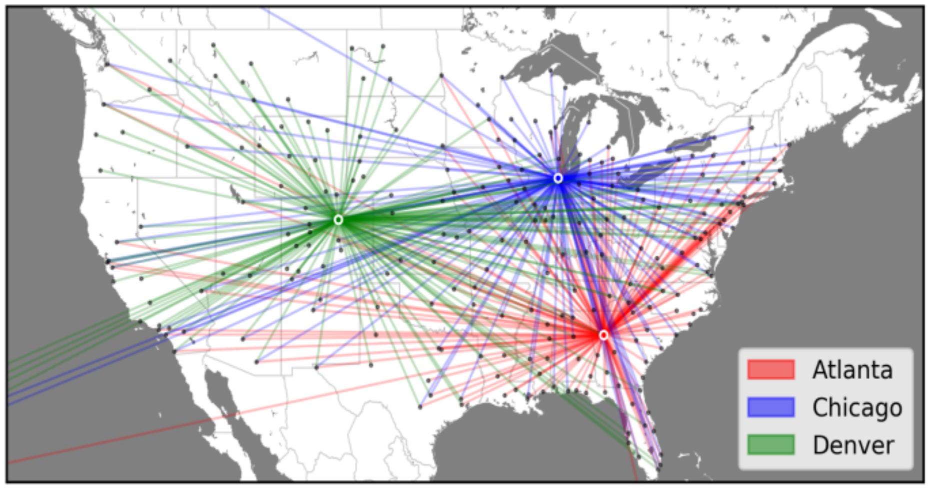
Popularity, heterogeneity and networks

Recap:

- Real networks are heterogeneous
- **In-links as a measure of popularity**
- The **heterogeneity** parameter as a **measure of distribution's broadness**

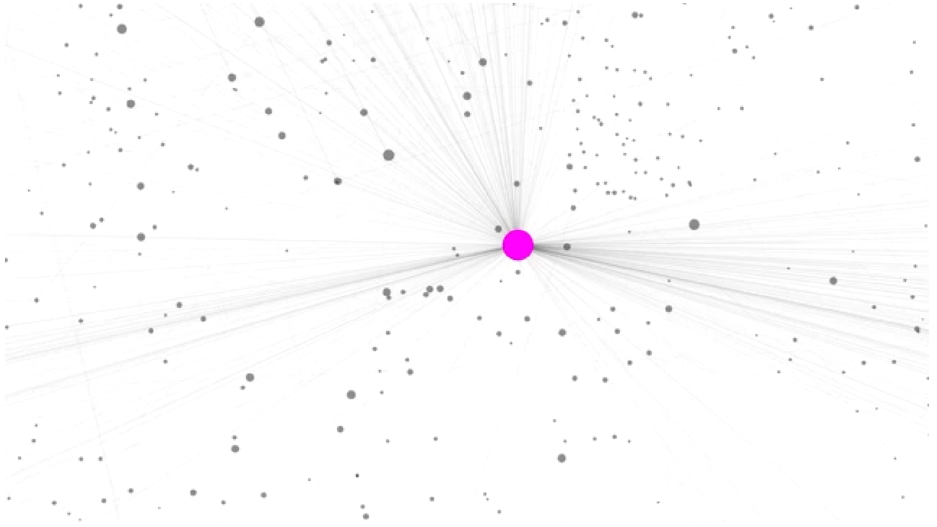
$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$$





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?
  - Have a look at the video for a dynamic view



UNIVERSITÀ DI TORINO

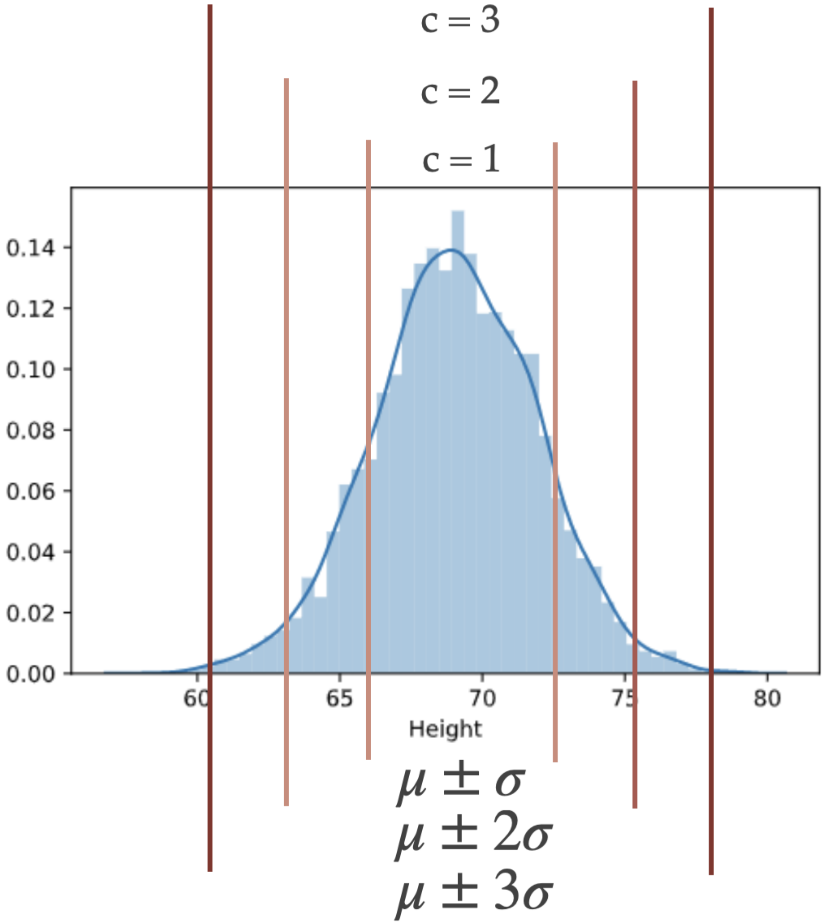
Looking for a popularity scale

- 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

UNIVERSITÀ DI TORINO

Example: Heights

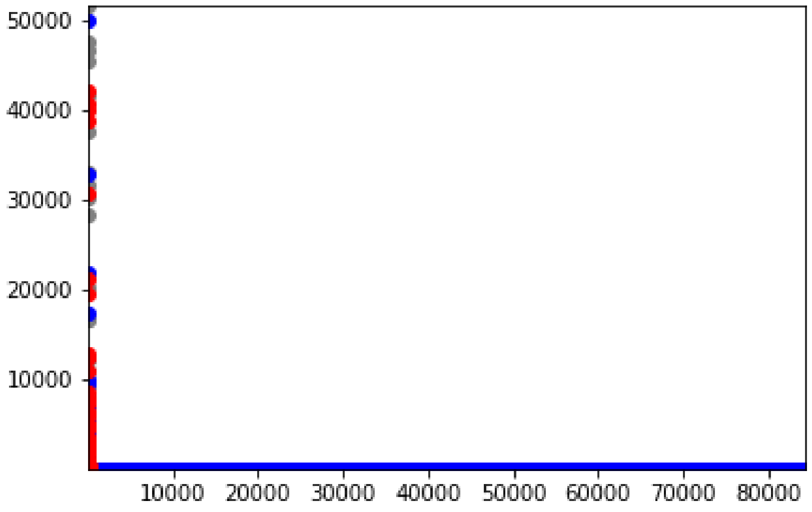
- If we look at people's height distributions
- mean: = 69.03 (feet)
- std: = 2.86
- The probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially with c
- Amazingly high persons are very unlikely



UNIVERSITÀ DI TORINO

Example: the Web (a sample)

- **Source dataset:**
    - 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

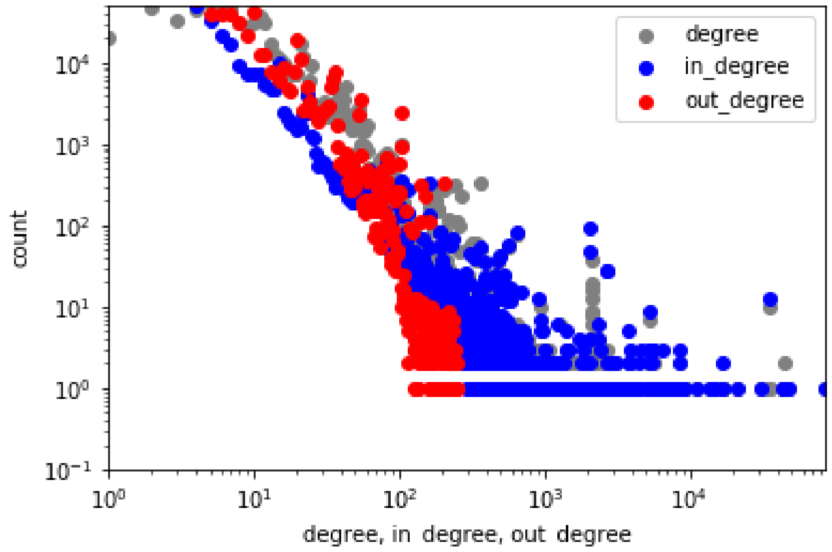


Example: the Web (a sample) in a log-log scale

- 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

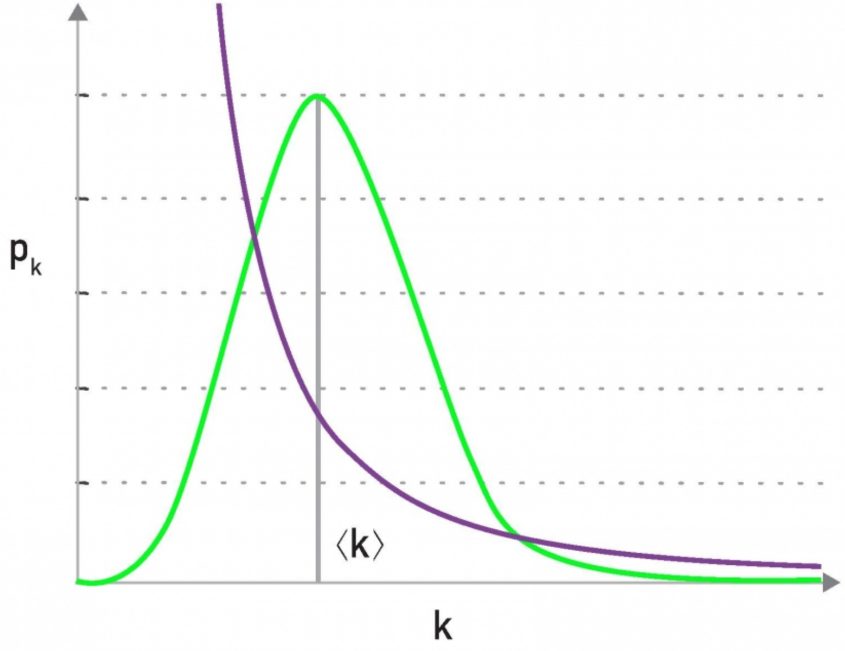
- The fraction of web pages that have k in-links is:

$$f(k) \approx \frac{1}{k^c} = k^{-c}$$

- $c = 2.1$
- in other networks $2 < 3$ very often
- $f(k) = ak^{-c}$: power law distribution
- you can calculate that in the $2 < 3$ 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 the 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 about 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 characteristics of power laws is useful - as much as looking for an underlying process that explains the emergence of hubs

UNIVERSITÀ DI TORINO

Power Laws

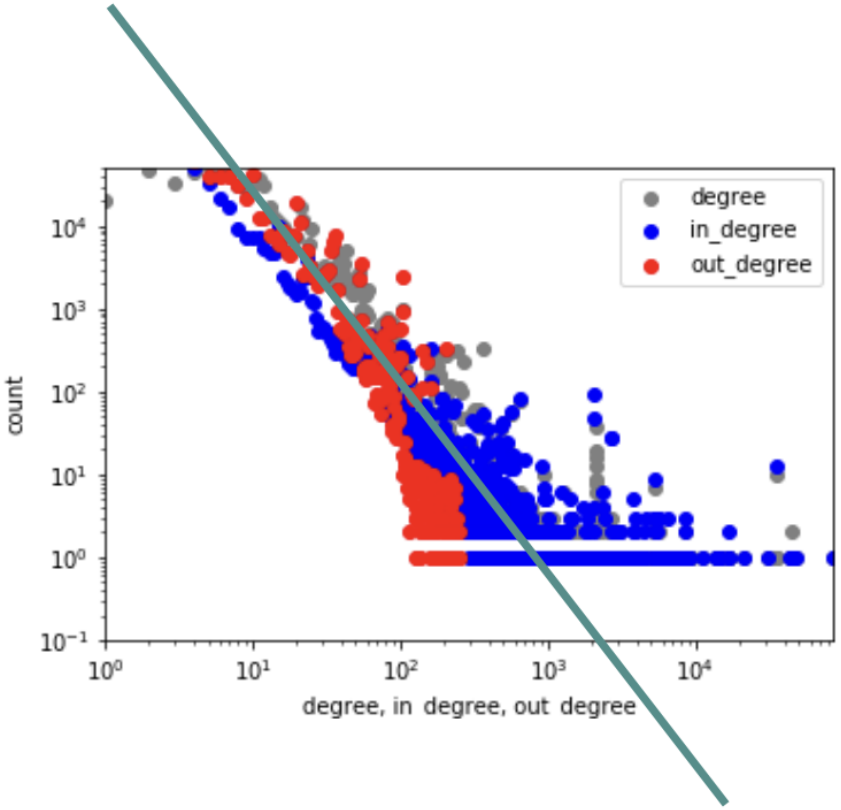
- $f(k) \approx \frac{1}{k^c} = k^{-c}$ decreases much more slowly as k increases
  - Pages with very large k are much more common than expected with the normal distribution
  - Emergence of hubs is likely
- This can be observed empirically in many domains

Fitting with a straight line

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

$$log(f(k)) = log(ak^{-c})$$
$$log(f(k)) = log(a) - clog(k)$$
$$y = log(a) - cx$$

- 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?

Rich-Get-Richer Models

Rich-Get-Richer models

- Lesson from cascades: we assume that people tend to copy the decision of people who acted before them

1. Nodes are created in a sequence: $1, 2, \ldots, N$
2. For each node j that joins the network, repeat:
   i. with probability $p \Rightarrow$ page i is selected uniformly at random, and a link (j, i) is created
   ii. with probability $1 - p \Rightarrow$ page i is selected uniformly at random, l is the page i is connected to, then a link (j, l) is created

- Keep the process simple: only one link is created at every step

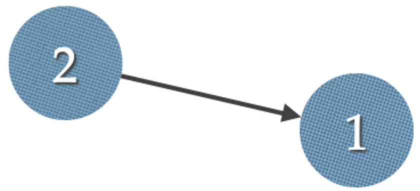
Rich-Get-Richer process

$p = 0.5$

Rich-Get-Richer process

$p = 0.5$

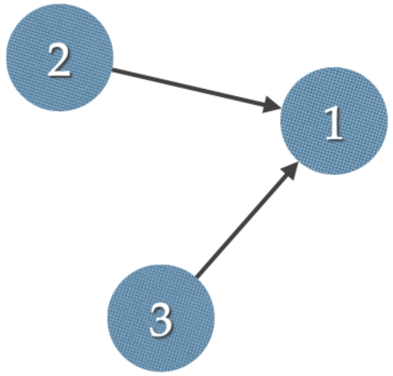
$rand = 0.4 < p$



Rich-Get-Richer process

$p = 0.5$

$rand = 0.3 < p$

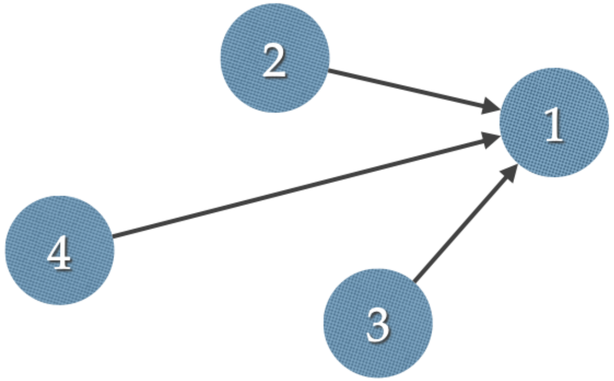


UNIVERSITÀ DI TORINO

Rich-Get-Richer process

$p = 0.5$

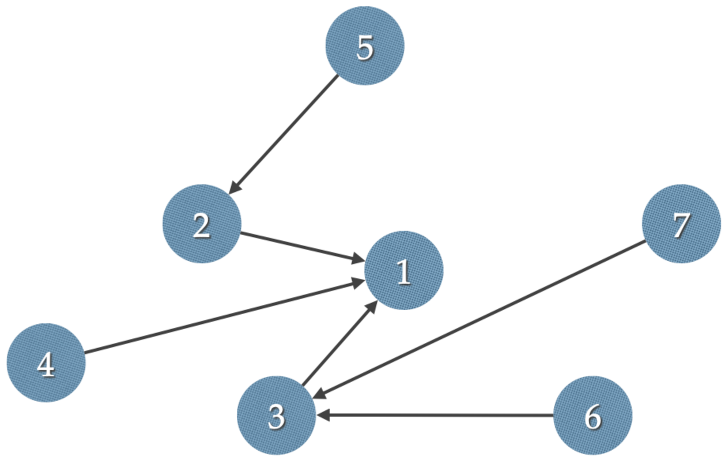
$rand = 0.6 > p$



UNIVERSITÀ DI TORINO

Rich-Get-Richer process

$p = 0.5$



Preferential attachment

- The rich-get-richer model is a generalization of the preferential attachment model by Barabasi and Albert
- with probability $p \Rightarrow$ the selection of the end-point is random
- with probability $1 - p \Rightarrow$ we can prove that rule 2.(ii) is equivalent to the following rule: "page j chooses a page with probability proportional to l's current number of in-links" $\Rightarrow$ that is preferential attachment
- This is a simple model: it does not explain everything, but it provides a natural explanation for the emergence of hubs
- Do not be surprised to observe skewed distributions in real data that resemble power law (usually from $k > k_{min}$ - other times exhibiting exponential cut-offs)
- Check: "preferential attachment" simulation with NetLogo

UNIVERSITÀ DI TORINO

The Unpredictability of Rich-Get-Richer Effects

The fragility of popularity

- Power laws are produced by **feedback effects**
- The initial stages of the process that gives rise to popularity is a **relatively fragile state**
- Focusing 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** and **hubs will emerge**
- **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

[The MusicLab experiment paper]

The MusicLab experiment

- MusicLab: a site where you could listen to songs and download your favorites unknown songs from unknown artists
    - different audio 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

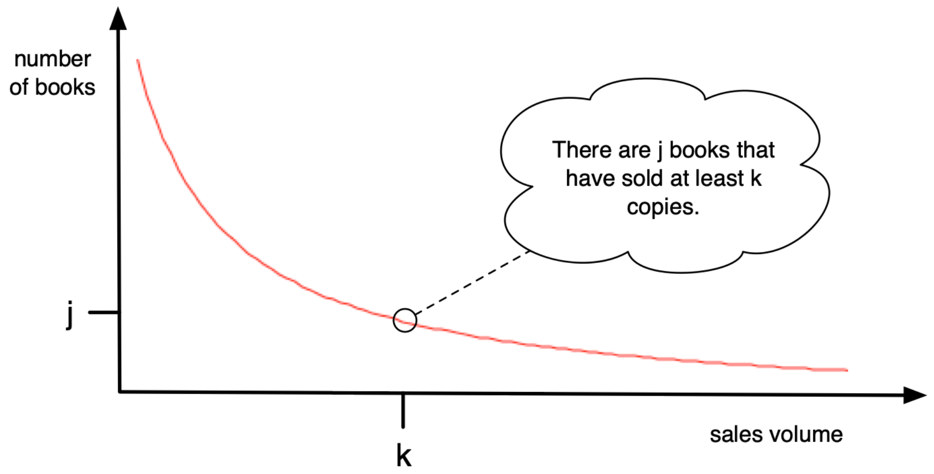
UNIVERSITÀ DI TORINO

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**
- Which area is bigger?
  - **unpopular vs popular items**
- As a function of k, what fraction of items have popularity exactly k?

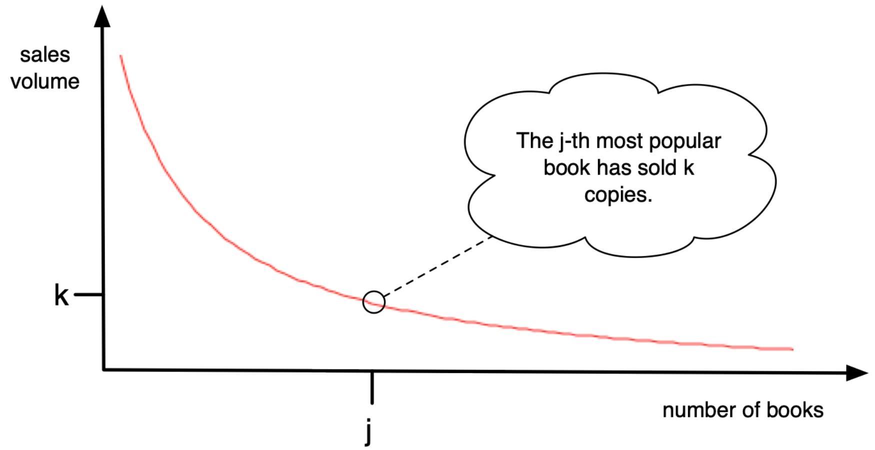


Focus on niches

Switch the axes:

- 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).

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



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 its rank
  - $k \approx j^{-b}$, with b usually close to 1

UNIVERSITÀ DI TORINO

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

- Reading material:
  - Zipf, Power-laws, and Pareto - a ranking tutorial (Adamic)
  - Power laws, Pareto distributions and Zipf's law (Newman)

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

Reading material

[ns2] **Chapter 18 (18.1 - 18.6) Power Laws and Rich-Get-Richer Phenomena**

Please check your general understanding of the topic completing the exercises at the end of the chapter

UNIVERSITÀ DI TORINO

Q&A