Social Network Analysis – Network Dynamics

Power Laws

Overview

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

Overview

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

Popularity

- Popularity is a phenomenon characterized by extreme imbalances:
  - Almost everyone goes through life known only to people in their immediate social circles
  - A few people achieve wider visibility
  - A very, very few attain global name recognition
- The same can be said of books, movies, or almost anything that commands an *audience*
- How to quantify these imbalances in popularity?
- Are they intrinsic to the idea of popularity?

Popularity of Web pages

- The number of in-links to a web page is a measure of the page's popularity.
- *As a function of k, what fraction of pages on the Web have k in-links?*
- Larger values of k indicate greater popularity, so this is precisely the question of how popularity is distributed over the set of Web pages.

Normal (Gaussian) Distribution
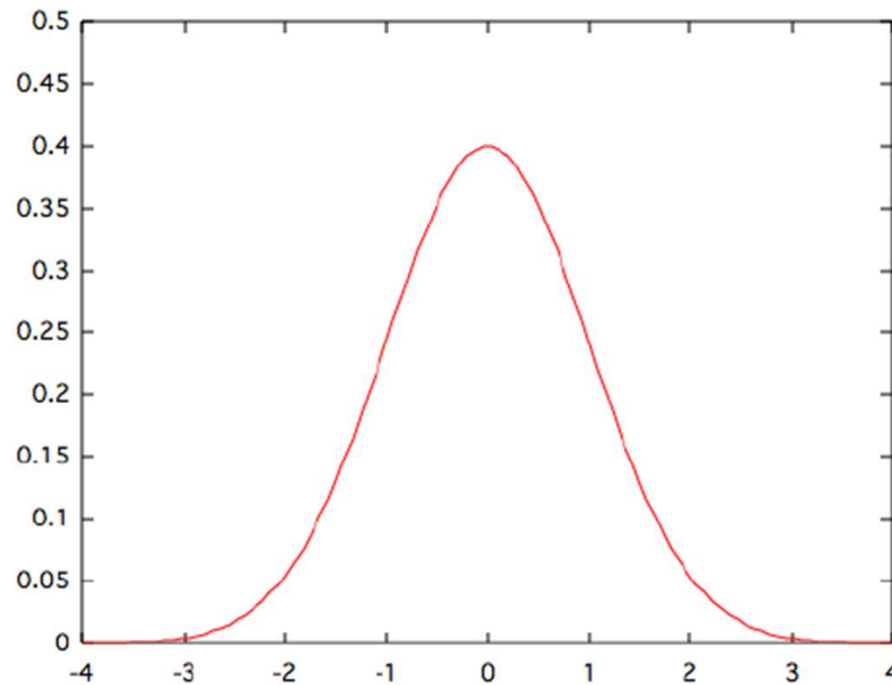
The density of values in the normal distribution

- A Normal Distribution is characterized by a "mean" and a "standard deviation"
- (0 and 1 in the above, respectively)
- The probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially in c

Central Limit Theorem

- A Normal distribution occurs in many real-life situations.
- Central Limit Theorem
  - *Take any sequence of small independent quantities.*
  - *Then in the limit, their sum (or average) will be distributed according to the normal distribution.*

How does it apply to Web Pages?

- If we model the link structure of the Web
  - Assuming that each page decides independently at random whether to link to any other given page,
  - Then the number of in-links to a given page is the sum of many independent random quantities
  - Hence we'd expect it to be normally distributed.

- So, our hypothesis… *The number of pages with k in-links should decrease exponentially in k, as k grows large*

Overview

- Popularity as a Network Phenomenon
- **Power Laws**
- Rich-Get-Richer Models
- The Long Tail
- The Effect of Search Tools and Recommendation Systems
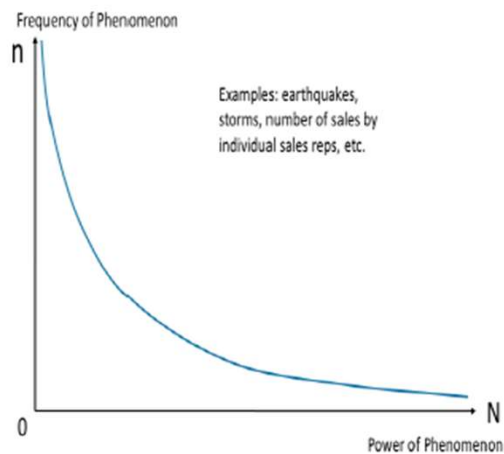
Distribution of Links on Web

- But when people actually measured the distribution of links on the web, they found something very different [1]
- *The fraction of pages with k in-links turned out to be approximately proportional to $1/k^2$*
- Why is this so different from the normal distribution?
- *$1/k^2$* decreases much more slowly as k increases, so pages with a very large number of in-links are much more common than we would expect with a normal distribution.

Power Laws

- Popularity seems to exhibit extreme imbalances, with very large values likely to arise
  - On the web, there are a large number of very popular pages
  - 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 $1/k^3$
- Normal distribution is widespread in a family of settings in the natural sciences
- Power laws seem to dominate in cases where the quantity being measured can be viewed as a type of popularity.

Power Laws

Basic Power Law

Frequency of Phenomenon

n

Examples: earthquakes, storms, number of sales by individual sales reps, etc.

0

N

Power of Phenomenon

© 2016 Alcera Consulting Inc.

https://exploitingchange.com/2016/09/14

- ❑ Power Law: a relative change in the value of one variable leads to the proportional change in the value of other variable
- ❑ Independent of the initial values of both the variable
- ❑ Mathematically,

$$y \propto x^{-b}, b \in \mathbb{R}$$

- ❑ Functions that follows power-law are scale-invariant
- ❑ Pareto Principle (or the 80/20 rule) in Economics: 80% of the outcomes are results of 20% of the causes
- ❑ Power law principle is also coined as Pareto Distribution

- • Scale invariance is a property of objects or laws that remain the same when the scale of length, energy, or other variables are multiplied by a common factor.
- • In other words, the behavior or structure of a system does not change regardless of the scale at which it is observed.
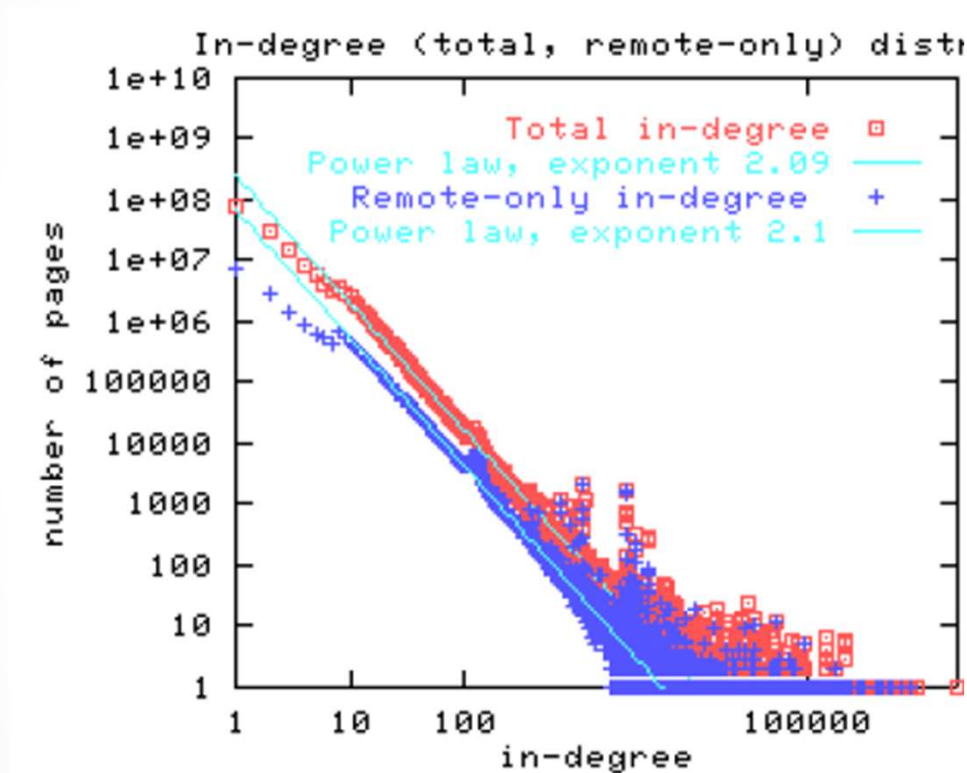
Power-law Distribution
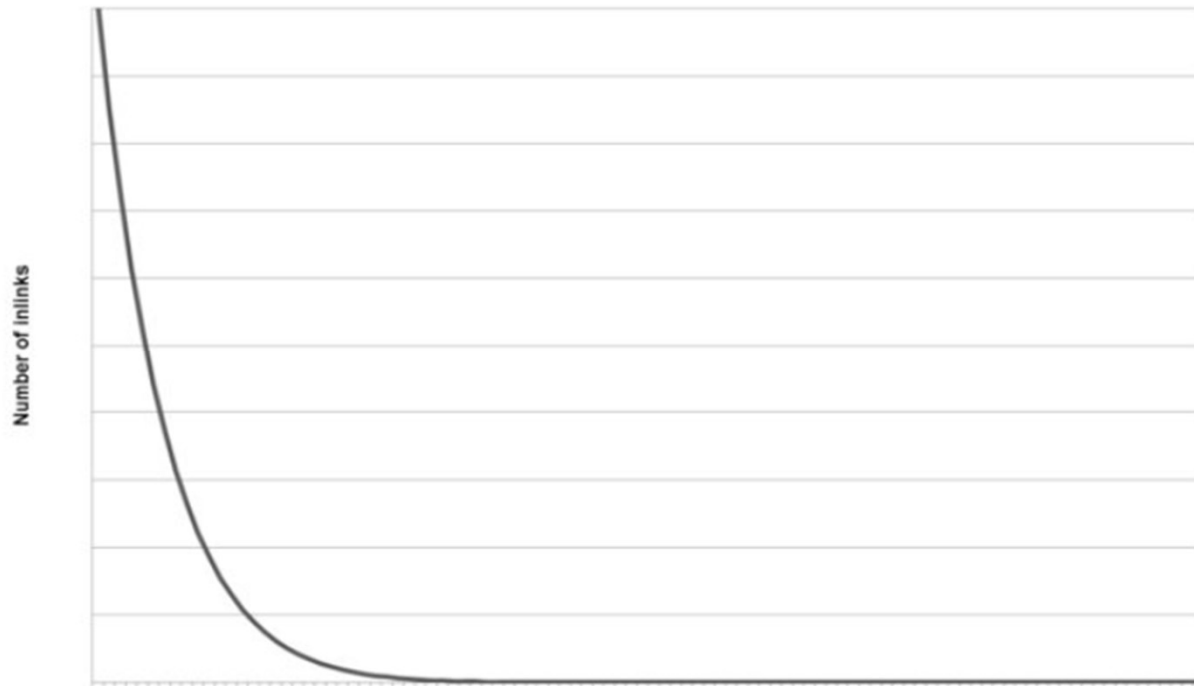
Typical Shape of a Power-law Distribution of Inlinks over Nodes in a Network

Testing for Power Laws

- Let $f(k)$ be the fraction of items that have value k
- Suppose you want to know whether the equation $f(k) = a/k^c$, approximately holds

- $f(k) = a/k^c$
- $\log f(k) = \log a - c \log k$
- $y = +c + (-m).x$

- A power law distribution shows up as a straight line on a log-log plot

Number of Web page in-links

Follows power law distribution [1]

Reason for Power Law Distribution

- If power laws are so widespread, we need a simple explanation
- Just as the Central Limit Theorem gave us a very basic reason to expect the normal distribution, we'd like something comparable for power laws.
- Just as normal distributions arise from many independent random decisions averaging out, …
- Power laws arise from the feedback introduced by correlated decisions across a population

Overview

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

Creation of Web Page Links

- A simple model for the creation of links among Web pages [2]:

1. Pages are created in order, and named 1; 2; 3; : : : ;N.
2. When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule (which is controlled by a single number p between 0 and 1)
   a) With probability p, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to this page i.
   b) With probability 1-p, page j instead chooses a page i uniformly at random from among all earlier pages and creates a link to the page that i points to.
   c) This describes the creation of a single link from page j; one can repeat this process to create multiple, independently generated links from page j (However, to keep things simple, we will suppose that each page creates just one outbound link.)

Copying previous decisions

- Part (2b) of this process is the key: author of page j copies the decision made by the author of page i

- The main result about this mode is that if we run it for many pages, the fraction of pages with k in-links will be distributed approximately according to a power law $1/k^c$

- The value of the exponent c depends on the choice of p [3]

- As p gets smaller, so that copying becomes more frequent, the exponent c gets smaller as well, making one more likely to see extremely popular pages.

Rich getting richer

- The copying mechanism in (2b) is really an implementation of the following "rich-get-richer" dynamics:
  - When you copy the decision of a random earlier page, the probability that you end up linking to some page y is directly proportional to the total number of pages that currently link to y.
- So, we can rewrite step 2(b) as:
  - 2(b) With probability 1 – p, page j chooses a page y with probability proportional to y's current number of in-links, and creates a link to y.
- This is called a "rich-getting-richer" rule because…
  - the probability that page y experiences an increase in popularity is directly proportional to y's current popularity.
  - This phenomenon is also known as preferential attachment [2], in the sense that links are formed "preferentially" to pages that already have high popularity.

Rich getting richer Models

- So, the copying model provides as operational story for why popularity should exhibit such rich-get-richer dynamics.

- The more well-known someone is, the more likely you are to hear their name come up in conversation, and hence the more likely you are to end up knowing about them as well.

- A page's popularity grows at a rate proportional to its current value, and hence exponentially with time.

- A page that gets a small lead over others will therefore tend to extend this lead

- The rich-get-richer nature of copying actually amplifies the effects of large values, making them even larger.

- The Central Limit Theorem states that small independent random values tend to cancel each other out

- However, the rich-get-richer nature of copying actually amplies the effects of large values, making them even larger

Other instances of the model

- Maybe power laws are not as surprising as they may have initially appeared!
- Rich-get-richer models suggest a basis for power laws in a wide array of settings, including some that have nothing at all to do with human decision-making.
  - The populations of cities have been observed to follow a power law distribution: the fraction of cities with population k is roughly $1/k^c$ for some constant c [4].
  - The number of copies of a gene in a genome approximately follows a power-law distribution [5].

Reasons for Power Laws

- These are still simple models designed just to approximate what's going on
- There are other classes of simple models designed to capture power-law behaviour that we have not discussed here.
- What all these simple models suggest is that when one sees a power law in data, the possible reasons why it's there can often be more important than the simple fact that it's there.

Emergence of popularity

- Once an items (book, song) is well-established, the rich-get-richer dynamics of popularity are likely to push it even higher…

- …but getting this rich-get-richer process ignited in the first place seems like a precarious process, full of potential accidents and near-misses.

- The dynamics of popularity suggest that random effects early in the process should play a role

- If history were to be replayed multiple times, it seems likely that there would 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.

Experiment on emergence of popularity

- Salganik, Dodds, and Watts created a music download site, populated with 48 obscure songs of varying quality written by actual performing groups [6]
- Visitors to the site were presented with a list of the songs and given the opportunity to listen to them
- Each visitor was also shown a table listing the current "download count" for each song — the number of times it had been downloaded from the site thus far.
- At the end of a session, the visitor was given the opportunity to download copies of the songs that he or she liked.
- What the visitors did not know is that they were actually being assigned at random to one of eight "parallel" copies of the site.
- The parallel copies started out identically, with the same songs and with each song having a download count of zero.
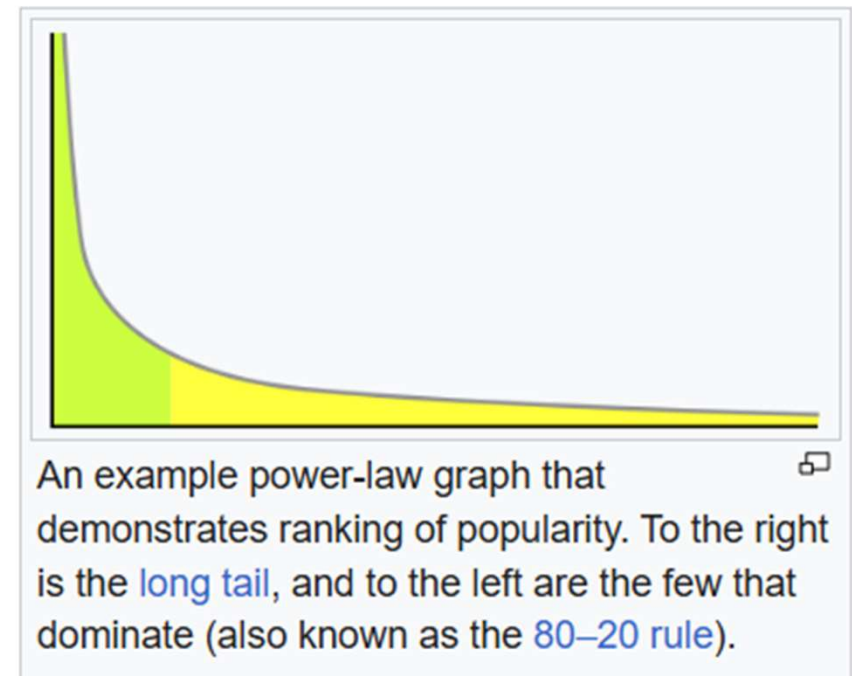
Experiment Results

- Each parallel copy then evolved differently as users arrived.
- This experiment provided a way to observe what happens to the popularities of 48 songs when you get to run history forward eight different times.
- The "market share" of the different songs varied considerably across the different parallel copies.
- Although the best songs never ended up at the bottom and the worst songs never ended up at the top.
- Salganik et al. also used this approach to show that, overall, feedback produced greater inequality in outcomes.
- They assigned some users to a ninth version of the site in which no feedback about download counts was provided at all.
- There was no direct opportunity for users to contribute to rich-get-richer dynamics.
- There was significantly less variation in the market share of different song
- The future success of a book, movie, celebrity, or Web site is strongly influenced by these types of feedback effects, and hence may to some extent be inherently unpredictable.
-

Overview

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

Popularity Distribution – Business Consequences

▪ The distribution of popularity can have important business consequences.

▪ Consider a media company selling books and music.

- Are most sales 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"

An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).
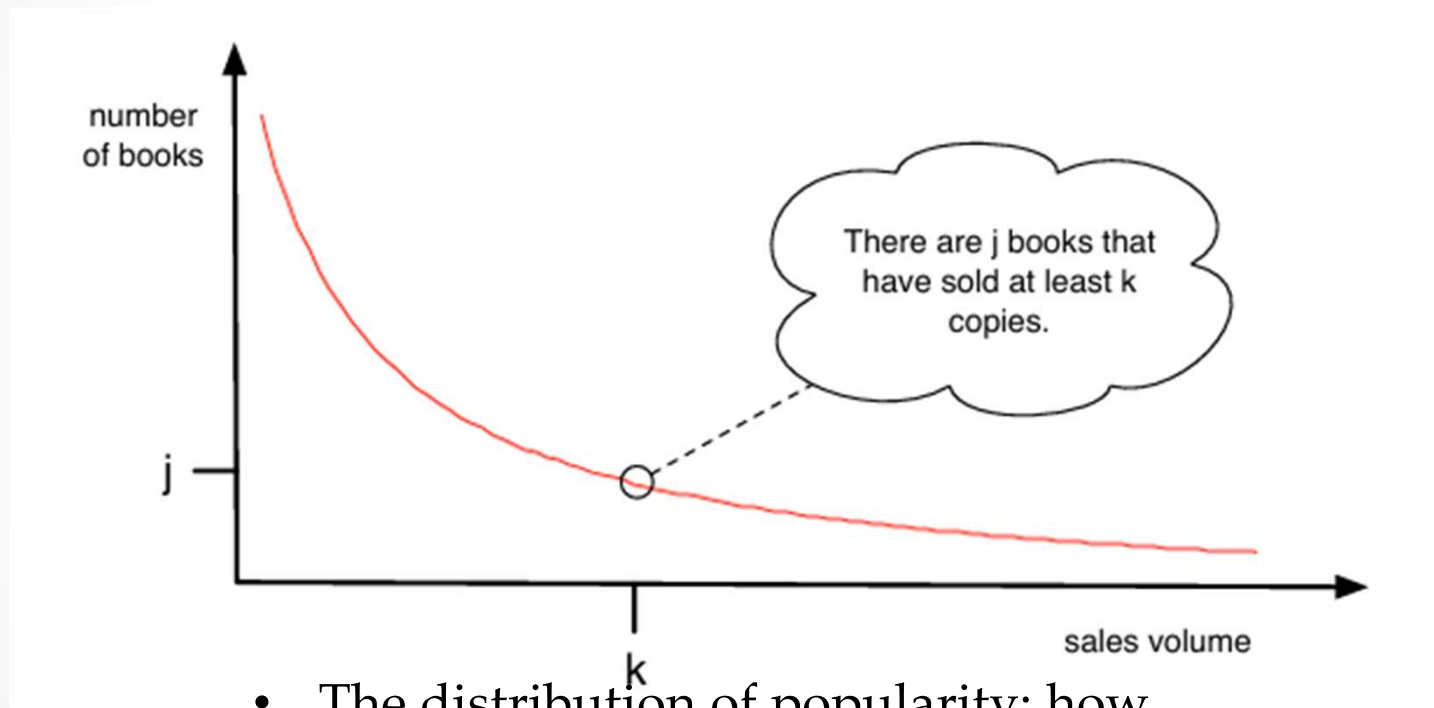
Long Tail

- In a 2004 article, "The Long Tail" [7], Chris Anderson argued that:
  - Internet-based distribution and other factors were driving the media and entertainment industries toward a world oin which the latter alternative would be dominant, with a "long tail" of obscure products driving the bulk of audience interest.
- This tension between hits and niche products makes for a compelling organizing framework.
- This is the fundamental models for companies like Amazon or Netflix
  - where the ability to carry huge inventories, makes it feasible to sell an astronomical diversity of products o even when very few of them generate much volume on their own.
- Ultimately, quantifying the importance of the **Long Tail** comes down to an analysis of power laws…

Discussion so far...

- Initially, we asked about Popularity.
- We started from a baseline in which we expected to see Gaussian distributions and tight concentration around the average.
- We observed that the number of highly popular items was much higher than this baseline would suggest.
- The observation of the power laws brought us to the "rich-getting-richer" models.
- The distribution of popular items can have important business consequences
- Now, we are looking at "long tails" and asking the opposite of the question that we started with.
- A sort of stereotype of the media business in which only blockbusters matter,…
- However, we're observing that the total sales volume of unpopular items, taken together, is really very significant.

Visualizing the Long Tail

number of books

There are j books that have sold at least k copies.

j

k

sales volume

- The distribution of popularity: how many items have sold at least k copies?
- As we follow the x-axis of the curve to the right, we're essentially asking, "As you look at larger and larger sales volumes, how few books do you find?"

Visualizing the Long Tail
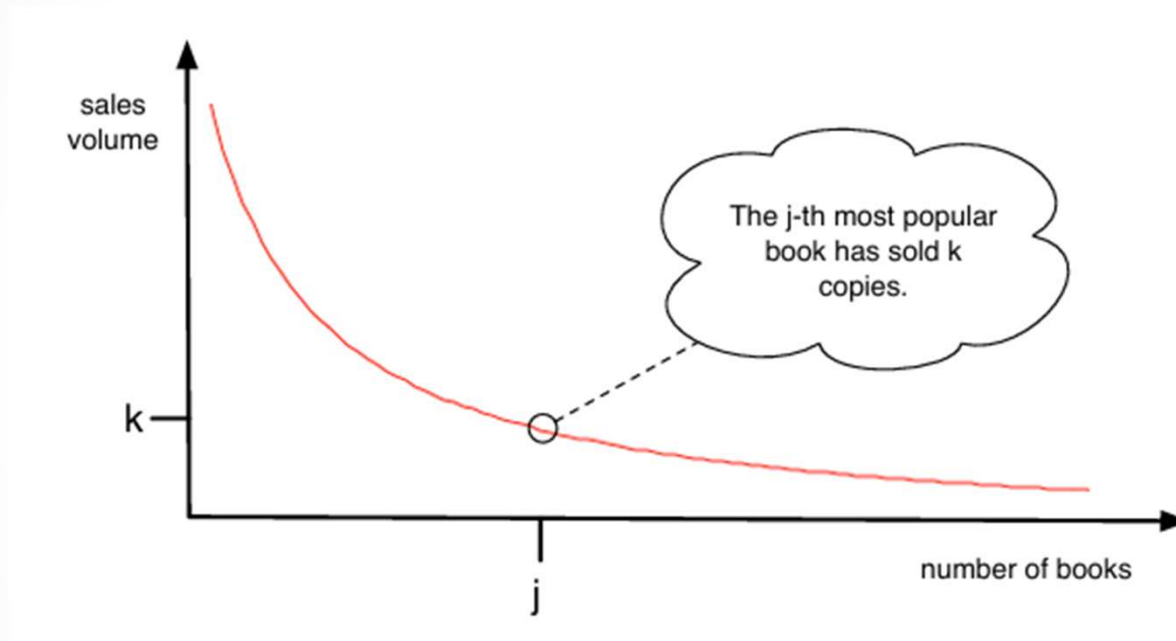
- Flipping the axis
- The distribution of popularity: how many copies of the jth most popular item have been sold?
- "As you look at less and less popular items, what sales volumes do you see?"

Trends in Sales Volume

- Now, we can easily discuss trends in sales volume

- Essentially, the area under the curve from some point j outward is the total volume of sales generated by all items of sales-rank j and higher

- So a concrete version of the hits-vs.-niche question, for a particular set of products, is *whether there is significantly more area under the left part of this curve (hits) or the right (niche products).*

- The debate over trends toward niche products becomes a question of whether this curve is changing shape over time, adding more area under the right at the expense of the left.

Zipf Plots

- The curves of this type, where the axes are ordered so that the variable on the x-axis is rank rather than popularity, are called Zipf plots.

- The linguist George Kingsley Zipf, who produced such curves for a number of human activities [8].

- He identified the empirical principle known as Zipf's Law, that the frequency of the jth most common word in English (or most other widespread human languages) is proportional to 1/j.

Overview

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

Effects of Search Engine

- *Are Internet search tools making the rich-get-richer dynamics of popularity more extreme or less extreme?*

- On one side of this question, we've seen that a model in which people copy links from uniformly random Web pages already gives an advantage to popular pages.

- Search engines such as Google are using popularity measures to rank Web pages, and the highly-ranked pages are in turn the main ones that users see in order to formulate their own decisions about linking.

- A similar argument can be made for other media in which a handful of the most popular items have the potential to crowd out all others.

- In simple models, this kind of feedback can accentuate rich-get-richer dynamics, producing even more inequality in popularity [9]

- However, there are other forces at work [10]
  - 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
  - enabling people to find unpopular items more easily
  - and potentially counteracting the rich-get-richer dynamics.

Effects of Recommendation Engine

- In order to make money from a giant inventory of niche products, a company crucially needs for its customers to be aware of these products, and to have some reasonable way to explore them [7]

- Viewed in this light, the types of recommendation systems that companies like Amazon and Netflix have popularised can be seen as integral to their business strategies.

- They are essentially 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.

- Ultimately, the design of search tools is an example of a kind of higher-order feedback effect

  - by causing people to process their available options in one way or another, we can:

    - reduce rich-get-richer effects
    - or amplify them
    - or potentially steer them in different directions altogether.

Exercises

1. Consider an on-line news site, such as cnn.com or nytimes.com, which consists of a front page with links to many di erent articles. The people who operate such sites generally track the popularity of the various articles that get posted, asking questions like the ones that we've seen in this chapter: "As a function of k, what fraction of all articles have been viewed by k people?" Let's call this the popularity distribution of the articles.

- Now suppose that the operators of such a news site are considering changing the front page, so that next to each link is a counter showing how many people have clicked on the link. (E.g., next to each link it will say something like, "30,480 people have viewed this story," with the number getting updated over time.)

1. What effect do you think this change will have on the behavior of people using the site?

2. Do you expect that adding this feature will cause the popularity distribution of the articles to follow a power-law distribution more closely or less closely, compared to the version of the site before these counters were added?

- Give an explanation for your answer

Answer

- On changing the front page of the online news site by mentioning the number of viewers for each news bulletin, people will find the popularity of the posted news easily so that they can choose to read the news with highest view rate.

- The behaviour of people will change from just randomly viewing the news post to select the appropriate news post with the help of the total number of viewers of a particular post.

- Adding the number of viewers in the front page of online news site increases the popularity of the site among the readers which in turn will lead to the imbalance in the distribution of articles published in the news site.

- This sudden imbalance will raise the popularity value with a rough proportion of $1/k^2$, where, k is the fraction of people who view the news post within certain time limit.

- So, as long as the number of viewers increases there will be an exponentially decaying power-law distribution function.

- The popularity distribution of articles will have close power-law distribution because of the increase in the number of viewers of the article.

References

1. Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Ra-jagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. In Proc. 9th International World Wide Web Conference, pages 309–320, 2000.

2. Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. Science, 286:509–512, 1999.

3. Bela Bollobas and Oliver Riordan. Mathematical results on scale-free random graphs. In Stefan Bornholdt and Hans Georg Schuster, editors, Handbook of Graphs and Networks, pages 1–34. John Wiley & Sons, 2005.

4. Herbert Simon. On a class of skew distribution functions. Biometrika, 42:425–440, 1955.

5. Stanislaw Cebrat, Jan P. Radomski, and Dietrich Stauffer. Genetic paralog analysis and simulations. In International Conference on Computational Science, pages 709– 717, 2004.

6. Matthew Salganik, Peter Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311:854– 856, 2006.

7. Chris Anderson. The long tail. Wired, October 2004

8. George Kingsley Zipf. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Addison Wesley, 1949.

9. Soumen Chakrabarti, Alan M. Frieze, and Juan Vera. The influence of search engines on preferential attachment. In Proc. 16th ACM-SIAM Symposium on Discrete Algorithms, pages 293-300, 2005.

10. S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. Proc. Natl. Acad. Sci. USA, 103(34):12684-12689, 2006.

Reading

1. David Easley and Jon Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World.

   https://www.cs.cornell.edu/home/kleinber/networks-book/

   - Chapter 18.1-18.6