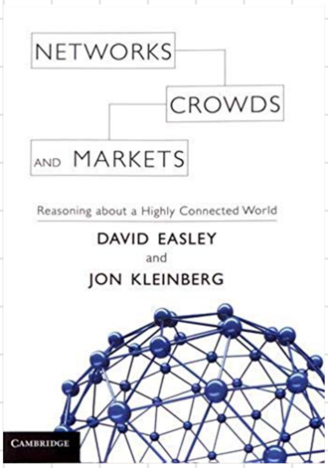
Lecture 14

Network Science

Power Laws and Rich-Get-Richer Phenomena

Today's topics

- 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



Chapter 18

"Power Laws and Rich-Get-Richer Phenomena"

Section 18.1 – 18.6

Popularity as a Network Phenomenon

Network Science $\longleftrightarrow$ popularity

popularity $\longleftrightarrow$ in-links

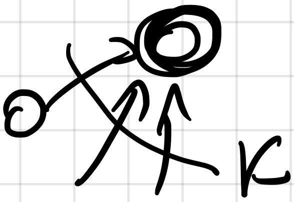
the web is a great domain for analysis

Characterizing popularity reveals imbalances (inequalities)

- almost everyone is popular for very few people
- very few people achieve very high popularity.
- very very few people achieve global popularity!

Why? Is this phenomenon intrinsic to the whole idea of popularity itself?

"As a function of k, what fraction of (web) pages have k in-links?"



larger k → greater popularity

First (and simple) hypothesis: normal distribution!

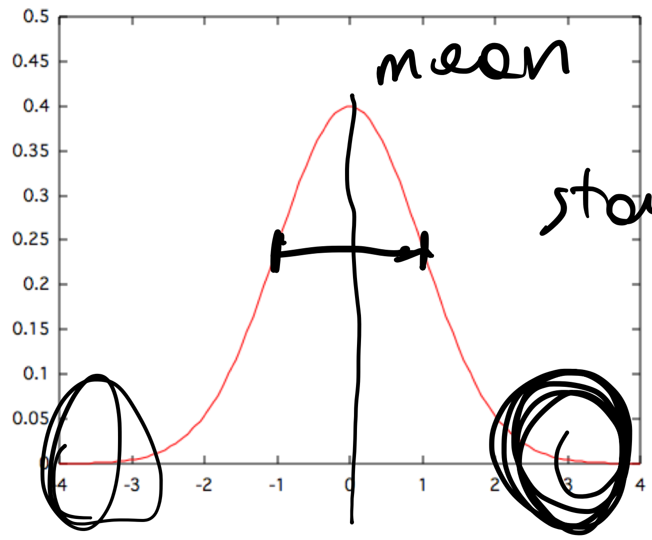


Figure 18.1: The density of values in the normal distribution.

standard deviation

they define a scale

the prob. of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially in c ⇒ big popularity is unlikely.

Central limit theorem

We take small independent random quantities, then in the limit their sum (or average) will be distributed according the normal distribution.

If we "believe" this model, then the number of Web pages with in-links should decrease exponentially in k, as k grows large

watch the video on moodle.

Power Laws

Empiral findings:
the fraction of web pages that have k links

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

$(c = 2.1)$

other networks $(2 < c < 3)$

$$f(k) = a\,k^{-c}$$

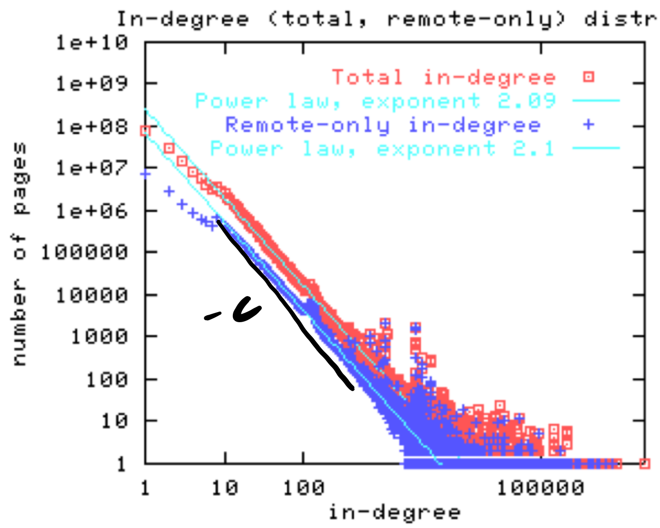
power law distribution

$\frac{1}{k^c}$: decreases much more slowly as k increases.

$\Rightarrow$ pages with very large k are much more common than expected with the normal distr.

$\Rightarrow$ emergence of "hubs" is likely.

this can be observed in many different domains.



$c = 2.1$

Figure 18.2: A power law distribution (such as this one for the number of Web page in-links, from Broder et al. [80]) shows up as a straight line on a log-log plot.

approximations of power laws are very common

$f(k) = e\,k^{-c}$ for some e and c (constants)

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

$\underbrace{\log f(k)}_{y} = \log e - c\,\underbrace{\log k}_{x}$

$y = m - c\,x$

$m \to$ y-intercept

$c \to$ slope

log-log plot

let's accept that power laws represent many phenomena -

<u>Why?</u>

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

we assume that people have the tendency to copy the decisions of people who acted before them.

1) Nodes are created in a sequence
   1, 2, ..., N
2) when J joins the net, then we will create a link
   (a) with prob. p, a link (j, i) is created uniformly at random
   (b) with prob. 1-p, page J chooses a page ℓ with probability proportional to ℓ's current number of "in-links"
   (c) repeated this process
   (Keep the process simple: only one link is created at every step).

"preferential attachment"
Barabasi, Albert 1999

simple model: this does not explain everything.

But it provides a natural explanation for the emergence of hubs.

Do not be surprised to observe power laws or skewed distributions with real data!

(watch 2nd video on moodle)

+ wikipedia notebook
+ net logo simul.

The Unpredictability of Rich-Get-Richer Effects

feedbacks effects ⇒ produce power laws

"Initial fluctuations" : unpredictable

We can predict that a power law can emerge after a while ⇒

⇒ we will have hubs!

BUT: what hubs?

Salganik, Dodds and Watts "musiclab" experiment

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]*

Hit songs, books, and movies are many times more successful than average, suggesting that "the best" alternatives are qualitatively different from "the rest"; yet experts routinely fail to predict which products will succeed. We investigated this paradox experimentally, by creating an artificial "music market" in which 14,341 participants downloaded previously unknown songs either with or without knowledge of previous participants' choices. Increasing the strength of social influence increased both inequality and unpredictability of success. Success was also only partly determined by quality: The best songs rarely did poorly, and the worst rarely did well, but any other result was possible.

10 FEBRUARY 2006 VOL 311 SCIENCE www.sciencemag.org

MusicLab: a site where you could download songs.

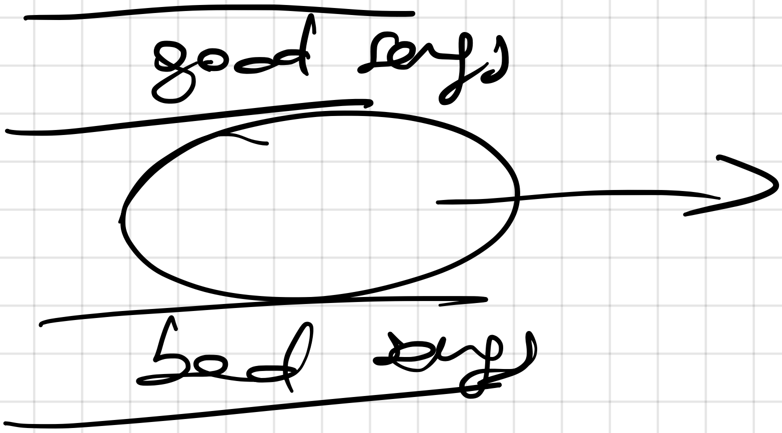
unknown songs from unknown artists; different qualities.

Visitors:
- different sessions
- they could listen "all" the songs
- they could download only "favorite" songs.

"download counts": measure of popularity.

good songs did not end up at the bottom

bad songs did not end up at the top



other songs? they resulted in mixed positions.

in some sessions: the order was established by means of "popularity"

social influence is important at the end of the process.

BUT initial fluctuations are unpredictable

The Long tail

popularity ⟺ power laws (often)

small sets of items → enormously popular

would you bet on "hits" or "niches"

Chris Anderson "the long tail"

Do not focus on hits but try to estimate the market sales of all the "niches"

Stereotype of media business: focus only on "hits"

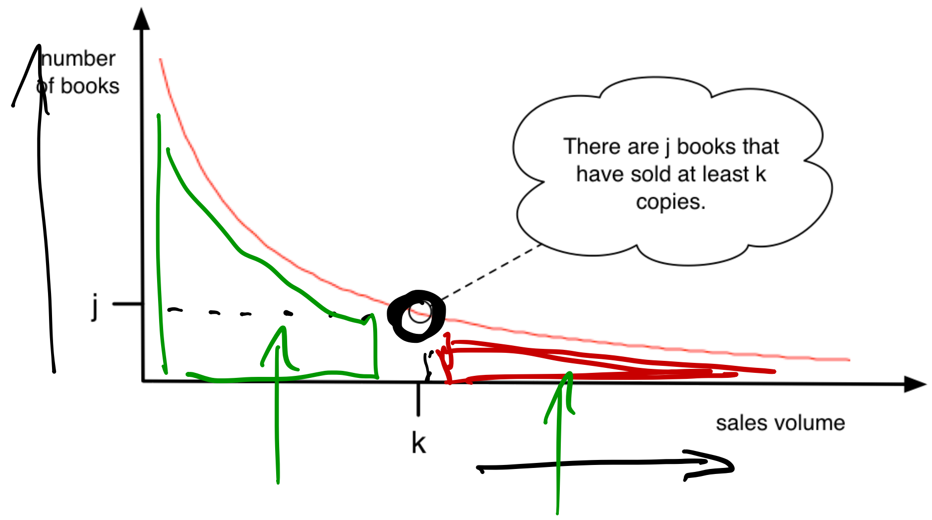


Figure 18.3: The distribution of popularity: how many items have sold at least k copies?

which area is bigger, unpopular vs popular items.

let's switch the axes.

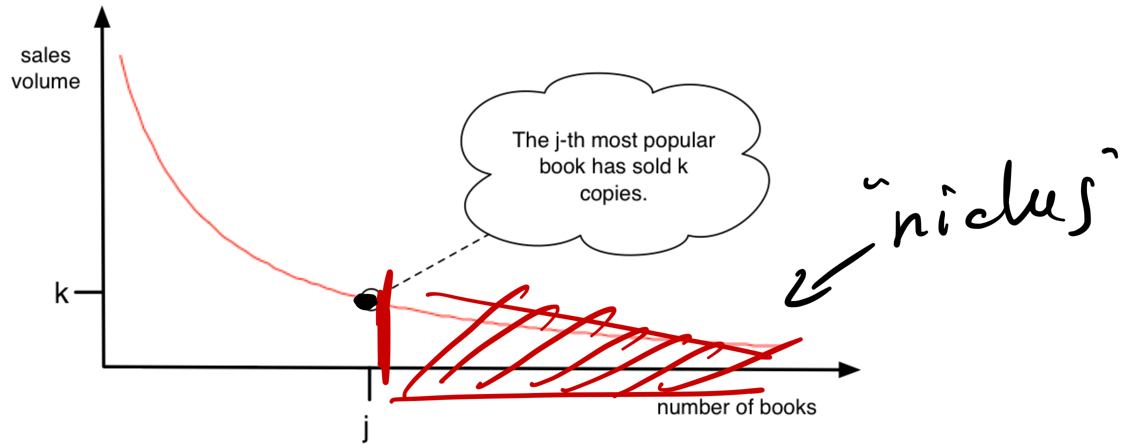


Figure 18.4: The distribution of popularity: how many copies of the j^{th} most popular item have been sold?

focus on the "long tail" of unpopular items

just compare the areas.

pareto distr.
power law dist.
zipf's law

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

Lada A. Adamic

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

Abstract

Many man made and naturally occurring phenomena, including city sizes, incomes, word frequencies, and earthquake magnitudes, are distributed according to a power-law distribution. A power-law implies that small occurrences are extremely common, whereas large instances are extremely rare. This regularity or 'law' is sometimes also referred to as Zipf and sometimes Pareto. To add to the confusion, the laws alternately refer to ranked and unranked distributions. Here we show that all three terms, Zipf, power-law, and Pareto, can refer to the same thing, and how to easily move from the ranked to the unranked distributions and relate their exponents.

The Effect of Search Tools and Recommendation Systems

Search tools make the RGR dynamics more evident

other aspects that make the effect less extreme

1) different queries → different google results

2) targeted and personalized search → unpopular items ranked first

3) Recomm. systems ⇒ "serendipity" → exploit "the big tail argument"

Complex effects in already complex systems.

Take Home Messages

1) popularity $\Longleftrightarrow$ power laws
2) Normal distributions do not explain much
3) RGR models (aka "Preferential attachment) provide some explanations:
   feedback and copying effects
4) though distributions can be predicted ("power law"), we do not know how to predict the success of a single "item"
5) Switch axes: "the long tail", that opens many opportunities for the media industries
6) this knowledge is applied to new systems and dynamics may change again (and again ...)