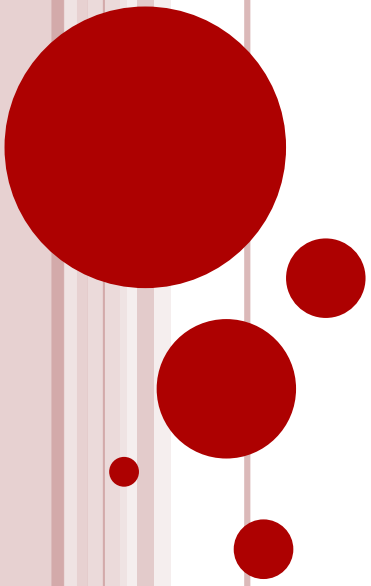
CORSO DI LAUREA
MAGISTRALE IN
INGEGNERIA INFORMATICA

SOCIAL NETWORKS ANALYSIS
A.A. 2021/22



POWER LAWS

Popularity as a Network Phenomenon

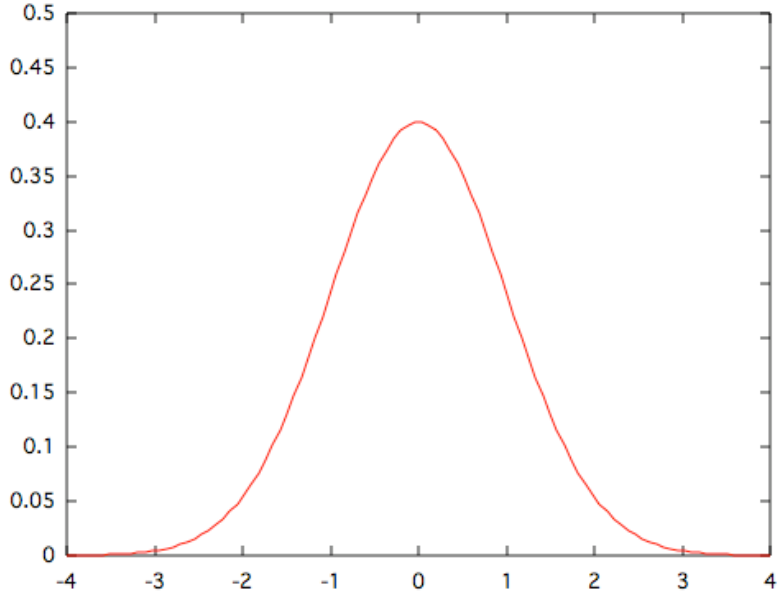
- In network settings we observe that a person's behavior/decisions depend on the choices made by other people
  - These coupled decisions can lead to outcomes very different from what we find when individuals make independent decisions
- In this lesson we apply this network approach to analyze the general notion of popularity
- Popularity is a phenomenon characterized by extreme imbalances
  - almost everyone is known only to people in their immediate social circles
  - a few people achieve wider visibility
  - very few attain global visibility
- The same could be said of books, movies, or almost anything that commands an audience
- How can we quantify these imbalances?
- Why do they arise?
- Are they somehow intrinsic to the whole idea of popularity?

Web Popularity

- Web is a concrete domain in which it is possible to measure popularity very accurately
  - We can take the number of in-links as a measure of the popularity of a page
  - Take a snapshot of the full Web and count the number of links to each page
- Early in the Web's history, people asked how popularity is distributed over the Web pages
- As a function of k, what fraction of pages on the Web have k in-links?

A Simple Hypothesis: Normal Distribution

- A natural guess for the distribution of popularity is the normal (Gaussian) distribution
  - used widely throughout probability and statistics
  - characterized by two quantities: a mean value, and a standard deviation around this mean



plot of the density of values in the normal distribution with mean 0 and standard deviation 1

Normal Distribution in Natural Sciences

- The normal distribution is ubiquitous across natural sciences
- The Central Limit Theorem says that the sum (or average) of any sequence of small independent random quantities, will be distributed in the limit according to the normal distribution
- Suppose we perform repeated measurements of a fixed physical quantity
  - assume variations in the measurements across trials are the cumulative result of many independent sources of error in each trial
  - then the distribution of measured values should be approximately normal

Normal Distribution in the Web Popularity Case

- Does the Normal Distribution apply in the case of Web pages?

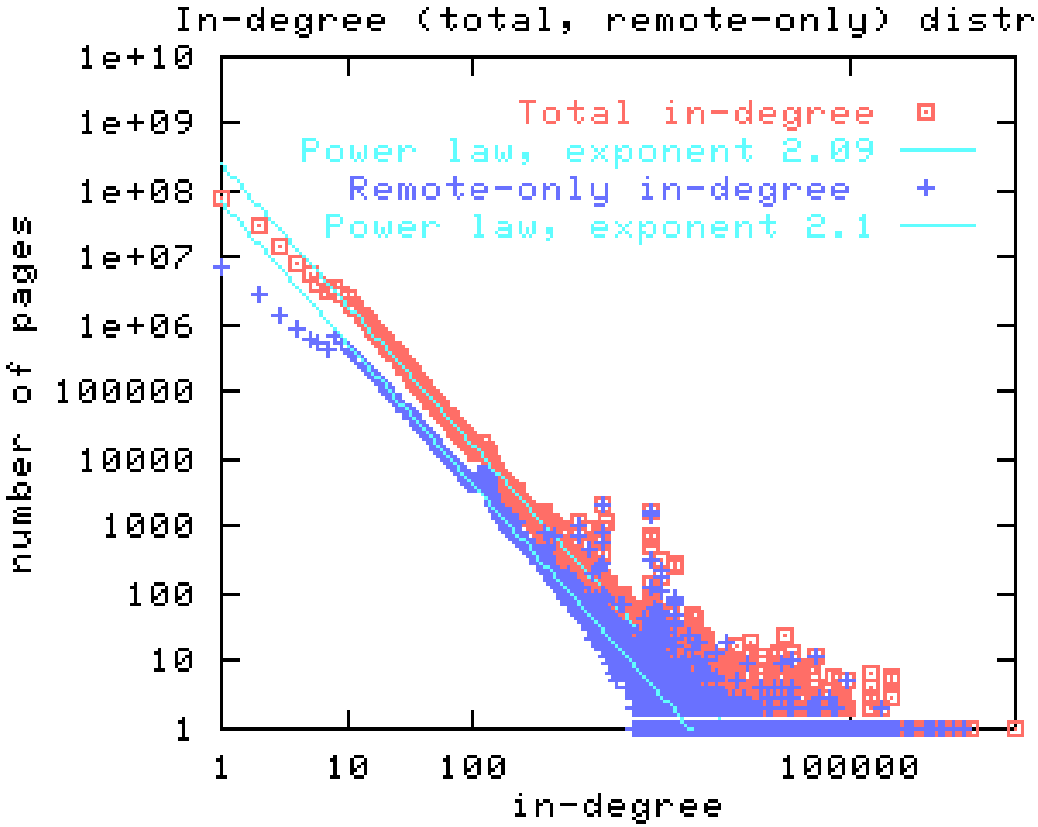
- Possible argument:
  - model the link structure of the Web assuming that each page decides independently at random whether to link to any given other page
  - the number of in-links to a given page is the sum of many independent random quantities and is normally distributed

- In this case the number of pages with k in-links should decrease exponentially in k
  - Several measurements proved this is wrong

POWER LAWS

- Studies over many different Web snapshots showed that the fraction of Web pages with k in-links is approximately proportional to $1/k^2$
  - pages with very large numbers of in-links are much more common than we'd expect with a normal distribution.
- A function that decreases as k^{-c} is called a power law
- Power laws seem to dominate in cases where the quantity being measured can be viewed as a type of popularity
  - The number of Web pages with k in-links is roughly proportional to $1/k^2$
  - The fraction of telephone numbers that receive k calls per day is roughly proportional to $1/k^2$
  - the fraction of books that are bought by k people is roughly proportional to $1/k^3$
  - the fraction of scientific papers that receive k citations in total is roughly proportional to $1/k^3$

How to Recognize a Power Law?



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

- A power law distribution shows up as a straight line on a log-log plot
  - c is the slope of the line and $\log a$ is the intercept on the y-axis

Why Power Laws are so Widespread?

- Why power laws are so widespread?
  - Ideas from the analysis of information cascades and network effects provide the basis for a very natural mechanism to generate power laws
- Normal distributions arise from many independent random decisions averaging out
  - Central Limit Theorem
- Power laws arise from the feedback introduced by correlated decisions across a population
  - It is an open research question to provide a fully satisfactory model of power laws starting from simple models of individual decision-making

Rich-Get-Reacher Models

- We can construct a model based on the observable consequences of decision-making in the presence of cascades

- We assume that people have a tendency to copy the decisions of people who act before them
  - They copy decisions of popular individuals with higher probability

A Simple Model of Web Page Creation

- To keep things simple, we suppose that each page creates just one outbound link
- Pages are created in order, and named 1, 2, . . , N.
- When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule
  - With probability p, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to this page
  - With probability $1 - p$, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to the page that i points to

Rich-get-Richer Dynamics Give Rise to Power Laws

- If we run the model for many pages, the fraction of pages with k in-links will be distributed approximately according to a power law $1/k^c$
  - c depends on the choice of p
  - as p gets smaller, (copying more frequent), the exponent c gets smaller as well, (more likely to see extremely popular pages)

- The key point of our model is that author of page j with probability (1-p) copies the decision of the author of page i
  - This copying mechanism is an implementation of a rich-get-reacher dynamics

WHY RICH-GET-REACHER?

- When you copy the decision of a random earlier page, the probability that you end up linking to some page i is directly proportional to the total number of pages that currently link to i
- We can write our copying process as
  - With probability $(1 - p)$, page j chooses a page i with probability proportional to i's current number of in-links, and creates a link to i
  - the probability that page i increases its popularity is directly proportional to i's current popularity
- This phenomenon is also known as preferential attachment
  - links are formed "preferentially" to pages that already have high popularity
- This copying model provides a reason for why popularity should exhibit such rich-get-richer dynamics
  - the more well-known someone is, the more likely you are to hear her name comes up in a conversation, and hence the more likely you are to end up knowing about them as well

Rich-get-Richer as a Basis for Power Laws

- Rich-get-Richer models can suggest a basis for power laws in a wide array of settings
  - Even settings that have nothing at all to do with human decision-making
- Some examples
  - The in-degree of Web pages follows a power law distribution
  - the population of cities follows a power law distribution
  - the number of copies of a gene in a genome approximately follows a power-law distribution
- Other models were designed to capture power-law behavior
  - e.g. power laws can arise from systems that are being optimized in the presence of constraints
    - not discussed here

The Unpredictability of Rich-Get-Richer Effects

- The rise to popularity for any other object of popular attention is a relatively fragile thing
  - Once any item is well-established, the rich-get-richer dynamics of popularity are likely to push it even higher
- The dynamics of popularity suggest that random effects early in the process play a fundamental role
- Replaying history multiple times
  - it seems likely that there would always be a power-law distribution of popularity each of these times
  - but it's far from clear that the most popular items would always be the same

An Interesting Experiment -- 1

- Salgankik, Dodds, and Watts performed an experiment that gives evidence of the fragility of the rich-get-richer phenomenon

- They created a music download site, populated with 48 obscure songs
  - Visitors were presented with a list of songs with a download count of each song, and given the opportunity to listen to them
  - At the end of a session, the visitor could download copies of the songs she liked

An Interesting Experiment -- 2

- They run 8 "parallel" copies of the server
  - Each copy started with the same initial configurations
  - Visitors randomly assigned to a copy and unaware of the parallel copies
- In different parallel copies popularity of songs varied considerably
  - But best songs never ended up at the bottom and worst songs never ended up at the top
- Another server was run without download counts
  - Songs' popularity significantly changed and they showed no power law
- The success of a book, movie, celebrity, or Web site is strongly influenced by these types of feedback effects
  - Hence may be inherently unpredictable

Information Cascades and Power Laws

- In information cascades people that are aware of earlier decisions made between two alternatives could end up in a cascade
  - Based on rational choices
- In the Rich-get-Richer model individuals are copying decisions of other (maybe randomly selected) persons
- The two models differ in several respects
  - The model for popularity should include choices among many possible options rather than just two options
  - In the copying model an individual observes only a limited part of the whole population (maybe only one randomly selected individual)
  - Imitation in information cascades derive from a model of rational decision-making, we don't have such a model for rich-get-richer dynamics

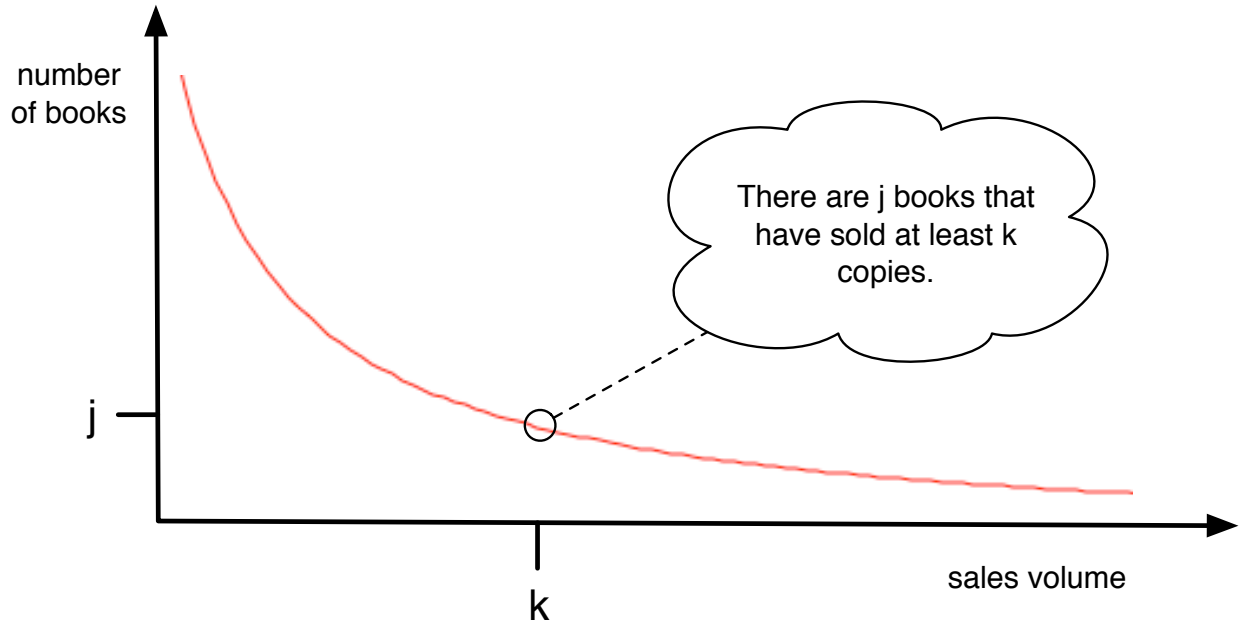
The Long Tail

- The distribution of popularity can have important business consequences, particularly in the media industry
- Are most sales of a media company (with a huge inventory) being generated by a small set of items that are enormously popular, or by a much larger population of items that are each individually less popular?
  - In the former case, the company is basing its success on selling "hits"
  - In the latter case, the company is basing its success on a multitude of "niche products"
- In 2004 Chris Anderson argued that Internet-based distribution and other factors are making the latter alternative dominant
- Huge success of companies like Amazon or Netflix confirmed the rightness of this argument
  - Huge inventories without restrictions of physical stores
  - Their volume of sells consists of a huge quantity of products, each one sold in a very small quantity

VISUALIZING THE LONG TAIL -- 1

Consider the following different question

- As a function of k, what number of items have popularity >= k?

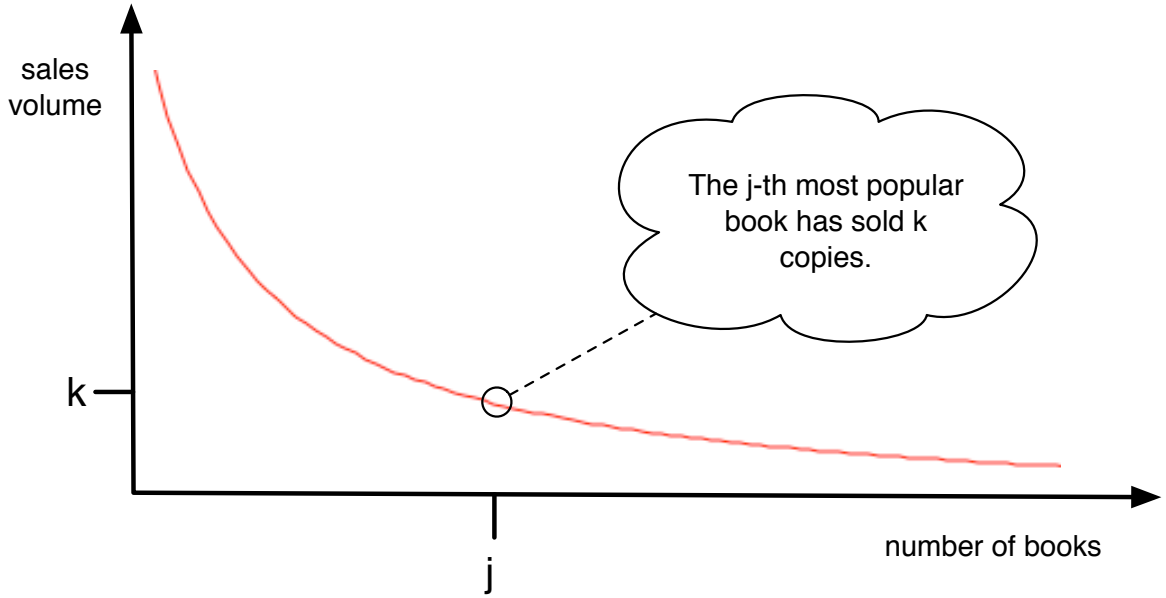


- We still have a power law

- When k increases the number of products with popularity >= k gets smaller and smaller
  - But there is a fraction of very popular items

VISUALIZING THE LONG TAIL -- 2

- As you look at less and less popular items, what sales volumes do you see?
  - Simply exchange axes



- Order products by "sales rank,"
- Look at the popularity of books as we move to smaller and smaller sales ranks
- The area under the right tail of the curve is the volume of sales due to niche products

Effects of Searching Tools on Popularity

- Are Internet search tools making the Rich-get-Richer dynamics of popularity more extreme or less extreme?

- People use search engines such as Google to find pages
  - Google is using popularity measures to rank Web pages and highly-ranked pages are the preferred alternatives for linking
  - This kind of feedback makes rich-get-richer dynamics even stronger, producing even more inequality in popularity

- Users type a very wide range of queries into Google
  - by getting results on relatively obscure queries, users are being led to pages that they are likely never to have discovered through browsing alone
  - Search tools used in this style enable people to find unpopular items more easily
  - This kind of feedback counteracts the rich-get-richer dynamics

Recommendation Systems

- In order to make money from a giant inventory of niche products, a company needs to make its customers aware of these products

- Companies like Amazon and Netfix adopted recommendation systems as important parts of their business strategies
  - search tools designed to expose people to items that may not be generally popular, but which match user interests as inferred from their history of past purchases